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

Vol. 131, No. 2B, 2022, 33–47; DOI: 10.26459/hueunijtt.v131i2B.6877

Feature's importance assessment for activation probability measure in topic's diffusion prediction

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 study, we aim to estimate the sigma coefficient in the activation probability calculation for a topic's diffusion prediction problem. In our previous studies, we proposed an aggregated activation probability combination of the metapath and text information, in which sigma is the characteristic coefficient of interest's similarity based on textual content. σ is a parameter that controls the rates of the influence of active probability based on the metapath and interest similarity on aggregated activation probability. In a previous study, we supposed the equal importance between the metapath and textual information, when $\sigma = 0.5$. However, for different datasets, this coefficient differs, depending on the meaning of the meta-path and the textual information. In this study, we continue to investigate the importance of the sigma coefficient for the effectiveness of the topic's diffusion prediction problem on the bibliographic network. We propose to utilize the two most common methods for feature selection: the ANOVA test and mutual information to obtain the significance of two features MP (metapath) and the IS (textual information). The experimental results show that the use of the feature selection methods to estimate the sigma coefficient is reliable and improves the predictive performance of the topic's diffusion compared with the standard assignment of 0.5.

Keywords: activation probability, bibliographic network, meta-path, sigma coefficient

1 Introduction

Information diffusion is the process of transferring information from one destination to another through interaction. Information includes rumors, ideas, diseases, etc. The process of propagating information can be described as a node that is considered active if it acts on the information. For example, a scientist is said to be "active" in "deep learning" since he has researched and published articles on this topic. Or, a customer is called "active" with a product "computer" at the time of purchase.

The propagation of information has been used in two types of networks: homogeneous networks [1–5] and heterogeneous networks [6–8]. A homogeneous network is a network that contains a single object type and a link type. The co-author network with an object author and 'co-author link' or an object user and link 'friendship' on a friend network are examples of

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

Received: 27–07–2022; Accepted: 26–09–2022

homogeneous networks. A heterogeneous network is a network with different types of objects and relationships. A bibliographic network is an example of a heterogeneous network which includes objects, such as authors, articles, places, and partnerships, and simultaneously different relationships between authors, such as co-author, participation in conferences, and collaboration in laboratories.

In our previous publication [9], we focused on exploiting the propagation of information in a heterogeneous network. We examined the topic prediction in a bibliographic network using a novel approach that combines external and intrinsic factors. The supervised learning technique was used to predict the diffusion of a given subject by combining dissimilar features with the dissimilar measurement coefficient.

First, we proposed a new method to estimate the activation probability from an active node to an inactive node by combining meta-paths and textual information (IS). The activation probability was estimated from the meta-path (MP) by using the Bayesian framework. Also, activation probability could be measured from the textual information with term frequency – inverse document frequency (TFIDF) and cosine distance or with topic modeling and distance measurements regarding the probability distribution. Subsequently, we proposed an aggregated activation probability (AAP) based on the activation probability from the meta-path and textual information. This probability came into play as an external factor in activating an inactive node switched to an active state. Finally, we suggested an intrinsic factor, which was the author's interest in the subject propagated.

Previous experimental results show that the aggregated activation probability with the combination of the meta-path and textual content improved the accuracy of topic broadcast prediction compared with the old activation probability, which only used the information meta-path or textual information separately. In addition, the amalgamation of activation probability and the author's interest in the topic achieved the highest accuracy.

Furthermore, the experimental results show that the use of topic modeling achieved better accuracy compared with TFIDF when estimating the activation probability based on textual information.

The aggregated activation probability is calculated based on the meta-path and textual information, where the sigma coefficient ($\sigma \in [0, 1]$) indicates the importance of the textual information. If the sigma is larger, it means that textual information has a greater impact on AAP and vice versa. However, in previous studies, we did not consider the importance of the meta-path and textual information in calculating AAP. The sigma coefficient in the AAP equation was assigned to 0.5 by default, which means that the meta-path and text are equally important. In reality, however, the meta-path and textual information with different datasets have a different importance, leading to dissimilar sigma coefficients and the effect on the activation process prediction. Therefore, in this study, we continue to consider the estimation of the sigma coefficient to improve the prediction performance of the prevalence of subjects in the bibliographic network.

We propose to use two common methods in feature selection: ANOVA F-test and mutual information to obtain the importance scores of two features: the meta-path and textual information. After that, we normalize them in [0, 1]. Sigma is the coefficient of the IS characteristic. For different datasets, this coefficient is different depending on the meaning of the meta-path and textual information. Experimental results show that estimating the sigma coefficient improves the accuracy of subject propagation prediction compared with the default setting at 0.5.

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 of disseminating information from one person or community to another in a network, also known as information dissemination, information propagation, and information spreading. Numerous studies have analyzed the diffusion of information, with particular emphasis on which information spreads fastest, which factors influence the dissemination of information, and which models should simulate and predict dissemination. These questions have played an essential role in understanding the phenomenon of diffusion. They have been resolved by research into smaller branches of information dissemination, including models of epidemic spread, impact analyses, and predictive models.

Most information dissemination studies have been conducted on homogeneous networks, where there is only one type of entity and one type of connection. However, in the real world, most networks are heterogeneous because of different kinds of objects and many network relationships. For example, a bibliographic network is a heterogeneous network of multiple entities, including authors, articles, places, and affiliations. There are many relationships between authors, for example, the principal author and co-authors. Our research aims to disseminate information in heterogeneous networks.

For studying prediction patterns in heterogeneous networks, there are two main methods of modeling and predicting the distribution of information. First, the diffusion process is modelled with models such as the linear threshold model (LT) [1, 2], the independent cascade model (IC) [3], the descending cascade model [4], the general threshold model [5], heat diffusion-based models [6], and others. In this way, some active nodes influence the inactive neighbours of the network to become active nodes. An inactive node in an IC can be infected by an active node with a certain probability. In LT, an inactive node is active if the total weight of its active neighbours is greater than or equal to a threshold. There are also several comprehensive models of IC, such as Homophily Independent Cascading Diffusion (TextualHomo – IC) [7] or Heterogeneous Probability Model – IC (HPM–IC) [8], which estimate the probability of infection based on textual information, where the probability of infection is calculated as a conditional probability based on information about the meta-path. There are also several comprehensive models of LT, including the Multiple Relational Linear Threshold Model – MLTM-R) [9] or the Probabilistic Model – LT (HPM–LT) [8].

These models propose a method to measure the probability of inactive nodes infected based on meta-path information or textual information. However, the intrinsic factors of inactive nodes or other features are not considered, for instance, the interest level of the nodes to the topic or each node's influence. Therefore, the second approach emerges by combining dissimilar futures.

The second approach is the use of supervised learning and deep learning to predict the dissemination of information over a heterogeneous network. The spread of a tweet on Twitter was studied with a supervised learning method [10] that combined user interests and the content similarity between an active user and an inactive user by using latent topic information. Furthermore, the information dissemination on Github by using supervised learning was investigated [11]. Furthermore, deep learning was used to predict information dissemination over a heterogeneous network [12]. The diffusion of the topics on the bibliographic network was studied as a first approach under diffusion models. Additionally, this problem was studied with the second approach under the deep learning method [12], but the supervised learning method was not used. Therefore, we focus on predicting the topic spread in the bibliographic network using the supervised learning method.

In our previous studies [13], we proposed a method for estimating the probability of activation based on the meta-path and textual information, namely the aggregated activation probability. We conducted experiments with TFIDF and topic modeling in estimating text information. The experimental results show that our method improved the accuracy of predicting the diffusion of the topic from the bibliographic network. Topic modeling, in particular, worked better than TFIDF. However, we did not evaluate the importance of the feature meta-path and textual information in calculating the AAP. Based on our previous investigations, in this study, we propose to use the feature importance estimation methods to estimate the sigma coefficient for the probability estimation of activation.

3 Preliminaries

3.1 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) [14] 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 Fig. 1a. The LDA generative process for a corpus *D* consisting of *M* documents, with N_i being their length and K denoting the number of topics, is as follows:

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

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

3.2. Choose a word $w_{i\,j}$ from a multinomial distribution with the 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 the 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 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 where the observed data are given. 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 [14] and Gibbs sampling.

3.2 Author-topic model

Author-topic model (ATM) [15] 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 authors who exhibit similar patterns. In 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 the topics are multinomial distributions over words. The ATM generative model is described with a graphical model in Fig. 1b and proceeds as follows:

Step 1. Choose a group of authors a_d, cooperating to write the document *d*

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

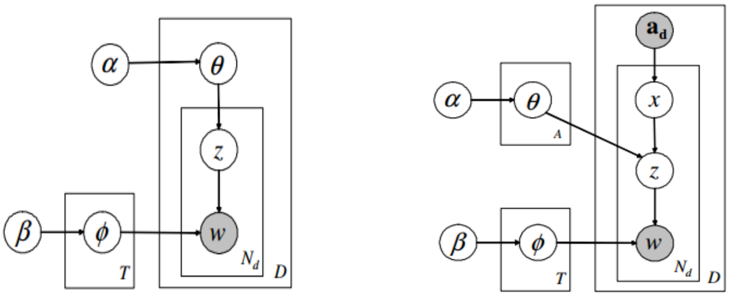
2.1. Associate distribution over topics θ_i from a Dirichlet distribution with parameter α.

2.2. Choose 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 the parameter θ_i

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



a. LDA b. ATM

Fig. 1. Topic modeling

Inference: For ATM, Gibbs sampling algorithm was proposed to learn the posterior distributions of the latent variables in the model where the observed data were given [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

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

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

$$\emptyset_{mj} = \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \tag{3}$$

$$\theta_{kj} = \frac{C_{kj}^{AT} + \alpha}{C_{kj'}^{AT} + T\alpha} \tag{4}$$

3.3 Feature selection methods

In building a machine learning model, all the variables in a dataset are rarely helpful for modeling. Adding redundant variables reduces the generalizability of the sample and can reduce the overall accuracy of the classifier. Also, adding more variables to the model increases the overall complexity of the model. Therefore, feature selection is an essential part of building machine learning models. In machine learning, there are many popular feature selection techniques, such as information gain (mutual information), ANOVA F-test, and Fisher's Score. The two most commonly used feature selection methods for numerical input data and categorical targets are the ANOVA F-test statistic and the information gain.

ANOVA F-test [16]**:** ANOVA means "analysis of variance" and is a test of parametric statistical hypothesis to determine whether two or more data samples come from the same distribution.

An F-statistic (or F-test) is a class of statistical tests that calculate the ratio of variance's values, such as the variance from two different samples or the explained and unexplained variance with a statistical test, like ANOVA. The ANOVA method is a type of F-statistic referred to here as an ANOVA F-test.

In particular, ANOVA is used when one variable is numeric, and the other is categorical, such as numerical input variables and a classification target variable in a classification task. The results of this test can be used for feature selection, where those features are independent of the target variable and can be removed from the dataset.

The scikit-learn machine library provides an implementation of the ANOVA F-test in the F classif() function. This function can be used in a feature selection, such as selecting the top k most relevant features (largest scores) via the SelectKBest class.

Information Gain [17]**:** Information gain calculates the reduction in entropy from the transformation of a dataset.

Entropy function: given a probability distribution of a discrete variable x can receive n different values x_1, x_2, ..., x_n. Suppose that the probability x receives value x_i is $p_i = p(x = x_i)$; $0 \leq p_i \leq 1$; $\sum p_i = 1$. The entropy of this distribution is defined as

$$H(p) = -\sum_{i=1}^{N} p_i \log(p_i) \tag{5}$$

The entropy function is used to measure the purity of the dataset. This function returns a minimum value if the data in the child node belongs to a class and gives a maximum value if the child node contains data with different classes. The information gain provides a way to use entropy to determine how a change to the dataset affects the purity of the dataset. Information Gain based on attribute x

$$IG(S, x) = H(S) - H(S|x) \tag{6}$$

where $IG(S, a)$ is the information for the dataset S for the variable x for a random variable; $H(S)$ is the entropy for the dataset before any change, and $H(S|x)$ is the conditional entropy for the dataset given the variable x.

In the context of feature selection, the information gain may be referred to as "mutual information" and calculate the statistical dependence between two variables. An example of using information gain (mutual information) for feature selection is the mutual_info_classif() scikit-learn function in Python.

The mutual information [18] is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable. The mutual information between two random variables X and Y can be stated formally as follows:

$$I(X; Y) = H(X) - H(X|Y) \tag{7}$$

where $I(X;Y)$ is the mutual information for X and Y; $H(X)$ is the entropy for X, and $H(X|Y)$ is the conditional entropy for X given Y. The result has the units of bits.

4 Our approach

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

$$AAP(u,v) = (1-\sigma) * P(u|v) + \sigma * IS(u,v) \tag{8}$$

$$AAP(u,\{v\}) = max_{M=1..n}(AAP(u,v)) \tag{9}$$

Equation (8) presents the aggregated activation probability from active node v to inactive node u. $P(u\,|\,v)$ is the activation probability estimated from the meta-path information. $IS(u, v)$ is the activation probability based on the textual content.

Equation (9) illustrates the aggregated activation probability of an inactive node u switched to 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 (10). $n^k_{v\to u}$ illustrates the path instances between nodes in meta-path k.

$$P(u|v) = \frac{\sum_{k=1}^{m} \alpha_k\, n^k_{v\to u}}{\sum_{k=1}^{m} \alpha_k\, \sum_{r\in nei_v} n^k_{v\to r}} \tag{10}$$

IS(u, v) is estimated on the basis of the textual content by using TF-IDF or topic modeling

$$IS(u,v) = Cos(T_u, T_v) = \frac{T_u . T_v}{||T_u||.\,||\,T_v||} \tag{11}$$

$$IS(u,v) = d_H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{K}(\sqrt{p_i}-\sqrt{q_i})^2)} \tag{12}$$

$$IS(u,v) = d_{KL}(P||\,Q) = \sum_{x\in X} P(x)\frac{P(x)}{Q(x)} \tag{13}$$

$$IS(u,v) = d_{JS}(P,Q) = \frac{1}{2}\sum_{i=1}^{K} p_i ln\frac{2p_i}{p_i+q_i} + \frac{1}{2}\sum_{i=1}^{K} q_i ln\frac{2q_i}{p_i+q_i} \tag{14}$$

where σ is a parameter that controls the rates of the influence of active probability on the basis of the meta-path and interest similarity on the aggregated activation probability. σ ∈ [0, 1], if the larger σ means that we focus on the text information and vice versa.

In this study, we propose using ANOVA F-test Feature Selection and Mutual Information Feature Selection to get the importance scores of two features of MP and IS. After that, we normalize them into [0, 1] and apply them to equation (8).

5 Experiments and results

5.1 Dataset

We used dataset ”DBLP-SIGWEB.zip”, which is originated from September 17, 2015, snapshot of the dblp bibliography database. This dataset contains all publications and authors records of seven ACM SIGWEB conferences. Furthermore, the dataset also contains the authors, chairs, affliations and additional metadata information of the conferences that are published in the ACM digital library.

5.2 Experiment setting

We will consider the spreading of each specific topic T. We conduct experiments with three topics: “*Data Mining*”, “*Machine Learning*”, and “*Social Network*”. Firstly, all active authors with topic T will be considered positive training nodes. We also sample equal-sized negative nodes corresponding to inactive authors.

In our experiments, we utilize classification methods as the prediction model. In training a dataset, the active author X activates topic T in the year y_{XT}, and we the extract features of X in the past period T_1 = [1995, y_{XT} – 1]. Besides, with inactive author Y, we extract features in the past period T_1 = [1995, 2014].

The purpose of this study is to compare the performance of spreading prediction with different values of sigma in the activation probability estimation. Therefore, firstly we estimate the best sigma coefficient.

In our previous study, we saw that ATM provided the best diffusion activation probability with topic “Data Mining”, and LDA for topics “Machine Learning” and “Social Network”. Therefore, in this study, we continue to implement experiments using ATM for topic “Data Mining” and LDA for topics “Machine Learning” and “Social Network”.

ANOVA F-test Feature Selection and Mutual Information Feature Selection are utilized to get the importance scores of MP and IS. After that, we normalize them into [0, 1]. The results of sigma estimation are shown in Table 1.

Table 1. Sigma coefficient estimation

Topic	ANOVA F-test Feature Selection	Mutual Information Feature Selection
Data Mining	0.5	0.3
Machine Learning	0.55	0.15
Social Network	0.5	0.6

The meaning of the meta-path and textual information is different in different datasets, and the lead-to-sigma coefficient changed.

With the sigma coefficient in Table 1, we perform experiments and evaluate the incremental performance improvement. These features and their corresponding sigma values are shown in Tables 2, 3 and 4.

Table 2 presents the features that are used for calculating AAP in the dataset with topic "Data Mining". When σ is 0, the AAP calculation does not consider textual information. Therefore, the feature considered for AAP is MP. When σ is 1, the AAP calculation considers only the textual information, ignoring the meta-path information. With the ANOVA F-test estimation method, σ is 0.5; that means that the AAP calculation considers the meta-path and textual information equally important. In the mutual information estimation method, σ is 0.3; that means we consider both the meta-path and the information, but they do not have the same importance: the meta-path is more important than the text information.

Similarly, we analyze the features corresponding to the sigma coefficient based on the results in Table 2 to dataset "Social Network" and "Machine Learning" (Tables 3 and 4).

Table 2. Feature with sigma on topic "Data Mining"

No.	Feature	Sigma
1	MP	0
2	IS(ATM)	1
3	AAP(MP + IS(ATM))	0.5
4	AAP(MP + IS(ATM))	0.3

Table 3. Feature with sigma on topic "Social Network"

No.	Feature	Sigma
1	MP	0
2	IS(LDA)	1
3	AAP(MP + IS(LDA))	0.5
4	AAP(MP + IS(LDA))	0.6

Table 4. Feature with sigma on topic "Machine Learning"

No.	Feature	Sigma
1	MP	0
2	IS(LDA)	1
3	AAP(MP + IS(LDA))	0.5
4	AAP(MP + IS(LDA))	0.55
5	AAP(MP + IS(LDA))	0.15

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

In addition, we implement experiments with different sigma values from 0 to 1 with a step of 0.05 to compare the results with the estimated sigmas above.

5.3 Results

The experimental results show that estimating sigma (σ) can improve the performance of prediction diffusion.

For the topic "Data Mining", we estimated the best sigma to be 0.5 or 0.3. The experimental results of classification (Table 5) show that the highest accuracy is obtained when σ = 0.5 with the RF classifier.

For the topic "Machine Learning", the best sigma was chosen to be 0.55 or 0.15. The classification results (Table 6) display that the highest accuracy reaches when σ = 0.15.

For the topic "Social Network", the best sigma was chosen at 0.5 or 0.6, and when σ = 0.6, the diffusion prediction picks (Table 7).

Table 5. Classification results-topic "Data mining"

Features	σ	Prediction Accuracy					
		SVM		DT		RF	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
MP	0	0.609	0.704	0.623	0.639	0.609	0.694
IS(ATM)	1	0.55	0.606	0.618	0.618	0.605	0.652
AAP(MP+IS(ATM))	**0.5**	0.555	0.555	0.6	0.6	**0.664***	**0.693***
AAP(MP+IS(ATM))	0.3	0.491	0.574	0.523	0.523	0.550	0.511

Table 6. Classification results-topic "Machine Learning"

Features	σ	Prediction Accuracy					
		SVM		DT		RF	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
MP	0	0.689	0.771	0.592	0.586	0.644	0.666
IS(LDA)	1	0.668	0.753	0.511	0.511	0.567	0.716
AAP(MP+IS(LDA))	0.5	0.665	0.769	0.667	0.667	0.677	0.722
AAP(MP+IS(LDA))	0.55	0.665	0.754	0.500	0.500	0.642	0.699
AAP(MP+IS(LDA))	**0.15**	**0.721**	**0.812**	0.557	0.557	**0.746***	**0.707***

Table 7. Classification results-topic "Social Network"

Features	σ	Prediction Accuracy					
		SVM		DT		RF	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
MP	0	0.613	0.691	0.614	0.586	0.589	0.623
IS(LDA)	1	0.62	0.664	0.643	0.643	0.625	0.694
AAP(MP+IS(LDA))	0.5	0.621	0.686	0.654	0.654	0.688	0.695
AAP(MP+IS(LDA))	**0.6**	**0.600**	**0.653**	0.563	0.563	**0.688***	**0.742***

We can see that the estimated sigma values improve the performance of propagation prediction compared with the default sigma of 0.5.

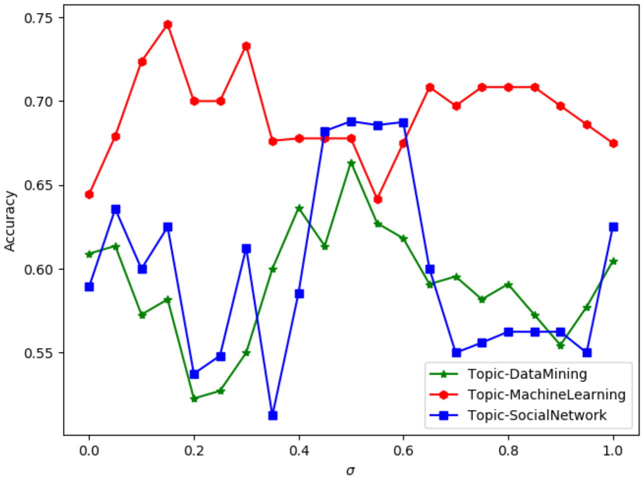


Fig. 2. Accuracy of classification with different sigma values

Fig. 2 demonstrates the accuracy of diffusion prediction for the sigma values from 0 to 1 with an increment of 0.05. We can see that the highest accuracy is obtained at the sigma value of 0.5, 0.15 and 0.5–0.6 for the topics “Data Mining”, “Machine Learning”, and “Social Network”, respectively. These results prove that using the feature’s selection methods to infer to sigma value is reliable and can improve the performance of propagation prediction.

6 Conclusion

In this paper, we continue our previous study by estimating the best sigma coefficient for calculating aggregated activation probability. We use the Latent Dirichlet Allocation model and the Author-Topic model to estimate the topic’s distribution of nodes and the distance’s measure related to probability distribution to measure textual information. The feature’s selection methods are reliable and can improve the performance of topic's spreading prediction compared with the standard assignment at 0.5.

References

1. Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, *83*(6), 1420-1443.
2. Macy, M.W.: Chains of Cooperation: Threshold Effects in Collective Action. American Sociological Review 56(6), 730–747 (1991), https://www.jstor.org/stable/
3. 2096252
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. 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.
6. 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).
7. Yang, H.: Mining social networks using heat diffusion processes for marketing candidates selection. ACM (2008), https://aran.library.nuigalway.ie/handle/10379/4164
8. 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.
9. Molaei, S., Babaei, S., Salehi, M. & Jalili, M. (2018). Information spread and topic diffusion in heterogeneous information networks. *Scientific reports*, *8*(1), 1-14.
10. 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).
11. 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)
12. Akula, R., Yousefi, N., Garibay, I.: DeepFork: Supervised Prediction of Information Diffusion in GitHub p. 12 (2019)
13. 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
14. 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.

15. Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

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

17. Kuhn, M. & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. CRC Press.

18. Mitchell, T. M. "Machine Learning McGraw-Hill International." (1997): 58.

19. Witten, Ian H., and Eibe Frank. "Data mining: practical machine learning tools and techniques with Java implementations." Acm Sigmod Record 31.1 (2002): 76-77.