MYCIN II: Design and Implementation of a Therapy Reference with Complex Content-Based Indexing

David K. Kim MS
Mayo Medical School, Rochester, Minnesota; kim.david@mayo.edu
Lawrence M. Fagan MD, PhD
Kathleen T. Jones
Daniel C. Berrios MD, MPH
Stanford Medical Informatics, Department of Medicine, Stanford University School of Medicine, Stanford, California; ffagan, kjones, berriosj@smi.stanford.edu
Victor L. Yu MD
Division of Infectious Disease, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania; vly+@pitt.edu

ABSTRACT

We describe the construction of MYCIN II, a prototype system that provides for content-based markup and search of a forthcoming clinical therapeutics textbook, Antimicrobial Therapy and Vaccines. Existing commercial search technology for digital references utilizes generic tools such as textword-based searches with geographical or statistical refinements. We suggest that the drawbacks of such systems significantly restrict their use in everyday clinical practice. This is in spite of the fact that there is a great need for the information contained within these same references. The system we describe is intended to supplement keyword searching so that certain important questions can be asked easily and can be answered reliably (in terms of precision and recall). Our method attacks this problem in a restricted domain of knowledge—clinical infectious disease. For example, we would like to be able to answer the class of questions exemplified by the following query: "What antimicrobial agents can be used to treat endocarditis caused by Eikenella corrodens?"

We have compiled and analyzed a list of such questions to develop a concept-based markup scheme. This scheme was then applied within an HTML markup to electronically "highlight" passages from three textbook chapters. We constructed a functioning web-based search interface. Our system also provides semi-automated querying of PubMed using our concept markup and the user's actions as a guide.

KEYWORDS: Information retrieval; electronic textbooks; drug therapy, computer assisted; hypermedia; abstracting and indexing; content-based markup (non-MeSH).

INTRODUCTION AND BACKGROUND

Physicians in clinical practice have an ongoing need for access to diverse sorts of information. They also face significant barriers in obtaining that information. Covell et al.1 found that a significant number (70%) of questions arising during an outpatient visit to an internist's office went unanswered. More than half of these questions would be potentially answerable with consultation of reliable, quick, up-to-date tertiary1 references. Notable problems include: Lack of adequate organization or indexing, concern that accessible textbooks were out of date, and lack of print sources that would answer the questions they asked.

What kinds of questions are these? In the Covell study, requests were dominated by practical management and therapeutic issues; approximately 70% were regarding treatment, diagnosis, and drug therapy. Other studies in a variety of practice settings2 3 4 support this claim.

There is a need for current, well-indexed, clinical diagnostic and therapeutic references. Many authoritative textbooks are available in CD-ROM form. These use graphical user interfaces with hierarchical navigation as well as textword search facilities. To use such tools, one must use keywords to "triangulate" passages of interest (for example, "eikenella", "endocarditis", and "therapy -OR- treatment"). It is our anecdotal experience that such tools are rarely used in practice, despite their intuitive promise. Existing CD-ROMs do not help enough to make it worth finding and

† In this context, tertiary refers to texts and edited compilations of review material.
using them in a clinical setting. Possibly contributing to this inadequacy are problems inherent in textword searches; even those relying on statistical weighting or geographical strategies suffer from several difficulties. These include a trade-off between precision and recall as well as matches lost to synonymy and pronominal references. Reliability of search results also depends upon searching skill, which undermines the benefit of using an authoritative reference such as a textbook. We have attempted to remove a fundamental contribution to these problems by pre-assigning semantic meanings to sections of text.

Two of the authors previously helped construct MYCIN, an expert system to assist in antimicrobial therapy choice. Experience with this formal knowledge base thus influenced the resulting hybrid approach, which attempts to impose some of the ontological ("definitional") structure of an expert system on a nearly free-text data source. Bernstein, et al., arguably pioneered the concept of computer-assisted compilation of a structured-text knowledge base in creation of the Hepatitis Knowledge Base. In our case, there is a constraint on the source text in terms of content--Editors collaborated with authors to produce a fixed topical outline for chapters to help ensure that answers to those questions deemed important would be present somewhere in the text. Authors were required to include specific recommendations in several areas (e.g. antimicrobial therapy). Bottom-line recommendations were stressed; in this sense the chapters are motivated by the idea of a hospital "consult" where a specialist is called in to render the best opinion possible for given data. We feel that such specialist opinion, even in canned form, will be useful as long as it is authoritative; the Covell study noted among internists that 69% of the questions they posed fell into a subspecialty outside their own.

In our system, ontological structure takes the form of a semantic markup. This markup indicates occurrence not only of certain facts or terms in the text, but also of the context in which they occur. The application of question and context models to assisted searching of the primary medical literature has been explored in depth by Purcell. MYCIN II builds on Purcell's work, notably in its starting assumption of coverage of a domain of knowledge by a limited number of question types.

The design of our system provides for relatively simple linking to external databases. Specifically, the system constructs and submits searches to Medline. This has been implemented previously by Cimino et al. A link to primary literature serves as an important supplement when considering the problem of potentially out-of-date textbook information. Providing abstracts in this way serves to provide "just-in-time information" at a minimal extra cost in the user's time.

**DESIGN CONSIDERATIONS**

Our facility is designed to efficiently steer a user to relevant passages of a text when asked one of a predetermined class of questions. We also wish to provide facile linking from these search results to external resources. By first building a model of what the user is interested in, we extend the text with a structured description of its contents. We use this description in the matching process in order to reduce the drawbacks mentioned above.

To this end, we first require adequate coverage of concepts by our list of questions. These questions do not, by design, cover all topics that are implied by the set of textwords in the document. The process by which these "canonical" classes of questions and the resulting markup are created must be such that a diligent polling of experts will yield a representative basis. *I.e., we must elicit those concepts whose identification in the text will allow the handling of the large majority of eventual questions.* In this way, we tacitly generate a minimal ontology for the domain at hand.

**System Description**

* Precision = (# of relevant hits)/(total # of hits) This corresponds intuitively with the clinical concept of specificity. Recall = (# of relevant hits)/(# of relevant hits in overall database) This corresponds intuitively with the clinical concept of sensitivity.

1 Gr. *Onto*– having to do with existence. An ontology, in our restricted sense, is a conceptualization or enumeration of terms, contexts, or properties which we wish to label and with which we wish to reason.
Figure 1 contains a block diagram of the overall system. The marked text takes the form of HTML with semantic markup applied within comment tags. The search engine and front-end interface were implemented in Perl using the Common Gateway Interface.

As discussed, authors and potential users of the book *Antimicrobial Therapy and Vaccines* were asked what sort of questions the system ought to answer. This list was analyzed to group like questions together into classes, for example: "*What antimicrobial therapy can be used for ORGANISM X in INFECTION TYPE Y?*" Passages that answer this question need to specify the drug to be used (DRUG="ceftriaxone") and must refer to treatment (INTENT=TREAT). These essential terms and contexts (e.g., DRUG and INTENT=TREAT) were extracted by inspection and used to define our semantic markup. We refer to these definable concepts as tag parameters below. Three chapters, two dealing with bacteria and one dealing with an antimicrobial agent, were marked by one of the editors for the initial version of this system.

Figure 2 shows a query screen from our current system and the parameter/value pairs it explicitly and implicitly generates. These will be used to search through the markup to identify matching text passages.

![User's Query Screen](image)

An example of a marked section of text is shown in Figure 3. It is taken from a section on *Eikenella corrodens* endocarditis. Answering passages are flanked by begin <!-- MARK -->...--> and end <!-- /MARK --> markers. The markup for both passages contains the tag values corresponding to the initial query, as well as the same specific drug names (DRUG="ceftriaxone", etc.) discussed. They differ in efficacy and order of recommendation—This regimen is considered best (EFF="clinbest") for penicillin allergic patients (CONdition="allergy to penicillin") but is first line (EFF="clinfirst") overall.

![Marked Text For Example](image)

Figure 3: Marked Text For Example

Figure 4 shows the results screen that the user sees after the local textbook search is executed. It is organized in a master-detail fashion; selecting **See Excerpt** for any match ("hit") shown on the left will display that portion of the document in the right-hand frame.

![Results Page](image)

Notice also that each match provides its own Medline query option. Clicking on this brings up a guided query page (Figure 5), containing options for query terms taken from the markup surrounding the passage of interest.

![Guided Medline Query](image)

Figure 5: Guided Query Screen
Options are pre-selected based upon the user's initial query. Upon submission, the user-selected query terms are mapped to search phrases for the target system, PubMed.

The system described provides the following facilities:
1. **Searching based on context and meaning.** The context is specified as a tag parameter and thus is not dependent upon document structure. To illustrate, a particular treatment recommendation could be mentioned in the Contraindications section rather than Therapy (e.g., tetracycline for a penicillin-allergic patient). The meaning of a word or phrase may be obscured from simple matching by variant vocabulary or pronominal reference, e.g. "clavulanic acid" or "it" in a sentence referring to "clavulanate".

2. **Facile linking to external, more current information resources.** This provides a "bootstrapping" function between from our system to PubMed without the user's needing to define a keyword search strategy.

**DISCUSSION**

Production of Markup. We began with three chapters and a large number of questions. Our anecdotal experience was that the markup schema required fewer changes as more questions were added. We believe that this rapid convergence is suggestive both of coverage and consistency.

To date, the process of applying the markup schema has not been prohibitive, requiring approximately 1-2 hours per chapter for an infectious disease expert. Related work is underway at Stanford to produce a tool to assist in this markup process, especially for sentence-by-sentence markup. The current version assists mechanically in the process, but the ultimate goal is to provide a partially-automated process that provides suggestions from the text itself via an information extraction system.

General Issues Regarding The Markup. In the present system, markup has taken the form of inline, human-readable, HTML-like annotation. Such markup would be conceptually equivalent to an external indexfile with tokenized references to vocabulary terms (e.g. "org269" instead of "Eikenella corrodens") and pointers "into" the source text. If a standardized inline markup (say, XML/SGML) were used, the full inline form might be retained through publication to allow ad hoc information processing. It would be directly usable and inspectable by the end user (perhaps in an SGML-aware browser) instead of serving solely as a vehicle for answering those question-types identified in advance. Such an exposure of the underlying content-based index could more easily allow interoperability with local systems, although this is not an absolute requirement (token translation tables could be used). Examples include:

- Use of patient descriptors from a computer-based patient record as query specifiers.
- Use of markup to link to local resources (pharmacy cost data, locally licensed clinical references, order entry gateways, etc.)

In such a case, the scope of the markup might have to be more carefully defined, thought-through, and validated against possible uses.

A second issue involves use of the representation choice as a conceptual model. Conceiving of the sets of markup parameters as statements or assertions, rather than as overlapping regions of text attributes, has been helpful in designing the markup schema and semantics. Potential ambiguities or implied contradictions were most easily recognized when reflexively asking the question, "What statements are we able to generate from this markup 'vector'?

It is important to identify such problems in advance. For example, we noted that separately-interpretable or "alternate" tagging of the same text is sometimes needed. In the inline-markup model, this corresponds to nested tagging without inheritance. In these cases the same text needs to be marked up in two (non-combinable) contexts. We do not explore here the many other issues of syntax and semantics encountered during this work.

Physical Density of Markup. The question naturally arises, "Should we mark every sentence, or even every sentence that answers a given question?" Instead, one could mark only occasional, exceptional, very important, or non-obviously placed passages. In MYCIN II, we have chosen consistent question-driven coverage--If a fact was specifically elicited from the authors, and/or falls into our list of canonical questions, all answering passages are marked. Physical granularity was usually at the sentence or paragraph level, but occasionally phrases were separately marked.

Breadth of Conceptual Coverage. We have attempted to construct a markup that anticipates all reasonable questions in our given field, not just all specific questions we have gathered (our minimal goal). For this reason, any mention of antimicrobial therapy (whether of a named or unnamed organism, whether it is a vague guideline or a specific dose recommendation) is marked INTENT=TRE and AGENT_TYPE=AMA (Antimicrobial Agent). This anticipates extensions and revisions of existing questions.

External Search Benefits of Markup. To pose a query to a journal database such as Medline, a user generally

II by "vector" we mean a tuple of parameter name/value pairs.
must enter or select search terms. Some systems (e.g. PubMed) keep a record of document descriptors, either from prior search requests or tacitly derived from returned documents. This allows for “bootstrapping” of new searches via such offerings as “see Related articles”, “Smart search”, etc. As shown above, such internal information is used by MYCIN II when bridging to a PubMed search (Figure 5). This makes use of the original information the user has entered as well as his or her response to subsequent prompts. These prompts are themselves based on search results deemed to be of interest. In this “navigational” picture, an interaction between the question posed and the markup underlying a selected example record creates search parameters for the external search. It should be mentioned that this was not an initially anticipated use of the markup but was readily adopted! We expect increased precision and recall also in the external (PubMed) linking and that this will be most noticeable in the cases of quick searches and less-experienced users.

SUMMARY

We have constructed a system with the following features: A fixed set of question types, an extensible embedded markup which provides content-based indexing, matching based on context and meaning, and the ability to answer questions and link to external information resources with the query structure intact.

Acknowledgments

We acknowledge Edward Shortliffe for developing the original MYCIN (and allowing us to use the name MYCIN II) and for helping create the laboratory environment in which both projects were accomplished. David Kim thanks Stanford Medical Informatics for graciously hosting him during this work. This work was supported in part by contract N44-C061025-001 from the National Cancer Institute to Lexical Technology Inc.

References


