Ranking User Influence in Healthcare Social Media

XUNING TANG
College of Information Science and Technology, Drexel University, PA, U.S.A.

and

CHRISTOPHER C. YANG
College of Information Science and Technology, Drexel University, PA, U.S.A.

Attribute to the revolutionary development of web 2.0 technology, individual users have become major contributors of web content in online social media. In light of the growing activities, how to measure a user’s influence to other users in online social media becomes increasingly important. This research need is urgent especially in online healthcare community since positive influence can be beneficial while negative influence may cause negative impacts to other users of the same community. In this paper, a research framework was proposed to study user influence within online healthcare community. We proposed a new approach to incorporate users’ reply relationship, conversation content and response immediacy which capture both explicit and implicit interaction between users to identify influential users of online healthcare community. A weighted social network is developed to represent the influence between users. We tested our proposed techniques thoroughly on two medical support forums. Two algorithms UserRank and Weighted in-degree are benchmarked with PageRank and in-degree. Experiment results demonstrated the validity and effectiveness of our proposed approaches.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval
General Terms: Algorithm, Measurement, Experimentation
Additional Key Words and Phrases: User Influence, Social Network, Social Media Analytics, Social Computing, Online Healthcare Community, Web Mining, Ranking Algorithm

1. INTRODUCTION

Web 2.0 technology brings revolutionary changes of our social life in the virtual world. In the last decade, a series of social networking sites such as Facebook, Twitter and LinkedIn has emerged and tightly connected web users from all over the world. Information is spreading out much faster than before throughout different online social communities. In recent years, besides the general social networking sites, the number of online healthcare social networking sites, such as MedHelp, PatientsLikeMe, Inspire, and HealthCentral, is increasing, which attract users with particular medical concerns. Attribute to Web 2.0 technology, e-patient can post a question on a discussion board or a message on their personal pages easily. As a result, online healthcare social networking sites become the popular platforms for e-patients to seek health-related information, share therapy, discuss medical issues, request or offer emotional support. In some cases, e-patients may identify side effects of new medications by information sharing before scientific reports are available. Patients living in remote areas may seek help from other e-patients when they do not have the local medical support.

A survey of US citizens conducted in December 2008 by Pew Internet Project (www.pewinternet.org) found that 61 percent of the respondents have searched online for health information [Domingo 2010]. In another similar survey conducted by iCrossing, 63 percent of the respondents reported performing searches at least once a month, 11 percent once a week and 12 percent two or more times a week [Domingo 2010]. In light of the growing activities in online healthcare communities, one major concern is the impact of these activities on the overall health outcome in a society. For instance, how does this user contributed information make impact on other users’ health conditions? Is there any inaccurate information disseminating throughout the online healthcare community? Through the
identification of influential users, the health professionals or public health workers can detect the information being disseminated by the influential users and how this information is propagated through the social media. Consequentially, they may determine whether intervention is needed to correct the inaccurate information or rumors being spread in social media. In addition, they can optimize the dissemination of health information in social media by initializing it through the influential users. Krulwich and Burkey developed a content-based system to find experts who answered certain questions by matching text of messages on bulletin boards [Krulwich and Burkey 1995]. Zhang et al. [Zhang et al. 2007] evaluated network-based approaches such as PageRank and HITS for expert finding in an online forum. However, limited previous works has integrated both content-based and network-based approaches. In addition, an influential user may not necessarily be an expert. Specifically, expertise represents the skill of answering some questions or conducting some activities. However, influence represents the capacity of drawing attention and producing impact. As a result, techniques of expert finding may not be directly applied to the problem of influential user identification. On the other hand, another line of studies focus on identifying seed users who can maximize the information propagation within a social network theoretically [Chen et al. 2009; Domingos and Richardson 2001; Domingos and Richardson 2002; Kempe et al. 2003; Kimura and Saito 2006; Leskovec et al. 2007]. These works are motivated by the viral marketing application. They assume that the social network structure and the influence probabilities on the edges of pairs of customers are given as inputs to the influence maximization problem. However, the influence probabilities are not always available unless prior knowledge of the relationships between actors is accessible. The problem we investigate here is extracting the influential users given the user interactions of a social media platform rather than the influence maximization problem.

In this work, we propose techniques to compute the user influence by combining content-based and network-based approaches. Moreover, we analyze several features of online medical support forum and then introduce an approach to construct a social network according to replying relationships and compute weights of edges by making use of forum features. Another major contribution of this study is that we motive two effective methods, UserRank and Weighted in-degree, to quantify user influence. Intensive experiments on two online medical support forums were conducted to evaluate our proposed techniques.

This paper is organized as follows. In the next section, we discuss the related works, and then highlight the research gap and propose our research question. In Section 3, we introduce our proposed techniques to solve the research problem defined in section 2. Experiment results and discussions are presented in Section 4. We conclude our work and discuss the future work in Section 5.

2. RELATED WORK

In the literature, there are a number of works related to influential user identification in social networks, namely, (i) influence maximization, (ii) social network construction and influence probability computation, and (iii) expert findings.

2.1 Influence Maximization
Since innovative ideas, opinions, and recommendations can propagate by “word-of-mouth” effect, social networks are an ideal platform for studying the information dissemination. In viral marketing, the objective is to utilize customers to recommend commercial products to their friends through the customer social networks. The influence maximization problem is defined as: given a customer social network and the influence probabilities between the customers for all edges on the network, we want to identify an initial set of users that will maximize the influence propagation such that all customers in the network will adopt the new product in the shortest time.

This influence maximization problem was first studied as an algorithmic problem from a data mining perspective by Domingos and Richardson [Domingos and Richardson 2001; Domingos and Richardson 2002]. In their work, the problem was considered in a probabilistic model of interaction, and heuristics were employed to select initial users. Kemp et al. tackled roughly the same problem by formulating it as a discrete optimization problem and proposing a greedy approximation algorithm that was applicable to several different models [Kempe, Kleinberg and Tardos 2003]. They proved that this optimization problem was NP-hard and showed that the greedy algorithm can guarantee the influence spread within (1-1/e) of the optimal influence spread. Experimental results of their studies also showed that their greedy algorithm significantly outperformed the classic centrality-based heuristics in influence spread. However, efficiency is a serious drawback of the greedy algorithm. To address this efficiency issue, a recent line of research works were developed to improve Kemp’s greedy algorithm based on “submodular property”. Kimura and Saito defined two shortest-path based independent cascade models within which influence spread can be efficiently computed [Kimura and Saito 2006]. However, these two shortest-path models were special cases of general independent cascade model studied in [Domingos & Richardson, 2001, 2002; Kempe, et al., 2003] so that the research problem remained unsolved. Leskovec et al. [Leskovec, Krause, Guestrin, Faloutsos, VanBriesen and Glance 2007] modeled a new problem namely outbreak detection and proved that influence maximization problem was a special case of outbreak detection problem. They proposed a “Cost-Effective Lazy Forward” (CELF) scheme by using the submodular property of the influence maximization objective to achieve 700 times speedup in selecting seed vertices comparing to basic greedy algorithm in [Kempe, Kleinberg and Tardos 2003]. Although Leskovec et al. achieved impressive improvement comparing to the basic greedy algorithm, as discussed in [Chen, Wang and Yang 2009], CELF still face serious scalability problem. To solve this problem, Chen et al. proposed new heuristics that have influence spreads close to the greedy algorithm while running at more than six orders of magnitude faster than the greedy algorithm [Chen, Wang and Yang 2009].

The objective of these studies is identifying seed users who maximize the information propagation within a social network given that the social network structure and the influence probabilities between any two connected customers are given as inputs. The required inputs are not always available in the real world problem. For example, given an online social networking platform, mechanisms must be developed to extract the social network and determine the influence probabilities. In addition, in this work, we are not interested in solving an optimization problem in order to maximize the influence in a social network but to
identify the influential users given the interaction patterns and user contributed content of an online social network. By identifying the influential users and examining the content contributed by these users, we are able to understand the impacts that they are making in the community and who they are influencing.

2.2 Social Network Construction and Influence Probability Computation

As discussed in the previous section, both of the social network structure and influence probabilities (contagion probability) in influence maximization problem are assumed to be given as input. A proper network structure and precise influence probabilities make a substantial impact on the final result. However, it is a non-trivial task of extracting social network structure and computing the influence probabilities precisely in many practical situations. There are many possible relationships between actors in a social network. Besides, not all relationships correspond to the same influential probability. To address these issues, a line of research works focusing on understanding the correlation between social affiliations and user behavior are conducted [Anagnostopoulos et al. 2008].

Anagnostopoulos et al. considered that one user can make influence on another user because of three reasons [Anagnostopoulos, Kumar and Mahdian 2008]. The first is called induction, where an action of a user is triggered by an action of another user. The second is homophily, which means similar users like to gather into clusters and make influence on each other. The third is confounding factors, which include common location, gender, school and several other external factors. Corresponding to each of these three reasons, we can construct social network and estimate influence probabilities differently.

- **Induction – based approach**

Before building the social network and estimating influence probabilities according to users’ actions, Anagnostopoulos et al. considered that it was important to verify that induction was the cause of social correlation [Anagnostopoulos, Kumar and Mahdian 2008]. To achieve this goal, they designed two delicate tests namely, shuffle test and edge-reversal test, which can identify induction as a source of social correlation, based on the assumption that timing of actions should be matter if induction is a likely cause of correlation. Although Anagnostopoulos et al. designed tests to identify cause of social correlation, they did not propose any techniques to quantify influence probabilities in the case of induction. Goyal et al. followed similar logic and developed techniques to model social correlation attribute to induction [Goyal et al. 2010]. They proposed to establish relationship between users by scanning user’s action log. If user A made an action ahead of user B, there will be an edge directing from A to B representing the influence from A to B. The influence probability between these two users is determined by common actions and time issues.

- **Homophily – based approach**

Besides considering induction effect, other researchers developed the framework to incorporate homophily effect. Crandall et al. raised the issue that induction can push systems toward uniformity of behavior, while homophily lead to fragmentation [Crandall et al. 2008]. To estimate the influence probabilities between two
users, they proposed a probabilistic model which made use of the interaction of induction and homophily effect. With similar idea, Matsuo and Yamamoto studied the bidirectional effects between user trust and rating behavior on E-commerce site [Matsuo and Yamamoto 2009]. They claimed that the rating of a user was influenced by the ratings of his/her trusted peers. Moreover, they considered that users tended to trust other users who have similar rating behavior with them. Following this observation, the authors proposed a theoretical model to reinforce both users’ rating score probabilities and trusting score probabilities. Although the authors did not directly address the same problem that we discussed above, to some extent, these trusting scores between users can be used to represent influence probabilities.

- **Confounding factor – based approach**

Singla and Richardson proposed to study the probability of relationship existing between two users by measuring their similarity [Singla and Richardson 2008]. They investigated the problem upon a collection of MSN Messenger logs. They found that users who were similar were more likely to chat with each other, where similarity was measured in terms of the matching of features such as zip code, age, gender, word and queries issued. According to their finding, social network could be constructed based on users with common features and the influence probabilities could be estimated by their similarity.

Although these research works proposed multiple techniques of generating social network and estimating influence probabilities, they did not directly deal with the problem of identifying influential users from a social network. Indeed, some of these approaches can be employed to provide inputs to the influence maximization problem introduced in Section 2.1. To address the problem of identifying influential users, we need to understand the properties of the online social networking environment and develop the mechanism to extract the social network structure, develop formulations to compute the influence probabilities, and develop algorithms to rank the influential users.

### 2.3 Expert Finding

The research in expert finding is highly relevant to the research problem of identifying influential users, although the definitions of expert and influential user are not completely the same. Expertise represents the skills of answering questions or conducting certain activities while influence corresponds to the capacity of drawing attention and producing impact. In the literature, methodologies of expert finding can be divided into three categories:

- **Content-Based Approach**

One straightforward approach of finding expert is content matching. The system extracts related texts contributed by a user and represents this user by a term vector. At the same time, expert finding query can be also built as a term vector so that the system extracts experts effectively by standard information retrieval techniques using vector space model. Krulwich and Burkey developed a content-based system, ContactFinder, to find the expert to answer certain questions by matching text of messages on bulletin boards [Krulwich and Burkey 1995]. However, due to the limitation of NLP techniques, it’s difficult for a
machine to really understand and evaluate given texts. Without addressing these limitations, the performance of the content-based approach is unstable and time consuming.

- **Network-Based Approach**

Aside from content-based techniques, another approach made use of link analysis techniques to investigate the social network with the purpose of extracting experts. Schwartz and Wood were first to analyze email flows to identify groups of individuals with common interests [Schwartz and Wood 1993]. They constructed a social network purely based on the history of communications among people, and then extracted implicit core network of specific topics from the social network. The algorithm returned a list of closely related people who were supposed to be experts of the given topic. Instead of extracting core network, Compbell et al. used a graph-based ranking algorithm to rank users’ expertise [Campbell et al. 2003]. They grouped emails into clusters according to their content, and then constructed a social network for each cluster of emails. HITS algorithm was used to rank users’ expertise levels within each social network. They demonstrated that network-based approach can outperform simple content-based approach for finding experts. Similar to Compbell’s idea, Jurczyk and Agichtein proposed to use HITS algorithm to discover authorities in question answering community [Jurczyk and Agichtein 2007]. The assumption behind their work was that a good asker asked good questions to attract answers from good answerers, while good answerers answered good questions asked by good askers. They presented experiments to show that HITS algorithm outperformed degree-based techniques in terms of extracting experts from a question answer community. Zhang et al. employed network-based approaches such as PageRank and HITS to find experts in an online forum [Zhang, Ackerman and Adamic 2007]. However, experiment result showed that complex algorithms, such as PageRank and HITS, did not outperform some relatively simple centrality measures, e.g. Z-Score and In-Degree centrality. In a very general sense, the technique developed by Tang et al. can be considered as a Network-Based approach [Tang et al. 2009]. They proposed the Topical Affinity Propagation (TAP) approach and represented the expert finding problem by a graphical probabilistic model. A topic factor graph (TFG) was introduced to incorporate topical similarity, interaction strength, representative users, and other network information into a unified probabilistic model. The TAP approach was utilized to maximize the probability of TFG and compute the user influence probabilities of each topic based on the network connectivity. Finally they took these influence probabilities as input and employed a modified PageRank algorithm to rank users’ expertise levels. However, efficiency is a serious concern of their approach. Estimating the value of all hidden variables of their model is very time consuming which is impractical for very large dataset.

- **Hybrid Approach**

Since content-based or network-based approach alone has its limitations, Bian et al. develop a hybrid approach which incorporated content-based and network-based approach into a unified framework [Bian et al. 2009]. Their objective was to extract reliable users and content in question answering community. They proposed to make use of pre-selected features and employ a logistic regression function to compute
message quality and reputation of asker and answerer, and then reinforce the result by a HITS-like algorithm. Although their techniques can improve the accuracy of searching over Question Answering (QA) archives for best answers, their approaches relied on QA features which were not available for identifying influential users from online forum, such as question starts, number of answers, best answers, total thumbs up/down and so on. Goyal et al. studied a similar problem namely leader discovery [Goyal et al. 2008]. In their work, the concept of leader was slightly different from expert or influential users. They defined leader as a user who leaded a set of users in a set of actions. They assumed that influence was propagated from user A to user B if and only if user A conducted an action ahead of user B. Examples of these actions include tagging resources, rating books, and buying products online. They developed an algorithm to identify leaders based on frequent pattern discovery. Experiment result showed that the proposed algorithm was efficient and effective. However, their definition of leaders may not be applicable to the problem of identifying influential users in an online forum. Furthermore, many reasons will trigger different users to perform similar actions so that leadership is merely one of the many explanations.

According to the literature review in this section, we have identified some important research gaps and limitations of current studies. First of all, most of the current studies of influence maximization focused on the viral marketing application, which is fundamentally different from an online forum scenario. We do not intend to optimize the influence in an online healthcare social network but to identify the influential users given the interaction patterns and user contributed content in the online community. As a result, the influence maximization techniques cannot be applied directly to identify influential users. Secondly, it is worth emphasizing that, different from expert finding, our work focuses on identifying influential users who are able to raise interesting messages and draw immediate and extended attentions. Although influence and expertise are two closely related concepts, there are distinctions between them which motivate different approaches to address the problem. For example, expertise represents the skill of answering questions so that “Best Answer” and “Thumbs up” should be effective indicators of user’s expertise in QA portal or online product review board. However, influential users are good at raising interesting messages and drawing widely attention which do not necessary equal to expertise or authority. As a result, these QA features such as “Best Answer” and “Thumbs up” may not be effective to indicate influential users. Thirdly, content-based approaches have their own advantage of analyzing the content of conversations among users, while they are insufficient to depict the complicated interactions among users. Similarly, network-based approaches are good at representing the complex interaction network but completely ignore the content of users’ conversation. To our best knowledge, there are limited works on incorporating content-based and network-based approaches to quantify the user influence and rank the influential users in an online community. In this work, we introduce the framework of influential user identification, present the mechanisms to construct the online social network and compute the user influence weights on the constructed network, and investigate algorithms for ranking influential users.

3. METHODOLOGY
In this section, we will first introduce the framework of identifying influential users in a given online forum. Secondly, an online healthcare community, MedHelp, will be briefly introduced, which is used as a test bed in this work. Thirdly, we will introduce how to extract a weighted social network based on the collected threads of a forum. Finally, we will motivate the UserRank and Weighted in-degree algorithm to rank influential users within an online forum.

3.1 Framework of Identifying Influential Users in Online Forum

Figure 1 demonstrates the framework of identifying influential users in an online forum. We divide the whole process into three steps: (1) Forum data collecting phase, (2) Social network constructing phase, and (3) User Influence quantifying and ranking phase.

In the forum data collecting phase, a crawler was built to collect all threads and replies on the discussion board of the forum of interest. In addition, a parser was built to parse and filter the collected data. For each thread, we generated a formatted thread record which consisted of TID (unique ID for each thread), Thread Title, Thread URL, Thread Initiator ID, Timestamp, List of Replier ID and Thread Content. For each reply of a thread, we created a formatted reply record which was composed of RID (unique ID for each reply), Thread Title, Thread URL, Replier ID, Timestamp and Reply Content. The formatted data was stored in a database which provided inputs to the social network constructing phase. In the social network constructing phase, reply relationships were first extracted to construct an unweighted social network. Selected features, such as conversation content similarity and response immediacy, were incorporated to compute the weights of the edges on the social network. In the user influence quantifying and ranking phase, automatic techniques were developed to calculate the user influence on the weighted social network.

Fig. 1. Framework for identifying influential users in an online healthcare support forum

3.2 MedHelp Overview and Data Acquisition
In this work, we focus on the MedHelp social networking site, which is one of the largest online healthcare communities in the world. Since its opening in 1994, nearly 3 million threads were posted in the community and it attracted over 8 million visitors every month. MedHelp members interact with each other either by personal profile pages or by public medical support forum. Each registered MedHelp member has his/her personal profile page, where he/she can post journal messages, receive notes from their friends, track his/her health condition and so on. Personal profile page can be set to private. In that case, only the host or his/her friends can visit. On the other hand, medical support forum is a public site that every MedHelp member can visit and leave messages. In the MedHelp online healthcare community, there are 189 medical support forums, each of which is corresponding to one medical issue such as alcoholism, swine flu or smoking addiction. MedHelp member can be a member of one or more medical support forums. Registered members can post a question or make an announcement to initiate a thread. Other members can make comments and specify who they are replying to in the thread. In this work, we focus on identifying influential users in the forum format but our proposed techniques can be easily extendable to other formats such as journals and notes on personal profile pages.

Since each forum corresponds to one disease, threads and replies of a forum are mainly focusing on the prevention, symptom, therapy and other information of the corresponding disease. Due to MedHelp’s social networking site’s nature, MedHelp members are usually offering a lot of social support and informational support through the forum.

In this work, our test bed consists of two medical support forums: Alcoholism forum and Smoking Addiction forum. We built our generic crawler to fetch all HTML pages in the discussion board of MedHelp forums. Then we built a parser to extract threads, replies, timestamp, and user information from each extracted HTML page. A formatted record is identified by its message ID. The timestamp and user ID who post the message are also recorded in our database. Each message is also associated with a thread ID. The detailed statistics of these two medical support forums will be presented in experiment section.

3.3 Extracting Social Network from an Online Forum

3.3.1 Social Network Topology

Social network is a convenient and effective way to represent user interactions. Each vertex of a social network represents a social actor. Two social actors who are interacting with each other are connected by an edge in a social network. Depending on the specific applications and interactions, a social network can be constructed in different way. For example, a social network can be extracted from a collection of emails where email sender and receiver are connected by an edge. Similarly, an edge of a social network can also correspond to questioner-answerer relationship in QA portal or phone caller-receiver relationship. However, these techniques cannot be applied to threads in an online forum scenario because user interaction is more intensive and complex in a forum.

A forum consists of a number of threads. A forum thread is composed of a number of messages as shown in Figure 2 (a). Different from a 1-to-1 phone caller-receiver relationship or a 1-to-many email sender-receiver relationship, an initial post of a thread can attract multiple replies and each reply may
attract many other replies, leading to a hierarchical structure. Figure 2 (b) represents a hierarchical tree transformed from the thread shown in Figure 2 (a). In this work, a social network is constructed by extracting the users and their interactions in a hierarchical tree of a thread based on three observations:

O1: Direct reply from a user to another user in a thread represents an interaction. In other words, when A replies to a message posted by B, there should be an edge connecting vertex A and vertex B in the social network to capture their interaction.

O2: Besides direct reply, indirect reply should be also taken into account. For example, when B replies to A and then C replies to B, it is possible that C is not only replying to B but also addressing to the message posted by A. The indirect reply can properly capture the interaction between C and A which can be missed if we only consider direct reply. In this case, three edges should be created in a social network corresponding to the interactions between A and B, B and C, and A and C.

O3: Most of the threads in MedHelp medical support forum were initialized by the members who were seeking information help. When B is replying the question posted by A, we consider that B is offering some kinds of support and influence to A. As a result, the direction of an edge should be made from the one who receives the reply to the one who makes the reply to confer authority.

Based on these observations, we can construct a social network by extracting users and their direct/indirect interactions from a hierarchical tree. Figure 2 (c) represents the corresponding social network extracted from the tree shown in Figure 2 (b). We can construct one social network for each thread in the forum, and
then integrate these social networks together by taking the union of these social networks to form the complete social network of the forum.

As introduced in section 2, a social network can be constructed by induction-based approach, homophily-based approach or confounding factor-based approach. In the induction-based approach, a basic assumption is that the action of individuals can influence their friends to act in a similar way. Under such an assumption, Goyal et al. proposed to construct an action propagation graph with directed edges from one user to another if one user made an action ahead of another user who made the same action later [Goyal, Bonchi and Lakshmanan 2010]. Examples of these actions include making a social tag on a Web object, purchasing a product on an electronic commerce site, and contributing an opinion on a consumer reviewer site. We argue that users making the same action are not necessarily induced by the one who make the action earlier but can be caused by other plausible reasons including coincidence. In our approach of constructing social networks of online forums, the induction is obvious since a user will not reply unless he/she has read the previous messages in a forum thread. Goyal et al. further proposed three types of models, static model, continuous time model and discrete time model, to capture influence probability from one user to another which we will discussed in more details in the next section. In the homophily-based approaches, a social network can be constructed by connecting individuals with similar characteristics, for instance like or dislike certain product. Similarly, in the confounding factor based approach, a social network can be built by connecting individuals who share common background factors, for example living in the same city or graduating from the same college. In the MedHelp forums, each forum is focused on a specific health problem or disease. Members of a specific health problem or disease forum are either patients or caregivers of a patient with the same disease. Occasionally, some health professionals may involve in the discussion. The messages contributed by the members are all related to specific issues of the disease such as treatments, medications, symptoms, etc. The homophily properties are embedded in the interactions of users when they discuss the specific issues. However, it is infeasible to utilize the confounding factor because the demographic information is usually unavailable in the user profile page mainly due to the privacy concerns. E-patients are more comfortable to discuss the health issues without being identified. In our work, in addition to connecting users together according to their interaction, we also propose mechanisms to compute weights on the social network which will be introduced in the next section.

3.3.2 Computing Influence Probability of a Social Network

By following the techniques proposed in section 3.1, we construct a social network, \( G = (V, E) \), in which \( V \) is a set of nodes corresponding to the members of a MedHelp forum and \( E \) is a set of edges corresponding to the interactions between members. As introduced in section 2, many previous studies computed user influence or authority based on the topology of an unweighted social network. They either simply employ the degree centrality of a node to represent its importance in a social network, or make use of PageRank-like algorithm to rank the authority of the users based on the assumption that a user with high authority
should have a large in-degree but a small out-degree. In-degree represents the number of head endpoints adjacent to a node while out-degree represents the number of tail endpoints adjacent to a node.

Although these techniques can identify important users in some applications, there are also limitations. First of all, most existing approaches treated each edge equally important which does not necessarily reflect the actual importance between users in reality. In an unweighted social network, two users are connected by an edge if they interacted with each other in one thread of the forum no matter how intensive this interaction is. An unweighted social network itself cannot differentiate intensive interaction between two users and the interaction that only happened one time in the forum. Secondly, several important forum features cannot be incorporated into an unweighted social network which may affect the performance of identifying influential users. For example, a user making a reply may not always continue the original conversation but raise a new issue which can divert the discussion topic in the rest of a thread. In this case, the one who receives reply should not confer authority to the one giving rely because of the irrelevance of the reply. However, an unweighted social network does not differentiate relevant/irrelevant replies, even though topic shift happens very often in an online forum. To address these two major limitations, we propose to compute weight for each edge to represent the influence probability from one user to another through an edge.

It is important to note that some existing works also took into account time information when computing influence probabilities. Goyal et al proposed three different type of model to quantify the influence probabilities on edges from one user to another. The first model is called static model in which the influence probability of user v on u is estimated by $\frac{A_{v-u}}{A_v}$ where $A_{v-u}$ denotes the number of actions propagated from v to u, $A_v$ represents the number of actions performed by user v. In their second model, namely continuous time model, they incorporate time effect by multiplying $\frac{A_{v-u}}{A_v}$ to a time decaying factor. In the third model, discrete time model, they approximate the continuous time model by assuming that the influence from user v to u constantly equals to $\frac{A_{v-u}}{A_v}$ within a time interval and it is equal to 0 after that interval. However, as we have discussed, Goyal et al’s assumption does not hold in an online healthcare forum since we cannot observe action propagation explicitly and the same action performed by two users can be contributed by factors other than induction or influence. In addition, Goyal et al. computed the influence probabilities based on the ratio of the number of actions made ahead of another user and the total number of actions made and the timing factor. In this work, we propose to incorporate both content similarity and response immediacy to compute the influence probabilities instead of simply counting the number of posted messages and replied messages.

In healthcare social media, patients and caregivers are mostly seeking or offering social supports including informational support and nurturant support [Chuang and Yang, 2010]. Participants in healthcare online forum are engaged on specific health issues in the discussions. Chuang and Yang [2010] found that 72.8% of MedHelp Alcoholism forum were seeking information support and 44.4% of that were seeking nurturant support. Participants who offer information supports may offer advice or referral to address
specific information needs of other participants. The content will carry through from one message to another message. Active participants in the healthcare online forum also offer emotional support to meet the immediate needs of other participants especially in certain health problems such as alcohol and smoking addiction. However, forums on entertainment, consumer products, and sports usually do not include any nurturant support. Although users in these forums express their opinions and share information on a common object such as a movie, a book, and a handheld device, they usually do not request for advice and referral of a particular problem such as medical treatment or drugs that are required to meet an immediate medical need. In fact, user opinions on a particular consumer product do not require an immediate response; however, a medical informational and nurturant support request requires a timely response. Opinions on a consumer product can be very diverse; however, an information support or a nurturant support for a healthcare issue must meet the specific content requirement. As a result, content similarity and response immediacy are two features that are important to measure the influence made by the healthcare social media users.

In this work, we develop a weight function by incorporating both content similarity and response immediacy to compute weights, leading to a weighted social network $G' = (V, E, W)$, where the node set $V$ and edge set $E$ are constructed by the same approach introduced in section 3.1, and the weight set $W$ corresponds to a collection of weights $\{w_{ij}\}$, for each edge in $E$.

Formally, given a forum, there are a collection of $N$ threads, $T_1, T_2, \ldots, T_N$, which consist of messages posted by $n$ users $v_1, v_2, \ldots, v_n$. Let $M_{k,l}$ to be the $l^\text{th}$ message of $T_k$, $V(M_{k,l})$ to be the user who posts $M_{k,l}$, and $\text{Time}(M_{k,l})$ to be the timestamp of message $M_{k,l}$. In addition, let $\text{Similarity}(M_{k,a}, M_{k,b})$ denotes the content similarity between messages $M_{k,a}$ and $M_{k,b}$, and $\text{ReImmediacy}(M_{k,a}, M_{k,b})$ represents the response immediacy between two messages. The weight $w_{ij}$ of edge $e_{ij}$ between $v_i$ and $v_j$ can be computed by:

$$W(v_i, v_j) = \frac{1}{M} \sum_{k=1}^{N} \sum_{a:b:v(M_{k,a})=v_j, \text{Time}(M_{k,a})=\text{Time}(M_{k,b})}^{\text{M}_{k,l}} \alpha(\text{ReImmediacy}(M_{k,a}, M_{k,b})) + (1 - \beta)\text{Similarity}(M_{k,a}, M_{k,b})$$

where $M$ is the number of threads that $v_j$ has replied to $v_i$, $\alpha$ and $\beta$ are parameters with values between 0 and 1. In this work, each message $M_{k,l}$ is represented by a vector of terms with high TFIDF value, where the TFIDF value of a given term $t_i$ in a specific document $j$ is defined as,

$$\text{tfidf}_{t_i,j} = \frac{n_{t_i,j}}{\sum_k n_{t_i,k}} \log \frac{|D|}{|\{d: t_i \in d\}|}$$

with $n_{t_i,j}$ denotes the number of term $i$ appearing in document $j$, $|D|$ represents the total number of documents in the corpus, and $|\{d: t_i \in d\}|$ equals to the number of documents that contain $t_i$.  

13
After that we use cosine similarity of two term vectors to represent the content similarity of two messages, which can be defined formally as,

\[
\text{Similarity}(M_{k,a}, M_{k,b}) = \frac{\sum_i (tfidf_{i,a} \times tfidf_{i,b})}{\sqrt{\sum_i tfidf_{i,a}^2} \times \sqrt{\sum_i tfidf_{i,b}^2}}
\]

Response immediacy, \(\text{RelImmediacy}(M_{k,a}, M_{k,b})\), can be defined in two different ways:

**Definition 1** (time difference): response immediacy between two messages can be defined by the time difference, \(\text{Time}(M_{k,b}) - \text{Time}(M_{k,a})\), of two messages.

**Definition 2** (path length): response immediacy between two messages can be defined by the length of the downward path from one message to another message in a thread tree. As an illustrative example, in figure 2 (a), the path length from the initial message posted by user A and the last message posted by user D equals to 3, which can be easily counted in figure 2 (b).

In our experiment presented in Section 4, we compared the impact of these distance functions on the proposed ranking algorithm.

It is worthy to note that there are other social media features that can also be considered in computing the weights. For example, the length of messages can be considered as the amount of effort contributed by a user, and therefore, a higher weight may be assigned to users who contribute longer messages. In this work, we do not intent to cover all features comprehensively but investigate the two features namely, content similarity and response immediacy. However, this work can be easily extended to incorporate other potential features.

### 3.4 Computing user influence

Given the weighted social network \(G'\) as discussed in Section 3.3, we investigate several algorithms to rank the influential users in \(G'\). Consider different possible in-link structures involving user A in the weighted social network as shown in Figure 3. In Figure 3 (a), user A makes influence on user B, C, and D. In Figure 3 (b), user A makes influence on user B only. In this case, since user A has larger Weighted in-degree in Figure 3 (a) than in Figure 3 (b), user A should has larger influence in Figure 3 (a) than in Figure 3 (b). On the other hand, in Figure 3 (c), user B is influenced by three users A, C, and D while user B is only influenced by user A in Figure 3 (b). In this case, in Figure 3 (b) user A would receive more credit from user B than in Figure 3 (c). According to these observations, we proposed two different approaches of computing user influence: **Weighted in-degree** and **UserRank**.
Weighted in-degree

In a directed graph, In-degree of a node is the number of head endpoints adjacent to this node. In graph theory, In-degree is widely used as a centrality measurement to quantify the importance of a node within a network.

\[ ID(v_j) = \sum_{v_l, e_{ij} \in E} 1 \]

With edge weights computed, Weighted in-degree is a straightforward way of computing user influence. Given a weighted social network, user’s influence score is equal to the sum of weights on all in-link edges of the network.

\[ WID(v_j) = \sum_{v_l, e_{ij} \in E} W(v_l, v_j) \]

UserRank

Since the user influence within a social network is similar to the web page popularity in a hyperlink network, we also extended PageRank[Page et al. 1997] algorithm and proposed a UserRank algorithm to quantify user influence in a weighed social network that we constructed.

In PageRank, the transition probability \( P(v_j | v_i) \) equals to 1/out-degree(\( v_i \)), where out-degree(\( v_i \)) is the number of out-links of \( v_i \). In PageRank algorithm, all web pages are initialized with a unique PageRank score, which equals to one over the total number of web pages of the whole hyperlink network. The PageRank scores are iteratively updated by the following formulation until convergence.

\[ PR(v_j) = (1 - d) + d \sum_{v_l, e_{lj} \in E} P(v_i | v_j) PR(v_j) \]
To incorporate the content similarity and response immediacy, in our proposed UserRank algorithm we employ a weighted social network. As a result, the transition probability is computed as follow:

\[ P(v_j | v_i) = \frac{W(v_i, v_j)}{\sum_{v_k \in E} W(v_i, v_k)} \]

The UserRank scores will then be computed as follow:

\[ UR(v_j) = (1 - d) + d \sum_{v_k \in E} \frac{W(v_i, v_j)}{\sum_{v_k \in E} W(v_i, v_k)} UR(v_i) \]

4. EXPERIMENT AND DISCUSSION

4.1 Experiment Datasets

<table>
<thead>
<tr>
<th>Statistics of Collected Datasets</th>
<th>Smoking Addiction Forum(^1)</th>
<th>Alcoholism Flu Forum(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Registered Members</td>
<td>352</td>
<td>446</td>
</tr>
<tr>
<td>Average Number of Friends for each Members</td>
<td>12.7</td>
<td>9.2</td>
</tr>
<tr>
<td>Total Number of Threads</td>
<td>554</td>
<td>737</td>
</tr>
<tr>
<td>Total Number of Replies</td>
<td>2339</td>
<td>5307</td>
</tr>
<tr>
<td>Average Number of Replies per Thread</td>
<td>4.3</td>
<td>7.2</td>
</tr>
</tbody>
</table>

1. URL: [http://www.medhelp.org/forums/Smoking-Addiction/show/159](http://www.medhelp.org/forums/Smoking-Addiction/show/159)
2. URL: [http://www.medhelp.org/forums/Alcoholism/show/158](http://www.medhelp.org/forums/Alcoholism/show/158)

As introduced in section 3, we built a crawler to fetch all HTML web pages from the discussion board of two medical support forums, including alcoholism forum and smoking addiction forum. A parser was also built to extract threads, replies, timestamps, and user information from each HTML page. Table I describes the statistics of the two collected datasets.

4.2 Evaluation

**Gold Standard**

To evaluate the performance of our proposed techniques, four human annotators were recruited to analyze these datasets in order to produce the gold standard. The gold standard was the ranking of influential users in each forum. The human annotators were asked to read through each thread in the forum and manually assign an influence score, between 0 (no influence) and 5 (strong influence), to each thread participant. The score was corresponding to the degree of influence from this user to other users within a thread. These influence score were assigned by following four basic principles: (1) the user who only posted spam message received zero point, (2) the more followers a user had, the higher point the user received, (3) the
higher quality of the message content posted by a user, the higher point the user received, and (4) the more impact the message made on the following messages, the higher point the user received. In the second principle, the number of follower of a user \( A \) in a given thread corresponds to the number of users who have replied, directly or indirectly, to user \( A \) in the thread and the replies are responding to the message posted by user \( A \). The final score of a user was computed as the average scores the user obtained from all of the threads in a forum. We computed the ranking of influential users by sorting the final scores of all users in descending order. Each medical support forum has a ranking of influential users as the gold standard (denoted as GS) in the experiment.

To determine the reliability of the rankings produced by the four annotators, we used the weighted Kappa measure to compute the inter-rater agreement. Weighted Kappa measure is a statistical measure extended from Kappa measure for computing the agreement between two ordered lists. It has a maximum value of 1 when there is perfect agreement between two raters and a value of 0 when the agreement is not better than by chance. In general, a weighted kappa measure value larger than 0.8 is considered a very good agreement between the two raters [Bakeman and Gottman 1997]. The average weighted Kappa measure was 0.813. The average weighted Kappa measure was computed by taking the mean of the weighted Kappa measures of all possible pairs of annotators. It demonstrated a strong agreement among the four annotators and the Gold Standard was reliable for the experiment. It is important to note that Ariel et al. [Ariel et al., 2000] stated that two to six judges can obtain substantial improvement practically and a recent study conducted by Bermingham and Smeaton [Bermingham and Smeaton, 2009] used an average of 3.6 annotators to annotate the TREC Blog corpus. In this experiment, four annotators were recruited and it was reasonably sufficient to produce a high quality gold standard.

Evaluation Metrics

In this experiment, we compared the Top-K influential user rankings produced by UserRank, PageRank, Weighted in-degree, and In-degree respectively with the Top-K ranking of GS to evaluate the performance of our proposed techniques. We employed Kendall’s tau measurement to compute the ranking similarity between two Top-K rankings. Kendall’s tau measurement can be used to quantify the similarity of two Top-K rankings according to the relative order of elements in two different rankings.

Given two top-k rankings \( \Gamma_1 \) and \( \Gamma_2 \), **Kendall’s tau** is defined as:

\[
K^0(\Gamma_1, \Gamma_2) = \sum_{(i,j) \in P(\Gamma_1, \Gamma_2)} \tilde{R}^0_{ij}(\Gamma_1, \Gamma_2)
\]

where \( P(\Gamma_1, \Gamma_2) \) is the set of all unordered pairs of distinct elements in \( D_{\Gamma_1} \cup D_{\Gamma_2} \), \( D_{\Gamma_1} \) and \( D_{\Gamma_2} \) are the domains of \( \Gamma_1 \) and \( \Gamma_2 \), and the value of \( \tilde{R}^0_{ij}(\Gamma_1, \Gamma_2) \) can be assigned either 0 or 1 based on the following conditions:

1) \( \tilde{R}^0_{ij}(\Gamma_1, \Gamma_2) = 0 \) if \( i \) and \( j \) exist in both top k rankings with the same relative order;
2) If \( i \) and \( j \) exist in both top \( k \) lists while the relative order \( i \) and \( j \) in one top \( k \) ranking is different from another;
3) \( \hat{R}_{ij}^0(\Gamma_1, \Gamma_2) = 0 \) if both \( i \) and \( j \) exist in one ranking while neither \( i \) nor \( j \) exist in another ranking;
4) \( \hat{R}_{ij}^0(\Gamma_1, \Gamma_2) = 1 \) if \( i \), but not \( j \), exists in one ranking and \( j \), but not \( i \), exists in another ranking;
5) \( \hat{R}_{ij}^0(\Gamma_1, \Gamma_2) = 1 \) if both \( i \) and \( j \) exist in one ranking with the relative order, \( i > j \), and only \( j \), but not \( i \), exists in another ranking;
6) \( \hat{R}_{ij}^0(\Gamma_1, \Gamma_2) = 0 \) if both \( i \) and \( j \) exist in one ranking with the relative order, \( i > j \), and only \( i \), but not \( j \), exists in another ranking.

Kendall’s tau measure takes the relative ranking orders of any two elements in the union of two top \( k \) lists into account. A lower value of Kendall’s tau measure corresponds to a closer pair of top \( k \) lists. An algorithm is considered the best when the Kendall’s tau measure between the ranking generated by this algorithm and the GS is the lowest.

4.3 Experiment Results

We conducted two experiments. In experiment A, we compared the performances of UserRank, PageRank, Weighted in-degree, and In-degree by evaluating the Top-10 rankings for each dataset. Kendall’s tau measurement was used to evaluate these techniques. We investigated the impact of two parameters, \( \alpha \) and \( \beta \), of the weight function on the performance. In experiment B, we evaluated UserRank and Weighted in-degree with different values of \( K \).

Experiment A

Alcoholism Forum

Figure 4 presented the result of Kendall’s tau between the top-10 rankings of GS and the four techniques, including UserRank, PageRank, Weighted in-degree and In-degree on alcoholism forum dataset. In this paper, we proposed two definitions for computing response immediacy. Figure 4 (a) presented the Kendall’s tau result computed by following definition 1 – time difference, while Figure 4 (b) showed the Kendall’s tau result computed based on definition 2 – path length. The experiment results demonstrated that: (1) both definitions of response immediacy were effective and they did not lead to significantly different results in alcoholism forum; (2) UserRank algorithm outperformed PageRank algorithm consistently which indicated the effectiveness of the weight function; (3) with similar reason, Weighted in-degree outperformed In-degree most of the time, except in extreme cases, where \( \beta \) equals to 0 or 1; (4) Weighted in-degree achieved the best performance when \( \beta \) equals to 0.2; (5) both UserRank and Weighted in-degree were not sensitive to the parameter \( \alpha \).
Fig. 4. Experiment result of alcoholism forum: (a) rank similarity based on Kendall’s tau measurement, where the response immediacy was employing definition 1 (D1) – time difference; (b) rank similarity based on Kendall’s tau measurement, where the response immediacy was employed definition 2 (D2) – path length.

Smoking Addiction Forum

Figure 5 presented the result of Kendall’s tau between the top-10 rankings of GS and the four techniques, including UserRank, PageRank, Weighted in-degree and In-degree on smoking forum dataset. Smoking addiction forum is smaller than alcoholism forum. The difference of forum size affects the performance of the proposed techniques. Figure 5 demonstrated that Weighted in-degree outperformed all other techniques most of the time. Moreover, UserRank algorithm had better performance when the response immediacy is computed by using time difference as measurement. It was interesting to note that UserRank performed worse than PageRank when the path length was employed as the response immediacy. The path length was not a good measurement of response immediacy partially due to the variation in the number of responses in threads. The situation was worse when the forum size was small.

Considering all experiment results above, we observed that: (1) Weighted in-degree outperformed other techniques in most cases and it was more computationally efficient than PageRank and UserRank; PageRank and UserRank algorithms compute the authority scores recursively, which take substantially more computing time to iterate until it converges, and therefore, are computationally expensive. (2) the proposed weight function was effective which was demonstrated by the better performance of Weighted in-degree and UserRank than In-Degree and PageRank; (3) the parameter $\beta$ made more substantial impact on Weighted in-degree than on UserRank; (4) Content similarity was relatively more important than response immediacy in the weight function (the best performance was achieved when $\beta$ was equal to 0.2), however completely ignoring the response immediacy (when $\beta$ was equal to 0) led to a worse performance; (5) UserRank did not outperform PageRank when the path length was employed in response immediacy, which
was opposite to the result we found in the Alcoholism forum dataset in both response immediacy measurements and the result we found in the Smoking Addiction forum dataset when time difference was employed in response immediacy. By comparing the different settings, we found such inconsistency occurred when the dataset was smaller (i.e Smoking Addiction Forum was smaller than the Alcoholism Forum) and the path length was employed for response immediacy. A plausible explanation was that the number of responses was relatively fewer in a smaller forum, and therefore, the path length might not be accurate in measuring the response immediacy. However, such inconsistency did not exist when time difference was employed for response immediacy. In Figure 6, we plotted the relationships between the path length and average time difference (in days) in Alcoholism and Smoking Addiction Forums. It showed the relationships between time difference and path length was relatively more inconsistent in the Smoking Addiction Forum than in the Alcoholism Forum. Such inconsistency may cause the inconsistent performance in ranking obtained by UserRank.

![Graph](image)

---

**Fig. 5.** Experiment result of smoking forum: (a) rank similarity based on Kendall’s tau measurement, where the response immediacy was employing definition 1 (D1) – time difference; (b) rank similarity based on Kendall’s tau measurement, where the response immediacy was employing definition 2 (D2) – path length.
Fig. 6. A plot of average time difference and path length in the Alcoholism Forum and Smoking Addiction Forum.

Experiment B

From experiment A, we observed that UserRank and Weighted in-degree outperformed PageRank and In-degree. The best performance was achieved when $\alpha$ was equal to 0.9, $\beta$ was equal to 0.2 and the response immediacy was computed based on the time difference. In this experiment, we compared the performance of Weighted in-degree, In-degree, UserRank and PageRank in terms of different values of $K$. We set $\alpha$ to 0.9, $\beta$ to 0.2 and the response immediacy as the time difference. Figure 7 presented the result of Kendall’s tau for $K = 5$ to 15. It demonstrated that Weighted in-degree consistently outperformed In-Degree and followed by UserRank and PageRank.

This result demonstrated that in-degree and Weighted in-degree performed consistently better than the recursive PageRank and UserRank algorithm respectively. In addition, by incorporating content similarity
and response immediacy, Weighted in-degree performed consistently better than in-degree and UserRank performed consistently better than PageRank.

5. CONCLUSION AND FUTURE WORK

In this study, we combined link-based approach and content-based approach to quantify user influence in an online medical support forum. We proposed effective techniques to construct social network according to replying relationships in forum threads. In order to capture user behavior and dynamic conversation, we also introduced a weight function, which incorporated content similarity and response immediacy, to assign weights to the links of the constructed social network. Using two online medical support forums with different size as test bed, we thoroughly evaluated our proposed technique with different parameters in discrepant settings. Since there is limited work considering influential user in a forum context, this study has two primary contributions. The first one is the approach that utilizes replying relationships of forum threads to construct social networks and then employs forum features such as content similarity and response immediacy to assign weights to network edges. Secondly, we motivated Weighted in-degree and UserRank to quantify user influence and evaluate their performance in two different online medical support forums.

Experiment results demonstrate that our proposed technique of constructing social network captures the interaction patterns and our weight function is effective in an online forum especially when the size of forum is large. The results show that Weighted in-degree performs better than In-degree and UserRank performs better than PageRank. It shows that the performance increase when we incorporate content similarity and response immediacy in weighted social network. The results also demonstrate that Weighted in-degree outperforms UserRank. Besides, Weighted in-degree is more efficient than UserRank. However, our proposed techniques have their limitation. They are sensitive to the size of forum. In other words, when the forum size is small, weight function cannot capture influence probability between pairs of users precisely, and therefore, leads to unstable and inaccurate results.

In the future, we shall consider incorporating and examining other forum features to further improve the performance of our technique. We shall investigate features that are particularly important for healthcare social media. For example, we shall utilize the ontology such as Unified Medical Language System (UMLS) and Consumer Health Vocabulary (CHV) to improve the content analysis component. We shall also utilize the user contributed tags related to symptoms, medications, and health conditions to enhance the link analysis.
REFERENCES


