Expert Finder on social networks based on Ontological Learning

Sharadchandra solunke[1],Nitin babaji khatate[2] ,Govinda sakharam chaudhari[3]
[1]Student, IT Department, PREC Loni, Maharashtra, India
[2]Student, IT Department, PREC Loni, Maharashtra, India
[3]Student, IT Department, PREC Loni, Maharashtra, India

ABSTRACT

In the digital age, there are multitude of sites that offer content of varied proportions and aspects which need to mined. The data can be any form with generalized data as well as specialized data. Websites like Facebook and LinkedIn generate a lot of data which can be classified as general knowledge, but on other hand websites like quora, twitter and stack exchange produce a lot of data which can aid in the process of expert finding. The process of finding an expert is based on two main facets- the knowledge base available on the subject and the similarity between the knowledge shared by an individual in correlation with the original knowledge base. The mapping of data is to be done using ontological learning process. Along with this process, the link analysis approach fosters the task of finding experts.

Keywords:- *Expertise finding, information retrieval, social network, crowdsourcing, semantic analysis, ontological learning, knowledge extraction.*

Introduction:

With the advent of internet in our personal space, people have started to opt on internet for every advice they need. This is the sole reason why communities like Quora, stack exchange have seen a great boom. What is more challenging aspect in this whole equation is finding the right expert for the right advice. Most of the top notch websites offer a expert tag to the premium users but they might not be the well versed experts for that matter. So what is necessary is to have a knowledge base of basics of a given subjects and the experts should be extracted only from the individuals who offer an advice to a greater degree of similarity with the existing knowledge base. This approach would work perfectly fine but as soon as the data provided by the user goes out of the context of the knowledge base, it is rather cumbersome to analyse or predict if an user is actual expert or not. Therefore newer approach also involves the process of link analysis. The link analysis is an important aspect here which provides much better accuracy and credential tags associated with a person. On social networks like LinkedIn, people endorse other people for some specialized skills. This skills are tagged by individuals who have worked in collaboration in some task or the other. In online communities, great volume of data related to queries posted by individuals is yet another important task that makes queries unseen by experts who have the capability to respond to them. Therefore, the average response time for a legitimate expert for responding to questions takes a bit longer. By using the expert finding methodology and making recommendation systems based on these methodologies, the questions can be delivered to individuals who are actually apt to respond them. Also, it is possible to discard naive questions from being visible to experts; so they won't waste their time and energy responding to those queries. The common core along all the expert finding methodologies is the use of ontological learning approach. The ontological learning is the process of analysing the characteristics of a particular object and mapping of those characteristics into behaviours of individual elements. Here in the expert finding problem, a particular set of skills are extracted from ontological profiling of individuals over a social network. These skills are ranked based on the index points that they are generated. The ontological learning is a spectral model for the statistical approach of characterisation of data and correlating them with individuals or objects. The proposed system will work on concept mapping, djikstra's algorithm and ranking for the skills.

We consider the task of ranking the participants of an online community according to their corresponding level of knowledge of a given topic; after generating these rankings, the top-x experts are shortlisted. These experts can actually a match variety of different needs, right from cross spanning to providing recommendations of certain products and subjective people and places; or performing trivial tasks. The classical approach to these kind of problems consists of basic profiling of the group members and performing matches of textual queries against such given set of profiles, and ranking the members according to the matches. However, the profile information in almost all of the existing social networks is quite limited, as most members of a social network just provide enough amount of information which is mandatory for registration purposes, and do not explicitly state their interests or skills in certain set of activities.

RELATED WORK:

In [1], the generalized approach which was previously aforementioned in the introduction is used. Here all the data provided at a single point of registration is given. This data needs to be updated from time to time and the reflection of changes requires time as they training and knowledge base is extracted only once. The data extraction is carried once and the changes reflect only during the point of training. Also maintenance of this methodology is cumbersome.

In [2], the approaches enlisted in at the point of time yield good results and they use Spearman rank coefficient for the accuracy of their matches. The proposed system of our project also is built on same pretext. The spearman coefficient is important and is generally integrated as a correlation factor for the managed correlation mapping between knowledge and expertise of a person. This is a healthy approach and our approach extends the same prophecy.

In [3], the crowd sourcing approach is designed based on a modified version of travelling salesman problem and is subject to all the pitfalls of that algorithms. The knowledge base required in this approach is extremely low, but the end results are not up to the mark. But the system works at great speed and excluding the deadlock issues which might occur once in a blue moon, the system is actually highly efficient.

Methodology:

Our proposed system has four core modules:

a) Information Extraction
b) Content Analysis
c) Social Network Analysis
d) User Ranking and Experts Finding

Now for instance let's assume “A" is a user of the online community where questions will be answered. So the first step here is extraction of users profile data by a crawler. Now the generalized criteria for crawler to check user is if the total number of posts by the certain user was higher than a given substantial limit then the user is considered as a candidate for the ranking and is further verified for being an expert. The crawler will deliver a set of user profiles as output. This profile will be then used as an input to another crawlers. The second crawler here gets the detailed information over user's posts which includes the questions and corresponding answers of the user. The second step as the most crucial step, extracts concepts out of user's posts using the concept map for lookup. After that, the distance between concepts given in the answer is subjected to a computation based on concept map, for calculating the distance we have used Dijkstra's algorithm in our system. The idea is to calculate the weight of the concepts in the user's answers based on the ideologies related to the question. Thereby we reach to a user's final score based on weighted concept.

In the next step, we create a network of all the users with respect to the questions and the answers they have been asked and the ones they have answered. Finally a ranking algorithm computes rank of the user in perspective with the field in question. The user's score is available at the end of the computation in the whole network. This network keeps on changing based on content analysis and semantic analysis. The users knowledge base to answer resemblance ration is a key factor here.

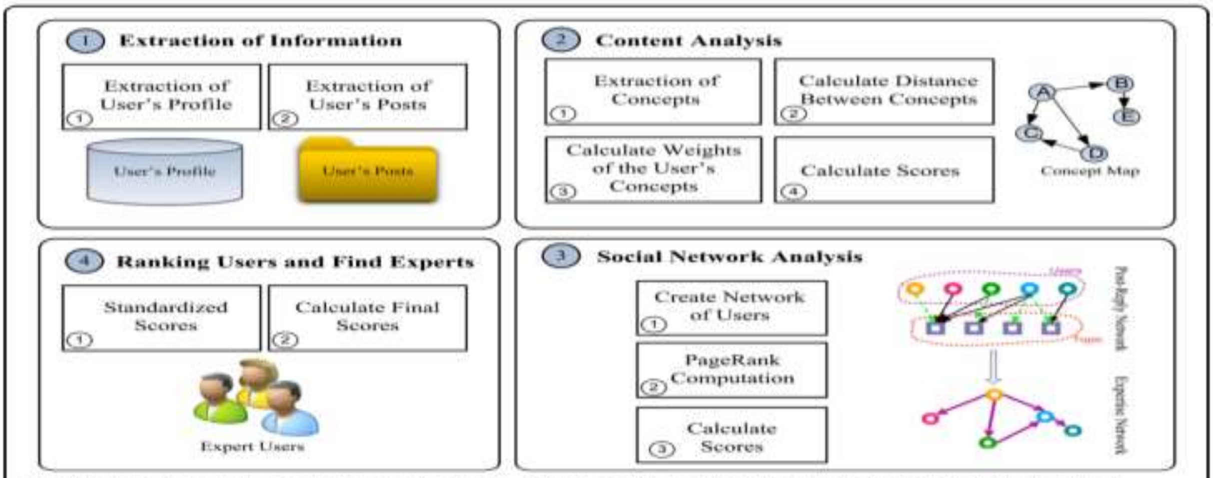

1 Extraction of Information
Extraction of User's Profile 1
Extraction of User's Posts 2
User's Profile
User's Posts
2 Content Analysis
Extraction of Concepts 1
Calculate Distance Between Concepts 2
Calculate Weights of the User's Concepts 3
Calculate Scores 4
A
B
C
D
E
Concept Map
4 Ranking Users and Find Experts
Standardized Scores 1
Calculate Final Scores 2
Expert Users
3 Social Network Analysis
Create Network of Users 1
PageRank Computation 2
Calculate Scores 3
Users
Topic
Post-Reply Network
Expertise Network

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

Fig- 1: Proposed model

Information Extraction:

At first, the structure of a social network like Quora was created. The user profiling is an integral part of the system, hence we had to create a profiles for about 10-20 individuals. For this concern, the URL rewriting and readdressed process through access logs are analysed. This process gives away great deal of information related to the users in concern. Following information is to be extracted in this step:

- User ID which can be used as a unique identifier
- User's advocacy level with a pointer
- User’s posts in form of textual dumps or sql dumps
- Total Number of user's posts
- Total Number of user questions

Content Analysis:

Initially, concepts in form of exchanged questions and answers of each user are extracted from the profiling. Since these concepts are supposed to be extracted and compared with a certain subject concepts map, it is absolutely necessary to extract all nodes from the subject concept map in preview form. For each node, other keywords with same meaning are considered as well. After the creation of a data structure for the concept map, concepts of the restored posts are extracted in accordance to the concept map. At the commencement of this stage, each user has a data structure which includes core concepts of each question and their keywords relevant to the certain response posted to that question.

To compute distance between the concepts deduced from the set of responses and the concepts deduced from set of questions should be extracted. In this concern, the smallest distance from one concept to every concept in the overall concept map is extracted. Thereby it is necessary to draw a graph which shows the relations between all the concepts. In an experimental procedure, using Dijkstra's algorithm, the shortest path between any two nodes in an undirected graph is calculated. The output here in this stage is a two-dimensional matrix which holds distance between the different concepts.

Now, the average distance between each concept with respect to all concepts in the question is calculated by equation given below:

$$AvgDist(R) = \frac{\sum_{Q \in Questions} (Dist(R,Q))}{N}$$

$$R, Q \in \text{Concepts of Concept Map}$$

Fig:-2 Formula for average distance

Where R is called the concept of a given response, Q is generalized concept of the certain question, Questions are all concepts in the question, Dist. (R,Q) is actually Euclidean distance between concept R in the response, and concept Q in the question and N is the number of the concepts in the question.

Computing Rank Scores:

The computation of ranks is the final step in the ontological extraction process. We have to take a calculated average of all the responses to a certain response in the map. The map is a logistic mapping and requires a blend of two techniques called a spearman mapping and average concept conjunction. The equation given below uses the same concepts.

Where Score(I) is the maximum score of user X, Messages are the subtle messages of user X, Responses are concepts that have been in the response of the message M, Rep (R) is the number of iterations of the base concept R in the response of the message M and Weight (R) is the consecutive weight of concept R in the response of the given message M.

The α and β are coefficients with correlating values between 0 and 1. α indicates the impacting factor of the number of concepts which are stored in the user response, and β indicates the impact of the distance between concepts in user consecutive response and the correlated concepts in question. In this paper, the overall optimum values for these coefficients are pre calculated. To achieve the optimum values state space is traversed by changing 0.01 intervals for all values of α and β. The optimum values obtained for these coefficientsare equally 0.5. By using these considerate coefficients, the best correlations found between the above scores is obtained from the proposed method and the scores provided by pre calculated community, is calculated.

Rank Computation:

In a online community it may be perfectly possible to have more than one link between two nodes, however in a given network which is deployed over the web, only one to one relation is possible in each direction. So in the modified implementation of our PageRank algorithm, weights in form of transfer matrix are computationally determined on the basis of the number of interlinks between two nodes.

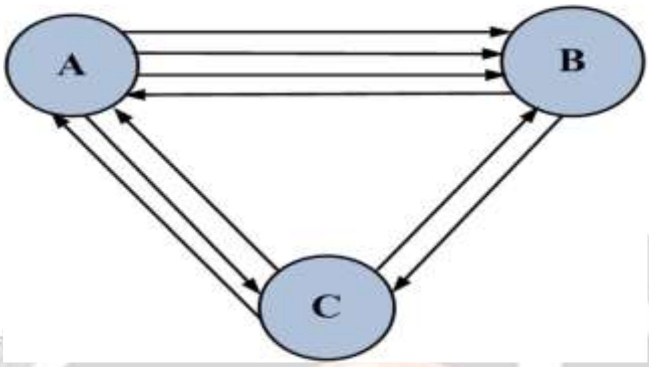
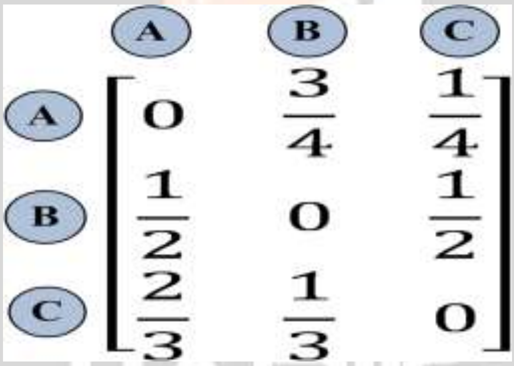


Fig:-3 Sample of expertise network



$$\begin{array}{c|ccc} & A & B & C \\ \hline A & 0 & \frac{3}{4} & \frac{1}{4} \\ B & \frac{1}{2} & 0 & \frac{1}{2} \\ C & \frac{2}{3} & \frac{1}{3} & 0 \end{array}$$

Fig:-4 Weight of sample score

Calculation of final scores:

When the network of the experts is created, a reaffirmation is needed so that the experts are given proper question which suit their expertise perfectly and do justice to their knowledge. The experts will be only shown the legitimate questions which will avoid the expert from wasting his time in writing answers which can be answered from a normal individual within that set of rank of questions. This can be considered by assigning questions ranks and merging of contextually similar questions. At final, users are ranked with final Score(p) values and top users are determined as expert users. The top hierarchy of users are assigned ranks based on the frequency of answer and its relevance with the question. The final score is calculated by summation of all scores of relevance in a certain area of interest.

Evaluation and results

To evaluate the proposed method, all the subsequent subsection of online forums is used.

Firstly, the number of responses for each of the given subsection was calculated and subsections which have the overall number of responses for them is less than a given threshold have been excluded. Spearman's correlation between our results and scores prepared by a certain online community was calculated separately for the 11 subsections and the entire java online community, the overall correlation was calculated by taking the average of

these correlations. The overall correlation was calculated with different values for weight of content analysis and social network analysis.

Conclusions:

In this proposed system, the importance and role of online answering communities along with social networks is brought into picture. In addition to emphasizing knowledge sharing and its position on web, yet another important concerns and challenges related to web communities were brought into picture, with a focused plan on one of the solutions to these challenges that was “Expertise Mapping and Expert Finding", related works done in this field of "Expert Finding and Mapping" were transcended, and a novel hybrid approach based on concept maps and PageRank which are a part of ontological learning were used for expert finding on web communities was presented. In the proposed method, similarity between concepts extracted by an Artificial Intelligence based measure.

Future Scope:

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

References:

[1] J. Zhang, J. Tang, and J.-Z. Li. Expert finding in a social network. In DASFAA, pages 1066–1069, 2007.

[2] Balog, Krisztian, et al. "Expertise retrieval." *Foundations and Trends in Information Retrieval* 6.2–3 (2012): 127-256.

[3] Bozzon, Alessandro, et al. "Choosing the right crowd: expert finding in social networks." *Proceedings of the 16th International Conference on Extending Database Technology*. ACM, 2013.

[4] Fang, Hui, and ChengXiang Zhai. "Probabilistic models for expert finding." *European Conference on Information Retrieval*. Springer Berlin Heidelberg, 2007.

[5] Zhang, Jing, et al. "A mixture model for expert finding." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2008.