ISSN 1870-4069

Topic Modeling Based Analysis of Professors Roles in Directing Theses

Mahdi Mohseni, Heshaam Faili

University of Tehran,
Department of Electrical and Computer Engineering,
Iran

{mahdi.mohseni, hfaili}@ut.ac.ir

Abstract. Topic-sensitive analysis of participants in scientific networks has been a hot research topic for many years. In this paper, we propose a topic modeling-based approach to analyze the influence of professors when they take on different roles (e.g., first supervisor, second supervisor and advisor) in directing theses. We explain how a topic distribution of theses obtained by Latent Dirichlet Allocation provides a basis to formulate the problem and how the unknown parameters are calculated. Results of experiments on a real-world dataset reveal some interesting facts about professor roles in supervising and advising theses. Our results are in contradiction with this traditional view that supervisors are more influential than advisor and first supervisor is a more effective role than second supervisor.

Keywords: Influence analysis, topic modeling, gradient descent.

1 Introduction

Influence analysis of professors in directing theses provides important information which can be used for ranking professors and modifying education regulations in academic institutions. Professors who supervise students in universities have the authority to determine the direction of theses. If supervision is performed by only one professor, he is the only supervisor and has all authority.

But when more than one professor has been involved in a thesis and they take on different roles, e.g., first supervisor, second supervisor and advisor, how can one analyze the influence of each role? One key approach is to make use of topic modeling to achieve topic distribution of theses as a basis for further analyses. Topic modeling has been utilized in this way in many researches.

Studying the problem of identifying influential users of the Twitter network, Weng et al. (2010) utilize topic modeling to identify topics that users of the network are interested in and accordingly a topic-specific network of Twitter users are created.

Then, a method which is an extension of PageRank is applied to find influential users. In (Ramage et al., 2009), topic modeling is applied to PhD dissertation abstracts to find which universities are leading and which are lagging. Underlying topics of the abstracts extracted using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are

pp. 151–159; rec. 2017-02-05; acc. 2017-04-12

provided to a temporal similarity measure to examine if a university lead or lag other universities.

In this paper, we intend to utilize a standard topic model, namely LDA (Blei et al., 2003), to analyze professor roles in directing theses in a real-world dataset. The dataset consists of graduate theses in the School of Electrical and Computer Engineering at University of Tehran. Based on regulations of this university, a professor can be a supervisor or an advisor in directing a thesis.

Moreover, each thesis can be directed by one, two or, in rare cases, three supervisors and zero, one or, again in rare cases, two advisors. Traditionally, it is supposed that supervisors have more influence on leading a thesis than advisors and similarly first supervisor is a more influential role than second supervisor, and so on. The results of proposed topic modeling-based analysis method contradict this traditional point of view about directing theses.

The innovations of this paper are as follows:

- Proposing a topic modeling-based approach to analyze the influence of professors in directing theses.
- Proposing an optimization function for influence identification of professor roles.
- Determining the influence of professor roles in the School of Electrical and Computer Engineering of Tehran University.

The rest of the paper is organized as follows: in section 2 related works are reviewed; section 3 describes the proposed methodology; experiments and analysis are presented in section 4; and finally, the paper is concluded in section 5.

2 Related Works

Topic models have been successfully applied to identify topics buried in text documents. Since the seminal paper on LDA by Blei et al. (2003), this model has been either a basis for many intuitive approaches or an inspiring method in proposing other topic modeling algorithms to analyze the influence of participants in scientific networks.

TwitterRank (Weng et al., 2010) is an extension of PageRank algorithm to measure the influence of users in Twitter. First, topics that twitterers are interested in are automatically identified according to the tweets they published using LDA model. Then, based on the topics distilled, topic-specific relationship networks among twitterers are constructed. Finally, TwitterRank algorithm is applied to measure the influence taking both the topical similarity between twitterers and the link structure into account.

Using standard topic models in microblogging environments is studied in (Hong et al., 2012). In this paper, a few schemes are proposed to train a topic model on twitter and to compare their quality and effectiveness through experiments on two tasks: predicting popular twitter messages and, classifying twitter users and corresponding messages into topical categories.

In (Hall et al., 2008), topic modeling is applied to ACL Anthology to analyze the development of ideas and trends in the field of computational linguistics over the course of the time. Topic distributions achieved by applying LDA to papers are used to determine the temporal distribution of topics over years by some post hoc calculations. Ramage et al. (2009) focus on this problem that based on the content of PhD dissertation abstracts, which universities are leading and which are lagging.

To answer this question, underlying topics of PhD dissertation abstracts are extracted using LDA and then by defining a similarity score which is based on topic similarity over years they examine how much a university matches the future or past of other universities. In a similar research, Shi et al. (2010) present an approach to analyze the lead-lag relationships of communities based on topic modeling and temporal information.

Based on this hypothesis that the context in which a cited document appeared in a citing document indicates how the cited authors have influenced the contributions by the citing authors, Kateria et al. (2011) propose two models, author link topic and the author cite topic models to simultaneously model the content of documents, and the interests as well as the influence of authors in certain topics.

Topic-Link LDA (Liu et al., 2009) has been proposed to jointly model the document topics and the social network among authors in one unified model. This model assumes the formation of a link between two documents as a combination of topic similarity and community closeness, which brings both topic modeling and community discovering in one unified model.

Author-Conference-Topic (ACT) (Tang et al., 2008) is a model to simultaneously extract topics of papers, authors, and conferences. Topical Affinity Propagation (TAP) (Tang et al., 2009) has been presented to model topic-level social influence on large networks. Inspiring ACT model (Tang et al., 2008), Kim et al. (2016) propose a model called Author-Journal-Topic (AJT) to take both authors and journals into consideration of topic analysis.

In the next section we will describe how we use topic modeling as the initial step of our approach to determine how professors affect directing graduate theses and how influential the role of each professor is.

3 Methodology

Our approach in analyzing the influence of professors in directing theses consists of several steps which will be explained in detail in this section.

3.1 Topic Modeling

In the first step, the topic distributions of thesis abstracts are needed to be extracted. To do that, LDA which is a standard topic model for extracting underling topics of documents is employed.

In this model, each document is represented by a multinomial distribution of topics and each topic is represented by a multinomial distribution over words. Parameters of the LDA model are α and β, the hyperparameters of Dirichlet distributions, and K, the number of topics. The resulted topic distributions are used for the succeeding computations.

3.2 Research Topics Identification

LDA is an unsupervised topic modeling approach to detect topics in documents. Consequently, all of detected topics in thesis abstracts do not characterize research fields and some of them are stop or common words. Hence, topics are examined manually to remove irrelevant ones. Then, the topic distribution of each document is normalized to make sure that the probability distribution is still valid. Remaining relevant topics each can be assigned a title to describe the corresponding research area that has been followed by researchers.

3.3 Problem Formulation

Education regulations of University of Tehran allow that a thesis is directed by more than one professor. According to that these roles are definable for involved professors: first, second and third supervisor, and first and second advisor. Two roles, third supervisor and second advisor, which rarely occur are ignored. It has been taken for granted that in directing theses supervisors are more influential than advisor and first supervisor is a more effective role than second supervisor. To scrutinize how influential these professor roles are two effective factors should be taken into account: 1) the role of involved professors and, 2) the match between topics of a thesis and the research record of professors.

If $\alpha(prof, thesis)$ is the influence of a professor role, with regard to different definable roles:

$$\alpha(\text{prof}, \text{thesis}) = \begin{cases} 1 & if\ thesis\ has\ only\ one\ supervisor \\ \alpha_1, \alpha_2 & if\ thesis\ has\ one\ supervisor\ and\ one\ advisor \\ \alpha_3, \alpha_4 & if\ thesis\ has\ two\ supervisors \\ \alpha_5, \alpha_6, \alpha_7 & if\ thesis\ has\ two\ supervisors\ and\ one\ advisor \end{cases}. \quad (1)$$

In which the influence of each role is defined for all possible states. The influence of supervisor is 1 if $thesis$ is supervised by only one professor. α_1 and α_2 are respectively the influence of supervisor and advisor when $thesis$ is directed by one supervisor and one advisor ($\alpha_1 + \alpha_2 = 1$).

If $thesis$ is supervised by two supervisors, α_3 and α_4 are the influence of the first and second supervisors ($\alpha_3 + \alpha_4 = 1$). Similarly, in cases that two supervisors and an advisor are involved in $thesis$, α_5, α_6 and α_7 are the influences of them, respectively ($\alpha_5 + \alpha_6 + \alpha_7 = 1$). Based on this definition, as long as $0 \leq \alpha_i \leq 1\ \forall\ i = 1, \ldots, 7$ and in each case the influences sum up to 1, these values are valid. In our experiments we initialize $\alpha_i\ \forall\ i = 1, \ldots, 7$ with regard to following conditions: $\alpha_1 \geq \alpha_2$, $\alpha_3 \geq \alpha_4$, $\alpha_1 \geq \alpha_2$ and $\alpha_5 \geq \alpha_6 \geq \alpha_7$.

Similarity of a thesis' topics and research interest of supervisor(s) and advisor is another factor in influencing a thesis. The more the topic distribution of a thesis with research records of a professor is matched, the more influence that professor has had in directing that thesis. $p^{year}(topic|prof)$ shows how much a professor has experience in a topic up to a specific year. If $p(topic|thesis)$ is the probability that an arbitrary thesis is about a specific topic, resulted from applying LDA to all theses, the cumulative experience of a professor in that topic in year $year$ is defined as:

$$f^{year}(topic|prof) = \sum_{thesis \in year} \alpha(\text{prof}, \text{thesis}) \times p^{year-1}(topic|prof) \times p(topic|thesis), \tag{2}$$

If $f^{year}(topic|prof)$ is normalized, $p^{year}(topic|prof)$, the experience of a professor in topic $topic$ in that year is obtained:

$$p^{year}(topic|prof) = \frac{f^{year}(topic|prof)}{\sum_{topic'} f^{year}(topic'|prof)}. \tag{3}$$

In formula (2) the unknown parameters, i.e. $\alpha_{\mathrm{i}}\ \forall\, i = 1, \dots, 7$, are required to be determined.

3.4 Finding Parameters Values

Research interests of professors might change by the passage of the time. However, sharp changes are rarely observed in professors' research directions. They usually get familiar with new areas that they have not worked on and gradually shift to them. This fact is our hypothesis in defining an objective function to determine the unknown parameters of formula (2), the influence of professor role. The objective function is defined as the changes of professors' experiences in consecutive years:

$$F = \sum_{year=2}^{n} \sum_{prof} \sum_{topic} (p^{year}(topic|prof) - p^{year-1}(topic|prof))^2. \tag{4}$$

By minimizing the above objective function with respect to $\alpha_{\mathrm{i}}\ \forall\, i = 1, \dots, 7$, these unknown parameters can be determined. As this objective function is complex, it is not easy to arrive at an analytical solution to solve it. So, we choose to find the parameters values by gradient descent. In an iterative procedure, the value of each parameter is changed by $\pm\varepsilon$ and F is calculated. In a direction that F is decreased the new value of the parameter is determined which is either raising or falling by ε. This will be repeated until there is no significant change in F value.

4 Experiments and Analysis

4.1 Dataset

The abstracts of graduate theses in the School of Electrical and Computer Engineering of University of Tehran between years 2006-2015 are used in our experiments. The number of theses in these years is 1813.

As it was mentioned before, the rare cases in which a thesis is directed by 3 supervisors or two advisors are removed from the dataset. The number of these theses is 36. The information of the remaining 1717 theses is shown in Table 1.

Table 1. Statistics of the dataset.

Item	Number
All theses	1777
Theses directed by only one supervisor	902
Theses directed by one supervisor and one advisor	542
Theses directed by two supervisor and no advisor	251
Theses directed by two supervisor and one advisor	82

4.2 Experiments

Preprocessing: first of all, the dataset should be normalized, tokenized and lemmatized. Moreover, stop words are needed to be removed. To do that, Persianp Toolbox (Mohseni et al., 2016), which is a free Persian text processing toolbox is used.

Topic Modeling: to obtain the topic distribution of documents and the word distribution of topics, JGibbLDA[1] which is a java implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling technique is used. The hyperparameters of Dirichlet distributions, α and β, are both set to 0.01 and the number of topics, K, to 100. The number of iterations is also 1500.

Research Topics Identification: According to 20 top words in the word distribution of each topic, an expert detects which topics consist of stop or common words and so to be removed and which ones are represented research areas and to be assign proper titles. In our experiments, from 100 topics, 79 are detected as research topics and other 21 as irrelevant. In other word, one can claim that research interests of professors in the Department of X at the University of Y can be partitioned into these 79 research areas according to an LDA-based analysis.

Finding Parameters Values: To find the influence of professor roles in directing theses (ref. to sections 3-3 and 3-4), we should initialize $\alpha_i \ \forall \ i = 1, \dots, 7$ (formula (1)) with proper values according to explained conditions. The independent parameters are α_1, α_3, α_5 and α_6 and other parameters are dependent. As assigning roles to professors has not been done recklessly and with no purpose, we initialize the value of each independent parameter in a way that the influence of none of roles is ignored. Moreover, to reach a comprehensive result, the initial value of each parameter is changed in an appropriate range by small rate δ and the experiment is repeated for each new parameter initialization. Finally, the mean and variance of resulted values for each parameter are reported. The change rate of parameter values in computing the gradient descent is set to 0.01. Table 2 summarizes the parameter setting of the problem.

To calculate the cumulative experience of professors (formula (2)), topic distribution of professors, i.e. $p^{year-1}(prof|topic)$, for $year = 1$ is not available. Also, if a professor has had no role in directing theses in the first few years his research record is unknown. To solve these problems, we can initialize the research record of a professor using one of these two strategies: 1) the mean of topic distributions of all theses in all years that a professor has been involved in or, 2) the mean of the topic distribution of theses in the first N years since a professor had been assigned his first role. In our experiments as long as $N \geq 3$ the results of both strategies are the same.

[1] http://jgibblda.sourceforge.net/

Table 2. Parameter setting.

Independent parameters	α_1, α_3, α_5 and α_6
Dependent parameters	α_2, α_4 and α_7
Range of values for parameters	$0.5 \leq \alpha_1 \leq 0.8$ $\alpha_2 = 1 - \alpha_1$ $0.5 \leq \alpha_3 \leq 0.8$ $\alpha_4 = 1 - \alpha_3$ $0.25 \leq \alpha_5 \leq 0.4$ $0.25 \leq \alpha_6 \leq 0.$ $\alpha_7 = 1 - \alpha_5 - \alpha_6$
Change rate to calculate the gradient descent (ε)	0.01
Change rate in initialization of values (δ)	0.05

Table 3. Mean and standard deviation (Stdv) of parameters.

Thesis is directed by	Role	Parameter	Mean	Stdv
One supervisor and one advisor	Supervisor	α_1	0.49	0.01
	Advisor	α_2	0.51	0.01
Two supervisors	1st Supervisor	α_3	0.50	0.00
	2nd Supervisor	α_4	0.50	0.00
Two supervisors and one advisor	1st Supervisor	α_5	0.29	0.02
	2nd Supervisor	α_6	0.30	0.01
	Advisor	α_7	0.41	0.02

By examining professors involved in all theses, we recognized that some professors have been involved in only one or few theses. They are professors who have been either new faculty members of Electrical and Computer Engineering Department or from other departments.

Since we don't want that the lack of information about research records of these professors biases our experiments, in optimization procedure (ref. section 3-4) we do not take professors whose name are repeated less than 3 time into account. In this way, new faculty members and professors from outside the department have no bad effect on our calculations.

Based on the aforementioned parameter setting and the initialization of parameters, $\alpha_i \ \forall\, i = 1, \ldots, 7$ is calculated. Table 3 shows the mean and standard deviation of resulted values.

4.3 Discussion

By analyzing the results presented in table 3, some interesting conclusions can be drawn. Above all, low standard deviation of obtained values shows that the proposed approach is robust to initialization of parameters. This is interpretable as dependability of the approach.

In cases that a thesis is directed by two professors, either by one supervisor and one advisor or by two supervisors, their influence are almost the same. This is in contradiction to this traditional point of view that supervisor role has more influence than advisor role and first supervisor is also more influential than second supervisor.

Another interesting conclusion is that when a thesis is supervised by two supervisors and one advisor, advisor role overshadows supervisor roles. This can be justified that when in spite of two supervisors it is still needed that another professor plays a role, it means the thesis covers a topic which is in the expertise of none of supervisors.

5 Conclusions

In this paper professors' roles in supervising or advising graduate theses in the Department of X at the University of Y were studied. Proposing a topic modeling-based approach, we analyzed how professors affect theses and how much influential their roles are when more than one professor has been involved in directing a thesis. For the future work, we aim to use our achievements here in order to analyze the professors' network to study the leading professors and discover the change pattern of research interests in the department.

References

1. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet allocation. Journal of machine Learning research, vol. 3, pp. 993–1022 (2003)
2. Hall, D., Jurafsky, D., Manning, C. D.: Studying the history of ideas using topic models. In: Proceedings of the conference on empirical methods in natural language processing Association for Computational Linguistics, pp. 363–371 (2008)
3. Hong, L., Davison, B. D.: Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics, pp. 80–88 (2010) doi: 10.1145/1964 858.196487
4. Kataria, S., Mitra, P., Caragea, C., Giles, C. L.: Context sensitive topic models for author influence in document networks. Twenty-Second International Joint Conference on Artificial Intelligence, vol. 22, no. 3, pp. 2274 (2011)
5. Kim, H. J., An, J., Jeong, Y. K., Song, M.: Exploring the leading authors and journals in major topics by citation sentences and topic modeling. In: Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 42–50 (2016)
6. Liu, Y., Niculescu-Mizil, A., Gryc, W.: Topic-link LDA: Joint models of topic and author community. In: Proceedings of the 26th annual international conference on machine learning, pp. 665–672 (2009) doi: 10.1145/1553374.155346
7. Mohseni, M., Ghofrani, J., Faili, H.: Persianp: A persian text processing toolbox. In: International Conference on Intelligent Text Processing Linguistics, Springer, vol. 9623, pp. 75– 87 (2016) doi: 10.1007/978-3-319-75477-2_4
8. Ramage, D., Manning, C. D., McFarland, D. A.: Which universities lead and lag? Toward university rankings based on scholarly output. In: Proceeding of NIPS Workshop on Computational Social Science and the Wisdom of the Crowds (2010)
9. Shi, X., Nallapati, R., Leskovec, J., McFarland, D., Jurafsky, D.: Who leads whom: Topical lead-lag analysis across corpora. NIPS Workshop (2010)
10. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su., Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 990–998 (2008) doi: 10.1145/ 1401890.14020

11. Tang, J., Sun, J., Wang, C., Yang, Z.: Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 807–816 (2009) doi:10.1145/1557019.155710
12. Weng, J., Lim, E. P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining, pp. 261–270 (2010) doi: 10.1145/1718487.17185