MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Building a Knowledge Base to Support a Digital Library

Eneida A. Mendonça, James J. Cimino

Department of Medical Informatics, Columbia University, New York, NY, USA

Abstract

As part of an effort to support searching of online medical literature according to individual needs, we have studied the possibility of using the co-occurrence of MeSH terms in MEDLINE citations to automate construction of a knowledge base of interrelated concepts. This study evaluates the relevance of the relationships between the semantic pairs generated by the extraction algorithm, and the clinical validity of the semantic types involved in the process. From the semantic pairs proposed by our method, a group of clinicians judged sixty percent to be relevant. The remaining forty percent were considered unimportant by clinicians. We believe our knowledge extraction method is appropriate for the task of retrieving information from the medical record in order to guide users during a search and retrieval process. Future directions include the validation of the knowledge, based on an evaluation of system performance.

Keywords:

Knowledge representation, knowledge acquisition, digital libraries, artificial intelligence

Introduction

The tremendous increase of medical knowledge resources in electronic form, particularly on the World Wide Web, has generated a great deal of interest. The increased availability of information does not make it easy for clinicians to filter large amounts of information and incorporate evidence into their clinical practice. Although the number of clinicians and medical students who routinely perform their own searches has increased, they still find difficult to keep up-to-date with the advances in medical science.[1,2] Information needs exist[3-5], but only a small portion are currently met.[6] Studies have also shown that information can help clinicians make better decisions in different clinical situations.[7] In summary, the current environment presents many resources, but raises questions about the quality of the information, information overload and access.

Digital libraries have been described by many as a "new way of carrying out library functions" [8], encompassing new types of information resources, approaches to acquisition, methods of storage and preservation, approaches to classification and cataloging, modes of interaction, and more reliance on electronic systems and networks. Digital libraries should provide more than just access to different resources. In the health care environment, clinicians require assistance in converting information needs into focused questions, in selecting the appropriate resource, retrieving the materials, critically appraising the information retrieved, and, finally, refining the strategy if necessary. This process is analogous to the search process recommended for the practice of evidence-based medicine (EBM).[9] EBM requires the ability to access, summarize, and apply information from the literature to day-to-day clinical problems.[10] The effective use of technology can be an important facilitator of the quality, and utility, in accessing and reviewing medical information on the Internet.[11]

Researchers have suggested that an informatics infrastructure is essential for evidence-based practice.[12,13] The integration of information with clinical applications is an opportunity and a challenge. It may facilitate the access to scientific evidence, clinical guidelines, and other decision tools, in a way that information retrieved from these sources is personalized based on the context of individual needs.[2,14] We believe that the development of personalized access to a distributed digital library can facilitate this process. One challenge in building such a system is the construction of a medical knowledge base to support the search of online medical literature according to individual needs. Such a task can be arduous; in part due to the extensive reviews of medical literature required.[15,16]

Previous research studies have introduced approaches to facilitate knowledge extraction from MEDLINE[16] and the UMLS.[17] The approach we propose is an automated knowledge extraction method from MEDLINE citations, based on the ideas introduced by Zeng and Cimino,[17] using with the search strategies by Haynes et al.[18] This approach involves the use of hierarchical and semantic links in the Medical Entities Dictionary (MED)[19] to identify additional terms which can be used to build specific patient-oriented queries. The extraction from MEDLINE citations is intended to enhance the knowledge contained in the MED by identifying additional relationships between concepts.

The value of these relationships is best understood through a simple clinical scenario. A patient comes to the hospital

complaining of painful tightness that originates in the midchest and radiates to the left arm. He reports the pain is relieved by nitroglycerin, which he used once. In the past few hours the severity and frequency of pain has increased, and is not relieved by rest. The patient is admitted to the hospital with a diagnosis of unstable angina. Past history revealed an uncomplicated myocardial infarction a few months ago. A medical student assigned to the case is reviewing the patient's prescriptions in the clinical information system and realizes she does not know much about the use of aspirin in such patients. The medical student selects the drug, and the system presents a list of questions related to aspirin. She then selects the question: "How effective is aspirin in preventing myocardial infarction?" The system was able to present such a question because of information known about the patient's conditions. The system interpreted the questions and used relationships in its knowledge base to retrieve a list of diagnosis, including diabetes mellitus, which was found in the patient's record.

The identification of relevant terms to build patient-oriented queries consists of the relationships between concepts and the logic for drawing valid connections. For instance, to connect myocardial infarction with CK-MB measurement (a laboratory test) requires knowledge of the relationships between myocardial infarction (MI) and heart diseases, and of relationships among intravascular CK test, creatine kinase, cardiac enzymes and heart disease. Figure 1 shows how a concept could be linked to related clinical data.

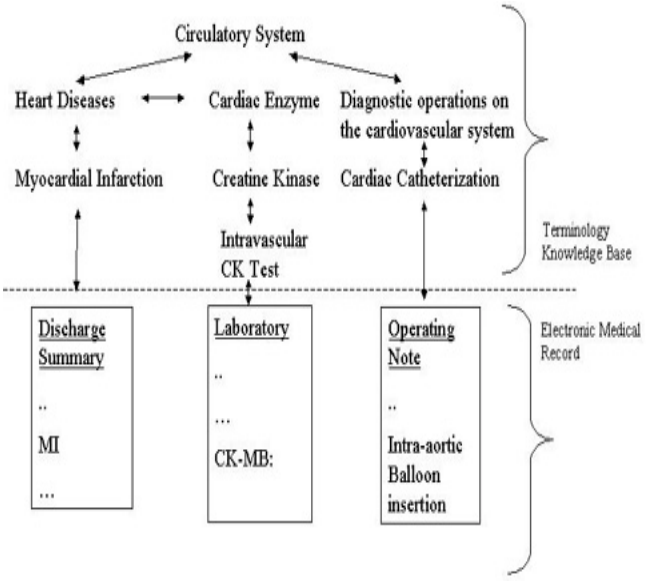

Figure 1. Linking medical concepts to patient data.

This paper describes the methodology used to build such a knowledge base to support the extraction of information from the medical record. The central idea of the approach we describe is to study the possibility of using the co-occurrence of MeSH terms in MEDLINE citations associated with the search strategies optimal for evidence-based medicine.[20]

Background

In a previous study, we described in detail the methodology used to build a table of MeSH terms that co-occur within MEDLINE citations.[20] We also described the methodology used to identify potential relationships that can be used in the construction of a knowledge base to support the extraction of concepts from the medical record. Briefly, this methodology consisted of:

1. The construction of a co-occurrence table of MeSH terms from MEDLINE citations using the four clinical query categories (therapy, diagnosis, etiology, and prognosis) with emphasis on specificity, and the creation a co-occurrence table of semantic types based on the MeSH pairs.

2. A statistical analysis to identify the *statistically* relevant pairs (those that occur more often than by chance) in each group.

3. A pilot study to evaluate the *clinical* validity of the information retrieved.

The statistical analysis involved a chi-square test with Yates correction for each pair generated. A Bonferroni correction was used to define the statistical level of significance because of the multiple testing hypotheses. A phi-coefficient was calculated for each pair. In the pilot study, a questionnaire was designed, which was completed by five physicians. Each questionnaire contained 40 pairs of semantic relationships (10 for each clinical category) and examples of MeSH heading pairs that matched to the semantic pair in question. An analysis of the validity of the information showed the results were suitable for the intended purpose, especially in the therapy group.

Methods

The current study focused on two questions: a) Did the relevant semantic pairs generated by this method capture relevant relationships between terms? b) Are the semantic types extracted by this method clinically relevant?

This study used the information extracted from MEDLINE citations collected in the previous study. It focused on the pairs identified as relevant for the therapy task. The therapy category was chosen because it showed the best performance in the previous study.

A similar questionnaire was designed, which was completed by 3 physicians. The questionnaire contained 144 pairs of semantic types (all semantic pairs found relevant in the first study with $p < 0.01$) and examples of MeSH heading pairs that matched to the semantic pairs in question. The examples were randomly selected from the list of pairs

generated. A brief explanation of the project was given to the physicians and they were asked whether the selected pairs were relevant to the specific clinical task. (See Figure 2)

For each pair, we thus have three different relevant scores based on the physicians' answers. From these scores, we assigned a relevance level to each pair. The pairs were than divided in two groups: relevant and non-relevant (based on the majority of physicians). A manual analysis was performed to determine the relevance of the semantic pairs and the reasons why particular semantic pairs were considered non-relevant.

If the patient has a **Diagnostic Procedure,**
Would you be interested in articles about related **Organ or Tissue Function?**
For example:

Heart Function Tests	Hemodynamics
Coronary Angiography	Coronary Circulation

[ ] Yes [ ] No

Figure 2. Example of a question

Results

For the clinical query category of therapy, we retrieved the most recent 1,000 citations from MEDLINE with subject area was "cardiovascular disease". The automated process generated 135,667 MeSH pairs. The generation of all possible semantic pairs based on the MeSH pairs increased the number of pairs to 195,096 pairs. The statistical analyses performed identified 2,559 unique semantic pairs, 92 unique semantic types and 144 (5.63%) pairs that occur with significant frequency. ($p < 0.01$, Bonferroni correction)

The analysis of the questionnaires showed that physicians identified 87 (60.42%) semantic pairs as relevant and 57 (39.58%) as non-relevant. The most frequent semantic types in those pairs are easily found in electronic medical records and were related to the specified task (literature review). Among them were "Disease or Syndrome", "Mental or Behavioral Dysfunction", "Diagnostic Procedure", "Pathologic Function", "Therapeutic or Preventive Procedure", and "Neoplastic Process". Figure 3 shows some examples of relevant pairs generated.

An analysis of each non-relevant pair was then performed in order to explore the semantic types in each of them. We found 20 (21.7%) semantic types that, when paired, were always deemed to be non-relevant. We also found nine (9.8%) semantic types that were judged relevant only when associated with another particular semantic type. For example, pairs containing the semantic type "Machine Activity" were considered relevant only when the second semantic type was "Anatomical Abnormality". The other semantic pairs identified in this category were: Age Group, Carbohydrate, Cell, Cell Function, Eicosanoid, Food, Lipid, and Tissue. Figure 4 shows the non-relevant semantic types and Figure 5 shows a few examples of semantic pairs that contain semantic types that were relevant only when associated to another particular semantic type (special pairs).

Diagnostic Procedure | Organism Attribute

Coronary Angiography	Obesity
Prenatal Diagnosis	Female

Mental or Behavioral Dysfunction | Mental Process

Stress, Psychological	Hostility
Alzheimer's Disease	Cognition

Amino Acid, Peptide, or Protein| Biologically Active Substance

Blood Coagulation Factors	Blood Proteins
Estrogen Receptors	Lipids

Diagnostic Procedure | Diagnostic Procedure

Electrocardiography, Ambulatory	Exercise stress test
Angiography, Digital Subtraction	Coronary Angiography

Disease or Syndrome | Disease or Syndrome

Asthma	Drug Hypersensitivity
Brain Ischemia	Ischemic Attack, Transient

Diagnostic Procedure | Sign or Symptom

Blood pressure determination	Hypertension
Pain Measurement	Migraine

Finding | Finding

Atrial Fibrillation	Tachycardia
Cardiac Output, Low	Diuresis

Disease or Syndrome | Pathological Function

Embolism	Hemorrhage
Pre-Eclampsia	Cerebral Hemorrhage

Injury or Poisoning | Pathological Function

Hematoma	Cerebral Hemorrhage
Myocardial Reperfusion Injury	Necrosis

Physiologic Function | Laboratory or Test Result

Blood Viscosity	Hematocrit level
Vascular resistance	Cardiac Output

Organ or Tissue Function | Organ or Tissue Function

Hemodynamics	Pulmonary Gas Exchange
Blood Coagulation	Fibrinolysis

Individual Behavior | Social Behavior

Patient Compliance	Social Behavior
Patience Acceptance of Health Care	Physician-Patient Relationships

Figure 3. Examples of relevant pairs

Discussion

The primary goal of this project was to explore an automated knowledge extraction method to determine its suitability for providing appropriate concept relationship knowledge. The previous experiment demonstrated that, compared to the amount of work required to build a knowledge base manually, the process was considerably faster and easier. The pilot study performed in order to evaluate the clinical validity of the information retrieved showed that the results were suitable for the intended purpose (literature retrieval), especially in the therapy group.

This experiment explored the relevance of semantic pair relationships generated and identified as relevant in the first study, and the clinical relevance of the semantic types in the pairs. Sixty percent of the semantic pairs were found to be relevant to a clinical context for literature review. Analyzing the semantic types of non-relevant pairs, we observed that it is possible to focus the extraction of information if we are more selective with the semantic pairs allowed. It was clear, for example, that certain semantic types are useful (e.g. Quantitative Concept and Qualitative Concept). Twenty percent of the semantic types in the sample were refuted (non-relevant types). Another ten percent were found relevant only when associated with another particular semantic type.

Taking the clinical case presented as an example, we can compare possible strategies, either using or not using the information extracted automatically by the system. If the medical student would search for myocardial infarction and aspirin in MEDLINE (PubMED), she would retrieve 572 articles. The addition of diabetes reduces the response to 22 articles. The same two strategies used to search EBM Reviews – Best Evidence would retrieve 126 and 21 articles, respectively, searching Cochrane Database of Systematic Reviews, 37 and 7.

We believe that the knowledge generated by the method described in this paper will be particularly useful for the task of retrieving relevant information from the electronic medical record in order to guide the users during the retrieval process and, consequently, improving search strategies and information retrieval. There are, however, a few limitations and concerns. The use of information extraction from MEDLINE citations raises concerns regarding the quality of indexing. Since information will be extracted from MESH terms assigned to citations, the quality of this information may differ depending on the specificity, exhaustivity, and consistency of indexing. Previous studies, for example, have found inconsistencies in human indexing in MEDLINE.[21]

Automated extraction of relationships may produce large quantities of information. In an ideal situation, a medical expert should review the validity of each relationship, but with such amount of information, this can be a very time-consuming or, perhaps, an impossible task. Another limitation of this study may be the generalizability of the method. The information extracted here was based on articles about cardiovascular diseases. Additional studies are needed to explore the use of this approach in different diseases or problems.

Although the knowledge extraction showed reasonable results, a final validation of the quality of the knowledge would require a full evaluation of the system.

Activity
Amino Acid Sequence
Body Location or Region
Body Part, Organ, or Organ Component
Chemical Viewed Structurally
Genetic Function
Geographic Area
Health Care Activity
Inorganic Chemical
Manufactured Object
Molecular Sequence
Nucleic Acid, Nucleoside, or Nucleotide
Occupational Activity
Plant
Qualitative Concept
Quantitative Concept
Research Device
Research Activity
Molecular Biology Research Activity
Temporal Concept

Figure 4. Non-relevant semantic types

Machine Activity:
Machine Activity | Anatomical Abnormality
Cardiac pacing, Artificial |Cardiomyopathy, Hypertrophic

Age group:
Age group | Organism Attribute
Adolescence | Obesity

Carbohydrate:
Carbohydrate | Injury or Poisoning
Heparin | Adverse Effects in the Therapeutic Use of Heparin

Food:
Food | Vitamin
Dietary Supplements | Beta Carotene

Figure 5. Examples of special pairs

Conclusion

This study demonstrates that it is possible to automatically extract useful medical knowledge from MEDLINE citations. The clear definition of relevant and non-relevant semantic types may allow us to limit the extraction to helpful information only. This study shows that sixty percent of semantic pairs generate by the process were judged to be relevant to the specific task proposed

(information retrieval). The extraction method may not generate totally accurate relationship information due to the problems described. However, we believe it is appropriate for the task of retrieving information from the medical record in order to guide users during a search and retrieval process.

Acknowledgments

This work was supported by a Center for Advanced Technology grant from New York State, a Digital Library Initiative grant from the National Science Foundation, and a CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) grant from Brazil.

References

[1] Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. Medical Decision Making 1995; 15(2):113-9.

[2] Hersh W. "A world of knowledge at your fingertips": the promise, reality, and future directions of on-line information retrieval. Acad Med 1999; 74(3):240-3.

[3] Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Annals of Internal Medicine 1985; 103(4):596-9.

[4] Timpka T, Ekstrom M, Bjurulf P. Information needs and information seeking behavior in primary health care. Scandinavian Journal of Primary Health Care 1989; 7(2):105-9.

[5] Shelstad KR, Clevenger FW. Information retrieval patterns and needs among practicing general surgeons: a statewide experience. Bulletin of the Medical Library Association 1996; 84(4):490-7.

[6] Gorman P. Information needs of physicians. Journal of the American Society for Information Science. 1995; 46:729-36.

[7] 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. Journal of the American Medical Association 1998; 280(15):1339-46.

[8] Fox EA, Akscyn RM, Furuta RK, Leggett JJ. Digital libraries. Communications of the ACM 1995; 38(4):23-8.

[9] Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based Medicine: How to Practice and Teach EBM. New York: Churchill Livingstone, 1997.

[10] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. British Medical Journal 1996; 312(7023):71-2.

[11] Silberg W.M., Lundberg G.D., Musacchio R.A. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor--Let the reader and viewer beware. Journal of the American Medical Association 1997; 277(15):1244-5.

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

[13] Rodrigues RJ. Information systems: the key to evidence-based health practice. Bull World Health Organ 2000; 78(11):1344-51.

[14] Cimino JJ. Linking patient information systems to bibliographic resources. Methods of Information in Medicine 1996; 35(2):122-6.

[15] Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. Journal of the American Medical Association 1987; 258(1):67-74.

[16] Cimino JJ, Barnett GO. Automatic knowledge acquisition from MEDLINE. Methods of Information in Medicine 1993; 32(2):120-30.

[17] Zeng Q, Cimino JJ. Automated knowledge extraction from the UMLS. Chute CG. Proceedings/AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus Inc., 1998: 568-72.

[18] Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. Journal of the American Medical Association 1994; 1(6):447-58.

[19] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. Journal of the American Medical Informatics Association 1994; 1(1):35-50.

[20] Mendonça EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. Proc AMIA Symp 2000; (20 Suppl):575-9.

[21] Funk ME, Reid CA. Indexing consistency in MEDLINE. Bulletin of the Medical Library Association 1983; 71(2):176-83.

Address for correspondence

Eneida A Mendonça, MD
Department of Medical Informatics, Columbia University
622West 168th Street, Vanderbilt Clinic 5th Floor
New York, NY 10032 USA
mendonca@dmi.columbia.edu