
Semantic Web Technologien zur Unterstützung des Zugriffs auf medizinischen Informationen

Heiner Stuckenschmidt, Universität Mannheim
Christian Herzog SyynX Solutions GmbH

Abstract

Viele Bereiche der Medizin, insbesondere die Medizinische Forschung und Medikamentenentwicklung benötigen effektiven Zugriff auf eine Vielzahl von Informationen in Form von Artikeln, Patenten und chemischen Zusammenhängen. Traditionelle Schlagwort-basierte Suchmaschinen stossen hier schnell an Ihre Grenzen. Eine Alternative bieten Ontologie-basierte Verfahren zur automatischen Indizierung sowie zur Suche und Navigation in grossen Informationsbeständen. Im Vortrag werden Erfahrungen mit der Entwicklung eines Ontologie-basierten Recherchesystems fuer medizinische Informationen geschildert und weiterführende Möglichkeiten der Technologie diskutiert.

1 Das Problem der Informationssuche

Viele Bereiche der Medizin sind auf die Verfügbarkeit von umfassenden Informationen über Krankheiten, Medikamente und Behandlungsmethoden angewiesen. Dies gilt insbesondere fuer die medizinische Forschung und die Medikamentenentwicklung. Führende Informationsanbieter wie Elsevier stellen inzwischen einen Großteil der benötigten Informationen in elektronischer Form zur Verfügung. Diese bechränken sich nicht nur auf Artikel in medizinischen Fachzeitschriften, sondern enthalten auch Informationen ueber Patente, chemische Stoffe und deren Nebenwirkungen, sowie Reaktionsketten chemischer Stoffe. Aufgrund der Vielzahl potentieller Informationsquellen setzt sich hier zunehmend das Prinzip der vertikalen Informationsbeschaffung durch, bei dem der Kunde nicht eine bestimmte Datenbank, sondern Informationen zu einem bestimmten Thema aus verschiedenen Datenbanken kauft. Voraussetzung fuer dieses Konzept sind effiziente themenbasierte Zugriffsmechanismen auf die entsprechenden Informationsbestände. Aufgrund von Komplexität und Heterogenität der Informationen aus unterschiedlichen Quellen stossen herkömmliche Schlagwort-basierte Suchmaschinen hier schnell an Ihre Grenzen.

Medizinisch wissenschaftliches Fachwissen ist sowohl im klinischen Behandlungsprozess als auch im wissenschaftlichen Arbeitsprozess unentbehrlich. Allerdings haben Arzt und Wissenschaftler unterschiedliche Anforderungen an die Art des Fachwissens, an die Präsentation des Wissens, die Art und Weise, wie das Wissen in den Arbeitsprozess integriert werden muss und an Aktualität und Validität des

Fachwissens. Während der Wissenschaftler große Publikationsmengen zu seinem Themengebiet überblicken, neueste Erkenntnisse und Trends frühzeitig entdecken, nach Möglichkeit mit intelligenten Methoden Wissen "zwischen den Zeilen" finden muss und vor allem assoziativ arbeitet, ist der klinisch tätige Arzt darauf angewiesen, im Arbeitsprozess ohne Zeitverzug und Zeitaufwand auf wenige ausgewählte behandlungsrelevante Daten zugreifen zu können. Diese müssen außerdem auf die speziellen Bedürfnisse des Patienten und die Arbeitssituation im Behandlungsprozess zugeschnitten sein. Beide Zielgruppen greifen auf den gleichen Wissenspool zu - aber mit deutlich unterschiedlichen Anforderungen, Herangehensweisen und Erwartungen.

In diesem Papier diskutieren wir die Verwendung von Technologien aus dem Bereich des Semantic Web zur Unterstützung der Versorgung mit medizinischen Informationen. In Kapitel 2 werden Anforderungen an die Realisierung einer solchen Unterstützung formuliert und kurz bestehende Ansätze erwähnt. Der Fokus liegt hierbei auf der automatischen Erschließung von medizinischen Dokumenten. Kapitel 3 stellt ein Recherchesystem vor, welches auf Grundlage der diskutierten Technologien die Inhaltsbasierte Suche in großen Mengen medizinischer Dokumente ermöglicht. In Kapitel 4 werden weitere Möglichkeiten der beschriebenen Technologien über eine Schlagwort-basierte Suche hinaus gehen vorgestellt und kurz auf einige bereits existierende Anwendungen dieser weiterführenden Möglichkeiten hingewiesen.

2 Semantic Web Technologien

Ein grundsätzliches Problem der Suche in Fachliteratur besteht darin, dass herkömmliche statistische Verfahren zur Bestimmung der Relevanz von Dokumenten im Hinblick auf eine bestimmte Fragestellung häufig suboptimale Ergebnisse liefern, da weder unterschiedliche Begriffe mit der gleichen Bedeutung erkannt werden. Im medizinischen Kontext finden sich viele Beispiele für diese Probleme. So gibt es weit mehr als 100 unterschiedliche Bezeichnung für das bekannte Schmerzmittel Aspirin. Auch wird der Kontext, in dem ein Begriff genannt wird berücksichtigt wird in der Regel nicht berücksichtigt. Um einen effizienteren Zugriff auf fachinformationen zu gewährleisten ist es daher notwendig, Informationsinhalte möglichst präzise in einer Form zu beschreiben, die von der Suchanwendung verstanden wird. Die geschieht in der Regel durch die Verwendung von Metadaten. Neben klassischen metadaten wie Autoren und Titel sind hierbei vor allem sogenannte inhaltsbeschreibende Metadaten von Bedeutung, mit deren Hilfe ein Dokument in einen thematischen Kontext gesetzt wird. Der erfolgreiche Einsatz solcher inhaltsbeschreibenden Metadaten hängt dabei von folgenden Faktoren ab:

- Standardisierte Sprachen zur maschinenlesbaren Darstellung von Metadaten
- Standardisierte Vokabulare zur inhaltlichen Beschreibung von Dokumenten

- Technologien zur automatischen Inhaltsanalyse von Dokumenten

Der Bereich der Semantic Web Forschung beschäftigt sich mit der Bereitstellung von Lösungen für diese Aspekte. Im folgenden wird kurz auf bestehende Lösungen eingegangen.

Im Rahmen der Semantic Web Initiative des W3C wurden eine Reihe von Sprachstandards für die Darstellung von Metadaten im WWW entwickelt. Die Grundlage dieser Sprachen bildet das Resource Description Framework RDF [Man04]. RDF bietet ein Konzept zur eindeutigen Referenzierung von Objekten, sowie ein Graph-basiertes Datenmodell zur Beschreibung der Eigenschaften von Objekten. Basierend auf diesem Modell wurden verschiedene Sprachen zur Darstellung von Wissen in Form von Ontologien entwickelt. Die einfachste dieser Erweiterungen sind RDF Schemata [Bri04]. Der RDF Schema Standard definiert eine Reihe von Eigenschaften mit einer speziellen Bedeutung. Diese können verwendet werden, um einfache Begriffssysteme in Form getypter Relationen sowie Klassen- und Relationshierarchen zu beschreiben. Inferenzen über diese Definitionen lassen sich mit Hilfe eines einfachen Produktionsregelsystems implementieren.

Diese Eigenschaften macht RDF zu einer idealen Grundlage für die Beschreibung von Metadaten unterschiedlicher Informationsobjekte und erfüllt demnach die Anforderung einer standardisierten, maschinenlesbaren Sprache für Metadaten. Dementsprechend dient RDF bereits heute als bevorzugtes Format zur Repräsentation von Metadatenstandards wie Dublin Core. Das Graph-basierte Datenmodell erleichtert hierbei zudem die Integration unterschiedlicher Standards in neue, massgeschneiderte Metadatenformate sowie die einfache Einbindung standardisierter Vokabulare zur Beschreibung von Informationsinhalten. Solche Vokabulare sind im Bereich der Medizin bereits seit einiger Zeit etabliert, so dass zumindest in diesem Anwendungsbereich die Anforderung nach einem standardisierten Vokabular erfüllt ist. Das Unified medical Language System UMLS [Bod04] umfasst zur Zeit mehr als 50 Terminologien aus unterschiedlichen Bereichen der Medizin. Hinzu kommen kommerzielle Thesauri, die von Informationsanbietern entwickelt wurden, um ihre Inhalte effizienter zu erschliessen.

Die wahrscheinlich größte Herausforderung stellt die automatische Analyse von Inhalten medizinischer Publikationen dar. Im Bereich des klassischen Information Retrievals wurden hier Methoden entwickelt, um die Relevanz eines Begriffes zur Beschreibung eines Dokumentes entwickelt. In der Regel wird die relative Anzahl der Vorkommen eines Wortes im Verhältnis zur Anzahl der Dokumente, in denen der Begriff vorkommt verwendet. Im medizinischen Kontext stellt sich hier das Problem der Fokussierung dieser Methoden auf medizinische Fachbegriffe. Abbildung 1 veranschaulicht dieses Problem. Links unten im Bild sind die Häufigkeiten und die Relevanz der nach herkömmlichen Methoden relevantesten Begriffe des auf der linken Seite gezeigten Textes abgebildet. Hierbei sind die drei Begriffe zwei Namen und ein

falsch geschriebenes Wort. Der Grund dafür, dass diese Begriffe als besonders relevant gewertet werden ist die Tatsache, dass sie nur in sehr wenigen anderen Dokumenten auftauchen und daher als sehr charakteristisch für das gegebene Dokument angesehen werden. Diesem Problem von bestehenden Ansätzen lässt sich durch die Verwendung eines medizinischen Thesaurus begegnen indem man die Bewertung der Relevanz auf im Thesaurus enthaltene medizinische Begriffe beschränkt.



Abbildung 1: Automatische Indizierung medizinischer Texte

Abbildung 1 zeigt den Effekt der Beschränkung auf medizinische Begriffe, die in diesem Fall aus UMLS stammen. Man sieht gut, dass in diesem Fall die als relevant angesehenen Begriffe (vor allem Artemisia und Malaria) den Inhalt des Dokuments deutlich sinnvoller beschreiben. Die entsprechende inhaltliche Beschreibung des Dokumentes kann nun verwendet werden, um eine semantische Suche auf der Grundlage von Thesaurusbegriffen zu realisieren. Im folgenden werden beispielhaft einige Systeme vorgestellt, die solche Suchlösungen realisieren und hierfür SynX Technologie zur automatischen Indizierung von Dokumenten auf der Grundlage medizinischer Thesauri verwenden.

3 Das DOPE Projekt

Ziel des DOPE Projektes war der integrierte und benutzerfreundliche Zugriff auf medizinische Publikationen aus unterschiedlichen Quellen. Besonderes Augenmerk wurde hierbei auf die Möglichkeit einer themenbasierten Suche in den vorhandenen Dokumentenbeständen gelegt. Zu diesem Zweck wurde der von Elsevier entwickelte und vermarktete medizinische Thesaurus EMTREE verwendet. EMTREE enthält ca 250.000 Begriffe zu unterschiedlichen Aspekten medizinischer Forschung (Krankheiten, Symptome, Medikamente, Behandlungsmethoden, etc.) die eine Grundlage für die

gezielte Formulierung thematischer Anfragen bilden. In diesem Zusammenhang wurden in enger Zusammenarbeit mit den Entwicklern von EMTREE fundierte Erfahrungen mit dem Aufbau, der Wartung und der Verwendung großer Thesauri in Suchanwendungen gesammelt, die für eine erfolgreiche Bearbeitung des beantragten Projekt notwendig sind. Unter anderem wurde der EMTREE Thesaurus zum Zweck einer einfacheren Integration in Web-basierte Anwendungen in den W3C Sprachstandard RDF übersetzt. Hierfür wurde eine an der freien Universität entwickelte Methode zur systematischen Konvertierung von Begriffshierarchien verwendet [Ass04]. Um die Verwendung des Thesaurus zu vereinfachen wurde das entsprechende RDF Modell mit Hilfe der von der Firma Aduna entwickelten RDF Datenbank Sesame [Bro02] über ein http-basiertes Web Interface zur Verfügung gestellt. Über dieses Interface können flexibel Anfragen an den Thesaurus gestellt werden. Die automatische Ergänzung von Suchanfragen mit Unter- bzw. Oberbegriffen eines bestimmten Terms wird hierbei auf Grundlage der standardisierten Semantik der Sprache RDF automatisch durchgeführt.

Um die in EMTREE enthaltenen Begriffe als Grundlage für einen themenbasierten Zugriff auf Dokumente verwenden zu können, bestand die Notwendigkeit, die Dokumente mit den entsprechenden Begriffen zu indizieren. Um die Skalierbarkeit und Erweiterbarkeit des Recherchesystem zu gewährleisten wurde Methoden zur automatischen Indizierung von Dokumentbeständen mit Begriffen aus einem gegebenen Thesaurus untersucht. Grundlage der Untersuchung waren 500.000 Volltextartikel aus dem Bestand von Sciencedirect (Elseviers Online Portal für wissenschaftliche Publikationen) sowie mehr als 10 Millionen Abstracts aus der Medline Datenbank, einer der wichtigsten Sammlungen medizinischer Publikationen. Die Indizierung wurde mit Hilfe eines Systems der Firma Collexis durchgeführt. Das System verwendet die relative Häufigkeit von Begriffen aus dem Thesaurus in den zu indizierenden Dokumenten, um die Relevanz eines Begriff im Hinblick auf das Thema eines Dokuments zu bestimmen. In Experimenten auf vergleichbaren Datensätzen wurde gezeigt, daß sich mit diesem Verfahren beider Suche nach Dokumenten der gleiche Grad der Vollständigkeit erreichen läßt wie mit den besten Methoden zur Freitextsuche. Durch die Einschränkung auf Thesaurusbegriffe läßt sich die Korrektheit der Suchergebnisse jedoch deutlich verbessern.

Im Rahmen eines durch den Elsevier Verlag finanzierten Projektes wurde ein System zum Thesaurus-basierten Zugriff auf indizierte Dokumentenbestände mit dem Namen DOPE implementiert [Stu04a]. Kern des Systems bildete eine erweiterte Version der Sesame Datenbank in der unterschiedliche Thesauri und Metadaten verschiedener Dokumentenbestände verwaltet werden. Basierend auf der Benutzeranfrage werden relevante Thesaurusbegriffe ermittelt und Metadaten der entsprechenden Dokumente an die Benutzeroberfläche zurückgegeben. Dokumente und deren Metadaten werden hierbei nicht direkt im System verwaltet, sondern über ein generisches Dokumentenmodell von externen Quellen (Metadaten-Servern oder Literaturdatenbanken) bezogen. Abbildung 1 illustriert die Architektur des Systems, die eine einfache Erweiterung um neue Thesauri und Informationsquellen ermöglicht.

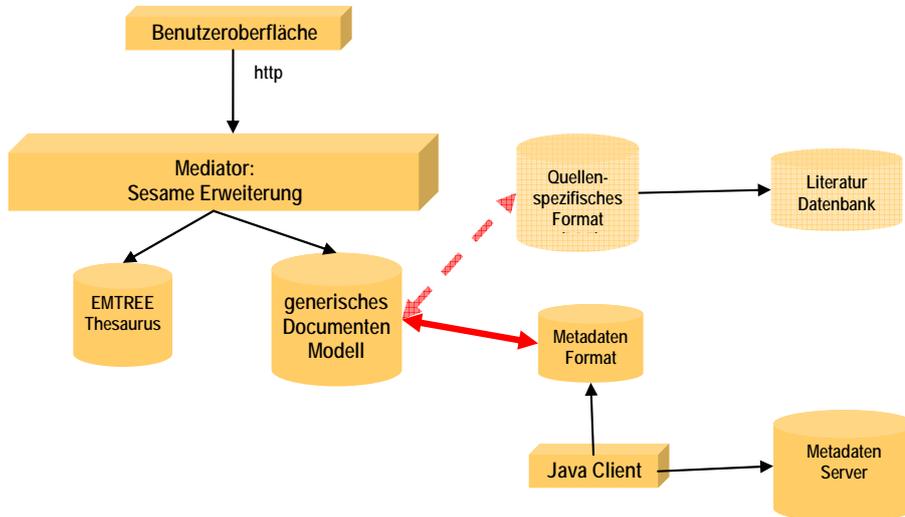


Abbildung 2: Architektur des DOPE Systems

Ein weiterer Aspekt, der im Rahmen der oben genannten Projekt eingehend untersucht wurde ist die gezielte Unterstützung des Benutzers durch die Implementierung geeigneter Suchstrategien und graphischer Darstellung der Suchergebnisse [Stu04b]. In diesem Zusammenhang wurde eine spezielle Benutzerschnittstelle für den thesaurusbasierten Zugriff auf grosse Dokumentenbestände entwickelt und durch die User-Centered Design Group des Elsevier Verlags in Benutzerstudien evaluiert. Die in der entwickelten Benutzerschnittstelle realisierte Suchstrategie beruht darauf, dass der Benutzer zunächst den Kontext der Suche durch die Angabe eines generellen Suchbegriffes (z.B. Malaria) einschränkt. Basierend auf den im Bezug auf dieses Thema relevanten Dokumenten schlägt das System dem Benutzer weitere Suchbegriffe zur Eingrenzung der Suche auf spezifischere Themen (z.B. Malaria in Verbindung mit einem speziellen Wirkstoff) vor, zu denen Dokumente vorhanden sind. Die entsprechenden Suchbegriffe werden hierbei dem Thesaurus entnommen und können durch Navigation in den entsprechenden Thesaurusstrukturen ausgewählt werden. Zur Darstellung der Suchergebnisse zu einer Kombination von Suchbegriffen wurde die von der Firma Aduna entwickelte Clustermap Technologie verwendet. Hierbei werden relevante Dokumente in einer interaktiven Karte mit zusätzlichen Interaktionsmöglichkeiten dargestellt. Abbildung 2 zeigt die Darstellung eines Suchergebnisses mit 4 Suchbegriffen im thematischen Kontext Aspirin.

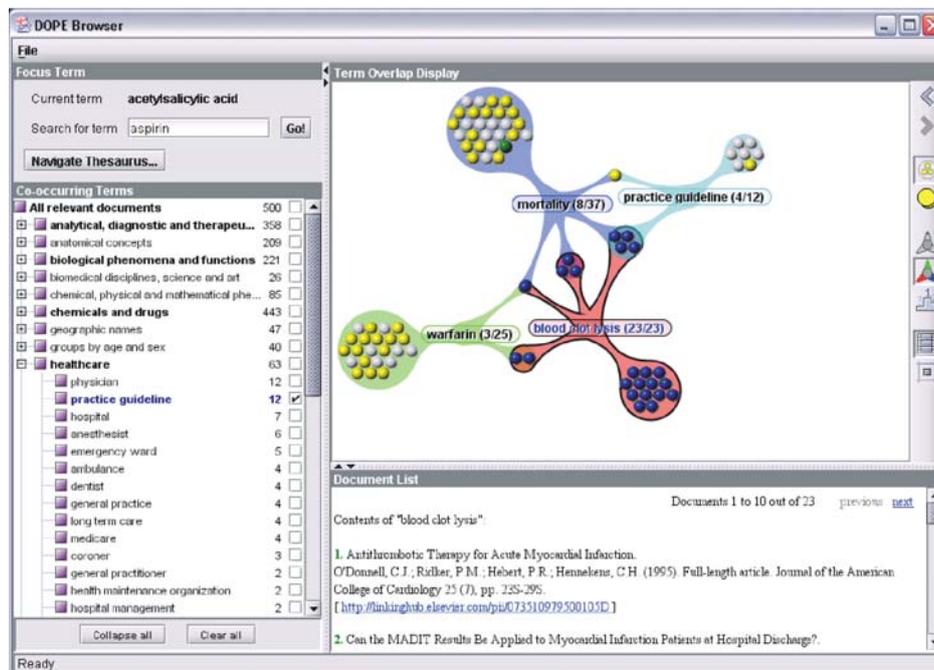


Abbildung 3: Benutzeroberfläche des DOPE Systems

In einer Studie wurde das beschriebene Suchinterface im Hinblick auf die Unterstützung des Benutzers bei der themenbasierten Suche nach Informationen untersucht. Hierzu wurde Besucher der Konferenz Drug Discovery Technology 2003 in Boston das System vorgeführt und den Probanden die Gelegenheit gegeben das System selbständig zu nutzen. Abschließend wurden die Probanden nach Ihrer Meinung zu Benutzerführung und Ergebnisvisualisierung befragt. Die Mehrzahl der befragten Personen äußerten sich positiv über das System. Als besonders nützlich wurde die automatische Generierung von Vorschlägen zur Einschränkung der Amfrage sowie die graphische Darstellung der Suchergebnisse und der Beziehungen zu den verschiedenen Suchbegriffen erwähnt. Das Prinzip der Thesaurus-basierten Suche wurde als sehr nützlich zur Erschließung unbekannter Dokumentenbeständen sowie zur Unterstützung der Lehre – insbesondere als Recherchesystem für Studenten und Doktoranden – eingeschätzt.

4 Weiterführende Ansätze

Mit dem SyynX mediator steht Wissenschaftlern aus dem Bereich Life Sciences und Ärzten ein neuartiger Recherchezugang auf Medline zur Verfügung, der neben der Suche mit einem oder mehreren Suchbegriffen auch weitergehende und komplexere Analysen ermöglicht. Möglich wird dies durch den Einsatz des thesaurusbasierten Indexierungssystem Collexis, mit dem die Jahrgänge 1996 - 2005 von Medline

indexiert wurden, um weitergehende Analysen auf der Basis der Medline-Daten durchführen zu können. Intuitive Darstellungen z.B. der Themenschwerpunkte der Abstracts in graphischer Form erlauben eine schnelle und zeitsparende Entscheidung, ob der Artikel wirklich von Interesse ist.



Abbildung 4: Aus Publikationen generiertes Expertenprofil

Im SynX mediator sind auf der Grundlage der auch im DOPE System verwendeten Technologien einige weitergehende Ideen realisiert, die einen flexibleren Zugang zu Information ermöglichen als die rein Keyword getriebene Suche im DOPE System. Im folgenden werden die wichtigsten dieser Funktionalitäten sowie ihre Anwendung kurz beschrieben.

- **Vorschlag von Fachtermini zur Sucheingrenzung** So werden dem Benutzer z.B. zu jeder Suche die Begriffe vorgeschlagen, die neben den gewählten Suchtermini noch in der Menge der Suchergebnisse vorkommen.
- **Analyse ganzer Texte** Ganze Texte können als Suchbegriff verwendet werden, um ähnliche Publikationen zu finden oder z.B. die Chancen eines Forschungsantrags abzuschätzen.
- **Expertenprofile der Autoren** Zu den gefundenen Experten wird ein Profil angezeigt, das Rückschlüsse auf die Aktivitäten der Person zulässt. Auf einen

Blick kann der Benutzer sich ein Bild über die Schwerpunkte des Experten machen.

- **Zeitreihenanalyse** Hat das allgemeine Interesse an den gewählten Suchbegriffen über die Jahre zugenommen? Mit dieser Darstellung können neue Forschungsrichtungen und Trends aufgespürt werden.

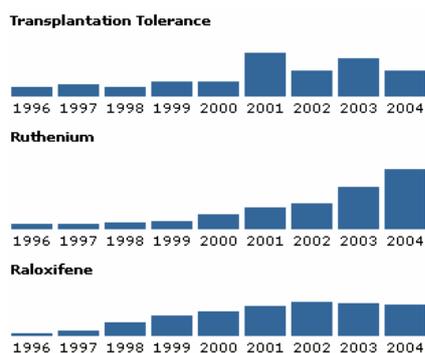


Abbildung 5: Analyse thematischer Trends

Auf der Basis dieser erweiterten Technologien wurden bereits einige innovative Anwendungen für die Unterstützung der klinischen Forschung.

SynnX Knowledge Cockpit - Schnelle Analyse großer Publikationsmengen Mit dem SynnX Knowledge Cockpit wurde für den Wissenschaftler im Bereich Medizin / Life Science eine Lösung geschaffen, die ihm erlaubt, große Publikationsmengen zu einem definierten Thema schnell und effizient inhaltlich zu analysieren und das für ihn und seinen wesentliche herauszufiltern. Datenbasis eines Knowledge Cockpits sind wissenschaftliche Publikationen, z.B. als PubMed-Abstracts vorliegend, die eine thematische Vertikale definieren und dabei bis zu bis zu 100.000 Publikationen umfassen können, z.B. zum Thema Alzheimer. Diese Abstracts werden mit Collexis-Engine in Fingerprints umgewandelt und unter einer ontologiebasierten Navigation zugänglich gemacht. Gleichzeitig werden aus den Abstracts und den Meta-Informationen Expertenprofile generiert, die einen klaren Überblick erlauben, wer mit wem und zu welcher Zeit weltweit zum Thema des Cockpits gearbeitet hat. In der aus den Fingerprints generierten Navigation können nun schnell relevante Publikationen identifiziert werden, z.B. die Publikationen die neben dem zentralen Thema eines Cockpits auch eine weitere Erkrankung und Medikament enthalten.

Im Bereich Knowledge Discovery wird für den Wissenschaftler aufbereitet, welche Begriffe erstmalig in einer Veröffentlichung mit dem Cockpit-Thema genannt werden und welche Fachbegriffe voraussichtlich mit dem Cockpit-Thema in Zukunft gemeinsame Relevanz entwickeln werden. Die Basis für die Analysen bildet ein Vektorraum aus 5 Millionen Konzeptvektoren abgeleitet aus den PubMed-Abstracts seit 1996. Dieser Dokumentenraum wird durch Analyse der Konzeptassoziationen

(gemeinsames Auftreten in mindestens einer Publikation) in einen Assoziationsraum gewandelt. Dieser Vektorraum, bestehend aus ca. 35.000 Dimensionen (den benutzten MeSH-Konzepten) und 60 Millionen (direkten) Assoziationen, kann nun dafür genutzt werden, um mittels Ähnlichkeitsanalyse neue Assoziationen zu finden, die in dieser Kombination so noch nicht in der Literatur gebildet wurden. Durch Interpolation der Ähnlichkeitsfaktoren über einen Zeitstrahl von 10 Jahren können feinste Bewegungen in diesem Raum analysiert werden und Wahrscheinlichkeiten für entstehende Assoziationen berechnet werden. Das Ergebnis sind gewichtete Vorschlagslisten von potentiellen (also indirekten) Assoziationen, die es einem Experten erlauben, z.B. Substanzen passend zu einem bestimmten Symptom zu entdecken oder eventuelle gewünschte oder unerwünschte Nebenwirkungen einer Substanz früher in seine Überlegungen einzubeziehen als es ohne diese Analyse möglich wäre.

SynnX Clinical Consult Application - passgenaues Wissen im klinischen Kontext

Mit der SynnX Clinical Consult Application werden dem Arzt für den jeweiligen Patienten relevante Fachpublikationen mit klinischer Relevanz und direkt im Arbeitsprozess angezeigt [Her04]. Möglich wird dies durch die direkte Anbindung an das Krankenhausinformationssystem (KIS) und die automatisierte Generierung eines "Patientenfingerprints" auf kodierten Informationen wie ICD-Codes, Altersgruppe und Geschlecht des Patienten. Dieser generierte Patientenfingerprint wird mittels der Collexis Core Engine mit den vorliegenden Fingerprints von wissenschaftlichen Publikationen verglichen. Die Veröffentlichungen mit einer definierten Ähnlichkeit werden dem Arzt im KIS oder im Intranet angezeigt. Dabei werden über Filter, die klinisch relevante Publikationen erkennen lassen, und über den Publikationstyp der Arzt mit behandlungsrelevanten Dokumenten versorgt (z.B. Cochrane Reviews als leitlinienähnliche Dokumente etc.) oder Case Reports auf PubMed. Der Behandler kann die Sicht durch die Ergänzung von angebotenen Schlüsselbegriffen verfeinern und anpassen. Mit der SynnX Clinical Consult Application, die in verschiedene KIS-Systeme integriert werden, werden dem Arzt im Arbeitsprozess ausgewählte wissenschaftliche Inhalte angeboten. Hauptaugenmerk liegt darauf, den Arzt im gedrängten klinischen Alltag nicht mit Informationen zu überfluten, sondern die Einbindung von Informationsangeboten in den Workflow den Arzt zur Auseinandersetzung mit Innovationen und Fachveröffentlichungen anzuregen

5 Diskussion

Zusammenfassend kann man sagen, dass die Verwendung von Prinzipien und Technologien des Semantic Web, insbesondere der Kombination von automatischer Indizierung und Thesaurus-basierter Suche zahlreiche Möglichkeiten bietet, den Zugriff auf medizinische Information zu verbessern. Die hier beschriebenen Techniken und Applikation zielen jedoch ausschliesslich auf medizinische Publikationen. In vielen Bereichen der medizinischen Forschung reichen solche Publikationen als

Informationsquelle jedoch nicht aus. Weitere wichtige Informationen liegen in Form strukturellen Informationen vor. Hierbei sind vor allem Patentdatenbanken und Datenbanken mit chemischen Informationen zu nennen. Die Herausforderung besteht vor allem darin, eine integrierte Suche, die sowohl Dokumente als auch strukturierte Informationen einbezieht. Ein Ansatzpunkt bietet hier die Verwendung von RDF als einheitliche Darstellungsform für Dokumentmetadaten und strukturelle Informationen sowie die Integration chemischer Informationen in bestehende Thesauri.

Referenzen

- [Ass04] Mark van Assem, Maarten R. Menken, Guus Schreiber, Jan Wielemaker and Bob Wielinga. A Method for Converting Thesauri to RDF/OWL. In: Proceedings of the Third International Semantic Web Conference (ISWC'04), Lecture Notes in Computer Science no. 3298, page 17-31, Hiroshima, Japan, November, 2004
- [Bae99] Ricardo A. Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley Longman Publishing, 1999.
- [Bod04] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 2004, Vol. 32, 2004 Oxford University Press.
- [Bri04] Dan Brickley, R.V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema W3C Recommendation 10 February 2004 (<http://www.w3.org/TR/2004/REC-rdf-schema-20040210/>)
- [Bro02] Jeen Broekstra, Arjohn Kampman, Frank van Harmelen. Sesame: An Architecture for Storing and Querying RDF and RDF Schema In Proceedings of the First International Semantic Web Conference (ISWC 2002), Sardinia, Italy, June 9-12 2002, pg. 54-68. Springer-Verlag Lecture Notes in Computer Science (LNCS) no. 2342.
- [Her05] Christian Herzog, Gianpiero Liuzzi, Mario Diwersy. SyynX solutions: practical knowledge management in a medical environment. Bremen: ACM Press, CIKM 2005: 556-559
- [Man04] Frank Manola, Eric Miller (Editors) RDF Primer W3C Recommendation 10 February 2004 (<http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>).
- [Stu04a] H. Stuckenschmidt et al. Exploring Large Document Repositories with RDF Technology - The DOPE Project IEEE Intelligent Systems, 19(3) 2004
- [Stu04b] H. Stuckenschmidt et al. A Topic-Based Browser for Large online Resources. Proceedings of the International Conference on Knowledge Acquisition, Modelling and Management, EKAW'04, Milton Keynes, England 2004.