Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm

Sumit Vashishta
Computer Science department
Samrat Ashok Technological Institute
Vidisha, M.P. INDIA

Dr. Yogendra Kumar Jain
Computer Science department
Samrat Ashok Technological Institute
Vidisha, M.P. 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.

I. 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?"

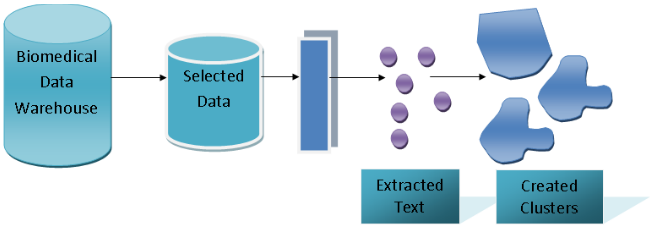


Figure 1: Text extraction from Biomedical literature base

II. 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 locations 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.

III. METHOD

Text mining is defined as the automatic discovery of new, previously unknown, information from unstructured textual data. This process is done in three steps: information retrieval, information extraction and data mining. A primary reason for using data mining for biomedical text is to assist in the analysis of collections of the available biomedical text. Biomedical data is vulnerable to co linearity because of unknown interrelations. The analysis in this paper will be augmented by using experiment-based approach.

Before data mining algorithms can be used, a target data set will be assembled. As data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns. Pre-process is essential to analyze the multivariate datasets before clustering or data mining. The target set is then cleaned. Cleaning removes the observations with noise and missing data.

The biomedical data available with us is first put into a data warehouse. Before putting the data in the data warehouse the keyword extraction algorithm is used to find out the keywords from the full text. This keyword extraction uses partial parser to extract entity names (gene, protein names etc). This parser uses linguistic rules and statistical disambiguity to achieve greater precision.

The data is then organized into clusters. Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. The clusters will be created based on the keywords extracted from our biomedical text. These clusters will be created using fuzzy C mean algorithm. The fuzzy c-means algorithm is one of the most widely used soft clustering algorithms. It is a variant of standard k-means algorithm that uses a soft membership function. Fuzzy C-Means (FCM) clustering algorithm is one of the most popular fuzzy clustering algorithms. FCM is based on minimization of the objective function Fm(u, c):

$$F_m(u,c) = \sum_{k-1}^{n} \sum_{i-1}^{c} (u_{ik})^m d^2(x_k, c_i)$$

FCM computes the membership *uij* and the cluster centers *cj* by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

where *m, the fuzzification factor* which is a weighting exponent on each fuzzy membership, is any real number greater than 1, *uij* is the degree of membership of *xi* in the cluster *j*, *xi* is the *i*th of d-dimensional measured data, *cj* is the dimension center of the cluster, *d2(xk,ci)*is a distance measure between object xk and cluster center ci, and ||*|| is any norm expressing the similarity between any measured data and the center.

The FCM algorithm involves the following steps:

1. Set values for c and *m*
2. Initial membership matrix U= [*uij*], which is U(0) (|i| = number of members, |j| = number of clusters)
3. At *k-step:* calculate the centroids for each cluster through equation (2) if k ≠ 0. (If k=0, initial centroids location by random)
4. For each member, calculate membership degree by equation (1) and store the information in U(k)
5. If the difference between U(k) and U(k+1) less than a certain threshold, then STOP; otherwise, return to step 3.

IV. 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.

V. Proposed Algorithm

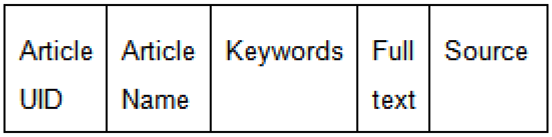
The proposed algorithm will take a complete list of all the biomedical articles and the output will be the XML files containing the clusters created using fuzzy c mean algorithm on keywords.

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. Use the fuzzy c-means algorithm to create clusters on keywords.
8. Save the article clusters in form of an XML file(containing articles IDs).

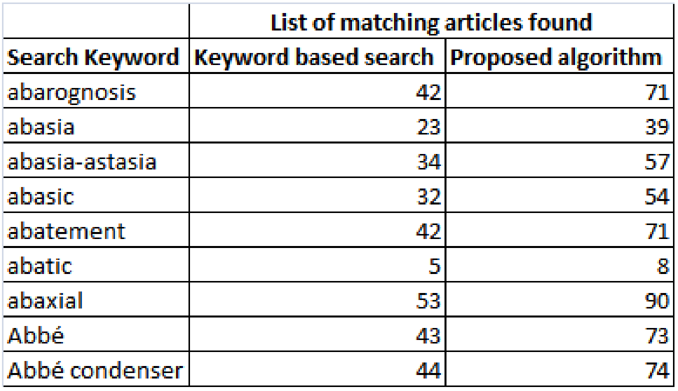
Note: The relation created step 6 will be used at the time of retrieval. Whenever the biomedical database is searched for any word the cluster containing the matching keywords is returned. The respective full text and other details corresponding to the returned cluster can be retrieved using this relation.

VI. 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 Figure 2.

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 3 illustrates the improvement achieved using the 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: Comparison of results using traditional and proposed algorithm

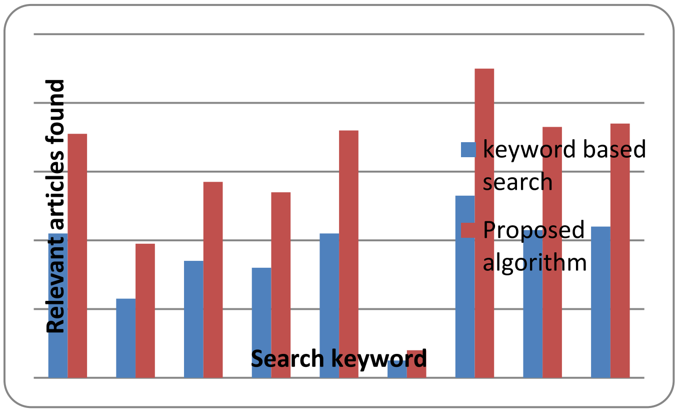


Figure 3: Improved text extraction using proposed algorithm

VII. 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, Vidhisha.

REFERENCES

[1] Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[2] Han, J., & Kamber, M., Data Mining Concepts and Techniques. CA : Morgan Kaufmann, 2001.

[3] Badgett RG: How to search for and evaluate medical evidence. Seminars in Medical Practice 1999, 2:8-14, 28.

[4] Richardson J: Building CAM databases: the challenges ahead. J Altern Complement Med 2002, 8:7-8.

[5] Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 0471228524. OCLC 50055336

[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.

[20] Kumar, V. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. *International Journal of Advanced Computer Science and Applications - IJACSA, 2*(3), 80-84.

[21] Jadhav, R. J. (2011). Churn Prediction in Telecommunication Using Data Mining Technology. *International Journal of Advanced Computer Science and Applications - IJACSA, 2*(2), 17-19.

[22] Devi, S. N. (2011). A study on Feature Selection Techniques in Bio-Informatics. *International Journal of Advanced Computer Science and Applications - IJACSA, 2*(1), 138-144.

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 IT from Rajiv Gandhi Technical University, Bhopal (M.P.) India in 2006.
E-mail-sumitvbpl@gmail.com

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.
Tel:+91-7592-250408,
E-mail: ykjain_p@yahoo.co.in