International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 01, January 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

# MINING OF BIOMEDICAL DATA: AN INTRODUCTION

**Nikita Gupta[1], Gunjan Pahuja[2]**

[1,2]*Dept. of Computer Science and Engineering, Dr. A.P.J. Abdul Kalam University, (India)*

## ABSTRACT

*World Wide Web is a huge deposit of medical information that includes text, audio, video etc. As the amount of biomedical information on web is increasing very rapidly, it is difficult to acquire the right information on web. The user usually depends on the search engines for finding the correct documents from web. As the search engine returns many pages in response to the query made by users. It is not possible for the user to go through all the resultant pages; so, the ranking algorithms are used by Search Engines to rank the resultant pages according to their relevance in context of the query made. Data Mining is non-trivial extraction of implicit, previously unknown and potentially useful information from data. In This Paper, We present some techniques use to mine the biomedical information from the web documents as the volume of biomedical research publications is increasing day by day. Biomedical text mining helps to get up to date biomedical information which can be required for further research or curing any harmful disease.*

***Keywords: Biomedical Information, Data Mining, HITS, PageRank, Text Mining, Weighted PageRank.***

## I. INTRODUCTION

Mining of Biomedical information becomes very major issue in last few years as the dynamic web is increasing exponentially. So to get a correct document from the thousands of available documents related to a subject is a very exhaustive and time consuming task. With the explosive growth in the volume of published biomedical research, it is very challenging to keep up to date with the underlying knowledge available [10]. Biomedical literatures are very much different from the other documents as biomedical words are not same as the normal English language words. Most of us generally use search engine like Google, Yahoo, and Bing etc to get the right and important information from World Wide Web [11]. Search Engines are the program that search documents from specified keywords and returns a list of documents where the keywords are found. These lines of results returned by search engine in the response of the query are referred to as the search engine resultant pages (SERPs).

In general query engine may return several hundreds and thousands of URL in response to the user query which includes a mixture of relevant and irrelevant information. Since, the user can read all the web pages returned by search engine. For the various searching methods almost two factors that distinguish the high quality page and the low quality page are relevance factor and the ranking factor. The relevance factor gives attention of the contents of web whereas the ranking factor gives the attention on the web structure not the content. Page

Ranking mechanism are used by search Engines for putting the important pages on top leaving less important at the bottom of the result list [3].

Broadly, web document ranking algorithm can be classified into two groups:

1) Content Based Ranking: The base of ranking web document is the text in it. The factors that influence it are give below:

i) The number of words matched with the query string.

ii) The frequency of the words i.e. the number of times the search string appears in the page. More the string appears in the page the better is the rank of the webpage.

iii) The location of words in the document i.e. query string should be found in the Heading of the page or in the heading paragraphs of the page or even near the head of the page.

2) Link or Connectivity Based Ranking: It works on the basis of link analysis technique. The view of the web as a directed graph where WebPages from the directed edges between these nodes. The two famous connectivity based methods are PageRank and Hyperlink Induced Topic Search (HITS).

## II. LITERATURE REVIEW

Manaswini et al. [4] presented study for classifying diabetic patients into two classes by using the powerful method is Artificial Neural Network (ANN) based classification model. For having better results Genetic Algorithm (GA) is used for feature selection and also to find out the number of neurons in single hidden layered model.

Carlos Ordonez, [12] in his work he shows that SQL implementation of the K- means algorithm works efficiently on the relational DBMS. It shows how to cluster large data sets defining and indexing tables to store and retrieve intermediate and final results, optimizing etc.,. The proposed K-Means implementation can cluster large data sets and exhibits linear scalability. The final implementation is a naïve translation of K means computation into SQL server as a framework to optimize the performance.

Data Clustering is one of the major task of data mining for exploring data and a technique for statistical data analysis used in many other fields like bioinformatics, pattern recognition etc.. Clustering is an unsupervised learning technique of grouping the data objects into group in such a way that objects in same group are more similar to each other than to those in other group. Clustering divides the dataset into similar and dissimilar dataset. MAFIA algorithm [2] is used to find out frequent patterns smoothly and efficiently than other algorithm. The main reason of this survey is too used by person for getting information about their disease, danger level, treatment etc.

A clustering algorithm partitions a data set into several groups based on the principle of maximizing the intra class similarity and minimizing the interclass similarity. K-means is a partitional unsupervised learning and iterative clustering algorithm in which data are moved among sets of clusters until the desired set is reached [3]. Within a cluster, a centroid represents a cluster, which is a mean point within cluster. Its main aim is to subset n observations into K clusters in which each observation belongs to the cluster with the nearest mean. The numerical attributes only works efficiently in K-means algorithm. The most popular clustering tool used in industrial and scientific applications is K-means algorithm.

HeTan [13] study that the domain knowledge required for biomedical text mining is present in the ontologies. Selecting an appropriate ontology that can integrate into the system rather than developing a new ontology from scratch is important for mining biomedical text. The Unified Medical Language System (UMLS) was designed by the National Library of Medicine [4]. It combines a large number of distinct terminologies into a single platform. The Unified Medical Language System (UMLS) was created and is maintained as a support for integration of biomedical textual annotations scattered in distinct databanks. There exist the numerous different controlled vocabularies especially of biomedical domain examples are SNOMED, SNOMED III, SNOMED-RT, MeSH etc. It allows translating a term among the various terminology systems and it may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts. The Metathesaurus is very large, Multi-purpose and Multi-lingual. The Metathesaurus contains information about the biomedical and health related concepts, their various names and associated codes with it. And also the relationship among them is defined by them.

## III. LINK BASED RANKING ALGORITHM REVIEW

### 3.1 PageRank Algorithm

Sergey Brin and Lawrence Page developed the algorithm known as PageRank while doing their PhD at Stanford University [14]. PageRank algorithm gives a more appropriate method to find the relevancy of the webpage. PageRank algorithm was used by one of the famous search engine Google to determine the importance of webpage. Functioning of PageRank algorithm depends upon the link structure of the webpage. In PageRank algorithm the rank of each page is determined individually not of a complete website. If the backlinks are from important page than that backlinks is given higher weighting than those coming from less important page. So, the link between one page to another page is considered as a vote. The vote is considered as an important aspect in ranking of page along with the importance of the one casting the vote.

The formula proposed by Page and Brin for computing the PageRank of an arbitrary page X (Page T1 to Tn pointing to it).

$$PR(X) = (1-d) + d\,(PR\,(T1)\,/C\,(T1) + \ldots.+ PR\,(Tn\,/C\,(Tn)) \qquad (1)$$

Here PR (Ti) is the PageRank of the Page Ti which links to Page X, C (Ti) is the number of outbound on page Ti and d is the damping factor. The value of damping factor lies between 0 to 1. PageRank of a page depends on the number of pages pointing to it.

### 3.2 Weighted PageRank Algorithm

The extension of the PageRank algorithm weighted PageRank (WPR) was given by Wempu Xing and Ali Ghorbani [15]. Weighted PageRank algorithm gives larger rank to more relevant pages instead of splitting the rank value of a page uniformly between its outlink pages. The value of each outlink page is proportional to its quality (i.e. the count of inlinks and outlinks). The quality of a page is defined in terms of number of inlinks and outlinks denoted by Win (r, s) and Wout (r, s) respectively.

Win (r, s) is the weight of link (r, s) computed on the basis of number of inlinks of page s and the number of inlinks of all citation pages of page r.

$$Win(r, s) = Iu / (\sum p \in R\,(v)\,Ip) \qquad (2)$$

Where Is and Ip shows the number of inlinks of page s and page p respectively. R(r) denotes the citation of page list page r.

Wout(r, s) is the weight of link (r, s) computed on the basis of number of outlinks of page s and the number of outlinks of all citation pages of page r.

$$Wout(r, s) = Ou / (\sum p \in R(v)\ Op) \qquad (3)$$

Where Os and Op shows the number of outlinks of page s and page p respectively. R(r) denotes the citation of page list page r.

Considering all the above parameters, the original PageRank formula will now be:

$$PR(u) = (1 - d) + d\sum_{v \in B(u)} PR(v)\, W^{in}_{(v,u)} W^{out}_{(v,u)} \qquad (4)$$

### 3.3 Hypertext Induced Topic Search

The WSM (Web Structure Mining) based algorithm developed by Kleinberg known as HITS (Hyperlink Induce Topic Search). In HITS, ranking of the webpage is done by the textual content against the given query. Kleinberg divides the webpage into two classes called Hubs and Authorities [6]. The set of Hub webpage contains links to relevant pages on that content and also too many authority pages. The Hub pages act as a resource list. A set of Authority pages contains the relevant content according to the query made by user. A page can be a Hub and a authority at the same time. The HITS algorithm treats world wide web as a directed graph G(V, E), where V is the vertices representing pages and E is the set of edges corresponding to the link. The implementation of HITS is mainly divided into two steps. The first step is sampling and the second step is iterative. In the sampling step, the relevant pages are collected for the given query. A sub graph of G is created which is high in authority pages. Here S is small and contains mainly the relevant and important pages. In the second step Hubs and Authorities are found using the output of sampling step using equations (5) and (6).

$$Hp = \sum Aq \quad \text{where } q \in I(p) \qquad (5)$$

$$Ap = \sum Hq \quad \text{where } q \in B(p) \qquad (6)$$

Where Hp and Ap is the Hub weight and Authority Weight respectively. I (p) and B (p) denotes the set of reference and referrer pages of page p.

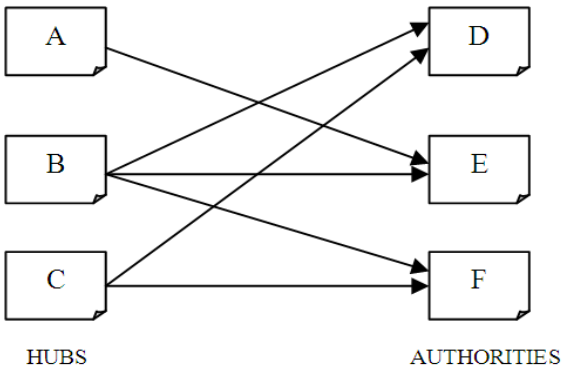


**Fig2: Hubs and Authorities [6]**

## IV. APPLICATION

**4.1** Medical records are written in different way from scientific articles, sequence annotations, or public health guidelines. Moreover, local dialects are not uncommon. Medical centers develop their own jargons and laboratories create their idiosyncratic protein nomenclatures. Text mining applications are tailored to specific types of text. In particular, for reasons of availability and cost, many are designed for scientific abstracts in English from Medline.

**4.2** Recently the amount of experimental data that is produced in biomedical research and the number of researches that are being published in this field have grown rapidly. In order to keep the developments up to date in their field of interest and to interpret the outcome of experiments in all available literature, researchers turn to use more and more to the use of automated text mining.

**4.3** Cancer is a malignant disease which involves abnormal cell growth which causes the great loss of human life. Biomedical text mining on cancer research is automatic, high throughput in nature and also error prone in nature. The information required for cancer treatment based on the type and stage of the disease, the size and place of the tumour and general health and medical history of the patients. In most cases, the main aim of the cancer treatment is to demolish or eliminate the cancer completely. If cancer is found and treated early, it can be cured.

## V. CONCLUSION

The goal of this research is to get the more appropriate biomedical documents by using some of the data mining techniques. According to the papers studied, sometimes the ranking algorithms used by search engines also not able to retrieve the relevant web document. Link based ranking algorithm gives importance to the links rather than the content of the webpage. Researchers can use the biomedical text mining techniques to get the relevant information. Biomedical text mining is greatly helpful in curing the harmful diseases. To fully utilise text mining, still there is a need to develop new tools and methods for highly complex text.

## REFERENCES

[1] 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.

[2] Manaswini Pradhan, Dr. Ranjit Kumar Sahu, "Predict the onset of diabetes disease using Artificial Neural Network ( ANN)," International Journal of Computer Science & Emerging Technologies ( EISSN: 2044-6004), pp.303-311, Vol. 2, No. 2, April, 2011.

[3] Kawsar Ahmed, Tasnuba Jesmin and Md. Zamilur Rahman, " Early Prevention and Detection of Skin Cancer Risk using Data Mining," International Journal of Computer Applications , Vol. 62, No. 4, January, 2013.

[4] Amandeep Kaur Mann and Navneet Kaur, "Survey Paper on Clustering Techniques," International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 2, No. 4, April, 2013.

[5] 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.

[6] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," J. ACM, Vol. 46, no. 5, pp. 604 -632, 1999.

[7] S. Chakrabarti ; B.Dom; D.Gibson; J. Kleinberg; R. Kumar; P. Raghavan; S. Rajagopalan,; A. Tomkins : Mining the Link Structure of the World Wide Web, IEEE Computer, Vol. 32, pp. 60-67, 1999.

[8] Taher H. Haveliwala,: Topic-Sensitive Page Rank: A Context-Sensitive Ranking Algorithms for Web Search", IEEE Transactions on Knowledge and Data Engineering Vol.15, No 4 July/August 2003.

[9] 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.

[10] 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.

[11] 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.

[12] Carlos Ordonez," Programming the K –means Clustering Algorithm in SQL," Proc. ACM Int'1 Conf. Knowledge Discovery and Data Mining, pp. 823-828, 2004.

[13] 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.

[14] 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.

[15] 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.

[16] 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.