Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 5 – No.10, September 2016 – www.caeaccess.org

Mining Health Data using Weighted Approach

P. Priyanga
Assistant Professor,
CSE Dept.,
K.S. Institute of Technology
Bangalore-560109

Naveen N. C., PhD
Professor, HOD, ISE Dept.,
Dayananda Sagar College of Engineering
Bangalore-560078.

ABSTRACT
Web Analytics (WA) is a vital area in the field of Data Mining (DM) that works with the principle of extracting interesting information or knowledge from the World Wide Web. WA is the measurement, collection, analysis and reporting of Internet data. The research in WA has led to the development of new techniques to generate automated topic hierarchies and web dictionaries. WA plays a major role in health care domain, to search health related information required from the web. Gathering knowledge about health has become a complex procedure for the majority of users. This confuses the users and consuming more time in overloaded data that continue to enlarge. Applications of DM to Web-page ranking helps Web search engines to find high quality web pages. In this paper Machine Learning (ML) methods for extracting knowledge from the large medical data on the Internet which is heterogeneous in nature of the web is proposed. The main objective is to develop a fast and efficient algorithm for real-time processing of big data and create knowledge out of the existing information in the web.

Keywords
Web Analytics; Health care; Big Data; Machine Learning

1. INTRODUCTION
World Wide Web data in medical area is rapidly increasing day by day that is large, diverse and dynamic. Traditionally, the healthcare industry needs improvement in using big data. People are looking for more medical information in Web. The available search engines cannot provide accurate information for medical related terms all the time because they do not consider its special requirements.

Managing and retrieving medical information are the main parts of health care domain. The huge access of the Web has drastically changed the way people receiving medical information. Every day, more Indians search for medical information on the Web rather than visiting doctors [1]. Doctors are also using web to facilitate diagnosis because of the difficulty in keeping up with the rapid development of medical knowledge [2]. [3] Around 79% of Internet users have used search engines for medical related information on the Web. Most of these users believe that get valuable information online, and were more willing to use Web search engines rather than going to a particular health-related Web site. Research reveals that more than 50% of these users would resort to the Web first for their next health question. All the information on the Web is not accurate, but still most doctors and patients believe that access to such online resources is beneficial. Before visiting doctor people would like to access medical information in Web to prepare better and also to understand information provided by doctor. Due to lack of doctors and increased number of patients, the appointment time with doctors keep increasing day by day, and this is expected to continue in the future as well.

Kosala et. al [5] has described the following problems when users access the web to retrieve information.

1. **Accessing Relevant Information** - Users browse or use the search engine to find specific information on the web. Currently available search tools have low precision that is due to irrelevance of the search results that makes users difficult to find the relevant information.
2. **Creating knowledge out of the information available on the web** - This problem is a data triggered process that assumes that users already have collection of web data and they want to extract potentially useful knowledge out of it which is a real challenge.
3. **Personalization of information** - When users interact with the web depending on the preferences they differ in the way how the contents and presentations are used.

There are several tools and techniques available to solve the above stated problems. Web Mining (WM), Information Retrieval (IR), and Natural Language Processing (NLP) could be more efficiently used to solve these problems. DM is the computational process of discovering patterns in large data sets that involves Artificial Intelligence (AI), ML, statistics and database system. DM aims at extracting information from a data set and transforms it into an understandable format for further use.

When user try to browse the web to identify information about a specific disease or health symptoms, the accuracy of the information can be improved by changing ranking system which is available in the web. This research work proposes re-ranking based on keywords. All the relevant key words for the specific search are considered when users try to search for a medical issue.

Figure 1 depicts the Web Analytics process which contains different stages i.e. collecting web data from different data set, preprocessing, store data in the required format, validate using ML techniques to improve the accuracy and retrieve the useful knowledge.

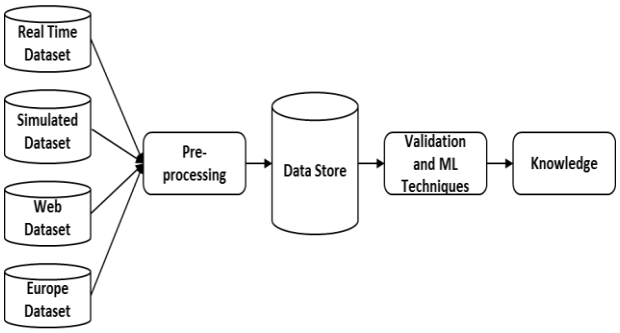

Figure 1. Web Analytics Process

Searching medical information in the Web has many unique needs that differentiate this from other Web search. In most of the cases a person searches for medical details to know about the diseases or symptoms. During this the searcher usually prefers to understand all kinds of information that is related to his search. The available medical Web search engines are optimized for precision and concentrate their search results on a few keywords [6]. This lack-of-diversity in using various keywords related to the topic which user wants to search [7, 8, 9], will not give accurate search results as expected.

In this research paper obesity, heart disease, arthritis and thyroid datasets are used as samples to evaluate the web search engine re-ranking framework. Further application of ML methods for extracting knowledge from the data on the Internet which is heterogeneous in nature of the web is proposed.

2. RELATED WORK

Web Analytics plays major role in Health care domain. Web search related to lab measurement, disease, diagnosis, symptoms, and chemical compound details of the drugs prescribed is growing rapidly. Across the globe it helps people to give basic understanding of health related problems.

Etzioni [11] proposed "Web mining" concept to extract information from huge data on the World Wide Web.

Kosala and Blockeel [5], suggest disintegrating web mining into the following sub tasks: Resource discovery, Information selection and Pre-Processing, Generalization, Analysis, Visualization.

Web Usage Mining(WUM) was first suggested by Chen et al. [12], Mannila and Toivonen [13] and Yan et al. [14]. Baraglia and Palmerini offered a WUM system called SUGGEST which optimizes the Web server performance by providing useful information.

Jasleen Kaur Bains [15] presented a wide insight on how the various Big Data analytics initiatives can improve healthcare worldwide. The paper also explains the different phases involved in Big Data Analytics (BDA) process and depicts its benefits and challenges with focus on healthcare industry.

Dalia Abdul Hadi AbdulAmeer [17] discusses big data term in medical sector and the way to analyze clinical big datasets that can discover knowledge to use in clinical prediction. Deployment and the new trends with big data and two paradigms on building real data infrastructure for future studies in on-line health care applications are also presented. The proposed framework figure-out how to analyze big data and how to discover knowledge from the extracted information. Big data analysis is expected to reveal the knowledge structure which can guide in better decision making.

Gang Luo, Chunqiang Tang et.al [18], presented MedSearch, a specialized medical web search engine for medical information retrieval. It help users to search information throughout the entire process of medical treatment. Medical search requires unique requirements and this design takes into consideration and supports queries written in plain English. It also accepts long queries, provides diversified search results, and suggests related medical phrases with proper ranking and annotation. Users with very little medical knowledge can have great advantage of these features and also helps to understand unfamiliar medical terminology. In addition, MedSearch can process long queries at a speed comparable to that of traditional web search engines in processing short queries.

Mayank Trivedi [19] described many search engines which are useful for health care professionals. The paper discusses the problem of inflation of medical information which medical community has been facing because the Internet is now the major worldwide medical communication tool. Improved search engines are required to avoid wasting time and get accuracy with efficiency. Although the complete list is not available related to search engines as number of search engines are added on the web daily.

Shubham Borikar et.al [20] proposed an overview of big data analytics, different technologies that are used in big data and its impact on healthcare domain to give few useful predictions based upon analyzing a variety of datasets. A model is provided which can be used for predictive analytics using DM and ML algorithms to predict the chances for a person to be prone to a specific disease.

Priyanka K et.al [21] proposed a brief introduction about how they can uncover additional value from health information used in health care organizations using a new information management approach called as big data analytics. Including big data analytics in health care sector gives stakeholders with new insights that have the potential to advance personalized care, improve patient outcomes and avoid unnecessary costs. They defined big data analytics and its characteristics, comments on its advantages and challenges in health care.

Gemson Andrew Ebenezer J and Durga S [22] reviewed the various big data analytics platforms and algorithms with its challenges. Although medical diagnoses applications use different algorithms they proved C5.0 approach produced more accuracy with less space requirement when volume of data is increased to millions or billions. It also has lower error rate and minimizes the predictive error. In case of big data, the C5.0 algorithm works much faster and provides the better accuracy with less memory consumption.

In response to this demanding and huge market need, Healthline [18], a popular Web search engine for medical information, came into existence in the year 2005. Shortly thereafter, Google announced its own medical Web search engine, Google Health [23], in May 2006. There were several other medical Web search engines [8, 25, 26]. While these systems have their own advantages, they mostly treat medical search in much the same way as traditional Web search.

Bharati Suvalka et.al [16] proposed to lay the foundations of the next generation of domain agnostic ML techniques which will be able to sum up the knowledge of ML successes across

domains in analytics framework. Their goal is to alter the difficult experimentation involved in using ML techniques that take years to master into a simple skill which will be easily adapted by practitioners across fields. They referred this science as "Meta-Machine Learning".

A ML algorithm's generalization capability depends on the dataset [27] which is why engineering a dataset's features to represent the data in a noticeable structure is important. However, feature engineering requires domain knowledge to generate appropriate features.

The objective of these papers in this field is to advance web services performance through the improvement of websites, contents, structure, presentation, and delivery.

RE-RANK FRAMEWORK

Re-rank framework automatically converts and stores huge set of text in a web database. Applications on the web extract this data and process in the database in a range of highly flexible tools.

The Re-Rank framework consists of four sections, which is used to process the web content to build the profile. The important module includes sentence detection, relevant sentence extraction, tokenizer, part-of-speech tagger and re-ranking of the websites. Figure 2 illustrates the Re-ranking framework of websites.

Like open NLP the sentence detector can detect whether a punctuation character marks the end of a sentence or not. In this case a sentence is mentioned as the white space trimmed character sequence. The first non-whitespace character is assumed to be the beginning of a sentence, and the last character (non-whitespace) is assumed to be a sentence end. Relevant sentence extraction module is used to detect sentences using Apache OpenNLP to check for relevance to the search key. These relevant sentences are then stored as an array of strings.

The OpenNLP tokenizer segments an input character sequence into tokens. Tokens are usually words, punctuation, numbers, etc. The Part of Speech (POS) tagger mark tokens with their corresponding word type based on the token itself and the context of the token. A token might have multiple POS tags depending on the token and the context. The OpenNLP POS tagger uses a probability model to predict the correct POS tag out of the tag set. In this implementation, nouns and adjectives are extracted for further computations.

Using the data obtained from previous steps, two sets of tables are created to store word frequency: The master table, which contains word frequency of all websites combined, and a website table which contains word frequency of the corresponding website. Each table is sorted in descending order of word counts and top 10 relevant words from each website are chosen and the website's weight is incremented corresponding to the word count. The search results are re-ranked according to the weights of the websites.

3. WEIGHTED RE-RANKING ALGORITHM

Although all the websites have been grouped based on their similar properties, it is not to understand which disease contains the useful information. DBSCAN algorithm used to rank the webpage by using the re-ranking algorithm. Disease names varies from web page to webpage, and the webpage containing content can be any of those.

Fortunately, most of the webpages stored some keywords in tags in order to accommodate web crawlers. So extracted the description of an article by parsing the tags in the webpage. The description usually contains a brief summary or just the few words of the article content. With that, it is able to calculate the similarity between each webpage and description using the weighted re-ranking algorithm. This algorithm calculating the weight the more likely is the webpage to contain the content text.

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 5 – No.10, September 2016 – www.caeaccess.org

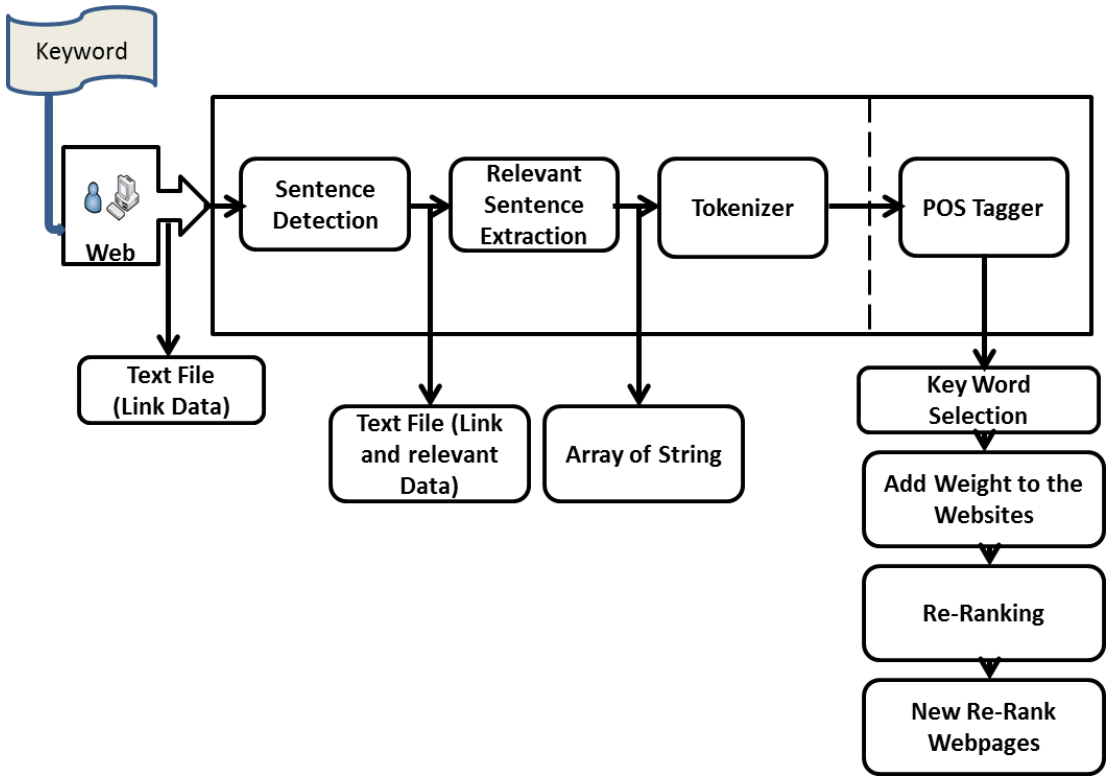

Figure 2. Re-Rank Framework

At first, tried to automatically order the websites by finding the rank to each webpage on a website. Then found that on some rare pages, the best webpage according to that website description is comment instead of the main text. Algorithm 1 calculates the weight across the entire website for re-ranking.

weight := total number of words in a website;

for each disease i of a website do

for each website j under that do

weight [i] := weight[i] * new rank of

website j;

end

end

Pick the disease(D) with the website highest similarity weight;

Label all the website of in the same disease D as 1;

Algorithm 1: Weight the google webpage with weighted method for re-ranking

This algorithm performs very accurately and removes the unwanted keywords and providing the new ranking to the websites which is most useful to the users.

4. EXPERIMENTAL RESULTS

The result shows that, the content of the diseases across multiple webpages, has multiple keywords. These keywords count are collected for this experiment. Other common elements, like similar sidebar, advertisement, unimportant words, etc., are also successfully removed. The algorithm is quite effective in discriminating the important keywords on the webpages.

Table 1: For Heart Disease, old and new Ranking for different websites

Web Pages	Google Rank	New Rank	%
www.heart.org	1	4	100%
www.mayoclinic.org	2	7	
www.medicalnewstoday.com	3	2	
www.bhf.org.uk	4	5	
www.medicinenet.com	5	3	
en.wikipedia.org	6	1	
www.webmed.com	7	6	

Table 2: **For Thyroid, old and new Ranking for different websites**

Web Pages	Google Rank	New Rank	%
www.webmed.com	1	5	83%
www.medicinenet.com	2	6	
www.womenshealth.gov	3	3	
en.wikipedia.org	4	1	
www.healthline.com	5	2	
www.endocrineweb.com	6	4	

Table 1: For Arthritis, old and new Ranking for different websites

Web Pages	Google Rank	New Rank	%
www.arthritis.org	1	5	100%
www.medicinenet.com	2	3	
en.wikipedia.org	3	1	
www.medicalnewstoday.com	4	2	
www.niams.nih.gov	5	6	
www.nlm.nih.gov	6	7	
www.healthline.com	7	4	

5. CONCLUSION

This research paper, to develop a frame work for Heart, thyroid and Arthritis diseases. This model collects the keywords from various webpages. This frame work is using the weighted re-ranking algorithm to find the new ranks against the existing search engine ranking which is useful for the users. Weighted Re-ranking algorithm can achieve perfect weighting for the websites by using several keywords. This work have limited up to seven google ranking websites to find the new ranking. This re-ranking method for these web sites have brought the good level of improvement. Accuracy also measured for the diseases like heart disease 100%, thyroid 83% and Arthritis 100% by using the algorithm.

Further work can be done to improve the user's usage of the search quality of the diseases by using Machine Learning techniques. Further research can be focused to increase the google ranking for more web pages. In future the proposed method needs to be compared with existing methods.

6. ACKNOWLEDGMENTS

The Authors would like to thank VGST (Vision Group on Science and Technology), Government of Karnataka, India for providing infrastructure facilities through the K-FIST Level I project at KSIT, CSE R&D Department.

7. REFERENCES

[1] C. Sherman. Curing Medical Information Disorder. http://searchenginewatch.com/showPage.html?page=3556491, 2005.

[2] 'Googling' Aids Difficult Diagnoses. http://www.e-healthinsider.com/news/item.cfm?ID=2258, 2006.

[3] M. Klein, H. Easley. Checking Medical Facts Online can be OK, but don't Become a 'Cyberchondriac'. The Journal News, June 26, 2006. http://www.thejournalnews.com/apps/pbcs.dll/article?AID=/2006062 6/NEWS03/606260311/1019.

[4] Healthline homepage. http://www.healthline.com.

[5] Kosala and Blockeel, " Web mining research: a survey, " SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol.2, 2000.

[6] A.Z. Broder. Identifying and Filtering Near-Duplicate Documents. CPM 2000: 1-10.

[7] J.G. Carbonell, J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. SIGIR 1998: 335-336.

[8] C. Zhai, W.W. Cohen, and J.D. Lafferty. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. SIGIR 2003: 10-17.

[9] B. Zhang, H. Li, and Y. Liu et al. Improving Web Search Results Using Affinity Graph. SIGIR 2005: 504-511.

[10] The National Coalition on Health Care. Facts on the Cost of Health Care. http://www.nchc.org/facts/2006%20Fact%20Sheets/Cost%20%202006.pdf, 2006.

[11] O. etzioni, The world wield web: Quagmire or Gold Mining. Communicate of the ACM, (39)11:65-68, 1996.

[12] Chen, M., Park, J. and Yu, P. "Efficient data mining for path traversal patterns," in IEEE Transactions on knowledge and data engineering, Volume 10, No.2, March/April 1998, 209-221.

[13] Mannila, H. and Toivonen, H. "Discovering generalized episodes using minimal occurrences," in International Conference on Knowledge and Data Mining, 1996, 146-151.

[14] Yan. T., Jacobsen. M, Garcia-Molina. H, Dayal. U, "From user access patterns to dynamic hypertext linking," in International World Wide Web conference on Computer networks and ISDN systems, Volume 28 Issue 7-11, Pages 1007-1014, 1996.

[15] Jasleen Kaur Bains, "Big Data Analytics in Healthcare-Its Benefits, Phases and Challenges," in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 4, April 2016. ISSN: 2277 128X

[16] Bharati Suvalka ,Sarika kandelwal , Sidharth Singh Sisodia, "Big Data Analytics using Meta Machine Learning," International Journal of Innovative Research in Science, Engineering and Technology , Vol. 3, Issue 8, August 2014. ISSN: 2319-8753

[17] Dalia AbdulHadi AbdulAmeer, "Medical Data Mining: Health Care Knowledge Discovery Framework Based On Clinical Big Data Analysis," International Journal of Scientific and Research Publications, Volume 5, Issue 7, July 2015. ISSN 2250-3153.

[18] Gang Luo, Chunqiang Tang, Hao Yang, Xing Wei, "MedSearch: A Specialized Search Engine for Medical Information Retrieval", CIKM'08, October 26–30, 2008.

[19] Mayank Trivedi, "A study of search engines for health sciences", International Journal of Library and Information Science Vol. 1(5) pp. 069-073 October, 2009.

[20] Shubham Borikar, Mohan Bhagchandani, Raunak Kochar, Ketansing Pardeshi, Manisha Gahirwal, " A Survey on Applications of Big Data Analytics in Healthcare" , International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-5 Issue-5, November 2015.

[21] Priyanka K, Prof Nagarathna Kulennavar, " A Survey On Big Data Analytics In Health Care", International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5865-5868. ISSN 0975-9646.

[22] Gemson Andrew Ebenezer J and Durga S , " Big data analytics in healthcare: a survey", ARPN Journal of Engineering and Applied Sciences , V0l. 10, No. 8, May 2015. ISSN 1819-6608.

[23] Google Health homepage. http://www.google.com/Top/Health. Curbside.MD homepage. http://www.curbside.md, 2008.

[24] SearchMedica - The GPs search engine. www.searchmedica.co.uk/searchmedica/EUIHomeAction.do, 2006.

[25] Medstory homepage. http://www.medstory.com

[26] Erik Cambria, Guang-Bin Huang, "Extreme Learning Machines," Published by the IEEE Computer Society, Nov/Dec 2013.