published as: H. Sack: Semantische Suche - Theorie und Praxis am Beispiel der Videosuchmaschine yovisto.com. In U. Hentgartner, A. Meier (Eds.): Web 3.0 & Semantic Web, HMD - Praxis der Wirtschaftsinformatik, Nr. 271, dpunkt Vlg. Heidelberg, 2010, pp. 13-25. (ISSN 1436-3011)

SEMANTISCHE SUCHE

Theorie und Praxis am Beispiel der Videosuchmaschine yovisto.com

Traditionelle Suchmaschinen stoßen im World Wide Web heute schnell an ihre Grenzen. Einerseits liefern Suchanfragen oft Listen mit Millionen von Ergebnisdokumenten zurück, so dass die Vollständigkeit des Suchergebnisses nicht mehr beurteilt werden kann. Andererseits finden sich darin zahlreiche nicht relevante Informationen, verursacht durch die Mehrdeutigkeiten, unterschiedlichem Kontext oder Pragmatik. Wünschenswert wären sowohl eine höhere Treffgenauigkeit und damit Qualität der erzielten Suchergebnisse sowie ein besserer Überblick über die Ergebnisse bzw. über den gesamten Suchraum. Abhilfe verspricht eine Suche, die sich am tatsächlichen Inhalt der durchsuchten Dokumente und dessen Bedeutung orientiert, anstatt wie heute üblich am Vergleich von Zeichenketten, wobei Kontext und Pragmatik berücksichtigt werden müssen. Im Semantic Web wird die Bedeutung multimedialer Dokumente mit Hilfe geeigneter Wissensrepräsentationen explizit gemacht. Werden diese semantischen Annotationen in den Suchprozess integriert, eröffnen sich neue Möglichkeiten, die Qualität der erzielten Suchergebnisse zu verbessern und speziell an Benutzerbedürfnisse anzupassen. Inhaltliche Zusammenhänge zwischen einzelnen Dokumenten können explizit gemacht werden und erlauben über Klassifikationen und Kategorisierungen neue Wege der Visualisierung des Such- und Ergebnisraumes hin zu einer explorativen Suche, die es dem Benutzer gestattet, die Suchergebnisse und damit im Zusammenhang stehende Informationen zu erkunden. Diese neuen Möglichkeiten der semantischen Suche sollen am Beispiel der Videosuchmaschine yovisto.com dargestellt werden.

Stichworte: Semantic Search, Information Retrieval, Semantic Web, Suchmaschinen, Videosuche

1. Einleitung

Am 13. März 2009 feierte das World Wide Web (WWW) am europäischen Kernforschungszentrum CERN seinen zwanzigsten Geburtstag. Der britische Physiker Sir Tim Berners-Lee, der als Vater des WWWs gilt, lieferte im März 1989 mit seinem kurzen Memorandum „Information Management: A Proposal“ [Berners-Lee 1989] das Startsignal für das weltumspannende, multimediale Informationssystem, dessen unglaublicher Erfolg sich aus seiner offenen Architektur herleitet. Dabei hält das rasante Wachstum des WWWs ungebrochen an und schon lange ist es auf eine Größe angewachsen, die ohne automatisierte Hilfsmittel nicht bewältigt werden könnte. Suchmaschinen wie Google sind es, die mit ihren automatisierten WebCrawlern das WWW durchforsten, relevante Dokumente erfassen und auswerten, und uns damit überhaupt erst den Zugriff auf die von uns benötigte Information ermöglichen. Im Juli 2008 vermeldete das Google Weblog, dass der Google WebCrawler mehr als eine Billion (10^{12}) Dokumente erfasst habe [Google 2008]. Diese Zahl betrifft dabei nur einen Teil der insgesamt über das WWW erreichbaren Dokumente und lässt zugriffsbeschränkte Informationen, die meisten dynamisch generierten Inhalte und die Informationsbestände der zahlreichen Intranets unberücksichtigt. Wären wir alleine auf handrecherchierte Web-Portale und Hyperlinkverknüpfungen angewiesen, keiner fände sich mehr im WWW zurecht.

So sind heute Suchmaschinen das probate Mittel zur Informationsrecherche im WWW. Doch die Vielzahl der Medien- und Dokumentenvarianten mit all ihrer Mehrdeutigkei-

ten und Komplexität sowie die schiere Masse an vorhandenen Informationsressourcen führen traditionelle WWW-Suchmaschinen schnell an ihre Grenzen.

2. Grenzen des traditionellen Information Retrievals im WWW

2.1 Information Retrieval im WWW

Die grundlegende Aufgabe einer Suchmaschine besteht in der Ermittlung von Dokumenten, die bezogen auf eine Suchanfrage als relevant eingestuft werden. Dokumente und Informationsressourcen werden dazu indexiert, d.h. in eine abstrakte Repräsentationsform gebracht, in der eine Ähnlichkeitsbestimmung zwischen den durchsuchten Informationsressourcen und der Suchanfrage den Ausschlag für die anschließende Ergebnisauswahl liefert. Um die Qualität der erzielten Suchergebnisse abschätzen zu können, werden üblicherweise einfache statistische Maßzahlen zur Hilfe genommen, die angeben, wie genau (Precision) und wie vollständig (Recall) die erzielten Suchergebnisse sind. Werden beim reinen Text Retrieval lediglich Sammlungen von Textdokumenten betrachtet, erweitert das multimediale Information Retrieval den Dokumentenbestand um Bild-, Audio- und Videodaten. Aus Textdokumenten lassen sich Deskriptoren (Metadaten) gewinnen, indem eine Untermenge der enthaltenen Terme ausgewählt wird. Dagegen ist die Ableitung inhaltlich aussagekräftiger Deskriptoren aus multimedialen Daten ohne Textanteile wesentlich schwieriger.

Low-Level Deskriptoren lassen sich direkt aus statistischen Analysen der Multimediadaten gewinnen, wie z.B. Aussagen über die Farbverteilung eines Bildes, der Verlauf der Lautstärke einer Audiosequenz oder die Helligkeitsdifferenzen zweier aufeinanderfolgender Einzelbilder einer Videosequenz. Low-Level Deskriptoren erlauben zunächst keine Aussagen über den Inhalt der betrachteten Daten. Sie eigen sich insbesondere für eine ähnlichkeitsbasierte Suche,bei der z.B. Bilddateien gefunden werden sollen, deren Inhalt nach visuellen Kriterien gemessen einem vorgegebenen Bild ähnlich sehen. High-Level Deskriptoren dagegen besitzen ein höheres Abstraktionsniveau und repräsentieren nicht direkt visuelle oder auditive Parameter, sondern vielmehr deren inhaltliche Interpretation. Eine inhaltsbasierte Suche lässt sich in multimedialen Daten am besten über High-Level Deskriptoren durchführen. Diese Metadaten können entweder manuell hinzugefügt oder mit Hilfe automatischer Analyseverfahren gewonnen werden.

Das WWW als verteiltes Hypermediasystem besteht aus miteinander über Hyperlinks vernetzten (multimedialen) Dokumenten. Suchmaschinen sammeln diese Dokumente mit Hilfe von WebCrawlern, die jedes erfasste WWW-Dokument auf darin enthaltene Hyperlinks untersuchen und diese weiterverfolgen. Die erfassten Dokumente durchlaufen dabei eine Prozesskette unterschiedlicher Analyseverfahren, an deren Ende für jedes Dokument eine Liste aussagefähiger Deskriptoren mit zugehöriger Relevanzgewichtung ermittelt wird, nach denen dann der gesamte Dokumentenbestand gezielt durchsucht werden kann. Die Vernetzung der WWW-Dokumente untereinander dient dabei als Indikator für die Relevanz der einzelnen Dokumente [Ribeiro-Neto et al. 1999, Brin et al. 1998].

Bilder, Audiodateien und insbesondere Video bestimmen heute mehr und mehr das Erscheinungsbild des WWWs. Dabei sind aktuelle Suchmaschinen bei der Suche in multimedialen Daten immer noch auf Textdaten angewiesen, die direkt über Analyseverfahren aus den Originaldaten gewonnen oder zusammen mit den Originaldaten abgespei-

chert vorliegen. Die automatische inhaltliche Analyse von multimedialen Dokumenten stößt heute schnell an ihre Grenzen. Noch immer ist das bild- und informationsverarbeitende System der menschlichen Wahrnehmung jeder automatisierten Analyse überlegen, insbesondere dann, wenn komplexe Transfer- und Abstraktionsleistungen notwendig werden. Daher werden die zur Informationssuche benötigten Metadaten heute meist noch manuell gewonnen.

2.2 Probleme traditioneller WWW-Suchmaschinen

Google als prominentester Vertreter der WWW-Suchmaschinen steht heute vielfach in der Kritik, da als Antwort auf eine Suchanfrage oft zuviele nicht relevante Ergebnisse präsentiert werden. Auf der anderen Seite kann heute kein Mensch mehr beurteilen, ob sich auch tatsächlich alle relevanten Ergebnisse in der Fülle der angebotenen Suchtreffer befinden. Verantwortlich dafür sind unter anderem sprachliche Mehrdeutigkeiten auf unterschiedlichen semantischen Abstraktionsebenen. Ein Wort selbst kann unterschiedliche Bedeutungen besitzen (Homonymie), so bezeichnet das Wort “Golf“ sowohl eine Sportart, eine Automarke als auch einen Meeresarm. Eine klassische schlüsselwortbasierte Suche nach dem Begriff „Golf“ resultiert in Ergebnisdokumenten, in denen das Wort “Golf“ in unterschiedlichen Bedeutungen verwendet wird, von denen nicht alle der vom Benutzer intendierten Bedeutung entsprechen. Aber selbst wenn auf eine einheitliche Bedeutung geschlossen werden könnte, so kann ein Wort auch in unterschiedlichem Kontext und mit unterschiedlicher Absicht (Pragmatik) vom Autor des Dokuments verwendet worden sein, die mit den Absichten des suchenden Benutzers nicht übereinstimmen muss.

Ebenso können Dokumente mit synonymen Wörtern, die denselben Begriff mit anderen Worten beschreiben über eine einfache schlüsselwortbasierte Suche nicht gefunden werden, da inhaltsrelevante Dokumente das eigentliche Suchwort nicht enthalten müssen. Neben solchen Synonymen können Begriffe auch mit Metaphern und anderen sprachlichen Ausdrucksmitteln umschrieben werden, ohne dass der inhaltlich damit bezeichnete Suchbegriff in einem an sich relevanten Dokument auftauchen muss. Daher liegt der Schluss nahe, dass vielmehr die inhaltliche Bedeutung (Semantik) eines Dokuments und nicht nur die darin verwendeten Zeichenketten im Vordergrund einer inhaltsbasierten Suche stehen müssen.

Das gängige Paradigma des Information Retrievals setzt unter anderem voraus, dass der Informationssuchende tatsächlich genau weiss, was er sucht. Ist dies jedoch nicht der Fall und möchte der Suchende lediglich einen Überblick über die vorhandenen Informationsressourcen zu einem bestimmten Themengebiet erlangen, wird dies im Falle der WWW-Suche nahezu unmöglich aufgrund der schieren Masse an Informationsressourcen im Suchraum. Eine mögliche Variante, zumindest einen besseren Überblick über die erzielten Suchergebnisse zu erhalten, bieten statistische oder auf maschinellen Lernen beruhende Clusteringverfahren, die eine Sortierung und Filterung der erzielten Suchergebnisse nach weiteren, inhaltlichen Kriterien ermöglichen (Facettierte Suche). Aber das Blättern im Katalog gleich einem Bummel durch die Auslagen der Schaufenster einer Einkaufspassage ist mit einer schlüsselwortbasierten WWW-Suchmaschine dennoch nicht möglich.

3. Semantische Suche

3.1 Semantic Web, Linked Data und semantische Technologien

Einen Ausweg aus dem Dilemma der Informationssuche im WWW verspricht die vor über 10 Jahren gestartete Initiative des **Semantic Web**. Tim Berners-Lee, der ursprüngliche Entwickler des WWWs, kommentierte 1998 dessen bisherige Entwicklung kritisch. Das Web sei, so Berners-Lee, als gigantisches Informationsuniversum konzipiert gewesen, das nicht nur für die zwischenmenschliche Kommunikation geschaffen wurde, sondern den Menschen durch die aktive Teilnahme und Mitarbeit automatisierter und autonomer Computerprogramme unterstützen solle [Berners-Lee 1998]. Um dies zu erreichen, müssen die Inhalte der Informationsressourcen des WWWs maschinell gelesen und korrekt interpretiert, d.h. verstanden werden. Während das klassische Information Retrieval zu diesem Zweck Analysetechniken aus der Linguistik und Statistik bemüht, um Rückschlüsse auf den Inhalt natürlichsprachlicher oder multimedialer Dokumente zu ziehen, d.h. die Bedeutung (Semantik) implizit aus der vorhandenen Information erschließen, beschreitet das Semantic Web einen grundsätzlich anderen Weg und setzt darauf, die Semantik der Information selbst explizit zu machen. Dazu müssen Informationsressourcen mit Hilfe semantischer Metadaten annotiert werden, die die Bedeutung der Inhalte selbst formalisieren und kodieren, damit diese maschinell gelesen und korrekt interpretiert werden können.

Die Repräsentation der Semantik erfolgt mit Hilfe sogenannter „Ontologien“. Der Begriff Ontologie stammt aus der Philosophie und zählt zur Disziplin der Metaphysik, die sich primär mit dem Sein, dem Seienden als solchem und mit den fundamentalen Typen von Entitäten beschäftigt. Im Gegensatz zur Erkenntnistheorie (Epistemologie) beschreibt die Ontologie die Welt, wie sie tatsächlich ist, und nicht, wie sie uns gefiltert durch unsere Sinnesorgane und durch unsere persönliche Erfahrung erscheint. In der Informatik reduziert sich der Ontologiebegriff auf eine rein technische Sichtweise und bezeichnet eine „*explizite, formale Spezifikation einer gemeinschaftlichen Konzeptualisierung*“, d.h. ein abstraktes Modell (Konzeptualisierung), das alle relevanten Begriffe innerhalb einer Domäne und deren Beziehungen untereinander abbildet, wobei die Bedeutung aller Begriffe vollständig definiert werden muss (explizit) in einer maschinenlesbaren Form (formal) und Konsenz unter den kommunizierenden Parteien über die Bedeutung der Ontologie herrschen muss [Gruber 1993].

Einfache Beispiele für Ontologien aus unserem täglichen Leben sind z.B. Thesauri, also Wörterbücher, in denen inhaltliche Zusammenhänge zwischen einzelnen Begriffen aufgezeigt werden, wie z.B. Ober- und Unterbegriffe, Spezialisierungen und Verallgemeinerungen, Synonyme und assoziativ verknüpfte Begriffe. Zum Eintrag „Hose“ wäre in einem Thesaurus die „Textilie“ als Oberbegriff und die „Kniebundhose“ als eine der vielen möglichen Spezialisierungen (Unterbegriff) angegeben. Die „Hose“ ist ein Teil der „Kleidung“ und besteht selbst aus verschiedenen Einzelteilen, wie z.B. „Hosenbund“, „Hosentasche“, „Gürtelschlaufen“, usw. Als „Kleidungsstück“ können weitere Begriffe, wie z.B. das „Bein“, die „Mode“ oder der „Schneider“ mit der „Hose“ assoziiert sein.

Neben Thesauri existieren einfachere Taxonomien und Partonomien, hierarchisch aufgebaute Wissensrepräsentationen, in denen Ober- und Unterbegriffe bzw. Teil-Ganzes Beziehungen baumartig aufeinander aufbauen. Weiter können an einzelne Begriffe Regeln oder Bedingungen geknüpft werden, deren Gültigkeit sich formal überprüfen lässt.

So definiert die einfache Regel „Wenn A die Schwester von B ist, und C die Tochter von A, dann ist B die Tante von C“ eine Bedingung, die erfüllt sein muss, wenn zwischen zwei Entitäten die Beziehung „ist Tante von“ definiert wird.

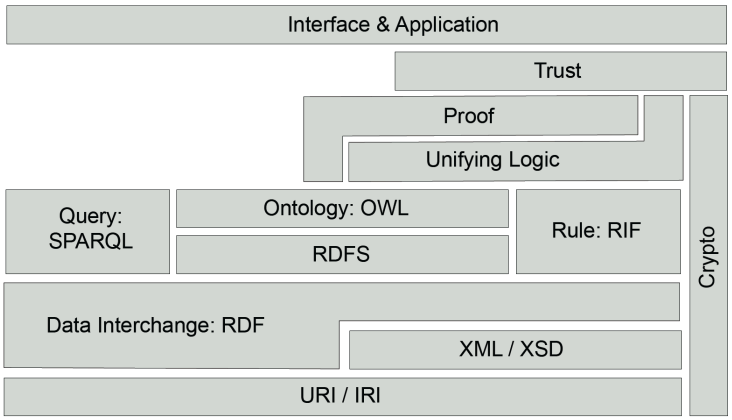

Abb. 1. Semantic Web Architekturmodell [Bratt 2007]

Im Semantic Web werden Ontologien mit Hilfe unterschiedlich ausdrucksstarker, formalsprachlicher Hilfsmittel realisiert, die hierarchisch in einer Schichtenarchitektur dargestellt werden können (siehe Abb. 1). Alle Ressourcen werden im Semantic Web mit Hilfe einer eindeutigen Adresse, eines Uniform Resource Identifiers (URI) identifiziert. Alle zur Wissensrepräsentation verwendeten Sprachen bauen auf der Extensible Markup Language (XML) als Sprache zur universellen Vokabulardefinition auf. Mit ihr lassen sich Klassen (XML Schema Definition Language) und Instanzen dieser Klassen (XML) definieren. Einfache Zusammenhänge zwischen Objekten lassen sich mit Hilfe des darauf aufbauenden Resource Description Frameworks (RDF und RDF Schema) festlegen. Ausdrucksstärkere Semantik über weitere Einschränkungen oder zu erfüllende Bedingungen und Abhängigkeiten zwischen Klassen und Instanzen können mit der Web Ontology Language (OWL) definiert werden. OWL selbst implementiert eine Beschreibungslogik (Description Logic), mit der eine formale Definition der im Semantic Web gebräuchlichen Wissensrepräsentationen erfolgt. Desweiteren wird ein Austauschdatenformat für logische Regeln definiert (Rules Interchange Format). In der Schichtenarchitektur folgen dann weitere Abstraktionsebenen mit logikbasierten Systemen, die es erlauben, aus dem vorhandenen Wissen neue Schlussfolgerungen zu ziehen oder auch dessen Konsistenz zu prüfen. Die Plausibilität des repräsentierten Wissens lässt sich anhand der Herkunft (Provenienz) überprüfen (Web of Trust Layer). Den Abschluss der Semantic Web Architektur bildet eine Anwendungsschicht, die die Schnittstelle zwischen Benutzer und Semantic Web festlegt und gestaltet. Die Standardisierung der einzelnen Architekturebenen ist bislang bis zur Ontologieschicht vorgedrungen (Stand Nov. 2009), alles darüberliegende ist, wie die Ontologieschicht selbst, noch Gegenstand der aktuellen Forschung. Das Semantic Web selbst ist als Ergänzung des bestehenden WWWs zu sehen, dessen Ressourcen um eine formale Beschreibung ihrer inhaltlichen Bedeutung erweitert werden, damit diese maschinell gelesen und weiterverarbeitet werden kann [Berners-Lee et al. 2001].

Eine besondere Stellung innerhalb des Semantic Webs nimmt aktuell der als „**Linked Data**“ bezeichnete Bereich ein. Dahinter verbergen sich semantisch annotierte Daten deren Grundlage strukturierte Datenbestände aus Datenbanken bilden, die bereits öffentlich über das WWW zugegriffen werden konnten. Zahlreiche Web-Portale und Websites basieren auf großen Datenbeständen, die mit Hilfe von Content Management Systemen und Datenbanken verwaltet werden. Allerdings sind diese Datenbanken jeweils unabhängig voneinander und zu unterschiedlichen Zwecken entwickelt worden. Die Semantik der darin befindlichen Daten wird lediglich implizit durch die darauf zugreifenden Programmen implementiert. Ein Ziel des Semantic Webs ist es jedoch, diese Semantik aus den Anwendungsprogrammen herauszuholen und mit in die Daten selbst zu integrieren, um so eine Unabhängigkeit dieser Daten von den jeweiligen Anwendungsprogrammen zu erreichen. Große Datenbestände strukturierter Daten werden als Linked Data in die semantische Markupsprache RDF überführt, damit sie interoperabel und unabhängig verwendet werden können. Damit lassen sich vormals nicht kompatible Datenbestände miteinander in Bezug setzen und gemeinsam in neuen Anwendungen nutzen. Die „DBPedia“ [DBPedia 2009], das Linked Data Pendant zur Online-Enzyklopädie Wikipedia, mit der große Mengen frei verfügbaren Wissens im Semantic Web zur Verfügung gestellt werden [Bizer et al. 2008], bildet den Nukleus eines oft als „Linked Open Data Cloud“ bezeichneten Netzwerks öffentlich verfügbarer Linked Data Ressourcen (siehe Abb. 2).

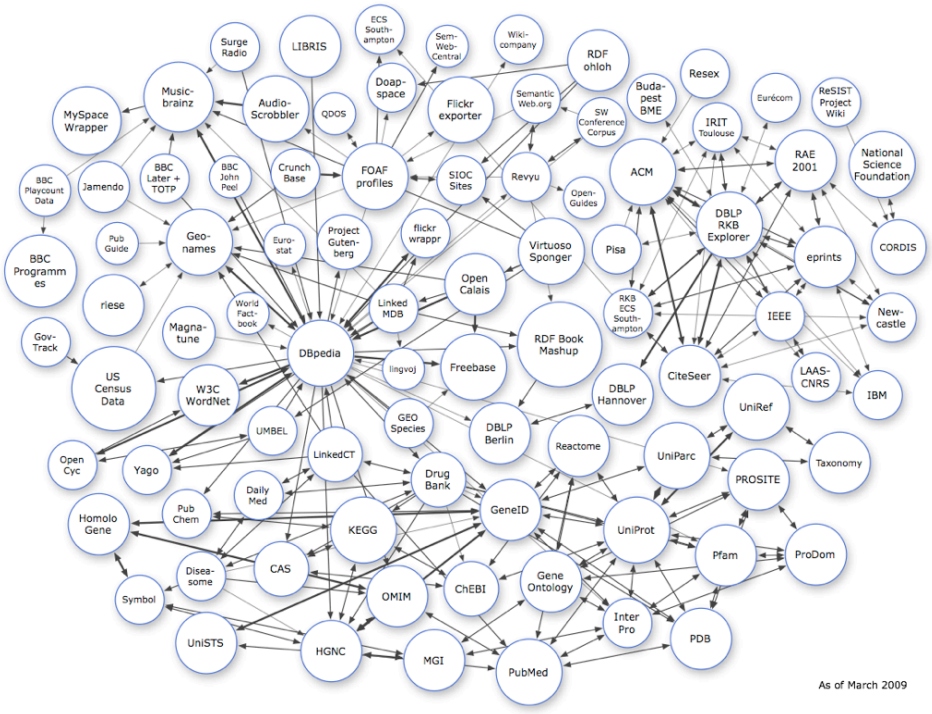

Abb. 2: Visualisierung der „Linked Open Data Cloud“ im Semantic Web [LOD 2009]

3.2 Semantisch unterstütztes Information Retrieval mit Hilfe von Ontologien und Linked Data

Explizite formale Semantik in Form von Semantic Web kompatiblen Metadaten können zur Verbesserung der Suche im WWW herangezogen werden. Dabei unterscheiden wir

die Nutzung von Ontologien im Sinne von Klassen, Beziehungen zwischen Klassen, sowie Einschränkungen, Bedingungen und Regeln, die an Klassen geknüpft sind, und die Nutzung einzelner Instanzen dieser Ontologien in der Form von Linked Data. Diese semantischen Metadaten können das klassische Information Retrieval auf die folgende Weise unterstützen:

- Sinnvolle und zielgerichtete Präzisierung und Erweiterung von Suchergebnissen (Query String Refinement)
- Herleitung von implizit vorhandener, verdeckter Information (Inference)
- Herstellung von Querverweisen und Assoziationen (Cross Referencing)
- Nutzung von semantischen Beziehungen zur Visualisierung und Navigation durch den Such- oder Ergebnisraum der Suche (Explorative Suche)

3.2.1 Präzisierung und Erweiterung von Suchergebnissen (Query String Refinement)

Im traditionellen Information Retrieval lassen sich im Kontext einer Suchanfrage einzelne Terme der Sucheingabe (Query String) mit Hilfe Boolescher Junktoren verknüpfen. Werden zwei Terme mit einem logischen „UND“ miteinander verknüpft, müssen beide Terme als Deskriptoren im Suchergebnis vorliegen, d.h. in einer textbasierten WWW-Suchmaschine müssen alle Ergebnisdokumente diese beiden Terme des Query Strings enthalten. Das Ergebnis entspricht der Schnittmenge der beiden Ergebnismengen, die jeweils durch einen der beteiligten Terme ermittelt wurde. Durch das Hinzufügen weiterer Einzelterme und deren logischer UND-Verknüpfung lässt sich das Suchergebnis also weiter einschränken, d.h. mit den richtigen Termen kann so eine **Präzisierung** des ursprünglichen Suchergebnisses erreicht werden. Verknüpft man auf diese Weise den mehrdeutigen Term „Bank“ mit dem Term „Finanzen“, werden die meisten Dokumente ausgefiltert, die den Term „Bank“ in einer seiner anderen Bedeutungen beinhalten.

Verknüpft man hingegen mehrere einzelne Terme über eine logische „ODER“ Verknüpfung, werden die im Ergebnis gelieferten Dokumente einen der beiden Terme bzw. sogar beide Terme enthalten. Die resultierende Dokumentenmenge entspricht der Vereinigungsmenge der Ergebnismengen, die den jeweiligen Einzeltermen zugeordnet sind. Durch Hinzunahme weiterer Einzelterme und deren logischer ODER-Verknüpfung lässt sich dementsprechend eine **Erweiterung** des Suchergebnisses erreichen.

Die Frage ist jedoch, welche Terme sind jeweils zur Präzisierung bzw. Erweiterung der Sucheingabe am besten geeignet? Wird als Einzelterm-Sucheingabe ein mehrdeutiger Term gewählt, enthält das Suchergebnis voraussichtlich viele nicht relevante Dokumente. Um diesen mehrdeutigen Term zu präzisieren, kann man bereits mit Hilfe eines Thesaurus bzw. einer Taxonomie eine entsprechende Filterung der Suchergebnismenge erreichen. Für den mehrdeutigen Begriff „Golf“ könnte so eine Präzisierung über die UND-Verknüpfung mit dem diskriminierenden Oberbegriff „PKW“ erfolgen. Dadurch werden nur noch Dokumente für das Ergebnis selektiert, die die Begriffe „Golf“ und „PKW“ enthalten, und daher „Golf“ im Sinne von „PKW“ konkretisieren. Die Wahrscheinlichkeit, dass sich Dokumente unter den Ergebnisdokumenten befinden, in denen der Begriff „Golf“ im Sinne von „Meeresarm“ oder „Sportart“ verwendet wird, ist meist

hinreichend gering. Ebenso können assoziierte Begriffe, wie z.B. „Straße“ oder „Führerschein“ für eine Präzisierung eingesetzt werden. Diese können über einen Thesaurus oder eine Domain-Ontologie ermittelt werden, die den Begriff „Golf“ im intendierten Sinne beinhaltet.

Eine semantische Suche kann in diesem Zusammenhang unterschiedliche Möglichkeiten der Präzisierung des gesuchten Begriffes vorschlagen oder die dazu ermittelten Suchergebnisse direkt in differenzierter Weise zur Darstellung bringen. Die konkrete Festlegung auf eine eindeutige Bedeutung erfolgt dann über die vom Benutzer durchgeführte Auswahl. In der gleichen Weise kann die semantische Suche eine Erweiterung des Suchergebnisses in unterschiedliche Bedeutungskontexte vorschlagen. Dies ist insbesondere dann von Vorteil, wenn sich unter den erzielten Suchergebnissen zu wenige bzw. keine für den Benutzer relevanten Ergebnisse befinden. In diesem Fall hilft eine gezielte Erweiterung des Suchraums durch die Beistellung synonymer oder bedeutungsähnlicher Begriffe aus einem Thesaurus oder einer geeigneten Domain-Ontologie. Um etwa die Suchphrase „Bank“ zu erweitern könnte man diesen mit synonymen Termen, wie z.B. „Kreditanstalt“ oder „Sparkasse“, mit assoziierten Termen, wie z.B. „Konto“ oder Kredit“, oder aber auch mit konkreten Ausprägungen (Instanzen) der Klasse „Bank“, wie z.B. „Raiffeisen“ oder „Hypobank“ erweitern, um die Anzahl der erzielten relevanten Suchergebnisse zu erhöhen.

3.2.2 Herleitung von implizit vorhandener, verdeckter Information (Inference)

Durch die Ergänzung des ursprünglichen Query Strings mit Termen, die aus relevanten Ontologien stammen, wird nicht nur eine zielgenauere Suche ermöglicht, sondern vielmehr auch eine assoziativ motivierte Suche, die anhand impliziter Zusammenhänge Naheliegendes erschließt und aufdeckt, und so dem Suchenden einen Einblick in vorhandene Informationen gewährt, die er über eine traditionelle Informationssuche nie gefunden hätte.

Prinzipiell unterscheidet man hier zwischen zwei unterschiedlichen Arten der Herleitung impliziter Information:

- Die häufigste Form ist dabei das **deduktive Reasoning**, bei dem aus explizit gespeicherten Fakten auf implizites Wissen geschlossen wird. Ist z.B. die Entität „Alice“ eine Instanz der Klasse „Mutter“, und die Klasse „Mutter“ eine Unterklasse der Klasse „Frau“, kann daraus gefolgert werden, dass „Alice“ ebenfalls eine „Frau“ ist.
- Umgekehrt werden mit Hilfe des **induktiven Reasonings** aus vorhandenem Faktenwissen allgemeinere Behauptungen aufgestellt. Sei z.B. die Entität „Alice“ eine Instanz der Klasse „Frau“ und sei „Alice“ mit der Entität „Franz“ über die Eigenschaft „hatKind“ verbunden. „Barbara“ sei ebenfalls eine Instanz der Klasse „Frau“, allerdings ohne eine Verbindung zu einer weiteren Instanz über die Eigenschaft „hatKind“. Dann kann eine neue Klasse „Mutter“ aus dem positiven Beispiel „Alice“ und dem negativen Beispiel „Barbara“ gelernt werden. Mit Hilfe einer Beschreibungslogik könnte der Sachverhalt folgendermaßen ausgedrückt werden:
    - $\text{Alice} \in \text{Frau}, \text{hatKind}(\text{Alice}, \text{Franz}), \text{Barbara} \in \text{Frau}$
      $\text{Mutter} \sqsubseteq (\text{Frau} \sqcap \exists \text{hatKind})$

3.2.3 Herstellung von Querverweisen und Assoziationen (Cross Referencing)

Auf ähnliche Weise lassen sich Querverweise und Assoziationen zwischen Instanzen oder Klassen ermitteln. Dabei geht es prinzipiell darum, zusätzliche Suchergebnisse bereitzustellen, die zwar den Suchbegriff nicht direkt enthalten, aber mit diesem inhaltlich im Zusammenhang stehen. Grundlage ist dabei ebenfalls wieder eine Domain-Ontologie bzw. Thesauri oder auch Kookurenzanalysen in repräsentativen Dokumentenkorpora, mit deren Hilfe Zusammenhänge zwischen zwei Entitäten gefunden werden, die sich nicht auf den ersten Blick erschließen. Dabei wird versucht, zwei Instanzen einer gemeinsamen Klasse bzw. Oberklasse zuzuordnen. Anhand gemeinsamer Ausprägungen von Eigenschaften, die diesen Instanzen über ihre Klassenzugehörigkeit zugeordnet werden können, lassen sich implizite Zusammenhänge als Querverbindungen entdecken.

Wird z.B. nach dem Einzelterm „Hemingway“ gesucht, muss zunächst erkannt werden, dass es sich dabei um den Namen eines bekannten US-amerikanischen Autors handelt (Instanzerkennung). Die konkrete Entität „Ernest Hemingway“ kann z.B. über die Linked Data Ressource „DBPedia“ einer Ontologie zugeordnet werden und gehört dort der Klasse „amerikanische Autoren“ an. Wenn also bekannt ist, dass „Ernest Hemingway“ ein „amerikanischer Autor“ ist, können weitere Instanzen dieser Klasse bestimmt werden, wie z.B. „Edgar Allan Poe“, deren gemeinsame Klassenzugehörigkeit eine assoziative Verbindung zwischen diesen ausdrückt (siehe Abb. 3). Mit dem Namen dieser Instanzen, deren Herleitung implizit erfolgte, kann die Suchanfrage erweitert werden.

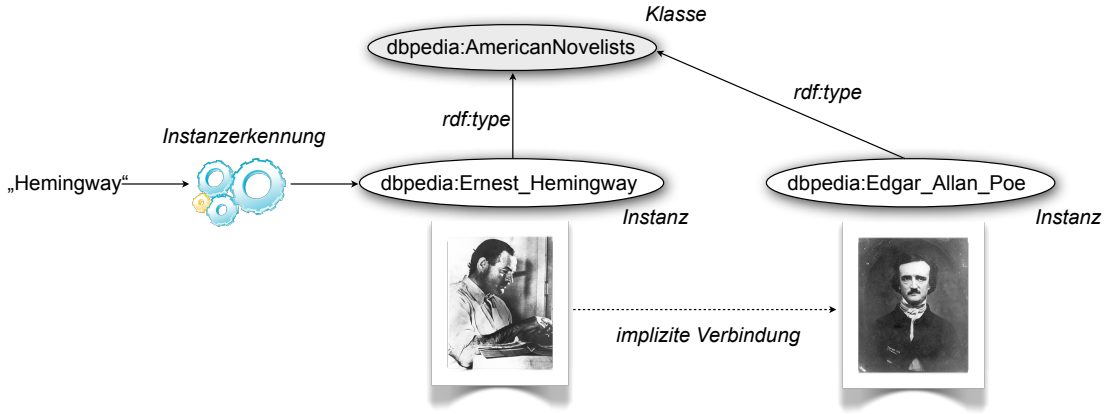

Abb. 3. Herstellung von Querverweisen mit Hilfe der Linked Data Ressource DBPedia (Fotos: wikipedia.org)

3.2.4 Explorative Suche als neues Suchparadigma

Aktuelle WWW-Suchmaschinen orientieren sich am klassischen Suchparadigma, d.h. um eine gesuchte Informationsressource zu finden, gibt der Benutzer in seiner Suchabfrage einen für die gesuchte Informationsressource charakteristischen Suchbegriff ein. Im Falle des textbasierten Information Retrievals werden Textdokumente zurückgeliefert, die den Suchbegriff tatsächlich enthalten. Der Benutzer geht also von der Annahme

aus, dass die zu findenden Informationsressourcen den verwendeten Suchbegriff beinhalten. Der Benutzer muss also in etwa wissen, was er sucht.

Bei der Suche in einem Bibliothekskatalog sucht der Benutzer nach Autoren, Titeln, Verlagen oder Schlagworten. Entweder kennt der Benutzer den Namen des betreffenden Autors bzw. den Buchtitel, oder aber er versucht sich an einer thematischen Zuordnung des von ihm gesuchten Werkes und schlägt diese im Schlagwortkatalog nach, in dem den Informationsressourcen von autoritativer Stelle, d.h. vom Autor, dem Verleger oder dem Bibliothekar Schlagwörter zugeordnet wurden. Da aber Benutzer und Schlagwortautor unterschiedlicher Auffassung über die treffende Zuordnung von Schlagwörtern sein können, ist diese Variante der Suche nicht immer zielführend.

Anders ist die Situation, wenn der Benutzer nicht genau weiß, was er sucht. Wenn er sich z.B. erst einmal einen Überblick über die zu einem Themengebiet vorhandenen Informationsressourcen bzw. über den gesamten Suchraum verschaffen möchte. Im WWW ist dies heute aufgrund der vorhandenen Dokumentenmenge unmöglich. In der Bibliothek dagegen hat der Benutzer die Möglichkeit, die Bücherregale selbst zu betrachten, in denen die vorhandenen Informationsressourcen entsprechend einer vorgegebenen Systematik eingeordnet wurden. So kann er innerhalb eines Themengebiets „herumstöbern“ und dabei zufällig auf Bücher stoßen, die ihn interessieren, wobei er sich dessen zuvor gar nicht bewusst war. Diese Möglichkeit der „zufälligen und glücklichen Entdeckung“ wird im Englischen auch als „**Serendipity**“ bezeichnet. Es geht also darum, Suchergebnisse zu entdecken, nach denen der Benutzer zunächst gar nicht gesucht hatte. Diese Art der zielgerichteten Erkundung des Suchraums ist uns aus unserem täglichen Leben vertraut und wird auch als „**explorative Suche**“ bezeichnet, die sich mit Hilfe von Domain-Ontologien und Linked Data Ressourcen realisieren lässt.

Über thematisch passende Domain-Ontologien werden Klassenzugehörigkeiten und Beziehungen von Klassen untereinander repräsentiert, die sowohl zwischen den Instanzen als auch zwischen den Klassen explizite und implizite inhaltliche Verbindungen herstellen. Diese werden dazu genutzt, die traditionelle Informationssuche zu erweitern und Wege entlang dieser Verbindungen aufzuzeigen, mit deren Hilfe der vorhandene Suchraum zielgerichtet erkundet werden kann [Sack 2005]. So wird die Informationssuche selbst zu einem Sucherlebnis, angetrieben durch die Neugierde des Benutzers, die sie auf der anderen Seite aber auch zu beflügeln vermag.

4. yovisto.com - eine semantische, akademische Videosuchmaschine

Ein Beispiel für die im vorangegangenen Kapitel beschriebene explorative Suche soll anhand der Videosuchmaschine yovisto.com aufgezeigt werden. Das in dieser Videosuchmaschine verwaltete Videomaterial beschränkt sich aktuell auf ca. 8.000 Aufzeichnungen universitärer Lehrveranstaltungen und wissenschaftlicher Vorträgen, schwerpunktmäßig in deutscher und englischer Sprache.

4.1 Inhaltsbasierte Videosuche in aufgezeichneten Lehrveranstaltungen

Die Videosuchmaschine yovisto ermöglicht eine zielgenaue und inhaltsbasierte Suche in den verwalteten Videoressourcen. Dies wird durch eine komplexe, automatisierte inhalt-

liche Analyse der Videoaufzeichnungen mit daraus gewonnenen zeitbezogenen Metadaten erreicht. Die Videoanalyse umfasst dabei folgende Technologien:

- **automatische Segmentierung**: Die Videoaufnahme wird in einzelne, inhaltlich kohärente Sequenzen unterteilt. Den einzelnen Sequenzen werden inhaltliche Metadaten zugeordnet.
- **intelligente Schrifterkennung** (Intelligent Character Recognition, ICR): Zu jeder erkannten Bildsequenz werden repräsentative Einzelbilder (Key-Frames) bestimmt, die den Inhalt der Sequenz möglichst gut repräsentieren. Universitäre Vorlesungen und wissenschaftliche Vorträge werden heute meist von textbasierten Präsentationen (Folien, Desktop-Präsentation, Tafelanschrieb, etc.) unterstützt, in denen die inhaltlich wichtigsten Punkte zusammengefasst werden. Diese Texte werden im Videobild identifiziert, mit Hilfe geeigneter Texterkennungsmethoden extrahiert und als Metadaten verwendet.
- **Audioanalyse**: Zusätzlich kann eine Spracherkennung (Automated Speech Recognition) verwendet werden, die eine (fehlerbehaftete) Transkription der gesprochenen Inhalte einer Sequenz erlaubt. Aufgrund der meist schlechten Aufnahmebedingungen (kein professionell ausgebildeter Sprecher, keine Studiobedingungen, Störgeräusche, etc.) enthalten die erkannten Texte zahlreiche Fehler und sind qualitativ im Vergleich zu den mittels ICR erzielten Metadaten nebenrangig.

Zusätzlich erlaubt yovisto eine benutzergenerierte, zeitbasierte Verschlagwortung (Tagging) der Videoinhalte, die nicht-autoritative zeitbezogene Metadaten für die Videosuche generiert. Bei der einfachen Videosuche wird auf eine Suchphrase hin der vorhandene Metadatenbestand durchsucht und dem Benutzer eine Auswahl relevanter Suchergebnisse präsentiert, die einen zielgenauen Zugriff auf die gewünschten Inhalte innerhalb der einzelnen Videos ermöglicht [Sack et al. 2006].

4.2 Einsatz semantischer Technologien zur explorativen Videosuche in yovisto.com

Zur qualitativen Verbesserung der Videosuche in yovisto.com wurde ein erster Prototyp zur explorativen Videosuche auf Basis semantischer Technologien implementiert [yovisto 2009]. Grundlage der semantischen Suche ist eine vorangegangene semantische Videoanalyse. Dabei werden aus den bereits vorhandenen textuellen Metadaten Schlüsselwörter ausgewählt, die einer Linked Data Entität zugeordnet werden können. Diese Abbildung wird automatisch vorgenommen und umfasst bislang noch keine Disambiguierung, so dass Homonyme mehreren Entitäten zugeordnet werden können. Eine Disambiguierung erfolgt entweder manuell durch einen Benutzer oder aber automatisch mit Hilfe einfacher statistischer Verfahren (Koreferenz- und Kontextanalyse, Clustering, Machine Learning). So können (textuelle) Schlüsselwörter mit strukturierten Daten und semantischen Wissensrepräsentationen ergänzt werden, die die Grundlage einer explorativen Suche bilden.

Zur explorativen Suche werden explizite und implizite inhaltliche Zusammenhänge einzelner Entitäten genutzt, d.h. zu den jeweils vorhandenen semantischen Metadaten einer Informationsressource werden weitere Metadaten bestimmt, die mit diesen inhaltlich zusammenhängen. Ist also z.B. ein Videosegment mit dem Schlüsselwort „Stephen King“ annotiert, wird über eine Verknüpfung mit den DBPedia-Daten die DBPedia-Entität des US-amerikanischen Autors „Stephen King“ bestimmt und mit dem Schlüssel-

wort verknüpft. Über die enzyklopädischen Daten der DBPedia werden zusätzliche Informationen, wie z.B. das literarische Genre des Autors („Fantasy“, „Science Fiction“) , sein Geburtsort („United States“) oder auch andere verwandte Autoren („Edgar Allan Poe“) sowie weitere assoziativ verbundene Entitäten (z.B. „Maine“, „Desperation“, „Author“, „Pseudonym“, etc.). Zusätzlich erfolgt ein automatischer Abgleich, ob zu diesen verknüpften Begriffen überhaupt Videosegmente in der zugrundeliegenden Datenbank vorhanden sind und wieviele Suchtreffer diesbzgl. erzielt werden können. Begriffe, zu denen keine Videosegmente gefunden werden können, werden sofort ausgefiltert.

Abb.4: Explorative Videosuche im semantischen yovisto-Prototyp mit dem Suchbegriff „Stephen King“ und einer Detailvergrößerung des explorativen Navigationselements

Eine einfache explorative Navigationshilfe wird in Abb. 4 dargestellt. Links neben der eigentlichen Trefferliste für den Suchbegriff „Stephen King“ werden weiterführende Suchbegriffe und die dazu vorhandene Anzahl an Informationsressourcen angezeigt, für die ihrerseits durch Anklicken erneut eine Suche ausgelöst werden kann. Dabei werden qualifizierte Assoziationen, bei denen die Beziehung zwischen Suchbegriff und assoziiertem Begriff benannt werden kann, von unqualifizierten Assoziationen ohne Nennung der verknüpfenden Beziehung unterschieden. So werden die mit dem ursprünglichen Suchbegriff in Bezug stehenden Begriffe als Navigationselement verwendet, mit dem eine explorative Suche im vorhandenen Gesamtdatenbestand durchgeführt werden kann [Waitelonis et al. 2009].

5. Ausblick

Die im vorangegangenen Kapitel vorgestellte Implementation einer explorativen Suche befindet sich bislang noch in einem prototypischen Entwicklungszustand. Standardevaluationen des Information Retrievals folgen ebenfalls dem bereits beschriebenen Paradigma, dass der Benutzer weiß, wonach er sucht. Um einen quantitativen Nachweis zur verbesserten Suchqualität der explorativen Suche zu erbringen, werden zu diesem Zweck aktuell spezielle Evaluationsverfahren entwickelt.

Darüberhinaus wird aktuell an einer weiterführenden Nutzung der vorhandenen semantischen Metadaten gearbeitet, bei der die erkannten Assoziationen und Beziehungen auch im Rahmen eines Vorschlagmechanismus zur benutzerunterstützten Schlüsselwortvergabe (Suggested Tagging) und zur Disambiguierung verwendet werden. Die prototypische Implementierung bietet noch großes Potenzial zur Verbesserung der Auswahl tatsächlich relevanter inhaltlicher Zusammenhänge und ebenso in der grafischen Aufbereitung und Darstellung der Ergebnisse, die dem Benutzer das spielerische Erkunden und zufällige Entdecken interessanter Suchergebnisse ermöglichen soll.

Semantische Technologien ermöglichen zusätzlich die Realisierung intelligenter personalisierter Vorschlagsmechanismen (Recommender Systems). Über inhaltliche Zusammenhänge der vom Benutzer bereits ausgewählten Suchergebnisse können diese mit inhaltlich „ähnlichen“ bzw. im Zusammenhang stehenden Ressourcen verknüpft und entsprechend des persönlichen Informationsbedarfs als neues Ergebnis vorgeschlagen werden. Derartige Techniken finden bereits vor allem im Musikbereich Anwendung. Interessant für den Benutzer ist aber nicht nur ein eventuell relevanter Vorschlag, sondern auch der Grund, warum ausgerechnet diese Ressource vorgeschlagen wurde. Die konzise Darstellung dieser oft komplexen Zusammenhänge (Story Telling) ist für den Benutzer hinsichtlich der Akzeptanz vorgeschlagener Suchergebnisse von besonderem Interesse.

Generell gewinnt die Suche in multimedialen Daten zunehmend an Bedeutung. Insbesondere mobile Endgeräte mit beschränkten Verarbeitungs- und Darstellungskapazitäten stellen dabei eine besondere Herausforderung dar, bei der neue Methoden der inhaltlichen Visualisierung von Einzelmedien bzw. von ganzen Suchergebnismengen auf engem Raum aber in übersichtlicher und rasch aufzufassender Weise besonders wichtig sind.

6. Literaturangaben

[Berners-Lee 1989] Berners-Lee, T.: Information Management - A Proposal, http://www.w3.org/History/1989/proposal.html (Zugriff am 28.11.2009)

[Berners-Lee 1998] Berners-Lee, T.: Semantic Web Roadmap, Sept 1998, http://www.w3.org/DesignIssues/Semantic.html (Zugriff am 30.11.2009)

[Berners-Lee et al. 2001] Berners-Lee, T.; Hendler, J.; Lassila, O.: The Semantic Web. Scientific American, (284)5:34-43, 2001.

[Bizer et al. 2008] Bizer, C.; Heath, T.; Idehen, K.; Berners-Lee, T.: Linked data on the web. In Proceedings of the 17th International Conference on World Wide Web (WWW), 1265-1266, ACM, 2008.

[Bratt 2007] Bratt, S.; "Semantic Web, and Other Technologies to Watch, slide 24", World Wide Web Consortium, 2007, http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24) (Zugriff am 30.11.2009)

[Brin et al. 1998] Brin, S.; Page, L.: The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, (30)1-7:107--117, 1998.

[DBPedia 2009] DBPedia: http://dbpedia.org/ (Zugriff 30.11.2009)

[Google 2008] The Official Google Blog: We knew the Web was Big..., 25.7.2008, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html (Zugriff am 30.11.2009)

[Gruber 1993] Gruber, T. R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies, 43(5-6):907-928, 1995.

[LOD 2009] Linked Data - Connect Distributed Data accross the Web, http://linkeddata.org/ (Zugriff am 30.11.2009)

[Ribeiro-Neto et al. 1999] Ribeiro-Neto, B.; Baeza-Yates, R.: Modern Information Retrieval, ACM Press / Addison-Wesley, 1999.

[Sack 2005] Sack, H.: NPBibSearch: An Ontology Augmented Bibliographic Search, in Proc. of SWAP 2005, the 2nd Italian Semantic Web Workshop, Trento, Italy, December 14-16, 2005, CEUR Workshop Proceedings, ISSN 1613-0073.

[Sack et al. 2006] Sack, H.; Waitelonis,J.: Integrating Social Tagging and Document Annotation for Content-Based Search in Multimedia Data, in Proc. of the 1st Semantic Authoring and Annotation Workshop (SAAW2006), Athens (GA), USA, 2006.

[Waitelonis et al. 2009] Waitelonis, J.; Sack, H.: Towards Exploratory Video Search by Using Linked Data, in Proc. of 2nd IEEE International Workshop on Data Semantics for Multimedia Systems and Applications (DSMSA2009), December 14-16, 2009, San Diego, California, 2009.

[yovisto 2009] Prototyp zur explorativen Suche in yovisto.com, http://testing.yovisto.com/ (Zugriff am 30.11.2009)