WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Introducing a New Method to Measure Preference Similarity of Social Networks' Users Using Text Mining

[1]Masoumeh Gholipour*, [2]Ahmad A. Kardan

[1]Master's Graduate, Department of Computer Engineering & Information Technology, AmirKabir University of Technology, Tehran, Iran
[2]Assistant Professor, Department of Computer Engineering & Information Technology AmirKabir University of Technology, Tehran, Iran
*S.masoumeh.gholipour@gmail.com
Doi: https://doi.org/10.55248/gengpi.5.0624.1551

ABSTRACT—

A social network is a social structure consisting of After appearing social networks, relation of human got new formation. They have dramatically changed any kind of relations like how people interact, the media perspectives and supporting human needs. Nowadays many people participate in social networks for different reasons. Considering the soaring increase of the users of social networks and its usage, we cannot ignore their impact. In many social network websites, users can interact with each other. Analyzing user's behavior for realizing preference similarity has great interest for researchers and faces different challenges. In this research we introducing a new method to measure preference similarity of social networks' users using Text Mining. Due to the variable expression of phrases in natural language processing, determining the similarity of text is difficult. Different methods has been introduced for this purpose.Meanwhile, the amount of information that is obtained from social networks are far beyond the information contained in a sentence. For this reason, the similarity of social networks users can be determined by combination of this available information. Accordingly, in this paper, a hybrid approach to assess similarities based on user profile data and contextual data that users share is suggested. For evaluating this method Facebook data has been used. The results of the proposed method have acceptable accuracy. At the end, the most similar user to the target user is given as well.

Keywords—component; Social Network, Measure Preference, Text Mining, Behavioral Model, Facebook

I. INTRODUCTION

A social network is a social structure consisting of nodes, which are connected by one or more specific types of dependencies. Social network analysis (SNA), which is a strategy to analyze this structure, views the society structure as a graph and the individuals and social relationships as nodes and edges. Nodes represent individual or organizational actors in the networks and edges are the relationships and connections between these actors. There can be different types of connections between nodes.

In most social network-based websites, users are able to perform collaborative activities. One of the most fundamental features of these websites is creating relationships between users with similar interests. In addition to discovering the possible relationships between users, such relationships expand the social network. Creating communities with similar interest and forming social groups is an important part of our social lives. Analyzing users` behaviors in a social network to discover these interests is one of the interest domains for information and communication technology researchers. This analysis is faced with different challenges. The most important challenge is the scale and complexity of the data related to users` behaviors. This analysis complexity is multiplied by different forms of communications between users, the dynamicity of social network data during different periods, and frequent changes in users` behaviors.

Most approaches determine similarity in social networks using the terms used in users` tags or posts. Words with the same meaning and different forms have been considered less frequently in these methods. Another drawback of these approaches is a lack of definition for a similarity measure in the social network. The similarity measures defined in these studies is mostly limited to a textual similarity measure between prepositions. According to different behavioral data of the users, a social network should redefine the similarity measures according to its users` behavioral and structural characteristics. In addition to investigating the structure of social networks to determine the similarity of users in social networks, this study exploits the texts created by users in form of tags or posts.

II. RESEARCH METHODOLOGY

Most methods, which determine similarity in social networks, investigate the terms used in users` tags or posts. finding similarities using synonym words with fewer different forms is an important task in analyzing social networks, which has applications in other domains, including information marketing,

e-commerce, security ,etc. in the domain of web and internet science, it can be applied to tasks like finding website users` interests and providing recommendations about favorite items. In e-commerce, recommender systems are one of its most important applications. More importantly, in security applications, it can be used to identify hidden terrorist and criminal groups, track the activities of these networks, remove the heads of their groups, or disrupt their activities. It can also be used as a friend recommender to advance political, belief, economic, etc. objectives. Therefore, due to its applications in different social, economic, political, defense, and security problems, finding similarities is one of the most important issues in social network analysis.

Currently, there are many methods to determine the similarity of users` interests in social networks. however, due to the importance of this subject and the vast applications it has particularly had recently, it is still an open problem and there are still efforts to propose more accurate and efficient solutions; specially since some applications like discovering and identifying criminal groups require higher accuracy and their errors are highly costly.

Consequently, recognition, determination, and selection of features, which can be effective in improving similarity finding, seem very important. It is also essential to select a less costly approach and algorithm that provide a better solution to determine similarities.

The methodology of this research in the first section is a library approach through studying scientific papers and research, technical reports, books, dissertations, research projects, and content in scientific and technical websites regarding social network analysis and similarity measures of user interests, as well as data mining techniques to determine similarity. In the second section, the existing methods are experimentally investigated through simulations and comparisons to identify their advantages and disadvantages, as well as existing gaps in this field. Moreover, a method is proposed based on a combination of users` profile data and related textural contents to recommend users similar to them.

III. IMPORTANCE OF FEATURE SELECTION IN SIMILARITY DETERMINATION

Feature selection is one of the problem in machine learning and statistical pattern recognition. This technique is used to deal with high dimensional data. The main problem is that high dimensional data takes more time to be processed. One of the ways to reduce computation time is to select features from the problem space, which are effective in determining solutions, and ignore the rest. Accordingly, data with fewer dimensions is achieved, which can achieve solutions similar to the original data after operations like classification. This problem is very important in applications like classification and clustering, since there are a large number of features in such applications, which are mostly useless or have low information gain. There will be no information problem if these features are not removed; however, they increase the computational load for the desired application and force us to store useless information in addition to useful data. the efficiency of the learning models is significantly reduced By removing these features from the dataset. Generally, feature selection methods try to select a subset of initial features to reduce the dimensions of the data. Sometimes, data analysis on a reduced space is performed better than on the original one. Feature selection aims to find the smallest subset of input features with the strongest classification property. In contrast to other dimension reduction methods (feature extraction approaches), feature selection methods maintain the original meaning of the features after reduction. These methods are mostly applied to datasets, which include a large number of features that make its processing difficult. In social networks, analysis and clustering is very difficult, since the amount of data required by the system to determine the similarity of users` interests is very large. Therefore, the amount of data to be processed must be reduced to remove or mitigate the complexity of relationships between some features. Since increasing the number of features increases the computational cost of the system, it is essential to design and implement them with the least number of features. On the other hand, it is very important to consider that an effective subset of features should be selected, which provide an acceptable efficiency for the system. The main goal of feature selection is reducing the dimension of feature vectors in clustering, such that an acceptable clustering rate is achieved. Under such conditions, the remaining features include appropriate information to distinguish between pattern clusters. Due to removing repetitive and irrelevant data, Feature selection in social networks simplifies the problem and makes similarity finding easier and more accurate for a particular user.

IV. METHOD FOR WEIGHTING FEATURES

Various methods have been reported for weighting features. These methods include term frequency based approaches (TF), methods based on the number of word repetitions in different documents (IDF), hybrid TF-IDF methods, methods based on the classification hierarchy information, and methods based on feature selection. TF-IDF, which is one of the most common features weighing method, was first introduced in information marketing and then used in classifying documents to weight features. This method uses a combination of TF and IDF methods. We also employ TF-IDF to determine the weight of keywords.

Assume D is the set of web documents. For each documents j in D, first all terms are extracted and the weight of each term i in document j is calculated as follows.

$$tf - idf_{i,j} = tf_{i,j} \times idf_i \quad (1)$$

Where, $tf_{I,j}$ is the number of occurrences of i in j and

$$idf_i = \log_2 \frac{|D|}{df_i} \quad (2)$$

Where, |D| is the number of members of D and df i is the number of documents containing i.

V.THE PROPOSED METHOD IN DETERMINING USERS` SIMILARITIES IN SOCIAL NETWORKS

A method is proposed to determine users` similarities in social networks, which is based on a combination of users` profile data and relevant textual content. This textural content includes materials shared or visited by users or any other type of user-related text. Figure (1) presents the stages of calculating the similarity between two users.

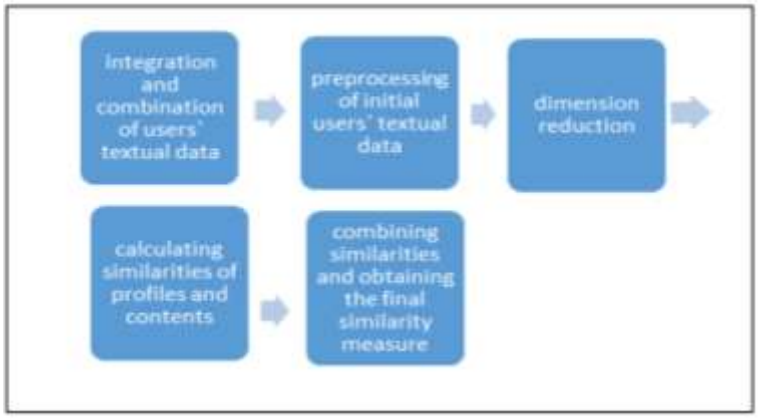

Fig. 1. The stages of calculating user similarity measure

According to figure (1), user-related textual data are first integrated. Regarding users` profiles, all fields in a profile can be aggregated as text and a single text is achieved for each user`s profile. Subsequently, in order to create integrated content for each user, all shared or visited materials can be integrated as a unified text.

At the next stage, for each user, a set of initial preprocessing operations are applied to the texts resulting from the previous stage and the initial textual vector is achieved. Accordingly, first stop words are removed and the noise in the text is identified and eliminated. Subsequently, the tf matrix is achieved and a SVD method, which was explained in the previous section, is applied to it to reduce dimensions. After dimension reduction, the Tf-idf matrix is obtained, which is ready to apply the similarity measure. A cosines similarity function is used to calculate the similarity measure. Assuming two vectors X and Y, the cosines measure is defined as equation (3).

$$cosine(X,Y) = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \quad (3)$$

We define a similarity function between two users as a combination of the similarity between their profiles and the similarity between the textual content related to those users. Assuming Tp is the integrated text of the user`s profile and Tc is the integrated textual content of the user, the final similarity measure of two users, as the combination of their profile and textural content similarities, is defined as equation (4).

$$sim(u_1,u_2) = \alpha * Cosine(T_{p1},T_{p2}) + (1-\alpha) * Cosine(T_{c1},T_{c2}) \quad (4)$$

Where, α is between 0 and 1 and it is adjusted depending on the user type. Therefore, the final similarity measure of two users is as follows.

$$sim(u_1,u_2) = \alpha * \frac{\sum_{i=1}^{n} Tp1_i Tp2_i}{\sqrt{\sum_{i=1}^{n} Tp1_i^2}\sqrt{\sum_{i=1}^{n} Tp2_i^2}} + (1-\alpha)\frac{\sum_{i=1}^{n} Tc1_i Tc2_i}{\sqrt{\sum_{i=1}^{n} Tc1_i^2}\sqrt{\sum_{i=1}^{n} Tc2_i^2}} \quad (5)$$

VI. The Algorithm for Recommending Similar Users

This section proposes an algorithm based on the users` similarity measure, which was introduced in the previous section, to find and recommend users similar to a specific user. Here, we want to find m most similar users to the current user and recommend them as output. The stages of this algorithm as follows.

Algorithm name: Recommending similar user Input: specified user and list of users and profiles Output: list of similar users

- Step1: Clustering users based on the proposed similarity measure using k-means algorithm.
- Step2: Finding the cluster corresponding to the user.
- Step3: Finding the similarity of cluster members with the user based on the proposed similarity measure.
- Step4: Sorting cluster members in descending order based on their similarity with the user.
- Step5: Returning the first m members of the sorted list as output.

In the proposed algorithm, users are first clustered using the k-means algorithm, which is an effective method for clustering textual data. Subsequently, it is specified which cluster the considered user belongs to. Since, clustering generated clusters such that their members are similar to each other and different from those of another cluster, it is only sufficient to seek users most similar to the considered users in its corresponding cluster.

In order to do so, the similarity of each user in the cluster is calculated based on the proposed measure to find their similarity with the considered user and find the m most similar user. It is sufficient to sort them in descending order, then select, and return m users at the top of the list, which have the most similarity with the user.

VII. Analysis and Evaluation

This section introduces the dataset used to run the proposed method. This study uses a set of data corresponding to Facebook users. This dataset contains the information of 2013 Facebook users collected by the Skill security organization in the scope of the movies and films. The dataset is downloadable from the URL in [10]. In this dataset, there a user ID for each user and there are a set of materials related to users` interests for each user ID.

TABLE I. THE SPECIFICATIONS OF THE INPUT DATASET

Feature	**Value**
Average count of files users talked about	3
Count of films discussed in all discussetions of users	1744
Maximum referencing of film by users	8
The Number of users	2013

Since the similarity computation results should be tangible and usable for the all system users, an expert has helped to evaluate the precision of the similarity between two users. As it was mentioned in the previous section, the similarity of two users uj and uk is obtained using the following equation.

(6)

$$sim(u_j,u_k) = \alpha * \frac{\sum_{i=1}^{n} Tp[j]_i Tp[k]_i}{\sqrt{\sum_{i=1}^{n} Tp[j]_i{}^2}\sqrt{\sum_{i=1}^{n} Tp[k]_i{}^2}} + (1-\alpha)\frac{\sum_{i=1}^{n} Tc[j]_i Tc[k]_i}{\sqrt{\sum_{i=1}^{n} Tc[j]_i{}^2}\sqrt{\sum_{i=1}^{n} Tc[k]_i{}^2}}$$

Where,

$$TP[j] = tf - idf\left(\bigcup_{i=1}^{m_j} ProfileText_i\right) \quad (7)$$

Where, vector TP[j] is the vector of aggregated texts for the profile of the jth user. Moreover, mj is the number of existing texts in the profile of the jth user and ProfileText1 shows the vector of each text in the profile of the jth user. Finally, the value of TP[j] is equal to the tf-idf vector resulting from the union of each text in the profile of the jth user. The following equation calculates the value of TC[j].

$$TC[j] = tf - idf\left(\bigcup_{l=1}^{r_j} ContentText_l\right) \quad (8)$$

Where, $TC_{[j]}$ is the vector of the collected shared texts for the j_{th} user. Moreover, r_j is equal to the number of texts shred by the j_{th} user and ContentText1 shows each text shared by the j_{th} user. Finally, $TC_{[j]}$ is equal to the TF-IDF vector of the union of each text shared by the j_{th} user.

As it was mentioned, the view of an expert is used to evaluate the prevision of the proposed distance measuring method. Assuming the expert has specified the similarity score of each two users between 0 and p, such that 0 indicates the lowest possible amount of similarity between users Uj and Uk and p is the highest amount of similarity. Since the similarity value is always between 0 and 1, the expert`s vote to the similarity of the two users uj and uk should be divided by p to achieve the accuracy of similarity calculations as the following equation.

$$SimAccuracy(u_j, u_k) = 1 - \left| \frac{exsim(u_j, u_k)}{p} - sim(u_j, u_k) \right| \quad (9)$$

here, $exsim(u_j, u_k)$ is the similarity between the two users uj and uk based on the expert`s opinion.

In order to more accurately evaluate the resulting solutions and achieve better coverage, the mean of the resulting accuracy values can be calculated for different pairs of users. Therefore, the following equation is achieved, where, m is the number of compared user pairs.

$$OverallSimAccuracy = \frac{\sum_m \left(1 - \left| \frac{exsim(u_j, u_k)}{p} - sim(u_j, u_k) \right|\right)}{m} \quad (10)$$

The most similar user is determined by sorting the users in the cluster of the considered user in descending order. Thus, in order to evaluate the most similar user, it suffices to show the unsorted users in the corresponding cluster to the expert and ask him to select the most similar user. Naturally, the closer the

expert`s selection is to the first member of the sorted cluster members, the proposed method can achieve a more accurate solution. Therefore, if the number of users in the cluster, which contains the considered user, is m and the user that is selected by the expert as the most similar user is p, the accuracy of our solution in the recommendation of the most similar user to a selected user section is calculated using the following equation.

$$RecommAccuracy(u_j) = 1 - \frac{p-1}{m-1} \quad (11)$$

TABLE II. :RESULTS OF SIMILARITIES CALCULATED USING THE PROPOSED METHOD

Two user's similarity	**Value**
U5 , U21	**0.29**
U15 , U32	**0.65**
U7 , U113	**0.78**
U6 , U22	**0.89**
U221 , U15	**0.39**
U176 , U193	**0.69**
U10 , U17	**0.52**
U2 , U43	**0.08**
U215 , U239	**0.57**
U239 , U76	**0.72**

Now, we used an expert who has 7 years of experience in social network analyzing to assist for determining the similarities of users. He participated in many similar projects and was the best candidate to do the comparison .for the users above, the expert was requested to give each pair of users a similarity score between 0 and 5, where 0 means having the least similarity and 5 shows the highest similarity. Table (3) presents the scores of the expert to each pair of users in the table above.

TABLE III. RESULTS OF THE EXPERT ABOUT SIMILARITY VALUES

Two user's similarity	**Value**
U5 , U21	3
U15 , U32	2

U7 , U113	5
U6 , U22	4
U221 , U15	1
U176 , U193	4
U10 , U17	2
U2 , U43	1
U215 , U239	3
U239 , U76	4

Therefore, the accuracy in calculating the similarity between users is computed as equation

Consequently, we achieved an accuracy of 84.8% in this section, which seems acceptable.

In order to evaluate the most similar user recommendation, users were first clustered according to the aforementioned method in the previous section to recommend the most similar user to a considered one. Table (4-2) presents the clustering results.

For clustering, the number of clusters was considered 2 to 20 and the Davies-Bouldin measure was used for each state to evaluate the quality of clustering. As a result, 17 clusters obtained the best solution. We must note that this operation was performed using the RapidMiner software operators.

Now, one of the members of cluster 11, which has seven members, was considered and the rest of the members were sorted in descending order based on their similarity. The following table presents the results.

TABLE IV. RESULTS OF CLUSTERING USERS

Order of similar users to U38	**Value**
U45	1
U89	2
U365	3
U873	4
U1536	5
U417	6

Without having any knowledge of this order, the expert selected user U89 as the most similar user to the selected one, which has the second position in the table above. Therefore, according to the proposed evaluation method in the previous section, the accuracy of this section is calculated as the following equation.

$$RecommAccuracy = 1 - \frac{2-1}{7-1} = 0.833 \quad (12)$$

VIII. COMPARISON OF THE PROPOSED METHOD AND A BASIC APPROACH BASED ON USERS` TEXTUAL CONTENT

This section compares the proposed method, which is a hybrid method based on users` profiles and shared textural content data, with a user similarity measuring approach, which is only based on the textural content shared by the users.

The accuracy of the proposed method for evaluating the users` similarities, which works based on a combination of users` profiles and textual data, was 84.8%. Whereas, in the similarity measuring method based on only textual data according to 9, the achieved accuracy was 77.7, which shows a 7.1% improvement by the proposed method in comparison to the approach based on textural data.

IX. CONCLUSIONS

This research proposed a method to measure similarity and analyze the behavior of users in social networks to determine the similarity of their interests and tastes. The proposed method used a combination of users` profiles and shared texts in a social network. The data collected from Facebook was used to evaluate the proposed method. In what follows, experimental results and evaluation of the obtained solutions are presented for the proposed method. The biggest challenge during this work was there are not too many datasets including both profile and shared text.

Results showed an accuracy of (85%) in measuring user similarity and (83%) in recommending the most similar user to a selected user. This shows that the proposed solution has acceptable results and can be employed in practical applications.

X. FUTURE WORKS

The proposed methods can be adapted to distributed environments (e.g. Hadoop) to use distributed computing infrastructures to perform and evaluate the proposed method on several million users.Using the proposed method to measure user similarity, studies can be defined in discovering communities and family networks in social networks and improve their group and community discovery methods.

REFERENCES

[1] D. Lin, "An Information-Theoretic Definition of Similarity", in Proceedings of the Fifteenth international Conference on Machine Learning, San Francisco, CA, 1998, pp. 296-304.

[2] G. Hirst and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms", in C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, pp. 305-332. Cambridge, Mass.: MIT Press, 1998.

[3] Tahmasebifard, H., Pouyan, M. M., & Mirzaagha, M. (2018). Latent Functions of Brands and Brandings. Business and Management Horizons, 6(1), 89.

[4] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone", in Proceedings of the 5th annual international conference on Systems documentation, 1986, pp. 24 – 26.

[5] S. Banerjee and T. Pedersen, "Extended gloss overlap as a measure of semantic relatedness" in Proc. of IJCAI˝03, Acapulco, 2003, pp. 805- 810.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis", Journal of the American society for information science, vol. 41, no. 6, 1990, pp. 391-407.

[7] Xun Zhou, Jing He, Guangyan Huang, Yanchun Zhang, "SVD-based incremental approaches for recommender systems", Journal of Computer and System Sciences, vol. 81, no. 4, June 2015, pp. 717-733.

[8] Pranab Kumar Dhar, Tetsuya Shimamura, "Blind SVD-based audio watermarking using entropy and log-polar transformation", Journal of Information Security and Applications, vol. 20, February 2015, pp. 74- 83.

[9] Grigorios Tzortzis, Aristidis Likas, "The MinMax k-Means clustering algorithm", Pattern Recognition, vol. 47, no. 7, July 2014, pp. 2505-2516.

[10] https://blog.skullsecurity.org/blogdata/fbdata.torrent