TECHNOLOGIE **INTERNET**

Semantische Suche

Auf dem Weg zur Bedeutungserschliessung grosser Informationsmengen

Internet-Suchmaschinen haben ihren Ursprung im Information Retrieval. Sie erstellen einen Index aus Schlüsselworten, der für Suchanfragen mittels syntaktischem Vergleich verwendet wird, und liefern eine nach Relevanz geordnete Trefferliste. Semantische Suchsysteme hingegen erkennen automatisch den Bedeutungszusammenhang eines Suchterms und schlagen alternative Begriffe vor. Der semantische Ansatz verspricht, die Suche besser, intuitiver und relevanter zu gestalten.

Andreas Dengel

Gibt man in eine konventionelle Suchmaschine den Begriff «Bank» ein, so erhält man einige Hundert Millionen Treffer in Form von Verweisen auf möglicherweise relevante Dokumente, die den Suchbegriff enthalten. Da die traditionelle Schlüsselwortsuche lediglich das Vorkommen der Suchbegriffe im Index überprüft, bleiben Akronyme oder Synonyme mit gleicher oder ähnlicher Bedeutung unberücksichtigt. Für die Anfrage nach «Bank» werden relevante Dokumente, die beispielsweise den Begriff «Geldinstitut» enthalten, nicht gefunden. Semantische Suchsysteme erlauben in solchen Fällen beispielsweise Rückfragen wie «Meinten Sie das Finanzinstitut oder die Sitzgelegenheit?». Dazu werden Bedeutungen der Anfrage, wie auch Aussagen der zu durchsuchenden Texte mithilfe kognitiver Algorithmen analysiert und es wird versucht, diese formal zu verstehen. Man kann die semantische Suche als einen Suchprozess betrachten, bei dem in jeder Phase der Suche semantisches Wissen verwendet wird. Dies betrifft die Anfragestellung genauso wie die (eigentliche) Suche und Ergebnisrepräsentation.

Grenzen klassischer Suche

Nehmen wir an, wir möchten die Adresse des Financial Service Unternehmens CarFS in Bern herausfinden und geben die Begriffe «Adresse», «CarFS», «Bern» in die Suchmaschine Google ein, so erhalten wir alle Dokumente, in denen die genannten Stichworte vorkommen. Da die Begriffe als Zeichenketten betrachtet werden, versteht die Suchmaschine weder, worum es geht, noch, dass die Anfrage ein ganz bestimmtes Faktum anfragt. Daher umfasst die Ergebnisliste auch ca. 71 Millionen potenziell relevanter Dokumente, mit Vorschlägen wie «Einkaufen in Bern», Autovermietungen, Busreisen, Ticketservices, usw. Wüsste die Suchmaschine, dass Bern eine Stadt in der Schweiz ist, so könnte sie einen Zusammenhang herstellen, dass es wohl um eine Adresse in Bern geht und hinter CarFS eine Person oder ein Unternehmen steht, das einen solchen Namen trägt. Entsprechend eingeschränkt kann dann eine gezielte Faktensuche stattfinden.

Formularbasierte Suche

Eine Form zur näheren Bestimmung der Bedeutung wird mit dem Begriff «Formularbasierte Suche» umschrieben. Sie beruht darauf, dass ein Benutzer seine Anfrage unter Verwendung von (Web-) Formularen eingeben kann. Das Formular umfasst eine Menge von bezeichneten Feldern, wie «Stadt», in die man einen dazugehörigen Wert wie «Bern» eingeben kann. Da mit der Eingabe die Semantik

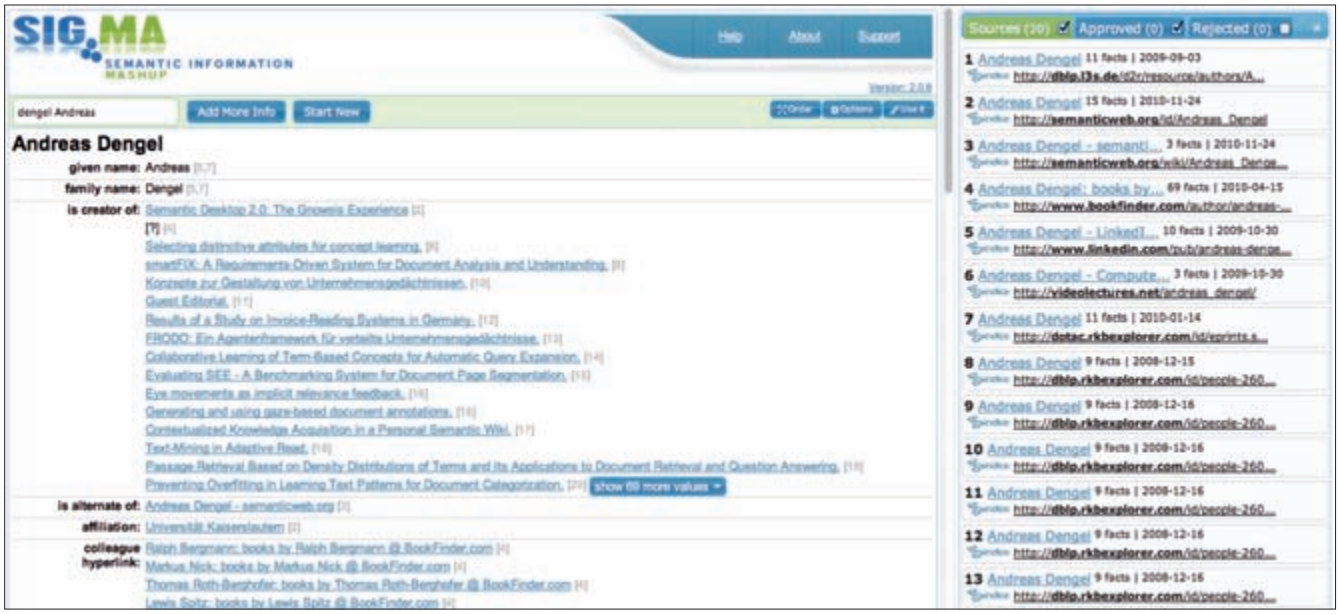


Bild 1 Semantikbasierte Schlüsselwortsuche mit SIG.MA.

TECHNOLOGIE **INTERNET**

bereits festgelegt ist, sind für diese Form der Suche keine aufwendigen Suchalgorithmen erforderlich, denn die gemachten Angaben reichen aus, um die Abfrage, z.B. als SPARQL-Ausdruck, zu konstruieren und auszuführen.

Klassifikationssysteme beschreiben Vokabular

An diesem Beispiel kann man bereits eine wichtige Dimension von Wissen erkennen, auf der semantische Suchsysteme aufbauen: Sie nutzen das Konzept der Klassifikation, das es erlaubt, gleichartige Ressourcen unter einem Begriff zusammenzufassen. Klassen sind dabei wie Variablen zu betrachten, d.h. sie können eine Menge von Werten annehmen, so wie in unserem Beispiel es neben Bern auch noch viele andere Städte gibt, wie Zürich, Basel oder Genf. Sie alle gehören damit zur Klasse «Stadt». Wenn man es genauer nimmt, so gehören sie sogar zur Klasse «Schweizer Stadt», die nur Werte von Städten annehmen kann, die auch in der Schweiz liegen. Man kann also sagen, dass «Schweizer Stadt» eine Unterklasse von «Stadt» ist.

An diesem erweiterten Beispiel wird nun noch klarer, welche Prinzipien einer formalen Semantik unterliegen. Eine Domäne (Vokabular) wird in Klassen (Begrifflichkeiten) gegliedert, diese werden in Form einer Klassenhierarchie (Taxonomie) angeordnet und am Ende werden jeder Klasse Werte zugeordnet. Diese Werte bezeichnet man als Instanzen einer Klasse (Bern ist eine Instanz von Stadt). Beides zusammen bildet eine Ontologie, die, formal beschrieben, das Hintergrundwissen für semantische Suchmaschinen darstellt.

Ontologien als formale Semantik

Im Internet stehen bereits eine ganze Reihe solcher Ontologien als Linked Open Data [1] zur Verfügung. Ein prominenter Vertreter ist DBpedia – die semantische Version der offenen, freien Online-Enzyklopädie Wikipedia – wo die als Datensammlung vorliegenden HTML-beschreibungen aus Wikipedia in RDF transformiert wurden. Das Resource Description Framework (RDF) [1] ist eine Standardsprache des W3C (World Wide Web Consortium) und dient der Beschreibung von Ontologien. Das zugrundeliegende Modell besteht aus den drei Objekttypen: Ressourcen, Eigenschaftselementen und Objekten. Jeweils eine Ressource, eine Eigenschaft

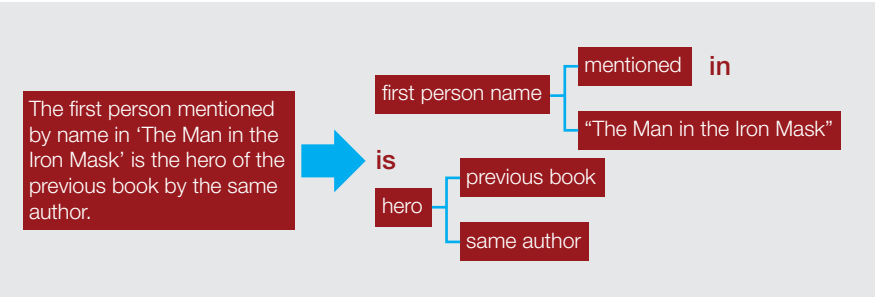


Bild 2 Anfragezerlegung bei Watson [6].

und ein Objekt bilden zusammen ein sogenanntes RDF-Tripel. Ein solches Tripel kann man sich wie einen elementaren Satz bestehend aus Subjekt, Prädikat und Objekt vorstellen, der eine Aussage über ein bestimmtes Objekt innerhalb einer Domäne macht. Die Aussage (das Tripel) etwa, dass Bern eine Stadt ist, lässt sich unter Zuhilfenahme des in DBpedia standardisierten und auf URIs (Uniform Resource Identifier) aufbauenden Vokabulars in RDF wie folgt beschreiben:

```
<http://dbpedia.org/resource/Bern>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/City>
```

Eine aktuelle semantische Beschreibung für Bern findet man beispielsweise unter der Web-Adresse http://dbpedia.org/page/Bern.

Formale Anfrageformulierung

Man kann die Nützlichkeit solcher Ontologien sehr schön am Beispiel von Suchmaschinen erkennen, deren Schwerpunkt in der Anfrageformulierung liegt. Entsprechende Suchmaschinen verwenden eine spezielle Sprache zur Formulierung von Anfragen, die sich der formalen Beschreibung von RDF bedient. SPARQL [2] beispielsweise ist eine solche formale Anfragesprache, mit der sich Suchanfragen der folgenden Form ausdrücken lassen:

```
PREFIX abc: <http://example.com/example-Ontologie#>
SELECT ?canton_capital ?canton
WHERE {
?x abc:cityname ?canton_capital.
?y abc:statename ?canton.
?x abc:isCapitalOf ?y.
?y abc:isInCountry abc:switzerland.
}
```

Im dargestellten Beispiel werden alle Kantonshauptstädte mit dem zugehörigen Schweizer Kanton ermittelt, sofern diese in der Hintergrundontologie enthalten sind. Hierzu erfolgt eine Abfrage mittels Zugriff auf die aus RDF bekannten Tripel. Dabei werden alle Variablenbelegungen für **?canton_capital** und **?canton** zurückgegeben, die auf die Muster dieser vier RDF-Tripel passen. Durch den verwendeten Präfix kann die Notation verkürzt und somit leserlicher gestaltet werden.

Semantikbasierte Schlüsselwortsuche

Aber auch reine Schlüsselwortsuche kann durch Semantik erweitert werden. Die ersten Suchmaschinen, die dieses Prinzip eingesetzt haben, verwendeten eine einfache Syntax, mit der man beschreiben konnte, welcher Klasse die Wörter einer Suchanfrage zugehörig sind. Durch erweiterte Anfragen wie organization:bank city:zuerich konnte man die Bedeutung der Suchterme präzisieren.

Ein aktuelles Beispiel für semantikbasierte Anfragesysteme ist die Suchmaschine SIG.MA (Semantic Information Mashup – http://sig.ma/). Sie durchstöbert in Webseiten eingebettete RDF-Annotationen, aggregiert die gefundenen Ressourcen und stellt diese strukturiert zur Verfügung. Gibt man etwa den Namen des Autors dieses Artikels in SIG.MA ein, so erhält man – wie in **Bild 1** ausschnittsweise dargestellt – eine Menge von Fakten (links) und dazu gehörigen Ressourcen (rechts), die über die entsprechende URI zugreifbar sind. Bei der Eingabe kann der Benutzer direkt mitverfolgen, wie sukzessive immer mehr Fakten gefunden und geprüft werden und die Ergebnisliste Schritt für Schritt erweitert wird. Bemerkenswert für SIG.MA ist, dass es nicht nur Fakten hinzufügt, sondern während der Suche auch solche Fakten wieder entfernt, deren Mehrdeutigkeiten aufgelöst werden konnten.

Question Answering

Eine erweiterte Form der semantischen Suche sind sogenannte Question-

TECHNOLOGIE **INTERNET**

Bild 3 «Click & Filter» bei facettierter Suche.

Answering-Systeme. Sie sind dafür geeignet, sprachlich ausformulierte natürliche Anfragen mithilfe einer vorher definierten Semantik zu verarbeiten und zu beantworten. Sie verwenden oft auch weitergehendes linguistisches Wissen, um eine Anfrage zu analysieren und in eine formale Anfrage zu übersetzen. Ebenso werden die gefundenen, formal beschriebenen Ergebnisse in natürliche Sprache umgeformt, bevor sie als Antwortsatz ausgegeben werden. Die Wissensbasis der Question-Answering-Tools ist aufwendig konstruiert und bietet detailliertes Wissen über die vorhandenen Instanzen und deren Beziehungen. Ein bekannter Vertreter von Question-Answering ist das von IBM entwickelte System Watson [3], das englische Fragen – wie in **Bild 4** dargestellt – versteht, die enthaltenen Wörter im Kontext der Anfrage einordnet und mit Hilfe einer umfangreichen Wissensbasis schnell und präzise beantwortet.

Da semantische Suchsysteme auf umfangreichem Wissen aufbauen, wäre es natürlich sinnvoll, wenn der Benutzer für die Anfragestellung das Vokabular der verwendeten Ontologien kennen würde. In vielen Fällen werden zur Unterstützung entsprechende grafische Benutzerschnittstellen angeboten, die z.B. vorhandene Klassen, wie Orte (Länder, Städte, Gemeinden), Ereignisse (Feste, Kongresse, Aufführungen), Personen oder Organisationen z.B. mittels Pulldown-Menü zur Verfügung stellen, sowie zugehörige Prädikate (Eigenschaftselemente), z.B. LivesIn, WorksFor, oder HasLocation anzeigen, mit der Aussagen über die dazugehörigen Instanzen gemacht werden können.

Facettiertes Browsing

Eine einfache, aber wirkungsvolle Variante dieser Form der semantischen Suche ist das sogenannte facettierte Browsing, das eine klassische Suche mit Klassifikationssystemen verbindet. Neben der bekannten Eingabe von Stichwörtern kann der Benutzer aus einer Reihe von gruppierten Klassen auswählen, die ihm helfen sollen, das zugrunde liegende Vokabular besser zu verstehen und seinen Informationsbedarf auf Grundlage des Informationsangebotes gezielter zu beschreiben bzw. zu nutzen. Solche polyhierarchischen Klassifikationssysteme nennt man auch Facetten. Facetten beschreiben i.d.R. das Vokabular unterschiedlicher Informationssichten, wie z.B. Was, Wer, Wo, usw. und erlauben durch Anklicken der entsprechenden Klasse in den Facetten dann Anfragen wie etwa «Privatdarlehen, Banken, Zürich» sehr intuitiv anzustossen.

Facettierte Suche arbeitet nach dem Prinzip «Click & Filter», d.h. das Anklicken bestimmter Klassen in den angebotenen Facetten schränkt die Menge der infrage kommenden Ergebnisse nach und nach ein. **Bild 3** zeigt ein Beispiel aus dem EU-Verbundvorhaben Mace [4], an dem das DFKI mitgewirkt hat und in dem Information zu Architektur in ganz Europa über Facetten zugänglich gemacht wird. Mace bietet Facetten zu verschiedenen Repositorien, Sprachen, Ressourcen, Architekturkategorien und Kompetenzen an. Durch Anklicken einer Klasse, z.B. der Sprache «English» wird der Suchraum der anderen Facetten eingeschränkt und gleichzeitig nach Anzahl der verfügbaren Ressourcen sortiert. Die Auswahl des Repositoriums «Dynamo» führt zur weiteren Einschränkung der verbleibenden Optionen. Bei diesem Vorgehen erhält der Benutzer bei jedem Suchschritt einen exzellenten Überblick darüber, welche Eigenschaften die Ergebnismenge besitzt bzw. welche Eigenschaften zur Verfeinerung der Suche weiter zur Auswahl stehen.

Visuelle semantische Suche

Die Visualisierung von Anfrageoptionen und Suchergebnissen ist durchaus sinnvoll, wenn man die inhärente Komplexität von Bedeutung betrachtet. Ausgewählte Visualisierungstechniken kombinieren daher unterschiedliche Darstellungsmetaphern wie Farbe, Grösse, Intensität, Abstand oder Richtung und helfen so, den Überblick über die Ergebnismenge nicht zu verlieren bzw. die Optionen für eine weitere Suche/Navigation zu erkennen. Auch an dieser Stelle kommen, ähnlich wie bei der facettierten Suche, Klassifikationssysteme zum Einsatz, um Wissen darzustellen und durch geeignete Interaktionsmöglichkeiten den Zugang dazu anzubieten. Bekannte Instanzen oder Dokumente werden je nach ihrer Relevanz in der Wissensbasis, der semantischen Nähe zur Anfrage, ihre Klassenzugehörigkeit oder eines betrachteten Beziehungstyps in einem Informationsraum dargestellt und können mit der Mouse-Over-Funktion weiter erschlossen werden.

TECHNOLOGIE **INTERNET**

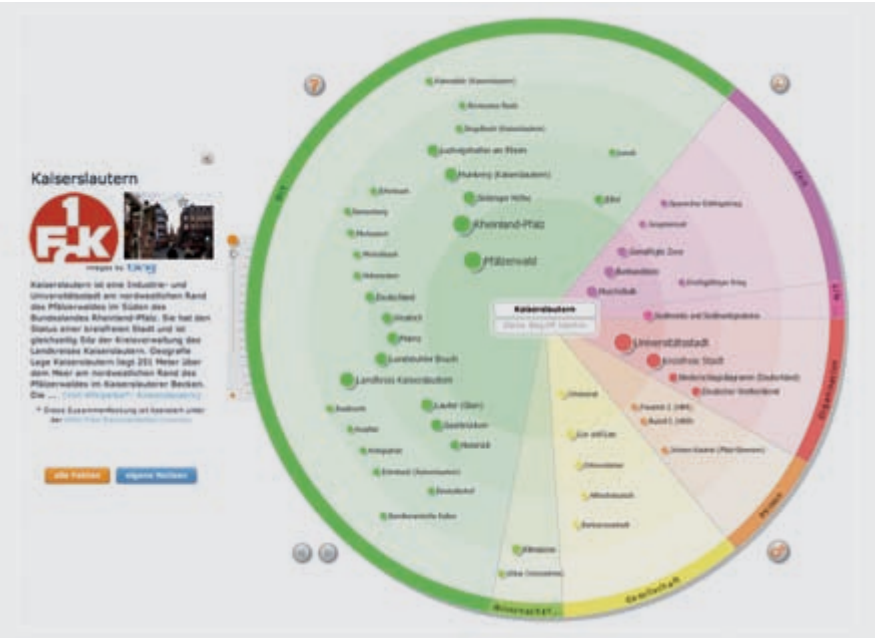
Résumé **La recherche sémantique**

En route vers l'exploitation sémantique de quantités d'informations importantes

Les moteurs de recherche sur Internet trouvent leur origine dans l'informatique documentaire (information retrieval). Ils créent un index de mots clés qui représente la quantité de documents considérée. L'index est destiné à répondre aux demandes à partir des termes recherchés en proposant une liste de résultats classés en fonction de leur pertinence. Ce processus s'effectue en réalisant une comparaison syntaxique des chaînes de caractères entrées avec les mots clés. Contrairement à ce mécanisme, les moteurs de recherche sémantique identifient automatiquement le contexte sémantique d'un terme recherché et proposent d'autres mots ou bien des concepts utilisés. D'après un grand nombre de spécialistes, les moteurs de recherche sémantique constituent la solution la plus prometteuse pour améliorer la recherche et la rendre plus intuitive et plus pertinente en fonction du sujet.

La recherche sémantique suscite un intérêt croissant, notamment en ce qui concerne le développement du Web sémantique et du Linked Open Data Cloud. Même Google œuvre actuellement afin de faire avancer les microformats en tant que variante d'entrée dans la recherche sémantique et de les utiliser de manière similaire à SIG.MA. Ainsi, l'association d'une reconnaissance contextuelle aux profils d'utilisateur et à des concepts de visualisation innovants est censée permettre la réalisation de progrès importants en matière de moteurs de recherche. No

Das System eyePlorer ist ein Beispiel für eine solche visuelle semantische Suche [5]. Es ordnet zur Anfrage verwandte Begriffe in einem Kreis um das Eingabefeld herum an, wobei Suchkategorien farblich hervorgehoben sind. Zieht man den Mauszeiger über einen der dargestellten Begriffe, erscheint ein kurzer Informationstext, der mittels Mausklick den zugehörigen Wikipedia-Artikel öffnet. Bild 4 zeigt ein Beispiel für die Anfrage «Kaiserslautern». Auf der linken Seite der Benutzeroberfläche wird eine Zusammenfassung der Suchergebnisse angezeigt, bestehend aus Bildern und Text sowie entsprechenden Quellenangaben.



Bilder: Andreas Dengel

Bild 4 Visualisierung mittels eyePlorer.

Fazit

Die semantische Suche wird insbesondere mit der Entwicklung des semantischen Web und der Linked Open Data Cloud zunehmend interessanter. Auch Google arbeitet daran, Microformate als Einstiegsvariante in die semantische Suche voranzutreiben und ähnlich wie bei SIG.MA zu nutzen, sodass zu erwarten ist, dass durch die Kombination mit Kontexterkennung und Benutzerprofilen sowie innovativen Visualisierungskonzepten grosse Fortschritte für Suchsysteme zu erwarten sind. Weiterführende Literatur findet man auch in [1].

Referenzen

[1] A. Dengel, Semantische Technologien – Grundlagen. Konzepte. Technologien. Spektrum Akademischer Verlag, Springer Berlin Heidelberg (Okt. 2011).
[2] http://www.w3.org/TR/rdf-sparql-query/
[3] http://www-05.ibm.com/de/pov/watson/
[4] http://portal.mace-project.eu/
[5] http://de.vionto.com/show/
[6] http://video.golem.de/pc-hardware/5297/ibm-erklaert-jeopardy-computer-watson.html

Angaben zum Autor

Prof. Dr. **Andreas Dengel** ist Standortsprecher und Wissenschaftlicher Direktor am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) in Kaiserslautern. Seit 1993 ist er Professor für Informatik an der TU Kaiserslautern, wo er den Lehrstuhl für Wissensbasierte Systeme leitet. Seit 2009 ist er auch Professor (Prof. h.c.) am Department for Computer Science and Intelligent Systems der Graduate School of Engineering an der Osaka Prefecture University in Japan.

DFKI GmbH, 67663 Kaiserslautern, Deutschland
andreas.dengel@dfki.de