LETTER InfluenceRank: Trust-Based Influencers Identification Using Social Network Analysis in Q&A Sites

GunWoo PARK^{†a)}, Student Member, SungHoon SEO^{††*b)}, SooJin LEE[†], and SangHoon LEE^{†c)}, Nonmembers

SUMMARY Question and Answering (Q&A) sites are recently gaining popularity on the Web. People using such sites are like a communityanyone can ask, anyone can answer, and everyone can share, since all of the questions and answers are public and searchable immediately. This mechanism can reduce the time and effort to find the most relevant answer. Unfortunately, the users suffer from answer quality problem due to several reasons including limited knowledge about the question domain, bad intentions (e.g. spam, making fun of others), limited time to prepare good answers, etc. In order to identify the credible users to help people find relevant answer, in this paper, we propose a ranking algorithm, *InfluenceRank*, which is basis of analyzing relationship in terms of users' activities and their mutual trusts. Our experimental studies show that the proposed algorithm significantly outperforms the baseline algorithms.

key words: knowledge sharing, question answering, social network analysis, identifying influencers

1. Introduction

Basically, people not only seek specific information but share subjective opinions and suggestions on their problemsolving and decision-making tasks based on closeness and trustfulness [1], and might be influenced by credible users more than by others [2]. Influence is defined as 'the act or power of producing an effect without apparent exertion of force or direct exercise of command' or 'the power or capacity of causing an effect in indirect or intangible ways' [3]. In this paper, we propose the InfluenceRank algorithm to identify credible users, to answer a given target question. InfluenceRank is basis of users' interactions in a Q&A site (e.g., Naver's Knowledge iN and Yahoo! Answers**), and we assess activity and trust which are considered as main factors to measure Influence value. We define a credible user as a trust-based active user who has high Influence value. Our assumption is that credible users issue relevant answers. Researchers in various fields [4]-[6] have applied Social Network Analysis (SNA) method to different types of networks. To measure Influence value, in this paper, we analyze characteristics of Q&A networks by using SNA.

Manuscript received February 24, 2011.

Manuscript revised March 23, 2012.

[†]The authors are with the Department of Information System, Korea National Defense University, Seoul, Korea.

a) E-mail: pgw4050@emerald.yonsei.ac.kr

- Motivation: Although recent commercial search engines use the information concerned with Question and Answering (Q&A), it is still difficult to acquire an appropriate content from numerous user-described answers in Q&A sites. Therefore, discovering credible users became a very hot issue in Q&A sites. One obvious method to find relevant answers is to return the best answers among answers returned by the answerers. By directly returning answers to a user, we hope to reduce the efforts required to locate good answers. This hopefully will also reduce the need to post duplicate questions in the Q&A site. The research [7], [8] on answer quality predicting problem has focused on the textual features or non-textual features of answers (or documents) in Q&A sites. However, so far, there is no metric to predict the quality of answers based on the users in Q&A sites. Our work is different from previous efforts in that we focus on not directly estimating the quality of answers but finding credible users for helping people find relevant answer in an open-domain community-based Q&A site.
- **Objective:** The goal of our work is to investigate how the expertise of users, or credible users, can be captured, and when combined with state-of-the-art information retrieval techniques whether the system is able to identify the group of credible users who are likely to provide relevant answers to given questions.

The rest of this paper is organized as follows: We demonstrate the factors to measure influence value, and then describe our proposed *InfluenceRank* algorithm in Sect. 2. Experimental results demonstrating the performance of our method are shown in Sect. 3. Finally, we draw conclusions in Sect. 4.

2. Proposed Algorithm

2.1 User Interaction-Based Network Construction

As shown in Fig. 1, a user posts questions, and answers or user feedback (*e.g.*, select, recommend) in the Q&A sites. A user can post multiple questions simultaneously, and there can be one or more answers in a question. The questioner

^{††}The author is with the Department of Computer Science, Columbia University, New York, USA.

^{*}Presently, with the Advanced Institute of Technology, KT Corp., Seoul, Korea.

b) E-mail: hoon@cs.columbia.edu

c) E-mail: pgw4050@kndu.ac.kr

DOI: 10.1587/transinf.E95.D.2343

^{**}http://kin.naver.com includes more than 80 million Q&A information in January 2008, and Yahoo benchmarked it for their initial launching of Yahoo! Answers service.



(Questioning, Answering and Recommending)

can choose the most relevant answer as the best one. We regard users as nodes, and their actions as edges. We can express the QAR (Questioner-Answerer-Recommender) network like Fig. 1 (b) based on the user interactions (see Fig. 1 (a)). We consider the directed graph G(V, E). Nodes in V represent the users, and each directed edge in E indicates relationship. As an example, user1 answers the question of user3, and comments on or votes the answer from user4.

2.2 Activity Measurement

We define a user's activity in a QAR network as the frequency of producing good quality question, sharing the best/good answer 'selected' or 'recommended' by other users. The activity can be measured by the number of the user behavior, such as the number of answers in own questions or his/her selected answers of other users' questions. The activity value of an i-th user, $ACT(u_i)$, is given by

$$ACT(u_i) = \frac{m' + \sum_{p=1}^{m} A_p}{m} + \frac{n' + \sum_{q=1}^{n} (R_q + S_q)}{n}$$
(1)

where *m* and *n* are the total number of questions and answers posted by i-th user, respectively. *m'* and *n'* are the total number of questions in which answers are posted and answers which is recommended or selected as a good answer, respectively. A_p is the number of answers per each question posted by i-th user. R_q and S_q are the total recommended and selected number over the each i-th user's answer by the answerer and the neighbor users, respectively.

2.3 Trust Measurement

We define a user's trust in a QAR network as the credibility of the knowledge based on the user interaction (*e.g.*, 'selecting' or 'recommending' the best/good answer). For measuring trust, we use the analysis of degree centrality which is frequently used when measuring authority and influence value [4]–[6]. Considering a graph $G := (V, E_A)$ and E_A the set of the directed connections $E = (e_{1,1}, e_{1,2}, \cdots, e_{i,j})$ between the users of the set V then the in-degree centrality, K_i^{in} , and out-degree centrality, K_i^{out} of a vertex V_i are the sum of the inbound connections and outbound connections to that vertex, respectively [4]–[6].

$$K_i^{in} = \sum_{j=1}^{n} e_{j,i}$$
 and $K_i^{out} = \sum_{j=1}^{n} e_{i,j}$ (2)

 $e_{i,j} = 1$ if there is an inbound link between i-th user and j-th user, and $e_{j,i} = 1$ if there is an outbound link between i-th user and j-th user with i-th user in the center. The indegree centrality and the out-degree centrality of a vertex make sense only in cases where a directional relationship is available. However, there exist reciprocal connections in the QAR network. In that case, the degree centrality is computed by the influence domain of the vertex. For a non directed graph $G := (V, E, \overline{E})$, the influence domain of a vertex V_i is the number or proportion of all other vertices which are connected by a path to that particular vertex [4]–[6].

$$\overline{d}_i = \frac{1}{N-1} \sum_{j=1} \overline{e}_{j,i} \tag{3}$$

On the above measure *E* represents the set of paths between the vertices V_i and V_j and N - 1 is the number of all available nodes in the QAR network. The total number of nodes N = |V| minus the node that is subject to the metric. According to these centralities in terms of prestige, we define indegree centrality and out-degree centrality of i-th user as a combination of the above two metrics. The trust indicator of a question of i-th user, $TQ(u_i)$, the in-degree centrality measured from the number of answers, and the trust indicator of an answer, $TA(u_i)$, the out-degree centrality, measured by the frequency of best-answer choice or good-answer recommendation. The degree centrality of i-th user encompasses the normalization of the in-degree and out-degree of the user by its degree of influence, which are given by [5]–[7]

$$TQ(u_i) = \frac{K_i^{in}}{\overline{d}_i}$$
 and $TA(u_i) = \frac{K_i^{out}}{\overline{d}_i}$ (4)

2.4 InfluenceRank Algorithm

We focus on the relationships between users, *i.e.*, that of a questioner whose question is posted an answer by an answerer, and an answerer whose answer is recommended as a good or best answer by questioner or neighbor users. As shown in Fig. 1 (b), the degree centrality in the QAR network can be used as a very meaningful factor in terms of credibility. Figure 2 shows the factors (*e.g.*, activity, trust) and approach for calculating influence value in a Q&A site. The more give an answer to an i-th questioner's question and the more questioners select or neighbor users recommend an answer as a best or good answer, the more trust-based influence increases. The *InfluenceRank* algorithm of i-th user, *IRank*(u_i), can be defined as follows:

$$IRank(u_i) = \alpha \left[\frac{m'}{\sum_{p=1}^{m} A_p} \times TQ(u_i) \right] + (1 - \alpha) \left[\frac{n'}{\sum_{q=1}^{n} (R_q + S_q)} \times TA(u_i) \right]$$
(5)

where m and n are the total number of questions and answers posted by i-th user in terms of activity. m' and n' are the total number of questions in which answers are posted and

Evaluation Criteria of Answer's Relevancy	Editor's Vote	Relevancy Level	Relevance Score (rel)
S1: Objective with certain basis (theoretic/scholarly source)	(S1 & S2 & S3)	Suitable ⁺	3
S2: Subjective but logically explained	(S1 & S2) or (S1 & S3) or (S2 & S3)	Suitable ⁰	2.5
S3: Complete, broad, well-organized description of the question	Only S1 or S2 or S3	Suitable ⁻	2
C1: Insufficient source	(C1 & C2 & C3)	Common ⁺	1.5
C2: objective but lack of details	(C1 & C2) or (C1 & C3) or (C2 & C3)	Common ⁰	
C3: Subjective with no basis but partially logical	Only C1 or C2 or C3	Common ⁻	0.5
U1: Very speculative or subjective with no basis			
U2: Abuse languages or spams contained	As meeting one of the evaluation criteria		
U3: Libel on someone particular, irrelevant answer to the question	(e.g., U1, U2, U3, U4, U5)	Unsuitable	0
U4: Contains very little, misleading information			
U5: Many inappropriate words, wrong spellings, or bad readability			





 (a) u_i's activity: the number of question, answer and recommendation, out-degree centrality



• a'l: The answer which is selected as a best-answer and recommended as a good answer • a'2: The answer which is recommended as a good answer

(b) *u_i*'s trust in terms of answer: the number of answers which is selected as a best-answer or recommended as a good answer, in-degree centrality



(c) *u_i*'s trust in terms of question: the number of questions in which answers are posted, in-degree centrality

Fig.2 Factors for calculating u_i 's influence value in a Q&A site.

answers which is recommended or selected as a good answer by neighbor users in terms of trust-based activity. A_p is the number of answers posted by other users per question, S_q is the total number of answers selected as a best-answer by corresponding user, and R_q is the number of the answer which is recommended as a good answer by other users per answer. $TQ(u_i)$ and $TA(u_i)$ are in-degree centrality and outdegree centrality in terms of question and answer, respectively (see Eqs. (2) and (4)). α is a balance parameter where $0 \le \alpha \le 1$.

3. Experimental Analysis

We evaluate our algorithm's effectiveness by using Normalized Discounted Cumulative Gain (NDCG) [10]. NDCG uses the observation that most of Web users refer to search results in top ranks, then counts relevance score (*rel*) of each ranks to discriminate their differences and make better ranking function. In order to apply NDCG, we need to check *rel* of each answer. 7 point measurement for *rel* is used to evaluate answers posted by users. Table 1 shows descriptions of evaluation criteria and seven-level of answer quality (*e.g.*, Suitable⁺, Suitable⁰, Suitable⁻, Common⁺, Common⁰, Common⁻ and Unsuitable). 50 editors were asked for handtagging each given answer as one of the evaluation criteria (*e.g.*, S1-S3, C1-C3, U1-U3) for *rel*. Higher NDCG@*k* score means the *InfluenceRank* algorithm shows better performance through identifying more credible users, called influencers.

- Baseline: We set three baseline methods (*e.g.*, HITS [7], PageRank [11], Point System which is now used in Naver's Knowledge iN [12]).
- Datasets: User interactions were collected from March 2009 to September 2011 in the Naver's Knowledge iN. We collected 18,372 users and 415,882 Q&A pairs from the 'Sports' category, due to 'Forum' type network which is useful to explain mutual activities among users. We use 291,117 Q&A pairs (70%) as training set to adjust parameter α, and 124,765 pairs (30%) as test set to show our algorithm's effectiveness. We organize three test sets which have similar ratio of answer(s) per question. Interestingly, more than 80% of Q&A data is in 3 to 5 answer cases, which means that we need to analyze the 3~5 answer cases as the important Q&A material.
- Parameter Estimation: Parameter was adjusted to 0.17 by Backpropagation Neural Network.

3.1 Structural Analysis of Q&A pairs

To confirm relevancy of category selected for experiment and measure Influence value, we analyze a sample networks using user connections, questions, answers and recommendation, then assess their rankings based on link analysis. To find efficient network structure for our study, we try to cluster main categories of Naver's Knowledge iN by using SNA. To analyze the structure of given user network, we select 50 users in each category. User connections based on their questions and answers are analyzed using UCINET, and then projected to 3 types of ego network [13]. Table 2 shows the clustering result, and Fig. 3 shows example of network diagrams.

Cluster Categories in Naver's Knowledge iN Factuality Computer/Communication, Game, Business/Economy Advice Shopping, Health/Medical, Home/Life, Education/Study Sports, Entertainment, Society/Culture, Travel/Leisure Discussion (b) Advice (a) Factuality (e.g., Game) (e.g., Shopping) (e.g., Sports) Fig. 3 Naver's Knowledge iN's ego network diagram example. ■ HITS ■ Point System ■ InfluenceRank 1 1 0.8

Clustering result.

Table 2



Fig. 4 Evaluation results by Average NDCG@k of three test data sets.

3.2 Evaluation Results

Figure 4 shows the evaluation results. Average improvements of NDCG@k show 14.36~19.94% over PageRank, 12.08~18.47% over HITS and 11.23 16.99% over Point System. Especially InfluenceRank shows the highest performance in the case of questions with 3 to 5 answers which constitute major part of test data set (82.7~93.8%) of data set). Average improvements of NDCG@3-5 are 18.32~32.41% over PageRank, 16.72~30.61% over HITS and 15.86~28.54% over Point System. Improvements of Maximum improvements are shown in NDCG@4, 28.4~33.59% over PageRank, 26.6~31.96% over HITS and 25.1~30.77% over Point System. Minimum improvements are shown in NDCG@10, 5.12~5.60% over PageRank, 4.93~5.24% over HITS and 3.98~4.15% over Point System. As a result, InfluenceRank cases signifycantly outperform all baseline methods in all cases Page Rank (12.15% average, 33.59% max in NDCG@4), HITS (15.28% average, 31.96% max) and Point System (13.96% average, 30.77% max).

In view of HITS, askers can be regarded as hubs, and answers who were chosen as the best can be authorities. Note that HITS is originally designed to rank nodes from directed non-weighted graphs. However, interactions between users in a QAR network can be assumed as weighted graph structure. Using HITS or PageRank on a QAR network may ignore a crucial information type: the magnitude of interaction between users. Therefore, HITS and PageRank could not be used to accurately identify credibility-based influencers. Though Point System uses trust factor, which relies on user voting results based on its popularity, but it still cannot filter off awarded points abnormally. *InfluenceRank* algorithm strengthens the trust factor by using social network structure within a Q&A site, and it shows better performance.

4. Conclusion

In this paper, we proposed *InfluenceRank* algorithm to measure influence value of users in Q&A sites. To measure influence value in terms of credibility, we analyzed trustbased social network according to user interactions related to questioning and/or answering. In our experiment, effectiveness of *InfluenceRank* by using NDCG@k based on editors' judgments shows significant improvements in comparison with baseline methods. We confirm that credible users' answers are more relevant than those of ordinary users' through the experiment. Therefore, it is possible to enhance search effectiveness by identifying and then recommending credible users to users or letting rank answers of credible users higher than other ordinary users by our algorithm.

References

- M. Subramani and B. Rajagopalan, "Knowledge-sharing and influence in online social networks via viral marketing," Commun. ACM, vol.46, no.12, pp.300–307, 2003.
- [2] J. Zhang, M. Ackerman, and L. Adamic, "Expertise networks in online communities: Structure and algorithms," Proc. 16th International Conference on World Wide Web, p.230, ACM, 2007.
- [3] http://www.merriam-webster.com/dictionary/influence
- [4] D. Fisher, M. Smith, and H. Welser, "You are who you talk to: Detecting roles in usenet newsgroups," Proc. 39th Annual Hawaii International Conference on System Sciences, 2006, HICSS'06, p.59b, IEEE, 2006.
- [5] J. Bollen, H. Van de Sompel, J. Smith, and R. Luce, "Toward alternative metrics of journal impact: A comparison of download and citation data," Information Processing & Management, vol.41, no.6, pp.1419–1440, 2005.
- [6] X. Liu, J. Bollen, M. Nelson, and H. Van de Sompel, "Co-authorship networks in the digital library research community," Information Processing & Management, vol.41, no.6, pp.1462–1480, 2005.
- [7] J. Kleinberg, "Authoritative sources in a hyperlinked environment," J. ACM (JACM), vol.46, no.5, pp.604–632, 1999.
- [8] Y. Zhou and W. Croft, "Document quality models for web ad hoc retrieval," Proc. 14th ACM International Conference on Information and Knowledge Management, pp.331–332, 2005.
- [9] L. Hoang, J. Lee, Y. Song, and H. Rim, "A model for evaluating the quality of user-created documents," Information Retrieval Technology, LNCS, vol.4993, pp.496–501, 2008.
- [10] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," ACM Trans. Inf. Syst. (TOIS), vol.20, no.4, pp.422– 446, 2002.
- [11] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Technical Report, Stanford Digital Library Technologies Project, 1998.
- [12] K. Nam, M. Ackerman, and L. Adamic, "Questions in, knowledge in?: A study of Naver's question answering community," Proc. 27th International Conference on Human Factors in Computing Systems, pp.779–788, ACM, 2009.
- [13] L. Adamic, J. Zhang, E. Bakshy, and M. Ackerman, "Knowledge sharing and yahoo answers: Everyone knows something," Proc. 17th International Conference on World Wide Web, pp.665–674, ACM, 2008.