Scientific Journal Impact Factor (SJIF) : 3.518

International Journal of Modern Trends in Engineering and Research

ISSN (ONLINE) : 2349 - 9745

ISSN (PRINT) : 2393 - 8161

A Survey of Page Ranking Features and Techniques

Anurag Soni [1]And Diamond Jonawal[2]

[1]M.Tech (CTA) Research Scholar RGTU University,[2]RGTU University, Bhopal, India

Abstract—Re-ranking of Web page is a method which has been extensively used to minimize the access latency of web pages on the internet. However, it is required for the users to visit the highly frequent prefetched web pages in their subsequent accesses, so that the limited bandwidth of the network and server resources can be efficiently utilized and the access delay can be avoided. Therefore, there is a need of a perfect re-ranking method during prefetching. User‘s navigational behavior on the web can be analyzed by using Markov Model. Multidamping is another technique for the ranking of the pages which applies the stochastic matrix. This paper gives detailed information for the page ranking, different features and methods.

Index Terms - Information extraction, Page prediction, Web re-ranking, Web mining, Web Usage Mining.

I. INTRODUCTION

With the increase in the use of the internet in day to day life, Importance of the web world is getting higher. As large amount of work is performed on the net for the transparency and speed, this introduces the load on the web sites for work. The availability of resources is limited and one has to manage things in available resources. So the other way of optimizing the performance of web sites is to learn the user’s behavior pattern for presenting the next page on the client end. The internet and the World Wide Web are significant sources of information retrieval these days and the users from different fields access the web frequently. The web logs are recorded information about usage of web by the users. Web Usage Mining is the process of analyzing web log files to generate useful patterns regarding user’s usage behavior. This process includes clustering, association rule mining and sequential pattern mining etc. To facilitate web page access by the users, web recommendation model is needed. Therefore, the interest in the analysis of user’s behavior on the web has been increasing rapidly. The aim of such analysis is to provide users with easier access to the required information at the right time and in the most suitable form.

Web Usage Mining Procedure: The first activity involved in the web usage mining procedure is preprocessing in the web log files. The second activity is mining algorithm. The use of such algorithm is to find out the pattern, and the final activity is analyzing the pattern which is mined by mining algorithm. Figure 1 below illustrates the procedure of web usage mining [2], [9], [11]. User’s timestamp and the period of session is gathered from his/her web log data and identified by way of preprocessing process. Mining algorithm is a method to discover the patterns and rules from the sequence pattern such as clustering algorithm, association rule and sequential pattern analysis.

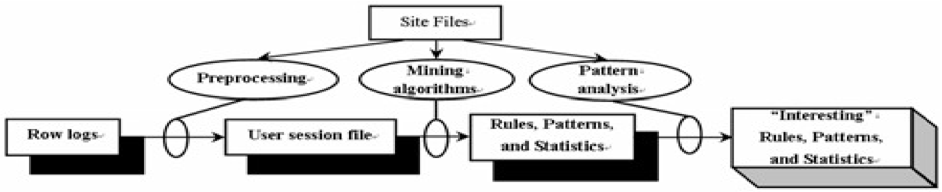

Figure 1: Web Usage Mining Procedure

There has been tremendous amount of work carried out on the IR, database, and topology, which construct the base for the web content mining and web structure mining. Web usage mining is a new area of interest and has gained lot of popularity in present time. Detailed information about usage mining is given in the next section on the basis of some up-to-date research works.

Web usage mining includes the automated discovery and analysis of patterns in data which is the results of the user's interactions with one or more web sites. The reason behind the discovery of web access patterns is to understand the navigation preferences followed by the user on the web and their behavior using some tools and techniques. These techniques are used effectively to help e-commerce businesses improvise their web sites in a better manner (Heer & Chi-2002). The focus of web usage mining is to get the model and analyze the user's behavioral patterns. It consists of three phases: (1) the web data is pre-processed, (2) discovery of patterns from pre-processed data and (3) analysis of the patterns (Srivastava et al. 2000) [15]. Of all these three phases only the later phase is performed in reality. The patterns which are discovered are represented as a group of pages that are frequently accessed by groups of users with same type of interests within the same web site.

In web prediction, primary challenges are in preprocessing as well as in prediction. Challenges of preprocessing include handling of large volume of data which is beyond the storage capacity of computer's memory, selection of optimum sliding window size, recognizing the login sessions, and searching/obtaining domain knowledge. Prediction challenges include lengthy training or prediction time, lower prediction accuracy, and memory constraints.

II. RELATED WORK

Based on the type of information used to make a particular prediction, the prediction algorithms can be broadly classified in two main groups [Domenech 06e]. The first of which includes algorithms that predict future accesses based on the previous access patterns. There can be two subgroups as follows: one group consists of algorithms that use Markov models [Padmanabhan 96, Palpanas 99, Domenech 06a, Zhu 02], and the other one with algorithms that make use of data mining techniques [Yang 03, Gunduz 03, Nanopoulos 03]. Large number of prediction algorithms based on Markov models are found in the literature and few models among them provide predictions with high precision but at the cost of extreme computation and lot of memory consumption. Even the data mining based algorithms also consume the resources .

The second group makes use of the algorithms that analyze the web content to make certain predictions. Some authors have proposed to combine the analysis of the content with usage profiles [Duchamp 99], others utilize neural networks to apply on keywords which are extracted from HTML content and few others observe similarities in context words around links in the HTML content [Davison 02]. The proposals are based on the object popularity and the association of hyperlinks, but the relationship among objects is not considered by them.

The commonly used method of search results presentation is a simple ranked list [2]. Possibly, this type of presentation method is suitable for unambiguous, similar search results. In practical, it is suitable when the search results are good and a user can easily find many relevant documents in the top ranked results.

Though, when the search outcomes are dissimilar (because of ambiguity or more than one aspects of a topic) which commonly happens during web search, the presentation using ranked list would be ineffective; in such a case, it would be finer to organize the search outcomes into clusters to make user's navigation into a particular interesting group more easier. People attempt to infer user objectives and purposes by predetermining some specialized classes and performing query classification accordingly. Lee et al. [6] classified user goals and queries in the following two categories as 'Navigational' and 'Informational'. Other related works concentrate on tagging queries but having some previously defined concepts for making improvement in the queries representation. Although, user's interest varies so much for different queries, therefore, exploring more suitable predefined search objective classes is not very easy and practical.

Approaches of arranging search outcomes based on text classification are examined in [7]. In this work, a text classifier is trained by using a web directory and then obtained search results are classified into the predefined categories. The authors have modeled and studied various class interfaces and they observed that class interfaces are more productive as compared to list interfaces. Although, predefined classes are mostly very general to show the finer granularity features of a query.

III. BACKGROUND

Web data mining is the method for applying data mining techniques on web data. Researches made in this field has the aim of helping e-commerce businesses in their decision making, assisting in the design of good web sites and assisting the users during navigation on the web. The web data mining mainly concentrates on three points: (1) Web structure mining (2) Web usage mining and (3) Web content mining. This classification is based on two factors namely, the purpose and the data sources.

Structure: If the web page has a direct link to another web page, or the web pages are linked in subsequent manner, then it is likely to find out the relationships between those web pages. The relationships can be categorized in the following types: as they can be related by similarity or philosophy, they may have similar contents and both can be in the same web server or may be created by the same person. Web structure mining is also used to find the tree structure or network of hyperlinks present in the web sites which come under a specialized domain on the internet. This technique would make query processing method easier and effective by checking the flow of information in the web sites of a particular domain.

Web documents contain lots of links and it makes use of primary data available on the web. Web structure mining has a real relation with the web content mining. These two techniques are combined often in an application. The objective of web structure mining is to produce structured detail about web sites and web pages in order to identify needed documents. The primary vision here is on link information, which is an essential point of web data. Web structure mining is used to impart the structure or schema of web pages which would make the web document classification easier and clustering based on its structure.

IV. WEB USAGE MINING

Web usage mining makes an attempt to collect useful information from the secondary data which is generated by the users during web surfing. Its aim is to find the techniques that can reveal user’s behavior while the users communicate over the Web. M. Spiliopoulou [14] discovered the potential strategic goals in each domain into mining aim as: finding the user’s behavior within the site, comparison between what is expected and what is actual usage of web sites, adjustment of the web site according to the interests of its users. There is no proper difference between the web usage mining and web content mining. To prepare data for web usage mining, the web content and the topology of the web site will be used in the form of the sources of information which interacts web usage mining with the web content mining and web structure mining. The clustering involved during the process of pattern discovery acts as a link between web content and structure mining from usage mining.

S. Chakrabarti’s [13] discovery gave detailed knowledge related to the application of the various techniques covering the areas of data mining, statistical pattern recognition, machine learning and analyzing hypertext. It provides latest knowledge about the emerging trends in the area of content mining discoveries. Zaiane & Han (2000) [5], made a focus on resource recovery on the web. The authors transformed the unstructured data available on the web into a structured data by using database technology.

V. TECHNIQUES

Markov Modal: The generation of Markov models of all orders and utilizing them all in forecasting is depicted as in [9]. It is important to notice that the function *predict* (*x, mk*) is presumed to forecast the next visited page of session *x* by utilizing the k_{th}- order Markov model *mk*. If the *mk* becomes unsuccessful, the $mk-1$ is deliberated using a new session *x'* of length $k - 1$ where *x'* is calculated by stripping the first page ID in *x*. This process recapitulates until prediction is achieved or prediction fails. For instance, given a user spell $x = (P1, P5, P6)$, forecast of all K_{th} model is done by asking third-order Markov model. If the forecasting using third-order Markov model becomes unsuccessful, then the second-order Markov model is asked on the spell $x\_ = x - P1 = (P5, P6)$. This process reiterates until getting the first-order Markov model. Therefore, the all K_{th}-order Markov model gets better forecast and it only becomes unsuccessful when all orders of the basic Markov models become unsuccessful in forecasting.

Predict_markov algorithm takes session and model number as input and then find most frequent page. If it generates more than one page then second feature will be predicted for the page selection which is keywords extracted from the web pages. Their similar functions take key_vector which is the collection of the keywords which is obtained from the previous page of the session, then compare the keywords of the pages in V vector. The most similar page will be the next target page of the session. This page is returned to the function.

Computing HITS Algorithm [12]**:** Two weights are assigned to every page P: a_p is a non-negative authority weight and h_p is non-negative hub weight. The invariant is retained that the weights of every kind are normalized so their squares sum to 1. Better authorities are those pages which have a larger ά value, and the pages having a larger h value are called "better" hubs. Numerically, the reciprocally emphasizing relationship among hubs and authorities may be shown as follows: (1) if p points to many pages with large ά values, then it should get a large h -value; (2) if many pages point to p with large h -values, then it should get a large ά-value. This motivates the definition of two operations on the weights, denoted by I and O. Given weights a_p and h_p, the I operation updates the ά -weights as follows;

$$a_p \longleftarrow \sum_{q:(q,p)\in\epsilon} h_q$$

In the same way, the O operation updates the h-weights as follows;

$$h_p \longleftarrow \sum_{q:(p,q)\in\epsilon} a_q$$

Thus, the I and O operations are the basic means by which hubs and authorities reinforce each other. To find the expected "equilibrium" values for the weights, the I and O operations can be applied alternatively, and then observe whether a fixed point is achieved.

SALSA: In [10], algorithm completes an irregular walk on the bipartite hubs and authorities' graph, back and forth among the hubs and authority sides. The random walk begins from some authority nodes chosen evenly at random and then continues by changing back and forth among backwards and forward steps.

The algorithm selects among one of the incoming links evenly at random at a node on the authority side of the bipartite graph, and goes ahead to a hub node on the hub side. When at node on the hub side the algorithm selects among one of the outgoing links evenly at random and goes ahead to an authority node. The authority weights are defined to be the static distribution of this random walk. Precisely, there are transition possibilities in Markov Chain of the random walk. Let if F_u is the set of pages to which 'u' is pointing and B_u is the set of pages which are pointing to u. Then

$$P_a(i,j) = \sum_{k:k \in B(i) \cap B(j)} \frac{1}{|B(i)|} \frac{1}{|F(k)|}$$

Multi-Damping Method: In [11], Let Y is an adjacency matrix for the graph of nodes. Where i represents the node after which j node is chosen by the surfers with probability p'.
P' = (Vj / V_total) = (number of logs contain j node after i node / total number of logs which contain i node).

$$Y\ (i,j) = p'$$

In this algorithm, first Zk is calculated which is the damping coefficient & G(µ) is the Google matrix. Stochastic matrix $S{:} = P + Y$. For a random web surfer about to visit the next page, the damping factor $\mu \in [0,\ 1]$ is the probability of choosing a link-accessible page. Alternately, with probability $1 - \mu$, the random surfer makes a transition to a node selected from among all nodes based on the conditional probabilities in vector *v*. As an example, for the case of Linear Rank for $k = 3$, the damping coefficients are $\xi 0$ =2/5 = 1– 3/5 , $\xi 1$ = 2/4*3/5 = 3/55 (1 – 2/4), $\xi 2$ = 2/4*2/5 = 3/5*2/4 (1 – 1/3)and $\xi 3$ = 2/4*1/5 = 1/3*2/4*3/5.
This clearly identifies $\mu 1$ = 1/3, $\mu 2$ = 2/4 and $\mu 3$ = 3/5 as the corresponding damping factors. M is damping factor = ($\mu 1$, ..., μk).
Require: Zk := {ξj ≥ 0, j = 0, ..., k} finite set of coefficients defining or approximating the functional ranking.
Normalize: If $\Sigma_{j=0}^{k} = \xi_j < 1$ then

$$\texttt{add\_cor}(\mathcal{Z}_k) \quad := \quad (\xi_0, \ldots, \xi_{k-1}, \xi_k + 1 - s)$$

$$Zk \leftarrow \text{add cor}(Zk)$$

end if.
Encode: Generate damping factors Mk, e.g. using recurrence.

$$\mu_j \quad = \quad 1 - \frac{1}{1 + \frac{\rho_{k-j+1}}{1-\mu_{j-1}}}, j = 1, \ldots, k,$$

Where $\rho_k \quad = \quad \frac{\xi_k}{\xi_{k-1}}$

$$\prod_{j=1}^{k} G(\mu_{k-j+1})v = \xi_k S^k v + p_{k-1}(S)v.$$

.

In order to evaluate this work, there are different parameters present for the different techniques. The best parameter which suits this work is the precision where it gives the value which is a measure of the prediction which is correctly identified by the proposed model to the all logs pass in the experiment. The other important measure is the Recall and F-score [16].

True Positive: When the system says page P1 and actual page is also P1.
True Negative: When the system says page P1 and actual page is P2.
False Positive: When the system says no page and actual page is also P1.
Precision = TP / (TP+ FP)
Recall = TP / (TP + TN)
F-score = 2 * Precision * Recall / (Precision + Recall)
Where TP = True Positive, TN = True Negative, FP = False Positive.

VI. CONCLUSION

As the chance of page re-ranking in the web network is totally dependent on the user but with the help of different pattern generated from the behavior of each user, it is possible to successfully generate a positive result in this direction with the involvement of different techniques. This paper presents different combination of features of the web mining for re-ranking of the web pages. There is no general method developed till now which can rank the pages efficiently. Different web has different requirement.

REFERENCES

[1] J. Domenech, J. Sahuquillo, J. A. Gil & A. Pont. The Impact of the Web Pre-fetching Architecture on the Limits of Reducing User's Perceived Latency. Proc. of the International Conference on Web Intelligence, 2006.
[2] D. Duchamp; Pre-fetching Hyperlinks. Proc. of the 2nd USENIX Symposium on Internet Technologies and Systems, 1999.
[3] Boldi, Paolo, et al. "Query suggestions using query-flow graphs."*Proceedings of the 2009 workshop on Web Search Click Data*. ACM, 2009.
[4] Z. Lu, H. Zha, X. Yang, W. Lin, Z. Zheng, ―A New Algorithm for Inferring User Search Goals with Feedback Sessions,‖ Proc. IEEE Transactions on Knowle dge and Data Engineering, pp. 502-513, 2013.
[5] Zaiane, Osmar R., and Jiawei Han. "Webml: Querying the world-wide web for resources and knowledge." *Proc. ACM CIKM'98 Workshop on Web Information and Data Management (WIDM'98*. 1998.
[6] Lee, Uichin, Zhenyu Liu, and Junghoo Cho. "Automatic identification of user goals in web search." *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005.
[7] Mele, ― Web Usage Mining for Enhancing Search Result Delivery and Helping Users to Find Interesting Web Content,‖ ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '13), pp. 765-769, 2013.
[8] Brian D. Davison, "A Web Caching Primer" IEEE INTERNET COMPUTING 2001.
[9] Mamoun A. Awad and Issa Khalil "Prediction of User's Web-Browsing Behavior: Application of Markov Model". IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 42, NO. 4, AUGUST 2012.
[10] R.Lempel, S.Moran, The stochastic approach for link-structure analysys (SALSA) and the TKC effect, Proceedings of the 9th International World Wide web Conference, 2000.
[11] Giorgos, Kollias, Efstratios Gallopoulos, and Ananth Grama "Surfing the Network for Ranking by Multidamping". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING- 2014.
[12] J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of ACM (JASM), 1999.
[13] Chakrabarti, Soumen. "Data mining for hypertext: A tutorial survey." *ACM SIGKDD Explorations Newsletter* 1.2 (2000): 1-11.
[14] Spiliopoulou, Myra, Lukas C. Faulstich, and Karsten Winkler. "A data miner analyzing the navigational behavior of web users." *Proc. of the Workshop on Machine Learning in User Modeling of the ACAI99, Greece*. Vol. 7. 1999.
[15] "Web usage mining: discovery and applications of usage patterns from Web data" by J Srivastava,R Cooley,M Deshpande,P Tan - *SIGKDD Explor. Newsl , 2000.*
[16] http://en.wikipedia.org/wiki/Sensitivity_and_specificity.