INFORMATION RETRIEVAL IN KLINISCHEN FREITEXTDOKUMENTEN

Spat S[1], Cadonna B[2], Rakovac I[1], Gütl C [3], Leitner H[4], Stark G[4], Beck P[1]

Kurzfassung

Die Anzahl der gespeicherten Freitextdokumente in Elektronischen Patientenakten (EPA) steigt. Um Behandlungsentscheidungen zu treffen ist das medizinische Krankenhauspersonal auf relevante Informationen in diesen Freitextdokumenten angewiesen. Da eine Suche in unstrukturierte Dokumenten sehr viel Zeit und Know-how beansprucht, stellt diese Arbeit den Prototyp eines klinischen Information Retrieval Systems zur effizienten und effektiven Suche nach relevanten medizinischen Informationen in klinischen Freitextdokumenten vor.

1. Einleitung

Die Menge der elektronisch gespeicherten Textinformationen steigt kontinuierlich. Vor allem die Einführung Elektronischer Patientenakten (EPA) stellt neue Anforderungen an die Suche nach medizinisch relevanten Informationen in klinischen Freitextbeständen.

Die Steiermärkische Krankenanstaltungenges.m.b.H. (KAGes), die Trägergesellschaft von 20 Spitälern der Steiermark mit über 6.000 Betten und 16.000 Mitarbeiterinnen und Mitarbeitern, führte im Jahr 2004 ein neues Krankenhausinformationssystem mit dem Projektnamen openMEDOCS ein. Ziel war es, die heterogenen IT-Systeme der einzelnen Spitäler durch ein einheitliches zentrales System zu ersetzen. Die Grundlage bildet das Softwarepaket *IS-H* von SAP sowie *i.s.h.med* von GSD and T-Systems. Den Kern von openMEDOCS bildet eine Elektronische Patientenakte, in der alle medizinischen Daten der KAGes-Patienten erfasst werden. [3,4] Ein beträchtlicher Teil dieser Informationen sind Freitextdokumente wie Arztbriefe oder Befunde. Die medizinischen Informationen dieser Dokumente haben große Bedeutung für die Entscheidungsfindung der Ärztinnen und Ärzte hinsichtlich der weiteren Behandlung ihrer Patientinnen und Patienten. Zur Unterstützung der Entscheidungsfindung stellt diese Arbeit einen kombinierten Ansatz - aus *Text Information Retrieval* und *automatisierter Textklassifikation* - zur Suche nach relevanten Informationen in klinischen Freitextdokumenten vor.

[1] Institut für medizinische Systemtechnik und Gesundheitsmanagement, JOANNEUM RESEARCH Forschungsgesellschaft mbH, Graz
[2] Fakultät für Informatik, Freie Universität Bozen, Italien
[3] Institut für Informationssysteme und Computer Medien, Technische Universität Graz
[4] Steiermärkische Krankenanstaltenges. m.b.H., Graz

2. Material und Methoden

Das Design des kombinierten Ansatzes aus Text Information Retrieval und automatisierter Textklassifikation wird in diesem Kapitel anhand des Prototyps eines medizinischen Information Retrieval Systems (MIRS) beschrieben.

Unstrukturierte klinische Freitextdokumente, die aus dem openMEDOCS System der KAGes extrahiert wurden, werden in einer vereinfachten elektronischen Patientenakte (EPA) im MIRS gespeichert. Für die Indexierung, die Suche, sowie für die Klassifikation werden diese Dokumente aus dieser EPA durch das *„DATENBANK-Modul“* angefordert. Das *„KLASSIFIKATIONS-Modul“* trainiert und evaluiert Algorithmen zur automatisierten Klassifikation von klinischen Freitextdokumenten in zuvor definierte Kategorien (hier: medizinische Fachbereiche). Das *„INDEXIERUNGS-Modul“* extrahiert Index-Terme aus den Dokumenten und speichert diese Information, zusammen mit weiteren Metadaten wie Datum der letzten Modifikation, dem Namen des Dokuments, sowie die durch die Dokumentenklassifikation ermittelten medizinischen Fachbereiche des Dokuments im *„INDEX“*. Das medizinische Personal definiert ihre Suchanfrage über das *„BENUTZEROBERFLÄCHEN-Modul“*. Die Benutzeranfrage, bestehend aus Termen und Metadaten, wird an den *„INDEX“* weitergeleitet und Dokumente werden nach Relevanz sortiert an den Benutzer zurückgeliefert. Durch die Wahl von medizinischen Fachbereichen in der Suchmaske, kann der Benutzer direkt Einfluss auf die Relevanz der zurückgelieferten Dokumente nehmen. *Abbildung 1*gibt einen Gesamtüberblick über das Design des MIRS.

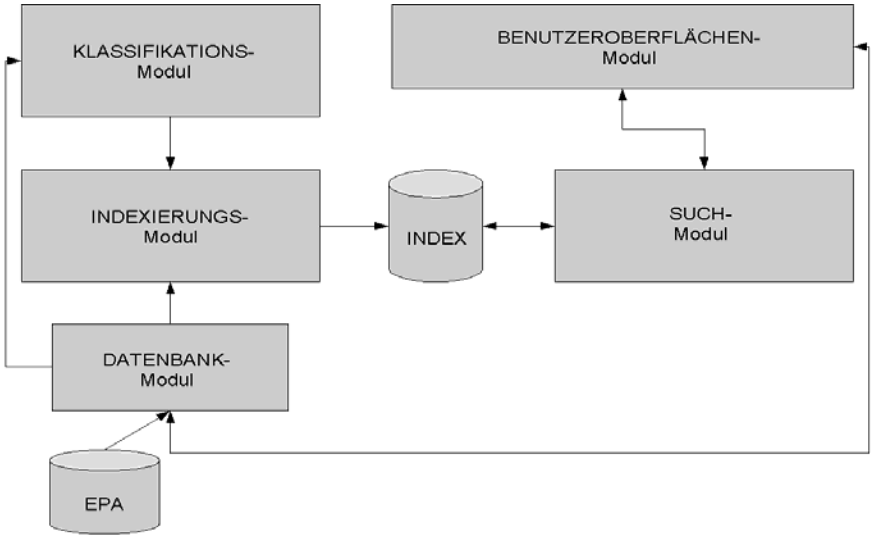

Abbildung 1: Design des MIRS Prototyps

2. 1. EPA

Da ein direkter Zugriff auf die klinischen Freitextdokumente der KAGes aus technischen und Datenschutzgründen nicht möglich war, wurde für die Speicherung eine vereinfachte elektronische Patientenakte (EPA) modelliert. Die KAGes extrahierte aus openMEDOCS 18.000 Freitextdokumente, die in der modellierten EPA gespeichert wurden. Insgesamt finden sich in dem extrahierten Dokumentensatz 26 unterschiedliche Dokumenttypen wie Arztbriefe oder Befunde aus acht medizinischen Fachbereichen (Chirurgie, Gefäßchirurgie, Interne Medizin, Neurologie, Anästhesie und Intensivmedizin, Radiologie und Physiotherapie). Dokumente wurden als reine Textdokumente zur Verfügung gestellt. Jedes Dokument kann einem anonymisierten Patienten zugewiesen werden.

2. 2. KLASSIFIKATIONS-Modul

Zur automatisierten Klassifikation von unstrukturierten klinischen Textdokumenten in medizinische Fachbereiche wurde ein *multi-label Klassifikationssystem* [8] basierend auf dem Open-Source Data Mining Framework WEKA [9] entwickelt. 1.500 zufällig ausgewählte Freitextdokumente aus den extrahierten klinischen Dokumenten der KAGes wurden von einem Domain Experten (Internisten) händisch kategorisiert und in einen Trainings- und einen Evaluierungsdatensatz geteilt. Vier unterschiedliche Klassifikationsalgorithmen (J48, SMO, k-NN, und Naïve Bayes [7,6,1,2]) wurden mit diesen Datensätzen trainiert bzw. evaluiert.

Obwohl die Information, aus welchem medizinischen Fachbereich ein Dokument extrahiert wurde, zur Verfügung steht, kann es vorkommen, dass das Dokument auch für andere Fachbereiche interessant ist. Die automatisierte multi-label Klassifikation erlaubt, diese Möglichkeit zu berücksichtigen und ein Dokument mehreren Fachbereichen zuzuordnen.

2. 3. INDEXIERUNGS- und SUCH-Modul

Für die Entwicklung des Such- bzw. Indexierungsmechanismus wurde das Open-Source Framework Apache Lucene [5] verwendet. Um die Relevanzberechung der gefunden Dokumente zu beeinflussen, bietet Apache Lucene die Möglichkeit einen *„boost factor“* zu setzen. Das bedeutet im Falle des MIRS, dass Dokumenten, die einem oder mehreren medizinischen Fachbereichen zugeordnet wurden, bei der Auswahl dieser Fachbereiche in der Suchmaske eine höhere Relevanz zugewiesen wird, als Dokumenten die nicht in diese Fachbereiche klassifiziert wurden.

2. 4. J2EE Webanwendung

Die Konzeption des medizinische Information Retrieval System als J2EE-Webanwendung erlaubt, neben Plattformunabhängigkeit, Modularität, Erweiterbarkeit und Datenschutzfunktionen, einen einfachen Zugriff auf das MIRS über einen Webbrowser.

3. Ergebnisse und Diskussion

3. 1. Automatisierte Dokumentenklassifikation

Wie bereits in Kapitel 2. 2 erwähnt, wurden für die Klassifikationsaufgabe vier Klassifikationsalgorithmen mit einem händisch kategorisierten Dokumentendatensatz von 1.500 klinischen Freitextdokumenten trainiert bzw. evaluiert. J48, ein Klassifikationsalgorithmus basierend auf einem Entscheidungsbaum, erreichte mit einer „F_1-measure“ von 0.89 das beste Ergebnis [8] und wurde im MIRS eingesetzt um alle 18.000 klinischen Freitextdokumente in eine oder mehrere der acht medizinische Fachbereiche (siehe Kaptitel 2. 1) zu klassifizieren.

3. 2. Beispielhafte Anwendung des MIRS

Im Folgenden wird anhand einer beispielhaften Anwendung des MIRS Prototyps der Effekt des *„relevance boosting“* dargestellt:

Eine Internistin interessiert sich für die Krankheitsgeschichte ihres Patienten mit der ID 12019922. Sie möchte klinische Textdokumente, in denen das Wort „Herz“ vorkommt, vorrangig lesen. Also gibt sie „Herz*“ in die Suchmaske ein. Der „*“ ist Platzhalter für beliebige weitere Zeichen nach

dem Wort „Herz“. Als Internistin ist sie an Dokumenten der Fachbereiche „Innere Medizin“ sowie „Chirurgie“ interessiert, daher setzt die Ärztin in der Suchmaske eine Marke für diese Fachbereiche.

Nach der Übermittlung der Suchanfrage, ermittelt das MIRS all jene Dokumente die das Wort „Herz“ beinhalten. Zusätzlich wird die Relevanz dieser Dokumente bezüglich der Suchanfrage berechnet. Im nächsten Schritt werden jene Dokumente, die vom Klassifikationsalgorithmus in die Kategorien „Innere Medizin“ bzw. „Chirurgie“ klassifiziert wurden mit einem höheren Relevanz-Faktor gewichtet, als jene, die nicht in diese Kategorien fallen. Höher gewichtete Dokumente erscheinen in der Ergebnisliste vor niedriger gewichteten. Dokumente, die zwar das Wort „Herz“ enthalten, aber in keine der beiden medizinischen Fachbereiche klassifiziert wurden, sind am Ende der Ergebnisliste zu finden.

Abbildung zeigt einen Ausschnitt aus der Ergebnisliste. Insgesamt wurden 24 Dokumente mit dem Wort „Herz“ im Text für den Patienten mit der ID 12019922 gefunden. Neben einer kurzen Vorschau auf den Inhalt des Dokuments, werden das Datum der letzten Modifikation, der Dokumenttyp sowie die Kategorien in die das Dokument automatisiert klassifiziert wurde, dargestellt. Die Spalte „Score“ zeigt die berechnete Relevanz des Dokuments bezüglich der Suchanfrage. Das Dokument, welches in beide Kategorien klassifiziert wurde, besitzt die größte Relevanz. Anschließend folgen Dokumente, die eine Kategorie aus der Suchanfrage enthalten. Am Ende der Ergebnisliste (nicht dargestellt) finden sich all jene Dokumente ohne Übereinstimmung der Kategorien.

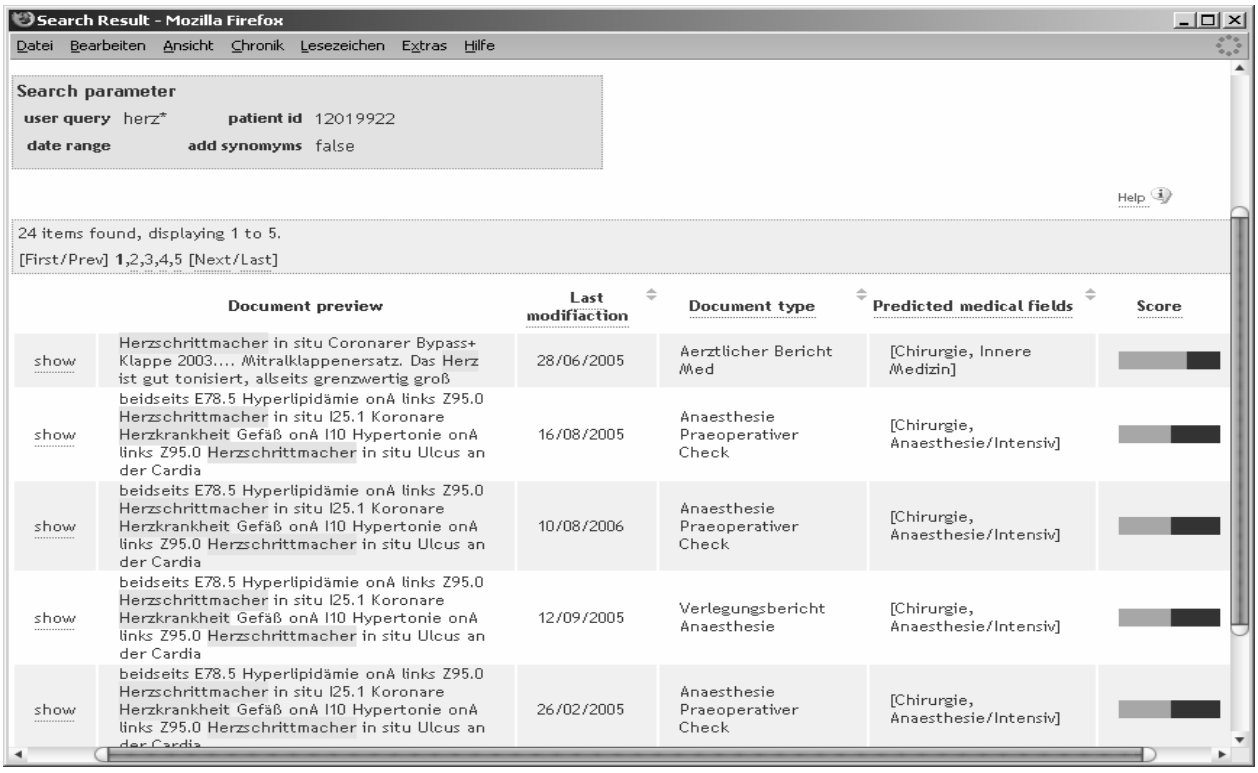

Abbildung 2: Ausschnitt aus der Ergebnisliste

Zur Untersuchung der praktischen Relevanz des MIRS Prototpys wurden fünf erfahrenen Klinikärztinnen und -ärzten vier unterschiedliche Suchaufgaben vorgelegt. Bei zwei der Aufgaben durften die Probanden medizinische Fachbereiche als *‚boost factor‘* setzen, bei den zwei anderen Aufgaben

nicht. Anschließend wurden die Probanden gebeten, einen Fragebogen auszufüllen, der Usability, Antwortzeiten des Systems, sowie den Einfluss der Fachbereichs-*,boost factor'* auf die Suche in der EPA abgefragte. In einer ersten qualitativen Analyse der Ergebnisse zeigt sich, dass die Probanden überwiegend eine Beschleunigung der Informationssuche – vor allem in Patientenakten mit vielen Dokumenten – feststellten. Weiters wurde angemerkt, dass der Einsatz von medizinischen Fachbereichen als *,boost factor'* eine feiner granulierte Suche ermöglicht und das dadurch fachbereichsspezifische Informationen schneller gefunden werden können.

4. Diskussion und Ausblick

Da das MIRS als Prototyp implementiert wurde, ist dessen Funktionalität auf eine einfache Indexierung von klinischen Freitextdokumenten und die Suche in diesen Dokumenten beschränkt. Weder strukturierte Daten wie Diagnosen, noch andere Medien, wie Bilder, wurden berücksichtigt. Eine erste qualitative Analyse des Prototyps deutet auf eine gute Akzeptanz der potentiellen Nutzer des MIRS hin. Eine detaillierte Untersuchung des Prototyps hinsichtlich Benutzerakzeptanz, der Qualität der Suchergebnisse und der mittleren Antwortzeiten des Systems sind Gegenstand zukünftiger Untersuchungen.

In Hinblick der Integration des MIRS in die Elektronische Patientenakte (EPA) von openMEDOCS, ist die Einbeziehung aller vorhandenen Metadaten der EPA, sowie die Suche und Darstellung in weiteren Datenbeständen (z.B. Diagnosen), als auch anderen Medien als Text (z.B. Bilder), von besonderer Bedeutung. Dem Benutzer sollen dadurch möglichst viele patientinnen- bzw. patientenbezogene Daten zur Verfügung stehen. Auch ist für zukünftige Untersuchungen interessant, welche zusätzlichen Informationen aus dem Dokumentensatz der EPA durch die automatisierte Textklassifikation gewonnen werden können.

5. Schlussfolgerung

Die Zunahme von unstrukturierten klinischen Freitextdokumenten bedarf „natural language processing (NLP)"-Techniken wie Text Information Retrieval oder der automatisierter Klassifikation von Freitextdokumenten um relevante medizinische Informationen in einem großen Datensatz zu finden. Basierend auf etablierte Open-Source Frameworks, bietet diese Arbeit einen kombinierten Ansatz beide Techniken in einem medizinischen Information Retrieval System zu nutzen.

6. Literatur

[1] AHA D.W., KIBLER D., ALBERT, M.K. Instance-based learning algorithms, Machine Learning; 6(1): 37-66, 1991.

[2] JOHN G., LANGLEY P. Estimating continuous distributions in bayesian classifiers, paper presented to Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, Vancoucer, Canada, 1995. Apache Lucene: http://lucene.apache.org/, Datum des letzten Zugriffs: 2008-2-1.

[3] KRAßNITZER M. EDV im Spital: Der Patient auf Knopfdruck, CliniCum, 7-8/2006, Online; http://www.medical-tribune.at/dynasite.cfm?dssid=4171&dsmid=74897 &dspaid=582979, Datum des letzten Zugriffs: 2008-2-1.

[4] LEITNER H. openMEDOCS erfolgreich eingeführt, G'sund.net, 50/2006, Online: http://www.gsund.net/cms/beitrag/10073293/2052790, Datum des letzten Zugriffs: 2008-2-1.

[5] Apache Lucene: http://lucene.apache.org/, Datum des letzten Zugriffs: 2008-2-1.

[6] PLATT J. Fast training of support vector machines using sequential minimal optimization, In B. Scholkopf B, Burges C and Smola A, Advances in Kernel Methods - Support Vector Learning, MIT Press, 1998.

[7] QUINLAN, J.R. C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[8] SPAT S., CADONNA B., RAKOVAC I., GÜTL C., LEITNER H., STARK G., BECK P. Multi-label text classification of German language medical documents. Stud Health Technol Inform. 2007;129:1460-1.

[9] WITTEN, I. H., FRANK E. Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann, San Francisco, 2005.