Enhanced Approach to Find Topic Experts in Forum

Mayur Jakate[1], Lata Ragha[2]
[1]PG Student, Department of Computer Engineering, Terna Engineering College, Nerul, Navi-Mumbai
[2]Professor and HoD, Computer Engineering, Fr. C. Rodrigues Institute of Technology, Vashi, Navi-Mumbai

Abstract- *Social media Networks and micro blogging are now used all over the world. It has become very difficult to identify which among this information are important for us. So, Expert finding has become a hot topic on social network. Information from Experts is considered to be trustworthy and relevant to satisfy our need but several attempts use the relations among users and list of label query which extracted from given query. In literature the authors used to find the expert from twitter corpus that is specific to twitter data. We propose a method that uses stack overflow API data set to find an expert of different topics which are at the pic of micro blogging. We use different traditional method to explore the topics and user among their relationship such as semi supervised algorithm; page rank algorithm, HITS Algorithm and LDA Algorithm Our work is divided into three stages such as data set formation, extract complete information form dataset and apply suitable algorithm to get desired output.*

Keywords- Expert search, micro-blogging, SSGR (Semi-Supervised Graph-based Ranking), graph-based ranking, Latent Dirichilet Allocation (LDA), Hyperlink-Induced Topic Search (HITS) algorithm and Term Frequency–Inverse Document Frequency (TF-IDF)

I. INTRODUCTION

Expert finding is one of the trading topics in social media with result to get the best profile of the person having relevant knowledge on the particular topic. Recently expert finding problem attract the attention of on social media such as forum like Stack overflow, it features questions and answers on a wide range of topics given by different people [1]. As we know knowledge sharing is one of the most important applications of online communities in virtual space of the Internet. For example in [2] and [3] number of factor that impact of knowledge sharing in online communities get identified.

In online communities we don't know knowledge level of user so value of answer and comment are unclear. This is the biggest challenge in micro blogging communities. By identifying knowledge level of each user we can extract the most valuable answer for the answer posted by the user.

It is also important that the question posted by user on forum not take too much time to respond by the other user, questions can be exposed to individuals who have adequate knowledge to respond them.so without wasting time answering simple question such a recommender system has been implemented in [4].

Several research [5] efforts have been made for to find an expert on given topics. However previous works usually studied the person local information and relationships separately and combined them in an ad-hoc approach. They utilize the link between follower relation, user-list relation and list-list relation, and provide the one directional approach to find the expert like they find the expert for the given topics the flow is first need to provide topic and then you will get the top- n expert in that given topic, in TREC'2005 and TREC'2006 have provided a common platform for researchers to empirically assess methods and techniques devised for expert finding.

In this paper, our focus is how to make use of person local information and relationships between persons in a unified approach. We proposed a propagation based approach for finding expert in a social network. The approach consists of two steps. In the first step, we make use of person local information to estimate an initial expert score for each person and select the top ranked persons as candidates. The selected persons are used to construct a sub-graph. In the second step, we propose a propagation-based approach, which propagates one's expert score to the persons with whom he/she has relationships. And we are investigate to extract meaningful information present in unstructured format by the concept of data mining, we collect the all unstructured data and store it in structure format by ETL (Extract-Transform-load) concept and apply some algorithm to such as graph base algorithm for connection between the information, and some ranking algorithm such as HITS (hyperlink-induced topic search) algorithm and TF-IDF (Term Frequency–Inverse Document Frequency) to define rank for the user based upon the positive vote count for the answer given by that person on specific topic.

II. LITERATURE SURVEY

[1] A virtual community is a social network of individuals who interact through specific social media, how virtual community is important with perspective of sharing

knowledge and reduces time to respond on the query asked by the user.

[2] Provide the brief information about the objective of professional virtual communities PVCs to encourage people to exploit or explore knowledge through websites. However, many virtual communities have failed due to the reluctance of members to continue their participation in these PVCs.

[3] Concentrate how Help-seeking communities have been playing an increasingly critical role in the way people seek and share information, here used Question Matching Engine (QuME) to identify similarity between question posts by different user.

[4] Problem of identifying influential users of micro-blogging services. Twitter, one of the most notable micro-blogging services, employs a social-networking model called "following", in which each user can choose who she wants to "follow" to receive tweets from without requiring the latter to give permission first.

[5] Page prediction method that is based on semantic classification of Web pages supported with Popularity based Page Rank (PPR) technique. As the first step, we use a model that basically uses Web page URLs in order to classify Web pages semantically. By using this semantic information, next page is predicted according to the semantic similarity of Web pages.

[6] HITS algorithm is a very popular and effective algorithm to rank documents based on the link information among a set of documents. However, it assigns every link with the same weight which results in topic drift. In this paper, we generalize the similarity of web pages and propose a query-induced similarity describing how a webpage is similar to another on a query topic.

[7] Combined approach which employs both unsupervised and semi-supervised learning paradigms. An unsupervised distance learning procedure is performed as a pre-processing step for improving the kNN graph effectiveness. Based on the more effective graph, a semi-supervised learning method is used for classification.

[8] In this paper, author evaluated two different approaches for documents ranking and further we checked this achieved results with other approach based on machine learning Firstly, Documents are ranked based on standard score calculation i.e using the tf-idf concept. Secondly, Documents are ranked based on Textiling approach [9]. Textiling is a technique in which text is divided into more than one paragraph units known as passages.

III. ISSUES IN PREVIOUS APPROACH

However, existing approaches [1], [3], [4] only partially utilize either follower relation or user-list relation alone, and their architecture is specifically related to Twitter data set only thus it is insufficient for Twitter expert finding problem to identify the expert when have different data set provided as input.

To this end, we propose universal architecture in which we provide any type of data set as input and get the result as top N number of expert related to topic.

IV. METHODOLOGY

Before move ahead we must need to understand the concept of Data mining. Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas. We are using the stack overflow data set which is presented in structure and unstructured format

Data Extraction

We consider Stack overflow data [9] where they provided the API (Application Programing Interface) for accessing the data which is available in JSON (Java script object notation) format. The data presented in JSON format is not readable so we have to extract data by using JSON reader.

Transform

Now we have data but not available in the required format, we have to transfer the data into desire format by applying rules [11] on data set.

Load

Next step is to store the data into data base. We consider target database as Microsoft SQL server

Now, we apply some algorithm to extract the information as per user query.

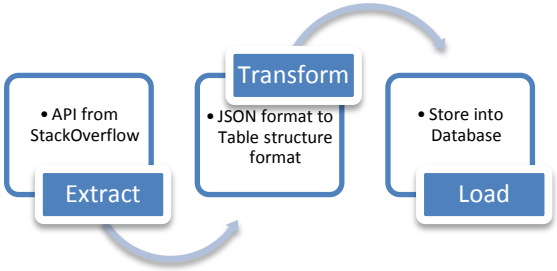

Fig: 1 Steps to convert unstructured data into structured format

For extracting labeled data there are various modeling methods such as:

Explicit Semantic Analysis (ESA)

ESA is a vectoral representation of text (individual words or entire documents) that uses a document corpus as a knowledge base. Specifically, in ESA, a word is represented as a column vector in the tf–idf matrix of the text corpus and a document (string of words) is represented as the centroid of the vectors representing its words.

Latent Semantic Analysis (LSA)

LSA is a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text.

Latent Dirichlet Allocation (LDA)

LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

LDA Algorithm

In LDA, each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA

If we have list of some phrase and we need to divide them as per their topic.

1. I like to eat chocolates and apple.
2. My kitten like warm milk
3. Dogs and cats are cute

Table-I. Document-Topic matrix

	T1	T2	T3	Tn
D1	1	0	1	0
D2	0	1	1	0
D3	1	0	1	1
Dn	1	0	0	1

In table-I and II, T means list of different topic and D means list of document and W means words.

Given above phrase we have some list of topic such as Eat, Animal, chocolates, cat, dog, milk. So we just put 1 in matrix if the document contain that topic else 0.

Table-II. Topic-word matrix

	W1	W2	W3	W..n
T1	1	0	0	1
T2	0	1	1	0
T3	0	1	1	1
T..n	0	0	1	0

In above matrix we calculate whether given topic contain that word or not.

Proportion of words in document d that currently assigned to topic t as P1= (Topics/Document)

Proportion of assigned topics t over all documents that come from this word as P2 = (Words/Topics)

Probability (p) of document contain topics given as

P = P1*P2.

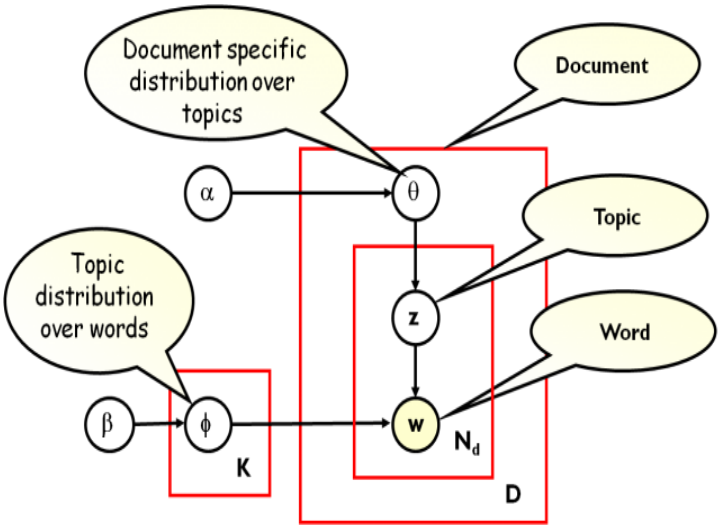

Fig 2: Plate notation representing the LDA model

Notations used in figure 4 are explained below:

- α is the parameter of the Dirichlet prior on the per-document topic distributions.
- β is the parameter of the Dirichlet prior on the per-topic word distribution.
- w is the specific words.
- ø is the topic distribution for document k.
- θ is the word distribution for topic k,

Graph-Based Ranking Algorithms

Graph-based ranking algorithms are essentially a way of deciding the importance of a vertex within a graph, based on information drawn from the graph structure. We present page rank and HITS algorithms

Let G = (V;E) be a directed graph with the set of vertices V and set of edges E, where E is a subset of V * V . For a given vertex V_i, let In(V_i) be the set of vertices that point to it (predecessors), and let Out(V_i) be the set of vertices that vertex Vi points to (successors).

Page rank algorithm

PageRank (PR) is an algorithm to rank websites in their search engine results. "Using PageRank, we are able to order search results so that more important and central Web pages are given preference. In experiments, this turns out to provide higher quality search results

$$PageRank\ of\ user = \Sigma \frac{PageRank\ of\ inbound\ user}{Number\ of\ links\ on\ that\ user}$$

HITS (Hyperlink-Induced Topic Search)

Hyperlink-Induced Topic Search (HITS; also known as hubs and authorities) is a link analysis algorithm that rates Web pages.
The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that they held, but were used as compilations of a broad catalog of information that led users direct to other authoritative pages.

In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs.

- Hubs: Consider as single User profile.
- Authorities: An authority is a User that many hubs link to.

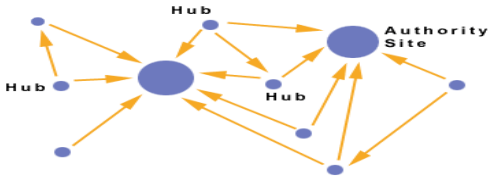

Fig 3: Hubs and Authority connection

We associate to each page *i* two numbers: an authority weight a_i, and a hub weight *h*i. We consider pages with a higher a_i number as being better authorities, and pages with a higher *h*i number as being better hubs. Given the weights $\{a_i\}$ and $\{h_i\}$ of all the nodes in *S*Q, we dynamically update the weights as follows:

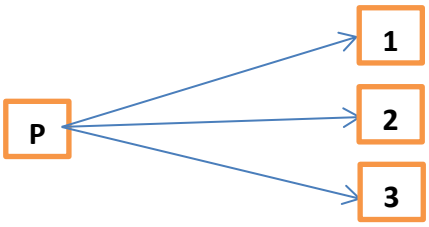

Fig 4: Authority of person P connected to different hubs

Authority of person a_p=the sum of h_i for all nods i pointing to p as a(p) = ∑p→qh(q)

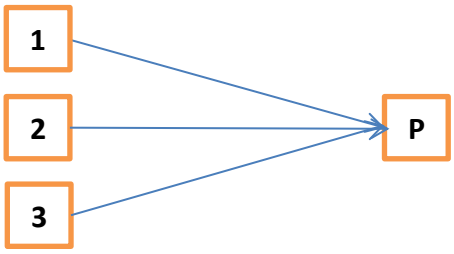

Fig 5: Different Hubs connected to person P

Hubs of person $h_{p=}$ the sum of a_i for all nods $_i$ pointed to by p
h (p) = ∑q→pa(q)

TF*IDF ALGORITHM

TF*IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF*IDF weight of that term.

Put simply, the higher the TF*IDF score (weight), the rarer the term and vice versa. The TF*IDF algorithm is used to weigh a keyword in any content and assign the importance to that keyword based on the number of times it appears in the document. More importantly, it checks how relevant the keyword is throughout the web, which is referred to as corpus.

For a term t in a document d, the weight Wt.d of term t in document d is given by:

$$Wt.d = TFt.d \log (N/DFt)$$

Where:

- TFt.d is the number of occurrences of t in document d.
- DFt is the number of documents containing the term t.
- N is the total number of documents in the corpus.

V. CONCLUSION

In this paper we addressed a problem of topic specific expert finding in forum. We integrated different algorithms and methods to identify topics and users from dataset. Our method aimed to assign similar ranking scores to the similar users, and meanwhile the ranking scores are subjected to the supervised information from the input data which is provided. Based on the computed ranking scores, we selected the top-N relevant users for any given topic.

REFERENCES

[1] Kardan A, Garakani M, Bahrani B (2010), "A method to automatically construct a user knowledge model in a forum environment", Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp 717–718

[2] Chen I (2007), "The factors influencing members' continuance intentions in professional virtual communities - A longitudinal study", Journal of Information Science, 33(4), pp 451–467

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

[4] J. Weng, E.-P. Lim, J. Jiang, and Q. He (2010) "Twitterrank: Finding topic-sensitive influential Twitterers", Proceedings of ACM International Conf. on Web Search Data Mining, pp. 261–270.

[5] Banu Deniz Gunel, Pinar Senkul (2013) ," Integrating Semantic Tagging with Popularity-Based Page Rank for Next Page Prediction", Computer and Information Sciences III.

[6] Weiming Yang (November 2016), "An Improved HITS Algorithm Based on Analysis of Web Page Links and Web Content Similarity", International Conference on Cyber worlds.

[7] Fabricio Aparecido Breve ; Daniel Carlos Guimarães Pedronette (2016), "Combined unsupervised and semi-supervised learning for data classification", IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP).

[8] Sanjeev Patel ; Kriti Khanna ; Vishnu Sharma (2016), "Documents ranking using new learning approach", International Conference on Computing, Communication and Automation (ICCCA)

[9] Link for API consider as dataset https://api.stackexchange.com/docs

[10] http://www.aclweb.org/anthology/J97-1003

[11] https://isi.edu/integration/papers/bowu12-iiweb.pdf