THE IMPROVE CONCEPT OF WEB PERSONALIZATION ON TASK BASED MODELING IN WEB USAGE MINING TECHNIQUES

[1]C.Ramkumar, [2]D.Prabhakaran
[1,2]Assistant Professor, Department of CSE, Sasurie College of Engineering, Tirupur, India
[1]c.ramkumar84@gmail.com, [2]prabha619@gmail.com

Abstract: Web usage Mining (WUM) is interesting patterns behavior that allow analyzing with help of website administrator. This technique is used to classify users and pages the content pages, analyzing user's behavior and ordered of URL's accessed. The research engine is providing the matching information to the users to satisfy their requirements. If anyone approach is fulfills the requirements of the user to personalize the information available on the web is known as web personalization. our approach is searching historic search engine logs to find out other users are performing same tasks to the current user and leverage their on-task behavior to discover web pages to promote in the current ranking. Here, providing richer models of the web page to the current user from the historic users search tasks used to improve likelihood of finding matching content and improve the concept of personalization. In this paper presents how to identify the same tasks to the current user's task from the web log files and ranking to the web pages.

Key Terms: Web Usage Mining, Preprocessing, Personalization, Web Log Files, Ranking Features, Task modelin

1. INTRODUCTION

The World Wide Web is a rich source of information, expand in size and complexity. An ultimate need of the search engine is that of predicting the user needs in order to improve the usability of a web site. Web Personalization can be defined as any action that adapts the information or services provided by a web site to an individual user, or a set of users, based on knowledge acquired by their navigational behavior, recorded in the Web Log files. Historic search interactions from a user over time can be used to personalize search results [3,4], but the focus there is either once again on query based matching [4] or creating common models of searcher interests across a variety of topics [3].

This paper is structured as follows. Section 2 describes Web Mining and Web Mining Categories, section 3 describes Preprocessing Stages, and Section 4 describes related work in task modeling, mining task-relevant search behavior and personalization. Sections 5 and 6 describe the identifying similar tasks and ranking features. Section 7 and 8 discusses these findings and implications and we conclude in Section 9.

2. WEBMINING

Web Mining is the Data Mining technique that automatically discovers or extracts the information from web documents. There are three areas in web data mining.

2.1 Web Content Mining

It is the process of extracting the information from the content of the web pages. Web content mining is related to data mining techniques. This technique is used in web content mining. It is also linked with text mining, because web data is mainly semi structured in nature.

2.2 Web Structure Mining

Web structure mining objective is producing structural summary about web sites and web pages. The covered on structure mining is therefore on link information that is an important feature of web data.

2.3 Web Usage Mining

It is used to determine interesting usage patterns from web data. With the purpose of understand and better serve the needs of web based application. It tries to make sense of data produced in web surfer's sessions/ behaviors.

3. PREPROCESSING

Web log data is usually diverse and voluminous in nature. This data must be assembled into a consistent, integrated and comprehensive view. A typical example of web log file is shown in Figure.1.

Figure 1: Common Web Log Format

A web server log file contains requests made to the webserver chronological order recorded. The main log

file formats are the Common Log Format (CLF) and extended. A common log format file is generated by the web server to keep track of the requests that happen on the web site. The following format standard log file shown in figure 2.

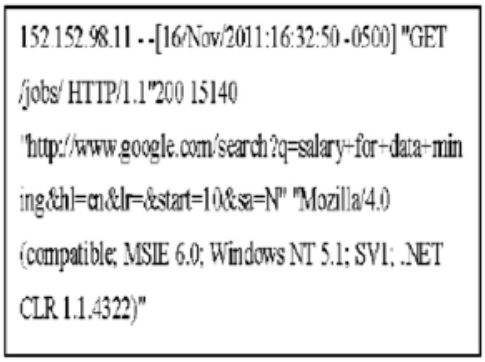

Figure 2: Example of Log File format

A Web log is a file to that Web server writes information every time a user requests a resource from that specific site. While user submit request to a web server that activity is recorded in web log file. Log files range between 1KB to 100MB.

Figure 3: Example of typical server log

A common log file is created in Web server to keep track of the requests that occur on a Web site. Figure 3 is shown the typical server log. The stages of preprocessing are shown in Figure 4.

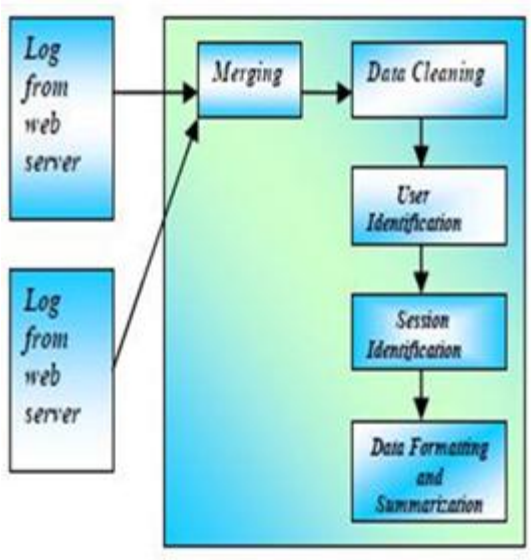

Figure 4: Stages of Preprocessing

The following steps are comprised: Data cleaning, identification of users, sessions, visits, data formatting, merging of log files from different web servers and summarization.

4. RELATED WORK

There are three relevant areas of related work: (1) task modeling (2) task modeling personalization of search engines (3) mining the search behaviors of other users. Monika Shoni, Rahul Sharma et al[1] provides framework for web personalization using web mining. K.R. Suneetha, et al. [2] website top errors, potential visitors of the site, are analyzed. Fang Yuan, et al. [5] mainly consider on analyzing visiting information from logged data to extract usage patterns, which can be classified on to three categories. Rekha Jain, et al. [6] discusses the page ranking algorithm for web mining. Ramya C, et al. [7] discusses stages of preprocessing and Web log files.

5. IDENTIFYING SIMILARTASKS

The first step is applying task identification in sessions. Before that, we have to use groupization concept to collect historic users' search tasks. To identifying similar tasks involves two processes. There are Groupization and Computing task Similarity.

5.1 Groupization

The task-based approach is used to the current users search history or all users search histories as so-called "groupization" (a variant of personalization here, other

users profiles are used to personalize the search experience [16]).The first step in applying our method is to identify tasks within search sessions. Now we describe the task identification Process. It involves in two steps: 1) Log data 2) Identifying task in sessions.

5.1.1 Log Data

The primary source of this study is a data set comprising from the anonymized set of users of the Google Search Engine. The logs contained a unique user identifier, a search session identifier, the query, the top-10 URLs returned by the search engine for that query, and clicks on the results. Logs were split into search sessions demarcated with a 30-minute inactivity timeout, such as that used in previous work[17].

5.1.2 Identifying tasks in sessions

In order to calculate inter-query similarities, QTC takes a supervised learning approach. It involves two process.

- Measure the similarity between query pairs
- Cluster queries into tasks based on similarities

Figure 5 explains how the sessions and tasks are divided. Here, with the help of query clustering QTC [8], this has the advantage of segmenting the interleaved tasks in a session.

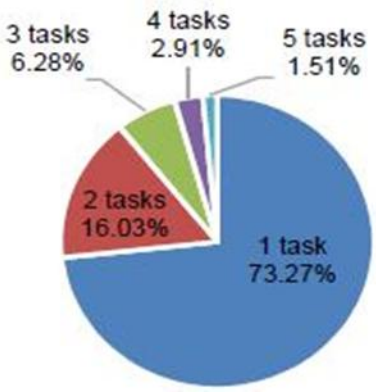

Figure 5: Number of search tasks in search sessions

Figure 5 illustrates the fraction of sessions containing between one and five search tasks. The figure shows that around 90% of sessions have one or two tasks; 73.3% sessions contain a single task and about 16.0% sessions contain two tasks. This shows that although most sessions comprise a single task, there are still a sizable number of sessions (over 25%) containing multiple tasks.

5.2 Computing Task Similarity

A key part of this process is finding other users attempting searching similar tasks. In this section describes how find similar task and features that we generate for ranking. There are number of ways to find similarity between a given pair of tasks.

5.2.1 Query Similarity

These similarity measures are based on comparing the queries that users issue in both tasks under consideration. Similarity in this case can be based on the exact terminology used in the queries (after normalization) and more generally, on the semantic similarity between the queries.

5.2.2 Syntactic Similarity and Semantic Similarity

Syntactic similarity illustrates the string match between the queries. Similarity can be computed based on the overlap between the tasks in terms of:

- The fraction of queries that are shared between tasks (i.e., the intersection divided by the union), and
- As the fraction of unique query terms that are shared between tasks.

While the queries may not overlap, but the semantic of queries may overlap. To address this, compute task similarity by measuring semantic similarity. Let $Q = q_{1.....}q_j$ be one query and $S = s_{1.....}s_I$ be another. The semantic similarity between two queries can be evaluated Depends on the IBM Model 1 [9, 10]. Treating Q and Sas two sequences of words, the IBM Model 1-based semantic similarity model is defined as:

$$P(S|Q) = \prod_{i=1}^{I} \sum_{j=1}^{J} P(s_i|q_j) P(q_j|Q)$$

Where P(q|Q) is the unigram probability of word qin query Q. The word translation probabilities P (s|q) are estimated on the query title pairs derived from the click through search logs, assuming that the title terms are likely to be the desired alternation of the paired query.

6. RANKING FEATURES

Web mining technique gives the extra information through hyperlinks where dissimilar documents are connected. There are number of algorithms proposed for ranking to the web pages.

6.1 Page Rank

This algorithm was developed by Brin and Page at Stanford University [11]. The Page Rank forms a probability distribution over the web pages so the total amount of Page Ranks of all web pages will be one. Page and Brin proposed a formula to calculate the PageRank of a page A stated as below.

$$PR\ A\ =\ 1-d\ +d\ PR\ T1\ \backslash C\ T1\ \ +\ PR\ Tn\backslash C\ Tn$$

Here PR (T_i) is the Page Rank of the Pages T_i that links to page A, C(T_i) is the number of outlinks on page T_i and d is damping factor. It is mainly used for stop other pages and also including too much influence.

6.2 Weighted Page Rank

This algorithm was projected by Wenpu Xing and Ali Ghorbani are used an extension of Page Rank algorithm The importance is assigned to weight values incoming and outgoing links.This is represented

$$W\ \ {}^{in}_{m,n}\ and W\ \ {}^{out}_{m,n}\ \text{respectively}$$

$$W\ {}^{in}_{(m,n)} = \frac{In}{\rho\epsilon R\ m\ Ip}$$

In is amount of incoming links of page n, Ip is amount of incoming links of page p, R (m) is the reference page list of page m.

$$W\ {}^{out}_{m,n} = \frac{On}{\rho\epsilon R\ m\ \ OP}$$

On is amount of outgoing links of page n, Op is number of outgoing links of page p, and then the weighted PageRank is given by formula in

$$WPR\ n\ = 1-d \sum_{m\in B(n)} WPR(m)\, W^{in}_{(m,n)} W^{out}_{(m,n)}$$

6.3 Re-ranking Feature

The inter-task similarity is computed using k (t, t') which is then multiplied against the click count for each URL in the top-10(i.e., w (t', u)). The result summed over all tasks in the historic data is used to generate the final feature value. In addition, we also compute Clicked Tasks Count, which is the total number of tasks for which a particular URL u is clicked. This measures URL popularity independent of task. Note that since Query Translation and Category Similarity KL are asymmetric.

7. EXPERIMENTS

Our log-based evaluation method focuses on a re-ranking task, assessing the extent to which the models promote clicked results. The re-ranking models attempt to promote observed satisfied result clicks(SAT clicks) toward higher rank positions in the result list. This allows offline assessment of models performance using judgments personalized to each user. This approach has been used to determine the effectiveness of various re-ranking methods [13, 14, 15].We define the clicks having less than 30 seconds dwelling time as quick backs. We consider three types of clicks in labeling user feedback in the logs: SAT clicks, quick back clicks, and no clicks

Table 1: Statistics of the weekly data for

Count	Training	Validation	Test
SAT Clicks	2,086,335	2,062,554	2,082,145
Quick back Clicks	417,432	408,196	413,496
Tasks	1,165,083	1,126,452	1,135,320
Queries per Tasks	2	2	2

- **Learning/evolution**

In each impression, if a URL received at least one SAT click, the URL is labeled with a 2; if a URL received only quick back clicks, the URL is labeled with a 1; if a URL was not clicked at all, the URL is labeled with a 0. This gives us a three level judgment for each top-10 URL for each query.

7.1 Measures

We measure ranking quality by mean average precision and mean reciprocal rank. In both cases, the mean is calculated over all the impressions in our test set. Mean

average precision (MAP) for a set of queries is the average mean of precision scores for all query. The average precision score is denoted as

$$Average\ Precision = \frac{\sum_{k=1}^{n} Precision\ \ k\ \ Rel\ (k)}{\sum_{k=1}^{n} Rel(k)}$$

where nis the number of URLs in the impression, usually 10,Rel(k)is an indicator function equaling 1 if the URL at rank kisarelevant document, zero otherwise, and Precision(k)is the precision at cut-off kin the ranked list .Mean reciprocal rank (MRR) for a query set is the average of the reciprocal ranks across all results, which is defined as

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i}$$

Where rank iis the rank of the first relevant URL in the ranking list ,and Nis the number of impressions in test. These measures are complementary in that MRR focuses on the rank of the first relevant document in the top 10, whereas MAP targets the rank of relevant results across the top 10 documents.

8. RESULT

Precision-Queries graph is done on average of user feedbacks. The graph shows that proposed Re Rank gives better results than PageRank and Weighted PageRank. We managed to provide high precision at the tasks increases.

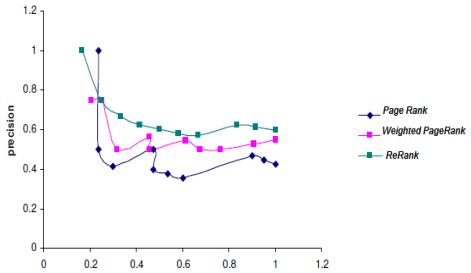

Figure 6: Queries Precision – Queries Plot

9. CONCLUSION

This work clearly demonstrates the value of considering search tasks rather than just search queries during personalization, as well as the benefit of groupization. In this chapter, attempted to present a complete view of the personalization process depends on web usage mining. The approaches we have detailed show web log files, identifying similar tasks, Re ranking features, can be leveraged effectively as an integrated part of a web personalization system. In future work involves use of a broader range of cohorts. Cohorts include location, browser, place, etc., and the development of more sophisticated and generalizable models of task behavior.

REFERENCES

[1] Monikashoni, Rahul Sharma and Vishal Shrivatsava. Framework for web Personalization, IJRET Oct 2012

[2] Suneetha, K. R., and Raghuraman Krishnamoorthi. "Identifying user behavior by analyzing web server access log file", IJCSNS International Journal of Computer Science and Network Security 9, no. 4 (2009): 327-332.

[3] Sontag, David, Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Susan Dumais and Bodo Billerbeck, "Probabilistic models for personalizing web search", In Proceedings of the fifth ACM international conference on Web search and data mining, pp. 433-442. ACM, 2012.

[4] Teevan, Jaime, Daniel J. Liebling, and Gayathri Ravichandran Geetha, "Understanding and predicting personal navigation", In Proceedings of the fourth ACM international conference on Web search and data mining, pp.85-94. ACM, 2011.

[5] Yuan, Fang, Li-Juan Wang, and Ge Yu. "Study on data preprocessing algorithm in web log mining." In Machine Learning and Cybernetics, 2003 International Conference on, vol. 1, pp. 28-32. IEEE, 2003

[6] Jain, Rekha, and Dr GN Purohit. "Page ranking algorithms for web mining", International journal of computer applications 13, no. 5 (2011): 22-25.