Asian Journal of Information Technology 15 (19): 3763-3769, 2016
ISSN: 1682-3915


# Academic Tweet Concept Based Co-author Recommendation

G. Manju and T.V. Geetha
Department of Computer Science and Engineering,
Anna University, Chennai, Tamil Nadu, India

**Abstract:** Researchers carrying out research and writing research article requires knowledgeable person in their topic to assist them in successfully publishing the study. Hence, this study presents a solution to this problem by recommending suitable co-authors for a particular topic. We identify co-researchers by incorporating researchers social similarity along with the traditional features like proficiency in a research area, semantic similarity of research interests and publication details. We have determined the social similarity of the researcher based on the Twitter social network. We determine the concept, social and difference topic similarity between the researchers and rank the co-authors using Lambda rank algorithm. We investigated the approach by carrying out experiments with datasets of academic publications in the area of computer science. The experimental results illustrates that the combination of social and semantic features provides better recommended list of co-authors, when compared to baseline approach.

**Key words:** Co-author recommendation, lambda rank, twitter, semantic relatedness, concept

## INTRODUCTION

There is a rapid increase in the number of researchers as well as new research topics. A researcher working on a research problem may not be an expert in all the topics associated with the problem. Henceforth, a researcher can communicate and work collaboratively with similar researchers to achieve mutual research benefits and eventually publish the article with similar researchers as the co-authors. This leads to the situation where a researcher search for suitable co-authors with similar research interests to accomplish the research and publish a research study. It has been observed that the information about the researchers is available in multiple sources such as research publications, academic homepages and social networking sites. As there is a huge volume of information available, finding a suitable researcher partners is a challenging task. Therefore, co-author recommendation has gained importance which aims to determine similar co-researchers on a research topic and recommend them. In the existing approaches to co-author recommendation, either semantic or social similarity between researchers is considered. These two dimensions of similarity are rarely integrated. In this study, we use a combination of semantic and social similarity between researcher profiles to identify relevant similar researchers. Profile of a researcher is a list of knowledge areas in which he/she has a higher level of expertise. To construct the profile, expert knowledge areas identified from publication corpus using generative language model and research interests extracted from academic homepages have been used. The conceptual similarity between the researchers profiles are computed using DB pedia ontology where as the social relatedness is identified based on citation graph and conversations in Twitter on a research topic. Given a research topic and a researcher, an aggregated similarity score of concept similarity, social similarity and difference topic similarity between a researcher seeking a co-author and the other researcher is computed and ranked using the Lambda rank algorithm.

**Literature review:** Related entity finding is a task of finding entities similar to an entity given as input (Fang and Si, 2015). An instance of 'similar entity search' for the academic domain, considering researchers as entities have been addressed (Balog and Rijke, 2007). The approach develops models for computing similarities between researchers based on expert profiles. The expert profiles are built from their academic publications and homepages. Expert profiles are the description of the areas in which a researcher has a higher level of expertise. The manual way of extraction of the profile information of experts is difficult and time consuming. Sujatha describes two models for extracting expert profiles; the first uses information retrieval techniques to obtain a set of relevant documents for a given knowledge area and aggregates the relevance of those documents that are associated with the given person. The second model represents both candidates and knowledge areas as a set of keywords and

**Corresponding Author:** G. Manju, Department of Computer Science and Engineering , Anna University, Chennai , Tamil Nadu, India

the skills of an individual are estimated based on the overlap between these sets. Fazel and Fox (2011) presents a technique for generating evolving expert profiles of individuals. The profile is composed of the skills and competencies collected using heterogeneous data from diverse sources of information. Self- declarations, completed learning activities and previous research experience are used to generate the initial profile. Gollapall *et al.* (2011) shows a method to apply information extraction methodologies and extract structured data from the web. This structured data is further used to build the profile of a researcher.

Similarity between researchers can be found using semantic relatedness or social relationship among them. Sujatha finds similarities between researchers based on their profiles. It uses OKAPI BM25, KL divergence and probabilistic modeling to compute the conceptual similarities between experts. Researcher social network extraction aims at finding, extracting and fusing the 'semantic' based profiling information of a researcher from the Web. Previously, social network extraction was often undertaken separately in an ad-hoc fashion. Tang *et al.* (2008) gives a formal solution to the entire problem. Specifically, it identifies the relevant documents from the Web by a classifier. It then proposes a unified approach to perform the researcher profiling using Conditional Random Fields (CRF).

Identifying collaborators for a research topic is another instance of 'similar entity search'. As there is huge information available on the Web about the researchers and their research activities, it has led to information overload. There is a need for personalized recommendations of the researchers they can potentially collaborate with for mutual research benefits. Li *et al.* (2003) investigates this recommendation problem from two independent dimensions. They are the social relations between the researchers and the common expertise between them. The social relation is built based on the email communication and common expertise is identified based on the conceptual similarity between them. Xu proposes a novel researcher recommendation approach which combines the two dimensions. It builds the common expertise based on the conceptual similarity computed using WordNet as ontology. The social relation is built based on the co-author graph and e-mail communication between the researchers. It improves the effectiveness of researcher recommendation by combining the two dimensions of common expertise and social relation. Heck *et al.* (2011) has used bibliographic coupling and co-citation similarity to determine researcher similarity and recommend authors. Nykl *et al.* (2014) used citation network to rank the authors by eliminating the self citations and distributing proportional parts of publication values to the authors. The authors aimed to discover team leaders in research teams by recommending ranked list of authors. An approach has been introduced for suggesting potential collaborators for solving innovation challenges online, based on their competence, similarity of interests and social proximity with the user (Stankovic *et al.*, 2012). The approach uses Linked Data to derive a measure of semantic relatedness. It is also used to enrich both user profiles and innovation problems with additional relevant topics, thereby improving the performance of recommendation.

There are various methods to find semantic similarity between concepts. Thiagarajan focuses on computing the similarity between very short texts of sentence length. It presents an algorithm that takes account of semantic information and word order information implied in the sentences. The semantic similarity of two sentences is calculated using information from a structured lexical database and from corpus statistics. The use of a lexical database enables to model human common sense knowledge and the incorporation of corpus statistics allows this method to be adaptable to different domains. Xu computes the semantic similarity of concepts using WordNet, a well-developed lexical network for English words. The basic idea is that the more information two concepts share in common, the more similar they are. A combination of node-based approach of information content calculation and edge based approach of edge counting scheme is presented to compute the similarity.

Social relation between researchers is computed in various ways. Li *et al.* (2003) identifies the social network of researcher based on the co-author graph information. It is based on the assumption that more the number of times the authors co-author on a research topic, they are more likely to be close in the researcher network. Xu represents link between researchers based on some kind of social relationships such as e-mail communication, taking part in a project, collaborate with a study, etc. In the existing approaches to expert recommendation, the similarity between researchers is found either based on semantic relationship or based on the social relationship. But these two are rarely combined to identify similarity between researchers. Also, the social relationship among researchers is based on email communication and bibliographic networks. The most common social networking sites like Twitter or Face book have not been used. Moreover, the ontology (WordNet) used to compute semantic similarity is not domain specific. Therefore in this research, a novel co-author recommendation system which integrates semantic and social similarity is introduced. The semantic relatedness

between concepts is computed using DBpedia a computer science domain specific ontology. The social relationship between researchers is obtained from Twitter social network. Based on the semantic and social relatedness, an aggregated score is computed and the researchers are ranked as suitable co-authors using the Lambda rank algorithm.

## MATERIALS AND METHODS

A researcher (seed researcher) performing a research activity requires research partners to find solution to their research problem. Co-authors are identified based on their expertise in a research area, similarity of research interests and social proximity with other researchers. Figure 1 shows the architecture of the co-author recommendation system. The system uses three profiles namely concept profile, social profile and difference topic profile of the seed researchers to recommend the suitable co-authors. The entire procedure involved in the co-researchers recommendation system is shown in Algorithm 1. The subsequent sections describes in detail about the steps involved in the system.

**Conceptual profiling:** Generally, the profile of the researchers describes the research interests of the researcher and his expertise level. The researchers profile is available in the home page of the researcher. Hence, we extracted the research interest of the seed researcher from their respective homepages using conditional random field (Tang *et al*., 2008). based on the extracted research interest of the researcher, concepts are extracted using Alchemy API. Moreover the researchers' participation in the social network is also substantially increasing. The researchers communicate academic related information with other researchers. Hence, apart from using the academic profile information available in homepages, the academic discussions carried out using social network is also considered for building conceptual profile of the researchers. In this research, we consider the Twitter social network. From Twitter, we identified the academic tweets of the researchers (Manju and Geetha, 2013) and extracted academic concepts from them using the Alchemy API. Further, a concept profile of the seed researcher is created by combining the academic profile based concepts and social network based concepts. The created concept profile is represented as a Concept Profile vector (CP).

**Social profiling:** In a researcher homepage, publication details of the researchers are available. The co-authors in each publication of the researcher possess social bonding with the researcher. Hence, the co-authors are retrieved for a seed researcher from his homepage using Conditional Random Field (Tang *et al*., 2008). Similarly, the followers of the seed researcher are extracted from the Twitter social network. Further, the social profile of the seed researcher is created as a combination of the academic partners (co-authors) and social network followers and represented as a Social Profile vector (SP).

**Difference concept profiling:** A research working on a research problem is associated with a concept. A concept vector is created for the research problem. The difference between the concept profile vector of the researcher and the concept profile vector of the research problem is determined and created as a difference concept profile. The profile is represented as a difference concept vector (DP).

**Candidate researcher profiling:** The researchers in the Social Profile (SP) of each seed researcher are considered as candidate researchers. The concept profile, the Social profile and difference concept profile of the candidate researchers are created.

**Profile similarity computation:** Our research aims to recommend researchers (co-authors) for a given research problem of the seed researcher. Hence we used three similarity measures namely semantic concept similarity, social similarity and difference topic similarity.

**Semantic similarity of concepts:** This similarity measure (SC) represents the semantic similarity between concept profiles of the seed researcher and candidate researchers. The semantic similarity computation is carried out based on the domain ontology, DBpedia. The similarity measure proposed by Jiang is used for determining the similarity.

**Social similarity:** Social Similarity (SS) represents the represents the similarity between the social profile of the seed researcher and candidate researchers. The similarity computation is carried out using the Cosine similarity measure.

**Difference concept similarity:** This similarity measure (DC) is concerned with the difference between the concept profile of the seed researcher and the candidate researchers.

**Score aggregation and ranking model:** The different similarity scores computed between the seed researcher and candidate researchers are aggregated by computing the product of the similarity scores as shown in Eq. 1:

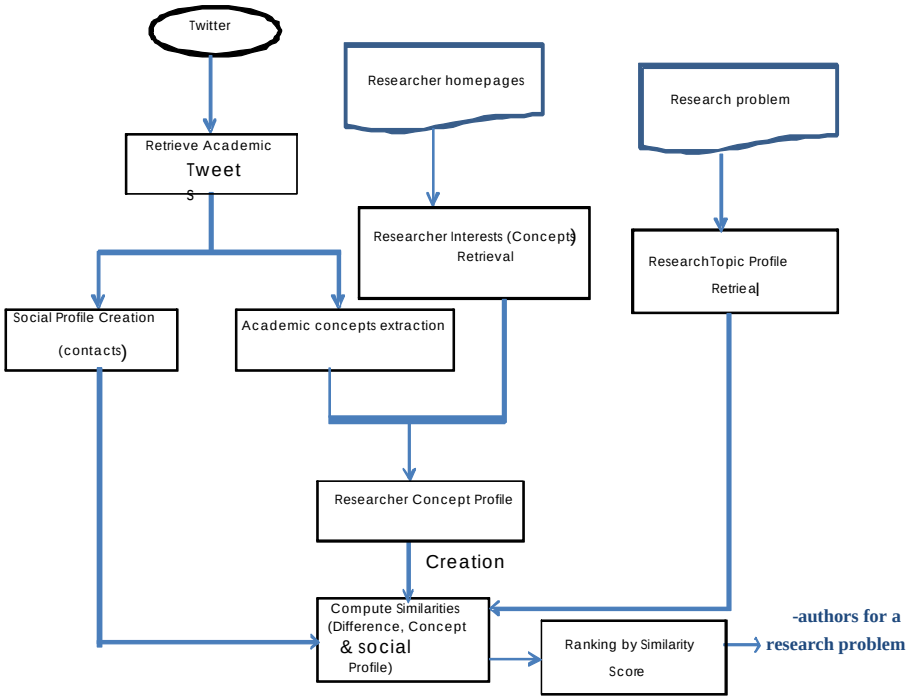


Fig. 1: Co-researchers recommendation

$$\text{Aggregated\_score} = SC \times SS \times DT \quad (1)$$

Further the aggregated score is ranked using the Lambda rank algorithm (Kavitha *et al*., 2014). This algorithm is basically a learning to rank algorithm that produces a ranking function as output based on the trained input. During the training stage, Lambda values are calculated between the seed researcher and each of its candidate researchers using the following equation:

$$\lambda = N\left(\frac{1}{1+e^{si-sj}}\right)\left(2^{li}\,2^{lj}\right)\left(\frac{1}{\log(1+i)} - \frac{1}{\log(1+j)}\right) \quad (2)$$

where, $S_i$ is the score of the researcher (co-author) ranked at position I and $S_j$ is the score of the researcher (co-researcher) ranked at position j. The researcher's (co-author) score is incremented or decremented by λ where the more relevant researcher (co-author) gets the positive increment. During testing, the learned ranking function is applied to determine the rank of a researcher (co-author) for a new research problem.

## RESULTS AND DISCUSSION

**Evaluation:** The evaluation of the recommendation process is done by conducting experiments with computer science academic publications and academic tweets from Twitter social network. The results show that combining semantic and social similarity between researchers provide better recommendations in identifying reviewers of a study and collaborators for a research problem. The following measures are used to measure the performance of the recommendation system.

**Algorithm 1: Recommend Co-researchers:**

Input: Profile of a research problem, a researcher (seed user), his tweets and contacts in Twitter
Output: List of Co-authors
Build Concept profile of the researchers
CP(researcher) = {(C1,w1),...,(Cn ,wn )}
Build Social profile of the researchers
SP(researcher) = {user1,...,usern}
Identify difference topics
Identify candidate co-authors
Form Concept profile vector and Social Profile Vector of the seed user and candidate co-authors
Calculate similarity measure
Semantic similarity of concepts (SC)
Social similarity (SS)
Difference concept similarity(DC)
Aggregated_Score = SC *SS *DC
Rank by similarity measure using Lambda rank

**Precison, recall and F1 score:** In classification tasks, the termstrue positives true negatives, false positives andfalse negativescompare the results of the classifier under test with trusted external judgments. The termspositive and negative refer to the classifier's

prediction (sometimes known as theexpectation ) and the termstrue andfalse refer to whether that prediction corresponds to the external judgment (sometimes known as theobservation). Precision and recall are defined as follows:

$$\text{Precision} = \frac{tp}{tp+fp} \quad (3)$$

$$\text{Recall} = \frac{tp}{tp+fn} \quad (4)$$

A measure that combines precision and recall is the harmonic meanof precision and recall, the traditional F-measure or balanced F-score is defined as:

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (5)$$

**Precision@k and Mean Average Precision (MAP):** Precison@k (P@k) computes the fraction of relevant experts retrieved in the top k position. It is used when a user wishes only to look at the first k retrieved domain experts and is defined as follows:

$$P@k = \frac{r(k)}{k} \quad (6)$$

where, r(k) is the number of relevant authors retrieved in the top k positions. Mean Average Precision for a set of queries is the mean of the average precision scores for each query. It is defined as follows:

$$MAP = \frac{\sum_{q=1}^{Q} \text{Ave } P(Q)}{Q} \quad (7)$$

**Conditional random field tagging model:** The academic research interests of the computer science researchers are extracted from their Google scholar page using Conditional Random Fields (CRF) tagging model. In order to evaluate the accuracy of the tagging process, model was trained with Google scholar pages of 1000 researchers. Five different labels are assigned to various sequences of texts in the homepage. Label '0' is used to the researcher name, label '1' is used to mark the qualification, label '2' to mark the university affiliation, label '3' is assigned for research interests, label 4 represents the citation index . The trained tagging model is used to test and retrieve the research interests of newly entered researcher homepages. The accuracy of the trained CRF model is measured using Precision, Recall and F-Score. Figure 2 shows the precision, recall and F-score values for the various labels tagged by the CRF tagged model. It is evident that the CRF tagging model tags the different sequence of texts from academic homepages with a higher accuracy. Particularly, the label 3 which is the research interests tag is marked with an accuracy of over 92%.

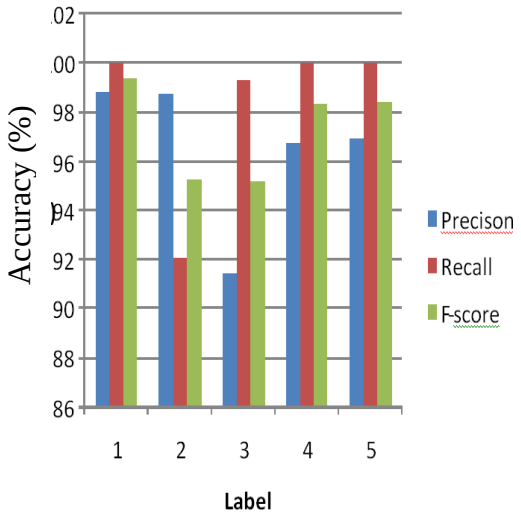


Fig. 2: Accuracy of CRF Tagging

**Semantic similarity (wordnet vs dbpedia):** The specific domain ontology defines the set of basic concepts comprising the vocabulary of the domain area and the relationships that exists behind these concepts. Other than using WordNet to capture the semantic relationships of researchers' expertise in the forms of words or phrases, the specific domain ontology (DBPedia) enables us to capture the domain knowledge of a specific area. An experiment is conducted to compare the similarity values found using WordNet and DBPedia. The WordNet similarity is calculated according to the method proposed by Lesk . It is based on the idea that the relatedness of two words is proportional to the extent of overlaps of their dictionary definitions. The similarity between concepts is calculated using DBPedia ontology as follows: A subset of DBpedia nodes representing a context of interest are identified. The nodes that are related by the relation dcterms:subject and skos:broader to the source node are collected. Wikipedia hyper textual links mapped in DBpedia by the property dbpedia-owl:wikiPageWikiLink is found and checked if the rdfs:label of given concept is

Table 1: Semantic Similarity (Wordnet vs Dbpedia)

| Concept pair | WordNet similarity | DBPedia similarity |
|---|---|---|
| 'Zend Engine', 'Zend Framework | '7.60 | 18.0966 |
| 'PHP', Delphi_for_PHP' | 4.23 | 9.2910 |
| 'PHP', 'JUnit' | 1.83 | 5.0915 |
| 'QPHP_Framework', 'PHP' | 3.67 | 6.5843 |
| 'Simple XML', 'PHP' | 4.51 | 6.1691 |

contained within the dbpedia-owl:abstract of related concepts. Table 1 shows a comparison of the semantic similarity values computed using WordNet and DBPedia.

**Co-author recommendation using different set of similarities:** The data set is extracted from Arnetminer.org,an academic search system which contains 1,436,990 authors and1,932,442 publications. We considered the following five sub-domains test cases data mining, theory, computer tetworks, visualization, database We used the recommendations by Arnetminer as ground truth and evaluated the effectiveness of using the different set of features in terms of P@k and MAP. The results are shown in the Table 2. The results show that the combined social and semantic similarity produce better accuracy compared to individual level of similarities and identifies better co-authors for a research problem.

**Lambda rank:** We evaluated the performance of Lambda rank algorithm in recommending co-authors for a research problem using NDCG measure. This measure is used to assess the relevance level of researchers based on their position in the rank list produced by the system for some topic. It works on the assumption that highly relevant researchers are more significant than marginally relevant ones. It is given by the following equation

$$NDCG_p = Z_p \sum_{i=1}^{p} \frac{2^{reli}-1}{\log(1+i)} \quad (8)$$

Where

$Z_p$ = The normalization constant

p = The truncation level (for example, if we want to retrieve only the top 6 authors of returned results, we consider p = 5)

$rel_i$ = The label of the ith positioned researcher

The experiment, we considered truncation level as '5' and evaluated the performance of Lambda rank. Figure 3 shows the NDCG results obtained before and after applying the Lambda rank. The number of items to rank

Table 2: Precision values for 4 different cross domains

| Sub-domains | Similarity | P@5 | P@10 | P@20 | MAP |
|---|---|---|---|---|---|
| Data | Concept | 0.3471 | 0.2500 | 0.2176 | 0.2715 |
| mining | concept+social | 0.4471 | 0.2735 | 0.2255 | 0.3153 |
| Computer | Concept | 0.3059 | 0.2020 | 0.1629 | 0.2236 |
| Networks | concept+social | 0.3529 | 0.2235 | 0.1528 | 0.2430 |
| Computer | Concept | 0.3412 | 0.2371 | 0.1794 | 0.2525 |
| Networks | concept+social | 0.3882 | 0.1980 | 0.1824 | 0.2562 |
| Visualization | Concept | 0.3118 | 0.1932 | 0.1762 | 0.2270 |
| | concept+social | 0.4632 | 0.2145 | 0.1987 | 0.2921 |

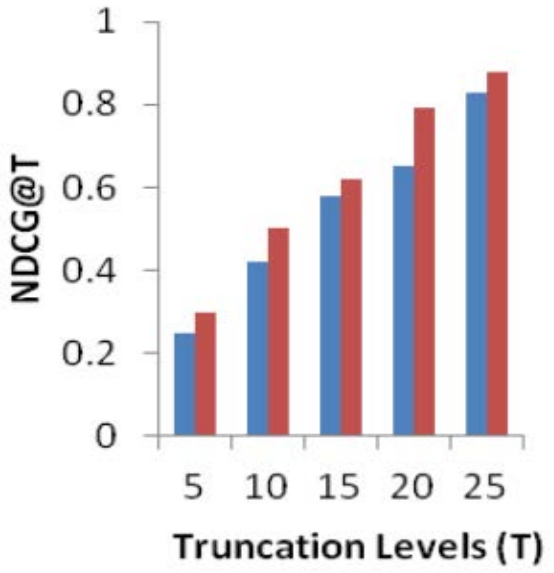


Fig. 3: Comparison of NDCG values before and after λ rank

after applying the learning to rank method. An optimized ranking of the researchers is obtained using the different ë values learnt using the ranking function. The ranked results illustrate that, learning to rank method produces better ranking results.

## CONCLUSION

In this research, we developed a system to recommend suitable co-authors for a research problem. We used the Concept profiles, Social profiles and Difference topic profiles to recommend co-authors. The concept profile has been created in a novel way by combining the academic researcher profile concepts and Academic Tweet concepts of the researcher. Similarly, the social profile is based on both the academic and social network followers. Further, the similarity between the profiles has been computed. In specific, we have determined semantically the concept profile similarity between researchers using the domain ontology,

DBpedia. The three similarity scores are aggregated and ranked using the Lambda rank algorithm. The proposed techniques are evaluated on two publicly-available datasets ACL corpus and Arnetminer. Lambda rank produced better ranked results of co-authors. The system can be extended by considering other researcher profile features like the impact of publication venue, researcher affiliation and designation. Furthermore, the system can be enhanced to handle the incremental updates of the researcher profiles and recommend the co-authors based on the updated profile. In addition, other than the academic tweets of the individual researcher even the academic groups in Twitter can be used as a source for computing the social relatedness of researchers.

## REFERENCES

Balog, K. and D.M. Rijke, 2007. Determining expert profiles (with an application to expert finding). IJCAI., 7: 2657-2662.

Fang, Y. and L. Si, 2015. Related entity finding by unified probabilistic models. World Wide Web, 18: 521-543.

Fazel, Z.M. and M.S. Fox, 2011. Constructing expert profiles over time for skills management and expert finding. Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, September 7-9, 2011, ACM, Graz, Austria, ISBN: 978-1-4503-0732-1, pp: 5-5.

Gollapalli, S.D., C.L. Giles, P. Mitra and C. Caragea, 2011. On identifying academic homepages for digital libraries. Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, June 13-17, 2011, ACM, Ottawa, Canada, ISBN: 978-1-4503-0744-4, pp: 123-132.

Heck, T., I. Peters and W.G. Stock, 2011. Testing collaborative filtering against co-citation analysis and bibliographic coupling for academic author recommendation. Proceedings of the 3rd ACM Workshop on Recommender Systems and the Social Web RecSys'11, October 23-23, 2011, ACM, Chicago, Illinois, USA., pp: 16-23.

Kavitha, V., G. Manju and T.V. Geetha, 2014. Learning to rank experts using combination of multiple features of expertise. Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), September 24-27, 2014, IEEE, New Delhi, India, ISBN: 978-1-4799-3078-4, pp: 1053-1058.

Li, Y., Z.A. Bandar and D. McLean, 2003. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowledge Data Eng., 15: 871-882.

Manju, G. and T.V. Geetha, 2013. Concept Similarity Based Academic Tweet Community Detection using Label Propagation. In: Mining Intelligence and Knowledge Exploration. Rajendra, P. and T. Kathirvalavakumar (Eds.). Springer International Publishing, Berlin, Germany, ISBN: 978-3-319-03843-8, pp: 677-686.

Nykl, M., K. Jezek, D. Fiala and M. Dostal, 2014. PageRank variants in the evaluation of citation networks. J. Inf., 8: 683-692.

Stankovic, M., M. Rowe and P. Laublet, 2012. Finding Co-Solvers on Twitter, with a Little Help from Linked Data. In: The Semantic Web: Research and Applications. Simperl, E., P. Cimiano, A. Polleres, O. Corcho and V. Presutti (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-642-30283-1, pp: 39-55.

Tang, J., J. Zhang, L. Yao, J. Li and L. Zhang et al., 2008. Arnetminer: Extraction and mining of academic social networks. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2008, Las Vegas, Nevada, USA., ISBN: 978-1-60558-193-4, pp: 990-998