Automated Information Extraction to Support Biomedical Decision Model Construction: A Preliminary Design

Xiaoli Li[a,b], TzeYun Leong [a,b]

[a] Singapore-MIT Alliance, 4 Engineering Drive 3, National University of Singapore, Singapore-117576
[b] Department of Computer Science, 3 Science Drive 2, National University of Singapore, Singapore-117543

Abstract—We propose an information extraction framework to support automated construction of decision models in biomedicine. Our proposed technique classifies text-based documents from a large biomedical literature repository, e.g., MEDLINE, into predefined categories, and identifies important keywords for each category based on their discriminative power. Relevant documents for each category are retrieved based on the keywords, and a classification algorithm is developed based on machine learning techniques to build the final classifier. We apply the HITS algorithm to select the authoritative and typical documents within a category, and construct templates in the form of Bayesian networks. Data mining and information extraction techniques are then applied to extract the necessary semantic knowledge to fill in the templates to construct the final decision models.

Index Terms — Data Mining, Decision model, information extraction,

I. Introduction

Decision analysis aids decision-making under uncertainty by systematically representing, analyzing, and solving complex decision models, which are mathematical frameworks with graphical representations [12]. With the rapid advancement of biomedical knowledge, a large quantity of new findings, methodologies, and insights are published and made available online. Decision model construction in biomedical decision analysis can be greatly facilitated by automatically deriving the relevant semantic knowledge from online biomedical resources.

MEDLINE [3] is the United States National Library of Medicine (**NLM)**'s [2] premier bibliographic database. It covers many biomedicine fields: nursing, dentistry, veterinary medicine, health care, and preclinical sciences. In **MEDLINE,** the citations and author abstracts of more than 4600 biomedical journals published in the United States and worldwide were available online. Each citation in **MEDLINE** is assigned one or more terminologies from NLM'S controlled vocabulary Medical Subject Heading (Mesh Term) as well as other attribute values such as publication types, author and title.

This work aims to automatically derive the semantic knowledge and evidence from a large biomedical repository, such as MEDLINE, to assist in building a substantial part of a decision model. Given a particular user's query (e.g., colorectal cancer), we first search the **MEDLINE** database to get relevant documents about this particular disease. We then design a text classifier without any manual labeling. It can classify the documents into predefined categories, e.g., symptoms, screening and diagnostic procedures, etc. Within each category, we find those documents which describe the important and comprehensive methods and steps. Templates for each category, in the form of Bayesian networks, can then be constructed with the help of the domain experts. Finally, Data mining and information extraction techniques are applied to extract the semantic knowledge for filling in the templates to construct the final decision models.

Xiaoli Li, Singapore-MIT Alliance, 4 Engineering Drive 3, National University of Singapore, Singapore-117576,
Tel +65 68744098; fax +65 67794580;
email: smalxl@nus.edu.sg

Tze Yun Leong, School of Computing, 3 Science Drive 2, National University of Singapore, Singapore-117543,
email: leongty@comp.nus.edu.sg

II. Related works

Various related research efforts exist in extracting information or knowledge from the **MEDLINE** database [7][9][10][11][12]. For example, Craven designed a Naïve bayes classification technique to classify the sentences of MEDLINE abstracts (rather than the whole abstracts) and then tries to extract out the relations in these sentences, e.g., subcellular-localization.

Although all these efforts try to extract knowledge from **MEDLINE**, our approach is targeted at a specific task, i.e., to support biomedical decision model construction. The proposed technique can automatically classify the medical literatures into predefined categories. Also, we are trying to learn the semantic knowledge from both free text (abstracts) and structured information (**MEDLINE** citations). Data mining and information extraction techniques are also used to discover new evidence and semantic knowledge.

III. The proposed technique

This section presents the proposed approach, which consists of two steps: (1) classify the medical literature; (2) find those authoritative and typical documents within a category, perform further knowledge extraction and construct the final decision

A. Medical literature classification

Since the **MEDLINE** database contains tens of millions of records, we focus on the dataset that has an indexing MeSH term of "Colorectal Neoplasm". We send a query with the following keywords "colorectal neoplasm, colonoscopy, diagnoses" to retrieve a set of documents. We classify them into predefined categories; each of the categories presents a particular aspect of the disease. The predefined categories are Symptoms, Diagnostic procedures, Treatment strategies, Screening techniques, Patient profile, Root Disease Causes, Environmental Factors (location, etc).

Traditionally, text classifiers are often built from manually labeled training examples. Our proposed technique does not require any manual labeling. In the past few years, researchers investigated the idea of using a small set of labeled examples and a large set of unlabeled examples to help in learning. Recently, the problem of building classifiers from positive and unlabeled examples (no negative examples) was also studied. These efforts all aim to reduce the burden of manual labeling. We propose a new approach that requires no manual labeling. Given a particular user's query (e.g., colorectal cancer), our approach first constructs an original noisy training set for each category by using the results from a search engine (e.g., search "colorectal cancer screening" using Google). Then an entropy based feature selection method is used to identify those important keywords for each category based on their discriminative power. By using the keyword set for each class we retrieve the relevant documents for each category. Finally, through learning from the information in the training set and the Medical Subject Heading (MeSH) terms in the documents, a classification algorithm is applied to build the final classifier.

Figure 1 shows the structure of our system. It illustrates four steps: 1) constructing an original training set by sending the queries to search engine; 2) selecting features from the original training set; 3) retrieving relevant documents from MEDLINE database; and 4) building final classifier by learning from the training set and MeSh terms. Our preliminary results show that the new approach is effective and promising.

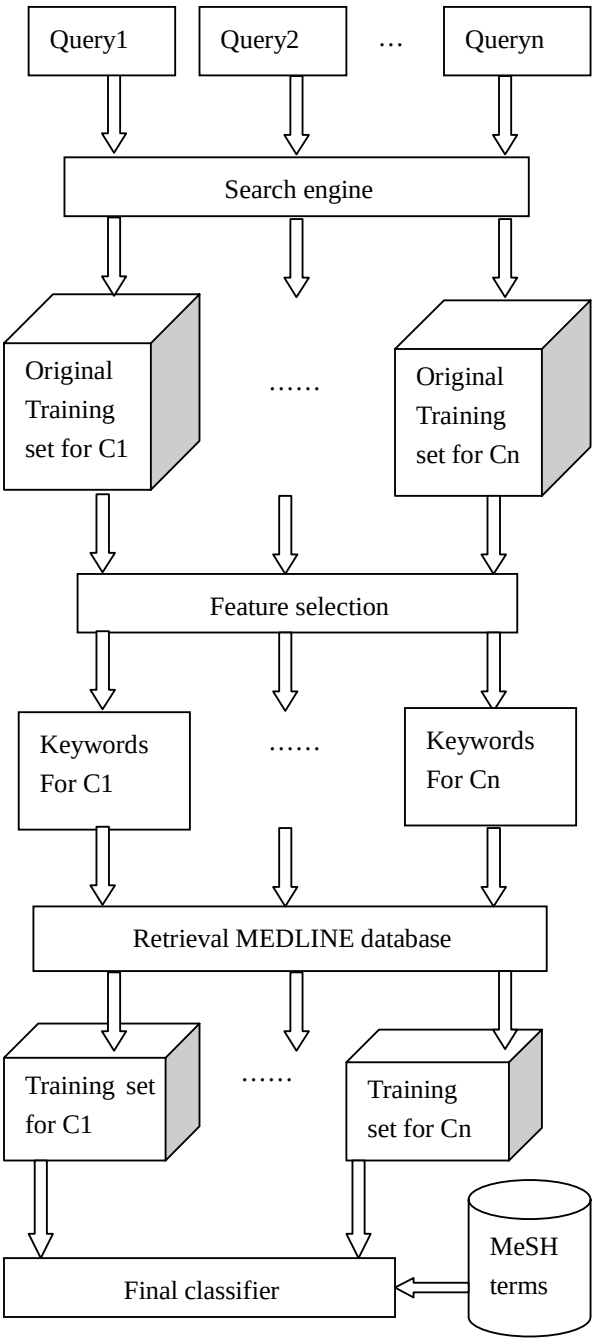

Figure 1. Structure of classification system

B. Perform further knowledge extraction and construct the final decision models

This phase is to find those authoritative and typical documents within a category to perform further knowledge extraction. For each category, for example, "colorectal cancer screening", we find those documents which describe the important and comprehensive screening methods and steps. The HITS algorithm, which is widely used in search topic information in information retrieval, can be used to analyze the citations to identify the relevant documents. Templates for each category, in the form of Bayesian networks, can then be constructed with the help of the domain experts. Data mining and information extraction techniques are then applied to extract the semantic knowledge for filling in the templates to construct the final decision models.

1) Find the authoritative literature in each category

For each category, for example, "screening techniques", we want to find those authoritative and typical documents in order to perform further knowledge extraction. Here we borrow the idea from the HITS algorithm. Note here citations of the articles correspond to the hyperlinks in web pages. The HIT algorithm was first introduced by Jon M. Kleinberg [12]. He assumes that a topic can be roughly divided into pages with good coverage of the topic, called authorities, and directory-like pages with many hyperlinks to useful pages on the topic, called hubs. And the goal of HITS is basically to identify good authorities and hubs for a certain topic which is usually defined by the user's query. Figure 2 depicts the HITS algorithm.

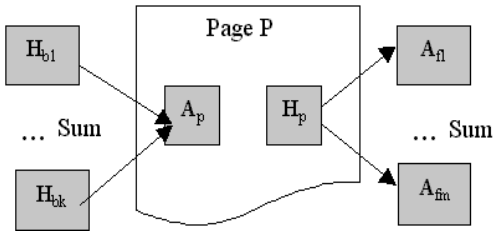

Figure 2. The HITS algorithm.

Given a user query, the HITS algorithm first creates a neighborhood graph for the query. The neighborhood contains the top *n* matched web pages retrieved from a content-based web search engine; it also contains all the pages these *n* web pages link to and pages that are linked to these *n* top pages (*n*=200 in their experiments).

Then, an iterative calculation is performed on the value of authority and value of hub. For each page p , the authority and hub values are computed as follows:

$$Ap \leftarrow \sum_{q(q,p)\in G} Hq$$

$$Hp \leftarrow \sum_{q(p,q)\in G} Aq$$

The authority value *Ap* of page p is the sum of hub scores of all the pages that points to p, the hub value *Hp* of page p is the sum of authority scores of all the pages that p points to (Fig.2). Iteration proceeds on the neighborhood graph until the values converge. Kleinberg claimed that the small number of pages with the largest authority converged value should be the pages that have the best authorities for the topic. And the experimental results support the concept.

In our work, we try to find the authoritative documents which include the comprehensive methods and steps. The information we extracted from these relevant articles can be used to help construct better quality decision models.

2) Mining semantic knowledge from Mesh Terms and full text

Based on previous work done in [13], we want to use data mining techniques to extract the semantic knowledge from 1) Mesh terms 2) full text. One of the most popular data mining applications is that of mining association rule, which was introduced in 1993 [2, 3, 4, 11]. It can be used to identify relationships among a set of items in a database.

A formal statement of the association rule problem is as follows:

Definition 1: Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of m distinct attributes, also called *literals*. Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that $T \subseteq I$. An *association rule* is an implication of the form $X \Rightarrow Y$, where $X, Y \subset I$, are sets of items called *itemsets*, and $X \bigcap Y = \phi$. Here, X is called the antecedent, and Y the consequent.

Two important measures for association rules, support (*s*) and confidence (α), can be defined as follows.

Definition 2: The *support (s)* of an association rule is the ratio (in percent) of the records that contain $X \bigcup Y$ to the total number of records in the database.

Therefore, if we say that the support of a rule is 5% then it means that 5% of the total records contain $X \bigcup Y$. Support is the statistical significance of an association rule.

Definition 3: For a given number of records, *confidence (α)* is the ratio (in percent) of the number of records that contain $X \bigcup Y$ to the number of records that contain X.

Thus, if we say that a rule has a confidence of 85%, it means that 85% of the records containing X also contain Y. The confidence of a rule indicates the degree of correlation in the dataset between X and Y. Confidence is a measure of a rule's strength. Often a large confidence is required for association rules.

Mining of association rules from a database consists of finding all the rules that meet the user-specified threshold support and confidence. The problem of mining association rules can be decomposed into two sub problems [11] as stated in Algorithm

3) Construct a decision model using the semantic knowledge

Templates for each category, in the form of Bayesian networks, can then be constructed with the help of the domain experts. The basic infrastructure of the final decision models can be constructed based on the semantic knowledge dreived from applying the data mining techniques. The templates can be substantially filled by using the information from: 1) semantic relations which are obtained by association rule mining; 2) Semantic networks from the UMLS; and 3) detail parameters and utility functions derived by using information extraction and natural language processing techniques Such partially filled templates would then go through the normal decision analysis cycles and iterations to derive the final decision models. An example decision model, in the form of an influence diagram, is shown in Figure 3. It is a structure of directed acyclic graph and can be used to

identify and represent the decision factors and the relations between them. One of the advantages is that it can facilitate the communication between physicians and decision analysts.

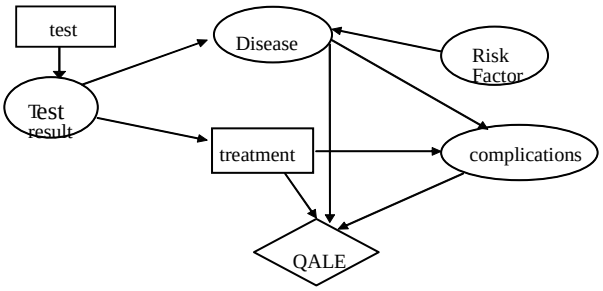

Figure 3. An influence diagram

IV. Discussion

This paper sketches the initial design of a technique which facilitates the construction of biomedical decision models automatically. Since biomedical knowledge is advancing rapidly, the ability to construct such models from the online resources are very important.

Our approach first classifies articles in the biomedical literature into predefined categories without manually labeled training documents. Then within a category our technique tries to find the important and comprehensive documents so that we can perform further knowledge extraction. The knowledge which we obtained by using Data mining and information extraction technique can be used to fill the templates toward constructing the final decision models.

References

1. http://www.cs.helsinki.fi/u/goethals/software/index.html#apriori.
2. http://www.nlm.nih.gov.
3. http://www.nlm.nih.gov/databases/databases_medline.html.
4. Agrawal, R., Imielinski, T. and Swami, A., Mining association rules between sets of items in large databases. in *In ACM SIGMOD Conference on Management of Data*, (Washington,D.C. USA, 1993), pages 207-219.
5. Agrawal, R. and Shafer, J. Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engineering.*, *8* (6).
6. Agrawal, R. and Srikant, R., Fast Algorithms for Mining Association Rules. in *Proc. of 20th Int'l Conference on Very Large Databases.*, (Santiago, Chile, 1994).
7. Carven, M. and Kumlien, J., Constructing biological knowledge bases by extracting information from text sources. in *In Proc. of the 7th International Conference on Intelligent Systems for Molecular Biology*, (1999).
8. Kleinberg, J., Authoritative Sources in a Hyperlinked Environment. in *Proc. 9th Ann. ACM-SIAM Symp. Discrete Algorithms,*, (New York, 1998), ACM Press, 668-677.
9. Mendonca, E.A. and Cimino, J.J., Automated knowledge extraction from MEDLINE citations. in *Proc.AMIA Fall Symposium*, (2000), 575-579.
10. Weeber, M. and Vos, R., Extracting expert medical knowledge from texts. in *In Working Notes of the Intelligent Data Analysis in Medicine and Pharmacology Workshop*, (1998), 183-203.
11. Zeng, Q. and Cimino, J.J., Automated knowledge extraction from the UMLS. in *Proceedings/AMIA Annual Fall Symposium.*, (Philadelphia, 1998), 568-572.
12. Zhu, A., Li, J. and Leong, T.Y., Automated knowledge Extraction for decision model construction: a data mining approach. in *Annual Symposium of the American Medical Informatics Association(AMIA)*, (USA, 2003).

Xiaoli Li, received his Ph.D. degree in Computer Software Theory from Chinese Academy of Sciences, Beijing, China. His research interests include Knowledge Discovery and Data Mining (Text and WEB Mining), Machine Learning, Information Retrieval, Bioinformatics, etc.

Tze Yun Leong, received her Ph.D. degree in Electrical Engineering and Computer Science from Massachusetts Institute of Technology (MIT), USA. She is an Associate Professor in the Department of Computer Science, School of Computing, at the National University of Singapore. She is also the Co-chair of the Singapore-MIT Alliance (SMA) Computer Science Program. She directs the Medical Computing Laboratory in the School and leads the Biomedical Decision Engineering Group, a cross-faculty, multidisciplinary research program at the university.