DATA MINING TECHNIQUES AND THEIR ROLES IN INCREASING THE SPEED OF SEARCH IN WEB DATA

[1]Ms. Dhanashree Kuthe, [2]Mr.P.Srinivas, [3]N.Venkateswaran
[1,2,3]Associate Professor
[1,2]Trinity College of Engineering and Technology, Karimnagar, Telangana, India.
[3]Jyothishmathi Institute of Technology And Science, Karimnagar, Telangana, India.

Abstract : The web is the source of all information in the different form. To get the appropriate contents by searching the data on the web is a challenging task. Data mining techniques can help to increase the speed of the search from billions of web pages over the internet. Search engines help to do the proper search for the user. Search engine optimization using web mining is an important aspect to increase the search speed. To make the information more accuracy in the web, different data mining techniques proposed. In this paper, various methods are compared, and the best data mining techniques for search optimization recommended. Results show that the performance of search en improved with the accuracy of finding the information in less time.

Index Term: Data Mining, web search, web data, data mining algorithm, Web Mining

1. INTRODUCTION

The web search is a computer system that provides information retrieval service with the demand of Internet users to quickly query information after the internet is produced [2-4]. His is like an information processing system, with a specific strategy, finding and understanding the data, extraction, organization and processing, and retrieval services for users, for navigation information. Fast information retrieval from the internet or billions of web pages is related to the mining of significant data to get useful results as per the user's query. Now, this web mining is the field of Artificial Intelligence where the use of Machine Learning, Natural Language Processing, Database Query, Graph theory, Optimization algorithms s used. One can use the combination of technologies mentioned above for improving the search results. Now a day’s search through image is also available. So the fields of Computer Vision is also applicable in finding the information related to that particular image.

In this paper, were mainly focused on textual query web data. The rapid development of the internet assures thousands of web pages added every day on the World Wide Web (WWW). So the task of finding the proper and desired search results is challenging. Most of the search engines like Google search includes the page rank algorithm where the number and quality of links to a page have been counted to determine a rough estimate of how important the website is. The underlying assumption is that more essential sites are likely to receive more links from other sites. With the increase in the use of Data Mining Machine Learning algorithms for efficient information retrieval, most of the search engines like Google combined their page rank algorithm with the ML algorithms to increase the search speed. Search engines are used ML pattern recognition to identify duplicate and unrelated content and avoid low-quality content. ML helps to identify new ranking signals to improve the quality of search results. The search engine will learn about the specific user’s preferences and would base its information on previous queries to current the most exciting information possible. By combining the page rank algorithms with data mining techniques will increase the speed of search.

Search Engine Optimization based on page rank rules, website domain name, keywords in search query and structure and content of the website. Non-commercial domain name have high rank unlike commercial site with com extension. The selection and extraction of keywords from the search query is the most critical task for fast search results. The structure of the place like the hierarchical relationship between the pages of the site and the DOM (Document Object Model) is essential to retrieve the proper webpage.

1.1 DATA MINING

It is the technique for valuable, valid and useful information retrieval from the pool of astronomical data. Data can represent structured like data in tabular form in most cases or unstructured like web data. There are five steps for information retrieval from the data, Data selection, pre-processing, (feature extraction) transformation, data mining algorithm and interpretation and evaluation. The problem is to find a useful analysis from the web data which has e trained from historical data of the user's search patterns. There are many algorithms like regression, classification, and clustering which re supervised and unsupervised respectively. We are comparing different algorithms for web data mining.
In this paper, Section 2 provides the study of existing approaches and their features; Section 3 contains the proposed data mining approach with Results and further improvements in Section 4 and five respectively.

2. LITERATURE REVIEW

In [1] authors proposed the search engine optimization approach using Support Vector Machine (SVM) algorithm. It is a classification algorithm to categorize the categories of web search. Results show accurate identification of types of information, obtain the ideal in the search engine optimization to reduce working time.

In [5] authors, proposed an approach based on SVM, Naive Bayes and Decision tree classification algorithm to classify the query which has treated as a question. For feature selection, they have used CRF based on a probabilistic model that addresses features an observation sequence.

In [6] authors proposed synonyms based search based on the tree data structure.

The effectiveness of a website depends on the users visit. This survey motivation is to the analysis of the related work so that new insights can be determined to find a better prediction of optimized cumulative traffic or the visit [7]. In [7] authors study and analyze data mining and optimization techniques. This analysis can determine the associated optimization concerning a different domain.

To increase the effectiveness of any algorithm hybrid and cross-domain approach would be useful. In [8] authors proposed a cross-domain approach to reduce the occurrence of search errors using a combination of data mining and semantic modelling techniques.

Users frequently use search systems on the Web as well as online social media to learn about ongoing events and public opinion on personalities. Prior studies have shown that the top-ranked results returned by these search engines can shape user opinion about the topic (e.g., event or person) searched. In case of polarizing topics like politics, where multiple competing perspectives exist, the political bias in the top search results can play a significant role in shaping public opinion towards (or away from) certain aspects. Given the considerable impact that search bias can have on the user, we propose a generalizable search bias quantification framework that not only measures the political bias in ranked list output by the search system but also decouples the bias introduced by the different sources—input data and ranking system [9].

In [11] authors, find the relationship between tweets, retweets, and twitters of political tweets. They formulate political leaning inference as a convex optimization problem that incorporates two ideas: (a) users are consistent in their actions of tweeting and retweeting about political issues, and (b) similar users tend to be retweeted by the same audience.

The biggest challenge in web mining is to remove noisy data information or unwanted information from the web page such as banner, video, audio, images, hyperlinks, etc. which is not related to a user query [12]. In [12] authors proposed a technique for removal of noise from the web page like a banner, audio, image, etc. An URL pattern extraction algorithm extracts the all relevant index pages from the web according to the user's query. Noisy Data Cleaner (NDC) algorithm is applied to remove the unwanted content from the retrieved web pages.

3. PROPOSED APPROACH

Web mining is categorizing into content mining, structure mining and usage mining. In this paper, we are focussing mainly on web content mining. The query is extracted, and the result of that query classified into image, text, audio, and video.

3.1 Problem Statement

To extract web content using web content data mining and divide the web pages into different categories using classification algorithm.

3.2 Proposed Model

There are three parts of the proposed model. The first part is the user interface for entering the query from the user. The query is pre-processed, and important keywords have extracted from the query. The extracted query is then fed to the database containing the web pages. All the related web pages are then collected. Different classification algorithms are used to extract the relevant information from the related pages and display the results. The flow of the system is shown in figure 1.

3.3 Database Used

Each sequence in the dataset corresponds to page views of a user during those twenty hours. Each event in the sequence corresponds to a user's request for a page. The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "MSN-news", and "msn-sports". Any page requests served via a caching mechanism were not recorded in the server logs and, hence, not present in the data [14].

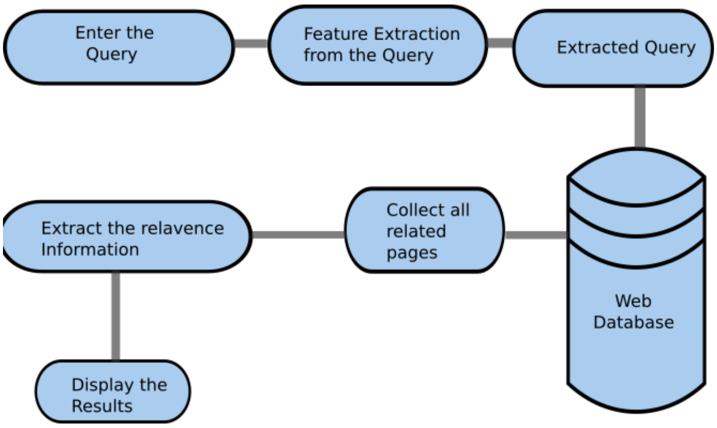

Enter the Query
Feature Extraction from the Query
Extracted Query
Web Database
Collect all related pages
Extract the relevance Information
Display the Results

Figure 3.1. System Flow

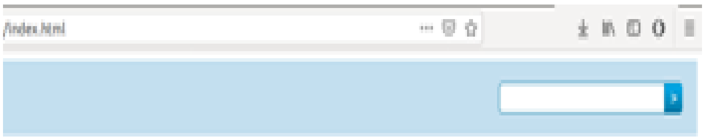
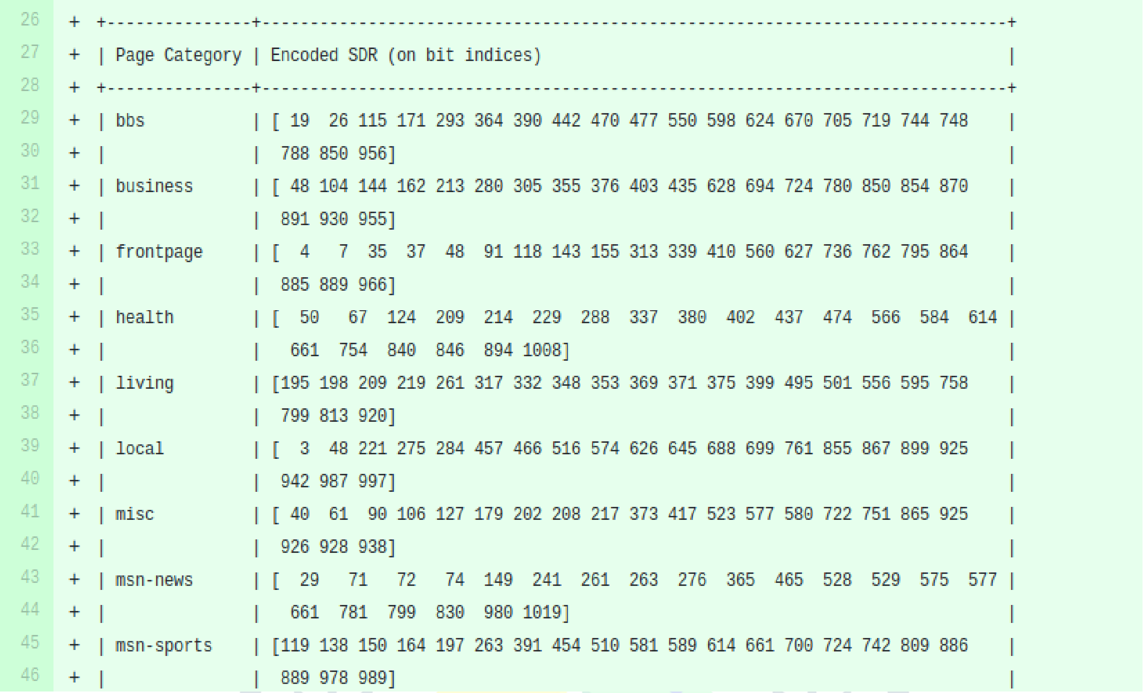


Figure 3.2. User Interface for Search

4. EXPERIMENTAL RESULTS

The results show the classification of web data into different categories.



```
+ +---------------+-----------------------------------------------------------------------------+
+ | Page Category | Encoded SDR (on bit indices)                                                |
+ +---------------+-----------------------------------------------------------------------------+
+ | bbs           | [ 19  26 115 171 293 364 390 442 470 477 550 598 624 670 705 719 744 748    |
+ |               |  788 850 956]                                                               |
+ | business      | [ 48 104 144 162 213 280 305 355 376 403 435 628 694 724 780 850 854 870    |
+ |               |  891 930 955]                                                               |
+ | frontpage     | [  4   7  35  37  48  91 118 143 155 313 339 410 560 627 736 762 795 864    |
+ |               |  885 889 966]                                                               |
+ | health        | [  50   67  124  209  214  229  288  337  380  402  437  474  566  584  614 |
+ |               |   661  754  840  846  894 1008]                                             |
+ | living        | [195 198 209 219 261 317 332 348 353 369 371 375 399 495 501 556 595 758    |
+ |               |  799 813 920]                                                               |
+ | local         | [  3  48 221 275 284 457 466 516 574 626 645 688 699 761 855 867 899 925    |
+ |               |  942 987 997]                                                               |
+ | misc          | [ 40  61  90 106 127 179 202 208 217 373 417 523 577 580 722 751 865 925    |
+ |               |  926 928 938]                                                               |
+ | msn-news      | [  29   71   72   74  149  241  261  263  276  365  465  528  529  575  577 |
+ |               |   661  781  799  830  980 1019]                                             |
+ | msn-sports    | [119 138 150 164 197 263 391 454 510 581 589 614 661 700 724 742 809 886    |
+ |               |  889 978 989]                                                               |
```

Results show different accuracy results for the separate classification algorithms.

S.N.	Algorithm Used	Classification Accuracy
1	Decision Tree	93 %
2	Random Forest	95 %
3	Naive Bayes	94 %
4	Support Vector Machine	93 %
5	KNN	88 %

Table 4.1 Classification Algorithm Accuracy

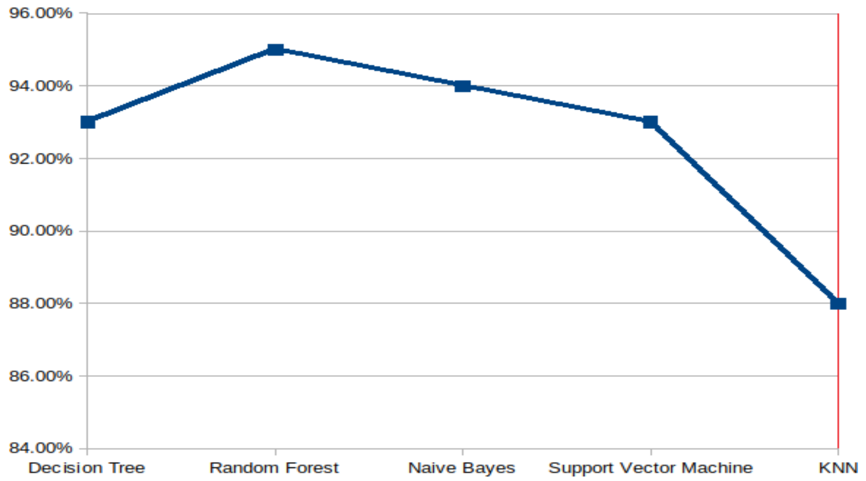


Fig. 4.1 Classification Algorithm Accuracy

5. CONCLUSION AND FUTURE WORK

Data mining techniques can help to increase the speed of the search from billions of web pages over the internet. Search engines help to do the proper search for the user. Search engine optimization using web mining is an important aspect to increase the search speed. o Improve the accuracy of finding information from the web, different data mining techniques proposed. In this paper, various methods are compared, and the best data mining techniques for search optimization suggested. Results show that the performance of search has improved with the accuracy of finding the information in less time. Clustering algorithm can e used for detecting the pattern from the web-based data.

6. REFERENCES

[1] Ronghua Chen, Research on the optimization strategy of web search engine based on data mining in Advances in Materials, Machinery, Electronics II, AIP Conference Proceedings Volume 1955, Issue 1, April 2018.

[2] Robert Cooley, Pang-Ning Tan, Jaideep Srivastava. Discovery of new usage patterns from web data [M]. Germany: Springer-Verlag, 2000.

[3] Andrea Garratt, Mike Jackson, Peter Burden, Jon Wallis. A Survey of Alternative Designs for a Search Engine Storage Structure [J]. Information and Software Technology, 2001, 43(11):661–677

[4] Pal S K, Talwar V, Mitra p. Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions[J]. IEEE Transactions on Neural Networks, 2002, 13(5):1163–1177.

[5] Gaurav Aggarwal Neha V. Sharma and Kavita, Enhancing Web Search Through Question Classifier in Proceedings of First International Conference on Smart System, Innovations and Computing pp 791-798, Part of the Smart Innovation, Systems and Technologies book series (SIST, volume 79), 2018.

[6] Palvi Arora Tarun Bhalla, A Synonym Based Approach of Data Mining in Search Engine Optimization in International Journal of Computer Trends and Technology (IJCTT) –volume 12number4–Jun 2014.

[7] Singh, Vinamrata; Patidar, Kailash; Sahu, Rajendra Prasad, A survey and analysis of page ranking through data mining and advanced techniques, International Journal of Advanced Technology and Engineering Exploration; Udaipur Vol.5, Iss.39, (Feb 2018): 37-42.

[8]Ekaterina Petrova, Pieter Pauwels Kjeld Svidt Rasmus Lund Jensen, In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modeling on Disparate Building Data, Advances in Informatics and Computing in Civil and Construction Engineering pp 19-26, 2018.

[9] Motahhare Eslami Johnnatan Messias Muhammad Bilal Zafar Saptarshi Ghosh, Search bias quantification: investigating political bias in social media and web search in Information Retrieval Journal pp 1-40, 2018.

[10] Van Couvering, E.(2010). Search engine bias: The structuration of traffic on the World-Wide Web. Ph.D. thesis, The London School of Economics and Political Science.

[11] Wong, F. M. F., Tan, C. W., Sen, S., & Chiang, M. (2016). Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering*, *28*, 2158.

[12] Pradeep Sahoo andRajagopalan Parthasarathy, An Efficient Web Search Engine for Noisy Free Information Retrieval, The International Arab Journal of Information Technology, Vol. 15, No. 3, May 2018.

[13] M. Álvarez, A. Pan, J. Raposo, F. Bellas, F. Cacheda, Extracting lists of data records from semi-structured web pages, Department of Information and Communication Technologies, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain, 11 Oct 2007.

[14] http://archive.ics.uci.edu/ml/datasets/msnbc.com+anonymous+web+data