Automated Knowledge Extraction for Decision Model Construction: A Data Mining Approach

Ai-Ling Zhu*, M.Sc., Jian Li**, M.Sc., Tze-Yun Leong***, Ph.D.

Medical Computing Laboratory
School of Computing
National University of Singapore,
3 Science Drive 2, Singapore 117543

Abstract

*Combinations of Medical Subject Headings (**MeSH®**) and Subheadings in **MEDLINE®** citations may be used to infer relationships among medical concepts. To facilitate clinical decision model construction, we propose an approach to automatically extract semantic relations among medical terms from **MEDLINE** citations.We use the Apriori association rule mining algorithm to generate the co-occurrences of medical concepts, which are then filtered through a set of predefined semantic templates to instantiate useful relations. From such semantic relations, decision elements and possible relationships among them may be derived for clinical decision model construction. To evaluate the proposed method, we have conducted a case study in colorectal cancer management; preliminary results have shown that useful causal relations and decision alternatives can be extracted.*

1 Introduction

Modern clinical practice emphasizes medical research evidence, physicians' clinical expertise, as well as patients' individual values (e.g. preference, concerns, and expectations). Decision analysis is an effective method for structuring clinical problems, encoding objective evidences, representing clinicians' subjective judgements and expressing patients' preferences to derive optimizing solutions in diagnostic, therapeutic, and prognostic management. The main task in decision analysis is to construct decision models, which are mathematical frameworks with a graphical representations. This work focuses on the construction of Influence Diagrams (***IDs***), a powerful tool in structuring complex problems and facilitating communication between the physicians and decision analysts.

Decision model construction is a knowledge intensive task, involving one or more decision analysts working closely with one or more domain experts to elicit the relevant structural and numerical parameters of the decision models. It is usually an arduous task to process and integrate all the knowledge needed for model construction. In addition, as biomedical knowledge is advancing rapidly, the domain experts may not always be up-to-date on all the latest advancements in a particular field during the model construction process. Hence it is desirable to automatically derive a summary of the useful relations among the medical concepts relevant to a decision problem, which can in turn help build a proper context of the knowledge required for full decision model construction.

MEDLINE contains over 12 million citations to biomedical journal articles, indexed by the Medical Subject Heading (***MeSH***) and Subheadings which manifest the topics and the relevant contexts of these articles [1]. Based on the co-occurrences of ***MeSH*** and Subheadings in ***MEDLINE*** citations, the semantic relationships among medical concepts can be deduced. For example, "disease A" may be caused by "drug B" if "disease A/etiology" often co-occurs with "drug B/ adverse effects" in the ***MEDLINE*** citations. Such semantic relations are helpful for identifying and establishing dependencies in ***IDs***.

We present a methodology of applying an association rule mining algorithm to automatically extract semantic relations among medical concepts from the ***MEDLINE*** database. Our methodology is based on a general unsupervised learning approach that uses background knowledge to guide the information extraction process. The extracted information aims to facilitate ***ID*** construction by providing a context for further knowledge elicitation in the modeling process.

1.1 Methodology

1.2 Influence Diagrams

An influence diagram (ID) is a directed acyclic graph that provides an intuitive way to identify and represent

* Current address: Bioinformatics Institute, 21 Heng Mui Keng Terrace, Singapore 119613, Email Address: ailing@bii.a-star.edu.sg.
** Email Address: jian_li_sg@yahoo.com.sg
*** Email Address: leongty@comp.nus.edu.sg

the decision factors, including decision alternatives, uncertainties, and objectives, and how they influence each other.

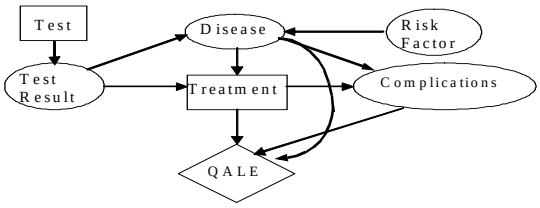

Figure 1: An influence diagram.

Figure 1 shows an influence diagram which contains two decision nodes, "Test" and "Treatment" denoted by squares/rectangles; four chance nodes, "Disease", "Risk Factor", "Test Result", "Complications" denoted by ovals; one value node, "quality of life expectancy (QALE)" denoted by the diamond; ten directed arcs representing the informational and probabilistic dependency relationships among these nodes. Some of the arcs also indicate causal relations: e.g., "Risk Factor" would cause "Disease", "Treatment" would cause "Complications", and so on. These relations can be extracted automatically from MEDLINE citations.

1.3 The System Architecture

We use the association rule mining approach to discover co-occurrences of ***MeSH*** main heading (***MH***) and subheading (***SH***) combinations in the form of {(***MH***A/***SH***B), (***MH***C/***SH***D)}, such as {(Colorectal Neoplasms/Pathology)*, (Liver Neoplasms/Secondary)*}, etc., in ***MEDLINE*** citations. After being filtered through a set of predefined semantic templates, these co-occurrences can be further explained as "cause and effect" relations (Colorectal Neoplasms ***cause*** Liver Neoplasms), and other kinds of semantic relations between medical entities, which could be used for ID construction.

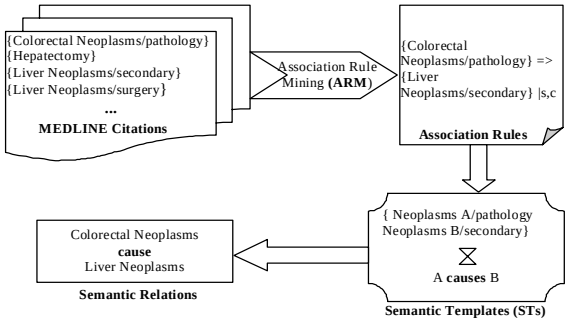

Figure 2: The System Architecture.

* Colorectal Neoplasms/pathology indicates that "Colorectum is the original site where neoplasms have developed", while Liver Neoplasms/secondary indicates that "Liver is the secondary location to which the neoplastic process has metastasized".

As shown in ***Figure 2***, the inputs of our system are ***MEDLINE*** citations, and the outputs are semantic relations. The system performance is measured by the percent of relations correctly extracted. In our proposed system architecture, there are two key components: Association Rule Mining (ARM) process and Semantic Templates (STs). ***Figure 2*** illustrates that the ARM is used to mine association rules in ***MEDLINE*** citations, and such association rules are filtered through a set of predefined STs to instantiate semantic relations.

1.4 Association Rule Mining [2]

The approach to mining association rules was introduced by Agrawal et. al. [2]. Given a set of transactions ***T*** (each transaction is a set of items), an association rule can be expressed in the form $X \Rightarrow Y$ (which is read as "**X** implies **Y**"), where ***X*** and ***Y*** are mutually exclusive sets of items.

The rule's statistical significance is measured by ***support***, and the rule's strength by ***confidence***. The ***support*** of the rule is defined as the percentage of transactions in ***T*** that contain both ***X*** and ***Y***, and may be regarded as ***P(*** $X \cup Y$ ***)***-the probability of $X \cup Y$ (**X** union **Y**).The ***confidence*** of the association rule is the ratio of the support of the itemset $X \cup Y$ to the support of the itemset ***X***, which roughly corresponds to the conditional probability ***P(Y|X)***.

The problem of mining association rules can be divided into two subproblems:

1. Find all combinations of items whose supports are greater than a user-specified minimum support (threshold). The combinations are called frequent itemsets.
2. Use the items from frequent itemsets to generate the desired rules. More specifically, the confidence of each rule is computed, and if it is above the confidence threshold, the rule is satisfied.

   For example:

   {(Colorectal Neoplasms/pathology), (Liver Neoplasms/secondary)} is a ***frequent itemset*** generated from Step1;

   The ***confidence*** of the rule-{(Colorectal Neoplasms/pathology)=> (Liver Neoplasms/secondary)}- is calculated as:

   support{(Colorectal Neoplasms/pathology), (Liver Neoplasms/secondary)} / support(Colorectal Neoplasms/pathology);

   If the confidence calculated is greater than the confidence threshold,

   Then {(Colorectal Neoplasms/pathology) => (Liver Neoplasms/secondary)} holds.

We apply the Apriori algorithm [3] in this work since it is the most efficient serial algorithm for discovering association rules in large databases. The implementa-

tion of the Apriori algorithm used is downloaded from [4].

1.5 Semantic Templates

MEDLINE citations are indexed by ***MeSH*** headings (***MHs***) and subheadings (***SHs***), which are coordinated to accurately convey the content of these journal articles: ***MHs*** are the preferred descriptors for subjects; ***SHs***, also called ***MeSH*** qualifiers, are used to express a certain aspect of a ***MH*** [1]. The process of indexing a journal article roughly consists of three steps: identifying the subjects and the relations among them as discussed in the article; searching the matched ***MHs*** for these subjects from ***MeSH*** vocabulary; expressing these subjects and the relations by the ***MHs*** and/or ***SHs***. We are interested in the reverse process: discovering the relations among medical concepts through examining the co-occurrences of ***MHs*** and subheadings in citations for and against a set of semantic templates.

The set of semantic templates are developed based on the National Library of Medicine (NLM)'s instructions for the correct coordination of MHs and SHs during the process of indexing articles [1]. The semantic templates cover four classes of ***MeSH*** terms in ***MeSH Tree***: Anatomy[A]; Diseases[C, Neoplasm[C04]]; Chemicals and Drugs[D]; Analytical, Diagnostic and Therapeutic Techniques and Equipment[E, Diagnosis[E01], Therapeutics[E02], Surgical Procedures, Operative[E04]].

In this preliminary study, 71 semantic templates are defined in the following form:

Semantic Relation ***A***: {(***MH 1*** category ***X*** **[/]** ***SH b***), (***MH 2*** category ***Y*** **[/]** ***SH c***)}.

which can be interpreted as a boolean expression:

If ***MH 1*** is from ***MH*** category ***X AND*** appears with/without ***SH b***, ***AND*** it associates ***MH 2*** which is from ***MH*** category ***Y AND*** appears without/with ***SH c***,
Then ***MH 1*** and ***MH 2*** are associated with Relation ***A***.

Table 1 lists some examples of the semantic templates designed in our implementation.

Table 1: Examples of semantic templates.

MH 1 category ***X*** **[/]** ***SH b***	***MH 2*** category ***Y*** **[/]** ***SH c***	Relation MH1 **A** MH2
Disease 1/CO	**Dis**ease 2/ET	Dis 1 **causes** Dis 2
Neoplasm 1/PA	**Neo**plasm 2/SC	Neo 1 **causes** Neo 2
Disease 1/CI	**Chem**ical 2/PO	Chem 2 **causes** Dis 1
Disease 1/DI	**Dia**gnosis 2	Dia 2 **diagnoses** Dis 1
Disease 1/DI	**Dia**gnosis 2/MT	
Disease 1/DI	**Dia**gnosis 2/IS	
Disease 1/DT	**Drug** 2/TU	Drug 2 **treats** Dis 1
Disease 1/SU	**Surg**ery 2/IS	Surg 2 **treats** Dis 1

Keys: CO(Complications), ET(Etiology), PA(Pathology), SC(Secondary), CI(Chemically Induced), PO(Poisoning), DI(Diagnosis), MT(Methods), DT(Drug Therapy), IS(Instrumentation), SU(Surgery), TU(Therapeutic Use).

The following example illustrates how the semantic templates are generated and instantiated. With the knowledge that indexers will assign the ***MeSH SH***: "Pathology" to the ***MeSH MH*** "Neoplasm 1" which can cause the other kind of ***MeSH MH*** "Neoplasm 2", and the ***MeSH MH*** "Neoplasm 2" developed due to metastases will be assigned the ***MeSH SH***: "Secondary", we may define the following template:

IF {(Neoplasm 1/Pathology), (Neoplasm 2/Secondary)}**,**
THEN Relation-Neoplasm 1 **causes** Neoplasm 2-exists.

MeSH Tree provides a taxonomy to define a subsumption or "isa" relationship, for instance, Colorectal Neoplasms is a Neoplasms, which can be used to judge the category of a ***MH*** term. Therefore, the template described above can be instantiated as:

IF {(Colorectal Neoplasms/pathology), (Liver Neoplasms/secondary)}, **THEN** "Colorectal Neoplasms **causes** Liver Neoplasms" holds.

2 Extracting Semantic Relations from MEDLINE: a case study

MEDLINE has about approximately 12 million records[1], and around $20000(\mathrm{NumberofMH}) \times 80(\mathrm{NumberofSH}) = 16\times10^5$ ***MeSH Terms***. For the first round of frequent set generation, the ARM algorithm needs to allocate as many as 16×10^5 counters and the cost of ***I/O*** will be expensive if the whole database is used as data source for extracting the relevances. Fortunately, this problem can be avoided because clinical decision analysis usually focuses on one target clinical problem each time.

In our case study, we were interested in discovering all the relations on "Colorectal Neoplasms". We searched ***MEDLINE*** using the following search strategy: "Colorectal Neoplasms" [MAJOR_MESH_NOEXP]. As a result, 12,488 records with 16 ***MeSH MH*** and ***SH*** combinations per record on average, were found and downloaded from [5]. This dataset was used as the input of ARM to generate the frequent set and association rules (co-occurrences).

In the experiment, 1,345 associate rules were generated for a minimum support of 0.4% and minimum confidence of 30%, in the form $X \Rightarrow Y(s,c)$, where s is the support and c the confidence.

{(Colorectal Neoplasms/pathology)=>(Liver Neoplasms/secondary) (9.0%,48.2%)}

{(Colorectal Neoplasms/etiology)=>(Smoking/adverse effects) (0.4%,46.0%)}

{(Colorectal Neoplasms/surgery)=>(Colectomy) (1.0%,44.6%)}

{(Colorectal Neoplasms/drug therapy)=>(Fluorouracil/ therapeutic use) (2.7%,51.8%)}

{(Colorectal Neoplasms/diagnosis)=>(Occult Blood) (3.2%,42.8%)}...

From 1345 associate rules, 33 semantic relations were generated, for example:

"Colorectal Neoplasms" **causes** "Liver Neoplasms"

"Smoking" **causes** "Colorectal Neoplasms"

"Colectomy" **treats** "Colorectal Neoplasms"

"Fluorouracil" **treats** "Colorectal Neoplasms"

"Occult Blood" **diagnoses** "Colorectal Neoplasms"...

We verified these relations through manually examining the indexed ***MeSH*** Terms and the abstracts in these citations. Of 33 possible relations, 32(97%) were found to be acceptable and 1 (3%) was unreasonable: "Postoperative Period **treats** Colorectal Neoplasms", generated mainly because "Postoperative Period" belongs to "Surgical Procedures, Operative" (E04) in ***MeSH Tree***. The accuracy rate can be improved via modifying our semantic templates (e.g. adding constraints).

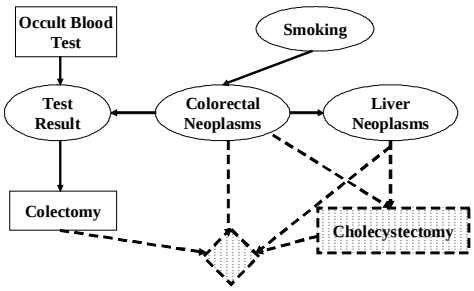

Figure 3: An influence diagram constructed from the relations mined in MEDLINE.

As shown in Figure 3, an influence diagram was constructed using the relations discovered in ***MEDLINE***. The dotted arcs and components indicate concepts and relations that were not derivable from the discovered concepts and relations, but may be relevant to the problem addressed. For instance, in the case study, a semantic rule of the form: *"Cholecystectomy **causes** Colorectal Neoplasms"*- was not admissible because the following rule: *{(Colorectal Neoplasms/etiology) => (Cholecystectomy/adverse effects) (0.16%,41.6%)}*, has a support (0.16%) below our minimum support (0.4%). Such cases can be further examined through adjusting minimum support, but some incorrect relations may be generated in the meantime. Therefore, setting the threshold of support and confidence requires balancing the trade-off between information overloading and crucial knowledge missing. For decision analysis, the minimum support and confidence can be set according to the scope of the target clinical decision problem.

In any case, the semantic relations discovered were useful in setting up the context or skeleton of an ID to be further examined and refined to model the decision problem. Some "unexpected" or "surprising" relations which may turn out to be important may also be discovered in the process.

3 Related Work

There are some efforts in extracting information or knowledge from the ***MEDLINE*** database. Craven applied machine learning techniques to map biomedical knowledge - some semantic relations among molecular biological objects- from ***MEDLINE*** text sources into structure representations (e.g. knowledge base) [6]. In Weeber and Vos's work, textual Intelligent Data Analysis (IDA) approach is adopted to extract information about adverse effects of drug combinations from ***MEDLINE*** [7]. In these efforts, the relationships between words are statistically identified, and the methods or systems are developed with the aim of extracting knowledge from ***free text*** sources (e.g., ***MEDLINE*** abstracts). Although these works shares with our work the common aim of extracting knowledge from ***MEDLINE***, the specific information sources and methods we use are different: association rule mining approach is adopted to discover semantic relations from structured data - ***MEDLINE*** citations.

Using the co-occurrences of ***MeSH*** Headings in ***MEDLINE*** citations to extract medical knowledge was first introduced by Cimino et. al. [8]. They developed a rule-based system to conduct the searches and then manually examine the searching results. In 1998, Zeng and Cimino improved on the method by automatically extracting disease-chemical knowledge from the UMLS MRCOC table [9], which provided the co-occurrence of ***MeSH terms*** designated as main points in the same journal article. The most recent study was performed by Mendonca and Cimino, in which a co-occurrence table similar to UMLS MRCOC was built to extract relevant semantic pairs for information retrieval [10]. Such relevant semantic pairs are generated on an abstract level using the UMLS semantic types, and only "relevance" relation is established in the semantic pairs. Our approach, on the contrary, is targeted at discovering specific relations such as causal relations among specific medical terms, for clinical decision model construction.

4 Discussion

We have introduced an approach to automatically ex-

tract medical knowledge from ***MEDLINE*** citations for decision model construction by applying an association rule mining algorithm. Our methodology has the following features:

First, as shown in our case study, the knowledge extracted from ***MEDLINE*** citations can be helpful for decision analysts to construct clinical decision models. Such knowledge as causal relations between medical concepts can be mapped into the chance nodes and decision nodes, even the dependences among them in IDs. The relations discovered are supported by published clinical literature, which facilitates the practice of evidence-based medicine.

Second, our approach can flexibly deal with target clinical decision problems of different scopes, and support generation of a variety of relations. The target clinical decision problems may have different requirements for the number and statistical strength of generated relations, which can be controlled via adjusting the minimum support and confidence. Moreover, since the ARM process can mine not only the co-occurrences of two concepts but those of three concepts and more, the relations among multiple concepts such as *"Cholecystectomy **treats** Gallbladder Neoplasms but **causes** Colorectal Neoplasms"* can be extracted if such semantic templates are predefined.

Third, we could benefit from the use of association rule mining algorithms in the following aspects. Our approach can perform efficiently in extracting medical knowledge from large databases such as ***MEDLINE,*** since many fast algorithms including serial and parallel algorithms have been well developed or are being improved for manipulating these massive databases [3][11]. On the other hand, ARM can produce useful data -- the counts of frequent itemsets, which allow us to apply different statistical measurements to effectively discover and evaluate association rules.

Initial results from a case study have demonstrated the usefulness of the proposed approach. Future research agenda for this work include the following: First, although the current knowledge captured can be converted into semantic network directly, such knowledge does not explicitly indicate probabilistic dependencies in ***IDs.*** We plan to address the challenging task of extracting and incorporating the numerical and probabilistic knowledge as priors into Bayesian Network learning algorithms to support learning of the probabilistic dependencies from data. Second, as our semantic templates are defined manually based on ***MEDLINE*** indexing routines, missing templates may occur due to the limitation of the knowledge engineer. We plan to define a more complete set of semantic templates via learning more about the indexing patterns in ***MEDLINE***. Finally, we plan to investigate other statistical techniques which can be used to effectively mine the knowledge via analyzing the support counts of frequent itemsets.

Acknowledgment

We would like to thank our other group members, especially Prof. Kim Leng Poh, for their valuable comments and suggestions. This research is supported by a Research Grant No. R-252-000-111-112/303 from the Agency of Science and Technology (A*Star) and the Ministry of Education in Singapore.

References

[1] http://www.nlm.nih.gov/databases/ databases_medline.html.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Conference on Management of Data*, pages 207–219, Washington, D.C. USA, May 1993.

[3] R. Agrawal, and J. Shafer. 1996. Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engineering*.8(6).

[4] http://www.cs.helsinki.fi/u/goethals/software/ index.html#apriori

[5] http://www.nlm.nih.gov

[6] M. Carven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proc. of the 7th International Conference on Intelligent Systems for Molecular Biology*, 1999.

[7] M. Weeber and R. Vos. Extracting expert medical knowledge from texts. In *Working Notes of the Intelligent Data Analysis in Medicine and Pharmacology Workshop*, 1998: 183–203.

[8] J. J. Cimino and G. O. Barnett. Automatic knowledge acquisition from medline. *Methods of Information in Medical*, 32(2):120–130, 1993.

[9] Q. Zeng and J. J. Cimino. Automated knowledge extraction from the UMLS. *Proceedings/AMIA Annual Fall Symposium*. Philadelphia: Hanley & Belfus Inc., 1998: 568-72

[10] E.A. Mendonca and J. J. Cimino. Automated knowledge extraction from MEDLINE citations. *Proc. AMIA Fall Symposium*, 2000: 575-579.

[11] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. of 20th Int'l Conference on Very Large Databases*, Santiago, Chile, September 1994.