Information Retrieval: Proactive Semantic Search Strategies

Simon Hoelzer MD[a, b], Ralf K Schweiger PhD[b], Joachim Dudeck MD[b]

[a] *H+ The Swiss Hospitals, Berne, Switzerland*
[b]*Institute of Medical Informatics, Justus-Liebig-University, Giessen,Germany*

Abstract

Information access and retrieval are essential to serve the delivery and application of evidence-based medicine. The eXtensible Markup Language (XML) provides a standard means to explicitly describe a document's structure and to identify meaningful elements inside textual narrations. We have developed an Information model to represent medical knowledge contained in Clinical Practice Guidelines, textbooks, patient information, articles, etc. that provides a transparent, granular, scalable representation of text-based medical information. Access to these processed and "enriched" electronic resources is achieved through a new concept using an XML search engine. This search engine exploits XML for improving the search quality. Search mechanisms that analyses questions posed by quantitative and semantic parameters are becoming increasingly important. When, and to what extent they will be deployed in this use or similar types of uses, will depend on the efforts being made towards the production, upkeep and updating of structured (XML) documents.

Keywords:
Clinical practice guidelines, quality of care, eXtensible Markup Language, information retrieval

1. Introduction

The increasing sources of medical knowledge need to be made efficiently available as and when required for individual care situations in order to press ahead with the application of evidence-based approaches in medicine. There are several elements that are critical for success in this: the quality of the source and its availability at the point of care should be seen as central among these.

Active decision support bump up against their limits if the clinical decision-making situation or the required knowledge base and framework are too complex [1,2]. In such a situation, however there is the possibility of drawing on selected knowledge items direct from their sources (textbooks, journal articles, guidelines, etc). With regard to representation of those items, efforts are being directed towards a single, uniform and comprehensive data model (XML schema) for structuring and semantic labelling [3].

2. Materials and Methods

Data and information in medicine are mainly represented in slightly structured or even unstructured, narrative text documents. It is nearly impossible to detect and handle relationships between data elements within narrative documents or to retrieve parts of documents that contain specific information. But information access and retrieval are essential to serve the delivery and application of evidence-based medicine. This way, the exploitation of medical information resources by electronic means is still limited [4,5] without an explicit structure. One possible exploitation, for example, is the quick access of the healthcare professional to the information of interest. A physician may not want to read a complete clinical practice guideline (CPG) if he is only interested in a specific part of the guideline. Irrelevant search results can be reduced to a minimum as soon as we insert meaningful structures such as diagnoses and therapies into clinical documents. The eXtensible Markup Language (XML) provides a standard means to explicitly describe a document's structure and to identify meaningful elements inside textual narrations. XML provides powerful concepts to represent the structure in narrative documents that is otherwise immanently hidden. It will play an important role to improve the management of Healthcare documents of any kind [6].

At the same time, additional information contained in the text either implicitly or explicitly, on say metatags or an attribution process, can be placed in the XML document. This includes enhancement with standardised, encoded information (for example MeSH or ICD coding), assigning clinically-relevant text properties (e.g. levels of evidence of individual recommendations) as well as linking external information sources. This document structure defined, inherent information can subsequently be drawn upon to look up case-relevant knowledge.

3. Results

If we want to communicate the meaning of XML documents we must standardize the underlying XML models, i.e. XML modelling and standardization will play a significant role in the future. This task will remain a challenge to standardization bodies. However, XML can support the implementation of a flexible and composite bottom up approach to this process. XML Schema, for example, suggests the construction of content models that can be shared and used as building blocks for creating new schemas. We have developed an Information model to represent medical knowledge contained in Clinical Practice Guidelines, textbooks, patient information, articles, etc. This core model provides a transparent, granular, scalable representation of text-based medical information. With this generic document-based approach it is possible to convert and map an XML file to the original text resource as well as other computable formats.

Concerning the electronic representation of clinical practice guidelines, the model allows for:

- Queries on content and structure by means of typical search items: diagnosis, codes, patient characteristics ... (proactively structured queries)
- Queries using data types: date, codes ...
- Linking data items of an electronic patient record to corresponding information resources
- Provision of metainformation about the text resource as well as about the quality of the recommendations within the text resource

- Customized presentation of the most relevant parts of a set of matching guidelines

4. Discussion

In this way, one can expect improved availability of clinically-relevant knowledge on a specific medical problem to meet the user‘s needs. Access to these processed and "enriched" electronic resources is achieved through a new concept using an XML search engine [7,8]. This search engine that exploits XML for improving search quality (relevance and completeness of URIs). It can handle Internet resources of any type (text/xml, image/jpeg, etc.) and uses XML Topic Maps to establish relationships between search items. Thus, documents in different formats (HTML, PDF, XHTML and XML) and with varying degrees of refinement with regard to their structure as well as semantics can be processed.

In the absence of any structure or semantic markup, as a last resort the search will be carried out similar to a free-text search using the search algorithms of normal search engines, that weight the referent links, the proximity, position and emphases (bold print, headers, metatags) of search terms. However, such purely textual searches severely restrict the automatic identification and extracting of relevant knowledge from within highly-structured sources.

Improvements can be achieved through a combination of the following approaches:

1. To increase the precision of search results, it is possible to define so-called proactively structured queries for individual clinical care situations (scenarios) in advance. In this way, the occurrence of target search terms can be inspected context-sensitively in pre-defined document structures (for example, the occurrence of a particular stage of a disease in the "Therapy" paragraph, the meaning of which has been defined in the XML data model). Thus, when selecting qualifying sources, it is ensured that only the really relevant sections of the complete document are inspected, and that the search term has the correct meaning assigned to it (examples: a] Stage of illness "III" will not be confused with paragraph number III; b] a document will only be included on the list of results if the contents thereof refer only to the primary diagnosis "Apoplexy", and not to apoplexy as a complication of medical intervention or the underlying illness). The marking of a section of text with a description of its function within the document results in a machine-readable semantic plan, which should ensure this document will also appear in the correct lists of results. At the next step, the scenarios and corresponding proactively-structured queries already defined will determine which sections of text are to be displayed, and in what layout. To do this, standard XML transformation and rendering tools (XSL, CSS) are deployed.
2. Furthermore, the search engine is not just able to associate different search terms by their frequency intervals in the text, but can also make use of both the structure (terms are found in the same document paragraph or are found inside a hierarchy of documents) as well as the standardised semantics described in the model above (the meaning of a tag‘s contents). In this way, the closeness of a relationship is interpreted using the document‘s structure or semantics, and weighted appropriately. In general, search terms can be input actively. However, the above approaches can be combined with the automated production of sets of search terms, or filling in a proactively structured query with search terms extracted from the electronic medical record. This process is currently supported when processing diagnostic terms by using synonymous terms taken from the ICD-10 diagnostic

thesaurus which, by contrast, can be compiled and updated in XML using the concept of topic maps [8]. It is possible here to call up an alternative term or the appropriate ICD-10 code for the search.

5. Conclusion

Search mechanisms that analyse questions posed by quantitative and semantic parameters are becoming increasingly important. When, and to what extent they will be deployed in this use or similar types of uses, will depend on the efforts being made towards standardisation, also future work on the production, upkeep and updating of documents. The efforts being made by W3C consortium, such as adopting the XHTML 2.0 standard, with a clear separation of structure and presentation, along with setting hierarchies, chapters and sub-chapters, could acquire a mediating role in the medium-term, and help drive forward the move towards highly-structured sources and their use. In the future, experience in practical use at various levels must be collated and drawn upon. By doing this, it become apparent to what degree structure and semantics are essential and can be usefully put to work.

6. Acknowledgments

This project is supported by the German Federal Ministry of Health.

7. References

[1] Bakken S. An Informatics Infrastructure is essential for evidence-based practice. J Am Med Inform Assoc 2001; 8(3):199-201.

[2] 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.

[3] Hoelzer S, Schweiger R, Dudeck J. Representation of practice guidelines with XML--modeling with XML schema. Methods Inf Med. 2002;41(4):305-12

[4] Roberts A. Analysing XML health records. XML Europe Conference Proceedings 2000, 377-380.

[5] Schweiger R, Tafazzoli A, Dudeck J. Using XML for flexible data entry in healthcare, example use for pathology. XML Europe Conference Proceedings 2000, 357-362.

[6] Schweiger R, Hölzer S, Altmann U, Rieger J, Dudeck J Plug and Play XML? A Healthcare Perspective J Am Med Inform Assoc. 2002 Jan;9(1):37-48.

[7] Intelligenter suchen als bisher mit LuMriX - Forschergruppe der JLU entwickelt neuartige Suchmaschine, Giessener Anzeiger vom 31.07.2002

[8] Simon Hoelzer, R.K. Schweiger, R. Liu, D. Rudolf, J. Rieger, J. Dudeck XML Representation of Hierarchical Classification Systems: From Conceptual Models to Real Applications Proc AMIA Symp 2002

Address for correspondence

Simon Hoelzer, MD
H+ The Swiss Hospitals, Lorrainestr.4A, CH-3000 Berne, Switzerland
e-mail: simon.hoelzer@hplus.ch
web: www.hplus.ch/main/Show$Id=481.html