Rafiei and Kardan *Human-centric Computing and Information Sciences* (2015) 5:10
DOI 10.1186/s13673-015-0030-5

Human-centric Computing and Information Sciences
a SpringerOpen Journal

RESEARCH Open Access

A novel method for expert finding in online communities based on concept map and PageRank

Majid Rafiei* and Ahmad A Kardan

* Correspondence: mrafiei@aut.ac.ir
Advanced E-Learning Laboratory, Department of Computer Engineering and IT, Amirkabir University of Technology, Tehran, Iran

Abstract

An online community is a virtual community where people can express their opinions and their knowledge freely. There are a great deal of information in online communities, however there is no way to determine its authenticity. Thus the knowledge which has been shared in online communities is not reliable. By determining expertise level of users and finding experts in online communities the accuracy of posted comments can be evaluated.

In this study, a hybrid method for expert finding in online communities is presented which is based on content analysis and social network analysis. The content analysis is based on concept map and the social network analysis is based on PageRank algorithm. To evaluate the proposed method java online community was selected and then correlation between our results and scores prepared by java online community was calculated. Based on obtained results Spearman correlation for 11 subcategories of java online community using this method is 0.904, which is highly an acceptable value.

Keywords: Expert finding; Online community; Concept map; Dijkstra's algorithm; Knowledge sharing; PageRank algorithm

Introduction

Online communities due to their unique features such as ease of access have become one of the main sources for resolving problems. Individuals from all over the world can freely share their questions or comments in online communities without any limitations regarding time/place constraints. Online communities are one of the important achievements of Web 2.0 technologies and have been welcomed by academic researchers and commercial organizations in recent years [1].

Knowledge sharing is one of the most important applications of online communities in virtual space of the Internet. So far, several studies concerning the reasons for knowledge sharing in online communities have been presented. For example in [2] and [3] number of factors that influence knowledge sharing in online communities have been identified. In most of these studies, theories in different domains of human science have been considered as a foundation for examining motivation of individuals to share knowledge in online communities. For example in [4] and [5], social cognitive theory; in [6], the combination of social cognitive theory and goal setting theory; in [7], the combination of social cognitive theory and trust and in [8] the combination of social

cognitive theory and social capital theory have been examined as a basis for the individual's motivation to share knowledge in online communities. In addition in [9] and [10], theory of justice, and in [11], theory of social capital have been considered to study the individual's motivation in knowledge sharing.

In online communities knowledge levels of users are unclear and therefore value of answers and comments are unknown. Therefore, this is one of the biggest challenges in online communities. By determining knowledge level of an individual user and finding experts in an online community, we can determine the answers which are more reliable.

In online communities, large volume of information related to questions posted by users is another important challenge that makes questions unseen by experts who have the ability to respond to them. Therefore, the response time for responding questions takes longer. By using expert finding methods and making recommender systems based on these methods, questions can be exposed to individuals who have adequate knowledge to respond them. In addition, it is possible to make simple questions unseen to experts; thus covering them to not waste their own times for answering simple questions. Such a recommender system has been implemented in [12].

In addition in online communities the volume of information related to posted answers are very large. By finding expert we can summarize this information. Obviously answers submitted by different users can be evaluated according to the expertise or knowledge level of sender. Therefore, only those answers are acceptable that their senders have adequate knowledge for answering them. Thus, if the knowledgeable individuals can be distinguished as experts, ones who are looking for answers are not confused with large amount of true and false answers.

As above-mentioned, in online communities obviously the methods of expert finding and determining expertise level of users are highly important since the valuable volume of information can be exploited. In general, there are two main approaches for finding experts. The first approach focuses on Social Network Analysis (SNA) and the second approach emphases on content analysis. We will describe these two approaches with details in section II.

Both of the above-mentioned approaches have some defects. For example in the approaches based on social network analysis, content of users' messages are not considered and users may send many irrelevant or empty messages. Subsequently this increases users' communications and may cause some mistakes in finding experts. Additionally, in the approaches based on content analysis, communications between individuals are not considered and there is no distinction among exchanged responses. However, expertise of an individual who responds to an expert may be more important than one who responds to a normal or beginner user. Thus, for higher accuracy the hybrid methods should be applied for expert finding.

In this study, a hybrid method for expert finding in online communities is presented; in our method content analysis is performed by analyzing concept map and social network analysis is based on PageRank algorithm. In content analysis for first time till now we have used from similarity between question and answer as a factor for evaluating user's expertise in online communities and in social network analysis we have designed a customized PageRank algorithm which works very good even alone. We have compared our hybrid method with content analysis and social network analysis as components of the proposed method, our method is better than the methods of content

analysis and social network analysis alone. In addition, we have compared our method with some basic methods and results indicate that our method works better.

In section II the related works which have been studied in related field is introduced. This is necessary for a better understanding of the presented method which is expressed in section III. The proposed method is presented in section IV, and is evaluated in section V. Finally in section VI, conclusion and future works is described.

Related studies

Before now, most researches in the field of expert finding had been done in organizations. However at present, more interests are shown for finding experts in virtual environment, especially in social networks and online communities [13]. Some studies described in this section are related to the organization and several others are related to the Internet.

Expert Finder Systems (EFS's) are considered as part of the CSCW systems (Computer Supported Cooperative Work). For example, they are designed to find people with special expertise or knowledge in online communities, who have ability to respond to a particular question. In addition, EFS's establish an important class of the recommender systems [12].

As mentioned in previous section, there are two main approaches for finding experts. The first approach focuses on Social Network Analysis (SNA) and the second approach emphases on content analysis. Considering the first approach so far network-based ranking methods and algorithms have been used to identify experts, for instance PageRank and HITS. In any network-based ranking algorithm individuals are considered as nodes and relationship between them are considered as links of a network. When information is exchanged between two nodes a link between them is shaped. For example, if person $\boldsymbol{A}$ responds to person $\boldsymbol{B}$ a link from $\boldsymbol{A}$ is drawn to $\boldsymbol{B}$. After creating all possible links between all individuals a network which is called the Expertise Network (EN) is established [14]. Right now network-based ranking algorithms are able to find important nodes and indicate the experts. For example in [14], they aimed to find different methods to identify and rank experts by shaping EN, and then the performance of these methods have been compared. In [15] and [16], SNA was used for finding experts. In [13], the aim was to find experts in Meta Filter online community using SNA approach. In [17] experts were found by means of SNA in Friendfeed online community. In [18], Thiago Baesso and his colleagues have analyzed some graph metrics and algorithms in order to finding experts in Java discussion group from an online community of Facebook.

The second approach for finding experts in online communities is focused on content analysis. In this approach, text mining techniques are utilized. For this purpose the content of messages sent by users are analyzed and based on information extracted from text messages, user's knowledge model or a probability model of the relationship between the user and the messages is generated. Knowledge model and probability model can be utilized to identify expert users. For example in [1], user's knowledge modeling has been used to identify experts. In [19-22], experts have been identified by using probabilistic models. In [23], we have used only content analysis for expert finding in online communities.

There are few studies that have used hybrid methods for finding expert. For example, in [24], a hybrid method has been used for finding experts in a social network of researchers. In [25], exchanged emails have been used to establish both above-mentioned approaches known as social network analysis and content analysis. In [26], Bozzon and his colleagues have introduced a method for finding expert in social networks based on text analysis and social network context. In addition in [27] and [28], by means of combining features of both approaches, experts have been identified.

It is mentionable that so far some of the works in the field of expert finding are about utilizing expert finding to support other applications. Designing recommender systems is one of the most important applications that utilize expert finding algorithms. For example, in [12], a recommender system for interface personalization of java online community is provided. In [28] a recommender system mechanism is provided for improving knowledge sharing in online forums. Another example of utilizing expert finding systems is to solve complex problems in organizations. In [29], such a system has been designed. In [30], Chen Yang and his colleagues have used expert finding method for expert recommendation in online scientific communities.

In our study a hybrid method is presented for expert finding in online communities. By means of this method experts can be detected with high accuracy. In our method content analysis is performed by analyzing concept map and social network analysis is based on PageRank algorithm. Details of the proposed method will be described in sections IV.

Basic concepts

For founding a better understanding of the proposed method, the basic concepts including: Java Online Community, Concept Map, Dijkstra's algorithm and PageRank algorithm will be presented in the following subsections.

Java online community

Java online community is part of Oracle corporation forum which is dedicated to the Java technology. According to obtained information in this study, in February 2013 Oracle forum had nearly one million users and almost two million and a half questions were raised and examined. These statistics clearly indicates that this online community is highly active.

In this community membership is free, and users after gaining membership can post their questions. Previous FAQ threads can be viewed without registration and they are visible to everyone. This forum has been segmented into 16 subsections; and each subsection corresponds to one of the Java technologies. This segmentation has led to exchange of highly specialized questions and answers.

Like most of online communities, java online community has its own scoring mechanism, in which the inquirer can use two types of labels for a respond submitted by others; including "Helpful" and "Correct". If submitted answer obtains "Helpful" label by the inquirer, respondent user receives 5 points, and if submitted answer obtains "Correct" label by the inquirer, respondent user receives 10 points. The points in each subsection of the online community are collected and so the rating of each user in the subsections is identified. Ten top experts in each subsection are ranked based on their

points. At the end, the total scores in all subsections are collected and the overall score is determined. Altogether, ten experts in online community according to these points are introduced, regardless of the section in which they act. In this study, we evaluate our proposed method by calculating correlation between scores obtained from our method and the scores related to ten experts in each subsection.

Concept map

Concept map is a graphical method for knowledge visualization. Moreover, this map is actually a graph which contains nodes to represent concepts and labeled links explaining the relationship among the concepts [31]. Concept map can be used as a network based resource for extracting semantic similarities which has a high precision [32]. Semantic similarity between two words can be obtained from the likeness of their meaning content. Concept mapping can also be used as a method to deduce students' conceptual model in a particular field of study [33].

In this work, Java technology concept map has been used as a knowledge base to extract the concepts exchanged in question-answer pair, and to calculate similarity between concepts in response and question as well.

Dijkstra's algorithm

Dijkstra's algorithm, conceived by computer scientist Edger Dijkstra in 1956 and published in 1959 [34], Dijkstra's algorithm is a graph search algorithm which solves the single-source shortest path problem for a graph with non-negative edge path costs, producing the shortest path tree.

For a given source node in the graph, the algorithm finds the path with the lowest cost (i.e. the shortest path) between that node and every other node. It can also be used to find costs of the shortest paths from a source node to a destination node by stopping the algorithm once the shortest path to the destination node has been determined.

In this study, we use Dijkstra's algorithm for finding the shortest path between semantically related concepts in Java concept map. In fact, the concept map is a weighted graph where all edges weight equally.

PageRank algorithm

PageRank is a link analysis algorithm that has been developed at Stanford University [35]. The PageRank algorithm is used for ranking pages in Web, and it is described by equation (1).

$$\boldsymbol{PR}(\boldsymbol{A}) = (1-\boldsymbol{d}) + \boldsymbol{d}\left(\frac{\boldsymbol{PR}(\boldsymbol{T}_1)}{\boldsymbol{C}(\boldsymbol{T}_1)} + \frac{\boldsymbol{PR}(\boldsymbol{T}_2)}{\boldsymbol{C}(\boldsymbol{T}_2)} + \ldots + \frac{\boldsymbol{PR}(\boldsymbol{T}_n)}{\boldsymbol{C}(\boldsymbol{T}_n)}\right) \quad (1)$$

Where $\boldsymbol{PR(A)}$ is the PageRank of page $\boldsymbol{A}$, $\boldsymbol{PR(T_i)}$ is the PageRank of pages $\boldsymbol{T_i}$ which link to page $\boldsymbol{A}$, $\boldsymbol{C(T_i)}$ is the number of outbound links on page $\boldsymbol{T_i}$ and $\boldsymbol{d}$ is a damping factor which can be set between $\mathbf{0}$ and $\mathbf{1}$.

Based on equation (1), PageRank does not rank web sites as a whole, but ranking is done for each page individually. Further, the PageRank of page $\boldsymbol{A}$ is recursively defined by the PageRanks of those pages which are linked to page $\boldsymbol{A}$. This algorithm continues computation until reaching to a desired convergence and ranking scores do not change.

Within the PageRank algorithm, the PageRank of a page T is always weighted by the number of outbound links $C(T)$ on page T.

Finally, the sum of the weighted PageRanks of all pages T_i is multiplied with a damping factor d which can be set between **0** and **1**.

Proposed method

In this section, framework of the proposed method is described. Our proposed framework has four main steps.

1. Information Extraction
2. Content Analysis
3. Social Network Analysis
4. User Ranking and Experts Finding

Figure 1, shows the proposed framework, in the following we briefly explain the proposed method by an example and then details for each step will be described.

Suppose "X" is a user in java online community, at first step user's profile information is extracted by a crawler, if number of user's posts was higher than 100, user is considered for ranking and verified as expert. Then, information of user's posts including questions and answers of user are extracted by another crawler. Crawlers have been implemented by Perl programming language.

At second step, concepts of user's posts are extracted with regard to concept map. Then, distance between concepts is calculated based on concept map, for calculating distance we have used Dijkstra's algorithm, for implementing this algorithm we have utilized MATLAB programming and used concept map as a graph. Weights of concepts in the user's response are calculated based on distance of

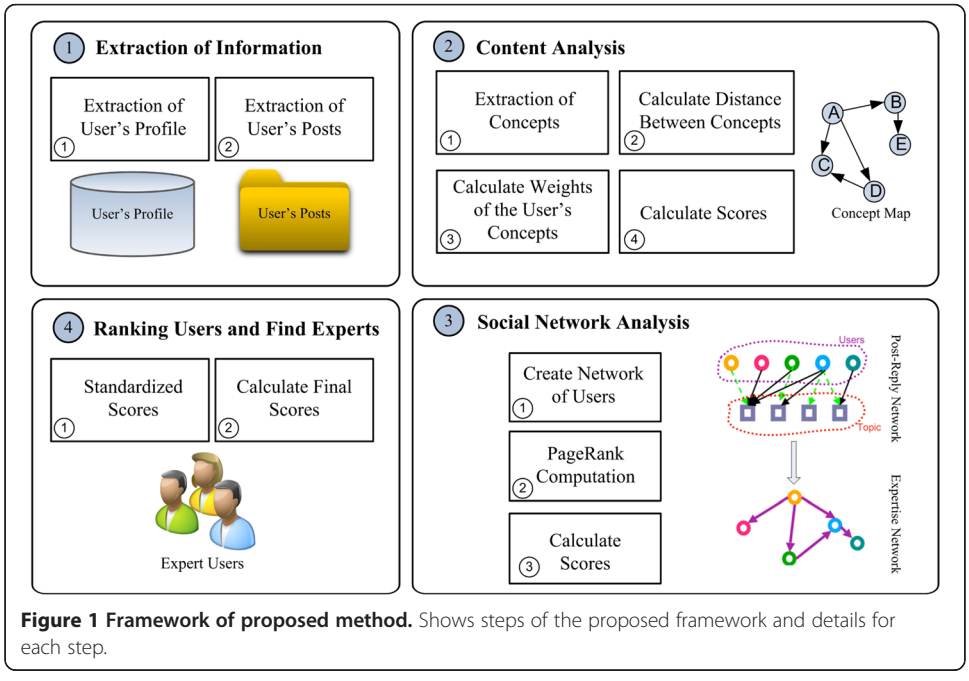

Figure 1 Framework of proposed method. Shows steps of the proposed framework and details for each step.

these concepts from question's concepts. Finally user's score is calculated based on weighted concepts.

At third step, network of users is created based on questions and answers of each user. Then, customized PageRank computations are done on network of users, for customized network computations we have used MATLAB programming. At the end of this step user's score in network will be available.

At final step, scores which reach from content analysis and social network analysis are standardized and then final score is calculated.

Information extraction

In this step, essential information relevant to the user's profile and the user's posts are extracted. Details of information extraction from each of the mentioned sources are addressed as follows.

Information extraction from User's profile

At first, the structure of web in Oracle online community was obtained. This is necessary to find addresses related to users' information, threads, user's messages and other desirable information. For this purpose, the URL addresses were examined and logic beside URLs was discovered. Then information related to the users was extracted. The most important information related to each user profile includes the following items:

- User ID that is a unique identifier
- User handle that is a unique name
- User level associated with the number of points
- User's posts
- Number of user's posts
- Number of user questions

Since analysis of the millions of messages gathered in the website is impossible and most of the users are not currently active, our analysis was focused on the members who were most active. For this purpose, first, the information file was converted to a format which could be entered as input to the ETL (Extract-Transform-Load). By using the ETL, information of users were entered into MSSQL Server to easily run query on it and extract desired users.

In this study, we have considered users as active users if they have more than 100 posts. The total number of active users in the Oracle forum till February 2013 was equal to 7749.

Extraction of user's posts

After extracting user information, threads of question/answer can be extracted. It was necessary to extract 2.5 million threads of question/answer because Java forum does not allow access to a specific user's posts. Therefore, first, all threads of question/answer must be evacuated and then those posts related to specific user can be extracted from this threads. In posts extraction, source code and other quotes should be removed.

Finally a data structure is created for each user that contains the following information:

- Thread ID that user submits a message to it
- Subject of thread
- If the post is kind of response, then who has received a response?
- The content of messages that have been sent by each user

Content analysis

In this step concepts of the users' posts are extracted and then dedicated score to each user is calculated based on concept map. Details of this step are described next.

Extraction of concepts

At first, concepts in exchanged questions and answers of each user are extracted. Since these concepts should be extracted and compared with java concept map, it is necessary extract all nodes of Java concept map in advance. For each node, other keywords with same meaning are considered as well.

After creating a data structure for the Java concept map, concepts of exchanged posts are extracted according to the concept map. At the end of this stage, each user has a data structure which includes concepts of each question and keywords relevant to response posted to that question.

Calculating distance between concepts in the concept map

To calculate distance between the concepts deduced from responses and the concepts deduced from questions should be extracted. In this regard, the shortest distance from one concept to each concept in the concept map is extracted. For this purpose it is necessary to draw a graph of relations between concepts. The number of concepts shaping the concept map was 211. After drawing a graph, by using Dijkstra's algorithm, the shortest path between any two nodes in an undirected graph was calculated. The output of this stage is a two-dimensional matrix that holds distance between concepts.

Calculating weights of the concepts in users' responses

At this stage, the average distance between each concept in response and all concepts in the question is calculated by equation (2).

$$\boldsymbol{AvgDist}(\boldsymbol{R}) = \frac{\sum_{\boldsymbol{Q}\in\boldsymbol{Questions}}(\boldsymbol{Dist}(\boldsymbol{R},\boldsymbol{Q}))}{N} \qquad R, Q \in \text{Concepts of Concept Map} \tag{2}$$

Where $\boldsymbol{R}$ is concept of the response, $\boldsymbol{Q}$ is concept of the question, $\boldsymbol{Questions}$ are all concepts in the question, $\boldsymbol{Dist}\ (\boldsymbol{R},\boldsymbol{Q})$ is distance between concept $\boldsymbol{R}$ in the response, and concept $\boldsymbol{Q}$ in the question and $\boldsymbol{N}$ is the number of concepts in the question.

In the numerator of equation (2), sum of distances related to the concept R in the response from all of the concepts in the question has been calculated.

$\boldsymbol{AvgDist}(\boldsymbol{R})$, has been calculated for all concepts of the response. Finally each concept in the response has been replaced with average distance of the concept from all concepts of the question.

Calculating ranking scores

At this stage, the final scores for users has been calculated based on equation (3).

$$\boldsymbol{Score}(\boldsymbol{I}) = \sum_{\mathbf{M}\in\mathbf{Messages}} \sum_{\mathbf{R}\in\mathbf{Responses}} \left(\boldsymbol{\alpha}.\mathbf{Rep}(\mathbf{R}) + \frac{\boldsymbol{\beta}}{\boldsymbol{AvgDist}(\boldsymbol{R})}\right)$$

$$\boldsymbol{R} \in \text{Concept of Concept Map} \tag{3}$$

Where ***Score(I)*** is score of user ***I***, ***Messages*** are messages of user ***I***, ***Responses*** are concepts in the response of the message ***M***, ***Rep (R)*** is the number of iterations of concept ***R*** in the response of the message ***M*** and ***Weight (R)*** is weight of concept ***R*** in the response of the message ***M***.

Based on equation (3), it is obvious that score of each user has been calculated based on two measures. One is the number of iterations of concept which is used by user in response. Another measure is similarity between concept in response and concept in question. Since the average distance of concept in response from the concepts in question is inversely proportional to similarity, thus AvgDist are used inversely in calculating score for users.

α and **β** are coefficients with values between **0** and **1.** **α** indicates the impact of the number of concepts which are in the user response, and **β** indicates the impact of distance between concepts in user response and concepts in question. In this study, the optimum values for these coefficients are calculated. To achieve the optimum values state space is searched by changing 0.01 intervals for **α** and **β**. The optimum values obtained for these coefficients are equally 0.5. By using these coefficients, the best correlation between the scores obtained from the proposed method and the scores provided by Java's online community, is calculated.

Social network analysis

In this step, we describe how the network is created among users and then how the PageRank algorithm operates on the network of users and scores are given to them.

Create network of users

Online communities usually have a discussion thread structure. A user posts a topic or question, and then some other users post an answer to a question or participate in a discussion. By using these posting/replying threads in a community, we can create a post-reply network of users, as shown in Figure 2.

When a user replying to a question or a topic, usually this indicates that the respondent user has a higher level of expertise on the subject than the person who asks the question. Connecting questioners to respondents by directional arrows from questioners to respondents makes a network which is called Community Expertise Network (CEN). The CEN created from Figure 2, is shown in Figure 3.

Social network analysis for ranking users in online communities is based on CEN which is created among users. Inbound link to a node in CEN indicates that the user linked to this node answers to the user who is on other side of the link. Whatever the number of inbound links to a node is more, indicates that linked user to that node has higher expertise.

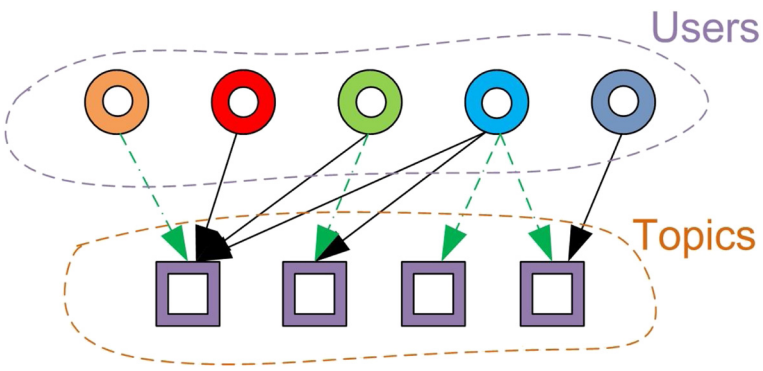

Figure 2 Post-reply network. In this figure dashed arrows indicate topics or questions and responses are shown in solid black arrows.

PageRank computation

In this stage, we describe the difference between the community expertise network and the network which is created in the web. We also describe how the PageRank algorithm is used for giving scores to users in a community expertise network.

In a community expertise network (CEN) it may be possible to have more than one link between two nodes, however in a network which is created in the web, only one link can be existed in each direction. This is the main difference between a CEN and a network which is created in the web. Based on this difference, PageRank algorithm should be modified and then can be used for users ranking in a CEN. In the modified PageRank algorithm, weights in transfer matrix are determined on the basis of the number of links between two nodes. This issue in more detail is explained next.

Suppose a CEN has been formed between three users A, B and C as shown in Figure 4. Transition probability matrix for Figure 4, is shown in Figure 5. Such as

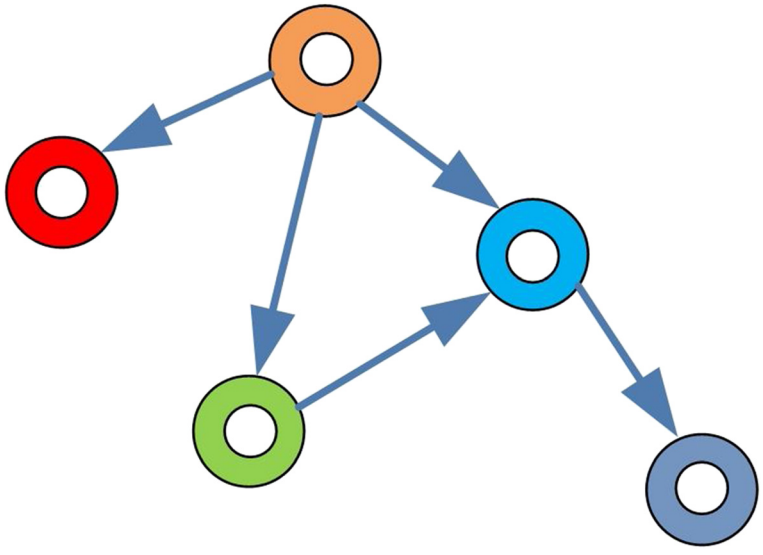

Figure 3 Expertise network. A Community Expertise Network (CEN) created from the Post-Reply Network.

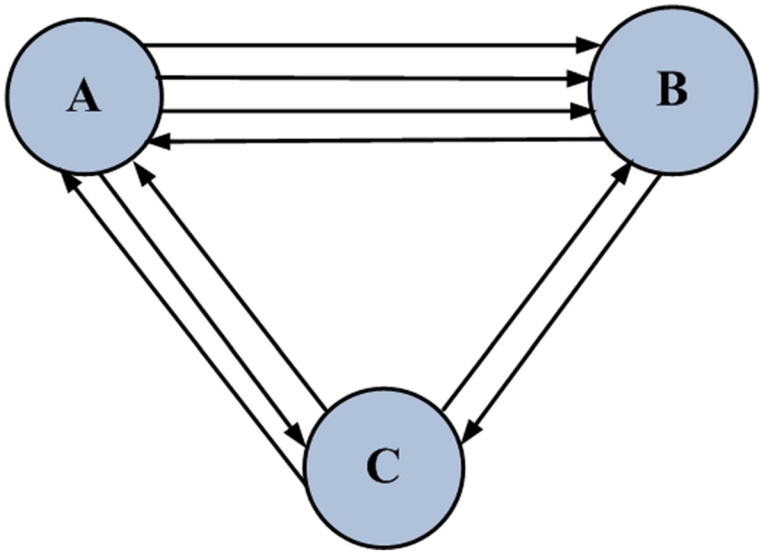

Figure 4 Sample of expertise network. A CEN has been formed between three users as nodes, questions and answers between users have been showed as arrows.

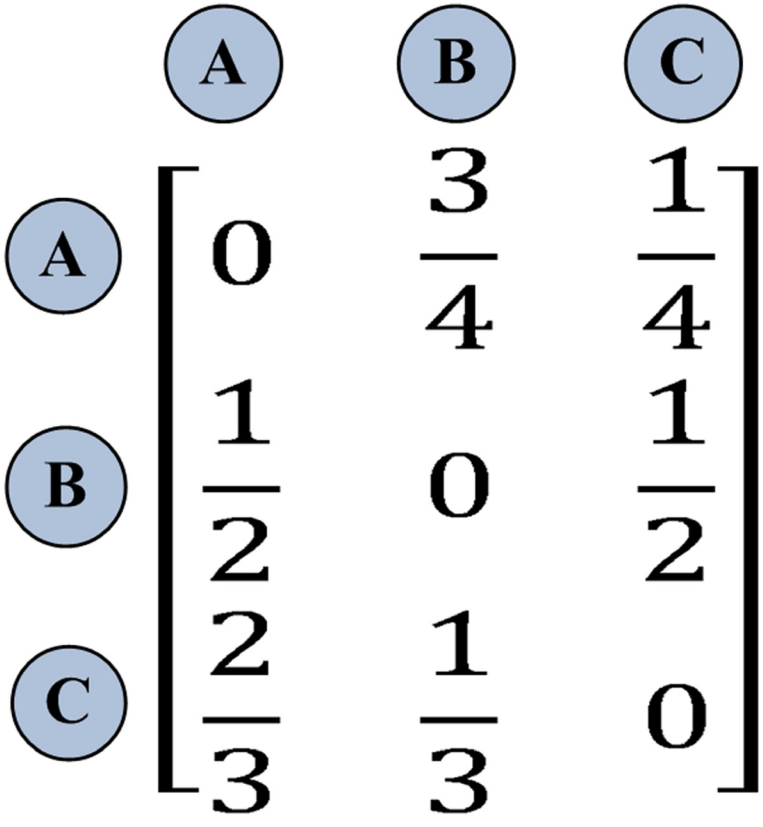

Figure 5 Weights for sample CEN. This matrix shows weights of arrows between users as nodes in the Sample CEN.

PageRank Algorithms row of the matrix replaced with $1/n$ if all values are zero, n is the number of nodes in the network.

Weight for row related to user x and column related to user y are calculated as equation (4). In equation (4), n is the total number of node or users and R_{xy} is the outbound links from x to y and R_{xi} is the outbound links from x to i.

$$Weight(xy) = \frac{R_{xy}}{\sum_{i=1}^{n} R_{xi}} \tag{4}$$

Number of nodes or users of communications in our data set of java online community, are 14,274 nodes.

Calculate scores

After creation of transition probability matrix based on CEN we multiple damping factor by all the matrix elements. Damping factor usually set to 0.8, we also have used this value.

In the PageRank algorithm $1-d$ is the probable of teleport operation. In teleport, surfer can jump into each node in the network. The destination of teleport operation is selected randomly. If a node doesn't have any output link, then the teleport action will be done with the probability of $1/n$ in which n is the number of nodes/users in the community expertise network, otherwise it will be done by probability of $1-d$. Finally the amount of $(1-d)/n$ will be added to all elements of transition probability matrix. After running PageRank algorithm scores will be determined.

Ranking users and find experts

In this step, scores of content analysis and social network analysis are combined and the final scores are obtained. Details of this step are described in the following.

Standardized scores

For combining link analysis and content analysis approaches, scores that are given to users must be standardized. Scores of users in network analysis based on PageRank algorithm are between 0 and 1. However scores in content analysis not limited to a certain range. Scores of content analysis will also being the range of 0 and 1, using equation (5). In this equation, Max and Min, respectively, are the highest and lowest scores.

$$Stand(value) = \frac{value - Min}{Max - Min} \tag{5}$$

Calculate final scores

After scores obtained with both approaches, were the same range, scores were combined to obtain a final score for each user, using equation (6).

$$Score(p) = T.Score(T) + N.Score(N) \tag{6}$$

In equation (6), $Score(T)$ is score of content analysis and $Score(N)$ is score of link analysis. T and N, respectively are considered as weight for content analysis and link

analysis. At final, users are ranked with ***Score(p)*** values and top users are determined as expert users.

Evaluation and results

To evaluate the proposed method, all the subsection of java online community is used. First, number of responses for each subsection was calculated and subsections which the number of responses for them is less than 3000 have been excluded. Finally, 11 subsections have remained.

Spearman correlation between our results and scores prepared by java online community was calculated separately for the 11 subsections and the entire java online community, the overall correlation was calculated by taking the average of these correlations. The overall correlation was calculated with different values for weight of content analysis and social network analysis in equation (6). Table 1, shows these correlations.

The results show that the hybrid method is better than the methods of content analysis or social network analysis alone. As you see in first row of Table 2, when only social network analysis is used average spearman correlation is 0.877 and when only content analysis is used as you see in last row of Table 2, average spearman correlation is 0.829. However, when weight of social network analysis is 0.8 and content analysis is 0.2, spearman correlation between our results and scores prepared by java online community reaches maximum value.

Table 2 shows detailed information for each subsection of java online community and entire java online community separately. Spearman correlation between our results and scores prepared by java online community has been presented for each subsection; these correlations are for best weights of social network analysis and content analysis.

In Table 2 the abbreviations are defined as:

- ***NQ***: Number of question
- ***NR***: Number of response
- ***NU***: Number of active users

Table 1 Spearman correlation for different weights

Weights of content analysis (t) and social network analysis (n)	Average spearman correlation for 11 subsections and entire java online community
T 0.0 - N 1.0	0.877142549
T 0.1 - N 0.9	0.901276748
T 0.2 - N 0.8	**0.904018451**
T 0.3 - N 0.7	0.903982375
T 0.4 - N 0.6	0.901998248
T 0.5 - N 0.5	0.901998248
T 0.6 - N 0.4	0.887171408
T 0.7 - N 0.3	0.890147598
T 0.8 - N 0.2	0.878188724
T 0.9 - N 0.1	0.864101422
T 1.0 - N 0.0	0.82930109

Bold numbers indicate that when "T 0.2-N 0.8" Spearman correlation between our results and scores prepared by java online community reaches the maximum value.

Table 2 Information about each subsection

Category	NQ	NR	NU	A(Q)	A(R)	P(80Q)	P(80R)	*SpCo*
*ALL (*Entire Java Online Community*)*	6465	345206	614	10.52	562.22	9.60	3.74	***0.96***
Database Connectivity	367	23456	254	1.44	92.34	33.85	3.14	***0.89***
Development Tools	308	4869	152	2.02	32.03	18.42	6.57	***1***
Java APIs	337	31096	105	3.20	296.15	12.38	3.80	***0.91***
Java Card	415	4145	28	14.82	148.03	10.71	7.14	***0.80***
Java Desktop	1657	54839	150	11.04	365.59	4.66	5.33	***0.88***
Java Developer Tool APIs	9	2719	37	0.24	73.48	13.51	5.40	-
Java Embedded	8	62	13	0.61	4.76	38.46	30.76	-
Java Enterprise & Remote Computing	447	40642	134	3.33	303.29	25.37	3.73	***1***
Java Essentials	2266	157398	264	8.58	596.20	12.5	5.68	***0.85***
Java HotSpot Virtual Machine	40	8299	51	0.78	162.72	37.25	1.96	***0.87***
Java Mobile	80	2000	30	2.66	66.66	13.33	0	-
Java Real-Time	0	12	4	0	3	100	25	-
Java Security	86	4868	50	1.72	97.36	26	2	***0.85***
Java TV	0	5	3	0	1.66	100	66.66	-
JavaFX	484	9689	53	9.13	182.81	18.86	22.64	***0.85***
Other Topics	58	5905	47	1.23	125.63	19.14	12.76	***0.95***

Bold numbers indicate the value of Spearman correlation in different categories.

- ***A(R)***: Average number of responses per user
- ***A(Q)***: Average number of questions per user
- ***P(80Q)***: Percentage of users who submit %80 of questions
- ***P(80R)***: Percentage of users who submit %80 of responses
- ***SpCo***: Spearman correlation between our results and scores prepared by java online community

For '*Java Developer Tool APIs*', '*Java Embedded*', '*Java Mobile*', '*Java Real-Time*', '*Java TV*', because the number of responses is less than 3000, the correlation is not valid for them and is not calculated.

In Table 2, the value of ***NR*** and ***A(R)*** are more important for our study, because the proposed method is based on user's responses and if these numbers are much higher, accuracy will be higher as well.

P(80Q) and ***P(80R)*** are also calculated to determine in which field users are more trended to work. As the values for the various subsections indicate that few users ask % 80 of the questions and submitted %80 of the responses.

Average percentage of users that submitted %80 of responses under 11 subsections is 6.79; hence, less than %10 of users can respond to %80 of questions and more than % 90 of users can respond to only %20 of questions. Average percentage of users that submitted %80 of questions under 11 subsections is 19.92. Therefore, less than %20 of users, submitted %80 of questions and more than %80 of users, submitted only %20 of questions. The Pareto's chart of these statistics shown in Figure 6 and indicates that most online users in this forum are the questioners and responders.

Our results are similar to ones obtained in [14] which has also been done on the java online community as well, indicating that few users respond most questions and most users answer a few questions.

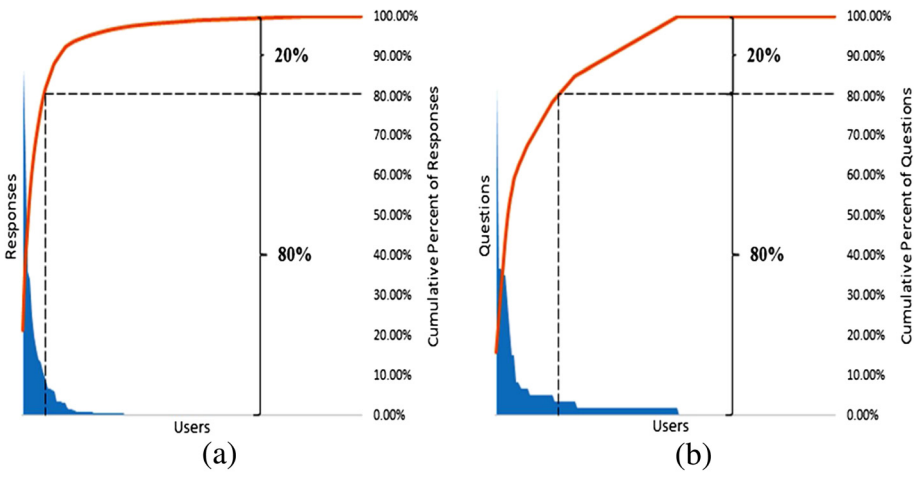

Figure 6 Pareto's chart. Part **(a)** shows users who submit %80 of responses and part **(b)** shows users who submit %80 of questions.

In above, we have compared our hybrid method with content analysis and social network analysis as components of the proposed method and in the following we will compare our hybrid method with other methods.

There are some basic methods which are used for comparison; these methods have been described in [14]. We briefly introduce these basic methods and comparison our hybrid method with these methods in the following.

AnswerNum: in this method experts have been identified with counting of answers of one user.

Indegree: in this method experts have been identified with counting of users that one user has sent answers for their questions.

Z-score: if one user makes $n = a + q$ posts, q is the number of questions and a is the number of answers, then Z-score has been calculated with equation (7).

$$\mathbf{z} = \frac{\mathbf{a} - \mathbf{q}}{\sqrt{\mathbf{a} + \mathbf{q}}} \tag{7}$$

If Z-score has been considered for the number of questions one user asked and answered, the method called **Z-number**, and if Z-score has been considered for the number of users one user replied to and received replies from, the method called **Z-degree.**

Figure 7 indicates comparison between our hybrid method and basic methods in different categories of java online community and Figure 8 indicates comparison between our hybrid method and basic methods in average of all categories of java online community.

As you see in Figure 7 our hybrid method is better than other methods in most of categories of java online community, in addition as you see in Figure 8 our method is better than other methods in average of all categories of java online community.

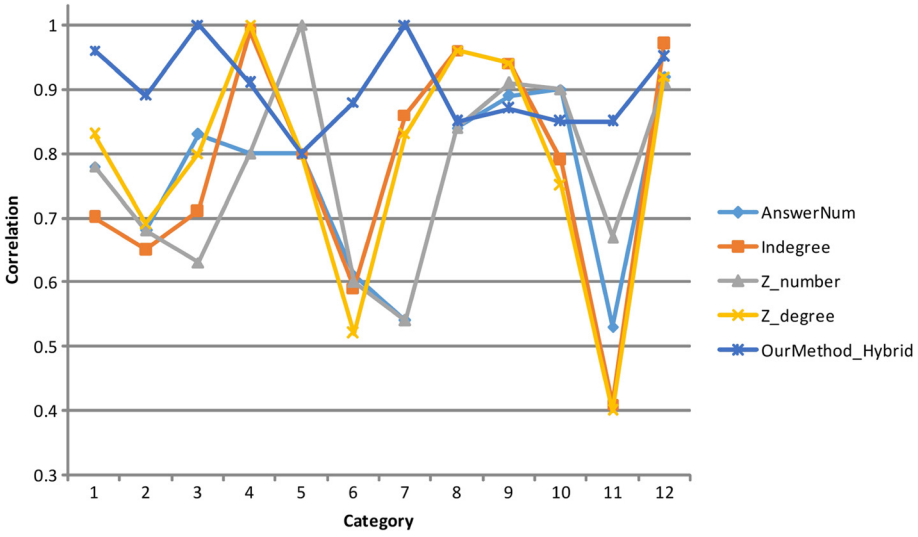

Figure 7 Comparison with other methods in different categories. This figure shows the results of comparison between our hybrid method and basic methods in different categories of java online community.

Conclusions and future works

In this study, the role and importance of online communities was discussed. In addition to understanding knowledge sharing and its position in online communities, important concerns and challenges in online communities were expressed, we focused on one of the solutions to these challenges that was "Expert Finding", related works done in the field of "Expert Finding" were expressed, and a novel hybrid method based on concept maps and PageRank for expert finding in online communities was presented. The proposed method is implemented and evaluated on Java online communities, and the

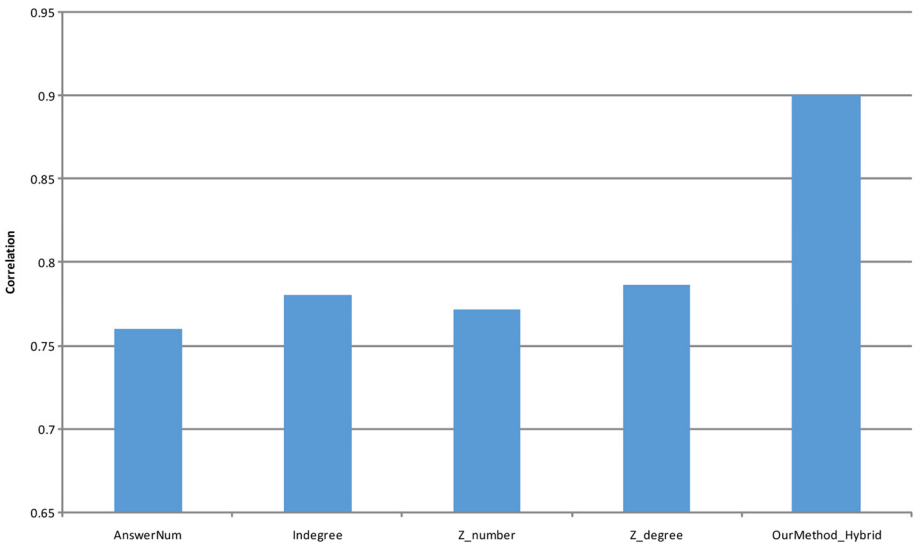

Figure 8 Comparison with other methods in average of all categories. This figure shows the results of comparison between our hybrid method and basic methods in average of all categories of java online community.

results revealed that the correlation between our results and scores prepared by java online community exceeds 0.9 which is highly a reasonable value.

This method is applicable to all online communities and only a concept map in the field of online community is needed to accomplish that. As aforementioned, concept map is a network based resource for extracting semantic similarities. Semantic network based measures for extracting semantic similarity, have a high precision and thus this method has high precision and reasonable results.

By finding expert with this method, accuracy of posted comments can be evaluated, shared knowledge in online communities will be reliable and response time will be reduced. In addition, large volume of information can be summarized. This method can be used for design expert recommender systems with high precision, as well.

In the proposed method, similarity between concepts extracted by a network based measure. In the future, other approaches can be used for extracting semantic similarity, such as corpus based or dictionary based approach. Also we can add other measures for enhance accuracy of proposed method. Moreover, combine content analysis and social network analysis approaches can be done differently.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The described method was developed through discussions collectively by both authors. MR carried out the implementation of the proposed method, analyzed the results and drafted the manuscript, which was critically revised by AK. Both authors read and approved the final manuscript.

Authors' information
1. Ahmad Kardan is the Assistant Professor of Computer Engineering and IT Department of Amirkabir University of Technology, Tehran, Iran.
2. Majid Rafiei finished his Master of Engineering degree in Information Technology from Amirkabir University of Technology in 2013, Tehran, Iran.

Received: 9 November 2014 Accepted: 22 March 2015
Published online: 10 April 2015

References
1. Kardan A, Garakani M, Bahrani B (2010) A method to automatically construct a user knowledge model in a forum environment, Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '10)., pp 717–718
2. Chen I (2007) The factors influencing members' continuance intentions in professional virtual communities - A longitudinal study. J Inf Sci 33(4):451–467
3. Lin H (2007) Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions. J Inf Sci 33(2):135–149
4. Chen C, Hung S (2010) To give or to receive? Factors influencing members' knowledge sharing and community promotion in professional virtual communities. Information and Management 47(4):226–236
5. Lin M, Hung S, Chen C (2009) Fostering the determinants of knowledge sharing in professional virtual communities. Comput Hum Behav 25(4):929–939
6. Kim J, Song J, Jones D (2011) The cognitive selection framework for knowledge acquisition strategies in virtual communities. Int J Inf Manag 31(2):111–120
7. Hsu M, Ju T, Yen C, Chang C (2007) Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. International J of Human-Computer Studies 65(2):153–169
8. Chiu C, Hsu M, Wang E (2006) Understanding knowledge sharing in virtual communities: an integration of social capital and social cognitive theories. Decis Support Syst 42(3):1872–1888
9. Fang Y, Chiu C (2010) In justice we trust: Exploring knowledge-sharing continuance intentions in virtual communities of practice. Comput Hum Behav 26(2):235–246
10. Chiu C, Wang E, Shih F, Fan Y (2011) Understanding knowledge sharing in virtual communities: An integration of expectancy disconfirmation and justice theories. Online Inf Rev 35(1):134–153
11. Chang H, Chuang S (2011) Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. Information and Management 48(1):9–18
12. Zhang U, Ackerman M, Adamic L, Nam K (2007) QuME: a mechanism to support expertise finding in online help-seeking communities, Proceedings of the 20th annual ACM symposium on User interface software and technology., pp 111–114
13. Kardan A, Omidvar A, Behzadi M (2012) Context based Expert Finding in Online Communities using Social Network Analysis. International J of Computer Science Research and Application 2(1):79–88

14. Zhang J, Ackerman M, Adamic L (2007) Expertise networks in online communities: structure and algorithms, Proceedings of the 16th international conference on World Wide Web (WWW '07)., pp 221–230
15. Gunilla W, Mariam G (2004) Explaining knowledge sharing in organizations through the dimensions of social capital. J Inf Sci 30(5):448–458
16. Gunilla W, Stefan E, Mariam G, Reija P, Pia S (2008) Information behaviour meets social capital: a conceptual model. J Inf Sci 34(3):346–355
17. Kardan A, Omidvar A, Farahmandnia F (2011) Expert Finding on Social Network with Link Analysis Approach, Electrical Engineering (ICEE), 19th Iranian Conference., pp 1–6
18. T. Baesso, S. Wolfgand, M. Siqueir, and L. C. Vasconcelos de Andrade (2014) Finding Experts on Facebook Communities: Who Knows More?, 5(2):7-19.
19. Liu X, Croft W, Koll M (2005) Finding experts in community-based question-answering services, Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM '05)., pp 315–316
20. Serdyukov P, Rode H, Hiemstra D (2008) Exploiting sequential dependencies for expert finding, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08)., pp 795–796
21. Balog K, Azzopardi L, Rijke M (2006) Formal models for expert finding in enterprise corpora, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)., pp 43–50
22. Balog K, Rijke M (2007) Determining expert profiles (with an application to expert finding), Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI'07)., pp 2657–2662
23. Kardan A, Rafiei M (2013) A novel method based on concept map for expert finding in online communities. Int J Nat Eng Sci 7(2):82–85
24. Zhang J, Tang J, Li J (2007) Expert Finding in a Social Network, Advances in Databases: Concepts, Systems and Applications. Springer, Berlin Heidelberg, vol. 4443, pp. 1066-1069
25. Campbell C, Maglio P, Cozzi A, Dom B (2003) Expertise identification using email communications, Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03)., pp 528–531
26. Bozzon A, Brambilla M, Ceri S, Silvestri M, Vesci G (2013) Choosing the right crowd: expert finding in social networks. In: *Proceedings of the 16th International Conference on Extending Database Technology* (EDBT '13). ACM, New York, NY, USA, pp 637–648
27. Serdyukov P, Rode H, Hiemstra D (2008) Modeling multi-step relevance propagation for expert finding, Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)., pp 1133–1142
28. Li Y, Liao T, Lai C (2012) A social recommender mechanism for improving knowledge sharing in online forums. Inf Process Manag 48(5):978–994
29. Tung Y, Tseng S, Weng J, Lee T, Liao A, Tsai W (2010) A rule-based CBR approach for expert finding and problem diagnosis. Expert Systems with Applications 37(3):2427–2438
30. Yang C, Ma J, Silva T, Liu X, and Hua Z (2014) A Multilevel Information Mining Approach for Expert Recommendation in Online Scientific Communities, The Computer Journal, first published online May 9, 2014 doi:10.1093/comjnl/bxu033
31. Novak JD and Canas AJ (2008) The Theory Underlying Concept Maps and How to Construct and Use Them, Technical Report IHMC CmapTools
32. Panchenko A (2013) A Semantic Similarity Measure Based on Lexico-Syntactic Patterns, doctoral dissertation, Université catholique de Louvain & Bauman Moscow State Technical University
33. Park U, Calvo RA (2008) Automatic Concept Map Scoring Framework Using the Semantic Web Technologies, Proceedings of the 2008 Eighth IEEE International Conference on Advanced Learning Technologies (ICALT '08)., pp 238–240
34. Dijkstra EW (1959) A note on two problems in connection with graphs. Numer Math 1:269–271
35. Page L, Brin S, Motwani R, and Winograd T (1998) The PageRank citation ranking: Bringing order to the web, Stanford Digital Library Technologies