Publication Venue Based Language Modeling for Expert Finding

Ali Daud and Sabir Hussain

Department of CS & SE, International Islamic University, Islamabad, Pakistan 44000

ali.daud@iiu.edu.pk, sabbir.iiui@gmail.com

Keywords: Expert Finding, Language Models, Co-Author Networks.

Abstract. Expert finding is hot topic discussed in co-author networks. Traditional language models only compares query terms with the documents of candidate for expert finding and ignores venues (conferences or journals) in which the paper is published. In this paper we propose novel influence language models which consider the importance of venues in which the papers of candidates are published. If the paper is published in a high level venue and another is published in a low level venue then these two papers should not have same weight-age for finding experts. The paper which is published in high level venue is more valuable than the paper which is published in low level venue. Experimental results show that our proposed models outperform the existing models.

1. Introduction

Finding the expertise of a candidate in specific fields is an important task in both academic and non-academic domains. It is one of challenging field in co-author networks. The idea in this paper is based on the fact that the man is known by the company he keeps. In expert finding context it means the better the venue in which an author is publishing paper will have more chances to be ranked as an expert. For example, in an organization for a particular project if we choose appropriate persons who have skills and knowledge about that project then the project will confidently be done successfully. But, how can we find the right persons for project? How to find the significant scientists for specific research areas? How to find an expert mentor? etc.

Much research work has been done to deal with these challenges. Especially, TREC platform [2] for finding experts in different domains has gained a lot of attention recently. The main solutions provided for this problem can be considered as three types, such as, (1) the language models based on text similarity [1,8] (2) graph based linkage methods based on co-author and co-citation relationships [5,7] and (3) latent topic layer based methods which exploits texts semantics [3,6].

Language models are used to find experts for a specific topic. These methods calculate scores for each of the candidate by just comparing query terms with documents terms of the candidate, without considering the venue of the document in which it is published. In our thinking importance of venue is much important in calculating scores for the publication. So we are proposing influence language models that use entropy to check how important the venue of the publication is, and then calculate the scores for that particular document according to frequency of query terms occurring in the document and importance of its venue. Results and discussions show that our proposed methods significantly outperformed existing language models.

2. Influence Models for Expert Finding

In influence language models we consider the importance of venues in which the paper is published. To check level of venue we calculate entropy of venue. Entropy means disorder-ness. High-level venues has low entropy as they have similar papers to specific topics and low level venues has high entropy as papers other than specific topics are also accepted. It is a known fact that if paper is published in high level venue automatically its will have large number of citations

and will be an important paper. So through entropy we were able to consider impact of number of citations for a paper.

$$Entropy\ (v) = \_\sum_{i=1}^{m} w_i log_2(w_i) \quad (1)$$

Where, w_i is the probability of $word_i$ in a venue v. Table 1 provides High level and low level venues with their entropies.

Table 1. Entropy of Publication Venues.

High Level Venues	Entropy	Low Level Venues	Entropy
SIGIR	1.74	CBMS	1.99
SIGMOD	1.71	CIARP	1.93
SIGSOFT	1.73	CODES	1.97

The venues that have entropy less than 1.8, we consider them high level and those that have entropy more than 1.85 are considered low level. In influence language models we multiply $P(q|e)$ with entropy but not simply multiply but first for high level venues we multiply entropy with 3 (as high level venues has smaller value of entropy but we have to give them high scores so we multiply entropy with 3) and then multiply the resulted entropy value with $P(q|e)$ and for the low level venues we first divide the entropy with 3 to decrease the entropy value and then multiply the resulted entropy value with $P(q|e)$. The venues for which the value of entropy is in between 1.8 and 1.85 we simply multiply $P(q|e)$ with entropy. Composite model proposed in [1] doesn't consider venues influence.

2.1 Influence Composite Language Model

In influence composite language model we find $P(q|e)$ same as composite model given in [1] then it is multiplied by the entropy of venue in which papers of candidate are published, and then according to the status of entropy we apply one of the following equations to find the scores for the expert.

Final equation if entropy is less than 1.8

$$P(q|e) = [\{\sum_{d_j \in De} P(d_j|e) \prod_{t_i \in q} P(t_i|d_j)\} * (Entropy * 3)] \quad (2)$$

Final equation if entropy is greater than 1.85

$$P(q|e) = [\{\sum_{d_j \in De} P(d_j|e) \prod_{t_i \in q} P(t_i|d_j)\} * (Entropy/3)] \quad (3)$$

Final equation if entropy is greater than 1.8 and less than 1.85

$$P(q|e) = [\{\sum_{d_j \in De} P(d_j|e) \prod_{t_i \in q} P(t_i|d_j)\} * (Entropy)] \quad (4)$$

2.2 Influence Hybrid Language Model

In influence hybrid language model we find $P(q|e)$ same as hybrid model given in [1] then it is multiplied by the entropy of venue in which papers of candidate are published, and then according to the status of entropy we apply one of the following equations to find the scores for the expert.

Final equation if entropy is less than 1.8

$$P(q|e) = [\{\prod_{t_i \in q} \sum_{d_j \in De.} P(t_i|d_j) P(d_j|e)\} * (Entropy * 3)] \quad (5)$$

Final equation if entropy is greater than 1.85

$$P(q|e) = [\{\prod_{t_i \in q} \sum_{d_j \in De.} P(t_i|d_j) P(d_j|e)\} * (Entropy/3)] \quad (6)$$

Final equation if entropy is greater than 1.8 and less than 1.85

$$P(q|e) = [\{\prod_{t_i \in q} \sum_{d_j \in De.} P(t_i|d_j) P(d_j|e)\} * (Entropy)] \quad (7)$$

where, *P(t/d)* is calculated using the Eq 5. Each candidate *e* has documents De = {dj} and each document d_j is treated separately and the results of all the documents of a candidate *e* are combined later. $P(d_j|e)$ shows the relationship of document d_j with the candidate *e* and the $P(t|d_j)$ is the probability with which query *q* terms are generated from the document for both the proposed models given in section 2.1 and 2.2.

3. Experiments

3.1 Experimental Setup

The dataset is taken from the DBLP online publication database [4]. Total number of paper and authors are 100,000 and 20,000, simultaneously. Standard text preprocessing steps are performed on title of papers and authors names such as, stop word removal, lower casing words, and removing words and authors less than three.

There is no standard ranked list of authors for queries typed by users for expert finding which limits to precision and recall performance evaluation metrics. We performed evaluation in terms of citations received on Google scholar by the author papers for all queries. The more the citations for top ten authors for queries of a method the more it is better. The state-of-the-art language models named composite and hybrid language models [1] are used as baselines in these experiments.

3.2 Results and Discussions

The abbreviations used are the following. Name (Author / Expert Name), Cit (Average) Citations, CoLM (Composite Language Model), InCoM (Influence Composite Model), HyLM (Hybrid Language Model), and InHM (Influence Hybrid Model).

Table 2. Results for query statistical analysis.

CoLM	**Name**	**Cit**	***InCoM***	**Name**	**Cit**	***HyLM***	**Name**	**Cit**	***InHM***	**Name**	**Cit**
0.99	farid n. najm	**213**	0.99	alin dobra	**224**	0.99	davidblaauw	**67**	0.99	davidblaauw	**67**
0.61	davidblaauw	**67**	0.51	stefanconrad	**394**	0.31	vladimirzolotov	**61**	0.41	jiaweihan	**264**
0.49	noel menezes	**51**	0.46	dawson r. engler	**116**	0.25	dennissylvester	**44**	0.40	peter johnson	**77**
0.47	vladimirzolotov	**61**	0.35	farid n. najm	**213**	0.24	peng li	**98**	0.37	dengcai	**111**
0.44	kavirajchopra	**63**	0.22	davidblaauw	**67**	0.17	farid n. najm	**213**	0.35	alexander aiken	**141**

Table 2 shows the top five expert's results (with number of citations and scores) for query statistical analysis. The author "Farid N. Najm"and "Alin Dobra" are top ranked experts for CoLM and InCoM models, respectively. Farid N. Najm have total of 26 papers and the frequency of query words statistical and analysis is 3,which means in 3 papers both of the query words are present where as "Alin dobra" has only 8 papers and in these papers frequency of both statistical and analysis is 1, though expert elected by influence composite language model have less frequency for query words but on the other hand the venues in which papers of Farid N. Najm are published are DAC,FPGA,ICCAD,ISLPED,ISQED in which 11,1,9,3,2 papers are published, respectively. When we calculate entropy for the above venues it results 1.90,1.92,1.95,1.99,1.96 for DAC,FPGA,ICCAD,ISLPED,ISQED, respectively, so it becomes clear that all of these venues are of low level as there disorder-ness rate is so high, the venues in which papers of "Alin dobra" are SIGMOD,FOCS,INFOCOM,PODS,VLDB in which 4,1,1,1,1 papers are published and the entropy of these venues is 1.71,1.68,1.92,1.71,1.74, respectively, except one which is INFOCOM all other venues have very smaller values for disorder-ness and are of high-level.

Above was all about composite language models now let's move towards hybrid language models, best expert chosen by simple hybrid language model and influence hybrid language model is same which is "David Blaauw" this happens because frequency for query words in the papers of "David Blaauw" is highest and the level of venues in which his papers are published is also high, but here is difference between the secondly chosen experts by both of the models according to simple hybrid model "Vladimir Zolotov" is expert and according to influence hybrid model "Jiawei Han" is expert. In papers of "Vladimir Zolotov" query word statistical occurs 8 times and analysis occurs 9

time where as in “Jiawei han’s” papers statistical occurs 1 time and analysis occurs 8 times but venues of “Vladimir Zolotov” are DAC,ICCAD,ISPD,ISQED which are all low level venues whereas venues of jiaweihan are SIGIR,SIGMOD,KDD,PKDD,SIGSOFT,VLDB which are all high level venues difference is clear from average number of citations of both of these experts average citations of “Vladimir Zolotov” are 61 and average citations of “Jiawei Han” are 264.
As experts elected by proposed models belong to high level venues that’s why there average citations are much higher than those belonging to low level venues, this happens because a man is known by the company he keeps.
Table 3 shows the overall results of 10 queries which are used in these experiments. It is clear that for both composite and hybrid modeling more number of average citations are obtained by top ranked authors due to the addition of conference influence.

Table 3. Average Citation Results for 10 queries.

LM	Average citations	LM	Average citations
CoLM	930.0	HyLM	715.5
InCoM	**968.25**	InHM	**965.5**

4. Conclusions

In this work we proposed influence language model on the basis of idea that better venues are topic specific and of higher quality. Multi-topic venues are not topic specific and are of lower quality as compared to the topic specific. It is proved from the results that influence language modeling outperforms existing language modeling methods.

5. Acknowledgements

The work is supported by Higher Education Commission (HEC), Pakistan startup research grant under Interim Placement of Fresh PhDs program 2011.

References

[1] K. Balog, L., Azzopardi, M.de., Rijke. Formal models for expert finding in enterprise corpora. *In Proceedings of the ACM SIGIR*, pp. 43–55, 2006.

[2] N. Craswell, A. P. D. Vries. Overview of the TREC-2005 enterprise track. *In Proceedings of TREC,* 2005.

[3] A. Daud, J. Li, L. Zhou, F. Muhammad. Temporal expert finding through generalized time topic modeling. *Knowledge-Based Systems (KBS),* 23(6): 615-625, 2010.

[4] DBLP bibliography database, http://www.informatik.uni-trier.de/~ley/db/.

[5] Y. Fu, R. Xiang, Y. Liu, M. Zhang, S. Ma. Finding experts using social network analysis. *In Proceedings of Web Intelligence*, pp. 77-80, 2007.

[6] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su. ArnetMiner: extraction and mining of academic social networks. *In Proceedings of the 14th ACM SIGKDD*, pp. 990-998, 2008.

[7] W. Wei, A. Bargiela, P. Barnaghi. Rational research model for ranking semantic entities. *Information Science Journal*, 181(13):2823-2840, 2010.

[8] J. Zhu, X. Huang, D. Song, S. R¨uger. Integrating multiple document features in language models for expert finding. *Knowledge and Information Systems Journal*, 23(1):29-54, 2009.