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

Vol. 132, No. 2B, 2023, 17–31; DOI: 10.26459/hueunijtt.v132i2B.7348

Effect of similarity measures in diffusion prediction on homogeneous and heterogeneous bibliographic network

Thi Kim Thoa Ho[1]*, Quang Vu Bui[2], Minh Duc Phan[1], Thi Uyen Trinh Le[3]

[1] Hue University of Education, Hue University, Vietnam
[2] Hue University of Sciences, Hue University, Vietnam
[3] Ham Nghi Secondary School, Tay Loc Ward, Hue City, Vietnam

Abstract. This paper evaluates the effect of similarity measures in predicting diffusion on homogeneous and heterogeneous bibliographic networks. The bibliographic network is analyzed within a homogeneous network and heterogeneous network, where a co-author relationship exists for the former, and multiple types of meta paths are considered for the latter. The supervised learning method is used to predict whether a node will be active with a topic or not. The features are extracted as the activation probability of a node, which represents the maximum of the activation probabilities of the neighbors of this node. In a homogeneous network, the activation probability from the activated node to the inactive node is measured based on one relationship co-author with basic similarity measures while it can be calculated based on diverse meta paths with dissimilar meta path-based similarity measures in the heterogeneous network. We performed our analysis on three different datasets. Our experimental results show that diffusion prediction in bibliographic networks provides better accuracy among heterogeneous networks than among homogeneous networks and that the Bayesian similarity measure provides the best efficiency.

Keywords: Social network, bibliographic network, information diffusion, meta path, meta path-based similarity measures, machine learning.

1 Introduction

Information diffusion is the process of transmitting information from one destination to another through interaction. Information includes rumors, ideas, diseases, etc. The information diffusion process It can be explained as follows a node considered active when it acts on information. For example, a scientist is said to be "active" in "deep learning" because he has researched and published papers on the topic. A customer is defined as “active” with a “computer” product at the time of purchase. An active node can activate inactive another through interaction, for example, scientist X feels excited and starts to research deep learning when he discusses with his colleague Y about published articles. The probability that Y activates X is called activate probability or infected probability. This probability can be measured based on dissimilar similarity measures.

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

Received: 24–10–2023; Accepted: 29–11–2023

The prediction of information propagation in the bibliographic network helps us to recognize the research tendencies of scientists and leads to applications such as collaboration recommendations, discovery of the research community, etc. Most studies about information diffusion have been conducted the homogeneous bibliographic network [1-6] by its simplicity in which objects are authors and are connected by co-author links. Nevertheless, bibliographic networks is actually heterogeneous network in which there are different useful meta path types that haven't been exploited in homogeneous network such as Author – Paper – Author (APA), or meta – path Author – Paper – Author – Paper – Author (APAPA) and so on. Therefore, the research tendencies transfer from the study on homogeneous network to heterogeneous network [7, 8, 9, 12, 13, 19, 20].

Although previous studies predicted the propagation of topics on a bibliographic network on both the homogeneous and heterogeneous network. Nevertheless, they were conducted independently and didn't compare the effectiveness between homogeneous network and heterogeneous network. Few types of meta paths are overused and there is no comparison between meta path-based similarity measures in heterogeneous networks. Therefore, in this study, we focus on exploiting typical meta path types in heterogeneous bibliographic networks and evaluating the effect of meta path-based similarity measures in predicting topic distribution as well as on the comparison of prediction performances in homogeneous networks.

The objective of our study is to predict whether an inactive author will be active on a specific topic or not using supervised learning method. The features as inputs to the supervised learning method are activation probability of node *u*. This probability is estimated by maximizing the activation probabilities from the active neighborhoods v to the inactive node *u*.

On the one hand, we consider a bibliographic network that is part of a homogeneous bibliographic network, where the node of the network is the authors and the relationship is "co-author". We use five well-known and frequently used distances to measure activate probability from active author to inactive author including Common Neighbor (CN), Jaccard Coefficient (JC), Adamic/Adar Index (AA), Preferential Attachment (PA), Shortest Path Length (SPL). These play a role as baseline measures to compare with meta path-based similarity measures in heterogeneous network.

On the other hand, we consider a heterogeneous bibliographic network where objects including authors, articles, venues, workspaces, ... and relations including co-authors, common co-authors, publish articles in the same venue, and so on. We analyze four typical meta paths including Author – Paper – Author (APA), Author – Paper – Author – Paper – Author (APAPA), Author – Paper – Venue – Paper – Author (APVPA) and Author – Affiliation – Author (AAFA). For each meta path type, we extract one feature namely "activate probability" for each inactive node by maximizing activate probabilities of active neighborhoods. We apply meta path-based similarity measures PathCount, PathSim, and Bayesian to estimate activate probability from active node to inactive node.

Experimental results show that the combination of bibliographic network analysis in a heterogeneous network and the application of meta path-based similarity distances to extract features for information diffusion prediction improves performance compared to similarity measurements in a homogeneous network. Moreover, the Bayesian distance measure provides the best efficiency among meta path-based similarity measures in diffusion prediction.

The structure of our article is as follows: section 1 describes the problem; section 2 summarizes related works; section 3 deals with preliminaries; our approach is proposed in section 4; the experiments, results and discussion are demonstrated at section 5; finally we conclude our study in section 6.

2 Related works

Information diffusion is the process of spreading information from one person or community to another in a network, called information dissemination, information propagation, and information spreading. Numerous studies have analyzed the information diffusion, with particular attention to which information spreads fastest, which factors influence information diffusion, and which models should simulate and predict diffusion. These questions have played an essential role in understanding the phenomenon of diffusion. These problems have been resolved through research into smaller branches of information dissemination, including epidemic spread models, impact analysis, and predictive models.

In our study, the diffusion of information in the bibliographic network is the diffusion of research topics among researchers. There are numerous studies on the diffusion of information in bibliographic networks. However, most studies on information diffusion have been conducted in homogeneous networks where there is only one type of entity and one type of connection.

To study prediction patterns on networks, there are two main methods of modeling and forecasting the information distribution. First, the diffusion process was modeled using diffsuion models such as the linear threshold model (LT) [1, 2], the independent cascade model (IC) [3], the descending cascade model [4] and the general threshold model [5] and heat diffusion-based models [6] and others. In this way, some active nodes influence the inactive neighbors of the network to become active nodes. 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 estimates 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].

The second approach is to use supervised learning and deep learning to predict the distribution of information across a bibliographic network. In this approach, diffusion prediction aims to predict whether an author will be active on a specific topic or not, based on observed existing connection information. Previous studies have used supervised learning and deep learning methods to predict topic distribution over a heterogeneous bibliographic network [12, 13, 19, 20]. However, we exploited only two types APA and APAPA and used Bayesian measures to estimate the similarity. Previous studies have used few types of meta paths in heterogeneous networks and have neither compared the effectiveness of meta path-based similarity measures nor compared to baseline similarity measures in homogeneous networks. Therefore, in this study, we focus on exploiting typical meta path types of heterogeneous bibliographic networks and testing the effect of meta path-based measurement distances for predict diffusion.

3 Preliminaries

3.1 Baseline similarity measures in homogeneous bibliographic network

In homogeneous network, there are several well-known and frequently used distance measures to estimate similarity between network's nodes such as Common neighbor (CN), Jaccard coefficient (JC), Adamic/Adar index (AA), Preferential Attachment (PA), Shortest Path length (SPL).

Common Neighbors (CN): The CN is one of the most prevalent metric used in similarity measurement by cause of its simplicity [21]. For two nodes, x and y, the CN is defined as the number of common neighbors of x and y. The greater number of common neighbors between x and y, the more similar they are. The definition of CN is as follows.

$$CN(x,y) = \Gamma(x) \cap \Gamma(y) \tag{1}$$

Jaccard Coefficient (JC): Jaccard coefficient normalizes the size of common neighbors. This similarity depends on two factors including the number of common neighbors and the total number of neighbors they have. Higher similarity are assumed for pairs of nodes that share a greater percentage of common neighbors to the total number of neighbors of both. This measure is defined as:

$$JC(x,y) = \frac{\Gamma(x) \cap \Gamma(y)}{\Gamma(x) \cup \Gamma(y)} \tag{2}$$

Adamic-Adar Coefficient (AA): The AA metric was initially proposed by Adamic and Adar to calculate the similarity between two web pages [22], after which it has been extensively utilized in social networks. The Jaccard coefficient is a factor in the formulation of the AA measure. However, common neighbors who have fewer neighbors carry more weight. It is defined as:

$$AA(x,y) = \sum_{Z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log\,|\Gamma(z)|} \tag{3}$$

Preferential Attachment (PA): The PA metric indicates that new node will be more likely to connect higher-degree nodes than lower ones [23]. It is defined as:

$$PA(x,y) = \Gamma(x).\Gamma(y) \tag{4}$$

Shortest Path Length (SPL): Paths between two nodes can also be used to calculate node pair similarities in addition to node and neighbor information. Finding a path between two vertices in a graph with the least amount of distance between them is known as the "shortest path" problem. The most significant algorithms for solving this problem include Bellman-Ford, A*, and Dijkstra's algorithms.

3.2 Meta path in heterogeneous bibliographic network

A meta path P is a generally defined path for a network TG = (A, R) in which A and R represent nodes and their relations, respectively [8, 9, 24]. The meta path is denoted by $\boldsymbol{A_1} \xrightarrow{\boldsymbol{R1}} \boldsymbol{A_2} \xrightarrow{\boldsymbol{R2}} \boldsymbol{A_3} \xrightarrow{\boldsymbol{R3}} \ldots \xrightarrow{\boldsymbol{Rl}} \boldsymbol{A_{l+1}}$, where l is an index indicating the corresponding meta-path. The accumulated relationship between A_1 and A_{l+1} is captured as R = R1oR2o...Rl, where o is the composition operator.

The length of P is the number of relations in P. Furthermore, we say a meta path is symmetric if the relation R defined by it is symmetric. A path $p = (a_1a_2...a_{l+1})$ between a_1 and a_{l+1} in network G follows the meta path P, if $\forall$i, $\phi(ai) = A_i$ and each link $e_i = \langle a_ia_{i+1}\rangle$ belongs to each relation R_i in P. We call these paths as path instances of P, which are denoted as p $\in$ P.

In heterogeneous bibliographic network, there are many meta path types including author – paper –author (APA), author – paper – author – paper – author (APAPA), author – paper – venue – paper – author (APVPA), APPA (author – paper – paper – author), AAFA (author – affiliation – author) and so on. Semantic meaning of meta paths is describes in Table 1.

Table 1: Semantic meaning of meta path

Meta path	Schema of meta path	Semantic meaning
APA	$A \xrightarrow{\textit{write}} P \xrightarrow{\textit{is written by}} A$	a_i and a_j are co-authors
APAPA	$A \xrightarrow{\textit{write}} P \xrightarrow{\textit{is written by}} A \xrightarrow{\textit{write}} P \xrightarrow{\textit{is written by}} A$	a_i and a_j are co-authors of common authors
APVPA	A $\xrightarrow{\textit{write}} P \xrightarrow{\textit{is publish at}} V \xrightarrow{\textit{publish}} P \xrightarrow{\textit{is written by}} A$	a_i and a_j publish articles in same venues
AAFA	$A \xrightarrow{\textit{work at}} AF \xrightarrow{\textit{is workplace of}} A$	*ai* and *aj* have same affiliation

3.3 Meta path –based similarity measures

In heterogeneous network, there are several meta path-based similarity measures such as pathcount [25], pathsim [9, 24] and bayesian [8, 13, 19, 20] used to estimate similarity two nodes.

PathCount (PC) : PathCount measures the number of path instances between two objects that follow a particular meta path, called PCR, where R is the relation specified by the meta path. A pathcount can be determined using the product of the adjacency matrices affiliated with each relation in the meta path.

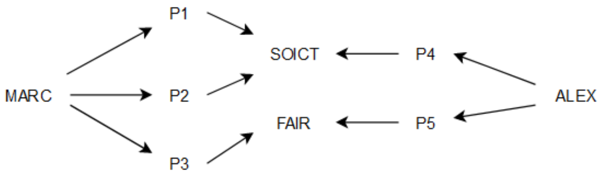

Fig. 1. A example meta path APVPA between two authors

Figure 1 demonstrate a example about meta path APVPA between Marc and Alex and three path instances respectively.

PathSim (PS) : A meta path-based similarity measure [9, 24]. Given a symmetric meta path E_k corresponding relation type k, PathSim between two objects can be defined at equation 5:

$$PS^{E_k}(u,v) = \frac{2|P_{(u,v)}^{E_k}|}{\left|P_{(u,u)}^{E_k}\right| + |P_{(v,v)}^{E_k}|} \tag{5}$$

where $P^{E_k}(u,v)$ is the set of path instances according to relation type k, originating from node $u$and ending at node v, $P^{E_k}(u,u)$ *is that between u and u, and* $P^{E_k}(v,v)$ *is that between* v *and* v.

Bayesian (BE) : A meta path-based similarity measure [8, 12, 19, 20]. Given a symmetric meta path E*k* according to relation type k, Bayesian similarity between two objects can be defined at equation :

$$P^{E_k}(u|v) = \frac{P^{E_k}_{(u,v)}}{P^{E_k}_{(v)}} = \frac{P^{E_k}_{v\to u}}{\sum_{r\in v} P^{E_k}_{v\to r}} \qquad (6)$$

where $P^{E_k}_{v\to u}$illustrate number of path instances from u to v in meta-path k.

4 Our approach

To forecast the propagation of topics in the bibliographic network, we use supervised learning techniques. Using the observed engagement data from the previous time T1, we predict whether the inactive author will engage with the topic in the future T2. An author is marked as active if he has published articles on a specific topic and vice versa.

In the training phase, we first investigate the group of authors X who were not active in the previous period T1, and then we extract their features. A training model is then built using machine learning techniques to maximize the probability of forming the activation and learn the best coefficients affiliated with the features. In the testing phase, we apply the trained model to the test set and evaluate the predicted accuracy against the actual results. Figure 2 shows how predicting topic prevalence using machine learning works.

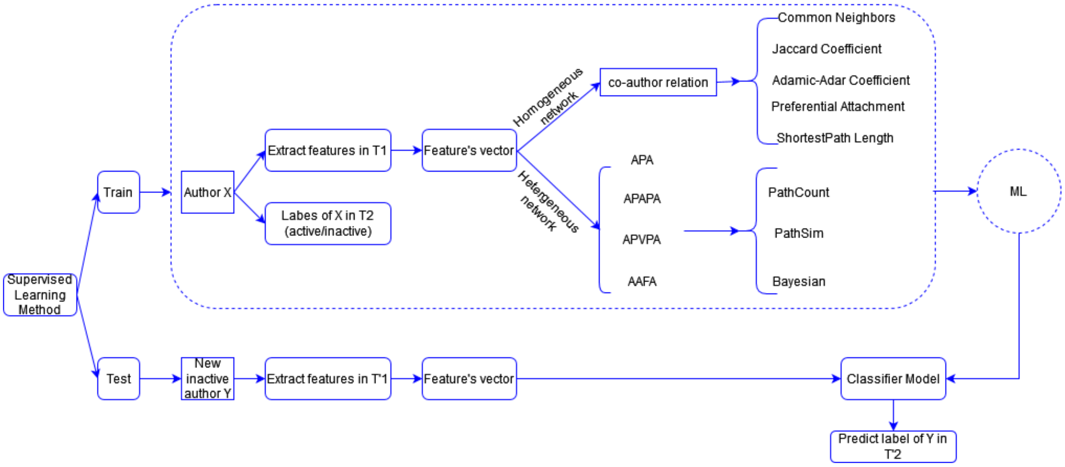

Fig. 2. Process of topic diffusion prediction on bibliographic network

We analyze bibliographic networks in terms of homogeneous network and heterogeneous network patterns. In order to estimate how similar two nodes are in a homogeneous network, we take into account networks with co-author relationships and employ five well-known and frequently used distance measures, including Common Neighbor (CN), Jaccard coefficient (JC), Adamic/Adar index (AA), Preferential Attachment (PA), and Shortest Path Length (SPL). The activation probability of an inactive node, node u, is an extracted feature for machine learning.

In equation 7, activation probability (P(u)) maximizes all similarity measures between node u and all nodes v (P(u,v)), where v is u's active neighbor. The distance measurements mentioned above can be used to calculate the P(u, v).

$$P(u) = max(P(u, v)) \qquad (7)$$

For a heterogeneous network, we analyze the network using four meta paths, and for each meta path, we use three meta path-based similarity measures, including PathCount (PC), PathSim (PS), and Bayesian (BE), to estimate similarity between two nodes.

The activation probability of an inactive node u follow meta path k ($P^{E_k}(u)$) is maximized for all similarity measures corresponding meta path k from node u to all nodes v ($P^{E_k}(u, v)$) where v is an active neighbor of u follow meta path k (see equation 8). The $P^{E_k}(u, v)$ can be estimated by PC, PS or BE.

$$P^{E_k}(u) = max(P^{E_k}(u, v)) \qquad (8)$$

Finally, there are three distinct feature sets. Using the distance measure PathCount, the first set of features is (P_{PC}^{APA}, P_{PC}^{APAPA}, P_{PC}^{APVPA}, P_{PC}^{AAFA}) correspond to activation probabilities of node u following the meta paths APA, APAPA, APVPA, AAFA. Given their similarities, we have feature's set (P_{PS}^{APA}, P_{PS}^{APAPA}, P_{PS}^{APVPA}, P_{PS}^{AAFA}) and (P_{BE}^{APA}, P_{BE}^{APAPA}, P_{BE}^{APVPA}, P_{BE}^{AAFA}) correspond to PathSim and Bayesian.

5 Experiments and results

5.1 Dataset

The dataset "DBLP-SIGWEB.zip" was used, which is a copy of the dblp bibliography database's snapshot from September 17, 2015. It includes metadata of publications from seven conferences: Hypertext and social media; Digital Libraries; Document Engineering; Web Science; Information and Knowledge and Management; Web Science and Data Mining; User Modeling, Adaptation and Personalization. This dataset consists of information about the authors, papers, affiliations, chair, conferences, keyworks, etc., suitable for heterogeneous network analysis.

5.2 Experiments Setting

We will take into account the spreading of each specific topic T. Three subjects are the focus of our experiments: "Data Mining", "Machine Learning" and "Social Network" because of their high frequency in dataset. First, all authors who are currently writing on topic T will be regarded as positive training nodes. Additionally, we select negative nodes of equal size that correspond to inactive authors.

We use classification techniques as the prediction model in our experiments. In the training data, active author X is activated by the topic T in year y_{XT}, we extract features of X in past time period T_1 = [1995, y_{XT} - 1]. In addition, inactive author Y, we extract features in past time period T_1 = [1995, 2014]. The features for machine learning are extracted and demonstrated in Table 2 and 3.

Table 2. Features for prediction on homogeneous network

No.	Feature
1	P_{CN}
2	P_{JS}
3	P_{AA}
4	P_{PA}
5	P_{SPL}

Table 3. Features for prediction on heterogeneous network

No.	Features
1	$P_{PC}^{APA}, P_{PC}^{APAPA}, P_{PC}^{APVPA}, P_{PC}^{AAFA}$
2	$P_{PS}^{APA}, P_{PS}^{APAPA}, P_{PS}^{APVPA}, P_{PS}^{AAFA}$
3	$P_{BE}^{APA}, P_{BE}^{APAPA}, P_{BE}^{APVPA}, P_{BE}^{AAFA}$

5.3 Results and Discussion

Tables 4, 5, 6, and Figures 3, 4, and 5 display the results of the experiments. The results showed that using meta path-based similarity measures and analyzing multiple meta paths in heterogeneous networks as features improved accuracy for diffusion prediction in bibliographic networks compared to baseline similarity measures in homogeneous networks. Additionally, experimental findings show that Bayesian distance measure offers the best efficacy on heterogeneous networks.

For the topic “Data Mining” (see Table 4 and Figure 3), we can see that using four features with PathSim and Bayesian distance show excellent effectiveness compared to PathCount as well as one feature with baseline similarity measure in homogeneous network.

For the topic “Machine Learning” (see Table 5 and Figure 4), combining four features with Bayesian reached the peak of accuracy in RF classification. Next, PathSim and PathCount demonstrate higher accuracy compared with one feature with baseline similarity measures in the homogeneous network.

For the topic “Social Network” (see in Tables 6 and Figure 5), we can see the advantage of combining four features with Bayesian and PathCount in RF classification and PathSim at SVM classification compared to baseline similarity measures in homogeneous network. In particular, Bayesian distance measure demonstrated the best contribution to prediction performance.

In short, combining meta path types in a heterogeneous network and meta path-based similarity measures to extract features improves the effectiveness of subject diffusion prediction. The explanation for these results is that there is a different semantic meaning behind the meta path, which is not taken into account in the homogeneous network. These meta-paths contain useful information for prediction.

Furthermore, the Bayesian measure is the most effective distance measure for predicting the distribution of topics in a heterogeneous bibliographic network. The Bayesian is based on conditional probability and non-symmetric while PathCount or PathSim are symmetric measures. We can see that PC(u,v) is equal to PC(v,u) or PS(u,v) is equal to PS(v, u). However, BE(u|v) is dissimilar from BE(v|u). Bayesian measure BE(u|v) is determined by two parts: (1) their connectivity defined by the number of path instances between them that follow the meta path k; (2) the influence of node v. Bayesian distance measure is suitable for measuring activate probability from node u to v since the implication behind it is that two similar objects are not only strongly connected, but also consider the aspect that a node with high influence has more power to spread information, but it is difficult to be triggered by another, and vice versa. This is why Bayesian brings the best prediction accuracy compared to PathCount and PathSim.

Information diffusion prediction in heterogeneous networks leads to improved performance. However, the calculation is more complicated than in the homogeneous network. In homogeneous network, the estimation of activation probabilities from u to v has complexity O(n) since we consider only co-author relation together with n basic distance measures which have complexity O(1). For PathCount and PathSim, the computational complexity is O(1) while it is O(n) for Bayesian distance measure. Therefore, in the heterogeneous network, the complexity of activation probabilities calculation from u to v is O(n^2) if we consider k meta path types and n distance measures with complexity O(1) like PathCount or PathSim. This complexity becomes O(n^3) if we consider k meta path types and n distance measures with the complexity O(n) as Bayesian.

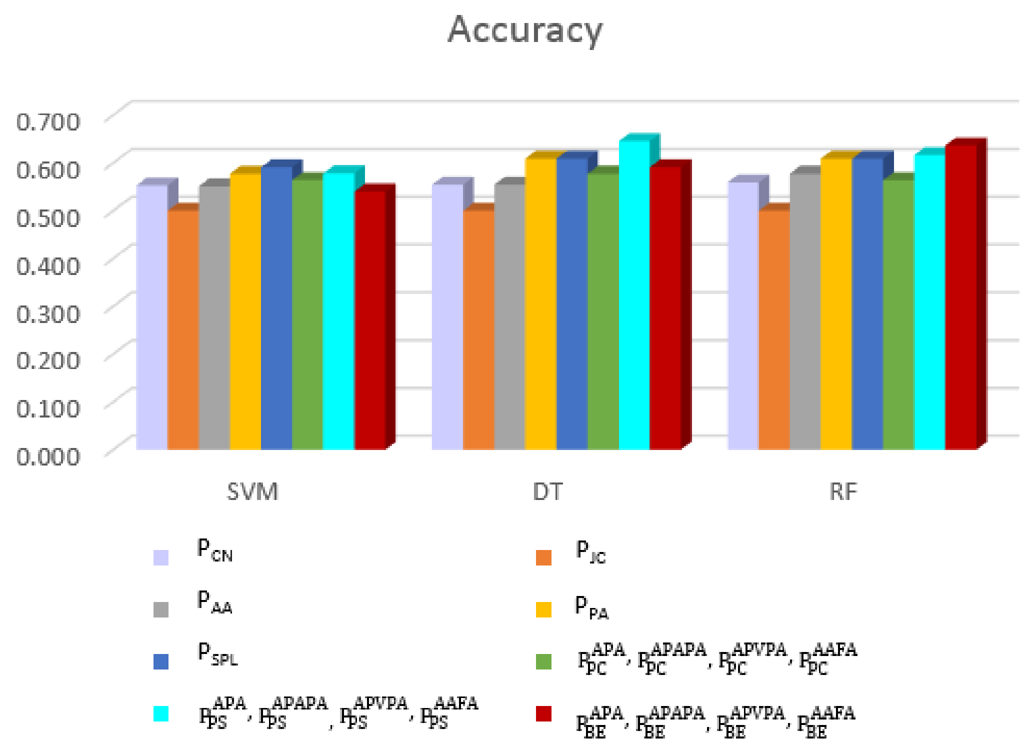

Fig. 3. Accuracy for topic diffusion prediction with topic "Data Mining"

Bảng 4. Classification results-topic "Data Mining"

Network	**Features**	**Prediction Accuracy**					
		SVM		**DT**		**RF**	
		ACC	AUC	ACC	AUC	ACC	AUC
Homogeneous network	P_{CN}	0.553	0.557	0.555	0.606	0.559	0.619
	P_{JS}	0.500	0.500	0.500	0.500	0.500	0.500
	P_{AA}	0.551	0.553	0.555	0.537	0.577	0.593
	P_{PA}	0.577	0.583	0.609	0.609	0.609	0.609
	P_{SPL}	0.591	0.603	0.609	0.609	0.609	0.609
Heterogeneous network	$P_{PC}^{APA}, P_{PC}^{APAPA} P_{PC}^{APVPA}, P_{PC}^{AAFA}$	0.564	0.606	0.577	0.560	0.564	0.606
	$P_{PS}^{APA}, P_{PS}^{APAPA} P_{PS}^{APVPA}, P_{PS}^{AAFA}$	0.578	0.591	**0.645**	**0.613**	0.616	0.648
	$P_{BE}^{APA}, P_{BE}^{APAPA} P_{BE}^{APVPA}, P_{BE}^{AAFA}$	0.540	0.601	0.591	0.591	**0.636**	**0.716**

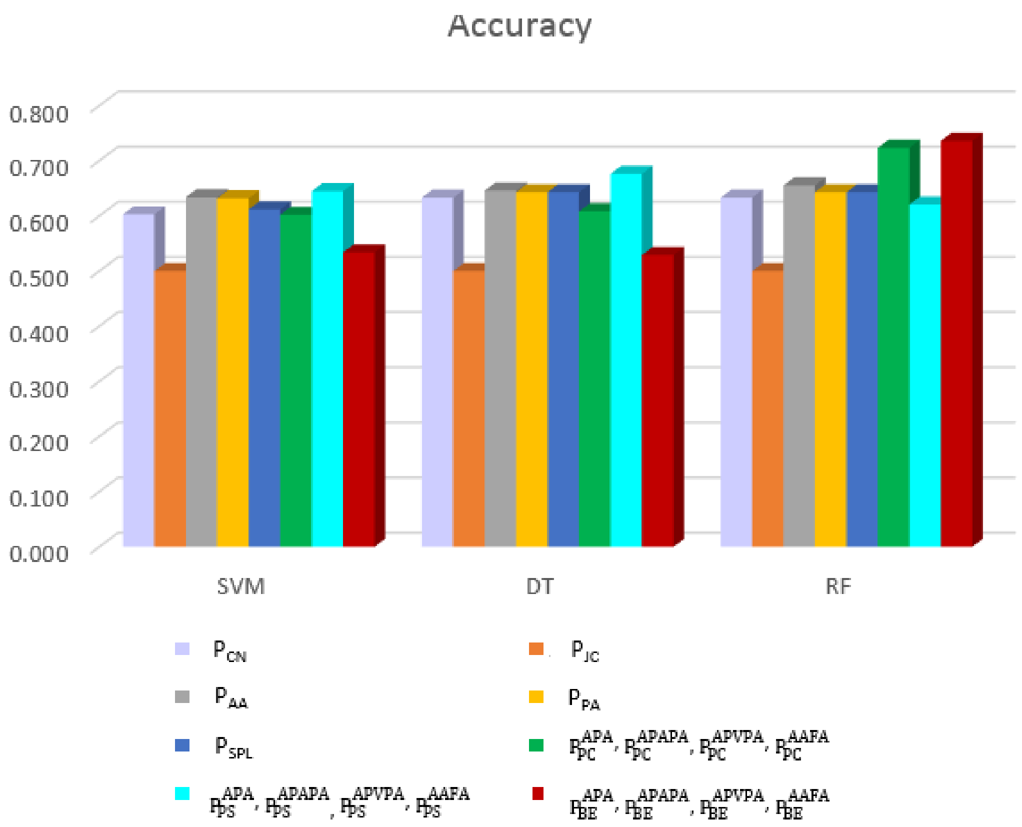

Fig. 4. Accuracy for topic diffusion prediction with topic "Machine Learning"

Bảng 5. Classification results-topic "Machine Learning"

Network	**Features**	**Prediction Accuracy**					
		SVM		**DT**		**RF**	
		ACC	AUC	ACC	AUC	ACC	AUC
Homogeneous network	P_{CN}	0.603	0.656	0.633	0.655	0.633	0.588
	P_{JS}	0.500	0.500	0.500	0.500	0.500	0.500
	P_{AA}	0.634	0.632	0.646	0.623	0.656	0.641
	P_{PA}	0.632	0.643	0.643	0.643	0.643	0.643
	P_{SPL}	0.612	0.633	0.643	0.643	0.643	0.643
Heterogeneous network	P_{PC}^{APA}, P_{PC}^{APAPA} P_{PC}^{APVPA}, P_{PC}^{AAFA}	0.602	0.615	0.608	0.607	**0.724**	**0.780**
	P_{PS}^{APA}, P_{PS}^{APAPA} P_{PS}^{APVPA}, P_{PS}^{AAFA}	0.645	0.607	**0.676**	**0.619**	0.621	0.654
	P_{BE}^{APA}, P_{BE}^{APAPA} P_{BE}^{APVPA}, P_{BE}^{AAFA}	0.534	0.617	0.529	0.518	**0.736**	**0.708**

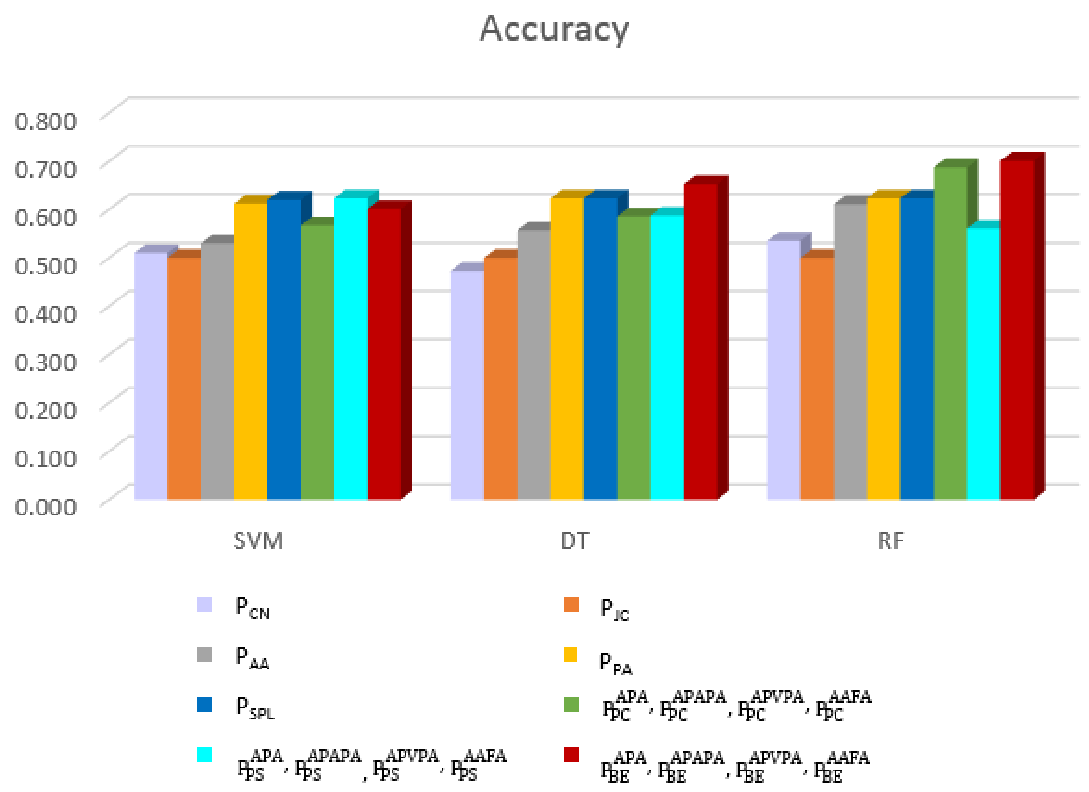

Fig. 5. Accuracy for topic diffusion prediction with topic "Social Network"

Table 6. Classification results-topic "Social Network"

Network	**Features**	**Prediction Accuracy**					
		SVM		**DT**		**RF**	
		ACC	AUC	ACC	AUC	ACC	AUC
Homogeneous network	P_{CN}	0.510	0.500	0.473	0.530	0.536	0.588
	P_{JS}	0.500	0.500	0.500	0.500	0.500	0.500
	P_{AA}	0.530	0.550	0.557	0.551	0.611	0.609
	P_{PA}	0.612	0.613	0.623	0.623	0.623	0.623
	P_{SPL}	0.620	0.589	0.623	0.623	0.623	0.623
Heterogeneous network	$P_{PC}^{APA}, P_{PC}^{APAPA}\ P_{PC}^{APVPA}, P_{PC}^{AAFA}$	0.567	0.570	0.586	0.578	**0.688**	**0.703**
	$P_{PS}^{APA}, P_{PS}^{APAPA}\ P_{PS}^{APVPA}, P_{PS}^{AAFA}$	**0.623**	**0.669**	0.588	0.559	0.561	0.542
	$P_{BE}^{APA}, P_{BE}^{APAPA}\ P_{BE}^{APVPA}, P_{BE}^{AAFA}$	0.601	0.632	0.653	0.670	**0.701**	**0.723**

6 Conclusion

In this study, we compared the performance of diffusion prediction in bibliographic network with homogeneous and heterogeneous network. We have shown that propagation prediction in a heterogeneous network achieves higher accuracy than in a homogeneous network. In addition, the Bayesian distance measure demonstrated the best effectiveness in estimating the activation probability compared to others. We believe that our work can provide significant insights into applications using the information dissemination process in a scientist's network. In the future, we will expand to leverage more types of meta paths and similar meta path metrics, topics and conduct further experiments on other networks.

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/2096252
3. 3. 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.
4. 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.
5. 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).
6. Yang, H.: Mining social networks using heat diffusion processes for marketing candidates selection. ACM (2008), https://aran.library.nuigalway.ie/handle/10379/4164
7. 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.
8. Molaei, S., Babaei, S., Salehi, M., & Jalili, M. (2018). Information spread and topic diffusion in heterogeneous information networks. *Scientific reports*, *8*(1), 1-14.
9. 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).
10. 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)
11. Akula, R., Yousefi, N., Garibay, I.: DeepFork: Supervised Prediction of Information Diffusion in GitHub p. 12 (2019)

12. 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

13. 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.

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

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

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

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

18. 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.

19. Ho, T. K. T., & Bui, Q. V. (2022). Topic diffusion prediction on bibliographic network: effect of topic modeling on activation probability measure. Hue University Journal of Science: Techniques and Technology, 131(2B), 49-63.

20. Ho, T. K. T., & Bui, Q. V. (2022). Feature's importance assessment for activation probability measure in topic's diffusion prediction. Hue University Journal of Science: Techniques and Technology, 131(2B), 33-47.

21. Newman M E J. Clustering and preferential attachment in growing networks. Physical Review Letters E, 2001, 64:025102

22. Adamic L A, Adar E. Friend and neighbors on the web. Social Networks, 2003, 25: 211-230

23. Barab´asi A L, Jeong H, N´eda Z, et al. Evolution of the social network of scientific collaborations. Physica A, 2002, 311: 590-614

24. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In: In VLDB' 11 (2011)

25. Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., & Han, J. (2011, July). Co-author relationship prediction in heterogeneous bibliographic networks. In 2011 International Conference on Advances in Social Networks Analysis and Mining (pp. 121-128). IEEE.