International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

An Improved Mining of Biomedical Data from Web Documents Using Clustering

Nikita Gupta[1], Gunjan Pahuja[2]

[1]Department of Computer Science, A.P.J. Abdul Kalam University, Lucknow, U.P., India

Abstract: ***Now a day's web is the main source of information in every field. Web is also expanding exponentially day by day. To get the relevant information is very time consuming and is not a very easy task. Mainly users go for the various search engines to search any information. But sometimes search engines are not able to give the useful results as most of the web documents are present in unstructured manner. Data mining is extraction of information from large database. Text mining uses the many techniques of data mining. In web, biomedical documents are also increasing at a very fast pace but most of them are unstructured text. These documents can be very helpful in diagnostics, treatment and prevention of any disease. There are we millions of documents on internet about a specific term so to obtain a relevant document is very difficult. The goal of this is to apply text mining techniques to retrieve useful biomedical web documents. Here a more efficient mechanism is proposed which uses the optimised K-means clustering algorithm where it can group the similar documents in one place. This approach will help the user to get all the relevant biomedical documents at one place. On comparing our approach with the original k-means algorithm and found that our algorithm on an average giving 99.06 % F-measure.***

Keywords: Data Mining, Biomedical Data, Clustering, K-means.

1. Introduction

Biomedical information search over search engines results in large number of query results. These results are not always relevant or sometimes fail to provide the information user looking for. Most of the explicit knowledge is stored in different types of documents but only a few people (often only the authors of the documents) know where to locate them. However, this information is often scattered among many web servers and hosts, using many different formats .The manual organization of these documents can be time-consuming to create and cannot usually be used in practice. Some websites use so many fake links to increase their importance. The famous search engine Google uses the ranking algorithm like PageRank [19] and HITS (Hypertext Induced Topic search) [20] to give the query results. Some web pages have too many links of advertisements, which may not relevant to the user. It is the user, and only the user who can say whether a given item of information is relevant to a query issued to a web search engine or to a private folder in which documents should be filtered according to the content.

Biomedical terms are different from the normal English language words [1]. So to get the domain specific information is very difficult now days. Recently Biomedical Information is very increasing very rapidly on web. Most of this information on web is stored in unstructured way. The need of automatically retrieval of useful knowledge from the huge amount of textual data in order to assist the human analysis is fully needed. Data Mining discovers the previously unknown and potentially useful information from data in databases. Text mining is the extension of data mining. Text mining has been one of the fastest growing research fields for the past few decades. There are various text mining techniques available named by Information retrieval, Information extraction, text classification, text Clustering and so on [3].

For application developers, this interest is mainly due to the enormously increasing need to handle larger and larger quantities of biomedical documents, a need emphasized by increased connectivity and availability of document bases of all types at all levels in the information chain. Biomedical text mining generally has five phases: Text Gathering, Text Preprocessing, Data Analysis, Visualization, and Evaluation [18]. Text gathering aims to get desired text on a certain topic; In text preprocessing tokenization, part of speech tagging stopwords removal etc is done ; In data analysis actual information extraction is done ; Visualization is how to visual the results obtained; At last evaluation is done. Clustering is an unsupervised technique used to discover new set of categories. Here the proposed system gives an optimized k-means algorithm to make cluster of the biomedical documents so that the documents would be more relevant and informative.

In this paper to identify the biomedical documents a UMLS (Unified Medical Language System) is used. This will conclude that if the document contains the biomedical term then the document is of biomedical domain. The proposed system finds out the biomedical terms in web documents and classifies the documents on the basis of it. Then the proposed system forms the cluster of similar and related documents using optimised K-means algorithm. Our approach will discard the non biomedical documents during classification of web documents.

2. Literature Survey

The Literature [16] describes that World Wide Web is the most worthful resources for information retrieval due to increase in amount of information available online. Web Mining technologies are the correct solution for discovering the relevant data. The paper provides introduction of Web mining as well as discuss the web mining categories.

Mostly search engines are used for obtaining the relevant information. Search Engines are the programs that search documents related to the queries keyword and returns the list of documents where the keyword found. Pooja et al. [4] describes that Google uses the two link-based ranking algorithms, PageRank [19] and HITS [20] with its advantages and disadvantages.

The literature [2] describes that the huge size of biomedical literature and its speedy growth in last few decades makes searching the relevant information a needy task. Obtaining and reading relevant information in literature is crucial for any researcher in life sciences. Fei et al. [5] discusses the text mining application in cancer research as cancer is a malignant disease and biomedical text has large value for its diagnostic, treatment and prevention. The paper also provides resources for cancer text mining and gives the general workflow of text mining in cancer system biology.

Ravi et al. [2] gives an efficient approach using NLP to get correct biomedical document from web by counting the number of biomedical terms in the documents with the help of UMLS ontology and finally ranking the documents according to them.

Wang et al. [11] proposed a similarity method for medical literature searches the normalised MEDLINE distance based on the biomedical literature knowledge source MEDLINE. This model is to refine the search process using user interests. User interests are analyzed to calculate semantic similarity among interest terms.

A new text mining algorithm to increase the performance of information retrieval system for Medline and Pubmed has been proposed by Sagar et al. [6]. The unique term are weighted by using novel global relevant weighing schema. The biomedical text retrieval is the techniques applied to biomedical resources to extract information required as published biomedical research is very large. Sumit et al. [9] proposed a technique of soft clustering data mining algorithm to increase the accuracy of biomedical text extraction by discovering the predictive relationships between the different pieces of extracted data.

Ontology is like dictionary or glossary but with more details like properties and interrelationships of the entities that exists for a particular domain. Tan et al. [12] showed that text mining of biomedical documents require substantial amount of domain knowledge. The paper proposes a model for choosing the most appropriate ontology for particular application.

3. Background

3.1 K-means Clustering Algorithm

K-means is a widely used popular partitional and unsupervised technique used for document clustering due to its simplicity [21]. K-means algorithm uses an iterative process in order to cluster database. If the algorithm is required to produce K clusters then there will be K initial means and K final means. If *K* is the desired number of clusters, then partitional approaches typically find all *K* clusters at once. K-means is based on the idea that a center point can represent a cluster.

K-means algorithm involves the following steps:

Input: Collection of HTML documents.

Output: Clusters of documents.

Algorithm

1) Load all documents.
2) Remove stopwords and html tags.
3) Set the value of k.
4) Select k points at random as cluster centers.
5) Calculate the cosine similarity based on TF/IDF for each documents.
6) Assign documents to their closest cluster center according to the cosine similarity.
7) Recalculate the new cluster center.
8) If no document was reassigned then stop, otherwise repeat from step 6.

3.2 UMLS

UMLS stands for the Unified Medical Language System is a collection of many controlled vocabularies in biomedical sciences [23]. UMLS is very large and contains terms related to biomedical and health. It is a wide range of general and specialized biomedical terminologies. It is used in variety of purposes like Information retrieval, Public Health Reporting, Clinical Health and Research. UMLS further provides facilities for natural language processing. It is intended to be used mainly by developers of system in medical informatics. The number of biomedical resources available to researchers is very large. Often the problem arises when the medical literature is searched as large volume of documents retrieved. The purpose of the UMLS is to enhance access to this literature by facilitating the development of computer systems that understand biomedical language.

3.3 Vector Space Model

In this work documents are represented using the vector-space model (VSM) for clustering algorithm. VSM is most widely used algebraic model for representing text documents as vector of identifiers. In this model, each document, *d*, is considered to be a vector, ***d***, in the term-space. Each document is represented by the (*TF*) vector, $d_{tf} = (tf1, tf2, \ldots, tfn)$, where *tfi* is the frequency of the *ith* term in the document. In addition, we use the version of this model that weights each term based on its *inverse document frequency* (IDF) in the document collection. The weight associated with each keyword determines the relevance of the keyword in the document. The similarity between two documents can be measured in various ways. There are a number of possible measures for computing the similarity between the documents. The most common one is the cosine measure, which is defined as:

Cosine-similarity $(\boldsymbol{d1}, \boldsymbol{d2}) = (\boldsymbol{d1} \bullet \boldsymbol{d2}) / \|\boldsymbol{d1}\| \|\boldsymbol{d2}\|$

Where • indicates the vector dot product and ||***d***|| is the length of vector ***d***. The strength of the similarity depends on the value of Θ. For K-means clustering, the cosine measure is used to compute which document centroid is closest to a given document.

4. Proposed Model

Following is our proposed system of mining the biomedical documents. Briefly describing the various phases as follows:

4.1 Document Collection

Firstly collect several number of HTML documents using search engines. Then the HTML documents are converted to simple text documents .Text Documents are used to perform the further process.

4.2 Documents Preprocessing

The collected documents are preprocessed to perform the effective documents search and to improve the efficiency of the system. There are abundant unnecessary texts on web documents like articles, be verbs, preposition, conjunction, and punctuations etc, which are categorized as stop words. There are large quantities of stop words appear in the document and typically have no use in this research. The stopword are removed and the documents are simplified.

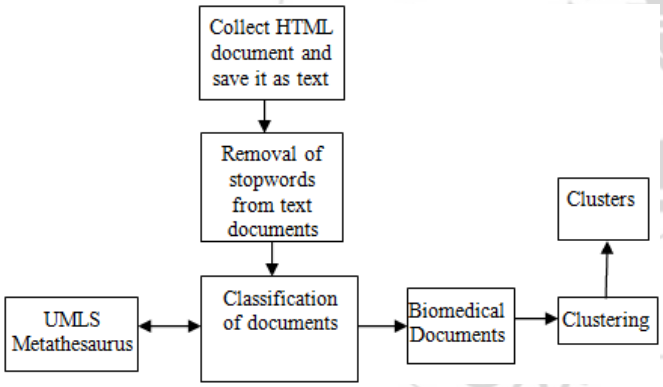

Figure1: Proposed System Arcitecture

4.3 Classification of Documents

Here in this phase the documents are classified into two categories i.e. biomedical documents and Non-biomedical Documents. This classification is done to extract the biomedical documents from all the documents. Non-biomedical documents are discarded. The classification of documents is done with the help of the UMLS.

4.4 Clustering of the Documents

The next step is to make cluster of similar and related documents. The improved K-means algorithm is used to make the clusters of biomedical documents.

5. Experimental Results

To obtain the similar and related biomedical documents we have collected approximately 100 web documents using Google search engine. After preprocessing, documents are classified into biomedical documents and non-biomedical documents as shown in fig.2.

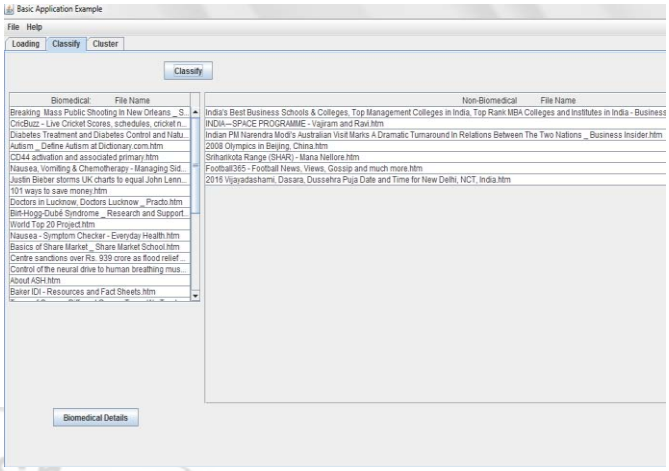

Figure 2: Classification of Documents

Non biomedical documents are discarded while Clusters of biomedical documents using K-means clustering. We tested the proposed system for both original K-means algorithm and improved K-means algorithm. We found that our improved approach takes less execution time. The comparisons of results are shown in fig.3.

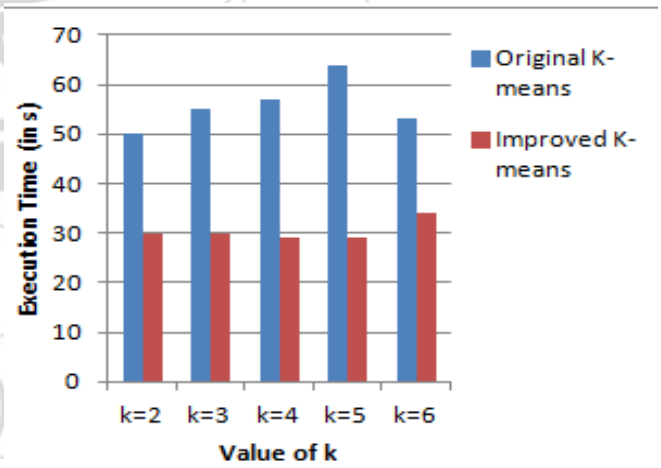

Figure 3: Showing Comparison of Execution time

Three performance measures precision, recall and F-measure have been calculated for both the approaches. The tested results of performance measures for demonstration are shown in table 1. The use of this is to obtain the similar documents in a group. So that finding the relevant information is not time consuming.

$$Precision = \frac{True\ Posiives}{(True\ Positives + False\ Negatives)}$$

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

$$F - Measure = \frac{2 * Recall * Precision}{Precision + Recall}$$

Table 1: Comparison of Performance Measures

Value of K	Original K-means			Improved K-means		
	Precision	Recall	F-measure	Precision	Recall	F-measure
2	0.796	0.783	0.789	0.998	0.979	0.988
3	0.805	0.786	0.796	0.998	0.990	0.994
4	0.772	0.761	0.767	0.998	0.987	0.992
5	0.789	0.778	0.783	0.998	0.981	0.989
6	0.779	0.772	0.776	0.998	0.983	0.990

6. Conclusion

The main aspect is retrieving the related biomedical documents in same group so that there is no need to go through all the documents. This paper presents improve the K-means algorithm so that it can perform clustering of biomedical documents in reasonable durations. It has been observed that selection of all words also affects the convergence time of K-means algorithm. The experimental results show that the execution time of the improved algorithm has become less or approximately half of the run time as in original. This shows a technique for selecting better initial points than random ones. We found that our approach is giving better performance results. This work can be further extended by ranking the documents in each cluster parallelization of K-means algorithm has been proposed as an improvement for the algorithm. In future we will consider the hierarchy of the biomedical documents in each cluster which will help the user to find all his documents in an organised and manner.

References

[1] Ravi Shankar Shukla, Kamendra Singh Yadav, Syed Tarif Rizvi, and Faisal Haseen, "An Efficient Mining of Biomedical Data from Hypertext Documents via NLP," Proc. of the 3rd Intell. Comput. (FICTA), Vol.1, pp. 651-658, 2014.

[2] Chung-Chi Huang and Zhiyong Lu, "Community challenges in biomedical text mining over 10 years: success, failure and the future," Briefings in Bioinformatics Advance Access, May 1, 2015.

[3] Dr. Shilpa Dang and Peerzada Hamid Ahmad," A Review of Text Mining Techniques Associated with Various Application Areas," International Journal of Science and Research (IJSR), Volume 4, no. 2, pp. 2461-2466, February, 2015.

[4] Pooja Devi1, Ashlesha Gupta, and Ashutosh Dixit," Comparative Study of HITS and PageRank Link based Ranking Algorithms," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, no. 2, February, 2014.

[5] Fei Zhu , Preecha Patumcharoenpol , Cheng Zhang , Yang Yang , Jonathan Chan , Asawin Meechai , Wanwipa Vongsangnak , and Bairong Shen, " Biomedical text mining and its applications in cancer research," Journal of Biomedical Informatics, Vol. 46, no. 2, pp. 200–211, 2013.

[6] S. Sagar Imambi and T. Sudha, "Extraction of Biomedical Information from MEDLINE Documents – A Text Mining Approach," International Journal of Science, Environment and Technology, Vol. 2, no. 2, pp. 267 – 274, April, 2013.

[7] Shally and Rejimol Robinson, "Survey for Mining Biomedical data from HTTP Documents," International journal of Engineering Sciences & Research Technology, pp.165-169, Feburary, 2013.

[8] Hui Yang and Yan Dong," Recognizing Hierarchically Related Biomedical Entities Using MeSH-Based Mapping," Tsinghua Science and Technology, Vol. 17, no. 6, pp. 609-618 December, 2012.

[9] Sumit Vashishta and Dr. Yogendra Kumar Jain," Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No. 4, pp. 77-80, 2011.

[10] Sazia Salah uddin and Rashedur M Rahman, "Mining Biomedical Data from Hypertext Documents, " Proc. of 14th Int. Conf. on Computer and Information Technology , ICCIT 2011, pp. 417-422, Dhaka, Bangladesh , December 22-24, 2011.

[11] WANG Yan, WANG Cong, ZENG Yi, HUANG Zhisheng, Vassil Momtchev, Bo Andersson, REN Xu ,and ZHONG Ning," Normalized MEDLINE Distance in Context-Aware Life Science Literature Searches", Tsinghua Science and Technology, Vol. 15, no. 6, pp.709-715, December, 2010.

[12] He Tan and Patrick Lambrix, "Selecting Ontology for Biomedical Text Mining," Workshop on BioNLP, pp. 55–62, Boulder, Colorado, June 2009, Association for Computational Linguistics, 2009.

[13] Saurav Sahay, Sougata Mukhaerjea, Eugene Agichtein, Ernest V. Garcia , Shamkant B. Navathe, and Ashwin Ram," Discovering Semantic Biomedical Relations Utilizing the Web," ACM-Transaction on knowledge Discovery from Data (TKDD), Vol. 2, no. 1, March, 2008.

[14] SougataMukherjea and Saurav Sahay, "Discovering Biomedical Relations Utilizing the World-Wide Web," Pacific Symposium on Biocomputing 11:164-175(2006).

[15] Rainer Malik and Arno Siebes, "CONAN: An Integrative System for Biomedical Literature Mining," Progress in Artificial Intelligence, 12th Portuguese Conference on Artificial Intelligence, EPIA 2005, pp. 248 – 259, Dec. 5-8, 2005.

[16] Miguel Gomes da Costa Junior and Zhiguo Gong," Web Structure Mining: An Introduction," Proc. of the 2005 IEEE Int. Conf. on Information Acquisition, pp. 590-595, Hong Kong and Macau, China, june 27-july 3, 2005.

[17] Wempu Xing and Ali Ghorbani, "Weighted PageRank Algorithm," Proc. of the second Annual Conference on Communication Networks and Services Research (CNSR'04), May 19-21, 2004.

[18] Brigitte Mathiak and Silke Eckstein, "Five Steps to Text Mining in Biomedical Literature," Proc. of the Second European Workshop on Data Mining and Text Mining in Bioimformatics, pp. 47-50, Pisa, Italy, September 24, 2004.

[19] Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry, "The Page Rank Citation Ranking: Bringing Order to the Web," Technical Report. Stanford InfoLab, 1999.
[20] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," J. ACM, Vol. 46, no. 5, pp. 604 -632, 1999.
[21] K. A. Abdul Nazeer and M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm," Proceedings of the World Congress on Engineering, WCE 2009, Vol.1, pp.308-312, London, U.K., July 1 - 3, 2009.
[22] H. Shatkay and R. Feldman, "Mining the Biomedical Literature in the Genomic Era: An Overview," Journal of Computational Biology (JCB), Vo.10, Issue: 6, pp. 821-855, July 5, 2004.
[23] Keith E. Cambell, Diane E. Oliver, and Edward H. Shortlife, "The Unified Medical Language System: Toward a collaborative Approach for Solving Terminologic problems," Journal of the American Medical Informatics Association, Vol.5, no.1, Jan / Feb, 1998.
[24] Https://www.nlm.nih.gov/research/umls/knowledge_sources/.

Author Profile

Nikita Gupta is a research scholar pursuing M.Tech in Computer Science & Engineering from JSS Academy of Technical Education, Noida, U.P., India. She has completed her B.tech in CSE from Mangalayatan University, Aligarh, U.P., India in 2012.