Expert Finding Systems in Question Answering Forums using Graph Theoretical Approach

G Madhu Kiran, K Rounak Reddy, Darisi Dileep Kumar,
Bindu K. R. and Latha Parameswaran

[#] *Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, India*
E-mail: mkgmadhu1994@gmail.com, rounak4991@gmail.com, dileepdarisi94@gmail.com, j_bindu@cb.amrita.edu, p_latha@cb.amrita.edu

***Abstract*— Question answering websites are one of the platforms for sharing knowledge, but to make it more effective it is important to ensure that the asker gets the perfect answer to his question within an optimal amount of time. To attain this it is important for us to suggest the list of experts in the question domain so that the asker can directly post the queries to the expert which minimizes the amount of time it usually takes and improves the quality of the answer the asker gets. In this paper, we use a novel hybrid approach to recommend a list of experts for the asker's question by calculating the knowledge score and the authoritative score which helps in deriving the expert score for each user.**

***Keywords*— Question-answering Forums, Question domain, Expert score.**

I. INTRODUCTION

The web contains massive amounts of data which is stored in the form of documents, hyperlinks, databases, etc. Although we have a lots of search engines like Google, Bing, etc., which can be used to retrieve required information, but the information returned by the search engine is based on keyword matching which gives us a large set of results which are in the form of Hyperlinks that may contain relevant or irrelevant information. This process consumes a lot of time searching for the information required. Thus, people have started using a question-answering forums to get a simple and tailored solution for their queries.

Question-answering forums is a platform where users post questions to get solutions from other users or browse through the already existing questions and answers given by users. But sometimes the user may not be able to get the appropriate response or won't get any response for a long time which makes it unreliable. Therefore, it is necessary to identify a list of experts for an incoming query so that the user can directly post the query to the set of experts who can provide a precise solution within an optimal amount of time.

The methods mentioned above have their pros and cons. When we use the authoritative scores given by the hits algorithm or page rank algorithm we get a set of experts who would answer the query quickly, but these users may not have good knowledge in the particular domain. And when we use the knowledge scores derived from knowledge profile and user reputation the experts suggested by this score have great knowledge in the domain of the query, but they may not answer the query in a stipulated amount of time.

The methods mentioned above have their pros and cons. When we use the authoritative scores given by the hits algorithm or page rank algorithm we get a set of experts who would answer the query quickly, but these users may not have good knowledge in the particular domain. And when we use the knowledge scores derived from knowledge profile and user reputation the experts suggested by this score have great knowledge in the domain of the query, but they may not answer the query in a stipulated amount of time.

The task of finding the expert is done using a method called social network analysis. It is a process of analysing the social network through networks or graphs. It is the most common method used in tasks like suggesting friends in Facebook, twitter etc. The conventional method to find an expert is to build a social network and then calculate the authority score for each user based on an algorithm.

Various methods exist for finding the experts in question answering forums. Larry Page et al. [1] proposed PageRank algorithm to rank the experts. A page is said to have more importance than other if it is linked by many pages with high importance. Jurczyk et al. [2] used HITS algorithm to identify good authorities and hubs for a given query by calculating the values like authority weight and hub weight. The hub weight is the aggregate of authoritative scores of the pages that are contained in a page. The authority weight is the sum of hub weights of the page that is pointed by various other pages. Zhang et al. [3] proposed an algorithm called Expertise Rank which is a modification of PageRank algorithm. The algorithm considers both the factors like how many other people one helped, whom he/she helped in ranking the experts present in the network.

Expert finding systems are evaluated using these methods namely Precision, MAP and MRR [4],[5],[6],[9].

Precision (P) is the ratio of retrieved and relevant documents related to a particular query to the count of entire documents that are retrieved.

$$Precision = \frac{\#(A\&B)}{\#(B)} \quad (1)$$

Where A = relevant items and B = retrieved items

MRR is the ratio of summation of the reciprocal ranks to the total number of queries

$$\text{MRR (Q)} = \frac{1}{|TQ|}\sum_{i=1}^{|TQ|}\frac{1}{r_i} \quad (2)$$

Where MRR is the Mean Reciprocal Rank, TQ is the total number of queries, r_i is the rank of the first relevant document returned by a particular query

MAP is the ratio of the average of the precision to the total number of queries

$$\text{MAP (Q)} = \frac{1}{|TQ|}\sum_{j=1}^{|TQ|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision\ (R_{jk}) \quad (3)$$

Where MAP is the Mean Average Precision, TQ is the set of queries, R_{jk} is the set of ranked retrieval results for a particular document d_k.

II. METHODOLOGY

A. *Framework*

Our proposed architecture diagram uses knowledge score and authoritative score to recommend a list of experts for a given query as shown in the Fig. 1.

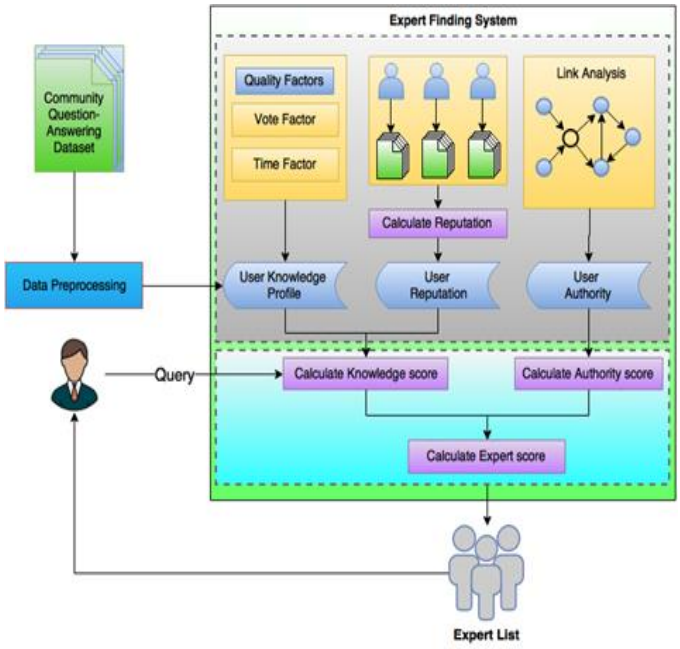

Fig. 1 Architecture Diagram

By considering the type of the given target question various priorities are assigned to the tokens which we get from the tokenizer is used for deriving the user knowledge profile based on the weights. The user knowledge profile is derived from certain quality factors like vote factor and time factor. The knowledge score for a particular user is derived from the knowledge profile which is compared to the given target question and then we consider additional factors like user reputation which is determined by the user's previous question answering records. By using PageRank algorithm we find the authoritative score of each user. For recommending an expert for a given target question we combine the user's knowledge score and authoritative score to produce a list of experts for the given particular query.

B. *PreProcessing*

The dataset used is a Yahoo Answers dataset. It consists of questions and answers posted by various users. Each question may have multiple responses from various users. First, we extract the data from the dataset and we preprocess the data. Preprocessing involves the tasks like removing the stop words. The question and answer pair for a user is analyzed using the TF-IDF approach [5] where high priority terms are extracted using the tokenizer package of NLTK. Then theses relevant terms are stemmed using the Potter Stemmer. Finally, we find the frequency of relevant terms. The processed terms and their frequency forms the knowledge profile of a user.

C. *Knowledge Profiles*

The knowledge profile [7] for a user in particular category is defined as

$$\text{KP} = \frac{\sum_{q_r} TV_{u_a,q_r} * V_{u_a,q_r} * T_{q_r}}{|C_{u_a}|} \quad (4)$$

Where q_r is the target question that corresponds to a category, u_a is the answer given by the user, V_{u_a,q_r} is the vote factor which is acquired from the number votes given to the user's answer for a particular question, T_{q_r} is the time factor, TV_{u_a,q_r} is the term vector of the Question and its answer given by the user and C_{u_a} is the set of queries answered by the user.

The vote factor for a user is calculated using the formula shown below

$$V_{u_a,q_r} = \frac{\#\,(votes\ to\ user\ QA) + \lambda}{\#\,(votes\ to\ all\ answers) + |TA^{q_r}| * \lambda} \quad (5)$$

Where $|TA^{q_r}|$the number of answers given to a particular question, λ is a correction factor for answers with zero votes. Generally, value of λ = 0.1 is considered when the answers did not receive any votes.

Time factor [8] for a question answer pair is calculated using the equation shown below

$$T_{q_r} = e^{-\delta(ct_{now} - t_{q_r})} \quad (6)$$

Where ct_{now} is the current date time, t_{q_r} is the time when the query is asked and δ is the tuneable parameter with a value of 1/365.

III. EXPERT RECOMMENDATION

A user knowledge score represents how much knowledge does a user has related to a particular query and how accurate he is in answering to the questions posted in that particular category. A user authority score determines how much the user is actively involved in answering to the questions posted. The user's expert score for a given target query is calculated using

Expert Score = μ* $KS_{u_a,q}$ + (1- μ) * $AS_{u_a,q}$ (7)

Where μ is a variable that controls the balance between knowledge score and authority score, $KS_{u_a,q}$ is the knowledge score and $AS_{u_a,q}$ is the authority score.

A. *Knowledge Score*

Knowledge score is derived from the user knowledge profile and user reputation. The knowledge score for a particular user and the target query is calculated by

User reputation = $\frac{\#\,(answers\ voted\ as\ best\ answers)}{\#\,(total\ answers\ posted\ by\ user)}$ (8)

Knowledge score = β * Sim (KP, TQ) + (1-β) * user reputation (9)

Where KP is Knowledge Profile, TQ is Target Question, β is a parameter which restores the balance between knowledge profile and user reputation, Sim (KP, TQ) is the cosine similarity of knowledge profile and the target query. The value of β varies from 6.5 to 8.5 depending on the relative importance between cosine similarity and user reputation.

B. *Authoritative Score*

PageRank algorithm. The users in the social network and their interactions are modelled as graph. To implement the graph we use Neo4j which is a highly scalable graph database. The graph consists of two elements called a node and a relationship. Each user is represented in the form of a node. The relationship between two nodes is represented in the form of an edge as shown in the Fig. 2.

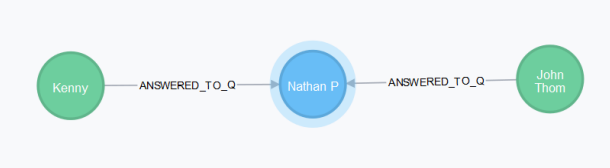

Fig. 2 Representing each user and the relationships between them in a graph

In PageRank algorithm the frequency for a particular page is calculated using the equation shown below.

$$PR\,(a_n) = \frac{1-d}{N} + d\sum_{a_o \in M(a_n)} \frac{PR\,(a_o)}{L(a_o)} \quad (10)$$

Where $a_1, a_2, a_3 \ldots a_n$ are the used pages, $M(a_n)$ is a set of pages that are linked to a_n, $L(a_o)$ is the set of tasks performed by page a_o ,d is the parameter whose value is taken as 0.85 and N is the page count.

IV. EXPERIMENTAL FINDINGS

Our dataset is taken from Yahoo Answers forum about Internet and its related queries. This dataset is used to find the values of variables - μ and β. Various evaluation techniques exist to validate the expert but it is advisable to validate the expert manually with the help of technical raters for a given target query. We then use various values for μ to derive the final expert score. The performance is maximized by setting μ = 0.85 and β = 0.8.

A. *Comparison of Different Models*

In this section, we compare the performance of our approach with other conventional modules through 50 target questions, and the results are shown in Fig. 3.We observed that the PageRank and HITS perform worse than the K_score and Hybrid model. Hence, K_score and Hybrid model are only suitable to find experts for a target query. From the results, we can also observe that the proposed Hybrid method that combines both the K_score, which considers user knowledge profile and user reputation, and the A_score, which is computed by PageRank algorithm, performs better than any other methods.

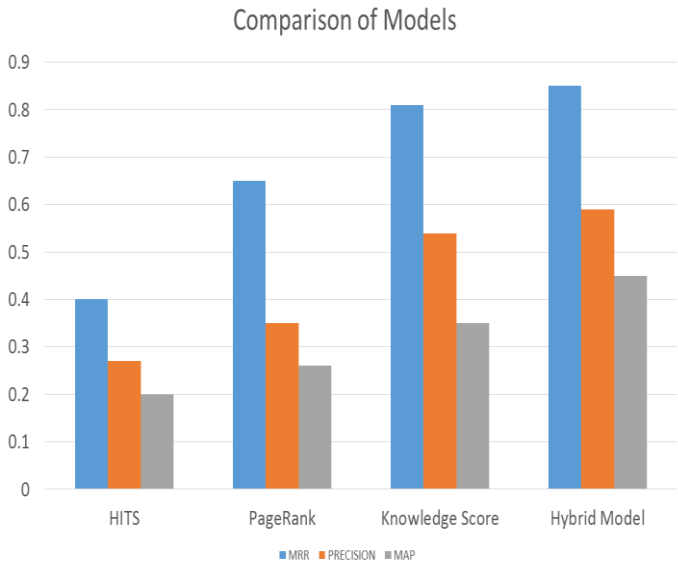

Fig. 3 Comparison of Models

V. CONCLUSION

In this paper, we use a novel hybrid approach to suggest some experts for a given target question and represent it in a graphical approach. The knowledge profile of a user indicates how much knowledge does a user has related to the query posted by the other user this involves the calculation of measures like vote factor, time factor. By considering the user reputation we can improve the knowledge score and considering the subject knowledge of a user to the target question in expert finding. The authority score is calculated by the link analysis which gives the participation of user in answering to a question based on the previously answered target question in a category. Furthermore, the knowledge score which is derived by the summation of user knowledge

profile and user reputation gives more efficient results compared to the results that are derived using only the authoritative score which uses PageRank or HITS algorithm.

ACKNOWLEDGMENT

We express our sincere thanks to Dr. Latha Parameswaran, Professor and Chairperson, Department of Computer Science and Engineering, for her support and encouragement.

We extend our gratitude to our guide Ms. Bindu K R for her valuable guidance. We would like to extend our gratitude to the Panel members Dr.Vidhya Balasubramanian, Mr.Arun Kumar C, Mr.Raghesh Krishnan K and Ms.Divya M for their extensive help and rigorous assessments which have helped us in achieving our objectives. We would also thank our project coordinators for their co-operation. We are grateful to all faculty members for their valuable guidance.

REFERENCES

[1] Page, L., Brin, S., Motwani, R. and Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1998.

[2] Jurczyk, P., and Agichtein, E. HITS on Question Answer Portals: Exploration of Link Analysis for Author Ranking. *Proc. of the SIGIR'07*, Amsterdam, the Netherlands, 2007.

[3] Jun Zhang, Mark, S. Ackerman and Lada Adamic. Expertise Networks in Online Communities: Structure and Algorithms. International World Wide Web Conference Committee (IW3C2), Canada, 2007.

[4] Macdonald, C., Hannah, D., and Ounis, I. High Quality Expertise Evidence for Expert Search. *Proc. of the 30th European Conf. on Information Retrieval*, 2008, LNCS, vol. 4956, pp. 283–295.

[5] Liu, X., Croft, W. B., and Koll, M. Finding Experts in Community-Based Question-Answering Services. *Proc. Of the CIKM'05*, Bremen, Germany, 2005.

[6] Serdyukov, P. and Hiemstra, D. Modeling Documents as Mixtures of Persons for Expert Finding. *Proc. of the 30th European Conference on Information Retrieval (ECIR 2008)*, Glasgow, UK, 2008, LNCS, vol.4956, pp. 309-320.

[7] Wei-Chen Kao et.al. Expert Finding in Question-Answering Websites. SAC'10. Sierre, Switzerland, 2010.

[8] Zhang, J., Ackerman, M.S., Adamic, L., and Nam, K.K.QuME: A Mechanism to Support Expertise Finding In Online Help-seeking Communities. Proc. of the UIST'07, Newport, Rhode Island, USA, 2007.

[9] Bindu, K.R., Nambiar, S.R., Chandran, J., Latha Parameswaran., Performance evaluation of algorithms for expert finding on an open email dataset (2015) International Journal of Applied Engineering Research, 10 (73), pp. 71-76.