Handbuch Internet-Suchmaschinen 2
D. Lewandowski (Hrsg.)

Semantic Search – State-of-the-Art-Überblick zu semantischen Suchlösungen im WWW

Ulrike Spree[a], Nadine Feißt[b], Anneke Lühr[c], Beate Piesztal[d], Nina Schroeder[e], Patricia Wollschläger[f]
Hochschule für Angewandte Wissenschaften Hamburg,
Department Information
Finkenau 35, 22081 Hamburg
[a]ulrike.spree@haw-hamburg.de, [b]nadine.feisst@haw-hamburg.de,
[c]anneke.luehr@haw-hamburg.de, [d]beate.piesztal@haw-hamburg.de,
[e]nina.schroeder@haw-hamburg.de, [f]patricia.wollschlaeger@haw-hamburg.de

Abstract. In diesem Kapitel wird ein Überblick über bestehende semantische Suchmaschinen gegeben. Insgesamt werden 95 solcher Suchdienste identifiziert und im Rahmen einer Inhaltsanalyse verglichen. Es kann festgestellt werden, dass die Semantische Suche sich wesentlich von den im Rahmen des Semantic Web propagierten Technologien unterscheidet und Semantik in den betrachteten Suchmaschinen weiter zu fassen ist. Die betrachteten Suchmaschinen werden in ein Stufenmodell, welches nach dem Grad der Semantik unterscheidet, eingeordnet. Das Kapitel schließt mit 8 Thesen zum aktuellen Stand der semantischen Suche.

Keywords. Suchmaschinen, Semantische Suche, Inhaltsanalyse.

Einleitung

Wolfgang Wahlster, Direktor des Deutschen Forschungszentrums für Künstliche Intelligenz, erklärte 2006 im Zusammenhang mit der Diskussion über ein europäisches Technologieentwicklungsprojekt[1] die Funktionsweise einer semantischen Suche mit folgendem Beispiel: Wenn ein Hobby-Golfer eine Aufnahme seiner Sportart in einer Foto-Community mit der Bezeichnung „Golf" tagge, sei für den Hobby-Golfer implizit durch den Kontext eindeutig, dass die Sportart gemeint ist. Für eine herkömmliche Suchmaschine hingegen sei nicht entscheidbar, ob die Sportart oder ein Automodell gemeint sei. Ein semantisches System könne solche Mehrdeutigkeiten auflösen und eine Klassifizierung nach verschiedenen Zusatzbegriffen oder Kategorien empfehlen [1]. Diesem eher bescheiden anmutenden Szenario stehen von Technologieanbietern geschürte hohe Erwartungen gegenüber. Verfolgt man die seit 2006 [2] [3] [4] vor allem in der Publikumspresse und in Technologieblogs geführte Diskussion zu semantischen

[1] Gemeint ist das 2006 geplante deutsch-französische Technologieprojekt Quaero, das eine Alternative zu Google entwickeln sollte. Auf dem Technologiegipfel in Potsdam wurde dann entschieden, statt eines gemeinsamen deutsch-französischen Projektes das Projekt in Frankreich unter dem Namen Quaero und in Deutschland mit dem Titel Theseus fortzuführen (Deutscher Bundestag 2006: Drucksache 16/4671).

Suchmaschinen, entsteht der Eindruck, dass auf dem stark Anbieter- orientierten Suchmaschinenmarkt viele Suchmaschinenentwickler derzeit das Label „semantisch“ als zusätzliches Marketinginstrument einsetzen und semantische Suchmaschine als Synonym für bessere, relevantere, schnellere, den Bedürfnissen der Nutzer unmittelbar entsprechenden Suchlösungen verwendet wird [5]. Das Spektrum der in diesem Zusammenhang einbezogenen Technologien und Anwendungen ist äußerst weit und heterogen.

Der vorliegende Beitrag unternimmt den Versuch einer Momentaufnahme des aktuellen Entwicklungs- und Diskussionsstandes auf dem Gebiet semantische Suchmaschinen. Der Beitrag basiert auf der Ausarbeitung einer studentischen Arbeitsgruppe, die im Rahmen eines forschungsorientierten Mastermoduls im Wintersemester 2010/2011 an der Hochschule für Angewandte Wissenschaften Hamburg einen empirischen Überblick über Art und Umfang im Web öffentlich zugänglicher, semantischer Suchmaschinen erarbeitet hat, wobei der Schwerpunkt auf allgemeinen Suchmaschinen lag. Einführend wird zunächst die aktuelle Diskussion der Begriffe „Semantic Web“ und „semantische Suche“ bzw. „Semantic Search“[2] knapp umrissen. Um einen möglichst umfassenden Einblick in den State of the Art zu erhalten, wurde auf eine enge definitorische Festlegung des Begriffs „Semantische Suche“ verzichtet. Stattdessen wurden bei der Recherche (siehe Abschnitt 4) bewusst weiträumig Suchdienste berücksichtigt, die sich entweder selber als semantisch beschreiben, oder von anderen als solche beschrieben werden. Nach einer gerafften Zusammenfassung des Forschungsstandes werden 63 im Web frei zugängliche Suchmaschinen auf der Grundlage eines zunächst groben, im Verlauf der Bearbeitung verfeinerten, Kategorienrasters in Bezug auf ihre Eigenschaften und die eingesetzten Technologien untersucht. Die inhaltsanalytische Untersuchung der semantischen Suchmaschinen stützt sich hauptsächlich auf die Selbstauskunft der Suchmaschinenbetreiber und wird ergänzt durch systematisches Testen der Anwendungen. Basierend auf den Ergebnissen der Inhaltsanalyse werden grundlegende Ansätze und Trends auf dem Gebiet der semantischen Suche identifiziert. Da es erklärtes Ziel der Untersuchung war herauszufinden, ob sich die gesichteten Suchmaschinen in Hinblick auf den Komplexitätsgrad der eingesetzten semantischen Methoden unterscheiden, wird eingangs ein Modell entwickelt, mit dem sich Grade der semantischen Komplexität unterscheiden lassen.

1. Begriffseingrenzung Semantic Web und Semantic Search

Semantik als Teildisziplin der Linguistik beschäftigt sich mit der „Erforschung der Vermittlung von Bedeutung durch Wörter und Sätze im alltäglichen Sprachgebrauch.“ [6, S. 100]. In der informationswissenschaftlichen Praxis und Forschung werden semantische Fragestellungen vor allem in Hinblick auf Beziehungen zwischen Zeichen und bezeichneten Gegenständen diskutiert [7, S. 19]. Traditionelle wissensorganisatorische Ansätze aus dem Bereich der Ordnungssysteme (Klassifikationen) und der Dokumentationssprachen (Thesauri) nutzen seit jeher die klassischen Beziehungen der Wortsemantik (Äquivalenzrelationen, Hierarchierelationen und Assoziationsrelationen) zur Disambiguierung von Bedeutungen und um Informationssysteme zu entwickeln, die vielfältige Zugriffspunkte (Access points) auf Informationen erlauben [7, S. 176 ff.].

[2] Die Begriffe „emantische Suche“ und „Semantic Search“ werden im vorliegenden Artikel synonym verwendet.

Ebenfalls auf linguistischen Erkenntnissen fußend werden – häufig im Gegensatz zu klassischen Verfahren der Wissensorganisation – Verfahren der natürlich-sprachlichen Sprachverarbeitung (Natural Language Processing; NLP) genannt. Dieser Ansatz der automatischen Sprachverarbeitung unterscheidet sich dadurch von älteren Verfahren der Sprachverarbeitung, dass nicht auf intellektuell erstellte Regeln zurückgegriffen wird, sondern Texte aufgrund allgemeiner Algorithmen, die häufig auf statistischen Inferenzen basieren, analysiert werden und beispielsweise automatisch indexiert oder klassifiziert werden [8]. Hier besteht also eine breite informationswissenschaftliche Vorarbeit, auf die semantische Suchmaschinen zurückgreifen können.

Nach Hitzler et al. basiert das Semantic Web auf der Idee, Informationen maschinenlesbar aufzubereiten, plattform- und anwenderunabhängig auszutauschen und zueinander in Beziehung zu setzen. Diese Fähigkeit bezeichnet man als Interoperabilität [9, S. 11]. Neben der Interoperabilität zeichnen sich semantische Anwendungen im Semantic Web durch Reasoning aus. Reasoning bedeutet das automatische Schlussfolgern von „neuen" Informationen aus gegebenen unter Anwendung von Prinzipien der formalen Logik [9, S. 11]. Im Vergleich zum World Wide Web (WWW) werden im Semantic Web die individuellen Bedürfnisse der Anwender stärker in den Vordergrund gestellt. Eine Leitidee des Semantic Web und damit der Semantischen Suche lautet: „Finde Wege und Methoden, Informationen so zu repräsentieren, dass Maschinen damit in einer Art und Weise umgehen können, die aus menschlicher Sicht nützlich und sinnvoll erscheint" [9, S. 12].

Gattanis Definition des semantischen Web ist technologieorientierter: „Semantic web is about annotating facets and attributes associated with web content and linking data. In other words, semantic web is about teaching machines to read web pages, which are designed to be read by humans" [10]. Jedoch auch hier nimmt der Aspekt der Maschinenlesbarkeit von Informationen eine zentrale Rolle ein. Um eine Suchanfrage richtig zu beantworten, sind Informationen von verschiedenen Webseiten notwendig. Mithilfe semantischer Technologien sind Suchmaschinen in der Lage, die Informationen verschiedener Webseiten zu konsolidieren, weil Daten plattform- und anwenderunabhängig zur Verfügung stehen [10].

Dieses Verständnis vertritt auch das W3C, das in seinen Publikationen das Semantic Web als eine Reihe von Technologien beschreibt, die applikationsunabhängigen Daten- und Informationsaustausch ermöglichen, weil Informationen geteilt und wiederverwendet werden können und jede Relation eindeutig gekennzeichnet ist [11]. Wenn im Zusammenhang mit dem W3C von Semantic Web oder semantischen Technologien gesprochen wird, geht es immer um ein festgelegtes Set von Standards, das in abgewandelter Form, in dem auf Berners-Lee zurückgehenden „Layer-cake"-Diagramm (Abbildung 1) visualisiert werden kann.

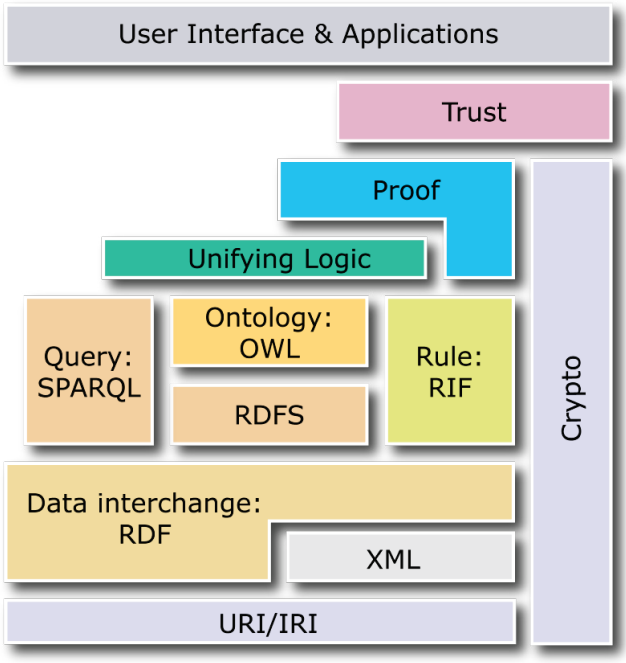

Abbildung 1: Semantic Web Layer Cake [12, S. 346]

Von den in diesem Zusammenhang aufgeführten Standards sind im Kontext der semantischen Suche vor allem das Prinzip der URI (Uniform Resource Identifier), RDF(S) und OWL sowie die Abfragesprache SPARQL von Bedeutung [12]. Gemeinsam ist den genannten Standards und Technologien (Details siehe Abschnitt 4.2), dass sie die Funktion haben, Prinzipien des WWW wie URIs und Hyperlinks von der Dokumentenebene auf die Datenebene zu erweitern [11]. Ivan Herman bringt diese Auffassung auf den Punkt, wenn er als eine wichtige Funktion semantischer Ansätze die Nutzung des „Web of Data“ als eine Art von „Content Management System“ für das WWW beschreibt, innerhalb dessen die Mitglieder der Community als Redakteure fungieren [12, S. 10]. Neuerdings wird in der Internet Community die Diskussion um die Erweiterung des bestehenden Web weniger unter dem Label Semantic Web und verstärkt unter dem Topos der ‚linked data’ geführt. Dadurch treten die im Begriff der Semantik mitschwingenden philosophischen Aspekte der Frage nach Bedeutung im ontologischen Sinn in den Hintergrund [13]. Das Prinzip der linked data zielt vorrangig darauf ab, Relationen zwischen Daten und zwischen Daten und Dokumenten auszudrücken und für die Wiederauffindbarkeit von Dokumenten und Fakten im Web nutzbar zu machen. Im Folgenden wird in geraffter Form ein Überblick darüber vermittelt, in wieweit diese Technologien bereits in der praktischen und theoretischen Auseinandersetzung mit dem Thema Websuche ihren Niederschlag finden.

2. Forschungsstand Semantische Suche

Die Themen Semantische Suche (semantic search) und semantische Suchtechnologien werden in populärwissenschaftlichen und journalistischen Webressourcen wie Blogs, Foren und Online-Journals intensiv diskutiert [14], [15], [16]. Das 2003 gegründete Technologieblog ReadWriteWeb liefert beispielsweise alleine 398 Ergebnisse zum Suchbegriff „semantic search“ (Interne Suche ReadWriteWeb. Stand 2011-03-08). Die fachwissenschaftliche Diskussion zu semantischen Suchlösungen verteilt sich neben wenigen Überblicksbeiträgen [2] [17] [18] [19] [20] [63] [64] auf eine Vielzahl von spezifischen Einzelthemen, wie die Entwicklung und Erprobung von, meistens domainspezifischen, Prototypen [21], die Analyse des Einsatzes von Ontologien im Information Retrieval, beispielsweise in Form ontologiebasierter Frage-Antwort-Systeme, die Entwicklung von Clustering-Algorithmen für die facettierte Darstellung von Suchergebnissen, die Integration verschiedener semantischer Ebenen in der Suche, Graph-basierte semantische Modelle oder die Visualisierung von Suchergebnissen. Mit öffentlich über das WWW zugänglichen allgemeinen Suchmaschinen hat sich die Fachwissenschaft bisher noch nicht systematisch auseinandergesetzt.[3] Einen knappen Überblick – allerdings von neueren Entwicklungen schon teilweise überholt – zu den eingesetzten Technologien und ihrem Zusammenwirken liefern Kassim & Rahmany [63] und Dong, Hussain & Chang [64].

Wie nicht anders zu erwarten, werden die eingesetzten Technologien in der populärwissenschaftlichen bzw. der journalistischen Diskussion nur grob umrissen und es wird nicht trennscharf zwischen statistischen und linktopologischen Verfahren und auf RDF basierenden Semantic-Web-Technologien im engeren Sinne unterschieden [20], [16].

Vielmehr wird Semantische Suche vor allem aus der Nutzerperspektive – ohne Berücksichtigung der dahinter liegenden Technologie – im Hinblick auf Anforderungen an eine optimierte Websuche beschrieben. Als semantisch werden dann solche Technologien beschrieben, die die Funktion erfüllen, die Intention der Suchanfrage zu erkennen[4] und aufzulösen und aufgrund dessen bedeutsame („meaningful“) Ergebnisse zu liefern. Semantisch suchen in diesem Sinne heißt, dass Suchmaschinen die Suchanfragen der User „verstehen” und in der Lage sind, Suchanfragen mit den persönlichen Vorstellungen und Einstellungen des Users zu verknüpfen, indem sie den Kontext der Suchanfrage erfassen und Synonyme in die Suche einbeziehen. In letzter Konsequenz sollen Suchmaschinen lernen wie Menschen zu denken, indem sie etwa den kognitiven Aufwand der Suche nach Synonymen vom Nutzer auf das System verlagern und einen Rahmen für die begriffliche Eingrenzung von Suchanfragen bieten [14] [15] [20]. Durch die Konzept- bzw. Begriffsorientierung anstelle der Orientierung am reinen Wortlaut soll eine „semantic invariance“ erreicht werden. Das heißt, Suchmaschinen sollten für Suchanfragen, die unterschiedlich formuliert wurden, aber inhaltlich gleichbedeutend sind, dieselben Suchergebnisse anzeigen [22, S. 2]. In diesen Zusammenhang gehört auch die Erwartung, dass Suchmaschinen in der Lage sein sollen, natürlichsprachig formulierte Suchanfragen (gemeint ist die natürliche Wortfolge etwa als Frageformulierung alternativ zum Einsatz Boolescher Operatoren) zu interpretieren [22, S. 2]. Es wird erwartet, dass semantische Suchmaschinen relevante(ere) Suchergebnisse liefern dadurch,

[3] Systematische Recherche mit den Keywords „search engines“ und „semantic“ in LISA (Library and Information Science Abstracts). ACM-DL und IEEE-Xplore.

[4] S. dazu auch den Beitrag zu „Query Understanding” in diesem Band.

dass sie Fakten extrahieren, bzw. auf einschlägige Textstellen („pertinent parts of texts") direkt verweisen [14].

Viele zusammenfassende Darstellungen betonen die Funktion von semantischen Suchmaschinen, den Nutzer bei der Konkretisierung von Suchanfragen zu unterstützen. Als Beispiel hierfür wird etwa die Archivierung aller bereits getätigten Suchanfragen und deren Ergebnisse und ihre Nutzung zur Lösung neuer Suchanfragen genannt [22]. Während Imielinski und Signorini die Optimierungsmöglichkeiten ausschließlich auf der Ebene der Suchanfrage sehen, geht Ewell davon aus, dass bereits am Index und somit der Wissensbasis angesetzt werden kann. Die erwarteten Optimierungsmöglichkeiten durch Einsatz semantischer Technologien liegen somit zum einen in den Verfahren, die dem User helfen, seine Suchanfrage zu erweitern oder einzugrenzen. Zum anderen kann der semantische Ansatz bereits beim Indexieren des Datenbestandes eingesetzt werden.

Zusammenfassend kann man sagen, dass das in der gesichteten Literatur ausgedrückte Verständnis von Semantischer Suche bzw. semantischen Ansätzen im Information Retrieval nicht auf die oben skizzierten RDF-basierten semantischen Technologien im engeren Sinne eingeschränkt ist, sondern auch statistische und probabilistische Ansätze einschließt. Aus dem Methodenspektrum des Semantic Web übernommene Ansätze beziehen sich auf Möglichkeiten der Organisation und Strukturierung von Daten und Informationen, Methoden zur Verbesserung der Interoperabilität von Daten und Kollektionen sowie den Einsatz plattform- und anwenderunabhängiger Techniken. Automatisches Reasoning, also das eigenständige Erschließen neuer Information aus dem Kontext heraus, spielt in den Darstellungen eine untergeordnete Rolle. Vielmehr ist die Diskussion um die Semantische Suche stark fokussiert auf die Optimierung von Suchanfragen und die anschließende Ergebnisdarstellung bzw. Visualisierung. Im Kern der Idee der Semantischen Suche steht die Anforderung, dass nicht die User die Suchmaschinen verstehen müssen, sondern die Suchmaschinen lernen sollten, die User bzw. deren Anfragen eindeutig zu interpretieren. Das Ziel sowohl des Semantic Web als auch der Semantischen Suche ist, die Informationsqualität für den User zu erhöhen und den Kontext von Informationen eigenständig zu erschließen.

3. Das Stufenmodell semantischer Komplexität

Motivation für die Entwicklung des Stufenmodells ist die heuristische Vorannahme, dass sich Suchmaschinen im Hinblick auf den semantischen Komplexitätsgrad – verkürzt: Grad an Semantik – der bei der Datenaufbereitung eingesetzten Technologien unterscheiden lassen. Das vorgestellte Stufenmodell (s. Abb. 2) ist angelehnt an die semantische Treppe von Blumenauer und Pellegrini [23], anhand der die semantische Reichhaltigkeit von Systemen bestimmt werden kann. Für die Ableitung des Modells wurden zunächst zum Einsatz kommende Arten der Datenaufbereitung bzw. -organisation identifiziert und zueinander in Verhältnis gesetzt sowie deren Auswirkungen auf die Daten definiert.

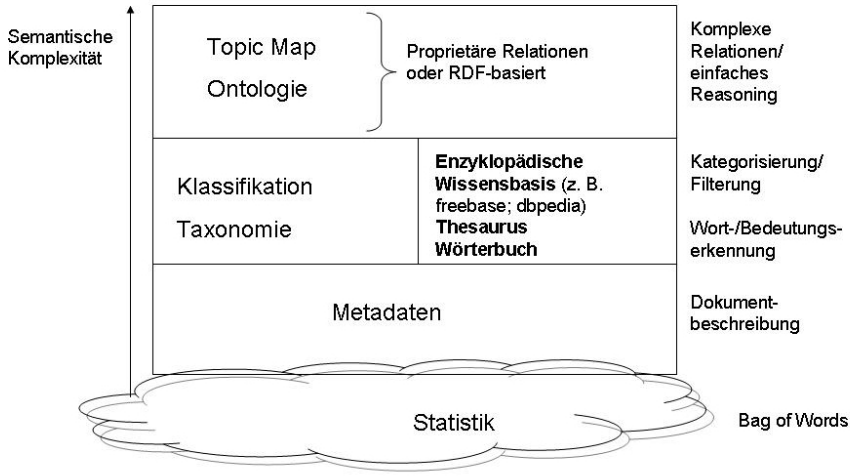

Abbildung 2: Stufenmodell zur Einordnung der semantischen Komplexität

Nicht anders als Websuchmaschinen allgemein, setzen semantische Suchmaschinen in der Regel auf der – auch unter dem Label „bag of words“ bekannten – Volltextrecherche von unstrukturierten Daten auf. Hierbei werden die Indices in der Regel bereits durch regelbasierte linguistische Verfahren (stemming; Stopwortlisten) bereinigt. Die Daten werden dann unter Anwendung statistischer und probabilistischer Verfahren (Indexerstellung, Termgewichtung, Clustering) so aufbereitet, dass der Vergleich der Repräsentation von Frage- und Dokumentinhalten die Ausgabe gewichteter Treffer ermöglicht. Im Rahmen der Information-Retrieval-Forschung wurde eine Vielzahl unterschiedlicher Algorithmen entwickelt, deren gemeinsames Kennzeichen ist, dass sie immer bessere Wege zu finden versuchen, durch die sich die Wahrscheinlichkeit erhöhen lässt, dass relevante Treffer angezeigt werden [24]. Textstatistische Verfahren werden durch die durch Googles PageRank-Algorithmus populär gewordenen linktopologischen Ansätze ergänzt, die aufgrund von Linkauswertungen die relative Bedeutung eines Dokumentes messen. Gemeinsam ist den genannten Ansätzen, dass sie – zunehmend komplexere – Algorithmen einsetzen, die auf empirisch abgeleiteten Vermutungen darüber basieren, welche statistisch erfassbaren Eigenschaften von Dokumenten (Worthäufigkeiten, Linkeigenschaften und -häufigkeiten) dazu geeignet sind, relevante Treffer zu liefern. Die Ansätze streben jedoch nicht an, Dokumente im Sinne der eindeutigen Identifikation von Aussagen zu “verstehen“.

Bereits als erste Stufe der „semantischen Veredlung“ der unstrukturierten Volltexte im Sinne einer Anreicherung mit expliziten Bedeutungsinformationen kann eine Beschreibung durch, bzw. Nutzung von – in der Regel auf Dokumentebene angesiedelten – beschreibenden Metadaten gelten [25] [26, S. 269 f.]. Typische Beispiele für dokumentspezifische Metadaten sind bibliografische Informationen (Autor, Titel) oder Informationen zur Medienart (Bild, Text).

Ein höherer Grad semantischer Komplexität kann durch den Einsatz verschiedener Werkzeuge erreicht werden, die eine Erkennung von Wörtern (Disambiguierung) und deren Bedeutungen innerhalb eines Dokuments (named entity recognition) erlauben.

Dazu bietet sich eine Vielzahl von Möglichkeiten zur Datenaufbereitung an. Diese reichen von der Einbindung einfacher, elektronisch aufbereiteter Wörterbücher (Formenwörterbücher) über komplexe computerlinguistische Tools mit ausgefeilten Relationsstrukturen. Mit der Bereitstellung umfangreicher, auf Thesaurusstrukturen [27] aufsetzender, elektronisch aufbereiteter Wortschätze für zahlreiche Sprachen wie WordNet (englisch) oder GermanNet (deutsch) stehen mächtige Hilfsmittel – teilweise sogar, wie im Fall von WordNet, frei im Web verfügbar – zur computerlinguistischen Datenanreicherung zur Verfügung [28]. Die an der Universität Magdeburg entwickelte Pilotanwendung „clever search“ ermöglicht vom Prinzip her[5] eine Disambiguierung einer Suchanfrage gegen die WordNet Datenbank, bevor sie zur Weiterverarbeitung an eine Suchmaschine geschickt wird [29].

Mit SKOS (Simple Knowledge Organization System) steht eine auf RDF basierende formale Sprache zur Verfügung, mit der sich kontrollierte Vokabulare, Thesauri und Klassifikationen für Semantic-Web-Anwendungen kodieren lassen [30]. Die im Rahmen der EUROPEANA eingesetzten kontrollierten Vokabulare liegen beispielsweise bereits alle in skosifizierter Form vor. Auf derselben Ebene der semantischen Komplexität, aber mit einem anderen Fokus, ist der Einsatz von mehr oder weniger stark formal logisch strukturierten Ordnungssystemen wie Nomenklaturen, Taxonomien oder Klassifikationen [31], [7, S. 176 ff.] einzuordnen. Hierbei kommen sowohl hierarchische als auch facettierte Ordnungssysteme [32] zum Einsatz. Dienen die Wörterbücher und Thesauri vorrangig der Disambiguierung von Wortbedeutungen und der named entity recognition, können Klassifikationen zur inhaltlichen und thematischen Filterung und Eingrenzung eingesetzt werden. Im Zusammenhang mit einer Datenvisualisierung ermöglichen hinterlegte klassifikatorische Strukturen beispielsweise ein häufig auch als Drill-down bezeichnetes Navigieren in Datenbeständen vom Allgemeinen zum Spezifischen. Eine Verbindung beider Ansätze stellt gewissermaßen die Einbeziehung von enzyklopädischen Wissensbasen wie Wikipedia dar, da hier zum einen über die Einbeziehung der Volltexte der lexikografischen Beiträge eine Bedeutungsanreicherung erreicht werden kann und zum anderen die systematischen Strukturen und Kategoriensysteme der Wikipedia genutzt werden können.

Applikationen auf der obersten Stufe des Stufenmodells bilden komplexe Beziehungen zwischen Begriffen und referenzierten Dokumenten oder Artefakten ab, die über hierarchische Beziehungen hinausgehen. Ein zunehmend wichtiger werdendes Instrument, um solche Beziehungen abzubilden, sind Ontologien. Durch den Einsatz von Ontologien kann das Wissen einer Domäne unter Verwendung einer formalen Sprache (Schema) formalisiert (damit auch maschinenlesbar) repräsentiert werden. Ontologien beschreiben Konzepte mit ihren Beziehungen innerhalb einer Wissensdomäne und unterstützen Maschinen dabei, Inhalte im Web interpretieren zu können [23, S. 12]. Durch den Einsatz von Ontologien kann beispielsweise die Konsistenz von Faktenaussagen gegen das formale Schema geprüft werden und neue noch nicht explizit ausgedrückte Beziehungen können per Inferenz abgeleitet werden. Das Schema in Verbindung mit einer Menge von Faktenaussagen bildet eine Wissensbasis (knowledge base). Wichtiges Einsatzgebiet von Ontologien ist das automatische Schlussfolgern. Dazu gehören Aufgaben wie die Bestimmung der Klassenzugehörigkeit eines Elementes, die Ableitung von äquivalenten Klassen oder die Überprüfung der Wissensbasis auf Inkonsistenzen. Im medizinischen Bereich könnte eine Ontologie beispielsweise von

[5] Die Anwendung läuft derzeit nur im Testbetrieb und funktioniert noch nicht stabil. Clever Search. URL: http://wdok.cs.uni-magdeburg.de/clever-search/

Art und Umfang der Symptome eines Patienten auf die Art der Krankheit schlussfolgern oder entdecken, dass unterschiedliche Symptome sich auf den Einfluss ein und desselben Virus' ableiten lassen [33].

Eine weniger stark formalisierte Form der Abbildung von komplexen Beziehungen sind Topic Maps [34], ein auf XML basierendes Datenformat, mit dem Synonyme und Zusammenhänge beschrieben werden können. Im Vergleich zu einer Ontologie kann eine Topic Map jedoch keine Eigenschaften vererben und keine regelhaften Zusammenhänge ausdrücken [35, S. 7]. Der Grad der Abbildung von komplexen Beziehungen ist bei einer Topic Map folglich etwas geringer als bei Ontologien.

4. State-of-the-Art-Überblick semantischer Suchmaschinen

Die folgende Darstellung liefert einen Überblick über das derzeitige Angebot an frei über das WWW zugänglichen, semantischen Suchmaschinen. Zur Aussagekraft der Überblicksdarstellung ist einleitend einschränkend zu sagen, dass allein aufgrund der extremen Flüchtigkeit des Suchmaschinenmarktes keine Vollständigkeit im Hinblick auf die tatsächlich angebotene Anzahl von semantischen Suchmaschinen erzielt werden kann. Trotz des explorativen Charakters der Übersicht lassen sich aber auf der Grundlage von insgesamt 95 gesichteten Tools Aussagen zu typischen Eigenschaften von semantischen Suchmaschinen sowie wichtigen Entwicklungstendenzen ableiten.

4.1. Recherche und Auswahl der Suchmaschinen

Die der Übersicht zugrunde liegenden Suchmaschinen wurden auf dem Weg einer systematischen Recherche (Dezember 2010) lokalisiert. Ausgangspunkt bildete eine Internetrecherche über Google und Google Scholar nach Suchmaschinen, Studien und allgemeinen Informationen zum Thema Semantic Search, Semantische Suche, semantische Suchmaschinen. Zusätzlich wurde in informationswissenschaftlichen Fachdatenbanken (LISA, LISTA, E-LIS), Blogs (z. B. Technorati, Google Blog Search) und dem „Social Web" (z. B. Mister Wong, delicious) recherchiert. Wie erwartet, erwiesen sich gerade „Social-Web"-Anwendungen als besonders zielführend für das Aufspüren semantischer Suchmaschinen. Vor allem wegen der hohen Aktualität des Social Web lassen sich auf diesem Wege im Vergleich zu Fachdatenbanken aktuellere Daten und Informationen erheben. Weiterhin können hier Suchmaschinen recherchiert werden, die noch nicht systematisch untersucht, aber bereits von Nutzern untereinander empfohlen werden. Das Ergebnis der Recherchephase war eine Sammlung von 95 semantischen Suchmaschinen. Es ist davon auszugehen, dass mit den 95 identifizierten Suchmaschinen nur ein (nicht weiter zu quantifizierender) Teil der tatsächlich verfügbaren Suchmaschinen nachgewiesen wird. Bei der vorangegangenen Recherche wurden ausschließlich deutsche und englische Quellen genutzt und Suchmaschinen anderer Sprachen finden keine Berücksichtigung.

4.2. Kategoriensystem und Kodierhandbuch

Als Ausgangsbasis für die weitere Untersuchung wurden in einem ersten Analyseschritt die identifizierten semantischen Suchmaschinen in ein Kategoriensystem eingeordnet, das in mehreren Schritten auf der Basis eines Pretests mit einzelnen Such-

maschinen verfeinert wurde.[6] Das in der Untersuchung eingesetzte Kategoriensystem besteht aus elf Kategorien, die bis zu neun verschiedene Ausprägungen haben (vgl. Tabelle 1). Mithilfe der verwendeten Merkmale lassen sich die Suchmaschinen sowohl hinsichtlich ihres inhaltlichen Schwerpunktes und ihrer Herkunft (Anbieter) als auch in Bezug auf die zum Einsatz kommenden (semantischen) Technologien beschreiben.

Tabelle 1: Kategoriensystem zur Beschreibung semantischer Suchmaschinen

Kategorie	Beschreibung
Thema	Eingrenzung des Themenbereichs, den die Suchmaschine abdeckt. Die Unterteilung erfolgt dabei in die Themenkreise „Allgemein", „Wissenschaftliche Themen" (wie Medizin, Biologie, Recht und Physik), „Arbeit" und „Sonstiges".
Anbieter	Unterschieden wird zwischen den Ausprägungen kommerziell/gewerblich, nicht kommerziell/Forschung und keine Angaben.
Typ	Unterschieden wird zwischen den beiden Ausprägungen Stand-alone-Suchmaschine und Metasuchmaschine. Eine Stand-alone-Suchmaschine baut einen eigenen Index auf, während eine Metasuchmaschine die Suchanfrage an andere Suchmaschinen weiterleitet.
Umfang	Beschreibt den Umfang der Wissensbasis, den die jeweilige Suchmaschine abdeckt. Entweder wird mit der Suchmaschine nur eine begrenzte Wissensbasis durchsucht oder sie erhebt den Anspruch, das gesamte WWW zu berücksichtigen. Die Kategorie Umfang ist mit der Kategorie Wissensbasen verknüpft. Hat eine Suchmaschine eine begrenzte Wissensbasis, so wird in der folgenden Kategorie die verwendete Wissensbasis angegeben.
Wissensbasis	Es werden verschiedene Ausprägungen der Wissensbasen (z. B. Wikipedia, PubMed) beschrieben, auf die die Suchmaschinen zugreifen.
Wissensorganisation	Kodiert werden die Methoden, mit denen die jeweilige Suchmaschine ihre Daten organisiert: Klassifikation, Ontologie, Thesaurus/SKOS, Taxonomie, Wörterbuch, Lexikon, Topic Map, Enzyklopädie sowie Metadaten.
Sichtbarkeit der Semantik	Es wird ausgewertet, ob die semantischen Instrumente, die von den einzelnen Suchmaschinen angewendet werden, für den User sichtbar bzw. nachvollziehbar sind oder nicht. Diese Kategorie steht dabei in enger Verbindung mit der Kategorie Optimierung der Suchanfrage. Handelt es sich bei der Optimierung um Funktionen, die eine Interaktion des Users erfordern und den Rückgriff auf z.B. ein kontrolliertes Vokabular vermuten lassen, so wird dadurch die Semantik für den User sichtbar.
Optimierung der Suchanfrage	Verfahren und Möglichkeiten, wie Suchanfragen durch Interaktionen des Nutzers mit dem System optimiert werden können. Ausprägungen dieser Kategorie sind die Autovervollständigung der Suchanfrage, die Möglichkeit zur Filterung der Suchergebnisse, die Autokorrektur der Rechtschreibung, der Einsatz einer Tagcloud, der Synonym-/Homonymabgleich, die Recherche mit ähnlichen und verwandten Begriffen, die Recherche in Kategorien oder auch Feedbackrückfragen bzw. Bewertungsmöglichkeiten durch die Suchmaschine. Falls keine der oben genannten Optimierungen stattfindet, wird die Ausprägung „es findet keine Optimierung statt" kodiert.
Lernfähigkeit	Es wird festgehalten, ob die Suchmaschine sich die Nutzeraktionen einer Session ‚merkt' und für die weitere Interaktion nutzt. Ein Beispiel für die Lernfähigkeit: Wird in einer ersten Suche mit dem Begriff „Golf" recherchiert und der User entscheidet sich dabei für die Sportart Golf, sollten innerhalb einer Sitzung bei der erneuten Recherche nach „Golf" direkt die Treffer zur Sportart Golf angezeigt werden. Wenn dies der Fall ist, kann die Suchmaschine als lernfähig bezeichnet werden. Lernfähigkeit wird hier im Sinne einer Art session-basierter Suche verstanden.

[6] Grundlage des Kategoriensystems bilden die im Forschungsüberblick bereits genannten Fachartikel und Studien. So hatten Hildebrand u. a. bereits 2007 für ihre Analyse suchebasierter Nutzerinteraktionen im Semantic Web einen Kriterienkatalog entwickelt, auf den für die vorliegende Studie zurückgegriffen werden konnte. Ergänzende Kriterien wurden aus weiteren Studien abgeleitet. Siehe [36] [16] [2] [17].

Kategorie	Beschreibung
Verfahren	Es wird festgehalten, welche technologischen Verfahren von den Suchmaschinen angewendet werden, um die gestellten Suchanfragen zu verarbeiten und anschließend relevante Treffer zu liefern.
Ergebnis-darstellung	Es wird festgehalten, wie die Ergebnisse der Suchanfragen dem User präsentiert werden. Ausprägungen in dieser Kategorie sind das Anzeigen der Treffer in einer einfachen Trefferliste mit Snippets[7], das Anzeigen der exakten Antwort auf die Suchanfrage mithilfe von Information Extraction [37, S. 15], die Anzeige der Treffer mithilfe von Passage Retrieval, das Clustern der Ergebnisse, das Anzeigen des Kontextes oder auch die Visualisierung der Antworten beispielsweise über eine Tagcloud [38, S. 71].

Da die Identifizierung der eingesetzten Verfahren ein Kernelement der Untersuchung bildet, werden im Folgenden die einzelnen Verfahren bzw. Ausprägungen dieser Kategorie einzeln beschrieben.

Linktopologische Verfahren
Ausgangspunkt linktopologischer Verfahren ist die Linkstruktur des Internet und die Annahme, dass sich daraus eine Bewertung von Dokumenten ableiten lässt. Es wird folglich angenommen, dass Verlinkungen nicht zufällig gesetzt werden, sondern ein Kriterium für die Wertigkeit eines Dokuments sind, weil auf dieses verwiesen wird [40, S. 117].

Computerlinguistische Verfahren
Diese beschreiben eine Identifikation von Indextermen auf Basis einer vorherigen linguistischen Analyse. Voraussetzung ist eine zuverlässige Worterkennung und ein homogener Dokumentbestand. Unterschieden werden diese Verfahren nach regel- und wörterbuchorientierten Verfahren [26, S. 104].

Regelbasierte Verfahren
Die linguistische Analyse wird auf Grundlage von Regeln in Form von Algorithmen durchgeführt. Ein Algorithmus ist eine Verarbeitungsvorschrift und definiert folglich die einzelnen Verarbeitungsschritte. Regelbasierte Verfahren sind jedoch nicht so genau wie Wörterbücher, mit denen sich jeder Einzelfall erfassen lässt [26, S. 104 f.].

Wörterbuchbasierte Verfahren
Die linguistischen Analysen werden auf Grundlage eines hinterlegten Wörterbuchs durchgeführt und sind damit Einzelfalllösungen. Die Behandlung von Komposita lässt sich durch Wörterbücher einfacher als durch das oben beschriebene regelbasierte Verfahren realisieren [26, S. 106].

Statistische und probabilistische Verfahren
Statistische Verfahren beschreiben Systeme, die auf Termgewichtungen beruhen. Hierbei findet ein Vergleich der Repräsentation von Frage- und Dokumentinhalt statt, der die Ausgabe gewichteter Treffer zur Folge hat. Die Gewichtung der Terme hängt dabei maßgeblich von der vom System geschätzten Wahrscheinlichkeit ab, inwiefern die Dokumenteinheit für die Suchanfrage relevant ist [26, S.106].

[7] Snippets sind kurze Beschreibungen der Treffer [39, S. 137].

Textmining
Die genannten computerlinguistischen und statistischen Verfahren werden auch unter der Bezeichnung Textmining für Algorithmus-basierte Verfahren zur Entdeckung von neuen, bisher nicht bekannten Informationen durch automatische Informationsextraktion aus un- oder schwach strukturierten Textdaten genutzt [41]. Hearst hatte 2003 noch einen grundsätzlichen Unterschied zwischen Textmining und Suche konstatiert, da der Nutzer bei der Suche nach etwas suche, das bereits bekannt sei und über das schon jemand geschrieben habe und die Anforderung der Suche gerade darin bestehe, für den gegenwärtigen Informationsbedarf irrelevante Informationen auszuschließen [42]. Textmining zielt hingegen eher darauf ab, auch Informationen anzubieten, nach denen nicht gesucht wurde, die aber relevant sein könnten. Ein klassisches Beispiel aus dem E-Commerce ist, dass einem Kunden bei der Suche nach Taschenlampen auch Informationen zu Batterien angeboten werden.

Named Entity Recognition
Named Entity Recognition (NER) ist ein Verfahren zur Erkennung von Bestandteilen (Named Entities) innerhalb eines unstrukturierten, in natürlicher Sprache verfassten Dokuments und dessen Zuordnung in eine Kategorie. Eine Entität bezeichnet dabei eine konkrete Ausprägung eines beliebigen Konzepts. Zum Beispiel ist Angela Merkel eine Entität des Typs Person. Die Named Entity Recognition setzt auf wörterbuchbasierte, regelbasierte oder statistische Verfahren auf. Das einfache Verfahren verwendet eine Wortdatenbank, welche alle relevanten Wörter und Phrasen enthält und ein Set von Regeln, wie das System verfahren soll, wenn die Entitäten im Volltext erkannt werden. Regelbasierte Verfahren erkennen Entitäten auf Basis von Heuristiken und linguistischem Wissen. Statistische Verfahren „lernen“ Entitäten zu erkennen und zu klassifizieren [43, S. 1].

RDF/OWL
Das Resource Description Framework (RDF) bezeichnet Standards des World Wide Web Consortiums (W3C) zur formalen Beschreibung von Informationen über Objekte, sogenannte Ressourcen, die durch eindeutige Bezeichner (URI – Uniform Resource Indentifier) identifiziert werden. URLs (URL – Uniform Resource Locator) sind eine Spezialform der URIs, mit der Webressourcen eindeutig benannt werden. Auf diese Weise ließe sich z. B. die im Anfangsbeispiel genannte Sportart Golf von dem Automodell Golf eindeutig dadurch unterscheiden, dass für jede der beiden Bedeutungen auf eine URL, unter der eine Website, auf der eine Begriffsdefinition zu finden ist, verwiesen wird. RDF ist ein Datenmodell, das einfache Aussagen über Sachverhalte in der Syntax Subjekt, Prädikat, Objekt (den sogenannten Triples; Für die linearisierte Darstellung hat sich der Begriff Graph etabliert) ermöglicht. In RDF-Schema (RDF(s) werden die Bedeutungen der einzelnen Sprachelemente von RDF eindeutig definiert. SPARQL ist eine Abfragesprache für Statements, die in RDF formuliert sind. SPARQL kann eingesetzt werden, um Abfragen über verschiedenen Datenquellen durchzuführen, unabhängig davon, ob sie lokal als RDF gespeichert sind oder ob über eine Middleware eine RDF-Sicht erstellt wird. Bei OWL (Web Ontology Language) handelt es sich um eine Sprache, mit der Ontologien (siehe Wissensorganisation) im Web formal repräsentiert werden können. Mit der Formulierung als W3C Recommendation im Jahre 2004 [44] [45] hat sich OWL als Standard durchgesetzt.

4.3. Kodierung

Auf Grundlage des Kodierhandbuches wurden 63 nach der Datenbereinigung (z. B. Ausschluss nicht (mehr) funktionierender Suchmaschinen und reiner Demoversionen) verbliebene Suchmaschinen evaluiert.[8] Für die Analyse wurden folgende Quellen zugrunde gelegt:

1. Metaanalyse von Begleitmaterialien und Literatur: Öffentlich zugängliche Informationen, die auf der Homepage der jeweiligen Suchmaschine (z. B. „About us"), in Handbüchern, Whitepapern und Dokumentationen sowie Sekundärquellen recherchiert wurden. Eine Aussage gilt erst dann als „gesichert", wenn sie explizit aus einer der aufgeführten Quellen nachgewiesen werden konnte. Auf diese Weise können Fehler oder Fehleinschätzungen vermieden bzw. vermindert werden, da z.B. die verwendeten Technologien einer Suchmaschine oftmals nicht eindeutig erkennbar bzw. nachvollziehbar waren und nicht eindeutig ermittelt werden konnten. Ergänzend wurden v. a. zur Erfragung der verwendeten Technologien einige Suchmaschinenanbieter per E-Mail kontaktiert. Das Ziel bestand darin, Informationen zu erhalten, die nicht aus dem Web entnommen werden konnten, jedoch für die Evaluation notwendig waren. Auf Basis der Metaanalyse wurden die Angaben zu den Kategorien „Anbieter", „Thema", „Typ", „Umfang", „Wissensbasen", „Wissensorganisation" und „Verfahren" ermittelt. Es wurden nur dann Angaben übernommen, wenn die Informationen ausdrücklich aus den o. g. Quellen zu ermitteln waren. Andernfalls wurde für die jeweilige Kategorie die Ausprägung „keine Angabe" gewählt.
2. Für die Kategorien „Sichtbarkeit der Semantik", „Optimierung der Suchanfrage", „Lerneffekt" und „Ergebnisdarstellung" wurden die jeweils zutreffenden Ausprägungen anhand von Testanfragen[9] ermittelt.

Als schwierig erwies sich vor allem die Datenerhebung zu den verwendeten Technologien. Sehr umfangreichen und detaillierten Texten standen kryptische oder vor allem für Marketingzwecke erstellte Texte gegenüber. Insgesamt waren nur wenige verlässliche Informationen zu den verwendeten Technologien zugänglich. Das Anschreiben der Suchmaschinenanbieter per E-Mail mit Bitte um weitere Informationen, v. a. zu den verwendeten Technologien, ergab nur wenige Rückmeldungen.

5. Eigenschaften semantischer Suchmaschinen – Ergebnisse der empirischen Analyse

Von ursprünglich 95 Suchmaschinen konnten 63 Suchmaschinen bei der Kodierung berücksichtigt werden.[10] Thematisch konzentrieren sich die Suchmaschinen auf zwei

[8] Die Evaluierung wurde von fünf Studentinnen des Masterstudiengangs Informationswissenschaft und -management an der HAW Hamburg durchgeführt. Methodisch wurde analog zur in der Usability-Evaluation häufig angewandten heuristischen Evaluation durch Usability-Experten durch ständigen Austausch der Kodierer sichergestellt, dass ein einheitliches Verständnis der gewählten Kategorien besteht.

[9] Bei deutsch- und englischsprachigen Begriffen wurden beliebige Suchanfragen gewählt und mit bewussten Tippfehlern eingegeben (z. B. „Angela Merckel" oder „Michael Jakson").Tag-Cluod: Je nach Suchmaschine wurde ein beliebiger Suchbegriff eingegeben. Synonymabgleich: „Karotte – Möhre" (bei deutschsprachigen Suchmaschinen), „student" – „pupil" (bei englischsprachigen Suchmaschinen), „flu" – „influenza" (bei medizinischen Suchmaschinen). Homonymabgleich: „Bank" (bei deutschsprachigen Suchmaschinen), „bow" (bei englischsprachigen Suchmaschinen). Feedbackanfragen: Je nach Suchmaschine wurde ein beliebiger Suchbegriff eingegeben.

Bereiche. Mit 29 Nennungen handelt es sich bei den meisten Angeboten um allgemeine Suchmaschinen, die ähnlich wie Google, Yahoo oder MSN universell ausgerichtet sind und nicht auf einzelne Themengebiete begrenzt sind. Den zweitgrößten Block bilden mit 21 Nennungen Suchmaschinen aus dem wissenschaftlichen Bereich, wobei hier ein deutlicher Schwerpunkt auf medizinischen und biologischen Themen liegt. Bei den übrigen Suchmaschinen handelt es sich um Spezialsuchmaschinen zu eng begrenzten Einzelthemen wie Kochen, Musik und Arbeitsplatzsuche.

48 Suchmaschinen werden von kommerziellen Anbietern und 13 von nicht-kommerziellen Anbietern bereitgestellt. Bei zwei Suchmaschinen konnten keine Angaben über den Anbieter gefunden werden. Die kommerziellen Angebote überwiegen damit deutlich. Bei den kommerziellen Angeboten handelt es sich häufig um Angebote von Softwareunternehmen, die die Webpräsenz als eine Art Produktschaufenster nutzen (Carrot Search, Hakia); fachspezifische Angebote im wissenschaftlichen Bereich lassen sich nicht selten auf Wissenschaftsverlage zurückführen, die die Suchtechnologien nutzen, um ihre Publikationen/Fachzeitschriften zu platzieren (Deepdyve). Alle drei nicht kommerziellen, allgemeinen Suchmaschinen (Sindice, Swoogle und SWSE) haben einen universitären Hintergrund und haben ebenfalls, wie aus den begleitenden Webauftritten ablesbar ist, die Funktion, öffentlichkeitswirksam auf die entsprechenden Forschungsprojekte im Zusammenhang mit Forschung zum Semantic Web hinzuweisen.

5.1. Typ, Umfang und Lernfähigkeit der Suchmaschinen

Mit 33 Suchmaschinen handelt es sich bei der Mehrzahl der Suchmaschinen um Metasuchmaschinen, die auf die Indices von anderen Suchmaschinen wie Google oder Yahoo zugreifen. 27 Suchmaschinen hingegen sind sogenannte Stand-alone-Suchmaschinen. Drei weitere sind eine Kombination aus Stand-alone- und Metasuchmaschine, da sie sowohl eine eigene Suche als auch den Zugriff auf andere Suchmaschinen anbieten. Setzt man die Untersuchungskriterien „Suchmaschinentyp“ und „Thema“ in Beziehung zueinander, fällt nur im Bereich Medizin/Biologie auf, dass es sich hier mit 12 Nennungen überwiegend um Meta-Suchmaschinen handelt. Bei den übrigen Themenkreisen ist das Verhältnis fast ausgewogen. Dies lässt sich damit erklären, dass für den Bereich Medizin bereits auf gut aufbereitete Datenbestände zugegriffen werden kann und zudem gut strukturierte Datenbestände wie die bibliografische Datenbank Medline und kontrollierte Vokabulare wie MeSH (Medical Subject Headings) genutzt werden können. Ein Beispiel für eine solche Anwendung ist die von der National Library of Medicine angebotene, puristisch auftretende Suche ask Medline über die Datenbanken Medline und PubMed.

Knapp die Hälfte der analysierten Suchmaschinen verwendet lediglich eine begrenzte Wissensbasis. 19 Anbieter geben als Datenbasis ausschließlich das World Wide Web allgemein an (WWW), davon zählen 15 zu den allgemeinen Suchmaschinen, und 3 nennen zusätzlich noch eine spezielle Wissensbasis. Mit semantischen Suchlösungen wird also offensichtlich sowohl für begrenzte als auch sehr offene Wissensbasen experimentiert. Bezogen auf den Suchmaschinentyp besteht bei den Metasuchmaschinen ebenfalls ein ausgewogenes Bild hinsichtlich der Nutzung von begrenzten Wissensbasen und des WWW. Bei den Stand-alone-Suchmaschinen haben 8 Suchmaschinen

[10] Gründe für das Herausnehmen von Suchmaschinen aus der Untersuchung sind u. a., dass die Suchmaschinen zur Zeit der Datenerhebung nicht funktionsfähig oder nicht uneingeschränkt öffentlich zugänglich waren (nicht funktionierende Registrierung) oder keine ausreichenden Informationen etwa zu den angewandten Verfahren ermittelbar waren.

den Anspruch, das gesamte WWW abzudecken, während 17 dieser Suchmaschinen eine begrenzte Wissensbasis nutzen. Stellt man die Ergebnisse zum Umfang und den Themen der Suchmaschinen gegenüber, werden deutliche Unterschiede sichtbar. Für 18 von 29 allgemeinen Suchmaschinen dient das WWW als auszuwertender Datenbestand, 2 weitere verwenden eine Kombination. Fast allen der übrigen 34 Suchmaschinen, die sich ausschließlich auf einen Themenkomplex beziehen (sei es Medizin, Arbeit oder Musik), liegen nur begrenzte Wissensbasen zugrunde. Nur 5 Suchmaschinen indexieren ausschließlich das WWW. Dieses Ergebnis ist nicht überraschend, da bei einer begrenzten Wissensbasis einfacher sichergestellt werden kann, dass nur zum Hauptthema passende Treffer ausgegeben werden. Zudem kann so bereits vor der eigentlichen Anfrage die Qualität bzw. der Inhalt des Datenbestands kontrolliert werden.

5.2. Lernfähigkeit der Suchmaschinen und Sichtbarkeit der Semantik

Bei den meisten Suchmaschinen ist kein Lerneffekt sichtbar. Lediglich bei 6 Suchmaschinen kann man davon ausgehen, dass sie sich vorangegangene Suchanfragen und Suchwege merken und als Grundlage für nachfolgende Suchanfragen nutzen. Sessionbasierte Suchen scheinen somit kaum eine Rolle zu spielen. Vielmehr setzen Funktionen wie Homonymabgleich (ist die Sportart Golf oder das Automodell Golf gemeint?) oder die Rechtschreibkorrektur eine Interaktion zwischen User und Suchmaschine voraus. Bei 51 der 63 Suchmaschinen wird die Semantik für den User zumindest teilweise sichtbar, da das System mit ihm in eine Interaktion tritt. Dies kann sowohl durch Funktionen wie „Meinten Sie …“ oder das Angebot von geclusterten Trefferanzeigen oder weiteren Sortierfunktionen erfüllt werden. Allerdings konnte anhand der Testanfragen nicht eindeutig festgestellt werden, durch welche Verfahren jeweils Vorschläge oder Cluster generiert werden. Die Entwicklung hin zu Suchmasken, die eine verstärkte Interaktion mit dem Nutzer fördern, ist im Übrigen ein allgemeiner Trend bei der Gestaltung von Userinterfaces von Suchmaschinen. So bietet beispielsweise Google in der letzten Zeit dem Nutzer verstärkt Manipulationsmöglichkeiten der durch die Einstiegsrecherche erzielten Trefferliste (Abbildungen 3 und 4).

Abbildung 3: Sucheingabe Google

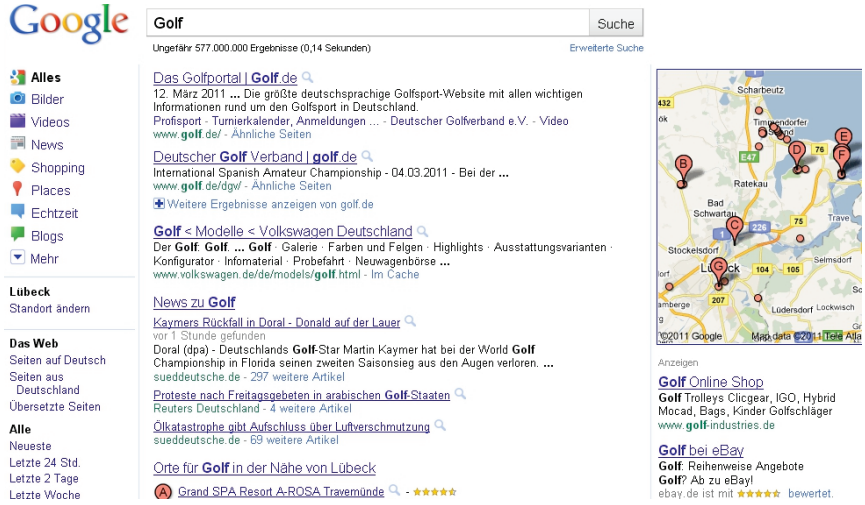

Abbildung 4: Nach Durchführung der Suche angebotene Filteroption bei Google

5.3. Optimierung der Suchanfrage und Ergebnisdarstellung

Bei den 63 analysierten Suchmaschinen werden nur bei zwei Suchmaschinen keine Möglichkeiten zur Optimierung der Suchanfrage angeboten.

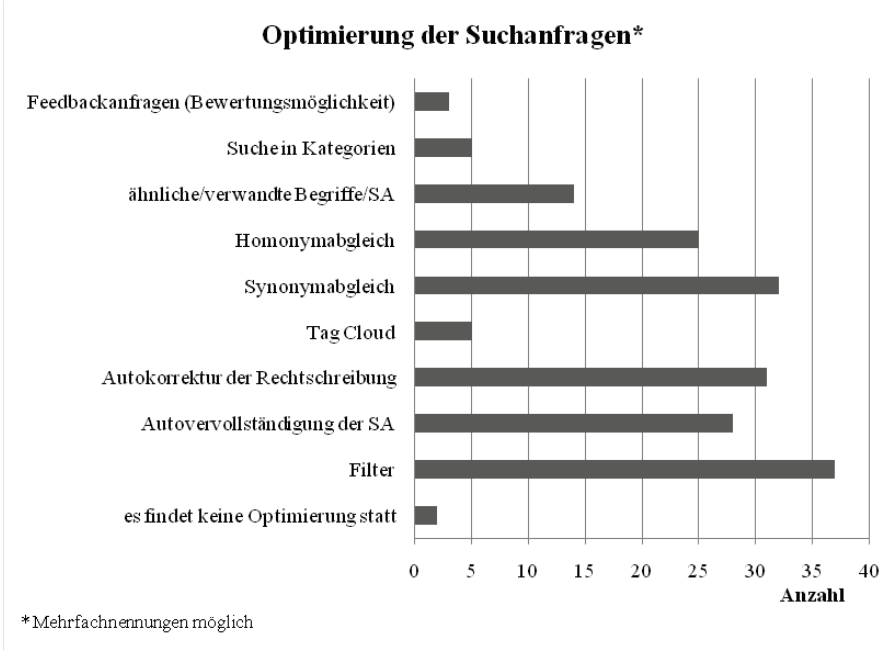

Abbildung 5: Optionen zur Optimierung von Suchanfragen

Am häufigsten werden dazu Filter eingesetzt (37 Nennungen). Danach folgen mit 32 und 25 Nennungen der Synonym- und Homonymabgleich, die Autokorrektur bei Rechtschreibfehlern mit 31 Nennungen und mit 28 Nennungen die Autovervollständigung der Suchanfrage (vgl. Abbildung 5). Hierbei muss in der Regel sowohl beim Filtern als auch bei der Autokorrektur und -vervollständigung der User nach Eingabe der Suchanfrage noch einmal aktiv werden, indem er seine Anfrage konkretisiert.

Nur fünf von 61 Suchmaschinen verwenden ausschließlich eine Methode zur Optimierung von Suchanfragen. Bei den übrigen finden sich unterschiedliche Kombinationen von 2 oder mehr Methoden. Auffällig ist, dass in den 3 häufigsten und 5 weiteren Kombinationen Filter eingesetzt werden.

Bei der Darstellung der Treffer fällt auf, dass die meisten der semantischen Suchmaschinen weiterhin auf die klassische Trefferliste setzen, bei der jeder Treffer durch einen Link zur Quelle und einen Snippet beschrieben wird. Bei den 63 Suchmaschinen ist diese Art der Ergebnisdarstellung 43-mal zu finden. Zudem werden 21-mal Cluster als Darstellungsform angeboten und 17-mal andere Formen der Visualisierung wie die Darstellung eines Treffers innerhalb eines semantischen Netzes (Beispiel EyePlorer; Abbildung 6) verwendet.

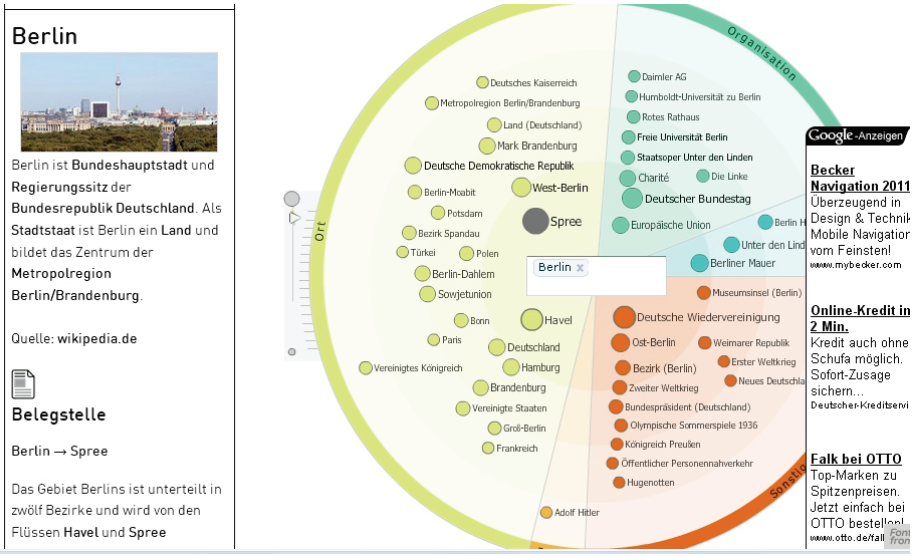

Abbildung 6: Ergebnisdarstellung als semantisches Netz bei EyePlorer.com

Auffallend ist, dass Methoden wie Information Extraction (11 Nennungen) und Passage Retrieval (9 Nennungen) nur selten genutzt werden, um exakte Antworten zu geben oder relevante Informationen aus verschiedenen Quellen zusammenzufassen. Insgesamt unterscheidet sich die Ergebnisdarstellung der untersuchten semantischen Suchmaschinen nicht wesentlich von den von klassischen Suchmaschinen bekannten Verfahren. Eine Ergebnisdarstellung in Form einer aggregierten Inhaltszusammenstellung aus Medienarten und Fakteninformationen, wie sie von der Suchmaschine Kosmix (Abbildung 7) gewählt wird, ist auch bei den semantischen Suchmaschinen eher die Ausnahme.

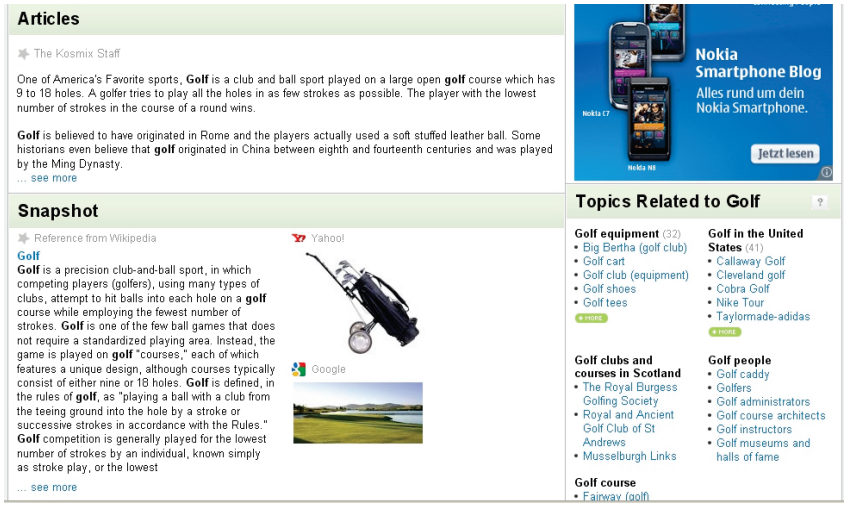

Abbildung 7: Aggregierte Ergebnisdarstellung bei Kosmix

33 Suchmaschinen nutzen zur Ergebnisdarstellung eine Kombination verschiedener Methoden. Hierbei fällt auf, dass man bei 23 Suchmaschinen auf eine Kombination in Zusammenhang mit einer Trefferliste setzt (Abbildung 8). Für 19 Suchmaschinen werden in der Kombination Cluster genutzt. Für die detaillierte Betrachtung der Kombinationen wird im Folgenden das gemeinsame Auftreten von zwei Darstellungsformen ausgewertet.

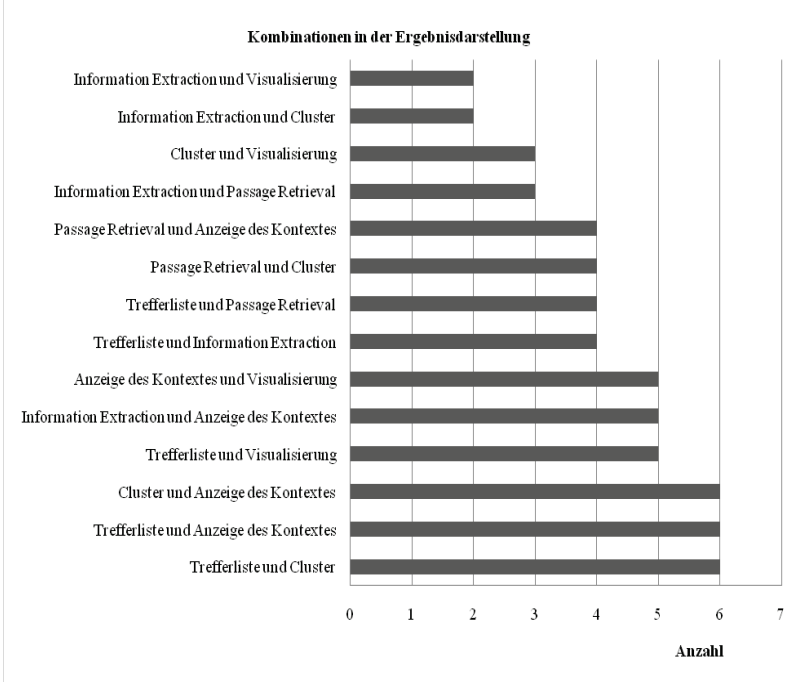

Abbildung 8: Kombinationen der Ergebnisdarstellung

Die drei am häufigsten vorgefundenen Kombinationen sind die Anzeige von Trefferliste, Cluster und Anzeige des Kontextes. Danach folgt die Kombination aus Trefferliste und Visualisierung sowie Anzeige des Kontextes und Visualisierung. Die Mehrzahl der untersuchten Suchmaschinen behält die Anzeige in Form der „klassischen“ Trefferliste bei und versucht diese durch Darstellungselemente wie Cluster und Kontextanzeige aussagekräftiger zu gestalten.

5.4. Wissensbasen, Datenaufbereitung und Verfahren

Wie bereits erwähnt, liegt gut der Hälfte der Suchmaschinen eine begrenzte Wissensbasis zugrunde. Daher wurde zusätzlich erfasst, um welche Wissensbasen es sich handelt. Hierbei ist es natürlich möglich, dass eine Suchmaschine verschiedene Datenbestände nutzt. Aufgrund der hohen Anzahl an unterschiedlichen Wissensbasen werden diese in Kategorien zusammengefasst.

Mit 19 Nennungen werten die meisten Suchmaschinen das WWW aus (s. Abbildung 9). Insgesamt zehnmal wird das WWW zudem für ein fokussiertes Webcrawling genutzt. Allerdings sind nach dem WWW zunächst Datenbestände anderer Suchmaschinen und fachspezifischer Datenbanken mit 15 und 12 Nennungen die wichtigsten Datengrundlagen. Nur 9 Suchmaschinen nutzen einen eigenen Datenbestand.

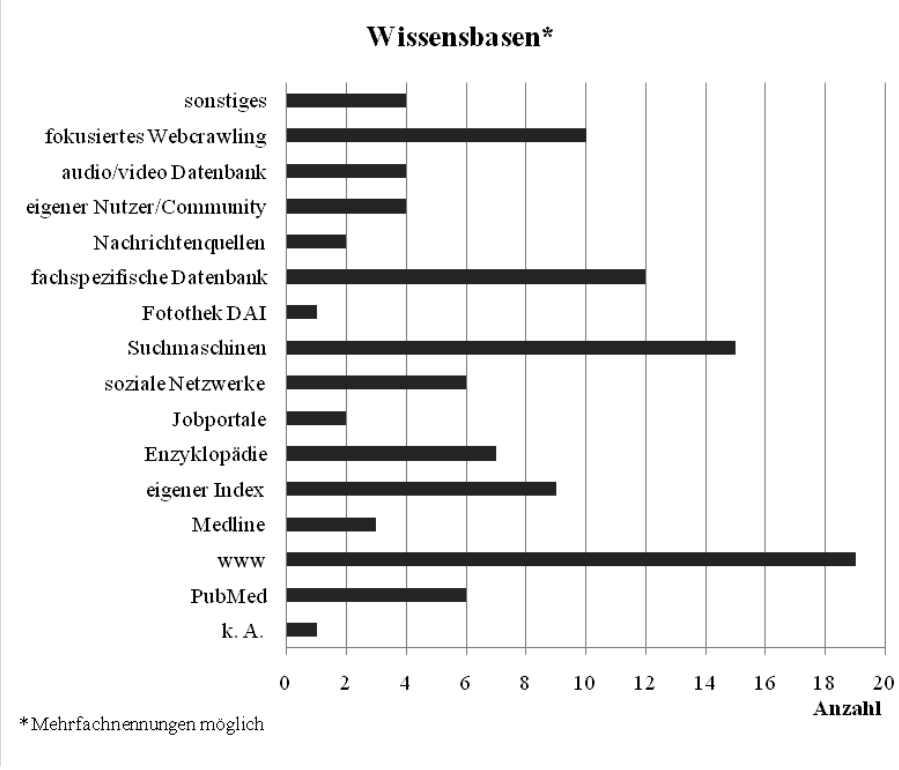

Abbildung 9: Genutzte Wissensbasen

Bezogen auf die Thematik der Suchmaschinen wird deutlich, dass das WWW am häufigsten als Wissensbasis für allgemeine Suchmaschinen genutzt wird. Fokussiertes Webcrawling wird hingegen häufig für Suchmaschinen verwendet, in denen nur Daten zu speziellen Themen wie Musik oder Hotels indexiert werden. Die wissenschaftlichen Suchmaschinen verwenden überwiegend fachspezifische Datenbanken als Wissensbasis, so z. B. PubMed in der Medizin. Das Ergebnis liefert ein weiteres Indiz dafür, dass semantische Suchmaschinen derzeit vorrangig entwickelt werden, um bereits aufbereitete Datenbestände effektiver und zielgruppengerechter auswerten zu können. Dies würde den hohen Anteil an wissenschaftlichen Suchmaschinen erklären, da für diese Wissenszweige in der Regel bereits umfassende Datenbestände, die zudem über bibliografische Metadaten und kontrollierte Vokabulare erschlossen sind, verfügbar sind.

In einem weiteren Schritt wird die Aufbereitung der Daten betrachtet (Abbildung 10). Allerdings gilt an dieser Stelle die Einschränkung, dass nicht für alle Suchmaschinen die Art der Datenaufbereitung erfasst werden konnte. Gesicherte Angaben zur Datenaufbereitung konnten nur bei 37 Suchmaschinen gewonnen werden. Mit 15 und 14 Nennungen werden bei semantischen Suchmaschinen am häufigsten Ontologien und Thesauri genannt. Allerdings ohne deutlichen Vorsprung gegenüber Klassifikationen und Taxonomien, die 10- und 9-mal kodiert wurden. Dass die Auswertung von Metadaten für nur 4 Suchmaschinen explizit als Methode zur Datenaufbereitung aufgeführt wird, ist vermutlich damit zu begründen, dass dies bei einer umfassenderen Auswertung lediglich als Teil eines anderen Verfahrens, beispielsweise ontologiebasierter Verfahren, verstanden wird oder so selbstverständlich ist, dass es nicht mehr eigens erwähnt wird.

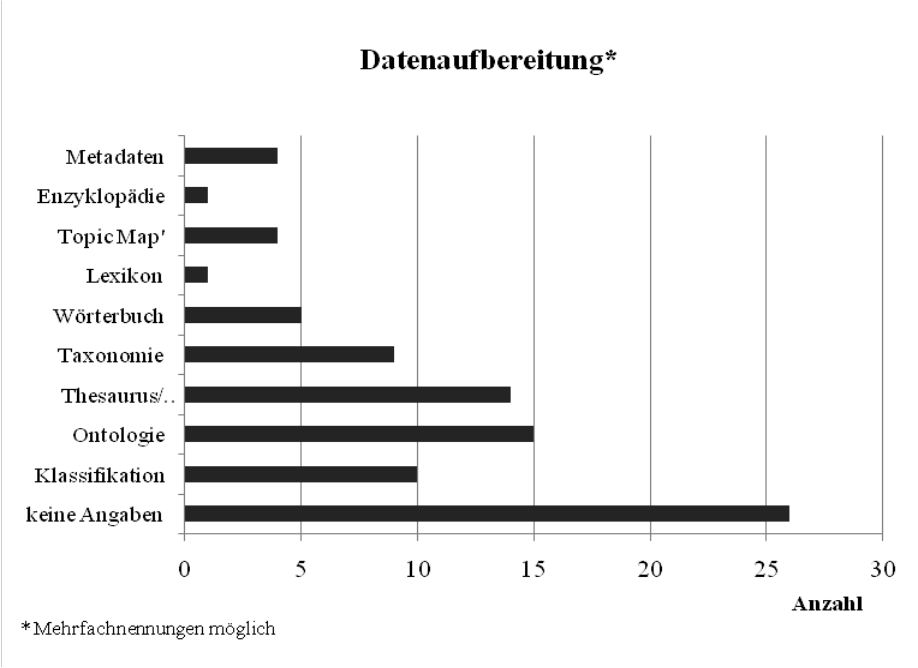

Abbildung 10: Formen der Datenaufbereitung in semantischen Suchmaschinen

Bei 17 Suchmaschinen konnte nur eine Form der Datenaufbereitung sicher bestimmt werden, bei den übrigen 20 Suchmaschinen wurde eine Kombination aus mindestens zwei Verfahren festgestellt. Allerdings verwenden von diesen 20 Suchmaschinen nur 4 mehr als zwei Arten der Datenaufbereitung. Betrachtet man die auftretenden Kombinationen genauer, wird deutlich, dass in der Regel zwei Methoden miteinander verknüpft werden (s. Tabelle 2). Während eine Methode zur Abbildung von Relationen dient, liefert die andere das kontrollierte Vokabular. Dies ist sowohl bei einer Kombination aus Ontologie und Thesaurus (5 Nennungen) als auch bei einer Kombination aus Klassifikation und Thesaurus (4 Nennungen) der Fall.

Tabelle 2: Kombinationen in der Datenaufbereitung, die mindestens zweimal erfasst wurden

Häufigkeit	Kombination
5	Ontologie und Thesaurus
4	Klassifikation und Thesaurus
3	Klassifikation und Taxonomie
3	Ontologie und Taxonomie
3	Thesaurus und Taxonomie
2	Klassifikation und Ontologie
2	Thesaurus und Wörterbuch
2	Thesaurus und Metadaten

Neben der Datenaufbereitung wurde bei der Analyse auch berücksichtigt, welche Verfahren zur Generierung der Suchergebnisse eingesetzt werden (Abbildung 11).

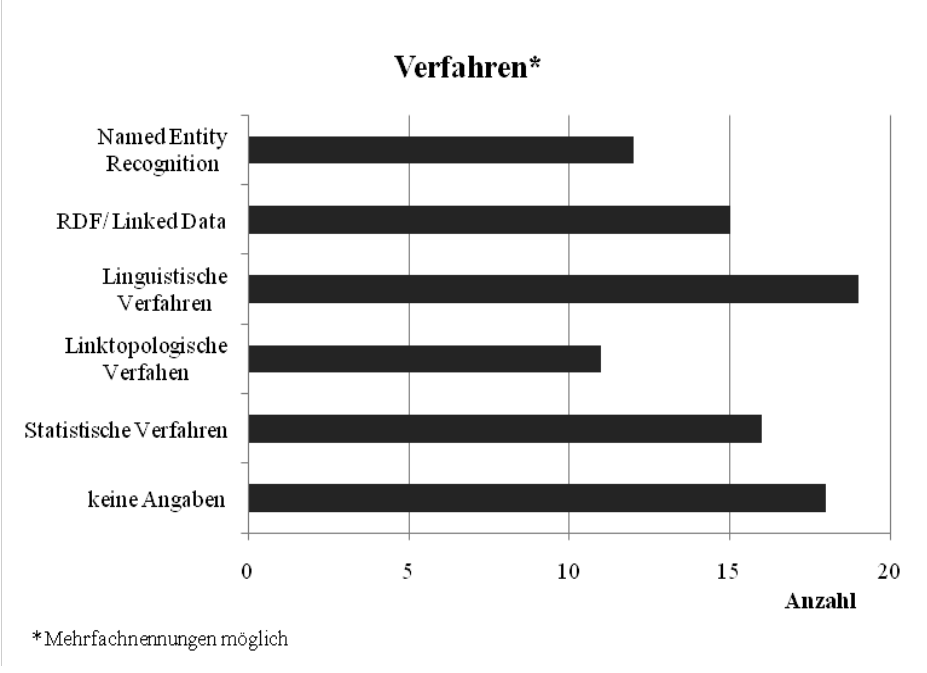

Abbildung 11: Für die Suche angewandte Verfahren

Die folgenden Auswertungen beziehen sich auf die Analyse von 45 Suchmaschinen, da nicht für alle Suchmaschinen die Verfahren ermittelt werden konnten. Natürlich ist es auch hier möglich, dass eine Suchmaschine mehrere Verfahren einsetzt. Betrachtet man die kodierten Verfahren im Einzelnen, so wird mit 19 Nennungen am häufigsten auf linguistische Verfahren zurückgegriffen. An zweiter Stelle werden bei 16 Suchmaschinen die statistischen Verfahren eingesetzt und erst dann folgt mit 15 Nennungen der Einsatz von RDF bzw. Linked Data. Named Entity Recognition und linktopologische Verfahren werden mit den Werten 12 und 11 fast gleich oft verwendet.

Da statistische Verfahren jedoch auch für nicht-semantische Suchmaschinen die wichtigsten Verfahren darstellen, erscheint es sinnvoll, sich die Kombination der ahren näher anzusehen, auch wenn nur für 21 Suchmaschinen ein Verfahren mit Sicherheit dokumentiert werden konnte. Die gängige Kombination von statistischen und linktopologischen Verfahren tritt häufig auf (6-mal; s. Tabelle 3). Insgesamt lässt sich bei dieser kleinen Gruppe von semantischen Suchmaschinen jedoch kein eindeutiger Schwerpunkt ausmachen, da fast alle Verfahren miteinander kombiniert werden. Auch wenn man die Verfahren in Beziehung zur Datenaufbereitung setzt, lassen sich keine Tendenzen ausmachen. Dies kann aber vor allem darauf zurückgeführt werden, dass diese Zusammenhänge auf Basis der Selbstauskunft der Suchmaschinenbetreiber getätigt wurden und lediglich bei 28 Suchmaschinen eindeutige Angaben zu den Verfahren identifiziert werden konnten. Mit neunmaligem Vorkommen ist nur eine Verbindung zwischen der Datenaufbereitung durch eine Ontologie und RDF-basierten Verfahren bzw. Linked Data auffällig. Gleiches gilt für die Verknüpfung von Ontologie und linguistischen Verfahren. Für die übrigen Arten der Datenaufbereitung ergibt sich, bezogen auf die Verfahren, ein ausgewogenes Bild.

Tabelle 3: Kombinationen der Verfahren, die mindestens zweimal erfasst wurden

Häufigkeit	Kombination
6	Statistische und Linktopologische Verfahren
5	Linktopologische Verfahren und Named Entity Recognition
5	Statistische und Linguistische Verfahren
4	Linguistische Verfahren und Named Entity Recognition
4	Linguistische Verfahren und Linked Data
3	Statistische Verfahren und Named Entity Recognition
2	Linked Data und Named Entity Recognition
2	Linktopologische Verfahren und Linked Data

Aufgrund der untersuchten Datengrundlage ist eine systematische Analyse, an welcher Stelle des Suchprozesses – etwas bei der Optimierung der Suchanfrage oder der Aufbereitung der Indices oder des Abfragealgorithmus – die semantische Analyse oder Informationsanreicherung ansetzt, nicht möglich.

6. Einordnung der Suchmaschinen in das Stufenmodell

In Bezug auf das entwickelte Stufenmodell ist es möglich, die 37 Suchmaschinen mit Angaben zu ihrer Datenaufbereitung in das Stufenmodell einzuordnen. Zunächst werden die verwendeten Maßnahmen einzeln gezählt und summiert. Wenn mehrere Maßnahmen zur Datenaufbereitung durchgeführt wurden, werden alle einzeln gezählt. Insgesamt wurden 58 Maßnahmen zur Datenaufbereitung von 37 Suchmaschinen

angegeben. Die Verwendung von Metadaten wird an dieser Stelle nicht berücksichtigt, bzw. mit 100 Prozent angenommen, da alle darüberliegenden Maßnahmen zur Datenaufbereitung die Verwendung von Metadaten voraussetzen.

Summiert ergibt dies die in Abbildung 12 dargestellte Verteilung innerhalb des Stufenmodells. Die exakten Werte sind in Tabelle 4 aufgeführt.

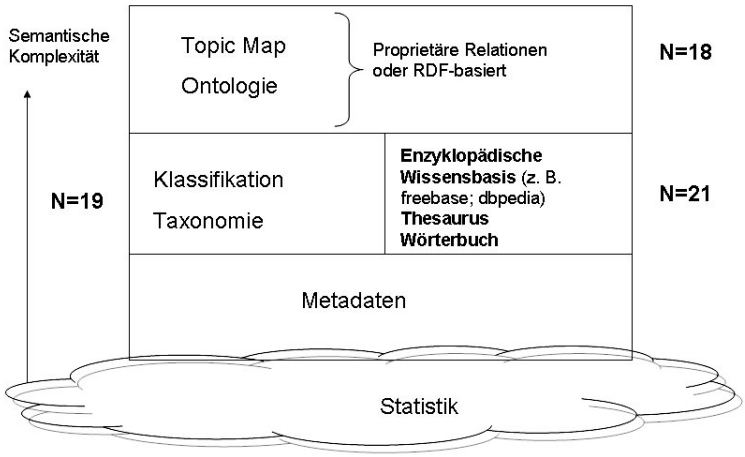

Abbildung 12: Anzahl der Maßnahmen im Stufenmodell in Prozent

Tabelle 4: Anzahl der angegebenen Maßnahmen zur Datenaufbereitung

Anzahl der Maßnahmen	Maßnahmen zur Datenaufbereitung
14	Ontologie
4	Topic Map
10	Klassifikation
9	Taxonomie
14	Thesaurus
5	Wörterbuch
2	Enzyklopädische Wissensbasis

Des Weiteren wurden die 37 Suchmaschinen in das Stufenmodell eingeordnet (Abbildung 13). Dabei wurden die Suchmaschinen anhand ihrer verwendeten Maßnahmen zur Datenaufbereitung mit dem jeweils höchsten Grad an Semantik eingeordnet. So wurde beispielsweise GoPubMed aufgrund der Verwendung eines Thesaurus und einer Ontologie in die oberste Stufe eingeordnet. Ausschlaggebend war hier die Datenaufbereitung durch eine Ontologie.

Knapp die Hälfte (18 Nennungen) der untersuchten Suchmaschinen bilden komplexe Beziehungen mithilfe von Ontologien und/oder Topic Maps ab. Gut ein Drittel der Suchmaschinen beschränkt sich auf die Nutzung einfacher hierarchischer oder systematischer Beziehungen in Form von Klassifikationen und Taxonomien. Die Erkennung von Wörtern bzw. Bedeutungen wird 8-mal explizit genannt. Die Auflistung der einzelnen Suchmaschinen je Stufe erfolgt in Tabelle 5 im Anschluss an die Visualisierung im Stufenmodell.

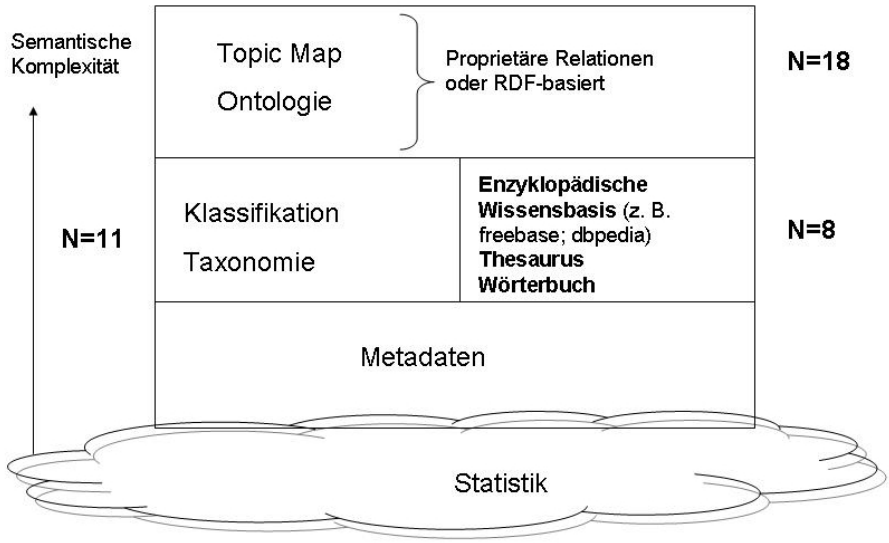

Abbildung 13: Anzahl der Suchmaschinen je Stufe

Tabelle 5: Eingruppierung der Suchmaschinen in das Stufenmodell

Stufe 1 **Abbildung komplexer Beziehungen**	**Stufe 2** **Abbildung einfacher Beziehungen**	**Stufe 3** **Erkennung von Wörtern und Bedeutungen**
Arachne	Arbeitsagentur	Bing
Cognition	DeepDyve	ClusterMed
EAGLi	DuckDuckGo	EyePlorer
Go3R	EBIMed	HealthMash
GoPubMed	Endeca	Hulbee
Hakia	Europeana	iHop
Kosmix	Healthline	Quintura.com
Nextbio	RightHealth	Releton
Pixolu	Semager	
SenseBot	TrialX	
Sindice	Yummly	
Swoogle		
TipTop		
TrueKnowledge		
TrustYou		
Viewchange		
WolframAlpha		
Yebol		

7. Drei semantische Suchmaschinen im Kurzporträt

Abschließend werden im Folgenden drei unterschiedliche Beispiele vorgestellt, die geeignet sind, die derzeit beobachtbaren Trends und Anwendungsfelder im Bereich semantische Suchmaschinen idealtypisch zu veranschaulichen:

- Bei Kosmix handelt es sich um eine algorithmische Suchmaschine, die vor allem auf avancierte Techniken der NLP setzt.
- GoPubMed ist eine domainspezifische Suchmaschine mit Schwerpunkt im medizinischen Bereich.
- Sindice gehört zum Typ der Linked-Data-Suchmaschinen, die auf Webseiten zugreifen, die als RDF-Triple aufbereitet sind bzw. in HTML eingebettete Microformate wie RDFa[11] nutzen.

7.1. Kosmix

Die semantische Suchmaschine Kosmix wurde im Jahr 2005 von Venky Harinarayan und Anand Rajaraman gegründet. Kosmix ist ein typisches Start-up- Technologieunternehmen, die Website dient beispielsweise auch zur Personal- und Kundenakquise. Wichtiges Geschäftsfeld von Kosmix ist das Angebot von technischen Lösungen zur Contentaggregation für Fachverlage [46]. Kosmix fungiert als allgemeine Metasuchmaschine, die sich nicht auf ein einzelnes Themengebiet konzentriert, sondern das gesamte Web durchsucht. In der Ergebnisanzeige von Kosmix tritt die klassische Trefferanzeige gegenüber der nach unterschiedlichen Medientypen und Informationsarten wie Presseartikel und Social-Web-Beiträgen, Videos, Informationen aus Nachschlagewerken, aufbereiteten Contentaggregation in den Hintergrund. Die Trefferanzeigen werden nach dem Prinzip der Information Extraction und des Passage Retrieval generiert. Zusätzlich werden die Treffer geclustert und deren Kontext dargestellt. Die Treffer werden von Kosmix dabei zum einen über Crawling generiert und zum anderen über eine föderierte Suche, über die Informationen aus dem Deep Web gefunden werden. Ein Ausschnitt einer Trefferliste ist in Abbildung 14 zu sehen.

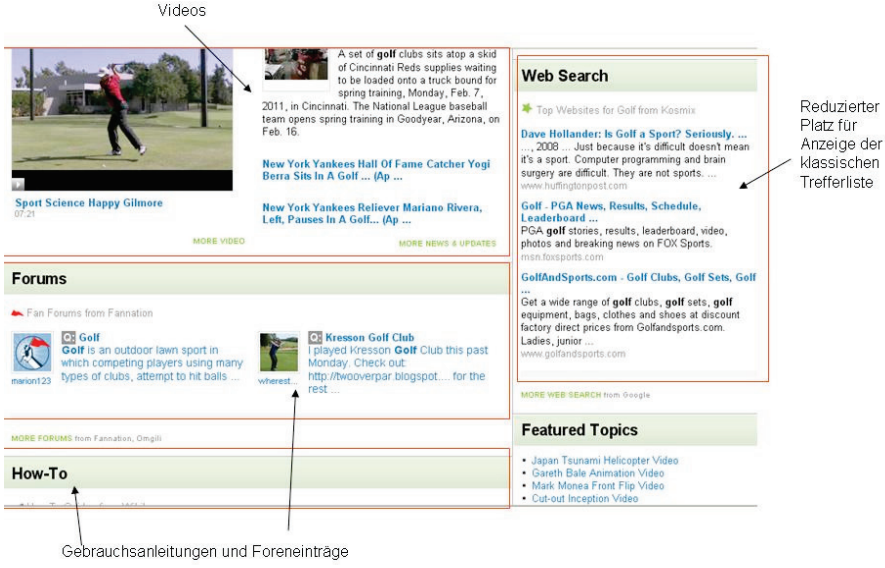

Abbildung 14: Ausschnitt einer Trefferliste für den Suchbegriff „Golf" (Quelle: http://www.kosmix.com/; Stand: 2011-03-14)

[11] RDFa ist ein serialisiertes RDF-Format, das RDF-Triple in HTML einbettet [13].

Die Suche über Kosmix bietet sich dann an, wenn nicht gezielt nach bestimmten Informationen gesucht wird, sondern wenn die Suchanfrage vor allem mit dem Ziel einen Überblick über ein Themengebiet zu erhalten, gestellt wird.

In die Gruppe der semantischen Suchmaschinen lässt sich Kosmix einordnen, da für die Bearbeitung und Beantwortung der Suchanfragen – so in den auf der Website bereitgestellten Informationen nachzulesen – eine Kombination aus Taxonomie (proprietäre offensichtlich teilautomatisiert erstellte Taxonomie, die ebenfalls automatisch aktualisiert wird) und dem Kosmix Categorization Service (automatische Einordnung der Suchergebnisse in die Taxonomie) erfolgt. Bei der automatischen Kategorisierung werden durch den Kosmix Categorization Service zunächst bei jeder gestellten Suchanfrage die Knoten (Oberbegriffe, Bündelung von Assoziationen zu einem Thema) in der Taxonomie bestimmt, die zu der Suchanfrage passen. Im Anschluss findet ein erstes Ranking der Datenquellen anhand der jeweiligen Entfernungen des Anfragethemas zu den einzelnen Knotenpunkten in der Taxonomie statt. Danach versendet Kosmix die gestellte Suchanfrage noch einmal in Echtzeit an alle Quellen und nimmt deren Antworten entgegen. Diese werden schließlich erneut weiter verarbeitet und dann dem User mit einem finalen Ranking zur Verfügung gestellt. Dieser hat dann noch die Möglichkeit, die Treffer zu filtern [47] [48] [49, S. 1-5] [50, S. 5-7].

7.2. GoPubMed

Die Suchmaschine GoPubMed, ein Projekt der Technischen Universität Dresden und Transinsight, ist eine Stand-alone-Suchmaschine für den Bereich Medizin und Biomedizin. Für die Beantwortung der Suchanfragen greift die Suchmaschine auf eine begrenzte Wissensbasis zurück. Diese setzt sich zusammen aus der Datenbank PubMed, den „Medical Subject Headings“ (kontrolliertes Vokabular der National Library of Medicine), der „Gene Ontology“ (biomedizinische Ontologie) und dem „Universal Protein Resource“ (kontrolliertes Vokabular für Proteine).

Die einzelnen Treffer werden als Trefferliste inklusive Snippets angezeigt und der Suchende hat die Möglichkeit, diese Treffer weiter zu filtern. Ein Ausschnitt einer Trefferliste ist in der Abbildung 15 zu sehen.

Abbildung 15: Ausschnitt einer Trefferliste für den Suchbegriff „cancer“ (Quelle: http://www.gopubmed.org/web/gopubmed/2?WEB10O00h00100090000; Stand: 2011-02-10)

GoPubMed ist ein gutes Beispiel dafür, dass auf dem Gebiet der semantischen Suchlösungen aus dem algorithmischen Bereich kommende Technologien wie statistische oder linktopologische Verfahren sich sinnvoll mit dem Zugriff auf mithilfe von RDF als Aussagen aufbereitete und eindeutig identifizierbare Daten kombinieren lassen. Durch das Zurückgreifen auf die genannten Wissensbasen ist während der Suche außerdem ein Synonymabgleich möglich sowie die Einbeziehung ähnlicher Begriffe in die Suche [51, S. 17, 21] [52] [53] [54] [55] [56] [57].

7.3. Sindice

Sindice ist ein Beispiel für einen Typ semantischer Suchmaschinen, der neuerdings in der Literatur auch unter der Bezeichnung Linked-Data-Suchmaschinen beschrieben wird [13]. Suchmaschinen dieses Typs crawlen Linked Data aus dem Web, indem sie den RDF-Links folgen und Abfragemöglichkeiten (z. B. als für den Endnutzer in Form einer Suchmaske aufbereiteten SPARQL-Schnittstelle) über aus unterschiedlichen Quellen aggregierte Daten ermöglichen. Nicht untypisch für diese Art von Suchtools wird Sindice im Rahmen einer Public-private-Partnerschaft von einem ganzen Konsortium öffentlicher (NUI Universität Galway, Fondazione Bruno Kessler) und privater Unternehmen (OpenLinkSoftware) entwickelt. In diesem Zusammenhang dient dann der Webauftritt, in den die Suchmaschine eingebettet ist, ganz ähnlich wie bei Kosmix als Schaufenster und zur Kontaktaufnahme für interessierte Forscher [58].

Sindice ermöglicht eine an Google oder Yahoo orientierte Keyword-basierte Suche. Nach Eingabe des Suchbegriffs erhält der Nutzer zunächst die übliche Trefferliste (Abbildung 16). Zusätzlich werden verschiedene Filtermöglichkeiten angeboten, die es ermöglichen, die Ergebnisse direkt nach dem Vorhandensein unterschiedlicher Linked-Data-Ressourcen (wie Dublin Core, RDF, FOAF, Ontologien) zu filtern (Abbildung 17).

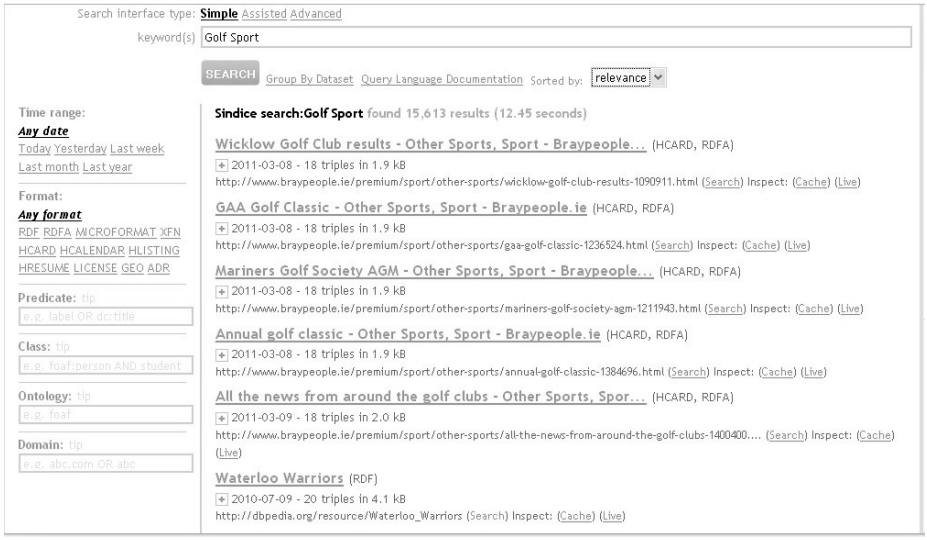

Abbildung 16: Trefferanzeige Sindice

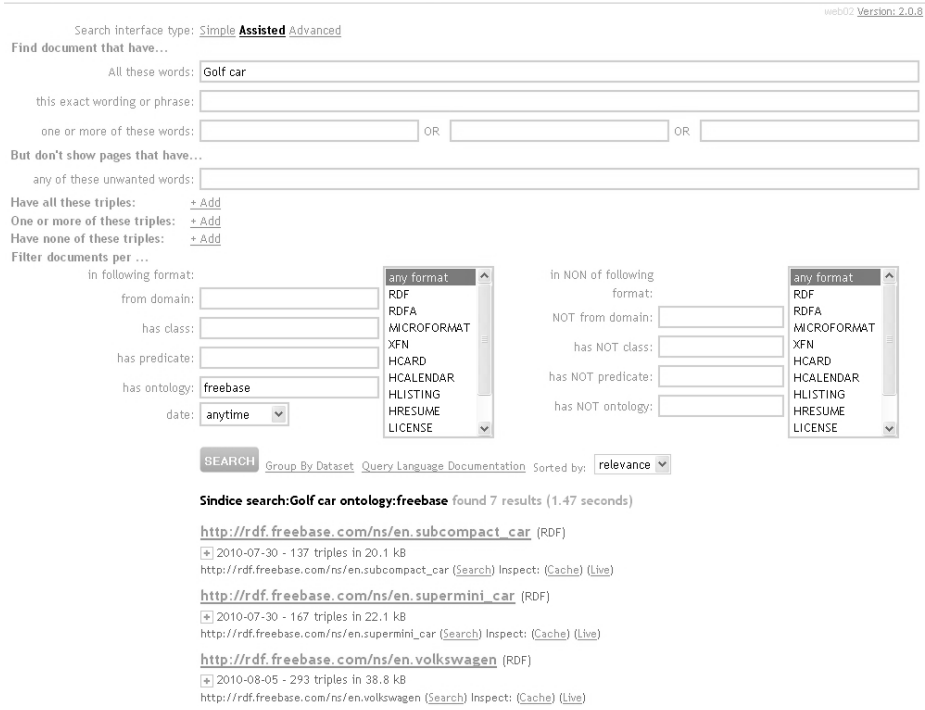

Abbildung 17: Suchmaske und Trefferanzeige von Sindice mit Möglichkeit der Einschränkung auf vorgegebene Ontologien. Im Beispiel freebase.

Der experimentelle Charakter von Sindice wird bereits daran deutlich, dass in der Trefferliste jeweils die genutzten Linked Data Typen wie RDFA oder RDF aufgeführt werden. Auch wenn versucht wird, die Eingabeoptionen durch als Suchmasken gestaltete Interfaces möglichst einfach zu gestalten, setzt diese Art der semantischen Suchmaschine eine Grundkenntnis semantischer Technologien und Standards, sowie in der erweiterten Suche die namentliche Kenntnis bestehender Ontologien, wie im obigen Beispiel (Abbildung 17) freebase, beim Nutzer voraus. Heath hebt die primäre Funktion von Sindice als Zugriffspunkt auf Schnittstellen/APIs für Linked-Data-Applikationen hervor [13].

8. Anstelle eines Fazits: Acht Thesen zum State-of-the-Art semantischer Suchmaschinen

Auf der Grundlage der gesichteten 63 Suchmaschinen lassen sich noch keine empirisch abgesicherten Aussagen zum derzeitigen State-of-the-Art der semantischen Suche machen. Die Datengrundlage ist jedoch ausreichend, um einige Thesen zu Trends, Anforderungen und Problemstellungen der semantischen Suche zu formulieren.

These 1: Vielfalt der angewandten Verfahren
Unter dem Label semantische Suchmaschine firmieren derzeit Suchmaschinen, die auf gänzlich unterschiedliche technische Verfahren und Darstellungsmöglichkeiten zurück-

greifen. Eine allgemein akzeptierte, einheitliche Definition für „Semantic Search“ hat sich bisher noch nicht durchgesetzt. Mit der Einführung des Konzeptes Linked Data werden auf RDF basierende Suchmaschinen zunehmend auch als Linked- Data-Suchmaschinen bezeichnet.

These 2: Trend zur Verknüpfung unterschiedlicher technischer Verfahren
Klassische statistische und linktopologische Verfahren und auf der Nutzung von RDF-Graphen basierende Verfahren schließen sich nicht gegenseitig aus, sondern werden häufig in Kombination eingesetzt. Dies zeigen gerade die auf begrenzte Datenbestände zugreifenden Spezialsuchmaschinen wie GoPubMed sehr gut.

These 3: Die Entwicklung semantischer Suchmaschinen ist ein intensiv bearbeitetes Experimentierfeld
Sowohl im privatwirtschaftlichen als auch im öffentlich geförderten Umfeld wird die Entwicklung semantischer Suchtechnologien derzeit intensiv gefördert.

These 4: Die unscharfe Definition des Begriffs erlaubt es, das Label „semantische Suche“ als unspezifisches Werbeargument einzusetzen
Die Tatsache, dass sich besonders viele kommerzielle Suchmaschinen im Internet als semantische Suchmaschinen bezeichnen, bestätigt die eingangs formulierte Annahme, dass es sich hierbei auch um ein „Marketing-Mittel“ handelt, um bereits seit Langem bekannte algorithmische Verfahren aus dem Umfeld des Textmining und Datamining für die Kunden attraktiv zu machen.

These 5: Semantische Suchmaschinen sind „konventionellen“ Suchmaschinen derzeit im praktischen Einsatz noch nicht überlegen
Bis auf wenige Ausnahmen kann den gesichteten semantischen Suchmaschinen derzeit noch keine Produktreife bestätigt werden. Vielmehr handelt es sich sowohl bei den kommerziellen als auch bei den mithilfe öffentlicher Förderung entwickelten semantischen Suchmaschinen vorrangig um Showcases für die angewandten Technologien. Was die Alltagstauglichkeit und die Qualität der Ergebnisse angeht, bestehen dennoch erhebliche qualitative Unterschiede.
Trotz innovativer technischer und konzeptioneller Ansätze ist derzeit im praktischen Einsatz im Bezug auf die Relevanz der Treffer noch keine Überlegenheit der semantischen Suche zu konstatieren.

These 6: Das an Googles Suchschlitz orientierte Paradigma der einfachen Keyword-Suche ist nicht dazu geeignet, die Potenziale der semantischen Suche voll auszuschöpfen
Fast alle gesichteten Suchmaschinen gehen von dem als Google-Suchschlitz bekannten Paradigma der Keyword-basierten Suche aus. Die ebenfalls durchgehend zu beobachtende Tendenz des Angebotes von unterschiedlichen Filteroptionen deutet jedoch schon darauf hin, dass diese Herangehensweise nicht unbedingt geeignet ist, das volle Potenzial der semantischen Suche auszuschöpfen. Eine Herausforderung für die Suchmaschinenentwicklung wird in der Bereitstellung von Nutzerschnittstellen bestehen, die geeignet sind, dem Nutzer die Potenziale der neuen Suchtechnologien zu vermitteln [13]. Eine Möglichkeit, dies zu tun wäre die Integration „expliziter Anweisungen“ [59]. Tatsächlich werden diese eher seltener, vor allem in den Linked- Data-Suchmaschinen, verwendet. Die für die Studie ausgewerteten Suchmaschinen experimentieren hingegen vorrangig auf dem Gebiet der Ergebnisdarstellung mit unterschiedlichen Formen der

Content Aggregation, verschiedensten Formen der Visualisierung und einer geclusterten Ergebnisanzeige.

These 7: Semantische Suche wird das bisherige Konzept der Suche grundlegend verändern bzw. ergänzen
Die Ergebnisse zu den gesichteten Suchmaschinen deuten darauf hin, dass sich das bisher vorherrschende Konzept der zielgerichteten Websuche (informationsorientierte Suche, navigationsorientierte Suche) in Zukunft um eher an Pushdiensten orientierten Modellen der Content Aggregation erweitern wird. Anbieter werden – ausgehend von Suchlösungen – semantische Suchdienste nutzen, um ihre Angebote portalartig in personalisierter Form den Nutzern anzubieten. Tendenziell wird das WWW bereits heute in dieser Form wie ein großes Content-Management-System genutzt. Die durch den Einsatz von Ontologien und Inferenzmechanismen prinzipiell mögliche Verbindung der Suche mit formal logischem Schlussfolgern – etwa mit dem Ziel, dem Nutzer stärker als bisher abgesicherte Fakteninformationen zu liefern – wird derzeit nur im Rahmen experimenteller Anwendungen auf begrenzten Wissensbasen erprobt.

These 8: Der Innovationsschub aus dem umfangreichen Forschungs- und Entwicklungsfeld der semantischen Suche ist so groß, dass auch ‚konventionelle' Suchmaschinen in Zukunft verstärkt semantische Technologien erproben werden.
Der kurze Blick auf die Integration von facettierten Darstellungsformen und Filteroptionen bei Google deutet bereits darauf hin, dass in Zukunft der Suchmaschinenmarkt sich nicht entlang einer Trennungslinie zwischen semantischen Suchmaschinen und konventionellen Suchmaschinen orientieren wird, sondern zu erwarten ist, dass semantische Technologien in Zukunft in unterschiedlicher Weise für die Suche genutzt werden.

Literatur

[1] S. Krempl, IT-Gipfel: Quaero heißt jetzt Theseus, *Heise* http://www.heise.de/newsticker/meldung/IT-Gipfel-Quaero-heisst-jetzt-Theseus-127968.html, 2006.
[2] M. Hildebrand, J. R. van Ossenbruggen & L. Hardman, An analysis of search-based user interaction on the Semantic Web, http://oai.cwi.nl/oai/asset/12302/12302D.pdf, 2007.
[3] K. Biermann, Den Maschinen die Welt erklären, http://www.zeit.de/digital/internet/2010-06/bing-microsoft-weitz, 2010.
[4] S. Winterschladen, Schlauer als Google, http://www.fr-online.de/panorama/schlauer-als-google/-/1472782/3305514/-/index.html, 2009.
[5] T. Declerck, Semantics – was ist das eigentlich? Bisherige Entwicklungen / Praktische Einsatzmöglichkeiten, *DFKI*, http://www.apa-it.at/cms/it/attachments/0/0/1/CH0317/CMS1254929768869/praesentation_t_declerck.pdf, 2011 (Abruf).
[6] D. Crystal, *Die Cambridge-Enzyklopädie der Sprache*, Campus Verl., Frankfurt a. M., 1993.
[7] W. Stock, *Information Retrieval*, Oldenbourg Verl., München, Wien, 2007.
[8] M. Bates, Models of natural language understanding, *Proceedings of the National Academy of Sciences of the United States of America*, **92** (22), 1995, 9977–9982.
[9] P. Hitzler, M. Krötzsch, S. Rudolph & Y. Sure, *Semantic Web*, Springer, Berlin, Heidelberg, 2008.
[10] A. Gattani, Web 3.0 and Semantic Search, http://blog.kosmix.com/?p=1210, 2010.
[11] W3C, What is the Semantic Web?, http://www.w3.org/RDF/FAQ, 2009.
[12] I. Herman, Tutorial on Semantic Web, http://www.w3.org/People/Ivan/CorePresentations/SWTutorial/Slides.pdf, 2011.
[13] T. Heath & C. Bizer, Linked Data: Evolving the Web into a Global Data Space, *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, 1-136. Morgan & Claypool, 2011.

[14] K. Ewell, Is "semantic search" and "the semantic web" the same thing?, http://commonsensical.wordpress.com/2007/05/24/is-%E2%80%9Csemantic-search%E2%80%9D-and-%E2%80%9Cthe-semantic-web%E2%80%9D-the-same-thing/, 2007.
[15] K. Ewell, The Search for Semantic Search, http://commonsensical.wordpress.com/2008/06/18/the-search-for-semantic-search/, 2008.
[16] S. Grimes, Breakthrough Analysis: Two + Nine Types of Semantic Search, http://www.informationweek.com/news/software/bi/showArticle.jhtml?articleID=222400100, 2010.
[17] C. Mangold, A survey and classification of semantic search approaches, *Int. J. Metadata, Semantics and Ontology*, **2** (1), 2007, 23-34, URL: ftp://ftp.informatik.uni-stuttgart.de/pub/library/ncstrl.ustuttgart_fi/ART-2007-09/ART-2007-09.pdf.
[18] W. Wei, M. Payam & A. J. Barnaghi, Search with Meanings: An Overview of Semantic Search Systems, *International Journal of Communications of SIWN* **3**, 2008, S. 76-82.
[19] T. Erich, M. Müller & G. Lübke, Anwendungsmöglichkeiten von Semantic Web bei Suchmaschinen. *Fachhochschule für Oekonomie und Management in Hamburg WinfWik*, http://winfwiki.wi-fom.de/index.php/Anwendungsm%C3%B6glichkeiten_von_Semantic_Web_bei_Suchmaschinen, 2009.
[20] T. Doszkocs, Semantic - Search Engines Mean Well, http://findarticles.com/p/articles/mi_hb3328/is_201007/ai_n54716428/?tag=content;col1, 2010.
[21] G. Kasneci, Searching and Ranking in Entity-Relationship Graphs, *Diss. Eng. Faculties of Natural Sciences and Technology of the Saarland University*, Saarbrücken, 2009.
[22] T. Imielinski & A. Signorina, If you ask nicely, I will answer : Semantic Search and Today's Search Engines, http://portal.acm.org/citation.cfm?id=1679885, 2009.
[23] T. Pellegrini & A. Blumauer, Semantic Web. *Wege zur vernetzten Wissensgesellschaft*, Springer, Berlin, 2006.
[24] W. B. Croft, D. Metzler & T. Strohman, *Search engines: information retrieval in practices*, Addison-Wesley, Boston, 2010.
[25] SUB Göttingen, Einführung in die Metadaten, http://www2.sub.uni-goettingen.de/intrometa.html, 2001.
[26] J. Bertram, *Einführung in die inhaltliche Erschließung. Grundlagen, Methoden, Instrumente*, Ergon-Verl., Würzburg, 2005.
[27] DIN 1463 Teil 1, *Erstellung und Weiterentwicklung von Thesauri, Einsprachige Thesauri*, Berlin, Beuth, 1988.
[28] C. Kunze, Semantische Relationstypen in GermaNet, In: Langer/Schnorbusch, *Semantik im Lexikon*, S. 161-178, 2005.
[29] Kruse & Naujoks, clever search, http://wdok.cs.uni-magdeburg.de/clever-search/, 2011 (Abruf).
[30] W3C, SKOS Simple Knowledge Organizsation Primer, http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/, 2009.
[31] DIN 32 705, *Klassifikationssysteme: Erstellung und Weiterentwicklung von Klassifikationssystemen*, 1987.
[32] M. A. Hearst, *Search User Interfaces*, Cambridge University Press, Cambridge, 2009.
[33] ISO 25964-1, *Thesauri and interoperability with other vocabularies* – Part 1: Thesauri for information retrieval, Draft Format for exchange of thesaurus data conforming to ISO 25964-1, 2010.
[34] ISO/IEC 13250:2002(E), *Topic Maps: Information Technology Document Description and Processing Languages*. 2nd edition, http://www.y12.doe.gov/sgml/sc34/document/0322_files/iso13250-2nd-ed-v2.pdf, 2002.
[35] P. Jaeger & U. Linck, Wissensmanagement heute – zwischen semantischen Suchtechnologien und Schwarmintelligenz, http://creativecommons.org/licenses/by-nc-nd/2.0/de/, 2009.
[36] H. Dietze & M. Schroeder, GoWeb: A semantic search engine for the life science web, *BMC Bioinformatics* **10**, 2009, 7, http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-435/paper01.pdf.
[37] M. Hess, Semantikanalyse-Verfahren, *Vorlesung an der Universität Zürich*, Institut für Computerlinguistik, https://files.ifi.uzh.ch/cl/hess/classes/sma/sma.0.1.pdf, 2007.
[38] T. Pellegrini & A. Blumauer, *Social Semantic Web: Web 2.0 – Was nun?,* Springer, Berlin, 2009.
[39] D. Koch, *Suchmaschinen-Optimierung: Website-Marketing für Entwickler*, Addison-Wesley Verl., München, 2007.
[40] D. Lewandowski, *Web Information Retrieval: Technologien zur Informationssuche im Internet*, DGI, Frankfurt am Main, 2005.
[41] G. Heyer, U. Quasthoff & T. Wittig, *Text mining: Wissensrohstoff – Text: Konzepte, Algorithmen, Ergebnisse*, W3L-Verl., 2008.
[42] M. A. Hearst, What is Text Mining?, http://people.ischool.berkeley.edu/~hearst/text-mining.html, 2003.
[43] J. Lang, Named Entity Recognition, *Universität Karlsruhe*, 2006.
[44] W3C, OWL Web Ontology Language Overview, *W3C Recommendation*, http://www.w3.org/TR/2004/REC-owl-features-20040210/, 2004.
[45] W3C, OWL 2 Web Ontology Language Document Overview, W3C Recommendation, http://www.w3.org/TR/2009/REC-owl2-overview-20091027/, 2009.

[46] Kosmix Corporation, Kosmix Publisher Solutions, http://www.kosmix.com/corp/publisher, 2011 (Abruf).
[47] Kosmix Corporation, About Kosmix, http://www.kosmix.com/corp/about, 2011.
[48] Kosmix Corporation, Kosmix, http://www.kosmix.com/, 2011.
[49] A. Rajamaran, Kosmix HighPerformance Topic Exploration using the Deep Web, http://www.vldb.org/pvldb/2/vldb09-1076.pdf, 2009.
[50] A. Rajamaran, Kosmix Exploring the Deep Web using Taxonomies and Categorization, http://sites.computer.org/debull/A09June/anand_deepweb1.pdf, 2009.
[51] A. Doms, GoPubMed – Ontologiebasierte Literaturrecherche für die Lebenswissenschaften, http://st.inf.tu-dresden.de/files/teaching/ss07/ring/Ringvorlesung%2022_06_07.pdf, 2007.
[52] Medizin Forum AG, Medline.de: medizinische Literaturdatenbanken, http://www.medline.de, 2010.
[53] National Center for Biotechnology Information, PubMed.gov, http://www.ncbi.nlm.nih.gov/pubmed.
[54] Science Media Centre, GoPubMed-PubMed browsing using ontologies, http://sciblogs.co.nz/code-for-life/2010/09/01/gopubmed-pubmed-browsing-using-ontologies/, 2010.
[55] The gene ontology, the gene ontology, http://www.geneontology.org/, 2011.
[56] Transinsight, gopubmed, http://www.gopubmed.org/web/gopubmed/2?WEB10O00h00100090000, 2011.
[57] Webwire, WWW goes WWWW: Searching is now sorted! *The Web 2.0 for real*, http://www.webwire.com/ViewPressRel.asp?aId=50204, 2007.
[58] About Sindice, http://www.sindice.com/main/about, 2011 (Abruf).
[59] J. Kalbach, *Usability von Semantischen Anwendungen*, Unveröffentlichter Vortrag im Rahmen eines Masterkurses an der HAW Hamburg am 05.01.2011 in Hamburg.
[60] Deutscher Bundestag, Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Grietje Bettin, Ekin Deligöz, Kai Gehring, weiterer Abgeordneter und der Fraktion BÜNDNIS 90/DIE GRÜNEN – Drucksache 16/4472 – Aktuelle Entwicklungen des Suchmaschinenprojektes, http://dipbt.bundestag.de/dip21/btd/16/046/1604671.pdf, 2007.
[61] K. Reichenberger, *Kompendium semantische Netze: Konzepte, Technologien, Modellierung*, Springer, Berlin, 2010.
[62] K.-U. Carstensen, C. Ebert, C. Ebert, S. Jekat, H. Langer & R. Klabunde, *Computerlinguistik und Sprachtechnologie: Eine Einführung*, 3. Aufl., Spektrum Akademischer Verl., Heidelberg, 2010.
[63] J.M Kassim & M. Rahmany: Introduction to Semantic Search Engine, International Conference on Electrical Engineering and Informatics, 2009: ICEEI '09 ; Bangi, Malaysia, 5 - 7 August 2009 , Vol. 2, IEEE, Piscataway, NJ, 2009, 380-386.
[64] H. Dong; F. K. Hussain; E. Chang: A Survey in Semantic Search Technologies, 2nd IEEE International Conference on Digital Ecosystems and Technologies, 2008: DEST 2008; 26 - 29 Feb. 2008, Phitsanulok, Thailand. IEEE, Piscataway, NJ, 2008, 403-408.

Anhang 1: Liste der gesichteten und in der Auswertung berücksichtigten Suchmaschinen

Suchdienst	**URL**
Arachne	http://www.arachne.uni-koeln.de/drupal/
Arbeitsagentur	http://jobboerse.arbeitsagentur.de/vamJB/ startseite.html?kgr=as&aa=1&m=1
askMEDLINE	http://askmedline.nlm.nih.gov/ask/ask.php
Bing	http://www.bing.com
Bio Portal	http://bioportal.bioontology.org/
Carrot Search	http://search.carrotsearch.com/carrot2-webapp/search
Chilibot	http://www.chilibot.net/
ClusterMed	http://demos.vivisimo.com/clustermed
Cluuz Search	http://www.cluuz.com/
Cognition	http://www.cognition.com/
Deepdyve	http://www.deepdyve.com/how-it-works
DuckDuckGo	http://duckduckgo.com
EAGLi	http://eagl.unige.ch/EAGLi/
EBIMed	http://www.ebi.ac.uk/Rebholz-srv/ebimed/
Endeca	http://www.endeca.com/en/home.html
Europeana	http://www.europeana.eu/portal/
Evri	http://www.evri.com
Exalead	http://www.exalead.com/search/

EyePlorer.com	http://eyeplorer.com/
factbites	http://www.factbites.com
Find the Best	http://findthebest.com
Freebase	http://www.freebase.com
Go3R	http://www.go3r.org/
GoPubMed	http://www.gopubmed.org
Hakia	http://www.hakia.com/
Healthline	http://www.healthline.com/
HealthMash	http://healthmash.com/
Hulbee	http://www.hulbee.com
iHop	http://www.ihop-net.org/UniPub/iHOP/
Jobanova	http://www.jobanova.de/
Kngine	http://www.kngine.com
Kosmix	http://www.kosmix.com/
LexisNexis	http://www.lexisnexis.de/
Liveplasma.com	http://liveplasma.com/
Medstory	http://www.medstory.com/Home.html
MnemoMap	http://www.mnemo.org
Musicovery	http://musicovery.com/
Nextbio	http://www.nextbio.com/b/nextbio.nb
Pixolu	http://www.pixolu.de/
Quintura.com	http://www.quintura.com
Releton	http://www.releton.com
RightHealth	http://www.righthealth.com/
Semager	http://www.semager.de
Semantifi	http://www.semantifi.com
SenseBot	http://www.sensebot.net/sense6.aspx
Sindice	http://www.sindice.com/
Spezify	http://www.spezify.com
Swingly	http://beta.swingly.com/login
Swoogle	http://swoogle.umbc.edu/
SWSE	http://swse.deri.org/
TextRunner Search	http://www.cs.washington.edu/research/textrunner/indexTRTypes.html
Tineye	http://beta.swingly.com/login
TipTop	http://feeltiptop.com
TrialX	http://trialx.com/
TrueKnowledge	http://www.trueknowledge.com
Truevert	http://www.truevert.com/
TrustYou	http://www.trustyou.com/
Vadlo	http://vadlo.com/
Viewchange	http://www.viewchange.org/
WolframAlpha	http://www.wolframalpha.com
Yummly	http://www.yummly.com/
Zibb	http://www.zibb.com/