Personalizing of Content Dissemination in Online Social Networks

Abeer ElKorany
Computer Science Department
Faculty of Computers & Information, Cairo University
5 Dr Ahmed Zoweil St., Orman, Giza, Egypt

Khaled ElBahnasy
Information Systems Department
Faculty of Computer & Information Sciences
Ain Shams University, Abbasia, Cairo,Egypt

***Abstract*— Online social networks have seen a rapid growth in recent years. A key aspect of many of such networks is that they are rich in content and social interactions. Users of social networks connect with each other and forming their own communities. With the evolution of huge communities hosted by such websites, users suffer from managing information overload and it is become hard to extract useful information. Thus, users need a mechanism to filter online social streams they receive as well as enable them to interact with most similar users. In this paper, we address the problem of personalizing dissemination of relevant information in knowledge sharing social network. The proposed framework identifies the most appropriate user(s) to receive specific post by calculating similarity between target user and others. Similarity between users within OSN is calculated based on users' social activity which is an integration of content published as well as social pattern Application of this framework to a representative subset of a large real-world social network: the user/community network of the blog service stack overflow is illustrated here. Experiments show that the proposed model outperform tradition similarity methods.**

Keywords—social network; content similarity measurement; Information retrieval; Information dissemination

I. INTRODUCTION

Nowadays, social networking has become an important part of online activities over the web. Social networks can be viewed as a structure which enables the dissemination of information through social interactions among individuals. The *analysis of the dynamics of such interaction* is a challenging problem in the field of social networks. Online social networks (OSN) such as Internet newsgroups, BBS, and chatrooms are interesting channels that enable its members to communicate and share activities in an easily accessible in anywhere any time trend. OSNs represent a new kind of information network that differs significantly from existing networks such as the Web. They are those network hosted by a web site where friendship represents shared interest or trust and online friends may have never met. When a user joins those networks, they could publish their own content, create links to other users in the network called "friends" or "acquainted". Virtual link is constructed between users with similar interests. User' generated information in OSNs has been characterized by either their provided published content in form of text, image or videos as well as network activities that are frequently changing over time. For example, web-blogging community is identified by its rich daily blog posts and a social network of bloggers who share, find, and disseminate content at a massive scale. Today a lot of active online social networks users complain that their streams have become too overloaded and hard to extract useful information from. To make use of the growing provided information flow within a social network and to keep being tuned with its related members, it is necessary to personalize the process of information distribution. Thus, it is necessary to control information propagation among users who share some common interest in OSN i.e finding similar users. Existing studies on user similarity focus on either link or content analysis. However, neither information alone is satisfactory in determining accurately the similarity between users. It is therefore important to unify the analysis of content and network activities about user in order to personalize content dissemination in social networks.

This research proposes a personalized user content dissemination framework that identifies the most appropriate user(s) to receive specific information (such as post) based on user similarity. Similarity between users within OSN is calculated based on users' social activity which is an integration of content published as well as social activities of users. Each of two sided of social pattern provides a partial indicator of the similarity. While content published by a user is used as an indicator of user interest, social pattern parameters tend to group people based on their similar networking behaviour. Accordingly, we identify, collect, and classify different users' online social activities that are used to construct their characteristics, and determine user' main preference. Next, users' preferences are used to detect similar users who are candidate to receive a specific stream (post). The vector space model is adapted to present user characteristics such that each user is represented by two vectors each contains a set of specific social activities of the user. The first vector represents the content published user through the *bag of word model* which is called content vector. Content vector represents the terms used in all the posts published by that user. While the second one, called social pattern vector, which represents the social pattern of user such as: user' contribution and influence attributes. The term weight of each vector is computed using different methods. For example, in content vector, all post published by each user is collected and TF-IDF is used to weight the terms. While, in the social pattern vector each social features is represented as term and weighted mean scheme is used to weight the terms. An aggregated linear model is then applied to combine similarity calculated by using each of those vectors. Thus, the process of content dissemination works as follow: first when a

target user post a stream, cosine similarity is applied to compute similarity between that post and content vector of all users and only the top ranked 20 users are used as input to the next phase. Next, social pattern vector of each of those 20 users is retrieved and cosine similarity is used again to compute similarity between social pattern vector of target user and the top 20 users in order to re-rank them. It is significant to mention that the proposed process is generally applicable to any knowledge sharing environment. However, our work is supported by a set of experiments and tests conducted. The experiment is applied on real world weblog system (question/answering) Stack Overflow dataset [1] a question answering web community that allows users to ask and answer questions about computer programming languages. A modified 5-fold cross validation method is used to insure the accuracy of the proposed model. This paper is organized as follow: section2 presents related works in three main areas such as social networking services, information filtering, and other methods used for user similarity. Section3: discuss the proposed model for personalization content dissemination and section4 explains the main components of the system used to identify candidate users to receive a specific post. Section5 discuss the experiments hold to measure the accuracy and efficiency of the proposed model on real dataset and finally section6 conclude the work and propose future work.

II. Related Work

Our model is closely related to models of information filtering and data analysis in social network

A. Social network service

Several Social Networking Services (SNS) are designed to facilitate communication, collaboration, and content sharing over a network of contacts [1]. They enable users to share profiles and personal information, media, event planning, communicate by email, send instant messages, share announcements, blog together, creation of interest groups, and meet online with their friends or even other new people. In SNS, the variety and quality of content is a key factor for success. Encouragement to produce content is therefore commonly observed in these networks. Therefore, analysis of two primary kinds of data in the context of social networks is widely increased. These data are:

Linkage-based and Structural Analysis: In linkage-based and structural analysis, analysis of the linkage behavior of the network is applied in order to determine important nodes, communities, links, and evolving regions of the network. Such analysis provides a good overview of the global evolution behaviour of the underlying network.

Adding Content-based Analysis: Many social networks such as *Flickr*, *Message Networks*, and *Youtube* contain a tremendous amount of content which can be leveraged in order to improve the quality of the analysis. For example, a photograph sharing site such as *Flickr* contains a tremendous amount of text and image information in the form of user-tags and images. Similarly, blog networks, email networks and message boards contain text content which are linked to one another.

B. Information filtering in social network

Information overloading has been a major problem in social media. Thus, social filtering systems are used to filter social media streams in order to overcome this problem . Several information filtering systems have been proposed such as in [2] which utilize text of documents that the user is interest in with other sources of information to identify social features of the users. These features are then used to detect others who are more likely to post relevant content. However, limitation on text size in micro-blogging services result in sparseness of data for text classification. Another systems suggested to use Wikipedia as an external source [3,4], or using web search engine results[5] to enhance text classification. Another approach that used topic modelling techniques [6] such as Latent Dirichlet Allocation [7] to classify texts based on universal corpus . Another work focused on short texts is [8], where a method for measuring the semantic similarity of texts, using corpus-based and knowledge-based measures of similarity is proposed.

C. User Similarity in social networks

There have been numerous efforts to calculate the user similarity for different objectives such as recommending people. Guy et al. [9] proposed a method based on various aggregated information about people relationships and focused only on people that the user is already familiar with. Thus, this method was not used for calculating the similarity with an unknown user such as suggest a new friend in the online social network. Terveen et al. [10] proposed a framework called *socialmatching* that match people mainly using their physical locations. Other system focused on applying the semantics of the location in order to calculate similarity between users by capturing the user's intention and interest and considering the similarity between different locations using the hierarchical location category[11]Other methods utilize similarity measurement in social network to recommend experts. McDonald et al. [12] proposed an expert locating system that recommends people for possible collaboration within a work place. An expert search engine was described in [13] which found relevance people according to query keywords. Those approaches are useful to find co-workers or experts in a specific domain however; they cannot be used for finding similar users in general.

III. Personalized Flow of Information in Online Social Network

OSN have exploded in popularity. They could be classified as into two categories, the networking oriented OSNs and knowledge-sharing oriented OSNs. The former, such as Facebook and LinkedIn, emphasizes more on the networking perspective, and the social relationship is the basis of these OSNs. Hence, they are called networking oriented OSNs. While the latter, such as blog networks, question answering networks, and viral video networks, emphasizes more on the knowledge or content sharing [14]. In knowledge-sharing OSNs, issues such as users' participation in network and their generated content are crucial to healthy growth of those networks. Information overloading is a major problem in such

[1] Stack overflow. http://stackoverflow.com

knowledge sharing environments [15]. Thus, the proposed framework shown in figure1 aims to overcome the information overloading by enhancing the distribution of content among users. It suggests the most relevant users that are candidate to receive a specific post from a target user. By identifying main features that characterize users and determining users preference, personalization of flow of information is achieved which is illustrated in this section.

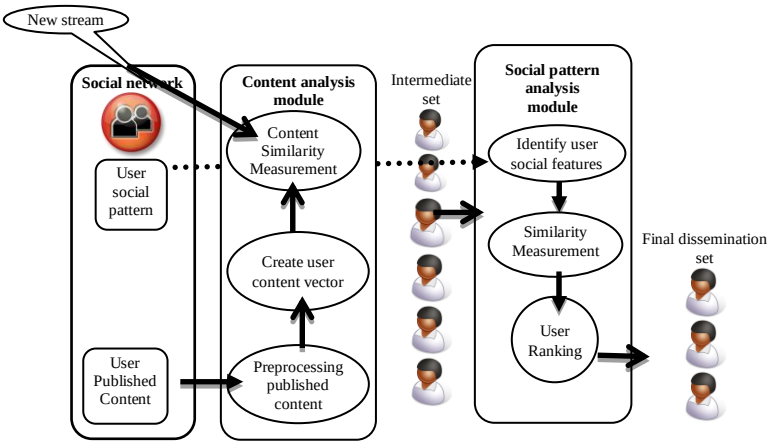

Fig. 1. Framework of Personalized Information flow

A. *Analysis of users' social contributions*

User contribution with a social network is changing over time in form of interaction and content provision resulting in variant network structure and text dissemination which evolve simultaneously and interrelated. User's activities in OSNs include authoring content, viewing, and networking. It has been also assumed that there is strong correlation between user active time and user contribution [16,17]. In general, user within the same OSNs could be classified as online and offline users. Online users are those who always active in writing posts, comments, view others publish content, provide feedback in form of like or dislike and many other activities. While offline users are those who just visit their homepages on daily or weekly basis without any contribution. Most of existing approaches for user and content recommendation rely only on network structure and relationship between users without considering the published content. Our proposed framework identifies user's preference in OSN by considering two main folds: content published by users and user social pattern. In blogs and question answering networks, there are two important elements in a shared content: text and hyper-links which provide related information from external sources such as web pages or images. Similar users are characterized by posting similar text and hyper-links. Therefore, posts of each user are parsed, and a bag of words is created for each user, keyword index is computed indicating that the more terms two users share, the stronger the tie between them. Second, other network features that distinguish users in OSN are considered which we classify as : user contribution and user influence are identified in order to impose a finer grained similarity between users. Each of them is expressed by a set of weighted features extracted from social activities of users which will be explained in details in next section. Unlike other approaches that only consider user pre-defined attributes such as demographical attributes, geographic location, and defined interests, our approach relies on extracting social activities and calculate their weight according to user contribution.

B. *Multilevel Model for detecting Relevance Users*

Recently, OSN started to be modelled with *rich structured data* that incorporate *semantics*. In such models edges between users are split to weighted links based several features such as: communication aspects (uni-direction or multi-direction), frequency of communication, and influence measure which are used to build up clusters of similar user. Therefore, we calculate similarity between users through aggregating communicative content of users in form of mutual published content with social activities. Our model of detecting relevance users is based on the assumption that if users *X* and *Y have n* similar published contents and similar social features, they have strong tie. In the proposed framework, adaptive vector space model is used to compute social similarity measurement of users. The vector space model is a standard and effective algebraic model widely used in information retrieval (IR) that use Cosine similarity to compute relevancy of documents with respect to a given query. Accordingly, our model works by first identifying content and social features, collecting required information, creating corresponding vectors. Each user is expressed by two vectors of elements corresponding to their published contents and social pattern respectively. A two-level model is then used to identify most candidate users to receive a specific post. Each level is responsible on computing similarity between users using different user social features (stored in different vectors) and a linear model is then used to combine similarity generated from each level. The main idea is based on utilizing model-based collaborative filtering by first finding users with similar content and then utilizes social features to map social characteristics of users and get the most neighbourhood users for the target user.

IV. Architecture of Personalized Information Diffusion

The main objective of the proposed model is to improve the dissemination of information in knowledge sharing network by identifying the most appropriate (similar) users to this information as well as its publisher. In social media, users are identified through their social content and participation. Therefore, the proposed personalization framework is decomposed of two main modules: content similarity measure and social patterns similarity measure.

1) User content similarity: using content of the posts and comments generated by users, similarity measurement is applied to get top ranked 20 users with respect to specific post. This short list of users represents the closest users to that stream.

2) User social pattern similarity: additional social information about top ranked 20 users is used to ensure similarity "coverage". Cosine similarity is then applied between social pattern vector of target user (who post that

stream) and the those users to rank the most neighbourhood users

Before content similarity phase takes place, pre-processing of the content representing the posts of users is applied as shown in figure1.

B. Pre-processing

Currently, there is a large volume of text data that is produced from the social communities such as blogs, tweets, and comments. Therefore, during this step, all blog- posts are aggregated from all users to form a corpus. Posts are parsed in order to clear all special characters, numbers, dates, stop words and single characters. This yields to construct a vocabulary that represents the set of words that have been used by the whole users of the social network within a specific time period. The final step of the preprocessing decomposes stemming, identifies hyper-links or code which may be included in a user posts, and then constructs content vector of each user.

C. User content similarity measurement

In linear algebra a *vector space* is a set *V* of *vectors* together with the operations of addition and scalar multiplication. A vector space model (VSM) is an algebraic model introduced a long time ago by Salton [18] in the information retrieval (IR) field. In a more general sense, a VSM allows to describe and compare objects using N-dimensional vectors. Each dimension corresponds to an orthogonal feature of the object (e.g. traditionally weight of certain term in a document). We adapt the vector space model so that it treats each user as a document and her posts as terms disregarding grammar and even word order. In IR field there are several approaches to calculate the weight of a term in a document. In our model, we apply the *term frequency-inverse document frequency (tf-idf).* Term Frequency (tf) assigns the weight to be equal to the number of occurrences of the term *t* in document *d*. While *IDFt* is obtained by dividing *N* by *DFt* and then taking the logarithm of that quotient, where *N* is the total number of posts generated by a user and *DFt* is the post frequency of *t*, i.e., the number of posts containing the term *t*. This approach identify the rarity of *t* in a given corpus Thus, if *t* is rare, then the posts containing *t* are more relevant to *t* which match with the idea we propose to find the target (most similar users) to specific set of words(post). Thus, content vector of a user would represent all words of her/his posts associated with TF-IDF weights ***wjt*** , where *wjt* is the weight of word *t* in post *j* and is calculated as follow:

$$wjt = tf(t, j) \log(\text{n posts}/df(t))$$

Then, when a user post a specific query (post) Cosine similarity is used as a standard measure estimating relevancy between this post and other users' content vector. The top 20 similar users to that post are selected to be used in the next phase.

D. Social patterns similarity measurement

In social knowledge sharing, in order to identify similar users, it is significant to consider the effect of network relationships. By aggregating communicative activities of users in form of social interaction, hidden relations between users are discovered, and hence a list of most similar users is generated. Unlike other approaches which considers only number of mutual friends [19], our proposed social pattern similarity measures consider all user social activities which we classify into two categories. The first category covers user contribution with the system while the second covers user influence with other community's members. User contribution is measured by the frequency of contribution to the knowledge sharing network (in our case blogs) which is: average number of posts and average number of comments a user provides. On the other hand, bloggers tend to *interact* with other bloggers by providing comment, like, or favorite in response to specific blog posts. Thus, we consider this type of information as a measure of user influence or trust relationship. "Trust relationships" are different from "social friendships" in many aspects. For example, when a user u_i likes a blog issued by another user u_j, user u_i probably will add user u_j to his/her trust list. The study of homophily has shown that people with similar interests are more likely to become connected, associate, and bond with other similar users [20]. Based on that hypothesis, we measure the user influence by the attributes that reflect the recognition he got from others in the same social network which may vary from social network to another. For example: Endorsements are a one-click system to recognize someone for their skills and expertise on LinkedIn, the largest professional online social network. In our case of stackoverflow, we map existing social features with the idea of "Trust relationships". Therefore, we utilize the following attribute:

`ViewCount`: Number of views of posts the users obtained
`FavoriteCount`: Number of users, who set a user's posts as favorite,
Vote: Count the number of votes for specific user' posts,

Each user is represents as vector associated with scores representing the average weight of her/his social features as terms. This was accomplished through combining all previous posts and comments of each user and calculated the average number of views, score, and favorite she obtained from others as well as the average number of posts and comments published by that user. Two users are similar if their vectors differs only a few coordinates. A high degree for a preference (term) of a user can be interpreted in a way that the other user repeatedly (frequently) confirms her preference [21].

V. EXPERIMENT

In this section we present several experiments to show how the proposed similarity measures model affect the dissemination of information in online social network. As mentioned earlier, the model aims to generated the minimum and appropriate users who can receive a specific information based on content-matching with that post and social pattern matching with the seed user who posted it. We applied this model on data dump provided by Stack Overflow[2] which is an online platform where users can exchange knowledge related to programming and software engineering tasks. This platform combines features of Wikis, Blogs and Forums, and aims to

[1] Stack overflow data dump http://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/. /. Accessed January 2013

provide free knowledge sharing between software developers worldwide. The collected data contains all the questions and answers posted on the web site between July 31, 2008 and March 31, 2009. All posts, both questions and answers, are scored, viewed, and voted by the users. Some questions are set as favorites by some users. We use a subset of the posts collection, obtained by filtering the document collection to select the 100,000 posts belongs to 2000 different users. Several experiments are applied, thus we split the set of users into 5 equally sized disjoint groups of users from G1 to G5 each of 400 users {Gi=1-5}. Two main experiments have been applied, the first one aims to prove the accuracy of the proposed model which the second one target the efficiency.

A. *Experiment Setup*

The first experiment has been applied using content-similarity module and trying different weighting scheme using the vector space mode. We used the TF and TF/IDF to weight the terms and according to table1, the similarity values obtained when using TF/IDF is better for top 20 users.

TABLE I. COMPARISON BETWEEN DIFFERENT CONTENT WEIGHTING SCHEME

	TF/IDF	TF		TF/IDF	TF
User1	0.6144	0.5987	User11	0.5947	0.2537
User2	0.6082	0.5973	User12	0.5942	0.2339
User3	0.6071	0.5879	User13	0.5928	0.2303
User4	0.6021	0.4740	User14	0.5923	0.2273
User5	0.6020	0.3762	User15	0.5912	0.2239
User6	0.5999	0.3605	User16	0.5910	0.2171
User7	0.5989	0.3538	User17	0.5902	0.2167
User8	0.5988	0.3316	User18	0.5898	0.1994
User9	0.5986	0.3180	User19	0.5897	0.1983
User10	0.5969	0.3148	User20	0.5893	0.1898

In order to ensure the accuracy of the proposed system, we adapt the cross validation. Thus, we create a training set and sets. The training set contains one group of users (G1=400 users) and other test sets are created based on different combinations between G1 and other groups. We use mathematical **combination** to produce several test sets. This combinations imply that if the set has *n* elements the number of *k*-combinations is equal to the binomial coefficient which can be written using factorials as:

$$\text{Number of generated groups} = \sum_{k=1}^{k=5} \frac{n!}{k!(n-k)} \qquad \text{Equ(1)}$$

As per equation1 and when having N=5, 31 subsets are generated. However, only 16 of them contain target group G1 like as follows:

$\{G_{1}\}$,
$\{G_1 \cup G_2\}$, $\{G_1 \cup G_3\}$, $\{G_1 \cup G_4\}$, $\{G_1 \cup G_5\}$,
$\{G_1 \cup G_2 \cup G_3\}$, $\{G_1 \cup G_2 \cup G_4\}$,$\{G_1 \cup G_2 \cup G_5\}$, $\{G_1 \cup G_3 \cup G_4\}$, $\{G_1 \cup G_3 \cup G_5\}$, $\{G_1 \cup G_4 \cup G_5\}$
$\{G_1 \cup G_2 \cup G_3 \cup G_4\}$, $\{G_1 \cup G_2 \cup G_3 \cup G_5\}$, $\{G_2 \cup G_3 \cup G_4 \cup G_5\}$,$\{G_1 \cup G_3 \cup G_4 \cup G_5\}$,
$\{G1 \cup G2 \cup G3 \cup G4 \cup G5\}\}$.

Next, we select a random post generated by a random user and apply content-similarity phase to get the top ranked 20 users relevant to that post as shown in first column in table2 using TF/IDF method. Then, we regularly increase the neighbourhood of a target $user_x$ and find out whether the system would produce the same set of candidate users and check the similarity values they obtained. Therefore, we use the same post for the same user all over other 15 groups and get similarity value for the same set of users. According to table2, similarity value obtained for those users remain almost the same however the number of neighbour is. This matches our assumption regarding distributing of information which should target interest of users however the size of community. Most relevant users to a specific stream (post) are filtered and posts propagate to specific users and thus we can overcome the problem of information overloading.

Next, in order to be able to detect whether the proposed similarity approach is able to reveal the most candidate users based on the content features, we apply the Mean Average Error (MAE). MAE is used here to measure the average absolute deviation between a predicted set of users among several neighborhood users.

Therefore, we repeat the previous experiment and get top 10 users for each of the 16 groups and get their similarity values. As shown in figure2 which represents the MEA between content similarities for top 10 users in all groups. According to the figure, the difference between similarity values is minor however the group size is which ensure the accuracy of user prediction.

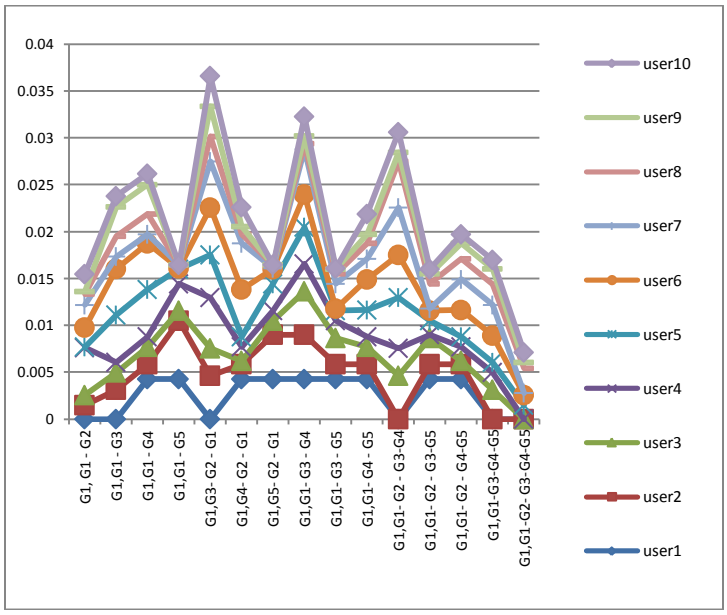

Fig. 2. MEA of content similarities (Top 10 users)

B. *Experimental evaluation criteria*

Other experiments have been applied to measure the efficiency of the proposed model.

Therefore, we first apply the content similarity module, get top 20 users similar to a random post, get their similarity values, and finally apply the social similarity module and get top 10 users. This experiment shows the effect of using addition social pattern features.

TABLE III. User Similarity (Top 10)

Model -rank	model-similarity value	Content-similarity -rank	content-similarity -value
1	0.99996	19	0.60820
2	0.99989	1	0.58928
3	0.99989	15	0.59993
4	0.99987	4	0.59016
5	0.99982	7	0.59225
6	0.99980	6	0.59118
7	0.99975	9	0.59423
8	0.99968	12	0.59860
9	0.99961	14	0.59886
10	0.99959	20	0.61443

As per table3, similarity between users and target user$_x$ who post the stream significantly improved when applying the whole model. Furthermore, ranking of top users has been changed when using the additional features which reflect the fact that we should consider network features of users when measuring similarity. For example, after adding network data user1 was ranked 19 using content similarity only while user 2 was having the first place Next, in order to measure the accuracy of the whole model, we repeat the first experiment l by applying the whole model and obtain similarity value of top 10 users. According to Figure3, similarity value has been improved after applying the social pattern level. Thus, we get the top candidate users generated from content-similarity module from table2, get their social similarity values and filter the top 10 users for all test groups. It is significant to mention that having applied the social pattern similarity module outperform using only content similarity.

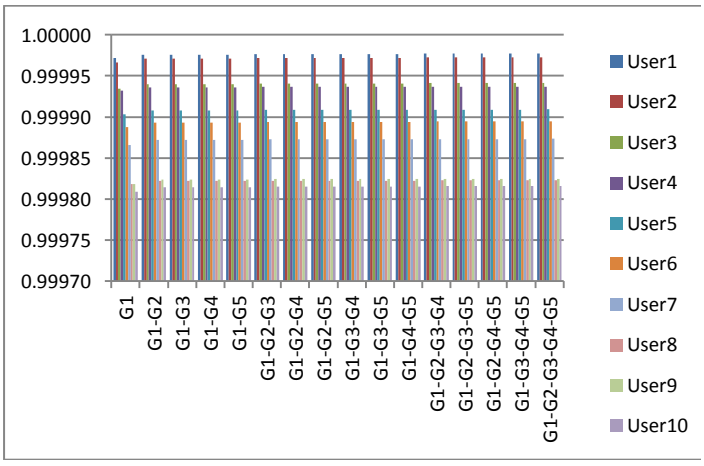

Fig. 3. User Social pattern similarity (Top 10)

Furthermore, longest common subsequence [LCS] method is used to measure the ranking used in our proposed model. Thus, we get the sequence of the top 10 users obtained from group1 containing 400 users, we apply the model all over other 15 groups and get the order of those users in each group as shown in figure4. According to figure4, there is no intersection points in the plotted area which mean that the order of predicted users remains the same however the size of the neighbourhood is.

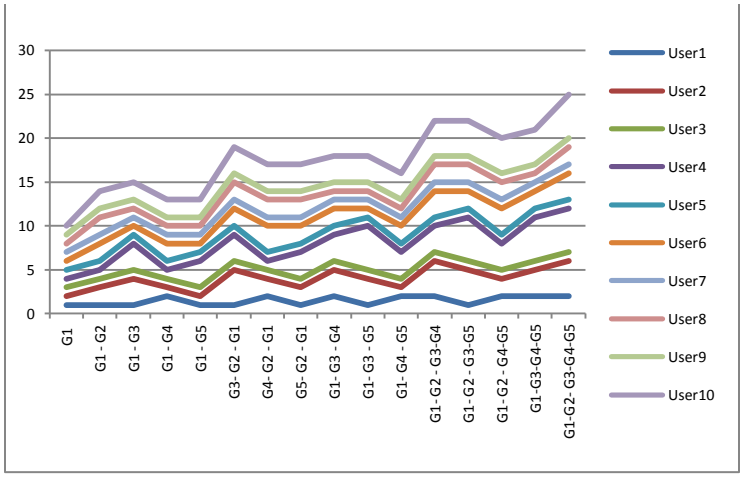

Fig. 4. longest common subsequence between users over groups (Top 10)

VI. CONCLUSION

This research proposes a model for enhancing the personalization of content dissemination among users in online social network with the aim of overcome the information overloading problem. The proposed similarity measurement model utilizes users' characteristics in social network. Such that when a target user posts a text, a set of most candidate receivers is generated and ranked by considering both similarity between this post and their interest as well as relevancy between social pattern of target user and others. Interest of users is extracted from content published by a user while social pattern is identified based on her social activities. Each user is represented by two vectors correspond to content and social pattern activities respectively. Different term weighting scheme was applied in order to weight social features and cosine similarity is then used to order the most similar users that is candidate to obtain that stream and to communicate with. The proposed model is applied on real dataset from stackoverflow and the experimental show that the accuracy of proposed method is precise as we obtain almost the same set of candidate users however the size of neighbourhood is. Furthermore, adding social pattern features in measuring similarity among user increase the similarity scores. In order to enhance the model, ontology could be used to measure content similarity among users. Furthermore, other topological features could also be used to rank users.

REFERENCE

[1] Cachia, R. (2008). Social Computing: The Case of Online Social Networking. IPTS Exploratory Research on Social Computing. JRC Scientific and Technical Reports.

[2] Güç, B. (2010). Information filtering on micro-blogging services (Doctoral dissertation, Swiss Federal Institute of Technology Zurich, Institute of Information Systems).

[3] Banerjee, S., Ramanathan, K., & Gupta, A. (2007, July). Clustering short texts using wikipedia. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 787-788). ACM.

[4] Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. Web Intelligence and Agent Systems, 7(2), 195-207.

[5] Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007, May). Measuring semantic similarity between words using web search engines. In Proceedings of WWW (Vol. 766).

[6] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In WWW '08: Proceeding of the 17th international conference on World Wide Web, pages 91–100, New York, NY, USA, 2008.

[7] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[8] Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the national conference on artificial intelligence (Vol. 21, No. 1, p. 775). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[9] Guy, I., Ronen, I., Wilcox, E.: Do you know?: recommending people to invite into your social network. In: International Conference on Intelligent User Interfaces, pp. 77–86 (2009) Terveen, L.G., McDonald, D.W.: Social matching: A framework and research agenda. ACM Trans. Comput. -Hum. Interact, 401–434 (2005)

[10] Lee, M. J., & Chung, C. W. (2011, January). A user similarity calculation based on the location for social network services. In Database Systems for Advanced Applications (pp. 38-52). Springer Berlin Heidelberg

[11] McDonald, D.W, 2003, Recommending collaboration with social networks: a comparative evaluation. In: Conference on Human Factors in Computing Systems, pp. 593–600 (2003)

[12] Ehrlich, K., Lin, C.Y., Griffiths-Fisher, V.: Searching for experts in the enterprise: combining text and social network analysis. In: International ACM SIGGROUP Conference on Supporting Group Work, pp. 117–126 (2007)

[13] Guo, L., Tan, E., Chen, S., Zhang, X., & Zhao, Y. E. (2009, June). Analyzing patterns of user content generation in online social networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 369-378). ACM.

[14] Adomavicius, G. and Tuzhilin, A. 2005. Personalization technologies: a process-oriented perspective. ACM. 48, 10, 83-90.

Guo, L., S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, analysis, and modeling of BitTorrent-like systems. In Proc. of ACM SIGCOMM IMC, 2005

[15] Leskovec, J L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In Proc. Of ACM SIGKDD, 2008

[16] Salton, G The smart retrieval system. Experiments in Automatic Document Processing, 1971

[17] Akcora, C. G., Carminati, B., & Ferrari, E. (2013). User similarities on social networks. Social Network Analysis and Mining, 1-21.

[18] Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social networks*, *25*(3), 211-230

[19] Wang, Shuaiqiang and Sun, Jiankai and Gao, Byron J and Ma, Jun}, Adapting vector space model to ranking-based collaborative filtering}, Proceedings of the 21st ACM international conference on Information and knowledge management}, pp{1487--1491}, 2012

TABLE II. USER CONTENT SIMILARITY (TOP 20)

	G1	G1 - G2	G1 - G3	G1 - G4	G1 - G5	G1 - G2 - G3	G1 - G2 - G4	G1 - G2 - G5	G1 - G3 - G4	G1 - G3- G5	G1 - G4 - G5	G1 - G2 - G3-G4	G1 - G2 - G3-G5	G1 - G3 - G4-G5	G1, G2,G4,G5	G1,G2 G3,G4,G5
user1	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144
user2	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082
user3	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071
user4	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021
user5	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020
user6	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999
user7	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989
user8	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988
user9	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986
user10	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969
user11	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947
user12	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942
user13	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928
user14	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923
user15	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912
user16	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910
user17	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902
user18	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898
user19	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897
user20	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893