Clustering Model Based on Web Behavior

[1]Milind Makhamle, [2]Mohammad.A.Mansuri, [3]Mohd.Tanvir Shaikh, [4]Er. Zainab Pirani
[1,2,3]Department of Computer Engineering
M.H.Saboo Siddik College of Engineering
Mumbai,India

***Abstract*— Web log mining is an emerging part of data mining. It provides invaluable information by discovering trends and regularities in web user's access patterns. Clustering based on access pattern is an important research topic of web usage mining. Knowledge obtained from web user clusters has been used in different fields of web mining technologies. This paper presents an algorithm for measuring similarities and automated segmentation of web users based on their past access patterns. The compatibility measures are based on content extracted from user's browser data. Furthermore it also provides a locality based clustering method for the people who are unknown to their most compatible friends.**

Keywords—Cluster, Model, Web, Activity, Report, Interest, Location, Compatibility, Matrix, Data Mining, Fuzzy Clustering, GeoIP , LDA, Sessions, History, Snippet, Links, IP.

I. INTRODUCTION

The Rapid web development and the increased number of available web searching tools push more and more organizations to put their information on the web and provide web-based services. In meantime, the continuous growth in the size and use of the Internet is increasing the difficulties in searching log information. Reductions on the Internet traffic load and user access cost is therefore particular important.

Cluster Model based on Web behavior is a user clustering technique based on user's browser history. Any browser the user uses has the functionality to store the data in the form of links. User clustering techniques has been applied on various social networking sites based on the data stored in the user profiles. As we look towards the clustering technique used by these sites the data stored in the users profile is edited by the user only. The information the user provides on the internet can be faked and may lead to clustering of users with fake information. Our project provides a way to avoid such clustering. The project deals with the users browsing pattern to recognize its interest,habits,plans, etc which is difficult to fake. It demonstrates that the benefit of using these techniques depends considerably on input data and on user's browsing habits.

There has been an increased demand for understanding of web-users due to the Web development and the increased number of web based applications. Based on different criteria, web users can be clustered and useful knowledge can be extracted from web user access pattern[7].Many applications can then benefit from the knowledge obtained.

II. EXISTING SYSTEM

There are various systems used over the internet for user clustering based on the search and browsing patterns of the users over the Internet. Some of the well knowing examples of such clustering techniques being used are Facebook Friend Suggestions, Twitter's "People you may know", Tagged's "Meet Me" and other Recommender Systems. The main disadvantage of such systems is that the information provided by the user can be faked which results in incorrect suggestions to the user. Thus the efficiency of a system is decreased and may endanger the user's life.

In this paper, we introduce an algorithm for data processing and clustering that will majorly depend on the user's web browsing pattern which will increase the efficiency and accuracy of a system to a greater extent.

III. PROPOSED SYSTEM

The Cluster Model designed can be represented in the given system block diagram. The System Block Diagram consists of the following blocks has been shown in fig 1. The proposed system shows the whole overview of our project which gives us the information about how the system is going to work throughout. It also shows us the input to the given system which is comparatively less with the existing systems in the market.

The input of the block diagram of our proposed system is divided into two fields that is IP address and the links visited from the user. This input fields illustrate that by considering the IP address of a user and the links through which the user visited can be taken into consideration for calculating the compatibility of a user. The system goes through various operations like Interest Detection, Clustering Sessions, User Interest Rating(UIR) calculation, Developing Interest Cluster Model(ICM), Interest Matrix Calculation, Distance by GeoIP, Location Matrix Calculation, Compatibility Matrix Calculation and Clustering Users. At the end the output produced by the system will display through various graphs like Interest Graph, Distance Graph, Compatibility Graph and Cluster Report. The System Block diagram is shown below.

Fig 1. System Block Diagram

A. *IP Address*

IP address acts an Input to the cluster model. IP address is used to determine the user identity and find the Location using GeoIP technique. IP address is stored in the database. Since IP address are unique to each user it can act as a primary key for determining the user.

B. *Links*

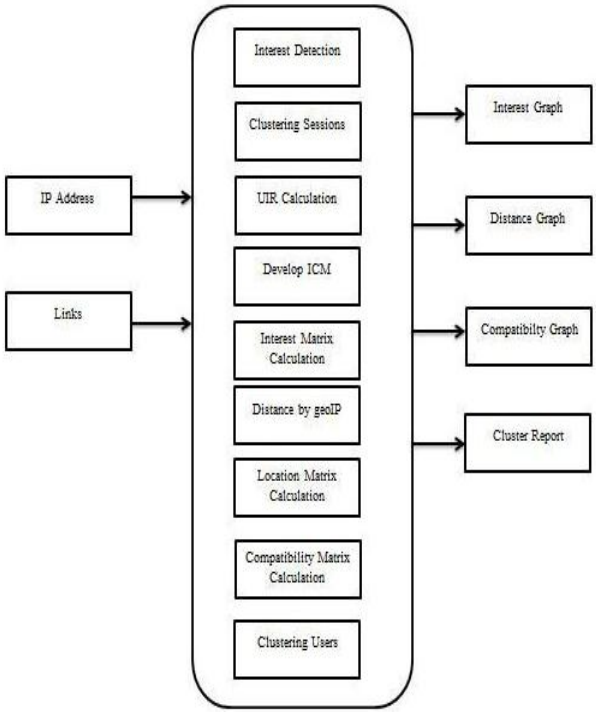
Links also acts as input to the cluster model. Links is a data set which can be represented as follows:

Links (url, snippet, timestamp, count)

Where, snippet is the combination of title and summary and url is uniform resource locator. [3]

C. *Interest Detection*

Interest detection is technique used to determine the Interest from user links more accurately the snippet of the links. Snippets are stored for determining the interest topic of each link. With this purpose, Latent Dirichlet Allocation (LDA) model is applied, which is an unsupervised machine learning method to identify latent topics from large data sets. [6].



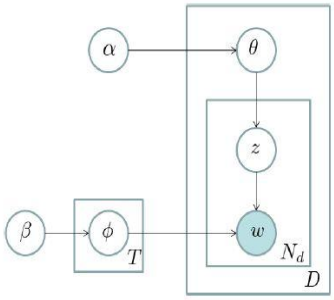


Fig 2. Graphical model of LDA

α is the parameter of the Dirichlet prior on the per-document topic distributions,

β is the parameter of the Dirichlet prior on the per-topic word distribution,

$\boldsymbol{\theta_i}$ is the topic distribution for document i,

ϕ_k is the word distribution for topic k,

z_{ij} is the topic for the jth word in document i, and

w_{ij} is the specific word.

JGibbLDA is a Java implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling technique for parameter estimation and inference. The input and output for JGibbLDA are the same format as GibbLDA++[3].

D. *Clustering Sessions*

Latent topics determined using LDA, are stored in the database and the links from which it is determined are categorized into sessions. Sessions are created to avoid multiple users on same browser. Clustering sessions helps us to avoid such problems and create ICM[4].

E. *UIR Calculation*

UIR stands for User Interest Rating it is value which ranges between 0 and 10. 0 <UIR<10

UIR is calculated based on the following formula,

No. of links visited (interest links) * 10 / Total no of links visited (1)

F. Develop ICM

ICM i.e. Interest Cluster Model is developed for each user which represents the user interest and their UIR for each user. Interests in ICM are sorted based on the UIR which represents the top most interest in ascending order[1].

G. *Interest Matrix Calculation*

ICM developed for each user is used for matrix calculation. Similarity based fuzzy clustering algorithm is applied to determine the similarity of two users. This algorithm uses the UIR of interest for any two users to determine the interest index. It uses the following formulae

$$\sigma(\mu_1,\mu_2)=\frac{|\mu_1\cap\mu_2|}{|\mu_1\cup\mu_2|} \quad (2)$$

Where $(\mu_1\cap\mu_2)(p)=\min\{\mu_1(p),\mu_2(p)\}$ (3)

And $(\mu_1 \cup \mu_2)(p) = \max\{\mu_1(p), \mu_2(p)\}$ (4)

μ_1, μ_2 are the interest index. [2]

H. Distance by GeoIP

Users IP address is use to determine the distance between any two users. Determining the distance between two users is done by the GeoIP technique. This technique uses user IP address to determine the longitude and the Longitude of the user. This coordinates are used to determine the distance by the radial distance formula.

I. Distance Matrix Calculation

Distance calculated by the GeoIP technique is stored in the Location matrix. Range conversion is done to limit the range of the distances calculated between 0 and 1. The formula for range conversion is

New Value (α) = 1 – ((Value - Min)/ (Max - Min)) (5)

Range conversion helps to convert long distance range to small range and then it is store in the Location Matrix.

J. Compatibility Matrix Calculation

Compatibility matrix is calculated as an average of Interest Matrix and Location Matrix. Compatibility matrix will determine the compatibility between any two users.

K. Clustering Users

K- Mean clustering is applied on the Compatibility matrix for each User to distinguish users in following category viz, Less Compatible, Moderate Compatible, More Compatible. This will help us determine the compatible users. [5].K- Mean clustering is applied on the Compatibility matrix for each User to distinguish users in following category viz, Less Compatible, Moderate Compatible, More Compatible. This will help us determine the compatible users. [5].

L. Interest Graph

K- Mean clustering is applied on the Compatibility matrix for each User to distinguish users in following category viz, Less Compatible, Moderate Compatible, More Compatible. This will help us determine the compatible users. [5].Following is an example which will show you the calculations of the interest graph, distance graph, compatibility graph and cluster report for a user U1.

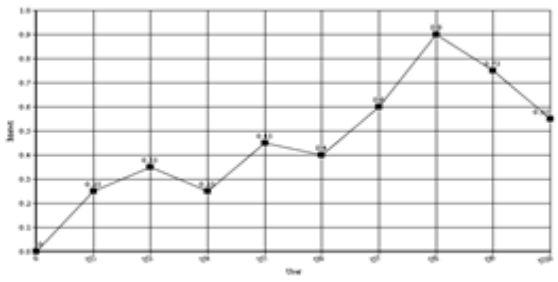


Fig 3. Interest Graph for U1

M. Location Graph

Graph will be generated based on the Location Matrix which will determine the distance for any user.x-axis of the graph will represent the user set and the y-axis represent the range between 0 and 1, where 1 is the closest and 0 is the farthest[6].

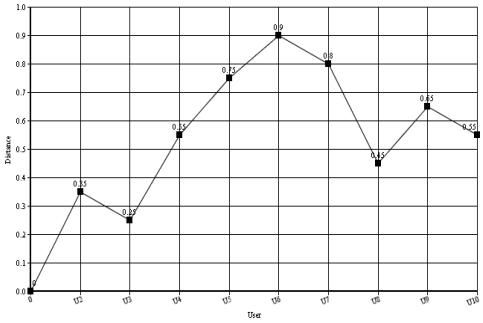


Fig 4. Location Graph for U1

N. Compatibility Graph

Graph will be generated based on the Compatibility Matrix which will determine the distance for any user. x - axis of the graph will represent the user set and the y – axis the range between 0 and 1 , where 1 is most compatible and 0 is no compatible[6].

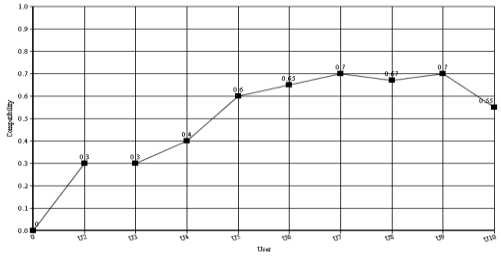


Fig 5. Compatibility Graph for U1

The above fig 5. shows you the compatibility graph for the user U1.The above given graph is plotted using the average values of the interest graph and the distance graph. Since the users having the same interest and within a local vicinity of user 1 will be more compatible, the users having same interest but are at a farther geographical location will be moderate compatible and the users having same interest around the globe will be having the less compatibility.

O. Cluster Report

Cluster report will be generated based on the K – mean clustering technique applied on the Compatibility Matrix. The report of a particular user will contain the set of user id divided in the following groups viz. Less Compatible, Moderate Compatible, More Compatible. This report will be stored in the database in the form of file. The report will contain 1 tables of above groups containing the User ID of different users. The report will be saved in the database. The following clustering report for user U1 is generated on the basis of compatibility graph as shown above.

The Table 1 given below will give us an overview of all the users that are compatible with the user U1 and the level of compatibility as well.

Table 1: Cluster Report

More Compatible	Moderate Compatible	Less Compatible
U6	U4	U2
U7	U5	U3
U8	U10	
U9		

IV. CONCLUSION

Clustering model technique thus improves the efficiency of a system through browsing history extraction which helps user to get a better compatibility according to its user interest as well as the location. Thus such an application can be used in various social networking or match making sites.

REFERENCES

[1] Bin Tan, Yuanhua Lv and ChengXiang Zhai. Mining long-lasting exploratory user interests from search history. Department of Computer Science, University of Illinois at Urbana-Champaign.

[2] Giovanna Castellano, A. Maria Fanelli, Corrado Mencar and M. Alessandra Torsello. 2007. Similarity-based Fuzzy clustering for user profiling. Computer Science Department, University of Bari, Italy.

[3] X. H. Phan, L. M. Nguyen and S. Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of WWW '08, 91-100.

[4] Xiao, J. & Zhang, Y. 2001. Clustering of web users using session-based similarity Measures. Proceedings of 2001 International Conference on Computer Networks and Mobile Computing.Beijing,China. IEEE. 223-228.

[5] D. M. Blei, A. Y. Ng and M. I. Jordan. 2003. Latent dirichlet allocation. In Journal of hine Learning Research, 3, 993-1022.

[6] J. A. Bondy, U. S. Murty. 2008. Graph Theory. Graduate Text in Mathematics Series SN: 0072-5285, USA, Spingers

[7] C. Shahabi, A. M. Zarkesh, J. Adibi, and V. Shah, Knowledge Discovery from Users Web Page Navigation, IEEE RIDE'97, 1997