Clustering Unanimous Web Users Based on Rating and User-signature

J.S. Kanchana[1], Dr.S. Sujatha [2]
[1]Associate Professor, K.L.N College of Engineering, Department of IT, Sivagangai, India.
[2]Assistant Professor, Anna University Regional Centre, BIT Campus, Department of MCA, Trichy, India.

***Abstract*—The new social networking era allows and makes users to use web as a place to find people with common interest. People do often search for someone who is of similar interest. People's interest can be found with the few options in web known as tagging and rating. A single user can rate a single item based on his likeness on that item. This paper describes a way on how to identify unanimous people with the tags and rating. A user-signature is constructed by establishing a set representing resource tag and its rating with an algorithm. User signature is generated as a text string specifying the category id and its rating average. Further, signatures of various users are compared and the scale of unanimity is found using an algorithm. A group of users are formed who are likeminded. The group is formed by applying a parameter based clustering with parametric group signature.**

I. Introduction

The web is a space for people to find many things. People not only search for information in web but also for people with common interest to share knowledge. Few of the facilities in web are tagging and rating. A people powered metadata is referred as Tag which may be descriptive or subjective. Tags are used in categorizing the items. Using the tag and the frequency of the usage of the resource, the resource is valued and its web presence is increased. The other way of finding efficient resource in web is with its user rating.

A. Adamic and E. Adar on their paper [1] discussed about how to identify similarity by analyzing text, links and mailing list. Considering these items, the usage of the items for two different users (User A and User B) may vary. Also they may have few usage resources in common. With the commonly used resources among the users the similarity is identified. Items/Resources that are unique to a few users are weighted more than commonly occurring items.

G. Smith, in his paper [2] elaborated that Tags are people powered metadata which is used to represent a particular resource available on the web. A tag is a text string which may be subjective or descriptive. A tag represents the type of the resource. People tag the resource on their own interest, which helps in identifying people's point of view on the resource.

C. Tanner, I. Litvin, and A. Joshi, on their paper [3] proposed that Similarity can be found using a matrix representation. With the help of the min hashing technique, in which user is represented in rows and the user interests are represented in column and the hashing algorithms are implemented to check the similarity between the users. This method is suitable if the users and the available resources are limited.

A. Nisgav and B. Patt-Shamir, on their paper [7] discussed that, similar users can be identified using the resource they see and how they see the resource in the web and how many times they see the resource. For example if a user tags a resource as "funny" and the same tag "funny" is done by other person for the same resource, which proves that they both have unanimous view on that resource. Calculating for many resources with many tags the unanimity can be identified.

Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou, on their paper [5], discussed that rating of a particular resource, for instance a movie; its rating can be evaluated using sentiment classification. The rating and the comments are semantically reviewed. The positive sentiment and the negative sentiment are compared for a particular resource and the rating is evaluated and summarized.

The proposed framework consists of a signature creator, an unanimity checker and a cluster creator. A new factor called "scale of unanimity" is introduced to estimate the unanimity. The scale of unanimity is found using the User signature which is created using the rating based algorithm. Finally the users are clustered using the parameter based algorithm by setting a parametric group signature. The parametric group signature is similar to the individual user signature, which identifies the group uniquely.

II. Proposed Work

Rating is usually given on a scale of 5 or 10, which is how the user rates the item present in the web. Based on how much users rated and how the users rated, the item's presence in the web is increased. Search engines which are available these days, use the rating as a major keyword to display the item in the search results. Whenever people search for particular categories of items, they search for the high rated item and also people rate it with their own view on the item. For example, Google play store which is an application

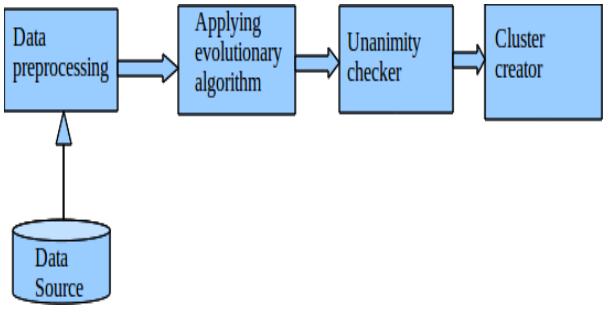
doi:10.4304/jetwi.6.3.359-363

store for android devices, sort its applications based on the rating provided by users.

This rating based searching idea paved a way to find unanimity using the categories and rating. Many websites use rating as a key factor to promote the resources. Every user has their view and rate an item present on the web based on their view. The items in the web are categorized with the help of tags. The tag/category is used to identify what type of resource. The rating depicts how the users rate/view the resource. This rating projects the user's point of view on the resource. So with the help of these tags and rating a user can be uniquely identified by generating a unique signature for him/her. The signature is a unique text string which uniquely identifies the interestingness and the mindedness of people over the resource available. The signature is generated using the Algorithm-I.

With these signature the unanimity, the like mindedness of people can be found. Since people with unanimous mind, rate the resources more or less on a similar level. So the signature which is created using the Rating and rating based algorithm, is used in finding the unanimity. The algorithm-II is used to check the unanimity among the different users/people on the web. On comparing the users with their signature a scale of unanimity is found.

The web is a place with huge number of people. So comparing two individuals is only good for few users. So a clustering mechanism is introduced which help in grouping likeminded individuals. A signature parameter is set and people who fall in those categories are clustered. A Parameter Based clustering algorithm-III is used to create cluster. Those parameters are used as a signature for the group.



Signature Creation

The user signature is created which uniquely identifies the characteristics of the user. A signature is created for valid users with the help of the category ID and the Rating given by the user. The algorithm-I is used to generate the signature. This rating based algorithm uses the average rating given by the user for each category which provides an efficient way of identifying the user's attraction towards that particular category of resources. The categories are limited under five common categories say A, B, C, D, E and for each category the user rating average is found. The signature is the text string which consists of Category ID and average rating.

signature(k)= 0i(catID C(i).AvRating r(i,k))

Avg_Rating r(i,k) = 0irate(i)n

where n is the no of ratings done by user 'k' and 'i' the category

ALGORITHM - I
begin
initialise user,catID;
for each user U
for each cateogory i
find average rating(U) for category(i);
if avgrating(U,i) <= 0
remove the category for user U;
else
generate String/Signature by catID.AvgRating;
end If
end for {category}
end for{user}
end

input : user table with userid and usersign column, user rating table with userid, books name, rating, books category.

Output : user signature

Process : For each user id in the user's rating is retrieved from user rating table. The books which are categorized under 5 common categories with which rating by each users for each category is considered and it average value is taken. A signature is created as a string with category id and average rating. Finally the text string (signature) is updated in the users table.

Unanimity Checking :

Unanimity is the like mindedness or the similarity among the users. The unanimity among the user is checked with the user signature. Scale of Unanimity evaluated using the algorithm-II. For example two users User U with Signature U, and user V with Signature V is considered. The Signature consists of the categories and the average rating for those categories by the user. On finding the difference between the average ratings for each category between the users, the non-unanimity among the users for that category is found. The unanimity value which is found, by deducing the non-unanimity value from 10 since the whole rating is done on a scale of 10. So the unanimity for each category is found in the

similar way. By finding the unanimity rate for each category and aggregating them together thereby finding the average of them provides the scale of unanimity among the users. The user U who checks the unanimity can set a particular threshold for the scale of unanimity. The users who fall in those threshold criteria are unanimous with the user U.

scale of Unanimity (i, j) = ^{0n} 10-$|r_{(i, n)}-r_{(j ,n)}|$N

where $r_{(i, n)}$ is the rating by user i, and $r_{(j, n)}$ is the rating by user j for the category 'n' and N is the no of categories

ALGORITHM - II

```
begin
 for user i and user j
    for each category n
        find r(i, n)
find r(j, n)
find |r(i, n)-r(j, n)|
Unim(n)=10-|r(i, n)-r(j, n)|   //since rating is done on a scale of 10
repeat for each category n
end for
SOU(i, j)=Unim(n)/n;
end for
end
```

Input : user table with userid and user signature

Output : unanimity value

process : Two users id's are received from the user and the signature for each user is retrived. With the signature (A(x)B(x)C(x)D(x)E(x)) is taken.. and the rating value for each category for each user is taken. Each rating value difference is calculated and then it is subtracted from overall range(10). This step is repeated for each category and the final average for each result is obtained. The resulting value is the unanimity value for the users.

Cluster Formation:

Since web is a place of many people we cluster many people together with a particular parameter which is the signature of that cluster. A cluster consists of unanimous individuals. A cluster is created by setting a particular parameter for individual categories and people who are satisfying the criteria are formed under a common cluster. Those parameters may be either set to high or low as to find liked cluster or disliked cluster. This is done with Parameter based clustering (i.e) algorithm-III.

Algorithm - III

```
begin
 for each category n
set parameter for p(n)
for each user A
   for each category n
       if (A(n)>p(n)
         in=1;
        else
in=0;
break;
end for {category}
if(in=1)
add to cluster;
else
go to next user;
end for {user}
end
```

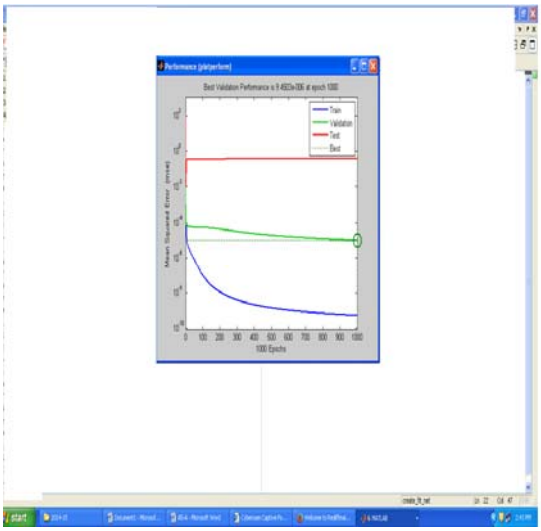
Input : user table with user id and user signature

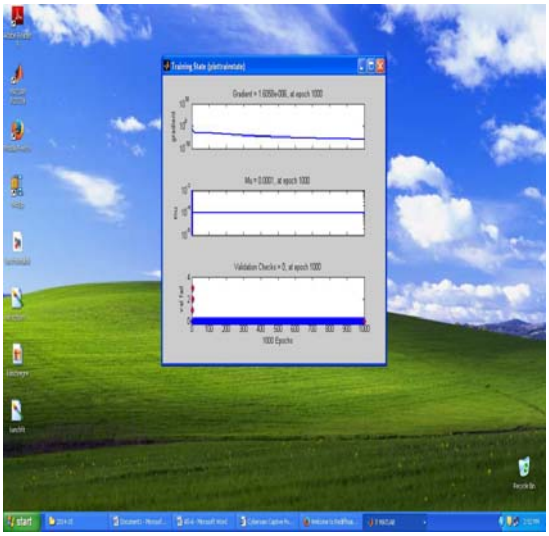
Output : A text file (cluster) with userID's

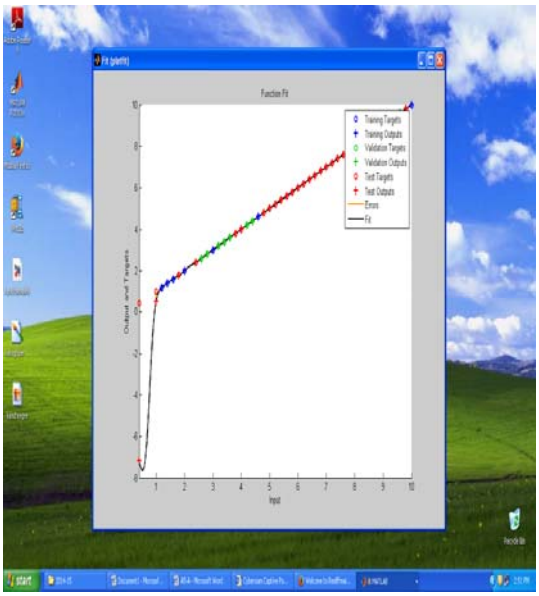
Process : The parameters for each category (A,B,C,D,E) is set. Signature for all users are retrived and checked with the parameter value for each category. If the user signature satisfies the parameter value i.e more than the parameter value, then it is added to the cluster. The user id's which satisfy these constraints are added to the text file which has the file name as cluster signature. Cluster signature is created with the same procedure as user signature.

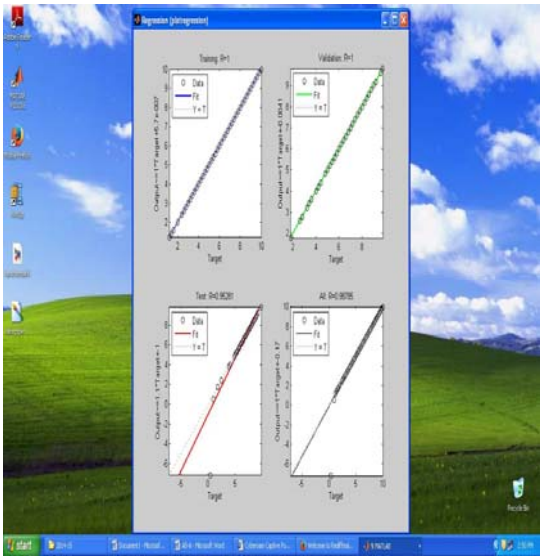
Experimental Setup:

The data set consist of 3 tables namely, a user information table (2,00,000+) Users , a books information table (2,00,000+) Books , a rating table (11,00,000 entries) . For efficient processing, the books information other that category and ISBN were eliminated since they are not used for current processing. The same way, after generating the signature, users who didn't have any signature are also eliminated since they cannot be used in comparing with other users. To verify the effectiveness of the algorithm, initially signature was with only 10,000 rating entries which could able to create signature for few 1000's of users and also few users remain unrated. Also the unanimity range was under 3 to 4. When 2,00,000 entries were considered, more user signature were created and also the unanimity level also increased gradually to the scale of 5 to 6. When all the 11,00,000 entries were considered, signature was created for more than 1,82,092 users and the unanimity among the users were increased. The number of people who fall under each cluster was also increased with the increase in the number of ratings considered.









III. Conclusion And Future Enhancement

The perception of the web has increased due to the introduction of new social platform which are in need of methods and tools to support users' and search for other user groups which communicates their own interests. The advent of social networks, web user groups and other user groups changed the ways of sharing information between the users. In this paper, rating plays a major role in web. Rating is a process that pulls the users for estimating the content in the web. Users who involved in rating will rate the resources available in the web, based on their perception. Users will represent their interests and likes using rating. This creates an opportunity for using user-signature to represent the users, since web is a place where people search for people who have similar interest with them. To find the unanimity among various users, algorithm-II is used. Also a new way of clustering the users is proposed as the parameter based clustering mechanism.

This paper proposed a way of creating signature only with five categories of books. In future more categories may be involved which increases the signature length. An effective methodology may be designed to manage more no. of categories. It deals with the books domain where the user rating for books is considered. It may be extended to various other domains like food, movies, television shows, Electronic gadgets etc. An effective change in the generation of signature or unanimity check algorithm which increases the scale of unanimity to an optimistic level.

References

[1] L. A. Adamic and E. Adar, "Friends and neighbors on the web," Social Netw., vol. 25, pp. 211–230, 2003.

[2] G. Smith, Tagging: People-Powered Metadata for the Social Web: New Riders, 2008.

[3] C. Tanner, I. Litvin, and A. Joshi, “Social networks: Finding highly similar users and their inherent patterns,” Social Netw., 2008.

[4] I. Guy, M. Jacovi, A. Perer, I. Ronen, and E. Uziel, “Same places, same things, same people Mining user similarity on social media,” in Proc.ACM Conf. Comput. Support. Coop. Work, Savannah, GA, 2010.

[5] A. Nisgav and B. Patt-Shamir, “Finding similar users in social networks, ” Theory Comput. Syst., vol. 49, pp. 720–737, 2011.

[6] P. De Meo, E. Ferrara, and G. Fiumara, “Finding similar users in Facebook,” in Social Networking and Community Behavior Modeling Qualitative and Quantitative Measurement, IGI Global, 2011, pp. 304–323.

[7] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou , “Movie Rating and Review Summarization in Mobile Environment”, Systems, Man and Cybernetics part C- application and reviews, vol 42, no 3, may 2012

[8] Ronald R. Yager and Marek Z. ” Looking for like minded individuals in social networking using tagging and E-fuzzy sets”. IEEE transaction on fuzzy systems, vol 21, No 4, August 2013