Engineering Universe for Scientific Research and Management
(International Journal)
ISSN(ONLINE) : 2319-3069
Vol. 7 Issue 7
July 2015

A Weighted Community Approach to Predict Http Behavior of End User

Ashwini Matange[1], Swati Dongre[2]
M. Tech – Second Year RITS[1]
Professor - CSE Department, RITS[2]
asmatange@gmail.com[1]

ABSTRACT

Rapid growth of web application has increased the researcher's interests in this era. All over the world has surrounded by the computer network. There is a very useful application call web application used for the communication and data transfer. An application that is accessed via a web browser over a network is called the web application. Web caching is a well-known strategy for improving the performance of Web based system by keeping Web objects that are likely to be used in the near future in location closer to user. The Web caching mechanisms are implemented at three levels: client level, proxy level and original server level. Significantly, proxy servers play the key roles between users and web sites in lessening of the response time of user requests and saving of network bandwidth. Therefore, for achieving better response time, an efficient caching approach should be built in a proxy server. This paper, use web page pre-fetching scheme have community detection [2], weighted rule mining concept [15] and Minkowski distance for fast and frequent web pre fetching .

Keywords

Web Services, Pre-fetching, Log file

1. INTRODUCTION

Web is a key resource in order to share the information along the world. It has large number of news, advertisements, global connectivity between people and lots of knowledge for the students. This massive use of Web or WWW makes it more important in the world of research. Researcher has the challenge to make the web applications more efficient. Many researchers work on it and give new idea in order to give the better results from the previous one. This dissertation is also puts its best foot forward in this era [1,12,13].

There is a huge need to improve the response time of server for web applications. Current Web has a massive repository due to increase its use suddenly. It has to focus on both the quality and quantity of web contents. Even, when the speed of Internet has improved with the reduced costs, the traffic is getting heavier. The enormous information makes it difficult to find the relevant information quickly. This led to the effort to improve the speed, by reducing the latency, make the web more relevant and meaningfully connected.[2,8,9] .

The Cache prefetching plays an important role in order to enhance the response time and make the application well-organized. The web prefetching is a technique in order to preprocess the user requests, before they are actually demanded. Therefore, the time that the user must wait for the requested documents can be reduced by hiding the request latencies. Prefetching is the method for reducing Latencies. The user always expects an interactive response, better satisfaction and quality of output. There are various approaches and algorithms have been proposed for improving the web performance [3,10,11].

The proposed work will use to predict fourth coming link to improve the user experiences and expedites users visiting speed. Predictive Web pre-fetching or link prediction refers to the method of deducing the upcoming page accesses of a client based on its past experience. In this work we demonstrate the frequent mining pattern which is obtain on the basis of input and on the basis of that caching and prefetching ratio is calculated. Thus we present a new idea for the interpretation of Web prefetching and web caching from the given usage items. The approach works on the basis is web mining with the combination of clustering approach.

This paper is divided into seven sections. First one is introduction in which give the brief description of work. The second section discusses the previous work related to the topic. The third section describes the approach used in the presented work. The next section describes the proposed architecture of the presented work. After this the simulation result has discussed. Finally paper concludes in the section eight's.

2. PREVIOUS WORK

Research over web mining and web pre-fetching is going very fast in last decades. Toufiq Hossain Kazi et.al [4] gives an Adaptive Resonance Theory (ART) based on pre-fetch technique namely ART1, use the bottom-up and top-down weights of the cluster-URL connections obtained from a modified ART1 algorithm to make pre-fetching decisions. A.B.M.Rezbaul Islam et.al [6] proposed a new and improved FP tree with a table and a new algorithm for mining association rules. This algorithm mines all possible frequent item set without generating the conditional FP tree. It also provides the frequency of frequent items, which is used to estimate the desired association rules, Whereas P. Sampath et.al[5] present an weight estimation process with span

time, request count and access sequence details. The user interest based page weight is used to extract the frequent item sets. Systolic tree is used to arrange candidate sets with frequency values. Due to the limited size of the systolic tree, a transactional database must be projected into smaller ones each of which can be mined in hardware efficiently. A high performance projection algorithm which fully utilizes the advantage of FP-growth is proposed and implemented. It reduces the mining time by partitioning the tree into dense and spare parts and sending the dense tree to the hardware. Systolic tree based rule mining scheme is enhanced for weighted rule mining process. Automatic weight estimation scheme is used in the system. With explosively growing number of Web contents including Digitalized manuals, emails pictures, multimedia, and Web services require a distinct and elaborate structural framework that can provide a navigational surrogate for clients as well as for servers. Due to the increasing amount of data Available online, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. So Sekhar Babu Boddu et.al [7] presents an introduction of Web mining as well as a review of the Web mining categories. Then we focus on one of these categories: the Web structure mining. Within this category, we introduce link mining and review two popular methods applied in Web structure mining: HITS and Page Rank.

3. PROPOSED ARCHITECTURE

The objective of the design process is to produce a set of detailed specification that describes the intended form of implementation for the software system. Software design is a process of problem-solving and planning for a software solution. The software system design is produced from the results of the requirement phase. Design can be regarded as a form of problem solving process that involves making extensive use of abstraction, including the separation of the logical aspects of the design from the physical aspects. This chapter consists of the algorithm or the method which is used in the project, the algorithm used, flowchart to represent the method and the description of the flowchart.

The increasing popularity of the World Wide Web in recent years, the substantial burden is imposed on the traffic on the Internet. The World Wide Web is considered as a distributed information system must provide access to shared data. The bulk of the research conducted to improve the response time of Web-based information system as distributed on geographical location. 'S of web caching and pre-Getting the two main approaches used in the response time significantly reduce user perception. Pre-caching scheme is ideal Getting prediction system can be made to the next (next issue) and applications pre-load the cache .The are pre-fetched objects stored in a local cache to reduce latency. This paper presents a study of algorithms for handling Web caching and pre-receipt.

Proposed web page pre-fetching scheme use community detection [2], weighted rule mining concept [15] and Minkowski distance for fast and frequent web pre fetching .Proposed Scheme use community detection concept for finding frequent page's efficiently without candidate set generation, whereas weighted mining concept are used to apply relative weight over each transaction after session and user identification . Minkowski use to assign that that relative weight over their relative position in transaction.

Assumption

H_L=*HTTP Server log file*
$H^R{}_L$ =*Number of Row in* $H^R{}_L$
$H^A{}_L$ = *Number of Attribute in* $H^R{}_L$
$R^W_{Yi,Xj}$ = *Relative weight Xj with respect to Yi*

Algorithm

{
For (i=1 to i<= $H^R{}_L$***)***
{
For (j=1 to j<= $H^A{}_L$***)***
{
$H^{R\,i}_{L\,j} = \frac{dH_{L_i}}{dj} =$
Token seperation with special charecter amplifier
Insert token $H^{R\,i}_{L\,j}$ *in log table in databaseas Ith row and Jth column*
}
}
For (i=1 to i<= $H^R{}_L$***)***
{
For (j=1 to j<= $H^A{}_L$***)***
{
Filter $H^R{}_L$ *record with supporting files*
}
}
For (i=1 to i<= $H^R{}_L$*)*
{

Case 1: $-$ *if* H^{Ri}_{IP} *not in list(distinct User)*
Add H^R *in list (distinct user)*

Case 2 : $-$ $H^{Ri}_{IP\ and\ \ os}$ *not in list(distinct User)*
Add H^R *in list (distinct user)*

Case 3: $-$ $H^{Ri}_{IP\ ,os\ and\ browwser}$ *not in list(distinct User)*
Add H^R *in list (distinct user)*

Case 4: $-$ $H^{Ri}_{IP\ ,os,browwser\ \ and\ \ referal\ \ uri}$ *not in list(distinct User)*
Add H^R *in list (distinct user)*

End case
delete H^R *from HTTP Log*
}
For (i=1 to i<=C)
{

- Evaluate $p(xi^{c_\beta}, yi^{c_\beta})$
- Evaluate $p(\text{comm}_{xi,yi})$
- Evaluate

$$p(xi^{c_\beta}, yi^{c_\beta} | \text{comm}_{x,y}) = \frac{p(\text{comm}_{xi,yi} | xi^{c_\beta}, yi^{c_\beta})\, p(xi^{c_\beta}, yi^{c_\beta})}{p(\text{comm}_{xi,yi})}$$

}}

$$d(i,j) = ((|x_{i1}| - |y_{j1}|)^p + (|x_{i2}| - |y_{j2}|)^p + \cdots .. (|x_{in}| - |y_{jn}|)^p)^{1/p}$$

- Evaluate

$$S_{xi,yi} = d(i,j) * \frac{p(\text{comm}_{xi,yi} | xi^{c_\beta}, yi^{c_\beta})\, p(xi^{c_\beta}, yi^{c_\beta})}{p(\text{comm}_{xi,yi} | xi^{c_\alpha}, yi^{c_\beta})\, p(xi^{c_\alpha}, yi^{c_\beta})}$$

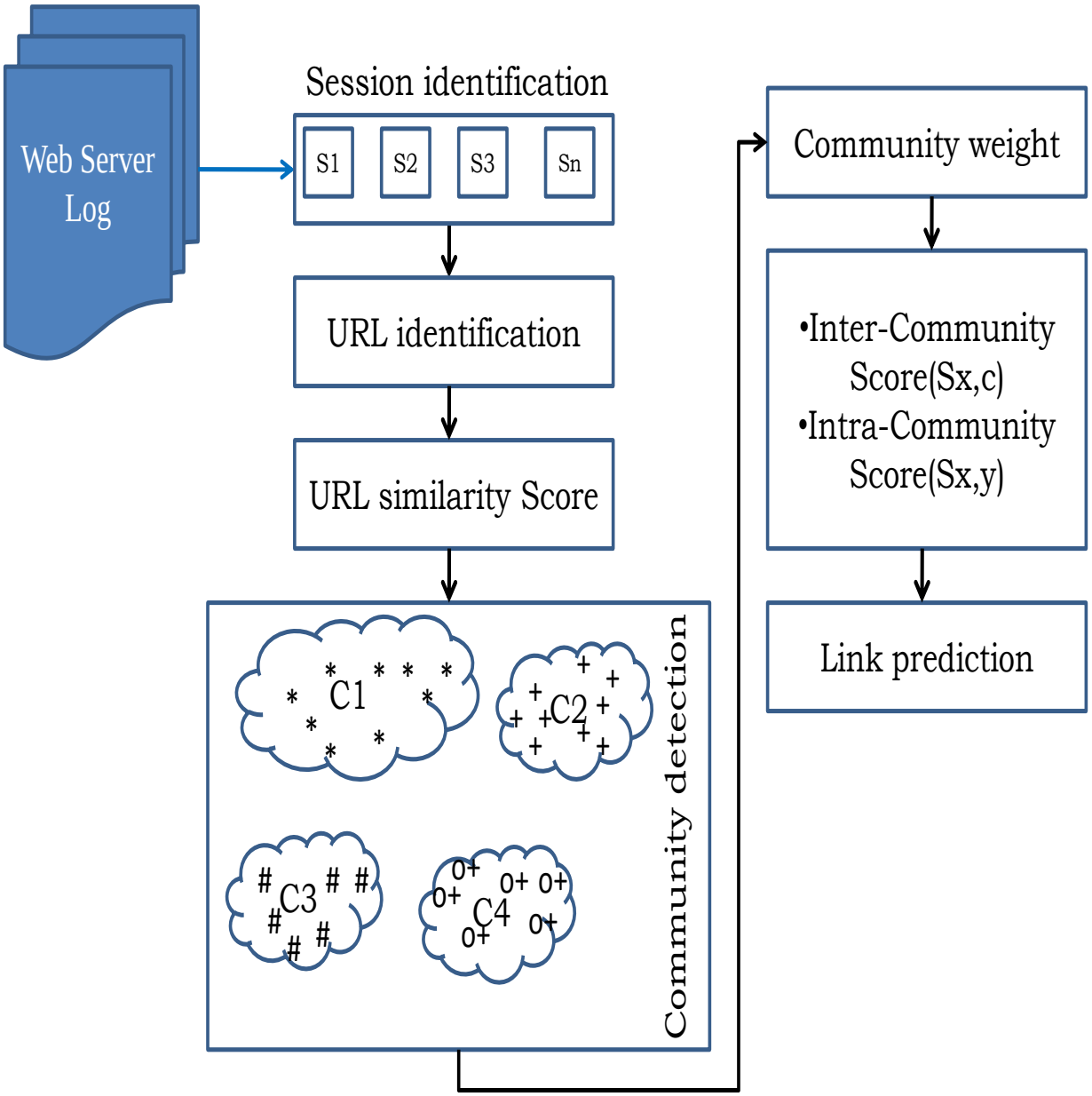

Figure: 1 Proposed Architecture

SIMULATION AND RESULTS

For simulation and result analysis a real time scenario of client server architecture having 30 clients and one server is taken as a scenario for verification of proposed work whole verification is done over MATLAB 10 and used My Sql for data base support.Here only define some steps which was necessary at time of implementation. The work is implemented in a single MATLAB file. In this work there is need of a web log file

- **Generate web log file:** In this work there was a small web site implemented in php in order to generate web log has used. After running this web site web has generate on web server.
- **Log Pre-Processing**- Along with that important information there is also some inconsistent data like noise, null value and other error information which is not so important for web personalization so in order improve web mining result its need to refine web log file before mining. Data cleaning, user and session identification, data integration and so on are main important part of log pre processing.
- **Data Cleaning:**- In proposed methodology data cleaning process use to recognized useful token and remove unwanted and redundant token then store in data base after normalization .
- **User & Session identification**:- User and session identification is very important step towards web personalization generally IP address is used to distinguish but when there is an proxy server then number of user having same IP address then some more attribute like browses information ,operating system and Refer URI field is used as per concern.
- **Community detection**: - Proposed algorithms used cluster based [2] concept to generate similar page cluster.
- **Community weight** -Proposed algorithms use Weight rule concept to assign relative weight to each page with respect to each other page. This relative weight is use to represent probability of page p request just after page q .
- **Minkowski distance** :- proposed algorithms use Minkowski distance to assign that that relative weight over their relative position in transaction. M is total number of unique transaction /page that had be identify in above step. for example relative weight of p wrt q is store at qth row and pth column then at time pre fetching if q page is called then at qth row the page having highest relative weight is to be pre fetched with page q.

In our simulation, we randomly divide the links of the original network into the train set, Etrain ,and the test set, Etest, in order to test the prediction accuracy of our method. This has been introduced in section 5.4. The train set contains 90% of links in e in the original network, and the remaining 10% of links are in the test set Etest. Here, we apply the overall metric, accuracy, to evaluate the prediction on the basis of following parameter.

- **True Positive Rate (TPR) or Sensitivity**

True positive rate tells about that how many times the object has been identified correctly

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})} \quad (1)$$

- **True Negative Rate (TNR) or Specificity**

The true negative rate tells about that how many times the object has been correctly rejected

$$\mathrm{TNR} = \frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FN})} \quad (2)$$

- **False Positive Rate (FPR)**

The false positive rate tells about that how many times the object has been incorrectly identified

$$\mathrm{FPR} = \frac{\mathrm{FP}}{(\mathrm{TP}+\mathrm{FP})} \quad .3)$$

- **False Negative Rate (FNR)**

The false negative rate tells about that how many times the object has been incorrectly rejected

$$\mathrm{FNR} = \frac{\mathrm{FN}}{(\mathrm{TN}+\mathrm{FN})} \quad (4)$$

- **Accuracy:** $Accruacy = \frac{TP+TN}{TP+TN+FP+FN} * 100$

Table 1: Accuracy Comparison

Page sequence	Community Approach Accuracy	Weighted Community Approach Accuracy
1	0.8556	0.8651
2	0.9607	0.9674
3	0.9639	0.9675
4	0.9017	0.917
5	0.8982	0.9221
6	0.9194	0.9275
7	0.8966	0.9166
8	0.9599	0.9728
9	0.9843	0.9893
10	0.9795	0.983

In experiment 1, we can easy predicate that by using our proposed approaches having higher accuracy .Whereas with the help of community detection and Minkowski distance along with relative weight concept (TPR,TNR,FPR and FNR) shows their maximum value

.

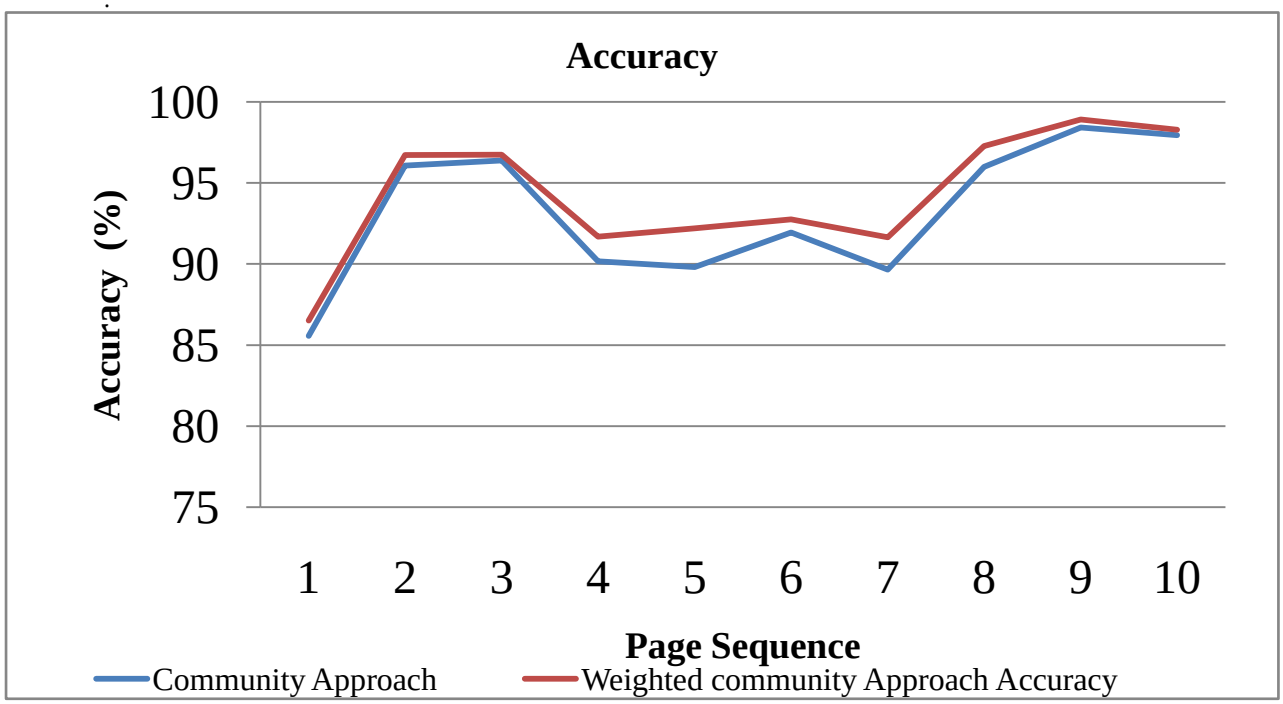

Figure 2:- Accuracy graph

However, in general, the method proposed in this paper is effective to pre-fetching web in the web environment. From this experiment 1 we conclude that our approach gives better method for the classification of the data as well minimum TPR,TNR,FPR and FNR. However, in general, the method proposed in this dissertation is effective to pre-fetching web in the web environment.

4. CONCLUSION

This Paper proposed a method using community detection and Minkowski distance along with relative weight concept in order to apply the pre-fetching in the web environment. Tests that have been conducted in this proposed work using the Minkowski distance shows that it gives better results as compare to previous work ie having higher accuracy as compare to previous one. The implementation also shows that it is easy to apply in order to pre-fetch the page of a web site.

5. REFERENCES

[1] R. Kosala and H. Blockheel, "Web Mining Research: A Survey", In SIGKDD Explorations, Volume 2, Number 1, pages 1-15, 2000.

[2] Fenhua Li , et.al "A Clustering-based Link Prediction Method in Social Networks" in ICCS . 14th International Conference on Computational Science, Elsevier-2014

[3] S. Chakrabarti, "Data mining for hypertext: A tutorial survey". ACM SIGKDD Explorations, 1(2):1-11, 2000.

[4] Toufiq Hossain Kazi, Wenying Feng and Gongzhu Hu, "Web Object Prefetching: Approaches and a New Algorithm", IEEE 2010, pp 115-120.

[5] P. Sampath, C. Ramesh, T. Kalaiyarasi, S. Sumaiya Banu and G. Arul Selvan, "An Efficient Weighted Rule Mining for Web Logs Using Systolic Tree", IEEE 2012, pp 432-436.

[6] Nizar R. Mabroukeh and C. I. Ezeife, "Semantic-rich Markov Models for Web Prefetching", IEEE 2009, pp 465-470.

[7] A.B.M.Rezbaul Islam and Tae-Sun Chung, "An Improved Frequent Pattern Tree Based Association Rule Mining Technique", IEEE 2011.

[8] Brijendra Singh and Hemant Kumar Singh, "Web Data Mining Research: A Survey", IEEE 2010.

[9] Kavita Sharma, Gulshan Shrivastava and Vikas Kumar, "Web Mining: Today and Tomorrow", IEEE 2011, pp 399-403.

[10] WANG Yong-gui and JIA Zhen, "Research on Semantic Web Mining" IEEE 2010, pp 67-70.

[11] R.Agrawal, and R.Srikant, "Fast algorithms for mining association rules", In VLDB'94, pp. 487-499, 1994 Borges and M. Levene,"A dynamic clustering-based markov model for web usage Mining", cs.IR/0406032, 2004.

[12] Zhu, J., Hong, J. and Hughes, J. G. (2002a) Using Markov Chains for Link Prediction in Adaptive Web Sites. In Proc. of Soft-Ware 2002: the First International Conference on Computing in an Imperfect World, pp. 60-73, Lecture Notes in Computer Science, Springer, Belfast, April.

[13] K.Ramu, Dr.R.Sugumar and B.Shanmugasundaram "A Study on Web Prefetching Techniques" Journal of Advances in Computational Research: An International Journal Vol. 1 No. 1-2 (January-December, 2012)