The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

Knowledge Representation in Digital Medical Documents for Efficient Information Access and Retrieval

Simon Hoelzer, Ralf Kurt Schweiger, Joachim Dudeck

Institute for Medical Informatics, University of Giessen, Germany
www.patientcare.de; sh@patientcare.de

Abstract:

Information access and retrieval are essential to serve the delivery and application of evidence-based medicine. The literature provides evidence for the effectiveness of computerization of medical knowledge for increasing compliance with current standards and improving patient outcomes. But all efforts in implementing knowledge-based functions remain limited in respect of either the prolonged exposure to live clinical use, the complexity of knowledge content supported, the dependencies on a technology platform and host system, or the proof of use and content sharing across multiple and varying environments. We have to face currently an information overload, especially in medicine. As a result, efficient ways of implementation of evidence-based care become more and more important. Most of the medical knowledge contained in textbooks, guidelines, journals as well as online resources is text-based. In order to present problem-specific information at the point of care the outlined concept relies on a document-based solution with the eXtensible Markup Language (XML). The adoption of XML as a standard for the publication and interchange of documents creates a great opportunity for better information retrieval. On the basis of structured data we are able to improve the search quality for clinical information which forms a crucial pre-requisite for the implementation of evidence-based care.

1 Introduction

Information access and retrieval are essential to serve the delivery and application of evidence-based medicine. Information about of the state-of-the-art medical care can be brought to the physician by different means and media. The literature provides evidence for the effectiveness of computerization of medical knowledge for increasing compliance with current standards and improving patient outcomes [1-3]. But the most effective concepts, such as knowledge-based functions for decision support or decision monitoring that are integrated in clinical information systems, are restricted by the efforts required for development and maintenance of the information systems and the limited number of implemented medical rules [4-6]. These reasons lower the acceptance by the developer as well as the intended end user (physician). Only a few systems that support guideline-based care in an automated fashion progressed beyond the prototype stage [7;8]. On the other hand, information in health care is, to a very large extent, transmitted and stored as unstructured or slightly structured text. Most of the medical knowledge contained in textbooks, guidelines, journals as well as online resources is text-based and is represented as text, figures and tables. It seems to be straight-forward to rely on a document-based solution in order to present problem-specific information at the point of care. For this reason, we have chosen a document-based approach using the eXtensible Markup Language (XML) for restructuring, searching and presenting available resources. The adoption of XML as a standard for the publication and interchange of documents creates a great opportunity for better information retrieval. XML has the ability to represent the semantics of data in a structured, documented, machine-readable form.

2 Objectives

We established a cooperative project with a medical internet portal / weekly journal (www.medical-tribune.de) and the editorial board of the most popular German-speaking textbook of Internal Medicine ("Innere Medizin", editor: G. Herold, current issue 2001). The objective is to provide an infrastructure to support the development of web-based services for retrieving and presenting medical knowledge in order to enhance the implementation of evidence-based care in the clinical routine. The information retrieved can be classified in one of two categories of measurable statistics. Whether the information retrieved is considered relevant, or whether all relevant material was retrieved. These two metrics of recall and precision serve to express information retrieval performance. Recall is percentage of total relevant documents retrieved from all documents. Recall refers to how much information is retrieved by the search. Total recall would locate every document that matched the search criteria in a database. Precision is the percentage of documents retrieved that the searcher is actually interested in. Precision focuses on the relevant, most useful items retrieved in the search. The goal of information retrieval is to provide the most precise or relevant documents in the midst of the recalled search results. By introducing semantics inherent in XML documents into the search process we can enhance the recall and precision of the query.

3 Methods

The last decade was strongly influenced by the development of web-based applications and infrastructure. Intranet-based solutions have been found to be an efficient method of integrating local and Internet information systems within a clinical environment [9-11]. Online catalogs and search engines (yahoo.com, lycos.com, google.com, hotbot.com) offer keyword searching and Boolean searching to assist in precision. With more data to search, the search engine can return more documents gaining greater recall. But a lot of work remain at the user to sift through the recalled documents to find useful ones. The use of metainformation, concept-based searching, relevancy weighing, probabilistic logic etc. are general approaches to overcome these drawbacks.
XML has emerged as a new standard for data representation and exchange on the Internet. It is believed that it will become a universal format for data exchange on the Web and that in the near future we will find vast amounts of documents in XML format on the Web. XML addresses the limitations of Hypertext Markup Language HTML in that it is (a) extensible (unlimited number of self-created tags), it supports (b) the specification of deep nested structures, e.g., needed to represent database schemas, and allows (c) for the checking of data for structural validity. Today XML is a World Wide Web Consortium (W3C) Recommendation. Today, we are restricted by the functionality of the current browser generation. For this reason the development of XML-based applications (with a standard browser front-end) is difficult. Technical detours at the client side using HTML instead of XML or CSS version 1 instead of XSL can't be avoided (see results).

4 Results

At the moment, the German textbook of Internal Medicine ("Herold") is available in a paperback version with approximately 800 pages. The textbook is used by most of the physicians of internal / general medicine as well as medical students in Germany, Switzerland and Austria. It is an up-to-date information source of current practice in

Internal Medicine. It is comparable with the "Harrison" textbook of Internal Medicine (www.harrisononline.com) that is also well-known outside the Anglo-American medical community, but far less comprehensive than the latter. The textbook ("Herold") is divided in chapters, sections, and paragraphs. The chapters (e.g., cardiology, hematology, endocrinology) contain different diagnosis-specific sections (e.g., ventricular tachycardia, anemia, diabetes). Depending on its content, each paragraph begins with a predefined, abbreviated header, such as therapy, classification, complications or prognosis. The source of the print version of this textbook consists of Microsoft Word'97 documents to each section.

Conceptual Modeling

We defined a reference data model for this text resource based on (a) the structure, (b) the content and notations of the paragraphs, and (c) other relevant information found in the text or layout objects. This XML model, represented by the XML schema, defines a flat hierarchy of all textbook elements that are needed for query and presentation of the medical content. In our model we differentiate, in general, between so called "medical content" elements and "search criteria". The medical content elements derive from the natural structure inherent in any medical information resource. They are used to define and name the content of the coherent units of a medical document, such as <diagnosis>, <pathology>, or <therapy>. The model orientates primarily on the meaning of the different textual parts (paragraphs, sections, etc.) found in these resources. Each "medical content" elements can be used as often as needed in the order defined by the individual document content and structure. Medical content elements have the following properties: Universal attributes to store implicit information, such as URLs, codes, or level of evidence. In addition, a flexible and optional set of further (sub-) elements in combination with the so called "search criteria" (elements such as : <code>, <characteristics>) to structure lower levels of the documents. The use of these elements is flexible and individually extendable.

Within the universal attributes as part of each medical content element we are able to

- link to other resources (http address - URLs according to the URI concept),
- link to corresponding text elements (idref concept of XML) of the same document (internal links),
- insert meaningful codes in accordance with controlled vocabularies (ICD, MeSH etc.),
- code the underlying scientific evidence of a recommendation, clinical statement etc. (evidence) and / or
- insert comments.

The search criteria and the "medical content" element <diagnosis> are designated to be the entry points for querying the XML documents or linking them to an electronic patient record. Codes that are used may clearly identify diagnoses, symptoms, comorbidities, complications (ICD), or classify medical procedures and interventions (ICPM or OPS 301 in Germany). The more coded information is available using the element <code> for explicit, in-text coding or the attribute <codes> for implicit, assigned coding the better an automatic exploitation or linking of XML resources will be supported. The element <characteristics> is used to identify and tag text units that describe patient characteristics (age, gender, ...) and specific features of disease (type, stage, entity, ...).

Conversion to XML

The following step consisted in restructuring the original textbook from its Microsoft Word format to XML according to the XML schema. The Word documents contain a lot of formatting information for the layout of the print version. By simply converting this

text in plain text (ASCII) and then in XML, we lose this valuable information. Especially the information of how to display headers, enumeration, and tables get lost. For this reason, we convert the Word documents in HTML. In several steps this HTML code has to be optimized (deleting Word Meta tags, Word XML tags, nested or empty tags, inline CSS styles etc.) and has to be converted in well-formed HTML (XHTML). Now we can add the XML markup in compliance with our reference data model (XML schema) which means to mix up HTML with XML markup.

Query and Presentation

Based on the XML schema a so called document manager (for details see: http://www.gca.org/papers/xmleurope2000/papers/s32-05.html) can provide the user with an HTML form to query existing XML documents. As already mentioned, each XML document represents a single diagnosis-specific section of the textbook. Once the relevant document(s) is (are) found, e.g., where <diagnosis> = "thyroid cancer" or "C73" (code of the International Classification of Diseases in Oncology), the content can be displayed according to the user's needs. In the next step, CSS is used for rendition and presentation of specific sections of the textbook. With stylesheets we have the capability to display (a) each distinct element with specific layout features (color, font, size, border, frame, etc.) and (b) supplement their content with additional information that is also defined by means of XML elements. Consequentially, we can use the CSS default specifications of HTML together with individual specifications for both HTML and XML tags.

5 Discussion

Our concept of XML structured medical knowledge can be seen as a consequent further development of text-based, especially HTML-based, approaches. HTML is restricted to the formatting of documents and doesn't support the specification of individual tags. The decisive advantage over HTML-based approaches is that we are able to query XML documents both on content and structure. By introducing the semantics inherent in XML into the search process recall and precision of a query are drastically improved resulting in a more efficient access to relevant medical knowledge (for details please consider: http://citeseer.nj.nec.com/385056.html). In order to take advantage of semantically structured XML documents, the query must have semantic structure as well. Schema restrictions, term context limitations, data type information, and other elements of structured queries can be applied. The use of so called proactively structured queries require a certain understanding of both the structure of the data model and the domain by the end user. With this informational background, the end user has the following potential benefits:

(1) In general, documents can be categorized by the schema. The schema of an XML document identifies its structure and its purpose. Results can be made more precise by limiting the search scope to only those documents matching a particular schema, that is, in our case, the schema of a medical textbook. (2) Ambiguous words can be distinguished by the context they appear in. The XML context (surrounding tags) can provide clues to help disambiguate meanings and potentially improve search precision. E.g., if the physician is searching for stage "II" disease, he doesn't want to get all of the documents that contain the term "II", which is frequently used as enumeration of a paragraph. (3) Structural proximity can be used instead of physical proximity to rank results. Information retrieval systems frequently use the proximity of the user's search terms in a document as a ranking method, either implicitly or explicitly (via quoted phrases and the NEAR operator). XML's structure offers more sophisticated proximity measures. The distance between the last word in one element and the first word in the next element is greater than the distance between adjacent words in the same element, even though their

physical proximity in the document is similar. Adjusting the ranking function (implicitly or explicitly) in this manner can improve precision and recall. (4) Portions of documents can be returned. Medical documents are often quite long, and in many cases only a small part of the document may be relevant to the user's query. By making document structure explicit, XML allows information retrieval systems to extract portions of documents. This can improve perceived precision by eliminating the user's need to scan through each result document to look for relevant material. Medical textbooks, guidelines, etc. are living documents and appropriately should be electronically based for ease of updating and access. XML data is as informative as the content the authors put in it. It is obvious that adding markup to medical texts has to be done by clinicians with their specific understanding of the content, context, and logic. Once users are aware of these benefits, they may be willing to invest in creating rich content using XML. At the moment, nobody is going to invest in creating rich XML content if there is no way to take advantage of the annotations, and nobody wants to invest in creating systems that take advantage of the annotations if there is no such data. The conceptual model orientates consequently on the user's information needs who is seeking for problem-relevant information in a textbook of internal medicine. There is a strong need for international development and harmonization of standard models for various medical resources, such as articles, textbooks or clinical practice guidelines with a set of core elements used in each of these resources.

The problems of information access and query management in digital medical resources don't appear to be close to a solution, because they involve the content of complex documents, which is inherently difficult to understand and model. The availability of large amounts of text resources has made this problem still more pressing. The Internet, for instance, has made navigation difficult and finding relevant information can be a challenge. In this case, an excess of information can be equivalent to an absence of information. It is therefore necessary to use tools that aggregate available knowledge and "put this data in order", namely to organize documents into intelligible and easily accessible structures and return answers at various levels of detail to support analytic and decisional activities. Unstructured clinical information is one of the major problems of electronic health care systems. Experts on medical informatics consequently postulate the insertion of more structure and more codes into clinical documents [12] and consider XML the enabling technology (e.g., http://www.gca.org/papers/xmleurope2000/papers/s38-03.html). It remains to be established that the health care community can learn to master the combined organizational and technological challenges of maintaining and managing over time large amount of changing, complex medical knowledge. The developing internet standards are beginning to provide a much more powerful set of platforms and media for exploiting the possibilities of the web as a knowledge medium [13]. XML may replace narrative text as a storage format and allows to structure the data in a stepwise fashion. On the basis of structured data we are able to improve the search quality for clinical information and the presentation which forms a crucial pre-requisite for the use of the growing amount of medical knowledge in the daily routine.

References

[1] Grol R, Grimshaw J. Evidence-based implementation of evidence-based medicine. Jt Comm J Qual Improv 1999; 25(10):503-513.

[2] Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review [see comments]. JAMA 1998; 280(15):1339-1346.

[3] Shiffman RN, Liaw Y, Brandt CA, Corb GJ. Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. J Am Med Inform Assoc 1999; 6(2):104-114.
[4] Hripcsak G. Writing Arden Syntax Medical Logic Modules. Comput Biol Med 1994; 24(5):331-363.
[5] Prokosch HU, Kamm S, Wieczorek D, Dudeck J. Knowledge representation in pharmacology. A possible application area for the Arden Syntax? Proc Annu Symp Comput Appl Med Care 1991;243-247.
[6] Tafazzoli AG, Altmann U, Wachter W, Katz FR, Holzer S, Dudeck J. Integrated knowledge-based functions in a hospital cancer registry-- specific requirements for routine applicability. Proc AMIA Symp 1999;410-414.
[7] Miller RA, Pople HE, Jr., Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982; 307(8):468-476.
[8] Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP system. J Med Syst 1983; 7(2):87-102.
[9] Cimino JJ, Socratous SA, Grewal R. The informatics superhighway: prototyping on the World Wide Web. Proc Annu Symp Comput Appl Med Care 1995;111-115.
[10] Cimino JJ, Socratous SA, Clayton PD. Internet as clinical information system: application development using the World Wide Web. J Am Med Inform Assoc 1995; 2(5):273-284.
[11] Cimino JJ. Intranet technology in hospital information systems. Stud Health Technol Inform 1997; 45:102-109.
[12] Cimino JJ. Data storage and knowledge representation for clinical workstations. Int J Biomed Comput 1994; 34(1-4):185-194.
[13] Gordon C, Gray JA, Toth B, Veloso M. Systems of evidence-based healthcare and personalised health information: some international and national trends. Stud Health Technol Inform 2000; 77:23-28.