In Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX 2005), August 2005, Salvador, Brazil

Representation of Information Needs and the Elements of Context: A Case Study in the Domain of Clinical Medicine

Jimmy Lin[1,3] and Dina Demner-Fushman[2,3]
[1]College of Information Studies
[2]Department of Computer Science
[3]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
jimmylin@umd.edu, demner@cs.umd.edu

1. INTRODUCTION

Information seekers today have an ingrained habit of issuing extremely short queries to retrieval systems. Although the dominance of Web search contributes to reinforcement of this behavior, the inability for retrieval systems to take advantage of more expressive query representations is also a culprit. Short queries pose a challenge to designers of information retrieval systems for a number of reasons: At the linguistic level, they often contain polysemous words that are difficult to disambiguate without appropriate contextual cues. At a higher level, short queries do not provide clues regarding the broader activities that give rise to the user's information need.

There is reason to believe that if a computer system were able to solicit richer queries from a user, it might be able to build a better model of the information seeking process, thereby leading to higher retrieval performance—for example, Hearst [8] has shown that faceted queries can be converted into simple post-filtering constraints to boost precision. Note, however, that "richer" queries do not necessarily mean natural language descriptions—in fact, language processing technology is not yet sufficiently advanced to capture the intricacies of free text for document retrieval. Instead, we believe that a more fruitful approach is to focus on structured representations of information needs that better capture the different aspects of the information seeking process.

The idea that "bags of words" are poor query representations is by no means new. Even before the invention of computers, librarians have long known that retrieval should be performed at the conceptual level. More recently, Belkin's work on anomalous states of knowledge [3] brought this to the attention of information retrieval researchers. The idea that IR systems serve to bring different cognitive representations "into alignment" has also been explored within the framework of cognitive information retrieval [9].

We recognize the difficulty in designing query representations for general purpose retrieval. It is unlikely that any single representation of information needs will be both sufficiently general to capture a wide variety of scenarios and sufficiently detailed so that it can be operationalized within a computer system. Faced with this challenge, the first step might be to explore these issues in a more restricted domain.

The domain of clinical medicine is very well-suited for experiments in building richer models of the information seeking process. Physicians' information needs are well-documented [4, 7, 6], particularly at the point of care. The real need for decision-support systems that deliver information in a timely fashion translates into significant potential impact in improving the quality of health care.

This paper presents a case study on how existing resources and knowledge in the clinical domain can be integrated into a rich computational model of the information seeking process. We present five elements—work task model, search task model, process model, problem structure, and domain model—that underlie a richer representation of clinicians' information needs. Although our work is situated within a specific domain, we believe that components of our approach can be generalized to other types of retrieval systems.

2. THE ELEMENTS OF CONTEXT

We identify five elements of context that serve to circumscribe the information seeking process, which we illustrate in the domain of clinical medicine.

2.1 Work Task Model

The work task model specifies the broader activities of a user; cf. [9]. In our particular domain, evidence-based medicine (EBM) [13] is a widely-accepted paradigm for medical practice that stresses the importance of evidence from patient-centered clinical research in the health care process. It defines a process for systematically reviewing, appraising, and applying findings from clinical research, suitably integrated with the physician's own expertise, to aid in the delivery of patient care. EBM identifies four general activities that physicians engage in:

- **Therapy:** Selecting effective treatments to offer patients, taking into account other factors such as risk and cost.
- **Diagnosis:** Selecting and interpreting diagnostic tests, while considering their precision, accuracy, acceptability, cost, and safety.

- **Etiology:** Identifying the causes for diseases, including iatrogenesis.

- **Prognosis:** Estimating the patient's likely course with time and anticipate likely complications.

It is important to note that these clinical tasks do not necessarily implicate the existence of information needs.

2.2 Search Task Model

When questions do arise in the process of practicing medicine, a search task is triggered. Information that is gathered in the search task ultimately informs the broader activities as defined in the work task.

There is, however, a tight coupling between work and search tasks. For therapy, the search task is usually *therapy selection*, e.g., determining which course of action is the best treatment for a disease. For diagnosis, there are two different scenarios: in *differential diagnosis*, a physician is considering multiple hypotheses regarding what disease a patient has; in *diagnostic methods selection*, the clinician is attempting to ascertain the relative diagnostic utility of different tests. For etiology, *cause determination* is the search task, and for prognosis, *patient outcome prediction*.

An explicit search task model allows a system to perform finer-grained retrieval at the sub-document level (essentially, question answering). For example, a physician with an information need related to therapy would be most interested in clinical findings about the effectiveness of various treatments. This might suggest the importance of identifying "outcome" statements, e.g., "Ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses." This, in turn, might suggest the importance of understanding comparative statements.

2.3 Process Model

Summarizing our models thus far, information needs arising in the work task trigger search tasks that lead to gathering of information. The process model specifies how this information is then integrated back into the original work task. In the domain of medicine, the process model defines how doctors act on clinical evidence pertaining to the patient problem at hand.

Within the framework of EBM, there exists several tools for appraising the strength of evidence presented in a study, i.e., how much confidence should a physician have in the results? As an example, the Strength of Recommendations Taxonomy (SORT) provides a basis for determining the potential upper bound on the quality of evidence:

- A-level evidence is based on consistent, good quality patient outcome-oriented evidence presented in systematic reviews, randomized controlled clinical trials, cohort studies, and meta-analysis.

- B-level evidence is inconsistent, limited quality patient oriented evidence on the same types of studies

- C-level evidence is based on disease-oriented evidence or studies less rigorous than randomized controlled clinical trials, cohort studies, systematic reviews and meta-analysis.

This process model has implications for information retrieval research: it suggests that the primary role of systems is to support the decision-making process. Indeed, this was a mistake made by artificial intelligence researchers in the 80's, and one of the major reasons that medical expert systems from that era did not gain widespread acceptance—doctors simply didn't want to be told what to do. This process model also suggests that topicality is only one of many factors that determine the usefulness of a particular result; other issues such as publication type, strength of evidence, and recency all influence relevance judgments within the context of a clinical situation.

2.4 Problem Structure

Physicians are trained to think about clinical questions in a particular way—this is the problem structure. Within the practice of evidence-based medicine, a well-formed clinical question is identified to have four key components [11]:

- **Problem** What is the primary problem or disease? What are the characteristics of the patient (e.g., age, gender, or co-existing conditions)?

- **Intervention** What is the main intervention (e.g., a diagnostic test, medication, or therapeutic procedure)?

- **Comparison** What is the main intervention compared to (e.g., no intervention, another drug, another therapeutic procedure, or a placebo)?

- **Outcome** What is the effect of the intervention? Were the patient's symptoms relieved or eliminated? Side effects reduced? Cost reduced? etc.

These elements are often referenced with the mnemonic PICO. As we will describe later, this structure can serve as the core of a richer representation of information needs.

2.5 Domain Model

Under the guidelines of evidence-based medicine, a physician is expected to integrate clinical evidence gained from a review of the literature with his or her knowledge and experience accumulated through practice. Naturally, this background knowledge will play an important role in the information seeking process. In order for computer systems to effectively complement and support physicians, they must have access to structures that explicitly encode ontological relations between conceptual entities within the domain.

Concepts in the medical domain have already been codified in the Unified Medical Language System® (UMLS) [10]. The 2004 version of the UMLS Metathesaurus contains information about over 1 million biomedical concepts and 5 million concept names from more than 100 controlled vocabularies. In addition, software for leveraging this ontology already exists: MetaMap [1] identifies medical concepts in free text, while SemRep [12] extracts relations between the recognized concepts.

In terms of specific applications for retrieval systems, domain knowledge can be employed to implement concept-based search and alleviate the vocabulary mismatch problem. Techniques such as query expansion, document expansion, or indexing of normalized forms may assist physicians in retrieving documents that share few or no terms with the query; see, for example, [2]. Furthermore, the ability to recognize concepts in documents serves as the foundation for finer-grained retrieval and document classification techniques discussed earlier.

3. INFORMATION NEEDS

The elements of context described in the previous section enumerate the types of knowledge that are important to building a more refined model of the information seeking process. In our view, they exist to support representations for capturing user information needs and subsequent processing of it by computer systems. For example, consider the following clinical question:

> In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever?

Taking into account both the problem structure and the search task, a representation of the question might be:

> **Search task:** therapy selection
> **Problem:** acute febrile illness in children
> **Intervention:** acetaminophen
> **Comparison:** ibuprofen
> **Outcome:** reducing fever

In our view, retrieval can be viewed as the process of "semantic unification" between such query structures and corresponding structures derived from text. This is simply the age-old idea of "retrieval at the level of conceptual structures". The contribution here, however, is that within restricted domains, it becomes possible to define exactly what these conceptual structures are.

To the above question, a computer system might supply the following response:

> Ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses.

Naturally, the physician would need to further assess the source, i.e., its strength of evidence, date of publication, etc., in order to arrive at the final decision about patient treatment (via the process model). In this hypothetical retrieval process, the domain model serves as the substrate that facilitates conceptual matching between the query representation and texts containing potentially relevant answers (citations from the MEDLINE database, for example).

We are currently building clinical question answering systems that realize the above scenario. Component technologies that underlie such a system are either already in place or under development. Preliminary results in fine-grained knowledge extraction, i.e., automatically identifying PICO elements in MEDLINE abstracts, are encouraging [5].

4. CONCLUSION

This paper identifies five elements of context that circumscribe the information seeking process: the work task model, the search task model, the process model, the problem structure, and the domain model. These components come together to inform a rich query representation that captures a user's information needs. Although we have only discussed applications in the domain of clinical medicine, we believe that many of these ideas can be generalized to other domains.

5. ACKNOWLEDGMENTS

We would like to thank Eileen Abels, Dagobert Soergel, and an anonymous reviewer for their helpful comments on earlier drafts of this paper. We would also like to thank Kiri for her kind support.

6. REFERENCES

[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceeding of the American Medical Informatics Association Annual Symposium*, pages 17–21, 2001.

[2] A. R. Aronson, T. C. Rindflesch, and A. C. Browne. Exploiting a large thesaurus for information retrieval. In *Proceedings of RIAO 1994: Intelligent Multimedia Information Retrieval Systems and Management*, 1994.

[3] N. J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5:133–143, 1980.

[4] D. G. Covell, G. C. Uman, and P. R. Manning. Information needs in office practice: Are they being met? *Annals of Internal Medicine*, 103(4):596–599, October 1985.

[5] D. Demner-Fushman and J. Lin. Knowledge extraction for clinical question answering: Preliminary results. In *Proceedings of the AAAI 2005 Workshop on Question Answering in Restricted Domains*, 2005.

[6] J. W. Ely, J. A. Osheroff, M. L. Chambliss, M. H. Ebell, and M. E. Rosenbaum. Answering physicians' clinical questions: Obstacles and potential solutions. *Journal of the American Medical Informatics Association*, 12(2):217–224, March-April 2005.

[7] P. N. Gorman, J. S. Ash, and L. W. Wykoff. Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of the Medical Library Association*, 82(2):140–146, April 1994.

[8] M. A. Hearst. Improving full-text precision on short queries using simple constraints. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR 1996)*, 1996.

[9] P. Ingwersen. Cognitive information retrieval. *Annual Review of Information Science and Technology*, 34:3–52, 1999.

[10] D. A. Lindberg, B. L. Humphreys, and A. T. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291, August 1993.

[11] W. S. Richardson, M. C. Wilson, J. Nishikawa, and R. S. Hayward. The well-built clinical question: A key to evidence-based decisions. *American College of Physicians Journal Club*, 123(3):A12–A13, November-December 1995.

[12] T. C. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, December 2003.

[13] D. L. Sackett, S. E. Strauss, W. S. Richardson, W. Rosenberg, and R. B. Haynes. *Evidence-Based Medicine: How to Practice and Teach EBM.* Churchill Livingstone, second edition, 2000.