Enhanced Information Retrieval from Narrative German-language Clinical Text Documents using Automated Document Classification

Stephan SPAT a,1, Bruno CADONNA a, Ivo RAKOVAC a, Christian GÜTL b, Hubert LEITNER c, Günther STARK c and Peter BECK a

a Institute of Medical Technologies and Health Management, JOANNEUM RESEARCH Forschungsgesellschaft mbH, Graz, Austria
b Institute for Information Systems and Computer Media, Graz University of Technology, Graz, Austria
c Steiermärkische Krankenanstaltengen. m.b.H., Graz, Austria

Abstract. The amount of narrative clinical text documents stored in Electronic Patient Records (EPR) of Hospital Information Systems is increasing. Physicians spend a lot of time finding relevant patient-related information for medical decision making in these clinical text documents. Thus, efficient and topical retrieval of relevant patient-related information is an important task in an EPR system. This paper describes the prototype of a medical information retrieval system (MIRS) for clinical text documents. The open-source information retrieval framework Apache Lucene has been used to implement the prototype of the MIRS. Additionally, a multi-label classification system based on the open-source data mining framework WEKA generates metadata from the clinical text document set. The metadata is used for influencing the rank order of documents retrieved by physicians. Combining information retrieval and automated document classification offers an enhanced approach to let physicians and in the near future patients define their information needs for information stored in an EPR. The system has been designed as a J2EE Web-application. First findings are based on a sample of 18,000 unstructured, clinical text documents written in German.

Keywords. EPR-CPR-EMR, Classification, Data analysis-extraction tools

Introduction

The amount of digital textual information in hospitals is increasing. Especially the implementation of Electronic Patient Record (EPR) systems, which
manage the patient related electronic information, have reinforced the demand for finding relevant information, particularly in unstructured clinical text documents.

Steiermärkische Krankenanstalten Ges.m.b.H. (KAGes), the governing body of Styrian hospitals, covers 20 hospitals with about 6,000 beds and 16,000 employees for over 1.2 million inhabitants. In order to displace heterogeneous IT-systems of numerous hospitals by offering an integrative Hospital Information System (HIS), in 2004, the roll-out of a new HIS – termed openMEDOCS – was conducted. openMEDOCS is the synonym for the implementation of the software packages IS-H from SAP and i.s.h.med from GSD and T-Systems. The core of openMEDOCS is an EPR system which stores all documents of patients generated in KAGes hospitals. [1-2]

Considerable amounts of managed data in the EPR system are narrative clinical text documents. Since the roll-out of openMedocs, the amount of clinical text documents has increased continuously. Since no effective retrieval tool is available yet, physicians spend a lot of time finding relevant information in patient records. Thus, efficient and effective retrieval of relevant patient-related information stored in these documents has gained importance. Consequently, a project leader of openMEDOCS asked for a prototype of a medical information retrieval system (MIRS).

Clinical text documents like discharge letters or reports are highly relevant for physicians’ decision-making processes regarding patient-care. These text documents are often narrative, non-structured, with few metadata. Approaches for standardizing narrative texts have been developed with HL7 CDA [3], however, they are mostly not implemented in clinical practice yet and do not necessarily include metadata needed for efficient information retrieval. Thus, text-mining may be a viable method for finding relevant information in non-structured documents [4-8].

In this paper, we describe an enhanced approach combining metadata generation by means of automated document classification with standard information retrieval techniques, which we implemented in a prototypical medical information retrieval system (MIRS).

1. Materials and Methods

In order to provide physicians with a tool to improve the search capabilities in the EPR system of openMEDOCS regarding efficiency and effectiveness of finding relevant patient-related information, this chapter describes the design of the prototypical MIRS.

Figure 1 illustrates the overall design: narrative clinical documents are stored in the modeled EPR system (‘EPRs’). For indexing, searching, and classification, these documents are gathered from EPRs by the ‘DB Component’.

The ‘Classification Component’ trains and evaluates classification models for classifying unseen narrative clinical text documents into an arbitrary number of predefined categories (i.e. medical fields).

The ‘Index Component’ extracts index terms from the documents and stores them together with additional metadata like document name, date of last modification as well as medical fields predicted by the ‘Classification Component’ in the ‘Index’.

Physicians define their information needs using the ‘User Interface (UI) Component’. The user question (search string and metadata) is analyzed and terms are extracted. Furthermore, the medical field categories generated by document classification can be used influencing the rank order of retrieved documents. In the
following, a query is written including this information. The query is used to search within the ‘Index’.

Finally, a ranked document list with snippets of the document content is returned to the user. If the snippet of the document presented to the user has aroused interest, the original document may be opened.

![Figure 1. Overall design of the MIRS.](image)

1.1. EPRs/DB Component

A simplified EPR system was modeled for the prototype to store and process text documents extracted from openMEDOCS’ EPR system. 18,000 narrative clinical text documents written in German were extracted. The extracted set comprises 26 different types of clinical documents (e.g. discharge letters, reports) from eight medical fields (surgery, vascular surgery, casualty surgery, internal medicine, neurology, anaesthesia and intensive care, radiology and physiotherapy). Documents were provided as plain text. Each document can be associated with a specific patient.

1.2. Classification Component

For the purpose of classifying documents in medical fields, a multi-label text classification system [9] using the open-source data mining framework WEKA [10] was developed. The classification system was trained with a sample of 1,500 text documents manually classified by a domain expert (physician).

Four different classification models were compared for the classification task: J48 – a tree based classifier, SMO – an implementation of support vector machines, k-NN – an instance based classifier, and Naïve Bayes – a classifier based on probabilistic theory [11-14]. For each model, a 10-fold cross validation was conducted. The F1-measure [15] was used as comparable evaluation measure.

Multi-label classification was used because information in a document can refer to multiple medical fields. For example, a person who had had an accident and had broken her leg was first in the radiology to X-ray her leg. Afterwards, she had an operation in the medical department of casualty surgery. All stages of treatment were documented in a single document generated by the department of casualty surgery but...
the information in the document refers to the medical fields ‘radiology’ and ‘casualty surgery’. Multi-label classification allows classifying a document in multiple medical field categories.

1.3. Index and Search Components

The open-source information retrieval framework Apache Lucene [16] was used for the implementation of the indexing and searching components of the medical information retrieval system (MIRS) prototype. Apache Lucene allows setting a ‘boost factor’ which influences the relevance score of retrieved information. If the user retrieves information with particular attention to specific medical fields, documents classified in these medical fields will have a high relevance score pushing the documents to the top of the results list.

1.4. J2EE Web Application

The system was designed as a J2EE web application which has the following advantages: (1) Platform Independence, (2) Multi-Access, (3) Modularity/Re-usability, and (4) Data Security. For future applications, the design allows patients to access their own medical records with an ordinary web-browser.

2. Results and Discussion

2.1. Automated Document Classification

Four different classification schemes have been compared for the classification task: J48, SMO, k-NN, and Naïve Bayes. Since evaluation showed J48 classification model to achieve best results with a F1 measure of 0.89 [9], it has been used to classify clinical documents in the developed MIRS prototype.

2.2. Sample Application of MIRS

In the following, a sample application of the MIRS prototype and especially the effect of ‘relevance boosting’ will be illustrated.

A physician is interested in particulars about medical treatments of the ‘herz’ (heart) of a patient with ID 12019922. The user enters the user query ‘herz*’ into the search mask. The use of the wildcard ‘*’ expands this term to all words which start with ‘herz’. Additionally, the physician is particularly interested in documents which belong to the medical fields ‘Chirurgie’ (surgery) and ‘Innere Medizin’ (internal medicine). Therefore, the user sets the medical field booster to ‘Chirurgie’ and ‘Innere Medizin’.

After the submission of the query settings, the MIRS prototype first looks for all documents which contain at least one word which starts with ‘herz’. In a second step, ‘relevance boosting’ applies, i.e. the MIRS calculates relevance scores for found documents resulting in higher scores for documents which were classified as ‘Chirurgie’ or ‘Innere Medizin’ and lower scores for those which were not classified in those classes. The corresponding list of results shows documents ordered by relevance.
score in descending order. It is important to note that documents which were not classified into one of the two classes are not excluded from the list of results because the classification might have been erroneous or incomplete. They can be found at the bottom of the list of results.

Figure 2 shows an extract of the list of results. The document at the first position in the list was classified as ‘Chirurgie’ and ‘Innere Medizin’, which results in the highest relevance score. The following documents were assigned just to one of the chosen classes resulting in lower relevance scores. On the bottom of the list (not shown) are documents which are not assigned to any of the classes chosen by the physician.

In a first evaluation, four searching tasks – conducting two tasks with and two tasks without ‘relevance boosting’ – were given to five clinicians who are using openMEDOCS in their practical routine. Qualitative feedback emphasizes that the MIRS could improve search speed and search quality in clinical text documents. The clinicians also underlined that ‘relevance boosting’ speed up searching, especially in patient records with many text documents.

The implemented prototype shows an approach to influence the rank order of retrieved document using automated multi-label document classification. Therefore, documents have been classified into medical fields with good results.

As the implementation of MIRS is prototypical, functionality is restricted to basic search and index functions. Neither structured data, like diagnosis codes, nor other media like X-ray radiographs were considered. Moreover, profound evaluation of the prototype regarding usability, retrieval quality, and response time was left for future work.
3. Conclusion

The increasing number of narrative clinical text documents demands for natural language processing techniques like information retrieval or automated document classification [4]. Based on well established open source frameworks, this paper presents an enhanced approach combining both in one medical information retrieval system.

References