Hue University Journal of Science: Techniques and Technology; pISSN 2588-1175 | eISSN 2615-9732

Vol. 131, No. 2B, 2022, 49–63; DOI: 10.26459/hueunijtt.v131i2B.6749

Topic diffusion prediction on bibliographic network: effect of topic modeling on activation probability measure

Thi Kim Thoa Ho[1]*, Quang Vu Bui[2]

[1] Hue University of Education, Hue University, Vietnam
[2] Hue University of Sciences, Hue University, Vietnam

Abstract. In this research, we propose using topic modeling to estimate activation probability for predicting topic diffusion on the bibliographic networks. We utilize the supervised method to predict the propagation of a specific topic. We propose a new method to calculate activation probability for an active node and an inactive node based on the meta-path and textual information using topic modeling. Firstly, based on textual information, topic modeling is suggested to measure activation probability, namely the textual information. Secondly, combining the meta-path and textual information, we propose a new method to estimate activation probability, namely the aggregated activation probability, in which the textual information is measured by topic modeling. We conduct experiments on dissimilar topics of the bibliographic network datasets. Experimental results demonstrate that topic modeling improves the accuracy of diffusion prediction compared with term frequency–inverse document frequency.

Keywords: activation probability, meta-path, bibliographic network, topic modeling

1 Introduction

Information diffusion is a process in which information is spread from one object to another through interactions. Information may be rumours, ideas, diseases, etc. The information diffusion process can be described as nodes that are considered active if they have already taken the action related to information. For example, a scientist is called ‘active’ with the topic “data mining” since he has carried out research and published articles on that topic, or a person ‘active’ with the virus covid-19 when he has been infected with this virus.

Information diffusion has been exploited in two kinds of networks: homogeneous networks [1–5] and heterogeneous networks [6–8]. Homogeneous networks are those containing only one type of object and one type of link. For instance, a co-author networks with an object author and a ‘co-author link’ or an object user and links ‘friendship’ on a friend network. Whereas, heterogeneous networks are those with dissimilar types of objects and relations. For example, a bibliographic network is a heterogeneous network with other objects, including authors, papers, venues, and affiliations, concurrently existing various relationships among

* *Corresponding: hothikimthoa@dhsphue.edu.vn*

Received: 29–03–2022; Accepted: 03–06–2022

authors, such as the co-author relation, the relation of common co-authors, the relation of participating in the same conference, and the relation in the same laboratory.

In our previous research [9], we concentrated on exploiting information propagation on a heterogeneous network. We studied topic prediction in the bibliographic network with a new approach that combined external factors and intrinsic factors. The supervised learning method was utilized to predict the spreading of a specific topic, where we combined the dissimilar features with the dissimilar measuring coefficient.

Firstly, we proposed a new method to estimate activation probability of an active node and an inactive node by combining the meta-path and textual information. Activation probability was estimated from the meta-path by using the Bayesian framework. Besides, the activation probability from textual information can be measured by using Term Frequency–Inverse Document Frequency (TFIDF) and cosine distance or using topic modeling and distance measures related to probability distribution. Finally, we proposed an aggregated activation probability (AAP) based on activation probability from the meta-path and textual information. This probability functioned as an external factor in activating an inactive node switched to an active state. Besides, we proposed an intrinsic factor that was the author's interest in the topic propagated. External and intrinsic factors were combined to predict the spreading of a specific topic. The experimental results show that aggregated activation probability with the combination of the meta-path and textual content enhanced the accuracy of the topic's diffusion prediction compared with the old activation probability that only used the meta-path information or textual information. Furthermore, the amalgamation between the aggregated activation probability and the author's interest in the topic obtained the highest accuracy.

Nevertheless, we only used TFIDF for measuring the activation probability based on the textual content and aggregated activation probability. In information retrieval systems, the Vector Space Model (VSM) [10] is a fundamental technique for textual analysis, where each document is represented by a word-frequency vector. However, VSM is highly dimensional because of the high number of unique terms in the text corpora and insufficient to capture all semantics. With VSM, we cannot capture the user's interest distribution on the topics. To overcoming this limitation, one of the possible solutions is to represent the text as a distribution of topics. This is the idea of topic modeling [11, 12] that identifies the distribution of latent topics in the text, which is useful in modeling interest distribution. The main idea of topic modeling is to create a probabilistic generative model for the corpus of text documents. Several methods of topic modeling, such as Latent Dirichlet Allocation (LDA) [11] and Author-Topic Model (ATM) [12], have been developed recently. Therefore, in this study, we continue using topic modeling to estimate both probabilities.

The experimental results demonstrate that utilizing topic modeling in measuring activation probability based on the textual content and aggregated activation probability provides higher accuracy in the topic's spreading prediction than using TFIDF.

The structure of our paper is organized as follows: Section 1 introduces the problem definition; Section 2 summarizes related works; Section 3 reviews preliminaries; our approach is proposed in Section 4; Section 5 illustrates experiments and results; we conclude our work in Section 6.

2 Related works

Information diffusion is the process by which a piece of information is spread from one individual or community to another on a network, also known as information propagation or information spreading. Recently, numerous researchers have investigated information diffusion, mainly concentrating on which information diffuses most quickly, which factors affect information diffusion, and which models are used to simulate and predict the propagation. These questions play a significant role in understanding the diffusion phenomenon. They have been answered by researchers in smaller branches of information diffusion, including epidemic spreading modeling, influence analysis, and predictive modeling.

The majority of research on information spreading has been conducted on homogeneous networks, where only one type of object and one type of link exist on the network. Nevertheless, in the real world, most networks are heterogeneous with various object types and multiple relations. For instance, a bibliographic network is a heterogeneous one that contains multiple objects, including authors, papers, venues, and affiliations. Besides, concurrently exist numerous relationships among authors, such as the co-author relation and the relation with a common co-author. Our study focuses on information propagation on heterogeneous networks.

For studying predictive models on heterogeneous networks, there were two main approaches for modeling and predicting information propagation. First, the spreading process is investigated with the linear threshold model (LT) [5, 14], independent cascade model (IC) [4], decreasing cascade model [2], general threshold model [1], heat diffusion-based model [15], etc. With this approach, some active nodes influence their inactive neighbours in the network and turn them into active ones. In IC, an inactive node can be infected by an active node with a certain probability. In LT, an inactive node is activated if and only if the total weight of its active neighbours is at least equal to a threshold. Besides, there are various expanded models based on IC, such as the Homophily Independent Cascade Diffusion model (TextualHomo–IC) [6] with an infected probability estimated based on the textual information or the Heterogeneous Probability Model–IC (HPM–IC) [7], where the infected probability is calculated according to a conditional probability based on meta-paths information. Besides, there are several expanded models based on LT, including Multi-Relational Linear Threshold Model–Relation Level Aggregation (MLTM–R) [8] or Probability Model–LT (HPM–LT) [7]. These models proposed methods to measure the infected probability of an inactive node based on meta-paths information or textual information separately. In addition, the influence factors from active neighbours were considered in the absence of the intrinsic factors of inactive nodes or other features, such as the interest level of the nodes to the topic or each node's influence. Therefore, the second approach appeared with the amalgamation of dissimilar features.

Utilizing supervised learning and deep learning to predict information spreading in a heterogeneous network is the second approach. Spreading a tweet on Twitter has been studied with the supervised learning method [16] that combines the user's interests and content similarity between an active user and an inactive user using latent topic information. Besides, information diffusion on Github has been studied by using supervised learning [17]. Furthermore, deep learning has been used to predict information propagation on a heterogeneous network [18]. Topic diffusion on a bibliographic network has been studied with the first approach by using dissimilar spreading models. This problem has also been investigated with the second approach by using the deep learning method [18]. However, the supervised learning method has not been utilized. Therefore, we focus on predicting topic diffusion on the bibliographic network by using the supervised learning method. Based on our previous results, in this study, we continue to propose using topic modeling to estimate activation probability with textual information instead of Term Frequency–Inverse Document Frequency.

3 Preliminaries

3.1 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) [11] is a generative statistical model of a corpus. In LDA, each document is considered as a mixture of different topics, and each topic is characterized by a probability distribution over a finite vocabulary of words. The LDA generative model is described with the probabilistic graphical model in Figure 1a. The LDA generative process for a corpus D consisting of M documents with a length of N_i each is as follows, where K denotes the number of topics:

Step 1: Choose distribution over topics $\theta_{i,\ i\ €\{1,\ \ldots,\ M\}}$ from a Dirichlet distribution with parameter α for each document.

Step 2: Choose the distribution over words $\varphi_{k, k\ €\{1,\ \ldots,\ K\}}$ from a Dirichlet distribution with parameter β for each topic.

Step 3: For each of the word position i, j, where j €{1, …,_N_i}, and i €{1, …, M}

3.1. Choose a topic z_{ij} from a Multinomial distribution with parameter θ_i

3.2. Choose a word $w_{i,j}$ from a Multinomial distribution with parameter $\varphi_{z_{ij}}$

The advantage of the LDA model is that interpreting at the topic level instead of the word level allows us to gain more insights into the meaningful structure of documents since noise can be suppressed by the clustering process of words into topics. Consequently, we can learn the topic distribution of a corpus, and then predict the topic distribution of an unseen document of this corpus by observing its words. The topic distribution can be used to organize, search, cluster, or classify the documents more effectively

Inference: The key problem in topic modeling is posterior inference. This refers to reversing the defined generative process and learning the posterior distributions of the latent variables in the model given the observed data. In LDA, this amounts to solving the following equation.

$$p(\theta, \emptyset, z|w, \alpha, \beta) = \frac{p(\theta, \emptyset, z, w|\ \alpha, \beta)}{p(w|\ \alpha, \beta)} \tag{1}$$

There are some inference algorithms available, including variational inference used in the original paper [11] and Gibbs sampling.

3.2 Author-topic model

The author-topic model (ATM) [12] is a generative model that represents each document with a mixture of topics, as in state-of-the-art approaches like LDA, and extends these approaches to author modeling by allowing the mixture weights for different topics to be determined by the authors of the document. The objective of the ATM model is to discover the patterns of word use and connect the authors who exhibit similar patterns. In the ATM, the words in a collaborative paper are assumed to be the result of a mixture of the authors' topics, where each author is associated with a mixture of topics, and topics are multinomial distributions over words. The generative model of ATM is described with a graphical model in Figure 1b as follows:

Step 1: Choose a group of authors a_d and cooperate to write the document *d*

Step 2: For each author $x \in a_d$:

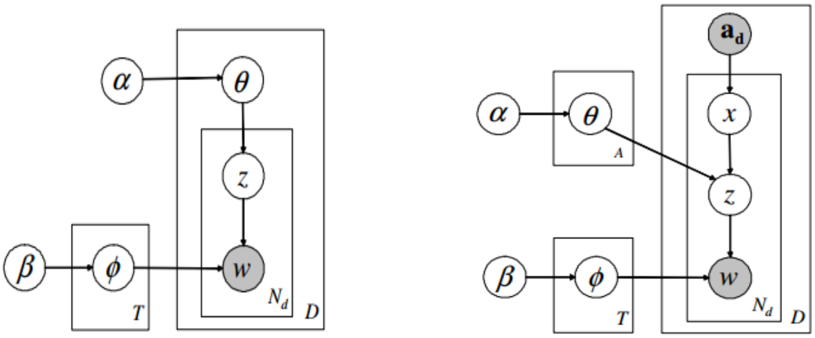
2.1. Associate a distribution over topics θ_i from a Dirichlet distribution with a parameter α.

2.2. Choose a distribution over words φ_j from a Dirichlet distribution with a parameter for each topic.

2.3. For each of the word position *i*, *j*:

2.3.1. Choose a topic z_{ij} from a multinomial distribution with parameter θ_i

2.3.2. Choose a word $w_{i,j}$ from a multinomial distribution with parameter $\varphi_{z_{ij}}$



a. LDA b. ATM

Fig. 1. Topic modeling

Inference: For the ATM, the Gibbs sampling algorithm was proposed to learn the posterior distributions of the latent variables in the model given the observed data [12]. In the author-topic model, we have two sets of latent variables: z and x. We draw each (z_i, x_i) pair as a block, conditioned on all other variables.

$$p(z_i = j, x_i = k | w_i = m, z_{-i}, x_{-i}, w_{-i}, a_d) \propto \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \frac{C^{AT}_{kj} + \alpha}{C^{AT}_{kj'} + T\alpha} \qquad (2)$$

where $z_i = j$ and $x_i = k$ representing the assignments of the ith word in a document to topic j and author k; $w_i = m$ representing the observation that the ith word is the mth word in the lexicon; z_{-i} and x_{-i} represent all topic and author assignments except the ith word, and C^{AT}_{kj} is the number of times author k is assigned to topic j, not including the current instance. $\sum_{m'} C^{WT}_{mj}$ is the number of times a word token w_i is assigned to a topic j across all docs.

Equation (3) presents the distribution of the words in a topic, and equation (4) is the distribution of topics in an author.

$$\emptyset_{mj} = \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \qquad (3)$$

$$A\theta_{kj} = \frac{C^{AT}_{kj} + \alpha}{C^{AT}_{kj'} + T\alpha} \qquad (4)$$

4 Our approach

We utilize supervised learning methods for predicting topic propagation on the bibliographic network. For a topic's spreading, we predict whether an inactive author activates with that topic in a future time T_2 based on available factors of the author in a past time T_1. All nodes of published papers in our particular topic of interest are tagged as active and vice versa. In the training stage, we, first, sample a set of authors X who have not been active in the past period T_1; then, extract the features. After that, the machine learning method is used to build a training model to learn the best coefficients associated with the features by maximizing the likelihood of relationship formation. In the test stage, we apply the trained model to the test set and compare predicted accuracy with the ground truth. The process of topic diffusion prediction using machine learning is described in Figure 2.

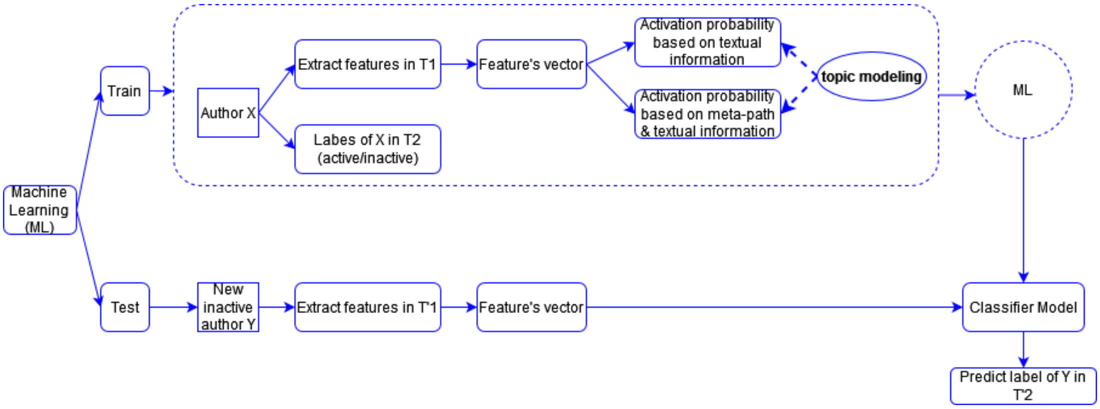

Machine Learning (ML)
Train
Test
Author X
Extract features in T1
Labes of X in T2 (active/inactive)
Feature's vector
Activation probability based on textual information
Activation probability based on meta-path & textual information
topic modeling
ML
New inactive author Y
Extract features in T'1
Feature's vector
Classifier Model
Predict label of Y in T'2

Fig. 2. Process of a topic diffusion prediction

We proposed a new method to estimate the activation probability of an active node and an inactive node based on the meta-path and textual content, namely aggregated activation probability.

$$AAP(u,v) = (1-\sigma) * P(u|v) + \sigma * IS(u,v) \tag{5}$$

$$AAP(u,\{v\}) = max_{M=1..n}(AAP(u,v)) \tag{6}$$

Equation (5) presents the aggregated activation probability of an active node v and an inactive node u. $P(u|v)$ is the activation probability estimated from meta-path information. $IS(u, v)$ is the activation probability of an active node v and an inactive node u based on the textual content.

Equation (6) illustrates the aggregated activation probability of an inactive node *u* switched into an active state by maximizing the aggregated activation probabilities from its active neighbours to it.

$P(u|v)$ is estimated by using the Bayesian framework in equation (7). $n_{v\to u}^{k}$ illustrates the path instances between nodes in the meta-path *k*.

$$P(u|v) = \frac{\sum_{k=1}^{m} \alpha_k \, n_{v\to u}^{k}}{\sum_{k=1}^{m} \alpha_k \, \sum_{r\in nei_v} n_{v\to r}^{k}} \tag{7}$$

$IS(u, v)$ is estimated based on the textual content by using the TFIDF or topic modeling. In our previous study, we used the TFIDF with the cosine distance. In this study, we continue to use topic modeling, specifically LDA and ATM.

Using LDA and ATM, we obtain the topic's distribution of authors. After that, we can use one of the distance measures related to the probability distribution to estimate the interest similarity between two authors, such as Hellinger distance (Eq. 8), KullbackLeibler Divergence (Eq. 9), and Jensen-Shannon divergence (Eq. 10).

$$IS(u,v) = d_H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{K}(\sqrt{p_i} - \sqrt{q_i})^2)} \tag{8}$$

$$IS(u,v) = d_{KL}(P||\,Q) = \sum_{x\in X} P(x)\frac{P(x)}{Q(x)} \tag{9}$$

$$IS(u,v) = d_{JS}(P,Q) = \frac{1}{2}\sum_{i=1}^{K} p_i ln\frac{2p_i}{p_i + q_i} + \frac{1}{2}\sum_{i=1}^{K} q_i ln\frac{2q_i}{p_i + q_i} \tag{10}$$

5 Experiments and results

5.1 Dataset

Experiments were conducted on the dataset "DBLP-SIGWEB.zip", which originated from the September 17, 2015, snapshot of the dblp bibliography database. This dataset contained all publications and author's records of seven ACM SIGWEB conferences. In addition, the dataset also contained the authors, chairs, affiliations, and additional metadata information of the conferences published in the ACM digital library.

5.2 Experiments setting

We considered the spreading of each specific topic *T* and performed experiments with three topics: "*Data Mining*", "*Machine Learning*", and "*Social Network*". Firstly, all active authors with topic *T* were considered as positive training nodes. Then, we sampled an equal-sized of negative nodes corresponding to inactive authors.

For our experiments, we utilized the prediction model as the classification method. In the training dataset, an active author *X* activates with topic *T* in the year y_{XT} , and we extracted the features of *X* in the past time T_1 = [1995, y_{XT} – 1]. Besides, for the inactive author *Y*, we extracted the features in the past time T_1 = [1995, 2014]. We conducted experiments with different features and evaluated the incremental performance improvement. These features are shown in Table 1.

Table 1. Features

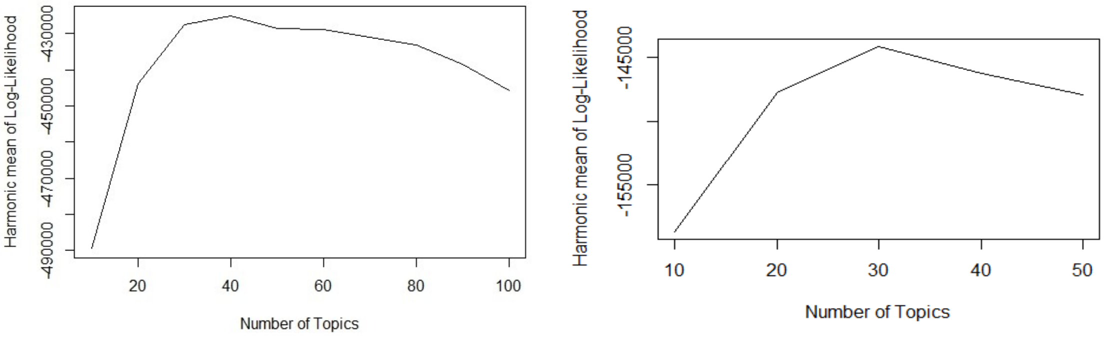
No.	Feature
1	IS(TFIDF)
2	IS(LDA)
3	IS(ATM)
4	AAP(MP+IS(TFIDF))
5	AAP(MP+IS(LDA))
6	AAP(MP+IS(ATM))

Table 2. Number of topics in corpus in each interval

Intervals	#topics
[1995 – Y] with Y = [1995, 1997]	10
[1995 – Y] with Y = [1998, 2001]	20
[1995 – Y] with Y = [2002, 2006]	30
[1995 – Y] with Y = [2007, 2008]	40
[1995 – Y] with Y = [2009, 2015]	50

The objective of this study is to compare the performance of spreading prediction when using topic modeling and TFIDF in activation probability estimation. Experiments (2) and (3) provide the activation probability estimation based on the textual content by using LDA and ATM. While experiments (5) and (6) give the aggregated activation probability based on the meta-path and textual content, in which *IS* was estimated by using LDA and ATM. Experiments (1) and (4) are considered baselines for comparing the prediction performance since they were conducted in the previous study.

For calculating *IS*, we collected the textual information from the keywords of the author's articles in the past interval T_1. To estimate the topic's probability distribution of authors by using topic modeling, we had to estimate the number of topics in the corpus. We defined the number of topics for the whole corpus based on the Harmonic mean of Log-Likelihood (HLK) [13]. For dissimilar intervals, the corpus is dynamic, leading to a change in the number of topics in the corpus. We set each interval starting from 1995 to year *Y* (*Y* in [1995, 2015]) to consider the transformation of the corpus over the years. We had to define the number of topics in the corpus in each interval. Firstly, we estimated the number of the whole corpus (from 1995 to 2015) by running HLK with the number of topics in the range of [10, 100] with a step of 10. We realized that the number of topics fell in the range of [30, 50] (Figure 3a). After that, for the corpus in each interval, we calculated the HLK with the number of topics in the range of [10, 50] with an increment of 10 and estimated the best number of topics (Table 2). Figure 3b illustrates the best number of topics in the corpus in the interval of [1995, 2006]. After defining the number of topics of the corpus, we estimated the topic probability distribution of authors using ATM and LDA and could use the distance measure in equations (4), (5), or (6) to estimate the interest similarity (IS).



a. # topics in whole corpus

b. # topics in interval [1995- 2006]

Fig. 3. Log-likelihood results

Two meta-paths were used to calculate $P(u|v)$, namely APA (Author-Paper-Author) and APAPA (Author-Paper-Author-Paper-Author). Parameter σ equal to 0.5 was set for AAP estimation.

Three classification algorithms, namely Support Vector Machine (SVM, Linear Kernel), Decision Tree (DT), and Random Forest (RF), were chosen for prediction.

5.3 Results

Experimental results show that using topic modeling to estimate IS and AAP can bring higher accuracy than with TFIDF. For the classification results of the topic "*Data Mining*" (Table 3), we can see that topic modeling gives higher accuracy than with TFIDF, in particular, IS(LDA) with the SVM classifier, IS(ATM) with the DT classifier, and IS(ATM) with the RF classifier. Besides, utilizing LDA and ATM for computing AAP also improves prediction's effectiveness in which AAP(MP+IS(ATM) with RF exhibits the highest accuracy.

Table 3. Classification results-topic "*Data Mining*"

Features	Prediction Accuracy					
	SVM		DT		RF	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
IS(TFIDF)	0.572	0.606	0.508	0.509	0.564	0.573
IS(LDA)	**0.614**	**0.607**	0.477	0.477	0.495	0.503
IS(ATM)	0.55	0.606	**0.618**	**0.618**	**0.605**	**0.652**
AAP(MP+IS(TFIDF))	0.582	0.664	0.555	0.555	0.627	0.691
AAP(MP+IS(LDA))	**0.600**	**0.644**	0.55	0.55	0.555	0.590
AAP(MP+IS(ATM))	0.555	0.555	**0.600**	**0.600**	**0.664***	**0.693***

The prediction results of the topics "*Machine Learning*" and "*Social Network*" are demonstrated in Tables 4 and 5. The advantage of LDA is that it can measure IS, in which AAP brings better performance than ATM. For the prediction of these two topics, higher accuracy is obtained by combining feature AAP(MP+IS(LDA) with the RF classifier.

Table 4. Classification results-topic "*Machine Learning*"

Features	Prediction Accuracy					
	SVM		DT		RF	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
IS(TFIDF)	0.668	0.753	0.511	0.511	0.567	0.716
IS(LDA)	**0.686**	**0.730**	**0.603**	**0.603**	**0.675**	**0.675**
IS(ATM)	0.557	0.613	0.551	0.556	0.547	0.546
AAP(MP+IS(TFIDF))	0.690	0.781	0.661	0.661	0.667	0.688
AAP(MP+IS(LDA))	0.665	0.769	**0.667**	**0.667**	**0.677***	**0.722***
AAP(MP+IS(ATM))	0.579	0.651	0.551	0.551	0.581	0.604

Table 5. Classification results-topic "*Social Network*"

Features	Prediction Accuracy					
	SVM		DT		RF	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
IS(TFIDF)	0.62	0.664	0.643	0.643	0.625	0.694
IS(LDA)	**0.638**	**0.620**	0.502	0.502	**0.625**	**0.695**
IS(ATM)	0.509	0.577	0.618	0.618	0.513	0.653
AAP(MP+IS(TFIDF))	0.621	0.686	0.654	0.654	0.688	0.695
AAP(MP+IS(LDA))	0.55	0.638	0.554	0.554	**0.688***	**0.711***
AAP(MP+IS(ATM))	0.55	0.593	0.5	0.5	0.523	0.538

Figures 4, 5 & 6 demonstrate the performance of our Random Forest classifiers on the topics "*Data Mining*", "*Machine Learning*", and "*Social Network*" by using different features.

Consequently, topic modeling can improve the effectiveness of the topic's diffusion prediction compared with TFIDF since we use it to estimate the topic's distribution of nodes.

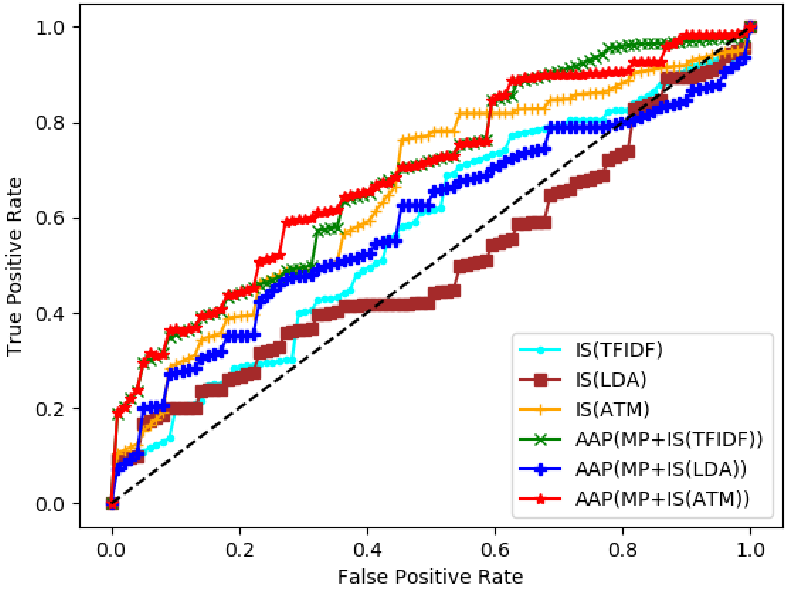


Fig. 4. ROC curve of Random Forest classifier with topic "*Data Mining*"

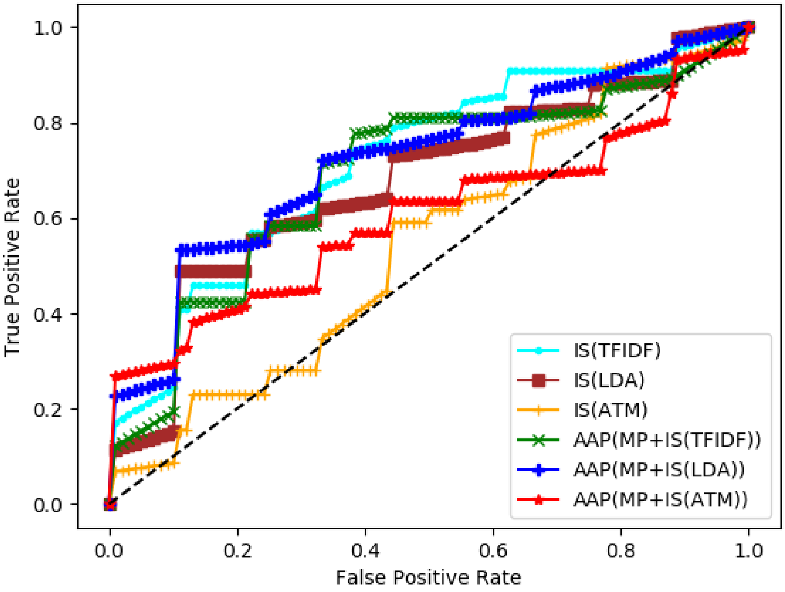


Fig. 5. ROC curve of Random Forest classifier with topic "*Machine Learning*"

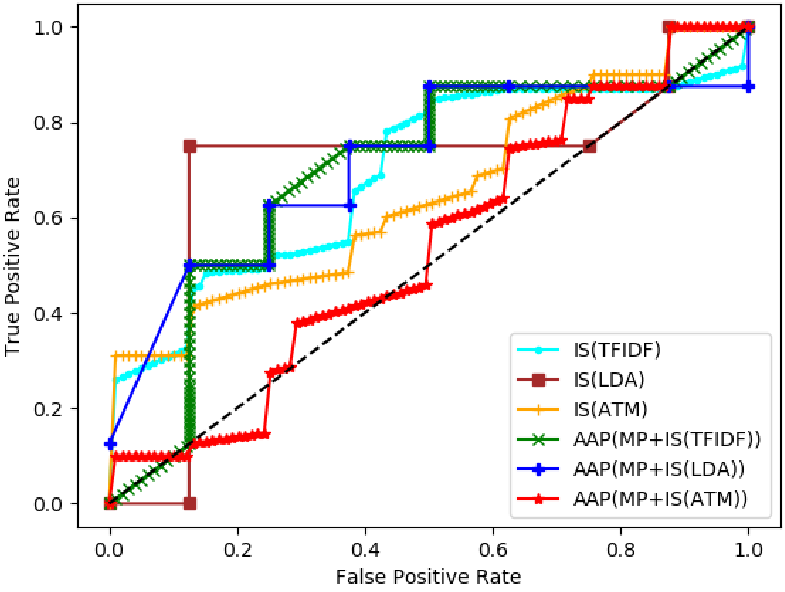


Fig. 6. ROC curve of Random Forest classifier with topic "*Social Network*"

6 Conclusion

In this work, we continued our previous investigation by utilizing topic modeling to estimate activation probability. We used the latent Dirichlet allocation and the author-topic model to estimate the topic's distribution of nodes and distance measures related to a probability distribution to measure interest similarity based on the textual content. Interest similarity was considered an activation probability. Furthermore, we applied interest similarity with topic modeling to calculate aggregated activation probability. Experimental results demonstrate that topic modeling can improve the performance of a topic's spreading prediction compared with the Term Frequency–Inverse Document Frequency technique.

References

1. Kempe, D., Kleinberg, J. & Tardos, É. (2003, August). Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 137-146).
2. Kempe, D., Kleinberg, J. & Tardos, É. (2005, July). Influential nodes in a diffusion model for social networks. In International Colloquium on Automata, Languages, and Programming (pp. 1127-1138). Springer, Berlin, Heidelberg.
3. Kimura, M. & Saito, K. (2006, September). Tractable models for information diffusion in social networks. In European conference on principles of data mining and knowledge discovery (pp. 259-271). Springer, Berlin, Heidelberg.
4. Goldenberg, J., Libai, B. & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing letters, 12(3), 211-223.
5. Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, *83*(6), 1420-1443.
6. Ho, T. K. T., Bui, Q. V. & Bui, M. (2018, September). Homophily independent cascade diffusion model based on textual information. In International Conference on Computational Collective Intelligence (pp. 134-145). Springer, Cham.
7. Molaei, S., Babaei, S., Salehi, M. & Jalili, M. (2018). Information spread and topic diffusion in heterogeneous information networks. *Scientific reports*, *8*(1), 1-14.
8. Gui, H., Sun, Y., Han, J. & Brova, G. (2014, November). Modeling topic diffusion in multi-relational bibliographic information networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 649-658).
9. Bui, Q. V., Ho, T. K. T. & Bui, M. (2020, November). Topic Diffusion Prediction on Bibliographic Network: New Approach with Combination Between External and Intrinsic Factors. In International Conference on Computational Collective Intelligence (pp. 45-57). Springer, Cham.
10. Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.
11. Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

12. Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smyth, P. (2012). The author-topic model for authors and documents. arXiv preprint arXiv:1207.4169.

13. Buntine, W. (2009, November). Estimating likelihoods for topic models. In Asian Conference on Machine Learning (pp. 51-64). Springer, Berlin, Heidelberg.

14. Macy, M.W.: Chains of Cooperation: Threshold Effects in Collective Action. American Sociological Review 56(6), 730–747 (1991), https://www.jstor.org/stable/2096252

15. Yang, H.: Mining social networks using heat diffusion processes for marketing candidates selection. ACM (2008), https://aran.library.nuigalway.ie/handle/10379/4164

16. Varshney, D., Kumar, S., Gupta, V.: Modeling Information Diffusion in Social Networks Using Latent Topic Information. In: Huang, D.S., Bevilacqua, V., Premaratne, P. (eds.) Intelligent Computing Theory. pp. 137–148. Lecture Notes in Computer Science, Springer International Publishing, Cham (2014).

17. Akula, R., Yousefi, N., Garibay, I.: DeepFork: Supervised Prediction of Information Diffusion in GitHub p. 12 (2019)

18. Molaei, S., Zare, H., Veisi, H.: Deep learning approach on information diffusion In heterogeneous networks. Knowledge-Based Systems p. 105153 (Oct 2019), http://www.sciencedirect.com/science/article/pii/S0950705119305076

19. G. Heinrich. Parameter estimation for text analysis. Technical report, 2004.

20. T. L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National academy of Sciences of the United States of America, 101(Suppl1):5235, 2004.