Hybride semantische Suche – eine Kombination aus Fakten- und Dokumentretrieval

Dipl.-Inform. (FH) Kinga Schumacher

Dissertation zur Erlangung des akademischen Grades

eines Doktors der Naturwissenschaften

(Dr. rer. nat.)

Eingereicht an der Mathematisch-Naturwissenschaftlichen Fakultät der

Universität Potsdam

Gutachter:
Prof. Dr. Christoph Meinel
Prof. Dr. Harald Sack
Prof. Dr. Prof. h.c. Andreas Dengel

Potsdam, den 15.12.2016

Online veröffentlicht auf dem
Publikationsserver der Universität Potsdam:
URN urn:nbn:de:kobv:517-opus4-405973
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-405973

Danksagung

Die vorliegende Dissertation wurde am Deutschen Forschungszentrum für Künstliche Intelligenz an den Standorten Kaiserslautern und Berlin unter der Schirmherrschaft von Prof. Dr. rer. nat. Dr. h.c. mult. Wahlster, Prof. Dr. Prof. h.c. Andreas Dengel und Dr. Norbert Reithinger fertiggestellt. Ihnen sowie den Mitarbeitern der Abteilung Wissensmanagement in Kaiserslautern und Intelligente Benutzerschnittstellen in Berlin möchte ich für die langjährige und überaus harmonische Zusammenarbeit danken. Mir ist bewusst, dass die sowohl von wissenschaftlicher Exzellenz als auch von Herzlichkeit geprägte Atmosphäre am DFKI für meine Entwicklung als Wissenschaftlerin von großer Bedeutung war.

Bei Professor Dr. Meinel möchte ich mich herzlich für seine Zusage mich zu promovieren bedanken. Insbesondere danke ich Professor Dr. Harald Sack für die Betreuung meiner Arbeit, seinen unermüdlichen Einsatz und für seine fortwährende Diskussionsbereitschaft meiner Dissertation.

Mein herzliches Dankeschön geht an Michael Sintek für ausgiebige fachliche Diskussionen und seine stete Hilfsbereitschaft, an Dr. Nicolas Flores-Herr für die kritische Betrachtung der Arbeit und die vielen hilfreichen Ratschläge sowie an Christian Reuschling für die technische Unterstützung. Ebenso möchte ich mich bei Wiebke Wenzel herzlich für das Durchlesen und die Korrektur meiner Arbeit bedanken.

Ein ganz besonderer Dank geht an meinen Mann, Michael Schumacher, an meinen Eltern, Budainé Csizsár Edit und Budai Mihály, sowie an meinen Geschwistern, Budai Márta Hanna und Budai Kornél, die meinen Weg als Wissenschaftlerin immer mit vollem Einsatz unterstützt haben und mir in jederzeit zur Seite standen.

Vorwort

Das Thema der vorliegenden Arbeit ist die semantische Suche im Kontext heutiger Informationsmanagementsysteme. Zu diesen Systemen zählen Intranets, Web 3.0-Anwendungen sowie viele Webportale, die Informationen in heterogenen Formaten und Strukturen beinhalten. Auf diesen befinden sich einerseits Daten in strukturierter Form und andererseits Dokumente, die inhaltlich mit diesen Daten in Beziehung stehen. Diese Dokumente sind jedoch in der Regel nur teilweise strukturiert oder vollständig unstrukturiert. So beschreiben beispielsweise Reiseportale durch strukturierte Daten den Zeitraum, das Reiseziel, den Preis einer Reise und geben in unstrukturierter Form weitere Informationen, wie Beschreibungen zum Hotel, Zielort, Ausflugsziele an.

Der Fokus heutiger semantischer Suchmaschinen liegt auf dem Finden von Wissen entweder in strukturierter Form, auch Faktensuche genannt, oder in semi- bzw. unstrukturierter Form, was üblicherweise als semantische Dokumentensuche bezeichnet wird. Einige wenige Suchmaschinen versuchen die Lücke zwischen diesen beiden Ansätzen zu schließen. Diese durchsuchen zwar gleichzeitig strukturierte sowie unstrukturierte Daten, werten diese jedoch entweder weitgehend voneinander unabhängig aus oder schränken die Suchmöglichkeiten stark ein, indem sie beispielsweise nur bestimmte Fragemuster unterstützen. Hierdurch werden die im System verfügbaren Informationen nicht ausgeschöpft und gleichzeitig unterbunden, dass Zusammenhänge zwischen einzelnen Inhalten der jeweiligen Informationssysteme und sich ergänzende Informationen den Benutzer erreichen.

Um diese Lücke zu schließen, wurde in der vorliegenden Arbeit ein neuer hybrider semantischer Suchansatz entwickelt und untersucht, der strukturierte und semi- bzw. unstrukturierte Inhalte während des gesamten Suchprozesses kombiniert. Durch diesen Ansatz werden nicht nur sowohl Fakten als auch Dokumente gefunden, es werden auch Zusammenhänge, die zwischen den unterschiedlich strukturierten Daten bestehen, in jeder Phase der Suche genutzt und fließen in die Suchergebnisse mit ein. Liegt die Antwort zu einer Suchanfrage nicht vollständig strukturiert, in Form von Fakten, oder unstrukturiert, in Form von Dokumenten vor, so liefert dieser Ansatz eine Kombination der beiden. Die Berücksichtigung von unterschiedlich Inhalten während des gesamten Suchprozesses stellt jedoch besondere Herausforderungen an die Suchmaschine. Diese muss in der Lage sein, Fakten und Dokumente in Abhängigkeit voneinander zu durchsuchen, sie zu kombinieren sowie die unterschiedlich strukturierten Ergebnisse in eine geeignete Rangordnung zu bringen. Weiterhin darf die Komplexität der Daten nicht an die Endnutzer weitergereicht werden. Die Darstellung der Inhalte muss vielmehr sowohl bei der Anfragestellung als auch bei der Darbietung der Ergebnisse verständlich und leicht interpretierbar sein.

Die zentrale Fragestellung der Arbeit ist, ob ein hybrider Ansatz auf einer vorgegebenen Datenbasis die Suchanfragen besser beantworten kann als die semantische Dokumentensuche und die Faktensuche für sich genommen, bzw. als eine Suche die diese Ansätze im Rahmen des Suchprozesses nicht kombiniert. Die durchgeführten Evaluierungen aus System- und aus Benutzersicht zeigen, dass die im Rahmen der Arbeit entwickelte hybride semantische Suchlösung durch die Kombination von strukturierten und unstrukturierten Inhalten im Suchprozess bessere Antworten liefert als die oben genannten Verfahren und somit Vorteile gegenüber bisherigen Ansätzen bietet. Eine Befragung von Benutzern macht deutlich, dass die hybride semantische Suche als verständlich empfunden und für heterogen strukturierte Datenmengen bevorzugt wird.

Abstract

The subject of this doctoral thesis is semantic search in the context of today's information management systems. These systems include intranets and Web 3.0 applications, as well as many web portals that contain information in heterogeneous formats and structures. On the one hand, they contain data in a structured form, and on the other hand they contain documents that are related to this data. However, these documents are usually only partially structured or completely unstructured. For example, travel portals describe the period, the destination, the cost of the travel through structured data, while additional information, such as descriptions of the hotel, destination, excursions, etc. is in unstructured form.

The focus of today's semantic search engines is to find knowledge either in a structured form (also called fact retrieval), or in semi- or un-structured form, which is commonly referred to as semantic document retrieval. Only a few search engines are trying to close the gap between these two approaches. Although they search simultaneously for structured and unstructured data, the results are either analyzed independently, or the search possibilities are highly limited: for example, they might support only specific question patterns. Accordingly, the information available in the system is not exploited, and, simultaneously, the relationships between individual pieces of content in the respective information systems and complementary information cannot reach the user.

In order to close this gap, this thesis develops and evaluates a new hybrid semantic search approach that combines structured and semi- or un-structured content throughout the entire search process. This approach not only finds facts and documents, it uses also relationships that exist between the different items of structured data at every stage of the search, and integrates them into the search results. If the answer to a query is not completely structured (like a fact), or unstructured (like a document), this approach provides a query-specific combination of both. However, consideration of structured as well as semi- or un-structured content by the information system throughout the entire search process poses a special challenge to the search engine. This engine must be able to browse facts and documents independently, to combine them, and to rank the differently structured results in an appropriate order. Furthermore, the complexity of the data should not be apparent to the end user. Rather, the presentation of the contents must be understandable and easy to interpret, both in the query request and the presentation of results.

The central question of this thesis is whether a hybrid approach can answer the queries on a given database better than a semantic document search or fact-finding alone, or any other hybrid search that does not combine these approaches during the search process. The evaluations from the perspective of the system and users show that the hybrid semantic search solution developed in this thesis provides better answers than the methods above by combining structured and unstructured content in the search process, and therefore gives an advantage over previous approaches. A survey of users shows that the hybrid semantic search is perceived as understandable and preferable for heterogeneously structured datasets.

Inhaltsverzeichnis

Kapitel 1
Einleitung - Warum hybride semantische Suche?

1.1 Semantische Suche heute

Semantische Technologien haben das Ziel Dinge präziser, d.h. auf einer höheren Abstraktionsebene als die Lexikalische zu identifizieren und zu beschreiben [Berners-Lee et al., 2001]. Somit wird ein Schritt von der Syntax in Richtung der Bedeutung, der Semantik der Inhalte gemacht. Informationen können effizienter gefunden, genutzt und auch über die Datenquellen hinaus verknüpft werden [Franklin et al., 2005, Beyer, 2011]. Dieses Potenzial wurde erkannt, wie folgende Beispiele zeigen: Die öffentliche Webpräsenz des Rundfunk-Unternehmens BBC basiert auf semantische Technologien[1]. Unternehmen und Organisationen, wie etwa die ING DiBa, IKEA, das Weiße Haus sowie das Bundesministerium für Wirtschaft und Energie (BMWi), nutzen diese Technologien für Informationsmanagement und setzen das semantikbasierte Content Management System Drupal[2] ein. Zahlreiche semantische Suchmaschinen wurden entwickelt. Bekannte Beispiele sind die Rezeptsuchmaschine Yummly[3], die Reisesuchmaschine FACT-Finder Travel[4], die Enterprise-Suchmaschine Exalead CloudView[5] oder die Websuchmaschine Sindice[6]. Auch die Suchmaschine Google setzt auf strukturierte Daten, die mithilfe von semantischen Technologien beschrieben sind[7].

Dennoch ist das Potenzial semantischer Technologien – besonders für Informationsexploration – nicht ausgeschöpft. Dr. Stefan Tweraser, der Country Director von Google Deutschland, Österreich und Schweiz wurde auf dem 6. Nationalen IT-Gipfel des BMWi gefragt, wie er den aktuellen Stand der Suche sieht. Er antwortete, dass wir gerade erst anfangen zu suchen[8]. Dies spiegelt sich in den heutigen *semantischen Suchmaschinen* wider. Während in semantisch angereicherten Systemen die Informationen in unterschiedlich stark strukturierter Form vorliegen, konzentrieren sich semantische Suchmaschinen *entweder* auf *semantische Dokumentsuche oder* auf *Faktensuche*. Semantische Dokumentsuche ist die Suche in un- und semistrukturierten Inhalten (z.B. Texte mit oder ohne HTML-/XML-Markup) mithilfe verfügbarer semantischer Informationen. *Faktensuche* bezeichnet die Suche in strukturierten Inhalten, die formal, mit semantischen Technologien beschrieben sind (RDF, RDFS, OWL). Beide Arten der Suchmaschinen haben Vor- und Nachteile: Strukturierte Daten sind insofern präziser als un- und semistrukturierte Inhalte, als dass strukturierte Daten explizit eine Eigenschaft (Attribut) und dessen Ausprägung (Wert) für ein Konzept angeben. Diese Informationen sind in un- oder semistrukturierten Texten nicht notwendigerweise explizit oder in wenigen aufeinanderfolgenden Sätzen enthalten, was die Informationsfindung erschwert. Demnach

[1] `http://www.w3.org/2001/sw/sweo/public/UseCases/BBC/` (06.01.2016)
[2] `http://drupal.org/` (06.01.2016)
[3] `http://www.yummly.com` (06.01.2016)
[4] `http://www.fact-finder.de/FACT-Finder-Semantic-Travel-Search.html` (06.01.2016)
[5] `http://www.3ds.com/products-services/exalead/` (06.01.2016)
[6] `http://www.sindicetech.com` (06.01.2016)
[7] `http://radar.oreilly.com/2009/05/google-announces-support-for-m.html` (06.01.2016)
[8] 6. Nationaler IT-Gipfel, Podiumsdiskussion im Forum Social Web zum Business Web, 6.12.2011.

kann die Faktensuche präzisere Ergebnisse liefern, sogar Fragen wie z.B. „In welchen Filmen von Garry Marshall spielt Julia Roberts eine Rolle?“ beantworten, wie auch das Beispiel in Abbildung 1.1 zeigt. Die Informationen müssen hierfür aber genügend detailliert, formal, in strukturierter Form vorliegen. Die Strukturierung bedeutet jedoch häufig einen großen Aufwand oder ist wegen der große Menge und Schnelllebigkeit der Daten gar nicht machbar.

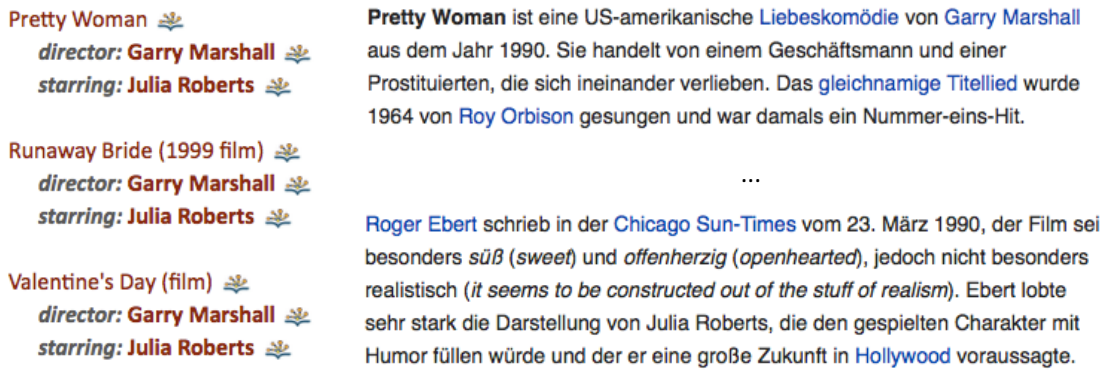


Abbildung 1.1: Fakten (links) und ein Dokument (rechts) zu der Frage „In welchen Filmen von Garry Marshall spielt Julia Roberts eine Rolle?“

Semantisches Dokumentretrieval bietet, da es weniger Modellierungs- und Extraktionsaufwand erfordert, ein breiteres Einsatzspektrum. Die semantischen Metadaten der Dokumente lassen sich mit aktuellen Informationsextraktionsverfahren bis zu einem gewissen Grad automatisiert erkennen. Gesucht wird aber nach unstrukturierten Dokumenten, die Ergebnisse sind somit vager als im Fall von Fakten und die Benutzer müssen in den Ergebnisdokumenten weiter suchen um ihr Informationsbedürfnis zu befriedigen. Dies führt zu der Frage:

> Wenn Informationen divers strukturiert vorliegen, sollten die Suchmaschinen nicht *beides* leisten und Suchanfragen so präzise beantworten, wie die zugrundeliegenden Daten es erlauben?

Diese Arbeit setzt sich mit dieser Frage auseinander und stellt die hybride semantische Suchmaschine mit den Namen SINFIO - *S*uche in den *Inf*orma*i*onsmanagementsystemen von M*o*rgen - vor, die die Lücke zwischen Fakten- und Dokumentsuche in jeder Phase des Suchprozesses schließt und somit eine umfassende Suchlösung in divers strukturierten Datenmengen bietet.

1.2 Fokus und Inhalt dieser Arbeit

Unterschiedlich stark strukturierte Datenmengen kommen in vielen verschiedenen Informationsmanagementsystemen vor. Intranets verfügen z.B. über Mitarbeiter-, Projekt- und Prozesswissen in strukturierter Form, die zugehörigen Dokumente sind jedoch unstrukturiert. Dokumente bezeichnen hierbei nicht notwendigerweise nur Textdokumente, sie können auch multimediale Inhalte, Tabellen usw. beinhalten. Genauso verhält es sich

mit den Daten auf dem persönlichen Computer. Sie sind ebenfalls unterschiedlich strukturiert und werden von verschiedenen Anwendungen verwaltet. Zahlreiche Webportale, z.B. für Reisen, Rezepte und digitale Bibliotheken beinhalten strukturierte Daten zu den textuellen Beschreibungen bzw. Dokumenten. Inhalte in Web 3.0-Anwendungen, die Konzepte des Web 2.0 mit semantischer Technologien kombinieren [Wahlster et al., 2006], sind miteinander verknüpft und stammen aus verschiedenen Datenquellen. Sie können strukturiert (z.B. Datenbank, RDF/S), semistrukturiert (z.B. HTML, XML) oder unstrukturiert (z.B. Text, Multimedia) sein. So enthalten beispielsweise semantische Wikis und Blogs unstrukturierte Text- und Multimediainhalte mit strukturierten semantischen Metadaten. Semantische soziale Netzwerke können strukturierte Angaben über die Personen und ihre Bekanntschaften sowie semistrukturierte oder unstrukturierte Inhalte, wie Statusmeldungen mit oder ohne Orts- und sonstigen auswählbaren Angaben beinhalten.

In solchen Informationsmanagementsystemen können *Informationen also in strukturierter, semistrukturierter und in unstrukturierter Form* vorkommen. Die strukturierten, semantischen Metadaten eines Dokumentes beschreiben nicht notwendigerweise die gesamten unstrukturierten Inhalte des Dokumentes, die zur Befriedigung des Informationsbedürfnisses der Benutzer beitragen können. Ebenfalls können Daten, wie z.B. Personendaten der Mitarbeiter oder Daten, die den finanziellen Rahmen eines Projektes beschreiben, nur in strukturierter Form vorliegen und nicht in unstrukturierten Dokumenten enthalten sein. Semistrukturierte Inhalte bestehen bereits aus unstrukturierten und strukturierten Teilen mit überwiegend unterschiedlichem Informationsgehalt. So geben z.B. die Tags $\langle title \rangle$ und $\langle img \rangle$ in HTML-Dokumenten an, dass es sich um den Titel bzw. ein Bild handelt. Dies führt zur folgenden *These*:

> *Für Datenmengen, die unterschiedlich stark strukturiert sind, aber gemeinsam, mit einer Suchanfrage durchsucht werden sollen, führt eine Kombination von strukturierten und unstrukturierten Inhalten im Suchprozess zu einem besseren Ergebnis, als wenn diese nicht kombiniert werden.*

Dabei stellen sich die Fragen:

- Ist es möglich Suchverfahren zu entwickeln, die sowohl Fakten als auch Dokumente durchsuchen und diese sinnvoll kombinieren können?
- Wie können solche Lösungen aussehen und welche Herausforderungen sind dabei zu lösen?
- Welche Anforderungen stellt solch eine Suchlösung an die Benutzerschnittstelle und wie können diese erfüllt werden?

Zur Präzisierung der These und der Fragen werden zuerst die theoretischen Grundlagen (Kapitel 2) beschrieben. Hierzu gehört die Klärung der Begrifflichkeiten und deren Zusammenhänge rund um das Thema semantische Technologien (Kapitel 2.1 und 2.2), die Definition der semantischen Suche (Kapitel 2.3) und der sogenannten hybriden semantischen Suche (Kapitel 2.4), die sowohl Fakten als auch Dokumente durchsuchen und finden kann. Ebenso wird auf die Grundlagen der Evaluierung von Suchsystemen eingegangen (Kapitel 2.5). Es wird ein Überblick der verwandten Arbeiten gegeben (Kapitel 3), im Einzelnen auf dem Gebiet der semantischen Suche (Kapitel 3.1), der ersten hybriden Suchansätzen (Kapitel 3.2), der Vorgehensweisen zur Gewichtung von Suchergebnissen (Kapitel 3.3) sowie der Evaluierung semantischer Suchsysteme (Kapitel 3.4). Auf Basis

der Grundlagen sowie des Stands der Technik wird die These präzisiert (Kapitel 4.1), eine formale Beschreibung des hybriden Suchproblems angegeben (Kapitel 4.2) und die Forschungsfragen formuliert (Kapitel 4.3). Anschließend wird der Lösungsansatz vorgestellt (Kapitel 5), wobei im ersten Schritt die Anforderungen aus der These und Forschungsfragen abgeleitet (Kapitel 5.1), dann das Lösungskonzept (Kapitel 5.2) und zuletzt die prototypische Realisierung unter den Namen SINFIO (Kapitel 5.3) beschrieben werden. Die Evaluierung umfasst das „Proof of Concept“ mittels eines frühen, einfachen Prototyps auf einer kleinen Datenmenge (Kapitel 6.1) und eine umfangreiche Evaluierung der These und Forschungsfragen sowohl mit Systemkennzahlen als auch aus Benutzersicht (Kapitel 6.2). Die Ergebnisse werden evaluierungsübergreifend diskutiert und die Forschungsfragen beantwortet (Kapitel 7.1). Die Arbeit schließt mit der Gesamtzusammenfassung und Ausblick (Kapitel 8).

1.3 Wissenschaftlicher Beitrag und Veröffentlichungen

Diese Arbeit stellt einen neuartigen Suchansatz vor, der unterschiedlich strukturierte Daten heutiger Informationsmanagementsysteme integriert und diese mit einer Suchanfrage durchsucht. Nach dem aktuellen Stand der Technik konzentrieren sich semantische Suchmaschinen entweder auf Faktensuche oder auf semantische Dokumentsuche. Die wenigen Suchansätze die beides ausführen tun dies entweder weitgehend voneinander unabhängig oder stark eingeschränkt. Demgegenüber durchsucht und kombiniert der in dieser Arbeit vorgestellte Ansatz SINFIO alle zur Verfügung stehenden Daten ohne Einschränkungen. Somit können Benutzeranfragen so präzise beantwortet werden, wie die zugrundeliegende Datenbasis es erlaubt. Dabei leistet die vorliegende Arbeit folgende Beiträge im Bereich der semantischen Suche und der Weiterentwicklung der Suchmaschinen Richtung der hybriden semantischen Suche:

- Die Arbeit gibt einen umfangreichen Überblick über die semantische Suche. Es werden Kategoriesysteme nach den wesentlichen Aspekten, die eine semantische Suchmaschine beschreiben, aufgestellt und beschrieben. Dies beinhaltet die Art der durchsuchten Daten, die Art der Anfragestellung an das System, die zugehörigen Suchansätze mit Schwerpunkt auf den verwendeten Suchalgorithmen und die damit verbundenen möglichen Ausprägungen der Suchmaschinenkomponenten. Da es für die semantische Suche, im Gegensatz zum traditionellen Information Retrieval, keine wohldefinierten, standardisierten Retrievalmodelle mit Rankingverfahren und Evaluierungsmethoden gibt, analysiert die Arbeit die unterschiedlichen Vorgehensweisen gesondert.
- Von bestehenden semantischen Suchlösungen abstrahierend gibt die Arbeit eine Definition der hybriden semantischen Suche an und stellt die formale Beschreibung des hybriden semantischen Suchproblems bereit. Die daraus resultierenden Anforderungen werden identifiziert, analysiert und ein konzeptionelles Modell zur Lösung entwickelt. Besondere Herausforderungen stellen sich dabei an die Anfragestellung und Ergebnisdarstellung sowie an das Ranking. Die wesentlichen Beiträge sind im Einzelnen:
  - Unterstützung der Benutzer bei der Anfragestellung: Durch die semantische Autovervollständigungskomponente können die Benutzer ihre Informationsbedürfnisse in natürlicher Sprache und ohne Kenntnis über die zugrundeliegende Wissensbasis formulieren. An das Suchsystem werden dabei jedoch zum Teil

oder vollständig formale Anfragen gestellt. Die Evaluierung zeigt, dass formal und somit präziser ausgedrückte Informationsbedürfnisse schneller befriedigt werden können als weniger präzise, informale Formulierungen.

  - Analyse der Suchansätze zur Fakten- und semantischen Dokumentsuche: Die Untersuchung der unterschiedlichen Suchansätze bezüglich der definierten Anforderungen gibt Aufschluss über die Vor- und Nachteile der Verfahren. Auf Basis der Analyse können geeignete Kombinationsmöglichkeiten identifiziert werden.
  - Hybride semantische Suchlösung inklusive Rankingstrategie: Eine formale Beschreibung der hybriden semantischen Suchlösung definiert die Vorgehensweise und liefert eine präzise Beschreibung der Rankingstrategie. Der Ansatz verarbeitet formale, informale und hybride Anfragen und beantwortet diese mit Fakten, Dokumenten und hybriden (Dokument mit Fakten) Suchergebnissen. Die erarbeitete Rankingstrategie ist in der Lage, die unterschiedlichen Ergebnistypen in Relation zueinander zu gewichten. Sowohl die ursprüngliche Strategie als auch ihre Erweiterung für große Datenmengen kann eine Rangordnung nahe der Idealen herzustellen.
  - Verständliche Ergebnisdarstellung: Die Benutzerakzeptanz solch einer hybriden Suchmaschine ist stark davon abhängig, wie leicht Suchanfragen formuliert und die Antworten interpretiert werden können. Die Interpretation von Ergebnislisten, die unterschiedliche Typen von Suchergebnissen aufweisen, sowie von hybriden Ergebnissen erfordern eine höhere kognitive Leistung seitens der Benutzer. Die in dieser Arbeit konzipierte Ergebnisdarstellung wurde als verständlich beurteilt und konnte sich auch gegen homogene Ergebnislisten durchsetzen, die nur Fakten oder nur Dokumente beinhalten.

- Gold Standard für hybride semantische Suche: Für die Evaluierung von SINFIO wurde ein Gold Standard erstellt.
- Die hybride semantische Suche verbessert die Retrievaleffektivität und wird von den Benutzern akzeptiert: Die Evaluierung des Suchansatzes SINFIO aus System- und Benutzersicht zeigt eine Verbesserung der Retrievaleffektivität gegenüber einer hybriden Suche ohne Kombination der formalen und informalen Inhalte im Suchprozess sowie gegenüber der Fakten- und der semantischen Dokumentsuche. Die Ergebnisse der Befragung zeigen deutlich, dass SINFIO von den Benutzern akzeptiert und bevorzugt wird.

Veröffentlichungen

Schumacher, K., Sack, H.: *SINFIO - A User-Friendly Hybrid Semantic Search Engine.* Einreichung an der 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017.

Schumacher, K., Forcher, B., und Tran, D. T.: *Semantische Suche.* In Semantische Technologien - Grundlagen. Konzepte. Anwendungen., S. 227–252., Spektrum Verlag, 2011.

Schumacher, K. und Sintek, M.: *Searching web 3.0 content.* In Proceedings of the 7th International Conference on Web Information Systems and Technologies, 2011.

Eichler, K., Hemsen, H., Neumann, G., Reithinger, N., Schmeier, S., Schumacher, K., und Seifert, I.: *DiLiA – the digital library assistant.* In Research and Advanced Technology for Digital Libraries, S. 534–537, 2010.

Grimnes, G. A., Adrian, B., Schwarz, S., Maus, H., Schumacher, K., und Sauermann, L.: *Semantic desktop for the end-user.* In i-com, Special Issue: Nutzerinteraktion im Social Semantic Web, Band 3, S. 25–32, 2009.

Forcher, B., Schumacher, K., Sintek, M., und Roth- Berghofer, T.: *Evaluating the intelligibility of medical ontological terms.* In Proceedings of the 5th Workshop on Knowledge Engineering and Software Engineering, 2009.

Schumacher, K., Sintek, M., und Sauermann, L.: *Combining fact and document retrieval with spreading activation for semantic desktop search.* In Proceedings of the 5th European Semantic Web Conference, volume 5021 of LNCS, S. 569–583, 2008.

Grothkast, A., Adrian, B., Schumacher, K., und Dengel, A.: *OCAS: Ontology-based corpus and annotation scheme.* In Proceedings of the High-level Information Extraction Workshop 2008, S. 25–35, 2008.

Sauermann, L., Kiesel, M., Schumacher, K., und Bernardi, A.: *Semantic desktop.* In Social Semantic Web – Web 2.0 - Was nun?, S. 337–362. Springer Verlag, 2008.

Schumacher, K.: *Four Methods for Supervised Word Sense Disambiguation.* In Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems, S. 317-328, 2007.

Kapitel 2
Theoretische Grundlagen

Kapitel 2 stellt die Grundlagen der semantischen Suche, der hybriden semantischen Suche und der Evaluierung von Suchmaschinen vor.

Zuerst werden die grundlegenden Begriffe im Bereich der formalen Wissensrepräsentation mit semantischen Technologien, wie Semantik und semantische Metadaten, Ontologien, Strukturiertheit und Grad der Formalisierung (Kapitel 2.1) sowie RDFS und RDF (Kapitel 2.2), geklärt. Danach wird die semantische Suche definiert (Kapitel 2.3.1), dem traditionellen Information Retrieval gegenübergestellt (Kapitel 2.3.2), die Kategorien semantischer Suchmaschinen vorgestellt (Kapitel 2.3.3) sowie anhand der konzeptionellen Architektur ein Überblick über die Vorgehensweisen in den verschiedenen Komponenten der Suchmaschine gegeben (Kapitel 2.3.4). Darauf aufbauend kann die Definition der hybriden semantischen Suche abgeleitet werden (Kapitel 2.4).

Evaluierungsmethoden und Kennzahlen für die Evaluierung von Information Retrieval Systemen werden beschrieben (Kapitel 2.5.1) und die Verfahren zur Beurteilung von Rankingverfahren vorgestellt (Kapitel 2.5.2). Im Gegensatz zum traditionellen Information Retrieval existieren noch keine standardisierten Evaluierungsmethoden und Rankingverfahren für semantische und insbesondere nicht für hybride semantische Suchmaschinen. Der Stand der Technik in diesem Bereich wird im Kapitel 3 vorgestellt.

2.1 Semantik, Ontologien, Metadaten, Strukturiertheit und Grad der Formalisierung

Das Semantische Web ist eine Erweiterung des World Wide Webs (WWW) und basiert auf dem Konzept, die Bedeutung der Informationen explizit, in einer maschinenverstehbaren Form formal zu beschreiben [Berners-Lee et al., 2001]. Um Daten maschinenlesbar und maschineninterpretierbar abzubilden, werden formale Beschreibungssprachen eingesetzt. Zu diesem Zweck entwickelte das W3C die standardisierten Repräsentationssprachen des Semantischen Webs: Resource Description Framework (RDF) [Manola et al., 2014], Resource Description Framework Schema (RDFS) [Brickley and Guha, 2004] und Web Ontology Language (OWL) [Group, 2012]. In dem Sinne steht der Begriff „Semantik“ in der Informatik für **formale Semantik** und meint, wie auch in der Linguistik, die formale Beschreibung der Bedeutung von künstlichen und natürlichen Sprachen [Nöth, 1990, Partee, 2011].

Die Standards erlauben uns **Ontologien** zu definieren, die wiederum zur formalen Beschreibung und somit zur maschinellen Interpretation von Inhalten verwendet werden. Ontologien sind nach Gruber definiert als [Gruber, 1993]:

> *„A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly. An ontology is an explicit specification of a conceptualization.“*

Ontologien definieren ein konzeptionelles Modell einer Domäne, das allen Akteuren bekannt ist und stellen ein kontrolliertes Vokabular zur Beschreibung der Konzepte der

Domäne zur Verfügung. Das Vokabular erlaubt eine eindeutige Kommunikation unter den Akteuren. Ontologien definieren üblicherweise Klassen von Objekten, beschreiben deren Eigenschaften sowie Relationen zueinander und definieren weitere Axiome über sie [Gruber, 1993, Hitzler et al., 2008a].

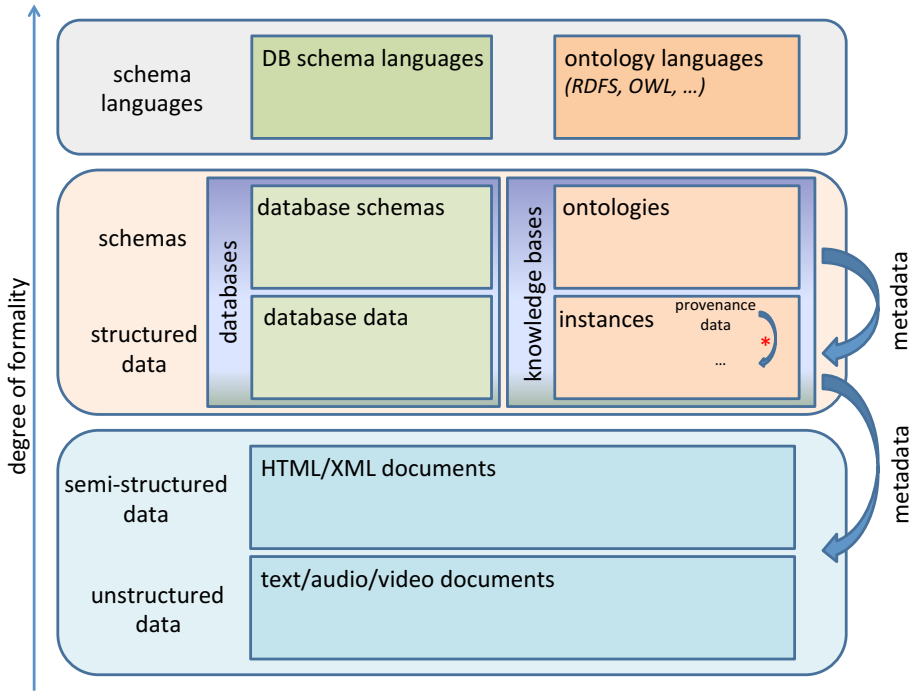


Abbildung 2.1: Übersicht der Zusammenhänge zwischen den Begrifflichkeiten [Bock et al., 2008]

Abbildung 2.1 veranschaulicht den Zusammenhang der verwendeten Begriffe, um die Bedeutung der Strukturiertheit der Daten, semantischen (Meta-)Daten, Ontologien, Instanzen (Individuen) und Schemen zu klären sowie sie anhand des Formalisierungsgrades einzuordnen. Der Begriff **Grad der Formalisierung** ist allgemeiner gefasst als die **Strukturiertheit**. Die Strukturiertheit bezieht sich auf die Daten selber und gibt an, ob sie eine bestimmte Struktur (z.B. als Attribut-Werte-Paare) aufweisen. So sind z.B. Inhalte von Datenbanken und formaler Wissensbasen strukturiert, HTML-/XML-Dokumente semistrukturiert, Textdokumente, Audio und Video unstrukturiert. Der Grad der Formalisierung ist bestimmt dadurch, wie stark die Formalisierung der Struktur durch formale Repräsentationssprachen ist. Werden keine formale Repräsentationssprachen verwendet, wie beispielsweise im Freitext, so sind die Inhalte unstrukturiert und informal.

Wie bei Datenbanken zwischen den Daten selber und dem Datenbankschema unterschieden wird, unterscheidet man in **formalen Wissensbasen** zwischen den Instanzen (Daten) und den Ontologien (Schemas). Für die Beschreibung der Instanzen wird RDF eingesetzt (s. Kapitel 2.2). Um ein Schema selber zu definieren, werden Schema-Sprachen verwendet. Für Ontologien sind es RDFS und OWL[1]. Diese werden Wissensrepräsentations- oder Ontologiesprache genannt. Sie basieren auf formaler Logik und erlauben daher logisches Schlussfolgern, d.h. aus dem bestehenden Wissen neues Wissen abzuleiten. RDFS ist eine weniger ausdrucksstarke Sprache als OWL. Im Ge-

[1] Wobei OWL ebenfalls als RDF kodiert wird.

gensatz zu OWL erlaubt RDFS beispielsweise keine Negation oder Disjunktion [Hitzler et al., 2008a]. OWL basiert auf Prädikatenlogik erster Stufe und erweitert das RDFS um weitere Sprachkonstrukte, die es ermöglichen, Ausdrücke ähnlich der Prädikatenlogik zu formulieren. Da OWL konzeptionell auf Beschreibungslogik (Fragmente der Prädikatenlogik erster Stufe[2]) basiert, werden dort die Ontologie TBox (terminologisches Wissen, d.h. Wissen über die Konzepte der Domäne) und die Instanzen ABox (assertionales Wissen, d.h. Wissen über konkrete Instanzen der Domäne) genannt [Hitzler et al., 2008b].

Wie mit Hilfe dieser Sprachen Daten formal beschrieben und von Maschinen interpretiert werden, wird im folgenden Kapitel beschrieben. An dieser Stelle werden die weiteren für diese Arbeit relevanten **Begriffe rund um die Semantik** erläutert.

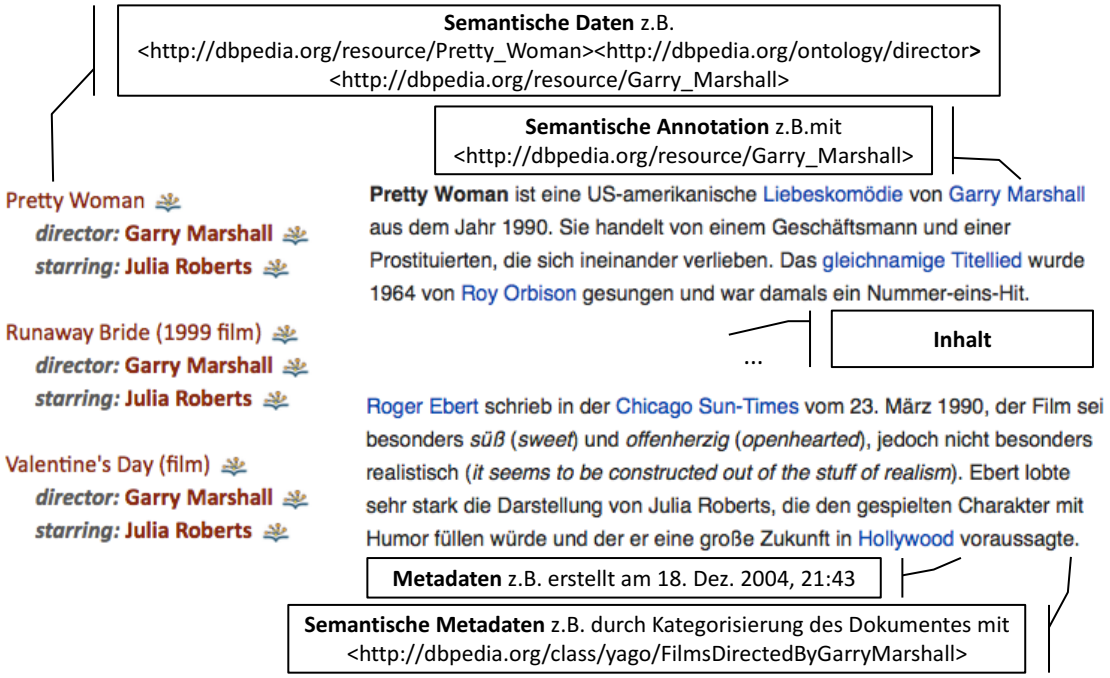


Abbildung 2.2: Begriffsklärung anhand Fakten (links) und eines Dokumentes (rechts) zu der Frage „In welchen Filmen von Garry Marshall spielt Julia Roberts eine Rolle?“

Metadaten sind, allgemein gefasst, Informationen über Merkmale anderer Daten. Sie sind strukturierte Informationen, die bestimmte Merkmale von Daten beschreiben, jedoch nicht die Daten selber beinhalten. Sie werden u. a. verwendet, um Daten zu erklären, zu lokalisieren, ihre Herkunft anzugeben, sie leichter auffindbar und verwaltbar zu machen [Greenberg, 2005]. Unter *semantischen Metadaten* verstehen wir Metadaten, die aus der formalen Wissensbasis (Ontologie oder Instanzbasis) stammen. Sie können beispielsweise Metadaten von Dokumenten, wie die formale Repräsentation des Autors oder eines Ortes sein, der im Dokument erwähnt wird. Es können aber auch Ontologien als Metadaten bezüglich der Instanzen oder Instanzen sogar als Metadaten anderer Instanzen angesehen werden. So sind z.B. die Herkunftsdaten (provenance data) einer Instanz selber als eine Instanz in der Wissensbasis abgebildet [Shadbolt et al., 2006, Carroll et al., 2005].

Unter *semantischen Daten* verstehen wir im Folgenden Daten, die formal beschrie-

[2]Beschreibungslogiken bezeichnen eine Gruppe von Logiken die zur Wissensmodellierung dienen. Sie sind üblicherweise Fragmente der Prädikatenlogik erster Stufe und sind entscheidbar [Hitzler et al., 2008b].

ben sind und die formale Beschreibungssprache durch Ontologien definiert wird. Der Begriff *Inhalt* wird, wenn nicht näher spezifiziert, für die Inhalte der Dokumente ohne die semantischen Metadaten verwendet. Unter *semantischer Annotation* verstehen wir mit dem Dokument verknüpfte semantische Daten, also die semantischen Metadaten des Dokumentes, die aussagen, dass eine bestimmte Passage ein bestimmtes Konzept aus der Wissensbasis repräsentiert [Handschuh and Staab, 2003, Uren et al., 2007]. Abbildung 2.2 veranschaulicht anhand des Beispiels aus Kapitel 1.1 die Begriffe.

2.2 Das Resource Description Framework

Strukturierte Informationen im Semantischen Web werden mit RDF formal beschrieben. Ein **RDF-Dokument** kann als ein gerichteter **Graph** aufgefasst werden, der aus einer Menge von Knoten und gerichteten, benannten (sogenannten qualifizierten) Kanten zwischen Knoten besteht. *Konzepte*, wie Personen, Orte, Projekte oder Veranstaltungen sowie Werte wie Datum, Name usw., bilden dabei die Menge der Knoten. Die Relationen zwischen den Konzepten bilden die Menge der gerichteten Kanten. Ein RDF-Dokument umfasst eine Menge von Tripeln, wobei ein *Tripel* einen *Fakt* repräsentiert[3]. Ein Tripel besteht aus dem Subjekt (Konzept), dem Prädikat (Property, die Eigenschaft) und dem Objekt (der Wert dieser Eigenschaft des Konzeptes) [Klyne et al., 2014]. Abbildung 2.3 zeigt einen einfachen RDF-Graphen, der genau einen Tripel enthält. Das Tripel sagt aus, dass Berlin die Hauptstadt von Deutschland ist. Es gibt verschiedene Syntaxen, wie Notation 3 (N3), N-Triple, Turtle und RDF/XML, um RDF-Graphen zu einfachen Zeichenketten zu serialisieren [Hitzler et al., 2008a]. In dieser Arbeit wird auf Turtle zurückgegriffen.

<http://dbpedia.org/resource/Berlin> <http://dbpedia.org/ontology/capital> <http://dbpedia.org/resource/Germany>

Abbildung 2.3: Einfacher RDF-Graph und der dargestellte Triple in Turtle-Notation

Gemäß RDF 1.1 Concepts and Abstract Syntax [Cyganiak et al., 2014] stehen Internationalized Resource Indentifiers (IRI[4]) und Literale für Dinge in der Welt und repräsentieren eine *Ressource*. Eine Ressource kann alles sein, was im Interesse der Modellierung steht. IRIs repräsentieren z.B. abstrakte Konzepte, Dokumente oder physische Dinge. *Literale* stehen für einen Wert, wie Zeichenkette, Zahl, Datum usw. Sie können typisiert sein, die Angabe des Datentyps[5] (z.B. String, Integer, Date, Boolean) erlaubt das korrekte Parsen der Werte.

Weiterhin erlaubt RDF die Definition von leeren Knoten (Blank-Nodes). Sie sind Knoten ohne IRI, die jedoch auch keine Literale sind. Sie kennzeichnen die Existenz von etwas ohne durch eine IRI ein konkretes Ding zu identifizieren. Tripel mit leeren Knoten sagen aus, dass etwas mit der angegebenen Relation existiert, ohne es explizit zu benennen [Klyne et al., 2014, Brickley and Guha, 2004]. Leere Knoten werden meist als

[3]Fakten sind also formal beschriebene strukturierte Daten. Ein Fakt ist eine Aussage über ein Konzept.
[4]IRIs sind eine Verallgemeinerung von URLs.
[5]Der Datentyp wird mithilfe eines Datentyp-IRIs wie z.B. `http://www.w3.org/2001/XMLSchema#integer` angegeben.

Helferknoten in einer lediglich strukturellen Funktion verwendet [Hitzler et al., 2008a].

RDF-Graphen bzw. Tripel sind syntaktische Aussagen einer Logik. Für ihre **Interpretation** wird auf die sogenannte modelltheoretische Semantik zurückgegriffen. Es wird eine Menge von Interpretationen definiert sowie festgelegt, wann eine Interpretation Modell eines Graphen ist oder anders ausgedrückt die Interpretation den Graphen erfüllt [Hitzler et al., 2008a]. Bei der Definition der modelltheoretischen Semantik für RDFS wird schrittweise vorgegangen: zuerst wird eine einfache Interpretation definiert, die dann durch Angabe weiterer Kriterien zur RDF-Interpretation und schließlich zur RDFS-Interpretation führt. Wie Abbildung 2.4 zeigt, sind RDFS-Interpretationen gültige RDF-Interpretationen, die wiederum gültige einfache Interpretationen sind [Hitzler et al., 2008a].

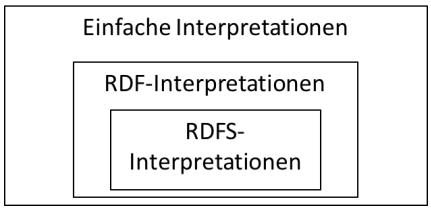


Abbildung 2.4: Beziehung der einfachen, RDF- und RDFS-Interpretationen [Hitzler et al., 2008a]

Der Ausgangspunkt von *(einfachen) Interpretationen* ist das Vokabular V, bestehend aus der Menge der IRIs und Literalen. Betrachtet man ein RDF-Tripel, so ist das Subjekt eine IRI oder ein leerer Knoten. Als Objekt können IRIs, leere Knoten oder Literale auftreten. Die Prädikate sind ebenfalls IRIs, die jedoch für eine Property, also eine binäre Relation stehen [Klyne et al., 2014][6]. Die leeren Knoten vorerst außer Acht gelassen teilt die modelltheoretische Semantik das Vokabular V in die Menge der Ressourcen (IRIs und Literale) und der Properties (Ressourcen, die in den Tripeln als Prädikat stehen) auf. Aus der Sicht des RDF-Graphen bilden die Ressourcen die Menge der Knoten und die Properties die Menge der Kanten. Die zwei Mengen sind nicht notwendigerweise disjunkt, da auch über Properties Aussagen getroffen werden können. Es werden Funktionen definiert, die:

- alle Ressourcen in diese zwei Mengen aufteilen;
- für jede Property diejenigen Ressourcen aus IR zuordnen, die durch diese Property verbunden sind. Diese Funktion bildet also ab, welche Ressourcen über welche Property verbunden sind.

Basierend auf diesen Mengen und Funktionen weist die Interpretationsfunktion I jedem Tripel $\langle s, p, o\rangle$ einen Wahrheitswert zu, wobei ein Tripel dann wahr ist, wenn alle s, p, o im Vokabular V enthalten sind und der Property p mithilfe der ersten Funktion s und o zugeordnet wurden. Ein Graph ist dann wahr, wenn jedes Tripel darin wahr ist. Um leere Knoten zu integrieren, wurde diese einfache Interpretation erweitert: ein leerer Knoten ist dann „gültig“, wenn es für ihn eine Ressource gibt, mit der er identifiziert werden kann [Hitzler et al., 2008a].

[6] Dies spiegelt eine vereinfachte Sicht wieder, die jedoch für diese Arbeit ausreichend ist. Es ist möglich auch über Literale Aussagen zu treffen (z.B., dass die Zahl 4 ein Integer ist). IRIs können somit auch für Literale stehen und solche Literale als Subjekt in einem Tripel vorkommen.

Die *RDF-Interpretationen* erweitern diese einfache Interpretation durch weitere Anforderungen, die sicherstellen, dass die oben genannten Regeln zur Tripelbildung, genauer gesagt was Subjekt, Prädikat und Objekt sein kann, eingehalten werden. Weiterhin werden axiomatische Tripel definiert, die als wahr ausgewertet werden müssen. Sie dienen beispielsweise dazu Ressourcen mit speziellen IRIs, wie die Ressource $\langle rdf : type \rangle$, als Property zu kennzeichnen (rdf steht für den Namensraum `http://www.w3.org/1999/02/22-rdf-syntax-ns#`).

Die *RDFS-Interpretationen* führen u. a. neue Klassenbezeichner, wie z.B. die Klasse Ressource, Literal und Datentyp sowie Properties, die es erlauben Klassen- und Propertyhierarchien zu definieren, ein. Diese hierarchischen Relationen, $\langle rdfs : subClassOf \rangle$ und $\langle rdfs : subPropertyOf \rangle$, sind transitiv. Weitere Klassen und Properties dienen der Abbildung von Kollektionen, Kommentaren usw. RDFS ermöglicht es, die Domäne (domain, also Subjekt) und den Wertbereich (range, also Objekt) einer Property durch die Angabe von Klassenbezeichnern zu charakterisieren. Diese beiden Eigenschaften sind ebenfalls Teil der RDFS-Interpretationen. Eine Reihe von axiomatischen Tripeln geben weitere Anforderungen an, wie z.B. dass die Domäne von $\langle rdf : type \rangle$ $\langle rdfs : Resource \rangle$ ist, oder dass die Domäne von $\langle rdfs : domain \rangle$ und $\langle rdfs : range \rangle$ $\langle rdf : Property \rangle$ ist[7].

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
PREFIX dbo: <http://dbpedia.org/ontology#>.
PREFIX dbr: <http://dbpedia.org/resource/>.

SELECT ?film WHERE {
?film rdf:type dbo:Film.
?film dbo:director dbr:Garry_Marshall.
?film dbo:starring dbr:Julia_Roberts.}
```

Abbildung 2.5: SPARQL-Abfrage nach den Filmen von Garry Marshall, in denen Julia Roberts die Hauptdarstellerin ist

Die standardisierte Anfragesprache für RDF ist die *S*PARQL *P*rotocol *A*nd *R*DF *Q*uery *L*anguage, kurz **SPARQL**. Sie basiert, vereinfachend gesagt, auf sogenannten Tripel-Pattern. Tripel-Pattern sind Tripel, in denen an beliebiger Stelle Variablen stehen können. Gesucht wird nach denjenigen Tripeln in einer gegebenen Menge von Tripeln, die zu dem Muster passen. Aus den Ergebnissen ergeben sich die möglichen Belegungen der Variablen [Aranda et al., 2013]. Abbildung 2.5 zeigt eine Beispielabfrage. Das Schlüsselwort „PREFIX“ ermöglicht es Kürzel für die verwendeten Namensräume, in dem Fall für den RDF-Namensraum und für einen eigens definierten Namensraum zu definieren. Hinter „SELECT“ stehen die Variablen. Im „WHERE“-Block befinden sich die Tripel-Pattern. Die Pattern drücken aus, welche Regeln die gesuchten Ressourcen erfüllen sollen: „?film“ soll eine Instanz der Klasse $\langle dbo : Film \rangle$ sein und es sollen Tripel geben, die Aussagen, dass $\langle dbr : Garry\_Marshall \rangle$ der $\langle dbo : director \rangle$ und $\langle dbr : Julia\_Roberts \rangle$ die Hauptdarstellerin, $\langle dbo : starring \rangle$, des Films „?film“ ist. Neben einfachen Tripel-Pattern erlaubt es SPARQL auch, komplexere Graphmuster wie Gruppen, Alternativen und Optionen zu erzeugen. Zudem bietet SPARQL Filter, Zähler, arithmetische Operationen, Vergleichsoperatoren, boolesche Operatoren und spezielle Prüffunktionen (z.B. ob etwas

[7] Eine detaillierte, formale Beschreibung der Interpretationen ist beispielsweise in [Hitzler et al., 2008a] zu finden.

ein leerer Knoten oder ein IRI ist) an, sowie die Möglichkeit mit Datenwerten zu suchen. Neben den Operatoren stehen auch Modifikatoren zur Verfügung, die zur Kontrolle der Form und Größe der Ergebnisliste dienen. So kann beispielsweise die Ergebnisliste anhand einer Eigenschaft sortiert, die maximale Anzahl der Ergebnisse angegeben oder redundante Ergebnisse entfernt werden.

2.3 Semantische Suche

Semantische Suche baut auf Verfahren des Information Retrievals (IR) auf und erweitert diese um formale Semantiken mit dem Ziel, die Suche zu verbessern. Eine allgemein anerkannte Definition für semantische Suche liegt noch nicht vor. Kapitel 2.3.1 diskutiert bestehende Definitionen und leitet, basierend auf den Ergebnissen und den bisher vorgestellten Grundlagen, eine Definition zur semantischen Suche her. Kapitel 2.3.2 vergleicht die semantische Suche mit dem traditionellen Information Retrieval und zeigt die Verbesserungen auf. Kategorien semantischer Suchmaschinen, basierend auf einer Analyse des Stands der Technik, werden im Kapitel 2.3.3 vorgestellt. Dabei zeigt sich, dass die verschiedenen Kategorien unterschiedliche Anforderungen an die Ausprägung der einzelnen Komponenten der Suchmaschine stellen. Die entsprechenden Lösungen und Methoden werden auf Basis der konzeptionellen Architektur von semantischen Suchmaschinen im Kapitel 2.3.4 beschrieben.

Teile dieses Kapitels, insbesondere 2.3.2, 2.3.3 und 2.3.4, stützen sich auf [Schumacher et al., 2011].

2.3.1 Definition

Es gibt viele Definitionen zur semantischen Suche. Beispielsweise beschrieb Dr. Riza C. Berkan, Entwickler der semantischen Suchmaschine Hakia und Gründer des gleichnamigen Unternehmens, die semantische Suche in [Berkan, 2007] wie folgt:

> *„...semantic search ought to be a system which understands both the user's query and the Web text using cognitive algorithms similar to that of the human brain, then brings results that are dead on target (right context) at first glance (not requiring to open the Web page for further investigation).“*

Diese Beschreibung schränkt die semantische Suche auf den Einsatz von kognitiven Algorithmen und Textsuche im Web ein und beinhaltet den Wunsch, die Antwort auf die Anfrage auf den ersten Blick erkennen zu können.

Simons et al. schreiben in Zusammenhang mit dem Semantischen Web und RDF über Intelligente Suche, die es erlaubt Dokumente auf der Ebene der Bedeutung durchzusuchen anstatt auf der syntaktischen Ebene [Simons et al., 2004]:

> *„...the ability to query documents based on their semantics, rather than on strings of characters that may occur in them or on their document syntax.“*

Duc Tran Thanh definiert [Thanh, 2011] semantische Suche, bezogen auf das Semantische Web als eine Suche, die die Daten im Semantischen Web verwendet:

> *„We use the term Semantic Web Search to refer to search solutions, which make use of the increasing amount of data on the Semantic Web.“*

Guha et al. sehen semantische Suche als Suchanwendung im Semantischen Web, die auf

Technologien des Information Retrievals basiert und die Daten im Semantischen Web ausnutzt [Guha et al., 2003]:

> *„Semantic search is an application of the Semantic Web to search. ... Semantic Search attempts to augment and improve traditional search results (based on Information Retrieval technology) by using data from the Semantic Web.“*

Die beiden letztgenannten Definitionen beschränken sich auf die Suche im Semantischen Web. Semantische Suche kann jedoch auch in anderen Umgebungen eingesetzt werden, wie z.B. in einer digitalen Bibliothek oder um einen Desktoprechner zu durchsuchen. Mehr auf den Suchprozess selber bezogen ist die Definition von Mangold [Mangold, 2007]:

> *„We define semantic search to be a document retrieval process that exploits domain knowledge.“*

Diese Definition beinhaltet jedoch auch nur semantisches Dokumentretrieval, aber keine Faktensuche. Eine verallgemeinerte Definition ist nach [Hildebrand et al., 2007]:

> *„We use the term Semantic Search when semantics are used during any part of the phases in the search process.“*

Da sie allgemein genug ist, um alle Arten semantischer Suchmaschinen abzudecken, wird an dieser Stelle diese Definition übernommen und lediglich - auf Basis der Präzisierung des Begriffes Semantik im Sinne der Informatik (s. Seite 7) - mit dem Wort „formal“ ergänzt. Die **semantische Suche** wird wie folgt **definiert** [Schumacher et al., 2011]:

> **Die semantische Suche ist ein Suchprozess, in dem in einer beliebigen Phase der Suche formale Semantik verwendet wird.**

Die folgenden Kapitel beschreiben welche Verbesserungen die Verwendung von formalen Semantiken gegenüber dem traditionellen Information Retrieval bewirken kann (Kapitel 2.3.2), wie sich semantische Suchmaschinen kategorisieren lassen (Kapitel 2.3.3) und auf welcher Art und Weise in den Phasen des Suchprozesses formale Semantiken eingesetzt werden (Kapitel 2.3.4).

2.3.2 Traditionelles Information Retrieval versus semantische Suche

Der Benutzer interagiert mit dem Suchsystem, um ein Problem zu lösen. Er verfügt über ein implizites mentales Modell des Problems, aus dem er die zur Lösung benötigten Informationen ableitet. In diesem Prozess macht er sein mentales Modell explizit, um die Anforderungen benennen zu können und drückt diese in der Form einer oder mehreren Suchanfragen aus, die das Suchsystem verarbeiten kann [Mizarro, 1997]. Die Suchmaschine sucht nach Inhalten, die zur Beantwortung der Suchanfrage beitragen.

Abbildung 2.6 verdeutlicht diesen Prozess am Beispiel der Dokumentsuche und vergleicht die traditionelle Schlüsselwortsuche mit der semantischen Suche. Mit traditioneller Schlüsselwortsuche ist hier der weit verbreitet eingesetzte **„Bag of Words“ Information Retrieval-Ansatz** gemeint, bei dem die Dokumente als eine ungeordnete Menge von Termen aufgefasst und in einer für die Suche optimierten abstrakten Repräsentationsform, im sogenannten (Voll-)Textindex abgebildet werden (vgl. [Ferber, 2003b, Salton et al., 1975]). Für die *Indexierung* werden die vorkommenden Terme (Token) aus der Dokumentmenge extrahiert (Tokenisierung), wobei diejenigen Terme die Menge der Indexterme bilden sollen, mit denen die Bedeutung der Dokumente beschrieben werden kann. Inhaltlich unwichtige Terme, die sogenannten Stoppwörter (z.B. Füllwörter, Arti-

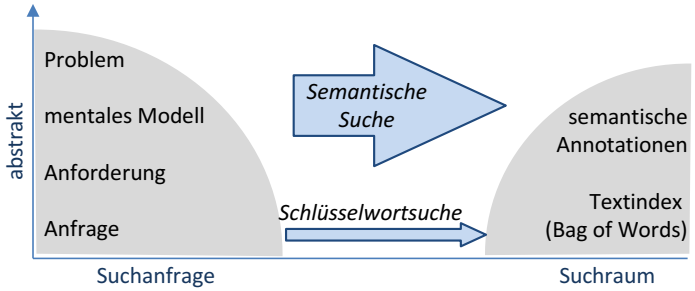


Abbildung 2.6: Schlüsselwortsuche *vs.* semantische Suche [Schumacher et al., 2011]

kel, Konjunktionen und Präpositionen in der deutschen Sprache) werden hierbei nicht berücksichtigt. Die verbliebenen Terme werden mit linguistischen Methoden, wie Normalisierung, Stemming und Lemmatisierung, vorverarbeitet. Das Ziel der Normalisierung ist es, verschiedene Schreibweisen desselben Termes auf Zeichenebene zu identifizieren, wie z.B. Groß-/Kleinschreibung, Akzente, zusammengesetzte Wörter mit oder ohne Bindestrich usw. Stemming und Lemmatisierung wird durchgeführt, um Deklinationen und Konjugationen von Termen zu entfernen und sie auf ihren Wortstamm zu bringen, wobei Nomen in Singular und Verben in ihre Präsensform gebracht werden. So können die unterschiedlichen Formen eines Wortes als das gleiche erkannt und als ein Term gewichtet werden. Dies führt oft zu Informationsgewinn, kann aber auch den Verlust relevanter Informationen verursachen [Manning et al., 2008a].

Nach dem Grundprinzip des Information Retrievals werden Dokumente und Anfragen in eine geeignete Repräsentationsform gebracht und miteinander verglichen. Konzepte hierzu beschreiben **Retrievalmodelle**. Die bekanntesten sind *das Vektorraummodell, das Boolsche Modell, das probabilistische Modell* und *das Kleinberg Modell* [Manning et al., 2008c, Ferber, 2003b][8]. Das Vektorraummodell lässt sich neben Information Retrieval auch für Dokumentklassifikation einsetzen, ist deshalb weit verbreitet [Manning et al., 2008d] und wird auch im vorgestelltem Lösungsansatz verwendet.

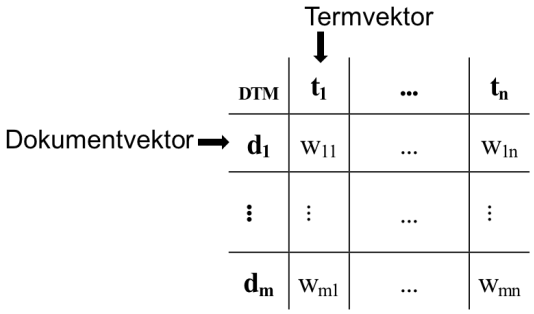


Abbildung 2.7: Dokument-Term-Matrix

Das *Vektorraummodell* (vector space model) basiert auf einem festen Vokabular, den sogenannten Indextermen, und auf einer endlichen Menge von Dokumenten. Das Modell definiert für jedes Dokument d_i aus der Menge der Dokumente $D = \{d_1, ..., d_m\}$ und für jeden Term t_j aus der Menge der Indexterme $T = \{t_1, ..., t_n\}$ ein Gewicht $w_{ij} \in \mathbb{R}$. Die Gewichte repräsentieren die Signifikanz der Terme bezüglich der Dokumente. Die Gewichte

[8] Die meisten Modelle schließen sich gegenseitig nicht aus, sie lassen sich kombinieren.

des Dokuments d_i lassen sich zu einem Vektor $d_i = (w_{i1}, ..., w_{in}) \in \mathbb{R}^n$ zusammenfassen. Dieser Vektor beschreibt das Dokument im Vektorraummodell und wird Dokumentvektor genannt. Analog bildet jede Spalte einen Termvektor [Ferber, 2003a] und die Dokument- und Termvektoren zusammen eine Dokument-Term-Matrix (s. Abbildung 2.7).

Es gibt mehrere Arten von Gewichtungsmethoden:

- Globale Methoden setzen dabei die Terme in den Vordergrund. So ist z.B. die Dokumenthäufigkeit (document frequency) $df(t_j)$, die Anzahl der Dokumente, in denen t_j vorkommt, eine globale Gewichtungsmethode.
- Lokale Methoden setzen die Dokumente in den Vordergrund. Die Termhäufigkeit (term frequency) $tf(d_i, t_j)$, die Anzahl Vorkommen von t_j in d_i, ist beispielsweise eine lokale Gewichtungsmethode.

Am häufigsten wird *tf-idf* (term frequency - inverse document frequency), mit:

$$tfidf(d_i, t_j) = tf(d_i, t_j)/df(t_j) \tag{2.1}$$

eingesetzt. Sie ist eine Kombination lokaler und globaler Gewichtungsmethoden, da sie die Terme mit großer Häufigkeit abhängig von der Dokumenthäufigkeit abschwächt, wodurch dann für jedes Dokument die unwichtigen Terme schwach und die wichtigen Terme stark gewichtet werden [Ferber, 2003a].

Die Anfragen werden entsprechend des Vorkommens der Indexterme ebenfalls durch Vektoren $q = (w_1, ..., w_n) \in \mathbb{R}^n$ repräsentiert. Wurden bei der Indexierung linguistische Werkzeuge eingesetzt, so werden dieselben Methoden auf die Anfrageterme angewendet. Für den Vergleich von Anfrage- und Dokumentvektoren werden Ähnlichkeitsfunktionen eingesetzt, $S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ weist jedem Paar von Dokument-Anfrage-Vektoren $d, q \in \mathbb{R}$ einen reellen Ähnlichkeitswert $s(d, q) \in \mathbb{R}$ zu. Als Ähnlichkeitsfunktion kann z.B. das Cosinus-Maß eingesetzt werden [Ferber, 2003a], [van Rijsbergen, 1979]:

$$cosinus\_similarity(d, q) = \frac{(d \cdot q)}{|| \, d \, || \cdot || \, q \, ||} \tag{2.2}$$

Es sei noch angemerkt, dass „Dokumente“ hierbei nicht notwendigerweise nur Textdokumente sind, es können auch Multimediadaten mit einbezogen werden. Für die Suche (mit textuellen Anfragen) werden jedoch meist nur textuelle Informationen herangezogen, wie die textuellen Metadaten von Bildern und Videos [Sack, 2010].

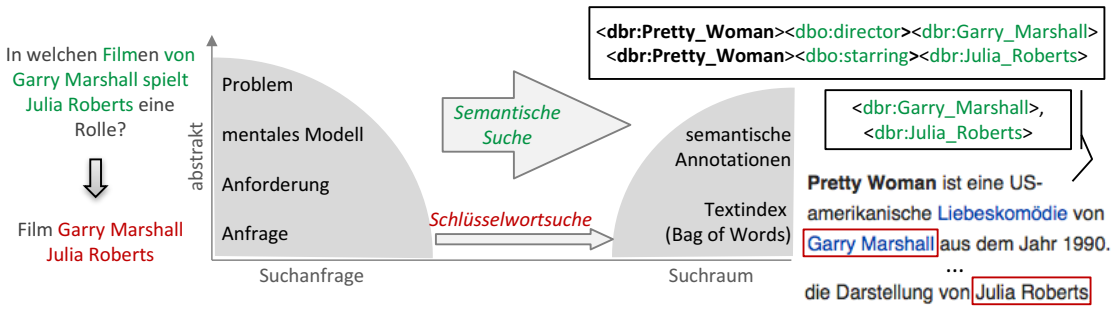


Abbildung 2.8: Beispiel zur Schlüsselwortsuche *vs.* semantische Suche

Der **Unterschied zwischen der traditionellen Schlüsselwortsuche und der semantischen Suche** liegt darin, dass die traditionelle Schlüsselwortsuche auf der syntak-

tischen Ebene, die semantische Suche auf der abstrakteren semantischen Ebene, d.h. auf der Ebene der Konzepte agiert.

Die *traditionelle Schlüsselwortsuche* mit dem „Bag of Words"-Retrieval Ansatz basiert auf dem syntaktischen Vergleich von Zeichenketten (pattern matching, Methoden hierzu werden im Kapitel 2.3.2, Seite 27, vorgestellt). Sie prüft das Vorkommen der Suchwörter (Schlüsselwörter) in den Indextermen und liefert die Dokumente zurück, die diese beinhalten. Dabei werden auch in der Suchanfrage die Stoppwörter entfernt und Stemming durchgeführt. Verschiedene Formen der Wörter werden also erkannt. Bei dem syntaktischen Vergleich verursachen Synonyme, Abkürzungen und Akronyme Probleme, da sie bei dem Abgleich mit dem Suchtermen nicht gefunden werden. Weiterhin können Homonyme nicht disambiguiert werden, wodurch die Ergebnismenge Dokumente zu allen Bedeutungen solch eines mehrdeutigen Wortes beinhaltet. Um dem entgegenzuwirken, werden Synonymlisten (also zusätzliches Wissen), die auch Abkürzungen und Akronyme enthalten, eingesetzt. Zur Disambiguierung werden in der Regel auf Statistik und Wahrscheinlichkeitsrechnung basierende Lernverfahren eingesetzt (z.B. [Schütze, 1998, Karov and Edelman, 1998, Purandare and Pedersen, 2004, Schumacher, 2007]), um verschiedene Bedeutungen zu identifizieren und die Dokumente entsprechend zuzuordnen. Automatische Verfahren, die ohne zusätzliches Wissen auskommen, disambiguieren jedoch nur etwa zu 60% richtig [Stokoe et al., 2003]. Websuchmaschinen verwenden üblicherweise das Wissen aus der persönlichen Historie, Lokalisierung[9] sowie geografische Informationen, um an erster Stelle Ergebnisse zu jenem Konzept zu liefern, das der Benutzer am wahrscheinlichsten sucht[10].

Die *semantische Suche* nutzt formal beschriebenes Wissen für die Suche, um dem mentalen Modell des Benutzers näher zu kommen, daher die Qualität der Suche zu verbessern [Schumacher et al., 2011]. Faktensuchmaschinen suchen in formalen Wissensbasen. Semantische Dokumentsuchmaschinen durchsuchen neben dem Inhalt auch die semantischen Annotationen bzw. semantischen Metadaten der Dokumente. Diese stellen maschineninterpretierbare Informationen zur Bedeutung von Termen, Ausdrücken in dem Dokument bzw. Wissen über das Dokument zur Verfügung[11] (vgl. Kapitel 2.1). Die semantische Suche führt in dem ersten Schritt ebenfalls einen syntaktischen Abgleich durch [Giunchiglia and Shvaiko, 2003], um passende Konzepte in der Wissensbasis und/oder in den damit verknüpften Dokumenten zu finden[12]. Die gefundenen Konzepte sind nicht mehr nur Zeichenketten, sie repräsentieren bestimmte Dinge: die IRI `http://dbpedia.org/resource/Pretty_Woman` repräsentiert beispielsweise den Film „Pretty Woman" und ist nicht nur eine Zeichenkette mit der Bedeutung „Pretty Woman" (vgl. Kapitel 2.2). Im nächsten Schritt wird ein sogenannter semantischer Abgleich durch-

[9] Anpassung der Software und Inhalte an die regionalen Gegebenheiten wie Sprache, kulturelle Unterschiede und technische Anforderungen [Hudson and Hall, 1997].

[10] Websuchmaschinen greifen jedoch zunehmend auf formal beschriebene, strukturierte Daten zurück (vgl. Kapitel 8.2). Google verwendet solches Wissen zur Disambiguierung, was beispielsweise bei der Eingabe von „George Bush" gut zu erkennen ist: bereits zum Zeitpunkt der Anfragestellung werden George Bush Jr. und George Bush Sr. zur Auswahl angeboten (`http://ebiquity.umbc.edu/blogger/2012/09/23/entity-disambiguation-in-google-auto-complete/`, 11.06.2016).

[11] Manche Faktensuchmaschinen nutzen auch die Inhalte der Dokumente, um Fakten zu finden. Semantische Dokumentsuchmaschinen können wiederum nur die formale Wissensbasis durchsuchen und die damit annotierte Dokumente als Ergebnis liefern. Beispiele für solche Suchmaschinen sind in der Stand der Technik-Übersicht im Kapitel 3.1 zu finden.

[12] Eine Ausnahme bilden die Suchmaschinen mit formalen Anfragen, wie im nächsten Kapitel beschrieben wird.

geführt, der entlang von Relationen das Wissen in der formalen Wissensbasis berücksichtigt [Giunchiglia and Shvaiko, 2003] und die „semantische“ Ähnlichkeit nicht mehr auf der Ebene der Syntax, sondern auf Basis solcher semantischen Relationen bestimmt. Die Vorgehensweisen hierfür sind vielfältig und werden im Kapitel 2.3.4.4 vorgestellt.

Das Beispiel in Abbildung 2.8 demonstriert den Unterschied zwischen der traditionellen Schlüsselwortsuche und der semantischen Suche. Gesucht wurde in Wikipedia und in DBpedia. Die formale Wissensbasis DBpedia wurde aus Wikipedia extrahiert und enthielt vorerst die Daten aus den Infoboxen der Wikiseiten. Sie wurde sukzessive erweitert, z.B. durch das Extrahieren der Kategorieinformationen und der Tabelleninhalte in den Wikipediaseiten [Auer et al., 2007]. Im Rahmen der Schlüsselwortsuche werden Dokumente aus der Wikipedia gefunden, die die Zeichenketten „Garry Marshall“ und/oder „Julia Roberts“ beinhalten. Es spielt jedoch keine Rolle, in welchem Zusammenhang diese beiden Namen in dem Dokument vorkommen. Die semantische Suche hingegen identifiziert durch die Zeichenketten beide Personen sowie die Konzepte ‘Regisseur (Direktor) eines Filmes zu sein’ bzw. ‘in einem Film eine Rolle zu spielen’[13] in DBpedia. Solche Konzepte sind eindeutig, und die Synonyme, Abkürzungen und Akronyme werden üblicherweise über die Wissensbasis abgedeckt [Schumacher et al., 2011].

2.3.3 Kategorien semantischer Suchmaschinen

Semantische Suchmaschinen lassen sich nach diversen Aspekten kategorisieren, am wesentlichsten sind die **Kategorien anhand der Art der zugrundeliegenden Daten**, **der Art der Anfragestellung** und damit verbunden anhand **vom Schwerpunkt des Suchansatzes**.

Anhand der **zugrundliegenden Daten** lassen sich semantisches Dokumentretrieval und Faktensuche unterscheiden [Hildebrand et al., 2007, Mäkelä, 2005]:

- *Semantisches Dokumentretrieval* erweitert die traditionelle Schlüsselwortsuche um semantische Technologien, verwendet also formal beschriebene strukturierte Informationen, um eine bessere Suche in unstrukturierten Inhalten anbieten zu können.
- *Faktensuche* agiert hingegen auf formalen Wissensbasen und liefert Konzepte[14] und Fakten (strukturiert) als Ergebnis.

Verfahren zum semantischen Dokumentretrieval setzen Algorithmen ein, die die Struktur des Graphen der Wissensbasis nutzen, um die Anfrageterme zu generalisieren oder zu spezifizieren, vordefinierte Kategorien zu bestimmen oder verwandte Inhalte zu finden [Hildebrand et al., 2007, Mäkelä, 2005, Sack, 2005]. Ansätze zur Faktensuche nutzen ebenfalls die Struktur der Wissensbasis. Sie werden durch die verfügbaren semantischen Relationen geleitet und führen eine strukturelle Interpretation der Anfrage durch. Ein Beispiel wären Algorithmen, die den kürzesten Pfad zwischen zwei Instanzen suchen [Hildebrand et al., 2007, Schumacher et al., 2008]. Die Suchansätze werden im Abschnitt 2.3.4.4 vorgestellt.

Lei et al. nahmen die Benutzerschnittstellen der Suchmaschinen als Basis und betrachteten, welche **Unterstützung der Benutzer bei der Anfragestellung** bekommt [Lei

[13] Der Übersichtlichkeit halber beinhaltet die Abbildung nicht die vollständige Menge an Tripeln. Es fehlt beispielsweise die Angabe, dass $< dbr : Pretty\_Woman >$ ein Film ist.

[14] Die Suche nach Konzepten wird in der Literatur auch Entity oder Data Retrieval genannt [Blanco et al., 2013, Strasunskas and Tomassen, 2010].

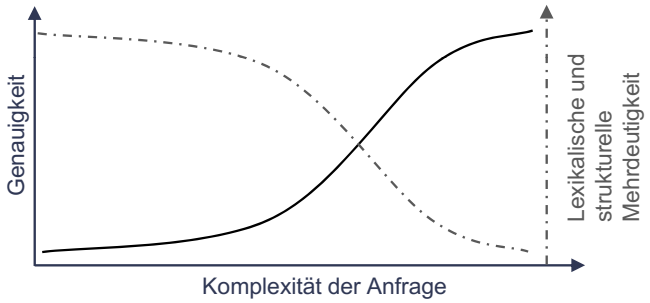


Abbildung 2.9: Genauigkeit der Suche *vs.* Komplexität der Anfrage bedingt durch die lexikalische und strukturelle Mehrdeutigkeit (aus [Schumacher et al., 2011])

et al., 2006]. Im Folgenden wird dieser Ansatz verfolgt, weil die Anfrageformulierung einen starken Einfluss auf den Schwerpunkt der Suchverfahren hat. Dies liegt daran, dass die Komplexität bzw. Präzision der Anfrage mit der Mehrdeutigkeit und daher indirekt auch mit der Komplexität des Suchansatzes in Wechselwirkung steht: Je *präziser die Anfrage* formuliert ist, umso weniger tritt das Problem der *lexikalischen und strukturellen Mehrdeutigkeit* auf, die Genauigkeit der Suchmaschine steigt [Hildebrand et al., 2007] (siehe Abbildung 2.9). Der Suchalgorithmus selber kann im Fall präziserer Anfragen einfacher gestaltet werden, ohne dass die Güte des Verfahrens, gemessen an der Genauigkeit und der Vollständigkeit der Suchergebnisse (Precision/Recall, mehr dazu im Kapitel 2.5.1), darunter leidet. Unter der lexikalischen Mehrdeutigkeit wird die Mehrdeutigkeit von Wörtern (Homonyme) verstanden, während die strukturelle Mehrdeutigkeit durch die Struktur eines Satzes bedingt ist. So ist z.B. im Satz „Anne beobachtet den Mann mit dem Fernglas“ nicht klar, ob Anne durch das Fernglas sieht oder der Mann es hält [Hildebrand et al., 2007]. Eine präzisere Anfrage bedeutet jedoch eine aufwändigere und kompliziertere Konstruktion durch den Benutzer, wodurch die kognitive Last bei der Suche steigt.

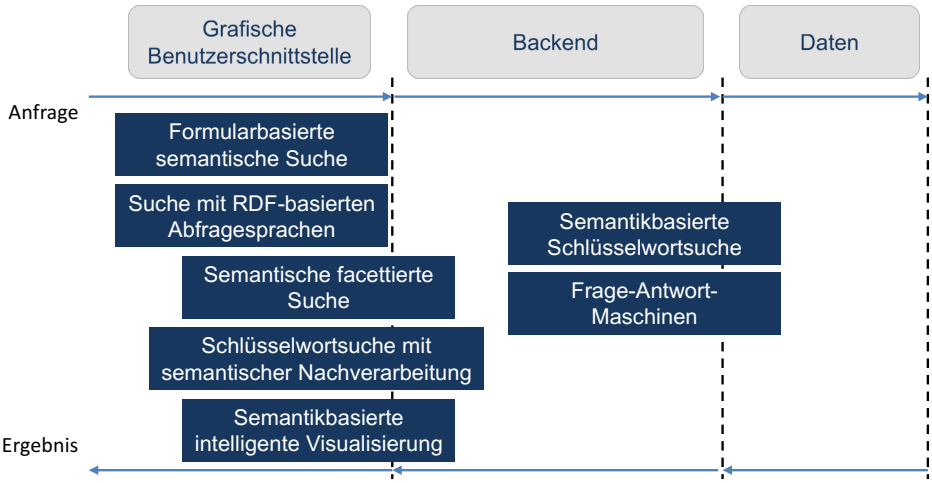


Abbildung 2.10: Kategorien Semantischer Suchmaschinen [Schumacher et al., 2011]

Wird eine formale Anfragegestellt, so muss diese Anfrage lediglich in die jeweilige Abfragesprache umgeformt werden. Ist die Anfrage nicht formal, so muss die Suchmaschine

sie mithilfe der zur Verfügung stehenden semantischen Informationen interpretieren, bevor diese in eine formale Abfrage übersetzt werden kann[15] (falls das Ergebnis nicht bereits im Rahmen des Interpretationsprozesses gefunden wurde, weil z.B. eine Instanz gesucht wurde). Folglich kann von der **Art der Anfragestellung** auf den **Schwerpunkt des Suchansatzes** geschlossen werden. Dabei lassen sich sieben Kategorien der semantischen Suche identifizieren, wobei die Schwerpunkte in der Anfragestellung (GUI), im Suchalgorithmus (Backend) und in der Verarbeitung der Ergebnisse und deren Darstellung (Backend, GUI) liegen.

Lei und Kollegen haben in [Lei et al., 2006] die folgenden vier Kategorien beschrieben: formularbasierte Suche, Suchmaschinen mit RDF-basierten Abfragesprachen, semantikbasierte Schlüsselwortsuche und Frage-Antwort-Maschinen. Es lassen sich jedoch noch drei weitere Kategorien identifizieren, die andere Ansätze verfolgen. Diese sind: semantische facettierte Suche, Suchmaschinen mit semantischer Nachverarbeitung und Suchmaschinen mit semantikbasierter intelligenter Visualisierung. Abbildung 2.10 skizziert die Kategorien und veranschaulicht ihre Zuordnung.

Im Folgenden werden die *sieben Kategorien* vorgestellt.

1. *Formularbasierte semantische Suchmaschinen* bieten Formulare an, um die Suchanfrage als Attribut-Werte-Paare basierend auf den verfügbaren Eigenschaften zusammenzustellen. Die meisten Suchmaschinen dieser Art erlauben auch die zusätzliche Eingabe von Schlüsselwörtern. Aus diesen Angaben wird die formale Abfrage konstruiert und ausgeführt. Es werden keine komplexen Suchalgorithmen benötigt. Ein Beispiel für formularbasierte Suchmaschinen ist SHOE [Heflin and Hendler, 2000]. Abbildung zeigt, wie in SHOE Werte für ausgewählte Attribute angegeben (links) und Kategorien ausgewählt (rechts) werden können 2.11 .

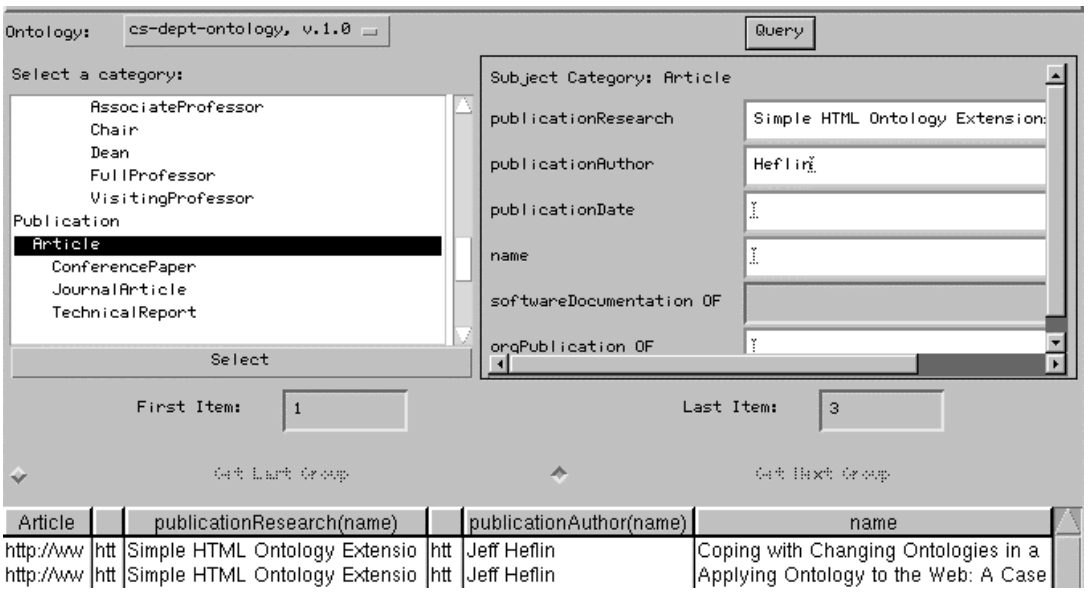


Abbildung 2.11: Formularbasierte semantische Suche mit SHOE [Schumacher et al., 2011]

[15] Viele Suchmaschinen verschmelzen diese zwei Schritte und leiten aus der Suchanfrage im Rahmen der semantischen Interpretation mehrere formale Abfragen ab. Die Suchansätze werden im Abschnitt 2.3.4.4 vorgestellt.

2. *Suchmaschinen mit RDF-basierten Anfragesprachen* bedienen sich einer formalen Beschreibung der Suchanfrage, wobei die Anfragesprache auf RDF basiert. Die Suchmaschine CORESE beispielsweise definiert eine Anfragesprache, in dem lediglich die SPARQL-Schlüsselwörter und Klammern weggelassen werden (vgl. Kapitel 2.2). Abbildung 2.12 zeigt als Beispiel die Suchanfrage nach Organisationen, die mit Humanwissenschaften zu tun haben und ihre Umsetzung als SPARQL-Abfrage [Corby et al., 2004]. Die Schlüsselwörter SELECT und WHERE, die SPARQL-spezifische Punktsetzung und Klammern fallen in der Anfragesprache weg. Diese Informationen lassen sich aus der Anfrage automatisch bestimmen, da Variablen durch ein Fragezeichen gekennzeichnet sind ($SELECT$?org ?rel ?$topic$), Punkte einen Tripel abschließen und die Klammern den Ausdruck hinter WHERE umfassen. Weiterhin ist das Prädikat (?rel) immer von Typ Property. Diese Angabe kann also auch bei der Übersetzung in eine SPARQL-Abfrage automatisch hinzugefügt werden. Das Beispiel macht jedoch deutlich, dass die Benutzer für solche Anfragesprachen das Vokabular der verwendeten Ontologien kennen müssen. Um dieses Problem abzuschwächen, werden häufig unterstützende grafische Benutzerschnittstellen angeboten, die z.B. die Klassen der Instanzen und passende Relationen auflisten.

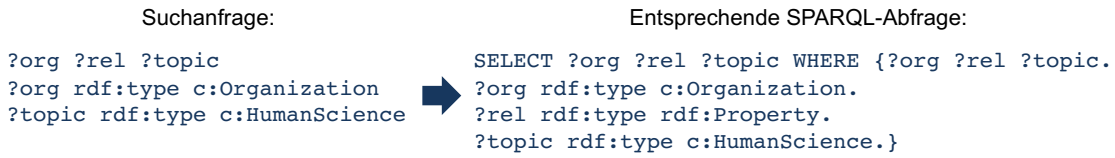


Abbildung 2.12: Beispiel für Suchanfrage und der zugehörigen SPARQL-Abfrage in CORESE [Schumacher et al., 2011]

3. *Die semantische facettierte Suche* verknüpft Suchen und Browsen miteinander, indem der Benutzer neben der Eingabe einer Suchanfrage die Ergebnismenge durch Auswahl von sogenannten Facetten filtern kann. Facetten sind typischerweise die semantischen Metadaten der Dokumente bzw. das Wissen, welches durch Textanalyse der Inhalte gewonnen wurde (z.B. Personen, Orte, Schlüsselwörter) [Hyvönen and Mäkelä, 2006]. Der Benutzer schränkt also durch Auswahl von Facetten Schritt für Schritt den Suchraum ein, wobei er nach jedem Schritt einen Überblick über die verfügbaren Eigenschaften bekommt. Der Suchalgorithmus implementiert also im wesentlichen eine Schlüsselwortsuche und eine Filterfunktion. Die semantische facettierte Suche kann als eine Weiterentwicklung der formularbasierten Suche angesehen werden. Der Unterschied liegt in der Dynamik durch die kontinuierliche Verfügbarkeit der Facetten und Klick-und-Filter-Funktion. Beispiele sind Piggy Bank, die im Rahmen des Projektes SIMILE[16] entwickelt wurde [Huynh et al., 2005], DynaQ [Reuschling et al., 2010] und die Wikipedia Faceted Search[17] [Hahn et al., 2010]. Abbildung 2.13 zeigt die Facetten für „nature science“ in Wikipedia Faceted Search.

[16] `http://simile.mit.edu/` (12.10.2011)
[17] `http://www.semantic-search.info` (11.06.2016)

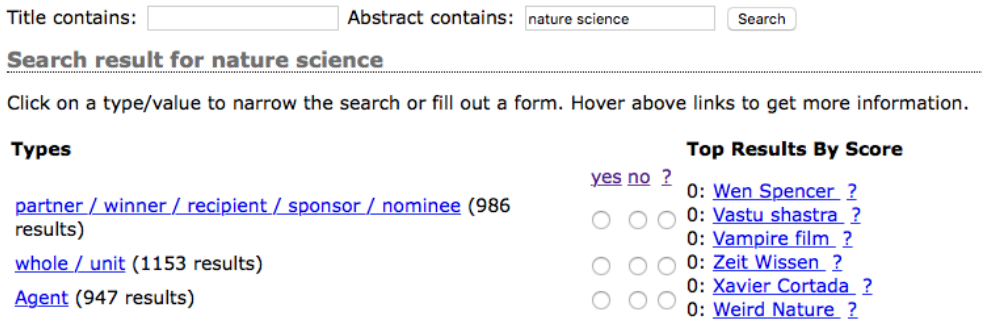


Abbildung 2.13: Semantische facettierte Suche mit Wikipedia Faceted Search

4. *Semantikbasierte Schlüsselwortsuchmaschinen* verbessern Schlüsselwortsuche durch das Einbeziehen der verfügbaren semantischen Daten. Die meisten semantischen Suchmaschinen verfolgen diesen Ansatz, da die Schlüsselwortsuche den Benutzern wohlvertraut ist. Die ersten semantikbasierten Schlüsselwortsuchmaschinen verwendeten gegenüber den Suchmaschinen mit RDF-basierter Anfragesprachen nur noch eine einfache Anfragesprache zur expliziten Beschreibung der Suchbegriffe, z.B. `person:Zuse` [Lei et al., 2006, Guha et al., 2003]. Die heutigen semantikbasierten Sschlüsselwortsuchmaschinen verzichten darauf und erlauben natürlichsprachige Eingaben. Die Benutzer brauchen weder eine Anfragesprache zu lernen, noch müssen sie das Vokabular der zugrundeliegenden Wissensbasis kennen. Der Ablauf lässt sich in zwei wesentliche Schritte aufteilen [Lei et al., 2006]:

1. Bestimmung der zur Abfrage passenden Konzepte in der Wissensbasis.
2. Auffinden der Instanzen, die mit den im Schritt 1 gefundenen Konzepten „eng verwandt“ (verknüpft) sind.

Ein Beispiel ist SIG.MA[18], die die semantischen Metadaten von Webseiten durchsucht, die Ergebnisse aggregiert und strukturiert darstellt [Tummarello et al., 2010] (s. Abbildung 2.14).

Abbildung 2.14: Die semantikbasierte Schlüsselwortsuchmaschine SIG.MA [Schumacher et al., 2011]

[18] `http://sig.ma/` (12.10.2011)

5. *Frage-Antwort-Maschinen* (Question Answering Tools) sind Faktensuchmaschinen, die in der Lage sind Fragen in natürlicher Sprache zu verarbeiten und zu beantworten. Sie verwenden linguistisches Wissen und Techniken des Natural Language Processing (NLP) [Chowdhury, 2003], um die Anfrage zu analysieren und in eine formale Abfrage zu übersetzen. Ebenso werden die gefundenen Fakten in natürlichsprachige Aussagen umgeformt. Die Erstellung der Wissensbasis dieser Suchmaschinen ist aufwändig, da detailliertes Wissen über die vorhandenen Instanzen und deren Beziehungen benötigt wird. Zudem müssen (für die jeweilige Sprache spezifische) linguistische Ressourcen für die Analyse der Anfrage und die natürlichsprachige Formulierung der Antworten vorliegen. Beispiele sind SmartWeb [Wahlster, 2008], AquaLog [Lopez et al., 2005] und Theseus Alexandria[19]. Abbildung 2.15 zeigt, wie Alexandria Fragen beantwortet: es wird die Antwort angegeben (links) und zusätzlich Fakten angezeigt (rechts), wobei die gefundenen Ressourcen mit einer kurzer Beschreibung versehen sind.

Abbildung 2.15: Die Frage-Antwort-Maschine Alexandria

6. *Schlüsselwortsuchmaschinen mit semantischer Nachverarbeitung* verschieben den Schwerpunkt auf die Ergebnisnachverarbeitung. Sie führen eine traditionelle Schlüsselwortsuche aus und extrahieren erst im Nachhinein das Wissen aus den gefundenen Dokumenten, wobei für die Extraktion auf Wissen aus formalen Wissensbasen zurückgegriffen wird. Anschließend stellen diese Suchmaschinen das Ergebnis anhand der gefundenen semantischen Metadaten strukturiert da. Ein Beispiel ist die Suchmaschine ALVIS [Buntine et al., 2005]. Wie in Abbildung 2.16 ersichtlich, werden semantische Metadaten zum Filtern (links) und die Ergebnisdokumente (rechts) angezeigt.

7. *Semantikbasierte Intelligente Visualisierung* unterstützt den Benutzer sowohl bei der Anfragestellung als auch bei der Erfassung der Ergebnismenge durch eine intelligente, interaktive Darstellung des Suchraumes bzw. der Ergebnismenge. Sie bietet explorative Suche [Marchionini, 2006]. Der Benutzer kann schneller als bei einer Schlüsselwortsuche einen Überblick über den Suchraum bzw. die Ergebnisse gewinnen und diese durch weitere Suchschritte bzw. Browsen verfeinern. Dies ist insbesondere dann hilfreich, wenn der Benutzer sich mit dem Thema nicht auskennt oder sich über das Suchziel noch nicht ganz im Klaren ist. Während die facettierte Suche das Filtern der Ergebnismenge durch

[19] `http://alexandria.neofonie.de` (06.01.2016)

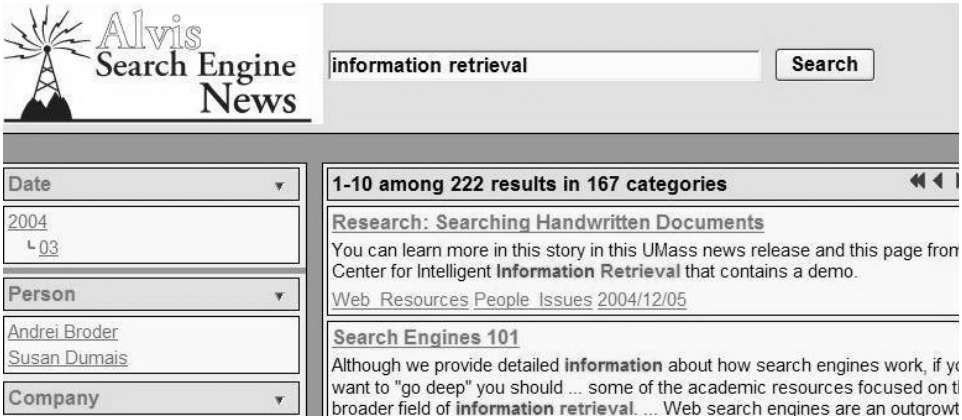


Abbildung 2.16: Die Schlüsselwortsuchmaschine mit semantischer Nachverarbeitung ALVIS [Schumacher et al., 2011]

Auswahl von Facetten forciert, erweitert die explorative Suche den Suchbereich, indem assoziierte Terme, Konzepte und Ressourcen aus dem gesamten Suchraum vorgeschlagen werden [Waitelonis and Sack, 2010]. Die Visualisierung basiert hierbei auf formalen Semantiken, häufig werden semantische Netze oder Kategorien, Klassen und Instanzcluster grafisch abgebildet. Ein Vertreter hierfür ist EyePlorer[20]. Wie in Abbildung 2.17 veranschaulicht, zeigt EyePlorer die gesuchte Ressource an (links), kategorisiert die Suchergebnisse (unterschiedliche Farben) und bietet innerhalb der Kategorien eine Wortwolke der häufigsten Ressourcen an (Punkte und Wörter im Kreis).

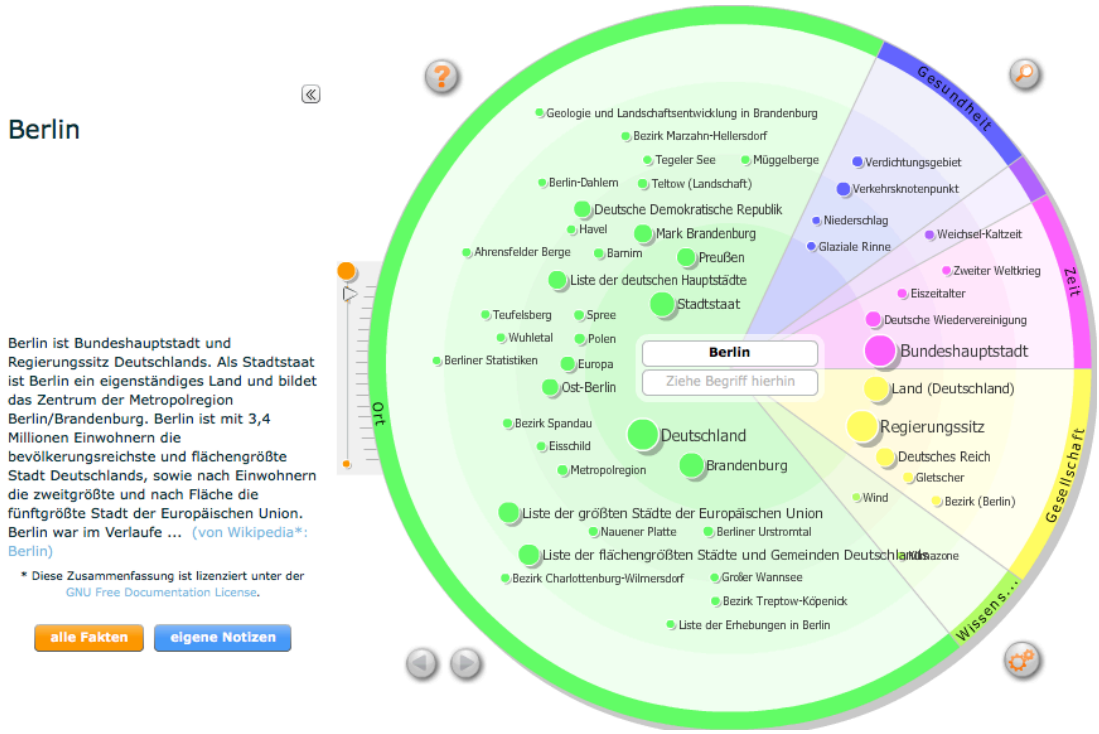


Abbildung 2.17: Semantikbasierte intelligente Visualisierung mit EyePlorer

[20] `http://eyeplorer.com/` (06.01.2016)

2.3.4 Konzeptionelle Sucharchitektur und Suchverfahren

Entsprechend der Definition der semantischen Suche kann formale Semantik in jeder Phase des Suchprozesses eingesetzt werden. Um die verschiedenen Vorgehensweisen und Suchverfahren zu erläutern, wird an dieser Stelle die allgemeine konzeptionelle Architektur semantischer Suchmaschinen aus Abbildung 2.18 herangezogen.

2.3.4.1 Architektur

Der Ausgangspunkt ist der Suchraum, der (auch) formal repräsentiertes Wissen enthält. In den meisten Fällen liegen formale Semantiken zum Zeitpunkt der Suche bereits vor. Aus diesem Grund kann das Erstellen bzw. Füllen der Wissensbasis als der **Offline-Teil** der Sucharchitektur bezeichnet werden [Schumacher and Sintek, 2011, Schumacher et al., 2011]. Das formal repräsentierte Wissen kann hierbei manuell abgebildet oder aus den Inhalten der Datenquellen extrahiert worden sein. Dies geschieht zum einen durch Entitätenextraktion, wobei die sogenannten Named Entities, im Folgenden Entitäten genannt, Dinge mit rigiden Namen wie Personen, Städte usw. sind [Kripke, 1972]). Zum anderen wird Relationextraktion, d.h. die Extraktion von Fakten aus Texten durchgeführt [Kosala and Blockeel, 2000, Nadeau and Sekine, 2007, Sarawagi, 2008]. Die Wissensbasis kann die Inhalte teilweise oder vollständig abbilden, sie beinhaltet z.B. die extrahierten Entitäten oder die Inhalte einer Datenbank, die komplett in einen RDF-Store überführt wurden. Ferner kann zusätzliches Wissen über die Inhalte der Datenquellen in der Wissensbasis liegen, wie z.B. Metadaten eines Dokumentes wie Autor, Erstelldatum oder Format [Schumacher and Sintek, 2011].

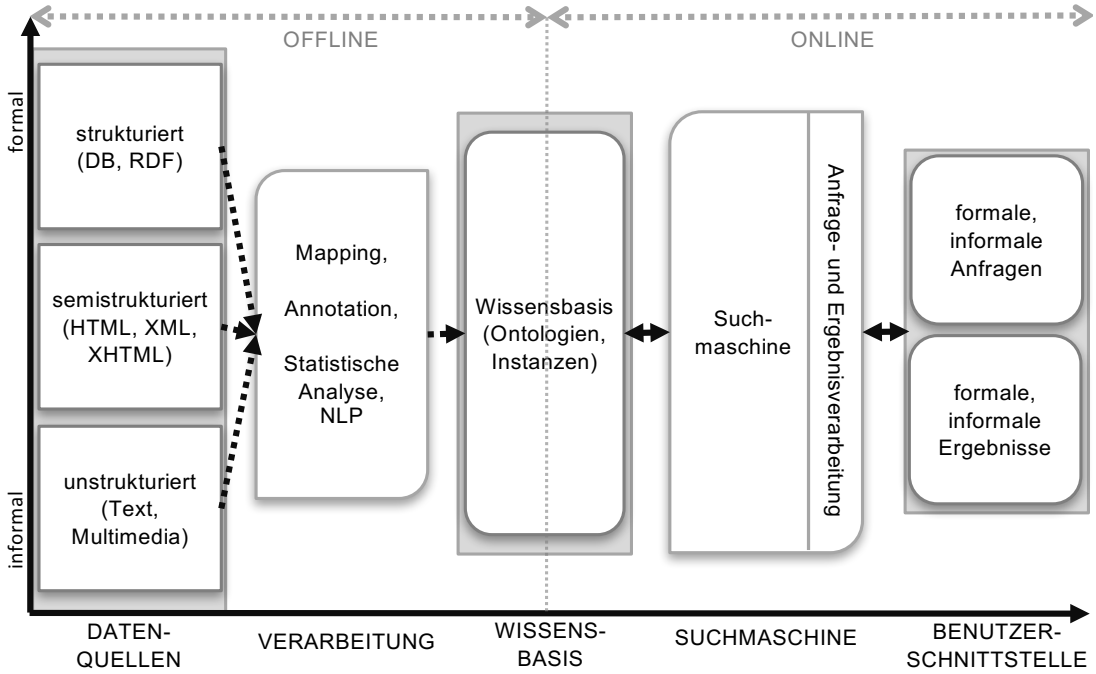


Abbildung 2.18: Architektur semantischer Suchmaschinen [Schumacher et al., 2011]

Das Wissen kann aus verschiedenen Datenquellen kommen, Inhalte sind beispielsweise Textdokumente, Webseiten, XML-Dokumente und Datensätze in Datenbanken. Die nötigen Verarbeitungsschritte, um sie in einer formalen Wissensbasis abzubilden, hängen

davon ab, wie stark strukturiert die Inhalte vorliegen. Mapping verschiedener Vokabulare können für semistrukturierte bzw. strukturierte Inhalte angewendet werden. Statistische Verfahren und Methoden der NLP eignen sich für unstrukturierte und semistrukturierte Inhalte. Annotation mit semantischen Konzepten lässt sich üblicherweise für alle Daten anwenden. Im Rahmen der Abbildung der Inhalte in der formalen Wissensbasis kann ein Zusammenhang zwischen Inhalten verschiedener Datenquellen hergestellt werden, z.B. weil sie dieselben Konzepte aus der formalen Wissensbasis beinhalten [Schumacher and Sintek, 2011].

Die Komponenten, die bei der Ausführung der Anfrage, also online, verwendet werden, bilden den **Online-Teil** der Architektur. Die Wissensbasis mit Ontologien, Instanzen und gegebenenfalls Regeln ist beiden Teilen zugeordnet: Sie wird meist offline gefüllt, es kann aber während des Suchprozesses aus dem vorhandenen Wissen durch logisches Schließen neues Wissen abgeleitet werden. Unter dem Begriff Suchmaschine ist an dieser Stelle der Suchalgorithmus, die Anfragevorverarbeitung sowie Ergebnisnachbereitung gemeint. Sie übernimmt die Anfragen der Benutzerschnittstelle, führt den Suchalgorithmus unter Zugriff auf die Wissensbasis aus und generiert die Ergebnismenge. Die Suchanfrage kann im Allgemeinen in einer formalen Anfragesprache (strukturiert) oder in natürlicher Sprache (nicht formal, unstrukturiert) gestellt werden. Die Repräsentationsform der Ergebnisse ist abhängig von der Beschaffenheit des Suchraumes, sie können formal (z.B. Fakten) oder informal (z.B. Dokumente) sein.

Suchmaschinen der Kategorie semantikbasierte intelligente Visualisierung (Seite 23) weichen von dieser Architektur insofern ab, als ein Großteil des Wissens erst nach der Ausführung der Anfrage (online) extrahiert wird.

In den folgenden Abschnitten werden die Beschaffenheit, die Aufgaben dieser Komponenten und die Ansätze zu ihrer Realisierung beschrieben.

2.3.4.2 Der Suchraum

Der **Suchraum** umfasst die Menge der Objekte, die durchsucht werden. Seine Beschaffenheit ist durch die Repräsentationsformen dieser Objekte bestimmt. Sind z.B. alle Informationen in einer formalen Wissensbasis abgebildet und es wird nach Fakten gesucht, so entspricht der Suchraum der Wissensbasis und es wird eine *Faktensuche* durchgeführt. Dies ist üblicherweise bei Frage-Antwort-Maschinen der Fall (Kapitel 2.3.3). Ergänzt die Wissensbasis jedoch Informationen, die in Form unstrukturierter Dokumente vorliegen, so macht die Wissensbasis nur einen Teil des Suchraumes aus. In dem Fall sind üblicherweise die Dokumente mit formalem Wissen angereichert und es wird *semantisches Dokumentretrieval* durchgeführt. Hierbei spielen zwei Faktoren eine wichtige Rolle: die Kopplung und die Struktur der Ontologie. Man unterscheidet zwischen zwei *Kopplungsarten* [Mangold, 2007]:

- Enge Kopplung: die semantischen Metadaten der Dokumente verweisen explizit auf die Konzepte der Ontologie und umgekehrt.
- Lose Kopplung: die Dokumente sind nicht an die Ontologien gekoppelt, die passende Ontologie muss im Suchprozess ermittelt werden.

Bezüglich der *Struktur der Ontologie* identifizierte Mangold die Properties als das relevante Element für semantische Dokumentsuche und unterscheidet drei Arten von Properties anhand ihrer Aussagekraft. Von stark bis wenig aussagekräftig gibt es domänenspezifi-

sche, standard- und anonyme Properties:

- Domänenspezifische Properties liefern Informationen über die Bedeutung/Art der Beziehung zweier Konzepte aus der Wissensbasis (z.B. capital, member_of).
- Standardproperties beschreiben Abhängigkeiten zwischen Konzepten, meistens sind es linguistische Beziehungen oder Kategoriezugehörigkeiten (z.B. synonym_of, instance_of).
- Anonyme Properties deuten lediglich an, dass zwei Konzepte etwas miteinander zu tun haben, die Verknüpfung hat jedoch keine Semantik.

Die Struktur der Ontologie ist nicht nur durch die Aussagekraft der Properties bestimmt. Die Klassen können sehr abstrakt oder (domänen-)spezifisch sein, z.B. „Person“ oder „Biologe“. Sowohl die Klassenstruktur, als auch die Struktur der Properties können stark hierarchisch oder flach sein.

Die Kopplung und die Aussagekraft der Properties haben einen großen Einfluss auf die **semantische Leistungsfähigkeit** (semantic power) der semantischen Dokumentsuchmaschinen. Je enger die Kopplung ist und je aussagekräftiger die Properties sind, umso spezifischer ist das für die Suche verfügbare Wissen; Suchanfragen können präziser beantwortet werden [Mangold, 2007]. Der Nachteil liegt jedoch darin, dass ein größerer Aufwand in die Erstellung der Wissensbasis und die Annotation der Dokumente mit den semantischen Metadaten investiert werden muss. Dies ist nur für Suchmaschinen einer spezielle Domäne mit einem eingeschränkten Suchraum machbar. Beispiele sind die FACT-Finder Travel Reisesuchmaschine[21] oder die semantische Rezeptsuchmaschine Yummly[22].

2.3.4.3 Anfrageverarbeitung

Formale **Anfragen** (z.B. Suchmaschinen mit RDF-basierten Anfragesprachen), müssen lediglich in eine konkrete Abfragesprache (z.B. SPARQL) übersetzt und ausgeführt werden. Ist die Suchanfrage hingegen in natürlicher Sprache (z.B. semantikbasierte Schlüsselwortsuchmaschinen) also *informal*, oder besteht sie aus formalen und natürlichsprachigen Teilen (z.B. formularbasierte semantische Suchmaschinen), so muss der Bezug zwischen den natürlichsprachigen Teilen und den Konzepten der Wissensbasis hergestellt werden. Hierzu führen die meisten semantischen Suchmaschinen einen *syntaktischen Abgleich* (syntactic matching) durch: Die informalen Anfrageterme werden mit dem textuellen Inhalt der Ontologien und der Instanzbasis syntaktisch verglichen, um die Konzepte zu finden, die gemeint sein könnten. Dieser Schritt basiert, wie auch die Schlüsselwortsuche in traditionellen IR-Systemen, auf dem lexikalischen Vergleich von Termen.

Der einfachste Weg ist zwei Terme auf die exakte Übereinstimmung zu prüfen (exact match). So werden jedoch Abweichungen in der Schreibweise, Tippfehler, unterschiedliche Flektionen, Formen eines Wortes usw. nicht entdeckt. Zur Minderung dieses Problems wurden verschiedene Methoden zur Berechnung der **lexikalischen Ähnlichkeit zweier Terme** entwickelt. Bekannte Beispiele sind die Hamming-Distanz, Edit-Distanz (Levenshtein-Distanz) und die N-Gram-Methode:

- Die *Hamming-Distanz* wurde ursprünglich für gleichlange Terme entwickelt. Die Hamming-Distanz zweier Zeichenketten ist die Anzahl der unterschiedlichen Stellen

[21] `http://www.fact-finder.de/FACT-Finder-Semantic-Travel-Search.html` (06.01.2016)
[22] `http://www.yummly.com/` (06.01.2016)

in ihnen [Baeza-Yates et al., 2011]. Zum Beispiel beträgt die Hamming-Distanz von „expresions“ zu „expression“ vier.

- Die *Edit-Distanz* ist die minimale Anzahl von Einfüge-, Lösch- und Substitutionsoperationen von Zeichen, um die erste Zeichenkette in die zweite umzuwandeln. Die Edit-Distanz von „expresions“ zu „expression“ ist zwei, da ein „s“ hinzugefügt und ein „s“ entfernt werden muss [Croft et al., 2010c, Baeza-Yates et al., 2011].
- Die *N-Gram-Methode* sieht vor, Texte in Fragmente der Länge n zu zerlegen, wobei Fragmente Buchstaben, Phoneme, Wörter und sonstige Einheiten sein können. Zum Vergleich von zwei Termen werden Buchstaben betrachtet. Zum Beispiel wird das Wort „expression“ in die Bigramme „_e“, „ex“, „xp“, „pr“, „re“, „es“, „ss“, „si“, „io“, „on“, „n_“ zerlegt. Die zu vergleichenden Terme werden als Vektoren von N-Grammen betrachtet und ihre Ähnlichkeit mithilfe Ähnlichkeitsfunktionen, wie das Jaccard- oder Dice-Maß, berechnet [Kondrak, 2005, van Rijsbergen, 1979]:

$$jaccard - similarity(d,q) = \frac{\mid ngrams(d) \cap ngrams(q) \mid}{\mid ngrams(d) \cup ngrams(q) \mid} \tag{2.3}$$

$$dice - similarity(d,q) = \frac{2 \cdot \mid ngrams(d) \cap ngrams(q) \mid}{\mid ngrams(d) \mid + \mid ngrams(q) \mid} \tag{2.4}$$

  Für die 2-Gramm-Zerlegung von „expresions“ und „expression“ beträgt das Jaccard-Maß $0,69$, das Dice-Maß $0,81$.

Der syntaktische Abgleich in der Wissensbasis fällt bei denjenigen semantischen Dokumentsuchmaschinen weg, die zuerst eine Schlüsselwortsuche im Dokumentindex durchführen und dann die semantischen Metadaten der Dokumente in die Suche mit einbeziehen.

Semantische Suchmaschinen bedienen sich häufig der **semantischen Anfragemodifizierung**, um die typischen Probleme bei der Suche mit informalen Anfragen, wie das der Homonyme und Synonyme, zu lösen. Dabei wird das Wissen aus der formalen Wissensbasis eingesetzt. Meist betrachten die Suchmaschinen linguistische Beziehungen bzw. den ontologischen Kontext eines Konzeptes. Den lokalen[23] ontologischen Kontext eines Konzeptes bilden die benachbarten Konzepte bis zu einer vorgegebenen Tiefe (Entfernung gemessen an Anzahl der Relationen im Pfad zwischen den Konzepten), wobei alle Relationen verfolgt werden [Dmitrieva et al., 2007]. Je nach Art der Relationen wird für die Anfragemodifizierung der gesamte oder nur ein Teil des ontologischen Kontextes berücksichtigt [Mangold, 2007].

Man unterscheidet zwischen der manuellen Anfragemodifizierung und der maschinellen Anfrageoptimierung.

Die *manuelle Anfragemodifizierung* wird durch den Benutzer durchgeführt, das System bietet ihm die Möglichkeit zur Spezifizierung der Suchwörter. So können beispielsweise zur Disambiguierung alle möglichen Bedeutungen angeboten werden (z.B. in [Nagypál, 2007]). Alternativ können die Klassen oder Kategorien der in Frage kommenden Instanzen zur manuellen Einordnung der Suchwörter herangezogen werden (z.B. in SCORE [Sheth et al., 2002]).

Die *maschinelle Anfrageoptimierung* bedient sich der automatischen Anfrageerweiterung, Anfragekürzung und Substitution der Anfrage [Mangold, 2007]:

[23] Der globale ontologische Kontext ist analog eine Menge von Konzepten, die aus mehreren Ontologien gesammelt wurden [Dmitrieva et al., 2007].

- Für die *Anfrageerweiterung* werden meist Terme, die aus dem ontologischen Kontext der gefundenen Konzepte abgeleitet wurden (z.B. die Namen der benachbarten Konzepte), eingesetzt. Die konjunktive Anfrageerweiterung schränkt die Ergebnismenge auf diejenigen Dokumente ein, die alle Terme bzw. Konzepte beinhalten. Die disjunktive Anfrageerweiterung erweitert die Ergebnismenge um die Dokumente, die einen der Terme bzw. Konzepte beinhalten.
- Die *Anfragekürzung* wird eingesetzt, wenn eine konjunktive Suchanfrage keine Ergebnisse geliefert hat. In dem Fall besteht die Möglichkeit dem Benutzer Teile der Anfrage zur Auswahl anzubieten oder ohne den Benutzer zu bitten eine neue Suche mit Teilanfragen durchzuführen. Die Kürzung konjunktiver Anfragen wirkt entgegengesetzt zur konjunktiven Anfrageerweiterung und kann zur Verbesserung der Vollständigkeit eingesetzt werden.
- Bei der *Substitution* werden bestimmte Terme der Anfrage durch verwandte Terme, die Konzepten aus der Ontologie entsprechen, ersetzt. Ebenso können Terme durch ihre Hyperonymex zur Verallgemeinerung oder durch Hyponyme zur Spezialisierung der Anfrage ersetzt werden.

Abbildung 2.19 gibt einen Überblick der Arten der semantischen Anfragemodifizierung.

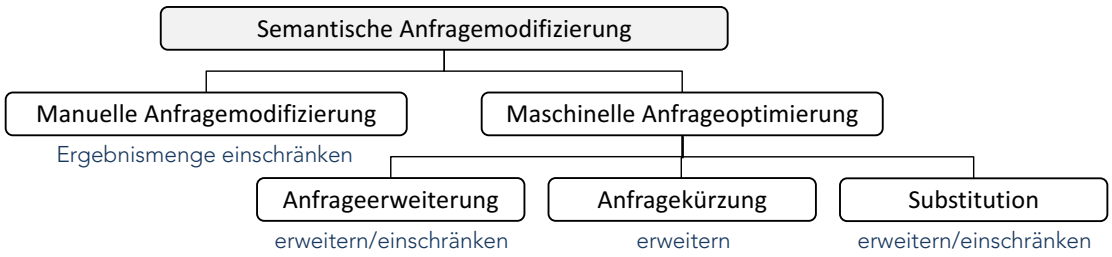


Abbildung 2.19: Arten der semantischen Anfragemodifizierung

[Mangold, 2007] unterscheidet innerhalb der maschinellen Anfrageoptimierung noch zwischen *query rewriting* und *graphbasierter Anfrageoptimierung*. In die erste Kategorie fallen die maschinelle Anfrageerweiterung, Anfragekürzung und Substitution der Anfrage, da die Anfrage explizit geändert wird. Unter graphbasierter Anfrageoptimierung ordnet Mangold Verfahren ein, die im Rahmen der Suche auf dem Graphen traversieren und durch Ausnutzung der Graphstruktur nicht nur die durch die Anfrageterme gefundenen, sondern auch die damit verknüpften Knoten betrachten. Dies kann als eine implizite Anfragemodifizierung angesehen werden, aber auch als der Ansatz für den semantischen Abgleich, wie es im folgenden Abschnitt vorgestellt wird.

2.3.4.4 Ansätze, Suchalgorithmen

Semantische Suchmaschinen führen einen **semantischen Abgleich** aus (s. Seite 17. Dabei werden Algorithmen angewendet, die nach dem syntaktischem Abgleich die Struktur des Graphen der formalen Wissensbasis ausnutzen, d.h. nicht mehr auf der syntaktischen Eben, sondern auf der Ebene der Ressourcen und semantischen Relationen agieren [Hildebrand et al., 2007, Giunchiglia and Shvaiko, 2003]. Dies geschieht ausgehend von den gefundenen Konzepten und/oder den semantischen Metadaten der gefundenen Dokumente. Handelt es sich um eine semantische Dokumentsuchmaschine, so muss eine Kopplung zwischen der Ontologie und den Dokumenten bestehen oder hergestellt worden sein.

Die *einfachsten Ansätze für den semantischen Abgleich* nutzen die verfügbaren Relationen aus, um zu den durch den syntaktischen Vergleich gefundenen Konzepten „verwandte“ Konzepte zu finden und diese zur Anfrageerweiterung einzusetzen. Häufig handelt es sich hierbei um linguistische Beziehungen (Synonyme, Hyperonyme usw.). Eine weitere Möglichkeit besteht darin, die benachbarten Konzepte eines gefundenen Konzeptes, oder die semantischen Metadaten eines gefundenen Dokumentes zur Anreicherung der Suchergebnisse einzusetzen. Solche Informationen können zum Verständnis der Ergebnisse beitragen.

Komplexere Suchalgorithmen lassen sich in die Kategorien Graphtraversierung, logisches Schließen (Reasoning), thesaurusbasierte, tripelbasierte und NLP-basierte Ansätze einordnen. Die **Ansätze** sind teilweise nur für Faktensuche, nur für semantisches Dokumentretrieval oder aber auch für beides geeignet [Schumacher et al., 2011, Hildebrand et al., 2007][24]. Im Folgenden werden diese Ansätze inklusive ihrer Zuordnung vorgestellt.

- *Tripelbasierte Ansätze* (auch statementbasiert genannt) eignen sich für die Faktensuche. Dabei werden die durch den syntaktischen Abgleich gefundenen Konzepte aus der formalen Wissensbasis (s. auch Kapitel 2.3.4.3, Seite 27) in die Menge der Properties und „Nicht-Properties“ aufgeteilt um die möglichen Subjekte, Objekte und Prädikate zu bestimmen. Daraus werden die möglichen Abfragetemplates konstruiert und in SPARQL-Abfragen überführt, um passende Tripel in der Wissensbasis zu finden. So können aus zwei „Nicht-Properties“ c_1 und c_2 die Templates $\langle c_1, ?, c_2\rangle, \langle c_2, ?, c_1\rangle$ generiert werden, wobei das Fragezeichen als Platzhalter für das gesuchte Prädikat dient [Goldschmidt and Krishnamoorthy, 2005].
- *Traversierung auf dem Graphen* eignet sich sowohl für Fakten-, als auch für semantische Dokumentsuche. Solche Ansätze setzen Graphalgorithmen ein, um die Struktur des Graphen für die Suche auszunutzen. Die Algorithmen lassen sich im Allgemeinen in zwei Kategorien einteilen: Für Faktensuche werden häufig solche eingesetzt, die Punkt-zu-Punkt Verbindungen suchen, für semantische Dokumentsuche eher diejenigen Verfahren, die von einem oder mehreren Knoten aus dem Graph explorieren ohne explizit einen Zielknoten zu haben. Ein Beispiel für ersteres sind Algorithmen, die den kürzesten Pfad suchen, um z.B. die Stärke der Beziehung/-semantischen Nähe zweier Konzepte zu bestimmen. Ein bekanntes Graphtraversierungsverfahren der zweiten Kategorie ist das sogenannte Spreading Activation (SA), eine Theorie aus den Kognitionswissenschaften [Anderson, 1983], die für Information Retrieval adaptiert wurde [Cohen and Kjeldsen, 1987, Crestani, 1997]. Der Algorithmus flutet den Graphen mit „Energie“ ausgehend von den initial aktivierten Knoten. Diese sind im Rahmen der semantischen Dokumentsuche (üblicherweise durch den syntaktischen Vergleich, s. Seite 27) identifizierte Konzepte oder durch die Schlüsselwortsuche gefundene Dokumente. Die Bedeutung der Kanten spielt bei dieser Kategorie der Graphtraversierung häufig eine besondere Rolle: Entweder werden die zu verfolgenden Relationen angegeben, z.B. TAP, [Guha et al., 2003], oder Kantengewichte zugeordnet [Hildebrand et al., 2007]. Die Gewichte können

[24] Graphtraversierung, logisches Schließen (Reasoning) und thesaurusbasierte Ansätze wurden durch die Analyse von 35 semantischen Dokumentsuchmaschinen mit natürlichsprachigem Zugang identifiziert [Hildebrand et al., 2007]. Tripelbasierte und NLP-basierte Ansätze konzentrieren sich mehr auf die Faktensuche und waren deshalb bei der Analyse nicht inbegriffen. Sie lassen sich jedoch durch weitere Analyse des Stands der Technik eingrenzen.

manuell (z.B. von einem Domänenexperten) bestimmt [Schreiber et al., 2006] oder automatisch, mit statistischen Verfahren, berechnet werden [Anyanwu, 2005].

- *Logisches Schließen* (Reasoning) leitet aus bestehendem Wissen neues Wissen ab (Inferenz), agiert also auf der Wissensbasis und eignet sich daher zur Faktensuche. Um Wissen abzuleiten, wird u. a. die Transitivität der hierarchischen Relationen ausgenutzt, da dadurch eine Generalisierung bzw. Spezialisierung der Konzepte erfolgt. Beispiele sind $\langle rdfs : subClassOf \rangle$, $\langle rdfs : subPropertyOf \rangle$ und die *part of*-Relationen, die Ganze-Teile-Beziehungen beschreiben. Sucht man beispielsweise in DBpedia nach Personen, die an einem Film beteiligt waren (Property $\langle coparticipatesWith \rangle$), findet man durch logisches Schließen auch den Regisseur (Direktor) des Films, wenn die Property $\langle director \rangle$ $\langle rdfs : subPropertyOf \rangle$ von $\langle coparticipatesWith \rangle$ ist. Ohne logisches Schließen würde man diesen Fakt nicht finden, falls er nicht explizit als Tripel in der Wissensbasis vorliegt. Im Rahmen des Reasonings wird die transitive Hülle der Tripel durchlaufen [Hildebrand et al., 2007]. Suchmaschinen, die Reasoning über die hierarchischen Relationen durchführen, sind z.B. SemSearch [Lei et al., 2006] oder Squirrel [Duke et al., 2007]. Neben den hierarchischen Relationen werden auch die Restriktionen der Domäne und des Wertebereichs der Properties (s. Kapitel 2.2) zum Reasoning genutzt. Zum Beispiel kann durch die Domänenrestriktion abgeleitet werden, dass das Subjekt in einem Tripel mit der Property $\langle director \rangle$ ein Film ist. Insbesondere die Suchsysteme, die aus verschiedenen Datenquellen Entitäten extrahiert haben, nutzen die sogenannten *OWL-Identity*-Relationen ($\langle owl : sameAs \rangle$, $\langle owl : differentFrom \rangle$, $\langle owl : allDifferent \rangle$), da diese die Identität von Konzepten aus unterschiedlichen Quellen ausdrücken. Ein Beispiel hierfür ist Flink [Mika, 2005]. Komplexeres RDFS/OWL-Reasoning als die Transitivität und Identität zu nutzen, kommt in der Regel nur in geschlossenen, sorgfältig modellierten, konsistenten Systemen zum Einsatz. Dies liegt zum Teil daran, dass die Grundlagen des logischen Schließens auf einer endlichen, statischen, konsistenten Faktenmenge mit vollständigem Regelwerk basieren, was in großen Wissensbasen meist nicht der Fall ist. Zudem muss das Regelwerk, die sogenannten Reasoning-Pattern (Axiome), manuell erstellt werden [Fensel and van Harmelen, 2007]. Deshalb arbeiten auf Reasoning basierende Suchmaschinen entweder nur mit sehr allgemeinen Pattern oder sie werden domänenabhängig, für eine handhabbare Menge von vordefinierten Anfragen, zu denen die entsprechenden Regeln existieren, eingesetzt.
- *Thesaurusbasierte Ansätze*, wie z.B. Squiggle [Celino et al., 2007], verwenden Thesauri als Wissensbasis für semantisches Dokumentretrieval. Der ontologische Kontext der Konzepte kann dabei Hyponyme, Hyperonyme, Meronyme, Synonyme, Partonomien usw. beinhalten. Für die Modellierung solcher Thesauri hat sich das Simple Knowledge Organization System, kurz SKOS[25], durchgesetzt [Hildebrand et al., 2007]. Häufig wird auch die WordNet Lexical Database[26] eingesetzt, beispielsweise um Synonyme und Meronyme der Anfrageterme abzufragen [Buscaldi et al., 2005]. Wie im Abschnitt 2.3.4.3 bereits beschrieben werden diese verwendet, um die Anfrage zu modifizieren.
- *Natural Language Processing (NLP)-basierte Verfahren* interpretieren natürlichsprachige Texte unter Einsatz vom (sprachabhängigen) linguistischen Wissen. Sie

[25] `http://www.w3.org/2004/02/skos/` (06.01.2016)
[26] `http://wordnet.princeton.edu` (06.01.2016)

eignen sich sowohl für die Faktensuche als auch für das semantische Dokumentretrieval. NLP-Technologien können dabei offline, für Wissensextraktion aus den Dokumenten und online, zur Analyse der Anfrage eingesetzt werden. Ziel der Offline-Anwendung ist es, Wissen aus den Dokumenten zu extrahieren bzw. diese semantisch zu annotieren[27], was auf unterschiedlichen Ebenen geschehen kann. Auf Wortebene wird eine morphologische Analyse durchgeführt, d.h. die Satzstruktur analysiert. So kann nicht nur die Reihenfolge der Wörter, sondern auch deren Rolle im Satz (Subjekt, Prädikat, Objekt, Zeitbezug, Raumbezug usw.) für die Suche herangezogen werden. Die nächste Stufe ist die semantische Analyse, was unter Einbezug der Morphologie die Erkennung von Entitäten (Shallow Analyse) und noch tiefer gehend, die Erkennung von Relationen beinhalten kann (Deep Shallow Analyse). Dieselben Analyseschritte werden auf die Suchanfrage angewendet, um diese in eine Anfragesprache oder direkt in eine formale Abfrage für die jeweilige Wissensbasis umformen zu können [Paşca, 2003, Hirschman and Gaizauskas, 2001, Frank et al., 2007]. Insbesondere Frage-Antwort-Maschinen setzen NLP für die Analyse der komplexen Anfragen ein und nutzen das gewonnene Wissen, z.B. darüber was als Subjekt, Prädikat und Objekt fungiert, um formale Anfragen zu erstellen und den semantischen Abgleich auf der Wissensbasis durchzuführen [Lei et al., 2006, Frank et al., 2007, Neumann and Xu, 2003, Croft et al., 2010a].

Semantischer Abgleich		**Kurzbeschreibung**	**geeignet für**
Einfache Ansätze	**Verwandte Konzepte**	Anfrageerweiterung mit verwandten Konzepten	semantische Dokumentsuche
Komplexere Ansätze	**Tripelbasiert**	Suchen nach Tripelpattern mit Abfragetemplates	Faktensuche
	Graphtraversierung	Explorieren den RDF-Graphen mithilfe von Graphalgorithmen	Faktensuche, semantische Dokumentsuche
	Logisches Schließen	Nutzen die RDF/S-Axiome und durchlaufen die transitive Hülle der Tripel, um neues (implizites) Wissen abzuleiten	Faktensuche
	Thesaurusbasiert	Verwenden Thesauri als Wissensbasis um mit Hyponymen, Hyperonymen usw. die Anfrage zu modifizieren	semantische Dokumentsuche
	NLP-basiert	Nutzen linguistisches Wissen zur Interpretation der Anfragen und Dokumenten	Faktensuche, semantische Dokumentsuche

Tabelle 2.1: Überblick der Ansätze zum semantischen Abgleich

Tabelle 2.1 gibt einen Überblick der Ansätze und ihrer Einsatzbereiche. Die Verfahren schließen sich gegenseitig nicht aus und werden häufig kombiniert eingesetzt. So eignen sich thesaurusbasierte Ansätze für die Anfrageerweiterung, während für den semantischen Abgleich z.B. tripelbasierte Verfahren oder Graphtraversierungsalgorithmen eingesetzt werden können. Ebenso kann NLP für die Analyse einer natürlichsprachigen Frage eingesetzt werden, der semantische Abgleich jedoch z.B. tripelbasiert geschehen. Zudem lassen

[27] Mit semantischen Metadaten versehen.

sich tripelbasierte Verfahren und logisches Schließen mit Dokumentretrieval ergänzen und für semantisches Dokumentretrieval einsetzen.

2.4 Hybride semantische Suche

Die hybride semantische Suche ist ein Spezialfall der semantischen Suche, die immer mehr erforscht wird. Die Motivation hierfür besteht darin, sowohl Fakten als auch Dokumente mit einer Suchanfrage durchsuchen zu können.

Die **ersten Ansätze Richtung hybrider semantischer Suche** kombinierten die Schlüsselwortsuche auf dem Dokumentindex mit der Suche auf der formalen Wissensbasis[28], jedoch ohne beides, Fakten- und semantische Dokumentsuche, zu unterstützen. So sind TAP [Guha et al., 2003], die Semantic-Web-Suchmaschine in [Rocha et al., 2004], KIM [Kiryakov et al., 2004] sowie CORAAL [Novacek et al., 2009] Suchmaschinen, die zwar neben dem Dokumentindex auch auf der formalen Wissensbasis suchen, sie tun dies aber unabhängig voneinander. TAP setzt das formale Wissen ein, um auf der Ergebnisseite zusätzlich zu den gefundenen Dokumenten auch Fakten über die Instanz anzugeben, die bei dem syntaktischen Abgleich mit der Suchanfrage gefunden wurde. Bei mehreren Instanzen versucht der Ansatz den Subgraphen in der Wissensbasis zu identifizieren, die die Suchanfrage abdeckt [Guha et al., 2003]. [Rocha et al., 2004] betrachten Webseiten als Instanzen einer Ontologie, nach denen eine schlüsselwortbasierte semantische Dokumentsuche ausgeführt wird. Eine Suche nach Konzepten, die keine Webseite repräsentieren, oder nach Fakten wird nicht unterstützt. KIM und CORAAL bieten alternativ eine Schlüsselwortsuche auf dem Dokumentindex oder eine Suche mit formalen Anfragen auf der Wissensbasis an, eine Kombination wird nicht unterstützt [Kiryakov et al., 2004, Novacek et al., 2009]. HyKSS durchsucht die formale Wissensbasis und die Dokumente. Die gefundenen Fakten werden aber nur zur Verbesserung der Reihenfolge der Ergebnisdokumente eingesetzt [Zitzelberger et al., 2014]. Mímir durchsucht ebenfalls beides, setzt strukturierte Inhalte jedoch lediglich als Filter ein, in die Ergebnisse fließen diese nicht ein [Tablan et al., 2015][29].

Diese Suchansätze fallen in die Kategorie semantisches Dokumentretrieval (vgl. Kapitel 2.3.3), sie erweitern die traditionelle Schlüsselwortsuche mit semantischen Technologien. [Rocha et al., 2004] repräsentieren zwar die Dokumente als Instanzen in der formalen Wissensbasis (die Dokumente sind instanziiert), gesucht wird jedoch nach Dokumenten (Webseiten) und nicht nach Konzepten oder Fakten.

Über das semantische Dokumentretrieval hinaus gehen **Suchmaschinen, die sowohl Faktensuche als auch Dokumentsuche durchführen und die Ergebnisse beider Sucharten in ihren Antworten berücksichtigen**.

PowerAqua[30] [Lopez et al., 2006] führt Faktensuche in der formalen Wissensbasis und semantisches bzw. traditionelles Dokumentretrieval auf dem Dokumentindex durch. Als Ergebnis liefert die Suchmaschine Fakten und zusätzlich relevante Dokumente. Hierdurch können Suchanfragen auch dann beantwortet werden, wenn keine Fakten zu der Frage

[28] Häufig „ontologiebasierte“ Suche genannt, wobei die Ontologie und die Instanzbasis durchsucht wird [Guha et al., 2003, Rocha et al., 2004, Kiryakov et al., 2004, Bhagdev et al., 2008].
[29] Diese Verfahren werden unter Stand der Technik, Kapitel 3.1 näher vorgestellt.
[30] `http://technologies.kmi.open.ac.uk/poweraqua/` (06.01.2016)

vorliegen, da in diesem Fall traditionelles Dokumentretrieval durchgeführt wird [Lopez et al., 2012, Uren et al., 2010, Fernandez et al., 2008].

Der Ansatz von [Bhagdev et al., 2008] unterstützt sowohl Dokument- als auch Faktensuche. Sie definieren die hybride semantische Suche als die Kombination von traditioneller Schlüsselwortsuche und der Möglichkeit, auch in den semantischen Metadaten der Dokumente zu suchen und logisches Schließen durchzuführen. Sie konkretisieren diese Aussage als den Einsatz von:

- semantischer Suche für die Teile der Suchanfrage, für die Metadaten vorliegen und
- Schlüsselwortsuche für alle anderen Teile der Anfrage.

Es gibt jedoch mehrere Probleme mit der Definition von [Bhagdev et al., 2008]. Wie im Abschnitt 2.3.4.4 beschrieben ist logisches Schließen nur ein möglicher Ansatz für den semantischen Abgleich, die Definition ist an dieser Stelle zu speziell. Weiterhin deckt die Kombination der traditionellen Schlüsselwortsuche mit der Möglichkeit, auch in den semantischen Metadaten zu suchen, auch das semantische Dokumentretrieval ab, was zu generell ist. Dabei wird unter Ausnutzung der semantischen Metadaten aber ohne Verwendung einer Faktensuche nach Dokumenten gesucht. Zudem besagt die Definition nicht, in welcher Art und Weise die beiden Verfahren kombiniert werden, sie können auch unabhängig voneinander ausgeführt werden[31].

Aus diesen Gründen wird an dieser Stelle die **hybride semantische Suche** wie folgt **definiert**:

> **Die hybride semantische Suche kombiniert traditionelle Schlüsselwortsuche oder semantisches Dokumentretrieval mit der Faktensuche, wobei diese nicht voneinander unabhängig, sondern miteinander verbunden durchgeführt werden und die Suchmaschine sowohl Fakten als auch Dokumente findet.**

Die hybride semantische Suche vereint also Fakten- und Dokumentsuche unter Ausnutzung des zur Verfügung stehenden formalen Wissens und der Dokumente. Dabei werden die Faktensuche und die Dokumentsuche nicht unabhängig voneinander ausgeführt, um lediglich sowohl Dokumente als auch Fakten als Ergebnis zu liefern. Vielmehr integriert ein hybrider Ansatz beide Sucharten, so dass die jeweiligen Teilergebnisse gegenseitig berücksichtigt werden und einen Einfluss auf die weiteren Suche nach Fakten bzw. Dokumenten haben. Die Berücksichtigung der Teilergebnisse kann an mehreren Stellen im Rahmen des Suchprozesses geschehen. Welche Vorgehensweisen es gibt wird im Kapitel 3.1, unter dem Stand der Technik hybrider semantischer Suchmaschinen, sowie im Kapitel 5, der Lösungsansatz SINFIO, genau vorgestellt. Eine formale Beschreibung des hybriden semantischen Suchproblems ist im Kapitel 4.2 zu finden.

Das **Ziel** der hybriden semantischen Suche ist es, die Suchanfragen sowohl dann beantworten zu können, wenn die Information nur formal (Fakten, Konzepte) als auch, wenn diese nur informal (Dokumente) vorliegen und zwar so präzise, wie es die zugrundeliegende Daten erlauben. Die semantische Dokumentsuche nutzt zwar das formale Wissen, um die Suche gegenüber traditionellen Retrievalverfahren zu verbessern, liefert jedoch nur Dokumente als Ergebnis. Die Benutzer müssen in den Dokumentinhalten weitersuchen, um ihr Informationsbedürfnis zu befriedigen. Faktensuche nutzt das vorliegende formale Wissen insofern mehr aus, dass sie Ergebnisse in Form von Fakten liefert, die

[31] Spätere Arbeiten zur hybriden semantische Suche übernehmen diese Definition.

üblicherweise präziser sind als Dokumente (vgl. Kapitel 1.1, Abbildung 1.1). Die hybride semantische Suche kombiniert beides, so dass sowohl Fakten als auch Dokumente bei der Suche berücksichtigt werden. Anfragen können so unter Umständen mit Fakten, die die gesuchte Information sofort ablesbar machen, beantwortet werden. Die Benutzer müssen keine Dokumente durchgehen, um ihr Informationsbedürfnis zu befriedigen. Liegt das Wissen nicht alleine (oder gar nicht) in Form von Fakten vor, so können Benutzer die gefundenen Dokumente zur Informationsfindung heranziehen.

2.5 Evaluierung von Suchmaschinen

Es gibt verschiedene Aspekte und Eigenschaften von Suchmaschinen, die evaluiert werden können. Die meisten davon stehen mit der Effektivität und der Effizienz einer Suchmaschine in Zusammenhang [Robertson, 1981]. Die *Effektivität* misst die Fähigkeit der Suchmaschine, die richtigen Ergebnisse zu finden, die *Effizienz* sagt aus, wie schnell sie diese findet [Croft et al., 2010b]. Effizienzmessungen lassen sich automatisiert durchführen und stellen keine besonderen methodischen Anforderungen. Die Feststellung der Effektivität ist jedoch insbesondere bei semantischen Suchmaschinen ein Problem, da die erprobten Verfahren für die Evaluierung traditioneller Information Retrieval (IR) Systeme häufig nicht eingesetzt werden können. Im Gegensatz zum IR mit definierten Retrievalmodellen gibt es zur Beschreibung von semantischen Suchsystemen keine allgemein anerkannten formalen Modelle. Dies liegt u. a. an der Diversität der Lösungen, wie bereits in den Kapiteln 2.3.3 und 2.3.4 dargelegt. So erfordert beispielsweise die Beurteilung von Frage-Antwort-Systemen andere Methoden als die der semantischen Dokumentsuche. Frage-Antwort-Systeme haben das Ziel, eine Antwort auf die Frage zu geben anstatt eine Reihe von Dokumenten zu liefern, in denen die Antwort enthalten ist. Dabei ist von Interesse, ob die Frage vollständig und korrekt beantwortet wurde. Dokumentsuchmaschinen haben hingegen das Ziel, Dokumente, die wahrscheinlich mehr zur Befriedigung des Informationsbedürfnisses beitragen, möglichst weit oben in der Ergebnisliste zu platzieren [Radev et al., 2002, Croft et al., 2010a]. Die Diversität der semantischen Suchlösungen erfordert auch unterschiedliche Rankingverfahren. Da hinter der Faktensuche und der semantischen Dokumentsuche verschiedene Verfahren liegen, ist auch das Vorgehen bei der Berechnung der Relevanz unterschiedlich[32]. Um Vorgehensweisen und Kennzahlen zu vergleichen und das Ranking einer semantischen Suchmaschine optimieren zu können, bedarf es einer vergleichenden Evaluation.

Die Kernaspekte für repräsentative Aussagen sind *Reliabilität, Validität* und *Objektivität*. Reliabilität ist die Zuverlässigkeit einer Evaluierung im Sinne der Reproduzierbarkeit. Validität ist die Eignung des Verfahrens und der Kennzahlen bezüglich der Zielsetzung der Evaluierung. Die Objektivität von Fragen bzw. Messverfahren ist dann gegeben, wenn die Antworten bzw. die Messwerte unabhängig vom Interviewer bzw. Prüfer sind [Kelly, 2009]. Auf die Resultate bezogen betrifft die Objektivität die Frage, inwieweit diese von den Personen und Geräten unabhängig sind, die in die Sammlung der Daten involviert waren [Janetzko, 2008]. Um die Validität zu gewährleisten, gibt es für viele verschiedene Fragestellungen etablierte Evaluierungsverfahren und Kennzahlen. Die Objektivität und Reliabilität sind jedoch stark davon abhängig, inwieweit benutzergenerierte Inhalte

[32] Ausnahmen bilden Verfahren, die formales Wissen auf traditionelle IR-Modelle abbilden. Solche Verfahren bedienen sich traditioneller Rankingverfahren, in unveränderter oder in leicht angepasster Form. Beispiele hierfür sind im Kapitel 3, auf Seite 50 und 56 zu finden.

involviert sind. Dieser Aspekt macht den wesentlichen Unterschied zwischen der *Evaluierung aus System- und Benutzersicht* aus. Bei Evaluierungen aus Systemsicht generiert der Untersuchungsprozess Zahlen. Häufigkeiten, Raten usw. werden anhand etablierter quantitativer Verfahren und Kennzahlen der deskriptiven Statistik gemessen und ausgewertet. Sie führen in der Regel zu repräsentativen, generalisierbaren Ergebnissen, wobei die Objektivität und daher auch die Reliabilität davon abhängt, inwieweit benutzergenerierte Daten (z.B. Relevanzurteil) involviert sind. Eine Evaluierung aus Benutzersicht ist von der Pragmatik der Benutzer geprägt. Benutzerzentrierte Methoden sind qualitativ[33]. Sie sind subjektiver als die quantitativen Methoden aus Systemsicht, wodurch die Reproduzierbarkeit und Generalisierbarkeit der Ergebnisse leidet [Kelly, 2009]. Da die Suchsysteme jedoch für die Benutzer entwickelt worden sind, ist die zentrale Frage, wie zufrieden der Benutzer damit ist. Um die Benutzerzufriedenheit zu messen, werden, neben den Evaluierungen aus Systemsicht, Benutzerstudien eingesetzt [Manning et al., 2008b].

In diesem Kapitel werden die Grundlagen für die Evaluierung der Effektivität von Information Retrieval Systemen aus Systemsicht (Kapitel 2.5.1), die Grundlagen für die Evaluierung von Rankingverfahren für Suchmaschinen (Kapitel 2.5.2) und die im IR verbreitet eingesetzten benutzerzentrierten Evaluierungsmethoden (Kapitel 2.5.3) vorgestellt. Sie bilden die Basis für die Analyse des Stands der Technik von semantischen Suchmaschinen im Hinblick auf Ranking (Kapitel 3.3) und Evaluierung (Kapitel 3.4).

2.5.1 Evaluierung der Effektivität von Information Retrieval Systemen aus Systemsicht

Die Evaluierung der Effektivität von IR-Systemen basiert auf der **Relevanz** der gefundenen Dokumente zu einer gegebenen Suchanfrage. Die Relevanz sagt aus, ob das Dokument zur Befriedigung des durch die Anfrage ausgedrückten Informationsbedürfnisses beiträgt. Ferber definiert die Relevanz in [Ferber, 2003b] als die Relation r mit:

$$r : D \times Q \rightarrow T \; mit \; d_i \in D = \{d_1, ..., d_m\} \; die \; Menge \; der \; Dokumente, \\ Q \; die \; Menge \; der \; Anfragen, \; T \; eine \; Menge \; von \; Wahrheitswerten \quad (2.5)$$

wobei T im Allgemeinen mit $\{0, 1\}$ *binär*, also nicht relevant oder relevant, ist.

Die binäre Relevanz basiert auf die Annahme, dass alle relevanten Dokumente für den Benutzer gleichermaßen wertvoll sind. Die Relevanz kann jedoch auch als eine kontinuierliche Variable aufgefasst werden: ein Dokument kann für verschiedene Personen bzw. in verschiedenen Kontexten im unterschiedlichen Ausmaß wertvoll sein [Kekäläinen, 2005]. Um dies auszudrücken, wird auf *nicht-binäre Relevanz* (graded relevance, [Robertson, 1977]) zurückgegriffen, die in Form von Kategorien, wie z.B. relevant, teilweise relevant und nicht relevant, zum Einsatz kommt [Kekäläinen, 2005].

Die Relevanzbeurteilung wird von Menschen vorgenommen und ist durch Pragmatik geprägt [Gordon and Pathak, 1999]. Sie ist abhängig von dem mentalen Modell des jeweiligen Benutzers, woraus das Informationsbedürfnis abgeleitet und in Form von Such-

[33] Sowohl für die Erhebung als auch für die Auswertung können quantitative Skalen und Kennzahlen verwendet werden. Diese Skalen bilden jedoch qualitative Masse z.B. anhand ausgesuchter Begriffe, Stufen oder Zahlen ab, um die Angaben unabhängig von der Wortwahl einzelner Personen und somit vergleichbar zu machen.

anfragen formuliert wurde (s. auch Kapitel 2.3.2).

Zur Beurteilung der Effektivität einer Suchmaschine aus Systemsicht gibt es mehrere **Kennzahlen**, die jedoch für unterschiedliche Zwecke geeignet sind.

Die zwei wesentlichen Kennzahlen, einsetzbar bei binärer Relevanz, sind die *Genauigkeit (precision, P)* und *Vollständigkeit (recall, R)*. Die Genauigkeit gibt den Anteil relevanter Dokumente unter den gefundenen Dokumenten an, die Vollständigkeit den Anteil relevanter Dokumente, die gefunden wurden (# steht für Anzahl):

$$P = \frac{\#(gefundene\ relevante\ Dokumente)}{\#(gefundene\ Dokumente)} \tag{2.6}$$

$$R = \frac{\#(gefundene\ relevante\ Dokumente)}{\#(relevante\ Dokumente)} \tag{2.7}$$

Das *F-Maß (F-Measure, F)* ist das gewichtete harmonische Mittel der beiden Kennzahlen:

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R},\ \beta^2 \in [0, \infty[\subset \mathbb{R} \tag{2.8}$$

Das balancierte F-Maß wird standardmäßig mit $\beta = 1$ berechnet, Genauigkeit und Vollständigkeit zählen gleichermaßen [Manning et al., 2008b].

Für nicht-binäre Relevanz eignen sich die generalisierte Genauigkeit und Vollständigkeit zur Effiktivitätsmessung. Sie adaptieren die Genauigkeit und Vollständigkeit, indem den Relevanzstufen Zahlenwerte (z.B. $1-5$ oder 1,00, 0,75, 0,50, 0,25 und 0,00) zugeordnet und für die Berechnung der Kennzahlen aufaddiert werden anstatt die relevanten Dokumente zu zählen. Sei D die Menge der Dokumente und $R \subseteq D$ die Menge der n Dokumente, die als Antwort auf eine Suchanfrage geliefert werden. Sei $r(d)$ der Relevanzwert des Dokumentes d bezüglich der Suchanfrage. Dann ist die *generalisierte, nicht-binäre Genauigkeit* (gP) und die *generalisierte, nicht-binäre Vollständigkeit* (gR) definiert als [Kekäläinen and Järvelin, 2002]:

$$gP = \frac{\sum_{d \in R} r(d)}{n} \tag{2.9}$$

$$gR = \frac{\sum_{d \in R} r(d)}{\sum_{d \in D} r(d)}. \tag{2.10}$$

Die Genauigkeit und die Vollständigkeit basieren auf der Relevanz der Ergebnisse, es spielt jedoch keine Rolle, wie diese geordnet sind. Um die Qualität einer Suchmaschine mit gerankten Ergebnissen durch eine Kennzahl auszudrücken und dabei *auch die Reihenfolge der Ergebnisse zu berücksichtigen*, eignen sich die Kennzahlen Mean Reciprocal Rank (MRR) und Mean Average Presicion (MAP).

Das *Mean Reciprocal Rank (MRR)* kann für die Evaluierung von Prozessen eingesetzt werden, in denen zu einer Anfrage eine geordnete Liste von Antworten geliefert werden. Dies ist bei Retrievalmodellen der Fall, die gerankte Ergebnisse liefern (sog. Ranked Retrieval), wie z.B. das Vektorraummodell. MRR für eine Suchanfrage aus der Menge der Suchanfragen $q_i \in Q$ ist der Kehrwert der Position der korrekten Antwort $rank_i$ in

der Ergebnisliste, also $1/rank_i$. Über die Menge von Suchanfragen wird das harmonische Mittel der Kehrwerte gebildet [Voorhees et al., 1999]:

$$MRR(Q) = \frac{1}{\mid Q \mid} \sum_{i=1}^{|Q|} \frac{1}{rank_i}. \tag{2.11}$$

Gibt es mehrere korrekte Antworten, so betrachtet MRR die Position der ersten korrekten Antwort. Sie eignet sich deshalb gut für Suchaufgaben mit einer wohlbekannten Antwort. Dies ist bei Frage-Antwort-Maschinen der Fall. Deren Ziel ist es, die korrekte Antwort auf die Frage zu geben [Radev et al., 2002, Baeza-Yates and Ribeiro-Neto, 2011d].

Sind mehrere richtige Antworten möglich, so kann auf den Mittelwert der durchschnittlichen Präzision, *Mean Average Presicion (MAP)* zurückgegriffen werden. Der Rang (rank) und damit die Rangordnung der Ergebnisse entspricht dem, wie wahrscheinlich die Suchmaschine auf Basis statistischer Auswertungen die Relevanz des Ergebnisses für den Benutzer einschätzt [Baeza-Yates and Ribeiro-Neto, 2011b]. Dabei berechnet das Rankingverfahren aus der Anfragerepräsentation und der Dokumentrepräsentation ein Gewicht in Form einer Abschätzung der Relevanz des Dokumentes bezüglich der Anfrage [Baeza-Yates and Ribeiro-Neto, 2011a]. Für eine Suchanfrage $q_j \in Q$ ist die durchschnittliche Präzision (Average Precision, AP) der Durchschnitt der Präzision der ersten n Ergebnisse, wobei n so groß ist, dass alle zu q_j relevanten Dokumente in der Menge enthalten sind. Die Berechnung von MAP betrachtet also die Präzision auf allen Recall-Ebenen. Sei Q die Menge aller Suchanfragen, $q_j \in Q$ die j-te Suchanfrage und $d_{1_j}, ..., d_{m_j}$ die dazu relevanten Dokumente, dann ist MAP definiert als [Manning et al., 2008b]:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{2.12}$$

wobei R_{jk} die Menge der gerankten Suchergebnisse von Platz 1 bis zu der Position des Dokumentes d_k ist. Wird ein Dokument nicht gefunden, so ist dessen Präzision mit dem Wert 0 mit einzuberechnen [Manning et al., 2008b].

Tabelle 2.3 gibt einen Überblick der Kennzahlen und ihren wichtigsten Eigenschaften. Die in der Tabelle zugeordneten Methoden werden im Folgenden vorgestellt.

Die **Evaluierungsmethoden** wurden mit der Zeit den wachsenden Dokumentmengen und damit verbundenen Herausforderung an die Handhabbarkeit angepasst.

Der Standardansatz nach der *Cranfield-Methode* sieht vor, die Menge der Dokumente für jede zur Evaluierung verwendeten Suchanfrage auf ihre Relevanz hin zu beurteilen, also einen sogenannten *Gold Standard* (auch Ground Truth genannt) zu erstellen auf dessen Basis die Genauigkeit und die Vollständigkeit berechnet werden können [Cleverdon, 1997]. Für die Evaluierung von IR-Systemen existieren solche Gold Standards, wie z.B. der TREC Web Track-Corpus[34]. Sie bieten eine thematisch kategorisierte Dokumentmenge, eine Menge von Anfragen (z.B. aus Anfragelogs großer Suchmaschinen zusammengestellt) und die Relevanzbeurteilung der Dokumente bezüglich der Anfragen an. Um das Problem der Pragmatik abzuschwächen, wird die Relevanz durch die Beurteilung mehrerer Experten bestimmt. Inwieweit die vorgenommenen Relevanzurteile übereinstimmen

[34] `http://trec.nist.gov/data/webmain.html` (06.01.2016

kann beispielsweise durch die *Kappa-Statistik* bestimmt werden. Die Kappa-Statistik berechnet die Rate der Übereinstimmung: Werte von 0,41 bis 0,6 sind als faire, von 0,61 bis 0,8 als substanzielle und Werte ab 0,81 als fast perfekte Übereinstimmung anzusehen [Viera et al., 2005][35]. Tabelle 2.2 stellt die Anzahl übereinstimmender und nicht übereinstimmender Urteile von zwei Personen bei k Relevanzstufen dar (Verallgemeinerung der Tabelle und Formel aus [Manning et al., 2008b]).

Person A/ Person B	**Relevanzstufe 1**	**...**	**Relevanzstufe n**	**Randhäufigkeit**
Relevanzstufe 1	h_{11}	...	h_{1k}	$\sum_{i=1}^{k} h_{1i}$
...	...	...	...	...
Relevanzstufe n	h_{k1}	...	h_{kk}	$\sum_{i=1}^{k} h_{ki}$
Randhäufigkeit	$\sum_{j=1}^{k} h_{j1}$	...	$\sum_{j=1}^{k} h_{jk}$	$\sum_{i=1}^{k}\sum_{j=1}^{k} h_{ij} = R$

Tabelle 2.2: Anzahl übereinstimmender und nicht übereinstimmender Relevanzurteile zweier Personen bei k Relevanzstufen

Anhand der Häufigkeiten lässt sich der Anteil übereinstimmender Urteile $P(A)$, die anteiligen Randwerte der Relevanzstufen $P(Relevanzstufe\ k)$ und daraus die Wahrscheinlichkeit, dass zwei Urteile zufällig übereinstimmen $P(E)$, berechnen.

$$P(A) = \frac{\sum_{i=1}^{k} h_{ii}}{R} \tag{2.13}$$

$$P(Relevanzstufe\ k) = (\sum_{i=1}^{k} h_{ki} + \sum_{j=1}^{k} h_{jk})/(R+R) \tag{2.14}$$

$$P(E) = \sum_{i=1}^{k} P(Relevanzstufe\ i)^2 \tag{2.15}$$

Der Kappa Wert ist dann definiert als:

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}. \tag{2.16}$$

Im Falle von mehr als zwei Personen wird eine Generalisierung der Kappa-Statistik eingesetzt [Fleiss, 1971]. Grundlage hierfür bietet eine Tabelle mit der Anzahl der Zustimmungen n_{ij} für alle N $(i = 1, ..., N)$ Fragen und alle k $(j = 1, ..., k)$ Relevanzstufen. Dann ist der Anteil der Zustimmung für die Relevanzstufe j bei n Relevanzurteile pro Frage:

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}. \tag{2.17}$$

[35] Es gibt keine als Standard angesehene Einteilung der Kappa-Werte. In [Landis and Koch, 1977] werden beispielsweise bereits Werte über 0,75 als eine exzellente und 0,4 als untere Grenze für eine faire Übereinstimmung betrachtet.

Der Anteil der übereinstimmenden Urteile $P(A)$ wird definiert als der Durchschnitt der Anteile übereinstimmender Urteile je Frage P_i:

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{2.18}$$

$$P(A) = \frac{1}{N} \sum_{i=1}^{N} P_i. \tag{2.19}$$

Die Wahrscheinlichkeit $P(E)$, dass die Urteile zufällig übereinstimmen, wird aus p_j berechnet:

$$P(E) = \sum_{j=1}^{k} p_j^2. \tag{2.20}$$

Der Kappa-Wert ist, wie in Formel 2.16 definiert $(P(A) - P(E))/(1 - P(E))$.

Die Kappa-Statistik erfüllt ihren Zweck, wenn die Randhäufigkeiten in der Tabelle 2.2 balanciert sind. Ist dies nicht der Fall, so kann eines der beiden Paradoxa auftreten [Feinstein and Cicchetti, 1990]:

- Ein hoher $P(A)$-Wert kann durch ein substanzielles Ungleichgewicht der Randhäufigkeiten (vertikal oder horizontal) und dadurch hohen $P(E)$-Wert drastisch gesenkt werden.
- Durch ein asymmetrisches Ungleichgewicht der Randhäufigkeiten wird ein höherer Kappa-Wert generiert als durch symmetrisches Ungleichgewicht, wobei eine nicht perfekte Symmetrie mehr zur Erhöhung beiträgt als eine perfekte.

[Cicchetti and Feinstein, 1990] schlagen vor, die Statistik für negative und positive Antworten getrennt zu berechnen. Bei ungerader Anzahl von Relevanzstufen, insbesondere wenn eine neutrale (0-)Stufe inbegriffen ist, liefert diese Vorgehensweise jedoch kein faires Beurteilungskriterium. Eine andere Lösung der Paradoxa basiert auf der Auffassung, dass in vielen Studien die Verteilung der Randhäufigkeiten als frei betrachtet werden sollte, da die Verteilung über die Kategorien nicht a priori bekannt sind (free-marginal Kappa). Die Personen sind also nicht daran gebunden, jeder Kategorie eine gewisse Anzahl von Fällen zuordnen zu müssen [Brennan and Prediger, 1981, Randolph, 2005]. Dies trifft auf die Relevanzbeurteilung bei IR-Systemen zu: Es ist nicht von vornherein bekannt, wie viele Dokumente einer Relevanzstufe zugeordnet werden. In solchen Fällen kann die Wahrscheinlichkeit, dass zwei Urteile zufällig übereinstimmen, durch $P(E) = 1/n$ berechnet werden. Die sogenannte multirater free-marginal Kappa ($mfmK$) ist definiert als [Randolph, 2005]:

$$mfmK = \frac{P(A) - 1/n}{1 - 1/n} \tag{2.21}$$

Die genannten Voraussetzungen sind erfüllt, wenn durch die Studie ein Gold Standard für die Suche erstellt werden soll: Die Zuordnung der Suchergebnisse zu den Relevanzstufen ist frei, eine Verteilung ist nicht bekannt oder vorgegeben.

In großen Dokumentmengen, die keine Relevanzbeurteilung aller Dokumente hinsichtlich der Testsuchanfragen beinhalten (wie es die Standard Cranfield-Methode fordert),

kann die Vollständigkeit nicht berechnet werden. Zudem kann bei großen Antwortmengen die Relevanz der Ergebnisse nicht vollständig beurteilt und deshalb auch die Genauigkeit nicht berechnet werden. Aus diesen Gründen wird häufig das sogenannte *Pooling* angewendet. Dabei wird ein Pool aus den ersten k Ergebnisdokumenten der zu vergleichenden Suchsysteme erstellt. Die Dokumente werden von mehreren geeigneten Personen (z.B. von Domänenexperten, Informationswissenschaftlern usw.) auf ihre Relevanz hin beurteilt [Manning et al., 2008b]. Die Effektivität der Suchmaschine wird auf Basis der Dokumente im Pool und ihrer Relevanzurteile berechnet, wobei Pooling auf der Annahme basiert, dass die meisten relevanten Dokumente sich in dem Pool befinden[36]. Die nicht darin enthaltenen Dokumente werden als nicht relevant betrachtet [Baeza-Yates and Ribeiro-Neto, 2011c, Manning et al., 2008b, Croft et al., 2010b].

Kennzahl	Methode	Relevanzart	Besonderheiten
Genauigkeit (P), Vollständigkeit (R) und F-Maß (F)	Cranfield	binäre Relevanz	Betrachtet den Anteil relevanter Ergebnisse (P) und Anteil gefundener relevanter Ergebnisse (R), F bildet P und R auf eine Kennzahl ab. Die Position wird nicht mit einbezogen, somit sind P, R und F sowohl für gerankte als auch für nicht gerankte Ergebnislisten geeignet.
Genauigkeit an der Stelle *k* (P@k)	Pooling, Cranfield	binäre Relevanz	Berechnet die Genauigkeit über die ersten *k* Ergebnisse. Die Position wird nicht mit einbezogen. Da jedoch nur die ersten *k* Ergebnisse betrachtet werden, ist diese Kennzahl nur für gerankte Ergebnislisten geeignet.
Generalisierte Genauigkeit (gP) und Vollständigkeit (gR)	Pooling, Cranfield	nicht-binäre Relevanz	Relevanzstufen werden auf Zahlenwerte abgebildet und fließen in die Berechnung mit ein.
Mean Reciprocal Rank (MRR)	andere (kann als ein Spezialfall der Cranfield-Methode angesehen werden)	Korrektheit (kann als binäre Relevanz betrachtet werden)	Betrachtet die Position der ersten korrekten Antwort, weitere korrekte Antworten werden nicht berücksichtigt. Nur für gerankte Ergebnislisten geeignet.
Mean Average Precision (MAP)	Cranfield	binäre Relevanz	Betrachtet die Genauigkeit der ersten *n* Ergebnisse, wobei *n* so groß sein muss, dass alle relevanten Dokumente enthalten sind. Bei nicht gerankten Ergebnislisten müssten alle Antwortdokumente betrachtet werden.

Tabelle 2.3: Übersicht der Kennzahlen

Tabelle 2.3 fasst zusammen, welche Kennzahlen bei welchen Methoden eingesetzt werden können. MRR basiert auf der Position der ersten korrekten Antwort und nicht auf die Genauigkeit, Pooling ist daher nicht relevant. MAP erfordert die Betrachtung der Präzision auf jeder Recall-Ebene, alle relevanten Dokumente werden mit einbezogen. Dies

[36] Die Begründung für die Vorgehensweise ist, dass die Benutzer sich in den meisten Suchsystemen tatsächlich nur die ersten hoch gerankten Ergebnisse anschauen [Croft et al., 2010b].

ist für Pooling nicht geeignet, vielmehr wird die sogenannte Präzision an der Stelle k (Precision at k), also die Präzision über die ersten k Ergebnisdokumente der jeweiligen Suchmaschine berechnet [Croft et al., 2010b, Baeza-Yates and Ribeiro-Neto, 2011d]. Bei mehrstufiger Relevanz können die generalisierte Genauigkeit und Vollständigkeit aus der Formel 2.9 eingesetzt werden. Sie werden, analog zu der nicht generalisierten Genauigkeit und Vollständigkeit aus Formel 2.6, über die ersten k Ergebnisse berechnet ($n = k$ in Formel 2.9).

2.5.2 Evaluierung von Rankingverfahren für Suchmaschinen

Relevante Ergebnisse hoch zu ranken, ist ein wesentlicher Erfolgsfaktor von Suchmaschinen. Zur Beurteilung von Rankingverfahren existieren mehrere Kennzahlen und Korrelationsmaße. Kennzahlen drücken die Performanz eines Rankingalgorithmus durch eine Zahl aus. Korrelationsmaße vergleichen die Korrelation zwischen zwei Rangordnungen. Werden sie zum Vergleich zwischen einer durch Rankingverfahren erzeugte Rangordnung und der idealen Rangordnung aus einem Gold Standard eingesetzt, so kann ebenfalls eine Aussage über die Performanz von Rankingalgorithmen getroffen werden.

Kennzahlen, die sich auch für nicht-binäre Relevanz eignen und die Nützlichkeit mit einschließen, dass ein Ergebnisdokument an einer bestimmten Position in der Ergebnisliste steht, sind *discounted cumulative gain (DCG)* und *normalized discounted cumulative gain (NDCG)*. Sie basieren auf zwei Annahmen [Järvelin and Kekäläinen, 2000]:

- Hochrelevante Dokumente sind nützlicher, wenn sie eine bessere Position in der Ergebnisliste haben, also höher gerankt sind;
- Hochrelevante Dokumente sind nützlicher als weniger relevante Dokumente, die jedoch nützlicher sind als nicht relevante Dokumente.

DCG wird auf Basis des *cumulative gain (CG)* berechnet. CG an der Rangposition p ist definiert als [Järvelin and Kekäläinen, 2000]:

$$CG_p = \sum_{i=1}^{p} rel_i \tag{2.22}$$

wobei rel_i die Relevanzstufe des Ergebnisses an der Position i bezeichnet, z.B. 2 für relevant, 1 für teilweise relevant und 0 für nicht relevant. Dieser Wert bezieht die Ordnung der Ergebnisse noch nicht mit ein. Um genau dies zu erreichen, benachteiligt DCG die hochrelevanten Dokumente, die nicht entsprechend gerankt sind, durch die Reduktion ihrer CG logarithmisch proportional zu ihrer Position. DCG an der Position p ist definiert als [Järvelin and Kekäläinen, 2000][37]:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_b(i+1)}. \tag{2.23}$$

Da die Ergebnislisten verschiedener Suchanfragen unterschiedlich lang sein können und ein Performanzvergleich bei unterschiedlichen Suchanfragen nicht alleine auf Basis von

[37] Die Formeln sowohl für CG als auch DCG sind die umgeformten, nicht-rekursiven Versionen der rekursiven Definitionen aus [Järvelin and Kekäläinen, 2000].

DCG durchgeführt werden kann, wird DCG über die Menge der Suchanfragen normalisiert (NDCG). Dazu werden die Ergebnisse nach ihrer Relevanz sortiert, so dass der höchstmögliche DCG bis Position p, das sogenannte ideale DCG (IDCG) erreicht wird. NDCG an der Position p ist dann:

$$NDCG_p = \frac{DCG_i}{IDCG_p}. \tag{2.24}$$

Um die durchschnittliche Performanz des Rankingalgorithmus einer Suchmaschine zu berechnen, kann der Durchschnitt der NDCG über alle Suchanfragen bestimmt werden. Die Schwierigkeit bei dem Einsatz von NDCG ist, dass keine ideale Ordnung der Ergebnisse hergestellt werden kann, wenn ausschließlich teilweise relevante Ergebnisse geliefert werden [Järvelin and Kekäläinen, 2000].

Die Qualitätsmessung von Rankingverfahren mit DCG und NDCG basiert also auf einem Gold Standard, indem eine mehrstufige Relevanzbeurteilung der Dokumente für eine Menge von Suchanfragen vorliegt. Liegt kein Gold Standard vor, so wird die Relevanzbeurteilung bis Position p zur Evaluierungszeit von den Benutzern vorgenommen (z.B. in [Al-Maskari et al., 2007]).

	Name	Besonderheiten
Kennzahl	**DCG**	Nur für gleichlange Ergebnislisten geeignet
	NDCG	Auch für unterschiedlich lange Ergebnislisten geeignet
Korrelationsmaß	**Spearman Koeffizient**	Auch für Ergebnislisten mit unterschiedlichen Ergebnissen geeignet
	Kendall Tau Koeffizient	Nur für Ergebnislisten mit denselben Ergebnissen geeignet

Tabelle 2.4: Übersicht der Kennzahlen und Korrelationsmaße zur Beurteilung von Rankingverfahren

Ranking Korrelationsmaße werden eingesetzt, um die Korrelation zwischen Rangordnungen zu messen. Auf diese Art können Rankingverfahren miteinander verglichen werden [Baeza-Yates and Ribeiro-Neto, 2011d]. Ist die Qualität eines Rankingverfahrens zu messen, so kann die Korrelation mit einem Gold Standard berechnet werden.

Einer der am häufigsten eingesetzten Maße ist der *Spearmans Rangkorrelationskoeffizient* S. Er basiert auf dem Unterschied der Relevanzen $s_{1,i}$, $s_{2,i}$ der Ergebnisdokumente an der Position i $d_{1,i}, d_{2,i}$ in der Ergebnisliste zweier Rankingverfahren R_1, R_2 und wird zu einer Suchanfrage über die ersten k Ergebnisse berechnet [Baeza-Yates and Ribeiro-Neto, 2011d]:

$$S(R_1, R_2) = 1 - 2\frac{\sum_{i=1}^{k}(s_{1,i} - s_{2,i})^2}{\frac{k(k^2-1)}{3}} = 1 - \frac{6\sum_{i=1}^{k}(s_{1,i} - s_{2,i})^2}{k(k^2-1)} \tag{2.25}$$

Dabei wird der maximale Wert von $\sum_{i=1}^{k}(s_{1,i} - s_{2,i})^2$, nämlich $\frac{k(k^2-1)}{3}$ zur Normalisierung eingesetzt. Die Multiplikation des normalisierten Terms mit 2 spannt den Wertebereich

$[0, 2] \in \mathbb{R}$ auf. Die Subtraktion aus 1 verschiebt ihn dann auf $[-1, 1] \in \mathbb{R}$. -1 bedeutet maximale Differenz, $+1$ eine perfekte Korrelation.

Häufig ist die Anzahl der Ergebnisse von R_1 und R_2 nicht gleich. In solchen Fällen wird die Menge der Ergebnisse S_{1+2} über beide Ergebnismengen bestimmt, $k = |S_{1+2}|$ gesetzt und R_1 sowie R_2, um die jeweils fehlenden Dokumente erweitert. Dabei werden die Dokumente am Ende der Ergebnisliste angefügt, ihre Reihenfolge ist durch ihren Rang bestimmt.

Ein seltener verwendetes, aber leichter interpretierbares Ranking Korrelationsmaß ist der *Kendall Tau Koeffizient* [Baeza-Yates and Ribeiro-Neto, 2011d]. Er basiert auf der Differenz der Relevanzen $s_{1,i} - s_{1,j}$ und $s_{2,i} - s_{2,j}$ von je zwei Dokumenten d_i, d_j für zwei Rankings R_1, R_2 bis Position k. Haben beide Differenzen dasselbe Vorzeichen, so sind d_i, d_j konkordant, sonst in beiden Rankings diskordant. Um den Korrelationskoeffezienten zu bestimmen, werden pro Ranking alle geordnete Paare über die ersten k Ergebnisse gebildet. Für die ersten 4 Ergebnisse in R_1, d_j, d_k, d_l, d_m, sind die $k(k-1)/2 = 6$ Ergebnispaare $(d_j, d_k), (d_j, d_l), (d_j, d_m), (d_k, d_l), (d_k, d_m), (d_l, d_m)$. Für jedes Paar aus R_1 und R_2 wird die Konkordanz bzw. Diskordanz zu dem entsprechenden Paar aus dem jeweils anderen Ranking bestimmt. Der Vorteil dieser Paarbildung ist, dass diese Eigenschaft über die Reihenfolge der Dokumente in einem Paar direkt ersichtlich ist: wenn z.B. (d_j, d_l) in R_1, aber (d_l, d_j) in R_2 liegt, so besteht Diskordanz. Sei C die Menge der konkordanten Paare, D die Menge der diskordanten Paare über beide Rankings. Der Kendall Tau Koeffizient r wird dann auf Basis der Wahrscheinlichkeit der Konkordanz $P(R_1 = R_2) = |C|/k(k-1)$ bzw. Diskordanz $P(R_1 \neq R_2) = |D|/k(k-1)$ berechnet [Baeza-Yates and Ribeiro-Neto, 2011d]:

$$r(R_1, R_2) = P(R_1 = R_2) - P(R_1 \neq R_2). \tag{2.26}$$

Dieser Koeffizient lässt sich nur auf Ergebnismengen einsetzen, die dieselben Ergebnisse beinhalten.

Tabelle 2.4 gibt einen Übersicht der vorgestellten Kennzahlen und Korrelationsmaße.

2.5.3 Benutzerzentrierte Evaluationsmethoden im Information Retrieval

Benutzerzentrierte Evaluationsmethoden beobachten die Benutzer bei der Interaktion mit dem System und/oder befragen sie nach den Aspekten, die zur Beantwortung der jeweiligen Forschungsfrage bzw. zum Test der Hypothese relevant sind. In diesem Kapitel werden die im IR am häufigsten verwendeten Methoden vorgestellt.[38]

Evaluierung mit **Fragebogen** ist ein häufig eingesetzte Methode. Fragebögen können geschlossene Fragen (mögliche Antworten sind vorgegeben) und offene Fragen (Antwort als freier Text) beinhalten. Mit geschlossenen Fragen erhält man quantitative Daten, die sich statistisch auswerten und vergleichen lassen. Offene Fragen liefern qualitative Daten, die üblicherweise stärker die eigene Meinung der Befragten widerspiegeln und Informationen dazu liefern können, warum eine Präferenz besteht. Sie erlauben eine tiefere Einsicht in die Gründe der jeweiligen Beurteilung und helfen die Antworten auf geschlossene Fragen zu kontextualisieren und zu interpretieren [Kelly, 2009]. Geschlossene Fragen bedienen sich typischerweise der Likert-Skala oder dem semantischen Differenzial. Die

[38] Für weitere Informationen siehe [Baeza-Yates and Ribeiro-Neto, 2011d, Kelly, 2009] und [Manning et al., 2008b].

Likert-Skala bietet den Befragten die Möglichkeit positive und negative Fragen auf einer ordinalen, rangsortierten Skala von 5-7 Stufen[39] von starker bis hin zu keiner Zustimmung zu beantworten. Die Auswertung erfolgt summativ. Im IR wird die Likert-Skala mit unterschiedlicher Anzahl von Stufen und unterschiedlichen ordinalen Wertesets eingesetzt [Kelly, 2009]. Das *semantische Differenzial* präsentiert Antonym-Paare mit einer nicht nummerierten Skala dazwischen, auf der die Benutzer frei ihre Einordnung einzeichnen können. Die Ergebnisse werden meist diskretisiert und statistisch ausgewertet. Dem Einsatz solch einer Skala wird der Vorteil zugesprochen, dass die Benutzer nicht durch die vorgegebenen Werte beeinflusst werden.

Die Formulierung der Fragen hat einen wesentlichen Einfluss auf die Validität der fragebogenbasierten Studien. Die Fragen müssen eindeutig sein und dürfen nicht polarisiert oder überladen werden. Bei geschlossenen Fragen ist auf die Benennung der Skalenpunkte zu achten, die Bezeichnungen müssen genügend Möglichkeiten anbieten, exklusiv und vollständig sein. Bei offenen Fragen ist die Platzierung der Fragen im Fragebogen ein weiterer wichtiger Einflussfaktor [Payne, 2014].

Für die Evaluierung der Benutzerfreundlichkeit (Usability) existieren zahlreiche standardisierte Fragebögen, die auch für Suchmaschinen eingesetzt werden können. Weit verbreitet sind die System Usability Scale (SUS) [Brooke, 1996], der Questionnaire for User Interface Satisfaction [Chin et al., 1988] und der SUMI Fragebogen [Kirakowski and Corbett, 1993].

Interviews werden im IR eingesetzt, um Fragebögen mit offenen Fragen auszufüllen, da die Befragten dazu neigen, in gesprochener Sprache längere Antworten zu geben [Kelly, 2009]. Der Vergleich von Interview, papierbasierter und elektronischer Fragebogenerfassung hat jedoch ergeben, dass trotz umfangreicheren Antworten kein Unterschied in den inhaltlichen Aussagen besteht [Kelly et al., 2008].

Eine Evaluierung durch **Analyse der Logdaten** ist attraktiv, da kein zusätzlicher Aufwand von Seiten des Benutzers erforderlich ist und eine Erfassung des Benutzerverhaltens in der natürlichen Interaktionsumgebung möglich ist [Kelly, 2009]. Durch Loggen können die Aspekte einer Suchmaschine untersucht werden, deren Beurteilung (oder der Vergleich verschiedener Variationen) aus der Benutzerinteraktion abgeleitet werden kann. Es kann beispielsweise festgestellt werden, welche Funktionalitäten wie häufig verwendet werden. Es können Suchanfragen, komplette Suchprozesse, sowie das Benutzerverhalten analysiert werden. Diese Methode eignet sich auch zur Effektivitätsmessung. Hierfür wird das sogenannte „Clicktrough Data“ gesammelt, wobei ein Datensatz aus der Suchanfrage, dem Ranking und der Menge der Ergebnisse, die der Benutzer angeklickt hat, besteht [Joachims, 2002]. Bei der Ergebnisinterpretation ist jedoch zu beachten, dass die Logdaten implizites Relevanzfeedback beinhalten. Daher sind sie, im Vergleich zur Evaluierung mit explizitem Relevanzfeedback, vager [Baeza-Yates and Ribeiro-Neto, 2011d]. Ein Nachteil der Logdatenanalyse ist, dass Aktivitäten außerhalb der Anwendung nicht mit erfasst werden können, die Intention hinter einer Aktion bleibt meist unbekannt [Kelly, 2009, Grimes et al., 2007].

[39] Traditionell werden 5 Stufen verwendet: „stimme völlig zu“, „stimme eher zu“, „unentschieden“, „stimme eher nicht zu“, „stimme überhaupt nicht zu“ [Likert, 1932].

Side-by-Side Panels bieten eine leicht umsetzbare benutzerzentrierte Alternative, um zwei Systeme zu vergleichen. Möchte man beispielsweise zwei Rankingfunktionen vergleichen, so werden jeweils die ersten k Ergebnisse zur selben Suchanfrage und im gleichen Design nebeneinander dargestellt und die Benutzerinteraktionen beobachtet. Es kann festgestellt werden, welches Rankingverfahren aus Sicht des Benutzers besser ist, jedoch nicht wie viel besser [Baeza-Yates and Ribeiro-Neto, 2011d].

A/B Tests werden eingesetzt, um zu evaluieren, ob bestimmte Änderungen eines Systems aus Benutzersicht positiv oder negativ sind. Hierfür wird eine repräsentative Menge an ausgewählten Benutzern zum veränderten System weitergeleitet und beobachtet, wie sie mit der Änderung umgehen. Diese Technik eignet sich insbesondere für die Untersuchung von Änderungen an der grafischen Benutzerschnittstelle von stark benutzten Webseiten, da dort auch kleinere Änderungen eher auffallen [Baeza-Yates and Ribeiro-Neto, 2011d].

Die bisher vorgestellten Methoden erfassen keine äußeren Einflüsse, sie verfolgen die Benutzerinteraktionen mit dem System oder befragen den Benutzer. Die Umgebung, die Interaktion von außen kann durch **Benutzerbeobachtung** (per Kamera oder durch Beisitzen) inkludiert werden. Solche Studien sind jedoch sowohl in der Ausführung, als auch in der Auswertung aufwändig. Zudem sind sie laborgebunden, daher interagieren die Benutzer weniger natürlich mit dem System.

Kapitel 3
Stand der Technik semantischer Suchmaschinen

Kapitel 3 verschafft einen Überblick über den Stand der Technik der semantischen Suche, der hybriden semantischen Suche sowie des Rankings und der Evaluierung semantischer und hybrider semantischer Suchmaschinen.

Die Kategorien semantischer Suchmaschinen sowie der verwendeten Suchalgorithmen wurden bereits in den theoretischen Grundlagen (Kapitel 2.3.3 und 2.3.4) vorgestellt. Kapitel 3.1 konzentriert sich auf Verfahren die eine besondere Vorgehensweise pflegen und betrachtet, wie formale und informale Inhalte dabei repräsentiert und durchsucht werden. Besonderes Augenmerk liegt auf Verfahren die formale und informale Inhalte in mindestens einer Komponente der Suchmaschine berücksichtigen. Diese Suchmaschinen sind jedoch entweder semantische Dokument- oder Faktensuchmaschinen. Kapitel 3.2 stellt zuerst die Ansätze Richtung hybride semantische Suche vor, die Dokumentretrieval und eine Suche in der formalen Wissensbasis kombinieren, jedoch ohne sowohl Dokument- als auch Faktensuche zu unterstützen. Anschließend werden die Suchmaschinen vorgestellt, die beides anbieten und nach der Definition im Kapitel 2.4 hybride semantische Suchmaschinen sind.

Die Evaluierung der Effektivität und die Rankingstrategien semantischer Suchmaschinen werden in diesem Kapitel gesondert behandelt. Im Gegensatz zum traditionellen Information Retrieval existieren für semantische, insbesondere für hybride semantische Suchmaschinen noch keine standardisierten Evaluierungsmethoden und Rankingverfahren. Während Kapitel 2.5 sich auf die Grundlagen der Evaluierung von IR-Systemen konzentriert, beschreiben Kapitel 3.3 und 3.4 die bestehende Vielfalt der Lösungen für das Ranking und die Evaluierung semantischer und hybrider semantischer Suchmaschinen.

3.1 Semantische Suchmaschinen

Hinsichtlich der These dieser Arbeit sind die Schritte des Suchprozesses semantischer Suchmaschinen von Interesse, in denen formale und informale Inhalte kombiniert werden können. Dies geschieht üblicherweise Online, **im Rahmen der Anfrageinterpretation bzw. Ausführung des Suchalgorithmus**. Es gibt jedoch Verfahren, die bereits **zum Zeitpunkt der Indexierung durch die Art der Abbildung des formalen Wissens** eine Kombination der formalen und informalen Inhalte durchführen (vgl. Abbildung 2.18 im Kapitel 2.3.4).

Semantische Suchmaschinen verarbeiten formale, informale oder hybride Anfragen, wobei verschiedene Möglichkeiten der semantischen **Anfrageinterpretationen** eingesetzt werden. *Formale Anfragen* werden typischerweise in RDF-basierten semantischen Suchmaschinen (z.B. CORESE [Corby et al., 2004] und CORAAL [Novacek et al., 2009]) eingesetzt, wobei sich auch manche der ersten schlüsselwortbasierten semantischen Suchmaschinen einer einfachen Anfragesprache bedienten (z.B. [Lei et al., 2006, Guha et al., 2003], vgl. Kapitel 2.3.3). Solche Anfragen müssen lediglich in eine konkrete Abfragesprache (z.B. SPARQL) übersetzt und ausgeführt werden, es bedarf keiner besonderen

Schritte zur Anfrageinterpretation (vgl. Kapitel 2.3.4).

Die meisten semantischen Suchmaschinen, insbesondere semantikbasierte Schlüsselwortsuchmaschinen (z.B. [Stojanovic et al., 2003], [Guha et al., 2003], [Tummarello et al., 2010], [Stoyanovich et al., 2010]), Frage-Antwort-Systeme (z.B. SmartWeb [Wahlster, 2008], Theseus Alexandria[1], START[2]) und schlüsselwortbasierte Suche mit semantischer Nachverarbeitung (z.B. ALVIS [Buntine et al., 2005]), erlauben *natürlichsprachige Anfragen*. Die Interpretation solcher Anfragen kann auf verschiedene Art und Weise geschehen, wie im Folgenden vorgestellt.

Die semantische Suche kann als die Aufgabe aufgefasst werden, *eine natürlichsprachige Anfrage in eine formale Abfrage zu übersetzen* und diese dann auf einer formalen Wissensbasis auszuführen. Ein Beispiel hierfür ist der Ansatz von [Tran et al., 2009], der Benutzeranfragen in formale konjunktive SPARQL-Abfragen übersetzt. Abbildung 3.1 zeigt ein Beispiel. Hierzu verwendet das System Hintergrundwissen aus der Wissensbasis, bestimmt zuerst aus den eingegebenen Schlüsselwörtern die formalen konjunktiven Abfragen und bietet diese dem Benutzer zur Auswahl an [Tran et al., 2009]. Für die Bestimmung der Abfragen wird die Graphdarstellung der Wissensbasis genutzt, das Verfahren sucht die ersten k zu der Anfrage passenden Teilgraphen und formuliert für jeden Einzelnen die formale Abfrage. Das System erlaubt daher zwar die Suche mit natürlichsprachigen Anfragen, der Benutzer muss jedoch im zweiten Schritt komplexe formale Abfragen interpretieren, um mit der Suche fortfahren zu können. Der Suchansatz wurde in ein Frage-Antwort-System integriert [Tran et al., 2009, Ladwig and Tran, 2010].

Example Keyword Query
2006 cimiano aifb
Example Conjunctive Query
(x, y, z). type(x, Publication) $\wedge$ year(x,2006)
$\wedge$ author(x,y) $\wedge$ name(y, P. Cimiano)
$\wedge$ worksAt(y, z) $\wedge$ name(z, AIFB)
Example SPARQL Query

```
SELECT ?x,?y,?z WHERE {
 ?x type Publication. ?x year 2006.
 ?x author ?y. ?y name 'P. Cimiano'.
 ?y worksAt ?z. ?z name 'AIFB'}
```

Abbildung 3.1: Beispiel Suchanfrage mit der berechneten konjunktiven Anfrage und der SPARQL-Abfrage [Tran et al., 2009]

Zahlreiche weitere Methoden zur Faktensuche und semantischem Dokumentretrieval setzen die Suche als *graphbasierte Anfrageverarbeitung* um (s. auch 2.3.4), bei der die semantische Interpretation der Anfrage auf der Struktur des Graphen basiert. Bekannte Beispiele sind BLINKS [He et al., 2007], XSearch [Cohen et al., 2003] und SemSearch [Lei et al., 2006]. Manche der Ansätze setzen dabei vordefinierte formale Anfragetemplates ein, wie z.B. die tripelbasierte Suchmaschine SemSearch [Lei et al., 2006] und das Frage-Antwort-System von [Sacaleanu et al., 2008]. Sie nutzen implizit die Struktur des Graphen. Den Ausgangspunkt bilden die durch den syntaktischen Vergleich identifizierten Konzepte, die in den Abfragetemplates eingesetzt werden. In welche Richtung die Suche weitergeht, d.h. welche weiteren Abfragen generiert werden, ist durch die gefundenen Tripel bestimmt.

[1] `http://alexandria.neofonie.de` (06.01.2016)
[2] `http://start.csail.mit.edu/index.php` (06.01.2016)

Andere Methoden zur semantischen Anfrageinterpretation setzen *NLP-Verfahren* ein. Sie nutzen Grammatikregeln für die Erkennung von Satzstrukturen und der Rolle der Terme in der Suchanfrage. Beispiele sind die Wikipedia-Suchmaschine Powerset [Converse et al., 2008] und die Frage-Antwort-Maschine Theseus Alexandria. Frage-Antwort-Systeme agieren auf einer detaillierten, mit hohem Aufwand erstellten Wissensbasis, wie es auch bei Powerset und Alexandria der Fall ist.

Suchmaschinen, die eine Kombination von strukturierten und unstrukturierten Inhalten in der Anfrage erlauben, also *hybride Anfragen* unterstützen, sind von besonderem Interesse. Dies ist üblicherweise bei semantischen Faceted Browsern der Fall, da sie häufig neben der Auswahl von Attribut-Wert-Paaren auch die Eingabe einer natürlichsprachigen Anfrage erlauben. Die Schlüsselwörter können für Dokumentsuche aber auch zum Durchsuchen der textuellen Inhalte der Wissensbasis eingesetzt werden. Die Facetten werden als Filter eingesetzt, sie filtern die Ergebnismenge anhand ihrer semantischen Metadaten. Beispiele hierfür sind der im Projekt SIMILE[3] entwickelter Faceted Browser Piggy Bank, das Faceted Wikipedia Search und die Akademische Videosuche Yovisto[4] [Sack, 2010, Waitelonis and Sack, 2010]. Piggy Bank durchsucht die semantischen Metadaten von Webseiten und stellt die Ergebnisse strukturiert dar [Huynh et al., 2005]. Faceted Wikipedia Search liefert den Namen und den Abstrakt der gefundenen Instanzen aus DBpedia [Hahn et al., 2010]. Da DBpedia aus Wikipediaseiten extrahiertes Wissen formal abbildet, handelt es sich hierbei um eine Faktensuche. Yovisto erlaubt facettiertes Browsen inkl. einer Schlüsselwortsuche basierend auf den semantischen Metadaten von Videos und liefert die gefundenen Filme zurück. Es wird also eine Faktensuche durchgeführt, die Videos, also Dokumente[5] als Ergebnis geliefert.

Tabelle 3.1 ordnet die vorgestellten Suchmaschinen entsprechend ihrer Kategorie und nach Art der Suchanfrage ein.

Kategorie	Anfrage		
	formal	informal	hybrid
RDF-basierte Abfragesprache	CORESE, CORAAL		
semantikbasierte Schlüsselwortsuche	[Guha et al., 2003], [Lei et al., 2006]	SIG.MA, [Stoyanovich et al., 2010]	
Frage-Antwort-System		Theseus Alexandria, SmartWeb	
semantische Nachverarbeitung		ALVIS	
Facettierte Suche			PiggyBank, Yovisto, Faceted Wikipedia Search

Tabelle 3.1: Beispiele für Suchmaschinen mit verschiedenen Anfragetypen

[3] `http://simile.mit.edu/` (12.10.2011)
[4] `http://www.yovisto.com` (06.01.2016)
[5] Dokumente bezeichnen nicht nur Textdokumente, sondern auch multimediale Inhalte, Tabellen usw. Diese sind unstrukturierte Inhalte und werden aus der Sicht der Suche als Dokumente betrachtet (s. auch 1.2).

Eine weitere Gruppe der semantischen Suchmaschinen bilden diejenigen Ansätze, die die **formalen und informalen Inhalte bereits zum Zeitpunkt der Indizierung kombinieren**.

Solch ein Ansatz wird in [Ruotsalo, 2012] vorgestellt. Die Grundlage für die Indizierung und das Ranking bilden die in RDF beschriebenen formalen Inhalte, wobei für die Prädikate, Subjekte und Objekte der Tripel einzelne Vektorräume erstellt und mit Hilfe von Mapping-Listen „verbunden“ werden. Durch die kleineren Vektorräume sollen präzisere Ergebnisse erzielt werden. Der Ansatz unterstützt vorrangig Suchanfragen, die Prädikate beinhalten.

In [Exeler et al., 2015] wird ein Ansatz beschrieben, dessen Ziel es ist, den durch Ambiguitäten verursachten Problemen bei der Suche entgegenzuwirken, jedoch ohne den Nachteil der Anfragemodifizierung, die in kleinen Dokumentmengen zu leeren Ergebnismengen führen kann. Hierzu wird eine Entitätenextraktion basierend auf DBpedia durchgeführt und die Dokumente mit den Entitäten annotiert. Es wird ein generalisiertes Vektorraummodell (GVSM) erstellt, das neben den Indextermen auch die IRIs der identifizierten Entitäten beinhaltet. Solch ein Vektorraummodell erweitert das traditionelle Vektorraummodell um das Konzept der Verwandtschaft zwischen zwei Termen [Tsatsaronis and Panagiotopoulou, 2009]. Der in [Exeler et al., 2015] vorgestellte Ansatz betrachtet die Indexterme nicht paarweise orthogonal. Die Termkorrelation wird basierend auf taxonomischen Inhalten berechnet. Die Relevanz einer Entität für ein Dokument basiert auf der Verbundenheit der Entitäten in dem aufgespannten RDF-Graph des Dokumentes. Vorgehensweisen zur Berechnung der Verwandtschaft sowie Verbundenheit werden im Kapitel 3.3.2 vorgestellt.

Einen von den bisherigen Beispielen abweichenden Ansatz verfolgt die Suchmaschine Semplore. Anstatt in der formalen Wissensbasis zu suchen, transformiert sie das „Web der Daten“ (Semantic Web[6]) in einen Textindex. Die Felder im Index repräsentieren vordefinierte Relationen zwischen einzelnen Instanzen und ihren Eigenschaften, die als freier Text bzw. als ein vordefiniertes Schlüsselwort zu interpretieren sind und nicht als ein Konzept, wie z.B. eine Person [Zhang et al., 2008, Wang et al., 2009]. Auf diesem Index basierend erlaubt Semplore eine facettierte Suche nach Konzepten und das Ausführen von hybriden Anfragen. Im Gegensatz zu den vorherigen Verfahren zum semantischen Dokumentretrieval handelt es sich hierbei um eine Faktensuche, die jedoch auf einem Textindex erfolgt. Durch die Art der Abbildung werden Konzepte gefunden, Relationen zwischen den Konzepten lassen sich alleine durch den Index nicht verfolgen.

Tabelle 3.2 fasst die vorgestellten Lösungen nach Suchansatz zusammen.

Bezüglich Ergebnistyp versus Suchart verfolgt nicht nur Semplore eine ungewöhnliche Vorgehensweise. Es gibt semantische Suchmaschinen, die *Dokumente als Ergebnisse liefern, jedoch eine Faktensuche durchführen.* Dies ist in manchen semantischen Wikis[7], wie z.B. im Semantic MediaWiki[8], und in Suchmaschinen, die auf mit RDFa und Mikroformaten[9] annotierte Dokumente ausgerichtet sind, der Fall. Dokumentsuche geschieht dabei durch das Durchsuchen der formalen semantischen Annotationen (Metadaten) in den Do-

[6] `http://www.w3.org/standards/semanticweb/data` (06.01.2016)
[7] Mit semantischen Metadaten angereicherte Wikis.
[8] `https://semantic-mediawiki.org` (06.01.2016)
[9] `dbpedia.org/`, `www.w3.org/TR/xhtml-rdfa-primer`, `microformats.org/` (06.01.2016)

Suchansatz		
graphbasiert	**angepasster Textindex und IR**	**NLP**
[Tran et al., 2009], [Ladwig and Tran, 2010], [Sacaleanu et al., 2008], BLINKS, Xsearch, SemSearch	Semplore, [Ruotsalo, 2012], [Exeler et al., 2015]	Theseus Alexandria, Powerset

Tabelle 3.2: Suchmaschinenbeispiele für die verschiedenen Suchansätze

kumenten. Googles strukturierte Suchergebnisse, die sogenannten Rich Snippets[10] und Googles Rezeptsuche[11] basieren ebenfalls auf den semantischen Annotationen in Webseiten.

3.2 Hybride semantische Suchmaschinen

Die im Kapitel 3.1 beschriebenen Suchlösungen konzentrieren sich entweder auf Dokumentsuche oder auf Faktensuche. Die Verfahren, die einen Mittelweg bieten oder beides leisten, werden im Folgenden vorgestellt. Für einen einfacheren Vergleich gibt Tabelle 3.3 einen Überblick ihrer wichtigsten Eigenschaften.

Frühe Ansätze in Richtung hybrider Suchverfahren kombinieren traditionelles Dokumentretrieval und die Suche in der formalen Wissensbasis, jedoch ohne sowohl Dokument- als auch Faktensuche zu unterstützen.

Die Suchmaschine *TAP* führt eine Schlüsselwortsuche auf dem Dokumentindex sowie eine „ontologiebasierte“ Suche, d.h. Suche auf der Wissensbasis[12] aus [Guha et al., 2003]. Sie erlaubt den Benutzern schlüsselwortbasierte Dokumentsuche, reichert jedoch die Ergebnisseite mit Fakten über die gefundene Entität an. Sucht man beispielsweise nach „Eric Miller“, so werden die gefundenen Dokumente aufgelistet und auf der rechten Seite eine Box mit den Fakten zu seiner Person angezeigt. Die Menge der Fakten ist determiniert als der Subgraph vorgegebener Pfadlänge um die Instanz „Eric Miller“ herum [Guha et al., 2003]. Die Dokumentsuche und die Suche in der formalen Wissensbasis werden voneinander unabhängig durchgeführt. Die Suchmaschine konzentriert sich auf Suchanfragen nach einer Entität. Da die Fakten von der Dokumentliste entkoppelt sind, stoßen Suchanfragen mit mehreren Entitäten sowie mehrdeutigen Suchanfragen an die Grenze der Darstellungsmöglichkeiten. Fakten zu mehreren Entitäten könnten zwar noch untereinander abgebildet werden, der Zusammenhang zwischen ihnen ginge jedoch verloren. Komplexere Graphen, die z.B. die Fakten über sowie die Verbindung unter mehreren Entitäten beinhalten, lassen sich in der Form nicht mehr übersichtlich darstellen.

[10] `http://www.google.com/support/webmasters/bin/answer.py?answer=99170` (06.01.2016)

[11] `http://ebiquity.umbc.edu/blogger/2011/02/26/google-recipe-search-exploits-semantic-web-data-in-rdfa/` (06.01.2016)

[12] Die Suche auf der formalen Wissensbasis wird häufig „ontologiebasierte“ Suche genannt, wobei die Ontologie und die Instanzbasis durchsucht wird [Guha et al., 2003, Rocha et al., 2004, Kiryakov et al., 2004, Bhagdev et al., 2008].

Rocha und Kollegen stellen in *[Rocha et al., 2004]* ein Verfahren vor, das eine schlüsselwortbasierte semantische Dokumentsuche erlaubt, die Webseiten sowohl in einem Dokumentindex abbildet als auch in einer formalen Wissensbasis instanziiert. Es wird zuerst eine traditionelle Schlüsselwortsuche in dem Index durchgeführt (syntaktischer Abgleich) gefolgt vom Spreading Activation (SA, s. Seite 30) auf dem semantischen Netz der Instanzbasis (semantischer Abgleich). Die Ergebnismenge der Suche bilden die aktivierten Webseiten-Instanzen, die nach Kategorien geordnet präsentiert werden. Eine Suche nach Konzepten, die keine Webseiten repräsentieren, oder nach Fakten, wird nicht unterstützt. Die Suchmaschine nutzt also die Verlinkung unter den Webseiten für semantische Dokumentsuche.

CORAAL bietet Faktensuche mit einer RDF-basierten Anfragesprache auf der Wissensbasis und alternativ eine Schlüsselwortsuche im Dokumentindex an. Die Ergebnisse sind dementsprechend entweder Fakten oder Dokumente [Novacek et al., 2009].

Die semantische Dokumentsuchmaschine *KIM* bietet eine Schlüsselwortsuche auf dem Dokumentindex oder, als Alternative, eine „ontologiebasierte“ Suche in der formalen Wissensbasis an. D.h. die Suchanfrage ist entweder natürlichsprachig (Schlüsselwortsuche) und es wird auf dem Dokumentindex gesucht, oder sie ist formal (formularbasiert realisiert) und die Wissensbasis wird durchsucht. Die Suche in der Wissensbasis dient ebenfalls zur Dokumentsuche, als Ergebnis bekommen die Benutzer nicht die Instanzen, sondern die damit annotierten Dokumente. Eine weitere Besonderheit ist, dass als Indexterme des Dokumentindexes die Instanzen der formalen Wissensbasis verwendet werden. Der Index repräsentiert also, welche Instanzen in welchem Dokument vorkommen und wie viel sie zum Inhalt des Dokumentes beitragen. Zusammengefasst bildet die semantische Dokumentsuchmaschine KIM die semantischen Metadaten der Dokumente sowohl formal als auch informal ab und unterstützt sowohl formale als auch natürlichsprachige Suchanfragen, jedoch jeweils voneinander unabhängig [Kiryakov et al., 2004].

Ansätze, die den Begriff hybride semantische Suche verwenden, aber semantisches Dokumentretrieval durchführen, sind HyKSS, Mímir und HS^3.

HyKSS durchsucht mit derselben Suchanfrage sowohl den Dokumentindex, als auch die formale Wissensbasis, wobei die Dokumente mit den Instanzen annotiert sind. Die Ergebnisse der Faktensuche dienen nach [Zitzelberger et al., 2014] dazu das Dokumentranking zu verbessern. Die gefundenen Instanzen werden als Filter betrachtet bzw. aus den gefundenen Fakten Bedingungen abgeleitet, die die Dokumente erfüllen sollen [Zitzelberger et al., 2014].

Mímir ist ein Framework für semantisches Dokumentretrieval, das hybride Suchanfragen bestehend aus Schlüsselwörtern und formalen Metadaten unterstützt. Die strukturierten Inhalte dienen auch in diesem Fall als Filter und werden über eine formularbasierte Benutzerschnittstelle eingegeben [Tablan et al., 2015].

Die Suchmaschine HS^3 verfolgt eine von den bisherigen Beispielen abweichende Strategie. Sie durchsucht neben der formalen Wissensbasis und dem Dokumentindex auch sogenannte Index Graphen. Diese Graphen bilden die semantischen Annotationen der Dokumente als eine Menge von Tupeln ab und sind sowohl mit dem Konzept als auch mit dem Dokument verknüpft. Die Suchmaschine nutzt die Index Graphen, um auch diejenigen Dokumente zu finden, die nicht alle Suchterme beinhalten, jedoch mit den durch die Terme gefundenen Konzepten verknüpft sind. Die Benutzerschnittstelle folgt dem Prinzip der facettierten Suche: je nach Auswahl werden passende Konzepte und Relationen angeboten [Gärtner et al., 2014].

Diese Suchansätze fallen in die Kategorie semantisches Dokumentretrieval (vgl. Kapitel 2.3.3). Sie erweitern traditionelle Schlüsselwortsuche mit semantischen Technologien. [Rocha et al., 2004] repräsentieren zwar die Dokumente als Instanzen in der formalen Wissensbasis (die Dokumente sind instanziiert), gesucht wird jedoch nach Dokumenten (Webseiten) und nicht nach Konzepten oder Fakten.

Im Sinne der Definition im Kapitel 2.4 auf Seite 34 sind PowerAqua, K-Search und CE^2 **hybride semantische Suchmaschinen**.

PowerAqua[13] [Lopez et al., 2006] erweitert die semantische Frage-Antwort-Maschine AquaLog [Lopez et al., 2005] um eine Komponente für semantisches bzw. traditionelles Dokumentretrieval. Der Dokumentindex beinhaltet dabei auch die semantischen Metadaten (die beinhalteten Entitäten) der Dokumente. PowerAqua führt eine Faktensuche auf der formalen Wissensbasis und zusätzlich eine Schlüsselwortsuche auf dem Dokumentindex durch. Wurden Fakten gefunden, so wird der Dokumentindex mit den Namen der durch die Suchanfrage identifizierten Ressourcen durchsucht. Als Ergebnis liefert die Suchmaschine vorrangig Fakten, listet aber zusätzlich auch die relevanten Dokumente auf. Sollten keine Fakten gefunden werden, so wird traditionelles Dokumentretrieval, also eine Schlüsselwortsuche auf dem Dokumentindex, durchgeführt. Somit können Suchanfragen auch dann beantwortet werden, wenn keine Fakten zu der Frage vorhanden sind [Lopez et al., 2012, Uren et al., 2010, Fernandez et al., 2008].

[Bhagdev et al., 2008] stellen einen hybriden semantischen Suchansatz vor, der traditionelle Schlüsselwortsuche sowie semantische Suche auf den Metadaten der Dokumente erlaubt und unter dem Namen *K-Search* implementiert wurde. Dabei gibt es jedoch keine Kombination der beiden Suchverfahren im Sinne des Nutzens von Teilergebnissen, lediglich die Endergebnisse der Schlüsselwortsuche und der Suche in den Metadaten werden zusammengefasst. Da im Rahmen der Zusammenfassung die Ergebnisse der beiden Sucharten eine Auswirkung auf die Auswahl der Endergebnisse haben, betrachten wir diesen Ansatz als hybride semantische Suche im Sinne der Definition aus Kapitel 2.4. K-Search erlaubt drei Arten von Suchanfragen:

- formal, d.h. konkrete Konzepte, Relationen, Instanzen
- informal, d.h. Schlüsselwörter, natürlichsprachige Suchanfrage
- hybrid[14], d.h. eine Mischung aus formalen und informalen Teilen.

Die formalen Anteile der hybriden Suchanfrage werden als Filter für die formale Wissensbasis angewendet. Danach erfolgt mit den informalen Anteilen der Suchanfrage eine Schlüsselwortsuche in den textuellen Eigenschaften (Literale) der gefilterten Konzepte. Es ist also nur eine Konjunktion der formalen und informalen Anteile der Suchanfragen möglich, wie z.B. die Suche nach Personen mit dem Vornamen „Michael“. Die Anfragen werden neben der Wissensbasis auch auf dem Dokumentindex ausgeführt. Das Zusammenführen der Ergebnisse der Fakten und der Dokumentsuche erfolgt anhand der semantischen Metadaten der Dokumente: Kommen die Konzepte eines gefundenen Tripels in einem Dokument vor, so werden das Dokument und der Triple als zusammengehörend betrachtet. Tripel, zu denen kein Dokument existiert, werden entfernt. Die Ergebnisse können die Benutzer wahlweise als eine Liste der Dokumente oder eine Liste der Fakten

[13] `http://technologies.kmi.open.ac.uk/poweraqua/` (06.01.2016)
[14] Hybride Suchanfragen werden bei [Bhagdev et al., 2008] “Schlüsselwort im Kontext“ genannt.

Suchmaschine	Suchanfrage	Fakten-/Dokumentsuche	Suchergebnis
TAP	informal	beides, jedoch voneinander unabhängig	Dokumente, zusätzlich Fakten zu der gesuchten Entität (unabhängig von den Dokumenten)
[Rocha et al., 2004]	informal	semantische Dokumentsuche	Dokumente (die in der Wissensbasis instanziiert sind)
CORAAL	informal oder formal	Faktensuche oder Dokumentsuche	Fakten oder Dokumente
KIM	informal oder formal	entweder Dokumentsuche oder Faktensuche	Dokumente
HyKSS	informal und formal (wird sowohl als formale, als auch als freier Text betrachtet)	semantische Dokumentsuche und Entitätensuche, voneinander unabhängig	Dokumente (die gefundenen Entitäten dienen zur Verbesserung der Rangordnung)
Mímir	informal, formal oder hybrid (eingeschränkt: formale Anteile werden als Filter eingesetzt)	semantische Dokumentsuche	Dokumente
HS3	formal und hybrid	semantische Dokumentsuche	Dokumente
PowerAqua *	informal	Faktensuche und zusätzlich semantische Dokumentsuche	vorrangig Fakten, zusätzlich relevante Dokumente
K-Search *	informal, formal oder hybrid (eingeschränkt: formale Anteile werden als Filter eingesetzt)	Faktensuche und semantische Dokumentsuche	Dokumente oder Fakten (nach der Suche zusammengeführt und entweder als Dokumentliste oder als Faktenliste präsentiert)
CE2*	informal, formal oder hybrid (eingeschränkt: nur konjunktive Suchanfargen, die sich als ein zusammenhängender Graph in Baumstruktur abbilden lassen)	Faktensuche und Dokumentsuche	Fakten und Dokumente als ein Graph
SINFIO *	*informal, formal und hybrid*	*Faktensuche und semantische Dokumentsuche*	*Fakten, Dokumente und hybrid, also Dokumente mit Fakten*

Tabelle 3.3: Überblick der Ansätze Richtung hybrider Verfahren. Die mit * gekennzeichneten Ansätze sind hybride semantische Suchmaschinen im Sinne der Definition im Kapitel 2.4.

anschauen, wobei für die Faktenliste auch eine Graphdarstellung zur Verfügung steht. Die Suchmaschine wurde mit 18.097 Berichten und 21 Suchanfragen aus der Domäne Raumfahrtdiagnostik evaluiert. Dabei zeigte der hybride Ansatz eine deutliche Verbesserung der Effektivität mit einem F-Maß von 84% gegenüber der Faktensuche mit 54% und der Schlüsselwortsuche mit 57%.

Die hybride semantische Suche CE^2 agiert auf einer Graphrepräsentation, die neben den Konzepten aus der Wissensbasis auch die Dokumente beinhaltet[15] [Wang et al., 2011]. Konzepte und Dokumente sind, wie üblich, anhand der semantischen Annotationen der Dokumente verknüpft. Die Suchanfragen der Benutzer werden in Teilanfragen zerlegt und die Abfragen für die Wissensbasis, den Dokumentindex sowie den davon getrennten Index der semantischen Annotationen der Dokumente konstruiert. Der Schwerpunkt von CE^2 liegt in der Abbildung der Suchanfrage in einer Abfrage der Form eines zusammenhängenden Graphen in Baumstruktur. Die Schlüsselwortsuche in den Dokumenten dient dabei zur Identifikation von Instanzen der Wissensbasis, die Teile der Baumstruktur bilden[16]. Die Suchergebnisse sind die Teilgraphen aus der Graphrepräsentation des Informationsraumes, die den Abfragegraphen erfüllen. Sowohl die Suchanfrage als auch die Ergebnisse können hybrid sein, wobei hybride Antworten als Graphen dargeboten werden (s. Abbildung 3.2). Der Ansatz ist auf konjunktive Suchanfragen eingeschränkt, die einen zusammenhängenden Graphen in Baumstruktur abbilden. Zusammenhänge über mehr als eine Kante zwischen zwei Knoten können nicht gefunden werden.

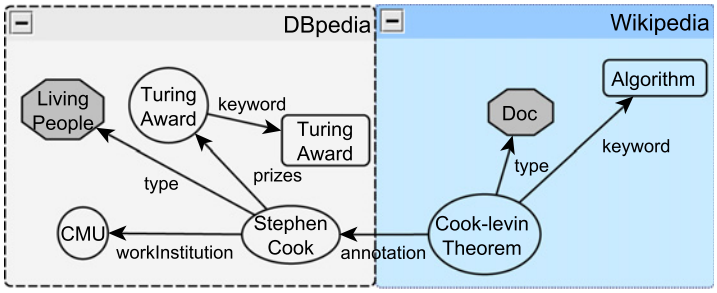


Abbildung 3.2: Hybrides Suchergebnis in CE^2 [Wang et al., 2011]

Manche Suchmaschinen liefern *hybride Ergebnisse, die aus Dokumenten und ihren semantischen Metadaten bestehen.* Sie reichern also die Ergebnisse der semantischen Dokumentsuche mit formalen Daten an. Wie bereits beschrieben stellt TAP [Guha et al., 2003] auf der Ergebnisseite Dokumente und zusätzlich Fakten zu der gefundenen Entität dar. PowerAqua liefert zusätzlich zu den Fakten auch relevante Dokumente, sie werden jedoch auch in diesem Fall voneinander unabhängig gelistet. K-Search reichert gefundene Dokumente mit ihren semantischen Metadaten an, die Oberfläche bietet jedoch entweder eine Dokumenten- oder eine Faktensicht an, beides zusammen wird nicht dargestellt.

Der dieser Arbeit zugrundeliegende Lösungsansatz *SINFIO* geht über die oben vorgestellten Methoden hinaus. SINFIO unterstützt als einzige Suchlösung sowohl Fakten- als auch Dokumentsuche, formale, informale und hybride Suchanfragen sowie Suchergebnisse und macht dies uneingeschränkt für alle möglichen Anfragetypen. Im Gegensatz

[15] CE^2 ist eine Weiterentwicklung von Semplore [Wang et al., 2009].
[16] Eine größere Rolle spielen die Dokumente nur für die Suchanfragen, die explizit nach Dokumenten zu einer Instanz fragen. In solchen Fällen muss lediglich der Annotationsbeziehung gefolgt werden.

zu K-Search setzt SINFIO die formalen Anteile der hybriden Suchanfragen nicht nur als Filter für die formale Wissensbasis ein, um den Suchraum einzuschränken. Sie werden für die Faktensuche eingesetzt und die gefundenen Fakten fließen in die weitere Suche mit ein. Während K-Search zwar die Ergebnisse zusammenführt, dem Benutzer jedoch entweder die gefundenen Dokumente oder die Fakten präsentiert, ist der hier beschriebe Ansatz in der Lage anfragebezogene hybride Ergebnisse, bestehend aus einem Dokument und an dem Dokument anknüpfenden Fakten, zu liefern. Verglichen mit CE^2 schränkt SINFIO die Menge der Suchanfragen nicht ein. Nicht nur konjunktive, sondern auch disjunktive Anfragen sind erlaubt und auch Pfade mit mehr als einer Kante zwischen zwei gefragten Konzepten (daher auch Verknüpfungen zwischen Teilgraphen) werden gefunden. Wie in Kapitel 5 detailliert beschrieben wird, spielen die Dokumente in den hybriden Ergebnissen von SINFIO eine größere Rolle als in CE^2.

3.3 Ranking semantischer und hybrider semantischer Suchmaschinen

Durch die formale Wissensbasis besteht die Möglichkeit das Ranking semantischer Suchmaschinen auf semantische Basis durchzuführen anstatt lexikalische Ähnlichkeiten zu berechnen. Das sogenannte *semantische Ranking* greift dabei auf die Struktur der Wissensbasis zurück [Stojanovic et al., 2001]. Viele semantische Suchmaschinen sehen vom Gebrauch des semantischen Rankings ab. Sie indizieren die Dokumente und/oder die semantischen Metadaten der Dokumente bzw. die Wissensbasis (vgl. Kapitel 3.1) und greifen auf die traditionellen IR-Rankingmethoden zurück (s. Kapitel 2.3.2). Beispiele sind die Suchmaschinen SHOE [Heflin and Hendler, 2000], PiggyBank [Huynh et al., 2005], KIM [Kiryakov et al., 2004], TAP [Guha et al., 2003] und Semplore [Zhang et al., 2008]. Dieses Kapitel konzentriert sich auf die Methoden des semantischen Rankings und gibt einen Überblick über die Strategien für die Faktensuche (3.3.1), die semantische Dokumentsuche (3.3.2) und die hybride semantische Suche (3.3.3). Liegt eine Evaluierung der Rankingverfahren vor, so wird sie ebenfalls beschrieben.

3.3.1 Ranking von Fakten

Das **Ranking von Ressourcen** (auch entity ranking genannt) wird für das sogenannte Entity oder Data Retrieval, also die Suche nach Konzepten in einer formalen Wissensbasis, benötigt. Dies bildet wiederum einen Baustein für das **Ranking von Fakten**. Wie im Kapitel 2.3.3 definiert, wird Data Retrieval als ein Fall der Faktensuche angesehen und nicht als eine getrennte Kategorie betrachtet.

Für das **Ranking von Ressourcen** wurden beispielsweise vektorraumbasierte Verfahren entwickelt, Heuristiken eingesetzt aber auch Ranking-Algorithmen für nicht semantische Suchmaschinen, wie z.B. Googles PageRank, adaptiert.

Corese, eine Suchmaschine mit RDF-basierter Anfragesprache, setzt ein *adaptiertes Vektorraummodell* ein. Der Ansatz konzentriert sich auf die Suche nach Konzepten mit formalen Anfragen und indiziert die Wissensbasis unter Einsatz einer adaptierten *tfidf*-Gewichtung. Für jedes Property wird ein eigener Vektorraum erstellt, die Zeilen und Spalten der Matrix bilden die Instanzen. Für die Berechnung der Ähnlichkeit der Suchanfrage zu den Tripeln setzt Corese das Kosinus-Maß ein, wobei auch dies für die Suche in

mehreren Vektorräumen angepasst wurde [Ruotsalo, 2012]. Implizit wird also die Struktur des Graphen genutzt, da die Indexierung und das Ranking auf Basis der vorhandenen Tripel geschieht.

Für die semantische Videosuchmaschine Yovisto wurde ein *auf Heuristiken basierendes Rankingverfahren* entwickelt, mit dessen Hilfe die verwandten Entitäten einer gegebenen Entität bestimmt und ihre „semantische Verwandtschaft“ gewichtet werden kann [Waitelonis and Sack, 2012]. Das System basiert auf das Ranking der Properties einer Entität, der Rang einer Entität wird dann als die Summe der Propertygewichte berechnet. Als formale Wissensbasis wurde DBpedia verwendet. Die eingesetzten Heuristiken betrachten:

- die Häufigkeit der Verwendung einer Property,
- die Tatsache, dass eine Property häufig Instanzen derselben Klasse verbindet,
- Properties zu Veranstaltungen und Orte, da diesen eine wichtige Bedeutung im Bezug auf die Inhalte der Suchmaschine zugeordnet wird,
- „duale“ Properties, d.h. Propertypaare, die zwei Instanzen in beide Richtungen verbinden (eine Verallgemeinerung von inversen Properties),
- Wikilinks (nicht qualifizierte Links zwischen Entitäten) ,
- Properties zu Listen, Kategorien und Ontologien.

Das Entitätenranking geschieht offline über die gesamte Wissensbasis. Die verwandten Entitäten werden den Benutzern zur Unterstützung der explorativen Suche dargeboten [Waitelonis and Sack, 2012].

[Hogan et al., 2006] adaptierten Googles *PageRank* für die Suche im Semantischen Web. Der PageRank-Algorithmus misst die relative Wichtigkeit von Webseiten basierend auf ihrer Verlinkung und bietet sich für das Ranking in denjenigen formalen Wissensbasen an, die sich ebenfalls als Graph abbilden lassen. Das Grundprinzip von PageRank ist: Je mehr eingehende Verbindungen auf einen Knoten verweisen, umso höher ist das Gewicht des Knotens und je höher das Gewicht der verweisenden Knoten ist, umso größer ist sein Einfluss auf das Gewicht des Knotens, auf den er verweist [Lawrence et al., 1999]. [Hogan et al., 2006] beschreiben ein Modell für das Ranking von einzelnen Ressourcen und Graphen (Menge von zusammenhängenden Fakten) aus dem Semantischen Web. Der Algorithmus analysiert aus Effizienzgründen nur die Suchergebnisse, liefert also ein lokales Ranking der Ergebnisse einer Suchanfrage, während der original PageRank-Algorithmus auf den gesamten Graph angewendet wird und ein globales Ranking bietet [Tang and Dwarkadas, 2004]. Als Grundlage für das lokale Ranking werden die Teilgraphen um diejenigen Ressourcen herum herangezogen, die durch den syntaktischen Abgleich gefunden worden sind. Der Teilgraph beinhaltet Pfade bis zur Länge n von der Ressource aus, wobei sowohl ausgehende als auch eingehende Verbindungen betrachtet werden. Somit werden komplette Teilgraphen gegebener Pfadlänge berücksichtigt und mithilfe des PageRank-Algorithmus der Rang der Ressource berechnet. Eine Erweiterung des Algorithmus erlaubt das Ranking eines RDF-Graphen unter Berücksichtigung der Herkunft als einen zusätzlicher Faktor. So können auch Graphen aus verschiedenen Quellen entsprechend dem Vertrauen, das der Quelle entgegengebracht wird, gerankt werden.

Wie folgende Beispiele zeigen verfolgen Rankingverfahren für **Fakt-Ranking** unterschiedliche Strategien um den Rang eines Ergebnisses zu berechnen. Dabei werden Konzepte, wie semantische Ähnlichkeit, semantische Assoziation und die Abschätzung des

Informationsgehaltes eines Ergebnisses betrachtet.

Im EU-Projekt SEmantic portAL, kurz SEAL[17], wurde ein Framework für die Realisierung semantischer Portale und für die formale Evaluierung von semantischen Technologien entwickelt. Die Architektur beinhaltet neben der Wissensbasis auch eine Komponente für Reasoning (s. Kapitel 31): die Ontobroker[18] Inferenzmaschine. Dokumentrepositories oder Textindizies sind nicht vorgesehen, denn SEAL berücksichtigt nur Faktensuche [Stojanovic et al., 2001]. Das Framework enthält eine Komponente, die Suchergebnisse basierend auf ihrer *semantischen Ähnlichkeit* zueinander in Rangordnung bringt. Um die semantische Ähnlichkeit zweier Konzepte zu bestimmen, wird die relative Position des Konzeptes in einer Hierarchie (Taxonomie, Kategoriensystem) auf der Basis der Semantic Cotopy betrachtet [Stojanovic et al., 2001]. Semantic Cotopy sammelt zu einem Konzept einer Wissensbasis alle weiteren Konzepte aus der selben Hierarchie und zwar sowohl nach oben zu generelleren Konzepten als auch nach unten zu den spezielleren Konzepten [Maedche and Staab, 2002]. Für die semantische Ähnlichkeit wird die Cotopy aufwärts (upwards cotopy, UC) der zu vergleichenden Objekte O_1, O_2 betrachtet. Die semantische Ähnlichkeit (object match, OM) wird dann aus der Anzahl Elemente in der Schnittmenge und Vereinigung der Cotopien berechnet:

$$OM(O_1, O_2) = \frac{\mid UC(O_1) \cap UC(O_2) \mid}{\mid UC(O_1) \cup UC(O_2) \mid}. \tag{3.1}$$

Die Vorgehensweise kann analog für Relationen eingesetzt werden, dabei wird die Hierarchie der Properties betrachtet. Die Generalisierung für konkrete Tripel (Fakten) T_1, T_2 ist das geometrische Mittel der semantischen Ähnlichkeiten des Subjektes, Prädikats und Objektes von T_1, T_2. Für mehrere Tripel wird der Durchschnitt berechnet [Stojanovic et al., 2001]. Der Rang eines Suchergebnisses ist dessen semantische Ähnlichkeit mit dem entsprechenden Subgraph aus der Wissensbasis, der benötigt wurde, um das Suchergebnis ableiten zu können. D.h., wenn ein Ergebnis besagt, dass Person X am Projekt P arbeitet, und Projekt P dem Topic T angehört, T jedoch durch die Transitivität der subTopicOf-Beziehung in zwei Schritten von T' abgeleitet wurde, so ist der Graph in der Wissensbasis ungleich dem Ergebnisgraphen. Er ist um die zwei subTopicOf-Relationen und zugehörigen Topics zwischen T' und T größer. Diese Definition der semantischen Ähnlichkeit zwischen Tripeln bewirkt, dass das Ergebnis umso höher gerankt wird, je weniger Inferenzschritte für das Ableiten des Tripels benötigt werden. Entspricht ein Ergebnis exakt einem Subgraphen der Wissensbasis, bekommt es den Rang 1. Ein Ergebnistripel jedoch, das einen größeren Subgraphen aufspannt, weil z.B. das Objekt im Ergebnistripel eine Unterkategorie des Gefragten ist, hat einen Rang < 1. Eine Evaluierung der Rankingmethode liegt nicht vor.

Weitere Ansätze zu Fakt-Ranking basieren auf dem Konzept der *semantischen Assoziation*. Die semantische Assoziation zweier Ressourcen wird anhand von Property-Sequenzen zwischen zwei Konzepten in der Wissensbasis bestimmt, wobei die in Abbildung 3.3 skizzierten drei Arten unterschieden werden:

- Direkter Pfad zwischen den Ressourcen r_1 und r_2 (möglicherweise über mehrere Ressourcen hinweg);
- Es besteht ein gleich langer, direkter Pfad zwischen r_1, r_n und r_2, r_m, so dass die Properties in jedem einzelnen Schritt dazwischen gleich sind oder in $< rdfs$:

[17] `http://www.seal-project.eu` (06.01.2016)
[18] `http://www.semafora-systems.com/de/produkte/ontobroker/` (06.01.2016)

$subPropertyOf >$-Relation zueinander stehen;

- r_1 und r_2 haben jeweils einen direkten Pfad zu einer Ressource r_m [Anyanwu and Sheth, 2003, Anyanwu et al., 2005].

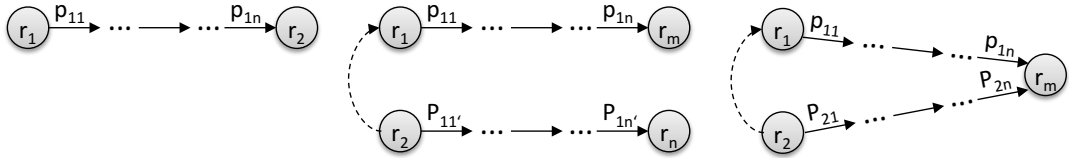


Abbildung 3.3: Semantische Assoziationen [Anyanwu et al., 2005]

Das *SemRank „Relevanz-Modell"* erwartet die Angabe von zwei Ressourcen als Input und gibt die nach Informationsgehalt gerankten Pfade zwischen diesen zurück [Anyanwu et al., 2005]. Das Modell greift neben der semantischen Assoziation auf informationstheoretische Grundlagen zurück und berechnet den Informationsgehalt eines Ereignisses auf Basis der Wahrscheinlichkeit, mit der das Ereignis auftritt. Ziel ist es abzuschätzen, wieviel Informationsgewinn das Ergebnis dem Benutzer bringt. Dies ist nach den Autoren eng mit der Wahrscheinlichkeit verbunden, dass der Benutzer die Existenz der Assoziation vermutet. Es soll also die Vorhersagbarkeit der Assoziation für das Ranking herangezogen werden. Je mehr ein Ergebnis vorhersagbar ist, umso höher wird es gerankt, wobei häufig vorkommende Relationen bevorzugt werden[19]. Zusätzlich betrachtet das Modell die Diskrepanz des Pfades im Vergleich zu der Ontologie: Ist ein Pfad die Instanziierung eines Subgraphen aus der Ontologie, so gibt es keine Diskrepanz. Sind jedoch Teile der Pfade nicht als Subgraph in der Ontologie wiederzufinden, so besteht eine Diskrepanz. Dies kommt dann vor, wenn Ressourcen auf dem Pfad mehrfach klassifiziert sind und mehrere Rollen in einem Ergebnis aufnehmen (z.B. eine Organisation, die als Organisation der Kunde einer anderen Organisation ist). Die Suchmaschine erlaubt es die Anfrage mit Schlüsselwörtern zu erweitern. Können durch den syntaktischen Vergleich weitere Ressourcen oder Properties (hier wird auch die Hierarchie anhand $\langle rdfs : subPropertyOf \rangle$ berücksichtigt) auf einem der Pfade gefunden werden, so wird der Rang der semantischen Assoziation erhöht.

[Aleman-Meza et al., 2005] entwickelten ein ähnliches Ranking-Modell, was ebenfalls auf semantischen Assoziationen basiert und die Suche nach Pfaden zwischen zwei Ressourcen erlaubt. Der Unterschied zu dem Ansatz in [Anyanwu et al., 2005] liegt darin, dass hier nicht nur die Properties auf den Pfaden eine Rolle spielen. Das Modell bezieht auch sogenannte Regionen mit ein, die aus einer Menge von Klassen und Properties der Ontologie bestehen und den „Kontext" der Suchanfrage definieren. Die Regionen werden von den Benutzern bestimmt. Bei einer Suche nach der Verbindung zwischen zwei Wissenschaftlern könnte der Benutzer einen Teil der Ontologie auswählen, der Projekte und Forschungsthemen modelliert. Neben der Spezifität der Pfade wird auch die Popularität einer Ressource und das Vertrauen (trust) mit einbezogen. Die Popularität wird anhand der Anzahl ein- und ausgehender Pfade berechnet. Je stärker ein Knoten vernetzt ist, umso populärer ist es. Das Vertrauen ist für einzelne Datenquellen definierbar und sagt aus, in welchem Maß davon ausgegangen werden kann, dass die Inhalte richtig sind. Das System wurde mit fünf Suchanfragen getestet und mit dem manuellen Ranking der ers-

[19] Basierend auf diesem Modell wird auch eine „Discovery-Mode" angeboten, bei dem für das Ranking gerade das Gegenteil angenommen wird und die am wenigsten vorhersagbaren Ergebnisse am höchsten Gewichtet werden.

ten 50 Ergebnisse durch fünf Benutzer verglichen. Für drei der fünf Suchanfragen rankte das System die Antwort am höchsten, die auch von den Benutzern als die Relevanteste beurteilt wurde. Für die restlichen zwei Suchanfragen war das von den Benutzern am höchsten gerankte Ergebnis unter den ersten fünf Suchergebnissen.

Die bisher vorgestellten semantischen Rankingverfahren berücksichtigen keine Unsicherheiten aus dem syntaktischen Vergleich, sie behandeln alle gefundenen Konzepte gleich. Es gibt jedoch graphbasierte Verfahren, die auch die **lexikalische Ähnlichkeit** der Suchterme zu den Konzepten als Grundlage für die Berechnung einsetzen.

Die Suchmaschine *BLINKS* führt die Faktensuche aus, indem sie den zusammenhängenden Graphen sucht, der alle Schlüsselwörter der Suchanfrage abdeckt. Die Rankingfunktion berechnet ein Gewicht für den Wurzelknoten des Graphen, das letztendlich als Rang des Graphen fungiert. Dieser Wert wird als eine Kombination von:

- der lexikalischen Ähnlichkeit der Schlüsselwörter zu den Labeln gefundener Konzepte,
- der Pfadlängen zwischen diesen Knoten zum Wurzelknoten und
- ähnlich zu PageRank, der Konnektivität, d.h. die Anzahl eingehender und ausgehender Pfade, der einzelnen Knoten berechnet.

Es werden also die lexikalische Ähnlichkeit, die Distanz der Knoten, die Größe und die Struktur des Graphen mit einbezogen [He et al., 2007].

Das graphbasierte semantische Rankingverfahren in *[Tran et al., 2009]* verfolgt einen ähnlichen Ansatz. Neben der lexikalischen Ähnlichkeit und der Pfadlänge wird auch die Popularität mit einbezogen, wobei die Popularität sich negativ auf den Rang des jeweiligen Knotens auswirkt und in diesem Fall die Häufigkeit der Anwendung einer bestimmten Relation zwischen den Instanzen bestimmter Klassen ausdrückt.

Die Suchmaschine *Semplore* führt eine Faktensuche aus, transformiert hierzu jedoch die formale Wissensbasis in einen Index. Die Felder im Index repräsentieren vordefinierte hierarchische Relationen zwischen Instanzen und ihren textuellen Eigenschaften [Zhang et al., 2008, Wang et al., 2009]. Auf dem Index basierend erlaubt Semplore eine facettierte Suche und das Ausführen hybrider Anfragen. Hybrid sind die Anfragen, deren Struktur formal angegeben ist, die einzelnen Elemente, wie z.B. die gesuchte Instanz, jedoch als freier Text angegeben werden können. Eine Suchanfrage ist ein gerichteter Graph, dessen Knoten entweder formal oder durch Schlüsselwörter benannt sind. Dabei gibt es eine Zielvariable, nach dessen Wert der Benutzer sucht. Beispielsweise gibt die Suchanfrage *„action films“* $\langle directedBy \rangle$ $\langle HongKongFilmDirector \rangle$ *and* $\langle starring \rangle$ *(*$\langle ChineseActor \rangle$ *and „martial")* einen Graphen an, wobei die eckigen Klammern die Konzepte der Wissensbasis und die Wörter in Anführungszeichen die natürlichsprachigen Anteile der Suchanfrage andeuten. *(*$\langle ChineseActor \rangle$ *and „martial")* gibt die Klasse und den Wert des gesuchten Knotens an [Wang et al., 2009]. Die frühe Version der Suchmaschine, in der das Ranking auf traditionellen IR-Modellen basierte [Zhang et al., 2008], wurde um ein relationsbasiertes Rankingverfahren erweitert, um gemischte Anfragen zu unterstützen [Wang et al., 2009]. Ausgangsbasis für den Rang eines Knotens ist das Gewicht, das die Schlüsselwortsuche auf dem Index liefert. Gesucht wird ein Graph, der der Struktur des durch die Query aufgespannten Graphen entspricht und die gefundenen Knoten beinhaltet. Um die Nähe der Konzepte zu berücksichtigen werden die Gewichte zu den benachbarten Knoten propagiert. Zudem fließt in das Ranking die Konnektivität der Knoten mit ein: Je mehr Nachbarn ein Knoten hat, umso höher wird er gerankt. Der

Rang des Knotens, der die Zielvariable repräsentiert (im obigen Beispiel „action films“) ergibt sich durch die Propagation der Gewichte unter Einbezug der oben genannten Faktoren, und wird während des Suchprozesses berechnet [Wang et al., 2009].

Die Ähnlichkeit der Suchwörter und der gefundenen Konzepte auf lexikalischer Ebene ist die Basis für das Ranking der Faktensuchmaschine *SemSearch*. Es handelt sich hierbei nicht um semantisches Ranking nach der Definition aus [Stojanovic et al., 2001], da die Ergebnisse nicht nach der Struktur der Wissensbasis gewichtet werden. Das Verfahren wird trotzdem vorgestellt, weil die Gewichtung in Bezug auf die **„Abdeckung“ der Suchanfrage** geschieht. SemSearch berechnet den Rang eines Ergebnisses auf Basis zweier Faktoren: Der lexikalischen Distanz zwischen einem Schlüsselwort und dem Label des gefundenen Konzeptes und der Anzahl im Ergebnis vertretenen Suchwörter [Lei et al., 2006][20].

Tabelle 3.4 gibt einen Überblick der vorgestellten Verfahren mit ihren Besonderheiten. Die Tabelle beinhaltet auch das Rankingverfahren in SemSearch, das kein semantisches Ranking ist jedoch die Abdeckung der Suchanfrage betrachtet, was in allen anderen Konzepten nicht explizit geschieht.

3.3.2 Ranking für semantisches Dokumentretrieval

Für das semantische Ranking in semantischen Dokumentsuchmaschinen werden ähnliche Aspekte herangezogen wie für das Faktenranking, da auch hier die Struktur der Wissensbasis mit einbezogen wird. Es werden angepasste IR-Modelle sowie Verfahren, die die Größe (Pfadlänge bzw. Distanz zwischen Knoten) sowie die Struktur (z.B. hierarchische Relationen, Klassenhierarchien) des Graphen nutzen, eingesetzt. Da es viele semantische Dokumentsuchmaschinen und daher zahlreiche Abwandlungen dessen gibt, wie die genannten Aspekte in das Ranking einfließen, werden im Folgenden je ein charakteristisches Beispiel sowie besondere Verfahren, die einen weniger verbreiteten Ansatz verfolgen, beschrieben.

[Vallet et al., 2005] adaptierten das **Vektorraummodell mit angepasster tfidf-Gewichtung** für semantisches Dokumentretrieval. Hierfür werden die semantischen Metadaten der Dokumente indexiert. Die Instanzen der Wissensbasis bilden die Menge der Indexterme. Das Gewicht eines Dokumentes zu einem Indexterm t ist die Anzahl der Vorkommen von t, geteilt durch die Anzahl der Vorkommen des am häufigsten vorkommenden Indexterms. Es wird also die Wichtigkeit des Konzeptes im Dokument abgeschätzt. Zur Normalisierung wird das Ergebnis noch mit $log(Anzahl\ Dokumente/Anzahl\ Dokumente\ mit\ t)$ multipliziert. Die Suchmaschine verarbeitet formale Anfragen, dabei bestimmt sie zuerst die passenden Instanzen und führt danach die Suche auf dem Index aus. Als Ähnlichkeitsmaß wird das Kosinus-Maß in einer leicht veränderten Form eingesetzt, mit dem Ziel auch dann größere Ähnlichkeitswerte zu bekommen, wenn ein Dokument eine Instanz im Vergleich zu anderen Instanzen nur wenige Male beinhaltet. Dies wirkt der Gewichtung in der Dokument-Instanz-Matrix entgegen.

[Khan et al., 2004] stellen ein Rankingverfahren vor, das eine **im Bezug auf die Such-**

[20] Die genaue Berechnungsvorschrift ist nicht angegeben.

	Suchmaschine bzw. Rankingverfahren	Ranking basiert auf	Kurzbeschreibung
Ressourcen-ranking	Corese [Ruotsalo, 2012]	Adaptiertes Vektorraummodell	Erstellt für jede Property einen eigenen Vektorraum und passt sowohl die tfidf-Gewichtung als auch das Kosinus-Maß für den Einsatz über mehrere Indizies an.
	Yovisto [Waitelonis and Sack, 2012]	Heuristiken	Gewichtet die "semantische Verwandtschaft" zwischen Entitäten, wobei verschiedene Aspekte von Properties betrachtet werden.
	[Hogan et al., 2006]	PageRank	Adaptiert PageRank und ist auch zum Gewichten von Graphen einsetzbar.
Fakt-Ranking	[Stojanovic et al., 2001]	Graphstruktur - semantische Ähnlichkeit	Für hierarchische Wissensbasen, wie z.B. Taxonomien, entwickelt. Semantische Ähnlichkeit wird basierend auf die Position des Konzeptes in der Hierarchie berechnet. Der Rang eines Tripels ist der geometrische Mittel der Gewichte des Subjektes, Prädikates und Objektes.
	[Anyanwu and Sheth, 2003]	Graphstruktur - semantische Assoziation	Die semantische Assoziation zweier Konzepte wird anhand der Reihe und Anordnung von Properties zwischen ihnen berechnet.
	SemRank [Anyanwu et al., 2005]	Graphstruktur - semantische Assoziation	Erweitert das Konzept der semantischen Assoziation mit der Abschätzung der Vorhersagbarkeit einer Assoziation sowie der Diskrepanz des Pfades im Vergleich zu der Ontologie.
	[Aleman-Meza et al., 2005]	Graphstruktur - semantische Assoziation	Erweitert das Konzept der semantischen Assoziation um den ontologischen Kontext der Suchanfrage und betrachtet auch die Popularität der Konzepte.
	BLINKS [He et al., 2007]	Lexikalische Ähnlichkeit und Graphstruktur	Die Gewichtung geschieht auf Basis der lexikalischen Ähnlichkeit der Suchwörter zu den Konzepten, Pfadlängen und der Konnektivität der Konzepte.
	[Tran et al., 2009]	Lexikalische Ähnlichkeit und Graphstruktur	Die Gewichtung geschieht auf Basis der lexikalischen Ähnlichkeit der Suchwörter zu den Konzepten, Pfadlängen und der Popularität der Konzepte.
	Semplore [Zhang et al., 2008, Wang et al., 2009]	Lexikalische Ähnlichkeit und Graphstruktur	Propagiert die Gewichte aus der Berechnung der lexikalischen Ähnlichkeit entlang den Pfaden und beachtet dabei auch die Konnektivität der Knoten.
	SemSearch [Lei et al., 2006]	*kein semantisches Ranking*	*Neben der lexikalischen Ähnlichkeit wird auch die Abdeckung der Suchanfrage pro Suchergebnis betrachtet.*

Tabelle 3.4: Überblick der Verfahren zu Ressourcen- und Fakt-Ranking

anfrage adaptierte Termhäufigkeitsgewichtung mit Distanz und semantischer Korrelation kombiniert. Die vorgestellte Audio-Suchmaschine wendet Texterkennung und anschließend ontologiebasierte Informationsextraktionsverfahren auf Audioinhalte an, um ihre semantischen Metadaten zu bestimmen. Die Wissensbasis beinhaltet die Domänenontologie (z.B. Sport) und die Instanzen (z.B. Teams) sowie Synonymlisten der Konzepte. Die semantischen Metadaten der Audiodateien werden für die Suche in einem Index abgelegt, wobei die Menge der Indexterme durch die Wissensbasis bestimmt wird, da sie aus den Labeln der Konzepte (Ontologie und Instanzen) gebildet wird. Die Suchmaschine führt den syntaktischen Abgleich mit den Suchtermen durch, während dessen auch die Synonyme berücksichtigt werden. Das Gewicht eines Konzeptes ist die Anzahl der Frageterme inkl. Synonyme, durch die es gefunden wurde, geteilt durch die Anzahl aller Terme inkl. Synonyme, die das Konzept beschreiben (die Label des Konzeptes). Zur Disambiguierung mehrdeutiger Schlüsselwörter wird die Distanz der Konzepte, die durch das Schlüsselwort gemeint sein könnten, zu den gefundenen Konzepten anderer Suchwörter herangezogen. Um die semantische Korrelation, also die Distanz zweier Konzepte in das Ranking einfließen zu lassen, werden die Gewichte propagiert. Diese Rankingmethode bezieht den lexikalischen Faktor nicht ein, der Rang der Ergebnisse ergibt sich aus der Anzahl der Terme inkl. ihrer Synonyme, die auf ein Konzept hindeuten, zusammen mit der semantischen Korrelation der Konzepte. Der Vergleich des Suchverfahrens mit traditionellem IR (in dem Fall *tfidf* und Kosinus-Maß, s. Kapitel 2.3.2) zeigte einen starken Anstieg der Vollständigkeit (von 55,6% auf 91,3%) und eine ebenfalls deutliche Verbesserung der Genauigkeit (von 67,5% auf 87,7%).

Eine **hierarchiebasierte Ranking-Methode** für das semantische Dokumentretrieval in digitalen Bibliotheken wird in [Stoyanovich et al., 2010] vorgestellt. Der Ansatz wurde auf PubMed[21], einer umfangreichen biomedizinischen digitalen Bibliothek evaluiert. PubMed verwendet den Media Subject Headings-Thesaurus (MeSH)[22] für die semantische Annotation der Dokumente. Der MeSH-Thesaurus ist polyhierarchisch, dasselbe Konzept kann an mehreren Stellen in der Taxonomie vorkommen (mit leicht unterschiedlicher Bedeutung im Kontext unterschiedlicher Themenbereiche, aber auch redundant). Ein Konzept wird deshalb so oft gezählt, wie es in der Hierarchie vorkommt. Seien t_1 und t_2 zwei Terme der Suchanfrage und n_{t_1} und n_{t_2} die Knoten mit dem Label t_1 und t_2. Die Ähnlichkeit von t_1 und t_2 ist dann die Anzahl Knoten in der Schnittmenge der Knotenmengen, die n_{t_1} bzw. n_{t_2} und jeweils alle darunterliegenden Knoten in der Taxonomie beinhalten [Stoyanovich et al., 2010]. Für die Normalisierung der Termähnlichkeit werden drei Wege vorgeschlagen:

- Anhand der Anzahl der Konzepte zu den Termen in der Suchanfrage;
- Anhand der Anzahl der Konzepte aus der Taxonomie, die im Dokument vorkommen;
- Anhand der Vereinigung beider Mengen.

Weiterhin wird die bedingte Ähnlichkeit definiert, die auch die Position des ersten Vorkommens eines Knotens (also eines Konzeptes aus der Taxonomie) in dem Dokument mit einbezieht. Je weiter oben das Konzept vorkommt, umso höher wird das Dokument gerankt. Schließlich definieren [Stoyanovich et al., 2010] die balancierte Ähnlichkeit als den Durchschnitt der bedingten Ähnlichkeiten der einzelnen Anfrageterme. Die Evaluie-

[21] `http://www.ncbi.nlm.nih.gov/pubmed` (06.01.2016)
[22] `http://www.nlm.nih.gov/mesh/` (06.01.2016)

rung vergleicht die Termähnlichkeit, bedingte Ähnlichkeit und balancierte Ähnlichkeit, eine statistisch signifikante Differenz ist jedoch nicht festzustellen. Ein Vergleich mit traditionellen Retrievalverfahren (z.B. *tfidf*-basiert) liegt nicht vor.

Dieses Modell definiert also ein mengenbasiertes Ähnlichkeitsmaß für schlüsselwortbasiertes semantisches Dokumentretrieval, das sich für Wissensbasen mit Baumstruktur eignet und die syntaktische Ähnlichkeit der Schlüsselwörter und Konzeptlabel in der Taxonomie nicht berücksichtigt. Der Ansatz lässt sich für Wissensbasen adaptieren, die überwiegend hierarchische Relationen bzw. Klassenstrukturen einsetzen, wie z.B. Teil-von-Relationen und Themen-Hierarchien.

Graphbasiertes semantisches Ranking, das die Wichtigkeit der Relationen mit einbezieht, wird in [Rocha et al., 2004] vorgestellt. Dabei werden Webseiten instanziiert, eine traditionelle Schlüsselwortsuche in den textuellen Inhalten der Instanzen und anschließend Spreading Activation (SA) auf dem Graph der Wissensbasis durchgeführt. Der Spreading Activation Algorithmus geht von initial aktivierten Knoten aus, in diesem Fall sind dies die gefundenen Dokumente, und flutet von diesen Knoten aus den Graphen mit „Energie“. Das Gewicht der benachbarten Knoten wird aus dem Gewicht des aktivierten Knotens und aus dem Gewicht der Kante dazwischen berechnet. Eine genaue Beschreibung des Algorithmus ist im Kapitel 5.2.3.2, Seite 97 zu finden, an dieser Stelle genügt es die grobe Verfahrensweise zu kennen. Semantische Netze für SA erfordern also die Zuordnung eines Kantengewichtes, das bestimmt, wie wichtig eine Relation für die jeweilige Domäne ist. Diese Zuordnung kann manuell erfolgen, z.B. von einem Domänenexperten, oder maschinell berechnet werden. [Rocha et al., 2004] schlagen die Berechnung der Verbundenheit, Spezifität und deren Kombination vor:

- Die *Verbundenheit* zweier Konzepte C_j, C_k als Gewicht der dazwischenliegenden Kante gibt an, wieviel Prozent der Konzepte, mit denen C_k durch eine Relation verknüpft ist, auch mit C_j verknüpft sind;
- Die *Spezifität* einer Kante zwischen den Konzepten C_j, C_k ist, basierend auf *tfidf*, invers proportional (die Wurzelfunktion wird eingesetzt) zu der Anzahl eingehender Kanten des Konzept C_k;
- Die Kombination ist das Produkt der Verbundenheit und Spezifität.

Diese drei Maße sind asymmetrisch, sie berücksichtigen die Richtung der Relationen. Als Startgewicht der initial aktivierten Knoten werden die Gewichte eingesetzt, die die traditionelle Schlüsselwortsuche liefert. Der SA-Algorithmus flutet den Graphen von diesen Knoten aus und bestimmt auf Basis dieser Gewichte sowie der Kantengewichte den Rang der benachbarten Knoten. Das Fluten kann durch den Einsatz von Bedingungen kontrolliert und auch gestoppt werden, z.B. wenn keiner der noch nicht weiter betrachteten aber bereits aktivierten Knoten ein Mindestgewicht erreichen. Das Ergebnis besteht aus einer Menge von gewichteten Knoten.

Tabelle 3.5 gibt einen Überblick der vorgestellten Verfahren.

Suchmaschine bzw. Rankingverfahren	Ranking basiert auf	Kurzbeschreibung
[Vallet et al., 2005]	Adaptiertes Vektorraummodell	Die Dokument-Instanz-Matrix beinhaltet die Abschätzung der Wichtigkeit eines Konzeptes für ein Dokument.
[Khan et al., 2004]	Distanz und semantische Korrelation	Berechnet die semantische Korrelation zwischen Suchanfrage und Konzept und verwendet die Distanz der Konzepte, die durch einen Suchterm gemeint sein könnten, zur Disambiguierung.
[Stoyanovich et al., 2010]	Hierarchiebasiert (Taxonomie)	Berechnet die Ähnlichkeit zweier Terme anhand des taxonomischen Kontextes der mit den Termen gefundenen Konzepte und kombiniert dieses Gewicht mit der Position des Termes in den Ergebnisdokumenten.
[Rocha et al., 2004]	Graphbasiert (SA)	Führt nach einer Schlüsselwortsuche auf dem Dokumentindex Spreading Activation auf dem semantischen Netz der Wissensbasis aus, die auch die Instanziierung der Dokumente beinhaltet. Dabei werden die Verbundenheit der Konzepte sowie die Spezifität der Pfade in Form von Gewichten in dem SA-Prozess berücksichtigt.

Tabelle 3.5: Überblick der vorgestellten Verfahren zum semantischen Ranking in semantischen Dokumentsuchmaschinen

3.3.3 Ranking hybrider semantischer Suchmaschinen

Dieses Kapitel stellt das Ranking von PowerAqua, K-Search und CE^2 vor, die im Sinne der Definition im Kapitel 2.4 hybride semantische Suchmaschinen sind.

PowerAqua, wie in Kapitel 3.1 bereits vorgestellt, führt neben der Faktensuche auch Dokumentretrieval durch. Die durch Faktensuche gefundenen Ressourcen werden nach ihrer Wichtigkeit gewichtet, die Berechnungsvorschrift ist jedoch nicht angegeben. Der Dokumentindex beinhaltet die semantischen Metadaten der Dokumente. Sie sind durch das eingesetzte Extraktionsverfahren gewertet, die Gewichte repräsentieren dabei die Stärke der Bedeutung des Konzeptes für das Dokument. Als Suchanfrage werden die Label der Ressourcen verwendet, die durch die Faktensuche gefunden wurden. Der Rang der einzelnen Dokumente ergibt sich nach dem Retrieval-Prinzip des Vektorraummodells (s. Kapitel 2.3.2) [Fernandez et al., 2008]. Ein Vergleich des Ansatzes mit traditioneller Schlüsselwortsuche zeigte in 56% der 20 Suchanfragen eine Verbesserung der durchschnittlichen Präzision der ersten 10 Ergebnisse, wobei die Verbesserung im Durchschnitt bei 5% lag.

Das Verfahren nutzt die Struktur des Graphen nur für die Faktensuche, wobei die Gewichte der gefundenen Fakten nicht in den weiteren Suchprozess einfließen. Es wird kein semantisches Ranking im Sinne von [Stojanovic et al., 2001] durchgeführt.

K-Search kombiniert Fakten- und semantische Dokumentsuche, indem die Ergebnisse beider Sucharten zusammenführt werden. Die formalen Anteile hybrider Anfragen werden als ein Filter für die Wissensbasis angewendet, der textuelle Inhalt der gefilterten Konzepte wird mit den informalen Teilen der Suchanfrage durchsucht. Das Zusammenführen der Ergebnisse basiert auf den semantischen Metadaten der Dokumente. Die Ergebnismenge

bilden diejenigen Dokumente, deren semantische Metadaten in den gefundenen Tripeln vorkommen. Alle Tripel ohne ein passendes Dokument werden aus der Ergebnismenge entfernt. Das Ranking der Suchmaschine K-Search basiert auf dem Vektorraummodell mit *tfidf*-Gewichtung. Die zusammengeführten Ergebnisse erhalten das Gewicht des Dokumentes aus der Schlüsselwortsuche. Die Fakten spielen keine Rolle, da eine formale Anfrage an die Wissensbasis nur solche Ergebnisse liefert, die die Anfrage erfüllen. Die Ergebnisse werden dem Benutzer wahlweise als eine geordnete Liste der gefundenen Dokumente, eine Liste der Tripel oder als Graph präsentiert [Bhagdev et al., 2008]. Das Ranking in K-Search ist kein semantisches Ranking, es basiert auf der Dokumentähnlichkeit im Vektorraummodell und nutzt die Struktur der Wissensbasis nicht.

Die hybride semantische Suchmaschine $\mathbf{CE}^2$ betrachtet die Wissensbasis und die Dokumente als einen Graph. Das initiale Ranking basiert auf den Gewichten, die durch die Schlüsselwortsuche geliefert werden, wobei nur den Prädikaten ein Gewicht ungleich 1 zugewiesen wird. Diese initialen Gewichte werden in dem Antwortgraphen von Knoten zu Knoten propagiert, wobei hier auch die Konnektivität der Knoten mit einbezogen wird. Wurden keine Prädikate gefunden, so bezieht das Ranking lediglich die Konnektivität der Knoten mit ein. Der Ansatz setzt semantisches Ranking ein, indem IR-basiertes Ranking mit Ranking basierend auf der Graphstruktur kombiniert wird.

Die in dieser Arbeit vorgestellte hybride semantische Suchmaschine SINFIO führt semantisches Ranking durch. Das Rankingverfahren bezieht die lexikalische Ähnlichkeit aus der syntaktischen Suche, *tfidf*-basierte Dokumentgewichte, die Distanz von Konzepten im semantischen Graphen (durch Spreading Activation) und die Abdeckung der Suchanfrage mit ein. Die genaue Vorgehensweise stellt Kapitel 5.2.3 vor. Eine Diskussion des Verfahrens in Bezug auf die Anforderungen an das Ranking sowie eine für große Datenmengen angepasste Variante wird im Kapitel 5.2.4 beschrieben.

Tabelle 3.6 gibt einen Überblick der vorgestellten Verfahren.

Suchmaschine	Ranking basiert auf	Kurzbeschreibung
PowerAqua	Fakt-Ranking und Ranking des Entitätenextraktions-verfahrens	Kombiniert den Rang, der im Rahmen der Faktensuche berechnet wird, mit der Wichtigkeit eines Konzeptes für ein Dokument. Letzteres wird durch das Entitätenextraktionsverfahren bestimmt.
K-Search	Vektorraum, tfidf-Gewichtung	Setzt gefundene Fakten als Filter für die Dokumentsuche ein. Die Gewichte der Ergebnisse sind die Dokumentgewichte, die anhand von tfidf-Gewichtung im Vektorrraummodell bestimmt werden.
CE2	Vektorraum und Graphstruktur	Instanziiert die Dokumente und betrachtet die Wissensbasis als ein Graph. Die Gewichte aus der Schlüsselwortsuche werden in dem Graph propagiert. Wenn keine Properties gefunden wurden, so wird auch die Konnektivität der Konzepte mit einbezogen.

Tabelle 3.6: Überblick der Rankingverfahren in hybriden semantischen Suchmaschinen

3.4 Evaluierung von semantischen und hybriden semantischen Suchmaschinen

Die Evaluierung semantischer Suchmaschinen ist ein offenes Problem, da keine Gold Standards und geeigneten Kennzahlen verfügbar sind [Uren et al., 2010, Strasunskas and Tomassen, 2010]. Aus diesem Grund wird häufig auf benutzerzentrierte Evaluierungsmethoden zurückgegriffen (z.B. in [Sure and Iosif, 2002, McCool et al., 2005, Todorov and Schandl, 2008, Elbedweihyhadija et al., 2012]). Hierdurch ist jedoch die Reproduzierbarkeit der Experimente (durch die Pragmatik) und die Vergleichbarkeit der Suchlösungen untereinander sowie mit traditionellen IR-Systemen nicht möglich. Zudem sind Versuche mit großen Datenmengen sehr kostenintensiv [Fernandez et al., 2009]. **Gold Standards und Verfahren zur Evaluierung von traditionellen IR-Systemen eignen sich nicht**, weil wesentliche Unterschiede bezüglich des Suchraumes, der Suchanfragen und der Suchergebnisse bestehen:

- Der *Suchraum* besteht aus der formalen Wissensbasis oder er beinhaltet neben den Dokumenten auch formales Wissen (s. 2.3.4.2 und 2.3.3). Gold Standards für traditionelle IR-Systeme lassen sich daher nicht für die Evaluierung einsetzen.
- Semantische Suchmaschinen unterstützen *Suchanfragen* verschiedener Formen und Komplexität. Die Leistung schlüsselwortbasierter semantischer Suchmaschinen verbessert sich im Allgemeinen mit der Anzahl der Suchterme. In traditionellen IR-Systemen werden jedoch meistens nur wenige Schlüsselwörter eingesetzt [Fernandez et al., 2009]. Formularbasierte Suchmaschinen, facettierte Suche und Suchmaschinen mit RDF-basierten Anfragesprachen unterstützen häufig komplexe, strukturierte Anfragen. Die meisten Frage-Antwort-Maschinen bieten die Möglichkeit, das Informationsbedürfnis in Form von natürlichsprachigen Fragen auszudrücken, z.B. SmartWeb [Wahlster, 2008], Alexandria[23] oder das START Natural Laguage Question Answering System[24]. Die Suchanfragen der Gold Standards eignen sich daher nicht für die Evaluierung aller semantischen Suchmaschinen, sie werden nicht von jedem Suchsystem unterstützt und nutzen die gebotenen Retrievalfeatures nicht aus.
- Entsprechend dem Suchraum können auch die *Suchergebnisse* verschiedenartig sein. Wird eine Faktensuche durchgeführt, so gibt es unter Umständen nur wenige relevante Ergebnisse, beispielsweise die vier Filme von Garry Marshall mit Julia Roberts. In solchen Fällen ist die Vollständigkeit nicht das geeignete Maß, um die Effektivität der Suchmaschine zu bestimmen. Vielmehr sollte der Fokus darauf liegen, wie gut die hoch gerankten Ergebnisse sind, z.B. durch die Berechnung der Genauigkeit an der Stelle (Rang) p [Croft et al., 2010b]. Einen Extremfall bilden Frage-Antwort-Maschinen, die genau eine Antwort liefern und weitere Bewertungskriterien benötigen (wird auf Seite 69 genauer vorgestellt).

Ein allgemeines Problem bei der Evaluierung semantischer Suchmaschinen ist, dass die Ergebnisse stark von der *Qualität der semantischen Daten* abhängen, dies wird implizit mit evaluiert. Wesentliche Faktoren dabei sind:

- die Abdeckung der Domäne seitens der Wissensbasis (durch Klassen und Instanzen),

[23] `http://alexandria.neofonie.de` (06.01.2016)
[24] `http://start.csail.mit.edu/index.php` (06.01.2016)

- die Komplexität des semantischen Modells, wie detailliert Relationen modelliert sind und (Fülle und Struktur der Relationen)
- die Konsistenz der Wissensbasis (im Sinne von konsistenten Aussagen) [Uren et al., 2010].

So liegt beispielsweise der medizinischen Suchmaschine RadSem eine manuell erstellte detaillierte Ontologie und eine von Experten korrigierte Instanzbasis zum menschlichen Skelett vor. RadSem erlaubt die Suche z.B. nach allen oder speziellen Knochen der Hand [Forcher et al., 2009]. Eine Suche in DBpedia hingegen kann keine solch präzise Information liefern, da die dafür notwendige Abdeckung der Domäne nicht vorhanden ist.

Gold Standards sind wichtig um eine komparative Evaluierung von Suchsystemen durchführen zu können. Das Problem der Vergleichbarkeit **semantischer Dokumentsuchmaschinen** untereinander und mit traditionellen IR-Systemen kann überwunden werden, indem *Gold Standards für IR-Systeme mit einer formalen Wissensbasis ergänzt werden.* Hierzu sind Anpassungen notwendig: es muss die Kopplung zwischen der Wissensbasis und den Dokumenten hergestellt, sowie modifizierte Anfragen für die jeweilige semantische Suchmaschine erstellt werden. Dann können die bestehenden Relevanzbeurteilungen für die Evaluierung herangezogen werden. Die semantischen Metadaten der Dokumente - und somit die Kopplung zwischen der Wissensbasis und der Dokumentmenge - wird mithilfe von Wissensextraktionsverfahren hergestellt (z.B. in [Fernandez et al., 2009, Perez-Aguera et al., 2010]). Die Suchanfragen werden so umformuliert, dass sie sich für eine semantische Suche auf dem jeweiligen Suchsystem eignen. Beispielsweise wird die Suchanfrage „Nirvana“ durch „What are the members of Nirvana?“ ersetzt [Fernandez et al., 2009]. Der Nachteil dieser Vorgehensweise liegt darin, dass die Qualität des Gold Standards stark von der Qualität der Wissensbasis sowie der Qualität der Annotationen der Dokumente abhängt [Fernandez et al., 2009]. Zudem decken verfügbare Ontologien bzw. Wissensbasen nur einen kleinen Teil der Anfragen aus bestehenden Gold Standards ab [d'Aquin et al., 2008]. Beispielsweise wurde bei der Evaluierung von PowerAqua [Lopez et al., 2006, Lopez et al., 2012] der TREC WT10G Corpus mit online verfügbaren Ontologien gekoppelt. Dabei konnten nur 20% der Anfragen abgedeckt werden [Uren et al., 2010]. Weiterhin kann die semantische Suche Dokumente finden, die im Gold Standard nicht beurteilt aber relevant sind. Die semantische Suche schneidet in diesem Fall schlechter als das IR-System ab [Fernandez et al., 2009].

Eine weitere Vorgehensweise ist es den *Gold Standard selber zu erstellen* [Castells et al., 2007]. Hierfür sollten geeignete Ontologien und Instanzbasen vorliegen, die die Themen der ausgewählten Testdokumente möglichst gut abdecken. Die Dokumente können mithilfe von Informationsextraktionsverfahren annotiert werden. Da die Relevanzbeurteilung der Dokumente hinsichtlich der Anfragen jedoch manuell erfolgt, kann ein Gold Standard lediglich auf kleinen Dokumentmengen mit wenigen Suchanfragen manuell erstellt werden. Zudem bestehen dieselben Nachteile, die bei der Erweiterung bestehender Gold Standards vorliegen.

Für die **Evaluierung der Faktensuche** existieren ebenfalls noch keine Gold Standards. Es werden Wissensbasen, wie z.B. DBpedia oder der Mooney Natural Language Learning Data Set[25] verwendet. DBpedia ist eine formale Wissensbasis mit Ontologien und Instanzen, sie beinhaltet keine Suchanfragen und Relevanzbeurteilungen. Die Such-

[25] `http://www.cs.utexas.edu/users/ml/nldata.html` (06.01.2016)

anfragen müssen aus Anfragelogs bestehender Suchmaschinen ausgewählt (z.B. DBpedia, Yahoo! Search oder Microsoft Life Search Query Log) und die Suchergebnisse manuell auf ihre Relevanz hin beurteilt werden. Die Mooney-Datenmenge enthält zusätzlich zu den Ontologien und der Instanzbasis auch eine Menge englischer natürlichsprachiger Suchanfragen sowie deren Repräsentation als logische Anfragen. Sie deckt die Domäne der geografische Daten, Jobdaten und Restaurants ab. Insbesondere die geografischen Daten werden häufig zur Evaluierung eingesetzt, da diese kein Expertenwissen erfordern (z.B. in [Elbedweihyhadija et al., 2012, Kaufmann and Bernstein, 2007, Wrigley et al., 2010]). Da beide Datenmengen selber keine Relevanzbeurteilung beinhalten, basiert die Evaluierung üblicherweise auf dem Relevanzurteil der Benutzer für die ersten k-Ergebnisse (Pooling, s. Seite 41).

Die Evaluierung von *Frage-Antwort-Maschinen* unterscheidet sich von anderen Faktensuchmaschinen, sobald diese auf sogenannte „Faktoide“ Fragen (factoid question) ausgerichtet sind. Faktoide Fragen sind Suchanfragen, die mit einem kurzen Satz oder mit einer kurzen Liste von Dingen beantwortet werden können [Agichtein et al., 2005], wie z.B. die Fragen „Wie lang ist die Donau?“ oder „Wie heißen die sieben kanarischen Hauptinseln?“. Gibt es eine korrekte Antwort, so kann MRR eingesetzt werden (s. Seite 38). In diesem Fall können die Kriterien Relevanz, Korrektheit, Prägnanz, Vollständigkeit und Begründung der Antwort herangezogen und daraus abgeleitet die Korrektheit, Exaktheit und Begründung der Suchergebnisse beurteilt werden. Genauigkeit und Vollständigkeit werden angepasst. Dabei ist die Genauigkeit die Anzahl der korrekten Ergebnisse geteilt durch die Anzahl der beantworteten Anfragen und die Vollständigkeit die Anzahl der korrekten Ergebnisse geteilt durch die Anzahl der zu beantwortenden Fragen [Allam and Haggag, 2012].

Für die **hybride semantische Suche** werden entweder kleine Gold Standards [Bhagdev et al., 2008, Wang et al., 2011] oder ein Informationspool bestehend aus größeren Dokumentmengen mit thematisch passender Wissensbasis erstellt [Lopez et al., 2006, Lopez et al., 2012], das jedoch keine Relevanzbeurteilung enthält. Die Evaluierung solcher Suchsysteme basiert auf Pooling, wobei die Relevanz der ersten k Ergebnisse von den Benutzern beurteilt wird. Um die Effizienz eines hybriden Suchansatzes zu zeigen, wurden die Lösungen entweder mit der Schlüsselwortsuche sowie der semantischen Dokumentsuche [Bhagdev et al., 2008, Lopez et al., 2012] oder mit Schlüsselwortsuche sowie Faktensuche [Wang et al., 2011] auf dem selben Informationspool verglichen.

Kapitel 4
These und Forschungsfragen

In diesem Kapitel wird auf der Basis der theoretischen Grundlagen abgeleitet, wie die Lücke zwischen der Fakten- und semantischen Dokumentsuche geschlossen werden kann. Die in der Einleitung formulierte These wird präzisiert (Kapitel 4.1), das hybride semantische Suchproblem formal beschrieben (Kapitel 4.2) und die damit verbundenen Forschungsfragen definiert (Kapitel 4.3). Das Kapitel lehnt sich an [Schumacher et al., 2008, Schumacher et al., 2011] sowie [Schumacher and Sintek, 2011].

4.1 These

Dieser Arbeit liegt die These zugrunde, dass die Suche in unterschiedlich strukturierten Datenmengen durch hybride semantische Suche, die Fakten- und Dokumentsuche kombiniert, verbessert werden kann.

Die hybride semantische Suche erweitert die Architektur aus Kapitel 2.3.4.1 um einen Textindex, wie in Abbildung 4.1 veranschaulicht. Die unterschiedlich strukturiert vorliegenden Informationen aus den verschiedenen Datenquellen werden also in einer *formalen Wissensbasis* (formal strukturiert, als Fakten) und in einem *Textindex* (informal, unstrukturiert, Dokumentrepräsentation nach dem „Bag of Words"-Ansatz[1]) abgebildet. Fakten und Dokumente sind *durch die formalen semantischen Metadaten der Dokumente vernetzt*[2]. Die Wissensbasis in Verbindung mit dem Textindex wird im Folgenden Informationspool genannt.

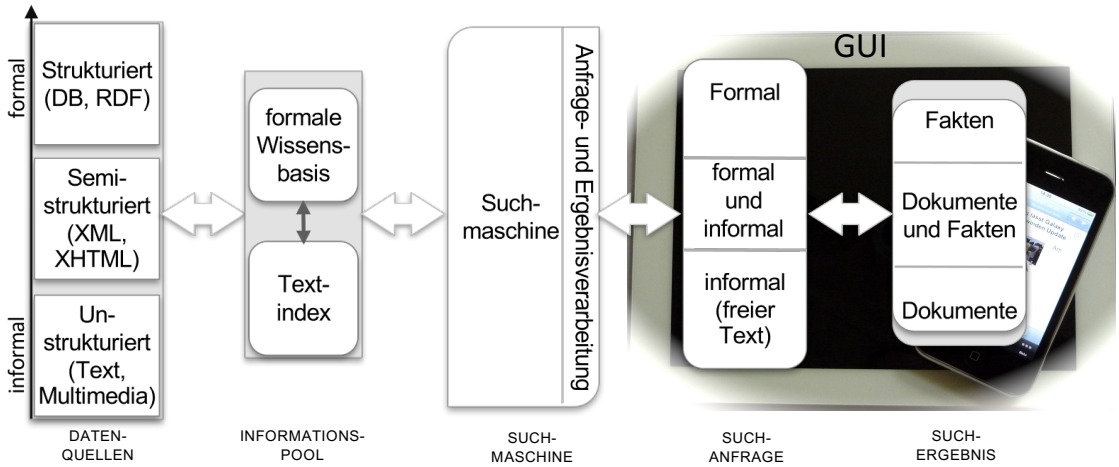


Abbildung 4.1: Architektur der hybriden semantischen Suchmaschine

Die Inhalte der formalen Wissensbasis und des Textindexes kommen aus unterschiedlichen Datenquellen, die diese unterschiedlich formal und strukturiert speichern (s. Abbil-

[1] Der „Bag of Words"-Ansatz wurde im Kapitel 2.3.2, auf Seite 14 beschrieben. Methoden zur Abbildung der Inhalte wurden im Kapitel 2.3.4.1 vorgestellt.

[2] Diese Architektur wird bereits in mehreren Web 3.0-Anwendungen eingesetzt, wie z.B. in NEPOMUK [Sauermann et al., 2007], Aletheia [Stieger and Aleksy, 2009] oder InfoSleuth [Nodine et al., 2000].

dung 4.1). Inhalte können folglich teilweise nur in strukturierter oder nur in unstrukturierter Form vorliegen, daher auch nur formal oder nur informal in dem Informationspool abgebildet sein. Auch wenn der Informationspool die Dokumente und als formales Wissen nur deren semantische Metadaten umfasst, decken die Metadaten selten den gesamten Inhalt der Dokumente ab, der für die Befriedigung des Informationsbedürfnisses der Benutzer relevant sein kann. Die Gründe hierfür sind:

- Es können nur Metadaten über das Dokument, wie Autor, Veröffentlichungsdatum und wenige Stichwörter für angesprochene Themen usw. vorliegen. Dies ist häufig in digitalen Bibliotheken der Fall (z.B. in DiLiA [Eichler et al., 2010]).
- Ebenfalls können die Metadaten unvollständig sein, bedingt durch die Extraktionsverfahren oder durch die Unvollständigkeit der zugrundeliegenden Ontologien. Informationsextraktionsverfahren bieten zwar eine hohe Genauigkeit, wenn es um einfachere Aufgaben, wie z.B. das Erkennen von Entitäten (Personen, Städte usw.) geht, sind jedoch weniger zuverlässig bei der Erkennung komplexerer Zusammenhänge und Aussagen (Fakten). So erreichen manche Verfahren ein F-Maß von über 90% bei der Entitätenerkennung [Ratinov and Roth, 2009]. Relationsextraktionsverfahren kommen jedoch nur für einfache binäre Relationen, bei denen das Nomen im Satz enthalten ist, auf eine über 80%ige Präzision. In komplexeren Fällen liegen sowohl die Präzision als auch die Vollständigkeit deutlich darunter und erreichen teilweise noch nicht einmal 60% [Etzioni et al., 2011, Fader et al., 2011]. Eine manuelle Annotation der Dokumente ist nur für kleine, mehr oder weniger statische Datenmengen machbar.
- Eine unvollständige Domänenontologie oder eine Modellierung aus einer bestimmten Sicht kann ebenfalls dazu führen, dass ein Teil der semantischen Metadaten fehlen [Bhagdev et al., 2008]. Relationen, Klassen und Instanzen, die nicht modelliert sind, können nicht erkannt werden. Die Modellierung der Filmwelt beispielsweise unterscheidet sich aus der Sicht eines Filmemachers von der des Publikums, da die beiden Gruppen an verschiedenen Aspekten interessiert sind.

Die formal und informal abgebildeten Inhalte enthalten also nicht dieselben Informationen. Folglich ist es möglich, dass eine *Suchanfrage* mit *Fakten*, mit *Dokumenten* oder auch mit *einer Kombination von Dokumenten und Fakten, also mit hybriden Suchergebnissen, beantwortet* werden kann.

Eine *Kombination von formalen und informalen Inhalten* sollte auch *bei der Anfrageformulierung* in Form von *hybriden Anfragen* ermöglicht werden. Denn formale Anfragen sind präziser als informale Anfragen, und je präziser die Anfrage ist, desto weniger tritt das Problem der Mehrdeutigkeit auf und umso höhere Genauigkeit kann die Suchmaschine erzielen (vgl. Kapitel 2.3.3).

Abbildung 4.1 skizziert die erweiterte Architektur mit dem Informationspool, den formalen, informalen und hybriden Suchanfragen sowie Suchergebnissen. Ein Beispiel für eine hybride Suchanfrage und ein hybrides Suchergebnis zeigt Abbildung 4.2, wobei in den Daten der vierten Europäischen Semantic Web Konferenz gesucht wurde. Die Suchterme „authors“ und „papers“ sind formal, d.h. mit diesen Termen lassen sich Konzepte in der Wissensbasis identifizieren. Der Ausdruck „semantic web technologies“ ist informal. Ein hybrides Suchergebnis besteht aus einem Dokument zu „semantic web technologies“ sowie der Fakten über das Dokument, wie die Autoren (`http://swrc.ontoware.org/ontology#author`) und dessen Typ „paper“ (`http:`

//data.semanticweb.org/ns/swc/ontology#Paper).

Abbildung 4.2: Beispiel für hybride Suchanfrage und hybrides Suchergebnis

Um hybride Anfragen verarbeiten und hybride Ergebnisse liefern zu können, muss der Suchalgorithmus in der Lage sein sowohl die formale Wissensbasis als auch den Textindex zu durchsuchen und dabei die Verknüpfungen, die durch die semantischen Metadaten der Dokumente gegeben sind, zwischen diesen zu nutzen. Die hybride semantische Suchmaschine integriert die unterschiedlich stark formalen Inhalte während des gesamten Suchprozesses: beginnend mit formalen, hybriden und informalen Suchanfragen über ein integriertes Suchverfahren bis hin zu formalen, hybriden und informalen Suchergebnissen. Dies präzisiert die allgemein formulierte **These** aus Kapitel 1.2. Die Verbesserung der Suche wird anhand der Retrievaleffektivität (s. Kapitel 2.5.1) im Vergleich des integrierten Ansatzes zur Fakten- und semantischen Dokumentsuche ohne eine Kombination der unterschiedlich strukturierten Inhalte gemessen:

Für unterschiedlich stark strukturierte Datenmengen ist ein integrierter Ansatz zur hybriden semantischen Suche, der formale und informale Inhalte während des gesamten Suchprozesses kombiniert, effektiver als eine semantische Dokument- und Faktensuche ohne Kombination der Inhalte.

So eine hybride semantische Suchlösung kann durch einen Ansatz erreicht werden, der die Faktensuche mit semantischer Dokumentsuche kombiniert, da das Suchen in einer formalen Wissensbasis grundsätzlich unterschiedliche Methoden erfordert als in einem Textindex [Schumacher et al., 2008, Schumacher et al., 2011, Schumacher and Sintek, 2011] (vgl. Kapitel 2.3.3).

Mit einer solchen hybriden Suchlösung können Suchanfragen so präzise beantwortet werden, wie es der zugrundeliegende Suchraum bzw. Informationspool erlaubt: Liegen die Informationen in der Wissensbasis vor, so liefert die Suche Fakten und der Benutzer braucht nicht in den Ergebnisdokumenten weiterzusuchen. Kann die Suchanfrage nicht allein durch Fakten beantwortet werden, so können die Fakten durch Dokumente, die die fehlenden Informationen beinhalten, ergänzt werden. Gibt es keine passenden Fakten, so wird durch semantisches Dokumentretrieval nach Dokumenten gesucht.

4.2 Das hybride semantische Suchproblem

Das hybride semantische Suchproblem besteht darin, zu einer Suchanfrage, die formal, informal oder hybrid sein kann, diejenigen Dokumente, Fakten und hybriden Ergebnisse zu finden, die zur Beantwortung der Anfrage relevante Informationen enthalten.

Sei R die Menge der Ressourcen in der formalen Wissensbasis, G_Σ die formale Wissensbasis (Ontologien und Instanzen) als Graph (Menge von Tripeln) und TI den Volltextindex. Dann ist der Informationspool I die Vereinigung der Wissensbasis und des Textindex, wobei der Textindex TI die Menge der darin abgebildeten Dokumente aus dem Suchraum (s. Kapitel 2.3.4) darstellt, also aus der Menge der Dokumentvektoren (s. Kapitel 2.3.2) besteht:

$$G_\Sigma = \{\langle s,p,o\rangle \mid s,p,o \in R, \langle s,p,o\rangle \textit{ ist ein Tripel in der Wissensbasis}\} \tag{4.1}$$

$$D = \{d|\ d \textit{ ist ein Dokument im Suchraum}\} \tag{4.2}$$

$$TI = \{\vec{d}|\ \vec{d} \textit{ ist die Repräsentation des Dokuments } d\ \in D \textit{ als Dokumentvektor im Textindex}\} \tag{4.3}$$

$$I = \{TI \cup G_\Sigma\} = \{\vec{d}, \langle s,p,o\rangle \mid \vec{d} \in TI, \langle s,p,o\rangle \in G_\Sigma\} \tag{4.4}$$

Für die hybride semantische Suche ist eine enge Kopplung notwendig: Die Metadaten der Dokumente verweisen explizit auf Ressourcen aus der formalen Wissensbasis und umgekehrt. Üblicherweise werden Dokumente instanziiert. Die Instanziierung eines Dokuments ist dessen formale Repräsentation in der formalen Wissensbasis. Die Instanz repräsentiert das Dokument als Ressource in der Wissensbasis, beinhaltet jedoch nicht notwendigerweise die Inhalte des Dokuments. Bezeichne $r_{d_i} \in R$ die Instanziierung eines Dokuments d_i in der formalen Wissensbasis:

$$\{\forall d_i \in D\ \exists r_{d_i} \in R|\ r_{d_i} \textit{ ist die Instanziierung von } d_i\}. \tag{4.5}$$

Sei Q die Menge der Suchanfragen. Dann besteht die Menge der Suchergebnisse SR bezüglich der Suchanfrage q aus Paaren von Suchergebnissen sr_q und deren Rang w_{sr_q}, so dass:

$$SR_q = \{(sr_q, w_{sr_q})|\ sr_q \in (G_\Sigma \cup D \cup SR_{hybrid}),\ w_{sr_q} \in [0,1] \subset \mathbb{R}\}. \tag{4.6}$$

Die Menge der hybriden Ergebnisse SR_{hybrid} besteht aus Paaren von einem Dokument und dem zugehörigen Tripelset, so dass Folgendes erfüllt ist:

$$SR_{hybrid} = \{(d_i, M_{d_i})|\ d_i \in D,\ M_{d_i} \subseteq G_\Sigma \textit{ ist ein zusammenhängender Graph oder eine Ressource},\ \exists\, \langle s,p,o\rangle \in M_{d_i} \textit{mit } s = r_{d_i} \vee o = r_{d_i},\} \tag{4.7}$$

Die hybriden Ergebnisse sind also mit einer Dokumentinstanz verknüpfte zusammenhängende Graphen (M_{d_i}), d.h. Fakten, die über das Dokument was aussagen und/oder zusätzliche Informationen dazu liefern. Lose Fakten würden zu zusammenhanglosen, daher unverständlichen Ergebnissen führen. Zudem ließe sich schwer bestimmen, welche

Fakten zu einem Dokument gehören, was also dem Benutzer als ein hybrides Ergebnis präsentiert wird.

Das hybride Suchproblem ist dann ein 3-Tupel der Suchanfragen Q, des Informationspools I und der hybriden Suchfunktion H:

$$(Q,\ I,\ H(Q,\ I)) \tag{4.8}$$

Die *hybride semantische Suchfunktion* H ordnet jedem $q \in Q$ die Menge der gewichteten Suchergebnisse mit einem Gewicht $w \in [0, 1]$ bezüglich q, SR_q zu:

$$H:\ (q,\ I) \rightarrow SR_q \tag{4.9}$$

Das von dem System berechnete Gewicht w soll ausdrücken, wie relevant das Suchergebnis ist, um das durch die Suchanfrage ausgedrückte Informationsbedürfnis zu befriedigen. Die Relevanz eines Dokuments kann jedoch nur der Benutzer beurteilen, der seine Suchanfrage auf Basis seines mentalen Modells mit einer bestimmten Intention (Pragmatik) formuliert hat (vgl. Kapitel 2.3.2 und 2.5.1).

4.3 Forschungsfragen

Semantische Dokumentsuche und Faktensuche erfordern unterschiedliche Suchverfahren aufgrund der unterschiedlichen Abbildung der Informationen (vgl. Kapitel 2.3). Um beides durchführen und miteinander kombinieren zu können müssen also verschiedenartige Verfahren eingesetzt werden. Daraus ergibt sich die Forschungsfrage:

F1 - Wie können ein Verfahren zur Faktensuche und ein Verfahren zur semantischen Dokumentsuche so kombiniert werden, dass die Kombination in der Lage ist, Fakten und Dokumente während des gesamten Suchprozesses abhängig voneinander zu durchsuchen?

Durch diese Forschungsfrage ergeben sich weitere Fragen, die in diesem Zusammenhang beantwortet werden müssen. Sowohl für die Faktensuche, als auch für die semantische Dokumentsuche existieren verschiedene Verfahren (s. 2.3.4.4), die sich entweder kombinieren lassen oder nicht. Weiterhin kann eine Kombination die Anpassung der Verfahren erfordern:

F2 - Welche Verfahren zur Faktensuche und zur semantischen Dokumentsuche eignen sich für solch eine Kombination und inwiefern müssen diese dazu adaptiert werden?

Dabei stellt sich die Frage, ob es Verfahren zur Faktensuche und zur semantischen Dokumentsuche gibt, dessen Kombination auch sinnvolle, zu der Suchanfrage passende hybride Ergebnisse finden kann oder hierfür ein drittes Verfahren eingesetzt werden muss:

F3 - Können die Verfahren zur semantischen Dokumentsuche und zur Faktensuche so kombiniert werden, dass sie auch hybride Suchergebnisse finden oder sollte hierfür ein weiteres Verfahren eingesetzt werden?

Ein hybrides Ergebnis ist dann sinnvoll, wenn es zur Beantwortung der Suchanfrage beiträgt, also relevant (s. Kapitel 2.5.1) und auch verständlich ist.

Verschiedene Suchverfahren verwenden zum Teil unterschiedliche Rankingfunktionen (Funktionen zur Relevanzgewichtung). Insbesondere werden für das Ranking von Fakten und das Ranking von Dokumenten unterschiedliche Methoden eingesetzt, die nicht notwendigerweise miteinander kombinierbar sind (s. Kapitel 3.3.2). Um die Relevanz der Ergebnisse einer Suchanfrage vergleichen zu können muss für den hybriden Ansatz eine geeignete Rankingfunktion gefunden werden:

F4 - Wie kann die Rankingfunktion für einen hybriden semantischen Suchansatz aussehen? Wie werden Fakten, Dokumente und insbesondere hybride Suchergebnisse bewertet?

Zudem muss die Komponente zur Anfrage- und Ergebnisverarbeitung zusammen mit der Benutzerschnittstelle die Darstellung und Interaktion mit hybriden Anfragen und Ergebnissen unterstützen. Insbesondere die Konstruktion von hybriden Anfragen und eine geeignete verständliche Darstellung der einzelnen Ergebnisse sowie der Ergebnisliste, die aus verschieden strukturierten Elementen bestehen kann, ist wesentlich für die Akzeptanz einer hybriden Suchmaschine. Daraus ergeben sich zwei weitere Forschungsfragen, die sich auf die Benutzerfreundlichkeit der Suchmaschine beziehen.

Formale Anfragen sind präziser als informale Anfragen, und präzisere Suchanfragen führen zu einer besseren Retrievaleffektivität [Hildebrand et al., 2007, Schumacher et al., 2011]. Um ohne Unterstützung formale Anfragen zu stellen müssten die Benutzer jedoch die zugrundeliegende Wissensbasis bzw. das Vokabular kennen (s. Kapitel 2.3.3). Die Forschungsfrage ist:

F5 - Wie kann der Benutzer bei der Anfragestellung unterstützt werden, so dass er ohne Kenntnis der zugrundeliegenden Wissensbasis Anfragen mit möglichst vielen formalen Anteilen stellt?

Die zweite Herausforderung bezüglich der Benutzerschnittstelle liegt darin, die unterschiedlichen Ergebnistypen (Fakten, Dokumente und hybride Ergebnisse) so darzustellen, dass diese sich dem Benutzer erschließen, d.h. verständlich sind. Dies gilt auch für die gesamte Ergebnisliste: Die Anordnung der Ergebnisse sowie die Repräsentation der Liste sollten die Interpretation und damit die Beantwortung der Suchanfrage unterstützen. Als Frage formuliert:

F6 - Wie können die Suchergebnisse, bestehend aus Fakten, Dokumenten und hybriden Ergebnissen, so dargestellt werden, dass diese und auch die Ergebnisliste verständlich sind?

Kapitel 5
Lösungsansatz

In diesem Kapitel wird vorgestellt, welche Anforderungen die These und die Forschungsfragen an den Lösungsansatz stellen (Kapitel 5.1) und wie diese erfüllt werden können (Kapitel 5.2). Die einzelnen Forschungsfragen und die zugehörigen Anforderungen werden im Hinblick auf die Suchverfahren und ein hybrides Suchsystem analysiert und mögliche Lösungen diskutiert. Der konkrete Lösungsansatz mit ausgewählten Verfahren wird formal beschrieben. Weiterhin wird die Realisierung des Lösungsansatzes und die damit verbundenen Entwurfsentscheidungen vorgestellt (Kapitel 5.3).

Die in diesem Kapitel erläuterten Konzepte wurden in den Veröffentlichungen [Schumacher et al., 2008], [Schumacher et al., 2011], [Schumacher and Sintek, 2011], [Grimnes et al., 2009], [Sauermann et al., 2008] publiziert.

5.1 Anforderungen

Aus der These und den Forschungsfragen ergeben sich **Anforderungen an das Lösungskonzept**. Diese lassen sich in zwei Gruppen unterteilen: Anforderungen an den gesamten Suchansatz (**F1-4**) sowie Anforderungen an die Benutzerschnittstelle (**F5-6**).

5.1.1 Anforderungen an den gesamten Suchansatz

Die These sowie die **Forschungsfragen F1, F2, F3** und **F4** beziehen sich auf die eingesetzten Suchverfahren, ihre Kombination und die semantische Leistungsfähigkeit (s. Seite 27) der Suchmaschine. Sie stellen die Anforderung geeignete Suchansätze zu finden, zu adaptieren und zu kombinieren, so dass:

A 1: der hybride Ansatz formale, informale und hybride Anfragen verarbeiten kann,

A 2: diese mit Fakten, Dokumenten und hybriden Suchergebnissen beantworten kann,

A 3: dabei nicht nur die vorhandenen Verknüpfungen zwischen dem Textindex und der Wissensbasis (Dokument - semantische Metadaten) ausnutzt, sondern auch diejenigen Dokumente finden kann, die nicht semantisch annotiert sind,

A 4: lexikalische und strukturelle Mehrdeutigkeiten in informalen und hybriden Anfragen auflösen kann

A 5: und die Umsetzung einer adäquaten Rankingfunktion erlaubt, welche die hybriden Ergebnisse und Ergebnislisten unterstützt.

5.1.2 Anforderungen an die Benutzerschnittstelle

Bei einer hybriden semantischen Suchmaschine bedarf es einer benutzerfreundlichen Bedienschnittstelle, die den Benutzer bei der Anfragestellung unterstützt. Bevor die konkreten Anforderungen aus der **Forschungsfrage F5** (Unterstürzung der Benutzer bei der Anfragestellung) abgeleitet werden, ist zu entscheiden, welche Art von Anfragen die

Suchmaschine unterstützen soll. Das Ziel ist es Anfragen mit möglichst vielen formalen Anteilen zu haben, ohne dass die Benutzer die zugrundeliegende Wissensbasis kennen müssen.

Suchmaschinen mit RDF-basierten Anfragesprachen (s. Seite 21), formularbasierten Suchmaschinen (s. Seite 20) und die facettierten Suche (s. Seite 21) sind für formale Anfragen ausgelegt. Sie bieten jedoch weniger Freiheiten für den Benutzer, als Suchmaschinen mit natürlichsprachigen Anfragen. RDF-basierte Anfragesprachen müssen die Benutzer erlernen, sie sind beschränkt durch die Ausdrucksmächtigkeit der Anfragesprache und benötigen Kenntnisse über die zugrundeliegende Wissensbasis. Formularbasierte Suchmaschinen geben einen Rahmen vor, um die Suchanfrage durch Auswahl verschiedener Metadaten zu bestimmen. Die Anfrage ist durch die Auswahlmöglichkeiten beschränkt, auch wenn häufig zusätzlich Schlüsselwörter angegeben werden können. Ähnlich verhält es sich bei der facettierten Suche, die eine Weiterentwicklung der formularbasierten Suche ist. Der Unterschied liegt in der Dynamik durch die kontinuierliche Verfügbarkeit der Facetten und Klick-und-Filter-Funktion. Schlüsselwortbasierte semantische Suchmaschinen bzw. Suchmaschinen mit natürlichsprachigen Anfragen erlauben den Benutzern, ihr Informationsbedürfnis frei zu formulieren. Eine Untersuchung der Benutzerakzeptanz bei der Anfragestellung für Faktensuche zeigte klare Präferenz für natürlichsprachige Eingaben, insbesondere für ganze Fragen, wobei das Informationsbedürfnis durch Schlüsselwortsuche am schnellsten befriedigt werden konnte [Kaufmann and Bernstein, 2007]. Eine andere Studie befragte Studenten, ob sie die Suche mit Schlüsselwörter oder mit ausformulierten natürlichsprachigen Fragen bevorzugen. Die Studienteilnehmer bevorzugten die schlüsselwortbasierte Suche und zwar unabhängig davon, wie die Qualität der Suchergebnisse beurteilt wurden [Reichert et al., 2005]. Der Unterschied in den Studienergebnissen kann an der Teilnehmergruppe, aber auch daran liegen, dass in der ersten Studie nach Fakten, in der zweiten nach Dokumenten gesucht wurde.

Basierend auf den Studienergebnissen ist das Ziel, *natürlichsprachige Anfragen* sowohl in Form von Schlüsselwörtern als auch als ganze Sätze bzw. Fragen, zu unterstützen. So eine Suchlösung kann als *eine Mischung von semantikbasierter Schlüsselwortsuchmaschine und Frage-Antwort-Maschine* für hybride semantische Suche angesehen werden: Sie verbessert die Schlüsselwortsuche durch das Einbeziehen der verfügbaren semantischen Daten, kann aber auch natürlichsprachige Fragen beantworten (vgl. Kapitel 2.3.3). Wie in Forschungsfrage F5 bereits formuliert, um möglichst präzise Suchanfragen zu haben sind möglichst viel formale Anteile erwünscht, jedoch ohne, dass die Benutzer eine Anfragesprache verwenden oder die Wissensbasis kennen müssen. Um die Möglichkeiten der natürlichen Sprache nicht einzuschränken, können die formalen und informalen Anteile der Suchanfrage in beliebiger Reihenfolge stehen. Die Anfragestellung unterstützende Komponente ist also so zu gestalten, dass:

A 6: der Benutzer die formalen Teile der Anfrage leicht identifizieren und auswählen kann;

A 7: nicht formale Anfragen und Anfrageteile ohne zusätzlichen Aufwand gestellt werden können;

A 8: eine beliebige Reihenfolge von formalen und informalen Anfrageteilen möglich ist.

Insbesondere bei semantischen Suchmaschinen ist eine verständliche Darbietung der Suchergebnisse einer der wichtigsten Aspekte hinsichtlich Benutzerakzeptanz [Elbedweihy et al., 2012]. Dies kann daran liegen, dass der Bezug der Ergebnisse zu der Suchanfrage

nicht immer direkt ersichtlich ist, weil z.B. Synonyme verwendet werden oder die Ergebnisse komplex sind (Fakten oder Dokumente mit ihren Metadaten). Die Präsentation von Fakten wird häufig als zu technisch empfunden [Elbedweihy et al., 2012]. Bei einer hybriden semantischen Suche mit Fakten, Dokumenten und hybriden Ergebnissen können die einzelnen Ergebnisse noch komplexer sein, zudem ist die Ergebnismenge bereits gemischt. Umso mehr gewinnt eine verständliche Darstellung der Ergebnisse gemäß **Forschungsfrage F6** (Verständliche Darstellung der Ergebnisse) an Bedeutung.

Fakten werden als Graphen (z.B. [Aleman-Meza et al., 2005, Bhagdev et al., 2008]) oder in Form vom strukturierten Text abgebildet (z.B. [Lei et al., 2006, Guha et al., 2003, Lopez et al., 2005]). Für Dokumente haben sich Titel und ein Textsnippet, das die Suchwörter beinhaltet (keywords in context), durchgesetzt, wie es auch in den gängigen Websuchmaschinen üblich ist [Hearst, 2011]. Diese Darstellungen unterscheiden sich jedoch deutlich voneinander. Befinden sich beide Darstellungsarten in derselben Ergebnisliste, so erfordert ihre Interpretation eine höhere kognitive Leistung von dem Benutzer, da sie beide verstanden und miteinander verbunden werden müssen [Mayer and Moreno, 2003, Sweller et al., 1998]. So verursacht beispielsweise die Interpretation einer Ergebnisliste, in der die Fakten als Graph und die Dokumente als Titel und Textsnippet dargestellt sind, eine höhere kognitive Last, als wenn beide Ergebnistypen als Text dargestellt werden. Zudem kann, durch die unterschiedlichen Ergebnistypen, die Anordnung der Ergebnisse in der Ergebnisliste das Verständnis und die Beurteilung der Antworten beeinflussen. Dabei stellt sich die Frage, ob eine, unabhängig von den Ergebnistypen nur anhand der Relevanz geordnete Ergebnisliste oder eine nach Ergebnistypen gruppierte Ergebnisliste bevorzugt wird. Zusammengefasst sind die Anforderungen an die Ergebnisdarstellung:

A 9: Verständliche Darstellung der Ergebnistypen (Fakten, Dokumente und hybride Ergebnisse),

A 10: so dass diese eine möglichst ähnliche Darstellungsform aufweisen.

A 11: Verständliche Darstellung der Ergebnisliste selber im Bezug auf die Anordnung der Ergebnisse.

5.2 Lösungskonzept

In diesem Kapitel wird, basierend auf den Anforderungen A1-A11, der Lösungsansatz vorgestellt.

Kapitel 5.2.1 stellt die Anfragestellung vor. Es beinhaltet sowohl den syntaktischen Abgleich gemäß der Anforderung 1 als auch die grafische Benutzerschnittstelle zur Unterstützung der Anfragestellung gemäß den Anforderungen 6-8, da diese in enger Zusammenhang zueinander stehen. Kapitel 5.2.2 stellt die an den Suchansatz gestellte Anforderungen 2-4 den bestehenden Verfahren zur Fakten- sowie semantischen Dokumentsuche gegenüber mit dem Ziel, kombinierbare Verfahren zu finden. Kapitel 5.2.3 beschreibt den semantischen Abgleich der hybriden semantischen Suchlösung mit den ausgewählten Verfahren. Dieses Kapitel beinhaltet auch den Rankingalgorithmus, da er ein integraler Bestandteil des Suchverfahrens ist. Kapitel 5.2.4 diskutiert das Ranking im Bezug auf Anforderung 5 und stellt eine zweite, für große Datenmengen erweiterte Version des Verfahrens vor. Kapitel 5.2.5 stellt die Ergebnisdarstellung entsprechend den Anforderungen 9-10 vor.

Die Diskussionen der bestehenden Lösungen im Bezug auf die möglichen Lösungswege sind im Folgenden eingerückt, um sie von dem Lösungsansatz selbst abzugrenzen.

5.2.1 Anfragestellung

Anforderung 1: Der hybride Ansatz soll formale, informale und hybride Anfragen verarbeiten können.

Anforderung 6: Die Benutzer sollen die formalen Teile der Anfrage leicht identifizieren und auswählen können

Anforderung 7: sowie nicht formale Anfragen und Anfrageteile ohne zusätzlichen Aufwand stellen können.

Anforderung 8: Die Benutzerschnittstelle soll die Stellung von Anfragen mit einer beliebigen Reihenfolge von formalen und informalen Anfrageteilen ermöglichen.

Suchanfragen werden in natürlicher Sprache, als Schlüsselwörter oder ganze Sätze, gestellt, jedoch mit so vielen formalen Anteilen, wie möglich. Die Benutzer sollen dabei kein Wissen über die formale Wissensbasis bzw. das verwendete Vokabular benötigen und das Gemeinte leicht identifizieren können. Hierzu eignet sich die **semantische Autovervollständigung**.

Die *Autovervollständigung* generiert dem Benutzer anhand der bereits eingetippten Zeichen und basierend auf dem zugrundeliegenden Vokabular automatisch Vorschläge für den jeweiligen Suchbegriff. Google setzt seit 2004 Autovervollständigung ein, ein aktueller Screenshot ist in Abbildung 5.1 zu sehen. Google Suggest erzeugt die Vorschläge dynamisch[1], d.h. die Vorschlagsliste wird beim jeden Hinzufügen oder Löschen eines Zeichens angepasst (dynamic query suggestion) [Hearst, 2009b]. Auf Google folgten alle großen Websuchmaschinen und heutzutage ist es ein selbstverständlicher Service [Baeza-Yates and Maarek, 2011]. Die dynamische Autovervollständigung bietet mehrere Vorteile: Sie hilft dem Benutzer Tippfehler zu vermeiden und das richtige Vokabular zu merken, reduziert den Aufwand bei der Eingabe und, als Nebeneffekt ermutigt sie den Benutzer längere, beschreibende Namen zu verwenden [Hyvönen and Mäkelä, 2006]. Die Realisierung muss jedoch effizient sein, die Vorschläge schnell (quasi verzögerungsfrei) zur Verfügung stellen und dabei den Benutzer nicht behindern. Und sie muss effektiv sein, d.h. möglichst hilfreiche, relevante Vorschläge anbieten [Baeza-Yates and Maarek, 2011]. Die Autovervollständigung basiert auf einem festen Vokabular, das in solch einer Datenstruktur abgebildet wird, die einen schnellen Abgleich mit dem Eingabepräfix zulässt. Wie bei den Information Retrieval Systemen für Dokumentsuche, eignet sich hierfür der Volltextindex (s. Kapitel 2.3.2). Das Vokabular bildet die Menge der Indexterme.

Die *semantische Autovervollständigung* schlägt passende Konzepte aus der Wissensbasis vor, deren Label jedoch, im Gegensatz zu der traditionellen Autovervollständigung, auf der lexikalischen Ebene nicht notwendigerweise ähnlich zu der Benutzereingabe sind [Osterhoff et al., 2012, Hyvönen and Mäkelä, 2006]. Synonyme, Abkürzungen,

[1] Weitere Autovervollständigungsvarianten sind *by request*, d.h. auf Aufforderung, z.B. durch das Drücken einer Taste, und *autoreplace*, das automatische Einsetzen einer Ergänzung anstatt eine Vorschlagsliste anzubieten [Hyvönen and Mäkelä, 2006].

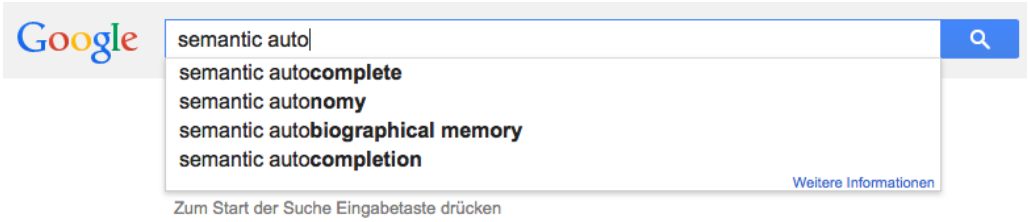


Abbildung 5.1: Screenshot der Google-Autovervollständigungskomponente (auf einem Desktoprechner)

Homonyme, Varianten in der Schreibweise, Singular/Plural und Phrasen müssen also erkannt werden, um Konzepte identifizieren zu können. Zudem könnte so eine Komponente auf den Präfix „Preside..“ gleich „George W. Bush“, eine Instanz der Klasse „President“, vorschlagen [Hyvönen and Mäkelä, 2006].

Die *Erkennung von Synonymen, Abkürzungen, Homonymen, Varianten in der Schreibweise, Singular/Plural und Phrasen*, um **semantische Autovervollständigung** durchführen zu können, ist in der hybriden semantischen Suchmaschine SINFIO folgendermaßen konzipiert:

- *Synonyme, Abkürzungen und verschiedene Schreibweisen* für das selbe Konzept müssen für eine effektive Autovervollständigung bereits im Index als zusätzliches Wissen vorliegen und nicht erst zum Zeitpunkt der Suche aus externen Quellen abgefragt werden. Hierfür wird im Rahmen der Indexierung zu den verfügbaren Labeln der Konzepte (Ontologie und Instanzen) das Online-Lexikon WordNet[2] nach Synonymen abgefragt und diese mit in den Index integriert. Werden keine Synonyme gefunden, so werden die Hyponyme aufgenommen, insofern diese nicht bereits als bestehendes Konzept in der Wissensbasis vorhanden sind. Fragewörter wie „wer“, „wen“, „wo“, „wohin“ usw. sind den entsprechenden Klassen der Ontologie, wie „Person“, „Location“, usw. manuell zugewiesen und ebenfalls indiziert.
- Die Erkennung von *Pluralformen* ist wichtig, da Wissensbasen Singular als Konzeptnamen verwenden. Sucht man aber beispielsweise die „Sternbilder“ oder „Hauptstädte afrikanischer Staaten“ mit dem Wissen, dass es mehrere gibt, so sollte dies auch im Plural erkannt werden. Hierfür können sogenannte Inflector-Bibliotheken eingesetzt werden, die sprachabhängig zu der Singularform eines Wortes dessen Plural und zum Plural den Singular generieren können. Eine weitere Möglichkeit besteht darin, ein Tool für die Indexierung und Abfragen des Indexes einzusetzen, das diese Funktionalität bereits integriert, wie z.B. Apache Lucene[3]. SINFIO basiert auf Lucene, eine externe Inflector-Bibliothek wurde nicht eingebunden.
- *Homonyme* werden durch Autovervollständigung disambiguiert, indem alle Bedeutungen dem Benutzer in einer erkennbaren Form angeboten werden. Welche Form erkennbar ist, wurde im Rahmen der Entwicklung der Benutzerschnittstelle evaluiert und wird in diesem Kapitel auf Seite 84 vorgestellt.
- Die Erkennung von *Phrasen* ist im Gegensatz zu der Autovervollständigung deshalb problematisch, weil Phrasen mehrere Konzepte beinhalten können. Es handelt

[2] `https://wordnet.princeton.edu` (06.01.2016)
[3] `http://lucene.apache.org` (06.01.2016)

sich also nicht um eine Präfixsuche. Beispielsweise könnten durch den syntaktischen Abgleich für „semantic web applications“ die Konzepte „semantic“, „web“, „semantic web“ und „web application“ gefunden werden. Deshalb müssen für die Erkennung von Phrasen Kombinationen der Suchterme durchlaufen werden und die Ergebnisse dem Benutzer entsprechend ihrem Rang präsentiert werden. Die Kombinationen sind diejenigen Permutationen der Terme, die deren ursprüngliche Reihenfolge in der Suchanfrage einhalten. Längere Phrasen sind höher gewichtet, um möglichst nah an der Benutzeranfrage zu liegen [Hyvönen and Mäkelä, 2006]. Insbesondere bei Suchanfragen mit formalen und informalen Anteilen in beliebiger Reihenfolge ist dies notwendig. Beispielsweise kann die Suchanfrage „semantic web in medical applications“, je nach Wissensbasis, als die Konzepte „semantic web“ und „medical applications“ oder als das Konzept „semantic web applications“ und freier Text „medical“ zu relevanten Ergebnissen führen. Wird eine Phrase eingegeben und wählt der Benutzer keinen der Konzeptvorschläge aus, so werden ab dem zweiten Term mögliche Kombinationen gesucht. Anfrageteile, die nicht durch die Auswahl eines Vorschlages bestätigt sind, sind informal, alle übernommenen Konzepte formal.

Die Suche nach den passenden Konzepten basiert auf einer *Termähnlichkeitsfunktion*. Die am häufigsten eingesetzten Funktionen sind die Hamming-, die Edit- und die N-Gram-Distanz (vgl. Kapitel 2.3.4.3). Die Hamming-Distanz zählt die unterschiedlichen Zeichen in einer Zeichenkette im Vergleich zu einer anderen. Fehlt durch einen Tippfehler ein Buchstabe, so werden alle darauffolgende Zeichen als falsch gezählt. Die Edit-Distanz hingegen betrachtet die Anzahl der Editieroperationen (Einfügen, Löschen und Ersetzen), die notwendig sind, um eine Zeichenkette in eine andere zu überführen und ist daher fehlertoleranter. Grzegorz Kondrak hat formal nachgewiesen, dass die Edit-Distanz ein Spezialfall der N-Gram-Distanz ist und gezeigt, dass die N-Gram-Methode mit $n \geqq 2$ bessere Ergebnisse liefert [Kondrak, 2005]. Deshalb wurde für den syntaktischen Vergleich die N-Gram-Methode ausgewählt. Am häufigsten wird $n = 3$ gewählt, da Trigramme sich für Terme diverser Längen am besten eignen [Kukich, 1992, van Berkel and Smedt, 1988]. Da sie jedoch für kurze Terme keine gute Ergebnisse liefern, werden häufig Bigramme für Terme bis 5 Zeichen und Trigramme für alle anderen Terme und Phrasen eingesetzt [Suen, 1979].

Die **Termähnlichkeit** zwischen den Suchtermen und den Konzeptlabeln wird basierend auf der N-Gram-Distanz mit einer Kombination von 2-Grammen bis 5 Zeichen und 3-Grammen für alle längeren Terme sowie der Dice-Ähnlichkeit[4] berechnet (s. Kapitel 2.3.4.3).

Die schematische Struktur des Konzeptindexes für die Autovervollständigung ist in der Tabelle 5.1 zu sehen. Sie beinhaltet den IRI, die Label, den spezifischsten Typ (Klasse) sowie Synonyme der Konzepte, wobei die Label und Synonyme in N-Gramme zerlegt sind.

Wie in der Formel 4.1 im Kapitel 4.2 definiert, bezeichnet $G_\Sigma = \{\langle s, p, o\rangle \mid s, p, o \in R, \langle s, p, o\rangle \; ist\; ein\; Tripel\; in\; der\; Wissensbasis\}$ die formale Wissensbasis inkl. Ontologien und Instanzen, Q die Menge der Suchanfragen. **Der syntaktische Abgleich** wird zwischen den Suchtermen bzw. Phrasen der Suchanfrage und den Labeln der Konzepte aus

[4] Sowohl die Jaccard- als auch die Dice-Ähnlichkeit liefern Werte in $[0, 1] \subset \mathbb{R}$ und sind für N-Gramme ausgelegt.

Merkmal	Wert
Label	⟨rdfs : Label⟩ der Ressource, N-Gram-Dekomposition
IRI	IRI der Ressource
Type	Ressource ist ⟨rdf : type⟩ von
Synonyms	Synonyme der Ressource, N-Gram-Dekomposition

Tabelle 5.1: Indexstruktur für die Autovervollständigung

der Wissensbasis (Ontologie und Instanzen), die um Synonyme, Abkürzungen, verschiedene Schreibweisen und Fragewörter erweitert ist, durchgeführt. Suchbegriffe bezeichnen im Folgenden die Suchterme und Ausdrücke, die durch den syntaktischen Vergleich identifiziert werden konnten. Teile der Suchanfrage, zu denen durch den syntaktischen Abgleich keine passenden Konzepte gefunden wurden, werden als einzelne Suchterme aufgefasst. Sei L_{G_Σ} diese Menge der Bezeichner der Konzepte und eine Suchanfrage q eine geordnete Liste von Suchbegriffen:

$$q = (t_1, ..., t_n), q \in Q, n \in \mathbb{N}. \tag{5.1}$$

Sei R_{G_Σ} die Menge der Ressourcen ohne Literale (RDFS definiert $\langle rdfs : Literal \rangle$ als Ressource [Brickley and Guha, 2004]):

$$R_{G_\Sigma} = \{r |\ \exists \langle r, p, o \rangle \in G_\Sigma, r\ ist\ kein\ Literal\}. \tag{5.2}$$

Dann ist das Ergebnis des syntaktischen Vergleichs für einen Suchbegriff $t_i \in q$ eine Menge von 3-Tupeln $(t_i, r, w_{t_i r})$:

$$M(t_i, G_\Sigma) = \{(t_i, r, w_{t_i r}) |\ t_i \in q,\ r \in R_{G_\Sigma},\ \exists \langle r, p, l_j \rangle \in G_\Sigma,\ l_j \in L_{G_\Sigma},\\ w_{t_i r} = dice - similarity(t_i, l_j), w_{t_i r} > H \in [0, 1] \subset \mathbb{R}\}. \tag{5.3}$$

Beispielsweise ist $M(t_i, G_\Sigma)$ für den Suchterm „author“ in der SWRC-Ontologie[5]: $M(author, swrc) = \{(author, \langle swrc : author \rangle, 1{,}0)\}$, da in der Ontologie nur die Property *author* zu dem Term „author“ gefunden wird. Werden mehrere Konzepte zu einem Suchterm mit einer lexikalischen Ähnlichkeit über einer vordefinierten Grenze H gefunden, so beinhaltet $M(t_i, G_\Sigma)$ je Konzept ein 3-Tupel. Diese Grenze H kann für jede Wissensbasis parametriert werden und bietet auch die Möglichkeit z.B. Properties gegenüber anderen Ressourcen zu priorisieren.

Für die Phrasensuche werden die Gewichte der potentiellen Phrasen aller Längen berechnet und jede Phrase mit den darin vorkommenden, durch den syntaktischen Abgleich ebenfalls als Konzept identifizierten Termen verglichen. Ist das Gewicht einer Phrase höher als der Durchschnitt der Gewichte der darin vorkommenden Terme, so wird angenommen, dass die Phrase gemeint war. Für das oben genannte Beispiel gilt: angenommen, dass Instanzen mit den Labeln $\langle semantic\ web\ application \rangle$, $\langle semantic\ web \rangle$ und $\langle semantic \rangle$ vorhanden sind. Es ist sofort ersichtlich, dass das Konzept $\langle semantic\ web\ application \rangle$ der Phrase „semantic web applications“ gleicht, das Gewicht ist in diesem Fall $1,0$, während alle anderen Phrasen ein Gewicht $< 1,0$ ergeben. Dies ändert sich sobald die

[5] Die Semantic Web for Research Communities, kurz SWRC-Ontologie ist für die Modellierung von Themen rund um die wissenschaftliche Arbeit entwickelt worden. `http://ontoware.org/swrc/` (06.10.2016)

Wissensbasis auch eine Instanz $\langle web\ application \rangle$ beinhaltet. In diesem Fall ergibt auch die Kombination von $\langle semantic \rangle$ und $\langle web\ application \rangle$ ein Gewicht von $1,0$. In solchen Fällen wird angenommen, dass die längere Phrase gemeint war. Erst wenn im semantischen Abgleich im Rahmen der Faktensuche keine Fakten gefunden werden, exploriert der Suchalgorithmus auch andere Phrasen. Das Gesamtergebnis des syntaktischen Vergleichs ist die Vereinigung der Teilergebnisse:

$$M(q, G_\Sigma) = \{\bigcup_{i=1}^{n} M(t_i, G_\Sigma)\}. \tag{5.4}$$

Die formalen Teile der Suchanfrage sind der Ausgangspunkt der Faktensuche. Die informalen Teile können ohne eine Verknüpfung zu formalen Daten nur für Dokumentretrieval eingesetzt werden. Da die Autovervollständigung nicht notwendigerweise benutzt wird, führt die Suchmaschine zuerst den syntaktischen Abgleich mit der formalen Wissensbasis aus. Nachteil hierbei ist, dass Konzepte gefunden werden können, die der Benutzer möglicherweise bewusst nicht ausgewählt hat.

Das Ergebnis ist eine vorverarbeitete Suchanfrage, in der die Suchbegriffe identifiziert wurden. Auf dieser Verarbeitungsstufe beinhaltet die Anfrage zu jedem identifizierten Suchbegriff eine der folgenden Erweiterungen:

- genau ein Konzept mit einem Gewicht von 1,0 aus der Autovervollständigung;
- mehrere Konzepte mit ihren N-Gram-Gewichten ($\in\in [0, 1] \subset \mathbb{R}$) aus dem syntaktischen Vergleich nach der Autovervollständigung;
- die Eigenschaft informaler Anfrageteil zu sein.

Die informalen Anteile der Suchanfrage werden für die Suche im Dokumentindex durch die Ergebnisse der Faktensuche erweitert. Dabei werden die Synonyme der identifizierten Konzepte für Anfrageerweiterung eingesetzt. Die genaue Vorgehensweise für den semantischen Abgleich wird in Kapitel 5.2.3 vorgestellt.

Grafische *Benutzerschnittstellen für Autovervollständigung* bieten die Vorschläge üblicherweise in Pop-Up-Menüs an [Hyvönen and Mäkelä, 2006]. Die Benutzer können durch einen Klick auf das Gemeinte schnell einen Vorschlag auswählen, die Auswahlliste behindert die Benutzer jedoch nicht. Häufig wird für eine noch einfachere Auswahl der am höchsten gewichtete Vorschlag bereits in die Eingabemaske geschrieben und kann durch die Pfeiltaste oder Enter übernommen werden (z.B. Google, s. Abbildung 5.1). Die Autovervollständigung wurde für mobile Endgeräte wie Smartphones und Tablets auf Betriebssystemebene umgesetzt, wie die T9-Funktion [Dunlop and Crossan, 2000]. Android integriert die Vorschläge in einen Balken als Aufsatz auf die Tastatur, iOS bietet jeweils ein Wort unter dem geschrieben Text an (s. Abbildung 5.2), das mit der Leertaste übernommen werden kann.

Die semantische Autovervollständigung sollte den Bezug des Vorschlags zum Suchbegriff kenntlich machen sowie die verschiedenen Bedeutungen eines Suchbegriffes ebenfalls erkennbar anbieten. Um dies zu erreichen, zeigen Lösungen den zugehörigen Ausschnitt aus der Konzepthierarchie oder gruppieren die Ergebnisse anhand ihrer Kategorie [Hyvönen and Mäkelä, 2006, Sinkkilä et al., 2008, Amin et al., 2009, Jung et al., 2009]. Beide Vorgehensweisen lassen sich jedoch nicht gut in Pop-up Menüs darbieten[6]

[6] [Jung et al., 2009] verwendet Pop-up-Menü, zeigt darin die Vorschläge nach abstrakten Typen, wie Person und Thema, gruppiert an. Es wird jedoch nur nach Instanzen gesucht.

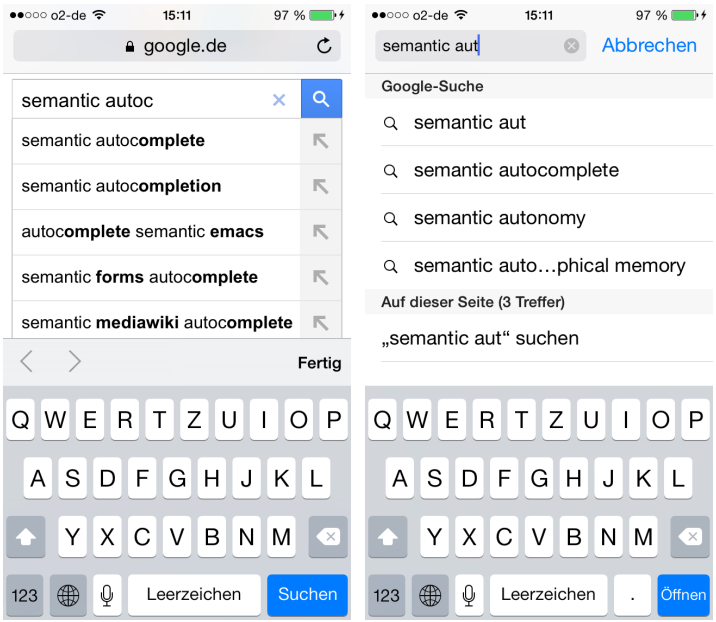


Abbildung 5.2: Screenshots der Autovervollständigung auf einem iPhone, auf der Google-Webseite (links) und in der Safari-Suchleiste (rechts)

und bedeuten einen Mehraufwand für den Benutzer. Die Suchmaschine NEXTBIO[7] gibt den Typ ($\langle rdf : type \rangle$, Klasse) der Instanzen an, um die Disambiguierung zu erleichtern [Hearst, 2011]. Diese Vorgehensweise ist geeignet für Domänensuchmaschinen, wobei je nach Komplexität und Hierarchiestruktur der Ontologie ein Domänenexperte festlegen sollte, welche Klassen dabei angezeigt werden. In domänenunabhängigen Suchmaschinen können ausgewählte abstrakte Konzepte, wie Person, Ort, Event oder die spezifischste Klasse einer Instanz in der Hierarchie, angezeigt werden.

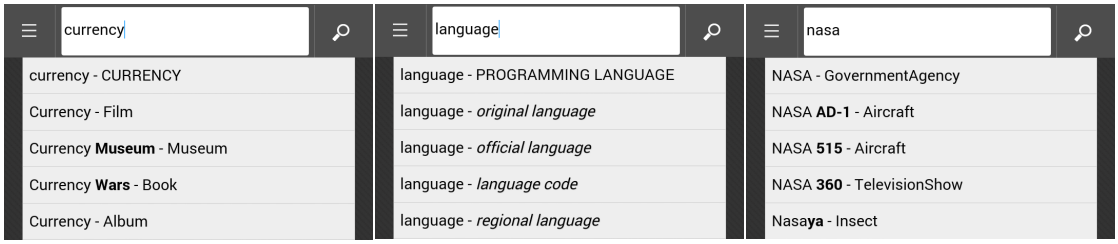


Abbildung 5.3: Vorschlagsliste der semantischen Autovervollständigung in SINFIO

Die vorgestellten **Benutzerschnittstellen für Autovervollständigung** eignen sich zur Disambiguierung, zeigen jedoch keinen Bezug zwischen den Vorschlägen und dem Synonym. Gibt der Benutzer einen Begriff ein, der in der Vorschlagsliste nicht wiederzufinden ist, steigt die kognitive Last, da die Informationen zusammengeführt und die Assoziation hergestellt werden muss (vgl. [Mayer and Moreno, 2003]). Aus diesem Grund wird in SINFIO bei der Autovervollständigung das Synonym, was der Benutzer gerade eintippen könnte, angezeigt. Der Vorschlag zum gerade eingegebenem Suchwort wird wie auch in gängigen Suchmaschinen (s. auch Abbildung 5.1 und 5.2) fett dargestellt ergänzt.

[7] `http://www.nextbio.com/b/nextbioCorp.nb` (06.01.2016)

Um die Interpretation des Vorschlags zu unterstützen, wird mit einem Bindestrich die Information zur vorgeschlagenen Ressource angefügt. Zur Disambiguierung von Instanzen ist dies die Typinformation, wobei hier entweder vorausgewählte Klassen aus der Klassenhierarchie oder die spezifischste Klasse in Frage kommen. Eine Unterscheidung der verschiedenen Typen (Instanz, Klasse, Property) wird visuell unterstützt: Propertynamen werden kursiv, Klassen in Großbuchstaben geschrieben, wobei die Schreibweise sich auch in der Ergebnisdarstellung wiederfindet. Der Benutzer braucht kein Wissen darüber was Klassen, Instanzen und Properties sind. Die visuelle Unterstützung zielt darauf ab, einen Bezug zwischen den Vorschlägen und deren Rolle in den Ergebnissen herzustellen, um eine Auswahl aus der Vorschlagsliste zu erleichtern. Abbildung 5.3 zeigt Beispiele. Um zu prüfen, ob solch eine Darstellung tatsächlich besser empfunden wird als eine Darstellung der Namen der gefundenen Konzepte, wurde eine entwicklungsbegleitende vergleichende Evaluierung durchgeführt. Zehn Studenten[8] haben jeweils 10 aus dem DBpedia-Query-Log ausgesuchte Suchanfragen gestellt und wurden befragt, welche Darstellung sie bevorzugen. Die Suchanfragen deckten Klassen, Instanzen und Properties sowie einzelne Suchterme und Phrasen ab. Alle Tester bevorzugten die hier vorgestellte Variante gegenüber der Darstellung, die anstatt des Synonyms den Namen des Konzeptes anzeigt.

5.2.2 Auswahl der Suchansätze

Tabelle 5.2 gibt einen Überblick der Suchansätze, bestehend aus einer kurzen Beschreibung des Ansatzes, die Angabe dessen, ob zusätzliches Wissen benötigt wird und ob der Ansatz sich für Faktensuche und semantische Dokumentsuche eignet. Für eine detailliertere Beschreibung siehe Kapitel 2.3.4.4.

Thesaurusbasierte Ansätze (s. Seite 31) sind der Vollständigkeit halber in der Tabelle mit aufgezählt. Sie sind auf einen Thesaurus als Wissensbasis ausgelegt, um durch den Kontext der Konzepte weitere geeignete Terme für die Suche zu finden. Sie eignen sich für Anfragemodifizierung. Auch dieser Lösungsansatz setzt die Synonyme der durch die Faktensuche gefundenen Konzepte zur Erweiterung der Anfrage für die Dokumentsuche ein. Da thesaurusbasierte Ansätze jedoch bei der Auswahl der Verfahren in Bezug auf die Anforderungen 2 und 3 keine Rolle spielen, werden sie hier nicht weiter betrachtet.

Anforderung 2: Der hybride Ansatz soll Suchanfragen mit Fakten, Dokumenten und hybriden Suchergebnissen beantworten können.

Das Verfahren zur Faktensuche liefert Entitäten und Fakten, die Dokumentsuche oder semantisches Dokumentretrieval Dokumente. Um auch hybride **Suchergebnisse** zu erhalten, wie in Formel 4.6 definiert, gibt es zwei Möglichkeiten:

- entweder werden die gefundenen Fakten und Dokumente miteinander kombiniert[9],
- oder diese als Ausgangspunkt für die weitere Suche eingesetzt.

Die Fakten enthalten jedoch nicht notwendigerweise Konzepte, die direkt mit einem Dokument verlinkt sind, es können über weitere Konzepte semantische Beziehungen bestehen. Abbildung 5.4 veranschaulicht die Möglichkeiten anhand der Graphdarstellung

[8] Für kleine Tests der Benutzerfreundlichkeit reicht es aus 5 Personen zu befragen [Nielsen, 2012a, Nielsen, 2000a].

[9] z.B. in K-Search, s. Seite 53.

Ansatz	Kurzbeschreibung	zusätzliches Wissen	Fakten-suche	sem. Dokument -suche
Thesaurus-basiert	Wissen aus Thesauri (wie Synonyme, Hyponyme, Hyperonyme) werden eingesetzt, um die Anfrage zu modifizieren.	Thesauri		✓
Graph-traversierung	Setzt Graphalgorithmen ein, um die Struktur des Graphen für die Suche auszunutzen. Hierzu gehört z.B. die Suche nach Punkt-zu-Punkt -Verbindungen (Faktensuche), aber auch Verfahren, die den Graphen in alle Richtungen explorieren (sem. Dokumentsuche).		✓	✓
Tripelbasiert	Erstellt aus den bereits gefundenen Properties, Instanzen und Klassen Anfrage-templates der Art ⟨c1, ?, c2⟩, um passende Tripel in der Wissensbasis zu finden.		✓	
Logisches Schließen	Reasoning nutzt die Transitivität von hierarchischen Relationen, Identitätsrelationen, Restriktionen durch Domäne- und Wertebereich von Properties, um neues Wissen abzuleiten. Für komplexere Reasoning-Aufgaben müssen aufgabenspezifische Reasoningpattern definiert werden.	Reasoning-Pattern/ Regeln	✓	
NLP-basiert	Setzt linguistisches Wissen zur morphologischen Analyse und Wissensextraktion aus den Dokumenten und der Anfrage ein. Das so gewonnene Wissen über die Rolle und Semantik der Begriffe wird für die Suche ausgenutzt.	ling. Wissen	✓	✓

Tabelle 5.2: Übersicht der Suchansätze

eines Ausschnittes aus der Wissensbasis. Die schattierten Knoten repräsentieren gefundene Konzepte und Dokumentinstanzen, die durchgehenden Kanten zeigen die Relationen zwischen ihnen. Die Konten A, B und C zeigen gefundene Fakten (bestehend aus zwei Tripeln) die nicht mit einem Dokument verbunden sind. Dokument 1 ist gefunden worden, bisher jedoch keine Metadaten dazu. Dokument 2 ist mit semantischen Metadaten (Fakten, Tripel mit Knoten D und E) abgebildet.

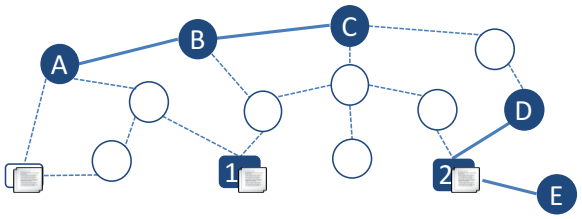


Abbildung 5.4: Beispiel für gefundene Konzepte (A, B, C, D, E), Dokumente (1, 2) und Fakten (A-B, B-C, 2-D, 2-E) im semantischen Graphen

Um Zusammenhänge zwischen nicht direkt verbundenen Fakten und Dokumente finden zu können, wird der Ansatz verfolgt, die Ergebnisse der Fakten- und Dokumentsuche als Ausgangspunkt für die weitere Suche zu verwenden. **Eine Kombination von Dokumenten und Fakten**, wie unter 4.7 definiert, ist *durch Verfahren möglich, die auf Basis der Ergebnisse der beiden Suchschritte weitersuchen können.* Betrachtet man die Graphdarstellung der Wissensbasis, so sind die Verfahren geeignet, die den zu einem Dokument gekoppelten zusammenhängenden Graphen, der die gefundenen Fakten beinhaltet, errechnen können. Nicht alle Ansätze sind geeignet:

- *Graphtraversierungsalgorithmen* (s. Seite 30) leisten genau das, was gefordert wird: Sie explorieren von einem oder mehreren Startknoten aus den Graphen mit dem Ziel, für die jeweilige Suchanfrage als relevant eingeschätzte Teilgraphen zu identifizieren. Instanziiert man die Dokumente in der Wissensbasis, so bilden die durch Faktensuche gefundene Konzepte und durch Dokumentsuche gefundene Dokumente die Menge der Startknoten. Die Wichtigkeit der Kanten kann in Form von Gewichten mit einbezogen werden, was bei den qualifizierten Relationen in formalen Wissensbasen ein wesentlicher Faktor ist.
- Der Einsatz von *tripelbasierten Verfahren* (s. Seite 30) wäre ebenfalls möglich aber ineffizient: Falls keine direkte Verknüpfung zwischen dem Dokument und den Fakten besteht, so müssten alle Pfade, die von dem Dokument oder von den Konzepten in den Fakten ausgehen, durch einzelne Abfragetemplates durchexploriert werden.
- *NLP-Verfahren* (s. Seite 31) konzentrieren sich auf Wissensextraktion und agieren vielmehr auf der Ebene natürlichsprachiger Sätze als in der formalen Wissensbasis. Sie setzen linguistisches Wissen ein, um sowohl die Dokumente als auch die Suchanfrage zu analysieren. Untersucht wird die Rolle der Terme in einem Satz, um Entitäten und Relationen erkennen und extrahieren zu können. Dabei können sich die Verfahren auf verschieden komplexe Strukturen konzentrieren, von einfachen Aussagen bis hin zu komplexen Sätzen z.B. mit Raum- und Zeitbezug.

  Da in diesem Lösungsansatz von einer bestehenden Wissensbasis ausgegangen wird, betrachten wir hier den semantischen Abgleich: Werden nur einfache Strukturen betrachtet, so entspricht der semantische Abgleich den tripelbasierten Ansätzen, wobei durch die Kenntnis der Rolle der Terme die Anzahl der Abfragen reduziert

werden kann. Komplexere Strukturen erfordern, dass diese in der formalen Wissensbasis abgebildet sind, was jedoch nicht immer der Fall ist. Von Nachteil wäre also, dass zusätzliches linguistisches Wissen benötigt wird und die Möglichkeiten durch die Struktur der Wissensbasis beschränkt sind.

- *Logisches Schließen* (s. Seite 31) konzentriert sich auf das Herleiten von neuem Wissen, indem Klassen- bzw. Propertyhierarchien verfolgt, Domäne und Wertebereichsrestriktionen von Properties und Identitätsrelationen genutzt oder auch die erstellten Domänenspezifischen Regeln angewendet werden. Mit diesen Mitteln lassen sich ebenfalls Zusammenhänge zwischen nicht direkt verbundenen Fakten finden.

Anforderung 3: Der hybride Ansatz soll nicht nur die vorhandenen Verknüpfungen zwischen dem Textindex und der Wissensbasis (Dokument - semantische Metadaten) ausnutzen, sondern auch nicht annotierte Dokumente finden.

Der hybride semantische Suchansatz soll nicht nur die **semantisch annotierte**, sondern auch **nicht mit den gefragten Konzepten annotierte**[10] oder **gar nicht annotierte Dokumente finden**, falls diese relevant sind. Deshalb wird die Dokumentmenge in jedem Fall durchsucht:

- Die vorhandenen Verknüpfungen zwischen der Wissensbasis und den Dokumenten (Dokument - semantische Metadaten) nutzen alle Verfahren zum semantischen Dokumentretrieval aus. Sie sind im Allgemeinen darauf ausgelegt durch Schlüsselwortsuche Dokumente zu finden und anhand der semantischen Metadaten der Dokumente die Wissensbasis zu explorieren, um weitere Dokumente zu finden. Auch Verfahren zur Faktensuche können die Verbindungen verfolgen, solange die Dokumente instanziiert, also in der formalen Wissensbasis repräsentiert sind. Es sind keine besonderen Schritte erforderlich und es besteht keine Einschränkung im Bezug auf die Auswahl der Verfahren.
- Um Dokumente zu finden, die relevant sein können aber nicht mit dem gefragten Konzept annotiert sind, führt das hybride Verfahren auch dann semantisches Dokumentretrieval aus, wenn nur Fakten gefunden wurden. Hierfür muss der Lösungsansatz auch von Fakten ausgehend die Dokumentmenge durchsuchen können. Hierdurch entsteht ebenfalls keine Einschränkung für die Auswahl der Verfahren, solange semantische Annotationen die Dokumente an die Wissensbasis koppeln.
- Nicht annotierte, also nicht an die Wissensbasis gekoppelte Dokumente können durch Schlüsselwortsuche gefunden werden. Werden also keine semantischen Daten zu einer Suchanfrage gefunden, so wird eine Schlüsselwortsuche auf dem Textindex durchgeführt.

Insgesamt stellt Anforderung 3 Restriktionen an den Gesamtablauf der Suche aber nicht an die konkreten Verfahren zur Fakten- und semantischer Dokumentsuche.

[10] Wie im Kapitel 4.1 beschrieben: die semantischen Metadaten decken selten die gesamte Inhalt der Dokumente ab, der für die Befriedigung des Informationsbedürfnisses relevant sein kann.

Anforderung 4: Der hybride Ansatz soll lexikalische und strukturelle Mehrdeutigkeiten in informalen und hybriden Anfragen auflösen können.

Die semantische Autovervollständigung (s. Kapitel 5.2.1) unterstützt die **manuelle Auflösung von lexikalischen Mehrdeutigkeiten**, indem dem Benutzer alle möglichen Bedeutungen eines Suchtermes zur Auswahl angeboten werden. Hierzu wird der Konzeptindex mit Wissen aus einem Thesaurus angereichert, also indirekt ein thesaurusbasierter Ansatz verfolgt. Wählt der Benutzer jedoch nichts aus, so kann der semantische Abgleich die Disambiguierung unterstützen.

Für die *automatische Disambiguierung lexikalischer Mehrdeutigkeiten* kann der Suchkontext eingesetzt werden. Als Kontext werden häufig die vorherigen Suchanfragen des Benutzers zur Modifizierung der Suchanfrage herangezogen [Guha et al., 2003, Stojanovic, 2003], es können aber auch weitere Informationen, die z.B. aus den gerade bearbeiteten Dokumenten oder Aufgaben des Benutzers extrahiert wurden [Schwarz, 2006, Grimnes et al., 2009], verwendet werden. Manche Systeme setzen hartkodierten Benutzerkontext in Form von Suchanfragekategorien, wie z.B. „location of...“, ein oder bieten die Möglichkeit, Kategorien bzw. Klassen als Kontext auszuwählen [Glover et al., 2001, Amaral et al., 2004, Hyvönen et al., 2003]. Die Zusatzinformationen helfen die Mehrdeutigkeiten aufzulösen, da durch sie weitere Konzepte und Dokumente gefunden werden und ihre Nähe zu den Konzepten, die die möglichen Bedeutungen repräsentieren, berechnet werden kann. Bei Suchanfragen aus mehreren Suchtermen können dies die weiteren Suchterme leisten.

Strukturelle Mehrdeutigkeiten bedeuten, dass die Struktur des Satzes oder eines komplexen Ausdruckes mehrdeutig ist (s. Kapitel 2.3.3). Beispielsweise ist in der Frage „Beobachtet Anne den Mann mit dem Fernglas?“ nicht klar, wer das Fernglas hat. Um so eine Mehrdeutigkeit manuell aufzulösen, müsste sie zuerst durch morphologische Analyse erkannt und die Interpretationsmöglichkeiten dem Benutzer zur Auswahl angeboten werden. Strukturell mehrdeutige Sätze (Anfragen) bleiben jedoch auch nach der morphologischen Analyse mehrdeutig und werden nur aufgelöst[11] [Etzioni et al., 2011, Fader et al., 2011], falls die Wissensbasis nur eine der Bedeutungen beinhaltet.

Sucht man *Verfahren* aus, *die die Struktur des Graphen ausnutzen, so kann die* **maschinelle Disambiguierung der lexikalischen und strukturellen Mehrdeutigkeiten** *unterstützt werden*, z.B. weil nur eine der Bedeutungen (Strukturen) in der Wissensbasis vorhanden ist oder weil andere gefundene Konzepte sich mit einer der möglichen Bedeutungen in engerem Bezug setzen lassen, d.h. sich in ihrem ontologischen Kontext befinden. Die Ansätze nutzen die Graphstruktur wie folgt:

- *Tripelbasierte Verfahren* generieren die Abfragetemplates basierend auf bereits gefundenen Konzepten. Erfüllt eine der möglichen Bedeutungen die Abfragen, so kann die lexikalische Mehrdeutigkeit aufgelöst werden. Bei längeren Suchanfragen mit mehreren zusammenhängenden Fakten kann an dieser Stelle das Ranking ebenfalls zur Disambiguierung beitragen, indem berücksichtigt wird, inwiefern eine Antwort die Terme der Suchanfrage „abdeckt“. Ähnlich verhält es sich bei strukturellen Mehrdeutigkeiten, wobei hier Abfragetemplates mit den unterschiedlichen Kombinationen generiert werden müssen. Für das Beispiel oben müssten also Abfragen

[11] Aus der Sicht der Suchmaschine, der Benutzer kann eine andere Bedeutung gemeint haben.

sowohl für „Anne“ und „Fernglas“ als auch für „Mann“ und „Fernglas“ ausgeführt werden. Sie eignen sich somit für die Disambiguierung bei der Faktensuche.

- Setzt man *Graphtraversierungsalgorithmen* ein, die einen Pfad zwischen zwei Knoten berechnen (Faktensuche), so kann die Länge des Pfades herangezogen werden, um lexikalische Mehrdeutigkeiten aufzulösen. Je kürzer der Pfad, umso engerer Zusammenhang kann angenommen werden: das Konzept, was näher an den durch andere Suchwörter gefundenen Konzepten wird ausgewählt. Strukturelle Mehrdeutigkeiten können graphbasierte Ansätze auflösen, falls nur eine der Bedeutungen in der Wissensbasis vorhanden ist. Wie bei tripelbasierten Verfahren sollte der Lösungsansatz alle Möglichkeiten betrachten. Dies geschieht bei den Verfahren automatisch, die von einer Menge von Startknoten heraus den Graphen explorieren (semantisches Dokumentretrieval). Graphbasierte Fakten- bzw. semantische Dokumentretrievalverfahren eignen sich also für die Disambiguierung bei der Faktenbzw. der Dokumentsuche. Für die Faktensuche sind tripelbasierte Verfahren effizienter, da sie nur zu der Suchanfrage passende Pfade verfolgen.
- *NLP-basierte Verfahren* können durch die Analyse der Anfrage, genauer durch den Kontext des Homonyms in der Anfrage, gegeben durch die weiteren Suchterme, lexikalische Mehrdeutigkeiten auflösen. Dies leisten jedoch auch tripelbasierte Verfahren und Graphtraversierungsalgorithmen im Rahmen des semantischen Abgleichs und kommen dabei ohne morphologische Analyse aus. Wie bereits erwähnt, kann die morphologische Analyse keine strukturellen Mehrdeutigkeiten auflösen. Nichtsdestotrotz können NLP-basierte Verfahren die Effizienz und Effektivität steigern, da die Rolle der Terme von vornherein bekannt ist und die Abfragen entsprechend zielgerichtet generiert werden können. Dafür muss jedoch linguistisches Wissen vorliegen und die Wissensbasis so strukturiert sein, dass die Rolle der Konzepte, im Sinne von Subjekt, Prädikat, Objekt, erkennbar ist.
- *Logisches Schließen* kann durch die Verfolgung von transitiven hierarchischen Relationen, Restriktionen der Domäne und des Wertbereichs von Properties sowie den OWL-Identity-Relationen zur Disambiguierung lexikalischer Mehrdeutigkeiten beitragen. Es können durch das logische Schließen Verbindungen zwischen den durch die Suchterme gefundenen Konzepte gefunden werden oder durch die Domäne-Wertebereich-Restriktionen bestimmte Interpretationsmöglichkeiten ausgeschlossen werden. Auch domänenspezifische Regeln können hierzu beitragen, indem nur passende Tripelketten verfolgt werden.

Auswahl der Verfahren

Tabelle 5.3 fasst die Ergebnisse der Analyse der Verfahren in Bezug auf die Anforderungen 2-4 zusammen. Dabei sind beide Kategorien der Graphtraversierungsalgorithmen vertreten, da sie sich entweder für Fakten- oder für semantisches Dokumentsuche eignen.

Im Bezug auf die **Anforderung 2** bestehen Einschränkungen auf die Auswahl der Verfahren durch die hybriden Ergebnisse. Diese können mithilfe von Graphtraversierungsverfahren, die für semantische Dokumentsuche eingesetzt werden, gefunden werden. Ebenso eignet sich logisches Schließen. Tripelbasierte Verfahren lassen sich zwar auch einsetzen, sie wären jedoch ineffizienter als diejenigen Graphtraversierungsverfahren, die den Graphen ohne einen konkreten Knoten als Ziel durchlaufen (s. Anforderung 2 auf Seite 86). **Anforderung 3** (s. Seite 89) kann durch alle Verfahren realisiert werden. **Anforderung 4** bezieht sich auf die Auflösung von Mehrdeutigkeiten. Lexikalische Mehrdeutigkeiten

Anforderung/ Suchansatz	tripel-basiert	Graphtraversierung		logisches Schließen	thesaurus -basiert	NLP-basiert
		FS	DS			
A2: Fakten, Dokumente und hybride Ergebnisse finden	✓	✗	✓	✓	✗	✗
A3: Verknüpfung zw. Wissensbasis und Dokumente nutzen	✓	✓	✓	✓	✓	✓
A4: Disambiguierung a) lexikalischer	✓	✓	✓	✓	✓	✗
b) struktureller Mehrdeutigkeiten	✓	✓	✓	✓	✗	✗

Tabelle 5.3: Überblick Anforderungen vs. Ansätze. FS ist Graphtraversierung für Faktensuche, DS für Dokumentsuche. ✓ steht für geeignet, ✗ für nicht geeignet.

werden anhand der Autovervollständigungskomponente manuell, durch den Benutzer aufgelöst. Zur Disambiguierung struktureller Mehrdeutigkeiten im Rahmen der Faktensuche eignen sich tripelbasierte Verfahren sowie logisches Schließen, im Rahmen der semantischen Dokumentsuche Verfahren der Graphtraversierung. NLP-Verfahren können lexikalische Mehrdeutigkeiten disambiguieren, strukturelle Mehrdeutigkeiten jedoch nicht bzw. nur durch den Einsatz von weiteren Verfahren (z.B. tripelbasiert) zum semantischen Abgleich (s. Anforderung 4 auf Seite 90).

Insgesamt lassen sich die Anforderungen durch folgende Kombinationen erfüllen:

a) Tripelbasiertes Verfahren für Faktensuche und ein Graphtraversierungsalgorithmus für semantisches Dokumentretrieval und für die hybride semantische Suche;
b) Logisches Schließen für Faktensuche und ein Graphtraversierungsalgorithmus für semantisches Dokumentretrieval und für die hybride semantische Suche;
c) Logisches Schließen für Faktensuche und traditionelle Schlüsselwortsuche in Dokumenten sowie logisches Schließen auf der Wissensbasis.

Logisches Schließen auf großen Datenmengen hat lange Laufzeiten. Bereits bei einer Datenmenge von 25 Millionen Tripeln kann die Ausführung einer Suchanfrage mit 3 Filtern über einer halben Minute liegen [Bizer and Schultz, 2008, Guo et al., 2005, Guo et al., 2004]. Die DBpedia Version 3.9 enthält 470 Millionen Tripel[12]. Aus diesem Grund wurde die Lösung *a*) gewählt: *SINFIO kombiniert ein tripelbasiertes Verfahren und ein Graphtraversierungsalgorithmus für die hybride semantische Suche.* Tripelbasierte Verfahren basieren ebenfalls auf Logik, indem sie die Ausdrucksmöglichkeiten bzw. die formale Logik von SPARQL nutzen und können als eine Instanziierung vom logischen Schließen angesehen werden. Die Mächtigkeit des Verfahrens hängt davon ab, welche Ausdrucksmöglichkeiten berücksichtigt werden. So können beispielsweise die Restriktionen durch Domäne und Wertebereich in die Abfragetemplates mit aufgenommen, hierarchische Relationen verfolgt oder Filter eingesetzt werden.

[12]Für detaillierte Angaben siehe `http://blog.dbpedia.org/?p=72` (07.10.2015).

5.2.3 Die hybride semantische Suchlösung

Die hybride semantische Suchlösung kombiniert Fakten- und Dokumentsuche basierend auf dem Graphtraversierungsverfahren Spreading Activation. Der semantische Abgleich geschieht dabei in mehreren Schritten, wie auch in Abbildung 5.5 veranschaulicht:

1. Faktensuche
2. Dokumentretrieval mit der erweiterten Suchanfrage
3. Aufbau des Spreading Activation Netzwerkes und Spreading Activation
4. Extraktion der Suchergebnisse

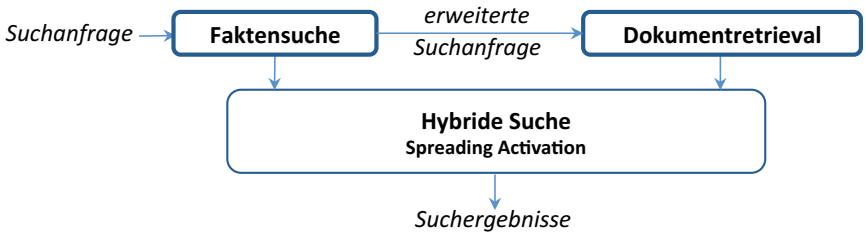


Abbildung 5.5: Übersicht der hybriden semantischen Suche

Werden Fakten gefunden, so wird hybride semantische Suche durchgeführt. Werden keine Fakten gefunden, so kann die Suchmaschine jedoch lediglich eine semantische Dokumentsuche durchführen.

In den folgenden Abschnitten wird die Faktensuche, die semantische Dokumentsuche und die hybride semantische Suche detailliert beschrieben. Da das Ranking integraler Bestandteil der Verfahren ist, wird dabei auch auf die Berechnung des Rankings der Ergebnisse eingegangen.

5.2.3.1 Faktensuche

Der Ausgangspunkt der Faktensuche ist die **Suchanfrage**, die

- formale Anteile mit einem Gewicht von 1,0 aus der Autovervollständigung,
- zu jedem Suchbegriff Konzepte aus der Wissensbasis mit ihren N-Gram-Gewichten (s. Formel 5.3) aus dem syntaktischen Vergleich nach der Autovervollständigung und
- natürlichsprachige Anteile (ohne Gewichtung, da durch den syntaktischen Vergleich keine passenden Konzepte gefunden wurden)

beinhalten kann.

Tripelbasierte Verfahren teilen die durch den syntaktischen Vergleich gefundenen Konzepte in die Menge der Properties p_i und Nicht-Properties n_j auf und generieren Abfragetemplates der Art $\langle n_j, p_i, ? \rangle$, $\langle ?, p_i, n_j \rangle$. Für die hybride Suchlösung wurde der tripelbasierte Ansatz von [Goldschmidt and Krishnamoorthy, 2005] adaptiert und erweitert, um auch Fakten, die aus mehr als zwei Tripeln bestehen, finden zu können. Für die formale Beschreibung des Prozesses seien die Mengen der Properties und Nicht-Properties wie folgt definiert:

$$P_{G_\Sigma} = \{p | \exists \langle p, rdf\colon type, rdf\colon Property \rangle \in G_\Sigma\} \tag{5.5}$$

$$NP_{G_\Sigma} = R_{G_\Sigma} \backslash P_{G_\Sigma}, \tag{5.6}$$

```
function semMatchOneTerm(q, G_I)
result = ∅
if ‖q‖ = 1, q = (t_1)
    for each r : (t_1, r, w_{t_1 r}) ∈ M(t_1, G_I)
        / * r is not a property * /
        if r ∈ NP_{G_I}
            / * r is a class: add instances of r to the results
            if r ∈ C_{G_I} : result = result ∪ {i ∈ G_I| ∃ ⟨i, rdftype, r⟩ ∈ G_Σ}
            / * r is an instance: add r to the results
            else result = result ∪ {r}
        / * r is a literal: add the triples with r to the results * /
        if r ∈ L_{G_I} : result = result ∪ {⟨s, p, r⟩ ∈ G_I| ∃(t_1, r, w_{t_1 r}) ∈ M(q, G_Σ)}
        / * r is a property: add all triples using r to the results * /
        if r ∈ P_{G_I} : result = result ∪ {⟨s, r, o⟩ ∈ G_I| ∃(t_1, r, w_{t_1 r}) ∈ M(q, G_Σ)}
return result
```

Abbildung 5.6: Semantischer Abgleich der Faktensuche für einen Suchbegriff

wobei R_{G_Σ}, wie in der Formel 5.2 definiert, die Menge der Ressourcen ohne Literale und Statements ist. Der semantische Abgleich im Rahmen der Faktensuche geschieht auf der Instanzbasis G_I. Analog zu $R_{G_\Sigma}, P_{G_\Sigma}$ und NP_{G_Σ} seien $R_{G_I} \subset R_{G_\Sigma}, P_{G_I} \subset P_{G_\Sigma}$ und $NP_{G_I} \subset NP_{G_\Sigma}$ die Menge der Ressourcen, Properties und Nicht-Properties in der Instanzbasis. Zusätzlich bezeichne $L_{G_I} \subset G_I$ die Literale in der Instanzbasis. Besteht die Suchanfrage aus einem Suchbegriff, so werden die gefundenen Instanzen, Instanzen der gefundenen Klassen sowie die Statements mit den gefundenen Properties und Literalen in die Ergebnismenge aufgenommen. Abbildung 5.6 gibt den Pseudocode des semantischen Abgleichs an. $M(q, G_\Sigma)$ repräsentiert die Menge der gewichteten Ergebnisse des syntaktischen Abgleichs und ist in Formel 5.4 definiert. Für die bessere Lesbarkeit des Pseudocodes wird noch die Menge der Klassen, $C_{G_I} \subset NP_{G_I}$, in der Instanzbasis definiert.

Beinhaltet die Suchanfrage mehrere Suchbegriffe, d.h. Suchterme/-Phrasen zu denen im Rahmen des syntaktischen Abgleichs Ressourcen gefunden wurden, so iteriert der semantische Abgleich zuerst über je zwei benachbarte Terme t_i, t_{i+1} und generiert danach Abfragen für alle Kombinationen der Ergebnisse aus der syntaktischen Suche $M(t_i, G_\Sigma)$ und $M(t_{i+1}, G_\Sigma)$. Dabei kann der Typ der Ressourcen berücksichtigt werden, um unnötige Abfragen zu vermeiden: unter Nicht-Properties wird zwischen Instanzen und Klassen, unter den Properties zwischen $\langle owl : ObjectProperty \rangle$ und $\langle owl : DatatypeProperty \rangle$ unterschieden. In kleinen Datenmengen besteht die Möglichkeit auch für zwei benachbarte Properties nach Tripeln zu suchen, die durch ein Konzept verbunden sind. Ebenso können Klassen durch ihre Instanzen ersetzt werden, falls die Abfragetemplates der Klasse mit den gefundenen Instanzen der benachbarten Terme keine Ergebnisse liefern. Dies kann jedoch nur für kleinere Datenmengen durchgeführt werden[13]. In großen Instanzbasen, wie z.B. DBpedia (ebenfalls für die Evaluierung verwendet, s. Kapitel 6.2) kann dies aus Effizienzgründen nicht durchgeführt werden (vgl. Kapitel 5.3). Solche Tripel können aber durch das Spreading Activation gefunden werden. Nachteil dabei ist, dass auch weitere Fakten gefunden werden können, die nicht Bestandteil der Suchanfrage sind und

[13] So eine kleine Datenmenge ist beispielsweise die Instanzbasis der Olympiade, die für eine Evaluierung verwendet wurde (s. Kapitel 6.1).

```
function semMatchFirst(q, G_I)
result = ∅
matched = ∅
if |q| > 1, q = t_1, ..., t_n
    / * for each pair of adjacent query terms * /
    for each (t_i, t_{i+1}) ∈ q, 1 ≤ i ≤ n − 1
        result(t_i, t_{i+1}) = ∅
        for each r_j : (t_i, r_j, w_{t_i r_j}) ∈ M(t_i, G_Σ)
            for each r_k : (t_{i+1}, r_k, w_{t_{i+1} r_k}) ∈ M(t_{i+1}, G_Σ)
                / * find suited triples in G_I * /
                result(t_i, t_{i+1}) = result(t_i, t_{i+1}) ∪ findStatements(r_j, r_k)
        / * if a query term is not matched * /
        if result(t_i, t_{i+1}) = ∅
            / * for each matched class of t_i * /
            for each r_c ∈ {r_c ∈ C_{G_I} | ∃(t_i, r_c, w_{t_i,r_c}) ∈ M(t_i, G_Σ)}
                / * replace the class by its instances and search again for triples * /
                for each r_i ∈ {∃ ⟨r_i, rdftype, r_c⟩ ∈ G_I}
                        for each r_k : (t_{i+1}, r_k, w_{t_{i+1} r_k}) ∈ M(t_{i+1}, G_Σ)
                        / * find suited triples in G_I * /
                            result(t_i, t_{i+1}) = result(t_i, t_{i+1}) ∪ findStatements(r_i, r_k)
            / * for each matched class of t_{i+1} * /
            for each r_c ∈ {r_c ∈ C_{G_I} | ∃(t_{i+1}, r_c, w_{t_{i+1},r_c}) ∈ M(t_{i+1}, G_Σ)}
                / * replace the class by its instances and search again for triples * /
                for each r_i ∈ {∃ ⟨r_i, rdftype, r_c⟩ ∈ G_I}
                        for each r_k : (t_i, r_k, w_{t_i r_k}) ∈ M(t_i, G_Σ)
                        / * find suited triples in G_I * /
                            result(t_i, t_{i+1}) = result(t_i, t_{i+1}) ∪ findStatements(r_k, r_i)
        if result(t_i, t_{i+1}) ≠ ∅
            matched = matched ∪ {t_i, t_{i+1}}
            result = result ∪ result(t_i, t_{i+1})
return (result, matched)
```

Abbildung 5.7: Semantischer Abgleich mit mehreren Suchbegriffen, Iteration über den Ergebnissen der syntaktischen Suche aller Suchbegriffe

nicht im Fokus des Benutzers liegen. Abbildung 5.7 zeigt den Ablauf Faktensuche für mehrere Suchbegriffe als Pseudocode.

In der zweiten Iteration werden die Ergebnistripel aus dem ersten Durchlauf und die Suchbegriffe, die bisher zu keinem Tripel geführt haben, ebenfalls paarweise der Reihenfolge nach betrachtet. Dabei werden Abfragen mit den Subjekten und Objekten der gefundenen Tripel und den Konzepten, die durch bisher nicht teilnehmenden $M(t_i, G_\Sigma)$ gefunden wurden, generiert. Der Prozess stoppt, wenn alle Suchbegriffe zu einem Ergebnis geführt haben oder keine neuen Tripel mehr gefunden werden. Abbildung 5.10 beschreibt den Prozess in Pseudocode. Durch die mehrfache Iteration werden auch Pfade über mehr als 2 Tripel gefunden, falls nicht vorher bereits eine Antwort gefunden wurde, die alle Suchbegriffe involviert. Die paarweise Verarbeitung von links nach rechts bewirkt, dass auch Aufzählungen von Instanzen, Klassen und Properties in der Suchanfrage verarbeitet werden können. Abbildung 5.9 veranschaulicht den Prozess anhand eines Beispiels.

Das **Ergebnis** des Prozesses können zusammenhängende Tripelmengen und einzelne

function $findStatements(r_j, r_k)$
/ * find RDF triples with r_j and r_k * /
if $r_j, r_k \in P_{G_I}$
    return $\{\langle s_1, r_j, o_1\rangle, \langle s_2, r_k, o_2\rangle \in G_I | o_1 = s_2 \vee s_1 = o_2 \vee s_1 = s_2\}$
else
    return $\{\langle s, p, o\rangle \in G_I | \{r_j, r_k\} \subseteq \{s, p, o\}\}$

Abbildung 5.8: Semantischer Abgleich, $findStatements(r_j, r_k)$

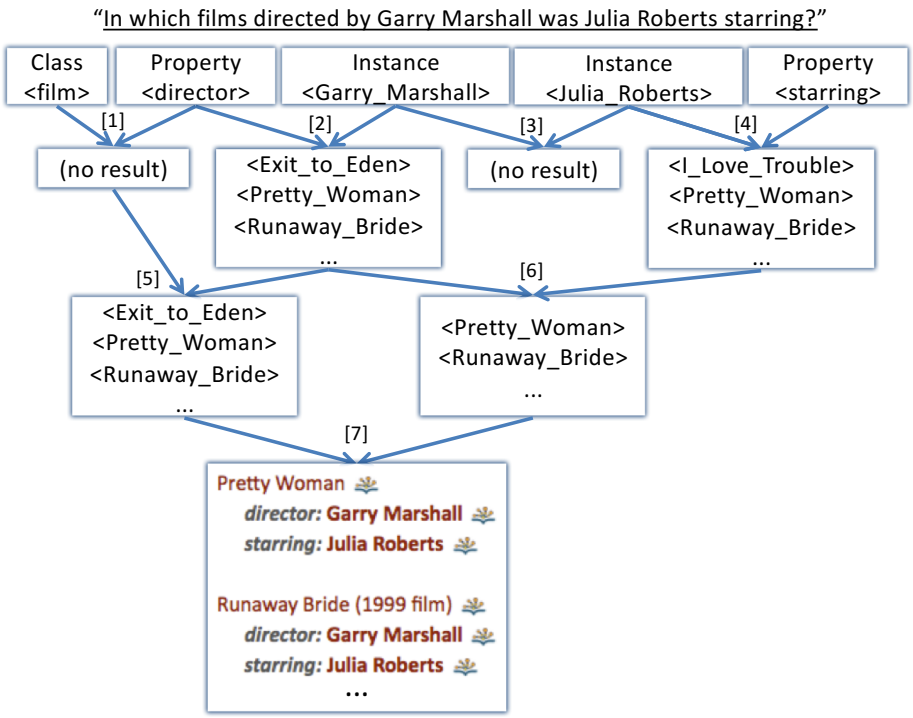


Abbildung 5.9: Faktensuche anhand eines Beispiels (basierend auf [Grimnes et al., 2009])

Instanzen sein. Im letzten Schritt sind daraus die Einzelergebnisse der Faktensuche zu bestimmen, indem (neben den Instanzen, falls welche vorliegen) Subgraphen als Antworten identifiziert werden. Um das Vereinen der Ergebnisse zu einem großen Graph zu vermeiden, werden hierbei nur Verbindungen über Instanzen betrachtet. Sind also zwei Instanzen nur über ihre Klasse verbunden, so bilden sie getrennte Ergebnisse. Die Suche nach den Filmen in Abbildung 5.9 ergibt Ergebnisse jeweils bestehend aus der Instanz eines Filmes und die Aussagen darüber, dass es sich um einen Film handelt sowie Garry Marshall der Regisseur und Julia Roberts die Hauptdarstellerin ist. Der Prozess wählt ein Tripel $\langle s_i, p_i, o_i\rangle$ und fügt jedes Tripel $\langle s_j, p_j, o_j\rangle$ mit $s_i = s_j \vee s_i = o_j \vee s_j = o_i \vee o_i = o_j,\ o_i \notin C_{G_i}$ hinzu, iteriert über die hinzugefügten Tripel, bis keine neuen Tripel mehr gefunden werden können. Das Ergebnis ist eine Menge von gewichteten Subgraphen M_{G_Σ} und gewichteten Instanzen.

Im Rahmen dieses Prozesses wird auch der **Rang der Ergebnisse** berechnet. Für die *Faktensuche* bietet die lexikalische Ähnlichkeit die Basis, um die möglicherweise intendierte Konzepte bereits in eine Rangordnung zu bringen. Der Rang einer Ressource entspricht ihrer lexikalischen Ähnlichkeit zum Suchbegriff (s. Formel 5.3). Für das Fakten-

function $semMatchSecond(q, G_I, result, matched)$
$/*$ iterate until all query terms are matched or no new triples found $*/$
$/*$ new is used to hold if an iteration returns new results $*/$
$new = true$
while $new \ \wedge (|matched| < |q|)$
  $new = false$
    for each $t_i \in q,\ t_i \notin matched$
      $/*$ for each adjacent term pair in q where one of them is not yet matched $*/$
      for each $t_j,\ j = i-1 \ \vee \ j = i+1$
        $part\_result = \emptyset$
        for each $r_i \in \{r_i |\ \exists(t_i, r_i, w_{t_i r_i}) \in M(t_i, G_\Sigma)\}$
          for each $r_j \in \{r_j |\ \exists(t_j, r_j, w_{t_j r_j}) \in M(t_j, G_\Sigma)\}$
            $part\_result = \ findStatements(r_j, r_i)$
            if $part\_result \neq \emptyset$
              $result = result \cup part\_result$
              $matched = matched \cup t_i$
              $new = true$
return $result$

Abbildung 5.10: Semantischer Abgleich, Iteration über die Ergebnisse aus Abbildung 5.7

ranking werden die Gewichte der gefundenen Ressourcen in einem Fakt bzw. in mehreren Fakten aufaddiert und zum Schluss durch die Anzahl der Suchbegriffe geteilt. Der Rang eines Graphes M_{G_Σ} ist:

$$SR_{q\_FactRetrieval} = \{(M_{G_\Sigma}, w_{M_{G_\Sigma}}) | \\ M_{G_\Sigma} \subset G_\Sigma,\ w_{M_{G_\Sigma}} = \frac{\sum_{r_j}^{(t_i, r_j, w_{t_i r_j}) \in M(t_i, M_{G_\Sigma})} w_{t_i r_j}}{|q|},\ w_{M_{G_\Sigma}} \in]0,1] \subset \mathbb{R}\} \quad (5.7)$$

Besteht der Graph aus einer Instanz, so entspricht der Rang des Graphen der lexikalischen Ähnlichkeit der Instanz zum Suchbegriff geteilt durch die Anzahl der Suchbegriffe.

5.2.3.2 Semantisches Dokumentretrieval

Liefert die Faktensuche keine Ergebnisse, so führt die hybride semantische Suchmaschine semantisches Dokumentretrieval durch. Im ersten Schritt erfolgt Schlüsselwortsuche auf dem Dokumentindex (syntaktischer Abgleich), für den semantischen Abgleich wird anschließend Spreading Activation durchgeführt.

Spreading Activation wird im Information Retrieval eingesetzt, um auf Basis von gefundenen Objekten durch die Exploration der Assoziationen in einem semantischen Netzwerk weitere relevante Information finden zu können [Crestani, 1997]. Die ontologischen Konzepte bilden die Knoten, die Properties die Kanten der Graphabbildung des Netzwerkes. Die Kanten sind üblicherweise gerichtet und gewichtet. Spreading Activation flutet den Graphen mit „Energie“, der Aktivierungsprozess startet bei initialen Knoten i mit einem Input-Gewicht I_i und propagiert die „Energie“ entlang der ausgehenden Kanten mit einem Output-Gewicht O_j zu den benachbarten Knoten j. Das

Input-Gewicht des Knotens j berechnet sich als

$$I_j = \sum_i O_i w_{ij}, \tag{5.8}$$

wobei w_{ij} das Gewicht der Kante von i zu j ist und die Stärke der Assoziation zwischen den Knoten ausdrückt. Als Output-Funktion $O_i = f(I_j)$ können verschiedene Funktionen eingesetzt werden, um z.B. eine lineare, schrittweise oder sigmoide Propagierung der Gewichte zu erreichen. Der Prozess der Aktivierung iteriert über neu aktivierte Knoten, flutet das semantische Netz mit Energie, bis die Stoppbedingung erfüllt ist. Ergebnis ist die Menge der aktivierten Knoten. Das Modell lässt sich mit weiteren Parametern, Bedingungen usw. erweitern und so für die jeweilige Aufgabe anpassen [Crestani, 1997].

Für die semantische Dokumentsuche wird das **semantische Netzwerk** wie folgt aufgebaut: Die Menge der Knoten N_{SN} bilden die Konzepte aus R_{G_I} ohne die Literale L_{G_I}, die Menge der Kanten E_{SN} die Properties zwischen diesen Knoten, wobei eine Kante durch ein Tripel und ein Gewicht beschrieben wird:

$$N_{SN} = \ R_{G_I} \backslash L_{G_I} \tag{5.9}$$

$$E_{SN} = \{(\langle s, e, o\rangle, w_d) | \exists \langle s, e, o\rangle \in G_I,\ s, o \in N,\ e \in P_{G_I},\ w_d \in]0, 1] \subset \mathbb{R}\} \tag{5.10}$$

Die Literale sind für das Spreading Activation nicht relevant, da sie nicht zur Entdeckung weiterer Assoziationen beitragen. Den Kanten wird ein Default-Gewicht w_d zugewiesen. Semantische Netze lassen sich als Adjazenzmatrix mit Kantengewichten abbilden, die Zeilen und Spalten bilden die Knoten $n \in N_{SN}$. Die Gewichtung der Kanten wird auch „Path-Constraint“ genannt, da es sich zum Bevorzugen bzw. Ausschließen von Pfaden eignet. Abbildung 5.11 beschreibt den Prozess.

```
1  matrix(|N_SN|, |N_SN|)
2  for each (⟨n_i, e, n_j⟩, w_d) ∈ E
3       if ∃ ⟨n_i, e, n_j⟩ ∈ G_I, ∃ ⟨n_j, e_inv, n_i⟩ ∈ G_I
4            matrix(n_i, n_j) = w_d
5       else if ∃ ⟨n_i, e, n_j⟩ ∈ G_I, ⟨n_j, e_inv, n_i⟩ ∉ G_I
6            matrix(n_i, n_j) = w_d
7            matrix(n_j, n_i) = w_d
8       else if ∃ ⟨n_i, e, n_j⟩ ∈ G_I, n_i ∉ C_{G_I}, n_j ∈ C_{G_I}
9            matrix(n_j, n_i) = w_d
```

Abbildung 5.11: Aufbau des semantischen Netzes als Matrix

Im Rahmen des Aufbaus des semantischen Netzes werden verschiedene Constraints berücksichtigt, um das Auffinden benachbarter Knoten zu ermöglichen, gleichzeitig aber ein unkontrolliertes Spreading zu vermeiden. Der Pseudocode in Abbildung 5.11 berücksichtigt, ob zu einer Porperty auch eine inverse Property, d.h. zu einer Kante $\langle n_i, e, n_j\rangle \in G_I$ auch eine Kante in eine andere Richtung, $\langle n_j, e_{inv}, n_i\rangle \in G_I$ existiert (Zeilen 3-4). Inverse Properties sind die Umkehrung von Properties im Sinne ihrer Domäne und Wertbereich. Definiert z.B. eine Ontologie die Eigenschaft $\langle starring\rangle$, die Domäne $\langle Work\rangle$ und den

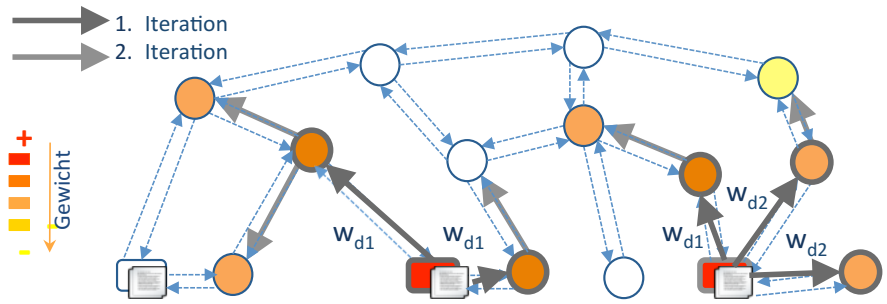


Abbildung 5.12: Spreading Activation, zwei Iterationen von den initialen Knoten (rot) aus

Wertbereich $\langle Actor \rangle$, dann hätte die inverse Property die Domäne $\langle Actor \rangle$ und den Wertbereich $\langle Work \rangle$ (vgl. Kapitel 2.2). Nicht vorhandene Inverse Properties werden beim Aufbau des des semantischen Netzes eingefügt, indem jede Kante in beide Richtungen eingetragen wird (Zeilen 5-7 in Abbildung 5.11). Die letzte Bedingung verhindert, dass von einer Instanz aus über ihre Klasse alle anderen Instanzen derselben Klasse geflutet werden. Andernfalls würde der SA-Algorithmus bei einer Suche nach dem Film „Pretty Woman“ die Instanzen aller anderen Filme finden. Die Instanzen einer durch den syntaktischen Vergleich gefundenen Klasse werden jedoch verbunden (Zeilen 8-9).

Die Menge der initialen Aktivierungsknoten für das **Spreading Activation (SA)** bilden die Instanzen der durch Schlüsselwortsuche gefundene Dokumente aus dem Dokumentkorpus D, ihr Input-Gewicht ist ihr Rang:

$$keyword\_search(q, D) = \{(d_l, w_{d_l}) | d_l \in D, w_{d_l} \in]0, 1] \subset \mathbb{R}\} \tag{5.11}$$

Allen anderen Knoten wird ein Gewicht von 0 zugeordnet. Die Aktivierung startet von den assoziierten Instanzen r_{d_i} der gefundenen Dokumente aus (s. Definition 4.5). Die Aktivierungsfunktion aus der Formel 5.8 wird durch den Auskühlungsfaktor (den sogenannten „loss of energy“ [Crestani, 1997]) α erweitert, um die Stärke der Aktivierung bei jeder Propagierung zu vermindern:

$$I_j = \sum_i O_i w_{ij} (1 - \alpha). \tag{5.12}$$

Um Endlosschleifen zu vermeiden ist sicherzustellen, dass jede Kante nur einmal betrachtet wird. Besteht zwischen zwei Knoten je eine Kante in beide Richtungen, so ist eine Aktivierung von einem zu dem anderen Knoten nur einmal auszuführen. In jeder Iteration wird der Knoten mit dem höchsten Gewicht als erstes betrachtet, propagiert wird entlang der noch nicht verarbeiteten Kanten. Gewichte bereits aktivierter Knoten werden nur mit höheren Gewichten überschrieben. Als Stoppbedingung wird das sogenannte *Activation-Constraint* eingesetzt: nur Knoten mit einem $I_j > activation\_threshold \in \mathbb{R}$ werden verfolgt. Wie in [Rocha et al., 2004] wurde die lineare Output-Funktion $O_i = I_i$ ausgewählt, da α bereits die Aktivierungsstärke bei jeder Iteration vermindert und Kantengewichte <1,0 vergeben. Zudem erreichten [Rocha et al., 2004] mit diesem Setup die besten Ergebnisse bei den Testläufen. Mit der Stoppbedingung zusammen ist O_i definiert als:

$$O_i = \begin{cases} I_i & I_i \geq activation\_threshold \\ 0 & I_i < activation\_threshold. \end{cases}$$

Abbildung 5.12 veranschaulicht den Spreading Activation-Prozess.

Durch die Wahl der Funktion O_i, der Kantengewichte w_{ij} und α kann das Verhalten des SA beeinflusst werden. Durch Kantengewichte kann die Wichtigkeit einer Property ausgedrückt werden. Der Faktor α, zusammen mit dem Activation Constraint, eignet sich zur domänenspezifischen Parametrierung. Je größer α ist, umso stärker werden kurze Pfade präferiert [Rocha et al., 2004]. Zudem gibt es noch weitere Constraints, die von der jeweiligen Domäne abhängig eingesetzt und parametriert werden können [Rocha et al., 2004]:

- Das Konzept-Constraint stoppt die Aktivierung von Knoten eines vordefinierten Typs aus.
- Das Fan-Out-Constraint stoppt die Aktivierung von Knoten aus, deren Anzahl ausgehender Kanten eine vordefinierte Grenze überschreitet.
- Das Distanz-Constraint stoppt die Aktivierung, wenn eine vordefinierte Distanz von dem initialen Knoten aus erreicht wurde.

Diese Constraints wurden im Rahmen der protoypischen Realisierung implementiert und können für die jeweilige Testdaten parametriert werden (s. Kapitel 5.3.1).

Das **Ergebnis** des Spreading Activation-Prozesses ist eine gewichtete Menge von Dokumentinstanzen. Die semantische Dokumentsuche liefert die Liste der zugehörigen Dokumente:

$$SR_{q\_SemanticDocumentRetrieval} = \{(d_i, w_{d_i})| \\ d_i \in D, w_{d_i} = w_{r_{d_i}} = matrix(w_{r_{d_i}}, w_{r_{d_i}}) \in]0,1]\mathbb{R}\} \quad (5.13)$$

Besitzt ein Dokument d_i keine semantische Metadaten, so entspricht der Rang des Dokumentes seines initialen Aktivierungsgewichtes $w_{d_i} = w_{r_{d_i}} = w_{d_l}$.

5.2.3.3 Hybride semantische Suche

Die Kombination der Ergebnisse der Faktensuche und der Dokumentsuche erfolgt ebenfalls über Spreading Activation. Abbildung 5.13 veranschaulicht die Vorgehensweise.

Synonyme und alternative Schreibweisen der durch Faktensuche gefundenen Ressourcen (vgl. Kapitel 5.2.1) werden für die **Anfrageerweiterung** eingesetzt[14] und anschließend *traditionelles Dokumentretrieval* auf dem Volltextindex (s. Def. 4.3) durchgeführt. Von einer Anfrageerweiterung mit dem ontologischen Kontext der gefundenen Ressourcen wird abgesehen, da der Kontext durch Spreading Activation automatisch exploriert wird. Eine zusätzliche initiale Aktivierung dieser Knoten in Wissensbasen, die nicht linguistische Beziehungen abbilden, mindert den Einfluss der Kantengewichte und kann zu einer niedrigeren Präzision der Ergebnisse führen.

Als Ausgangspunkt für die Kombination liegen bereits gefundene gewichtete Fakten bzw. Ressourcen (*result* in Formel 5.10) und gerankte Dokumente ((d_l, w_{d_l}) in Formel 5.11) vor. Um das **semantische Netz für Spreading Activation** aufzubauen, werden

[14] Spreading Activation eignet sich auch für implizite maschinelle graphbasierte Anfrageoptimierung (vgl. Kapitel 2.3.4.3). So könnten jedoch nur die Dokumente gefunden werden, die mit den entsprechenden Konzepten annotiert sind.

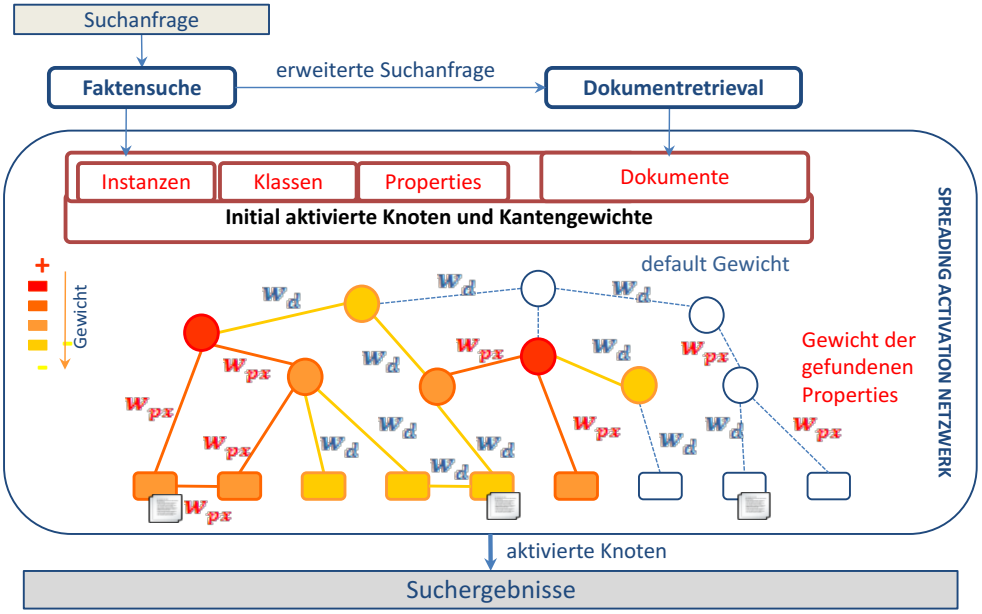


Abbildung 5.13: Übersicht der hybriden semantischen Suche [Schumacher et al., 2008]

die gefundenen Ressourcen in die Menge der Instanzen I_{FR}, Klassen C_{FR} und Properties P_{FR} unterteilt:

$$C_{FR} = \{c \in C_{G_I} | \exists (t_k, c, w_{t_k,c}) \in M(q, G_\Sigma)\} \tag{5.14}$$

$$P_{FR} = \{p \in P_{G_\Sigma} | \exists (t_k, p, w_{t_k,p}) \in M(q, G_\Sigma)\}. \tag{5.15}$$

Die Instanzmenge wird dabei um die Instanzen aus den gefundenen Fakten erweitert:

$$I_{FR} = \{i \in R_{G_I} \backslash (C_{G_I} \cup P_{G_I}) | \exists (t_k, i, w_{t_k,i}) \in M(q, G_\Sigma) \ \vee \ \exists \langle s, p, o \rangle \in result, \ i = s \vee i = o\}. \tag{5.16}$$

Die Menge der initialen Aktivierungsknoten bilden die assoziierten Instanzen r_{d_i} der gefundenen Dokumente d_i, die Menge der Instanzen I_{FR} und Klassen C_{FR}, Inputgewicht ist jeweils ihr Gewicht aus der Suche auf dem Dokumentindex bzw. Faktensuche. Das Gewicht der durch die Faktensuche gefundenen und hier neu hinzukommenden Instanzen ist das Gewicht des Subgraphen M_{G_Σ}, dessen Teil die Instanz bildet:

$$w_i = \{w_{M_{G_\Sigma}} | M_{G_\Sigma} \in M(q, G_\Sigma), \langle i, p, o \rangle \subseteq M_{G_\Sigma} \vee \langle s, p, i \rangle \subseteq M_{G_\Sigma}\}. \tag{5.17}$$

Die Gewichte der durch den syntaktischen Vergleich gefundenen Properties $p \in P_{FR}$ ersetzen das niedrigere Defaultgewicht w_d der entsprechenden Kanten e (s. Abbildung 5.14).

Für das **Spreading Activation** werden dieselben Funktionen und Constraints eingesetzt wie beim semantischen Dokumentretrieval (s. Kapitel 5.2.3.2).

Das **Ergebnis** der Aktivierung ist eine Menge von gewichteten Knoten, die auch Dokumentinstanzen enthält. Der letzte Schritt besteht aus der Extraktion von zusammenhängenden (aktivierten) Graphen, die, falls vorhanden, mit der durch die Faktensuche

for each $e \subseteq E$
  if $\exists(t_k, e, w_{t_k,e}) \in M(q, G_\Sigma),\ e \in P_{G_I}$
    $w_e = w_{t_k,e}$
  else
    $w_e = w_d$

Abbildung 5.14: Network Edges Setup for the Combined Approach

gefundenen $\langle owl : DatatypeProperty \rangle$ erweitert werden. Das Gewicht eines Subgraphen M_{G_Σ} ist der Durchschnitt der Gewichte der darin vorkommenden Knoten:

$$SR_{q\_HybridSearch} = \{(M_{G_\Sigma}, w_{M_{G_\Sigma}})| \\ M_{G_\Sigma} \subset G_\Sigma,\ w_{M_{G_\Sigma}} = \frac{\sum_{r_j}^{r_j \in M_{G_\Sigma}} matrix(r_j r_j)}{|M_{G_\Sigma}|},\ w_{M_{G_\Sigma}} \in]0,1]\mathbb{R}\} \quad (5.18)$$

5.2.4 Ranking

Anforderung 5: Der hybride Ansatz soll die Umsetzung einer adäquaten Rankingfunktion, welche für hybriden Ergebnisse und Ergebnislisten geeignet ist, unterstützen.

Semantisches Dokumentretrieval und Faktensuche setzen unterschiedliche Rankingverfahren ein (vgl. Kapitel 3.3). Für die Umsetzung einer **adäquaten Rankingfunktion für hybride semantische Suche, welche hybride Ergebnisse und Ergebnislisten unterstützt**, bedarf es also eines Rankingmodells, das unterschiedliche Rankingfunktionen miteinander verbinden kann. Dies beinhaltet die *Anforderung*, dass die Rankingfunktion der Faktensuche und der Dokumentsuche Werte im selben Intervall liefern oder dass die Unterschiede in den Wertebereichen zum Zeitpunkt der Verbindung passend angeglichen werden. Die Verbindung ist durch das Spreading Activation gegeben, wonach alle Ergebnisse nach dem selben Verfahren gerankt sind.

Für das **Dokumentretrieval** wird das Vektorraummodell mit *tfidf*-Gewichtung eingesetzt (vgl. Formel 5.11). Der Wertebereich des Rangs ist durch das Ähnlichkeitsmaß bestimmt. Bei Vektoren mit positiven Gewichten, wie die Dokument- und Termvektoren, liefert das Kosinus-Maß reelle Zahlen im Intervall $[0, 1] \subset \mathbb{R}$ (vgl. Formel 2.2).

Für die **Faktensuche** bietet die lexikalische Ähnlichkeit die Basis, um die durch den syntaktischen Abgleich gefundenen Konzepte in eine Rangordnung zu bringen. Wie im Kapitel 5.2.1 vorgestellt, wird die lexikalische Ähnlichkeit bereits zur Anfragestellung basierend auf der N-Gram-Methode und der Dice-Maß mit Werten im Intervall $[0, 1]$ berechnet.

Semantisches Dokumentretrieval wird dann ausgeführt, wenn keine Fakten gefunden worden sind. Der Rang der Ergebnisse ergibt sich aus dem Spreading Activation.

Bei der **hybriden semantischen Suche** fließen die Ergebnisse der Fakten- und traditionellen Dokumentsuche in das Spreading Activation als initiale Gewichte ein, und zwar jeweils aus dem Intervall $[0, 1] \subset \mathbb{R}$. Die Endergebnisse bekommen den Durchschnitt der Gewichte der teilnehmenden Knoten als Rang (für die genaue Rechenvorschrift s. Kapitel

5.2.3).

Bei großen Datenmengen können Suchbegriffe zu langen Ergebnislisten führen, insbesondere, wenn der Benutzer bei der Eingabe des Suchwortes kein Konzept aus der Autovervollständigungsliste auswählt[15]. Um dieses Problem abzuschwächen, wurde für SINFIO eine **zweite, angepasste Version des Rankingverfahrens** entwickelt, das die Popularität der Konzepte mit einbezieht.

Die Popularität, auch Konnektivität genannt, lässt sich global, in Relation zu der Gesamtkonzeptmenge (z.B. [Aleman-Meza et al., 2005]), oder lokal, bezüglich der Ergebnismenge (z.B. [He et al., 2007]) berechnen. Für SINFIO wurde die globale Methode aus [Aleman-Meza et al., 2005] eingesetzt, um die Vergleichbarkeit der Ergebnisse unterschiedlicher Suchanfragen gewährleisten zu können. Die Popularität wird für die Entitäten (aber nicht für Klassen oder Instanzen, die für eine Kategorie stehen) berechnet und bezüglich ihrer Klasse ($\langle rdf : type \rangle$) normalisiert, da es Klassen gibt, die in Relation zu anderen Klassen tendenziell eine hohe Konnektivität haben (z.B. Country). Bezeichne $c_{k,i},\ 1 \leq i \leq n$ eine Entität der Klasse k, wobei der Klasse n Entitäten angehören. Bezeichne $connectivity_{c_{k,i}}$ die Anzahl ein- und ausgehender Kanten der Entität $c_{k,i}$ und sei $max(connectivity_k)$ die Anzahl ein- und ausgehender Kanten der Entität mit der höchsten Konnektivität. Dann ist die Popularität von $c_{k,i}$:

$$pop_{c_{k,i}} = \frac{connectivity_{c_{k,i}}}{max(connectivity_k)}. \tag{5.19}$$

Die Popularität fließt im Rahmen des Spreading Activation in das Ranking ein, indem auf das Gewicht von jedem aktivierten Knoten c einmalig pop_c aufaddiert wird. Da die Dokumente im Activation Network instanziiert und ebenfalls (durch ihre semantischen Metadaten) vernetzt sind, kann ihre Konnektivität ebenfalls bewertet werden. Das Gewicht zum Zeitpunkt $t + 1$ ist:

$$w_{c_{t+1}} = w_{c_t} + pop_c. \tag{5.20}$$

Die Popularität ist daher ein Boost-Faktor und hat genau dann eine besondere Auswirkung auf das Ranking, wenn der Benutzer bei der Eingabe der Anfrage kein Konzept ausgewählt hat.

Welches der beiden Rankingverfahren eine stärkere Korrelation mit dem Gold Standard aufweist und besser für den hybriden Ansatz geeignet ist, wird in einer vergleichenden Evaluation untersucht (s. Kapitel 6.2.3).

Zusammengefasst führt SINFIO semantisches Ranking durch, das die lexikalische Ähnlichkeit, den Rang der Dokumente, die Distanz von Konzepten (durch Spreading Activation) und die Abdeckung der Suchanfrage mit einbezieht und die Anforderung 5 erfüllt.

[15] So gibt es z.B. in DBpedia v3.9 über 600 Ergebnisse, die „Berlin“ im Label beinhalten.

5.2.5 Die Benutzerschnittstelle

Anforderung 9: Die Darstellung der Ergebnistypen (Fakten, Dokumente und hybride Ergebnisse) soll verständlich sein,

Anforderung 10: so dass diese eine möglichst ähnliche Darstellungsform aufweisen.

Anforderung 11: Die Darstellung der Ergebnisliste selber im Bezug auf die Anordnung der Ergebnisse soll ebenfalls verständlich sein.

Die hier beschriebenen Ergebnisse sind nach dem Prinzip des User Centered Designs (UCD) [Abras et al., 2004] entstanden: Das Design, Testen und Anpassen der Benutzerschnittstelle erfolgte in mehreren Iterationen. Nach jedem Design- bzw. Redesign-Schritt wurden die Komponenten in Form von disfunktionalen oder funktionalen Prototypen Benutzerbefragungen (nach [Nielsen and Landauer, 1993, Nielsen, 2000b, Nielsen, 2012b]) unterzogen. Die Ergebnisse flossen in den Designprozess zurück.

Um die kognitive Last bei der Interpretation der Ergebnisse zu mindern, sollten die unterschiedlichen Ergebnistypen verständlich und mit möglichst ähnlicher Struktur abgebildet werden (vgl. Kapitel 5.1.2).

Für **Textdokumente** wird der Titel und ein zusammengesetztes Textsnippet, das die Suchwörter mit ihren textuellen Kontext im Dokument zeigt („keywords in context“) eingesetzt. Die Suchwörter sind durch fette Formatierung hervorgehoben. Diese Darstellung basiert auf mehrere Benutzerstudien [Tombros and Sanderson, 1998, White et al., 2003, Luhn, 1960], sie ist den Benutzern durch die gängigen Suchmaschinen bekannt und hat sich gegenüber andere Darstellungsformen durchgesetzt.

Fakten werden entweder als Graphen dargestellt oder als strukturierter Text abgebildet (vgl. Kapitel 5.1.2). Die Graphdarstellung hat für hybride semantische Suche zwei Nachteile:

- die Struktur unterscheidet sich stark von der Struktur der textuellen Dokumentdarstellung, und
- eine Verbindung mit der Dokumentdarstellung bei hybriden Ergebnissen ist schwer grafisch abzubilden.

Für die erste Version des Oberflächenprototypen wurden beide Varianten implementiert und eine Benutzerbefragung mit 5 Studenten[16] durchgeführt. Die Studenten konnten frei 10 Suchanfragen stellen, die zugrundeliegende Wissensbasis beinhaltete die Fakten und Dokumente der vierten Europäischen Semantic Web Konferenz. Sie wurden gefragt, welche Darstellung sie bevorzugen und warum. Die Studenten präferierten die textuelle Darstellung, da diese sich besser in die Ergebnisliste eingegliedert hatte und leichter zu verstehen war. Abbildung 5.15 zeigt jeweils ein Beispiel für beide Darstellungsarten.

In die textuelle Darstellung der Fakten können Dokumente eingegliedert werden, indem sie mit Titel und Textsnippet als das oberste Element angezeigt und die zugehörigen Fakten darunter gelistet werden. Abbildung 5.16 zeigt ein Beispiel, ein Konferenzpapier mit Autoren, aus einer der ersten Design-Versionen.

Im Rahmen dieser Arbeit wurde die grafische Benutzerschnittstelle sowohl für PCs als auch für SmartPhones konzipiert. Der Design-Prozess startete mit der Analyse von

[16] Für kleine Tests der Benutzerfreundlichkeit reicht es aus 5 Personen zu befragen [Nielsen, 2012a, Nielsen, 2000a].

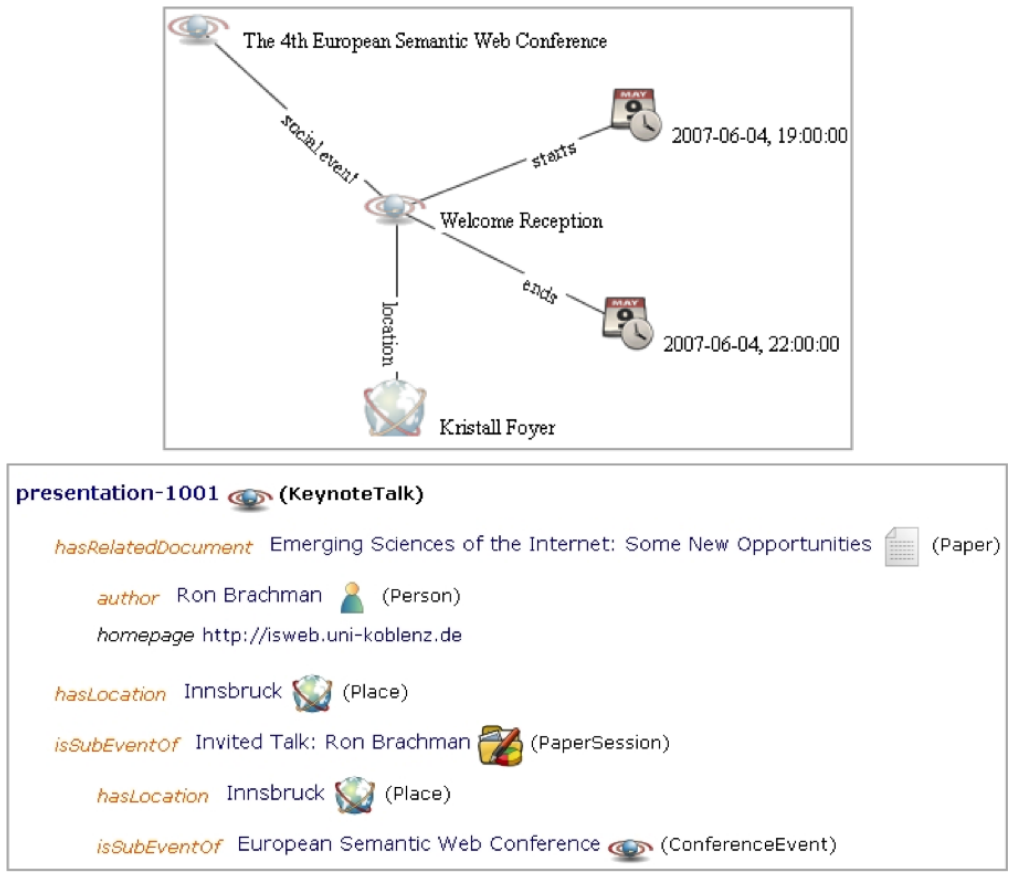


Abbildung 5.15: Graphdarstellung (oben) vs. textuelle Darstellung (unten) von Fakten

bestehenden Suchinterfaces, wie Google, Alexandria, DuckDuckGo, Yahoo, die Amazon App, Google Play, Yassu, Flipboard, Aurora Universal Search, You Version - Die Bibel, Recipe Search, Yellow Pages, WolframAlpha usw., um für eine hybride Suchmaschine geeignete Konzepte zu sammeln. Die Analyse und das Design baute auf bestehende Designrichtlinien, Vorschlägen und Ergebnissen aus Benutzerstudien (u. a. [Hearst, 2009a, Nielsen and Budiu, 2013, Blackler et al., 2005]). Auf Basis der Ergebnisse entstand der Design-Styleguide (Gestaltungsrichtlinien) und Wireframes (konzeptioneller Entwurf). Abbildung 5.17 zeigt die Wireframes des mobilen Interfaces für die Startseite, auf dem eine zufällige DBpedia-Instanz gezeigt wird, und für die Favoriten.

Abbildung 5.16: Darstellung von einem hybriden Ergebnis

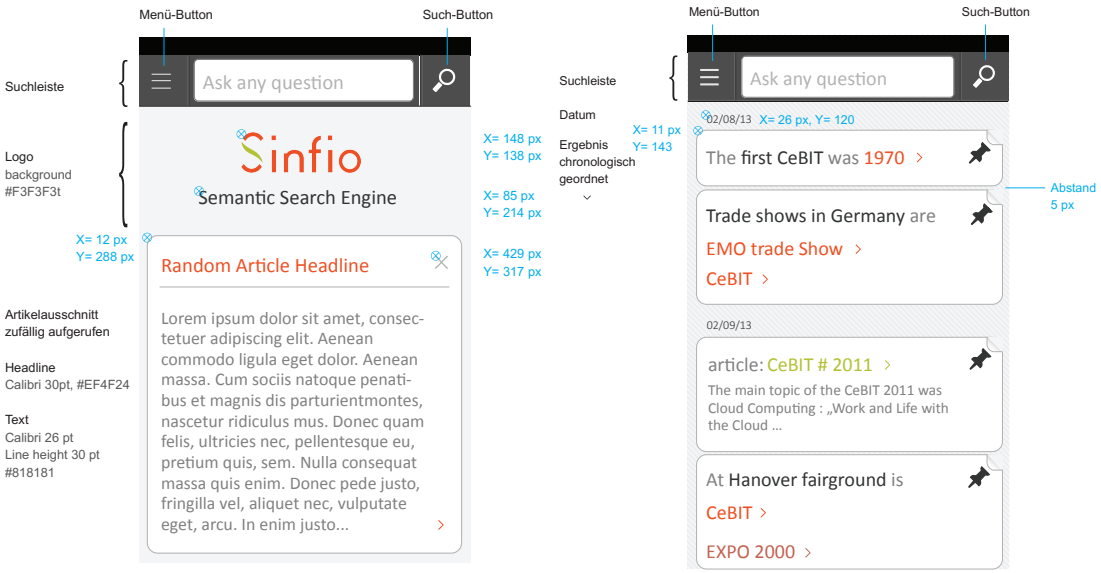


Abbildung 5.17: Styleguide für die Startseite (links) und die gespeicherten Suchergebnisse (rechts)

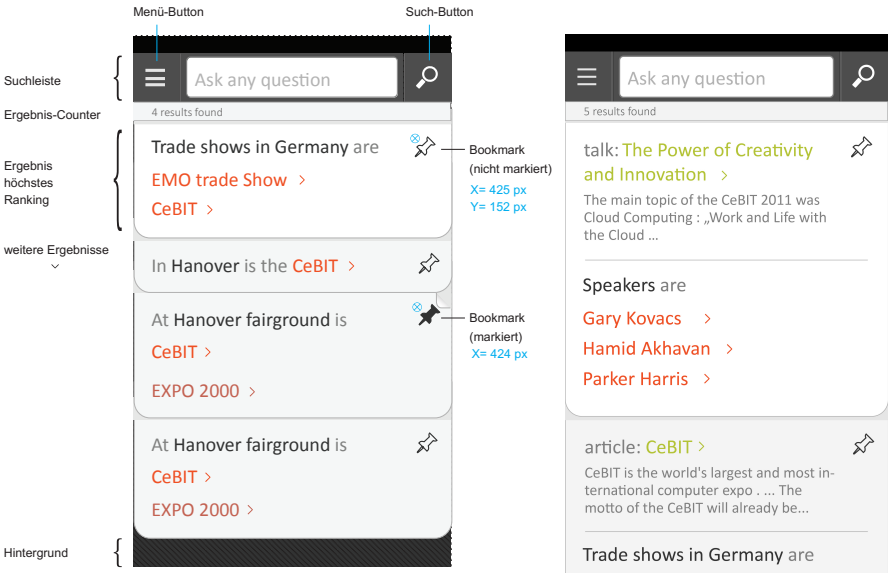


Abbildung 5.18: Styleguide für die Darstellung von Fakten (links) sowie von hybriden Ergebnissen und Dokumenten (rechts)

Den Entwurf für die Darstellung von Fakten, hybriden Suchergebnissen und Dokumenten zeigt die Abbildung 5.18. Für eine einfache Unterscheidung werden Farbcodes verwendet: Dokumenttitel werden grün, Instanzen orange dargestellt und beide mit den entsprechenden Ressourcen verlinkt. Der Typ des Dokuments, falls verfügbar und in der jeweiligen Datenbasis sinnvoll, wird dem Dokument vorangestellt. Die Suchterme werden wie in den Textsnippets der Dokumentdarstellung hervorgehoben.

Je ähnlicher die Struktur der einzelnen Antworten ist, umso mehr abstrahiert die Darstellung von der Diversität der Antworttypen, die Komplexität wird vor dem Benutzer verborgen. Eine gemischte Ergebnisliste, die nach dem Rang der Ergebnisse geordnet ist, ist nach [Mayer and Moreno, 2003] leichter zu interpretieren. Die Suchmaschine bietet jedoch auch eine Ansicht an, in der die Ergebnisse nach Ergebnistyp sortiert sind. Dies ermöglicht den Schnellzugriff auf Ergebnisse, falls die Benutzer beispielsweise genau wissen, wonach sie suchen bzw. an einem bestimmten Ergebnistyp interessiert sind.

Das Design für Browser auf PCs unterscheidet insofern von diesen Entwürfen, als dass die Größe und Abstand der Ergebnisse angepasst und die für Smartphone typische Elemente entfernt wurden. So ein Element ist beispielsweise das ⟩ hinter den Dokumenten und Konzepten, die einen Link zum Antippen kennzeichnen anstatt der auf den PCs üblichen Linkhervorhebung durch Unterstreichen (dauerhaft oder sobald die Maus über den Text streift). Ebenso wird auf dem PC aufgrund der großen Anzeigefläche kein Container hinter die Einzelergebnisse gelegt, sie werden durch genügend Abstand getrennt. Weiterhin wird ein längeres Textsnippet angezeigt, als auf dem Smartphone.

5.3 Prototypische Realisierung

Der Prototyp wurde als eine Webanwendung nach Client-Server-Architektur mit Java Backend, JSP und HTML5-Frontend realisiert. Für den Dokument- und den Autovervollständigungsindex wurde die Suchbibliothek Apache Lucene[17] bzw. der Lucene-Aufsatz Solr[18] eingesetzt, da sie performant im Sinne der Verarbeitungsgeschwindigkeit und bereits gut erprobt ist. Als Tripel-Store kam der Open Source Virtuoso Server[19] zum Einsatz.

Dieses Kapitel beschreibt die Entwurfsentscheidungen im Rahmen der prototypischen Realisierung für große Datenmengen, wie z.B. DBpedia und Wikipedia, und ihre Auswirkungen auf die hybride semantische Suche. Betroffen sind die drei Suchkomponenten semantische Autovervollständigung, Faktensuche und das Spreading Activation (Kapitel 5.3.1). Darüber hinaus wurde das Ranking im Rahmen der Dokumentsuche durch die eingesetzten Tools beeinflusst (Kapitel 5.3.2) und eine datenmengenspezifische Entscheidung in Bezug auf die Berechnung der Popularität getroffen (Kapitel 5.3.2).

[17] `https://lucene.apache.org/core/` (06.01.2016)

[18] Solr erweitert Lucene um eine REST-Schnittstelle, die Webzugriff und Ein- und Ausgaben in verschiedenen Formaten (JSON, XML, CSV, binary) anbietet. `http://lucene.apache.org/solr/` (06.01.2016)

[19] `http://virtuoso.openlinksw.com/` (06.01.2016)

5.3.1 Suchkomponenten

Entwurfsentscheidungen für die **semantische Autovervollständigung** müssen bezüglich der Erkennung von Pluralformen und der N-Gram-Methode getroffen werden. Da Lucene eine Inflector-Bibliothek anbietet, war für Ersteres keine externe Bibliothek notwendig. Die N-Gram-Methode wird im Abschnitt 5.3.2 in Zusammenhang mit weiteren rankingspezifischen Entscheidungen behandelt.

Das tripelbasierte Verfahren zur **Faktensuche** führt semantischen Abgleich durch, indem es über je zwei benachbarte Suchterme t_i, t_{i+1} iteriert und Abfragen für alle Kombinationen der Ergebnisse aus der syntaktischen Suche $M(t_i, G_\Sigma)$ und $M(t_{i+1}, G_\Sigma)$ ausführt (s. Kapitel 5.2.3.1). Das Konzept sieht bei kleinen Datenmengen vor, auch für zwei benachbarte Terme, durch die je eine Property gefunden wurde, Abfragen zu generieren. Es wird also nach zwei Tripeln gesucht, die beide Properties beinhalten und durch ein Konzept verbunden sind, jedoch nur die Properties bekannt sind (s. Abbildung 5.19). Ebenso sollen Klassen durch ihre Instanzen ersetzt werden, wenn sie sonst bei der tripelbasierten Suche zu keinen Ergebnissen führen (vgl. Kapitel 5.2.3.1). Solche Abfragen sind aufwändig. Sie konnten auf der Instanzbasis der Olympischen Spiele mit ca. 40.000 Instanzen und 770.000 Tripel angewendet werden und wurden für die erste Evaluierung (s. Kapitel 6.1) implementiert. In großen Instanzbasen wie DBpedia (ebenfalls für die Evaluierung verwendet, s. Kapitel 6.2) können Abfragen mit zwei Properties und die Ersetzung von Klassen durch ihre Instanzen zu großen Ergebnismengen führen, dadurch wird die Suche ineffizient. In der DBpedia v3.9 werden 675.879 Personen mit Namen und Geburtsdatum gefunden. Das Teilergebnis der Faktensuche, in dem die Klasse Person durch ihre Instanzen ersetzt wurde, enthielt über 1.350.758 Tripel. Die Ausführungszeit der Abfrage zur Substitution liegt zwischen $30 - 60$ Sekunden[20]. Aus diesem Grund wird auf den Einsatz von Abfragetemplates mit zwei Porperties, sowie auf die Substitution der Klassen durch ihre Instanzen verzichtet. Im Rahmen des Spreading Activation-Prozesses können die entsprechenden Fakten gefunden werden.

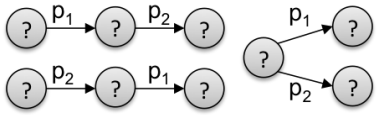


Abbildung 5.19: Gesuchte Tripel mit je zwei Properties

Für das **Spreading Activation** mussten in Bezug auf inverse Properties und den Einsatz von Constraints Entscheidungen getroffen werden. Bei dem Aufbau des Netzes wurden für DBpedia zu jedem Property ohne ein Inverses auch die inverse Kante mit aufgenommen, um Relationen in beide Richtungen finden zu können. Die vier Constraints (Stop-, Konzept-, Fan-Out- und Distanz-Constraint, s. Kapitel 5.2.3.2) wurden implementiert jedoch nur das Stop- und das Fan-Out-Constraint parametriert. Das Konzept-Constraint stoppt die Aktivierung von bestimmten Typen von Knoten und eignet sich für domänenspezifische Instanzbasen. Das Distanz-Constraint beschränkt die Distanz des Spreadings von den initial aktivierten Knoten heraus. In dem Lösungsansatz sollen jedoch hoch gewichtete Pfade verfolgt werden, um auch komplexere Zusammenhänge erkennen

[20] Auf 2 GHz Intel Core i7, 8GB RAM.

zu können. Die Aktivierung wird anhand des Stop-Contsraints kontrolliert, das Fluten des Graphen stoppt an den Knoten, deren Gewicht eine gegebene Grenze unterschreitet. Die Ergebnismenge ist die Menge der Knoten, deren Gewicht eine Schranke überschreitet.

Der Lösungsansatz geht davon aus, dass mit einer Instanz gar keine oder viele Dokumente verknüpft sein können und analog können auch Dokumente alleine stehen oder mit vielen Instanzen verknüpft sein. Ein Sonderfall entsteht, wenn DBpedia und Wikipedia als Datenquellen eingesetzt werden, da jedes Konzept sowohl durch eine Instanz als auch eine Wikipediaseite repräsentiert wird. In diesem Fall ist festzulegen, nach welchem Kriterium entschieden wird, ob die DBpedia-Instanz oder die Wikipediaseite dem Benutzer präsentiert werden soll. Um diese Entscheidung treffen zu können, wird im SA-Netzwerk zwischen einem Dokument (genau genommen der Instanziierung des Dokuments, s. Kapitel 4.2 und 5.2.2) und der zugehörigen Instanz unterschieden und die Kante dazwischen im Rahmen des SA-Prozesses ebenso behandelt, wie eine reguläre Kante. Hierdurch entstehen unterschiedliche Gewichte und als Ergebnis wird der Knoten mit dem höheren Gewicht ausgewählt.

5.3.2 Ranking

5.3.2.1 Dokument- und Fakt-Ranking

Lucene bietet eine effiziente Suche auf unstrukturierten Daten und viele Vorteile durch ihre zahlreichen Indizierungs- und Suchparameter, Erkennung von Pluralformen, verschiedene Rankingfunktionen, Autovervollständigungsfunktion und vieles mehr. Zudem ist Lucene javabasiert und daher plattformunabhängig, gut dokumentiert und erprobt. Das **Ranking in Lucene** weicht jedoch von dem Standard-Verktorraummodell ab. Dies liegt an der Umsetzung der Parametrierungsmöglichkeiten und der Kombination des Vektorraummodells mit dem Booleschen Modell. Das Vektorraummodell macht das Herz der Suchmaschine aus, das Boolesche Modell wird verwendet um die Logik der Suchanfrage auswerten zu können[21]. Als Ergebnis liegen die Gewichte der Ergebnisdokumente nicht im Intervall $[0, 1] \subset \mathbb{R}$. Der Rang eines Ergebnisdokuments ist absolut, er kann nur mit dem Rang der Ergebnisse aus der selben Suche verglichen werden. Dies gilt auch für die Präfixsuche im Rahmen der semantischen Autovervollständigung. Im Folgenden werden die Auswirkungen des Lucene-Rankings auf die Implementierung der Suchlösung und mögliche Lösungen diskutiert.

Der **Vergleich der Ergebnisse von verschiedenen Suchanfragen** bezüglich ihres Rankings ist für den Benutzer in den meisten heutigen Suchmaschinen nicht möglich, da der Rang der Ergebnisse nicht dargestellt wird. Zudem basieren weitere Suchbibliotheken entweder auf Lucene (ElasticSearch[22], SearchBlox[23], OpenSearchServer[24]) oder kombinieren ebenfalls das Vektorraummodell mit dem Booleschen Modell und haben deshalb die selben Einschränkungen (Xapian[25]). Aus diesen Gründen wurde die Dokumentsuche trotz dieses Nachteils mit Lucene umgesetzt. Die Gewichte der Ergebnisdokumente wur-

[21] `http://lucene.apache.org/core/3_6_0/scoring.html` (06.01.2016)
[22] `https://www.elastic.co` (06.01.2016)
[23] `http://www.searchblox.com` (06.01.2016)
[24] `http://www.opensearchserver.com` (06.01.2016)
[25] `http://xapian.org` (06.01.2016)

den lokal, durch das höchste Gewicht normalisiert, um Werte im Intervall $[0, 1] \subset \mathbb{R}$ zu erhalten und somit die Kombinierbarkeit der Ergebnisse der Fakten- und Dokumentsuche gewährleisten zu können (vgl. Kapitel 5.2.4). Der Einsatz hat jedoch die Auswirkung, dass ein Vergleich des Ranges der Ergebnisse für verschiedene Suchanfragen auch in der hybriden semantischen Suche nicht mehr möglich ist.

Das Problem des Intervalls kann gelöst werden, indem für die gefundenen Dokumente das Standard-Vektorraummodell umgesetzt wird. Die $tfdf$-Gewichtung kombiniert lokale und globale Gewichtung: die Termhäufigkeit tf und die Dokumenthäufigkeit df (vgl. Kapitel 2.3.2). Da alle Suchterme und alle Dokumente, die die Suchterme beinhalten, in der Ergebnismenge liegen werden dieselben Gewichte vergeben, wie auf der gesamten Dokumentmenge. Die Menge der Indexterme kann für eine Teilmenge der Dokumente anders ausfallen als auf der gesamten Dokumentmenge. An dieser Stelle wird die Ähnlichkeit mit einer Suchanfrage berechnet, weshalb der Unterschied der Indexterme den Rang nicht beeinflusst. Eine Vergleichbarkeit der Ergebnisse mit den Ergebnissen anderer Suchanfragen, deren Gewichte anhand anderer Indexterme berechnet wurden, ist jedoch nicht mehr gegeben. Dies kann dann gewährleistet werden, wenn ein Standard-Vektorraum mit der $tfidf$-Gewichtung implementiert und statt Lucene eingesetzt wird, was im Rahmen dieser Arbeit nicht geleistet werden kann. Die Indexierung inklusive der Datenstrukturen und Parametrierung, die verschiedenen Zugriffsmethoden inklusive der Erkennung von Pluralformen und Präfixsuche für die Autovervollständigung sowie die Suche mit $tfidf$ müssen hierfür realisiert werden.

Die **semantische Autovervollständigung** wurde ebenfalls auf Basis eines Lucene-Indexes umgesetzt. Lucene bietet auch N-Gram-Berechnung an. Diese lässt sich jedoch nicht ausreichend parametrieren und die Ähnlichkeitswerte liegen nicht im Intervall $[0, 1] \subset \mathbb{R}$. Dieses Problem kann durch die Implementierung einer N-Gram-Methode mit Bigrammen für Terme bis zu 5 Zeichen und Trigrammen für längere Terme sowie dem Dice-Maß umgegangen werden (vgl. Kapitel 5.2.1). Damit die Dokument- und Faktengewichte auf derselben Basis (angepasstes Vektorraummodell in Lucene) berechnet werden, wurden auch an dieser Stelle die Lucene-Gewichte auf $[0, 1] \subset \mathbb{R}$ normalisiert.

Eine dierekte Vergleichbarkeit der Gewichte aus der Präfixsuche und der Dokumentsuche ist nicht gegeben, da sie auf unterschiedlichen Vektorräumen berechnet werden. Sie befinden sich lediglich im selben Intervall und lassen sich kombinieren. Im Rahmen der hybriden semantischen Suche SINFIO ist eine direkte Vergleichbarkeit jedoch nicht erforderlich: Die Fakten und Dokumente werden nicht zu einer geordneten Ergebnisliste zusammen sortiert sondern Spreading Activation durchgeführt. Im Rahmen des Spreading Activations besteht die Möglichkeit, Dokumente oder Fakten stärker oder schwächer mit einfließen zu lassen, indem die entsprechenden Kantengewichte angepasst werden. SINFIO wurde ohne eine Anpassung evaluiert, die Gewichte der initialen Aktivierungsknoten entsprechen den Gewichten, die sich aus der Faktensuche sowie der Schlüsselwortsuche im Dokumentindex ergeben.

5.3.2.2 Popularität

Für die Berechnung der Popularität (s. 5.2.4) bedarf es einer **Auswahl der Ebene in der Klassenhierarchie**, die für die Normalisierung eingesetzt wird[26]. Um die geeignete Hierarchieebene bestimmen zu können, wurde die Klassenhierarchie der DBpedia-Version 3.9[27] in Hinsicht auf die Klassen-Instanzenzugehörigkeit statistisch ausgewertet. Im Fokus stand dabei die Abdeckung der Instanzen je Hierarchieebene. Im Folgenden werden die Ebenen ohne die Root-Klasse $\langle owl : Thing \rangle$ betrachtet. Abbildung 5.20 zeigt einen Ausschnitt aus der Klassenhierarchie der DBpedia-Ontologie[28]. Die Klassen erster Ebene sind überwiegend unspezifisch, wie Activity, Agent, Altitude, AnatomicalStructure usw. Insgesamt 27 der 49 Klassen dieser Ebene haben keine Unterklassen, wobei 20 davon keine Instanzen und die restlichen 7 insgesamt 25.491 Instanzen haben. Diese wären durch die Klassen zweiter Ebene nicht abgedeckt. Klassen der zweiten Hierarchiebene sind beispielsweise Game, Sports, Person, Organisation, Device, Event, PopulatedPlace usw. Hier haben bereits 86 Klassen keine Unterklassen, 33 davon ohne und 53 mit insgesamt 1.285.218 Instanzen, die durch die Klassen dritter Ebene nicht abgedeckt wären. Die Instanzen der Klassen zweiter Ebene decken 99, 4% die Gesamtheit der Instanzen ab, während die Klassen dritter Ebene lediglich 49% abdecken.

Abbildung 5.20: Ausschnitt aus der DBpedia-Klassenhierarchie der DBpedia-Ontologie

Wegen der besseren Abdeckung gegenüber der Klassen dritter Ebene wurden für die Evaluierung die Klassen der zweiten Hierarchiebene herangezogen und mit den 7 Klassen (25.491 Instanzen betreffend) aus der ersten Hierarchiebene erweitert.

[26] Die Berechnung auf Basis der spezifischsten Klasse der Instanzen würde zu einer unfairen Bewertung führen, da eine vollständige Klassifizierung entlang der Klassenhierarchie freiwillig geschieht und die Instanzen somit unterschiedlich präzise klassifiziert sind.

[27] Die DBpedia 3.9 Ontologie, `http://oldwiki.dbpedia.org/Downloads39#dbpedia-ontology` (07.10.2015).

[28] Stand 07.10.2015, `http://mappings.dbpedia.org/server/ontology/classes/` (07.10.2015).

Das Einbeziehen der Popularität erhöht die Gewichte der Knoten im Rahmen des Spreading Activation-Prozesses um einen Wert $pop_{c_{k,i}} \in [0,1] \subset \mathbb{R}$ (vgl. Kapitel 5.2.4). Das maximale Knotengewicht vor dem Spreading Activation erhöht sich von 1 auf 2 und erfordert die **Anpassung des Stop-Constraints**. Die Stoppbedingung ist parametrierbar umgesetzt. Sie ist ein fester Wert, der an die jeweilige Datenmenge angepasst werden kann. In Zusammenhang mit dem Auskühlungsfaktor α und den Kantengewichten bestimmt sie, inwieweit die Pfade verfolgt werden und unterstützt daher eine domänenspezifische Anpassung des Spreading Activation-Prozesses (s. Kapitel 5.2.3.2). Da das initiale Maximalgewicht durch die Popularität verdoppelt wird, wird in diesem Fall auch der Wert des Stop-Constraints verdoppelt.

5.3.3 Grafische Benutzeroberfläche und Ergebnisdarstellung

Der im Kapitel 5.2.5 vorgestellter Entwurf für die Benutzerschnittstelle wurde mit wenigen Änderungen umgesetzt. Abbildung 5.21 zeigt die für DBpedia und Wikipedia um-

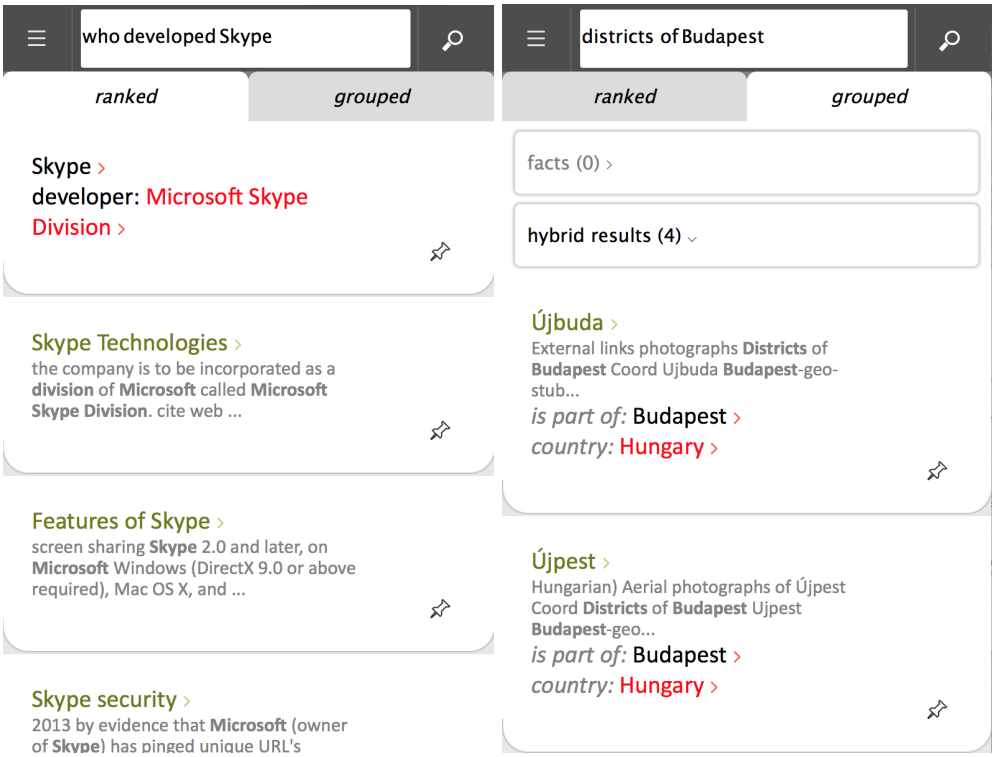


Abbildung 5.21: Ergebnisdarstellung auf dem Smartphone, links nach Rangordnung, rechts gruppiert

gesetzte Darstellung der unterschiedlichen Ergebnistypen, wobei links die gerankte und rechts die nach Ergebnistyp gruppierte Darstellung der Ergebnisliste zu sehen ist.

Abbildungen 5.22 und 5.23 zeigen die beiden Darstellungsmodi der Ergebnisliste auf dem PC für die Suchanfrage „tango music instruments“. Die smartphonespezifische Interaktionselemente sind entfernt, die Links werden unterstrichen, sobald der Mauszeiger sich darüber befindet. Neben dem Farbcode wurde auch das Logo von Wikipedia und DBpedia hinter die Dokumenttitel bzw. Namen der Konzepte eingefügt, um dem Benutzer zu

zeigen, wohin die Links führen. Beim Klick auf einen Link werden die entsprechenden Webseiten (Wikipedia-Seite bzw. DBpedia-Seite) geöffnet.

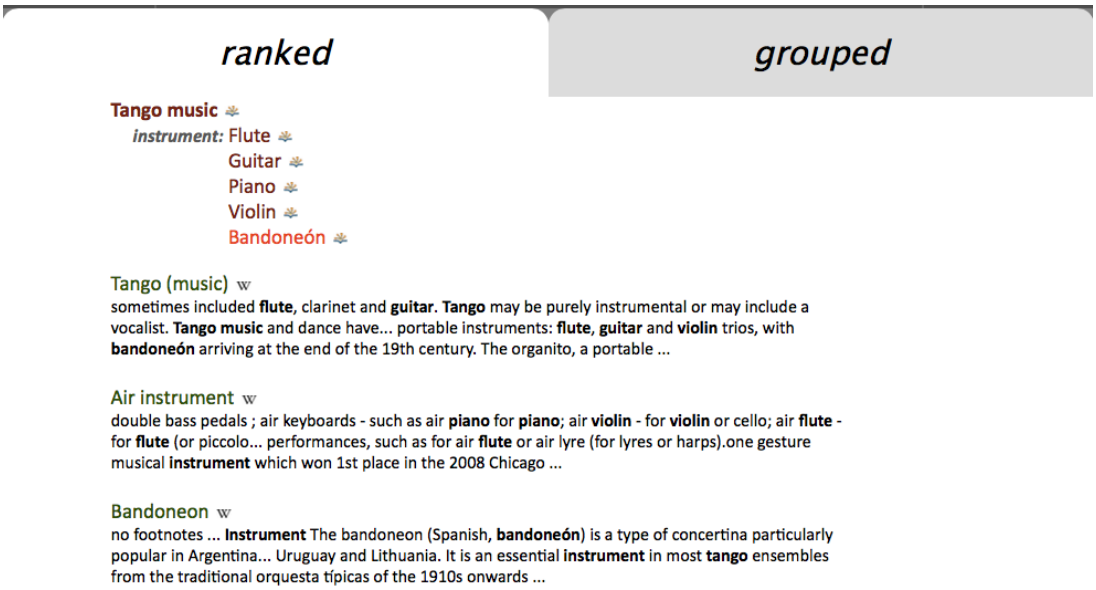


Abbildung 5.22: Darstellung nach Rangordnung auf dem PC

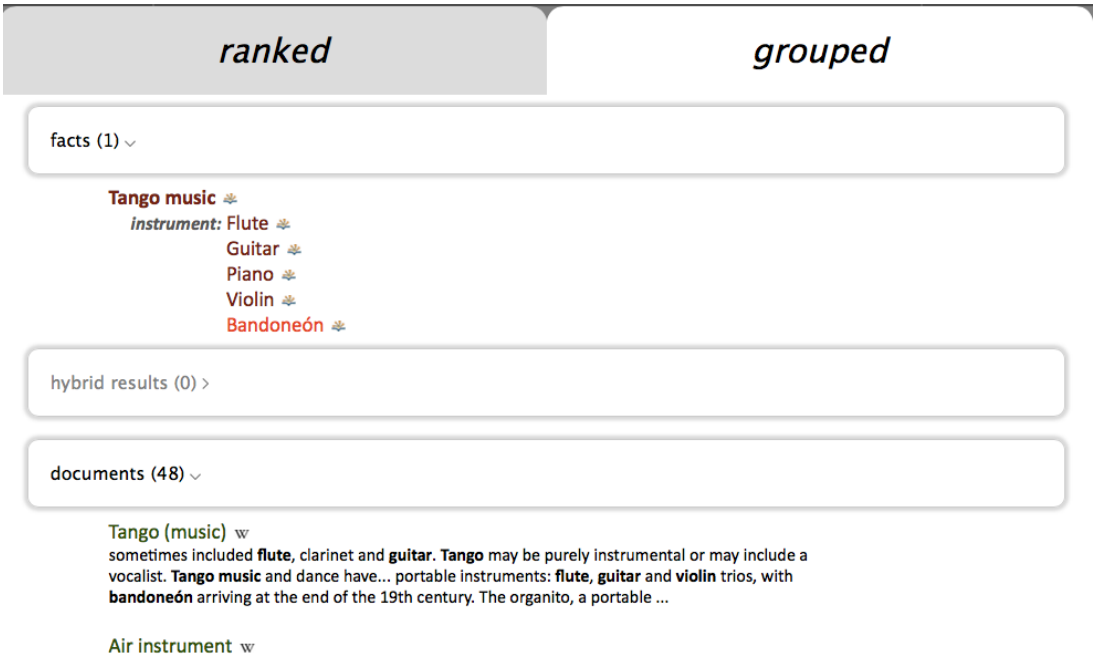


Abbildung 5.23: Gruppierte Darstellung auf dem PC

Kapitel 6
Evaluierung

Kapitel 6 beschreibt die Evaluierungen der Suchmaschine SINFIO. Um den Lösungsansatz zu prüfen, wurde zuerst im Rahmen der Entwicklung eine *Proof Of Concept-Evaluierung* auf einem kleinen Gold Standard über die Olympischen Spiele im Jahr 2004 durchgeführt. Die Versuchsplanung und die Ergebnisse beschreibt Kapitel 6.1. Eine *umfangreiche Evaluierung* erfolgte auf der englischen DBpedia und Wikipedia auf Basis eines Gold Standards und ist im Kapitel 6.2 beschrieben. Kapitel 6.2.1 beschreibt die Daten, die für die Evaluierung eingesetzt worden sind. Kapitel 6.2.2 stellt die Evaluierung der These aus System- sowie aus Benutzersicht vor. Kapitel 6.2.3 behandelt die Evaluierung der Rankingverfahren, Kapitel 6.2.4 die Evaluierung der Nutzung der semantischen Autovervollständigung und Kapitel 6.2.5 die Evaluierung der Ergebnisdarstellung in Bezug auf die Darstellung der einzelnen Ergebnistypen sowie auf die Darstellung der Ergebnislisten. Die Ergebnisse der Effizienzmessung sind im Kapitel 6.2.6 beschrieben.

Die statistische Signifikanz wurde mit dem Randomisierungstest bestimmt (s. [Smucker et al., 2007, Webber, 2010]). Student's Zweistichproben-t-Test wird zwar wegen seiner Robustheit bevorzugt in IR eingesetzt, er geht jedoch von einer Normalverteilung der Stichproben aus und eignet sich daher nur für den Vergleich von Mittelwerten. Stichproben sind in diesem Fall die Differenzen zwischen den Ergebnissen der untersuchten Kennzahlen je Suchanfrage und Suchsystem. Eine Normalverteilung kann bei dem Vergleich von Fakten-, semantische Dokument- und hybride semantische Suche nicht erwartet werden. Der Randomisierungstest erhebt keine Annahmen über die Verteilung und kann bei allen hier relevanten Kennzahlen eingesetzt werden [Smucker et al., 2007, Webber, 2010]. Der Signifikanztest der Systemvergleiche basiert auf der Nullhypothese, dass kein Unterschied zwischen den beiden Systemen besteht. Der berechnete p-Wert gibt an, wie wahrscheinlich es ist, dass die beobachtete Differenz der Systemkennzahlen entsteht, wenn die Nullhypothese wahr ist. Wird ein p-Wert unter einer Grenze α erreicht, so betrachtet man die Differenz der Systeme als statistisch signifikant und die Nullhypothese als unplausibel. Für α wird typischerweise 0,05 gewählt [Webber, 2010].

Für die Korrelationsanalyse unter den Ergebnissen der verschiedenen Evaluierungen wurde der Standard Pearson (Produkt-Moment-)Korrelationskoeffizient, das normalisierten Maß der linearen Korrelation, berechnet (s. [Johnson et al., 1992, Steiger, 1980]). Der Wertebereich des Koeffizienten ist $[-1, 1] \in \mathbb{R}$. Werte nahe 1 indizieren, dass die Variablen eine positive, Werte nahe -1, dass diese eine negative lineare Korrelation zueinander haben. Der Wert 0 deutet auf keine Korrelation hin. Der Pearson-Korrelationskoeffizient basiert auf der Annahme, dass eine lineare Korrelation zwischen zwei Zahlenreihen besteht und deutet bei einem nichtlinearen Zusammenhang auf keine Korrelation hin. Die durchgeführten Vergleiche betrachten die Ergebnisse unterschiedlicher Untersuchungen, zwischen denen eine lineare Korrelation vermutet wird. Werte nahe 0 wurden mit der Spearman-Korrelation geprüft, da dieser keine Annahmen voraussetzt und robust gegenüber aller Arten der Korrelation ist [Johnson et al., 1992, Steiger, 1980]. Der Wertebereich der Spearman-Korrelation ist ebenfalls $[-1, 1] \in \mathbb{R}$.

Eine gemeinsame Auswertung und Zusammenfassung der Ergebnisse aller Evaluierungen folgt im Kapitel 7.

6.1 Proof of Concept

Das Ziel dieser Zwischenevaluierung war zu prüfen, ob die hybride semantische Suche effektiver sein kann als die Schlüsselwortsuche bzw. die semantische Dokumentsuche.

Für die Evaluierung wurde der Ontology-Based Corpus and Annotation Scheme, kurz OCAS2008, verwendet. Der **OCAS2008-Gold Standard** wurde für die Evaluierung von ontologiebasierter Wissensextraktion sowie semantischer Suche entwickelt. Er besteht aus [Grothkast et al., 2008, Schumacher and Sintek, 2011]:

- einer Domänenontologie der Olympischen Spiele (s. Abbildung 6.1);
- einer Instanzbasis mit über 40.000 Instanzen und etwa 770.000 Tripeln mit Themenbezug zu den Olympischen Spiele 2004;
- einem Dokumentkorpus aus 121 Nachrichtenartikeln (etwa 31.000 Wörter) aus dem ABC[1] und BBC[2] Nachrichtenarchiv;
- 8 Suchanfragen, ausgewählt aus einer Menge von von Suchanfragen, die reale Informationsbedürfnissen ausdrücken;
- manuell annotierten binären Relevanzbeurteilungen bezüglich der Suchanfragen.

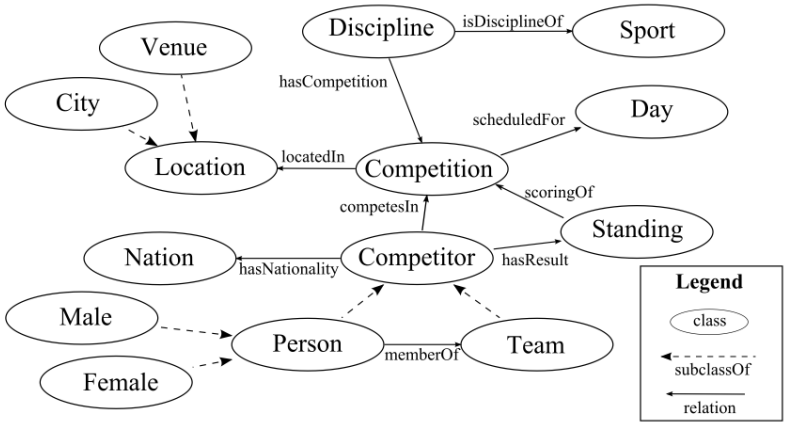


Abbildung 6.1: Die OCAS-Ontologie

Die Dokumente wurden aus 2000 Artikeln mithilfe eines Retrievalsystems ausgewählt, so dass alle 18 Tage und alle 32 Disziplinen der Olympischen Spiele abgedeckt sind. Die Annotation der Dokumente mit Instanzen, Relationen und die Relevanzbeurteilung zu den Suchanfragen wurde von 6 Oberschülern durchgeführt, die sich weder mit Wissensextraktion noch mit semantischer Suche auskennen [Grothkast et al., 2008]. Für die Evaluierung wurden aus der Menge von 77 Suchanfragen, die von den annotierenden Personen vorgeschlagen wurden, 8 ausgewählt [Schumacher and Sintek, 2011]:

q1: standings of Australian;
q2: disciplines with gold for Australians;
q3: teams of South Korea;
q4: when did Chinese win gold;
q5: places of competitions in cycling;

[1] `http://www.abc.net.au/olympics/2004/news_archive.htm` (01.09.2008)
[2] `http://news.bbc.co.uk/sport2/hi/olympics_2004/` (01.09.2008)

q6: in which disciplines did British sportsmen compete;

q7: who competed in swimming;

q8: which nations have a women football team.

Die Suchanfragen basieren auf realen Informationsbedürfnissen und weisen unterschiedliche Komplexitäten auf. Sie reichen von einfachen Suchanfragen, deren Ergebnis beispielsweise aus einer Menge von Instanzen besteht, bis hin zu komplexen Anfragen, die nur durch Fakten und/oder Dokumente beantwortet werden können. Da die Nachrichtenartikel üblicherweise eine Zusammenfassung der Ereignisse bezüglich einer Nation oder einer Disziplin beinhalten, ist es möglich die Suchanfragen mit einem oder wenigen Artikeln zu beantworten. Insgesamt erfüllt der OCAS2008 Gold Standard die Anforderungen an die Evaluierung von Retrievalsystemen (vgl. 2.5.1).

Basierend auf dem Gold Standards wurde **die Genauigkeit, die Vollständigkeit und das F-Maß** (s. Kapitel 2.5.1) **für traditionelle Schlüsselwortsuche** mit Lucene, **semantische Dokumentsuche** wie im Kapitel 5.2.3.2 beschrieben und die **hybride semantische Suche SINFIO** berechnet. Die Evaluierung erfolgte ohne eine domänenspezifische Gewichtung der Properties und unter Einsatz des Stop- sowie Fan-Out-Constraints für das Spreading Activation (s. Kapitel 5.2.3.2). Das Ranking wurde mit dem Verfahren ohne die Popularität durchgeführt (s. Kapitel 5.2.4). Abbildung 6.1 zeigt die Ergebnisse [Schumacher and Sintek, 2011].

	Schlüsselwortsuche			**Semantisches Dokumentretrieval**			**Hybride semantische Suche**		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
q1	0.2752	1.0	0.4316	**0.2632**	1.0	0.4167	0.6042	1.0	**0.7533**
q2	0.0354	1.0	0.0683	0.0345	1.0	0.0667	**0.069**	1.0	**0.129**
q3	0.3421	0.8125	0.4814	0.2192	1.0	0.36	**0.4688**	1.0	**0.6383**
q4	0.1563	1.0	0.2703	0.0408	1.0	0.0784	**0.2134**	1.0	**0.3517**
q5	0.2858	1.0	0.4444	0.0317	1.0	0.0615	**0.4**	1.0	**0.5714**
q6	0.125	0.2857	0.1739	**0.5**	1.0	0.6667	0.4615	1.0	**0.6315**
q7	0.0338	0.6667	0.0645	0.0256	1.0	0.05	**0.4**	1.0	**0.5714**
q8	0.0138	1.0	0.0274	0.0435	1.0	0.0833	**0.25**	1.0	**0.4**
Durch-schnitt	0.1584	0.8456	0.2452	0.1448	1.0	0.223	**0.3584**	1.0	**0.5058**

Tabelle 6.1: Genauigkeit (Precision, Vollständigkeit (Recall) und F-Maß (F-Measure) der Schlüsselwortsuche, des semantischen Dokumentretrievals und der hybriden semantischen Suche

Die Ergebnisse der *semantischen Dokumentsuche* sind weniger genau, als die der Schlüsselwortsuche, dafür verbessert sich jedoch die Vollständigkeit. Dies liegt am Spreading Activation und dessen Parametrierung. Es werden auch Artikel gefunden, die die Suchanfrage nur zum Teil oder gar nicht beantworten, weil z.B. ein relevanter Ort darin vorkommt, es jedoch nicht um die gefragte Disziplin (q5) geht. Dadurch sinkt die Genauigkeit. Gleichzeitig werden Artikel gefunden, die z.B. nur das Synonym eines Suchtermes beinhalten aber durch semantische Metadaten mit anderen relevanten Dokumenten verknüpft sind, wodurch die Vollständigkeit steigt [Schumacher and Sintek, 2011].

Die *hybride semantische Suche* zeigt die besten Ergebnisse, der Durchschnitt des F-Maßes verdoppelt sich gegenüber der Schlüsselwortsuche und der semantischen Dokumentsuche. Die initialen Aktivierungsknoten sind durch die Faktensuche präziser und das Problem des zu weitgehenden Spreadings, wie oben bei der semantischen Dokumentsuche

beschrieben, wird durch die Gewichtung der durch den syntaktischen Abgleich gefundenen Properties gemindert. Entsprechend konnte die größte Verbesserung bei denjenigen Suchanfragen erzielt werden, die Properties beinhalten (q2, q6 und q7). Sind nur Klassen und/oder Instanzen mit hoher Konnektivität gefunden worden, so ist die Verbesserung gegenüber der semantischen Dokumentsuche und der Schlüsselwortsuche geringer. Beispiele hierfür sind q4 und q5: durch „when“ wird eine Klasse gefunden, durch „Chinese“ eine Instanz mit hoher Konnektivität (q4), durch „places“ und „competitions“ ebenfalls jeweils eine Klasse (q5). Die Verbesserung des F-Maßes gegenüber der traditionellen Schlüsselwortsuche ist in diesen Fällen etwa 30%. Der hybride Ansatz SINFIO beantwortete die Anfragen q2, q3, q5 und q7 mit Fakten, eine weitere Suche in den Dokumenten fiel damit weg [Schumacher and Sintek, 2011].

Die Ergebnisse zeigen, dass der hybride Suchansatz gegenüber der Schlüsselwortsuche und der semantischen Dokumentsuche die Effektivität steigern kann. Die Verbesserung des F-Maßes der hybriden Suche ist sowohl gegenüber der semantischen Dokumentsuche als auch gegenüber der Schlüsselwortsuche statistisch signifikant (p=0,001185 bzw. p=0,01776)[3].

6.2 Evaluierung auf großen Datenmengen

Ziel dieser Evaluierung ist, sowohl die These als auch die Lösungen zu den Forschungsfragen, die durch Versuche untersucht werden können, zu prüfen. Zuerst werden die zugrundeliegenden Daten vorgestellt (Kapitel 6.2.1). Um sowohl eine objektive, generalisierbare Beurteilung als auch die Beurteilung der Benutzer einzubeziehen, wurde die These mit Systemkennzahlen sowie mit benutzerzentrierten Methoden evaluiert (Kapitel 6.2.2). Die Evaluierung der Rankingverfahren (F4) basiert auf ein Rankingkorrelationsmaß (Kapitel 6.2.3). Die semantische Autovervollständigung (F5) wurde in Benutzertests aufgabenbasiert untersucht (Kapitel 6.2.4). Die Ergebnisdarstellung (F6) wurde ebenfalls benutzerzentriert, anhand Side-by-side Panel und Fragebogen, evaluiert (Kapitel 6.2.5).

6.2.1 Daten

Die These geht davon aus, dass *die strukturiert und unstrukturiert abgebildeten Informationen verlinkt sind* und *sich nicht vollständig gegenseitig abdecken* (s. Kapitel 4.1). Zur Evaluierung bieten sich Daten aus der **Linked Open Data-Cloud**[4] (LOD-Cloud) und **dazu passende frei verfügbare Dokumente** aus dem Internet an. Die meisten Datenmengen konzentrieren sich auf eine Domäne und beschreiben Personen, geografische Daten, Musik, Fachwissen aus der Medizin usw. Die oben genannte Annahme erfüllen beispielsweise folgende formale Wissensbasen und Dokumentmengen:

- der BBC Wildlife Finder und die Artikel von World Wildlife[5],
- die US Kongress-Datenbasis GovTrack.us und Nachrichtenartikel über das Gesche-

[3] Im Rahmen der Proof-of-Concept Evaluierung erfolgte kein Vergleich mit anderen semantisches Suchmaschinen auf Basis des beschriebenen Gold Standards. Ziel dieser Evaluierung war festzustellen, ob eine hybride semantische Suche Vorteile gegenüber der verglichenen Sucharten bringt.

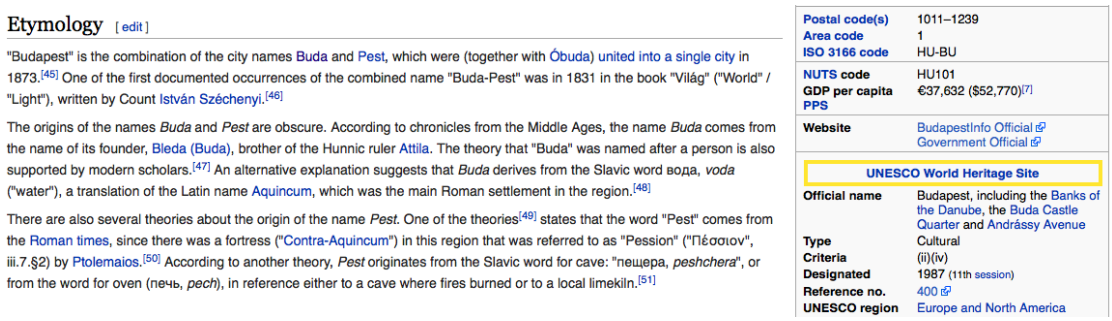
[4] `http://lod-cloud.net/` (06.01.2016)

[5] `http://www.bbc.co.uk/wildlifefinder/`, `http://www.worldwildlife.org/` (06.01.2016)

hen und die Politiker z.B. aus dem Library of Congress[6],

- die DBpedia und Wikipedia.

Die DBpedia ist domänenunabhängig, eine der größten formalen Wissensbasen in der LOD-Cloud und wird deshalb häufig zur Evaluierung von semantischen Suchmaschinen eingesetzt (s. Kapitel 3.4). DBpedia und Wikipedia als Datenbasis bieten sich an, da die erforderliche Verlinkung vorhanden ist. Es besteht aber auch eine deutliche Überschneidung an Informationsgehalt, da DBpedia aus Wikipedia extrahiert wurde. Die Wikipediaseiten wurden instanziiert und Fakten aus den Seiteninhalten extrahiert. Die Extraktionsverfahren konzentrieren sich auf die formale Abbildung der Informationen aus den Infoboxen, Kategorieinformationen, Geokoordinaten und Links zu externen Webseiten sowie Bilder[7]. Infoboxen sind die Boxen mit strukturierten Daten auf der rechten Seite der Artikel (s. Abbildung 6.2). Da diese häufig nicht oder nur spärlich ausgefüllt werden, werden für die Extraktion von Attribut-Werte-Paaren aus dem Text Tools, wie beispielsweise das iPopulator[8] eingesetzt [Dustin et al., 2010, Lange et al., 2010]. Die Extraktionsverfahren sind jedoch nicht vollständig. Die Artikel enthalten weitere Information, die zurzeit noch nicht automatisch extrahiert werden. Ein Beispiel hierfür ist die Information im Wikipediaartikel über Budapest: „'Budapest' is the combination of the city names Buda and Pest, which were (together with Óbuda) united into a single city in 1873"[9]. Dabei sind „Buda" und „Pest" mit den entsprechenden Wikipediaseiten verlinkt (s. Abbildung 6.2). Diese Information ist in der DBpedia-Instanz ⟨Budapest⟩[10] nicht zu finden. Die Instanz ⟨Budapest⟩ ist zwar mit den Instanzen ⟨Buda⟩ und ⟨Pest⟩[11] verlinkt, da die Links zwischen Wikiseiten in die DBpedia übernommen werden, die Relation ist jedoch nicht qualifiziert. Bei einer Suche nach der Zusammensetzung des Namens „Budapest" oder nach den Stadtteilen von Budapest werden sie nicht gefunden. Im Text der Wikipediaseite befinden sich noch zahlreiche weitere Fakten, beispielsweise die Theorien über die Namensherkunft, die nicht durch Wikilinks mit anderen Seiten verlinkt sind. Diese Informationen sind im DBpedia nicht vorhanden.



Etymology [edit]

"Budapest" is the combination of the city names Buda and Pest, which were (together with Óbuda) united into a single city in 1873.[45] One of the first documented occurrences of the combined name "Buda-Pest" was in 1831 in the book "Világ" ("World" / "Light"), written by Count István Széchenyi.[46]

The origins of the names *Buda* and *Pest* are obscure. According to chronicles from the Middle Ages, the name *Buda* comes from the name of its founder, Bleda (Buda), brother of the Hunnic ruler Attila. The theory that "Buda" was named after a person is also supported by modern scholars.[47] An alternative explanation suggests that *Buda* derives from the Slavic word вода, *voda* ("water"), a translation of the Latin name Aquincum, which was the main Roman settlement in the region.[48]

There are also several theories about the origin of the name *Pest*. One of the theories[49] states that the word "Pest" comes from the Roman times, since there was a fortress ("Contra-Aquincum") in this region that was referred to as "Pession" ("Πέσσιον", iii.7.§2) by Ptolemaios.[50] According to another theory, *Pest* originates from the Slavic word for cave: "пещера, *peshchera*", or from the word for oven (печь, *pech*), in reference either to a cave where fires burned or to a local limekiln.[51]

Postal code(s)	1011–1239
Area code	1
ISO 3166 code	HU-BU
NUTS code	HU101
GDP per capita PPS	€37,632 ($52,770)[7]
Website	BudapestInfo Official Government Official
UNESCO World Heritage Site	
Official name	Budapest, including the Banks of the Danube, the Buda Castle Quarter and Andrássy Avenue
Type	Cultural
Criteria	(ii)(iv)
Designated	1987 (11th session)
Reference no.	400
UNESCO region	Europe and North America

Abbildung 6.2: Ausschnitt aus dem Wikipediaartikel über Budapest

Die **englische DBpedia** Version 3.9 und die zugehörige **englische Wikipedia**[12] erfüllen die Anforderungen und wurden **als Datenbasis** für die Evaluierung der These und

[6] `http://www.govtrack.us`, `http://chroniclingamerica.loc.gov/newspapers/` (06.01.2016)
[7] `http://wiki.dbpedia.org/services-resources/datasets/data-set-39` (04.11.2015)
[8] `https://hpi.de/naumann/projects/completed-projects/ipopulator.html`, (04.11.2015)
[9] `https://en.wikipedia.org/wiki/Budapest`, (04.11.2015)
[10] `http://dbpedia.org/page/Budapest` (04.11.2015)
[11] `http://dbpedia.org/page/Buda`, `http://dbpedia.org/page/Pest` (04.11.2015)
[12] `http://oldwiki.dbpedia.org/Downloads39` (07.10.2015)

der Forschungsfragen ausgewählt. Die DBpedia fungiert als formale Wissensbasis, die zugehörige Wikipedia-Version liefert die Dokumente. Die Verlinkung zwischen Fakten und Dokumenten ist gegeben durch die Fakten über die zugehörigen DBpedia-Instanzen[13], die aus den Wikipediaseiten extrahiert worden sind (s. auch Kapitel 5.3.1). Die englische DBpedia v3.9 enthält 4 Millionen Konzepte (⟨owl:Thing⟩) und 470 Millionen Tripel[14]. Die zugehörige englische Wikipedia-Version beinhaltet 4,2 Millionen Artikel auf über 30 Millionen Seiten[15].

Im Kapitel 3.4 wurde auf ein *allgemeines Problem* bei der Evaluierung semantischer Suchmaschinen hingewiesen: Die Ergebnisse hängen stark von der *Qualität der unterliegenden semantischen Daten* ab, denn dies implizit mit evaluiert wird. Die wesentlichen Faktoren sind die Abdeckung der Domäne, die Komplexität des semantischen Modells bzw. Detailliertheit der Relationen sowie Konsistenz der Wissensbasis. Da DBpedia über keine detaillierten Modelle zu den verschiedenen Domänen verfügt, fällt die Abdeckung der Domänen, im Vergleich z.B. zu RadSem mit einer manuell erstellten detaillierten Ontologie und von Experten korrigierten Instanzbasis zum menschlichen Skelett [Forcher et al., 2009], gering aus. Dies ist bereits aus dem Ausschnitt der DBpedia-Klassenhierarchie in Abbildung 5.20 auf Seite 111 ersichtlich. Zudem kann aufgrund der Genauigkeit der automatisierten Verfahren, die unter 100% liegt, die Konsistenz der Daten nicht gewährleistet werden.

6.2.2 Evaluierung der These

6.2.2.1 These und Testhypothesen

Dieser Arbeit liegt die *These* zugrunde, dass die Suche in divers strukturierten Datenmengen durch hybride semantische Suche verbessert werden kann, wobei eine Verbesserung anhand der Retrievaleffektivität gemessen wird (s. Kapitel 2.5):

Für unterschiedlich stark strukturierte Datenmengen ist ein integrierter Ansatz zur hybriden semantischen Suche, der formale und informale Inhalte während des gesamten Suchprozesses kombiniert, effektiver als eine semantische Dokument- und Faktensuche ohne Kombination der Inhalte.

Die These impliziert die *Testhypothese 1*:

Die hybride semantische Suche erreicht eine höhere Retrievaleffektivität als die semantische Dokument- und die Faktensuche zusammen, jedoch ohne einer Kombination der formalen und informalen Inhalte im Suchprozess.

Die These und die Testhypothese erfordern eine *komparative Evaluierung* der *hybriden semantische Suche SINFIO* und der *semantischen Suche nach Fakten und Dokumenten ohne eine Kombination der Inhalte*. Letzteres ist die Ausführung von Faktensuche und semantischer Dokumentsuche, wobei die Ergebnisse nach ihrem Rang geordnet dargestellt

[13] Property ⟨foaf:isPrimaryTopicOf⟩ verbindet eine DBpedia-Instanz mit der zugehörigen Wikipediaseite.
[14] Für detaillierte Angaben s. `http://blog.dbpedia.org/?p=72` (07.10.2015).
[15] Für detaillierte Angaben s. `https://en.wikipedia.org/wiki/Wikipedia:Statistics` und `https://stats.wikimedia.org` (07.10.2015).

werden. Dabei werden formale und informale Inhalte in keinem Schritt des Suchprozesses kombiniert. Im Gegensatz zu der hybriden semantischen Suche kann weder die Suchanfrage hybrid sein noch kann die semantische Anfrageerweiterung für die semantische Dokumentsuche eingesetzt werden. Die Ergebnisse sind entweder Fakten oder Dokumente, jedoch keine hybride Ergebnisse.

Weiterhin soll die hybride Suche auf unterschiedlich stark strukturierten Datenmengen gegenüber der semantischen Dokumentsuche den Vorteil erbringen, dass die Benutzer ihr Informationsbedürfnis schneller befriedigen können, weil die Antwort häufiger bereits in der Ergebnisliste ablesbar ist. Es müssen keine weiterführenden Links geöffnet und in den Inhalten weitergesucht werden (s. Kapitel 4.1). Die *Testhypothese 2* lautet:

Gegenüber der semantischen Dokumentsuche können Benutzer durch die hybride Suche ihr Informationsbedürfnis schneller befriedigen.

Dies kann durch eine *komparative Evaluierung* der *hybriden semantischen Suche* und der *semantischen Dokumentsuche* geprüft werden, indem eine Benutzerbefragung bezüglich der Ablesbarkeit der Ergebnisse aus der Ergebnismenge und des schnelleren Findens der Antwort durchgeführt wird.

Gegenüber der Faktensuche besteht der Vorteil, dass auch Dokumente gefunden werden. Kann also das Informationsbedürfnis nicht alleine durch Fakten (strukturierte Daten) befriedigt werden, so können hybride Ergebnisse und Dokumente (unstrukturierte Daten) zum Finden der Antwort beitragen (s. Kapitel 4.1). *Testhypothese 3* kann folgendermaßen formuliert werden:

Gegenüber der Faktensuche kann die hybride semantische Suche stärker zur Befriedigung des Informationsbedürfnisses der Benutzer beitragen.

Dies kann ebenfalls durch eine *komparative Evaluierung* geprüft werden. Dabei werden die *hybride semantische Suche* und die *Faktensuche* verglichen. Der Unterschied ist jedoch nicht durch eine Frage an den Benutzer prüfbar. Für diese Testhypothese ist relevant, welche Suchmaschine als besser empfunden wird, wie die Ablesbarkeit der Antwort aus der Ergebnisliste ist sowie ob und bei welcher Suchmaschine die Antwort schneller gefunden wurde.

Wie bereits im Kapitel 2.5 beschrieben, kann die Evaluierung *aus Systemsicht*, mittels quantitativer Methoden sowie *aus der Sicht der Benutzer* geschehen. Quantitative Verfahren aus Systemsicht, die auf etablierte Kennzahlen basieren, führen in der Regel zu repräsentativen und generalisierbaren Ergebnissen. Benutzerzentrierte Methoden basieren hingegen auf benutzergenerierten Daten und sind von der Pragmatik der Benutzer geprägt. Sie sind subjektiv, worunter die Reproduzierbarkeit und Generalisierbarkeit der Ergebnisse leidet [Kelly, 2009]. Da Suchsysteme jedoch für die Benutzer entwickelt werden, ist die Zufriedenheit der Anwender ein wesentliches Qualitätsmerkmal. Aus diesem Grund wurde die These sowohl aus Systemsicht als auch aus Benutzersicht mit etablierten Methoden und Kennzahlen (vorgestellt im Kapitel 2.5) evaluiert. Vorgehensweise und Ergebnisse stellen die Kapitel 6.2.2.2 und 6.2.2.3 vor.

6.2.2.2 Evaluierung aus Systemsicht

Durch die Diversität der Ergebnisse und Ergebnislisten der zu vergleichenden Suchmaschinen gibt es Kriterien, die die Auswahl an geeigneten Kennzahlen einschränken. Diese Kriterien werden im ersten Abschnitt diskutiert und die Auswahl sowie die Evaluierungsmethode angegeben. Der zweite Abschnitt beschreibt den erstellten Gold Standard. Anschließend werden die Evaluierungsergebnisse vorgestellt und analysiert.

Methode und Kennzahlen
Die Messung der Effektivität aus Systemsicht erfolgt nach dem Standardansatz der **Cranfield-Methode mit Pooling** (vgl. Kapitel 2.5.1). Da die Benutzer sich häufig nur die erste Ergebnisseite anschauen, wurde $k = 10$ gewählt [Baeza-Yates and Ribeiro-Neto, 2011c]. Die Berechnung der Kennzahlen basiert also auf der Relevanz der Suchergebnisse in einem Pool, bestehend aus den ersten 10 Ergebnissen der hybriden semantischen Suche, der Faktensuche und der semantischen Dokumentsuche. Für die hybride semantische Suche wurden beide Rankingverfahren (s. Kapitel 5.2.4) berücksichtigt, um diese vergleichen zu können.

In Bezug auf die Auswahl der geeigneten **Kennzahlen** sind drei Aspekte zu berücksichtigen:

- Die drei Suchmaschinen liefern *unterschiedlich lange Ergebnislisten*. Insbesondere die Anzahl der Ergebnisse der Faktensuche kann häufig unter k liegen und es sind mehrere korrekte Ergebnisse möglich. In diesem Fall eignen sich nur die Kennzahlen Genauigkeit und Vollständigkeit für die Messung der Retrievaleffektivität. Sie basieren auf der Anzahl der gefundenen relevanten Dokumente in den ersten k Ergebnissen und der Anzahl relevanter Dokumente im Dokumentpool. Die Kennzahl MRR betrachtet die Position der ersten korrekten Antwort, sie ist also auf die Evaluierung von Suchanfragen mit einer wohlbekannten Antwort ausgerichtet. Sind mehrere richtige Antworten möglich, so kann MAP eingesetzt werden. Hierfür müssen jedoch alle relevanten Dokumente bekannt sein und k ist für jede Suchanfrage so groß zu wählen, dass alle relevanten Ergebnisse darunter liegen. Deshalb ist MAP nur mit einem Gold Standard berechenbar, in dem die Relevanz für alle (oder eine hinreichend große Anzahl) Ergebnisse vorliegt.
- Binäre Relevanzurteile sind für unterschiedliche Ergebnistypen weniger geeignet als für homogene Ergebnismengen, da sie keine Differenzierung der Relevanz und daher der Qualität der Ergebnisse erlauben. Insbesondere wenn die Ergebnisse Fakten, Dokumente und hybrid, also Dokument mit Fakten, sein können, ist die Frage, ob ein Suchergebnis die Suchanfrage vollständig oder nur teilweise beantwortet, bedeutungsvoll. Aus der Sicht der Beurteiler ist die Möglichkeit einer Differenzierung der *Relevanzstufen* wesentlich, eine „ja/nein“-Entscheidung kann in vielen Fällen nur schwer getroffen werden [Robertson, 1977, Robertson and Belkin, 1978]. Die geeignete Anzahl der Relevanzstufen für IR wurde u. a. in [Tang et al., 1999] diskutiert. Die Untersuchung der Konfidenz der Relevanzurteile von 2 bis 11 Stufen konnte keine optimale Stufenanzahl hervorbringen. Die Autoren schlagen jedoch sieben Stufen vor, da über 6 Relevanzstufen ein leichter Anstieg der Konfidenz verzeichnet wurde. Meist werden jedoch aufgrund der einfacheren Beurteilung die drei Relevanzstufen „relevant“, „teilweise relevant“ und „nicht relevant“ eingesetzt

[Kekäläinen, 2005]. Die geeignete Anzahl Relevanzstufen ist von der Fragestellung abhängig, sie müssen ein verständliches und sinnvolles Antwortset auf die Frage bilden [Robertson, 1977]. Im Fall der hier beschriebenen Evaluierung stellt sich die Frage, inwieweit ein Suchergebnis die Suchanfrage beantwortet. Dies impliziert die Frage an den Beurteilern *„Beantwortet dieses Suchergebnis die Suchanfrage?“*. Der Antwortset *„ja“, „teilweise“, „weder ja noch nein“, „eher nein“* und *„nein“* deckt 5 Stufen ab und erlaubt dem Benutzer auszudrücken, inwieweit die Suchanfrage beantwortet wird. Zudem kann durch die Stufe „weder ja noch nein“ zum Ausdruck gebracht werden, dass ein Ergebnis zwar etwas mit der Suchanfrage zu tun hat, eine Relevanzentscheidung jedoch nicht getroffen werden kann.

- Die Genauigkeit und Vollständigkeit sind für binäre Relevanz ausgelegt.

Für mehrstufige Relevanz kann die *generalisierte Genauigkeit und Vollständigkeit* eingesetzt werden (s. Seite 37). Dabei werden der Relevanzstufen Zahlenwerte (z.B. 1,00, 0,75, 0,50, 0,25 und 0,00 oder 5, 4, ..., 1) zugeordnet und für die Berechnung der Kennzahlen aufaddiert anstatt die relevanten Dokumente zu zählen [Kekäläinen and Järvelin, 2002]. Diese Kennzahlen eignen sich für unterschiedlich lange Ergebnislisten mit mehrstufiger Relevanz, da weder die Position noch die Anzahl der Ergebnisse den Wert beeinflussen. Die Vergleichbarkeit unter den Suchmaschinen bleibt bestehen.

Für die Evaluierung wurden also die *Relevanzstufen „ja“, „teilweise“, „weder ja noch nein“, „eher nein“, „nein“ mit den Werten 1,00, 0,75, 0,50, 0,25, 0,00* eingesetzt und die *generalisierte Genauigkeit und Vollständigkeit über die ersten* 10 *Ergebnisse* und über *20 Suchanfragen* berechnet. Die Auswahl der Suchanfragen und die Erstellung des Gold Standards ist im folgenden Kapitel beschrieben.

Der Gold Standard

Zur Evaluierung der Effektivität von Suchmaschinen sollten **Suchanfragen** verwendet werden, die reale Informationsbedürfnisse ausdrücken [Gordon and Pathak, 1999]. Die Suchanfragen sollten zudem zu der jeweiligen Datenbasis passen, das formulierte Informationsbedürfnis sollte aus den Daten befriedigt werden können. Diese Voraussetzungen erfüllt die „Test Collection for DBpedia Search“ von [Balog and Neumayer, 2013]. Die Sammlung beinhaltet 485 natürlichsprachige Suchanfragen, bestehend aus Schlüsselwörtern mit 1-2 Termen bis hin zu komplexen Fragen. Sie ist aus 6 Query-Sets für Entitäten- bzw. Faktensuche, Frage-Antwort-Maschinen und semantische Dokumentsuche in der englischen DBpedia und Wikipedia zusammengestellt [Balog and Neumayer, 2013]. Die Kollektion deckt sowohl Faktensuche, als auch semantische Dokumentsuche ab und ist für eine Evaluierung von hybrider semantischer Suche und einen Vergleich mit Fakten- und semantischer Dokumentsuche bzw. Faktensuche sowie semantische Dokumentsuche geeignet. Insgesamt 430 der Suchanfragen (5 von 6 Sets) stammen aus Query-Logs[16], sie drücken reale Informationsbedürfnisse aus.

Nachteil der Testkollektion ist, dass sie durch das Zusammensetzen aus verschiedenen Query-Sets keine natürliche Verteilung in Bezug auf die Komplexität der Suchanfragen widerspiegelt. Eine zufällige Auswahl von Anfragen für die Evaluierung hätte zur Folge, dass Aussagen über die Güte des Suchansatzes nur auf die dadurch entstandene Verteilung beschränkt getroffen werden könnten. Eine Generalisierung hinsichtlich der

[16] Die 55 Suchanfragen aus INEX_XER sind von den Teilnehmern des INEX 2009 Entity Ranking Tracks für Entitätensuche erstellt (`http://www.l3s.de/~demartini/XER09/`, 02.11.2015).

realweltlichen Benutzung wäre nicht möglich. Die reale Verteilung der Komplexität von Suchanfragen in DBpedia-Query-Logs wurde jedoch in einer empirischen Studie auf Basis von 5.166.272 Einträgen ermittelt. Diese Menge ist statistisch relevant in Bezug auf die Anzahl der Suchanfragen und auf die Heterogenität der Personen, von denen sie stammen [Möller et al., 2010]. Die Komplexität einer Suchanfrage wurde anhand der Anzahl der Triple-Pattern, also der Anzahl Tripel in der zugehörigen SPARQL-Abfrage (mit oder ohne Variable) gemessen [Arias et al., 2011]. Die 430 Suchanfragen aus der „Test Collection for DBpedia Search“ wurden mit derselben Methode kategorisiert und die Anzahl der Anfragen pro Kategorie für eine Evaluierung mit 20 Suchanfragen berechnet. Tabelle 6.2 fasst die Verteilung aus [Arias et al., 2011] und die Anzahl auszuwählender Suchanfragen pro Komplexitätskategorie zusammen. Dabei wurden die Komplexitätstufen ab 4 Triple-Pattern zusammengefasst, da diese nur 2% der Suchanfragen ausmachen und basierend auf der Verteilung nur eine Anfrage aus dieser Kategorie ausgewählt wird wird.

Anzahl Tripel-Pattern	Verteilung in DBpedia-Query-Logs	Anzahl auszuwählender Anfragen
1	66%	13
2	5%	1
3	27%	5
>=4	2%	1

Tabelle 6.2: Verteilung und Anzahl auszuwählender Suchanfragen nach Komplexität

Neben der Repräsentation realer Informationsbedürfnisse und einer natürlichen Verteilung der Suchanfragen sollten diese die Features, hilfreiche Funktionalitäten der zu vergleichenden Suchmaschinen abdecken [Gordon and Pathak, 1999]. Bei einer komparativen Evaluierung der hybriden semantischen Suche SINFIO mit Fakten- und semantischer Dokumentsuche bedeutet dies, dass die Suchanfragen die Suche nach Fakten, nach Dokumenten und auch hybride Ergebnisse abdecken sollten. Aus diesem Grund wurden zuerst aus jeder Komplexitätskategorie eine der natürlichen Verteilung entsprechende Anzahl von Suchanfragen zufällig ausgewählt und anschließend auf die Abdeckung der genannten Kategorien geprüft. Mit diesem Verfahren wurde die folgende Menge an Suchanfragen bestimmt:

q01: tango music instruments
q02: What is the longest river?
q03: What is the capital of Canada?
q04: Give me all companies in Munich.
q05: What is the currency of the Czech Republic?
q06: air wisconsin
q07: austin powers
q08: What are the official languages of the Philippines?
q09: keith urban
q10: overeaters anonymous
q11: boroughs of New York City
q12: five great epics of Tamil literature
q13: john lennon, parents

q14: Give me all launch pads operated by NASA.
q15: What is the largest city in Australia?
q16: Which country does the creator of Miffy come from?
q17: Through which countries does the Yenisei river flow?
q18: In which films directed by Garry Marshall was Julia Roberts starring?
q19: Give me all cars that are produced in Germany.
q20: Give me all actors starring in movies directed by and starring William Shatner

Dieselben Suchanfragen eignen sich nicht notwendigerweise für jedes Suchsystem, da die Suchsysteme unterschiedliche Arten von Anfragen unterstützen können. Gegebenenfalls müssen sie entsprechend den Eigenschaften der zu untersuchenden Suchmaschine umformuliert werden, jedoch ohne den Sinn zu verändern oder ein System zu bevorteilen (vgl. Kapitel 3.4). Da jedoch für jede der Suchmaschinen eine umfangreiche Anfragevorverarbeitungskomponente implementiert wurde, Pluralformen sowie Ausdrücke erkennt (vgl. Kapitel 5.2.1) und die Faktensuche zwar die Reihenfolge der Terme berücksichtigt aber durch mehrfache Iterationen auch nicht nebeneinander stehende Terme kombiniert (vgl. Kapitel 5.2.3.1), ist dies nicht notwendig.

Für die Bestimmung des **Gold Standard**s wurde für jede Suchanfrage ein Pool aus den ersten 10 Ergebnissen der Faktensuche, semantischer Dokumentsuche und der hybriden semantischen Suche erstellt. Der Pool beinhaltet damit auch die ersten 10 Ergebnisse der Fakten- und semantischen Dokumentsuche zusammen. Die Relevanzurteile wurden mithilfe einer Onlineumfrage, erstellt mit LimeSurvey[17], von insgesamt 14 Personen durchgeführt. Die Umfrageteilnehmer waren Informationswissenschaftler und weitere Akademiker, die aufgrund ihrer beruflichen Laufbahn reichlich Erfahrung in Recherchearbeiten haben und eine Relevanzbeurteilung vornehmen können (vgl. [Manning et al., 2008b]). Die Teilnehmer waren weder mit dem Inhalt dieser Arbeit vertraut noch selber im Bereich semantischer bzw. hybrider semantischer Suche tätig. Abbildung 6.3 zeigt ein Ausschnitt aus der Umfrage.

Die Korrelation der Relevanzurteile wurde mit der generalisierten Kappa-Statistik für unbalancierte Randwerte berechnet, der sogenannten *multirater free-marginal Kappa*, da die Verteilung über die Relevanzstufen nicht a priori bekannt ist (s. Kapitel 2.5.1, Formeln 2.18, 2.19 sowie 2.21). Der *Kappa-Wert* über die 20 Pools von Ergebnissen der Suchanfragen *beträgt 0,593* mit einem Standardfehler von 0,007. Der erreichte Kappa-Wert hat daher eine hohe statistische Signifikanz (vgl. [Fleiss, 1981]) und liegt an der oberen Grenze des Intervalls der fairen Übereinstimmung ([0,41 - 0,6], vgl. [Viera et al., 2005] und Kapitel 2.5.1).

Eine Zwischenauswertung der Relevanzbeurteilungen von 7 Personen im Vergleich zu der Auswertung über die 14 Antwortsets ergab, trotz Verdoppelung der Teilnehmer, lediglich eine Verbesserung um 0,8 Punkte. Mit Werten im Bereich 0,2-0,38 verzeichneten die Suchanfragen q06, q09 und q10 die niedrigste Korrelation. Sie sind, verglichen zu den anderen, eher allgemein gehalten und ihre Kappa-Werte haben sich durch die steigende Anzahl der Relevanzbeurteilungen nicht verbessert. Die höchsten Kappa-Werte im Bereich 0,72-0,78 sind bei q05, q16 und q18 zu verzeichnen. Alle drei Anfragen konnten

[17] LimeSurvey erlaubt die Angabe von HTML-Snippets und CSS anstatt natürlichsprachiger Fragen. Daher konnten die Suchergebnisse in ihrer ursprünglicher Form eingesetzt werden. `https://www.limesurvey.org` (06.01.2016)

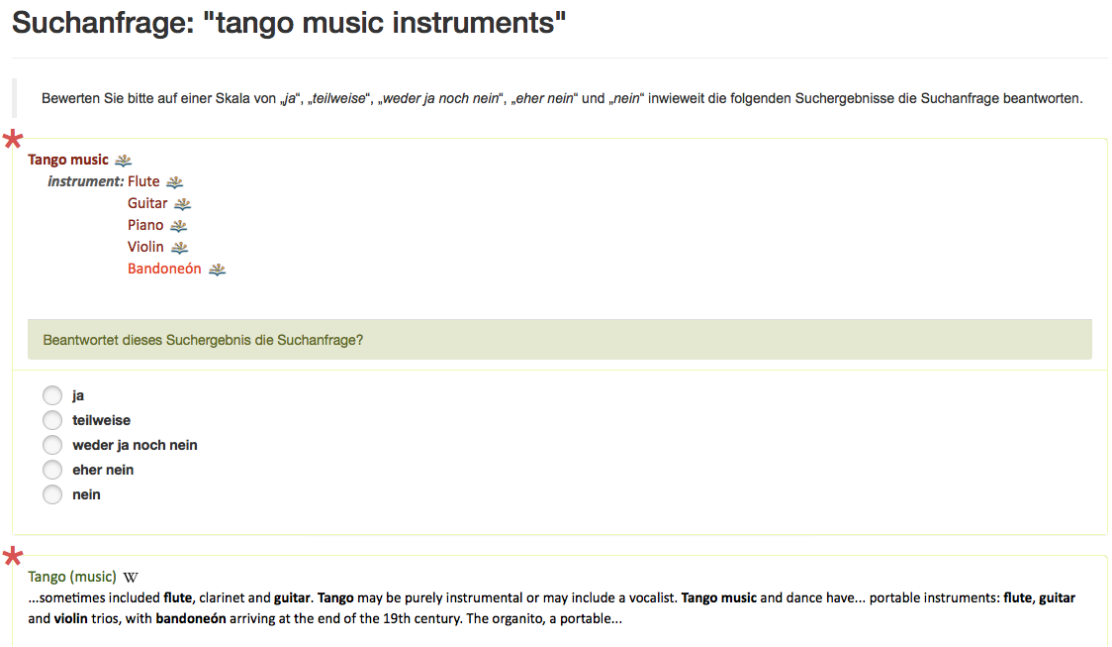


Abbildung 6.3: Abfrage der Relevanzurteile

mit Fakten vollständig beantwortet werden. Auch in diesem Fall gab es keine wesentliche Verbesserung des Kappa-Wertes zu verzeichnen. Dieses Verhalten deutet darauf hin, dass durch eine Hinzunahme von 5-10 weiteren Personen, wie es im Rahmen dieser Arbeit möglich wäre, keine wesentliche Erhöhung zu erwarten ist.

Die Evaluierung ist zuverlässig im Sinne der Reproduzierbarkeit (Reliabilität), sie ist mit geeigneten Methoden und Kennzahlen durchgeführt (Validität) und unabhängig vom Prüfer (Objektivität) (vgl. Kapitel 2.5).

Ergebnisse

Tabelle 6.3 fasst die generalisierte Genauigkeit, generalisierte Vollständigkeit und das daraus berechnete F-Maß über die ersten 10 Ergebnisse zusammen. Verglichen wird die hybride semantische Suche SINFIO mit der Fakten- und semantischer Dokumentsuche ohne eine Kombination der Inhalte, wobei für die hybride Suche sowohl das Ranking ohne Popularität als auch das Ranking mit Popularität (s. Kapitel 5.2.4) betrachtet wurde. Das F-Maß ist mit $\beta = 1$ berechnet, Genauigkeit und Vollständigkeit zählen gleichermaßen (s. Seite 37).

Hybride Ergebnisse können ein Dokument und zugehörige Fakten beinhalten, die im Falle der semantischen Dokument- und Faktensuche getrennt vorkommen und auch alleinstehend als ein relevantes Ergebnis im Pool sein können. Um dies bei der Berechnung der Vollständigkeit zu berücksichtigen[18], zählen die in den hybriden Ergebnissen vorkommenden Fakten und Dokumente als gefunden, d.h. in der Ergebnismenge enthalten. Der Umkehrschluss gilt jedoch nicht, da hybride Ergebnisse als solches als relevant be-

[18] Fließt in den Nenner, Summe der Relevanzwerte im Pool, mit ein. Bei der Berechnung der Genauigkeit kommt dies nicht zum tragen, da die ersten k Ergebnisse im Pool vorkommen und ihre Relevanz bekannt ist.

Such-anfrage	Hybride semantische Suche, Ranking ohne Popularität			Hybride semantische Suche, Ranking mit Popularität			Fakten- und semantische Dokumentsuche		
	gPrecision	gRecall	F-Measure	gPrecision	gRecall	F-Measure	gPrecision	gRecall	F-Measure
q01	0,2750	1,0000	0,4314	0,2750	1,0000	0,4314	0,2000	0,7273	0,3137
q02	0,1750	1,0000	0,2979	0,1750	1,0000	0,2979	0,0000	0,0000	0,0000
q03	0,2250	0,7500	0,3462	0,2250	0,7500	0,3462	0,2250	0,7500	0,3462
q04	0,7500	1,0000	0,8571	0,7500	1,0000	0,8571	0,7500	1,0000	0,8571
q05	0,5000	0,8333	0,6250	0,5000	0,8333	0,6250	0,3000	0,5000	0,3750
q06	0,3333	0,6667	0,4444	0,3333	0,6667	0,4444	0,2000	0,6667	0,3077
q07	0,7750	0,9688	0,8611	0,7500	0,9375	0,8333	0,6250	0,7576	0,6849
q08	0,3750	1,0000	0,5455	0,3750	1,0000	0,5455	0,0750	0,2000	0,1091
q09	0,8000	1,0000	0,8889	0,8000	1,0000	0,8889	0,8000	1,0000	0,8889
q10	0,7750	1,0000	0,8732	0,7750	1,0000	0,8732	0,7750	1,0000	0,8732
q11	0,3250	0,7647	0,4561	0,1000	0,2000	0,1333	0,1750	0,3500	0,2333
q12	0,6500	1,0000	0,7879	0,0000	0,0000	0,0000	0,6250	0,9615	0,7576
q13	0,4750	1,0000	0,6441	0,4750	1,0000	0,6441	0,3250	0,6842	0,4407
q14	0,6429	1,0000	0,7826	0,6429	1,0000	0,7826	0,6750	0,9000	0,7714
q15	0,3250	0,8125	0,4643	0,3250	0,8125	0,4643	0,2750	0,5500	0,3667
q16	0,3500	1,0000	0,5185	0,3500	1,0000	0,5185	0,1000	0,2857	0,1481
q17	0,3500	1,0000	0,5185	0,3500	1,0000	0,5185	0,9167	0,7857	0,8462
q18	0,3000	1,0000	0,4615	0,3000	1,0000	0,4615	0,3000	1,0000	0,4615
q19	0,7500	1,0000	0,8571	0,7500	1,0000	0,8571	0,7500	1,0000	0,8571
q20	0,7500	1,0000	0,8571	0,7500	1,0000	0,8571	0,7500	1,0000	0,8571
Durch-schnitt	**0,4951**	**0,9398**	**0,6485**	0,4501	0,8600	0,5909	0,4421	0,7059	0,5437

Tabelle 6.3: Generalisierte Genauigkeit (gPrecision), Vollständigkeit (gRecall) und F-Maß (F-Measure, $\beta = 1$)

urteilt worden sind. Nur die Fakten oder das Dokument in einem hybriden Ergebnis ist alleinstehend nicht notwendigerweise relevant. Dies ist insbesondere für die Relevanzstufe „ja“, das Ergebnis beantwortet die Suchanfrage, der Fall. Um die hybride Suche nicht zu bevorteilen, wurde die Vollständigkeit für die Fakten- und semantische Dokumentsuche ohne Einbezug der hybriden Ergebnisse berechnet. Die bei der hybriden Suche gefundenen Fakten und Dokumente, die kein hybrides Ergebnis bilden, wurden jedoch mit einbezogen.

Die Werte der generalisierten Genauigkeit und Vollständigkeit sind von den Zahlenwerten der Relevanzstufen abhängig, daher sind sie nur in Relation unter den betrachteten Suchverfahren interpretierbar. Bei Relevanzstufenwerten 1,00, 0,75, 0,50, 0,25, 0,00 und $k = 10$ beträgt das Maximum jeweils 1,0.

Die *höchste* **Genauigkeit** erreichen die Suchanfragen q04, q07, q09, q10, q19 und q20, wobei 3 davon (q04, q19, q20) nur Fakten und eine (q07) 8 hybride Ergebnisse unter den ersten 10 Ergebnissen beinhalten. In diesem Fall erzielt die hybride semantische Suche mit Ranking ohne Popularität die höchste Genauigkeit und auch die höchste Vollständigkeit. In den anderen Fällen sind die Werte gleich, da Fakten und Dokumente nicht zu hybriden Ergebnissen kombiniert wurden und die gefundenen Fakten keinen Einfluss darauf hatten, welche (top 10) Dokumente gefunden wurden.

Die *niedrigsten Genauigkeiten* treten bei q01, q02 und q03 auf. Die Suchanfragen q01 und q03 werden jeweils durch ein Ergebnis als Fakten beantwortet, hinzu kommen nur

wenige (1-2) relevante Dokumente unterschiedlicher Relevanzstufen. Im Falle von q02 gibt es insgesamt nur 2 Ergebnisse mit einer Relevanzstufe höher als 0 im Pool. Diese werden nur von den hybriden Suchmaschinen gefunden. Die Anzahl der nicht relevanten Ergebnisse unter den ersten 10 Antworten überwiegt also in allen drei Fällen, die Präzision ist deshalb sehr niedrig. Bei q01 und q02 ist die Performanz der hybride Suche besser, bei q03 erreichen alle Suchmaschinen dieselben Werte.

Die Unterschiede zwischen den zwei hybriden Suchmaschinen werden in Zusammenhang mit der Evaluierung der Rankingverfahren, im Kapitel 6.2.3, analysiert.

Das höchste **F-Maß** mit 0,6485, also die *höchste Retrievaleffektivität* erreichte die hybride semantische Suche mit Ranking ohne Popularität. Das F-Maß der Fakten- und semantischen Dokumentsuche liegt mit 0,5437 über 10 Prozentpunkte darunter. Die Verbesserung des F-Maßes ist mit p =0,10 statistisch nicht signifikant (bei einem α von 0,05). Die Wahrscheinlichkeit dessen, dass die Differenz von 10 Prozentpunkte entsteht, wenn beide Systeme gleich gut sind, beträgt 10% (s. Seite 115).

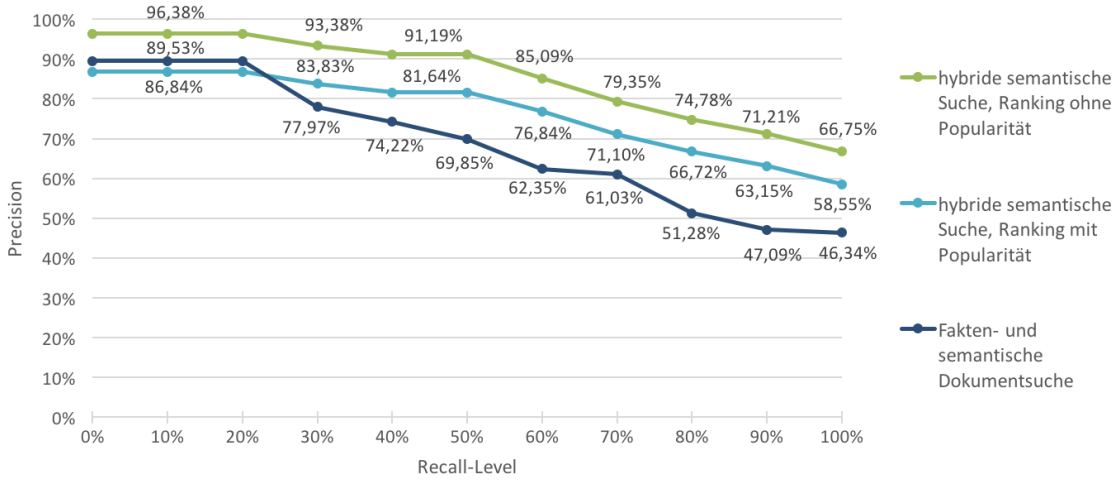


Abbildung 6.4: Interpolierte, generalisierte Precision-Recall Kurve

Abbildung 6.4 zeigt die Precision-Recall-Kurve (Genauigkeit-Vollständigkeit-Diagramm), berechnet auf Basis der generalisierten Genauigkeit und Vollständigkeit über alle Relevanzstufen sowie Suchanfragen [Kekäläinen and Järvelin, 2002] und interpoliert auf die 11 Standard-Recall-Ebenen 0%, 10%..., 100% [Baeza-Yates and Ribeiro-Neto, 2011d]. Die hybride semantische Suche mit dem Rankingverfahren ohne Popularität erreicht auf jeder Recall-Ebene die höchste Präzision. Je höher der Vollständigkeit, umso größer ist der Unterschied zu der Fakten- und semantischen Dokumentsuche. Die hybride semantische Suche, Ranking mit Popularität performt weniger gut, sie liegt erst ab 30% Recall über der Fakten- und semantische Dokumentsuche. Ein Vergleich der beiden Rankingverfahren sowie die Analyse der Unterschiede erfolgt in Zusammenhang mit ihrer Evaluierung im Kapitel 6.2.3.

Hybride Ergebnisse kommen bei den Suchanfragen q07, q11 und q14 vor. Für q07 und q11 erreicht die hybride Suche mit Ranking ohne Popularität die höchste generalisierte Genauigkeit und Vollständigkeit. Lediglich für q14 liefert die Fakten- und semantische Dokumentsuche zusammen eine höhere Genauigkeit, jedoch eine niedrigere Vollständigkeit.

Such-anfrage	Faktensuche			Semantische Dokumentsuche		
	gPrecision	gRecall	F-Measure	gPrecision	gRecall	F-Measure
q01	1,0000	0,3636	0,5333	0,1000	0,3636	0,1569
q02	0,0000	0,0000	0,0000	0,1750	1,0000	0,2979
q03	1,0000	0,3333	0,5000	0,2000	0,6667	0,3077
q04	0,7500	1,0000	0,8571	0,3000	0,4000	0,3429
q05	1,0000	0,1667	0,2857	0,2000	0,3333	0,2500
q06	0,0000	0,0000	0,0000	0,4000	0,6667	0,5000
q07	0,7500	0,0909	0,1622	0,6250	0,7576	0,6849
q08	0,7500	0,2000	0,3158	0,0000	0,0000	0,0000
q09	1,0000	0,1250	0,2222	0,7750	0,9688	0,8611
q10	0,0000	0,0000	0,0000	0,7750	1,0000	0,8732
q11	0,7500	0,1500	0,2500	0,1000	0,2000	0,1333
q12	0,0000	0,0000	0,0000	0,6500	1,0000	0,7879
q13	0,7500	0,4737	0,5806	0,1000	0,2105	0,1356
q14	0,7500	0,8000	0,7742	0,3750	0,5000	0,4286
q15	1,0000	0,2000	0,3333	0,1750	0,3500	0,2333
q16	1,0000	0,2857	0,4444	0,0000	0,0000	0,0000
q17	1,0000	0,2857	0,4444	0,8750	0,5000	0,6364
q18	0,3000	1,0000	0,4615	0,0000	0,0000	0,0000
q19	0,7500	1,0000	0,8571	0,0000	0,0000	0,0000
q20	0,7500	1,0000	0,8571	0,0000	0,0000	0,0000
Durch-schnitt	0,6650	0,3737	0,4785	0,2913	0,4459	0,3523

Tabelle 6.4: Generalisierte Genauigkeit (gPrecision), Vollständigkeit (gRecall) und F-Maß (F-Measure) der Faktensuche und der semantischen Dokumentsuche

Neben der Fakten- und Dokumentsuche wurde auch die **Retrievaleffektivität der Faktensuche** und **der semantischen Dokumentsuche** berechnet, Tabelle 6.4 zeigt die Ergebnisse. Das F-Maß der Faktensuche mit 0,4785 und das F-Maß der semantischen Dokumentsuche mit 0,3523 liegt unter dem F-Maß der hybriden semantischen Suche (Ranking ohne Popularität 0,6485, Ranking mit Popularität 0,5909). Die Verbesserung des F-Maßes ist in diesem Fall sowohl im Vergleich zu der Faktensuche ($p<0{,}004$ bzw. 0,026) als auch im Vergleich zu der semantischen Dokumentsuche ($p<0{,}0006$ bzw. 0,006) statistisch signifikant.

Die höhere Retrievaleffektivität der hybriden semantischen Suche im Vergleich zu der Fakten- und semantische Dokumentsuche **bestätigt die These und die Testhypothese 1 aus Systemsicht**. Die Verbesserung ist jedoch mit p =0,10 *statistisch nicht signifikant*: es besteht 10% Wahrscheinlichkeit, dass die gemessene Differenz auch dann entsteht, wenn beide Suchmaschinen gleich gut sind.

6.2.2.3 Evaluierung aus Benutzersicht

Mit Pooling und der generalisierten Genauigkeit sowie Vollständigkeit kann die Güte des Suchverfahrens aus Systemsicht gemessen werden. Das sagt aber nichts darüber aus, wie zufrieden die Benutzer mit den Suchmaschinen sind. Es bleibt ebenso unbekannt, wie

häufig durch die gefundenen Fakten und hybriden Ergebnisse das Suchbedürfnis schneller befriedigt werden konnte. Um dies zu untersuchen, wurde eine Benutzerbefragung durchgeführt. Der hybride Suchansatz kam mit dem Rankingverfahren ohne Popularität zum Einsatz, da dieses Verfahren bei der Untersuchung der Effektivität aus Systemsicht besser abgeschnitten hat.

Methode

Zum Vergleich von zwei Suchsystemen eignen sich **Side-by-side Panel** (vgl. Kapitel 2.5.3). Dabei werden die ersten k Suchergebnisse von zwei Suchmaschinen nebeneinander dargestellt und die Evaluierer gefragt, welche Ergebnisse sie besser finden [Baeza-Yates and Ribeiro-Neto, 2011d]. Durch ein einheitliches Design und eine zufällige Anordnung soll sichergestellt werden, dass die Suchmaschinen nicht identifizierbar sind. Dies kann in diesem Fall jedoch nicht vollständig erreicht werden, da die unterschiedlichen Suchmaschinen in einigen Fällen an der Art der Antworten erkennbar sind. Die Beschreibung zu der Umfrage gibt jedoch keine Auskunft darüber, welche Suchmaschine im Fokus steht und keiner der Teilnehmer hat Kenntnisse über den Inhalt dieser Arbeit. Die hybride Suchmaschine SINFIO kann nicht als Untersuchungsgegenstand erkannt werden.

Um generalisierbare Aussagen treffen zu können, sollten Personen unterschiedlicher Alters- und Berufsgruppen sowie etwa gleicher Anteil beider Geschlechter befragt werden. Abbildung 6.5 zeigt ein Side-by-side Panel aus der Umfrage.

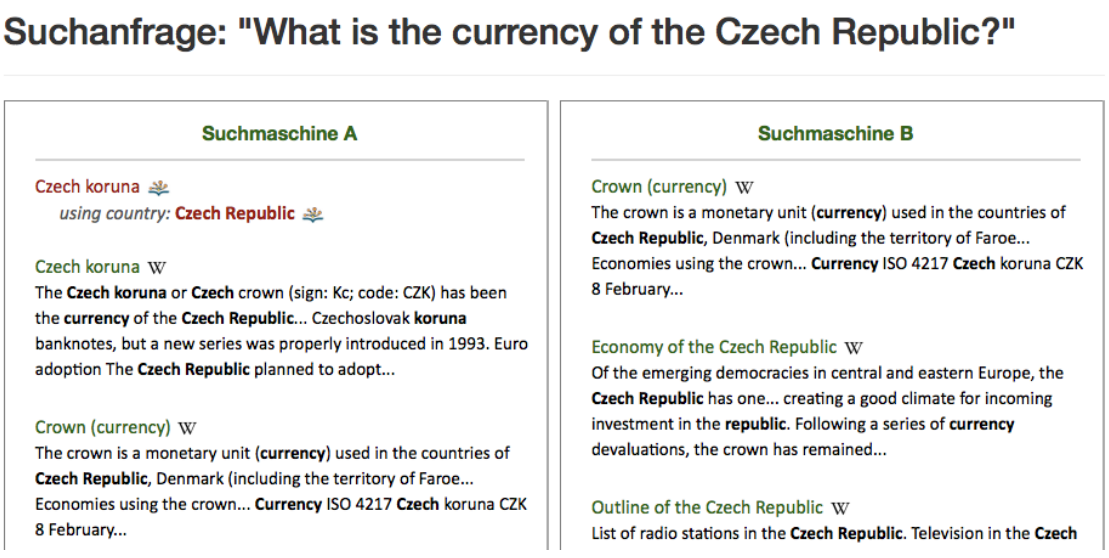


Abbildung 6.5: Side-by-side Panel für die Erhebung der Effektivität aus Benutzersicht

Für die benutzerzentrierte Evaluierung der *These und der Testhypothese 1* ist von Interesse, *ob die Benutzer die Ergebnisse der hybriden semantischen Suche oder der Fakten- und semantischen Dokumentsuche zusammen besser finden* und warum. *Testhypothesen 2 und 3* erfordern *Vergleiche mit der semantischen Dokumentsuche und der Faktensuche, sowie die Frage ob und in welcher Suchmaschine die Antwort schneller gefunden werden konnte* (vgl. Kapitel 6.2.2.1). Um zu erfahren, ob tatsächlich Zeit eingespart werden konnte, weil *die Antwort aus der Ergebnisliste abgelesen werden konnte*, werden die Benutzer auch zu diesem Punkt befragt. Abbildung 6.6 zeigt die Fragen der Umfrage.

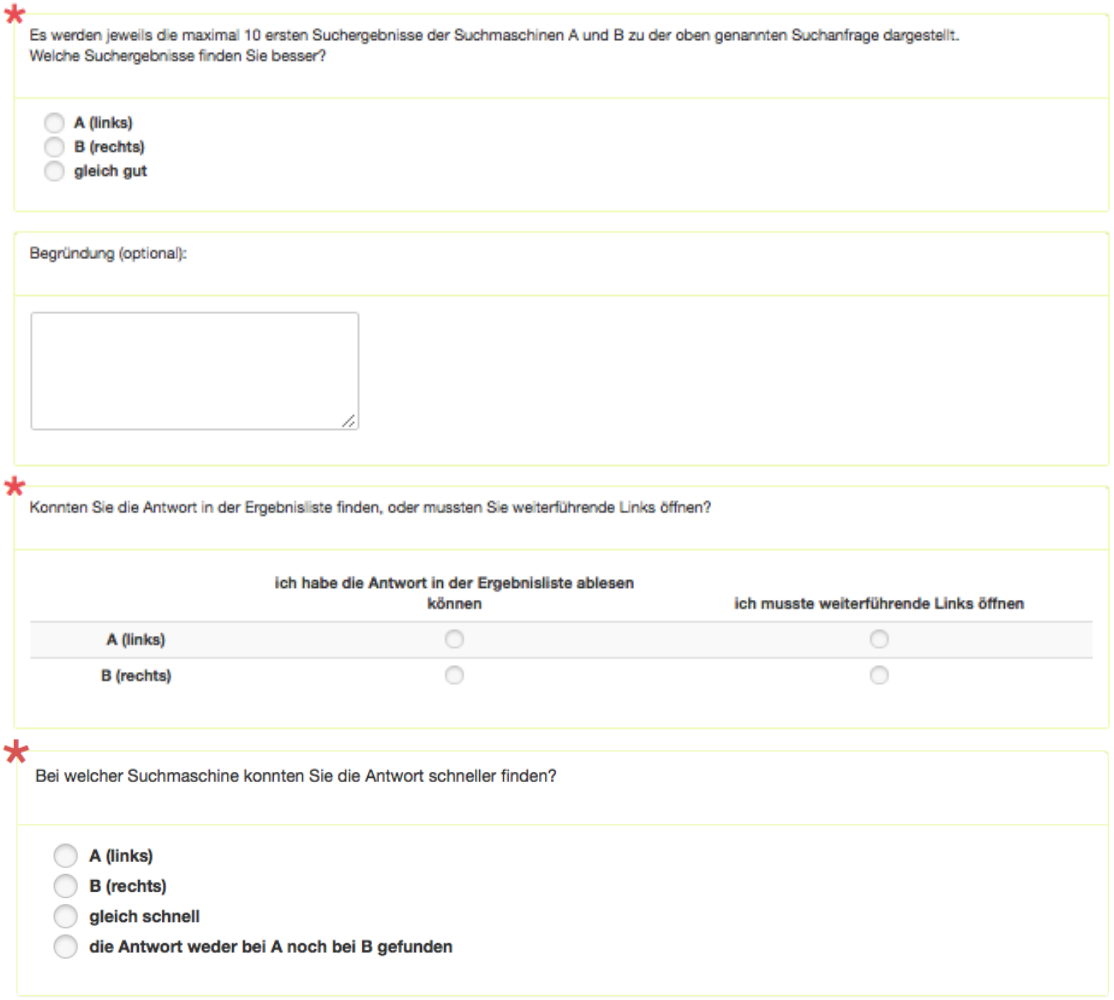


Abbildung 6.6: Fragen für die Erhebung der Effektivität aus Benutzersicht

Um die Testhypothesen zu prüfen, wurde eine **Online-Umfrage**[19] mit Side-by-side Panel zu den 20 Suchanfragen durchgeführt, wobei für jede Suchanfrage die hybride semantische Suche (Ranking ohne Popularität) mit Fakten- und semantischer Dokumentsuche zusammen, mit der Faktensuche und mit der semantischen Dokumentsuche verglichen wird. Die insgesamt 20 Teilnehmer der Umfrage - 12 weiblich (60%), 8 männlich (40%) - entstammen den Berufen Kartograph, Computerlinguist, Künstler, Handwerker, PR-Berater, technischer Manager, Pädagoge, Bankkauffrau, Sekretärin, Ökonom, Journalist, Jurastudent und Psychologe mit, einer der Teilnehmer war Informatiker. Das Alter wurde in 10-Jahres-Intervallen abgefragt, 20% der Teilnehmer waren zwischen 20 und 30, 55% zwischen 31 und 40, 20% zwischen 41 und 50 und 5% über 50 Jahre alt. Eine der realen Welt entsprechende Verteilung wurde damit angenähert, aber nicht erreicht. Im Jahr 2014 haben die Frauen 51%, die Männer 49% der Bevölkerung im Deutschland ausgemacht[20]. Die Suchmaschinennutzung pro Altersgruppe ist nach einer anderen Altersgruppeneinteilung verfügbar. Demnach machen Personen im Alter von 14 bis 29 Jahre 29,44%, von

[19] Durchgeführt mit LimeSurvey, `https://www.limesurvey.org` (06.01.2016).
[20] `http://de.statista.com/statistik/daten/studie/161868/umfrage/entwicklung-der-gesamtbevoelkerung-nach-geschlecht-seit-1995/` (06.01.2016)

30 bis 49 Jahre 42,43%, von 50 bis 64 Jahre 19,30% und über 65 Jahre 8,83% der regelmäßigen Suchmaschinennutzer im 2015 in Deutschland aus[21].

Ergebnisse

Im Folgenden werden die Ergebnisse der einzelnen Fragen vorgestellt und im Hinblick auf die These und Testhypothesen analysiert.

Zur Überprüfung der *These* bzw. der zugehörigen *Testhypothese 1* (s. Kapitel 6.2.2.1) dient die Frage **„Welche Suchergebnisse finden Sie besser?“**. Tabelle 6.5 listet die Ergebnisse je Vergleich und Suchanfrage auf. Die kursiven Einträge markieren Fälle, in denen die ersten $k = 10$ Ergebnisse beider Suchmaschinen gleich sind und deshalb kein Vergleich stattgefunden hat. Sie wurden als „gleich gut“ gewertet.

Im Durchschnitt über alle Suchanfragen beurteilten die Teilnehmer die hybride semantische Suche und die Fakten- und semantische Dokumentsuche zu 51,75% als gleich gut, 44,5% fanden die hybride semantische Suche, 3,75% die Fakten- und semantische Dokumentsuche besser. Die Differenz zugunsten der hybriden semantischen Suche ist statistisch signifikant (p =0,0001). Insgesamt in 16 Fällen fand keiner der Teilnehmer die Fakten- und semantische Dokumentsuche besser, und nur in einem Fall lag die Fakten- und semantische Dokumentsuche vor der hybriden Suche. Die zwei Suchmaschinen liefern in 7 Fällen dieselbe Ergebnisliste, nämlich wenn:

- keine Fakten gefunden wurden;
- die mit den gefundenen Fakten erweiterte Suchanfrage (vgl. Kapitel 5.2.3.3) zu denselben ersten 10 Wikipediaseiten führt wie die semantische Dokumentsuche und diese nicht zu hybriden Ergebnissen kombiniert werden.

Betrachtet man jeweils nur die Suchanfragen mit unterschiedlichen Ergebnislisten[22], so haben 68,46% der Teilnehmer die hybride semantische Suche, 5,77% die Fakten- und semantische Dokumentsuche besser gefunden, 30,77% beide gleich gut (s. Abbildung 6.7). Dies ist kein generalisierbares Ergebnis, zeigt jedoch, dass bei Suchanfragen mit unterschiedlichen top 10 Ergebnissen die hybride semantische Suche als deutlich besser empfunden wird.

[21] Berechnet aus `http://de.statista.com/statistik/daten/studie/354307/umfrage/regelmaessige-taetigkeiten-im-internet-nach-altersgruppen/`, `http://de.statista.com/statistik/daten/studie/3101/umfrage/internetnutzung-in-deutschland-nach-altersgruppen/` und `http://de.statista.com/statistik/daten/studie/1351/umfrage/altersstruktur-der-bevoelkerung-deutschlands/` (06.01.2016).

[22] In diesem Fall ist ein Vergleich der Ergebnisse unter den drei Suchmaschinenvergleichen nicht durchführbar, da diese über unterschiedliche Anfragemengen entstanden.

Es werden jeweils die maximal 10 ersten Suchergebnisse der Suchmaschinen A und B zu der oben genannten Suchanfrage dargestellt. Welche Suchergebnisse finden Sie besser?									
Suchanfrage	**hybride vs. sem. Dokumentsuche**			**hybride vs. Faktensuche**			**hybride vs. Fakten- und sem. Dokumentsuche**		
	hybride sem. Suche	sem. Dokumentsuche	gleich gut	hybride sem. Suche	Faktensuche	gleich gut	hybride sem. Suche	Fakten- und sem. Dokumentsuche	gleich gut
q01	75,00%	15,00%	10,00%	70,00%	0,00%	30,00%	70,00%	0,00%	30,00%
q02	20,00%	20,00%	60,00%	95,00%	0,00%	5,00%	95,00%	0,00%	5,00%
q03	100,00%	0,00%	0,00%	40,00%	35,00%	25,00%	10,00%	10,00%	80,00%
q04	80,00%	5,00%	15,00%	*0,00%*	*0,00%*	*100,00%*	*0,00%*	*0,00%*	*100,00%*
q05	90,00%	0,00%	10,00%	45,00%	30,00%	25,00%	50,00%	15,00%	35,00%
q06	20,00%	10,00%	70,00%	85,00%	0,00%	15,00%	*0,00%*	*0,00%*	*100,00%*
q07	40,00%	40,00%	20,00%	100,00%	0,00%	0,00%	60,00%	0,00%	40,00%
q08	100,00%	0,00%	0,00%	60,00%	10,00%	30,00%	100,00%	0,00%	0,00%
q09	35,00%	10,00%	55,00%	70,00%	5,00%	25,00%	*0,00%*	*0,00%*	*100,00%*
q10	*0,00%*	*0,00%*	*100,00%*	*0,00%*	*0,00%*	*100,00%*	*0,00%*	*0,00%*	*100,00%*
q11	65,00%	20,00%	15,00%	90,00%	0,00%	10,00%	75,00%	0,00%	25,00%
q12	*0,00%*	*0,00%*	*100,00%*	100,00%	0,00%	0,00%	90,00%	0,00%	10,00%
q13	70,00%	15,00%	15,00%	70,00%	0,00%	30,00%	90,00%	0,00%	10,00%
q14	75,00%	15,00%	10,00%	10,00%	45,00%	45,00%	10,00%	45,00%	45,00%
q15	95,00%	5,00%	0,00%	35,00%	15,00%	50,00%	90,00%	0,00%	10,00%
q16	100,00%	0,00%	0,00%	35,00%	5,00%	60,00%	70,00%	5,00%	25,00%
q17	80,00%	10,00%	10,00%	45,00%	0,00%	55,00%	80,00%	0,00%	20,00%
q18	100,00%	0,00%	0,00%	*0,00%*	*0,00%*	*100,00%*	*0,00%*	*0,00%*	*100,00%*
q19	90,00%	0,00%	10,00%	*0,00%*	*0,00%*	*100,00%*	*0,00%*	*0,00%*	*100,00%*
q20	90,00%	0,00%	10,00%	*0,00%*	*0,00%*	*100,00%*	*0,00%*	*0,00%*	*100,00%*
Durchschnitt - alle Anfragen	**66,25%**	8,25%	25,50%	**47,50%**	7,25%	45,25%	44,50%	3,75%	**51,75%**
Durchschnitt - nur unterschiedl. Ergebnislisten	**73,61%**	9,17%	17,22%	**63,33%**	9,67%	27,00%	**68,46%**	5,77%	25,77%

Tabelle 6.5: Auswertung der Antworten zu der Frage „Welche Suchergebnisse finden Sie besser?“

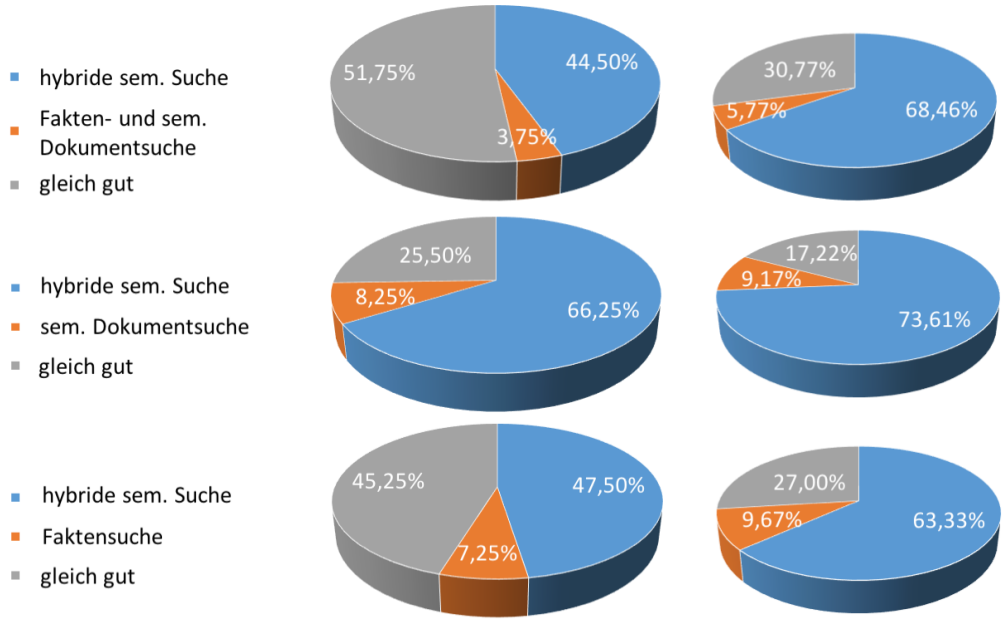


Abbildung 6.7: Überblick der Suchmaschinenvergleiche aus Benutzersicht, links über alle Suchanfragen, rechts nur über die Suchanfragen mit unterschiedlichen top 10 Ergebnissen der jeweiligen Suchmaschinen

Der Vergleich mit der semantischen Dokumentsuche sowie mit der Faktensuche zeigt ebenfalls den Vorteil der hybriden semantischen Suche. Zu 66,25% wurde die hybride Suche besser, und zu 91,75% gleich gut oder besser beurteilt, als die semantische Dokumentsuche. Lediglich zu 8,25% wurde die semantische Dokumentsuche als besser empfunden. Die Differenz ist statistisch signifikant (p =1,969e^{-8}). Der Vorteil gegenüber der Faktensuche fällt kleiner aus, hier fanden 47,50% die hybride Suche besser, 92,75% gleich gut oder besser. Für die Faktensuche stimmten die Teilnehmer jedoch nur zu 7,25%. Das Ergebnis ist statistisch signifikant (p =0,00006). Eine tiefergehende Analyse der Daten wird weiter unten, in Zusammenhang mit der Auswertung der anderen zwei Fragen beschrieben.

Die Ergebnisse bestätigen die These sowie die Testhypothese 1: *Auch die Benutzer beurteilen die hybride semantische Suche als effektiver als die Fakten- und semantische Dokumentsuche zusammen*, jedoch ohne eine Kombination der formalen und informalen Inhalte im Suchprozess. Die Verbesserung der hybriden semantischen Suche gegenüber der Fakten- und Dokumentsuche ist *statistisch signifikant*.

Die Ergebnisse der Befragung **„Konnten Sie die Antwort in der Ergebnisliste ablesen oder mussten Sie hierfür weiterführende Links öffnen?“** sind in der Tabelle 6.6 gelistet. Die Vergleiche mit denselben Ergebnislisten fehlen, denn sie lassen sich nicht aus den vorgenommenen Beurteilungen ableiten. Da die Suchanfrage q10 für alle Suchmaschinen dieselben Ergebnisse (Wikipediaseiten) liefert, fiel sie aus der Bewertung heraus.

Die Beurteilung für die *hybride semantische Suche* fiel je nach Vergleichssuchmaschine unterschiedlich aus: Die Werte liegen zwischen 84,17% - 79,67% für das Finden der Antwort in der Ergebnisliste und zwischen 15,83% - 20,33% für das Öffnen weiterführender Links. Sie übertrifft in jedem Vergleich die jeweils andere Suchmaschine. Gemessen an der Anzahl der Fälle, in der die Antwort in der Ergebnisliste abgelesen werden konnte, ist

der Unterschied zwischen der hybriden semantischen Suche und der Fakten- und Dokumentsuche (p =0,049), der Faktensuche (p =0,016) sowie der semantischen Dokumentsuche (p =1,961e^{-6}) statistisch signifikant. Im Vergleich zu der Fakten- und semantischen Dokumentsuche, die 62,31% für das Ablesen der Ergebnisse erreichte, erhielt die hybride semantische Suche mit 81,15% nahezu 20% mehr Zustimmung.

In Bezug auf die *Testhypothese 2* (s. Kapitel 6.2.2.1) ist der *Vergleich mit der semantischen Dokumentsuche* interessant. Hier konnte nach Angaben der Benutzer lediglich in 36,94% der Fälle die Antwort aus der Ergebnisliste abgelesen werden, in 63,06% mussten weiterführende Links geöffnet werden. Eine Überprüfung der Hypothese ist aber nur in Zusammenhang mit der Auswertung aus Tabelle 6.7 möglich. Sie listet je Suchmaschinenvergleich für die Suchanfragen die Anzahl Antworten auf die Frage **„Bei welcher Suchmaschine konnten Sie die Antwort schneller finden?“** auf. Die Frage konnte mit „gleich schnell“, „Antwort nicht gefunden“ oder durch die Auswahl einer der Suchmaschinen beantwortet werden. Abbildung 6.8 fasst das Ergebnis in Diagrammen zusammen.

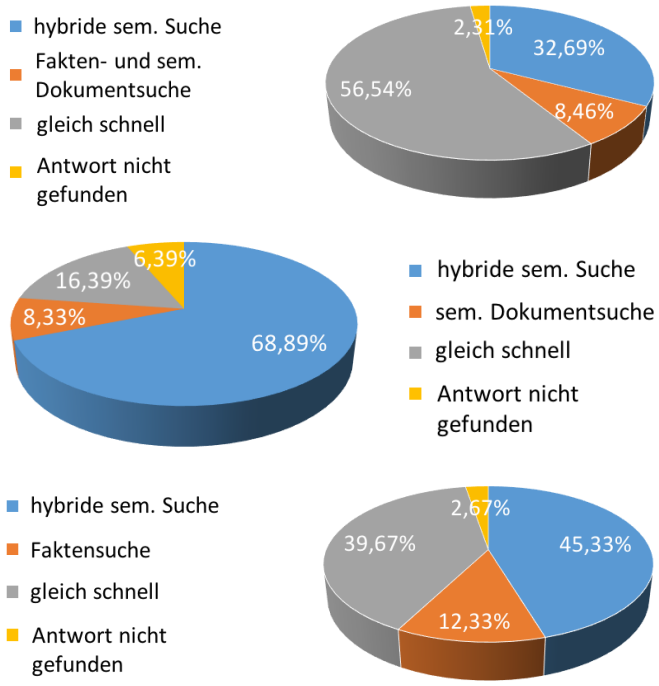


Abbildung 6.8: Überblick der Auswertung der Frage „Bei welcher Suchmaschine konnten Sie die Antwort schneller finden?“

Die Antwort konnte zu 68,89% durch die hybride Suche und zu 8,33% durch die semantische Dokumentsuche schneller gefunden werden. Zu 16,39% war die Antwort gleich schnell, zu 6,39% nicht zu finden.

Hier sei noch angemerkt, dass auch in diesem Fall die drei Suchmaschinenvergleiche untereinander nicht verglichen werden können. Die Werte basieren zwar auf dem Durchschnitt der Antworten, es sind jedoch nicht dieselben Suchanfragen involviert. Dies zeigt sich deutlich an den Zahlen für „Antwort nicht gefunden“: Durch die Suchanfragen q04, q19 und q20 ist der Wert bei dem Vergleich zwischen der hybriden Suche und der semantischen Dokumentsuche mit 6,39% etwa doppelt so hoch wie in den anderen zwei

Vergleichen.

Die **Testhypothese 2 ist aus Benutzersicht bestätigt**: *die Benutzer konnten ihr Informationsbedürfnis durch die hybride semantische Suche schneller befriedigen als durch die semantische Dokumentsuche.* Der Unterschied ist *statistisch signifikant* (p =1,207e^{-9}).

Die hybride semantische Suche wurde auch gegenüber der Fakten- und semantischen Dokumentsuche sowie der Faktensuche positiv bewertet. Hier konnten die Antworten nach Benutzerangaben häufiger gleich schnell gefunden werden (56,54% sowie 39,67%) als im Vergleich zu semantischer Dokumentsuche (16,39%). Die hybride semantische Suche liegt jedoch mit 32,69% zu 8,46% im Vergleich zu der Fakten- und semantischen Dokumentsuche (p =0,0068) und mit 45,33% zu 12,33% im Vergleich zu der Faktensuche (p =0,0042) statistisch signifikant vorne.

Insgesamt bestätigen die Ergebnisse die *Testhypothese 3* (s. Seite 121). Die hybride Suche wurde mit 47,5% gegenüber 7,25% für Faktensuche als die bessere beurteilt, zu 45,25% wurden beide Suchmaschinen als gleich gut bewertet. In 79,67% der Fälle konnte bei der hybriden Suche das Ergebnis aus der Ergebnisliste abgelesen werden. Bei der Faktensuche gelang das nur zu 51%. Das Finden der Antwort wurde bei der hybriden Suche mit 45,33% nahezu viermal so oft als schneller empfunden, als bei der Faktensuche mit 12,33%. Im Durchschnitt fiel die Beurteilung zu 39,67% auf „gleich schnell“ gefunden und zu 2,67% auf „Antwort nicht gefunden“. Diese Ergebnisse deuten darauf hin, dass die Dokumente und die hybriden Ergebnisse aussagekräftiger sind als nur Fakten. Diese Annahme wird von den Begründungen der Benutzer, warum sie welche Suchmaschine besser finden, bestätigt. Die häufigste Antwort ist, dass die hybride Suche eine präzise Antwort und noch weitere Informationen dazu beinhalten, die einem die Antwort liefern und gleich die Möglichkeit, dies zu prüfen oder weitere Informationen dazu einzuholen. Die Ergebnisse werden als aussagekräftiger beurteilt, auch im Falle der Suchanfragen mit hybriden Ergebnissen (q07, q11, q14).

Die **Testhypothese 3 ist aus Benutzersicht bestätigt**: *Die hybride semantische Suche liefert durch die Integration von Dokumenten und hybride Ergebnisse mehr Informationen als die Faktensuche.* Das Informationsbedürfnis kann durch die hybride semantische Suche *statistisch signifikant* schneller befriedigt werden.

Konnten Sie die Antwort in der Ergebnisliste finden, oder mussten Sie weiterführende Links öffnen?												
Such-anfrage	**hybride sem. Suche**		**sem. Dokumentsuche**		**hybride sem. Suche**		**Faktensuche**		**hybride sem. Suche**		**Fakten- und sem. Dokumentsuche**	
	Ergebnis-liste	weiterf. Links	Ergebnis-liste	weiterf. Links	Ergebnis-liste	weiterf. Links	Ergebnis-liste	weiterf. Links	Ergebnis-liste	weiterf. Links	Ergebnis-liste	weiterf. Links
q01	85,00%	15,00%	60,00%	40,00%	95,00%	5,00%	70,00%	30,00%	95,00%	5,00%	75,00%	25,00%
q02	40,00%	60,00%	35,00%	65,00%	30,00%	70,00%	0,00%	100,00%	40,00%	60,00%	0,00%	100,00%
q03	100,00%	0,00%	80,00%	20,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%
q04	70,00%	30,00%	50,00%	50,00%								
q05	95,00%	5,00%	75,00%	25,00%	95,00%	5,00%	95,00%	5,00%	95,00%	5,00%	95,00%	5,00%
q06	90,00%	10,00%	90,00%	10,00%	85,00%	15,00%	20,00%	80,00%				
q07	85,00%	15,00%	75,00%	25,00%	90,00%	10,00%	0,00%	100,00%	90,00%	10,00%	75,00%	25,00%
q08	90,00%	10,00%	5,00%	95,00%	95,00%	5,00%	80,00%	20,00%	95,00%	5,00%	60,00%	40,00%
q09	85,00%	15,00%	80,00%	20,00%	70,00%	30,00%	10,00%	90,00%				
q10												
q11	80,00%	20,00%	20,00%	80,00%	70,00%	30,00%	0,00%	100,00%	75,00%	25,00%	10,00%	90,00%
q12					40,00%	60,00%	0,00%	100,00%	40,00%	60,00%	35,00%	65,00%
q13	80,00%	20,00%	20,00%	80,00%	80,00%	20,00%	30,00%	70,00%	80,00%	20,00%	30,00%	70,00%
q14	65,00%	35,00%	25,00%	75,00%	50,00%	50,00%	70,00%	30,00%	50,00%	50,00%	60,00%	40,00%
q15	90,00%	10,00%	20,00%	80,00%	95,00%	5,00%	95,00%	5,00%	95,00%	5,00%	80,00%	20,00%
q16	100,00%	0,00%	10,00%	90,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	95,00%	5,00%
q17	95,00%	5,00%	15,00%	85,00%	100,00%	0,00%	95,00%	5,00%	100,00%	0,00%	95,00%	5,00%
q18	95,00%	5,00%	0,00%	100,00%								
q19	75,00%	25,00%	0,00%	100,00%								
q20	95,00%	5,00%	5,00%	95,00%								
Durch-schnitt	**84,17%**	15,83%	36,94%	**63,06%**	**79,67%**	20,33%	**51,00%**	49,00%	**81,15%**	18,85%	**62,31%**	37,69%

Tabelle 6.6: Auswertung der Antworten zu der Frage nach der Ablesbarkeit der Antwort aus der Ergebnisliste

Bei welcher Suchmaschine konnten Sie die Antwort schneller finden?												
Such-anfrage	**hybride vs. sem. Dokumentsuche**				**hybride vs. Faktensuche**				**hybride vs. Fakten- und sem. Dokumentsuche**			
	hybride sem. Suche	sem. Dokument-suche	gleich schnell	Antwort nicht gefunden	hybride sem. Suche	Fakten-suche	gleich schnell	Antwort nicht gefunden	hybride sem. Suche	Fakten- und sem. Dokument-suche	gleich schnell	Antwort nicht gefunden
q01	80,00%	15,00%	5,00%	0,00%	30,00%	20,00%	50,00%	0,00%	5,00%	0,00%	95,00%	0,00%
q02	20,00%	10,00%	50,00%	20,00%	100,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%
q03	100,00%	0,00%	0,00%	0,00%	5,00%	35,00%	60,00%	0,00%	5,00%	0,00%	95,00%	0,00%
q04	50,00%	0,00%	10,00%	40,00%								
q05	85,00%	0,00%	15,00%	0,00%	0,00%	40,00%	60,00%	0,00%	5,00%	0,00%	95,00%	0,00%
q06	10,00%	15,00%	75,00%	0,00%	85,00%	0,00%	15,00%	0,00%				
q07	30,00%	30,00%	40,00%	0,00%	100,00%	0,00%	0,00%	0,00%	30,00%	30,00%	40,00%	0,00%
q08	100,00%	0,00%	0,00%	0,00%	30,00%	5,00%	65,00%	0,00%	30,00%	5,00%	70,00%	0,00%
q09	25,00%	15,00%	60,00%	0,00%	65,00%	15,00%	15,00%	5,00%				
q10												
q11	70,00%	20,00%	10,00%	0,00%	95,00%	0,00%	5,00%	0,00%	60,00%	20,00%	20,00%	0,00%
q12					85,00%	0,00%	10,00%	5,00%	60,00%	0,00%	40,00%	5,00%
q13	75,00%	15,00%	10,00%	0,00%	65,00%	0,00%	30,00%	5,00%	70,00%	0,00%	30,00%	0,00%
q14	65,00%	15,00%	5,00%	15,00%	10,00%	50,00%	15,00%	25,00%	15,00%	45,00%	15,00%	25,00%
q15	95,00%	5,00%	0,00%	0,00%	0,00%	20,00%	80,00%	0,00%	20,00%	0,00%	80,00%	0,00%
q16	100,00%	0,00%	0,00%	0,00%	5,00%	0,00%	95,00%	0,00%	20,00%	0,00%	80,00%	0,00%
q17	80,00%	10,00%	10,00%	0,00%	5,00%	0,00%	95,00%	0,00%	5,00%	10,00%	75,00%	0,00%
q18	100,00%	0,00%	0,00%	0,00%								
q19	70,00%	0,00%	0,00%	30,00%								
q20	85,00%	0,00%	5,00%	10,00%								
Durch-schnitt	**68,89%**	8,33%	16,39%	6,39%	**45,33%**	12,33%	39,67%	2,67%	32,69%	8,46%	**56,54%**	2,31%

Tabelle 6.7: Auswertung der Antworten zu der Frage „Bei welcher Suchmaschine konnten Sie die Antwort schneller finden?“

Im Folgenden wird eine gemeinsame **Analyse der Ergebnisse aller drei Fragen** und die Auswertung der insgesamt 167 **Begründungen** zu der Antwort auf die Frage **„Welche Suchergebnisse finden Sie besser?“** durchgeführt. Ziel der Analyse ist zu erfahren, in welchen Fällen und warum die hybride semantische Suche besonders gut oder besonders schlecht abschneidet.

Bei den Suchanfragen q02 und q08 wurde die hybride semantische Suche gegenüber der Fakten- und semantischen Dokumentsuche mit 95% respektive 100% der Stimmen als besser beurteilt. Die Gründe hierfür sind unterschiedlich:

- Für q02, „What ist the longest river?“ bestehen die ersten 10 Ergebnisse der Fakten- und semantische Dokumentsuche aus Flüssen aus DBpedia ohne Zusatzinformation, wie Längenangabe oder sonstiges. Die hybride Suche liefert an erster Stelle das Wikipediadokument „List of rivers by length“.
- Für q08 ist zwar der Fakt, dass die offizielle Sprache der Philippinen die Philippino-Sprache ist, bei beiden Suchmaschinen vorhanden, die weiteren Dokumente sind jedoch nur bei der hybriden Suche relevant (s. Abbildung 6.9). Die Ergebnisse der Fakten- und Dokumentsuche wurden als falsch und nicht aussagekräftig (q02) sowie überwiegend falsch (q08) beurteilt.

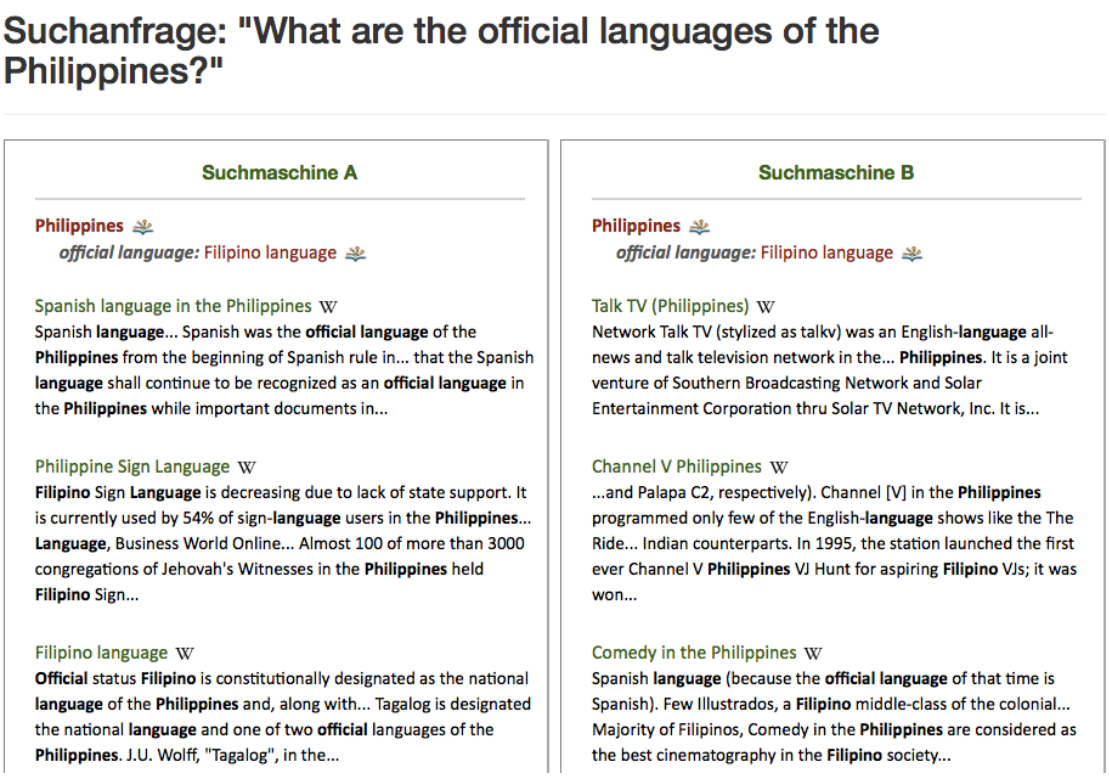


Abbildung 6.9: Side-by-side Panel der Suchanfrage q08 für den Vergleich der hybriden Suche mit Fakten- und semantischer Dokumentsuche

Schlecht abgeschnitten hat die hybride semantische Suche bei den Suchanfragen, die für die Fakten- und semantische Dokumentsuche dieselben Ergebnisse liefern ($q04$, $q06$, $q09$, $q10$, $q18$, $q19$, $q20$) und bei $q03$. Bei allen diesen Suchanfragen wurde sie jedoch genauso gut bewertet wie die Fakten- und semantische Dokumentsuche zusammen. Letzteres wurde lediglich bei $q14$ mit 45% zu 10% besser bewertet. Hier finden ein Teil

der Umfrageteilnehmer die Ergebnisse der Fakten- und Dokumentsuche übersichtlicher.

Für q03, q08, q16 und q18 bewerteten alle Befragten die hybride semantische Suche und in keinem Fall die semantische Dokumentsuche besser. In diesen Fällen konnte die Antwort in der Ergebnisliste der hybriden Suche zu 100% abgelesen werden und ebenfalls von allen Umfrageteilnehmern schneller gefunden werden. Die Gemeinsamkeit dieser Suchanfragen ist, dass die hybride Suche sie präzise, mit Fakten, beantwortet und weiterführende relevante Dokumente dazu liefert. Die semantische Dokumentsuche liefert häufig weniger relevante Dokumente und die Antwort ist nicht so klar und schnell ablesbar.

Mit 40% der Stimmen dafür, dass die hybride Suche besser ist, 40% für die semantische Dokumentsuche und 20% für gleich gut erzielte q07 das schlechteste Ergebnis. In der Ergebnisliste der hybriden Suche konnten 85%, in der Ergebnisliste der semantischen Dokumentsuche 75% der Teilnehmer die Antwort ablesen. Insgesamt 40% der Teilnehmer gaben an, dass das Ergebnis gleich schnell gefunden werden konnte, und jeweils 30% markierten die hybride Suche bzw. die semantische Dokumentsuche. Aus drei der fünf Begründungen geht hervor, dass die hybride Suche als besser empfunden wurde, weil die Ergebnisse besser diversifiziert und eindeutiger sind. Einer der Teilnehmer spricht sich für die Dokumentsuche aus, da daraus schneller hervorgeht, dass es sich beim Austin Powers um einen fiktiven Charakter handelt. Eine gleich gute Bewertung wird damit begründet, dass beide die Suchanfrage beantworten, da diese etwas allgemein formuliert ist.

In 2 Fällen (q07, q12) bewerteten alle Befragten die hybride semantische Suche besser als die Faktensuche. Für beide Suchanfragen mussten bei der Faktensuche zu 100% weiterführende Links geöffnet werden, um die Antwort zu finden. Für q07 (hybride Ergebnisse) konnte die Antwort allen Teilnehmern zufolge durch die hybride Suche schneller gefunden werden. Für q12 gaben dies 85% an, 10% der Personen sagten gleich schnell und 5% der Teilnehmer fanden die Antwort bei keiner der Suchmaschinen. Die Suchergebnisse der hybriden Suche für q07 wurden umfangreicher und aussagekräftiger beurteilt, die Ergebnisse der Faktensuche für q12 als nicht relevant und unbrauchbar. Dies liegt daran, dass hier nur ein Konzept aus DBpedia geliefert wurde, aus dem nicht ersichtlich ist, in welcher Beziehung es zu den „five great epics of Tamil literature“ steht.

Die Faktensuche übertrifft nur für die Suchanfrage q14 die hybride semantische Suche. In diesem Fall konnte die Antwort bei der Faktensuche zu 70%, bei der hybriden Suche nur zu 50% aus der Ergebnisliste abgelesen werden. Dementsprechend fanden 50% der Teilnehmer die Antwort bei der Faktensuche und nur 10% bei der hybriden Suche schneller. Hier wird die Faktensuche als übersichtlicher empfunden.

Die Benutzer präferieren überwiegend die hybride semantische Suche gegenüber der Faktensuche wegen der zusätzlichen Informationen durch Dokumente. *Gegenüber der semantischen Dokumentsuche* wird die hybride Suche bevorzugt, weil die Antwort eindeutig und präzise an oberster Stelle steht und durch die rote Farbe hervorgehoben ist. Zudem werden Dokumente, die durch die hybride Suche gefunden wurden, als relevanter empfunden. *Gegenüber der Fakten- und semantischen Dokumentsuche* wird die hybride Suche bevorzugt, weil die Ergebnisse ausführlicher und besser im Sinne von relevanter, aussagekräftiger empfunden werden. Wenige Begründungen nennen zwei *negative Punkte in Bezug auf die hybride Suche und die Faktensuche* mit nur Fakten in den ersten 10 Ergebnissen (q04, q18, q19 und q20): Die Sortierung sowie die Vollständigkeit der Antworten ist unklar.

6.2.3 Evaluierung der Rankingverfahren

Die **Forschungsfrage 4** beschäftigt sich mit dem Ranking des hybriden Suchverfahrens: *Wie kann die Rankingfunktion für einen hybriden semantischen Suchansatz aussehen? Wie werden Fakten, Dokumente und insbesondere hybride Suchergebnisse bewertet?* Die zwei Rankingverfahren, die mögliche Lösungen auf die Fragen liefern und an dieser Stelle untersucht werden, sind bereits im Kapitel 5.2.4 zusammenfassend und in dem Kapitel 5.2.3 formal beschrieben. Die prototypische Umsetzung behandelt Kapitel 5.3.2.

An dieser Stelle wird untersucht, *welches der beiden Verfahren sich besser für die hybride semantische Suche* auf der englischen DBpedia und Wikipedia *eignet*. Weiterhin wurden die zwei Verfahren bei der Erstellung des Gold Standards berücksichtigt und bereits bei der Evaluierung der Effektivität aus Systemsicht mitbetrachtet. Die Ergebnisse werden ebenfalls in diesem Kapitel diskutiert.

Methode
Um die Qualität von Rankingverfahren bestimmen zu können, wird die Korrelation zum Gold Standard berechnet (s. Kapitel 2.5.2). Die am häufigsten eingesetzte Ranking-Korrelationsmaße sind der Spearman-Koeffizient (s. Seite 43) und der Kendall-Tau-Koeffizient (s. Seite 44). Der Kendall-Tau-Koeffizient lässt sich nur dann einsetzen, wenn die Ergebnismengen dieselben Ergebnisse beinhalten. Da dies hier nicht der Fall ist, wurde für diese Evaluierung der Spearman-Koeffizient über die ersten $k = 10$ Ergebnisse der 20 Suchanfragen berechnet.

Spearman Koeffizient	Hybride semantische Suche	
	Ranking o. Popularität	Ranking m. Popularität
q01	0,9932	0,9932
q02	0,9932	0,9902
q03	0,9890	0,9890
q04	1,0000	1,0000
q05	0,9818	0,9818
q06	0,9500	0,9500
q07	0,9932	0,9951
q08	1,0000	1,0000
q09	0,9992	0,9992
q10	0,9985	0,9985
q11	0,9871	0,9716
q12	0,9947	0,9697
q13	0,9902	0,9902
q14	0,9993	0,9993
q15	0,9867	0,9867
q16	1,0000	1,0000
q17	0,9879	0,9879
q18	1,0000	1,0000
q19	1,0000	1,0000
q20	1,0000	1,0000
Durchschnitt	**0,9922**	0,9901

Tabelle 6.8: Spearman Koeffizienten der beiden Rankingverfahren

Ergebnis
Tabelle 6.8 fasst die Ergebnisse zusammen. Der Wertebereich des Koeffizienten ist $[-1, 1] \in \mathbb{R}$, wobei 1 die perfekte **Korrelation** und -1 die maximale Differenz bedeutet.

Das Ranking ohne Einbezug der Popularität schneidet mit einem Spearman-Koeffizienten von 0,9922 besser ab, als das Ranking mit Popularität, das mit einem Koeffizienten von 0,9901 etwas darunter liegt. **Beide Rankingverfahren weisen mit einem Spearman-Koeffizienten über 0,99 eine sehr hohe Korrelation zum Gold Standard (1,0) auf.** Die Korrelation ist sowohl beim dem Verfahren ohne Popularität (p=0,0042) als auch bei dem Verfahren mit Popularität (p=0,0019) *statistisch signifikant.*

Der Unterschied von 0,0021 zwischen den beiden Rankingverfahren ist statistisch nicht signifikant (p=0,6928), macht sich bei der **Retrievaleffektivität** aber bemerkbar: Die generalisierte Präzision des hybriden Verfahrens mit Ranking ohne Popularität liegt 4,5%, der generalisierte Recall 8% über den Werten des Verfahrens mit Ranking ohne Popularität.

Im Rahmen der Effektivitätsmessung aus Systemsicht erreichte das Ranking ohne Einbezug der Popularität mit einem F-Maß von 0,6485 das bessere Ergebnis. Das Ranking mit Popularität lag mit 0,5437 einen Zehntel darunter. Das auf Popularität basierende Verfahren schneidet bei den Suchanfragen q07, q11 und q12 schlechter, bei allen anderen Suchanfragen genauso gut ab, wie das Ranking ohne Einbezug der Popularität. Die Suchanfragen q07 und q11 liefern hybride Ergebnisse. Der kleine Unterschied von 0,03 im Falle von q07 entsteht durch Veränderung der Rangordnung und der Anordnung der Fakten in den einzelnen hybriden Ergebnissen. Für die Suchanfrage q07 ist der Unterschied des F-Maßes mit 0,32 deutlicher. Hier wurden durch das Einbeziehen der Popularität Konzepte, die mit New York City in Relation stehen, aber nicht die Bezirke sind, höher gewichtet. Die relevanten Konzepte, nämlich die Bezirke, sind nicht mehr alle unter den ersten 10 Ergebnissen. Der Einbezug der Popularität führt bei q12 zu einer deutlichen Verschlechterung: das F-Maß beträgt 0,00. Das relevante Konzept, das die Suchanfrage beantwortet, hat eine niedrige Popularität als Konzepte, die ebenfalls gefunden werden jedoch nicht relevant sind. Hierdurch werden nicht relevante Ergebnisse höher gewichtet und das relevante Konzept ist nicht mehr unter den ersten 10 Ergebnissen.

6.2.4 Evaluierung der Anfragestellung

Die **Forschungsfrage F5** bezieht sich auf die Eingabe der Suchanfrage: *Wie kann der Benutzer bei der Anfragestellung unterstützt werden, so dass er ohne Kenntnis der zugrundeliegenden Wissensbasis Anfragen mit möglichst vielen formalen Anteilen stellt?* Um die Anfragestellung zu unterstützen, wurde eine semantische Autovervollständigungskomponente, wie im Kapitel 5.2.1 vorgestellt, entwickelt. An dieser Stelle wird untersucht, *in welchem Maß die Benutzer die Komponente verwenden.*

Methode
Ziel der Evaluierung ist zu erfahren, ob die semantische Autovervollständigung verwendet wird und dadurch Suchanfragen gebildet werden, die möglichst viele formale Anteile beinhalten. Dies kann **auf Basis von vorgegebenen Suchanfragen und durch Logging der Verwendung der Komponente** untersucht werden. Für die Evaluierung

wurden 20 Benutzern 5 Fragen gegeben, die sie mit Hilfe der hybriden semantischen Suchmaschine SINFIO beantworten sollten. Die Fragen basieren auf Suchanfragen aus der „Test Collection for DBpedia Search“ (s. Kapitel 6.2.2.2), sie drücken reale Informationsbedürfnisse aus:

1. In which year did Steve Jobs found Apple?
2. Through which countries flows the Danube?
3. How many books wrote Stieg Larsson?
4. Which launch pads are operated by NASA?
5. Who were the parents of John Lennon?

Während der Ausführung der Suchanfragen kann mitgeloggt werden, wie oft Vorschläge der semantischen Autovervollständigung übernommen oder ignoriert werden, wie viele Suchanfragen die Teilnehmer ausführen, um eine Frage zu beantworten und wie lange sie zur Beantwortung der Frage brauchen. Für den letzten Punkt wurden sie gebeten einen Button zu drücken, sobald sie die Antwort haben. Durch Analyse der Logdaten kann so die Benutzung der Autovervollständigungskomponente und die Dauer der Aufgabenlösung bestimmt werden.

Die Evaluierung wurde mit 9 Frauen und 11 Männern im Alter zwischen 21 und 35 Jahren durchgeführt. Alle Teilnehmer waren Studenten aus den Fachbereichen Jura, Informatik, Linguistik, Design, Physik, Mathematik, Germanistik und Lehramt. Keiner der Teilnehmer hat weitreichende Kenntnisse über die zugrundeliegende Wissensbasis, wobei alle Teilnehmer Wikipedia kennen und drei der Teilnehmer auch bereits von DBpedia gehört, es aber nicht verwendet, haben. Vor der Benutzung wurden die Teilnehmer in einem persönlichen Gespräch darüber informiert, dass dieser Test zur Evaluierung einer neuartigen Suchmaschine dient, die die englische DBpedia und Wikipedia durchsucht und dementsprechend Ergebnisse aus beiden zurück liefert. Auch ein Beispiel für ein hybrides Ergebnis wurde gezeigt. Danach führten sie die Aufgaben aus und wurden im Anschluss gefragt, wann und warum sie die Autovervollständigung nicht verwendet haben.

Ergebnisse

Abbildung 6.10 fasst pro Aufgabe zusammen, zu welchem Anteil Autovervollständigungsvorschläge durchschnittlich übernommen wurden. Der Anteil liegt zwischen 56 und 95%,

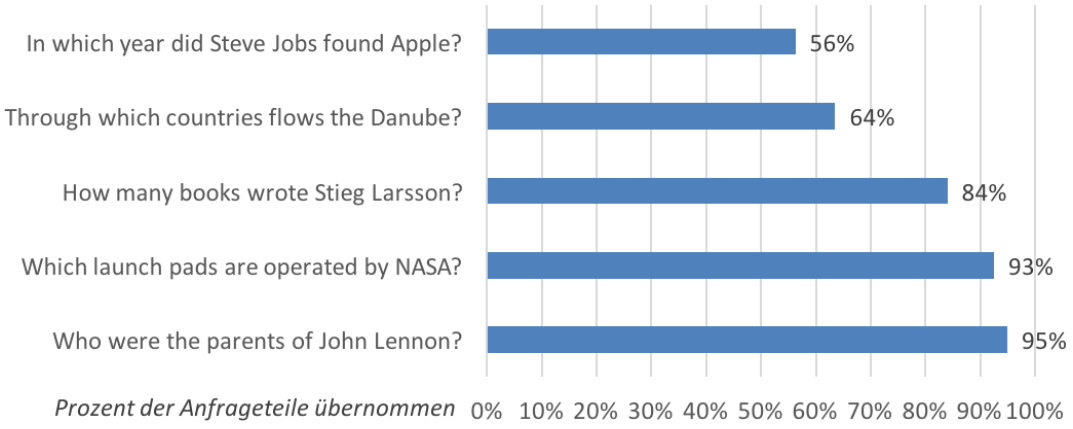


Abbildung 6.10: Durchschnittlicher Anteil der aus den Autovervollständigungsvorschlägen übernommenen Anfrageteile pro Frage

die Komponente wurde überwiegend verwendet. Die Teilnehmer gaben an die Komponente nur dann nicht verwendet zu haben, wenn sie nicht entscheiden konnten, welche der Vorschläge das meint, was sie fragen wollten oder keiner der Vorschläge exakt gepasst hat. Als Beispiel wurde häufig das Wort „found“ genannt. In diesem Fall sind die obersten Elemente der Vorschlagsliste „founded by“, „foundation pace“ und „founding person“. Keiner dieser Vorschläge bezieht sich auf die Zeit, sie könnten aber helfen, die Aufgabe zu lösen. Strategien in diesem Fall waren entweder keinen der Vorschläge anzunehmen oder das entsprechende Wort aus der Suchanfrage rauszulassen und dann mit den erhaltenen Ergebnissen weiter zu arbeiten.

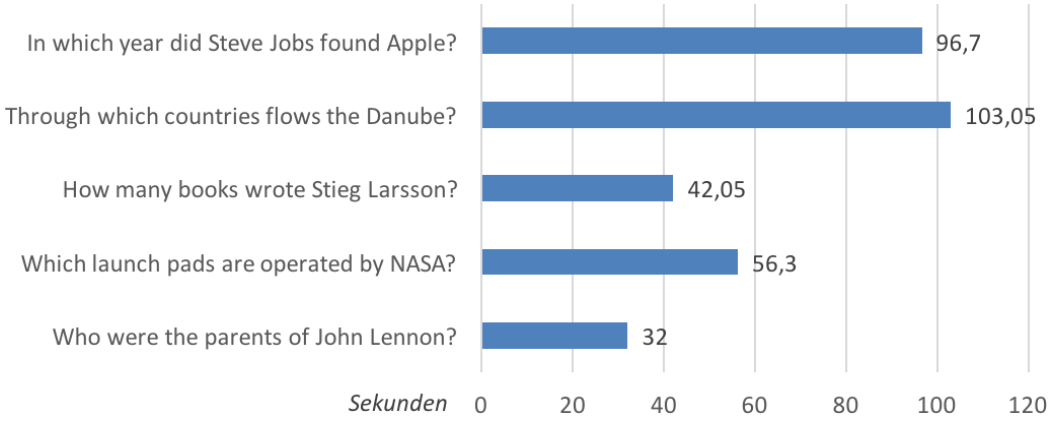


Abbildung 6.11: Durchschnittlicher Zeitaufwand zur Beantwortung der Fragen

Die durchschnittlichen Zeiten zur Beantwortung der fünf Fragen fasst Abbildung 6.11 zusammen. Die Aufgaben 1 und 2 wurden von sechs Teilnehmern durch zwei, durch die anderen vier Teilnehmer durch eine Suchanfrage gelöst. Diese Aufgaben haben mit 96,7 und 103,05 Sekunden die meiste Zeit in Anspruch genommen. In diesen Fällen war die Antwort auch nicht in der Ergebnisliste abzulesen, hierfür mussten Wikipediaseiten geöffnet werden, in denen die Antwort jedoch schnell zu finden war. Für die Fragen 3, 4 und 5, die von allen Teilnehmern mit nur einer Suchanfrage beantwortet wurden, lag die durchschnittliche Antwortzeit zwischen 32,0 und 56,3 Sekunden.

Die Ergebnisse zeigen, dass **die semantische Autovervollständigung von den Benutzern angenommen wird und sie ohne Kenntnisse über die zugrundeliegende Wissensbasis eingesetzt werden kann**. Dies führt dazu, dass **Suchanfragen mit möglichst vielen formalen Teilen gebildet werden**. Zudem zeigt die starke, *statistisch signifikante*, negative lineare Korrelation zwischen der Übernahme der Vorschläge und der Antwortzeit mit einem Koeffizienten von -0,91, dass **diejenigen Fragen, die überwiegend formal formuliert wurden, am schnellsten beantwortet werden konnten**.

6.2.5 Evaluierung der Ergebnisdarstellung

Die **Forschungsfrage F6** richtet sich auf die Ergebnisdarstellung: *Wie können die Suchergebnisse, bestehend aus Fakten, Dokumenten und hybriden Ergebnissen, so dargestellt werden, dass diese und auch die Ergebnisliste verständlich sind?* Die Konzepte für die

Benutzerschnittstelle, Darstellung der Ergebnistypen und der Ergebnisliste sind im Kapitel 5.2.5, ihre prototypische Realisierung im Kapitel 5.3.3 vorgestellt. An dieser Stelle wird untersucht, *ob die Benutzer die Darstellung der Ergebnisse verständlich finden* und *ob die Ergebnislistendarstellung nach Rangordnung oder die gruppierte Darstellung bevorzugt wird.*

Methode

Die **Verständlichkeit der Ergebnisdarstellung** kann durch **Benutzerbefragung** untersucht werden. Um die Darstellung der hybriden Suche zu den anderen Suchmaschinen in Relation setzen zu können, wurden auch für diese Evaluierung **Side-by-side Panels** verwendet (s. Abbildung 6.5). Ebenso wie bei der Befragung zur Retrievaleffektivität, deckte die Umfrage alle 20 Suchanfragen ab. Verglichen wurde die hybride semantische Suche (Ranking ohne Popularität) mit der Faktensuche, der semantischen Dokumentsuche sowie der Fakten- und semantischen Dokumentsuche. Vergleiche mit denselben Ergebnislisten wurden nicht mit einbezogen. Abbildung 6.12 zeigt die Frage und Antwortmöglichkeiten. Auch an dieser Stelle hatten die Benutzer die Möglichkeit, eine Begründung anzugeben.

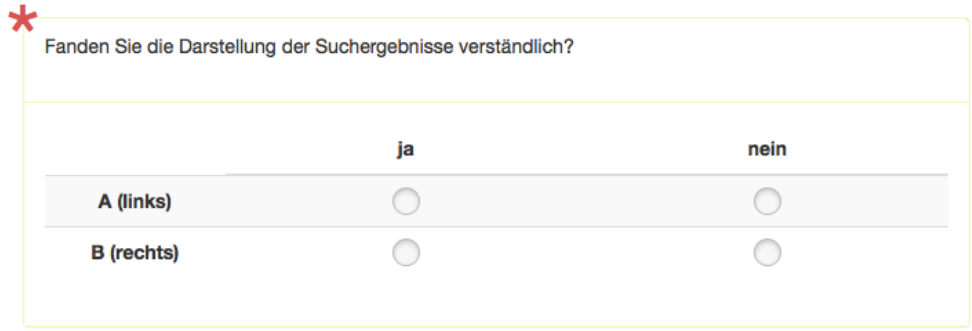


Abbildung 6.12: Frage zur Verständlichkeit der Ergebnisdarstellung

Die Evaluierung wurde als Teil der Umfrage zur Effektivität durchgeführt, für die demografischen Daten der 20 Umfrageteilnehmer finden sich in Kapitel 6.2.2.3.

Der **Vergleich der beiden Darstellungsmodi** wurde ebenfalls mittels **Side-by-side Panels** für die 20 Suchanfragen durchgeführt, Abbildung 6.13 zeigt ein Beispiel. In dieser **Benutzerbefragung** wurden die ersten 80 Suchergebnisse involviert, da in diesem Fall die Länge der einzelnen Abschnitte in der gruppierten Sicht ein wichtiger Faktor für die Meinungsbildung sein kann. Mehr Suchergebnisse konnten im LimeSurvey aus technischen Gründen nicht umgesetzt werden. Die Benutzer wurden nach ihrer Präferenz gefragt, die Frage zeigt Abbildung 6.14. Eine Begründung konnte ebenfalls angegeben werden.

An der Umfrage zum Vergleich der beiden Darstellungsarten haben ebenfalls 20 Personen teilgenommen. Die 12 weiblichen und 8 männlichen Befragten hatten die Berufe Künstler, Handwerker, PR-Berater, technischer Manager, Pädagoge, Bankkauffrau, Sekretärin, Ökonom, Journalist, Jurastudent, Kartograph, Computerlinguist, Controllerin. Einer der Teilnehmer hat als Beruf Pensionär angegeben. Mit 40% überwiegt der Anteil der 31-40 jährigen Teilnehmer, zwischen 20 und 30 Jahre alt waren 15%, zwischen 41 und 50 30% und über 50 genau 15% der Personen.

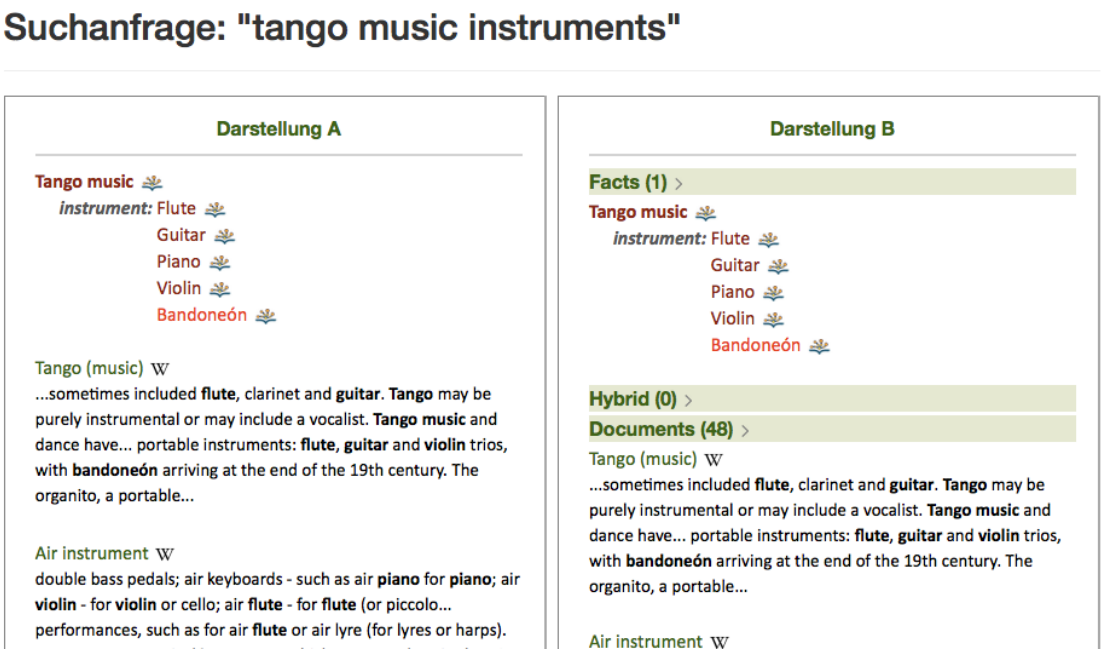


Abbildung 6.13: Frage zum Vergleich der zwei Ergebnislistendarstellungsarten

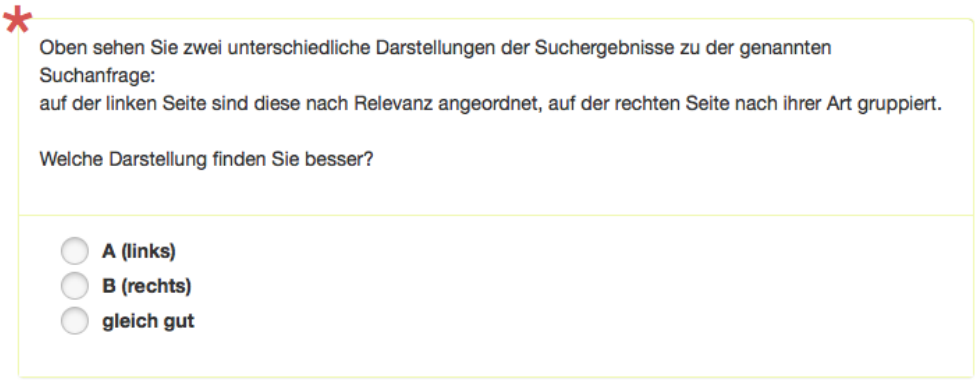


Abbildung 6.14: Frage zum Vergleich der zwei Ergebnislistendarstellungsarten

Ergebnisse

Tabelle 6.9 listet die Ergebnisse der Befragung **„Fanden Sie die Darstellung der Suchergebnisse verständlich?“** für jede Suchanfrage und jeden Vergleich auf, Abbildung 6.15 zeigt einen Überblick als Balkendiagramm.

Die *Ergebnisdarstellung der hybriden semantische Suche* wurde im Durchschnitt über alle drei Vergleiche zu 91,38% als verständlich bewertet. Sie wurde sowohl im Vergleich zu der Fakten- und semantischen Dokumentsuche mit 89,44% zu 62,5% als auch im Vergleich zu der Faktensuche mit 92% zu 69,33% sowie zu der semantischen Dokumentsuche mit 92,69% zu 73,46% besser bewertet. Alle Unterschiede sind statistisch signifikant, die p-Werte betragen 0,0014, 0,0068 sowie 0,00002. Am wenigsten verständlich fanden die Benutzer die semantische Dokumentsuche, gefolgt von der Faktensuche. Die 22 Begründungen sind vielfältig und beziehen zum Teil auch die Qualität der Suchergebnisse mit ein. Insgesamt wird die auffällige rote Schrift für präzise Antworten (Fakten) als positiv genannt, die „Antwort“ fällt damit ins Auge. Fakten und hybride Ergebnisse sind

übersichtlich dargestellt. Die Dokumentdarstellung wurde nicht kommentiert, Titel und Textsnippets folgen jedoch der üblichen Darstellungsart der meist verwendeten Websuchmaschinen.

In mindestens einem der drei Vergleiche fanden alle Umfrageteilnehmer die Ergebnisdarstellung der hybriden semantischen Suche verständlich, wenn diese die Antwort als Fakten enthält, dazu jedoch relevante Dokumente liefert (q03, q05, q06, q08, q15, q16, q17). In einem Fall wurde zudem eine Suchanfrage, die in den ersten 10 Ergebnissen nur Fakten liefert (q18), zu 100% als verständlich bewertet.

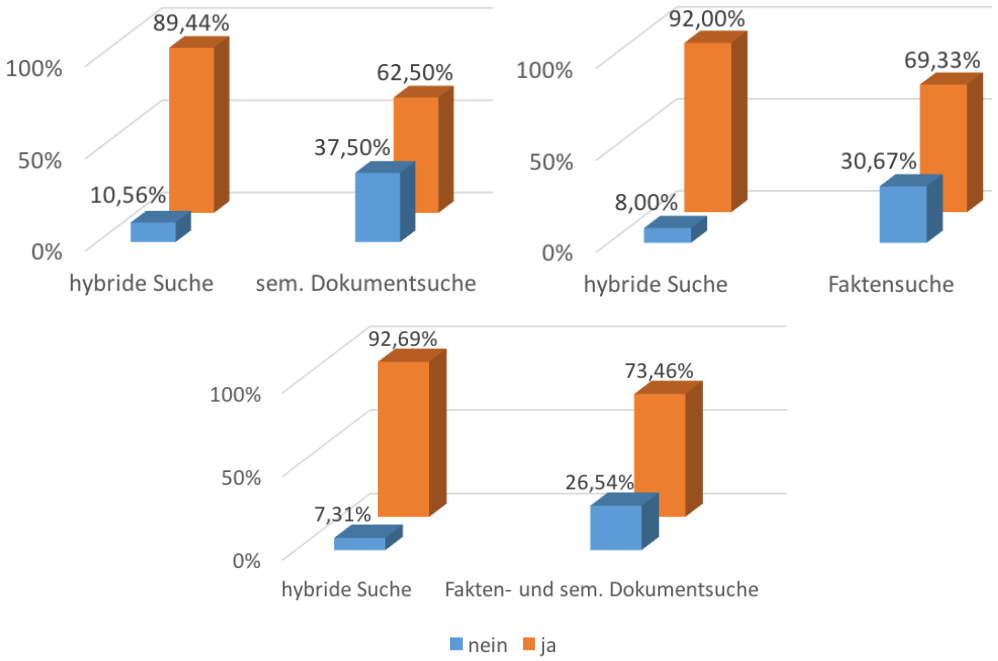


Abbildung 6.15: Überblick der Auswertung der Frage „Fanden Sie die Darstellung der Suchergebnisse verständlich?“

Die Suchanfragen mit *hybriden Ergebnissen* (q07, q11 und q14) wurden zu 91,7%, 83,33% bzw. 66,67% als verständlich empfunden. Sie liegen in zwei Fällen unter dem Durchschnitt. Die Suchanfrage q14, „Give me all launch pads operated by NASA.“, wurde auch insgesamt am schlechtesten bewertet. Die Kommentare deuten auf zwei Ursachen hin:

- gegenüber der Faktensuche wird die Ergebnisdarstellung als unübersichtlich empfunden,
- die Vollständigkeit der Antwort ist nicht ersichtlich.

Letzteres ist nur im weiteren Sinne ein Darstellungsproblem: Durch die Struktur der Wissensbasis (keine inverse Properties vorhanden) und der Suchlogik sind die gefundenen Fakten nicht als ein Ergebnis zusammengefasst. Dies ist auch bei vielen anderen Suchanfragen (q04, q07, q13, q18, q19 und q20) der Fall und wurde nicht negativ angemerkt oder bewertet. Die Begründungen insgesamt zeigen jedoch, dass die Benutzer sich bei den „Give me all...“ Suchanfragen über die Vollständigkeit der Ergebnisse Gedanken machen. Da in den Umfragen jedoch nur die ersten 10 Ergebnisse dargestellt sind, kann dies natürlich nicht geprüft werden.

Fanden sie die Darstellung der Suchergebnisse verständlich?												
Such-anfrage	**hybride sem. Suche**		**sem. Dokumentsuche**		**hybride sem. Suche**		**Faktensuche**		**hybride sem. Suche**		**Fakten- und sem. Dokumentsuche**	
	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein
q01	85,00%	15,00%	75,00%	25,00%	95,00%	5,00%	90,00%	10,00%	95,00%	5,00%	80,00%	20,00%
q02	75,00%	25,00%	80,00%	20,00%	95,00%	5,00%	20,00%	80,00%	95,00%	5,00%	20,00%	80,00%
q03	100,00%	0,00%	55,00%	45,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%
q04	70,00%	30,00%	50,00%	50,00%								
q05	95,00%	5,00%	80,00%	20,00%	100,00%	0,00%	95,00%	5,00%	100,00%	0,00%	90,00%	10,00%
q06	100,00%	0,00%	100,00%	0,00%	95,00%	5,00%	50,00%	50,00%				
q07	90,00%	10,00%	90,00%	10,00%	95,00%	5,00%	20,00%	80,00%	95,00%	5,00%	80,00%	20,00%
q08	100,00%	0,00%	30,00%	70,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	75,00%	25,00%
q09	95,00%	5,00%	95,00%	5,00%	95,00%	5,00%	65,00%	35,00%				
q10												
q11	80,00%	20,00%	65,00%	35,00%	85,00%	15,00%	40,00%	60,00%	85,00%	15,00%	60,00%	40,00%
q12					95,00%	5,00%	35,00%	65,00%	95,00%	5,00%	80,00%	20,00%
q13	70,00%	30,00%	35,00%	65,00%	75,00%	25,00%	55,00%	45,00%	75,00%	25,00%	50,00%	50,00%
q14	75,00%	25,00%	65,00%	35,00%	55,00%	45,00%	75,00%	25,00%	70,00%	30,00%	70,00%	30,00%
q15	100,00%	0,00%	50,00%	50,00%	95,00%	5,00%	95,00%	5,00%	95,00%	5,00%	75,00%	25,00%
q16	100,00%	0,00%	55,00%	45,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	90,00%	10,00%
q17	95,00%	5,00%	65,00%	35,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	85,00%	15,00%
q18	100,00%	0,00%	55,00%	45,00%								
q19	85,00%	15,00%	30,00%	70,00%								
q20	95,00%	5,00%	50,00%	50,00%								
Durch-schnitt	**89,44%**	10,56%	**62,50%**	37,50%	**76,67%**	6,67%	**57,78%**	25,56%	**66,94%**	5,28%	**53,06%**	19,17%

Tabelle 6.9: Auswertung der Antworten zu der Frage „Fanden Sie die Darstellung der Suchergebnisse verständlich?“

Hier sei noch angemerkt, dass die Verständlichkeit der Ergebnisdarstellung implizit auch bei der Bewertung der Qualität der Suchmaschinen inbegriffen ist. Bei allen drei Fragen - welches der Suchsysteme als besser empfunden wurde, die Ablesbarkeit der Antwort aus der Ergebnisliste sowie ob und in welchem System die Antwort schneller gefunden werden konnte (Ergebnisse s. Kapitel 6.2.2.3) - spielt es eine wichtige Rolle, ob und wie schnell die Benutzer die Ergebnisse interpretieren konnten.

Die Ergebnisse der Umfrage zeigen, dass **die Benutzer die Ergebnisdarstellung der hybriden semantischen Suche (zu 91,38%) verständlich finden**. Die Unterschiede zu der Fakten- und Dokumentsuche, der Faktensuche sowie der semantischen Dokumentsuche sind *statistisch signifikant.*

Suchanfrage	Welche Darstellung finden Sie besser?		
	nach Relevanz geordnet	nach Art gruppiert	gleich gut
q01	70,00%	20,00%	10,00%
q02	65,00%	15,00%	20,00%
q03	75,00%	10,00%	15,00%
q04	55,00%	30,00%	15,00%
q05	70,00%	5,00%	25,00%
q06	75,00%	10,00%	15,00%
q07	75,00%	20,00%	5,00%
q08	70,00%	10,00%	20,00%
q09	75,00%	15,00%	10,00%
q10	75,00%	10,00%	15,00%
q11	80,00%	15,00%	5,00%
q12	85,00%	5,00%	10,00%
q13	65,00%	10,00%	25,00%
q14	55,00%	25,00%	20,00%
q15	70,00%	20,00%	10,00%
q16	65,00%	15,00%	20,00%
q17	70,00%	15,00%	15,00%
q18	60,00%	10,00%	30,00%
q19	50,00%	35,00%	15,00%
q20	45,00%	30,00%	25,00%
Durchschnitt	**67,50%**	16,25%	16,25%

Tabelle 6.10: Ergebnisse der Umfrage zum Vergleich der Ergebnisdarstellungsarten

Tabelle 6.10 zeigt die Ergebnisse der Umfrage **„Welche Darstellung finden Sie besser?“**. Die Umfrageteilnehmer bevorzugten die Ergebnislistendarstellung nach Relevanz geordnet, mit 67,5% wurde diese als besser befunden. In 16,25% der Fälle wurde die gruppierte Darstellung besser und ebenfalls in 16,25% die zwei Darstellungsmodi als gleich gut bewertet. Der Unterschied ist statistisch signifikant ($p = 0$). Die gruppierte Darstellung für keine der Suchanfragen besser gefunden wurde als die nach Relevanz sortierte, nicht gruppierte Ergebnisliste. Betrachtet man die Präferenzen per Suchanfrage, so wurden diese für q20 am gleichmäßigsten auf die drei Antwortmöglichkeiten verteilt.

Die ersten 10 Ergebnisse der Suchanfrage sind Fakten, wodurch der Unterschied zwischen den beiden Darstellungsarten am geringsten ausfällt, wie der Screenshot in Abbildung 6.16 zeigt.

Abbildung 6.16: Side-by-side Panel zur Ergebnisdarstellung der Suchanfrage q20

Auf die Frage „Warum?“ gab es 22 Antworten. Für die nach Rangordnung sortierte Listendarstellung spricht, dass hier keine Boxen aufgeklappt werden müssen und keine Kategorien extra interpretiert werden müssen. Eine Unterauswahl fanden die Benutzer, die eine Begründung angegeben haben, unnötig. Die gruppierte Darstellung wurde der Auswertung von Kommentaren und Ergebnissen zufolge dann bevorzugt, wenn die Ergebnismenge aus Fakten besteht, die auch einzeln bereits lang sind (q04, q19, q20).

Die Benutzer bevorzugen diejenige Ergebnisdarstellung, die die Ergebnisse nach ihrer Relevanz geordnet listet und nicht nach Ergebnistyp gruppiert (zu 67,5%). Der Unterschied zu der gruppierten Darstellung ist *statistisch signifikant*.

6.2.6 Effizienz

Die Laufzeit der hybriden semantischen Suche SINFIO wurde auf einem MacBook Air mit 2 GHz Intel Core i7 CPU und 8GB RAM gemessen, wobei der Zugriff auf den semantischen Autovervollständigungs- und den Volltextindex sowie auf den Jena-Store (s. Kapitel 5.3) über ein 1 Gigabit Intranet geschah. Die Solr-Instanzen mit den Indizes und der Jena-Store liefen in einer virtuellen Maschine auf einem PC mit Pentium Dual-Core 2,50 GHz CPU, die virtuelle Maschine hatte 4GB RAM.

Das Balkendiagramm in Abbildung 6.17 zeigt die **durchschnittliche Laufzeit über 10 Messungen je Suchanfrage**. Die gesamte Laufzeit per Suchanfrage wird grün dargestellt. Um den Einfluss der Datenzugriffszeiten, also die Laufzeiten der Solr-Instanzen, der Jena-Store-Zugriffe inklusive der Transportzeit über das Intranet bestimmen zu können, wurden diese bei jedem Durchlauf gesondert mitgemessen. Die blauen Balken zeigen die Laufzeiten ohne diese Zugriffszeiten.

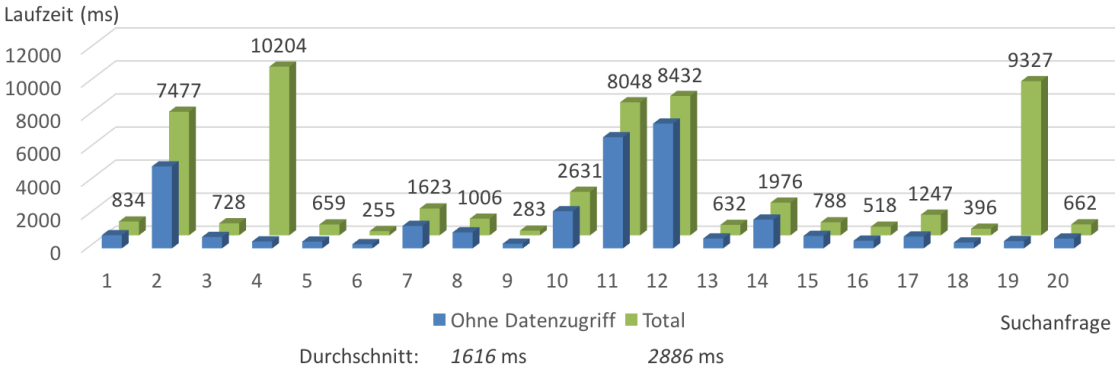


Abbildung 6.17: Diagramm der durchschnittlichen Effizienz je Suchanfrage, mit und ohne Datenzugriffszeiten

Die durchschnittliche gesamte Laufzeit ab Absenden der Suchanfrage bis vollständigen Generierung der Ergebnisseite mit 2.886 ms ist für eine Suchmaschine hoch. Ohne Datenzugriff reduziert sich der Durchschnitt auf 1616 ms (56%). Am langsamsten werden die Suchanfragen q02, q04, q11, q12 und q19 verarbeitet. Q04 und q19 liefern eine große Menge an Fakten, alleine der Zugriff auf DBpedia macht 96% der Laufzeit aus. Bei den Suchanfragen q02, q11 und q12 ist der Unterschied deutlich kleiner, hier macht der Zugriff auf die Datenquellen 44%, 17% und 11% der Gesamtlaufzeit aus. In diesen Fällen nimmt das Spreading Activation viel Zeit in Anspruch, da sehr viele Knoten, die jedoch nicht so nah beieinander liegen, dass sie im Rahmen der Aktivierung verbunden werden, in den Prozess als Aktivierungsknoten eintreten. Dadurch erhöht sich die Anzahl involvierter Kanten und Knoten.

Die Laufzeit kann verbessert werden, indem der Jena-Store und die Solr-Instanzen auf einem leistungsstarken Server gelegt werden und der Prozess des Spreading Activation für die Fälle parallelisiert wird, in denen nicht zusammenhängende Teilgraphen geflutet werden. In Abhängigkeit von den initialen Aktivierungsknotengewichten, Kantengewichten, dem Auskühlungsfaktor α und der Stoppbedingung kann berechnet werden, wieweit die Aktivierung um einen Aktivierungsknoten herum sich maximal ausbreiten kann. Gibt es keine Überschneidungen unter diesen Teilgraphen, so lässt sich das Spreading Activation parallelisieren (s. Kapitel 5.2.3.2 und 5.2.3.3). Eine weitere Möglichkeit besteht darin, die entfernten Tripel während des Aktivierungsprozesses inkrementell zu importieren, wie in [Marie, 2014, Marie et al., 2013] beschrieben. Die Vorgehensweise bietet neben der Steigerung der Effizienz den Vorteil, dass an eine Domäne anpassbare Routinen zur Selektion der zu verfolgenden Knoten im Rahmen des Aktivierungsprozesses (anstatt zum Zeitpunkt des Netzaufbaus) umgesetzt werden können. So können diese Routinen die zum Betrachtungszeitpunkt erreichten Gewichte berücksichtigen und zu weiterem Effizienzgewinn führen.

Kapitel 7
Diskussion der Ergebnisse

Im Kapitel 6 wurden die Methoden und Ergebnisse der einzelnen Evaluierungen vorgestellt. Kapitel 7 fasst die für die These und Testhypothesen relevanten Ergebnisse der Evaluierung aus Systemsicht und Benutzersicht zusammen und untersucht die statistischen Zusammenhänge unter ihnen (Kapitel 7.1). Des Weiteren werden die Forschungsfragen beantwortet und die Antworten mit den entsprechenden Evaluierungsergebnissen gemeinsam betrachtet (Kapitel 7.2). Anschließend wird das Gesamtergebnis zusammengefasst (Kapitel 7.3).

Die hier vorgestellten Korrelationsanalysen wurden mit den im Kapitel 6 beschriebenen Methoden durchgeführt.

7.1 Diskussion der These und Testhypothesen

Die **These** besagt: **Für unterschiedlich stark strukturierte Datenmengen ist ein integrierter Ansatz zur hybriden semantischen Suche, der formale und informale Inhalte während des gesamten Suchprozesses kombiniert, effektiver als eine semantische Dokument- und Faktensuche ohne Kombination der Inhalte.**

Die in dieser Arbeit vorgestellte hybride semantische Suchmaschine SINFIO bietet eine Lösung für das im Kapitel 4.2 definierte hybride semantische Suchproblem. Sie kombiniert formale und informale Inhalte im gesamten Suchprozess, von der Suchanfrage über den Suchalgorithmus bis hin zu den Suchergebnissen. SINFIO erreicht sowohl aus Systemsicht als auch aus Benutzersicht eine höhere **Retrievaleffektivität** als die Fakten- und semantische Dokumentsuche zusammen, jedoch ohne eine Kombination der formalen und informalen Inhalte. Die Verbesserung aus Systemsicht, gemessen anhand des F-Maßes (basierend auf der generalisierten Genauigkeit und Vollständigkeit), ist nicht statistisch signifikant. Mit 10%iger Wahrscheinlichkeit kann die Differenz auch dann entstehen, wenn beide Suchmaschinen gleich gut sind. Das Ergebnis der Benutzerbefragung ist jedoch statistisch signifikant. Die These sowie die Testhypothese 1 sind bestätigt.

Durch die Benutzerstudie wurden auch die Testhypothesen 2 und 3 bestätigt, beide Ergebnisse sind statistisch signifikant:

- Die Umfrageteilnehmer konnten ihr Informationsbedürfnis durch die hybride semantische Suche schneller befriedigen als durch die semantische Dokumentsuche.
- Die hybride semantische Suche liefert durch die Integration von Dokumenten und hybriden Ergebnissen mehr Informationen als die Faktensuche. Das Informationsbedürfnis kann schneller befriedigt werden.

Sowohl bei den Systemkennzahlen als auch bei allen Fragestellungen der Benutzerstudie erreichte die hybride semantische Suche bessere Ergebnisse, als die Faktensuche, die semantische Dokumentsuche und die Fakten- und semantische Dokumentsuche zusammen. Die ersten 10 Ergebnisse von SINFIO erreichten ein höheres F-Maß und wurden von den Benutzern als besser bewertet. Die Antwort konnte am häufigsten aus der Er-

gebnisliste abgelesen werden und wurde mehrheitlich schneller gefunden. Alle Ergebnisse der Benutzerbefragung waren statistisch signifikant.

Die qualitativen Ergebnisse der Benutzerstudie geben eine Auskunft darüber, wann und warum eine Suchmaschine besser bewertet wurde:

- Die Benutzer präferieren die hybride semantische Suche gegenüber der Faktensuche wegen der zusätzlichen Informationen, die durch Dokumente in die Ergebnisliste einfließen.
- Gegenüber der semantischen Dokumentsuche wird die hybride semantische Suche bevorzugt, weil die Antwort häufig eindeutig sowie präzise an oberster Stelle steht und durch die rote Farbe hervorgehoben ist. Die durch die hybride Suche gefundenen Dokumente werden zudem als relevanter empfunden.
- Gegenüber der Fakten- und semantischen Dokumentsuche wird die hybride Suche bevorzugt, weil die Ergebnisse als ausführlicher, aussagekräftiger und besser im Sinne der Relevanz, empfunden werden.

Im Bezug auf die Ergebnisse der hybriden Suche werden in wenigen Begründungen zwei negative Punkte genannt: Wenn die (ersten 10) Ergebnisse nur Fakten sind, ist sowohl die Sortierung als auch die Vollständigkeit der Antwort unklar. Dies kommt bei Suchanfragen vor, deren Ergebnis eine Auflistung von Konzepten ist, wie z.B. die Antworten auf „Give me all...“-Suchanfragen. Die unklare Sortierung ist dadurch verursacht, dass die Fakten mit dem selben Gewicht gefunden und deshalb in der Reihenfolge dargestellt werden, in der sie bei der Abfrage an den RDF-Store geliefert werden. Die syntaktische Ähnlichkeit eignet sich in diesen Fällen nicht zur Differenzierung, denn alle Fakten basieren auf demselben, durch den syntaktischen Abgleich gefundenem Konzept, sowie einer spezifischen Property oder Propertykette und haben daher das gleiche Gewicht. Die Vollständigkeit ist nicht prüfbar, da nur die ersten 10 Ergebnisse in die Evaluierung eingeflossen sind.

Unter Betrachtung aller Evaluierungsergebnisse stellt sich die Frage, **welcher der erfragten Faktoren den größten Einfluss darauf hat, dass eine Suchmaschine als besser beurteilt wird**. Für eine umfassende Untersuchung wurde neben der Ablesbarkeit der Antwort und ihr schnelleres Finden auch die Frage „Fanden Sie die Darstellung der Suchergebnisse verständlich?“ mit einbezogen. Die Korrelationsanalyse zwischen den positiven Ergebnissen der Frage „Welche Suchergebnisse finden Sie besser?“ und den positiven Ergebnissen der oben genannten Fragen für den Vergleich der hybriden Suche mit der Fakten- und Dokumentsuche ergab folgende Zusammenhänge:

- eine mittlere lineare Korrelation (mit einem Koeffizienten von 0,54) zu der Frage, bei welcher Suchmaschine die Antwort schneller gefunden werden konnte;
- keine Korrelation (-0,08) zu der Frage, ob die Antwort in der Ergebnisliste abgelesen werden konnte;
- eine schwache Korrelation (0,23) zu der Frage, ob die Ergebnisdarstellung verständlich war.

Korrelationskoeffizienten geben nur einen statistischen, aber keinen kausalen Zusammenhang wieder. Deshalb lassen sich die Werte nur unter der Annahme interpretieren, dass die untersuchten Fragen bei einem Vergleich der Suchergebnisse zweier Suchmaschinen und der Beurteilung dessen, welche besser sind, eine Rolle spielen. Unter dieser Annahme deuten die Ergebnisse darauf hin, dass die Geschwindigkeit, in der die Benutzer eine Antwort auf ihre Suchanfrage finden, unter diesen drei Faktoren am wichtigsten ist.

Einen schwachen Einfluss auf die Beurteilung der Güte der Suchmaschine hat die Verständlichkeit der Ergebnisdarstellung. Ob die Suchergebnisse einer Suchmaschine besser empfunden werden, hängt nicht von der Ablesbarkeit der Antwort in der Ergebnisliste ab. Es gibt weder einen linearen noch einen anderen Zusammenhang, auch der Spearman-Koeffizient liegt nah bei 0. Dies kann bedeuten, dass die Benutzer sich bei der Beurteilung der ersten 10 Suchergebnisse nicht auf das Finden der Antwort konzentrieren, sondern die Ergebnisse betrachten und deren Güte insgesamt beurteilen.

Eine nennenswerte lineare oder nichtlineare **Korrelation zwischen den Ergebnissen der Benutzerbefragung und der Systemkennzahl F-Maß** konnte in keinem Fall nachgewiesen werden. Dies zeigt den Unterschied zwischen der objektiven und der subjektiven Beurteilung der Suchmaschine: Die Beurteilung der Benutzer schließt mehr Faktoren als nur die (subjektiv empfundene) Relevanz der Suchergebnisse mit ein.

7.2 Antworten auf die Forschungsfragen

F1 - Wie können ein Verfahren zur Faktensuche und ein Verfahren zur semantischen Dokumentsuche so kombiniert werden, dass die Kombination in der Lage ist, Fakten und Dokumente während des gesamten Suchprozesses abhängig voneinander zu durchsuchen?

Kapitel 5.2.3 beschreibt den Lösungsansatz für solch eine hybride semantische Suche mit einem tripelbasierten Verfahren zur Faktensuche und dem Graphtraversierungsverfahren Spreading Activation für semantisches Dokumentretrieval sowie zur Kombination der Fakten- und der Dokumentsuche. SINFIO ist in der Lage formale, informale und hybride Suchanfragen zu verarbeiten, formale und informale Inhalte im Suchprozess einzubinden und die Suchanfragen mit Fakten, Dokumenten und hybriden Ergebnissen zu beantworten. SINFIO erreichte sowohl aus System- als auch aus Benutzersicht eine höhere Retrievaleffektivität als die Fakten- und Dokumentsuche zusammen, jedoch ohne eine Kombination der Inhalte im Suchprozess.

F2 - Welche Verfahren zur Faktensuche und zur semantischen Dokumentsuche eignen sich für solch eine Kombination und inwiefern müssen diese dazu adaptiert werden?

Die Analyse der Suchverfahren zur Fakten- und semantische Dokumentsuche hinsichtlich den Anforderungen ergab drei Kombinationsmöglichkeiten (vgl. Kapitel 5.2.2):

a) Tripelbasiertes Verfahren für Faktensuche und ein Graphtraversierungsalgorithmus für semantisches Dokumentretrieval und für die hybride semantische Suche;
b) Logisches Schließen für Faktensuche und ein Graphtraversierungsalgorithmus für semantisches Dokumentretrieval und für die hybride semantische Suche;
c) Logisches Schließen für Faktensuche und traditionelle Schlüsselwortsuche in Dokumenten sowie logisches Schließen auf der Wissensbasis.

Da Logisches Schließen auf großen Datenmengen ineffizient ist (vgl. Kapitel 5.2.2) wurde die Lösung *a*) umgesetzt. Für die Faktensuche in SINFIO wurde der tripelbasierte

Ansatz aus [Goldschmidt and Krishnamoorthy, 2005] adaptiert. Die Anpassungen beinhalten die Erweiterung des Ansatzes zu einem einem iterativen Verfahren, das auch aus mehr als zwei Tripeln bestehende Fakten finden kann (s. Kapitel 5.2.3.1). Weiterhin wird die Reihenfolge der Terme bei der Bildung der Abfragen berücksichtigt. Durch diese Änderungen spielt die Nähe der Suchterme zueinander eine Rolle. Falls benachbarte Terme jedoch keine Ergebnisse liefern, so betrachtet das iterative Verfahren die nächsten Suchterme, wodurch auch Zusammenhänge zwischen nicht benachbarten Termen gefunden werden. Für die Graphtraversierung wurde der Spreading Activation-Algorithmus ausgewählt, da er sowohl für semantisches Dokumentretrieval als auch für eine Kombination der Ergebnisse der Fakten- und Dokumentsuche geeignet ist, sowie ein adäquates Ranking erlaubt. Der Spreading Activation-Algorithmus selbst wurde nicht angepasst. Hier spielt der Aufbau des semantischen Netzes sowie die Auswahl und Parametrierung der Constraints die wesentliche Rolle (vgl. Kapitel 5.2.3.2 und 5.2.3.3).

F3 - Können die Verfahren zur semantischen Dokumentsuche und zur Faktensuche so kombiniert werden, dass sie auch hybride Suchergebnisse finden oder sollte hierfür ein weiteres Verfahren eingesetzt werden?

Die ausgewählten Verfahren sind in ihrer Kombination geeignet hybride Ergebnisse zu finden ohne hierfür ein weiteres Verfahren einzusetzen. Der Spreading Activation-Prozess agiert auf dem semantischen Netz, das sowohl Konzepte der Wissensbasis als auch Dokumente (instanziiert) enthält. Hierdurch werden im selben Prozess Fakten und Dokumente sowie ihre Verknüpfungen untereinander berücksichtigt. Die Ergebnisse sind Teilgraphen aus dem semantischen Netz, bestehend aus Knoten, die ein vordefiniertes Gewicht erreichen und den dazwischen liegenden Kanten (vgl. Kapitel 5.2.3.3).

F4 - Wie kann die Rankingfunktion für einen hybriden semantischen Suchansatz aussehen? Wie werden Fakten, Dokumente und insbesondere hybride Suchergebnisse bewertet?

Eine Herausforderung für hybride semantische Suchansätze besteht darin, ein geeignetes Rankingverfahren zu finden. Die Gründe hierfür sind im Kapitel 5.2.4 dargelegt und werden hier nochmal kurz zusammengefasst:

- die meisten Modelle zum Dokumentretrieval liefern gerankte Ergebnisse, eine Suche in einer formalen Wissensbasis jedoch nicht;
- das Rankingverfahren muss hybride Ergebnisse und Ergebnislisten unterstützen, also unterschiedliche Rankingfunktionen miteinander verbinden. Hierzu muss die Rankingfunktion der Faktensuche und der Dokumentsuche Werte im selben Intervall liefern, um miteinander vergleichbar zu sein oder die Unterschiede in den Wertebereichen müssen zum Zeitpunkt der Kombination von Fakten und Dokumenten angeglichen werden.

Die in dieser Arbeit entwickelten Rankingverfahren erfüllen diese Anforderungen. Beide Verfahren setzen eine termähnlichkeitsbasierte Rankingstrategie für die Faktensuche ein, die neben der lexikalischen Ähnlichkeit zu den Suchtermen auch die Abdeckung der Antwort bezüglich der Suchterme in der Suchanfrage berücksichtigt. Hierdurch wird eine anfragespezifische Rangordnung unter den Fakten erreicht. Die semantische Dokumentsuche liefert gewichtete Ergebnisse, die Gewichte liegen im selben Intervall ($[0, 1] \subset \mathbb{R}$). Eine

direkte Vergleichbarkeit der Gewichte aus der Präfixsuche zum syntaktischen Abgleich in der Wissensbasis und der Dokumentsuche, d.h. des syntaktischen Abgleichs auf dem Dokumentindex, ist nicht gegeben, da die Gewichte auf unterschiedlichen Vektorräumen berechnet werden. Sie befinden sich lediglich im selben Intervall und lassen sich kombinieren. Die Teilergebnisse werden jedoch nicht direkt zu einer geordneten Ergebnisliste zusammengefasst. Sie fließen in den Spreading Activation-Prozess ein, der die Gewichtung der Knoten im semantischen Netz weiter propagiert (vgl. Kapitel 5.2.3 für die formale und 5.2.4 für die informale Beschreibung der Verfahren). Im Rahmen der Spreading-Activation besteht die Möglichkeit, Dokumente oder Fakten stärker oder schwächer mit einfließen zu lassen, indem die entsprechenden Kantengewichte angepasst werden.

Im Rahmen dieser Arbeit wurden zwei Verfahren entwickelt. Eines der Verfahren erweitert das ursprüngliche Verfahren, indem es die Popularität der Knoten mit einbezieht. Die Popularität wird an der Anzahl ein- und ausgehender Kanten, jedoch bezogen auf die Klasse der Konzepte betrachtet, gemessen. Hierdurch sollen Konzepte, die stärker vernetzt sind und deshalb angenommen wird, dass diese populärer sind als andere Konzepte, höher gerankt werden (vgl. Kapitel 5.2.4). Mit einem Spearman-Koeffizienten von 0,9922 für das Ranking ohne Popularität und 0,9901 für das Ranking mit Popularität weisen beide beide Verfahren eine hohe Korrelation zu der idealen Rangordnung auf[1] (vgl. Kapitel 6.2.3).

Die Evaluierung der These aus Systemsicht berücksichtigte beide Rankingverfahren. Zwischen den Werten des Spearman-Koeffizienten und des erreichten F-Maßes der einzelnen Suchanfragen besteht mit 0,41 für das Ranking ohne und mit 0,36 für das Ranking mit Popularität jeweils eine schwache lineare Korrelation. Die Kennzahlen generalisierte Genauigkeit und Vollständigkeit beziehen die Position der relevanten Ergebnisse nicht explizit mit ein, sie basieren auf der Anzahl der gefundenen und nicht gefundenen relevanten Ergebnisse, während der Korrelationskoeffizient auf der Position der relevanten Ergebnisse in der Ergebnisliste basiert. Da eine systembedingte Korrelation zwischen dem Spearman-Koeffizient und dem Recall besteht, liegt die niedrige Korrelation an der generalisierten Genauigkeit. Eine Analyse der Ergebnisse der einzelnen Suchanfragen zeigt, dass es häufig nur wenige relevante Ergebnisse gibt[2], wodurch die Genauigkeit niedrig ausfällt. Dies ist jedoch stark von den Inhalten der zugrundeliegenden Wissensbasis abhängig, eine Schlussfolgerung auf die hybride semantische Suche selber kann nicht gezogen werden. Der Unterschied in den Koeffizienten beider Suchverfahren schlägt sich in der Retrievaleffektivität leicht nieder: Die generalisierte Präzision des hybriden Verfahrens mit Ranking ohne Popularität liegt 4,5%, der generalisierte Recall 8% über den Werten des Verfahrens mit Ranking ohne Popularität.

F5 - Wie kann der Benutzer bei der Anfragestellung unterstützt werden, so dass er ohne Kenntnis der zugrundeliegenden Wissensbasis Anfragen mit möglichst vielen formalen Anteilen stellt?

Eine Unterstützung bei der Anfragestellung bietet die semantische Autovervollständi-

[1] Der Wertebereich des Koeffizienten ist $[-1, 1] \in \mathbb{R}$, 1 bedeutet die perfekte Korrelation und -1 die maximale Differenz.

[2] Zu 9 der 20 Suchanfragen gibt es weniger als 6 relevante Ergebnisse im Pool.

gung in SINFIO. Die Komponente schlägt bei der Eingabe der Suchanfrage Konzepte aus der Wissensbasis vor. Der Benutzer braucht keine Kenntnisse über die Wissensbasis und muss keine formale Anfragesprache lernen. Trotzdem kann die Anfrage, falls die Vorschläge angenommen werden, formal und daher präziser als eine natürlichsprachige Anfrage formuliert werden (vgl. Kapitel 5.2.1).

Die semantische Autovervollständigung wurde mit 10 Personen ohne Kenntnisse über die Wissensbasis evaluiert. Die Aufgabe bestand darin fünf vorgegebene Fragen durch Suchen mit SINFIO zu beantworten. Das System hat die Übernahme und das Ignorieren von Vorschlägen gezählt und die Zeit, die zur Beantwortung der Frage benötigt war, gemessen. Die Ergebnisse zeigten, dass:

- die Vorschläge überwiegend angenommen werden (und nur dann ignoriert werden, wenn sie nicht eindeutig dem entsprechen, was der Benutzer fragen möchte);
- je mehr formale Anteile die Anfrage hat, umso schneller wird die Antwort gefunden.

Die semantische Autovervollständigung der Suchmaschine SINFIO unterstützt daher die Benutzer dabei, Suchanfragen in natürlicher Sprache und mit möglichst vielen formalen Anteilen zu stellen, ohne dass sie Kenntnisse über die formale Wissensbasis benötigen. Zudem zeigt die starke, statistisch signifikante, Korrelation von -0,91 zwischen der Übernahme der Vorschläge und der Antwortzeit, dass formale Anteile in der Suchanfrage zu einer schnelleren Beantwortung der Frage des Benutzers führen (vgl. Kapitel 6.2.4).

F6 - Wie können die Suchergebnisse, bestehend aus Fakten, Dokumenten und hybriden Ergebnissen, so dargestellt werden, dass diese und auch die Ergebnisliste verständlich sind?

Die Benutzerschnittstelle der hybriden semantischen Suchmaschine SINFIO basiert auf Erkenntnissen aus den Kognitionswissenschaften und wurde nach dem Prinzip des benutzerzentrierten Designs entwickelt.

Zur Evaluierung der Darstellung der Suchergebnisse wurden den 20 Umfrageteilnehmern die ersten 10 Ergebnisse von je zwei Suchmaschinen nebeneinander präsentiert und gefragt, ob sie die Darstellung verständlich finden oder nicht. Die Teilnehmer fanden die Darstellung der hybriden semantischen Suche im Durchschnitt über alle Vergleiche zu 91,38% verständlich. Die Unterschiede zu Fakten- und Dokumentsuche, Faktensuche und semantische Dokumentsuche sind statistisch signifikant (vgl. Kapitel 6.2.5).

Ebenfalls 20 Benutzer wurden gefragt, ob sie die Darstellung nach Rangordnung oder die nach Ergebnistyp gruppierte Darstellung besser finden. Die Präferenz fiel auf die nach Rangordnung organisierte Darstellungsart, der Unterschied ist statistisch signifikant. Die Teilnehmer fanden, dass die Kategorien der gruppierten Darstellung extra interpretiert werden müssen. Eine Unterauswahl und der dadurch verursachte zusätzliche Mausklick/-Tap empfanden sie als unnötig (vgl. Kapitel 6.2.5).

Bezüglich der Ergebnisdarstellung stellt sich die Frage, welcher Zusammenhang zwischen der Verständlichkeit der Suchergebnisse und der Beurteilung der Suchmaschine nach weiteren Gesichtspunkten besteht. Dies kann durch statistische Analyse der Ergebnisse der unterschiedlichen Evaluierungen untersucht werden.

Die Ergebnisse der Benutzerbefragung und die Systemkennzahl F-Maß je Suchanfrage weisen keine nennenswerte Korrelation auf, die Werte des Spearman-Koeffizienten für die

hybride semantische Suche SINFIO liegen im Intervall [-0,2, 0]. Zwischen der Verständlichkeit der Darstellung der Suchergebnisse und[3]:

- dem Finden der Antwort in der Ergebnisliste besteht mit einem Koeffizienten von 0,78 ein hoher linearer Zusammenhang;
- dem schnelleren Finden der Suchergebnisse besteht mit 0,34 ein schwacher linearer Zusammenhang;
- den Antworten, welche Suchergebnisse besser empfunden werden, besteht mit 0,27 ebenfalls ein schwacher linearer Zusammenhang.

Korrelationskoeffizienten zeigen nur einen statistischen, aber keinen kausalen Zusammenhang. Diese Zusammenhänge erlauben daher keine weitere Interpretationen. Die Begründungen der Umfrageteilnehmer deuten jedoch darauf hin, dass die Verständlichkeit der Ergebnisdarstellung implizit auch bei der Bewertung der Qualität der Suchmaschine eine Rolle spielt.

Insgesamt zeigt die Evaluierung, dass die Ergebnisdarstellung in SINFIO als verständlich empfunden wird. Zudem ziehen die Benutzer die Organisation der Ergebnisse nach Rangordnung einer nach Ergebnistyp gruppierter Darstellung vor.

7.3 Gesamtergebnis

Die hybride semantische Suche SINFIO, die formale und informale Inhalte im gesamten Suchprozess kombiniert, verbessert die Retrievaleffektivität gegenüber einer Suche, die zwar beide Arten der Inhalte durchsucht, diese jedoch nicht miteinander kombiniert. Mit diesem Ansatz wurde gezeigt, dass Verfahren zur Faktensuche und semantische Dokumentsuche verbunden werden können, so dass die Suchmaschine in der Lage ist, Fakten und Dokumente passend zu der Suchanfrage zu durchsuchen und zu kombinieren. Dies beinhaltet auch die Umsetzung eines Rankingverfahrens, das die unterschiedlichen Rankingfunktionen der verwendeten Verfahren integriert und eine hohe Korrelation zur idealen Rangordnung erreicht.

Durch die semantische Autovervollständigung, die von den Benutzern überwiegend verwendet wird, formulieren Benutzer ihr Informationsbedürfnis wie gewohnt in natürlicher Sprache und stellen trotzdem Suchanfragen mit formalen Teilen. Die Benutzer finden die Ergebnisdarstellung verständlich und können die Ergebnisse auch ohne Vorkenntnisse an hybrider Suche und der Suchmaschine SINFIO interpretieren. Insgesamt zeigen die Untersuchungen, dass die hybride semantische Suche SINFIO von den Benutzern verstanden, akzeptiert und gegenüber der Fakten- und semantischen Dokumentsuche ohne Kombination von Fakten und Dokumenten, der Faktensuche sowie der semantischen Dokumentsuche bevorzugt wird. Das Informationsbedürfnis der Benutzer kann durch SINFIO schneller befriedigt werden.

[3] Da die hybride semantische Suche der Fakten- und Dokumentsuche, der Faktensuche sowie der semantischen Dokumentsuche gegenübergestellt wurde, sind für die hybride Suche jeweils drei Korrelationskoeffizienten berechnet worden. Hier werden die Werte aus dem Vergleich hybride Suche und semantische Dokumentsuche verwendet. Sie sind am aussagekräftigsten, da die Dokumentsuche bzw. die Darstellung von Dokumenten allgemein bekannt ist.

Kapitel 8
Zusammenfassung und Ausblick

Diese Arbeit adressiert die Frage, wie die Lücke zwischen Fakten- und Dokumentsuche geschlossen werden kann, um strukturiert sowie semi- und unstrukturiert vorliegende Inhalte und die Zusammenhänge unter diesen Inhalten für die Suche ausschöpfen zu können. Sie stellt die These auf, dass eine Schließung dieser Lücke mittels eines hybriden semantischen Suchverfahrens die Effektivität der Suche in solchen heterogenen Datenmengen verbessern kann. Wie solch ein Verfahren entwickelt wurde, wie ihre Performanz ist und wie hybride semantische Suche von den Benutzern akzeptiert wird, fasst Kapitel 8.1 zusammen. Basierend auf den Ergebnissen der Evaluierung beschreibt Kapitel 8.2 die Möglichkeiten zur Verbesserung der in dieser Arbeit entwickelten hybriden semantische Suchlösung SINFIO.

8.1 Zusammenfassung

Der Anwendungskontext der Arbeit sind heutige Informationsmanagementsysteme, deren Inhalte unterschiedlich stark strukturiert vorliegen. Reiseportale geben beispielsweise durch strukturierte Daten den Zeitraum, das Ziel und den Preis einer Reise an und stellen in semi- und unstrukturierter Form weitere Informationen, wie Beschreibungen zum Hotel, Zielort, Ausflugsziele in der Umgebung zur Verfügung. Gleichzeitig konzentrieren sich semantische Suchmaschinen entweder auf die Suche nach strukturierten, formal beschriebenen (*Faktensuche*), oder nach semi- bzw. unstrukturierten, informal beschriebenen (*semantische Dokumentsuche*) Inhalten. Sie schöpfen die verfügbaren Informationen nicht aus, so dass Zusammenhänge sowie sich ergänzende Inhalte nicht den Benutzer erreichen. Diese Lücke kann durch eine hybride semantische Suche geschlossen werden, die unterschiedlich strukturierte Inhalte während des Suchprozesses kombiniert (Kapitel 1).

Das theoretische Rahmenwerk beschreibt die semantischen Technologien sowie Information Retrieval inklusive der Retrievalmodelle, Rankingverfahren und Evaluierungsmethoden. Die zwei Themenbereiche sind auf dem Gebiet der semantischen Suche zusammengeführt. Semantische Suchmaschinen lassen sich anhand verschiedener Aspekte kategorisieren, wie z.B. nach den zugrundeliegenden Daten, der Art der Anfragestellung oder der Suchverfahren. Alle diese Eigenschaften spielen bei der Leistungsfähigkeit der Suchmaschine eine Rolle. In Anbetracht der These war insbesondere die Beschaffenheit des Suchraumes von Interesse, da er einen wesentlichen Einfluss auf die Eignung von Suchverfahren hat: Bedingt durch die Unterschiede in der Struktur und in den Datenhaltungsformaten von Fakten und Dokumenten erfordern Fakten- und Dokumentsuche grundsätzlich verschiedene Vorgehensweisen. Vor diesem Hintergrund wird die *hybride semantische Suche* in dieser Arbeit als eine Suche definiert, die *traditionelle Schlüsselwortsuche oder semantisches Dokumentretrieval mit der Faktensuche kombiniert, wobei diese nicht voneinander unabhängig, sondern miteinander verzahnt durchgeführt werden und die Suchmaschine sowohl Fakten als auch Dokumente findet* (Kapitel 2).

Die Analyse des Stands der Technik von semantischen Suchmaschinen und den wenigen hybriden semantischen Suchansätzen gibt einen Überblick der unterschiedlichen

Betrachtungs- und Vorgehensweisen auf diesem Gebiet. Semantische Suche wird von vielen als die Aufgabe gesehen, eine natürlichsprachige Suchanfrage in eine formale Abfrage zu überführen. Dies ist jedoch nur eine Möglichkeit. Eine Suche kann durch mehrere formale Abfragen durchgeführt werden, was den Vorteil bringt, dass Teilergebnisse zur Präzisierung der weiteren Abfragen eingesetzt werden können. Ebenso ist es möglich, bei den Datenhaltungsformaten anzusetzen und traditionelle Retrievalmodelle einzusetzen. Entsprechende Ansätze bilden Fakten in einem Textindex ab, d.h. in derselben Repräsentationsform wie Dokumente, und adaptieren das erprobte indexbasierte Vektorraummodell für die Suche. Die vorgestellten Vorgehensweisen erfordern unterschiedliche Rankingstrategien. Formale Wissensbasen unterstützen meist keine unscharfe Suche. Sie liefern Fakten, die alle Bedingungen der Abfrage erfüllen und daher auch gleich „gut“ sind. Deshalb setzen viele entweder das Vektorraummodell ein, wodurch auch das Ranking definiert ist, oder sie beziehen die Struktur des durch die formale Wissensbasis aufgespannten Graphen ein. Letzteres wird semantisches Ranking genannt. Hybride semantische Suchansätze adaptieren Verfahren und Rankingstrategien der Fakten- und semantischen Dokumentsuche. Sie führen beides durch, tun dies jedoch entweder weitgehend unabhängig voneinander oder schränken die Suche auf bestimmte Suchanfragetypen ein (Kapitel 3).

Auf Basis der theoretischen Grundlagen und der Analyse des Stands der Technik lässt sich die *These* wie folgt formulieren: *Für unterschiedlich stark strukturierte Datenmengen ist ein integrierter Ansatz zur hybriden semantischen Suche, der formale und informale Inhalte während des gesamten Suchprozesses kombiniert, effektiver als eine semantische Dokument- und Faktensuche ohne Kombination der Inhalte.* Eine formale Beschreibung gibt eine präzise Definition des hybriden semantischen Suchproblems. Abstrahierend von den Anforderungen an Information Retrieval Systeme sind die Herausforderungen an solch eine hybride semantische Suchlösung abgeleitet. Drei Forschungsfragen betreffen die Möglichkeit, Fakten- und semantische Dokumentsuchverfahren zu kombinieren, so dass formale, informale und hybride Suchanfragen und Suchergebnisse unterstützt werden. Eine Forschungsfrage setzt sich mit dem Ranking auseinander. Zwei Forschungsfragen beziehen sich auf die Benutzerschnittstelle mit den Schwerpunkten Anfragestellung und Ergebnisdarstellung (Kapitel 4).

Den Schwerpunkt der Arbeit bildet die Entwicklung der hybriden semantischen Suchlösung SINFIO. Die mit den Forschungsfragen verbundenen Anforderungen geben einen Rahmen für das systematische Erarbeiten möglicher Lösungswege vor. Eine semantische Autovervollständigungskomponente sorgt dafür, dass Benutzer ihr Informationsbedürfnis in natürlicher Sprache formulieren, aber trotzdem teilweise oder vollständig formale Anfragen an das System stellen können. Eine Analyse der Eignung von Suchansätzen zur Fakten- und semantischen Dokumentsuche zeigt die Kombinationsmöglichkeiten von Suchverfahren auf. Die Beschreibung der hybriden semantischen Suchlösung stellt vor, wie SINFIO die ausgewählten Verfahren integriert und das semantische Ranking umsetzt. Die erarbeitete Lösung kombiniert formale und informale Inhalte im gesamten Suchprozess. SINFIO verarbeitet formale, informale sowie hybride Suchanfragen und liefert formale, informale sowie hybride Suchergebnisse. Im Gegensatz zu anderen hybriden Suchlösungen wird die Fakten- und semantische Dokumentsuche in Abhängigkeit voneinander ausgeführt und es bestehen keine Einschränkungen hinsichtlich der Suchanfragen. Entscheidendes Kriterium für die Erfolgschance neuartiger Systeme ist die Benutzerakzeptanz. Insbesondere bei solch einer hybriden Suchmaschine ist es eine Herausforderung

dies zu erreichen: Hybride Suchanfragen, Suchergebnisse und Ergebnislisten sind komplexer als natürlichsprachige Suchanfragen und die von Websuchmaschinen bekannten Ergebnislisten. Um die kognitive Last bei der Ergebnisinterpretation zu mindern und eine verständliche Darstellung zu bieten, setzt SINFIO auf eine ähnlich strukturierte, textuelle Darstellung aller Inhalte und verzichtet auf die für Fakten übliche Graphdarstellung (Kapitel 5).

Eine entwicklungsbegleitende Proof-of-Concept Untersuchung zeigt, dass die hybride semantische Suche effektiver sein kann als Faktensuche oder semantische Dokumentsuche. Eine umfangreiche Evaluierung der These und der Forschungsfragen erfolgte auf Basis von DBpedia und Wikipedia sowohl aus System- als auch aus Benutzersicht. Der im Rahmen der Arbeit erstellte Gold Standard beinhaltet die Relevanzbeurteilung der ersten 10 Suchergebnisse der verglichenen Suchmaschinen zu 20 Suchanfragen, die reale Informationsbedürfnisse ausdrücken. Sowohl die Systemkennzahl F-Maß, als auch die Benutzerbefragung bestätigen die These: *Die hybride semantische Suche führt zu einer höheren Retrievaleffektivität als die Fakten- und Dokumentsuche ohne eine Kombination der formalen und informalen Inhalte.* Zudem bestätigen die Ergebnisse, dass Benutzer ihr Informationsbedürfnis durch die hybride semantische Suche schneller befriedigen können als durch die semantische Dokumentsuche oder die Faktensuche. Die zwei erarbeiteten Rankingstrategien sind für die hybride semantische Suche geeignet, denn sie weisen eine hohe Korrelation zu der idealen Rangordnung auf. Die semantische Autovervollständigung kann ohne Kenntnisse der zugrundeliegenden Wissensbasis eingesetzt werden und wurde von den Umfrageteilnehmern angenommen. Je mehr formale Anteile eine Suchanfrage hatte, umso schneller kann das Informationsbedürfnis befriedigt werden. Die Benutzer finden die Ergebnisse verständlich. Insgesamt zeigen die Untersuchungen, dass *die hybride semantische Suche SINFIO von den Benutzern verstanden und angenommen wird. Sie wird gegenüber der Fakten- und semantischen Dokumentsuche, der Faktensuche sowie der semantischen Dokumentsuche bevorzugt. Das Informationsbedürfnis der Benutzer kann durch die hybride semantische Suche SINFIO schneller befriedigt werden als durch die anderen Suchlösungen* (Kapitel 6 und 7).

Fazit: Die hybride semantische Suche stellt für ihre Nutzer die beste Methode dar, ihr Informationsbedürfnis aus unterschiedlich strukturierten Datenmengen zu befriedigen: Sie ist effektiver als die reine Faktensuche, semantische Dokumentsuche oder Fakten- und Dokumentsuche ohne Kombination der Inhalte.

8.2 Ausblick

Nach Auswertung der Evaluierungsergebnisse lassen sich Verbesserungspotenziale beim Ranking der Fakten sowie in der Effizienz erkennen. Zudem ist es möglich, die Suche um Kontextualisierung und Personalisierung zu erweitern. Die Erweiterungen können die Effektivität der Suchmaschine steigern.

- **Ranking**: Im Rahmen der Benutzerbefragung fiel der Kommentar, dass die Sortierung von Fakten unklar ist. Die Anmerkung bezog sich auf eine Ergebnisliste, die nur aus Fakten bestand. Der Grund liegt in der unscharfen Suche in RDF-Stores. Die gefundenen Fakten sind nicht gewichtet und werden in einer zufälligen Reihenfolge zurückgeliefert. Je nach Suchanfrage kann die Ergebnisreihenfolge auch nach dem Spreading Activation als zufällig erscheinen. Auch das konnektivitätsbasierte Ranking ändert dies nicht immer, da die Konnektivität nicht notwendigerweise die Popularität eines Konzeptes widerspiegelt (vgl. Kapitel 5.2.4). Um die tatsächliche Popularität eines Konzeptes mit einzubeziehen, besteht die Möglichkeit eine der großen Suchmaschinen, wie Google, abzufragen. Der Popularitätswert kann dann in Abhängigkeit der Anzahl der Treffer berechnet werden.
- **Effizienz**: Die Laufzeit von SINFIO ist im Wesentlichen von zwei Faktoren beeinflusst: der niedrigen Verarbeitungsgeschwindigkeit von RDF-Stores und der Laufzeit des Spreading Activation Prozesses. Die Verarbeitungsgeschwindigkeit kann nicht ohne Weiteres beeinflusst werden. Das Spreading Activation, welches für die Kombination der Teilergebnisse aus der Fakten und Dokumentsuche eingesetzt wird, lässt sich effizienter implementieren. Eine Effizienzsteigerung kann z.B. erreicht werden, indem die entfernten Tripel während des Aktivierungsprozesses inkrementell importiert werden anstatt das semantische Netz vor dem Prozess aufzubauen (vgl. [Marie, 2014, Marie et al., 2013]).
- **Effektivität**: Wie nützlich ein Suchergebnis für einen Benutzer ist, hängt von dessen Informationsbedarf ab. Dieser ist wiederum vom Kontext des Benutzers, seinen persönlichen Präferenzen, Interessen und seinem Wissenstand abhängig [Schumacher et al., 2011].
  - **Kontext**: Der Benutzerkontext kann die geografische Position, der thematische Bezug, die aktuellen Wetterdaten usw. sein. Sie spielen je nach Suchziel mehr oder weniger eine Rolle. Im Rahmen von SINFIO, die sich nicht auf eine bestimmte Domäne konzentriert, ist der thematische Kontext am interessantesten. Um ihn dynamisch zu bestimmen, wird üblicherweise die Suchhistorie des Benutzers ausgewertet. Schlüsselwörter sowie Konzepte der letzten Suchanfragen werden üblicherweise zur Erweiterung der aktuellen Suchanfrage eingesetzt [Mangold, 2007]. In dem hybriden Suchansatz kann der thematische Kontext des Benutzers an mehreren Stellen mit einbezogen werden. So können vorher gesuchte Konzepte anhand ihrer Nähe zu den Konzepten, die mit der aktuellen Suchanfrage gemeint sein könnten, die Reihenfolge der Autovervollständigungsvorschläge beeinflussen. Wurde beispielsweise zuerst nach Programmiersprachen und dann nach Java gesucht, so würde die Sprache Java vor der Insel Java in der Vorschlagsliste erscheinen. Weiterhin können die Ergebnisse der letzten Suchanfragen in die Gewichtung im Rahmen des Spreading Activations mit in die Suche einfließen. Es besteht die Möglichkeit, während eines Suchprozesses die Knotengewichte derjenigen Konzepte

und Dokumente zu erhöhen, die in den vorherigen Suchen involviert waren. Alternativ kann das Spreading Activation-Netzwerk so gestaltet werden, dass die Gewichtung der Ergebnisse einer Suche nicht gelöscht, sondern vor jeder neuen Suche abgeschwächt wird. So bleiben frühere Suchen im Gedächtnis des Netzes enthalten, haben jedoch weniger Einfluss, je weiter sie zurückliegen. In beiden Fällen passt sich die Suche an den thematischen Kontext des Benutzers an.

– **Personalisierung**: Das Einsetzen der persönlichen Präferenzen, Interessen und des Wissensstandes des Benutzers kann als ein Spezialfall der Kontextualisierung angesehen werden. Semantische Suchmaschinen modellieren diese Informationen üblicherweise in der Wissensbasis und greifen bei der Suche darauf zurück, um das Ranking der Ergebnisse anzupassen. SINFIO bietet durch den Einsatz des Spreading Activation-Netzes die Möglichkeit, die Wissensbasis an den Benutzer anzupassen, indem Gewichte dauerhaft und dynamisch an den Benutzer angepasst werden. So können Konzepte bzw. Teilgraphen, die häufig erfragt werden, ein initiales Gewicht erhalten. Zusammen mit der Kontextualisierung führt dies für jeden Benutzer zu einem personalisierten und an den aktuellen Kontext angepassten Suchraum.

Die hybride Suche hat das Potential für einen breiten Einsatz, wie etwa in großen Websuchmaschinen. Wie in dieser Arbeit gezeigt wurde, verbessert sie die Retrievaleffektivität, führt zu einem schnelleren Finden der Antwort und wird von den Benutzern gegenüber der semantischen Dokumentsuche bevorzugt. Kombiniert mit dem Wissen aus der enormen Menge an Nutzungsdaten großer Suchmaschinen, kann durch hybride Verfahren die Entwicklung der Suche in strukturierten, semi- und unstrukturierten Inhalten vorangetrieben werden.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[Abras et al., 2004] Abras, C., Maloney-Krichmar, D., and Preece, J. (2004). User-centered design. In *Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*. Publications.

[Agichtein et al., 2005] Agichtein, E., Cucerzan, S., and Brill, E. (2005). Analysis of factoid questions for effective relation extraction. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 567–568. ACM.

[Al-Maskari et al., 2007] Al-Maskari, A., Sanderson, M., and Clough, P. (2007). The relationship between ir effectiveness measures and user satisfaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 773–774. ACM.

[Aleman-Meza et al., 2005] Aleman-Meza, B., Halaschek-Weiner, C., Arpinar, I. B., Ramakrishnan, C., and Sheth, A. P. (2005). Ranking complex relationships on the semantic web. *Internet Computing, IEEE*, 9(3):37–44.

[Allam and Haggag, 2012] Allam, A. M. N. and Haggag, M. H. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences*, 2(3).

[Amaral et al., 2004] Amaral, C., Laurent, D., Martins, A. F., Mendes, A., Pinto, C., et al. (2004). Design and implementation of a semantic search engine for portuguese. In *Proceedings of the LREC*.

[Amin et al., 2009] Amin, A., Hildebrand, M., van Ossenbruggen, J., Evers, V., and Hardman, L. (2009). Organizing suggestions in autocompletion interfaces. In *Advances in Information Retrieval*, pages 521–529. Springer.

[Anderson, 1983] Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–295.

[Anyanwu, 2005] Anyanwu, K. (2005). Semrank: ranking complex relationship search results on the semantic web. In *Proceedings of the 14th International World Wide Web Conference*, pages 117–127. ACM Press.

[Anyanwu et al., 2005] Anyanwu, K., Maduko, A., and Sheth, A. (2005). Semrank: ranking complex relationship search results on the semantic web. In *Proceedings of the 14th International Conference on World Wide Web*, pages 117–127. ACM.

[Anyanwu and Sheth, 2003] Anyanwu, K. and Sheth, A. (2003). P-queries: enabling querying for semantic associations on the semantic web. In *Proceedings of the 12th International Conference on World Wide Web*, pages 690–699. ACM.

[Aranda et al., 2013] Aranda, C. B., Corby, O., Das, S., Feigenbaum, L., Gearon, P., Glimm, B., Harris, S., Hawke, S., Herman, I., Humfrey, N., Michaelis, N., Ogbuji, C., Perry, M., Passant, A., Polleres, A., Prud'hommeaux, E., Seaborne, A., and Williams, G. T. (2013). SPARQL 1.1 Overview, w3c recommendation 21 march 2013. Website.

[Arias et al., 2011] Arias, M., Fernández, J. D., Martínez-Prieto, M. A., and de la Fuente, P. (2011). An empirical study of real-world sparql queries. In *Proceedings of the 1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011) at the 20th International World Wide Web Conference*.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., and Ives, Z. (2007).

Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference*, pages 11–15. Springer.

[Baeza-Yates and Maarek, 2011] Baeza-Yates, R. and Maarek, Y. (2011). Web retrieval. In *Modern Information Retrieval*, pages 449–517. Pearson, 2nd edition.

[Baeza-Yates et al., 2011] Baeza-Yates, R., Navarro, G., and Ziviani, N. (2011). Documents: Languages & properties. In *Modern Information Retrieval*, pages 203–254. Pearson, 2nd edition.

[Baeza-Yates and Ribeiro-Neto, 2011a] Baeza-Yates, R. and Ribeiro-Neto, B. (2011a). Ir models. In *Modern Information Retrieval*, pages 57–61. Addison Wesley, 2nd edition.

[Baeza-Yates and Ribeiro-Neto, 2011b] Baeza-Yates, R. and Ribeiro-Neto, B. (2011b). The ir system. In *Modern Information Retrieval*, pages 5–8. Addison Wesley, 2nd edition.

[Baeza-Yates and Ribeiro-Neto, 2011c] Baeza-Yates, R. and Ribeiro-Neto, B. (2011c). *Modern Information Retrieval*. Pearson, 2nd edition.

[Baeza-Yates and Ribeiro-Neto, 2011d] Baeza-Yates, R. and Ribeiro-Neto, B. (2011d). Retrieval evaluation. In *Modern Information Retrieval*, pages 131–176. Addison Wesley, 2nd edition.

[Balog and Neumayer, 2013] Balog, K. and Neumayer, R. (2013). A test collection for entity retrieval in dbpedia. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 737–740. ACM.

[Berkan, 2007] Berkan, R. C. (2007). Semantic search: an antidote for poor relevancy. *ReadWriteWeb. com archive, published on May*, 29.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):34–43.

[Beyer, 2011] Beyer, M. A. (2011). Information management in the 21st century is about all kinds of semantics. Technical Report G00215806, Gartner.

[Bhagdev et al., 2008] Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., and Petrelli, D. (2008). Hybrid Search: effectively combining keywords and semantic searches. In *Proceedings of the 5th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC'08, pages 554–568. Springer.

[Bizer and Schultz, 2008] Bizer, C. and Schultz, A. (2008). Benchmarking the performance of storage systems that expose sparql endpoints. *World Wide Web Internet And Web Information Systems*.

[Blackler et al., 2005] Blackler, A. L., Popovic, V., and Mahar, D. P. (2005). Intuitive interaction applied to interface design. In *Proceedings of the International Design Congress – IASDR*.

[Blanco et al., 2013] Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., and Tran, D. T. (2013). Repeatable and reliable semantic search evaluation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 21(0).

[Bock et al., 2008] Bock, J., Chaves, F., Forcher, B., Grimm, S., Henß, J., Kleb, J., Sintek, M., Tserendorj, T., and Volz, R. (2008). System overview. Project Deliverable.

[Brennan and Prediger, 1981] Brennan, R. L. and Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699.

[Brickley and Guha, 2004] Brickley, D. and Guha, R. V. (2004). RDF Vocabulary Description Language 1.0: RDF Schema. Website.

[Brooke, 1996] Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

[Buntine et al., 2005] Buntine, W., Valtonen, K., and Taylor, M. (2005). The alvis document model for a semantic search engine. In *Proceedings of the 2nd Annual European Semantic Web Conference*, pages 1–2.

[Buscaldi et al., 2005] Buscaldi, D., Rosso, P., and Arnal, E. S. (2005). A wordnet-based query expansion method for geographical information retrieval. In *Working notes for the CLEF workshop*.

[Carroll et al., 2005] Carroll, J. J., Bizer, C., Hayes, P., and Stickler, P. (2005). Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web*, pages 613–622. ACM.

[Castells et al., 2007] Castells, P., Fernandez, M., and Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19(2):261–272.

[Celino et al., 2007] Celino, I., Turati, A., Valle, E. D., and Cerizza, D. (2007). Squiggle – a semantic search engine at work. In *Proceedings of the 4th European Semantic Web Conference*.

[Chin et al., 1988] Chin, J. P., Diehl, V. A., and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, pages 213–218. ACM.

[Chowdhury, 2003] Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1):51–89.

[Cicchetti and Feinstein, 1990] Cicchetti, D. V. and Feinstein, A. R. (1990). High agreement but low kappa: Ii. resolving the paradoxes. *Journal of clinical epidemiology*, 43(6):551–558.

[Cleverdon, 1997] Cleverdon, C. (1997). The cranfield tests on index language devices. In *Readings in Information Retrieval*, pages 47–59. Morgan Kaufmann Publishers Inc.

[Cohen and Kjeldsen, 1987] Cohen, P. R. and Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management*, 23(4):255–268.

[Cohen et al., 2003] Cohen, S., Mamou, J., Kanza, Y., and Sagiv, Y. (2003). XSEarch: A semantic search engine for XML. In *Proceedings of the 29th International Conference on Very Large Data Bases*, pages 45–56.

[Converse et al., 2008] Converse, T., Kaplan, R. M., Pell, B., Prevost, S., Thione, L., and Walters, C. (2008). Powerset's natural language wikipedia search engine. Technical report, Powerset, inc., California.

[Corby et al., 2004] Corby, O., Dieng-Kuntz, R., and Faron-Zucker, C. (2004). Querying the semantic web with corese search engine. In *Proceedings of the 5th ECAI/PAIS*, volume 16, page 705.

[Crestani, 1997] Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482.

[Croft et al., 2010a] Croft, B., Metzler, D., and Strohman, T. (2010a). Beyond bag of

words. In *Search Engines: Information Retrieval in Practice*, pages 455–489. Addison-Wesley Publishing Company.

[Croft et al., 2010b] Croft, B., Metzler, D., and Strohman, T. (2010b). Evaluating search engines. In *Search Engines: Information Retrieval in Practice*, pages 301–342. Addison-Wesley Publishing Company.

[Croft et al., 2010c] Croft, B., Metzler, D., and Strohman, T. (2010c). Queries and interfaces. In *Search Engines: Information Retrieval in Practice*, pages 191–236. Addison-Wesley Publishing Company.

[Cyganiak et al., 2014] Cyganiak, R., Wood, D., and Lanthaler, M. (2014). RDF 1.1 Concepts and Abstract Syntax, w3c recommendation 25 february 2014. Website.

[d'Aquin et al., 2008] d'Aquin, M., Motta, E., Sabou, M., Angeletou, S., Gridinoc, L., Lopez, V., and Guidi, D. (2008). Toward a new generation of semantic web applications. *Intelligent Systems, IEEE*, 23(3):20–28.

[Dmitrieva et al., 2007] Dmitrieva, J., Bei, Y., and Verbeek, F. J. (2007). Ontological context visualization. In *Proceedings of the OWLED*, volume 258 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Duke et al., 2007] Duke, A., Glover, T., and Davies, J. (2007). Squirrel: An advanced semantic search and browse facility. In Franconi, E., Kifer, M., and May, W., editors, *The semantic web: research and applications.*, volume 4519 of *Lecture Notes in Computer Science*, pages 341–355. Springer.

[Dunlop and Crossan, 2000] Dunlop, M. D. and Crossan, A. (2000). Predictive text entry methods for mobile phones. *Personal Technologies*, 4(2-3):134–143.

[Dustin et al., 2010] Dustin, L., Böhm, C., and Naumann, F. (2010). Extracting structured information from wikipedia articles to populate infoboxes. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, pages 1661–1664.

[Eichler et al., 2010] Eichler, K., Hemsen, H., Neumann, G., Reithinger, N., Schmeier, S., Schumacher, K., and Seifert, I. (2010). DiLiA – the digital library assistant. In *Research and Advanced Technology for Digital Libraries*, pages 534–537. Springer.

[Elbedweihy et al., 2012] Elbedweihy, K., Wrigley, S. N., Ciravegna, F., Reinhard, D., and Bernstein, A. (2012). Evaluating semantic search systems to identify future directions of research. In *Proceedings of the 2nd International Workshop on Evaluation of Semantic Technologies*, number 843 in CEUR Workshop Proceedings, pages 25–36.

[Elbedweihyhadija et al., 2012] Elbedweihyhadija, K., Wrigley, S. N., and Ciravegna, F. (2012). Evaluating semantic search query approaches with expert and casual users. In *Proceedings of the International Semantic Web Conference (2)*, volume 7650 of *Lecture Notes in Computer Science*, pages 274–286. Springer.

[Etzioni et al., 2011] Etzioni, O., Fader, A., Christensen, J., Soderland, S., and Mausam, M. (2011). Open information extraction: The second generation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, pages 3–10. AAAI Press.

[Exeler et al., 2015] Exeler, C., Waitelonis, J., and Sack, H. (2015). Linked data annotated document retrieval.

[Fader et al., 2011] Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545. Association for Computational Linguistics.

[Feinstein and Cicchetti, 1990] Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

[Fensel and van Harmelen, 2007] Fensel, D. and van Harmelen, F. (2007). Unifying reasoning and search to web scale. *IEEE Internet Computing*, 11(2):94–96.

[Ferber, 2003a] Ferber, R. (2003a). Das vektorraummodell. In *Information Retrieval*, chapter 3.6, pages 61–82. Dpunkt Verlag.

[Ferber, 2003b] Ferber, R. (2003b). *Information Retrieval*. Dpunkt Verlag.

[Fernandez et al., 2008] Fernandez, M., Lopez, V., Sabou, M., Uren, V. S., Vallet, D., Motta, E., and Castells, P. (2008). Semantic search meets the web. In *Proceedings of the IEEE International Conference on Semantic Computing*, ICSC '08, pages 253–260. IEEE Computer Society.

[Fernandez et al., 2009] Fernandez, M., Lopez, V., Sabou, M., Uren, V. S., Vallet, D., Motta, E., and Castells, P. (2009). Using trec for cross-comparison between classic ir and ontology-based search models at a web scale. In *Semantic Search 2009 Workshop at the 18th International World Wide Web Conference*.

[Fleiss, 1971] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

[Fleiss, 1981] Fleiss, J. L. (1981). The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2:212–236.

[Forcher et al., 2009] Forcher, B., Schumacher, K., Sintek, M., and Roth-Berghofer, T. (2009). Evaluating the intelligibility of medical ontological terms. In *Proceedings of the 5th Workshop on Knowledge Engineering and Software Engineering, KESE*.

[Frank et al., 2007] Frank, A., Krieger, H.-U., Xu, F., Uszkoreit, H., Crysmann, B., Jörg, B., and Schäfer, U. (2007). Question answering from structured knowledge sources. *Journal of Applied Logic*, 5(1):20–48.

[Franklin et al., 2005] Franklin, M., Halevy, A., and Maier, D. (2005). From databases to dataspaces: A new abstraction for information management. *SIGMOD Rec.*, 34(4):27–33.

[Gärtner et al., 2014] Gärtner, M., Rauber, A., and Berger, H. (2014). Bridging structured and unstructured data via hybrid semantic search and interactive ontology-enhanced query formulation. *Knowledge and Information Systems*, 41(3):761–792.

[Giunchiglia and Shvaiko, 2003] Giunchiglia, F. and Shvaiko, P. (2003). Semantic matching. *The Knowledge Engineering Review*, 18(03):265–280.

[Glover et al., 2001] Glover, E. J., Lawrence, S., Gordon, M. D., Birmingham, W. P., and Giles, C. L. (2001). Web search—your way. *Communications of the ACM*, 44(12):97–102.

[Goldschmidt and Krishnamoorthy, 2005] Goldschmidt, D. E. and Krishnamoorthy, M. (2005). Architecting a search engine for the semantic web. In *AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications*.

[Gordon and Pathak, 1999] Gordon, M. and Pathak, P. (1999). Finding information on the world wide web: The retrieval effectiveness of search engines. *Information processing & management*, 35(2):141–180.

[Greenberg, 2005] Greenberg, J. (2005). Understanding metadata and metadata schemes. *Cataloging & Classification Quarterly*, 40(3-4):17–36.

[Grimes et al., 2007] Grimes, C., Tang, D., and Russell, D. M. (2007). Query logs alone are not enough. In *Workshop on query log analysis at WWW*. Citeseer.

[Grimnes et al., 2009] Grimnes, G. A., Adrian, B., Schwarz, S., Maus, H., Schumacher, K., and Sauermann, L. (2009). Semantic desktop for the end-user. *i-com, Special Issue: Nutzerinteraktion im Social Semantic Web*, 3:25–32.

[Grothkast et al., 2008] Grothkast, A., Adrian, B., Schumacher, K., and Dengel, A. (2008). OCAS: Ontology-based corpus and annotation scheme. In *High-level Information Extraction Workshop*, pages 25–35. ECML PKDD 2008.

[Group, 2012] Group, W. O. W. (2012). OWL 2 Web Ontology Language. Website.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition - Special issue: Current issues in knowledge modeling*, 5(2):199–220.

[Guha et al., 2003] Guha, R., McCool, R., and Miller, E. (2003). Semantic search. In *Proceedings of the 12th International Conference on Word Wide Web*.

[Guo et al., 2004] Guo, Y., Pan, Z., and Heflin, J. (2004). An evaluation of knowledge base systems for large owl datasets. In *International Semantic Web Conference*, pages 274–288. Springer.

[Guo et al., 2005] Guo, Y., Pan, Z., and Heflin, J. (2005). Lubm: A benchmark for owl knowledge base systems. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2):158–182.

[Hahn et al., 2010] Hahn, R., Bizer, C., Sahnwaldt, C., Herta, C., Robinson, S., Bürgle, M., Düwiger, H., and Scheel, U. (2010). Faceted wikipedia search. In *Business Information Systems*, pages 1–11.

[Handschuh and Staab, 2003] Handschuh, S. and Staab, S. (2003). *Annotation for the semantic web*, volume 96. IOS Press.

[He et al., 2007] He, H., Wang, H., Yang, J., and Yu, P. S. (2007). BLINKS: Ranked keyword searches on graphs. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 305–316. ACM.

[Hearst, 2009a] Hearst, M. (2009a). Mobile search interfaces. In *Search user interfaces*, pages 297–305. Cambridge University Press.

[Hearst, 2009b] Hearst, M. (2009b). Query specification. In *Search user interfaces*, pages 91–119. Cambridge University Press.

[Hearst, 2011] Hearst, M. (2011). User interfaces for search. In *Modern Information Retrieval*, pages 21–55. Pearson, 2nd edition.

[Heflin and Hendler, 2000] Heflin, J. and Hendler, J. (2000). Searching the web with SHOE. In *AAAI Workshop on AI for Web Search*, pages 35–40.

[Hildebrand et al., 2007] Hildebrand, M., Ossenbruggen, J., and van Hardman, L. (2007). An analysis of search-based user interaction on the semantic web. Report, CWI, Amsterdam, Holland.

[Hirschman and Gaizauskas, 2001] Hirschman, L. and Gaizauskas, R. (2001). Natural language question answering: the view from here. *Natural Language Engineering*, 7(04):275–300.

[Hitzler et al., 2008a] Hitzler, P., Krötzsch, M., Rudolph, S., and Sure, Y. (2008a). Einfache ontologie in rdf und rdf schema. In *Semantic Web: Grundlagen*, pages 36–88. Springer.

[Hitzler et al., 2008b] Hitzler, P., Krötzsch, M., Rudolph, S., and Sure, Y. (2008b). Formale semantik in owl. In *Semantic Web: Grundlagen*, pages 163–198. Springer.

[Hogan et al., 2006] Hogan, A., Harth, A., and Decker, S. (2006). Reconrank: A scalable ranking method for semantic web data with context. In *Proceedings of the 2nd Workshop on Scalable Semantic Web Knowledge Base Systems*.

[Hudson and Hall, 1997] Hudson, R. and Hall, P. (1997). *Software without Frontiers: A Multi-Platform, Multi-Cultural, Multi-National Approach*. John Wiley & Sons, Inc.

[Huynh et al., 2005] Huynh, D., Mazzocchi, S., and Karger, D. R. (2005). Piggy bank: Experience the semantic web inside your web browser. In *Proceedings of the International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 413–430. Springer.

[Hyvönen and Mäkelä, 2006] Hyvönen, E. and Mäkelä, E. (2006). Semantic autocompletion. In *The Semantic Web–ASWC 2006*, pages 739–751. Springer.

[Hyvönen et al., 2003] Hyvönen, E., Saarela, S., and Viljanen, K. (2003). Ontogator: combining view-and ontology-based search with semantic browsing. *Information Retrieval*, 16:17.

[Janetzko, 2008] Janetzko, D. (2008). Objectivity, reliability, and validity of search engine count estimates. *International Journal of Internet Science*, 3(1):7–33.

[Järvelin and Kekäläinen, 2000] Järvelin, K. and Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 41–48. ACM.

[Joachims, 2002] Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142. ACM.

[Johnson et al., 1992] Johnson, R. A., Wichern, D. W., et al. (1992). *Applied multivariate statistical analysis*, volume 4. Prentice hall Englewood Cliffs, NJ.

[Jung et al., 2009] Jung, H., Lee, M.-K., You, B.-J., and Kim, D.-W. (2009). Comparative evaluation of reliabilities on semantic search functions: Auto-complete and entity-centric unified search. In *Proceedings of the 5th International Conference on Active Media Technology*, AMT '09, pages 104–113. Springer.

[Karov and Edelman, 1998] Karov, Y. and Edelman, S. (1998). Similarity-based word sense disambiguation. *Computational linguistics*, 24(1):41–59.

[Kaufmann and Bernstein, 2007] Kaufmann, E. and Bernstein, A. (2007). How useful are natural language interfaces to the semantic web for casual end-users? In *Proceedings of the ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 281–294. Springer.

[Kekäläinen, 2005] Kekäläinen, J. (2005). Binary and graded relevance in ir evaluations—comparison of the effects on ranking of ir systems. *Information Processing & Management*, 41(5):1019–1033.

[Kekäläinen and Järvelin, 2002] Kekäläinen, J. and Järvelin, K. (2002). Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129.

[Kelly, 2009] Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224.

[Kelly et al., 2008] Kelly, D., Harper, D. J., and Landau, B. (2008). Questionnaire mode effects in interactive information retrieval experiments. *Information Processing & Management*, 44(1):122–141.

[Khan et al., 2004] Khan, L., McLeod, D., and Hovy, E. (2004). Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal*, 13(1):71–85.

[Kirakowski and Corbett, 1993] Kirakowski, J. and Corbett, M. (1993). Sumi: the software usability measurement inventory. *British Journal of Educational Technology*, 24(3):210–212.

[Kiryakov et al., 2004] Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, 2:49–79.

[Klyne et al., 2014] Klyne, G., Carroll, J. J., and McBride, B. (2014). RDF 1.1 Concepts and Abstract Syntax, w3c proposed recommendation 09 january 2014. Website.

[Kondrak, 2005] Kondrak, G. (2005). N-gram similarity and distance. In *Proceedings of the 12th International Conference on String Processing and Information Retrieval*, pages 115–126.

[Kosala and Blockeel, 2000] Kosala, R. and Blockeel, H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 2(1):1–15.

[Kripke, 1972] Kripke, S. A. (1972). *Naming and necessity*. Springer.

[Kukich, 1992] Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439.

[Ladwig and Tran, 2010] Ladwig, G. and Tran, D. T. (2010). Combining keyword translation with structured query answering for efficient keyword search. In *Proceedings of the 7th Extended Semantic Web Conference*.

[Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

[Lange et al., 2010] Lange, D., Böhm, C., and Naumann, F. (2010). Extracting structured information from wikipedia articles to populate infoboxes. Technical Report 38, Hasso-Plattner-Institut für Softwaresystemtechnik an der Universität Potsdam.

[Lawrence et al., 1999] Lawrence, P., Sergey, B., Rajeev, M., and Terry, W. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

[Lei et al., 2006] Lei, Y., Uren, V. S., and Motta, E. (2006). Semsearch: A search engine for the semantic web. In *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management*, pages 238–245.

[Likert, 1932] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

[Lopez et al., 2012] Lopez, V., Fernández, M., Motta, E., and Stieler, N. (2012). Poweraqua: Supporting users in querying and exploring the semantic web. *Semantic Web*, 3(3):249–265.

[Lopez et al., 2006] Lopez, V., Motta, E., and Uren, V. S. (2006). Poweraqua: Fishing the semantic web. In *The Semantic Web: research and applications*, pages 393–410. Springer.

[Lopez et al., 2005] Lopez, V., Pasin, M., and Motta, E. (2005). Aqualog: An ontology-portable question answering system for the semantic web. In *The Semantic Web:*

Research and Applications, pages 546–562. Springer.

[Luhn, 1960] Luhn, H. P. (1960). Key word-in-context index for technical literature (kwic index). *American Documentation*, 11(4):288–295.

[Maedche and Staab, 2002] Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge engineering and knowledge management: Ontologies and the semantic web*, pages 251–263. Springer.

[Mäkelä, 2005] Mäkelä, E. (2005). Survey of semantic search research. In *Seminar on Knowledge Management on the Semantic Web*.

[Mangold, 2007] Mangold, C. (2007). A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontologies*, 2(1):23–34.

[Manning et al., 2008a] Manning, C. D., Raghavan, P., and Schütze, H. (2008a). Determining the vocabulary of terms. In *Introduction to Information Retrieval*, pages 22–35. Cambridge University Press.

[Manning et al., 2008b] Manning, C. D., Raghavan, P., and Schütze, H. (2008b). Evaluation in information retrieval. In *Introduction to Information Retrieval*, pages 151–176. Cambridge University Press.

[Manning et al., 2008c] Manning, C. D., Raghavan, P., and Schütze, H. (2008c). *Introduction to Information Retrieval*. Cambridge University Press.

[Manning et al., 2008d] Manning, C. D., Raghavan, P., and Schütze, H. (2008d). The vector space model for scoring. In *Introduction to Information Retrieval*. Cambridge University Press.

[Manola et al., 2014] Manola, F., Miller, E., and McBride, B. (2014). RDF 1.1 Primer, w3c recommendation 25 february 2014. Website.

[Marchionini, 2006] Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46.

[Marie, 2014] Marie, N. (2014). *Linked data based exploratory search*. PhD thesis, Université de Nice Sophia-Antipolis.

[Marie et al., 2013] Marie, N., Corby, O., Gandon, F., and Ribière, M. (2013). Composite interests' exploration thanks to on-the-fly linked data spreading activation. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 31–40. ACM.

[Mayer and Moreno, 2003] Mayer, R. E. and Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52.

[McCool et al., 2005] McCool, R., Cowell, A. J., and Thurman, D. A. (2005). End-user evaluations of semantic web technologies. In *Proceedings of the ISWC 2005, Workshop on End User Semantic Web Interaction*.

[Mika, 2005] Mika, P. (2005). Flink: Semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, 3:211–223.

[Mizarro, 1997] Mizarro, S. (1997). Relevance: the whole history. *Journal of the American Society for Information Science*, 48(9):810–832.

[Möller et al., 2010] Möller, K., Hausenblas, M., Cyganiak, R., Handschuh, S., and Grimnes, G. A. (2010). Learning from linked open data usage: Patterns & metrics. In *Proceedings of the 2nd Web Science Conference*.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity

recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[Nagypál, 2007] Nagypál, G. (2007). *Possibly imperfect ontologies for effective information retrieval.* PhD thesis, Karlsruhe Institute of Technology.

[Neumann and Xu, 2003] Neumann, G. and Xu, F. (2003). Mining answers in german web pages. In *Proceedings of the International Conference on Web Intelligence (IEEE/WIC)*, pages 125–131. IEEE.

[Nielsen, 2000a] Nielsen, J. (2000a). Why you only need to test with 5 users. *Jakob Nielsen's Alterbox.*, 19.

[Nielsen, 2000b] Nielsen, J. (2000b). Why you only need to test with 5 users. Website.

[Nielsen, 2012a] Nielsen, J. (2012a). How many test users in a usability study. *Nielsen Norman Group*, 4.

[Nielsen, 2012b] Nielsen, J. (2012b). How many test users in a usability study. *Jakob Nielsen's Alertbox.*

[Nielsen and Budiu, 2013] Nielsen, J. and Budiu, R. (2013). *Mobile Usability.* New Riders.

[Nielsen and Landauer, 1993] Nielsen, J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, pages 206–213. ACM.

[Nodine et al., 2000] Nodine, M., Fowler, J., Ksiezyk, T., Perry, B., Taylor, M., and Unruh, A. (2000). Active information gathering in InfoSleuth. *International Journal of Cooperative Information Systems*, 9(1/2):3–28.

[Nöth, 1990] Nöth, W. (1990). Semantics of semiotics. In *Handbook of semiotics*, pages 103–114. Indiana University Press.

[Novacek et al., 2009] Novacek, V., Groza, T., and Handschuh, S. (2009). Knowledge-based search for oncological literature. In *Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems*, pages 1–8. IEEE.

[Osterhoff et al., 2012] Osterhoff, J., Waitelonis, J., and Sack, H. (2012). Widen the peepholes! entity-based auto-suggestion as a rich and yet immediate starting point for exploratory search. In *GI-Jahrestagung*, pages 1039–1046.

[Partee, 2011] Partee, B. (2011). Formal semantics: Origins, issues, early impact. *The Baltic International Yearbook of Cognition, Logic and Communication*, 6(0).

[Paşca, 2003] Paşca, M. (2003). Open-domain question answering from large text collections. *Computational Linguistics*, 29(4):665–667.

[Payne, 2014] Payne, S. L. B. (2014). *The Art of Asking Questions: Studies in Public Opinion, 3*, volume 3. Princeton University Press.

[Perez-Aguera et al., 2010] Perez-Aguera, J. R., Arroyo, J., Greenberg, J., Perez-Iglesias, J., and Fresno, V. (2010). INEX+DBPEDIA: A Corpus for Semantic Search Evaluation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1161–1162. ACM.

[Purandare and Pedersen, 2004] Purandare, A. and Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the 8th Conference on Computational Natural Language Learning*, volume 4, pages 41–48.

[Radev et al., 2002] Radev, D. R., Qi, H., Wu, H., and Fan, W. (2002). Evaluating web-based question answering systems. *Ann Arbor*, 1001:48109.

[Randolph, 2005] Randolph, J. J. (2005). Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Online Submission*.

[Ratinov and Roth, 2009] Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155. Association for Computational Linguistics.

[Reichert et al., 2005] Reichert, M., Linckels, S., Meinel, C., and Engel, T. (2005). Student's perception of a semantic search engine. In *Proceedings of the CELDA*, pages 139–147.

[Reuschling et al., 2010] Reuschling, C., Agne, S., and Dengel, A. (2010). Dynaq - faceted search for document retrieval. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. ACM.

[Robertson, 1977] Robertson, S. E. (1977). The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304.

[Robertson, 1981] Robertson, S. E. (1981). The methodology of information retrieval experiment. In *Information retrieval experiment*, pages 9–31. Butterworths.

[Robertson and Belkin, 1978] Robertson, S. E. and Belkin, N. J. (1978). Ranking in principle. *Journal of Documentation*, 34(2):93–100.

[Rocha et al., 2004] Rocha, C., Schwabe, D., and Aragao, M. P. (2004). A hybrid approach for searching in the semantic web. In *Proceedings of the 13th International Conference on World Wide Web*, WWW, pages 374–383. ACM.

[Ruotsalo, 2012] Ruotsalo, T. (2012). Domain specific data retrieval on the semantic web. In *Proceedings of the 9th International Conference on The Semantic Web: Research and Applications*, ESWC'12, pages 422–436. Springer.

[Sacaleanu et al., 2008] Sacaleanu, B., Orasan, C., Spurk, C., Ou, S., Ferrandez, O., Kouylekov, M., and Negri, M. (2008). Entailment-based question answering for structured data. In *Posters and Demonstrations, 22nd International Conference on Computational Linguistics*, pages 29–32.

[Sack, 2005] Sack, H. (2005). Npbibsearch - an ontology augmented bibliographic search. In *SWAP*, volume 166 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Sack, 2010] Sack, H. (2010). Semantische suche – theorie und praxis am beispiel der videosuchmaschine yovisto.com. In *Web 3.0 & Semantic Web*, number 271 in HMD - Praxis der Wirtschaftsinformatik, pages 13–25. dpunkt Vlg. Heidelberg.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

[Sarawagi, 2008] Sarawagi, S. (2008). Information extraction. *Foundations and trends in databases*, 1(3):261–377.

[Sauermann et al., 2008] Sauermann, L., Kiesel, M., Schumacher, K., and Bernardi, A. (2008). Semantic desktop. In *Social Semantic Web – Web 2.0 - Was nun?*, pages 337–362. Springer.

[Sauermann et al., 2007] Sauermann, L., van Elst, L., and Dengel, A. (2007). PIMO – a framework for representing personal information models. In *Proceedings of the I-SEMANTICS*, pages 270–277.

[Schreiber et al., 2006] Schreiber, G., Amin, A., van Assem, M., Boer, V. D., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., Niet, M. D., Omelayenko, B., van Ossenbrug-

gen, J., Siebes, R., Taekema, J., Wielemaker, J., and Wielinga, B. (2006). Multimedian e-culture demonstrator. In *Semantic Web Challenge at the 5th International Semantic Web Conference*.

[Schumacher, 2007] Schumacher, K. (2007). Four methods for supervised word sense disambiguation. In *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems*, volume 4592 of *LNCS*, pages 317–328. Springer.

[Schumacher et al., 2011] Schumacher, K., Forcher, B., and Tran, D. T. (2011). Semantische suche. In *Semantische Technologien – Grundlagen. Konzepte. Anwendungen.*, pages 227–252. Spektrum Verlag.

[Schumacher and Sintek, 2011] Schumacher, K. and Sintek, M. (2011). Searching web 3.0 content. In *Proceedings of the 7th International Conference on Web Information Systems and Technologies.*

[Schumacher et al., 2008] Schumacher, K., Sintek, M., and Sauermann, L. (2008). Combining fact and document retrieval with spreading activation for semantic desktop search. In *Proceedings of the 5th European Semantic Web Conference*, volume 5021 of *LNCS*, pages 569–583. Springer.

[Schütze, 1998] Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

[Schwarz, 2006] Schwarz, S. (2006). A context model for personal knowledge management applications. In *Modeling and retrieval of context*, pages 18–33. Springer.

[Shadbolt et al., 2006] Shadbolt, N., Berners-Lee, T., and Hall, W. (2006). The semantic web revisited. *IEEE intelligent systems*, 21(3):96–101.

[Sheth et al., 2002] Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., and Warke, Y. (2002). Managing semantic content for the web. *Internet Computing, IEEE*, 6(4):80–87.

[Simons et al., 2004] Simons, G. F., Lewis, W. D., Farrar, S. O., Langendoen, D. T., Fitzsimons, B., and Gonzalez, H. (2004). The semantics of markup: Mapping legacy markup schemas to a common semantics. In *Workshop on NLP and XML: RDF/RDFS and OWL in Language Technology*, NLPXML '04, pages 25–32. Association for Computational Linguistics.

[Sinkkilä et al., 2008] Sinkkilä, R., Mäkelä, E., Hyvönen, E., and Kauppinen, T. (2008). Combining context navigation with semantic autocompletion to solve problems in concept selection. *SeMMA*, 346:61–68.

[Smucker et al., 2007] Smucker, M. D., Allan, J., and Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 623–632. ACM.

[Steiger, 1980] Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

[Stieger and Aleksy, 2009] Stieger, B. and Aleksy, M. (2009). Utilization of knowledge management for service business processes improvement. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, IEEE Computer Society, pages 171–175.

[Stojanovic, 2003] Stojanovic, N. (2003). On analysing query ambiguity for query refinement: The librarian agent approach. In *Conceptual Modeling-ER*, pages 490–505.

Springer.

[Stojanovic et al., 2001] Stojanovic, N., Maedche, A., Staab, S., Studer, R., and Sure, Y. (2001). Seal: a framework for developing semantic portals. In *Proceedings of the 1st International Conference on Knowledge Capture*, pages 155–162. ACM.

[Stojanovic et al., 2003] Stojanovic, N., Studer, R., and Stojanovic, L. (2003). An approach for the ranking of query results in the semantic web. In *The Semantic Web – ISWC 2003*, pages 500–516. Springer.

[Stokoe et al., 2003] Stokoe, C., M. P. Oakes, M., and Tait, J. (2003). Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 159–166. ACM.

[Stoyanovich et al., 2010] Stoyanovich, J., Mee, W., and Ross, K. A. (2010). Semantic ranking and result visualization for life sciences publications. In *Proceedings of the IEEE 26th International Conference on Data Engineering (ICDE)*, pages 860–871. IEEE.

[Strasunskas and Tomassen, 2010] Strasunskas, D. and Tomassen, L. S. (2010). On variety of semantic search systems and their evaluation methods. In *Proceedings of the International Conference on Information Management and Evaluation*, pages 380–387.

[Suen, 1979] Suen, C. Y. (1979). N-gram statistics for natural language understanding and text processing. *IEEE transactions on pattern analysis and machine intelligence*, (2):164–172.

[Sure and Iosif, 2002] Sure, Y. and Iosif, V. (2002). First results of a semantic web technologies evaluation. *Common Industry Program at the federated event: ODBASE*, 2.

[Sweller et al., 1998] Sweller, J., van Merrienboer, J. J., and Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review*, 10(3):251–296.

[Tablan et al., 2015] Tablan, V., Bontcheva, K., Roberts, I., and Cunningham, H. (2015). Mímir: An open-source semantic search framework for interactive information seeking and discovery. *Web Semantics: Science, Services and Agents on the World Wide Web*, 30:52–68.

[Tang and Dwarkadas, 2004] Tang, C. and Dwarkadas, S. (2004). Hybrid global-local indexing for efficient peer-to-peer information retrieval. In *NSDI*, volume 4, pages 16–16.

[Tang et al., 1999] Tang, R., Shaw, W. M., and Vevea, J. L. (1999). Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science*, 50(3):254–264.

[Thanh, 2011] Thanh, D. T. (2011). *Process-oriented Semantic Web Search*. PhD thesis, Karlsruher Institut für Technologie (KIT), Fakultät für Wirtschaftswissenschaften, Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB).

[Todorov and Schandl, 2008] Todorov, D. and Schandl, B. (2008). Small-scale evaluation of semantic web-based applications. Technical report, University of Vienna,.

[Tombros and Sanderson, 1998] Tombros, A. and Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10. ACM.

[Tran et al., 2009] Tran, D. T., Wang, H., Rudolph, S., and Cimiano, P. (2009). Top-k

exploration of query candidates for efficient keyword search on graph-shaped (rdf) data. In *Proceedings of the 25th International Conference on Data Engineering (ICDE)*.

[Tsatsaronis and Panagiotopoulou, 2009] Tsatsaronis, G. and Panagiotopoulou, V. (2009). A generalized vector space model for text retrieval based on semantic relatedness. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 70–78. Association for Computational Linguistics.

[Tummarello et al., 2010] Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R., and Decker, S. (2010). Sig.ma: live views on the web of data. In *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 1301–1304.

[Uren et al., 2007] Uren, V. S., Lei, Y., Lopez, V., Liu, H., Motta, E., and Giordanino, M. (2007). The usability of semantic search tools: a review. *The Knowledge Engineering Review*, 22(04):361–377.

[Uren et al., 2010] Uren, V. S., Sabou, M., Motta, E., Fernandez, M., Lopez, V., and Lei, Y. (2010). Reflections on five years of evaluating semantic search systems. *International Journal of Metadata, Semantics and Ontologies*, 5(2):87–98.

[Vallet et al., 2005] Vallet, D., Fernández, M., and Castells, P. (2005). An ontology-based information retrieval model. In *The Semantic Web: Research and Applications*, pages 455–470. Springer.

[van Berkel and Smedt, 1988] van Berkel, B. and Smedt, K. D. (1988). Triphone analysis: a combined method for the correction of orthographical and typographical errors. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 77–83. Association for Computational Linguistics.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth-Heinemann, 2nd edition.

[Viera et al., 2005] Viera, A. J., Garrett, J. M., et al. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363.

[Voorhees et al., 1999] Voorhees, E. M. et al. (1999). The trec-8 question answering track report. In *Proceedings of the TREC*, volume 99, pages 77–82.

[Wahlster, 2008] Wahlster, W. (2008). SmartWeb — Ein multimodales Dialogsystem für das semantische Web. In *Informatikforschung in Deutschland*, pages 300–311. Springer.

[Wahlster et al., 2006] Wahlster, W., Schwarzkopf, E., Sauermann, L., Roth-Berghofer, T., Pfalzgraf, A., Kiesel, M., Heckmann, D., Dengler, D., Dengel, A., and Sintek, M. (2006). Web 3.0: Convergence of web 2.0 and the semantic web. *Technology Radar*, II:1–23.

[Waitelonis and Sack, 2010] Waitelonis, J. and Sack, H. (2010). Exploratory semantic video search with yovisto. In *Proceedings of the 4th IEEE International Conference on Semantic Computing (ICSC*, pages 446–447. IEEE Computer Society.

[Waitelonis and Sack, 2012] Waitelonis, J. and Sack, H. (2012). Towards exploratory video search using linked data. *Multimedia Tools and Applications*, 59(2):645–672.

[Wang et al., 2009] Wang, H., Liu, Q., Penin, T., Fu, L., Zhang, L., Tran, D. T., Yu, Y., and Pan, Y. (2009). Semplore: A scalable IR approach to search the Web of Data. *Journal of Web Semantics*, 7(3):177–188.

[Wang et al., 2011] Wang, H., Tran, D. T., Liu, C., and Fu, L. (2011). Lightweight integration of ir and db for scalable hybrid search with integrated ranking support. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):490–503.

[Webber, 2010] Webber, W. E. (2010). *Measurement in information retrieval evaluation*. PhD thesis, The University of Melbourne, Department of Computer Science and Software Engineering.

[White et al., 2003] White, R. W., Jose, J. M., and Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing & Management*, 39(5):707–733.

[Wrigley et al., 2010] Wrigley, S. N., Reinhard, D., Elbedweihyhadija, K., Bernstein, A., and Ciravegna, F. (2010). Methodology and campaign design for the evaluation of semantic search tools. In *Proceedings of the 3rd International Semantic Search Workshop*, SEMSEARCH '10, pages 1–10. ACM.

[Zhang et al., 2008] Zhang, L., Liu, Q., Zhang, J., Wang, H., Pan, Y., and Yu, Y. (2008). Semplore: An IR approach to scalable hybrid query of semantic web data. *The Semantic Web*, pages 652–665.

[Zitzelberger et al., 2014] Zitzelberger, A. J., Embley, D. W., Liddle, S. W., and Scott, D. T. (2014). Hykss: Hybrid keyword and semantic search. *Journal on Data Semantics*, pages 1–17.