ISSN:0975-9646

Performance Enhancement in Collaborative Filtering Technique by Removing Shilling Effect

Pooja, Max Bhatia

Department of Computer Science and Technology
Lovely Professional University, Phagwara

Abstract: Web page content mining is traditional searching of Web pages with the help of content, while Search results mining is a further search of pages found from a previous search. Web content mining has an approach that is shilling effect in which rating can be done and gives improper results. In this work we proposed an algorithm which gives appropriate values than shilling effects.

Keywords, web mining, visual methods, shilling effects

1. INTRODUCTION

Web content mining is a description and detection of useful data from the web contents/information/documents. There are two views on web content mining: view of data retrieval and data base [7]. The aim of Web content mining according to information retrieval based on content is to help the process of data filtering or finding data for the user which is usually performed based on extraction or demand of users, while according to the view of data bases it means attempt for modelling the information on web and its combination such that most of the expert needs for searching the data can be executed on different kind of data mode. Web content mining is used to analyze the content of Web pages as well as results of Web searching. The content may include text as well as graphics data [6]. Web content mining is divided into Web page content mining and search results mining. Web page content mining is traditional searching of Web pages with the help of content, while Search results mining is a further search of pages found from a previous search. We can use method as concept synonyms, hierarchies, user profiles and analyzing the links between pages to improve web content mining [12]. Web content mining uses data mining schemes for effectiveness, efficiency and scalability. Web content mining can be divided into:

- Agent-based approach
- Database based approach

Agent-based approach: Agents are software that performs the content mining. For example search engines. Intelligent search agents use techniques like user profiles or knowledge concerning specific domains. Data filtering utilizes IR techniques, knowledge of the link structures to retrieve a categorize documents. Personalize Web agents use information about user preferences to direct their search.

Database-based Approach: This approach views the Web data as related to a database. There have been techniques that view the Web as a multilevel database and there have been many query languages that target the Web.

2. REVIEW OF LITERATURE

S.Karthikeyan (2011) this paper [1] they proposed three different kind of algorithms for removing non-content blocks from the web pages. By Removal of non-informative content blocks from web pages can achieve a significant storage and time saving. Content-Extractor uses simple heuristics to detect primary content blocks. If there are web pages whose elements have the same style but different contents which are non-content blocks, then the algorithm would not be able to detect that. A style tree is an efficient technique to detect content blocks but constructing these style trees is a complex task. In Content-Extractor there is no overhead of constructing Style trees. This algorithm reduces storage requirements and provide fast search.

Ashish Jain (2013) they call the new method [2] as "Intelligent Search Method (ISM)". Their research is based on this new method. In this method they propose to index the web pages using an intelligent search strategy. It is very important to interpret the meaning of the search query and then index the web pages based on the interpretation. The new method can be integrated with any of the Page Ranking Algorithms to produce better and relevant search results. The different algorithms used for link analysis like PageRank (PR), Weighted PageRank (WPR), Hyperlink-Induced Topic Search (HITS) and CLEVER algorithms are discussed and compared in this paper. The aim of this research was to discover an efficient and better system for mining the web topology to identify authoritative web pages.

R. Etemadi N. Moghaddam (2008) they proposed a new algorithm [3] which has been represented to cluster web pages based on data content. The new method has been suggested based on the expressions and key words existing in web pages and their bit display a vector and using a newer similarity criterion obtained from Cosine and Jaccard similarity criterion. To evaluate the efficacy of proposed algorithm some pages with five subjects of architecture of computer, computerized networks, operating system, parallel processing and software engineering have been investigated and after preparing a suitable data bed the represented method has been simulated separately through two similarity criteria of Cosine and that of represented in this pager and has been evaluated using Dunn index. The results obtained from simulation represented high

efficiency of the proposed algorithm in separating web pages and their clustering. in most of the problems related to clustering web pages proposed algorithm can be used.

Swe Swe Nyein (2011) this research paper [4] they proposed an algorithm proposed that extract the main content from the web documents. This algorithm is based on Content Structure Tree (CST). Firstly the proposed model use HTML Parser to construct DOM (Document Object Model) tree from which construct Content Structure Tree (CST) which can easily segregate the main content blocks from the other one. The proposed technique then introduce cosine similarity analysis to evaluate which parts of the CST (Content Structure Tree)tree represent the less important and which parts shows the more important of the page. The proposed technique can define the ranking of the documents using similarity values and extracts the top ranked documents also as more relevant to the query.

Manne suneetha (2011 paper [5] aims at organizing the web search results into clusters facilitating quick browsing options to the browser providing an excellent interface to results precisely. Suffix tree clustering creates comparatively more informative and accurate grouped results. Image Repetition is a basic problem during image searching in any search engine. This can be minimized by using the L-Point Comparison algorithm a specially worked out method in field of Information Retrieval systems which is also discussed with a practical example. Search result clustering and Web mining as a whole have very promising perspectives; constantly improving Information storage and exchange media allow us to record almost anything. From our project, the user can easily access the results in the form of clusters so that he can browse the relevant contented cluster. Avoidance of repetition of images is an excellent we can further study to face the problem with almost every image search engine.

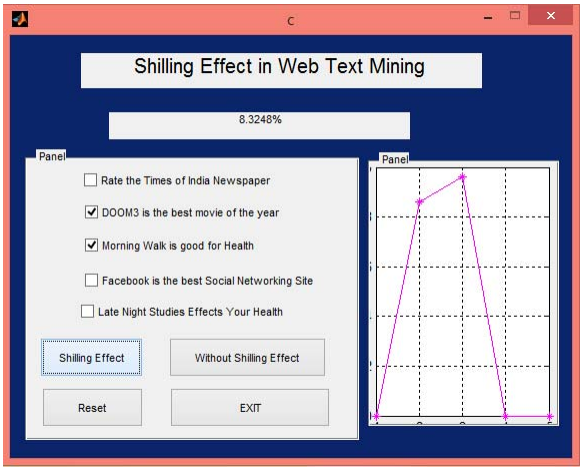
3. VISUAL METHODS

Visual approaches work the most similar to how a human segments a page, i.e. they operate on the rendered page itself as seen in a browser. They have thus the most information available but are also computationally the most expensive. They often divide the page into separators, such as lines, white-space and images, and content and build a content-structure out of this information [8]. They can take visual features such as background colour, font size and type and location on the page into account. The World Wide Web serves as a huge, widely distributed, global information service centre for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services. However, Internet web pages typically contain a large amount of non-informative content such as advertisements, search and filtering panel, headers, footers, navigation links, and copyright notices, etc. Most clients and end-users search for only the informative content and the need of Informative Content Extraction from web pages becomes evident[11] . To extract the informative content of the web page correctly, the informative content and the non-informative content of the web page must be known clearly. The informative content is the main content of the web page that gives some information to the user e.g., articles about technology, health or education, etc. The non-informative content of the web page contains fixed description noise such as site logos, copyright notices, privacy statements, etc., service noise, irrelevant services, such as the weather, stock or market index, etc., navigational links, advertisements, header, footer and so on. To distinguish between the informative and non-informative content in a web page, it needs to segment the web page into semantic blocks. There are several kinds of methods for Web Page Segmentation [10].

4. PROPOSED METHODOLOGY

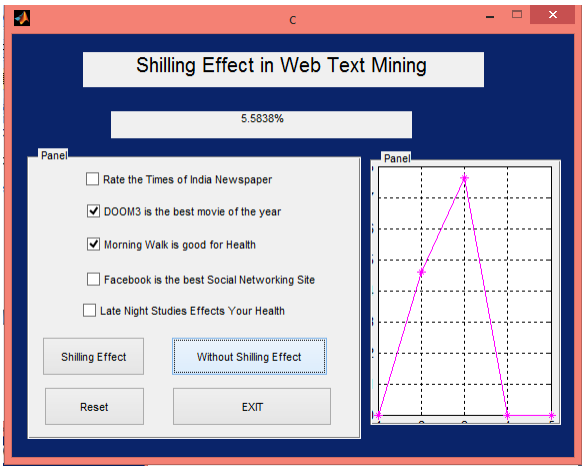
The huge data is available on the internet, to get the relative information from the available data, it is a challenging task. To gather the useful data from the rough data which is available on the web, the technique of web mining come into existence. The technique of web content mining, web structural mining comes under the web mining, to gather the relative content from the web, used technique is web content mining. Web personalization is a technique in which the user gets the information from the web in a relevant way. It is generally seen as a vital application in the field of data mining. Experts categorize the approaches of providing these personalized services into two. They are (1) Content based filtering and (2) Collaborative filtering. Content based filtering is purely based on the content of the information available in the web which mainly depends on the categories under which the products fall and the quality of the products in each category. For example, in an e-learning website, the learning products are classified and categorized based on the type and quality of information they contain. In contrast, the collaborative filtering approach entirely depends on the users. In classical collaborative filtering, the user's ratings and their similarity predictions are given more priority rather than the category and the quality of the information. The main drawback faced in the content based filter is that, extracting the attributes falling under the various categories becomes tedious. Thus, collaborative filter is generally preferred over content based filtering technique. However, collaborative filtering also has disadvantages. Collaborative filters suffer from a serious problem called the shilling effect. The term shilling effect refers to the improper guidance by the users. In some situations where anyone can give their reviews on a particular product, some people give excellent ratings to their own product or the product which they use. The same user may rate his competitors' product with minimal ratings. As a result, due to personal influence, the shilling effect directly or indirectly leads to gray sheep problem and cold-start problem. In this work, Novel technique will be proposed which solve the problem of shilling effect and enhance the reliability and efficiency of the collaborative filtering in web content mining.

5. RESULTS



Interface 1

The interface1 shown above is shown that overall rating of the application with shilling effect.



Interface 2

Interface2 shown in above in which the overall rating of the application is shown without shilling effect.
By comparing both the results it shows that the rating of application is better without shilling effects.

CONCLUSION

The main objective of this research paper is to discuss various challenges and technique of web content mining. We also focused on visual methods. We believe that proposed algorithms discussed in this paper will give benefit for various research scholars. Its experimental results show that proposed technique gives better as compared to the existing one.

REFERENCES

[1] R. Gunasundari and S. Karthikeyan, "Removing non-informative blocks from the web pages," in *Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on*, 2010, pp. 810-814

[2] S. Ajoudanian and M. Jazi, (2009) "Deep Web Content Mining," World Academy of Science,Engineering and Technology 49.

[3] R. Etemadi, N. Moghaddam (2008)"*An Approach in Web Content Mining for Clustering Web Pages*" Department of Electrical and computer engineering, Islamic Azad University branch of bonab,Tabriz, Iran.

[4] S. S. Nyein, "*Mining Contents in Web Page Using Cosine Similarity* "International Conference on Knowledge Discovery & Data Mining.

[5] M. Suneetha, S. Fatima, and S. Pervez, "Clustering of web search results using Suffix tree algorithm and avoidance of repetition of same images in search results using L-Point Comparison algorithm," in *Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference on*, 2011, pp. 1041-1046.

[6] B. Mobasher, "Data mining for web personalization," in *The adaptive web*, ed: Springer, 2007, pp. 90-135.

[7] R. Xie and C. Li, "Lexicon construction: A topic model approach," in *Systems and Informatics (ICSAI), 2012 International Conference on*, 2012, pp. 2299-2303.

[8] R. Etemadi and N. Moghaddam, "An approach in web content mining for clustering web pages," in *Digital Information Management (ICDIM), 2010 Fifth International Conference on*, 2010, pp. 279-284.

[9] A. Jebaraj and Ratnakumar, *"An Implementation Of Web Personalization Using Web Mining Techniques"* Journal of Theoretical and Applied Information Technology

[10 B. Liu and K. Chen-Chuan-Chang, "Editorial: special issue on web content mining," *Acm Sigkdd explorations newsletter*, vol. 6, pp. 1-4, 2004.

[11] F. L. Gaol, "Exploring The Pattern of Habits of Users Using Web Log Squential Pattern," in *Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on*, 2010, pp. 161-163.