Context-Based Expert Finding in Online Communities Using Ant Colony Algorithm

Mojtaba Sharifian
Department of electrical, computer and it engineering, Qazvin BranchIslamic Azad University, Qazvin, Iran
sharifian68@gmail.com

Neda Abdolvand*
Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran
n.abdolvand@alzahra.ac.ir

Saeedeh Rajaee Harandi
Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran
Sa.Rajaeeharandi@gmail.com

Abstract

Online communities are the most popular interactive environments on the Internet, which provide users with a platform to share their knowledge and expertise. The most important use of online communities in cyberspace is sharing knowledge. These communities are a great place to ask questions and find answers. The important challenges of these communities are the large volume of information and the lack of a method to determine their validity as well as expert finding which attracted a lot of attention in both industry and academia in. Therefore, identifying persons with relevant knowledge on a given topic and ranking them according to their expertise score can help to calculate the accuracy of the comments submitted on the internet. In this research, a model for finding experts and determining their domain expertise level by the aid of statistical calculations and the ant colony algorithm in the MetaFilter online community was presented. The WordNet Dictionary was used to determine the relevance of the user's questions with the intended domain. The proposed algorithm determines the level of people's expertise in the intended field by using the pheromone section of the Ant colony algorithm, which is based on the similarity of the questions sent by the users and the shared knowledge of the users from their interactions in the online community.

Keywords: Online Communities; Experts Finding; Ant Colony Algorithm; Word Net.

1- Introduction

Online communities are among the most important achievements of Web 2.0 technologies that have been noticed by researchers and business organizations due to their large volume of valuable raw data [1]. The most important use of online communities in cyberspace is sharing knowledge. These communities are a great place to ask questions and find answers [2] the unique features of these communities such as ease of access and lack of time and space constraints for accessing them make them as one of the most important sources of problem solving [3]. Since, there is no distinction between the levels of the users' knowledge in these communities; there is uncertainty about the value of the answers and comments submitted in these communities. In fact, the questioner does not know how much he/she can trust the answers sent by other users [4]. Considering the growth of information generated by users in question-and-answer forums (QA forums) and the need for speed and precision in finding the true answer, finding expert people in is necessary. Therefore, the important issue is finding the level of expertise of people in different fields as well as the level of users' trust in experts to use their answers in various issues '[2]. Given that only a small fraction of users are responsible for answering a significant number of questions, it is difficult to find potential experts [4]. Therefore, by using expert advisory methods, especially the recommender systems questions can be provided to knowledgeable people [5].

Expert finding is a challenging problem that has been considered in both industry and academia in the past few years. Identifying people with relevant knowledge about the issue and ranking them according to their expertise is a challenging task [6-7].

In general, there are two main approaches to expert finding: the content analysis of user messages and analysis of social networks between people. Each has some flaws. For example, in the social network analysis, the content of messages sent by users is not considered. In the content analysis approach, communication between people is not

considered and there is no difference between the answers given to people with different levels of expertise, while the knowledge of a person responding to an expert user is different from that of a user responding to a beginner [8]. Given that there are many queries in QA forums, and there is no way to determine the accuracy of the answers, users cannot trust the answers [4]. Therefore, the most important issue is how to identify and evaluate the experts as well as their level of expertise in professional communities with high precision and trust.. Hence, this study aims to rank the experts in different fields based on their answers as well as to determine their performance in QA forums. Therefore, for the first time in this field, using the Ant colony algorithm and considering the quality of communications and users' trust and reputation, a new model is proposed, which is a good solution to solve problems and challenges of online communities with high accuracy

The research literature is reviewed first; then, the research method and the proposed algorithm are presented. Finally, the conclusion and recommendations for future studies are expressed.

2- Literature Review

2-1- Expert Finding Systems

Expert finding systems are part of Computer Supported Cooperative Work (CSCW) systems, and one of the most important classes of recommender systems [9]. Given the importance of expert finding in both industry and academia, many approaches came into existence [10]. In general, two main approaches are used for expert finding, including content analysis and social network analysis. The former focuses on analyzing the content of user messages to find experts in online communities. In this approach, text-mining techniques and data retrieval algorithms are used to analyze messages sent by users. Then, based on the information extracted from the message, a user knowledge model is created using user's knowledge modeling techniques or a probabilistic model of the communication between the user and the messages is created. Expert people can be identified by using the knowledge model or probabilistic model. For example, Liu et al. [11] have identified experts using the data retrieval algorithms.

In latter, the features of the social network structure are used to estimate the people expertise. In this method, based on the relationships between people in a network, a graph is made [12], and the diffusion method [13] and the similarity of users [5] are used to weigh the edges and people, and sometimes the PageRank algorithm is used to score nodes and edges [2-13,14].

The study of Yuan et al. [15] the recent solutions for expert finding were classified into four different categories

of matrix factorization based models (MF-based models), gradient boosting tree based models (GBT-based models), deep learning based models (DL-based models) and ranking based models (R-based models). The results of the study indicated that MF-based models outperform other categories of models in the crowd-sourcing situation. However, their results indicated that different algorithms can perform well on different data sets and different problems.

The study of Dadkhaha et al. [16] used business intelligence tools to find potential experts as well as the analytic hierarchy process for assigning weights to both selection criteria and potential experts.

The study of Zhang [17] has tested a set of network-based ranking algorithms including Indegree, AnswerNum, Z-Score, PageRank, and HITS. These simulations have been used to find a few simple simulation rules governing dynamic network queries and answers. Kardan et al. [2] have also proposed a SNPageRank method for the accurate and comprehensive evaluation of the experts' knowledge. Their proposed method was specifically implemented and tested on FriendFeed's social network. However, this method was unable to identify experts in different areas. Zhao et al. [18] have proposed a method to improve the performance of expert finding in CQA systems using users' social networks. They also developed two iterative procedures, GRMC-EGM and GRMC-AGM, to solve the optimization problem. Their proposed algorithm had better performance in comparison to the state-of-the-art expert finding algorithms.

Besides, Cifariello et al. [19] proposed an unsupervised search engine for expert finding in academia. This system combines classical language modeling techniques, based on text evidences, with the Wikipedia Knowledge Graph, via entity linking. In another study, Rampisela et al. [20] used the thesis abstract and fasilkom UI students' metadata to find experts in the Faculty of Computer Science in Universities Indonesia. They used combination of word2vec and doc2vec to model semantic information, and indicated the comparable performance of the embedding models in retrieving experts using expertise queries in both Indonesian and English languages.

Given that people's posts are in different fields, they may have different levels of knowledge. So, Stephens-Martinez et al. [14] proposed the CEF method by changing the PageRank algorithm to run online forums in the CEF. The CEF algorithm can determine the level of users' knowledge in each domain. The relationship between topics was also considered in this study. Omidvar et al. [21] have proposed a novel algorithm to determine people's expertise level in AskMe forum using WordNet dictionary and social network analysis. Their proposed method outperforms other expert finding algorithms. Fu et al. [22] proposed a model that considers the relationship between users as well as the link between topics to

determine the expertise of each user. In addition, Brusilovsky and Millán [13] identified the people who had the highest rank in the list of experts. Then, they compiled a list of experts by calculating the level of communication of other people on the web pages or their emails with these experts. The algorithm used in this study was the diffusion algorithm, an important feature of which is the selection of some sample class of people. The problem with this research is it ignores the dynamic change of users who have been selected as experts.   The two mentioned approaches for finding experts are incomplete. For example, in the social network analysis, malicious users can expose their communications network by sending numerous unrelated or empty messages, and the algorithm automatically detects these users as experts. Moreover, in the content analysis approach, communication between people is not considered, and there is no difference between answers given to people with a different level of expertise. Therefore, in some studies, a combination approach has been used to find the experts. For example, Karimzadegan et al. [10] by using the hierarchical relationships within the organization have tried to identify the similarities of people in the organizational environment and improve their ranks. Their method was more helpful to new users who have not had much work so far. But they did not considered the topic similarity between users. In order to solve this problem, Ziaimatin et al. [23] have proposed a model that uses the PageRank algorithm. They indicated that the relevance and proximity of people who interact in different fields is of paramount importance. Zhou et al. [24] have proposed a different method that, in addition to counting the number of user answers in the expert network, also specifies the number of answers to the topics. Then, using the PageRank algorithm, calculates the similarity between the questioner and the respondent. The evaluations carried out in this study indicated that the proposed model works better than the content based models introduced in earlier studies. Rafiei and Kardan [25] have also proposed a hybrid method for expert finding in java online communities, which was based on content analysis and social network analysis. They indicated the better performance of their proposed method on expert finding. In the study [26] the LDA label model was proposed to determine the distribution of the user subject. In this study, features like post content, answer votes, proportion of the best answers and user relationships were considered to find the correct answer to the new question. Moreover, EL-Korany [3] proposed a method that combines two content-based approaches and social network analysis and uses ratings and the reputation of users. In this model, user profiles are first created using content analysis methods, then user rating posts are used to compute user reputation and to combine this rating with the content analysis step. In addition, using network methods, user ratings are

computed in the QA network, and the output of the two steps is combined. Their proposed method improved the accuracy of the expert finding. Wang et al. [27] have also proposed a novel expert finding algorithm, ExpertRank, which evaluates expertise based on both document-based relevance and one's authority in his or her knowledge community. They used PageRank algorithm to evaluate one's authority, and explored three different expert ranking strategies of linear combination, cascade ranking, and multiplication scaling that combine document-based relevance and authority. Their results indicated the better performance of the proposed algorithm than others. Gui et al. [6] proposed a new method for expert finding in heterogeneous bibliographical networks based on two aspects of textual content analysis and authority ranking. They indicated that the proposed framework outperforms existing methods.

2-2- Ant Colony Optimization Algorithms

Ant Colony Optimization (ACO) algorithms are constructive metaheuristics which inspired by the natural behavior of ants in finding the shortest paths between their nest and food sources by depositing pheromone on the ground in an iterative process. ACO uses the pheromone model to provide a set of solutions. The pheromone model is a set of values obtained by a reinforcement type of learning mechanism. Therefore, the pheromone model can be used to increase the probability of constructing good solutions from the components [28-29].

Different ACO algorithms have been used in the studies to solve the NP-hard problems. For example Anuradha et al. [30] used AntRank algorithm for ranking the web pages. In the same study, Setayesh et al [31] proposed a PageRank algorithm based on ACO. The results of simulation indicated that in their proposed algorithm, ratings are closer to real data, and more distinct ratings are generated.

As the research studies indicate, in earlier studies the Ant colony algorithm has not been used to find experts. In addition, all the basic algorithms for finding experts that are usually used for evaluation are not based on queries and just rank experts. The importance of using the Ant colony algorithm is that the existence of a possible component allows the ants to make various answers, and so they will answer more questions than greedy heuristics. In addition, the use of heuristic information, which, in many cases, helps the ants move towards more promising answers. Most importantly, the experience of ants during search by pheromone affects the answer construction in the next iterations of the algorithm and creates a dynamic learning process. In addition, the use of an ant colony gives more stability to the algorithm and the interaction of the population of agents (ants) together helps to solve the problem more effectively. Therefore, for the first time in this field, using the Ant colony algorithm and considering

the quality of communication and users' trust and reputation, a new method is proposed, which similar to [32] methodology uses some features of users such as ranking of posts to improve the algorithm's accuracy.

3- Proposed Algorithm

In this research, a new algorithm called the Expertise AntRank (EAR) is presented to find experts in each domain using the MetaFilter dataset, an online forum for discussing various issues [33]. In this algorithm, a new approach is proposed to solve the existing challenges in finding experts using statistical calculations and ant colony algorithm in online communities. In addition to expert finding, this algorithm indicates the level of expertise in different fields. The proposed algorithm consists of four phases: I. data preparation, II calculating the share of users' knowledge, III calculating the level of attraction between people based on the field, IV implementing the ant colony algorithm.

After data preparation, the Bayes' theorem is used to obtain the share of each user's knowledge who participated in the question and answer process, which its application is changed based on the conditions of this research. WordNet's dictionary, a processing program based on the psychology of language rules and encodes concepts as sets of synonymous words, is also used to determine the relevance of user-submitted questions to the domain of interest, [34]. In addition, changes were made to the PageRank algorithm in order to calculate the degree of attraction between people by considering the activity records and the similarity of these records with the users requested field. Finally, inspired by the pheromone part of the ant colony algorithm, users are treated as a route, questions are considered as ants, the level of attraction between people in the specialized network is treated as innovative information, and knowledge share of users in each question is considered as pheromone. . At the end of the ants' tour, the experts are identified and ranked by the amount of pheromone placed on each path.

Phase I: Data Preparation

The data preparation is done through the extract, transform and load (ETL) process. First, the required information is extracted from the HTML files pertaining to the MetaFilter website (Metafilter.com). Next, the extracted data is examined for any incorrect entries, and all users whose total number of questions and answers is less than 20 are deleted from the database. In addition, all related information, such as the answered questions and corresponding answers, all unanswered questions, and all questions that have not been assigned a keyword are deleted from their records. Then, the cleaned data is uploaded to the MetaFilter online database. Table 1 shows

statistical information about the data in the designed database.

Table 1. Statistical data of the MetaFilter online community

Number of Questions	Number of Answers	Number of Intended Answers	Number of Tags	Number of Users
229401	543699	2412653	951118	43676

Phase II: Calculation of User Knowledge

In online communities, each user can ask a new question, which will be answered by different users. Each question may receive more than 10 answers, all of which are somehow correct. But, the best answer is always chosen, which indicates that the proposed solution is more efficient and better. Once the best answer is given to a question, and the process of knowledge transfer ends, and the distribution of knowledge or the calculation of the share of each person's knowledge in response to the particular question is completed. To calculate the share of each user's knowledge in the QA process, the likelihood probability relationship based on the Bayes theory is adopted, which is modified based on the research conditions (equation 1):

$$K_{U_i P_j} = \frac{Votes(U_i.P_j)+1}{\sum_{k=1}^{N} Vote(P_{jk})+N} \qquad (1)$$

Where $K_{U_i P_j}$ is the knowledge gained by each user i in the question or post j, and $Votes(U_i, P_j)$ is the votes of the user i in question j. In addition, $\sum_{k=1}^{N} Vote(P_{jk})$ is the total number of votes earned in all (*N*) questions and answers.

For example, as table 2 indicates a question that asked by a user received 5 positive votes. Moreover, there are also 4 answers with different scores., The knowledge share of each user who contributed to this QA process is calculated using equation 1 (see Table. 2).

Table 2. An example of user knowledge calculation

Type of Post	Number of Votes	Best Answers	Users' share of knowledge
Question	5	-	0.24
Answer	4	Yes	0.2
Answer	1	No	0.08
Answer	0	No	0.04
Answer	10	No	0.44

At the end of this phase, a user-post table is created, in which the knowledge share of each person is specified from each post. The information of this table is used as a pheromone in the ant colony algorithm.

Phase III: Calculation of the Attractiveness

In the PageRank algorithm, nodes are web pages that are connected by page links [35]. In this study, nodes are memers of the online community instead of web pages, and and links are based on the responses sent between them. Then according to the questions and links among people

, the Chief Executive Network (CEN) is drawn [36]. At CEN, each user is linked to those who answered his/her questions, and each link is given a certain weight depending on the number of answers and the field of expertise. User who answers another user's question is thus deemed to have more information on that subject. Therefore, a link is drawn from the question to the respondent, by the questioner validates the respondent. To determine the relevance of a question to a subject, the distance between all the labels of that question to the intended field is calculated using the OSS distance function. The OSS function takes two concepts as inputs, and returns a numeric range of zero to one using the WordNet ontology. The closer the output number is to one, two concepts are more similar to each other [15]. Then, the average distance of the labels is considered as the relevance of the question to the subject and the weight of each edge is calculated using equation 2:

$$W_{AB} = \sum_{p=1}^{N_{AB}} \left(\frac{\sum_{t=1}^{N_p}(Distance(T_t.C))}{N_p} \right) \qquad (2)$$

Where N_{AB} is the number of answers of user B to user A, N_p is the number of the labels of question P, C denotes the domain of the desired expertise, T is its label, and distance is the OSS distance function that is used to calculate the distance between field C and label T.

The graph proximity table is used to construct a possible transfer table. If there is a non-zero row in the graph proximity matrix, all the entries of that row are placed in the probabilistic transfer matrix as $1/N$, where N is the number of members in the online community. Then, for the other rows, the value of each cell is calculated by using equation 3:

$$\frac{W_{ij}}{\sum_{z=0}^{z=N} W_{iz}} \qquad (3)$$

Where, W_{ij} is the weight of the link from person i to person j. using Equation 3, the weighted average of each relationship is calculated with respect to other relationships of that person. After making changes to the proximity table, all table cells are multiplied by 1-a, where the value of a is the probability of a mutation. In the mutation, the searcher can jump to any node in the graph. The destination of the mutation is also randomly selected. If the number of nodes in the graph is N, the mutation may move the searcher to any node (even to the current node) in the graph with probability of 1/N.

Phase IV: Implementing the Ant Colony Algorithm

At this phase, inspired by the pheromone part of the ant colony algorithm, users are treated as a route, questions are considered as ants, the level of attraction between people in the specialized network is treated as the information, and knowledge share of users in each question is considered as pheromone. First, a weighted

graph G = (N, A) is considered, in which N is a set of nodes or members of an online community with n = |N| members, and A is the set of edges. The dij is assigned to each (i, j) ϵ A edge, which indicates the level of attraction between members i and j. Ants can be considered as a possible construction procedure that produces an answer by moving on the graph G = (N, A). The ants do not move arbitrarily on the graphs, but rather use a construction policy that is subsidiary of the Ω constraints. The effect of pheromone T_i is considered for node i, which has a long-term memory of the search process and is updated by ants. A heuristic value n is also considered, which includes pre-problem information or information obtained at the time of problem solving from a source other than ants. Due to this feature, the algorithm, in addition to having an order based on ants' observation, also uses other desirable information that is outside of their observations to make decisions in finding more appropriate paths. The pseudocode of the metaheuristic behavior of the algorithm is given below:

Start the EAR metaheuristic process
Adjust the activity of the ants
Place each ant on a graph point
Place the initial value of the pheromone and calculate the heuristic values
Calculate the probability of an ant moving to the neighbouring points
Keep moving the ants to complete a tour for each ant
End setting activities
Update the pheromone
Pheromone placement based on the quality of the answer quality function
Pheromone Evaporation
End of pheromone update
End of the EAR's metaheuristic process

First, the variables and parameters of the algorithm, the number of program execution (termination condition), the initial pheromone value, α and β values, and the evaporation coefficient of the input are set. Then the probability of transition to neighboring nodes is calculated and the ants are randomly placed on the nodes by marking all the nodes as unappointed nodes. Then, each ant must be assigned to a primary city. To do this, we assign each ant to a random primary node. Finally, each ant makes a complete tour. At each step of answer making, the ants select the next node to navigate based on based on the constraint and the probability of transition. The only constraint on this issue is that all nodes should be met at most once. This constraint forces the ants to choose the next node through the unselected nodes during the answer construction process (that is, the possible neighborhood N^k_i of the ant k in node *i* and includes all nodes that have not yet met). When the ants are placed in randomly selected nodes, at each step of answer construction, an answer, the ant K uses a probabilistic selection law, called the relative random law, to select the next node. The probability of

selecting node j by ant k which is located on the node i is calculated by using equation 4:

$$p_{ij}^{k} = \frac{[T_i]^{\alpha}[n_{ij}]^{\beta}}{\sum_{i \epsilon N_I^K}[T_i]^{\alpha}[n_{ij}]^{\beta}} \; ; \; if \; j \epsilon \; N_i^k \qquad (4)$$

Where n is a heuristic value that is the inverse of users' desire ($n = 1/d_{ij}$) (which was calculated in the third phase. T_j is the amount of pheromone on the node j at any moment. Besides, α and β are two parameters that determine the effect of the dependent pheromone route and the heuristic information. N_i^k is the possible neighborhood for the ant k, (the probability of selecting a node out of N_i^k is zero).

After calculating the transition probability of all neighboring nodes, the rotary method is used to select the next node to navigate. The rotary method generates a random number between zero and one to select each node, and the mentioned number corresponding node of each interval that the given number is located in will be selected.

When all the ants have made their tours, the pheromone routes are updated. In this way, first, the amount of pheromone is reduced from all nodes by a fixed factor (evaporation), and then the amount of pheromone that is calculated in the second phase is added on the nodes through which the ants passed in their tours. Pheromone evaporation is calculated by using equation 5:

$$T_i \leftarrow (1 - p)T_i \qquad (5)$$

Where the evaporation rate of pheromone is $0 > p <= 1$, parameter p is used to prevent indefinite accumulation of pheromone pathways and enables the algorithm to forget the wrong decisions that have already made. If a node is not selected by ants, the amount of pheromone assigned to it decreases exponentially in repetitions. After evaporation, all the ants leave the pheromones on the nodes they have passed through their net. This means that the level of expertise decreases over time. The amount of pheromone that the ant k leaves on the nodes is calculated by using equation 6:

$$\Delta T_i = \begin{cases} \frac{K_i}{C^k} \; if \; node \; (i) belongs \; to \; T^K \\ 0 \qquad\qquad\qquad Otherwisw \end{cases} \qquad (6)$$

Where, C^k, the fitness of the tour T^k that is made by the ant k, is calculated as the sum of the edge sizes belonging to T^k, and K_i is the share of user knowledge. The pheromone update is calculated by using equation 7:

$$T_i(t + 1) = (1 - p).T_i(t) + \Delta T_i(t) \qquad (7)$$

The condition for terminating the algorithm is the number of iterations that are taken from the input. The appropriate answer can be reached by repeating the algorithm several times and keeping the best answers. That is, at the end of

the last iteration, a user-pheromone table is generated, which is sorted by pheromone columns to determine expert users and their level of expertise.

4- The Proposed Algorithm Evaluation

The proposed EAR method, along with other methods such as Indegree, Z-degree, Z-number, AnswerNum, ExpertiseRank and SNPageRank was run through MetaFilter's online community database.

The Indegree method is used to count the number of users that have helped others; Z-degree is the Z-score of the number of people whom the user has asked or answered; Z-number is the Z-score of the number of questions and answers that a user submitted; AnswerNum is the simplest way to identify experts by counting the number of their answers to the questions; ExpertiseRank provides the ranking of users in terms of decreasing the level of expertise in online communities. It counts the number of people who have helped the user as well as the number of users who have answered to each of these people [37]; SNPageRank is based on PageRank algorithm and is used for the accurate and comprehensive evaluation of the experts' knowledge in a social network [2].

The results are then compared under three different tests in three domains: travel, music, and internet. Fifty questions are selected for each field, ensuring that each question had an answer. The best answer for each question is specified by the person who originally asked the question. The selection of set of questions is done three times for each field, with the next steps being performed each time.

- First, questions with fewer than five answers are selected. This is done in such a way that expert-oriented approaches have a better chance of being selected by the appropriate respondent. In the next step, questions with 5 to 10 answers are selected. In the last step, questions with 10 or more answers are selected.
- The set of selected questions are then deleted from the database.
- All algorithms are run on a data set in order to rank the users.
- The best responder is selected for each question in the test based on the ranking of the algorithms. Finally, it is confirmed whether or not the selected participant gives the best answer.

The accuracy of each algorithm is calculated according to equation 8:

$$Accuracy = \frac{N_1}{N_1 + N_2} \qquad (8)$$

Where N_1 is the number questions in which the best respondent is correctly selected, N_2 is the number of questions in which the best respondent is not selected correctly. According to the results, the proposed method

has better results than other methods in finding the best respondent for each question. Figure 3 indicates the average accuracy of the proposed expert finding in three different test areas. As figure 1 demonstrates, the proposed algorithm with the average accuracy of 93.33% has better results in finding the best respondent than other algorithms in all three domains. Subsequently, two SNPageRank and ExpertRank algorithms yielded better results than other methods. Of course, the ExpertRank algorithm is able to detect the best music answer better than the SNPagerank algorithm. It is also able to identify the best responder for 35 of the 50 survey questions.
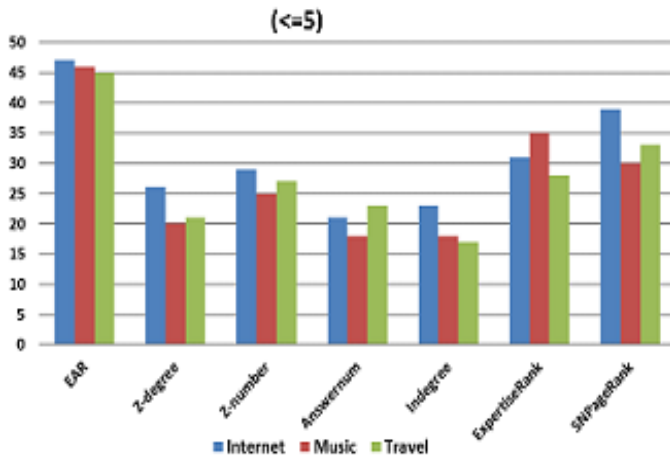

Figure 1. The results of the test for questions with 5 answers

Figure 2 indicates the accuracy of algorithms to find the best answer for questions with 5 to 10 answers. According to the figure, as in the previous test, the proposed algorithm is able to find the best respondent better than other algorithms in all three domains with the average accuracy of 74%. Of course, the results of the SNPageRank algorithm are close to the proposed algorithm. In addition, the results of Z-number method are better than those of ExpertiseRank algorithm and place in the third rank among the methods.
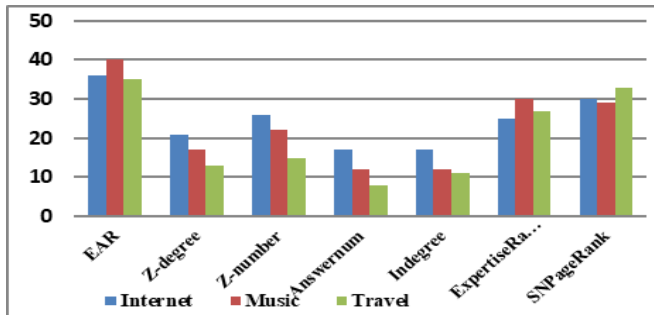

Figure 2. The results of test for questions with 5 to 10 answers

Figure 3 indicates the accuracy of the algorithms to find the best answer for the questions with more than10

answers. As with previous tests, the proposed method with the average accuracy of 50% has obtained better results than other methods.
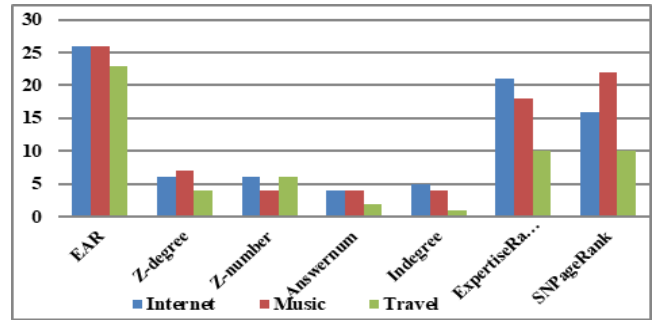

Figure 3. The results of the test for questions with more than 10 answers

Based on the results, the expert-oriented methods have better results in the first test, and the results of the second test are better than the third test. This is due to h increased number of respondents in this test compared to the first and third test. This reduces the probability of finding the best respondent in the expert-oriented approaches.

In the next step, the Spearman correlation method is used to compare the performance of proposed method with basic methods. To calculate the Spearman correlation value, n which is the raw score that people may have obtained in each subject, are presented by two symbols of X_i ($X_{1+}X_{2+...}+X_n$) and Y_i ($Y_{1+}Y_{2+...}+Y_n$). The value of d_i which is the difference between the ranks of each observation is then calculated as ($d_{i=}x_{i-}y_i$). In the absence of equal rankings, equation 9 is used [38]:

$$p = 1 - \frac{6 \sum id_i^2}{n(n^2 - 1)} \qquad (9)$$

Where X_i and Y_i are two points that user i received in two different rankings, and x_i and y_i are the rankings that person i received as a result of getting X_i and Y_i points respectively. As such, d_i is the difference in the ranks a user received in two subjects, and n is the total number of people.

In this evaluation, first all users who not answered any questions or their in-degree is equal to zero, are removed from the dataset. Then the first 1,000 users with a highest number of best answers are separated from the remaining users. The reason for this separation is that after observing the points earned by others, users with a rank of 1,000 onwards have similarly low scores. Thus, it can be said that the competition in expertise is among the first 1,000 users. Of these 1,000 users, 50 are randomly selected and Spearman correlation is calculated on them based on the total number of their best answers and output scores. The results indicate a correlation of 75.34%, which is a better result than other methods (Figure 4).
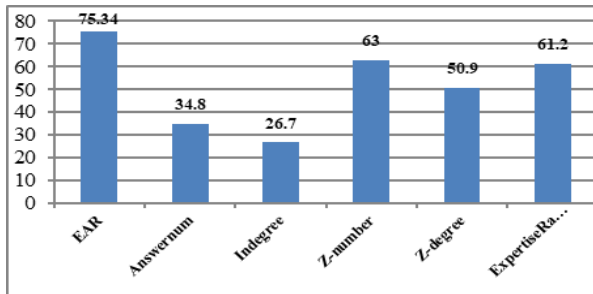
Figure 4. The accuracy of the proposed EAR method compared with basic methods of experts finding

One of the problems of the proposed method is its high computational complexity and its high temporal complexity compared to the basic methods. The reason for this complexity is that the proposed algorithm is context-based, meaning that all the basic expert finding algorithms are not query-based, and only rank experts. However, this advantage of the proposed method added the computational load of calculating the similarity between words to the algorithm.

5- Discussion and Conclusion

Today, people from all over the world can freely share their questions or opinions in online communities and social networks. Given that the number of answers to most questions is high and it is not possible to determine the accuracy of the received answers, online community users cannot trust on the answers of other users. Moreover, in some online communities such as MetaFilter, the expertise level of members is unclear. This requires the proper participation of all users, as well as the provision of accurate comments. In this study, the role and importance of online communities in knowledge sharing was discussed, and for the first time in this field, a model for finding experts and determining their domain expertise level by the aid of statistical calculations and the ant colony algorithm in the MetaFilter online community was presented. This algorithm uses the pheromone part of the ant colony algorithm to determine the level of people expertise in the intended field based on the similarity of users' questions with the studied field and the share of the users' knowledge of their interactions in the online community. The MetaFilter online community data was used to evaluate the proposed method.

The results of the proposed EAR method were compared with other methods like Indegree, Z-degree, Z-number, AnswerNum, ExpertiseRank, and SNPageRank in three different domains of internet, travel and music. Based on the results, the accuracy of the proposed method in finding the best respondent for each question was much better than other methods proposed by [2, 15-19]. In addition, the results of Spearman correlation indicated the correlation of 75.34% which was a better result than other methods. Therefore, the proposed method can be used in recommender systems, in such a way that people can submit their questions to the online community. Then, instead of exposing the question to everyone, it is suggested only to those who can provide appropriate answers.

The novelty of this research is the introduction of a new context-based expert finding method for and determining their level of expertise in each domain using the Ant colony algorithm in the MetaFilter online community. Since the WordNet dictionary has ontology, it was used to select the distance between concepts and to identify the background of members' knowledge. Moreover, the pheromone part of the ant colony algorithm was used to determine the level of users' expertise in each domain.

None of the methods used for expert finding can determine the level of individuals' knowledge in different fields. To achieve this goal, users' opinions should first be categorized and then the previous methods should be applied to different categories. Therefore, in order to change, improve and complete the identification model of experts provided in this research, a recommender system can be designed for online communities to answer any questions asked by users who have enough knowledge to answer them. Users can now see all the questions asked by other. Because the number of questions is too high, the large number of specialized questions that only a few users can respond, may remain unanswered, or take the long time to respond. Therefore, a recommender system could be designed and developed for QA forums to employ EAR algorithm for expert finding in order to show each question intelligently to users who are more knowledgeable for answering the question.

The biggest problem of a user in using online communities is the time it takes to review the entire content of a conversation, some of which reach more than hundreds of massages. Therefore, it seems desirable to summarize conversations in order to provide sufficient knowledge for user without losing important information. So far, various summarization methods have been proposed for multiple documents, but none of them consider the main feature of online communities, which is multiple writers. Therefore, the most important issue in generating a desire summary will be to identify experts and use their opinions in the summary form. In fact, the proposed method should be used for expert finding first, and then the summarization algorithm should be designed, which uses the writings of the experts in the process of summarizing.

Using concept map, a graph-like tool for knowledge representation, in each scientific domain, the semantic similarity between each term can be determined. These maps represent a meaningful relationship between the concepts of a subject, and can be used to determine the

similarity of concepts. In fact, for specialized concepts such as Java programming language terms that are not found in the WordNet dictionary, a concept map created by experts for this programming language can be used. [28].

Moreover, some distance-definition functions, such as the method presented in [39] uses the meanings of the words in WordNet to determine the distance between concepts. Such methods can be used in the proposed EAR method to investigate their effects on the accuracy of the proposed algorithm.

In the proposed EAR method, the keywords of the questions are used to determine the relevance of the questions to the field of knowledge. Questions in the MetaFilter online community are tagged, but in many online communities such as YA, this does not matter. Therefore, keyword extraction methods can be used to tag questions and answers in online communities.

To improve the accuracy and efficiency of the proposed algorithm in finding experts, future studies might determine the difficulty of question by using factors such as time elapsed after asking a question and the number of answers, and add it as a coefficient to the equation of knowledge-level computing.

In addition, future studies can use other methods of ensemble intelligence and compare their results with the proposed algorithm. Besides, future studies can apply the proposed model to other online communities and compare the results.

References

[1] T., Lappas, K., Liu, & E, Terzi. "A survey of algorithms and systems for expert location in social networks." In Social Network Data Analytics (pp. 215-241). Springer, Boston, MA. 2011.

[2] A., Kardan, A., Omidvar and F., Farahmandnia "Expert finding on social network with link analysis approach." In *2011 19th Iranian Conference on Electrical Engineering*, pp. 1-6. IEEE, 2011.

[3] A., El-Korany. "Integrated Expert Recommendation Model For Online Communities." Arxiv preprint arxiv: 1311.3394. 2013.

[4] A., Daud, M., Ahmad, M. S. I., Malik & D., Che. "Using machine learning techniques for rising star prediction in co-author network." Scientometrics, Vol. 102, No. 2. 2015. pp. 1687-1711.

[5] L. A., Adamic, J., Zhang, E., Bakshy, & M. S., Ackerman. "Knowledge sharing and yahoo answers: everyone knows something." In Proceedings of the 17th international conference on World Wide Web (pp. 665-674). ACM. 2008, April.

[6] H., Gui, Q., Zhu, L., Liu, A., Zhang, & J., Han. "Expert Finding in Heterogeneous Bibliographic Networks with Locally-trained Embeddings." arXiv preprint arXiv: 1803.03370. 2018.

[7] M., Neshati, Z., Fallahnejad, & H., Beigy. "On dynamicity of expert finding in community question answering."

Information Processing & Management, Vol. 53, No. 5. 2017. pp. 1026-1042.

[8] X., Ding, B., Liu & P. S., Yu "A holistic lexicon-based approach to opinion mining." In Proceedings of the 2008 international conference on web search and data mining (pp. 231-240). ACM. 2008.

[9] M., Karimzadehgan, R. W., White & M., Richardson. "Enhancing expert finding using organizational hierarchies." In European Conference on Information Retrieval (pp. 177-188). Springer, Berlin, Heidelberg. 2009.

[10] S. B., Sriramoju. "Heat Diffusion Based Search for Experts on World Wide Web." International Journal of Science and Research (IJSR), https://www. Ijsr. Net/archive/v6i11/v6i11. Php, Vol. 6, No. 11. 2017. pp. 632-635.

[11] X., Liu, W. B., Croft & M., Koll. "Finding experts in community-based question-answering services." In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 315-316). ACM. 2005, October.

[12] S., Ravi, B., Pang, V., Rastogi & R., Kumar. "Great Question! Question Quality in Community Q&A." ICWSM, Vol.14, No. 2014. pp. 426-435.

[13] P., Brusilovsky, & E., Millán. "User models for adaptive hypermedia and adaptive educational systems." In The adaptive web (pp. 3-53). Springer, Berlin, Heidelberg. 2007.

[14] K., Stephens-Martinez, M. A., Hearst & A., Fox. "Monitoring moocs: which information sources do instructors value?" In Proceedings of the first ACM conference on Learning@ scale conference, (pp. 79-88). ACM. 2014.

[15] S., Yuan, Y., Zhang, J., Tang, W., Hall, & J. B., Cabotà. "Expert finding in community question answering: a review." *Artificial Intelligence Review*, Vol. 53, No. 2, 2020, pp. 843-874.

[16] M., Dadkhaha, M., Lagziana, F., Rahim-niaa, & K., Kimiafar. "The potential of business intelligence tools for expert finding." *Journal of Intelligence Studies in Business*, Vol. 9, No. 2. 2019. pp. 82-95.

[17] J., Zhang, J., Tang & J., Li. "Expert finding in a social network." In International Conference on Database Systems for Advanced Applications, (pp. 1066-1069). Springer, Berlin, Heidelberg. 2007.

[18] Z., Zhao, L., Zhang, X., He & W., Ng. "Expert finding for question answering via graph regularized matrix completion." IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 4. 2015. pp. 993-1004.

[19] P., Cifariello, P., Ferragina, & M., Ponza. "Wiser: A semantic approach for expert finding in academia based on entity linking." *Information Systems*, Vol. 82 No.2019, 2019. pp. 1-16.

[20] T. V., Rampisela, & E., Yulianti. "Academic Expert Finding in Indonesia using Word Embedding and Document Embedding: A Case Study of Fasilkom UI. In *2020 8th International Conference on Information and Communication Technology (ICoICT)* pp. 1-6. IEEE. (2020, June).

[21] A., Omidvar, M., Garakani, & H. R., Safarpour. "Context based user ranking in forums for expert finding using WordNet dictionary and social network analysis." Information Technology and Management, Vol. 15, No. 1. 2014. pp. 51-63.

[22] F., Yupeng, R., Xiang, Y., Liu, M., Zhang, and Sh., Ma. "Finding experts using social network analysis." In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 77-80. IEEE Computer Society, 2007.

[23] H., Ziaimatin, T., Groza, G., Bordea, P., Buitelaar, & J., Hunter. "Expertise Profiling in Evolving Knowledge curation Platforms." GSTF Journal on Computing (JoC), Vol. 2, No. 3. 2018. pp. 118-127.

[24] G., Zhou, S., Lai, K., Liu & J., Zhao. "Topic-sensitive probabilistic model for expert finding in question answer communities." In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1662-1666). ACM. 2012.

[25] M., Rafiei & A. A., Kardan. "A novel method for expert finding in online communities based on concept map and PageRank." Human-centric computing and information sciences, Vol.5, No. 1. 2015. pp. 10-28.

[26] H., Li, S., Jin & L. I., Shudong. "A hybrid model for experts finding in community question answering." In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference on (pp. 176-185). IEEE. 2015.

[27] G. A., Wang, J., Jiao, A. S., Abrahams, W., Fan & Z., Zhang. "Expert Rank: A topic-aware expert finding algorithm for online knowledge communities." Decision Support Systems, Vol. 54, No. 3. 2013. pp. 1442-1451.

[28] Z., Dorrani, & M. S., Mahmoodi "Noisy images edge detection: Ant colony optimization algorithm." Journal of AI and Data Mining, Vol 4, No 1. 2016. pp. 77-83.

[29] M. B., Dowlatshahi & V., Derhami. "Winner Determination in Combinatorial Auctions using Hybrid Ant Colony Optimization and Multi-Neighborhood Local Search." Journal of AI and Data Mining, Vol. 5, No 2. 2017. pp. 169-18.

[30] G., Anuradha, G. L., Devi & M. P., Babu. "Antrank: An Ant Colony Algorithm for Ranking Web Pages." International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Vol. 3, No. 2. 2014. pp. 208-212.

[31] S., Setayesh, A., Harounabadi & A. M., Rahmani. "Presentation of an Extended Version of the PageRank Algorithm to Rank Web Pages Inspired by Ant Colony Algorithm." International Journal of Computer Applications, Vol. 85, No. 17. 2014. pp. 7-10.

[32] D. R., Liu, Y. H., Chen, W. C., Kao & H. W., Wang. "Integrating expert profile, reputation and link analysis for expert finding in question-answering websites." Information processing & management, Vol. 49, No. 1. 2013. pp. 312-329.

[33] L., Silva, L., Goel & E., Mousavidin. "Exploring the dynamics of blog communities: the case of MetaFilter." Information Systems Journal, Vol. 19, No.1. 2009. pp. 55-81.

[34] S. H., Hashemi, M., Neshati, & H., Beigy. "Expertise retrieval in bibliographic network: a topic dominance learning approach." In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (2013, October), pp. 1117-1126.

[35] R., Navigli. "Word sense disambiguation: A survey." ACM Computing Surveys (CSUR), Vol. 41, No.2. 2009. pp. 10-79.

[36] M., Rowe. "Mining User Development Signals for Online Community Churner Detection." ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 10, No. 3, 2016. pp. 1-29.

[37] M., Neshati, D., Hiemstra, E., Asgari, & H., Beigy. „Integration of scientific and social networks." World Wide Web, Vol. 17, No. 5, 2014. pp. 1051-1079.

[38] A. D., Well & J. L., Myers. *Research design & statistical analysis*. Psychology Press. USA. 2003.

[39] D., Petkova & W. B., Croft "Hierarchical language models for expert finding in enterprise corpora." International Journal on Artificial Intelligence Tools, Vol. 17, No.1. 2008. pp. 5-18.

Mojtaba Sharifian holds a Master's in Information Technology Engineering from the Islamic Azad University of Qazvin, Qazvin, Iran. He is interested in research in the field of information systems and technology, including innovation in ICT, business intelligence, electronic commerce, and big data analytics and cloud computing.

Neda Abdolvand is an Associate Professor in the Alzahra University and was a Postdoctoral Research Fellow at Tarbiat Modares University. She holds a PhD and MS in Information Technology from the Tarbiat Modares University and a Postgraduate Certificate in Information Systems from Melbourne University. She is interested in research in the field of information systems and technology, including innovation in ICT, business intelligence, electronic commerce, big data analytics and cloud computing. She has been a member of the Association for Information Systems since 2009.

Saeedeh Rajaee Harandi holds a Master's in Information Technology Management from the Alzahra University, Tehran, Iran. She is interested in research in the field of information systems and technology, including innovation in ICT, business intelligence, electronic commerce, big data analytics and cloud computing.