German Medical Knowledge Guide – Prozessorientierte automatisierte Analyse großer Textkorpora

Diwersy M[1], Herzog C[1], Kamphusmann T[2], Koch O[2], Vollmer G[2]
[1]*Syynx Solutions GmbH*
[2]*Fraunhofer Institut für Software- und Systemtechnik*
thomas@kamphusmann.org

Einleitung und Fragestellung

Sowohl für Patienten als auch für Ärzte und Entscheidungsträger der Verbände und in der Politik ist die Identifikation von medizinischen Experten mit vertretbarem Aufwand und auf der Basis vergleichbarer und transparenter Kriterien nicht möglich. Für Patienten stellen dabei Mund-zu-Mund-Propaganda, intransparente und tlw. zweifelhafte »Ratings« in Publikumszeitschriften den üblichen Lösungsweg dar. Ärzte, Funktionäre und Politiker verlassen sich in dieser Situation im Wesentlichen auf gewachsene Kollegennetzwerke. So unterschiedlich die Motivation für die Suche nach ausgewiesenen Experten im Einzelfall ist, lassen alle drei Lösungsstrategien ein operatives Kriterium für »Expertise« und damit ein transparentes Verfahren der Suche vermissen. In dieser Situation stellt die Entwicklung handhabbarer Kriterien für Expertise und deren operative Umsetzung in Recherchesysteme einen notwendigen Schritt dar.

Auf der anderen Seite steht mit den wissenschaftlichen Veröffentlichungen in den Bereichen Medizin und Life-Sciences, wie sie z.B. in PubMed verfügbar sind, ein umfassender Datenpool vor. Dieser kann, mit nur unerheblichen Einschränkungen, als vollständige Dokumentation der abgeschlossenen und eingeschränkt auch der aktuell laufenden wissenschaftlichen Arbeiten angesehen werden. Dieser Datenpool kann somit als Grundlage für den Nachweis von medizinischen Experten dienen, wenn es gelingt, ihn mit vertretbarem Aufwand und in hoher Qualität zu analysieren.

Die Nutzung von Analysen großer Korpora ist dabei nicht trivial, da sie zwar bei präzisen Fragestellungen belastbare Ergebnisse hinsichtlich der Dokumentenlage liefern, nicht aber den Nutzungskontext, also die Verwendung der Ergebnisse in Kommunikations- und Entscheidungsprozessen berücksichtigen. Damit fehlt jedoch die dynamische Komponente der Nutzungsprozesse, die wesentlich den Wert der Textanalysen bestimmt.

Material und Methoden

Damit stellen sich eine Reihe von Fragen an die Auswertung wissenschaftlicher Literatur, die einerseits auf eine inhaltliche Systematisierung, andererseits auf einen differenzierten Personen- und Institutionennachweis hinauslaufen. Kernfrage ist dabei, wie inhaltliche Konzepte, Informationen zu Autoren und Relationen zwischen diesen aus unstrukturierten Texten und ihren Metadaten gewonnen werden können und wie diese unterstützend in den Arbeitsprozessen von Ärzten, Forschern, Politikern und ggf. weiteren Gruppen eingesetzt werden können. Dies setzt eine hochqualitative systematische Analyse großer Textkorpora voraus, die ohne effiziente Automatismen nicht zu leisten ist.

Um diese Fragen zu klären, wurde durch SyynX Solutions auf der Basis der Collexis Fingerprint Core Engine als Ontologie-gestützter Indexierungslösung (zu den Grundlagen s. [2]) ein erster Prototyp des »German Medical Knowledge Guides« umgesetzt. In diesem werden für ca. 250.000 Publikationen Profile errechnet. Diese können aufgrund der großen in PubMed vorliegenden Publikationsmenge als repräsentativ für die einschlägigen wissenschaftlichen Aktivitäten aus Deutschland angesehen werden. Der Ansatz ist dabei grundsätzlich auf alle hinreichend großen Textkorpora übertragbar. Um die Analyse dieser Datenbasis im Hinblick auf die Expertise der Autoren einerseits und die räumliche Verteilung andererseits leisten zu können, mussten sowohl fachliche als auch geographische Ordnungssysteme in die Anwendung integriert werden. Hierfür wurden der MeSH (Medical Subject Headings), das UMLS (Unified Medical Language System) und ein geographischer Städtethesaurus für Deutschland integriert.

Bezüglich der Einbettung medizinischer Informationen stellen Prozessmodelle ein probates Mittel dar, um die Nutzungskontexte von Informationen systematisch zu erfassen und zu analysieren. Hierbei geht es in diesem Zusammenhang nicht um die Formalisierung von Behandlungsprozessen (clinical pathways), sondern um die wesentlich schlechter zu formalisierenden Kommunikationsprozesse im wissenschaftlichen und politischen Umfeld.

Ergebnisse

Mit diesem Prototypen [1] konnte gezeigt werden, dass eine vollautomatische Analyse großer Textkorpora wie PubMed als Grundlage für den personenbezogenen und regionalen Nachweis medizinischer Experten grundsätzlich möglich ist. Mit der ontologiebasierten Indexierung steht hierfür ein Werkzeug bereit, das nicht allein eine hinreichende Präzision bereitstellt, sondern auch durch die Operationen auf konzeptueller Ebene in der Lage ist, Synonymie, Homonymie und Ähnlichkeitsbeziehungen sachgerecht zu behandeln. Damit wird der Nachweis von einschlägigen wissenschaftlichen Publikationen als Ergebnismenge von sachlogischen Anfragen auf der Basis der MeSH / UMLS Ontologie ermöglicht.

Hinsichtlich der Prozesse haben zwei Studien [5 6] gezeigt, wie einerseits spezifisch wissenschaftliche Kommunikation, andererseits der Informationsbedarf niedergelassener Ärzte systematisch erfasst werden können und wie diese Ergebnisse in die Gestaltung von wissenschaftlich / medizinischen Informations- und Kommunikationssystemen eingehen können. Diese Ergebnisse geben deutliche Hinweise auch auf die funktionale Gestaltung von Systemen, in denen die personelle und / oder organisatorische Konzentration von Expertise für wissenschaftliche und politische Entscheidungsprozesse genutzt werden soll.

Diskussion

An diesem Prototyp wurde jedoch auch deutlich, dass der angestrebte Nachweis von Experten und Kompetenzkonzentration im Sinne organisatorischer Zusammenhänge nicht mit dem Nachweis von fachlich einschlägigen Publikationen gleichgesetzt werden kann. Auch zeigte sich, dass die Auswertung der Publikationsnachweise hinsichtlich personen- oder organisationsgebundener (und damit räumlich lokalisierbarer) Expertise nicht trivial ist. Vielmehr sind gerade bei einer derartigen Analyse von großen Korpora weitere Schritte notwendig, um die primären Analyseergebnisse für Anwendungen, wie den einleitend skizzierten nutzbar machen zu können.

Hierzu gehören vor allem Methoden, die Experten nicht allein als Namen repräsentieren, sondern als komplexere Konzepte unter Einbezug inhaltlicher Schwerpunkte und ihrer Position in Autorennetzwerken. Dieser Ansatz stellt nicht nur eine Lösung des Homonymieproblems bei Namen dar (zu einer vergleichbaren Problemstellung vgl. [4]), sondern liefert eine oft gewünschte Kontextualisierung der Personen, die ausgehend von mengenorientierten Recherchen navigierende Zugriffsformen ermöglicht. Hierfür ist eine weitergehende Auswertung sowohl der Publikationsmetadaten, insb. Autorennamen und Affiliationfeld, als auch des Abstracts hinsichtlich der thematischen Einordnung sowie der Verortung in Wissenschaftsprozessen nötig (s. hierzu [3]). Dies stellt insbesondere aufgrund der zeitlichen Abhängigkeit von Teilen dieser Informationen (Wechsel der Anstellung, Abschluss von Projekten) eine Herausforderung dar, die für die Nutzbarkeit von Anwendungen wie dem lokalen, organisationalen und persönlichen Nachweis von Expertise essentiell ist.

Mit diesem dynamischen Anteil innerhalb der Grundinformationen steht eine dynamische Komponente in der Nutzung dieser Informationen in Verbindung. Der Wert desselben Rechercheergebnisses unterscheidet sich deutlich je nach Nutzungskontext. Aus diesem Grund scheint eine eingehendere Erhebung und Analyse der sich mit den technologischen Möglichkeiten ändernden Nutzungsprozessen dringend geboten. Insbesondere [5] hat gezeigt, dass hierfür die traditionellen Methoden der Geschäftsprozesserhebung und –modellierung nur bedingt tauglich sind. Entsprechend muss, exemplarisch bei der Entwicklung eines über den Prototypen hinausgehenden »German Medical Knowledge Guide« die »guidance« in Form von Kommunikations- und Nutzungssterotypen präziser entwickelt und in der Einbettung des GMKG in die operativen Systeme der Anwender berücksichtigt werden.

Nicht zuletzt stellen sich auch Fragen der Darstellung der Rechercheergebnisse. Hierfür sind Darstellungsformen notwendig, die eine multidimensionale Sicht erlauben, in der nicht allein ein einzelnes Ordnugnskriterium wie z.B. alphanumerische Sortierung nach Nachnamen oder zeitliche Sortierung nach Erscheinungsdatum ausgewertet wird, sondern komplexe Relevanzmaße in Relation zur Fragestellung (z.B. nach potenziellen Gutachtern im Gegensatz zu potenziellen Kooperationspartnern) und Nutzungszusammenhang als Grundlage der Darstellung entwickelt werden.

Literatur

[1] Herzog C, Liuzzi G, Diwersy M. SyynX solutions: practical knowledge management in a medical environment. Bremen: ACM Press,CIKM 2005: 556-559
[2] van Mulligen EM, Diwersy M, Schmidt M, Buurman H, Mons B. Facilitating networks of information. Proc AMIA Symp. 2000;:868-72.
[3] Stuckenschmidt H, Siberski W, Mulligen E. Towards Mapping-Based Document Retrieval in Heterogeneous Digital Libraries. In: Goble C, Kesselman C, Sure Y, Hrsg. Semantic Grid: The Convergence of Technologies. Dagstuhl Seminar Proceedings, Dagstuhl, 2005.
[4] Schijvenaars BJ, Mons B, Weeber M, Schuemie MJ, van Mulligen EM, Wain HM, Kors JA Thesaurus-based disambiguation of gene symbols. BMC Bioinformatics 2005 Jun 16, 6:149
[5] Walter R u.a. Kommunikationsbeschleuniger in der virtuellen Wissenschaft. ITA-Studie, Bonn bmb+f, wird veröffentlicht im Frühjahr 2006
[6] Koch O, Reuter C, Vollmer G. Bedarfsgerechte Unterstützung von Ärzten an ihrem Arbeitsplatz über informationslogistische IT-Anwendungen. Projektabschlussbericht. Dortmund Fraunhofer ISST, Dez 2005.