ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Effective Web Log Mining using Combination of K-NN Algorithm and Apriori Algorithm.

Mohammad Rafey*, Mani Kant, and Mohana Prasad K.

Department of Computer Science Engineering, Sathyabama University, Chennai, Tamil Nadu, India.

ABSTRACT

Web log Mining is known as the application of Data Mining which is used to generate certain pattern of World Wide Web and online navigational pattern are certainly crucial for generating up of websites. We start web log mining with data cleaning process and finish the task by finding some encapsulated knowledge. Our beneficence in this project hold four distinct components, first we are generating a graph for each users. Second we are showing how much time user has spent on particular sites for session identification. Third we are using density based algorithm for navigational pattern prediction and finally, we are generating related item base search for each user.

Keywords: web log mining, graph, session identification, navigational pattern prediction, item base search

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Now a day's many companies and Organizations are depending mostly on Websites to communicate with their client's. Keeping present clients and eye catching potential in mind which push these organizations and companies to create their websites more functional and operative. To attain this task, some review works to be performed. We can perform this review task in two optional ways. First, we can ask specific website user about their browsing experience and depend on the review which we received we can improve website structure. Second, client's navigational history is automatic recorded and analyzed and accordingly tuning of website is done. From the above two mentioned review task second one is the best option, because it doesn't depend on client's manual input.

Our beneficence in this project hold four distinct components, first we are generating a graph for each particular users, using KNN and Apriori Algorithm. Two graphs will be generated one on user side and second one on admin side, using these graph both user and admin can identify which site they have visited more number of times. Secondly we are displaying time which means how much time particular user have spent on that particular sites. Finally we are displaying related item search for user, for example if user search for online shopping then all the sites related to online shopping will be displayed to the user. For all the above mentioned task we are using KNN Algorithm and Apriori Algorithm.

PROPOSED SYSTEM

In our proposed work we generating two graph one on client side and other on admin side using Apriori algorithm and KNN Algorithm. In client side, client can view whatever website they have visited recently and how many times they have visited that particular website based on these one graph will be generated on client side and on admin side overall graph will be generated for all users, admin will have the record for all the website which user have visited and based on these one more graph will be generated on admin side which client's can't access.

Apriori Algorithm: Apriori is known as Seminal Algorithm. This algorithm uses central knowledge of regular item sets properties. In this algorithm K- item sets are used to explore (K+1)-item sets. First, we will scan the database for the set of regular item sets by doing this we can count number of each item in database, and compose those items which satisfy minimum number of support. The resulted set will be denoted as L1.

Next, using L1 we will find L2 that is the set regular second item sets, using which again we will find L3 and it will go so on, until we find no more regular k-Item sets. To find each L(K) it requires full database scan, to improve this we are using Apriori Property using this Apriori Property action we can reduce Search space. But still it will consume more number of time for scanning full database if incase database is large or content more number of data, so to overcome this issue we are using apriori property that is "The Join Step and Prune step" this two property action help us to reduce the number of time consume to scan the entire database.

Apriori Algorithm

$F_{(i)}$=(Frequent item sets of cardinality 1);
for(k=1;$F_{(k)}$!=0,K++)do begin
$C_{(k+1)}$=apriori-gen($F_{(k)}$);//New candidates
for all transaction t belong to Database do
begin
$C'_{(t)}$=subset($C_{(k+1)}$);//candidates contained in t
for all candidates c belong to $C'_{(t)}$ do
c.count++;
end
$F_{(k+1)}$={C belong to $C_{(k+1)}$ |c.count>=minimum support}
end
end

ISSN: 0975-8585

K-Nearest-Neighbor(K-NN) classifiers: This classifiers is mostly used in Pattern recognition; it is basically based on knowledge by analogy, which means comparison of test tuple with other training tuples which are similar to it. There is a space called n-dimensional pattern space in which almost every training tuples will be stored. When we give an tuple which is unknown, K-NN Classifier will search for a pattern space for training tuple k which is nearest to the unknown type.

K-Nearest-Neighbor Algorithm
Input: D, the set of k training objects, and test objects z=(x',y') **Process:** Compute d(x',x),the distance between z and every object,(x,y) belong to D Select D_z subset of D,the set of k closest training object to z. **Output:**y'=argmax summation$_{(x,y)belong\ to\ D}$, I(v=y)

Use Case Diagram

In figure 1 we have shown use case diagram for user and admin side, in the following diagram we have six cases i.e Register, Login, Internet Activity, search. Store database and finally generation of graph. Out of these six cases user have accessible to only four cases i.e. Register, login Internet activity and search and admin will have accessible to three i.e. Login, Store database and graph generation.

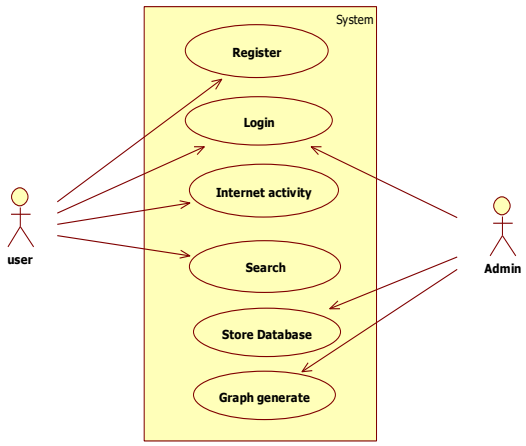

Figure 1: Use Case Diagram

Module Description

Graph Generation: In this module the graph is generated on user and admin side using KNN Algorithm . Two graphs will be generated one on user side and second one on admin side, using these graph both user can identify which site they have visited more number of times and admin can view which site is overall viewed more number of times by the user. Graph is divided into two axis X-axis and Y-axis; on X-axis Number of site user viewed will be display and on Y-axis number of times user have viewed those site will be visible.

"Generating of Max height graph"

For generating maximum height in a bar graph we are using a logic that if the new value of particular bar is greater than the previous one then it will replace the old bar diagram.

```
findMax : function(columns)
{var result = 0;
for (var i in columns)
{
if (columns.hasOwnProperty(i))
 {
```

ISSN: 0975-8585

```
var max = 0;
columns[i].forEach(function(value)
{                                                    max += value.value;});
if (max > result)
{
result = max;}
return result;
}
```

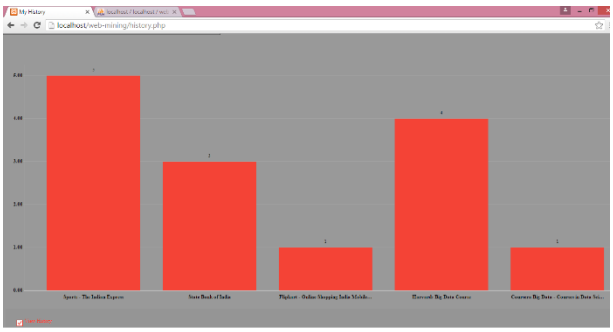

Fig 2: Admin side graph generation.

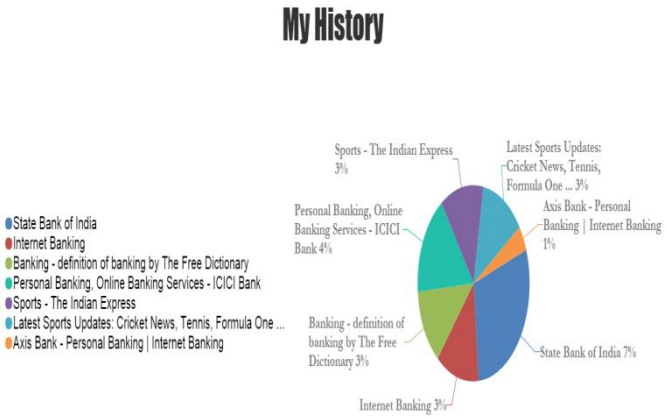

Fig 3 : User side graph generation

Display of date and time : In this module using Apriori algorithm we are displaying the time spent by user on particular site. This module will display when the user has access that particular site.

My History

Content	Views	Last Viewed(DATE&TIME)
State Bank of India	7	2016-03-09 10:36:05
Internet Banking	3	2016-03-09 10:36:30
Banking - definition of banking by The Free Dictionary	3	2016-03-09 10:36:41
Personal Banking, Online Banking Services - ICICI Bank	4	2016-03-09 10:36:35
Sports - The Indian Express	3	2016-03-09 10:37:26
Latest Sports Updates: Cricket News, Tennis, Formula One ...	3	2016-03-09 10:37:09
Axis Bank - Personal Banking | Internet Banking	1	2016-03-09 10:36:47

Fig: 3.1

More number of views : In this module using Apriori Algorithm we are displaying number of views which means which site user have viewed more number of time by using Apriori algorithm, same as apriori algorithm logic it will scan entire database and then it will be displayed the total number of views for that particular sites.

ISSN: 0975-8585

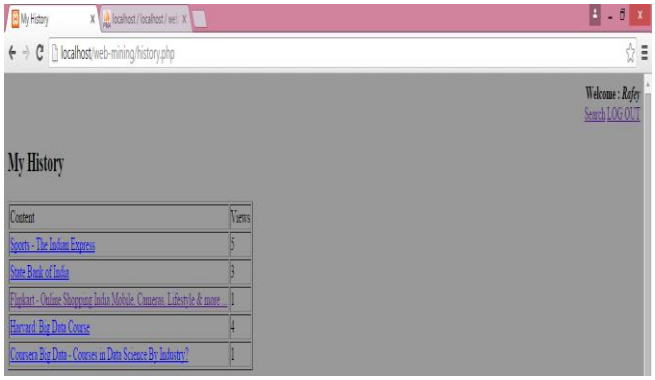

Fig: 3.2

Related item search: In this module using apriori algorithm we are finding the related item search for each user, here apriori algorithm is going to scan entire database and when user will search something it is going to display related item search for user.

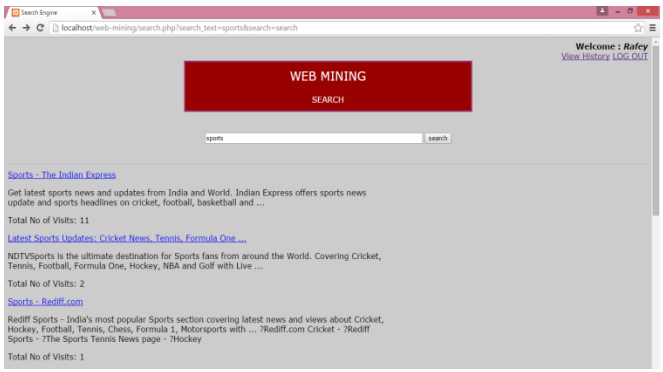

Fig 4: Related item search

Collaboration diagram

In Figure 2 we have shown a collaboration diagram for Graph generation in which user first have to register and then he can login to there id's after login user can search for particular website, what ever user will search in the search engine that all will get stored in the database next time when user will login again then they can view graph generation for there search .In this collaboration diagram we also admin part admin will register and login to there id and then they can view overall graph generation because they have the accessible to database where user won't have accessible to database for security purpose.

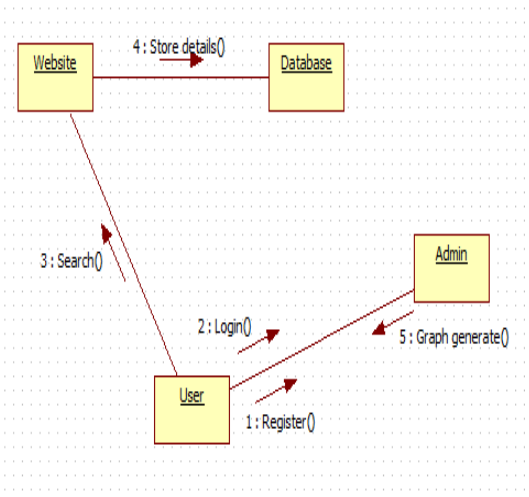

Figure 5: Collaboration Diagram for Graph generation

CONCLUSION

In this project we are generating graph and pi-chart on Admin and user side. In the existing module only search technique has been implemented, so we overcome this issue and generating graph in our manuscript. We are implementing our project using K-NN Algorithm and Apriori algorithm. We are using K-NN Algorithm for Graph and pi-chart generation and Apriori algorithm for Searching technique. The Main advantage of this project is generation of Pi-Chart and Graph generation.

REFERENCES

1. Mohana Prasad .k, Dr. Sabitha .R, "Meta Physical Algorithmic Representation for Flawless Clustering" Journal of Theoretical and Applied Information Technology *(JATIT)*, ISSN : 1992-8645, Volume 76,page no:82-87.
2. Mohana Prasad .k, Dr. Sabitha .R,"Evolution Of An Algorithm For Formulating Efficient Clusters To Eliminate Limitations**"** *International Journal of Applied Engineering Research (IJAER)*, ISSN 0973 – 4562, volume 9,issue 23,pp:20111=20118.
3. Mohana Prasad .k, Dr. Sabitha .R," Yoking of Algorithms for Effective Clustering", Indian Journal of Science and Technology,ISSN: 0974-6846 volume 8(22),IPL0269,sepetember 2015,pp1-4
4. Agosti .M, Nunzio .G.M.D, Web log mining: a study of user sessions, in:Proceedings of the 10th DELOS Thematic Workshop on Personalized, June2007.
5. Ankerst .M, Breunig .M, Kriegel .H,Sander .J, OPTICS: ordering points to identify the clustering structure, Sigmod Record 28 (2) (1999) 49–60.
6. Berendt .B, Mobasher .**B**, Nakagawa .M, Spiliopoulou .M, The impact of site structure and user environment on session reconstruction in web usage analysis, Webkdd 2002 – Mining Web Data For Discovering Usage Patterns and Profiles (2003) 159–179.
7. Berendt .B, Mobasher .B, Spiliopoulou .M, Wiltshire .J, Measuring the accuracy of sessionizers for web usage analysis, in: Paper presented at the Workshop on Web Mining, First SIAM Internat. Conf. on Data Mining, Chicago, IL, 2001.
8. Catledge .L, Pitkow .J, Characterizing browsing strategies in the world-wideweb,Computer Networks and Isdn Systems (1995) 1065–1073.
9. Chen M,Park .J,Yu .P, Data mining for path traversal patterns in a web environment, in: Proceedings the 16th International Conference on Distributed Computing Systems, 1996, pp. 385–392.
10. Cooley .R, Mobasher .B, Srivastava .J, Data preparation for mining world wide web browsing patterns, Knowledge and Information Systems. 1 (1) (1999).
11. Eirinaki .M, Vazirgiannis .M, Web mining for web personalization, ACM Trans Inter Tech. (3) (2003) 1–27.
12. Ester .M, Kriegel H.P, Sander .J, Xu .X, A density-based algorithm for discovering clusters in large spatial database with noise, in: Proceedings of the 2nd Int. Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August 1996.