Framework for Incorporating Social Networks with Recommender Systems -- An Implementation of Profiling Users

Aqsa Bajwa
Department of Computer Engineering
College of Electrical and Mechanical Engineering,
National University of Sciences and Technology
H-12, Islamabad, Pakistan
aqsa.mbajwa@gmail.com

Usman Humayun Mirza
Department of Computer Engineering
College of Electrical and Mechanical Engineering,
National University of Sciences and Technology
H-12, Islamabad, Pakistan
usman572@gmail.com

Usman Qamar
Department of Computer Engineering
College of Electrical and Mechanical Engineering,
National University of Sciences and Technology
H-12, Islamabad, Pakistan
usmanq@ceme.nust.edu.pk

ABSTRACT
Social media is the hype of our era. It has been effectively utilized by people across the world to share their thoughts, interests, likes, moods and dislikes. Interestingly social media sites have become a huge data source of people all around the globe. In this paper we propose to use this information overload about people for recommendation systems. We suggest utilizing social media for mining details about the users. A clustering technique has also been suggested for profiling user's information and how it can be used for making effective recommendations to the users.

Keywords:
Recommender systems, customer profiling, social media, user clustering

1. INTRODUCTION
Recommender systems have played a vital role in bridging the gap between the customers and the retailers in the world of e-commerce. They have evolved the way retailers reach out to their potential customers and provide them with what they want. Similarly recommender systems have also changed the way customers find what they want on the internet[1]. Successful personalization applications depend on knowledge about the customers personal preferences and behavior[3][5]. In general, most recommendation techniques fall in three fields: rules-based recommendation, collaborative filtering recommendation, and learning agent recommendation techniques[2]. A hybrid filtering technique is also commonly used in data analysis and pattern discovery. In this paper we argue that a new dimension of user data collection should be added to the way the recommender systems work. We believe that user's social profiles are much more authentic information source than the ratings, feedback or registration forms filled in by the users on any online retail store. Studies show that a person spends about 23 minutes on a social networking site per day [8]. Studies show that social signing on is a preferred sign in option over signing in as guest user and creating a new account [9]. User are not willing to answer surveys or fill tedious forms provided by retailers. What is worse, most of them seem to provide incorrect information in these forms (76% of consumers fill incorrect information [9]). In this paper we have put forward such a system that runs independently of any specific retail site and gathers products information all across the internet. For user information, the systems suggest mining their social profiles across all social sites to gather information related to their interests, likes, dislikes, status updates etc. Once the product and the user data have been collected, we suggest applying different clustering techniques on both data to form a structured dataset of products and users. Recommendations to the users can then be made about their products of interests and likes.

2. RELATED WORK
Let's have a look at some systems whose architecture has inspired us to suggest our own.

2.1 Graph Model
Zan Huang, Wingyan Chung, and Hsinchun Chen demonstrated the workings of commonly used recommender systems through a graph model. In this model, the data is shown in graphical form in two layers - one layer comprising the data of the products and the second layer comprising the data of the users. In both layers the individual products and individual users are represented by the nodes of the graph. The similarity among the users is shown by weighted links between them and the same has been done for the products. The transactions between the user layer and the product layer are captured by the interlayer links. The transations are basically user's click stream, browsing history, purchase history, and so on and so forth[7].

2.2 Improvisation on the Graph Model
Lekha G. Rao and Siddharth C. Ravi KanthRao suggested a model in concept that adds a new dimension to the way the recommender systems work in general. It adds the use of social media to gather user information so that the

recommendations can be made based on the user's interests. It suggested a Open Id/ Single Sign so that the retailer system can have an access to the customer's social profiles across different social sites. Of course it depends on the customer's / social sites privacy policy settings. But with this access, the retailer's recommender systems can use user's interests, likes, dislikes, joined groups and status updates, etc., to generate recommendations to the users that are based on the users themselves [6].

3. PROPOSED ARCHITECTURE

The proposed model in fig. 1 takes a layered approach from the graph model and incorporates the user information from the social media as described in the later system. However, in the proposed architecture, the second layer will consist of users and each entity will define a group of users and not a single user. The users will be clustered together on the basis of their similarities with one another and the information about these similarities will be mined from social medium. The product layer of the proposed architecture will also be clustered on the basis of the similarities between the products. The relevant user cluster is then mapped onto the related/ required product cluster. That means that every user in one cluster will be displayed all products grouped in the product cluster which is linked with the user cluster. This link is established on the basis of user information mined from their social profiles. Another main characteristic of the model is that it represents an independently working recommender system that mines information about the products as well as the users from the internet. It is not a retailer site specific system; it is in fact capable of working across different retailer sites and applying a clustering technique for grouping together the same products. Similarly, the user information is in fact a gather from the user's social profiles maintained across different social sites. However, for such type of a system to work, the implementation of a single signing on or open id is a must. When the users sign in, the information about their likes and dislikes, interests and hobbies is mined from the social network sites and on the basis of this information, the users are assigned particular cluster. This clustering represents all users with similar interests in one cluster. Similarly, the product layer is also clustered in the same way that the similar products are grouped together in one cluster. This similarity index can be as broad as the diversity in the data allows and can be as compact as the implementer's wish.

4. CLUSTERING OF THE PRODUCTS LAYER

This paper focuses on the implementation of the user layer but to understand how the clustering of the user layer is done, it is important to have an idea how the products layer has been implemented. For the product layer, jobs were considered as products and jobs data from LinkedIn was extracted and clustered using the concept hierarchical structure. To implement the proposed architecture, the user data are needed to be taken from the user's social profiles. Due to the strict privacy policies of these social sites, it was impossible to mine the user profile data from social sites like face book, LinkedIn, etc. So as a solution we had to limit the implementation and use jobs as products and faculty data from a local university was mined as user profile data as a proof of the concept.

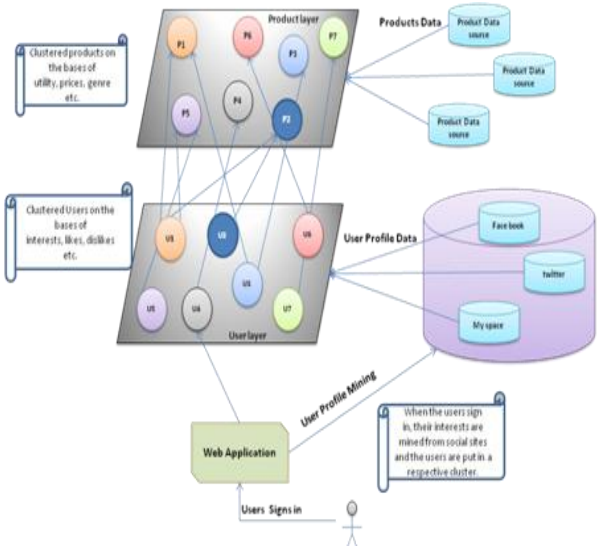

Fig 1. Proposed Architecture

5. IMPLEMENTATION OF USER'S LAYER

Implementing the user layer for the above explained architecture was particularly difficult since we tried to get data using web crawlers and site API (where applicable) but in vain. So, as a solution Faculty data from a local university was mined instead to implement the user layer. The suggested clustering technique was then applied on these data and the results were gathered accordingly.

5.1 Data Acquisition and Preparation

User profiles of faculty data were extracted in excel format. In order to perform data analysis, the data was migrated to a database and a master user profile table was created. From the available attributes, the following attributes were considered for applying the clustering technique:

i. Location--describes the current location of the user.
ii. Years of experience--Total work experience user has in his field.
iii. Field of interest --Field to which the user belongs by profession.
iv. Priority --what the user's priority is while looking for a job. The values considered for this field at the moment are location, career level and field of interest.

5.1.1 Data Cleansing

Once the data was populated in the database, data cleanser was responsible to perform the following operations on the data in order to clean the data from any anomaly and to bring it into a structured format.

a) *Redundancy checker:* Duplicate rows were identified and removed.
b) *Missing field checker:* If the fields selected for clustering were missing, then those rows were

removed. This is the limitation brought in by the selected dataset.

c) *Spelling checker:* Spelling anomalies were identified and corrected.

d) *Abbreviations checker:* Abbreviations such as IT, HR were indentified and expanded.

All these operations were carried out using SQL queries and APIs for spell check and abbreviation check. After the cleansing process, a total of 300 rows of data of user profiles were available ready to be clustered.

5.1.2 Data Profiler

This module consists of two rule engines, primary and secondary. **Primary rule engine** is responsible for applying the first set of rules on the cleansed data set, where as the **Secondary rule engine** is responsible for applying the second set of rules on the data clustered by the first rule engine.

There are two main objectives that we need to achieve using this module:

- To cluster the same users together using the primary rule engine on the basis of 'Priority' attribute.
- To divide every cluster formed by the primary rule engine into smaller clusters on the basis of user's experience level.

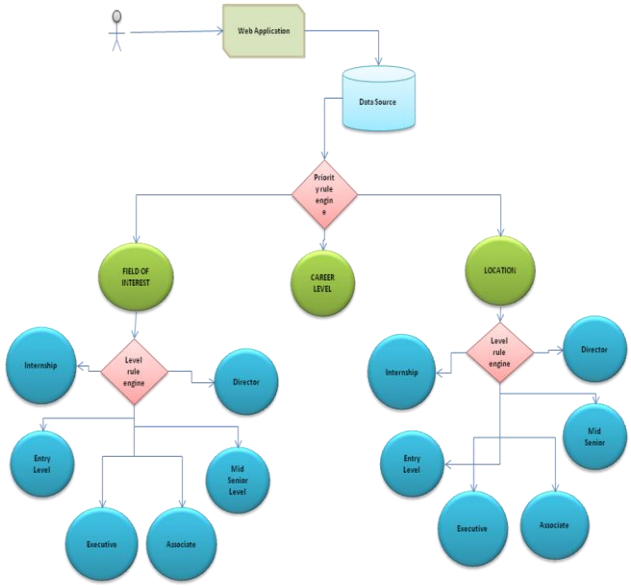

Fig. 2. User layer detailed design

a) ***Primary Rule Engine****:* The primary rule engine divides the complete dataset into numerous clusters based on the value of 'priority' attribute set by the user. Three options for priority have been short listed after careful data analysis. Location, Field of interest and career level. The users who have set the same priority are clustered together. Now when users are clustered together on the basis of the priority, one important thing to notice is that there is a lot of diversity in the data, e.g., users clustered together on the basis of location set as a priority can actually belong to any location in the world, can have any major field and can be at any career level in their professional life. So, this requires further disintegration of the data into smaller clusters. The reason for clustering the users with the same priority is that we now know one thing about the users in this group for sure, e.g., we know that the users belonging to the group where location is a priority, we have to show those jobs to the user that are at the same location as the user. Similarly, for the users in the group where field of interest is the priority, we can show only those jobs to the users that belong to that particular field to which the users themselves belong. This sums up the basic level of segmenting the users.

b) ***Secondary Rule Engine***

The secondary rule engine breaks down each cluster from the previous steps into smaller, more logical subset clusters. After the thorough evaluation of the user attributes available, the number of years of experience was selected for the second level of clustering. For this purpose a careful study of the jobs data was conducted to see the years of experience related to the requirements of a job. The following boundaries were drawn:

- Internship-- No previous work experience
- Entry Level/Officer -- between 1-3 years
- Executive-- between 3-5 years
- Associate-- between 5-7 years
- Mid Senior Level --between 7-10 years
- Director -- 10 years plus

Fig. 2 shows the detailed design of the user layer. Once these boundaries were set, the above range for rules were applied to data in each cluster and 6 sub-clusters were formed. When the user signs in to our web application, the primary rule engine is run and the user falls into a cluster based on the priority set. Once done, the second rule engine runs and further places the user into one of the clusters at the leaf nodes based on the number of years of experience of the user.

6. INTEGRATION WITH THE JOBS LAYERS

Only integrating the two layers together will reveal the working of the proposed architecture in real life. So, the purpose of integration is to show how job recommendations will be made to the user as per their profile information. Each leaf node in fig. 2 will be mapped onto the second level of the jobs clustering hierarchy shown in fig. 3. Once the profile is mapped onto the corresponding cluster in the second level of the job hierarchy, all the job positions on the third level under the mapped clustered will be displayed to the user.

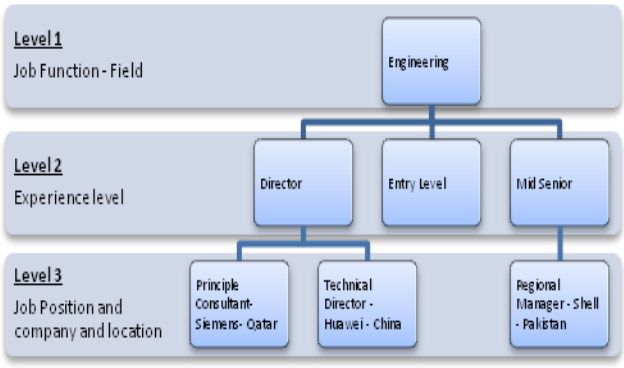

Fig.3. Schema Concept Hierarchy

7. RESULTS

Table 1 shows the clusters are formed on a data of 300 user profiles by the primary rule engine. Since only three options for priority were considered that is why three clusters were formed.

Priority	Number of users
Career Level	125
Field of Interest	93
Location	83

Table 1: Clusters formed by primary rule engine

Now let's have a look at the total number of clusters formed by the secondary rule engine for each data cluster formed by the primary rule engine. Tables 2, 3 and 4 show the subclustering of the three primary clusters on the basis of the rules defined in the secondary rule engine.

Experience/ career level	No of instances
Associate	33
Director	10
Entry level	10
Executive	30
Internship	18
Mid-Senior level	24

Table 2: clusters formed by secondary rule engine when 'Career Level' was set as priority

Experience/ career level	No of instances
Associate	26
Director	6
Entry level	5
Executive	14
Internship	18
Mid-Senior level	24

Table 3: clusters formed by secondary rule engine when 'Field of Interest' was set as priority

Experience/ career level	No of instances
Associate	25
Director	4
Entry level	9
Executive	19
Internship	9
Mid-Senior level	17

Table 4: clusters formed by secondary rule engine when 'Location' was set as priority

8. CONCLUSION

Our main contribution in this work was promoting the idea of using social media as a tool for collecting user data which in turn can be provided to the recommender system to make effective recommendations to the users. The idea was to have a system that revolves around the users' interests and likes and dislikes instead of systems that work on user click stream or browsing history. We have also defined a clustering technique to be applied on the mined user data. The results generated by using the defined technique have been shared. Since the data consisted of 300 rows, the validation of the results was done manually on a sample size of 15 rows. Currently we are working on collecting further data in order to extensively evaluate the approach using more attributes.

REFERENCES

[1] Choochart Haruechaiyasak, Chatchawal Tipnoe, Sarawoot Kongyoung, Chaianun Damrongrat, Niran Angkawattanawit, "A Dynamic Framework for Maintaining Customer Profiles in E-Commerce Recommender Systems"

[2] li niu, xiao-wei yan, cheng-qi zhang, shi-chao zhang, " product hierarchy-based customer profiles for Electronic commerce recommendation", 2002 IEEE

[3] Browne, J., Higgins, P., and Hunt I., "E-business Principles, Trends, Visions", E-business Applications, Technologies for Tomorrow's Solutions, J. Gasos & K.D. Thoben (Eds.), Springer-Verlag, 2003, pp. 3-16.

[4] Choa, Y.H., and Kimb, J.K., "Application of Web Usage Mining and Product Taxonomy to Collaborative Recommendations in E-commerce", Expert Systems with Applications, 2004, 26, pp. 233-246.

[5] Eirinaki, M., and Vazirgiannis, M., "Web Mining for Web Personalization", ACM Transactions on Internet Technology, 2003, 3(1), pp. 1-27.

[6] Lekha G. Rao, Siddharth C. Ravi KanthRao, "Connecting the Dots: Retailer, User and Social Sites", 2011 IEEE

[7] Zan Huang, "A Graph Model for E-Commerce Recommender Systems", February 2003

[8] http://blog.rescuetime.com/2011/10/03/facebook-and-youtube-dominate/

[9] http://www.janrain.com/consumer-research-social-signin