

# ConWeaver – Automatisierte Wissensnetze für die semantische Suche

Dr. Andrea Dirsch-Weigand  
Ingrid Schmidt  
Birgit Rein  
Richard Stenzel  
Dr. Thomas Kamps

*Fraunhofer Institut Integrierte Publikations-  
und Informationssysteme  
Dolivostraße 15  
64293 Darmstadt*

## **Abstract Deutsch**

*Google ist zum Inbegriff einer Suchmaschine geworden. Doch ist in Fachkreisen klar, dass Volltextsuchmaschinen wie Google auch deutliche Schwächen aufweisen und deshalb für die effiziente Suche in Fachportalen, Intranets und Enterprise-Content-Management-Systemen oft nicht ausreichen.*

*Weil Volltextsuchmaschinen nur mit dem genauen Wortlaut suchen, finden sie einerseits Informationen nicht, die zwar dem Inhalt, nicht aber den genauen Formulierungen der Suchanfrage entsprechen. Bezeichnungsalternativen, sprachlichen Varianten sowie fremdsprachliche Benennungen werden nicht als bedeutungsgleich erkannt. Andererseits entstehen unpräzise Suchergebnisse, weil gleichlautende Bezeichnungen unterschiedlicher Bedeutung nicht unterschieden werden.*

*Diese Probleme geht die semantische Suchmaschine ConWeaver an, die das Fraunhofer Institut Integrierte Informations- und Publikationssysteme (Fraunhofer IPSI) in Darmstadt entwickelt hat. An Stelle eines Volltextindexes setzt sie ein Wissensnetz als Suchindex ein. Die Software ConWeaver baut dieses Wissensnetz im Unterschied zu den meisten anderen ontologiebasierten Softwareprodukten weitgehend automatisch auf.*

## **Abstract Englisch**

*Google has become the embodiment of a search engine. But information professionals are aware that full text search engines like Google reveal evident deficiencies and are often not the best solution for efficient information search in technical information portals, intranets and enterprise content management systems.*

*Full text search engines are glued to the exact wording of a search phrase. On the one hand, they don't find information with different wording but same content because they don't recognize alternative denominations, phrasing or foreign-language translations to be synonym to the original wording. On the other hand, search results are not precise because full text search engines don't make a difference between homonymic but semantically different words.*

*The semantic search engine ConWeaver has been developed by the Fraunhofer Integrated Publication and Information Institute (Fraunhofer IPSI) at Darmstadt to address this problem. ConWeaver search solutions use a semantic network as search index in stead of a full text index. Unlike most other ontology-based software products ConWeaver generates this semantic network in an automated information extraction process.*

## **Semantische Suche versus Volltextsuche**

Kennen Sie solche Situationen? Beim Herumtollen mit Ihren Kindern landet ein Ellbogen versehentlich unsanft in Ihrem Auge und prompt ist ein Äderchen geplatzt. Was tun? Sie suchen Rat im Internet und geben in die Suchmaschine Google die Anfrage *Erste Hilfe bei Gefäßverletzung am Auge* ein. Sie erhalten 100 Treffer angezeigt, doch können sie mit den Treffern auf der ersten beiden Seite nichts anfangen: Sie verweisen auf Gefäßverletzungen an anderen Körperteilen oder behandeln andere Augenverletzungen.

Warum findet Google hier so wenige relevante Treffer? Als Volltextsuchmaschine „klebt“ Google am genauen Wortlaut einer Anfrage. Die alternative Formulierung *rotes Auge* hätte bessere Ergebnisse erbracht, wie jeder leicht selbst nachprüfen kann. Abgesehen davon, dass vielen Nutzern dieses Defizit nicht bewusst ist, sind die allermeisten Anwender nicht bereit, einem Suchinteresse mit allen möglichen Formulierungen nachzugehen. Vor allem für Unternehmensanwendungen besteht der Anspruch, dass passende Information schon im ersten Formulierungsanlauf gefunden werden muss.

So effizient können nur semantische Suchlösungen arbeiten. Im Unterschied zu Volltextsuchmaschinen erkennen sie, dass *Blutgefäßverletzung am Auge* gleichbedeutend ist mit *Gefäßverletzung am Auge* oder *geplatzt es Äderchen im Auge* und bedeutungsähnlich mit *Rotem Auge*. Semantische Suchlösungen liefern ein vollständigeres und zugleich präziseres Ergebnis als eine Volltextsuche.

ConWeaver ist eine Software, mit der sich semantische Suchlösungen für Portale und Intranets entwickeln lassen. Kern dieser Suchlösungen ist ein automatisiert erstelltes Wissensnetz, das als globaler semantischer Suchindex über alle angeschlossenen Datenquellen dient. ConWeaver-Lösungen erstellen und pflegen ein solches Wissensnetz weitgehend automatisch, damit eine hohe Aktualität gewährleistet ist und die Kosten gering bleiben.

## **Beispiel: Suche in einem Gesundheitsportal**

Ein Beispiel soll im Folgenden verdeutlichen, was die semantische Suche mit automatisierten Wissensnetzen leistet:

Eine Krankenkasse richtet im Internet ein Gesundheitsportal ein. In diesem Fachportal sollen Nutzer - vom medizinischen Laien bis zum Pflegefachpersonal - hochqualitative Auskunft zu allen Gesundheitsthemen erhalten und an die Ärzte und Kliniken verwiesen werden, mit denen die Krankenkasse bevorzugt zusammenarbeitet.

Die Datenbasis bilden einerseits hoch strukturierte Datenquellen: eine Medikamentendatenbank, eine Ärztedatenbank und eine Krankenhausdatenbank sowie eine Klassifikation zu Krankheiten. Zudem sind Listen mit Verbänden und Selbsthilfegruppen im Gesundheitssektor einzubinden. Andererseits sollen schwach oder kaum strukturierte

Informationsressourcen in das Portal integriert werden: verschiedene Lexika zu Gesundheitsfragen und ein umfangreicher Bestand an Volltexten aus Fachzeitschriften.

Im Unterschied zu bestehenden Gesundheitsportalen<sup>1</sup> soll die Portaloberfläche nicht nur eine gemeinsame „Eingangshalle“ für die angebotenen Quellsysteme sein, die im Weiteren dann doch jeweils wieder eine spezielle Suche erfordern. Das Portal soll vielmehr eine semantische Suche bieten, die die Suchergebnisse aus den unterschiedlichen Quellen mit einer einzigen Anfrage in einem integrierten Suchergebnis zusammenführt und inhaltlich strukturiert. Durch diese Suchfunktionalität soll das Portal sozusagen die Arbeit eines Referenten leisten, der jedes Anfragethema mit einem gut sortierten Dossier beantwortet.

Für eine solche Suchlösung werden zunächst die vorhandenen Datenquellen mit Hilfe eines weitgehend automatisch erstellten Wissensnetzes integriert. Für dieses Wissensnetzes wird im Anschluss die semantische Suche konfiguriert, die Anfragen in vielfältigen laien-, fach- und fremdsprachlichen Formulierungen verarbeitet und ein integriertes Suchergebnis in der gewünschten inhaltlichen Struktur erbringt.

### ***Automatisierter Aufbau eines Wissensnetzes***

Ein Wissensnetz ist eine geordnete und formalisierte Zusammenstellung von Konzepten, individuellen Ausprägungen von Konzepten, Bezeichnungen und Beziehungen. Durch die Darstellung eines inhaltlichen Zusammenhangs mit Hilfe von Relationen werden Konzepte und ihre individuellen Ausprägungen für Menschen und Maschinen erst verständlich und erhalten eine spezifische Bedeutung, eine Semantik (daher oft auch die Bezeichnung *semantisches Netz* für ein Wissensnetz).

Ein Wissensnetz umfasst drei logische Schichten.

Im Kern besteht es aus einem Konzeptnetz zu einem bestimmten Wissensgebiet. Dieses Konzeptnetz wird auch oft Ontologie genannt. Es beschreibt die jeweiligen abstrakten Konzepte eines Wissensgebietes. In unserem Beispiel sind solche Konzepte Begriffe wie KRANKHEIT oder WIRKSTOFF.

Dieses Konzeptnetz wird durch eine zweite Schicht mit individuellen Ausprägungen der Konzepte und konkreten Faktendaten erweitert. Individuelle Ausprägung des Konzeptes MEDIKAMENT ist beispielsweise *Benuron* als ein ganz konkretes Medikament gegen Fieber und Schmerzen. Fakten sind zum Beispiel der Preis oder die Verschreibungspflichtigkeit von *Benuron*. Solche Fakten werden als Eigenschaften von Individuen (Attribute) modelliert.

Die dritte Schicht des Wissensnetzes besteht aus einem Themennetz, das Themenvokabular mit seinen sprachlichen und inhaltlichen Zusammenhängen und Strukturen verzeichnet (Abb.1).

---

<sup>1</sup> z.B. [www.gesundheitpro.de](http://www.gesundheitpro.de), [www.das-gesundheitsportal.com](http://www.das-gesundheitsportal.com), [www.meine-gesundheit.de/](http://www.meine-gesundheit.de/), [www.deutschlandmed.de](http://www.deutschlandmed.de), [www.medizin-forum.de/](http://www.medizin-forum.de/), [www.netdokter.de](http://www.netdokter.de)

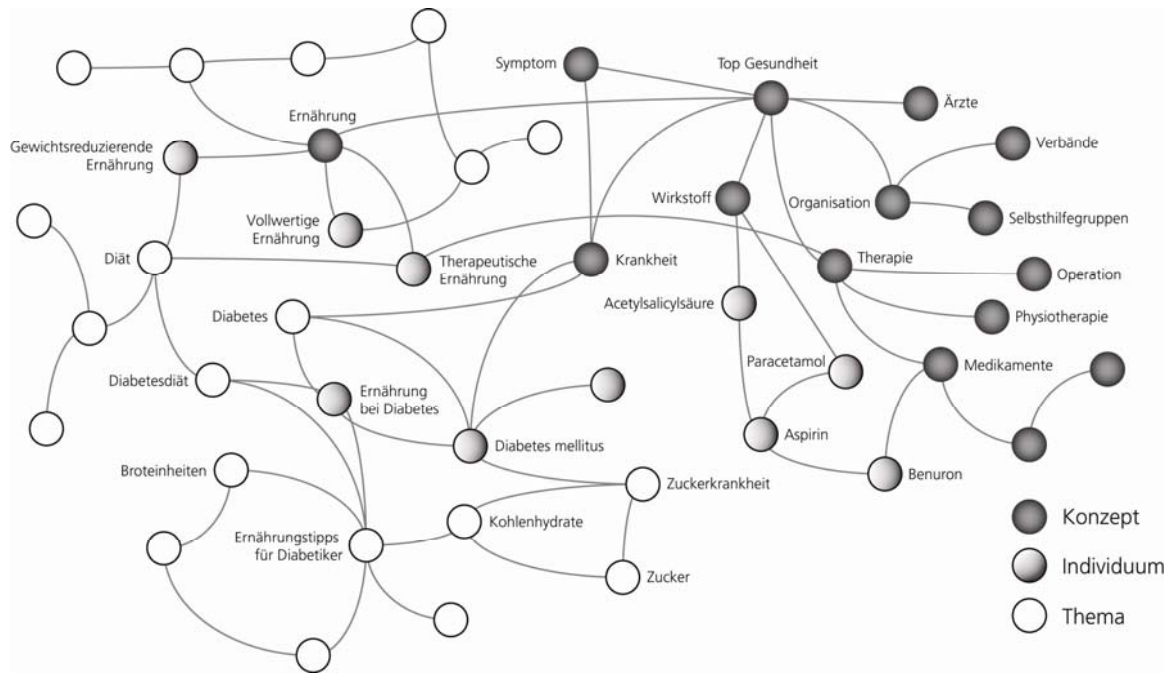


Abb. 1: Wissensnetz

Das Wissensnetz wird stufenweise aufgebaut.

In einem ersten Schritt werden die wichtigen Konzepte des relevanten Sachgebietes identifiziert und in einer Hierarchie aus Ober- und Unterkonzepten angeordnet. Solche Konzepte lassen sich aus vorhandenen Wissensstrukturen wie Fachklassifikationen und Datenbankschemata, Verzeichnisstrukturen in Portalen und Dokumentablagen herleiten. In unserem Beispiel sind die Medikamentendatenbank, Ärztedatenbank und Krankenhausdatenbank sowie die Listen und die medizinischen Klassifikation relevante Ressourcen. Die Konzepte werden nicht nur hierarchisch klassifiziert, sondern auch mit weiteren bedeutungstragenden Beziehungen untereinander verknüpft. Das Konzept KRANKHEIT wird zum Beispiel über die Beziehung *hatSymptom* mit dem Konzept SYMPTOM verbunden. Damit wird aus der Konzepthierarchie ein Konzeptnetz, aus einer Klassifikation wird eine Ontologie. Das Konzeptnetz wird intellektuell erstellt und umfasst bis zu 100 Knoten.

In einem zweiten Schritt werden den Konzepten individuelle Ausprägungen der Konzepte zugeordnet. Ausprägungen von Konzepten sind beispielsweise *Benuron* oder *Aspirin* als Individuen von MEDIKAMENT. Als MEDIKAMENT sind für sie bereits auf der Konzeptebene Eigenschaften wie *Preis* oder *Verschreibungspflichtigkeit* definiert, deren Werte Fakteninformationen zu den Individuen liefern. Die Individuen erben aber nicht nur die Eigenschaften, die für das jeweilige Konzept definiert wurden, sondern auch alle Relationen. Auf diese Weise wird zum Beispiel die Relation *Benuron* (Individuum von MEDIKAMENT) *hatWirkstoff Paracetamol* (Individuum von WIRKSTOFF) und die Fakteninformation *Benuron – Verschreibungspflicht: nein* im Wissensnetz modelliert.

Die Individuen mit ihren Eigenschaftswerten kommen aus den Datenbankfeldern der Ärzte-, Krankenhaus- und Medikamentendatenbanken. Das Mapping von Datenbankstrukturen und Konzepten erfolgt intellektuell. Der Import der Feldinhalte aus den Datenbanken läuft danach vollautomatisch ab. Insgesamt wird dieser Teil des Wissensnetzes also halbautomatisch erstellt. Die importierten Individuen ergeben erfahrungsgemäß einige zehntausend Knoten im Netz.

Im dritten Schritt werden Konzeptnetz und Individuennetz mit spezifischem thematischem Vokabular erweitert. Dieses Vokabular stammt – soweit vorhanden – aus Terminologien:



- inhaltlicher Zusammenhang nicht benachbarter Terme:  
*Anschwellen der* auf dem Foto sichtbaren *Bauchspeicheldrüse*
- benachbarte Terme, die nicht inhaltlich zusammenhängen:  
in der exokrinen Drüsenfunktion werden vom *endokrinen Drüsenanteil* Hormone direkt ins Blut abgegeben

Jede Phrase beschreibt für sich ein Thema und wird mit ihrer normalisierten Bezeichnung (zur Normalisierung siehe weiter unten) als ein Netzknoten eingetragen.

Aus dieser initialen Sammlung der spezifischsten Themen in einem Sachgebiet wird durch weitere linguistische Verfahren eine vernetzte Themenhierarchie aufgebaut. Dafür werden die Phrasen in einer syntaktischen Analyse rekursiv in immer kürzere Phrasen bis hin zu den Einzelwörtern zerlegt. Das Ergebnis ist eine Struktur aus Ober- und Unterbegriffen sowie aus Differenzbegriffen (Abb.3).



Abb. 3: Ober-Unter- und Differenz-Begriffs-Strukturen

*Muskel* ist der Oberbegriff zu *Wadenmuskel*, dieser wiederum zu dreiköpfigem Wadenmuskel. *Wade* ist dagegen der unterscheidende Begriff (Differenzbegriff), der aus einem *Muskel* einen *Wadenmuskel* im Unterschied zu einem *Bauchmuskel* macht. Ober-Unterbegriffsbeziehungen und Differenzbeziehungen beschreiben bereits die Semantik und den Bedeutungskontext eines Themenbegriffes und werden deshalb eins zu eins in das Themennetz übernommen.

Wichtig ist, dass dabei verschiedene syntaktische Varianten einer Phrase auf dieselbe Struktur abgebildet werden.

Eine entsprechende Zerlegung findet auch für Komposita statt. Hier werden statt der syntaktischen Regeln die Wortbildungsregeln eingesetzt.

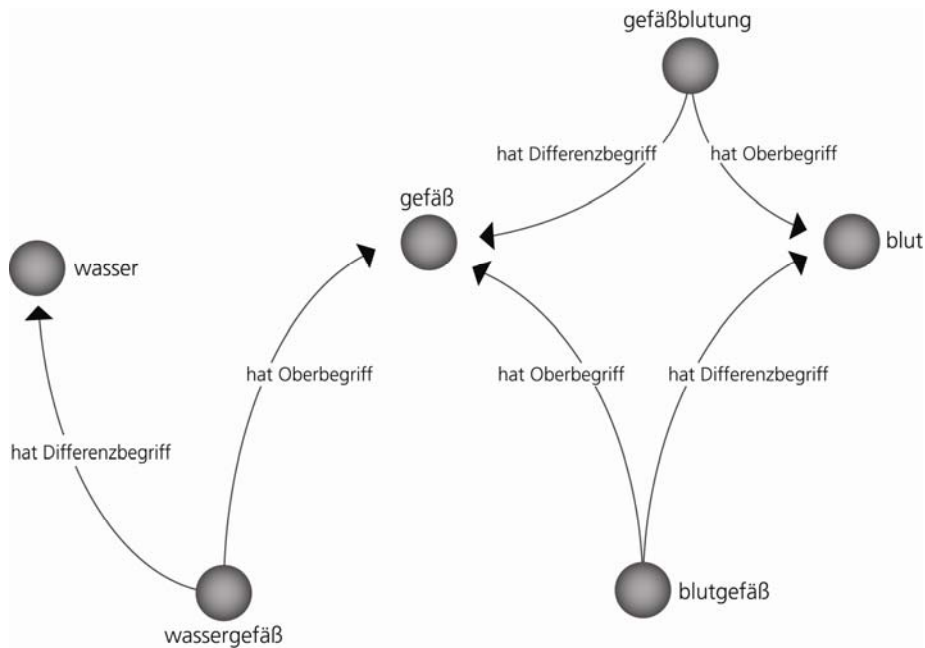


Abb. 4: Kompositazerlegung im Netz

Auf diese Weise wird automatisch ein Themennetz aus mehreren hunderttausend Themenkonzepten aufgebaut, die in ihrer Bedeutung durch Ober- und Unterbegriffsrelationen und durch Differenzrelationen, also eine Art Verwandtschaftsbeziehung, beschrieben sind. Das Themennetz weist damit eine ganz ähnliche Struktur auf wie ein Thesaurus, hat aber den großen Vorteil, dass es vollautomatisch aufgebaut wird.

### *Formulierungsvarianten für Themenkonzepte verzeichnen*

Im vorhergehenden Abschnitt ist deutlich geworden, wie durch die linguistischen Analyseverfahren verschiedene Formulierungsvarianten von Phrasen auf identische Themenkonzepte abgebildet und durch ihre Positionierung im Themennetz vereindeutigt werden.

Umgekehrt ist es aber auch wichtig, dass alle Formulierungsvarianten von Themenkonzepten weiterhin für die Indexierung von Individuen aus dem Individuennetz und für die Suche zur Verfügung stehen. Sie bilden das Zugangsvokabular zu den sprachlich normalisierten Themenkonzepten.

Aus diesem Grund werden einerseits alle empirisch aus den Texten erhobenen Phrasen als Alternativbezeichnungen für Konzepte mit in das Themennetz aufgenommen. Themenkonzept und Bezeichnungsalternative werden über die Relation *hatBezeichnungsvariante* verbunden.

Andererseits generiert man zusätzliches Zugangsvokabular, indem man aus den normalisierten Konzeptbezeichnungen alle Varianten ableitet, die morphologisch möglich sind.

Die morphologische Normalisierung oder Lemmatisierung während des Netzaufbaus hat bewirkt, dass alle konkret in den Texten auftretenden Wortformen auf eine Grundform, ein Lemma, zurückgeführt wurden. Die Beugungsformen *Blutgefäßes*, *Blutgefäße*, *Blutgefäßen* wurden so beispielsweise auf die Grundform *Blutgefäß* reduziert.

	Singular	Plural	
Nominativ	<i>Blutgefäß</i>	<i>Blutgefäße</i>	→ <b>Blutgefäß</b>
Genitiv	<i>Blutgefäßes</i>	<i>Blutgefäße</i>	
Dativ	<i>Blutgefäße</i>	<i>Blutgefäßen</i>	
Akkusativ	<i>Blutgefäß</i>	<i>Blutgefäßen</i>	

Alle Themenkonzepte sind mit ihrer normalisierten Benennung im Netz verzeichnet. Die Repräsentation des Themas *Schwellung der Bauchspeicheldrüse nach Trauma* ist entsprechend *bauchspeicheldrüse schwellen trauma*.

Mit Hilfe einer morphologischen Derivationsgrammatik werden nun zu jeder Grundform alle grammatikalisch möglichen Ableitungen gebildet, also zu *schwellen* die Ableitungen *Schwellung, geschwollen, schwillt, schwellend* usw., zu *trauma* der Plural *Traumata*. Diese Ableitungen werden in einem Normalisierungsindex als Zugangsvokabular für die Vorzugsbegriffe *schwellen* und *trauma* hinterlegt.

Gibt ein Nutzer nun die Anfrage *Geschwollene Bauchspeicheldrüse nach Traumata* ein, die in dieser Form noch nicht als Bezeichnungsvariante im Themennetz hinterlegt, weil sie nicht in der Textbasis auftauchte, so erfolgt über diese Liste automatisch die morphologische Normalisierung der Anfrage auf die Form *schwellen bauchspeicheldrüse trauma*. In dieser „Übersetzung“ deckt sich die Anfrage nun vollständig mit dem Konzept *bauchspeicheldrüse schwellen trauma*. Alle damit indexierten Individuen im Netz (Dokumente, Ansprechpartner, Therapien...) werden als Antwort ausgegeben. Da durch dieses Verfahren alle Formvarianten von Wörtern und Phrasen als äquivalent erkannt werden, steigt der Recall für die Suche deutlich.

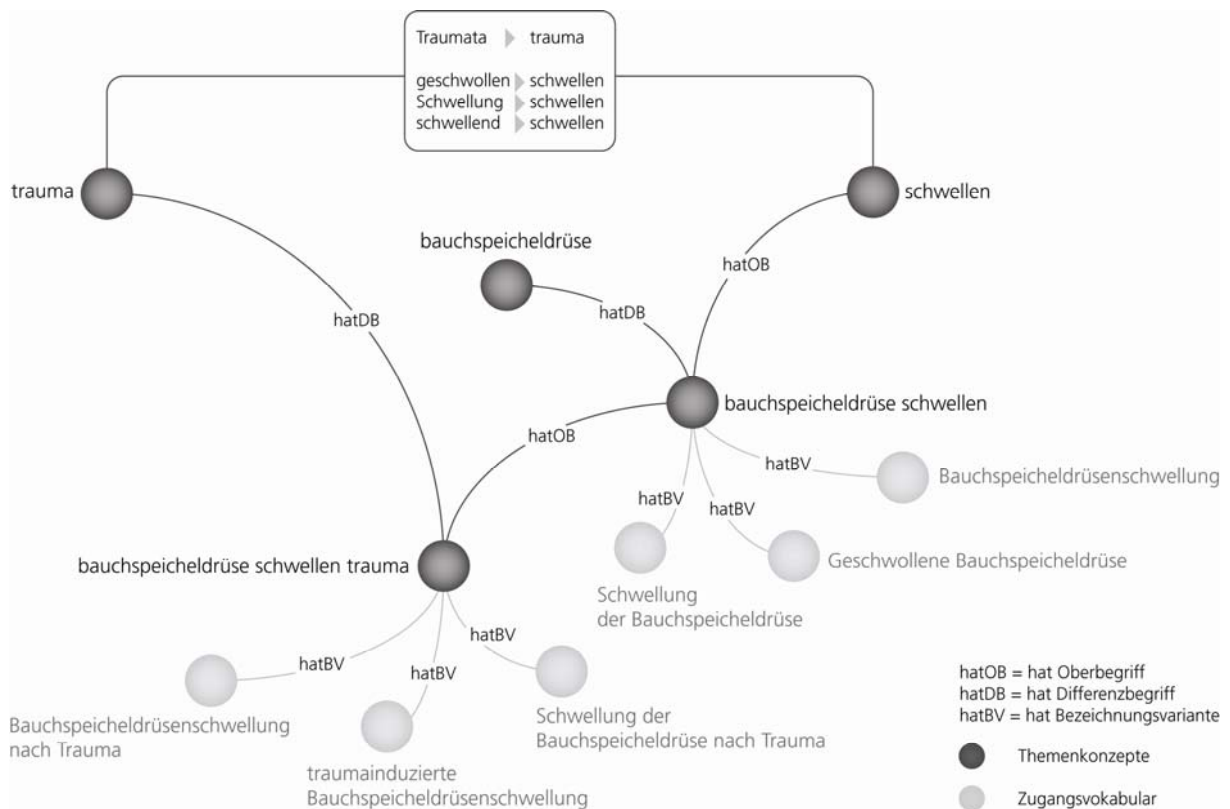


Abb. 5: Themenkonzepte und Zugangsvokabular

Synonyme, die im Unterschied zu den oben beschriebenen Bezeichnungsvarianten keinerlei sprachliche Verwandtschaft zu den Themenkonzepten haben, sowie fremdsprachliche Übersetzungen kommen über die automatische Auswertung von Wörterbüchern und Lexika oder maschinelle Übersetzungsverfahren ins Themennetz. So stammt die Synonymiebeziehung *Bauspeichedrüse hatFachsynonym Pankreas* aus einem medizinischen Lexikon, die Synonymiebeziehung *Bauchspeicheldrüse hatEnglSynonym pancreas* aus einem deutsch-englischen Fachwörterbuch.

### Statistische Analyse auf der Textebene

Der Netzaufbau mit Methoden der linguistischen Testanalyse wird durch statistische Text-Mining-Verfahren ergänzt. Diese Methoden schließen aus korrelierenden Auftretenshäufigkeiten von Termen in Texten eines Sachgebietes auf einen thematischen Zusammenhang. Treten Terme signifikant häufig zusammen auf, so sind sie wahrscheinlich inhaltlich verwandt.

Diese Analysemethode ermittelt keine so sicheren und differenzierten thematischen Zusammenhänge wie die linguistischen Verfahren. Doch sie erweitert Themenkonzepte um ganze Wortfelder. Diese Wortfelder können als Erweiterungsvokabular für die Suche genutzt werden, wenn der Recall unbefriedigend ist. Die Suche wird damit von bedeutungsgleichen Suchbegriffen auf thematisch assoziierte Begriffe ausgedehnt. Diese Begriffe können beispielsweise als Empfehlungen für weitere Suchthemen präsentiert werden.

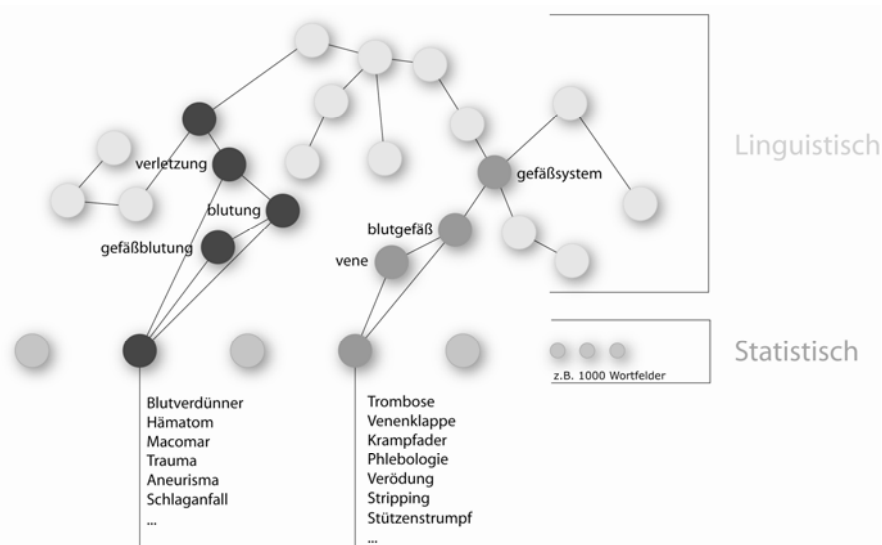


Abb. 6: Statistische Wortfeldindexierung

### Ergebnis

Das Ergebnis der linguistischen und statistischen Informationsextraktion sind drei verschiedene Indexe:

- Das umfangreiche Themennetz mit den sprachlich normalisierten Themenkonzepten und ihren semantischen Verknüpfungen. Diese Themeneinträge beschränken sich nicht auf einfache Konzepte wie

*Bauchspeicheldrüse*. Vielmehr werden auch komplexe Themenkonzepte wie *Bauchspeicheldrüse schwellen Trauma* im Netz verzeichnet.

- Einen Normalisierungsindex, der die Übersetzung zwischen auftretenden Wortformen und normalisierten Themenbezeichnungen leistet.
- Eine Wortfeldliste, die jedem Themenbegriff eine Vielzahl ähnlicher, thematisch verwandter Begriffe zuordnet.

Es ist konzeptionell und praktisch keine Problem, die Normalisierungs- und Wortfeldliste auch im Themennetz zu repräsentieren. Die Suche läuft aber performanter ab, wenn sie als eigene kleine Indexe ausgegliedert sind.

## ***Semantische Suche auf dem Wissensnetz***

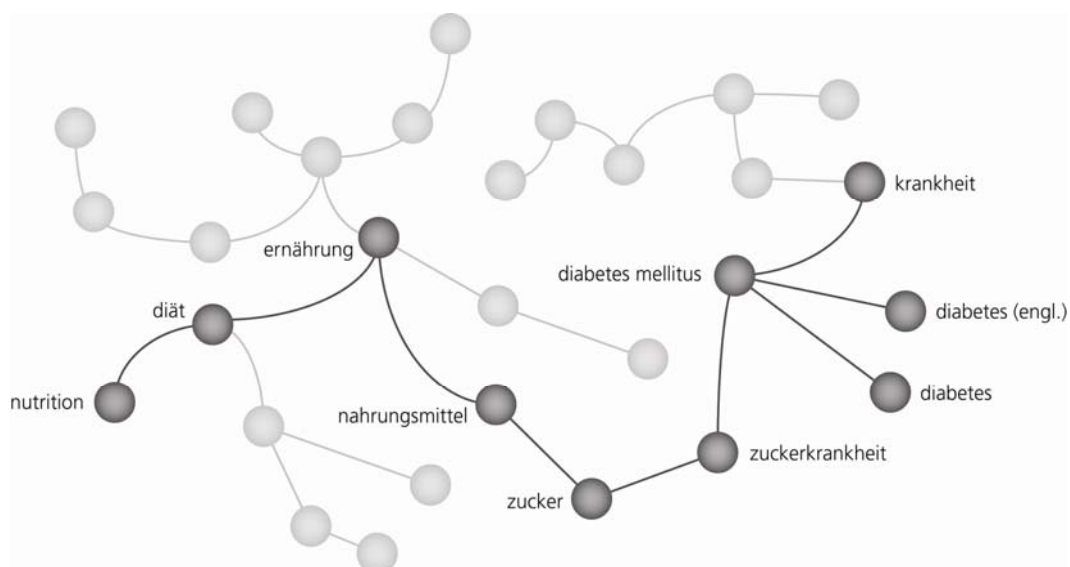
Was kann die semantische Suche mit automatisierten Wissensnetzen in unserem Beispiel leisten?

Die semantische Suche ermöglicht die Suche mit Synonymen und Übersetzungen, die Suche mit thematischen Verknüpfungen und die schlussfolgernde Suche nach impliziten Informationen.

### *Suche mit Synonymen und Übersetzungen*

Die typische Anfrage eines medizinischen Laien *Ernährung bei Zucker* liefert in unserem Gesundheitsportal eine Liste von einschlägigen Lexikonartikeln und Fachbeiträgen zurück, die sich mit den Themen *Ernährung im Rahmen von Diabetes Mellitus*, *Diabetikerdiät*, *Nahrungsmittel für Diabetiker*, *Schlemmertipps für Zuckerkrankte* oder *Diät bei Diabetes* befassen.

Relevante Dokumente werden also gefunden, obwohl sie nicht die umgangssprachlichen Formulierung *Zucker im Text aufweisen*, sondern den fachsprachlichen Term *Diabetes mellitus* bzw. *Diät* statt *Ernährung*. Die Bedeutungsgleichheit dieser Terme ist - wie im vorhergehenden Abschnitt beschrieben - im Wissensnetz hinterlegt. In der Suche werden diese Einträge wie die Siehe-Auch-Hinweise in einem Lexikon benutzt.



## Suche mit thematischen Verknüpfungen

Die oben genannte Anfrage wird aber nicht nur mit einer Dokumentliste wie in Google beantwortet, sondern der Nutzer erhält zusätzlich in eigenen Rubriken eine Zusammenstellung von Kontaktdaten und Profilen zu Ärzten, Kliniken und Verbänden, die auf Diabetologie spezialisiert sind, sowie gegebenenfalls eine Kurzbeschreibung der wichtigsten Medikamente. Diese Fakteninformationen stammen aus den Datenbanken, die über das Wissensnetz thematisch mit den Dokumentbeständen verknüpft sind. So deckt der Nutzer mit einer einzigen Anfrage alle wichtigen Facetten der Domäne Gesundheitsinformation ab. Textinformationen und Faktenauskunft kommen zusammen.

Die thematischen Verknüpfungen mit Hilfe des Wissensnetzes dienen nicht nur dazu, verschiedene Daten und Dokumente logisch zu integrieren. Sie sind auch wichtig, wenn eine Suchformulierung zu wenige Ergebnisse liefert. Dann werden thematisch benachbarte – aber nicht gleichbedeutende - Suchbegriffe unter der Rubrik Weiterführende Themen vorgeschlagen, die aus der Formulierungssackgasse herausführen (Abb.8): Ein thematisch assoziiertes Themenkonzept zu *Diabetes mellitus* ist beispielsweise *Metabolisches Syndrom*, ein „ähnliches“ Themenkonzept zu *Ernährung* ist *Übergewicht*. Mit diesen neuen Suchbegriffen kann der Nutzer in eine thematische Exploration einsteigen und sein Frageinteresse nach und nach präzisieren.

The image shows two side-by-side search results for the query 'Ernährung Zucker'. On the left is a Google search result, and on the right is a search result from the 'Gesundheitsportal'.

**Google Search Results:**

- Search bar: Ernährung Zucker
- Buttons: Suchen, Erweiterte Suche, Einstellungen
- Section: Web
- Result 1: **Zucker Ernährung** (www.cma.de) - 100% pflanzlich Wissenwertes über **Zucker** gibt's kostenlos bei der CMA.
- Result 2: **Zucker und seine Namen** (www.medizinfo.de/ernaehrung/zucker.htm) - Falsche **Ernährung** kann weitreichende Folgen haben. ... **Zucker** hat viele Namen, von denen nur einige leicht zu erkennen sind. ...
- Result 3: **Kohlenhydrate** (www.medizinfo.de/ernaehrung/kohlenhydrate.htm) - Falsche **Ernährung** kann weitreichende Folgen haben. Nahezu alle Zivilisationskrankheiten ...
- Result 4: **Südzucker AG - Zucker und Ernährung** (www.suezucker.de) - **Zucker** ist ein sicheres ...
- Result 5: **Ernährung Zucker** (www.heilpraktiker-links.de) - Ernährung **Zucker**. Krankheit der Haut "Ernährung **Zucker**"
- Result 6: **Praktische Tipps zur gesunden Ernährung** (www.inform24.de) - Tipps zur gesunden **Ernährung**. Basierend auf dem Ernährungskreis ... Der Verbrauch von **Zucker** ist (soweit wie möglich) zu reduzieren, denn **Zucker** ist ein ...
- Result 7: **ernaehrung.zucker - Vitanet.de - Prävention & Gesundheit** (www.vitanet.de) - Bei Vitanet finden Sie Informationen zum Thema **ernaehrung.zucker** sowie

**Gesundheitsportal Search Results:**

- Search bar: Ernährung bei Zucker
- Buttons: Suchen
- Section: Ihr Thema
- Result 1: Ernährung im Rahmen von Diabetes mellitus. Medizinische Rundschau 5/2005, Diabetikerdiät im Alltag. Apotheken Umschau 3/2006, Schlemmertipps für Diabetiker. Meine Familie und ich 7/2004
- Section: Der passende Arzt
- Result 1: Dr. Horst Kicher, diabetologisch tätiger Allgemeinarzt, 64293 Darmstadt...
- Result 2: Dr. Hannelore Franz, Diabetologe GDG, 64342 Seeheim...
- Result 3: Heinz Lammer, Schwerpunktpraxis Diabetes, 64431 Weinheim...
- Section: Die wichtige Klinik
- Result 1: Diabetes-Klinik Bad Nauheim GmbH
- Result 2: Diabetes-Klinik Bad Mergentheim
- Result 3: Knappschafts-Krankenhaus Sulzbach
- Section: Nichtärztliche Hilfe
- Result 1: Ulrike Kämmer, Ernährungsberatung Diabetes, Frankfurt am Main
- Result 2: Susanne Sollig, Diätassistentin Schwerpunkt Diabetesberatung, Bensheim
- Result 3: Diätsschule Kleinkeith, Gernsheim
- Section: Wichtige Medikamente
- Result 1: Antidiabetika
- Result 2: Insulin
- Section: Weiterführende Themen
- List: Metabolisches Syndrom, Übergewicht, Diabetes Typ I, Diabetes Typ II, Altersdiabetes

Abb. 8: Suche im Gesundheitsportal versus Suche mit Google

## Suche mit Schlussfolgerungen

Über seine Funktion als terminologischer und globaler Index hinaus ermöglicht das Wissensnetz auch die Suche mit Schlussfolgerungen.

Ein wichtiges Suchinteresse im Gesundheitsportal ist das nach alternativen preisgünstigeren Medikamenten. Deshalb ergibt die Suche mit *Aspirin* nicht nur Dokumente, die sich ganz konkret mit *Aspirin* beschäftigen. Diese Dokumentmenge wird einerseits ergänzt durch Beiträge, die sich mit dem Wirkstoff des *Aspirin*, der Acetylsalicylsäure, befassen. Und

andererseits werden in der Rubrik *Wichtige Medikamente* wirkstoffgleiche und wirkungsähnliche Medikamente zu *Aspirin* aufgeführt.

Das Ergebnis kommt durch folgende Ableitung aus dem Wissensnetz zustande:

- (1) Aspirin *enthält Wirkstoff* Acetylsalicylsäure.
- (2) ASS Ratiopharm *enthält Wirkstoff* Acetylsalicylsäure.

→ Also ist ASS Ratiopharm ein Alternativmedikament zu Aspirin.

- (1) Aspirin *enthält Wirkstoff* Acetylsalicylsäure.
- (2) Acetylsalicylsäure *gehört Zur Arzneimittelklasse* Schmerzmittel.
- (3) Paracetamol *gehört Zur Arzneimittelklasse* Schmerzmittel.
- (4) Benuron *enthält Wirkstoff* Paracetamol.

→ Also ist Benuron ein Alternativmedikament zu Aspirin.

### *Suchkonfiguration*

Welche Suchkonfiguration steckt hinter diesen semantischen Suchleistungen?

Der Großteil der in der Praxis auftretenden Suchanfragen bezieht sich nicht auf Konzepte, wie z. B. KRANKHEIT, MEDIKAMENT oder ARZT als solche, sondern auf deren individuelle Ausprägungen. Nicht das Konzept KRANKHEIT soll als Treffer im Suchergebnis auftauchen, sondern *Diabetes mellitus* als eine individuelle Ausprägung von KRANKHEIT. Auch das Vokabular, das im Themennetz hinterlegt ist, ist als Suchergebnis nicht von Interesse. Es dient lediglich als Zugangswerkzeug für die Ermittlung von relevanten Individuen in der semantischen Suche.

Der Ablauf der semantischen Suche lässt sich im Wesentlichen in vier konfigurierbare Hauptteile untergliedern:

1. Sprachliche Normalisierung
2. Einsprung in das Wissensnetz
3. Navigation auf vordefinierten Relationenpfaden
4. Filterung und Kategorisierung

### Sprachliche Normalisierung und Einsprung in das Wissensnetz

Wie jede Suchfunktion beinhaltet auch die semantische Suche die Methode des Stringabgleichs: Sind Suchstring und Indexierungsstring identisch, so wird das indexierte Objekt als Treffer angezeigt. Der reine Stringabgleich versagt bei Schreibfehlern, Schreibungsvarianten (ä – ae; ss - ß) und Beugungsformen. Deshalb werden sowohl alle Sucheingaben während des Suchprozesses als auch alle Bezeichnungen für Konzepte, Individuen und Themen schon beim Netzaufbau einer sprachlichen Normalisierung unterzogen.

Das heißt konkret: Auflösung von Umlauten, einheitliche Anpassung von Groß- und Kleinschreibung, Entnahme von Bindestrichen, Rückführung von gebeugten Wortformen auf die Grundform, Zerlegung von Komposita in ihre Simplizia. Stoppwörter werden im Vorfeld entnommen. Auch findet eine Rechtschreibkorrektur Anwendung.

Mit diesem Verfahren gelingt es, schon in der ersten Suchstufe möglichst viele Übereinstimmungen zwischen Suchtermen und Bezeichnungen von Individuen und Themen im Netz zu finden, also möglichst viele Netzeinsprünge mit gleicher inhaltlicher Bedeutung aber unterschiedlicher Schreibweise zu identifizieren.

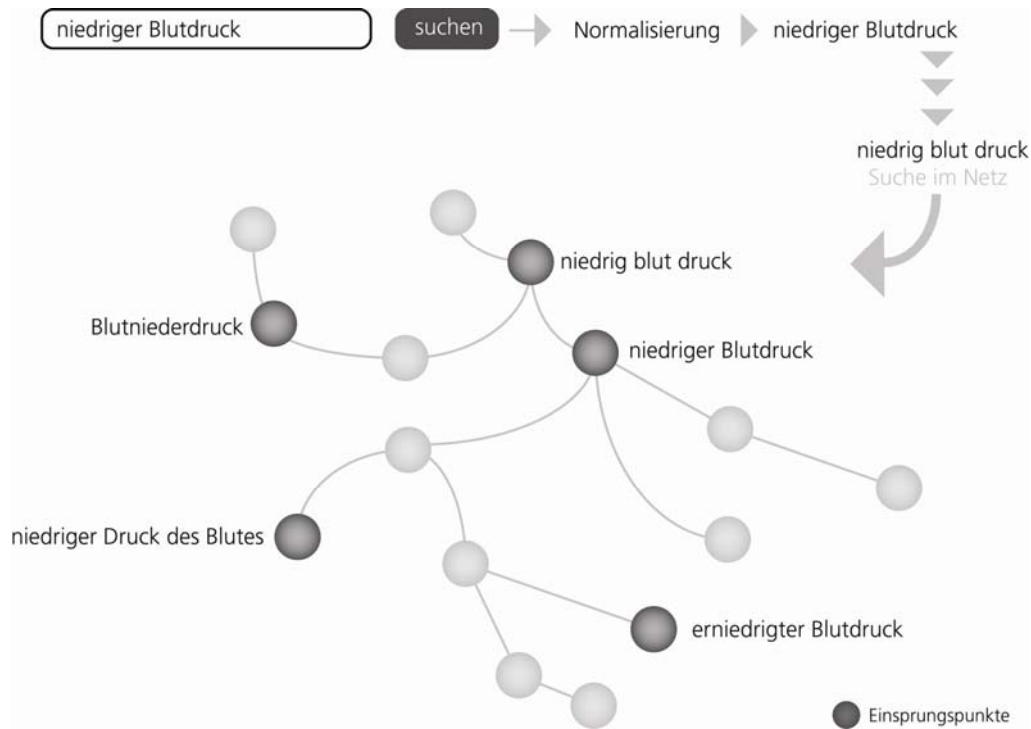


Abb. 9: Normalisierung und Einsprung in das Wissensnetz

### Navigation auf vordefinierten Relationenpfaden

An die Identifizierung von Netzeinsprünge schließt sich die Navigation auf vordefinierten Relationenpfaden an. Diese Pfade werden durch Hintereinanderschalten von Relationen gebildet. Dabei werden Relationen, die sprachliche Zusammenhänge abbilden, und Relationen, die inhaltliche Beziehungen darstellen, miteinander kombiniert.

Grundsätzlich stehen zu Beginn der Pfade sprachliche Relationen wie z. B. die Relation *hatSynonym*, *hatBezeichnungsvariante* oder *istOberbegriffVon*. Ausgehend von den gefundenen sprachlichen Relationen werden danach inhaltliche Relationen und Relationen zwischen individuellen Ausprägungen durchlaufen wie die Relation *istUrsacheVon*.

Beispiel:

Ist das Netzeinsprungsobjekt *niedriger Blutdruck*, so kann über den Relationenpfad

*hatSynonym >> istOberbegriffVon >> hatUrsache*

das Ergebnis *leptosomer Körperbau* ermittelt werden.

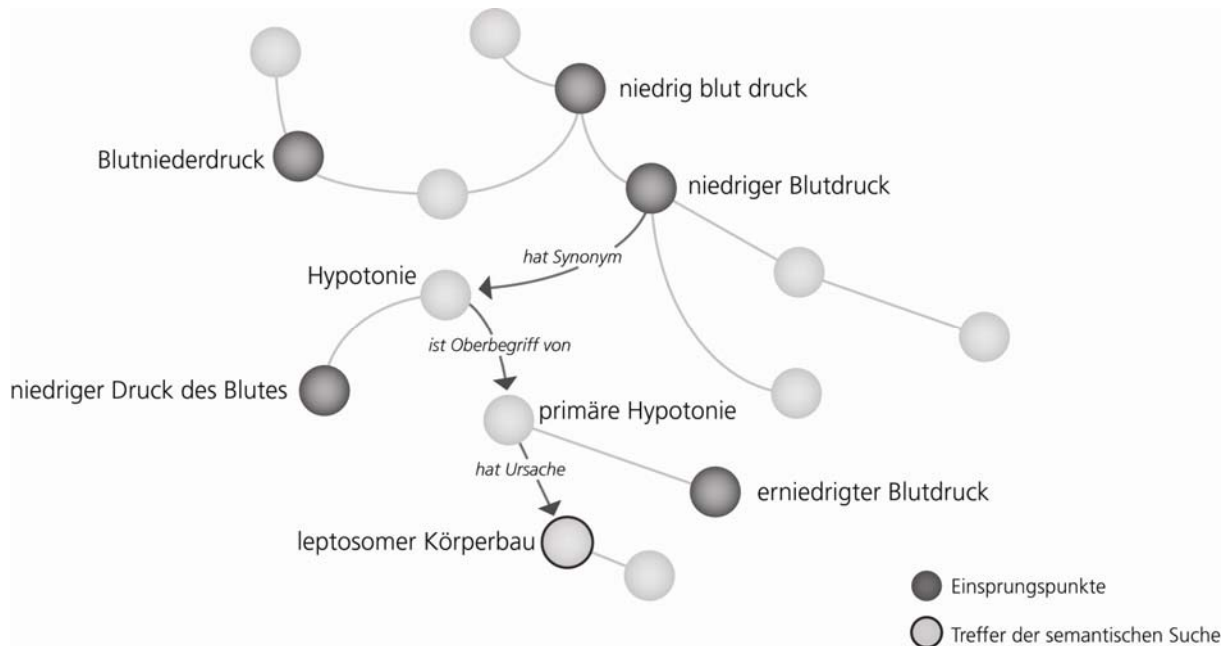


Abb. 10: Suche auf Relationenpfaden

## Filterung und Kategorisierung

Die sprachliche Normalisierung und das Ablaufen von sprachlichen und inhaltlichen Relationenpfaden führen unter Umständen zu einer sehr hohen Trefferzahl im Netz.

Ein Nutzer, der sich für *Hypotonie* interessiert, möchte in unserem Beispiel individuelle Ausprägungen von URSACHE, DOKUMENT, ERNÄHRUNG oder MEDIKAMENT finden, nicht aber das Vokabular rund um den Fachbegriff *Hypotonie*. Da aber genau dieser Wortschatz auch im Netz modelliert ist, ermittelt die Suche als Treffer auch viele gleichbedeutende oder verwandte Terme zu *Hypotonie*. Eine Filteroption auf Basis von Konzepten erlaubt das Ein- oder Ausblenden solcher unerwünschten Treffer. Lassen sich Treffer nicht auf die gewünschten Konzepte wie MEDIKAMENT oder ARZT zurückführen, werden sie aus dem Endergebnis der Suche herausgefiltert.

Am Ende des Suchprozesses steht eine Kategorisierung der gefilterten Treffermenge. Zur bessern Übersichtlichkeit für den Nutzer wird diese Treffermenge in Ergebnisrubriken wie *Der passende Arzt*, *Die richtige Klinik*, *Das wichtige Medikament* usw. einsortiert. Diese Kategorisierung beruht ebenfalls auf den Konzepten, denen die Treffer als individuelle Ausprägungen angehören.

## Weitere Einsatzszenarien für die Semantische Suche

Semantische Suchlösungen zeigen einen besonders hohen Nutzen, wenn unterschiedliche Datenwelten integriert werden sollen: hoch strukturierte Daten, wie sie typischerweise in Datenbanken vorliegen, und unstrukturierte Daten, wie sie Textdokumente darstellen. Hochstrukturierte Daten und Textdokumente treffen besonders häufig in Lotus-Notes-Umgebungen zusammen. Deshalb bietet ConWeaver eine spezielle Lösung für Lotus Notes an.

Semantische Suchfunktionen sind der Hauptbestandteil von Systemen für Enterprise Search. Als Suchmodul werden sie aber auch in Anwendungen für Enterprise Application Integration, Enterprise Content Management, Content Syndication und e-Commerce eingesetzt.

### ***Weiterführende Literatur***

Aramatzis, A.T.; T. Tsoirs; Koster, C.H.A.; van der Weide, Th.P.: Phrase-based Information Retrieval. In: Information Processing & Management, 34 (6), December 1998, pp. 693-707

Evans, D.; Zhai, C.: Noun-phrase Analysis in Unrestricted Text for Information Retrieval. In: Proceedings of the 34<sup>th</sup> Annual meeting of Association for Computational Linguistics. Santa Cruz, University of California, June 24-28, 1996, pp. 17-24

---

Ansprechpartner:

Dr. Andrea Dirsch-Weigand  
Fraunhofer Institut Integrierte Publikations- und Informationssysteme  
Dolivostraße 15  
64293 Darmstadt  
Fon 06151 – 869 4827  
Fax 06151 – 869 6968  
[dirsch@ipsi.fhg.de](mailto:dirsch@ipsi.fhg.de)