Efficient Retrieval of Text for Biomedical Domain using Expectation Maximization Algorithm

Sumit Vashishtha[1], Dr. Yogendra Kumar Jain[2]

[1] MTECH Scholar, Computer Science Department
Samrat Ashok Technological Institute, Vidisha, M.P, INDIA

[2] Head, Computer Science Department,
Samrat Ashok Technological Institute Vidisha, MP, INDIA

Abstract

Data mining, a branch of computer science [1], is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. Biomedical text retrieval refers to text retrieval techniques applied to biomedical resources and literature available of the biomedical and molecular biology domain. The volume of published biomedical research, and therefore the underlying biomedical knowledge base, is expanding at an increasing rate. Biomedical text retrieval is a way to aid researchers in coping with information overload. By discovering predictive relationships between different pieces of extracted data, data-mining algorithms can be used to improve the accuracy of information extraction. However, textual variation due to typos, abbreviations, and other sources can prevent the productive discovery and utilization of hard-matching rules. Recent methods of soft clustering can exploit predictive relationships in textual data. This paper presents a technique for using soft clustering data mining algorithm to increase the accuracy of biomedical text extraction. Experimental results demonstrate that this approach improves text extraction more effectively that hard keyword matching rules.

Keywords-: *Data mining, Biomedical text extraction, Biomedical text mining,Maximization algorithm*

1. Introduction

This paper aims to use data mining techniques to extract text from biomedical literature with reasonably high recall and precision. In recent years, along with development of bioinformatics and information technology, biomedical technology grows rapidly. With the growth of the biomedical technology, enormous biomedical databases are produced. It creates a need and challenge for data mining. Data mining is a process of the knowledge discovery in databases and the goal is to find out the hidden and interesting information[3]. The technology includes association rules, classification, clustering, and evolution analysis etc. Clustering algorithms are used as the essential tools to group analogous patterns and separate outliers according to its principles that elements in the same cluster are more homogenous while elements in the different ones are more dissimilar [2]. Furthermore, data mining algorithms do not need to rely on the pre-defined classes and the training examples while classifying the classes and can produce the good quality of clustering, so they fit to extract the biomedical text better. A major challenge for information retrieval in the life science domain is coping with its complex and inconsistent terminology. In this paper we try to devise an algorithm which makes word-based retrieval more robust. We will investigate how data mining algorithms based on keywords affects retrieval effectiveness in the biomedical domain. We will try to answer the following research question in this paper "How can the effectiveness of word-based biomedical information retrieval be improved using data mining algorithm?"

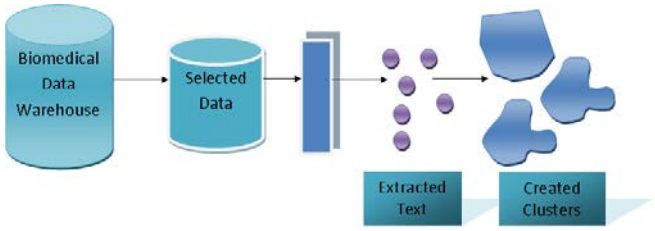

Figure 1: Text extraction from Biomedical literature base

2. Background

Biomedical text extraction refers to text mining applied to texts and literature of the biomedical and molecular biology domain. It is a rather recent research field on the edge of natural language processing, bioinformatics, medical informatics and computational linguistics.

There is an increasing interest in text mining and information extraction strategies applied to the biomedical and molecular biology literature due to the increasing number of electronically available publications stored in databases.

The main developments in this area have been related to the identification of biological entities (named entity recognition), such as protein and gene names in free text, the association of gene clusters obtained by microarray experiments with the

biological context provided by the corresponding literature, automatic extraction of protein interactions and associations of proteins to functional concepts (e.g. gene ontology terms). Even the extraction of kinetic parameters from text or the subcellular location of proteins have been addressed by information extraction and text mining technology.

The optimal retrieval of a literature search in biomedicine depends on the appropriate use of Medical Subject Headings, descriptors and keywords among authors and indexers. We hypothesized that authors, investigators and indexers in four biomedical databases are not consistent in their use of terminology in Complementary and Alternative Medicine.

The increasing research in Complementary and Alternative Medicine and the importance placed on practicing evidence-based medicine require ready access to the biomedical scientific literature. The optimal retrieval of a literature search in biomedicine depends on the appropriate use of Medical Subject Headings, descriptors and keywords among authors, indexers, and investigators [4]. It has been recognized that available online databases for biomedical domain differed in their thesaurus construction and indexing procedures, making effective and efficient searching difficult [5].

In this paper we try to employ an algorithm that extracts the biomedical texts fro the biomedical database based on the some data mining algorithm. Our approach first identifies the keywords contained in the biomedical database and then clustering these keywords to group all the text that fall into the category of the given keyword i.e. if that keyword is being used for searching the returned cluster for that particular keyword will contain all the text corresponding to that keyword.

3. Method

The main goal of the proposed system is to find valuable information of biomedical domain from the web. In the next step these information is exploited to mining user navigation pattern based on Expectation Maximization (EM) algorithm.

We adopted the usage mining system to exploit user navigation pattern based on the EM algorithm. According to different function, the system is partitioned into two main modules; Data pretreatment and navigation pattern mining.

Data pretreatment in a web usage mining model, aims to reformat the original web logs to identify all web access sessions. There are several tasks in this module of the system. The Web server usually registers all users' access activities of the website as Web server logs. Due to different server setting parameters, there are many types of web logs, but typically the log files share the same basic information, such as: client IP address, request time, requested URL, HTTP status code, referrer, etc. Data cleaning and filtering methods are the next step in this module. Not every access to the content should be taken into consideration. We need to remove accesses to irrelevant items, redundant items, accesses by Web crawlers, and failed requests. Data cleaning also identifies Web robots and removes their request. Web robots (also called spiders) are software tools that scan a web site to extract its content. In the web usage mining algorithm, to get knowledge about each user's identity is not necessary. However, a mechanism to distinguish different users is still required for analyzing user access behavior. A user session is a delimited set of pages visited by the same user within the duration of one particular visit to a Web site. Session identification is carried out using the assumption that if a certain predefined period of time between two accesses is exceeded, a new session starts at that point. Sessions can have some missing parts. This is due to the browser's own caching mechanism and also because of the intermediate proxy-caches. The missing parts can be inferred from the site's structure. Web usage data is prepared for applying navigation patterns mining algorithms by doing these pretreatment tasks.

Mining of navigation patterns is the main task of the proposed system. The main objective of this module is mining of user's navigation pattern. A user navigation pattern is common browsing characteristics among a group of users. Since many users may have common interests up to a point during their navigation, navigation patterns should capture the overlapping interests or the information needs of these users. In addition, navigation patterns should also be capable to distinguish among web pages based on their different significance to each pattern. In the web usage mining systems, the large majority of methods that have been used for navigation pattern mining from Web data are clustering methods. Clustering aims to divide a data set into groups that are very different from each other and whose members are very similar to each other. In this paper, a partitioning method is used for clustering of user's navigation patterns. Expectation maximization (EM) is a clustering algorithm that works based on partitioning methods. The EM is a memory efficient and easy to implement algorithm, with a profound probabilistic background.

Expectation maximization (EM) is a well-known algorithm used for clustering in the context of mixture models. This method estimates missing parameters of probabilistic models. Generally, this is an optimization approach, which had given some initial approximation of the cluster parameters, iteratively performs two steps: first, the expectation step computes the values expected for the cluster probabilities, and second, the maximization step computes the distribution parameters and their likelihood given the data. It iterates until the parameters being optimized reach a fix point or until the log-likelihood function, which measures the quality of clustering, reaches its maximum. To simplify the discussion we first briefly describe the EM algorithm. The algorithm is similar to the K-means procedure in that a set of parameters are re-computed until a desired convergence value is achieved. The parameters are re-computed until a desired convergence value is achieved. The finite mixtures model assumes all attributes to be independent random variables. A mixture is a set of N probability distributions where each distribution

represents a cluster. An individual instance is assigned a probability that it would have a certain set of attribute values given it was a member of a specific cluster. In the simplest case N=2, the probability distributes are assumed to be normal and data instances consist of a single real-valued attribute. Using the scenario, the job of the algorithm is to determine the value of five parameters, specifically:

1. The mean and standard deviation for cluster 1
2. The mean and standard deviation for cluster 2
3. The sampling probability P for cluster 1 (the probability for cluster 2 is 1-P)

And the general procedure states as follow:

1. Guess initial values for the five parameters.
2. Use the probability density function for a normal distribution to compute the cluster probability for each instance. In the case of a single independent variable with mean μ and standard deviation σ, the formula is:

$$f(x) = \frac{1}{(\sqrt{2\pi}\sigma)e^{\frac{-(x-\mu)^2}{2\sigma^2}}}$$

In the two-cluster case, we will have the two probability distribution formulas each having differing mean and standard deviation values.

3. Use the probability scores to re-estimate the five parameters.
4. Return to Step 2.

The algorithm terminates when a formula that measures cluster quality no longer shows significant increases.

4. Proposed Model

Clustering is the process of organizing objects into groups whose members are similar in some way. It can be considered the most important unsupervised learning problem which deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Hard clustering is the techniques in which any pattern can be in only one cluster at any time.

Soft clustering is the technique which permits patterns to be in more than one cluster at any time. There are various clustering approaches that can be applied to cluster the biomedical keywords extracted from full text articles, some of them are k-means, k-median, Hierarchical Clustering Algorithm, Nearest Neighbor Algorithm etc. Here we are using modified fuzzy C mean clustering algorithm.

Here the proposed algorithm is responsible for extracting keywords present in the full text biomedical article store these keywords in a relation. Then the actual work of algorithm begins, it starts clustering of keywords.

The algorithm initially picks some keywords that are extracted. It groups the full text articles based on these keywords. It means each cluster contains only those articles which contain that keyword as their part. Then it starts using fuzzy C mean clustering to combine the clusters together on some similarity measure. Here we combine two clusters if their similarity measure is greater than or equal to a specified threshold value.

The proposed Algorithm repeats this process until no more changes are made to the clusters. Finally the proposed algorithm stores all the clusters in an xml file. Here our motive to extract all the full text articles which may be relevant for the user providing the search string, for this out of all clusters the cluster with largest number of articles is our target.

5. Proposed Algorithm

The proposed algorithm will take a complete list of all the biomedical articles and the output will be the XML files containing the clusters created using fuzzy c mean algorithm on keywords.

Input: List of full text biomedical articles.

Output: XML files containing the created clusters.

Algorithm

1. Read the next article in the list of biomedical text
2. Read the full text article
3. Extract the keywords from the article using KEA algorithm
4. Refer to the biomedical lexicon and discard the irrelevant keywords
5. Put the data in following relation so that the full text can be retrieved later using keywords only.

Article UID	Article Name	Keywords	Full text	Source

6. Go to step 1 and repeat till all the articles in the list of biomedical articles are processed.
7. Apply navigation pattern mining using Expectation maximization.
8. Store a text file of all clusters formed.
9. Determine the cluster having maximum item count.
10. Cache all the URLS found in the cluster containing the maximum number of weblog entries.

Note: The relation created step 10 will be used at the time of retrieval. Whenever the biomedical database is searched for any word the cluster containing the matching keywords is returned. The respective full text and other details corresponding to the returned cluster can be retrieved using this relation.

6. Result

The experiments were performed on the test application developed in .Net 2.0. The database contains all the article entries populated manually from the web resources like "http://www.medilexicon.com" and few more, starting with letter 'A'.

The search was performed using the traditional keyword based search algorithm and compared with the proposed algorithm. The snapshot for asset of search results is shown in Table 1.

Given the same data for text extraction, the proposed algorithm seems to be retrieving approximately 69% more relevant search results than the keyword based searching. Figure 2 illustrates the improvement achieved using the proposed algorithm.

Table 1. Comparison of results using traditional and proposed algorithm

	List of matching articles found	
Search Keyword	Keyword based search	Proposed algorithm
abarognosis	42	71
abasia	23	39
abasia-astasia	34	57
abasic	32	54
abatement	42	71
abatic	5	8
abaxial	53	90
Abbé	43	73
Abbé condenser	44	74

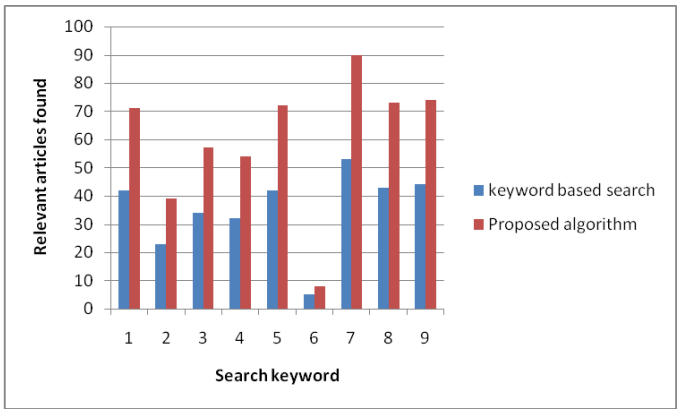

Figure 2: Improved text extraction using proposed algorithm

7. Conclusion

Extraction of text from biomedical literature is an essential operation. Given that there have been many text extraction methods developed; this paper presents a novel technique that employs keyword based article clustering to further enhance the text extraction process. The development of the proposed algorithm is of practical significance; however it is challenging to design a unified approach of text extraction that retrieves the relevant text articles more efficiently. The proposed algorithm, using data mining algorithm, seems to extract the text with contextual completeness in overall, individual and collective forms, making it able to significantly enhance the text extraction process from biomedical literature.

ACKNOWLEDGMENT

This research is supported by the Computer Science and Engineering department, SATI, Vidisha.

REFERENCES

[1] Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". Retrieved 2010-12-09.
[2] Berkhin, P., Survey of Clustering Data Mining Techniques. TechnicalReport, Accrue Software, San Jose, CA, 2002.
[3] Han, J., & Kamber, M., Data Mining Concepts and Techniques. CA : Morgan Kaufmann, 2001.
[4] Badgett RG: How to search for and evaluate medical evidence. Seminars in Medical Practice 1999, 2:8-14, 28.
[5] Richardson J: Building CAM databases: the challenges ahead. J Altern Complement Med 2002, 8:7-8.
[6] Miller, H. and Han, J., (eds.), 2001, Geographic Data Mining and Knowledge Discovery, (London: Taylor & Francis).
[7] Manu Aery, Naveen Ramamurthy, and Y. Alp Aslandogan. Topic identification of textual data. Technical report, The University of Texas at Arlington, 2003.
[8] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.
[9] Cecil Chua, Roger H.L. Chiang, and Ee-Peng Lim. An integrated data mining system to automate discovery of

measures of association. In Proceedings of the 33rd Hawaii International Conference on System Sciences, 2000.

[10] George Forman. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res., 3:1289-1305, 2003.

[11] Rayid Ghani. Combining labeled and unlabeled data for text classification with a large number of categories. In IEEE Conference on Data Mining, 2001.

[12] George Karypis and Eui-Hong Han. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical report TR-00-0016, University of Minnesota, 2000.

[13] Jerome Moore, Eui-Hong Han, Daniel Boley, Maria Gini, Robert Gross, Kyle Hastings, George Karypis, Vipin Kumar, and Bamshad Mobasher. Web page categorization and feature selection using association rule and principal component clustering. In 7th Workshop on Information Technologies and Systems, 1997.

[14] Sam Scott and Sam Matwin. Text classification using wordnet hypernyms. In Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, 1998.

[15] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.

[16] Andreas Weingessel, Martin Natter, and Kurt Hornik. Using independent component analysis for feature extraction and multivariate data projection, 1998.

[17] Robert Nisbet (2006) Data Mining Tools: Which One is Best for CRM? Part 1, Information Management Special Reports, January 2006.

[18] Dominique Haughton, Joel Deichmann, Abdolreza Eshghi, Selin Sayek, Nicholas Teebagy, & Heikki Topi (2003) A Review of Software Packages for Data Mining, The American Statistician, Vol. 57, No. 4, pp. 290–309.

[19] R. Agrawal et al., Fast discovery of association rules, in Advances in knowledge discovery and data mining pp. 307–328, MIT Press, 1996.

Authors Profile

Sumit Vashishta is a research scholar pursuing M.Tech in Computer Science & Engineering from Samrat Ashok Technological Institute Vidisha M.P India. He secured degree of B.E. in CSE (HONS) from Rajiv Gandhi Technical University, Bhopal (M.P.) India in 2006.

Dr. Yogendra Kumar Jain presently working as head of the department, Computer Science & Engineering at Samrat Ashok Technological Institute Vidisha M.P India. The degree of B.E. (Hons) secured in E&I from SATI Vidisha in 1991, M.E. (Hons) in Digital Tech. & Instrumentation from SGSITS, DAVV Indore(M.P), India in 1999. The Ph. D. degree has been awarded from Rajiv Gandhi Technical University, Bhopal (M.P.) India in 2010. Research Interest includes Image Processing, Image compression, Network Security, Watermarking, Data Mining. Published more than 40 Research papers in various Journals/Conferences, which include 15 research papers in International Journals.