IMPROVING WEB SEARCH RESULTS IN WEB PERSONALIZATION

Karthika.S[1], Ambika.K[2]

[1]*PG Student, University College of Engineering (BIT Campus), Tiruchirappalli*
[2]*Assistant Professor, University College of Engineering (BIT Campus), Tiruchirappalli*

Abstract

Web mining is the application of data mining which is useful to extract the knowledge. There are three divisions: they are web usage mining, web structure mining, and web content mining. Web personalization is one of the areas of web usage mining which deals with optimizing information by monitoring user interaction history of usage, user profiles. In traditional systems they acquire user feedbacks and ratings explicitly. Personalization concepts involved in order to provide the proper content for the web user who queried to obtain the search results. But the users interest and attractiveness may change frequently. Here in this new system ratings can be done by counting and ranking accordingly to user activity, user profile, dwell time, user clicks and their preference with user interest without the users knowledge. This will give the effective search results while browsing.

Keywords: *Personalization, usage mining, content mining.*

----------***----------

1. INTRODUCTION

Nowadays web users have been increased tremendously. Web users usually have very short life span in web portals while browsing and there is a big issue on getting the exact information what the web users actually wants. Here the website developers and publishers take the responsibilities of providing the efficient search results. Thus they are stepping towards to optimize the web content. Inorder to provide the optimized search results to the users, who are in need of the proper results. In content optimization, personalization is an important feature of focusing the individual user interest rather than the common approach like "one size fits all".

Personalized concepts involved in collecting, monitoring and storing present and past status of user interaction. This helps the client to get the right content at each iteration of searches. During each time of searching there will be improvisation of search results by updating user interactions as specified in personalization concepts.

There are two remarkable approaches such as content-based filtering and collaboration based filtering. In content based approach a user profile is generated for storing the basic user details ,their search actions and also items ratings that done by the user.

The collaborative filtering method is like recommending the user about a site where the site popularity can be accessed by some other user preferences. They usually recognize under the ratings and user commonalities.

The main aim of this system is to reduce manual interruptions in ratings and recommendations of websites. This system provides the concept of ranking by combining user interests obtained by CTR (Click Through Rates), user profiles (history based), recommended model, preference, dwell time calculation. This system is the combination of all preceding systems.

Mining Types

- Web usage mining
- Web content mining
- Web structure mining

1.1Web Usage Mining

Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access.

1.2 Web Content Mining

Web content mining, also known as text mining, is generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query.

This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query.

1.3 Web Structure Mining

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided a into two kinds: Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage

2. ADVANTAGES

Usage mining allows companies to produce productive information pertaining to the future of their business function ability. Some of this information can be derived from the collective information of lifetime user value, product cross marketing strategies and promotional campaign effectiveness. The usage data that is gathered provides the companies with the ability to produce results more effective to their businesses and increasing of sales. Usage data can also be useful for developing marketing skills that will out-sell the competitors and promote the company's services or product on a higher level.

Usage mining is valuable not only to businesses using online marketing, but also to e-businesses whose business is based solely on the traffic provided through search engines. The use of this type of web mining helps to gather the important information from customers visiting the site. This enables an in-depth log to complete analysis of a company's productivity flow. E-businesses depend on this information to direct the company to the most effective Web server for promotion of their product or service.

This web mining also enables Web based businesses to provide the best access routes to services or other advertisements. When a company advertises for services provided by other companies, the usage mining data allows for the most effective access paths to these portals. In addition, there are typically three main uses for mining in this fashion.

3. RELATED WORK

J.A. Konstan, B.N. Miller, D. Maltz, suggested that the system depicts some advantages as follows [1],[2].

Traditional personalization has two approaches they are content based filtering and Collaborative filtering. In the first method, user profile is generated based on content descriptions of content items which was rated explicitly by the users.

In collaborative filtering, second method is the most used method of analyse user rating and commonalities and recommendations items, which are same similar tastes among the users.

The main drawback is limited capability to recommend contents than rated by users. Collaborative filtering may not be appropriate since it suffers from the suffer from the start problem.

This base paper they have overcome that existing disadvantage by counting the user clicks and views rather than ratings that has been done explicitly. Second, accurate understanding of user action become one of essential factor achieve good recommendations performance.

3.1 Counting Feedbacks to Set Popularity

D.kelly depicts the advantage over using the personalization is such that implicit feedback technique is like making user to put more effort in order to obtain feedback explicitly.[3][4] And by utilizing this feedback count, the content will be optimized.

This has been overcome by using implicit feedback gather indirect from the user by monitoring behaviours of user during searching. Here the relevance feedback without much effort of user.

The main advantage is counting reading time, scrolling that is interaction with the document. If they read for more time that article is rated as interesting as opposed not (printing, saving, and bookmarking) is considered to be an interaction that automatically rates the articles.

The amount of time spent with relevant document and spent with irrelevant document are all similar sometimes. Hence, there is a chance of mistakes happen. In 561 documents with 6 subjects a small number less than 1% were displayed multiple times. 240(43%) are identified as relevant and 321(57%) were as irrelevant.

The major drawback is considering relevant document as irrelevant and vicversa. By simply judging the length of time spent by user with the document. It is because the user is unable to understand a document; they may spend a lot of time with the document. By considering such constraints they said to be a drawback.

This base paper they have been considering dwell time interpretation also made significant attention. Here the dwell time means stay time, the amount of time spent in the document by focussing user logs. And also based on click behaviours.

[5]D. Agarwal, founded some advantages that This approach is based on tracking per article performance in near time, through online models. Search engines are automated ranking algorithm. Usually most familiar links are highlighted [6].

There are few disadvantages, they are the articles sometimes may have short lifetime and also frequently changing behaviour. And also thousands and millions of user may visit the web portal within milliseconds. It is hard to capture the state and user profiles within short lifespan.

W. Chu and S.T. Park depicts about depicts user action can be predicted automatically whenever user changes their view or interest on websites/web pages as an advantage [7][8]. It also helpful for improving view on websites. User can get sophisticated views over the web pages. The web content over the web pages can be easily and frequently modifiable and it alters its structure accordingly to the user view.

Unconditionally predicts on user interest without considering are dealing the real truth with the user. Since user views are subjected to change dynamically, prediction mostly ends in failure.

The use of machine learning approach that is to synchronise and analyse user views and add dynamically to the user context profile. The problem can be overcome by creating segmentation.

4. PROPOSED SYSTEM

In this proposed system, it has been trying to reduce the manual work, by the way of updating automatically. The user interest over the websites can be captured without the user's feedback and enforcement of user to give ratings about the websites.

The main aim of this system is to reduce manual interruptions in ratings and recommendations of websites.

This system provides the concept of ranking by combining user interests obtained by CTR (Click Through Rates), user profiles (history based), recommended model, preference, dwell time calculation. This system is the combination of all preceding systems.Hence implicit feedback technique is like making user to put more effort in order to obtain feedback explicitly. And by utilizing this feedback count, the content will be optimized.

This has been overcome by using implicit feedback gather indirectly from the user by monitoring behaviours of user during searching. Here the relevance feedback without much effort of user.

The main advantage is counting reading time, scrolling that is interaction with the document. If they read for more time that article is rated as interesting as opposed not (printing,saving,and bookmarking) is considered to be an interaction that automatically rates the articles.

Hyper Clique Keyword Rank Swapping Algorithm:

RV-Ranking Value
IR-Initial Value
W_i-Websites
i->Order according to the popularity and ranking (1,2,3,……n)
N_v-No of views
T_i-Time duration
K->Keyword
Let W_{IR}=0 [initially]

For (i=0;i<50;i++)
If K==Website Content (W_i)
Display W_i
Else
For (j=0;j<50;j++)
Match K->W_i
While T_i>0 and N_v>0
Do N_v&&T_i->S
End while
If S>RV(W_i)
swap RV(W_i)=S
End for
End if

Hyper clique Keyword Rank Swapping Algorithm describes about swapping the rank value of the websites. It checks for the user clicks and update the rank value. And also verifies whether the previous rank value of the website is greater than new rank value that is been calculated. After that it swaps the new rank value of the website with the old one.

5. COMPARISON

Table 1 shows comparison of all the three algorithms

Table 1: Comparison of algorithms

Algorithm	Page Rank	Weighted Page Rank	HITS	Rank Swapping Algorithm
Mining technique used	WSM	WSM	WSM & WCM	WUM
Working	Computes scores at indexing time. Results are sorted according to importance of pages.	Computes scores at indexing time. Results are sorted according to Page importance.	Computes hub and authority scores of n highly relevant pages on the fly.	Updates websites with ranking. Provides proper search results implicitly according to the users usage
I/P Parameters	Backlinks	Backlinks, Forward links	Backlinks, Forward Links & content	Back and forward links
Limitations	Query independent	Query independent	Topic drift and efficiency problem	Key word dependent

6. CONCLUSIONS

Thus the system is done with the status of producing a new facility that it can provide a intermediate sophisticated user views on websites. This might be giving a different view on collecting the user's feedback. The process of providing a new environment to the user for better searching to be done.It has been considering dwell time interpretation also made significant attention. Here the dwell time means stay time, the amount of time spent in the document by focussing user logs and also based on click behaviours.

User can get web search results accordingly to their own interest but they may be expressed explicitly. In future geographic locations and more user click behaviors can be analyzed and tailored into this system for better performance. By this way of improving optimization will provide outstanding web services in future.

REFERENCES

[1] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Her locker, L.R. Gordon, and J. Riedl, "Grouplens: Applying Collaborative Filtering to Usenet News," Comm. ACM, vol. 40, pp. 77-87, 1997.

[2] D.Agarwal, B.-C. Chen, and P. Elango, "Spatio-Temporal Models for Estimating Click-Through Rate," Proc. 18th Int'l World WideWeb Conf. (WWW), 2009.

[3] D. Kelly and N.J. Belkin, "Reading time Scrolling and Interaction: Exploring Implicit Sources of User Preferences for Relevance Feedback," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2001

[4] D.Agarwal, B.-C. Chen, and P. Elango, "Fast Online Learning through Offline Initialization for Time-Sensitive Recommendation," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD).

[5] D. Agarwal, B.-C. Chen, P. Elango, N. Motgi, S.-T. Park, R. Ramakrishnan, S. Roy, and J. Zachariah, "Online Models for Content Optimization," Proc. Neural Information Processing Systems Foundation Conf. (NIPS), 2008.

[6] E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.

[7] J.L. Herlocker, J.A. Konstan, A. Borchers, and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 1999.

[8] T. Hofmann and J. Puzicha, "Latent Class Models for Collaborative Filtering," Proc. 16th Int'l Joint Conf. Artificial Intelligence (IJCAI), 1999.

[9] W. Chu and S.T. Park, "Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models," Proc. 18th Int'l World Wide Web Conf. (WWW), 2009.

[10] R. Jin, J.Y. Chai, and L. Si, "An Automatic Weighting Scheme for Collaborative Filtering," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2004.

[11] W. Chu, S.T. Park, T. Beaupre, N. Motgi, and A. Phadke, "A Case Study of Behavior-Driven Conjoint Analysis on Yahoo! Front Page Today Module," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD).

[12] J. Teevan, C. Alvarado, M.S. Ackerman, and D.R. Karger, "The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Directed Search," Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI), 2004