EFFICIENT ALGORITHM FOR MINING ON BIO MEDICAL DATA FOR RANKING THE WEB PAGES

Ponmani E, Suresh K S, Martina M, Ananthakrishnan S
School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India

ABSTRACT

Information in the internet is evolving in terms of high volume through different sources. Extracting tuples from HTML pages has been an important issue in various web applications such as web data integration, e-commerce market monitoring, and mash ups that repurpose and selectively combine existing web data services. Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information. Text Mining uses many applications of Data Mining. Text Mining is the discovery of unknown information by automatically extracting and relating the information from different resources. Text is classified based on the content that is used for mining. It is done based on comparing the text documents with the database. In the existing system, techniques like named entity recognition, information retrieval, information extraction and knowledge discovery are used for text mining. Google used page rank method to retrieve and rank the documents. However, Google rank may not provide the documents with the most relevant information. In the proposed system, information retrieval is used to collect many web documents and pre-processing the web documents and extract the text data. Then a word is identified as bio medical entity or not by using a Database with medical keywords. The page containing more bio medical words is ranked first. More relevant documents can be obtained by re ranking the documents using medical database.

Keywords: Algorithm, Bio-Medical Data, Ranking


1. INTRODUCTION

The existing tuple extraction system utilizes spatial relationships among elements. There are some drawbacks using these spatial relationships. This system is also a semi-automated system where user interference is required. The main objective of our paper is to extract the tuples present in the real time HTML documents like Flipkart, Amazon etc., by using a concept called Web scraping and to provide the necessary information to the user.
scraping is a computer software technique of extracting information from websites using the DOM structure. Web scraping is closely related to web indexing, which indexes information on the web using a web crawler and is a universal technique adopted by most search engines. In contrast, web scraping focuses more on the transformation of unstructured data on the web, typically in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. Web scraping is also related to web automation, which simulates human browsing using computer software. So, finally our method could develop an application in which user, instead of going to many websites and analyzing the product's cost, they can just check this interface and make a decision. Our method is designed in such a way that it is a user friendly one and its advantage over the existing techniques is to show the results in a less time comparatively. Rather than sending the requests sequentially, the script is designed to send the requests in parallel and thereby saving a lot of time.

2. RELATED WORK
The internet search will be through Google and other search engines to find useful information. Web contains many documents regarding a particular topic. Among those documents finding the useful document is a difficult task. Medical documents are a bit different from the normal documents. It is known that Google is a very popular search engine. Google uses Page Rank method to retrieve and rank the documents. The drawback of this ranking by using hyperlinks is that it may not provide us proper documents that contain useful information. In this paper, it ranked the documents so that the documents with useful information are ranked first. Database containing biomedical terms is used to rank the documents. The documents with more number of biomedical terms are ranked first. Thus the ranking of documents is based on the count of biomedical terms. Ranks obtained from the page rank algorithm used by Google and based on count of biomedical words are compared. Text Mining is the discovery of unknown information by automatically extracting and relating the information from different resources. Text is classified based on the content mining. It is done based on comparing the text documents with the database. Text Mining involves a number of different phases. Information Retrieval includes retrieving the textual resources for a particular area for mining. The retrieved document can be analysed based on classification by comparing with the database. Some specific comparison methods are used to extract the information from the huge Data content. This method includes extracting and ranking the already ranked documents.

3. PROPOSED METHODOLOGY

3.1 PAGE RANK ALGORITHM
Page Rank Algorithm is used by Google search engine to rank websites. This algorithm measures the importance of each website among many other websites. Parameters that page rank algorithm consider are number of qualitative links. Page rank algorithm gives weight age to each and every hyperlink before ranking the websites. Page rank algorithm uses Jenkins hash method in ranking the documents. Jenkins hash functions are a collection of multi byte keys. Page Rank algorithm considers incoming links while ranking of WebPages. Those incoming links may contain advertisements that are fake and no way related to the query given by the user. Those incoming links that contain fake advertisements can mislead the ranking of algorithm. Page Rank algorithm is based on the web graph. It takes web pages as nodes and hyperlinks as edges. The page Rank of a page is defined recursively and depends on the number and page rank metric of all pages that link to it.
A page that is linked to by many pages with high page rank receives a high rank itself. Thesis rank is generated by preprocessing in which a stop word removal procedure is applied.

3.2 RANKING BASED ON THESIS RANK ALGORITHM

The algorithm used in this paper is an extension to the existing page rank algorithm. The proposed algorithm is working based on the model in which it is mainly focusing on removing STOP words.

3.2 STEPS IN THESIS RANK ALGORITHM

Page Rank algorithm uses Jenkins Hash method to rank the documents. A code that is synced to the page rank button in the GUI runs and the rank is given for different URL s. Generally Google gives page rank in the range of 0-10 based on the relevant content in the document. Higher the rank of a page, the page is more informative. User collects the documents regarding a particular disease and the URLs are given in the GUI. A code in the backend runs...
for collecting the HTML documents and these HTML documents are converted into the text documents.

![Diagram](https://via.placeholder.com/150)

**Figure 2 Steps in Ranking**

The Converted text documents are stored in the local system. Pre-processing of a document includes removing the stop words from the text documents. To preprocess the documents, a code runs in the backend. The List of stop words in the English language are taken and included in the code. The preprocessing is done to get more accurate information from the documents. After preprocessing the documents are again stored in the local system. *The biomedical words related to a particular disease are stored in a database. A code in the backend runs in which pre-processed text documents are compared with the database and classification is done. This classification separates the biomedical and non-biomedical words. After classifying the text documents the total count of biomedical words in each document is calculated and the document with highest count is given rank 1. Accordingly all the documents are classified and ranked. This is called re-ranking. In GUI a button named thesis rank is present which is linked to this code. Finally Ranks are displayed to the user based on the count. The URL which has the lowest rank will contain more relevant information.*

**5. RESULTS AND ANALYSIS**

The thesis rank algorithm is implemented with sample URL of bio medical data. The output of the algorithm is depicted in the following figures 3,4,5,6.

![Figure 3](https://via.placeholder.com/150)

**Figure 3 Taking URLs as input**

![Figure 4](https://via.placeholder.com/150)

**Figure 4 URLs entered by user**
3. Page rank

Figure 5 Ranking after classification

Figure 6 Comparative analysis

6. CONCLUSION

The goal of this paper is to apply comparison methods to retrieve more important biomedical web documents. The rank given to the URLs by the thesis algorithm after comparing with the bio-medical database has more relevant information than the order of priority given by page rank algorithm. Hence the users of the search engine are benefited by getting the web pages in the first page that has more accurate information.

REFERENCES


Efficient Algorithm for Mining on Bio Medical Data for Ranking the Web Pages


[14] Mr. Shinde Santaji Krishna, Dr. Shashank Dattatraya Joshi, Unsupervised Approach to Deduce Schema & Extract Data from Template Web Pages, International Journal of Computer Engineering and Technology (IJCET), Volume 5, Issue 11, November (2014), pp. 57-64