International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 02, Issue: 02 (MAR-APR, 2016)

A Detail Survey of Page Re-Ranking Various Web Features and Techniques

Vipin Kumar Jain[1], Prof. Gaurav Soni[2], Prof. Rajesh Nigam[3]
Computer Science & Engineering Department
Technocrats Institute of Technology & Science, Bhopal, India
[1]er.vipinjain@gmail.com, [2]gauravsoni.rits@gmail.com, [3]rajeshrewa37@gmail.com

***Abstract*— as the internet market is growing a lot, so fast response sites are highly desirable. In order to reduce this latency time website page re-ranking is one of demanding research area. In order to understand the user behavior website data need to be update regularly with proper study. This paper has introduced various required features of web mining with techniques for ranking the pages as per user interest. Here techniques explained in paper are multi-damping, markov, stochastic matrix, etc.**

Keywords*— *Information Retrieval, Page Re-ranking, web mining.

I. INTRODUCTION

A day by day increase use of the internet Importance of the web world is high, so large amount of work is done on net for the transparency and quick. As the importance introduce load in the sites for work and with limited sources one has to manage things in available resource. So other way of optimizing sites is to learn the user behavior pattern for presenting the next page on the other side of the server that is client end. The web is an important source of information retrieval now-days, and the users accessing the web are from different backgrounds. The usage information about users is recorded in web logs. Analyzing web log files to extract useful patterns is called web usage mining. Web usage mining approaches include clustering, association rule mining, sequential pattern mining etc., to facilitate web page access by users, web recommendation model is needed. So the Interest in the analysis of user behavior on the Web has been increasing rapidly. This increase stems from the realization that added value for Web site visitors are not gained merely through larger quantities of data on a site but through 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 which used is to find out the pattern, whereas the final activity is analyzing the pattern which is mined by mining algorithm. The detailed processing of web usage mining procedure is shown the figure2 as follows [2] [12] [14]: users' timestamp and the session period is accumulated from web log data and recognized through preprocessing process. Mining algorithm is a method to find out the rules and patterns from the sequence pattern, for example, association rule, clustering algorithm, and sequential pattern analysis.

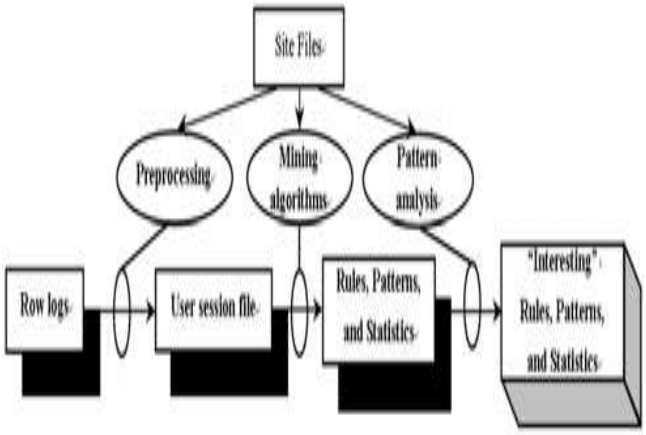

Fig. 1 Different stages of web pattern mining.

There are has been tremendous amount of works which has been carried out on the IR, Database, Intelligent Agents and Topology, which gives 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 current time. Detailed information about usage mining would be provided 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 results in the user's interactions with one or more Web sites. Web access patterns are discovered to understand the users' navigation preferences and behavior by focusing on 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 users' behavioral patterns. It consists of three phases: Pre-processing of Web data, pattern discovery and pattern analysis, Srivastava et al. (2000). Of all these three phases only the latter phase is performed in reality. The patterns which are discovered are represented as group of pages that are frequently accessed by groups of users with same type of interests within the same Web site. In Web prediction, main challenges are in both preprocessing and prediction. Preprocessing challenges include handling large amount of data that cannot fit in the computer memory, choosing optimum sliding window size, identifying sessions, and seeking/extracting domain knowledge. Prediction challenges include long training/prediction time, low prediction accuracy, and memory limitation.

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. The first of which includes algorithms that predict future accesses based on the previous access patterns. To distinguish in two subgroups can be: one consists of algorithms that use Markov models and the other one with algorithms that makes use of data mining techniques. Large number of prediction algorithms based on Markov models are found in the literature and some of them provide high precision predictions but at the cost of extreme computation and lot of memory consumption. The data mining based algorithms consume the resources even. 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, others apply neural networks to keywords extracted from HTML content and some others detect similarities in context words around links in the HTML content. The proposals are based on the object popularity and the association of hyperlinks, but they do not consider the relationship among objects. The most common strategy of presenting search results is a simple ranked list [2]. Intuitively, such a presentation strategy is reasonable for non -ambiguous, homogeneous search results; in general, it would work well when the search results are good and a user can easily and many relevant documents in the top ranked results. As web log feature is use for the web mining different navigation but this contains various patterns so perfect utilization of the feature is possible by grouping. Here clusters of user pattern are arranged as per the navigation. By grouping in cluster prediction accuracy result is increase. This help in predefining the class for increasing the query performance.

Lee et al. [6] consider user goals as ―Navigational‖ and ―Informational‖ and categorize queries into these two classes. Other works focus on tagging queries with some predefined concepts to improve feature representation of queries. However, since what users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.

Methods of organizing search results based on text categorization are studied in [7]. In this work, a text classifier is trained using a Web directory and search results are then classified into the predefined categories. The authors designed and studied different category interfaces and they found that Category interfaces are more effective than list interfaces. However predefined categories are often too general to reflect the finer granularity aspects of a query.

III. BACKGROUND

Web data mining is the method for applying data mining techniques on Web data. Research made in this filed has the aim of helping e-commerce businesses in their decision making, assisting in the design of good Web sites and assisting the user when navigating the Web. The World Wide

Web data mining has a major focus on three issues namely: Web structure mining, Web content mining and Web usage mining. The classification is based on two factors namely, the purpose and the data sources.

Structure: If the Web page is being is linked to another Web page directly, or the Web pages are next to each other or neighbors, then we would like to discover the relationships among those Web pages. The relations can be categorized in any one of the types, such as they 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 in the Web sites of a particular domain. This technique would make query processing method easier and effective by checking the flow of information in Web sites of a particular domain. Web documents consists lot 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 goal 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.

Web Usage Mining: - Web usage mining makes an attempt to find useful information from the secondary data derived from the interactions of the users while surfing on the Web. Its aim is to find the techniques that can reveal user 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 Web site usage, adjustment of the Web site 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 as the information sources 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 bridge to Web content and structure mining from usage mining.

S. Chakrabartis [19] discovery gave detail knowledge on the application of the techniques from machine learning, statistical pattern recognition, and data mining to analyzing hypertext. It is essential to have knowledge of the emerging trends in content mining research. Zaiane & Han (2000), made a focus on resource recovery on Web. The authors transformed the unstructured data available on Web into a structured data by using database technology.

IV. TECHNIQUES

Markov Modal: In [12] web log feature is utilize to generate different orders of the web markov modal. Here as per user current user web page movement prediction of next page is done by utilizing morkov modal which give required page. Here as per the length of the user markov orders are use so storage of different size of markov modal help in different stages of the proposed work. In case of higher order markov modal if this fail then lower order markov modal will handle the situation and send the next possible page. So this step of finding the next page in lower order is continuing until possible next page is not obtained. In order to understand this consider an example, let us assume a user session s = *{P*1, *P*5, *P*6*}*, prediction of all-*K*th model is performed by consulting third-order Markov model. If the prediction using third-order Markov model fails, then the second-order Markov model is consulted on the session $x\_ = x - P1 = <P5, P6>$. This process repeats until reaching the first-order Markov model. Therefore, unlike the basic Markov model, the all-*K*th-order Markov model achieves better prediction [10], and it only fails when all orders of the basic Markov models fail to predict.

Predict_markov algorithm take session and modal number as input 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.

International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 02, Issue: 02 (MAR-APR, 2016)

There similar function take key_vector which is the collection of the keywords which is obtain 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 return to the function.

Computing HITS Algorithm [15]: In this algorithm two types of values are assigned on each page first is positive non zero weight and other is again a positive non zero hub weight. Here value of each weight are so assigned that by taking an square of the number it remain below or equal to 1, So proper normalization of each value is done. Here page have high weight is consider as the important page or rank of the page is higher as compare to other existing page. Numerically, the mutually reinforcing relationship between hubs and authorities can be expressed as follows: if p points to many pages with large ά values, then it should receive a large h –value. In similar fashion if p is pointed to by many pages with large h -values, then it should receive a large ά-value. This motivates the definition of two operations on the weights, denoted by I and O. Given weights a p and hp, the I operation updates the ά -weights as follows, similarly the O operation updates the h-weights as follows

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

Thus, I and O operations are the basic means by which hubs and authorities reinforce one another. To find the desired "equilibrium" values for the weights, one can apply the I and O operations in an alternating fashion, and see whether a fixed point is reached SALSA: In [13] a random walk algorithm is proposed where bipartite hubs and authorities web graph is develop, and then proper movement is note by changing the web pages one by one. Here some of important nodes are choosing for the start of random walk, selection of those are done randomly. Now movement in walk is done by switching from one hub node to another hub node. Here selection of nodes is depending on the authority weight which is distributed as per the importance of hub. So markov modal will calculate the required weight probabilities. Let Fu be the set of pages u points to and Bu the set of pages that point to u.

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

Multi-Damping Method: In [14] Let Y is an adjacency matrix for the graph of nodes. Where i represent the node after which j node is chosen by the surfers with (P') probability. 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 calculate which the damping coefficient is & 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 μ ∈ [0, 1] is the probability of choosing a link-accessible page. Alternately, with probability 1 − μ, 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 ξ0 =2/5 = 1− 3/5 , ξ1 = 2/4*3/5 = 3/55 (1 – 2/4), ξ2 = 2/4*2/5 = 3/5*2/4 (1 – 1/3) and ξ3 = 2/4*1/5 = 1/3*2/4*3/5. This clearly identifies μ1 = 1/3, μ2 = 2/4 and μ3 = 3/5 as the corresponding damping factors. M is damping factor = (μ1... 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 (z_K) :=($\xi_{0\ldots}$, ξ_{k-1}, ξ_k +1-s)

Zk ← add cor (Zk)

End if

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

$$\mu_{j=1} - \frac{1}{1+\frac{p_k - j+1}{1+\mu_{j-1}}}, j = 1 \ldots k,$$

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

K $\prod$ $\mathbf{G}(\mu_K - J + 1)v = \xi_k S^K v + p_k\text{-}1(S)v_{J=1}$

V. EVALUATION PARAMETER

In order to evaluate this work there are different parameter present for the different techniques. The best parameter which suit this work is the precision where it give the value which is a measure of the prediction which is correctly identify by proposed model to the all the logs pass in the experiment. The other important measure is the Recall and F-score. True Positive (TP): When the system says page P1 and actual page is also P1. True Negative (TN): When the system says page P1 and actual page is also P2. False Positive (FP): 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).

VI. CONCLUSIONS

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

REFERENCES

[1]. Brian D.Davison, "A Web Caching Primer" IEEE Internet Computing 2001.

[2]. J. Dom Enech, 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.

[3]. D. Duchamp. Pre-Fetching Hyperlinks. Proc. Of The 2nd Use nix Symposium on Internet Technologies and Systems, 1999.

[4]. P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The Query Graph: Model and Applications. In Cikm, 2008.

[5]. 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 Knowledge And Data Engineering, Pp. 502-513, 2013.

[6]. I. 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 Retrival (Sigir '13), Pp. 765-769,2013.

[7]. Mamoun A. Awad, 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.

[8]. R. Lempel, S. Moran, the Stochastic Approach for Link-Structure Analysis (Salsa) and the Tkc Effect, Proceedings of the 9th International World Wide Web Conference, 2000.

[9]. Giorgos Kollias, Efstratios Gallopoulos, Ananth Grama "Surfing the Network for Ranking by Multi-damping". IEEE Transactions on Knowledge and Data Engineering 2014.

[10]. J. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Journal of Acm (Jasm), 1999.