IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Finding Topic Experts in the Twitter dataset using LDA Algorithm

Ashwini Anandrao Shirolkar[1*], **R. J. Deshmukh**[2]

[1,2]Department of Technology, SU, Kolhapur, India

[*]*Corresponding Author: ashwinishirolkar8@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 17/Aug/2018, Published: 31/Aug/2018

***Abstract*—**Expert finding which aims to identifying people with the relevant expertise or practices on a given topic query. In blogging services like Twitter, the expert analysis problem has gained big attention in social media. Twitter is a new type of media giving a publicly available way for users to publish 140-character short messages (i.e., tweets). However, earlier systems cannot be directly applied to twitter expert finding difficulty. They generally rely on the supposition that all the documents linked with the candidate experts receive implicit knowledge related to the expertise of individuals. Whereas it might not be directly allied with their expertise, i.e., who is not an expert, but may publish/re-tweet a substantial amount of tweets including the topic words. Recently, several attempts use the relations among users and twitter list for expert finding. Nevertheless, these strategies only partly utilize such relationships. To address these issues generate a probabilistic method to jointly exploit three types of relations (i.e., follower relation, user-list relation and list-list relation) for finding experts. LDA algorithm is applied to finding topic experts. LDA is based upon the concept of searching for a linear combination of variables (predictors) that best separates two classes (targets). Semi-supervised Graph-based Ranking approach (SSGR) to offline measure the global authority of users. Then, online compute the local relevance between users and the given query. Then order all of the users & find top-N users with the highest ranking scores. Therefore, the proposed method can jointly exploit the different types of relations among users and lists for improving the precision of finding experts on a given topic on Twitter.

***Keywords*—**Expert finding, Semi-supervised, Graph-based ranking approach, LDA, Sentiment Analysis, Hashtag, Twitter.

I. INTRODUCTION

The purpose of the data mining method is to extract knowledge from a data set and transform it into an acceptable structure for additional use. Data mining is widely used in diverse areas. There is a number of the commercial data mining system available today, and yet there are many challenges in this field. Data mining is widely used in intrusion detection, biological data analysis, scientific applications, social networking.

Mining in social networking is an essential data mining task with broad applications. Expert finding problem has gained increasing concentration in social media, it builds a weighted graph by considering both the topical similarity between two users and followers graph, and then employ page Rank algorithm to find topic-specific influential users. Pal et al. [4] extract user's characteristics from the follower graph, and users posted tweets, and then employ a Gaussian-based mixture model to cluster users for ranking. The Twitter Rank and pal's work only consider user-user single relation.

Gosh, et al. [2] proposes to utilize Twitter List to analyze the attributes of Twitter users. In their resulting work, they develop a system named Cognos [2] to infer the topic based expertise of users by appropriating only user-list relation in Twitter List, which takes the wisdom Twitter crowds. Cognos represents each user by the meta-data of Twitter lists that contain the user and then employs a similarity measure to measure the similarity score between each user and topical query, which is used to rank users for search.

The proposed method jointly exploits three types of relations (i.e., follower relation, user-list relation and list relation) for finding experts. LDA algorithm is used to finding topic experts.LDA is based upon the concept of searching for a linear combination of variables (predictors) that best separates two classes (targets). A Semi-Supervised Graph-based Ranking approach (SSGR) to offline calculates the global authority of users. In SSGR, employ a normalized laplacian regularization term to jointly explore the three relations, which is subject to the supervised information derived from Twitter crowds. Then online compute the local relevance between users and the given query. By taking advantage of the global authority and local relevance of users, rank all of the users and find top-N users with the highest ranking scores. The suggested approach jointly exploits the various types of relations. Two types of information to target Twitter expert finding problem, namely: i) Local Relevance:- the similarity between users published tweets and the given query, and ii)Global Authority:- the global expertise scores of users on a given topic in Twitter. The proposed approach successfully combines different types of user-related information (i.e., the crowdsourced Lists information, follower graph and users profiles) into a unified ranking framework for accurately inferring the topical expertise of users.

Next, we present a literature overview in section II. Proposed approach to addressing the topic-specific expert finding problem explained in figure 1 and section III..Specifically, it consists of two components, namely, an offline graph-based ranking to learn the global authority of each candidate and an online ranking to select top-N relevant experts on the given query. Section IV used to explain algorithm for expert finding. In section V semantic similarity calculation method for tweeter analysis explained. Result and discussion and conclude the paper further in VI and VII respectively.

II. Related Work

"Cognos: crowd sourcing exploration for topic experts in microblogs"S. Ghosh, N. Sharma, N. Ganguly and K. Gummadi [2] Propose and investigate a new methodology for discovering topic experts in the popular Twitter social network. The methodology relies on the wisdom of the Twitter crowds; it leverages twitter lists, which are often individual users to include experts on topics that interest them and whose metadata (list names and descriptions) provides valuable semantic cues to the experts' domain of expertise. In this list the knowledge to build Cognos a system for finding topic experts in Twitter. Cognos infer a user's expertise more accurately than a state-of-art system that relies on the user's bio or tweet content. Cognos scales well due to a built-in mechanism to update its expert's database with the new user.

"Twitter-rank: Finding topic sensitive influential twitters". Weng, E. P. Lim, J. Jiang, and Q. He [3] focuses on the 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 decide whom she wants to "follow" to receive tweets from without requiring the latter to give permission first. The Twitter-rank algorithm an extension of the Page-Rank algorithm used to measure the influence of users on Twitter. Twitter-rank measures the similarity between users and the link structure into account. The Twitter-rank works in two steps. First, it employs the Latent Dirichlet Allocation (LDA) model to detect the topics of individuals based on their tweets. Second, for each topic, it builds a weighted graph by considering both the topical similarity between two users and the follower graph, and then employ the Page-Rank algorithm to find topic-specific influential users.

"Identifying topical authorities in micro blogs" A. Pal and S. Counts [4] focus on the problem of identifying topic authorities in micro blogs. Hundreds of millions of users produce content in micro blogging systems such as Twitter. This diversity is a notable strength but also presents the challenge of finding the most exciting and authoritative authors on any given topic. In this proposes an approach that extracts user's features from the follower's graph and users posted tweets, and then employ a Gaussian-based mixture model to cluster users for ranking. Gaussian mixture model used to group users into two or more clusters. The primary motivation for the clustering was to reduce the size of the target cluster (i.e., the cluster containing the most authoritative users). This further makes the subsequent ranking of users more robust because it is less sensitive to outliers such as celebrities.

III. PROPOSED SYSTEM

The proposed work figure 1 shows system architecture, addressing the topic-specific expert finding problem consist of two components.

1. Offline graph-based ranking algorithm to learn the global authority of each candidate and
2. An online ranking model to select top-N relevant expert on the given query. In particular, each term t in twitter is treated as a potential topic.

LDA is based upon the concept of searching for a linear combination of variables (predictors) that best separates two classes (targets). To capture the notion of expertise defined the following score function.

$$Z = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_d x_d \tag{1}$$

$$S(\beta) = \frac{\beta^r \mu_1 - \beta^r \mu_2}{\beta^t C \beta} \tag{2}$$

$$S(\beta) = \frac{Z_1.Z_2}{Varience \quad of \quad Z \quad within \quad the \quad group} \tag{3}$$

$$S(\beta) = C^-1(\mu_1 - \mu_2) \tag{4}$$

$$C = \frac{1}{n_1 + n_2}(n_1 C_1 + n_2 C_2) \tag{5}$$

Where, β - Linear Model Coefficient; C_1, C_2 - Covarience Matrices and μ_1, μ_2 - Mean Vector.

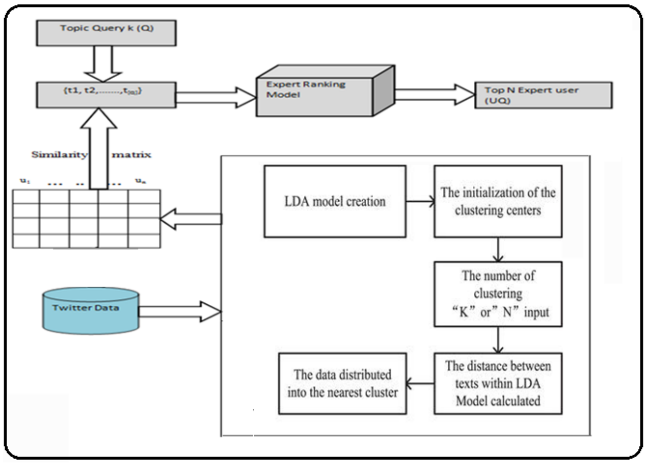
Topic Query k (Q)
{t1, t2,........,t|Q|}
Expert Ranking Model
Top N Expert user (UQ)
Similarity matrix
u1 ... . un
LDA model creation
The initialization of the clustering centers
The number of clustering "K" or" N" input
The distance between texts within LDA Model calculated
The data distributed into the nearest cluster
Twitter Data

Figure 1 System Architecture

[1]http://saedsayad.com/lda.htm

Given the score function, the problem is to determine the linear coefficients that maximize the score which can be done by the following equations.

One way of evaluating the effectiveness of the discrimination is to calculate the [1]**Mahalanobis distance** between the two groups. A distance is more significant than three means that in two averages differ by more than three standard deviations. It means that the overlap (probability of misclassification) is quite small.

$$\Delta^2 = \beta^T \mu_1 - \mu_2 \qquad (6)$$

Δ : Mahalanobis distance between two groups.

Finally, a new point is classified by projecting it onto the maximally separating direction and classifying it as *C1* if:

$$\beta^T = \left(X - \left(\frac{\mu_1 + \mu_2}{2} \right) \right) > -log \frac{p(C_1)}{p(C_2)} \qquad (7)$$

IV. EXPERT FINDING STRATEGY USING LDA

```
Input    R: Authority matrix;
Q: Given topic query;
U: User set
Result: Topic-specific expert set $U_Q$
        Set f ←0;
        Foreach $u_k$ ε u do
        If|Q| = 1 do
        $f_k \leftarrow \frac{H(u_k, Q)}{\Sigma_j H(u_k, Q)'}$; break;
        For i←1 to |Q|-1 do
        $f_k \leftarrow f_k +$ Pr($u_k | t_i$) Pr($u_k | t_{i+1}$) K($t_i, | t_{i+1}; d^{u_k}$);
        $U_Q \leftarrow \overset{N}{u \varepsilon \underline{u}} argmax f_i$ ($f_i \varepsilon f$);
        Return $U_Q$;
```

The expert ranking algorithm uses a topic query, authority matrix and user set as input parameters. An output of the algorithm is top N experts that are most relevant to a query.

Where, *UQ* denotes the retrieved top-N experts that are most relevant to query Q.

Where, *UQ* denotes the retrieved top-N experts that are most relevant to query Q.

$$\vartheta\,(\mu_k, t_i) = R_i, k$$

This is an entry of the authority matrix R. It indicates the global authority of user u_k on t_i, computed by SSGR.

LDA subject model having three layers Bayesian generative model—text-topic-word. The essence of LDA is to find topic structure of text using feature of words co-occurrence in text. In generation process, each text is represented as mixture distribution of subjects, and each subject is a probability distribution over words. Leading a hyper-parameter α into the model's document-topic probability distribution, thus the new model obeys Dirichlet distribution. Then Griffiths and etc apply Dirichlet prior distribution to another parameter, which β makes the LDA subject model come into being a completed model.

$$p(W_m, Z_m, \Theta_m, \Phi | \alpha, \beta) = \prod_{n=1}^{N} p(W_m, n | \Phi_{2m}, n), p(Z_m, n | \Theta_m) p(\Theta_m | \alpha), p(\theta | \beta)$$

We often set Hyper-parameter 0.1, K is= β 50/ K,= α number of topics.

Table1 : Hyper-parameter over number of topics.

Symbol	Meaning	Symbol	Meaning
α	Hyper-parameter of θ	Wm, n	word
β	Hyper-parameter of φ	M	Text No
θ_m	Text-topic probability Distribution	N	word No.
φ_k	Topic-word probability Distribution	K	Topic No.
$Z_{m,n}$	Distribution of words in a topic		

V. SEMANTIC SIMILARITY CALCULATION METHOD OF WORD IN TWEETER DOMAIN

Running LDA topic model and doing Gibbs sampling on the document corpus *D*, we get K topics hidden in the documents and topic-word probability distribution Φ.
The element $\varphi_{sk,wt}$ of Φ shows the probability of word W_t belongs to topic S_k $(1 \leq k \leq K)$.
K Topics build a feature space:

$$V = (s_1, s_2, s_3, \ldots, s_k) \qquad (9)$$

So the word *w1* and *w2* distribution vector in *K* topics feature space is:

$$V_{w1} = (\varphi_{s1,w1}, \varphi_{s2,w1}, \varphi_{s3,w1}, \ldots\ldots, \varphi_{sk,w1})$$

$$V_{w2} = (\varphi_{s1,w2}, \varphi_{s2,w2}, \varphi_{s3,w2}, \dots\dots, \varphi_{sk,w2}) \quad (10)$$

The semantic similarity calculation of two words w1 and w2

$$sim(w_1, w_2) = \cos(\mathbf{v}_{w_1} \oplus \mu\mathbf{v}_{n_1}, \mathbf{v}_{w_2} \oplus \mu\mathbf{v}_{n_2}) \quad (11)$$

the value of eq. (11) is higher; the similarity of two words w1, w2 is more approximate, vice versa.

VI. EXPERIMENTS AND RESULTS

A. *Experimental Setup*

Experiments were performed on a standalone machine with a 2.30 GHz Intel Core 3 processor running Windows 7 and 2GB of free RAM. To measure the performance we have taken twitter dataset, Twitter API which is 3.78 GB in size. The dataset is a subset of Twitter. It contains 284 million following relationships, 3 million user profiles, and 50 million tweets.

B. *Result Analysis*

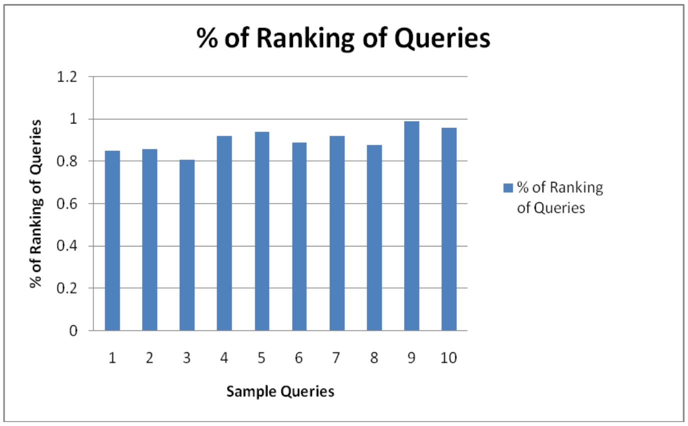

Figure 2 Topic expert relevance analysis

The graph in the above figure 2 shows the rank of relevance judgments over the top 10 results for the 55 sample queries, out of which (78.7%) judged the result (topical expert has shown by proposed method) to be relevant to the query.

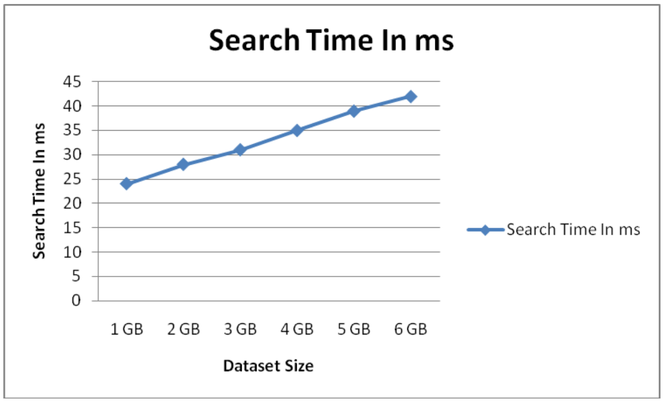

Figure 3 Time Complexity

The graph in the above fig. 3 shows the time complexity of the proposed approach used in our twitter expert finding system. The system uses constant time performance for all dataset ranging from 1GB, 2GB.....6GB.

VII. CONCLUSION AND FUTURE SCOPE

The paper elaborates the topic-specific expert finding in twitter successfully integrates different types of user-related information into a unified ranking framework for accurately inferring the topical expertise of users using the LDA model. By using twitter information, we have attempted to find experts. The ranking of an expert is done by the supervised information from the Twitter crowds. In the ranking scores, the top N users are selected as the topic experts to a particular topic.

ACKNOWLEDGMENT

This paper work was guided & supported by Mrs. R. J. Deshmukh. I would like to thank the guide, anonymous reviewers for their valuable and constructive comments on improving the paper.

REFERENCES

[1] Wei Wei, Gao Cong, Chunyan Miao, Feida Zhu, and Guohui Li "Learning to Find Topic Experts in Twitter via Different Relations" IEEE transactions on knowledge and data engineering, vol.28, no 7, July 2016.

[2] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi, "Cognos: Crowdsourcing search for topic experts in microblogs," in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2012, pp. 575–590.

[3] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential Twitterers," in Proc. ACM Int. Conf. Web Search Data Mining, 2010, pp. 261–270.

[4] A. Pal and S. Count, "Identifying topical authorities in microblogs," in Proc. ACM Int. Conf. Web Search Data Mining, 2011, pp. 45–54.

[5] Z. Zhao, L.-J. Zhang, X.-F. He, and W. Ng, "Expert finding for question answering via graph regularized matrix completion," IEEE Trans. Knowl. Data Eng., vol. 27, no. 4, pp. Retrieval, 2012, pp. 575–590.

[6] A. Pal and J. A. Konstan, "Expert identification in community question answering: Exploring question selection bias," in Proc. ACM Conf. Inf. Knowl. Manag., 2010, pp. 1505–1508.

[7] X. Liu, W. B. Croft, and M. Koll, "Finding experts in community-based question-answering services," in Proc. ACM Conf. Inf. Knowl. Manag., 2005, pp. 315–316.

[8] W. Wei, B. Gao, T.-Y. Liu, T.-F. Wang, H.-G. Li and H. Li. "A ranking approach on a large-scale graph with multidimensional heterogeneous information," IEEE Trans. Cybern., vol. Pp, no. 99, pp. 1–15, Apr. 2015.

[9] G. Demartini, D. E. Difallah, and P. Cudr_e-Mauroux, "Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 469–478.

[10] J. Lehmann, C. Castillo, M. Llamas, and E. Zuckerman, “Finding news curators in Twitter,” in Proc. Int. Conf. World Wide Web, 2013, pp. 469–478.

[11] K. Balog, L. Azzopardi, and M. De Rijke, “Formal models for expert finding in enterprise corpora,” in Proc. 29th Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 4

[12] A. Pal and J. A. Konstan, “Co-occurrence-based diffusion for expert search on the web,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 5, pp. 1001–1014, May 2013.

[13] M. David and A. Andrew, "Expertise modeling for matching papers with reviewers," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2007, pp. 500–509.

[14] H. Deng, I. King, and M.-R. Lyu, "Formal models for an expert finding on DBLP bibliography data," in Proc. Int. Conf. Data Mining, 2008, pp. 163–172.

[15] P. Serdyukov, H. Rode, and D. Hiemstra, “Modelling multi-step relevance propagation for expert finding,” in Proc. ACM Conf. Inf. Knowl. Manag., 2008, pp. 1133–1142.

Authors Profile

Ms. Ashwini shirolkar , received her B.Tech. degree from Department of technology, SUK,India. She is prsuing M.tech degree . from,Department of Technology at SUK.,Her research areas include Data Mining, Personalized,recommendation.

*M*rs.R J Deshmukh has pursed Bachelor of Engineering from Shivaji University, Kolhapur,India and Master of Technology from Shivaji University,Kolhapur,India in year 2009. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Technology, Department of Computer Science and Technology, Shivaji University,Kolhapur, Maharashtra,India. She has published more than 10 research papers in reputed international journals and conferences including IEEE,Elsvier,Springer and it’s also available online. Her main research work focuses on Data Mining.