International Journal of Natural and Engineering Sciences 8 (3): 18-21, 2014
ISSN: 1307-1149, E-ISSN: 2146-0086, www.nobel.gen.tr

Experts Ranking in Online Communities Using Combination of Text and Link Analysis

Adel ZERAAT* Mehdi GHAZANFARI Mohamad FATHIAN
Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

*Corresponding author:
E-mail: ad.zeraat@gmail.com

Received: August 24, 2014
Accepted: October 06, 2014

Abstract
Online communities are a question answering environment where individuals can express their opinions freely. Quality of the information generated in the online community is dependent on the individual expertise level. If the individual has higher level of expertise, shared knowledge in online community is valuable and reliable. Experts ranking methods used for determining expertise level of individuals and evaluating the accuracy of shared knowledge. In this study, a novel hybrid method for ranking experts in online communities is presented. This approach incorporates users' relationship and content of users' answers to finding expert users in an online community. This method is applicable to all online communities and only corpus in the field of online community is needed to accomplish that. We evaluated our proposed method on Java online community and Cryptography section of StackExchange online community. Correlation between scores of our method and scores of expert users introduced in both online communities exceeds 0.8, which is highly a reasonable value.
Keywords: Expert ranking, online community, knowledge sharing, social network, semantic similarity.

INTRODUCTION

Online community is a collective group of entities, individuals or organizations that come together through an electronic medium to interact in a common problem or interest space [1]. Professional online community is an online social network in which people with common interests, goals or practices interact to share information and knowledge [2].

Nowadays, knowledge is considered as a source of competitive advantage for individuals. Knowledge sharing is an activity which an individual imparts his or her expertise to another individual. One of the most important applications of online communities is knowledge sharing.

There is a great deal of knowledge shared in online communities but it is very challenging to determine their accuracy. Unknown values of responses, long time of response time, wasting time of experts with simple questions and large volume of information are main challenges in online communities.

By expert finding, questions are represented to experts and answers are represented to questioners based on the individual expertise level; subsequently experts spend their time just to answer the questions that others are incapable of responding to as well as large volume of information in online communities can be summarized and questioners is not confused with large number of responses. In additions accuracy of answers can be determined.

As aforementioned, it is clear that expert finding is critical issue in online communities and can be very useful for utilization mass of shared knowledge. In this study, a new hybrid method for expert finding in online communities is presented. Proposed method recognizes expert users by combining network-based and content-based approaches. For this purpose a social network based on replying relationships is constructed and weights of edges are computed by using threads content analysis between users. This paper is organized as follows. In the next section, we discuss the related studies in this field. In section 3, basic concepts are expressed. In section 4, we introduce our proposed method. Experiment results are presented in section 5. Conclusion of the study and future works is described in section 6.

Related Studies

According to the literature review, methods of expert finding can be divided into three categories includes: network-analysis approach, content-analysis approach and hybrid approach. These approaches are described as follows.

Network-analysis approach focuses on link analysis techniques in order to identify experts. For this purpose graph-based ranking algorithms such as PageRank and HITS have been used to rank users' expertise. In these algorithms individuals are considered as nodes and relationship between them is considered as edges in a graph. Exchange information between two individuals forms an edge between them. In [3], HITS algorithm, based on email of people in the organization was used to rank users' expertise. In [4], experts were discovered using HITS algorithm in question answering community. In [5], network-based approaches such as PageRank and HITS were employed to finding experts in an online community.

Content-analysis approach first extracts keywords of a user and represents that user by a term vector, subsequently experts are extracted by standard information retrieval techniques using vector space model. In [6], a content-based system has been developed for finding experts. In [7], experts can be identified with information retrieval techniques. In [8], an expert finding method based on assumption of sequential dependence between a candidate expert and the query terms in the scope of a document was presented. In [9], two models based on probabilistic language modelling techniques were proposed which have been successfully applied in other Information Retrieval tasks.

Both network-analysis and content-analysis approaches for finding experts ignore some properties of the data. Methods using network-analysis approach ignore the content of users' threads in order that someone can send irrelevance responses and obtain more scores, because he or she has many links. Whereas content-analysis methods do not consider relations between individuals in order that expertise of a user who answers experts' questions will be ignored.

As aforementioned, since network-analysis or content-analysis approach alone has its limitations, hybrid approach incorporates these approaches to finding experts. In [10], experts were identified by using person local information and relationships between persons in a unified approach. In [11], a novel approach was proposed by combining features of both above-mentioned approaches. In [12], a model with combining features of network-analysis and content-analysis approaches was proposed which recommends the most helpful experts.

In this study, we use the advantages of network-analysis and content-analysis approaches by combining features of both approaches. For this purpose we propose a novel method to construct the weighted social network in an online community which semantic similarity of questions and answers between two individuals has been used for calculating weights of edges. Subsequently score for individuals were calculated and users were ranked within an online community.

Basic Concepts

Online communities

Online communities are interactive environment in which people can express their opinions freely. Java online community and StackExchange online community are typical of these communities.

Until February 2013, Java forum has nearly one million users and almost two million and a half questions in the forum. These statistics clearly indicates that this online community is highly active. Java online community is divided into 16 subsections which each subsection corresponding to one of the Java technologies. The community introduces top participants in each subsection which helps individuals to identify experts. The primary way to gain score is by posting useful answers, so that the questioner can use two types of labels for each response. If questioner choice "Helpful" label for an answer, respondent user receives 5 points, and if questioner select "Correct" label, respondent user receives 10 points. Java online community introduces 10 top users based on these points.

The primary way to gain points in StackExchange online community is by posting good questions and useful answers, so that 5 points for proper question, 10 points for proper answer and 15 points for accepted answer be considered.

Semantic similarity

Semantic similarity between two words is obtained based on the likeness of their meaning content. There are three approaches for the extraction of semantic similarity includes: semantic network-based, definition-based and corpus-based. These approaches are described as follows.

Semantic network-based measures use a semantic network in order to calculate similarities. These measures are often referred as knowledge-based measures. In this approach, using WordNet, Concept Map or other resources, semantic similarities are extracted. Semantic network-based measures such as [13, 14], have a high precision but limited coverage.

Definition-based measures derive similarity scores from a set of explicit term definitions. These measures are also known as dictionary-based measures. In this approach, using the definitions in dictionaries or the web (Wikipedia) or other resources and based on Vector Space Model, semantic similarities are extracted. Definition-based measures such as [15, 16], have a high precision but limited coverage, same as Semantic network-based measures.

Corpus-based measures derive similarity scores from a text corpus. In this approach, using the lexical and dependency patterns in corpus, semantic similarities are extracted. Corpus-based measures such as [17, 18], provide extensive coverage but lower precision.

In this study, the aim is to cover all forums, thus third approach has been selected. This approach does not require any resources such as WordNet or dictionaries, instead semantic similarities are extracted based on a number of corpus.

Lexico-syntactic patterns

A lexico-syntactic pattern relies on lexical information and syntactic categories. Lexico-syntactic patterns are generalized linguistic structures which indicate semantic relationships between terms. In [19], 18 patterns are presented, as follows:

Such {NP=hyper} as {NP=hypo}, {NP=hypo}[,] and/or {NP=hypo}.

{NP=hyper} such as {NP=hypo}, {NP=hypo}[,] and/or {NP=hypo}.

{NP=hypo}, {NP=hypo}[,] or other {NP=hyper}.

{NP=hypo}, {NP=hypo}[,] and other {NP=hyper}.

{NP=hyper}, including {NP=hypo}, {NP=hypo}[,] and/or {NP=hypo}.

{NP=hyper}, especially {NP=hypo}, {NP=hypo} [,] and/or {NP=hypo}.

{NP=hyper}: {NP=hypo}, [{NP=hypo},] and/or {NP=hypo}.

{NP=hypo} is DET ADJ.Superl {NP=hyper}.

{NP=hyper}, e. g., {NP=hypo}, {NP=hypo}[,] and/or {NP=hypo}.

{NP=hyper}, for example, {NP=hypo}, {NP=hypo}[,] and/or {NP=hypo}.

{NP=syn}, i. e.[,] {NP=syn}.

{NP=syn} (or {NP=syn}).

{NP=syn} means the same as {NP=syn}.

{NP=syn}, in other words[,] {NP=syn}.

{NP=syn}, also known as {NP=syn}.

{NP=syn}, also called {NP=syn}.

{NP=syn} alias {NP=syn}.

{NP=syn} aka {NP=syn}.

Lexico-syntactic patterns are used for extraction of semantic relations from text of corpus.

PROPOSED METHOD

As aforementioned, for calculating weight of edges in a social network, semantic similarity of question and answer between two individuals has been used. At the beginning, information of corpus are extracted which are used to calculating semantic similarity. For this purpose, a collection of 56 e-books related to Java technology and 45 e-books related to cryptography was collected; subsequently, lexico-syntactic patterns are applied to the input corpus and all the concordances matching have been

retrieved. The total number of extracted relations for Java and cryptography e-books was 27400 and 17766 respectively.

After extracting related words, word pairs are ranked using equation 1 which is presented in [19].

$$Sim_{ij} = \sqrt{p_{ij}} \cdot \frac{2 \cdot \mu_b}{b_{i*} + b_{*j}} \cdot \frac{P(c_i, c_j)}{P(c_i) \cdot P(c_j)} \quad (1)$$

In equation 1:

Phrase p_{ij} is a number of patterns which word pairs are related, indicating the word pairs extracted by several patterns are more similar than those extracted only by a single pattern.

Phrase $\frac{2 \cdot \mu_b}{b_{i*} + b_{*j}}$ penalizes terms that are related to many words, where $\mu_b = \frac{1}{|C|} \sum_{i=1}^{|C|} b_{i*}$ is an average number of related words per term and $b_{i*} = \sum_{j:e_{ij} \geq 1} 1$ is a number of concordances containing word c_i and e_{ij} is equal to the frequency of extractions between pair c_i, c_j. Also, $b_{*j} = \sum_{i:e_{ij} \geq 1} 1$.

Phrase $\frac{P(c_i, c_j)}{P(c_i) \cdot P(c_j)}$ penalizes relations to general words, where $P(c_i, c_j) = \frac{e_{ij}}{\sum_{ij} e_{ij}}$ is the extraction probability of the pair c_i, c_j and $P(c_i) = \frac{f_i}{\sum_i f_i}$ is the probability of the word c_i and f_i is the frequency of c_i in the corpus. Also, $P(c_j) = \frac{f_j}{\sum_j f_j}$.

For the next step, information relevant to the user's profile and the user's posts of online community are extracted. The formatted data was stored in a database which provided inputs to calculating scores of users.

After extracting information of online community, social network of given online community was constructed. For this purpose, we analyzed threads in the online community and extract interactions between individuals. For calculating weight of edges we used semantic similarity of question and answer between two individuals. For each answer and question, similarity is calculated using equation 2.

$$SRQ(Response, Question) = \frac{2 \times \sum_{R \in Response} \left(\sum_{Q \in Question,\, Sim_{RQ} > 0} Sim_{RQ} \right) / Num_Q}{Count_{R \in Response} + Count_{Q \in Question}} \quad (2)$$

In equation 2:

$Response$: Keywords in the text of response.

$Question$: Keywords in the text of question.

Sim_{RQ}: Semantic similarity between keyword R in the response and keyword Q in the question.

Num_Q: The number of keywords in the question which related to response.

$Count_{R \in Response}$: The number of keywords in the response.

$Count_{R \in Question}$: The number of keywords in the question.

Weight of edge between individual i and j is calculated using equation 3.

$$W(v_i, v_j) = \frac{\sum_{R \in Responses_{ji}} SRQ(R, Q_R)}{Count_{R \in Responses_{ji}}} \quad (3)$$

In equation 3:

$Responses_{ji}$: All responses of j to i. Someone may have more than one response to another.

$SRQ(R, Q_R)$: Similarity of response R and question Q related to R.

$Count_{R \in Responses_{ji}}$: The number of responses which j answered to i.

After constructing weighted social network of online community, we can calculate the score for individuals and rank users. Final user's score is calculated using equation 4.

$$UserScore(A) = \sum_{\forall v_i:\, e_{iA} \epsilon E} W(v_i, v_A) \quad (4)$$

These scores have been calculated for all users of each subsection in Java online community and Cryptography section of StackExchange online community and individuals who have the most scores have been recognized as experts.

EVALUATION AND RESULTS

To evaluate the proposed method, Spearman correlation between the results provided by our method and Java online community and Cryptography section of StackExchange online community is used. For this purpose, first number of responses for each subsection of Java online community was computed and subsections which the number of responses is more than 3000 have been considered. The number of remained subsections was 11. Subsequently, final user's score is calculated using equation 4. Afterward Spearman correlations were calculated separately for the 11 subsection as well as for the entire Java online community. Overall correlation is computed by taking the average of correlations which was equal to 0.86.

Table 1 describes the statistics of the collected dataset from Java online community and results of our proposed method. In this table, the abbreviations are defined as:

NQ: Number of Question.
NR: Number of Response.
NU: Number of active Users.
Sp: Spearman Correlation.

Table 1. Information about subsections of Java forum

Category	NQ	NR	NU	Sp
ALL	6465	345206	614	**0.98**
Database Connectivity	367	23456	254	**0.77**
Development Tools	308	4869	152	**1**
Java APIs	337	31096	105	**0.93**
Java Card	415	4145	28	**1**
Java Desktop	1657	54839	150	**0.71**
Java Enterprise & Remote Computing	447	40642	134	**0.75**
Java Essentials	2266	157398	264	**0.73**
Java HotSpot Virtual Machine	40	8299	51	**0.89**
Java Security	86	4868	50	**0.88**
Java FX	484	9689	53	**0.80**
Other Topics	58	5905	47	**0.89**

Correlation between scores of our method and scores of expert users introduced in Cryptography section of StackExchange online community exceeds was equal to 0.81.

To evaluate the performance of our proposed method, we compared our method with Indegree technique which was explained in [20]. Indegree algorithm use non-weighted network for finding experts. Average of spearman

correlations for Indegree in Java online community was calculated equal to 0.78 and in Cryptography section of StackExchange online community was calculated equal to 0.8 which indicate the validity of our proposed method.

Conclusions and Future Works

In this study, a new hybrid method for expert finding in online communities is presented. We combined content-based and network-based approaches to recognize expert users in an online community. We proposed new method for constructing weighted social network based on replying relationships in threads of online community and thread content analysis between users. The proposed method is applicable to all online communities and corpus in the field of online community is needed. This method covers all forums and does not require any resources such as WordNet or dictionaries.

Using Java and StackExchange online communities as test bed, we thoroughly evaluated our proposed method and results demonstrated the validity of our proposed method so that correlation between scores of our method and scores of expert users introduced in these online communities exceeds 0.8, which is highly an acceptable value.

By finding experts with this method, we can determine what answers are more reliable and the response time is reduced. In addition, large volume of information in online communities can be summarized. Thus, those who seek to find answers in online communities are not confused with large volume of information. As a result, this method can be used in order to better exploit the valuable volume of information contained in online communities.

In the future, we can consider other network-based ranking algorithms such as PageRank to ranking users. In the proposed method, semantic similarity between keywords is obtained by corpus-based approach. In the future, other approaches can be used for extracting semantic similarity, such as network-based or definition-based approach.

REFERENCES

[1] Plant R. Online communities. *Technology in Society* 2004; 26(1): 51–65.

[2] Chao-Min C, Meng-Hsiang H and Eric T. Understanding knowledge sharing in virtual communities: an integration of social capital and social cognitive theories. *Decision Support Systems* 2006; 42(3): 1872-1888.

[3] Campbell C, Maglio P, Cozzi A and Dom B. Expertise identification using email communications. In: *Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03)*, 2003, pp. 528-531.

[4] Jurczyk P and Agichtein E. Discovering authorities in question answer communities by using link analysis. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM '07)*, 2007, pp. 919-922.

[5] Zhang U, Ackerman M, Adamic L and Nam K. QuME: a mechanism to support expertise finding in online help-seeking communities. In: *Proceedings of the 20th annual ACM symposium on User interface software and technology*, 2007, pp. 111-114.

[6] Krulwich B and Burkey C. ContactFinder: Extracting indications of expertise and answering questions with referrals. In: *Symposium on Intelligent Knowledge Navigation and Retrieval*, 1995, pp. 85-91.

[7] Liu X, Croft W and Koll M. Finding experts in community-based question-answering services. In: *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM '05)*, 2005, pp. 315-316.

[8] Serdyukov P, Rode H and Hiemstra D. Exploiting sequential dependencies for expert finding. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08*, 2008, pp. 795-796.

[9] Balog K, Azzopardi L and Rijke M. Formal models for expert finding in enterprise corpora. In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)*, 2006, pp. 43-50.

[10] Zhang J, Tang J and Li J. Expert Finding in a Social Network. In: Kotagiri, Ramamohanarao and Krishna, P.Radha and Mohania, Mukesh and Nantajeewarawat, Ekawit. *Advances in Databases: Concepts, Systems and Applications: Springer Berlin Heidelberg*. Springer, 2007, pp. 1066-1069.

[11] Serdyukov P, Rode H and Hiemstra D. Modeling multi-step relevance propagation for expert finding. In: *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)*, 2008, pp. 1133-1142.

[12] Li Y, Liao T and Lai C. A social recommender mechanism for improving knowledge sharing in online forums. *Information Processing and Management* 2012; 48(5): 978-994.

[13] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the 14th international joint conference on Artificial intelligence, 1995*, pp. 448-453.

[14] Gurevych I. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the Second international joint conference on Natural Language Processing (IJCNLP'05)*, 2005, pp. 767-778.

[15] Zesch T, Müller C and Gurevyc I. Using wiktionary for computing semantic relatedness. In: *Proceedings of the 23rd national conference on Artificial intelligence*, 2008, pp. 861-866.

[16] Navarro E, Sajous F, Gaume B, Prévot L, ShuKai H, Tzu-Yi T, Magistry P and Chu-Ren H. Wiktionary and NLP: improving synonymy networks. In: *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web '09)*, 2009, pp. 19-27.

[17] Snow R, Jurafsky D and Ng A. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems (NIPS) 2005*; 17:1297–1304.

[18] Sang E and Hofmann K. Lexical patterns or dependency patterns: which is better for hypernym extraction?. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL '09)*, 2009, pp. 174-182.

[19] Panchenko A, Morozova O and Naets H. A semantic similarity measure based on lexico-syntactic patterns. In: *Proceedings of KONVENS 2012*, 2012, pp. 174–178.

[20] Zhang J, Ackerman M and Adamic L. Expertise networks in online communities: structure and algorithms. In: *Proceedings of the 16th international conference on World Wide Web (WWW '07)*, 2007, pp. 221-230.