Biomedical Research 2017; Special Issue: *ISSN 0970-938X*
www.biomedres.info

A computational model to analyze medical web pages for optimized search results.

Sethuraman J*, Manikandan R, Sekar KR

School of Computing, SASTRA University, Thanjavur, India

Abstract

With the phenomenal growth of World Wide Web and the huge number of web sites with so many web pages, finding relevant information becomes quite difficult for the web users. Web site promoters also need to ensure that their web site is on the top of the search results for better market share. Finding the right combination of keywords and placing them at the appropriate places may help in bringing the web site on the top of search results when a search engine searches the web sites and creates indexes. Over the years web site optimization research has reported encouraging results. Despite promising results, data mining techniques for web sites classification have hardly been applied. An appropriate implementation of classification of web pages may help designing the web pages accordingly. Here in this work data mining using R's web scraping technique has been used for performing extraction of web pages and also pair of java tools called "website" and "webscrap" were developed to automatically get pages and for further analysis of the downloaded pages. This is the first attempt to analyze medical related nomenclature using.

Keywords: Search engine optimization, Diseases.

Accepted on July 20, 2017

Introduction

With the popularity of World Wide Web and the every business needs their presence of web sites and web based applications. Search engines bring out lot of web pages in response to users query. Web developers develop lot of web sites rich in content but inappropriate placements of information result in pages not appearing in Search engine Result pages. This study focuses on web sites and web applications classifications depending on the medical terms which form the contents in it. Web sites classification is very important for the future development of web sites. With the right classification web developers can emulate web site development and in turn bring the web site on top of the search engine results.

Literature Review

The main objective of this section is to provide a complete review of different classification algorithms and also techniques which were applied on different web pages containing medical terms. The accomplishment rate of classification algorithms on different web pages is also discussed [1]. A study has been done on the role of the prominence of a name in search results [2].Trend surveillance has been attempted using an algorithm for better search results.

Classification algorithms

The classification algorithms comparison is given in Table 1

Table 1. *Classification algorithms comparative table.*

Algorithm	Result
Decision Trees	takes More CPU time
Naïve Bayes	used in text classification
Support Vector Machine	More number of iterations

Existing work

To improve the accuracy and efficiency of the classification process data cleaning was done by smoothing process by filling up missing values for attributes like heading, title etc. Average imputation scores over other techniques like common-point imputation. Öztürk et al. presented a new defect clustering method using k-means++ and got good performance on large-data sets [1]. Mavridis et al. developed a crawler using search engine APIs and evaluated results against established metrics [2]. Fang et al. proposed an adaptive trend surveillance method based on frequency of terms used in the document in their work [3]. Mavridis et al. work probes on semantic analysis of web content which used a mechanism which employs latent dirichlet allocation for the semantic analysis of web conent [4]. Aizpurua et al. has examined the relationship between UX attributes and web accessibility based on

perception and conformance to guidelines [5]. Fdez-Glez et al. has developed a novel web spam filtering framework to take advantage of multiple classification schemes and algorithms [6]. Egri et al. mainly focused on measuring the significance of time, speed, reducing bounce rate etc. [7]. Tamah has focused on a tool to compare the suggestions of different search engines for the same search query [8]. Jansen et al. proposes a new framework to facilitate comparison of search results and also implications for the design of Web information retrieval systems are analyzed [9]. Mehta has analyzed the search engine optimization trends in the year 2015 [10].

Relevance analysis like correlation analysis and attribute subset selection was done to remove some unnecessary attributes thereby improving the performance of classification process. The originally 35 attributes were extracted through an online tool. Out of which 9 attributes were found to be redundant. They were removed from the original set through relevancy analysis. The actual time spent on relevance analysis was 0.05 seconds for a data set of 10000 records. This is quite small when compared to the time it would have taken for the original data set with 35 attributes. Search engine was used with a keyword "the most common diseases in India". The result showed that there are thirteen communicable diseases quite common in India. This list was obtained by a site called yourarticlelibrary.com which was the first site appeared in the Google search. To confirm with the list top ten search results were used and all the sites have the following list of diseases in common.

Table 2. *The thirteen communicable diseases quite common in India.*

Sr. No.	Disease	Keyword
1	Malaria	Malaria treatment in Chennai
2	Typhoid	Typhoid symptoms
3	Hepatitis	Hepatitis vaccination centers in Chennai
4	Jaundice	How to prevent jaundice
5	Leptospirosis	What is leptospirosis
6	Diarrhoeal diseases	Treatment for diarrhoeal diseases in Chennai
7	Amoebiasis	What causes amoebiasis
8	Cholera	How to avoid cholera
9	Brucellosis	How brucellosis spreads
10	Hookworm Infection	What is hookworm Infection
11	Influenza	Indication of Influenza
12	Filariasis	Filariasis meaning
13	Tuberculosis.	Is Tuberculosis curable

The above list of diseases were used to form different keywords as shown in Table 2 which were used in to grab a list of web sites using our own tool call Grab which was written in java. The result was about 20 web search results in each category of keywords. Those results were used one by one in R tool with a script written to extract the content in each page. Some pages were not extracted by RCurl tool. After the application of every keyword in Table 2 top 10 web site listings were generated for example for example "what is leptospirosis" was applied and the Google result pages produced the following output. Likewise every keyword was applied and thirteen tables of data were generated. A list of web sites was obtained. The actual data obtained and the full code is shown in www.tanjoregoogle.com. Every web site name was taken and the following R code snippet was applied to perform web scraping.

The necessary package like RCurl, tm were installed in R

Require ("Curl")

mydata<-getURL ("http://www.chp.gov.hk/en/content/9/24/3056.html",ssl.verifypeer=FALSE)

class (mydata)

Architectural Diagram

The architectural diagram is as shown in Figure 1.

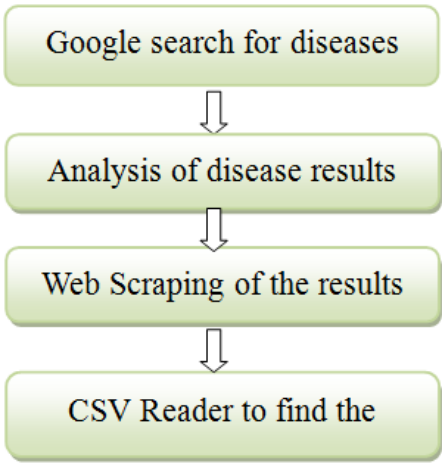

Figure 1. *Architectural Diagram.*

Frequency detection algorithm

Step 1: Find the most common diseases in India through search engine where a is the set of words and are part of medical terms

{a | T (a)}

T is a function which filters medical terms

An → ∞

Step 2: Relevant to every disease form keywords to fetch appropriate web pages Kj where j ranges from 1 to m

Step 3: Use web scraping to fetch the contents of every web page

Step 4: A new parser was developed to fetch web pages to identify keywords with most frequency

An application of classification algorithms of SVM (Support Vector machine) and NaiveBayes both predicted an accuracy of upto 73.33%.

The following code snippet is by using R for statistical analysis.

SVM

predsvm=predict (svm_model, mydata)

predsvm

1-1 TO 5, 2-1 TO 5, 3-1 TO 5, 4-1 TO 5, 5-1 TO 5, 6-5 TO 10, 7-5 TO 10, 8-5 TO 10, 9-5 TO 10, 10-10 TO 15, 11-10 TO 15, 12-10 TO 15, 13-10 TO 15, 14-10 TO 15, 15-10 TO 15

Levels: 1 TO 5, 10 TO 15, 5 TO 10

NaiveBayes

library("e1071")

pred_model=predict (bayes_model, mydata)

pred_model=predict (bayes_model,mydata)

pred_model

1-1 TO 5, 2-1 TO 5, 3-5 TO 10, 4-1 TO 5, 5-1 TO 5, 6-5 TO 10, 7-5 TO 10, 8-5 TO 10, 9-5 TO 10, 10-5 TO 10, 11-10 TO 15, 12-10 TO 15, 13-10 TO 15, 14-10 TO 15, 15-10 TO 15

Levels: 1 TO 5 10 TO 15 5 TO 10

Results and Conclusion

The results obtained clearly shows that the frequency of terms alone is not enough to bring a page to the top of search engine result pages. Other metrics like back links, social network presence and the factors like likes in the social media also play role in bringing the web page to the top. Future work can be extended by analyzing the impact of every domain characteristic in bringing the web page to the top.

References

1. Öztürk MM, Cavusoglu U, Zengin A. A novel defect prediction method for web pages using k-means++. Expert Syst Appl 2015.
2. Mavridis T, Symeonidis AL. Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms. Eng Appl Artific Intell 2015; 41: 75-91.
3. Fang ZH, Chen CC. A novel trend surveillance system using the information from web search engines. Decision Support Syst 2016; 88: 85-97.
4. Mavridis T, Symeonidis AL. Semantic analysis of web documents for the generation of optimal content. J Eng Appl Aritific Intell 2014; 35: 114-130.
5. Aizpurua A, Harper S, Vigo M. Exploring the relationship between web accessibility and user experience. Int J Human-Comput Studies 2016; 91: 13-23.
6. Fdez-Glez J, Ruano-Ordas D, Méndez JR, Fdez-Riverola F, Laza R, Pavón R. A dynamic model for integrating simple web spam classification techniques. Expert Syst Appl 2015; 42: 7969-7978.
7. Egria G, Bayrak C. The role of search engine optimization on keeping the user on the site. Procedia Comput Sci 2017; 36: 335-342.
8. Al-Shammari ET. Towards search engine optimization: Feedback collation. Procedia Computer Science 2015; 62: 395-402.
9. Jansen BJ, Pooch U. A review of web searching studies and a framework for future research. J Assoc Informat Sci Technol 2001; 52: 235-246.
10. Mehta S. Search engine optimization trends in 2015. Int J Sci Eng Res 205; 6: 548-549.

***Correspondence to**

Sethuraman J

School of Computing

SASTRA University

India