Expert Systems With Applications 123 (2019) 328–344

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

W-MetaPath2Vec: The topic-driven meta-path-based model for large-scaled content-based heterogeneous information network representation learning

Phu Pham, Phuc Do*

University of Information Technology (UIT), VNU-HCM, Viet Nam

ARTICLE INFO

Article history:
Received 15 May 2018
Revised 3 October 2018
Accepted 4 January 2019
Available online 15 January 2019

Keywords:
Heterogeneous information network
Representation learning
Topic similarity
Large-scaled network
Apache Spark

ABSTRACT

Recently, heterogeneous network representation learning has attracted a lot of attentions due to its potential applications. Our works in this paper are concentrated on how to leverage the output of network representation learning by combining with the topic similarity between nodes in content-based heterogeneous information network (C–HIN). These unique challenges come from the shortage of topic similarity evaluation between text-based nodes which limit the accuracy of the similarity search as well other network mining tasks. Moreover, the massive sizes of current real-world network also raises challenges for traditional standalone-based heterogeneous network analysis models. Different from previous network representation learning models, such as: Node2Vec or Metapath2Vec, our proposed W-MethPath2Vec model uses the topic-driven meta-path-based random walk mechanism for generating heterogeneous neighborhood of nodes as the learning features. Then, these learning nodes' features are used to train the learning model which is used for solving various heterogeneous network mining tasks such as: node similarity search, clustering, classification, link prediction, etc. The W-MethPath2Vec model enables the simultaneous modeling of structural and topic correlations between nodes in heterogeneous networks. Moreover, the W-MethPath2Vec model is implemented in the Apache Spark-based distributed framework which enables the capability of handling large-scaled networks. We tested our W-MethPath2Vec model with the previous state-of-the-art approaches in the real-world datasets to demonstrate the effectiveness of our proposed model.

1. Introduction

Most of the real-world information networks are heterogeneous, where the nodes and relations are of different types. In recent years, heterogeneous information network (HIN) analysis and mining have been thoroughly studied and applied in multiple disciplines. The common HINs, such as: "World Wide Web" (WWW), social networks (Facebook, Twitter, etc.) are naturally complex and very large in size (billions of nodes and links) (Sun & Han, 2012; Sun & Han, 2013; Shi et al., 2017). Similarity searching is one of the most important task in information network mining. It supports to explore the set of relevant nodes from networks. Measuring the similarity between nodes is also considered as the basis of many other data mining tasks, such as: clustering, classification, recommendation, etc. Meta-path is an important concept of most HIN mining techniques (Sun & Han, 2012). It is defined as a sequence of relations between node types which supports to distinguish the semantics among paths connecting two nodes in a network. There are several meta-path-based approaches for solving primitive tasks of HIN mining such as: similarity search (PathSim (Sun et al., 2011), HeteSim (Shi et al., 2014), etc.), ranking and clustering (RankClus (Sun et al, 2009a, 2009b), NetClus (Sun, Yu & Han, 2009), etc.). These approaches have gained notable attentions in HIN mining. Recently, researchers have intensively focused on studies related to nodes and relationships representation learning for information networks. Many algorithms have been proposed, such as: Node2Vec (Grover & Leskovec, 2016), Metapath2Vec (Dong, Chawla, & Swami, 2017), etc. Information network embedding approach can be widely applied to resolve multiple HIN mining tasks, such as: node similarity search (Sun et al., 2011; Zhang, et al., 2015), clustering, classification (Gupta, Kumar, & Bhasker, 2017), link prediction, etc. In short, the network embedding techniques support to transform the network nodes and edges into low-dimensional space of feature vectors. From these generated feature vectors of nodes and edges, we can easily process the similarity measure related tasks by using out-of-the-shelf distance measure algorithms (Euclid distance, cosine similarity, etc.)

* Corresponding author.
E-mail address: phucdo@uit.edu.vn (P. Do).

https://doi.org/10.1016/j.eswa.2019.01.015
0957-4174/

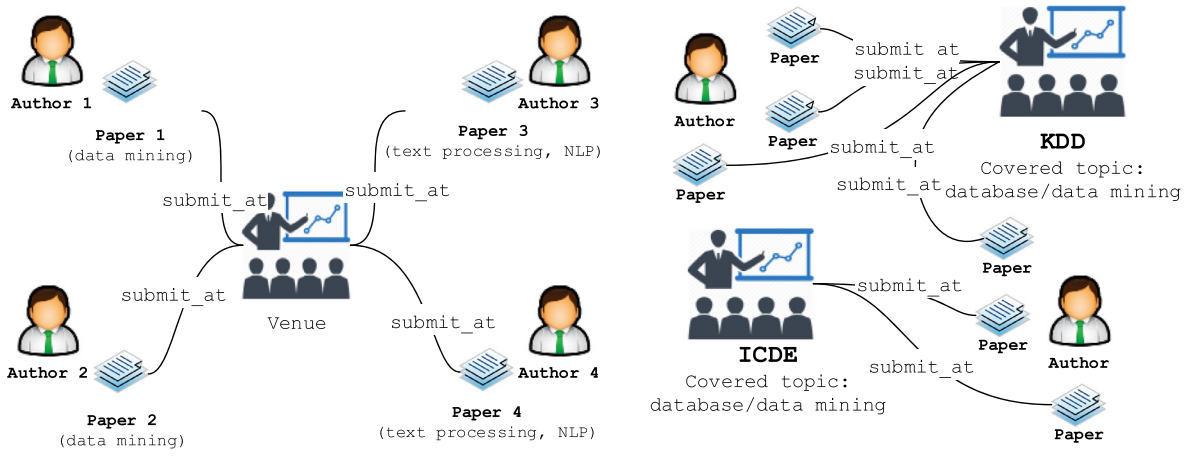

Fig. 1. Illustration of problems related to topic similarity of content-based nodes and less-linked nodes in DBLP bibliographic network.

Moreover, network embedding approach is capable for working effectively on large-scaled heterogeneous networks with millions of nodes. Because it take only one time for constructing the learning model. The embedding network model also can be applied reinforcement learning for the future data changes which takes less time than re-learning the overall network.

1.1. Problem definitions

1.1.1. Topic similarity evaluation between nodes

However, there are several challenges of both embedded and non-embedded network similarity measurement such as thorough evaluations on the topic similarity of text-based nodes such as "*paper*" in bibliographic networks (*DBLP, DBIS*, etc.) or "*comments*", "*posts*", etc. in social networks (*Facebook, Twitter*, etc.). Discovering topics of nodes in content-based heterogeneous information network (content-based HINs) is considered as an important task. Topic evaluation over network's nodes is widely applied in multiple systems like as building news or friend recommendation systems based on users' interactions on social networks. There are typical works of discovering users' topics of interests via analyzing their associated nodes like as: "*comments*", "*viewed tweets*", etc. (Michelson & Macskassy, 2010; Xu, et al., 2011). The fact is that thoroughly evaluating the topic of nodes in the network can leverage the output accuracy of the similarity search task. For example, in DBLP network, it is much more accurate for clustering "*Jiawei Hans*" with the other authors who work on "data mining". Or recommending possible new co-authorships for "*Christopher Manning*" with authors who are interesting on "*natural language processing*" and "*information retrieval*", etc. are much more meaningful than other authors who not mainly focus on these two fields. As illustrated in Fig. 1-A, a common case study in DBLP such as finding top-k similar authors with meta-path A-P-V-P-A. The meta-path A-P-V-P-A indicates the relationships of two authors who usually submit their works at the same set of venues/conferences. The assumption is that, a specific venue always has multiple tracks and each track covers different topic/sub-topic. So, it is unfair to rank "Author 1", who mainly works on "data mining" field, has the same similarity score with "Author 3" and "Author 4", who mainly work on "*text processing/NLP*" field. The "Author 1" should more similar to "Author 2", who has the same interest with "Author 1" in "*data mining*", than the left two authors.

Combining the topic attributes with nodes' relationships in similarity measure can help to improve the quality of the outputs. Additionally, by combining with topic similarity in networked data mining, we can also tackle the problem of less-linked nodes problem. Most of both homogeneous-based and heterogeneous-based similarity measure models are all considered as link-based approaches. These link-based approaches are mostly relied on the links between nodes for analyzing the similarity. Or we can say that the more nodes are connected is the more they are relevant to each other. Mostly depending on relationships between nodes leads to the drawbacks of the failure in examining less-linked nodes but in fact they are very similarity in the other aspects which do not clearly present as the network relationships. For example, two scientists who both are interesting on the "*database/data mining*" but they are rarely submit their works at the same venues/journals, as illustrated in Fig. 1-B.

1.1.2. Effective random walk mechanism via weighted paths

For most of previous network representation learning models like as: Node2Vec or Metapath2vec, work on the unweighted network which means all existed paths between two nodes are binary relations (1 for existing relation and 0 for otherwise). The walker needs to travel all the paths between two nodes in order to calculate the transitional probability (π). These computed transitional probabilities are used to rank the similarity level of destination nodes with the given source node. In some case, traveling through all existed paths between two nodes is not considered as efficiency way, especially with very large networks (Vahedian, Burke, & Mobasher, 2017). Between two nodes, there some relations are considered as important whereas the others are not. For example, such as in DBLP (as illustrated in Fig. 2), "Author 1" is known as the most active researcher who mostly contributes his works on "*data mining*" (3 papers) field, but sometime he also focuses on "*big data*" field (1 paper). It is obvious that the paths which connect "Author 1" to "Author 2" and "Author 3" are more important than the paths which connect to "Author 4" and "Author 5". Even these paths are all weighted as 1 (binary relations) but in the semantic aspect, the relationships between "Author 1", "Author 2" and "Author 3" are more stronger than the others.

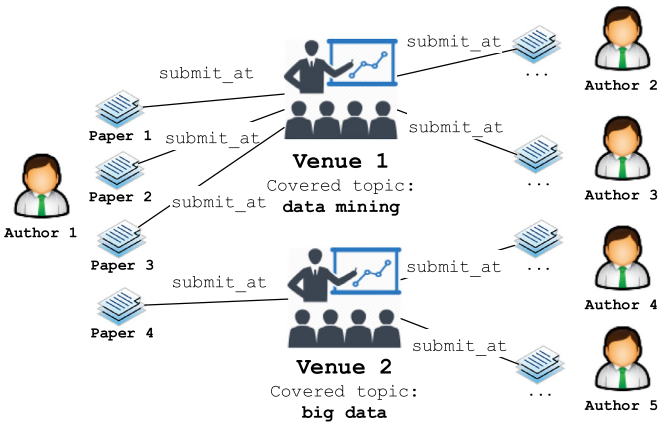

Fig. 2. Illustration of DBLP network with unweighted paths between authors.

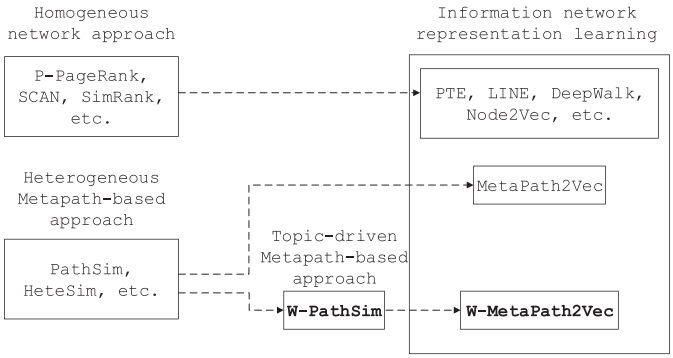

Fig. 3. The developing flow of our proposed W-MetaPath2Vec model.

Identifying which paths are more important than the others between two nodes is critical for reducing the efforts of transitional probabilities calculation. In order to evaluate the importance of paths, we need a mechanism to assign the weight for each path. Then, only paths satisfy the weight value threshold (σ) are selected for analyzing. In network representation learning, with fixed given walk length (l), for each node we only need to examine around $|l|$ amount of most important paths to generate its set of neighborhoods. Examining all possible paths which connected two nodes in this case is not really necessary. Only important paths should be taken in consideration. By limiting the number of paths which are needed for evaluation, we can leverage the performance of overall node random walk processes.

1.1.3. Big networked data challenges

Last but not least, most of the existed real-world information networks are very large in size with number of nodes can be up to billions, such as: Facebook, WWW, etc. Most of the traditional approaches of network analysis are designed to work on the standalone-based environment. It is definitely hard or impossible handle the big networked data resources like as: Facebook, Twitter, etc. with a single computer. The massive sizes of these networks beyond the capabilities of current heterogeneous network mining approaches. Therefore, we need to find a new solution for dealing with the challenge of large-scaled networks. One of the most common approach for big networked data processing is the distributed computing framework, like as: Apache Hadoop, Spark. Apache Spark is considered as a best choice for massive data handling, due to its capabilities of graph-paralleled processing, such as: GraphX, GraphFrames, etc. The GraphFrames framework can effectively support for handling common graph analysis task such as: path finding, node traversal (BFS, DFS), etc. in the manner of large-scaled networks.

1.2. Our contributions

Our overall works in this paper are mainly focused on studies of heterogeneous network representation learning problems as well as introducing the novel approach of W-MethPath2Vec model. The W-MetaPath2Vec is a topic-driven model which aims to capture distinctive features of nodes in heterogeneous network following the predefined meta-path(s). The topic similarities are obtained by evaluating the text-based nodes which are associated with investigated nodes following defined meta-paths. For example, like as "*paper*" nodes between "*author*" nodes with meta-path(s): A-P-A (author-paper-author), A-P-V-P-A (author-paper-venue-paper-author), etc. or "*comment*" nodes between "*user*" nodes with meta-path(s): U-C-P-U (user-comment-post-comment-user), etc. This topic-driven meta-path similarity measure has been introduced in our previous works, called W-PathSim model (Pham et al., 2018). Fig. 3 shows the relationship of our previous studies (the W-PathSim model) with the current proposed model. Through experiments in real-world DBLP bibliographic networks, we have proved that our proposed W-PathSim outperforms the traditional PathSim model. The W-PathSim model leverage the meta-path-based similarity measurement by combining with the topic similarity between nodes in content-based HINs. From previous achievements, in this paper, we introduce the W-MetaPath2Vec model which is an extension our previous works for topic-driven heterogeneous network representation learning.

This extended topic-driven skip-gram model supports to guide the process of extracting nodes' features. Then, these extracted features are used to train the learning model. Next, the W-MetaPath2Vec is implemented in Apache Spark-based GraphFrames distributed environment which enables to handle large-scaled heterogeneous networks. The ultimate goal of W-MethPath2Vec is to maximize process of node embedding via both link-based and topic-based evaluation in content-based heterogeneous networks. Our contributions in this paper can be summarized as five-folds, include:

- First of all, we introduce the application of LDA topic model in discovering the topic distributions of content-based nodes over content-based network. Then, these topic distributions are used for the processes of evaluating topic similarity between nodes following defined meta-paths.
- Secondly, we propose the topic-driven meta-path-based random walk mechanism which is used for generating neighborhoods of a specific node. These neighborhood nodes are used to train the network learning model. In our proposed random walk mechanism, the walker is restricted to travel to the other neighbor nodes through not only the defined meta-path(s) but also the level of topic similarity of their associated nodes. Evaluating the topic similarity while conducting the node walk makes W-MetaPath2Vec different from the previous approaches.
- Thirdly, in our proposed W-MetaPath2Vec model the walker is guided to travel within the most important paths only. These paths are selected base on their weights of topic similarity. Only paths which their weight scores satisfy the (σ) threshold are chosen for calculating the transitional probability (π). By limiting the number of paths are needed for examining by defined (σ) threshold, the W-MetaPath2Vec model can help to leverage the time-consuming performance of node embedding processes but do not influence the output accuracy.
- Next, we implement the W-MetaPath2Vec under the Apache Spark-based GraphFrames distributed graph computing frame-

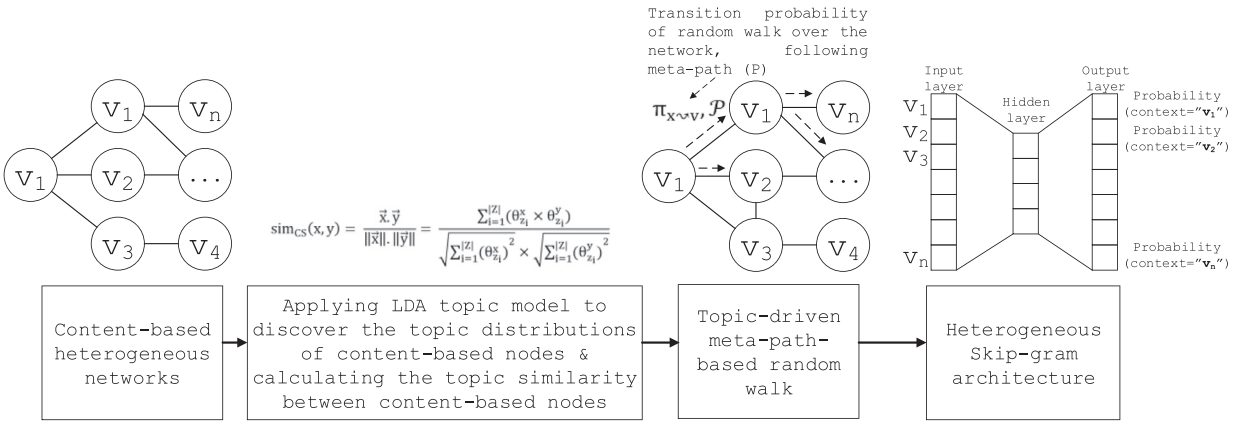

Fig. 4. The flowchart illustrates overall processes of W-Metapath2Vec model.

work in order to leverage the performance of proposed model in the context of large-scaled networks.

- Finally, we demonstrate the experimental studies on our proposed W-MethPath2Vec model with other state-of-the-art algorithms, include: DeepWalk, Node2Vec, LINE and MetaPath2Vec on the real-world DBLP/DBIS datasets. The experimental results show that the W-MethPath2Vec model is efficient for improving the quality of heterogeneous network representation learning as well as scalable for large-scaled networks with millions of nodes.

The overall processes of our proposed W-Metapath2Vec model are illustrated in Fig. 4. From the given content-based HINs, the LDA topic model is applied to extract topic distributions from the text-based nodes such as: papers in DBLP networks. After that, the topic distributions between text-based nodes are used to support the process of calculating the transitional probability (π) between nodes following defined meta-path. This is called topic-driven meta-path random walk mechanism. Finally, we applied the heterogeneous skip-gram architecture of Metapath2Vec to train the model. The network's nodes which are embedded as n-dimensional vectors can be used to solve multiple network analysis tasks, such as: node similarity search, clustering, classification, link prediction, etc. The rest of our paper is organized in four main sections. In the second section, we discuss about the previous works and preliminaries. In the third section, we formally describe about the background concepts, methodology and implementation of our proposed W-MethPath2Vec model. In forth section, we demonstrate the experimental studies on W-MethPath2Vec model. In this section, we present detailed information about datasets usage, testing scenarios, methods and evaluation metrics. We also give discussions about the output results in this section. The final section contains our conclusions about the W-MethPath2Vec approach and our future improvements. (Fig. 5–7)

2. Preliminaries & related works

2.1. Heterogeneous information network analysis

The natural principle of data is interconnected which called information networks. Interactions between data node are critical paradigm of modern information infrastructure and mining (Sun & Han, 2012). Heterogeneous information networks are becoming prevalent and widely applied in several real-world applications.

From the past most of the information network mining techniques are considered homogeneous-based approach. In homogeneous network all nodes and links are considered as a same type. Many well-known algorithms are mainly applied for homogeneous networks, like as: Personalized PageRank (PPR) (Jeh & Widom, 2003), SimRank (Jeh & Widom, 2002). However, most of real-world networks contain multiple types of nodes and link, called heterogeneous information network (HIN) (Fig 5). Recently, the baseline of HIN mining has been carefully considered and studied to solve the problem of multi-typed nodes and links. In HIN analysis, meta-path (Sun, et al., 2011) is an important concept, proposed by Sun et al. (2009a, b). A Meta-path presents the semantic meaning of a specific relation pattern which connect two nodes in the network. From the baseline of meta-path concept, many algorithms have been proposed for solving common heterogeneous network analysis tasks. For example in PathSim (Sun et al., 2011) model, it supports to calculate similarity score of two same-typed nodes (e.g. author-author, user-user). The similarity score between two nodes are calculated based on number of path instances between them. These path instances are identified base on a specific meta-path (e.g. author-paper-author, user-comment-user, etc.).

2.2. Information network representation learning

Recently, network representation learning is considered as most recent emerged approach in knowledge discovering from networked data. Since the proposal of Word2Vec (Mikolov et al., 2013a, b), the skip-gram model has significantly advanced the studies on information network embedding. The principal concept of information network analysis and mining starts from the exploration of the network structural properties. There are common properties of network nodes and links which are usually used in social network and community analysis algorithms, such as: between-ness, triangle count, modularity, etc. In fact, these traditional network structural properties require a lot of informative, discriminating of specific domain knowledges as well as computing resource. Therefore, many researchers had tried to find the new effortless way for effectively extracting network's structural properties. From the emergence of representation learning, there are several network embedding techniques have been extensively studied and proposed in order to automatically discover the network's structural properties. Then, these discovered properties are mapped into a low-dimensional space. Recently, information network representation leaning has gained tremendous attentions from researchers due to its wide applications. The outputs of network representation can be used as the input for multiple network mining tasks, such as: node similarity search, clustering, classification, link prediction, etc.

Information network node embedding. Formally, the ultimate goal of information network representation learning is to learn a function: $V \rightarrow \mathbb{R}^d$,with V is set of nodes in a network, denoted as a graph: $G=(V, E)$. The function (f) support to map each node to

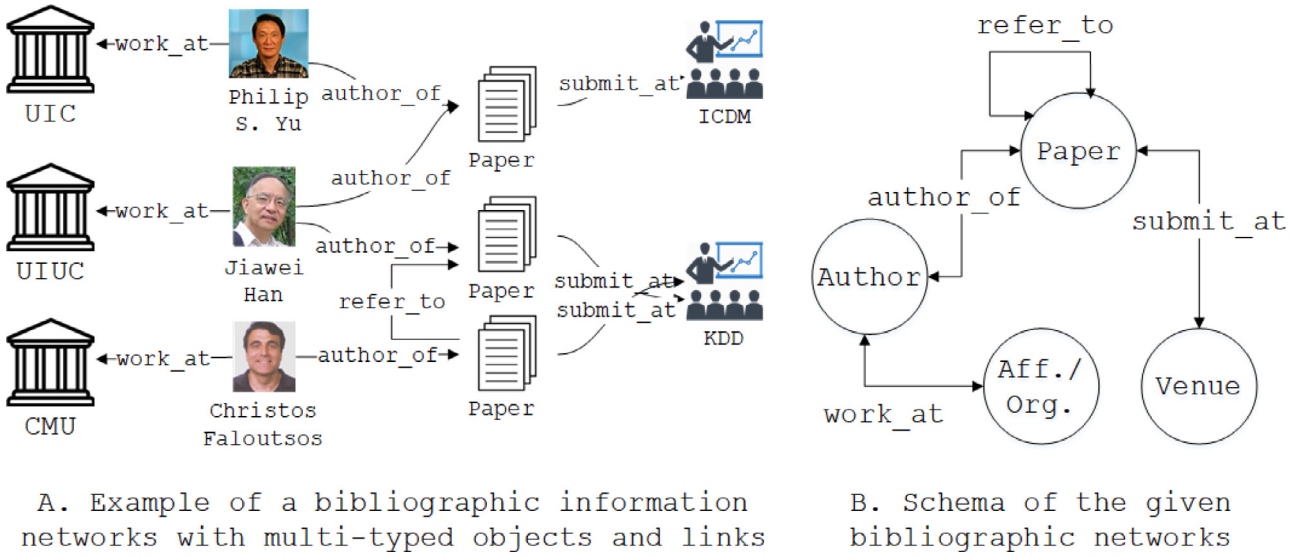

Fig. 5. Example of heterogeneous network and network schema.

a d-dimensional vector ($d \ll |V|$) which captures distinctive properties, also called features, of that node. In network representation learning, the principle of node feature representation uses the node's neighborhood as the distinctive features. This approach is largely influenced by the proposed skip-gram model of Mikolov et al., 2013a, b). The proposal of skip-gram model in the Word2Vec approach, which aims to learn continuous feature representations for words within specific word-contexts.

Supervised vs. unsupervised representation learning. Like as any supervised machine learning approach which requires a lot of informative as well expert knowledge resources (Bengio, Courville, & Vincent, 2013). The supervised learning models depend on labelled dataset to construct feature vector representation for network nodes and edges. The supervised feature learning approaches (Rasmus, et al., 2015; Chen, et al., 2016) can provide high accuracy outputs. The challenge in network feature learning is defining an objective function to ensure the balances in computational efficiency as well as the output accuracy (Grover & Leskovec, 2016). In fact, the supervised approaches mainly encounter problems related to high cost of constructing the annotated dataset. Moreover, with the complexity of heterogeneous network, supervised approaches also face with excessive increase in the number of features from network nodes and edges that need to be considered. In heterogeneous networks, multiple type of nodes and edges carry different distinctive features. Therefore, it seem impossible to define all the features which are needed for the networks learning process. Instead of constructing handcrafted network feature design, several unsupervised network representation learning models have been proposed, such as: LINE, DeepWalk, Node2Vec, Metapath2Vec, etc. These models enable to automatically extract the distinctive features of nodes by considering the paths traversed via random walks (Fig. 6) over the networks.

Homogeneous vs. heterogeneous network representation learning. In the past, many network embedding techniques (Fig. 7) like as LINE, DeepWalk, Node2Vec, etc. are designed to work on homogeneous networks only. These models treat all nodes and edges of the networks as a singular type. To challenge problems of heterogeneous nodes and links in network, novel approaches have been introduced like as Metapath2Vec (Dong et al., 2017). The Metapath2Vec model proposes the meta-path-based random walk mechanism. This novel meta-path guided random walk mechanism enable to restrict the walks to follow the defined the meta-path(s). Dong et al. (2017) also proposed a new mechanism for computing the transitional probability between two nodes via meta-path-based walks. The Metapath2Vec model supports to differentiate different types of nodes and links hence it leverage the accuracy of heterogeneous network embedding output.

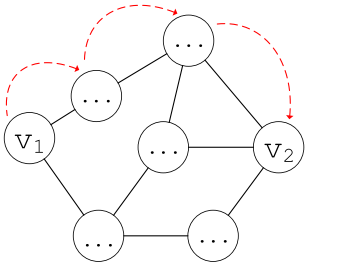

Fig. 6. Similarity measurement between two nodes on the information networks via evaluating the number of paths between nodes.

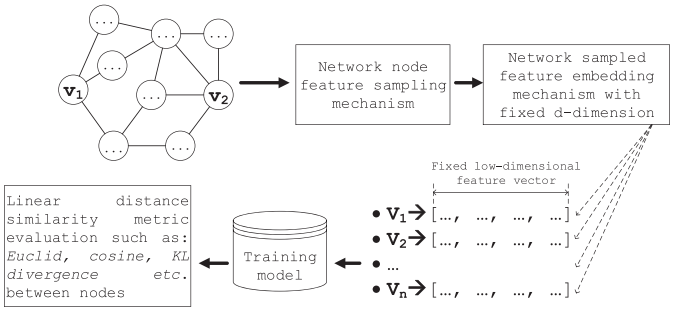

Fig. 7. Similarity measurement between two nodes on the information network via network embedding approach. Nodes are represented as feature vectors.

2.3. Related works & motivations

Information network representation attracts lots of attentions in recent years. From the past, many studies have been devoted to information network representation. Most of these studies (Grover & Leskovec, 2016; Tang, et al., 2015; Tang, Qu, & Mei, 2015) are focused on homogeneous networks which only contain single-typed nodes and links. There is no doubt that large number of f social and information networks are heterogeneous in nature which involve diversity of types in nodes and links. In fact, the important

Table 1
Differences between labelled topics and latent topics which are represented by set of keywords/concepts.

Human-named topic	Latent topic represented by set of keywords/concepts
artificial intelligence	*"cognitive", "deep learning", "neural network", "robotics", "computer vision", etc.*
data mining	*"association rules", "pattern mining", "graph mining", "database", "frequent item-set", etc.*
natural language processing	*"text processing", "stop-words", "machine translation", "treebank", "dependency tree", etc.*

baseline of network embedding technique is aimed to preserve the proximity between a node and its neighborhood (context) (Grover & Leskovec, 2016). With homogeneous embedding approaches, it is hard to distinguish the difference in type of specific node and its neighborhoods as well as their linked relationships. Therefore, we can't apply representation learning models that are specially designed for homogeneous networks.

To study such information networks with multiple types of nodes and links, (Dong et al., 2017) proposed Metapath2Vec model which uses meta-path-based random walk method to generate heterogeneous neighborhoods with network semantics for various types of nodes. It leverages a heterogeneous skip-gram model to perform node embedding in information networks. Experiments in multiple heterogeneous networks show that proposed Metapath2Vec model are able to outperform state-of-the-art embedding models in various heterogeneous network mining tasks. Some studies have inspired the meta-path-based random walk technique of Dong et al., such as: HERec model (Shi, et al., 2018) for solving recommendation task, PME model (Chen, et al., 2018) is an embedding model for handling link prediction task in heterogeneous network. However, the Metapath2Vec model is considered as a link-based approach which are designed to mostly concentrate on the paths between nodes to generate the set of heterogeneous neighborhoods. It fails to recognize the other aspect such as topic similarity between nodes following the used meta-paths. Most of real-world heterogeneous networks contain large amount of content-based nodes such as "*papers*" in bibliographic networks (DBLP, DBIS, etc.), "*comments*", "*posts*", etc. in social networks (Facebook, Twitter, etc.). These content-based nodes not only are rich in information but also have the similarity in their covered topics. Topic is one of the most important sources of information for text-based nodes, therefore we can use the topic similarity between nodes in combine with meta-path-based random walk technique to leverage the process of node embedding in heterogeneous network. Moreover, recent studies also encounter the shortage of thorough evaluation on very large-scale network analysis approach. Recent models, such as: LINE, PTE, Node2Vec, Metapath2Vec, etc. are only designed to work on standalone-based computing environment which is incapable to work on very large-scale social networks, like as: Facebook, Twitter, etc. Therefore, it is necessary to develop a new model which can be implemented in distributed computing environment for handling large-scale networks. These are the major differences between our works in this paper with previous state-of-the-art embedding models.

3. Methodology and implementation

In this section, we introduce three main approaches of our W-MetaPath2Vec model which includes:

- The approach of applying LDA model in discovering topic distributions from the given content-based heterogeneous information networks.
- Next, we present the mechanism of topic-driven meta-path-based random walk which is used to extract neighborhood nodes from a given source node. These extracted neighborhoods play as learning features which are used to feed the network learning model.
- Finally, we proposed the implementation of W-MetaPath2Vec model under the Apache Spark-based GraphFrames distributed computing environment. By designing to work on distributed computing environment of Apache Spark, the W-Metapath2Vev mode is scalable for large-scaled content-based heterogeneous networks with millions of nodes.

3.1. Applying LDA topic model for topic discovering in content-based heterogeneous network

Most of real-world network nodes usually have rich information, like as text content or meta-data descriptions. For example, website nodes in WWW have their webpages' content and HTML meta-data descriptions, the published papers in bibliographic networks (DBLP, DBIS, etc.) or comments, posts of users in Facebook, Twitter. Therefore, we come up with a proposal of enriching the information network representations from both network structure and topic attribute of text-based nodes.

In this section, we introduce an approach of using topic model to discover the latent topic from the text-based nodes. The text-based nodes are very popular in most real-world information networks, for example paper nodes in DBLP, DBIS. The LDA topic model is one of the most common approach for discovering latent topics in text. From the past, multiple researches have been focus on topic modeling approach, such as: latent semantic indexing (LSI) (Deerwester, et al., 1990), probabilistic latent semantic indexing (pLSI) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003; BLEI, 2012) model. The most well-known topic model is LDA due to its high performance in accuracy and easy implementation. From given text corpus, the LDA topic model can effectively enables to capture two probabilistic distributions, include:

- The words (w: $w \in W$) distribution over latent topics ($z \in Z$), denoted as: (ϕ_w^z)
- The latent topics (z) distribution over documents (d: $d \in D$), denoted as: (θ_z^d).

Latent topic representation. In topic model, a latent topic is defined as a set of keywords which sufficiently explains the meaning of that topic. Different from the human-labelled topic, which incorporate directly to specific human named labels, such as: "*artificial intelligence*", "*data mining*", "*natural language processing*" etc. (as illustrated in Table 1). Most of the topic model approaches for text corpus use the group of keywords to describe about specific topic or knowledge domain.

Binary vs. weighted links between nodes in information networks. Most of the current information network links/relationships such as: DBLP bibliographic network or social networks like as: *Facebook, Twitter,* etc. are merely the binary relations.

It means that [1] for existed link and [0] for otherwise, such as pairwise relationship of: $author \xrightarrow{co\_worker} author$, $author \xrightarrow{author\_of} paper$, $paper \xrightarrow{submit\_at} venue$, etc. (as shown in Fig. 8-A). Therefore, similarity search strategies on information network is considered as link-based rather than attribute-based approaches.

As aforementioned problems related to link-based similarity evaluation, this type of approach can't evaluate the non-linked

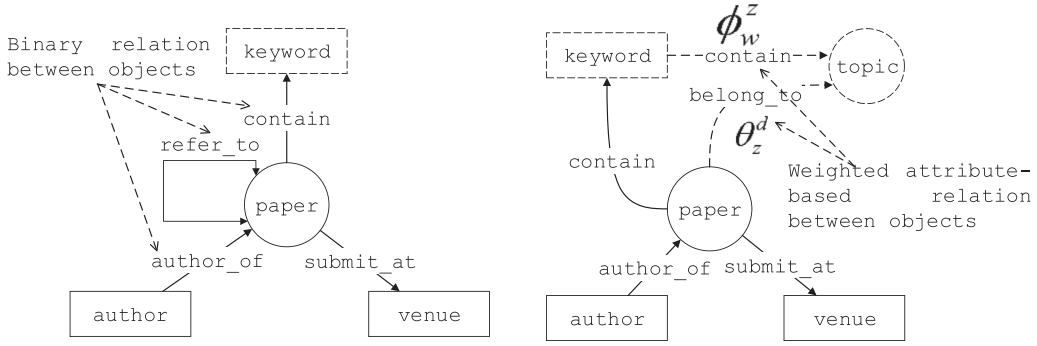

Fig. 8. Illustration of difference between traditional content-based HINs and schema-enriched content-based HINs by using probabilistic LDA topic modeling.

nodes, such as relevant authors who are interesting on the same researching fields but never submit their works to the same venues/journals. Hence, we need to solve this problem by enriching the network's schema of given information network. We add extra weighted attributes such as topic in the information network.

Moreover, the binary-based relation also leads to problem related to thoroughly evaluate the level of similarity between linked nodes, for example, such as a citing set of a specific paper, which represented as a semantic path: $paper \xrightarrow{refer\_to} paper$, assuming that the source paper is mainly focused on the hot issues related to "information retrieval" topic, more than 80% of the references are covered "information retrieval" papers, only the left 20% referred papers are on the other topics, hence it seem unfair when we treat the set of 80% same-topic cited papers as the same with the left 20% other-topic cited papers.

Enriching schema in content-based heterogeneous via LDA topic modeling. In general, the LDA topic model supports to produce the generative models of extracted latent topic distributions over the set of paper nodes, donated as: $P(z_i|d_j) = \theta^{d_j}_{z_{(i,\ i\in|Z|)}}$, with ($z_i$: $z_i \in Z$) . It presents for the probabilistic distribution of (i)th topic over document (d_j: $d_j \in D$). Then, each paper node in the given content-based HINs can be represented as a weighted vector $\vec{d}$ as following (as shown in Eq. (1)):

$$\overrightarrow{d_j} = \begin{bmatrix} P(z_1|d_j) \\ \cdots \\ P(z_{(i,\ i\in|Z|)}|d_j) \end{bmatrix} = \begin{bmatrix} \theta^{d_j}_{t_1} \\ \cdots \\ \theta^{d_j}_{z_{(i,\ i\in|Z|)}} \end{bmatrix} \quad (1)$$

Where,

- $\overrightarrow{d_j}$, represents for the distributed topic vector weight of (j)th document.
- $P(z_{(i,\ i\in|Z|)}|d_j)$, represents for the probabilistic distribution (i)th extracted latent topic (z) over the (j)th document.

LDA topic model helps to obtain the distributions of latent topics over the paper nodes in the given content-based HINs. Therefore, the information network's schema has been enriched (as illustrated in Fig. 8-B).

3.2. Topic-driven meta-path-based random walk

In this section, we introduce the approach of topic-driven meta-path-based random walk procedure for guide the random walkers within schema-enriched HINs.

Topic-driven similarity as weighted link-based attribute via cosine similarity. By applying LDA topic model in the given content-based HINs, we can obtain the probabilistic distributions of latent topics over the given "*paper*" nodes, or in the other ways, each paper node are now represented as a weighted vector $\vec{d}$, with $|Z|$-dimensions. Hence, we can apply available vector distance measurement, such as: cosine similarity, Euclid distance, etc. to compute the similarity weights between two paper nodes.

In this study, we use the cosine similarity to compute the similarity score, denoted as: $sim_{CS}(x, y)$ of two paper nodes (x) and (y) as following equation (as shown in Eq. (2)):

$$sim_{CS}(x,\ y) = \frac{\vec{x}.\vec{y}}{\|\vec{x}\|.\|\vec{y}\|} = \frac{\sum_{i=1}^{|Z|}(\theta^x_{z_i} \times \theta^y_{z_i})}{\sqrt{\sum_{i=1}^{|Z|}\left(\theta^x_{z_i}\right)^2} \times \sqrt{\sum_{i=1}^{|Z|}\left(\theta^y_{z_i}\right)^2}} \quad (2)$$

Where,

- Z, is set of extracted latent topics which discovered from a content-based HIN.
- $\theta^x_{z_i}$ and $\theta^y_{z_i}$, represent for the probabilistic distributions of latent topic (i)th over (x) and (y) paper node, respectively.

Within a given topic-driven similarity score between nodes, we attach these similarity score as the transitional probability for the meta-path-based random walks. The new transitional probability (π) for specific meta-path, denoted as symmetric path: $\mathcal{P} = V_s \to \cdots \to V_k \to \cdots \to V_e \leftarrow \cdots \leftarrow V_{k^-} \leftarrow \cdots V_{s^-}$ with (k) is the content-based nodes. The walker start from node (x) and stop at node (v) with the transitional probability (π) is computed as following (as shown in Eq. (3)):

$$\pi_{x\sim v},\ \mathcal{P} = \begin{cases} with\ e(x,v) \notin E \begin{cases} with:\ \phi(x) = \phi(v), & \dfrac{\sum_{i,i\in|E(x\sim v)|}\frac{1}{|N(V_i)|} + sim_{CS}(k, k^-)}{\lambda} \\ with:\ \phi(x) \neq \phi(v), & 0 \end{cases} \\ with\ e(x,v) \in E, & \frac{1}{|N(x)|} \end{cases}, \quad (3)$$

Where,

- N(x), is a set of directed neighbourhood nodes of given starting node (x), noticing that the set of neighbourhood nodes must be same-typed with the ending node (v).

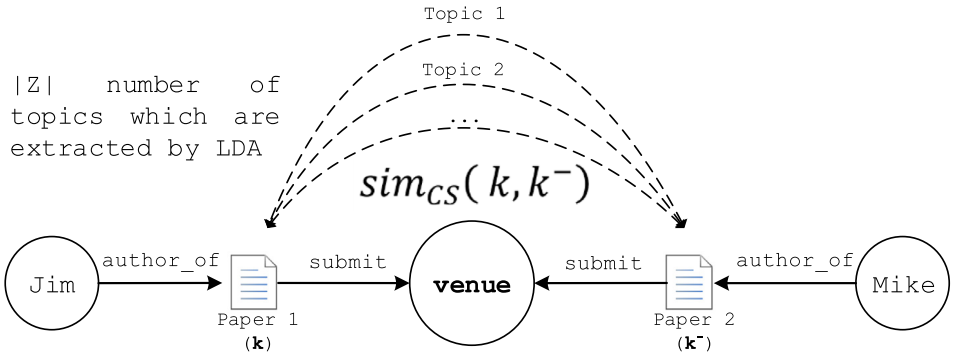

Fig. 9. W-MetaPath (APVPA) Jim & Mike author had published their paper at common venues / conferences.

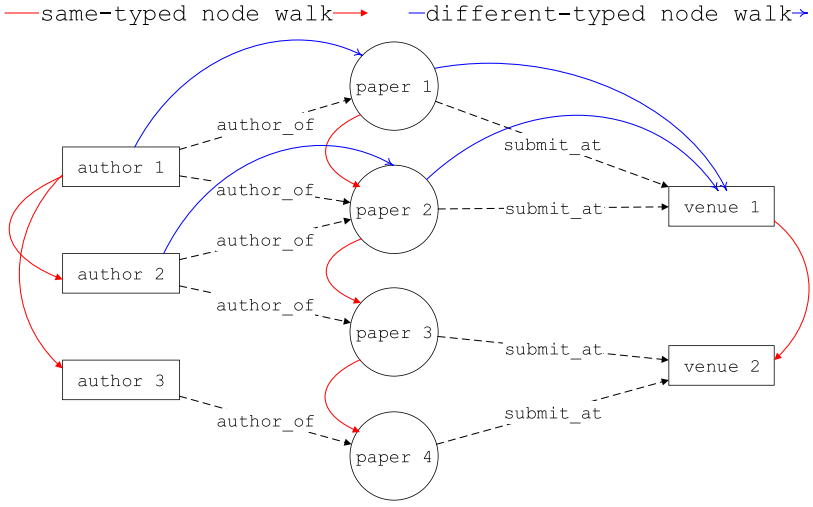

Fig. 10. Illustration of topic-driven random walk with meta-path "APVPA". (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

- $\frac{1}{|N(x)|}$, is the transitional probability of general node walk. For any walk starts at node (x) and reaches to the next node (v). In this type of walk, we don't care about the node type of (x) and (v).
- e(x, v)$\notin$E and e(x, v) $\in$ E, presents for the non-existed any directed link between node (x) and node (v), and otherwise (existed at least 1 directed link), respectively.
- $\sum_{i,i\in|E(x\sim v)|}\frac{1}{|N(V_i)|}$, is presented for the total transitional probability for the walker to travel from node (x) to node (v).
- $sim_{CS}(k,k^-)$, is the cosine similarity between two symmetric content-based nodes (such as paper node in meta-path A-P-V-P-A or V-P-A-P-V), following the given meta-path ($\mathcal{P}$) (as illustrated in Fig. 9).
- λ, the global normalizing constant.

Most meta-paths are defined as a symmetric way (Sun & Han, 2013; Shi et al., 2017; Sun et al., 2011), which means the node's type of a specific node (V_k) at (k) position on the left side is the same with the a node (V_{k^-}), at (k^-) opposite position, or: $\phi(V_k) = \phi(V_{k^-})$. For example, common meta-paths in DBLP, such as: $author \rightarrow paper \leftarrow author^-$, $author \rightarrow paper \rightarrow venue \leftarrow paper^- \leftarrow author^-$, etc.

Topic driven heterogeneous meta-path-based random walk. The mechanism of our topic-driven meta-path-based random walk in schema-enriched HINs is simple. The walker is guided to follow the sequential order of specific defined meta-path, take "A-P-V-P-A" as an example. The walker travels to the pairwise nodes which belong to the same type such as: "author"-"author", "paper"-"paper", etc. (as illustrated by the red-colored lines in Fig. 10). For the starting node as (x) and ending node as (v), with: $\varphi(x) = \varphi(v)$, the is controlled by two type of transitional probability, the first one is general node walk and the second one is topic-driven similarity node walk. In fact most of the common meta-path-based same-typed node walks are undirected, means that the walker rarely starts and ends directly to same-typed nodes.

For example, in order to form a walk of author's node like as: ["author 1", "author 2", "author 3"] following the given meta-path "A-P-V-P-A", the walker need to travel multiple full directed meta-path-based walks, such as: [**"author 1"**, "paper 1", "venue 1", "paper 2", **"author 2"**], [**"author 2"**, "paper 3", "venue 2", "paper 4", **"author 3"**], etc. (as illustrated in Fig. 10). Because this approach is the meta-path-based random walk, hence the walker are restricted to follow strictly the sequential order of meta-path ($\mathcal{P}$). Different from Node2Vec model, the walker is not allow to take the shortcuts such as: [**"author 1"**, "paper 2", **"author 2"**]. This restriction ensures the semantic relationships between same-typed nodes still be captured during the random walk process.

Overall processes of this type of walk is described in Algorithm 1. First of all, the transitional probability (π) of general node walk from (x) to (v) is computed, because this is an undirected walk, hence there might be several paths between these nodes. The general transitional probability is the sum of transitional probabilities of all nodes between them, as: $\sum_{i,i\in|E(x\sim v)|}\frac{1}{|N(V_i)|}$. For computing the transitional probability of topic-driven similarity node walk, we only need to calculate the cosine similarity (as shown in Eq. (2)) of two content-based nodes

Algorithm 1 Pseudo code for undirected same-typed node random walk.

Input: the heterogeneous network, denoted as: $G=(V, E)$, with pre-defined meta-path ($\mathcal{P}$), with starting node (u) with type is (t), with the defined walking length is denoted as: (l) and defined threshold (σ).
Output: the set of same-typed neighborhood nodes via undirected random walk which incorporating to node (u).

1: **Function** $W - MetaPath\_RandomWalk\_(u, G, \mathcal{P}, l)$:
2: **Create**: $walks = [u]$, neighborhood_nodes = []
3: **Do** meta-path ($\mathcal{P}$) traversal, **for all**: (v_{next}) in (V_t): #condition for this node traveling is: $\varphi(u) = \varphi(v_{next})$.
4: **Compute**: $sim_{CS}(k, k^-)$ of two nodes (u) and (v_{next}) (Eq. 2).
5: **If** $sim_{CS}(k, k^-) \geq \sigma$ **then**:
6: **Compute**: $\pi_{u \sim v_{next}}$, $\mathcal{P}$, #the value of $\pi = \frac{\sum_{i, i \in |E(u \sim v)|} \frac{1}{|N(V_i)|} + sim_{CS}(k, k^-)}{\lambda}$ (following the Eq. 3).
7: **Add**: neighborhood_nodes $\leftarrow \langle v_{next}, \pi_{u \sim v_{next}}, \mathcal{P} \rangle$
8: **End if**
9: **End do**
10: **Sorting:** neighborhood_nodes (reversed order by π)
11: **For** i = 1 **and** i $\rightarrow (l-1)$:
12: **Add:** $walks \leftarrow$ neighborhood_nodes[i]
13: **End for**
14: **Return** *walks*
15: **End function**

(k and k^-) – within this meta-path-based walk. For example: sim_{CS}(*paper* 1, *pa*per 2) in ["author 1", "paper 1", "venue 1", "paper 2", "author 2"] walk and sim_{CS}(*paper* 3, *pa*per 4) in ["author 2", "paper 3", "venue 2", "paper 4", "author 3"] walk etc.

The main purpose of applying topic-driven transitional probability is to encourage the walker to visit the nodes which are considered as similar in their covered topics.

Starting from each node (u) in the given network (G), the walker travels through paths in order to reach a specific next same-typed node (v_{next}). All paths must follow the pattern of defined meta-path ($\mathcal{P}$) (line 3). At each pairwise nodes (u, v_{next}), we compute the similarity score $sim_{CS}(k, k^-)$ of two given nodes (u) and (v_{next}), if the $sim_{CS}(k, k^-)$ (Eq. 2) (line 4) score satisfy (large or equal) the defined threshold (σ), it will pass to the next stage. We restrict the walker to travel to unsatisfied paths in order to reduce the time for completing the overall random walk process. These unsatisfied paths are considered less important than the other due to they do not present high topic similarity level between two (u) and (v_{next}) than the other paths. In next stage, the transitional probability (π) (Eq. (3)) of (u) and (v_{next}) is calculated and stored (line 6 and 7). This process is loop until there is no more linked node via meta-path ($\mathcal{P}$) with node (u). Finally, we sort the set of extracted neighborhoods of node (u) and get top-|l| neighborhoods base on the computed transitional probability (π).

The output results for this walk mechanism are the set of same-typed nodes which are linked to a given node (u) with (l) length. For example, starting with venue nodes in DBLP network, the set of neighborhoods which associate with source node are generated as following:

- [***KDD***], WWW, VLDB, PAKDD, ICDM, SDM, …
- [***CVPR***], NIPS, ICCV, ICML, SIGGRAPH, …
- [***ACL***], NAACL, LREC, COLING, NIPS, …

Another example of starting with author nodes, the set of neighborhoods which associate with source node are generated as following:

- [***Jiawei Han***], Philip S. Yu, Jian Pei, Ming-Syan Chen, Christos Faloutsos, Hans-Peter Kriegel, …
- [***Christopher D. Manning***], Bing Liu, David McClosky, John Bauer, Jenny Rose Finkel, Steven Bethard, …

As shown from the two examples, the set of generated neighborhoods of each node via our random walk mechanism depends not only on how much these nodes are connected but also their similarity in the covered topics. Like as KDD venue node which covers in database/data mining domain, tends to associate with other same-field venue nodes (WWW, VLDB, PAKDD, etc.). Or in example with author nodes, Christopher D. Manning author tends to associate with the other authors who are mostly interesting on "*natural language processing*" and "*information retrieval*" fields.

Our study in this paper is mainly focused on the content-based heterogeneous network hence this type of random walk is only applied for meta-paths which must contain at least one content-based node. The fact is that, in content-based information network mining, common meta-paths which mostly applied are all have the content-based node inside, therefore, it is promising that this type of topic-driven walk can widely applied in different cases, such as in content-based social networks: $user \overset{reply\_to}{\rightarrow} comment \overset{reply\_to}{\leftarrow} user^-$, $user \overset{like}{\rightarrow} post \overset{like}{\leftarrow} user^-$, etc.

Heterogeneous negative sampling. In content-based information network mining, the topic-driven meta-path-based random walk strategy ensures that the semantic relationships between different-typed nodes can be properly incorporated into Skip-gram model which have been introduced from previous works (Grover & Leskovec, 2016;Dong et al., 2017). The W-MetaPath2Vec is aimed to distinguish the context of a specific node (u) which is conditioned on its types when sampling its set of neighborhood nodes in given node context, (c_u). Inspiring from the proposals of Node2Vec (Grover & Leskovec, 2016), MetaPath2Vec (Dong et al., 2017) and PTE (Tang et al., 2015), our W-MetaPath2Vec model is applied to normalize the transitional probability of specific node type (t). There are two main approaches of applying soft-max function for different-typed node (directed walk) and same-typed node (undirected walk).

For the set of same-typed nodes, the soft-max function is defined as following (Eq. (4)) this approach is inspired from the idea of MetaPath2Vec (Dong et al., 2017) where supports to obtain the set of multinomial distributions for neighborhood nodes which are same type with (u) in the output layer of the skip-gram model:

$$p(c_u|u;\theta) = \frac{e^{X_{c_u} \times X_u}}{\sum_{v \in V_t} e^{X_v \times X_u}}, \ \phi(u) = t \tag{4}$$

Where,

- V, V_t, are represented for the set of nodes of overall given network and set of nodes which are same type with given node (u), where: $\varphi(u) = t$, respectively.
- X_v, X_u, X_{c_u}, are represented for the feature vector of node (v), (u) and node context (c_u), respectively.

In Table 2, we summarize the different approaches of top-k similarity search in DBLP information networks by using different approaches, includes: PathSim, LINE, Metapath2Vec and our

Table 2
Case studies of top-k similarity searching in heterogeneous network via different approaches.

PathSim (Sun et al., 2011)	LINE (Tang et al., 2015)	MetaPath2Vec (Dong et al., 2017)	W-MetaPath2Vec
Top-5 relevant authors for "**Christos Faloutsos**"			
Jiawei Han	Jian Pei	Churu Aggarwal	Jian Pei
Jianyong Wang	Ke Wang	Jian Pei	Churu Aggarwal
Huan Liu	Hui Xiong	Philip S. Yu	Jiawei Han
Churu Aggarwal	Vipin Kumar	Hui Xiong	Hanghang Tong
Jieping Ye	Eamonn J. Keogh	Vipin Kumar	Ke Wang
Top-5 relevant venues/conferences for "**PKDD**"			
ICDM	SIGMOD	KDD	IJCAI
SIGMOD	PKDD	ICDM	SIGKDD
DMKD	ICDM	SDM	AAAI
SDM	SDM	DMKD	ML
KDD	SIGKDD	PAKDD	ICML

Algorithm 2 Pseudo code for undirected same-typed node random walk which is implemented under Spark GraphFrames.

Input: the heterogeneous network, denoted as: $G=(V, E)$, with pre-defined meta-path ($\mathcal{P}$), with starting node (u) with type is (t), with the defined walking length is denoted as: (l) defined threshold (σ).
Output: the set of neighborhood nodes via directed random walk which incorporating to node (u).

Construct: GF (GF_Vertices ← V, GF_Edges ← E)
Generate: motif pattern: $motif\_P$, from meta-path: $\mathcal{P}$
Function $Distributed\_W-MetaPath\_RandomWalk$(u, GF, $motif\_P$, l):
  Create: $walks=[u]$, neighborhood_nodes = [ ]
  Do: path_instances = **motif_finding**(u,*,$motif\_P$)
  For path_instance(p) **in** path_instances:
    Set: v_{next} = p.get_end_node() *#directly get the end node because motif finding can ensure all path instances must follow the meta-path (types of source and target nodes are the same)* $\varphi(u)=\varphi(v_{next})$
    Compute: $sim_{CS}(k,k^-)$ of two nodes (u) and (v_{next}) (Eq. (2)).
    If $sim_{CS}(k,k^-) \geq \sigma$ **then**:
      Compute: $\pi_{u\sim v_{next}}$, $\mathcal{P}$, #the value of $\pi = \frac{\sum_{i,i\in|E(u\sim v)|} \frac{1}{|N(V_i)=GF.out\_degree(u)|} + sim_{CS}(k,k^-)}{\lambda}$ (following the Eq. (3)).
      Add: neighborhood_nodes ← $\langle v_{next}, \pi_{u\sim v_{next}}, \mathcal{P}\rangle$
    End if
  End for
  Sorting: neighborhood_nodes (reversed order by π)
  For i = 1 and i → ($l-1$):
    Add: $walks$ ← neighborhood_nodes[i]
  End for
  Return $walks$
End function

proposed W-MetaPath2Vec model. The main difference of our topic-driven approaches (W-PathSim, W-Metapat2Vec) in compare with previous link-based approach, like as: PathSim, Line, Metapath2Vec, etc. is that we take the advantage of topic similarity between paper nodes following A-P-V-P-A and V-P-A-P-V meta-paths to leverage the authors and venues similarity searching task. For the A-P-V-P-A meta-path, the author similarities are influenced the topic distributions of their submitted papers at specific venues. And for the V-P-A-P-V meta-path, the venue similarities are also depended on the content of their published papers which are written by a set of common authors.

3.3. Distributed W-MetaPath2Vec under Apache Spark GraphFrames

Apache Spark GraphFrames distributed graph computing model. Among distributed graph data processing model, GraphFrames is most easy framework for implementation via Python programming language. GraphFrames logically uses two separate DataFrame to represent the vertices and edges of the graph. Each DataFrame structure might also stores distinctive attributes of nodes and edges. Different from other graph distributed processing models of Apache Spark, GraphFrames supports motif finding mechanism. Motif finding support to do complex path searching between two nodes which involve with multiple relations and nodes inside. This feature is suitable for solving the problem of finding all path instances between two nodes following defined meta-paths. For example, in DBLP with the meta-path: A-P-V-P-A, we can find all the path instances between two authors, like as: "*Jiawei Han*" and "*Philip S. Yu*", with following defined motif finding pattern:

```
APVPA_motifs = DBLP_GraphFrames.find("(a1)
-[e1]->(p1); (p1)-[e2]->(v); (a2)-[e3]->(p2);
(p2)-[e4]->(v)")
    · filter(a1.full_name = ''Jiawei Han'')\
    · filter(a2.full_name = ''Philip S. Yu'')\
    · filter(v.node_type = ''venue'')\
    · filter(e1.relation_type = ''author_of'')\
    · filter(e2.relation_type = ''submit_at'')\
    · filter(e3.relation_type = ''author_of'')\
    · filter(e4.relation_type = ''submit_at'');
```

The GraphFrames model is implemented under the Apache Spark environment, therefore it can be capable for handling huge number of nodes and relations. The motif finding in GraphFrames can leverage the process of meta-path-based graph traversal between nodes which support to find all path instances between them. We re-implement the topic-driven meta-path-based random walk of W-Metapath2Vec model as following (as described in Algorithm 2):

4. Experiment and discussions

In this section, we conduct thorough empirical studies in order to demonstrate the effectiveness of W-MethPath2Vec model. The section is divided into two main parts, include:

- In the first part, we evaluate the accuracy of W-MetaPath2Vec model with previous network embedding models by solving

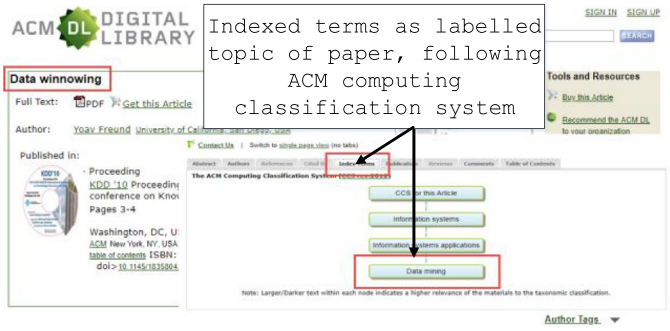

Fig. 11. Example of an ACM papers which is indexed by topics following ACM CCS-2012.

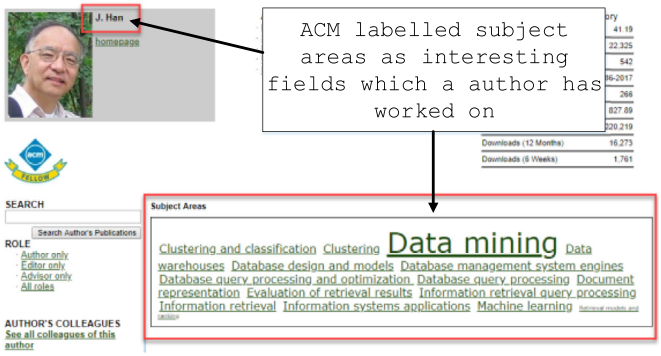

Fig. 12. Example of an ACM indexed author which is indexed by subject areas.

network analysis tasks include: node similarity searching, clustering and classification.
- In the second part, we perform the experiment on the scalability of W-MetaPath2Vec with Metapath2Vec model in the context of large-scaled heterogeneous networks.

4.1. Experimental dataset usage

4.1.1. DBLP, AMiner & ACM datataset (DAC-Dataset)

About the experimental dataset, we used the real-world dataset of DBLP bibliographic network (DBLP - Computer Science Bibliography, n.d.) in combination with AMiner dataset (AMiner Dataset, n.d.) to test the performance our proposed W-MethPath2Vec model as well as comparing with the other approaches. About the experiment dataset usage, includes:

- Full DBLP bibliographic network[1]: contains *2.060.445* authors, *4.101.409* papers, *5.377* venues/conferences and *1.570* journals.
- AMiner abstract content dataset[2]: contains over *1.572.277* abstracts of DBLP indexed papers.

ACM topic labelled papers and authors. From full dataset of DBLP bibliographic network, we carefully selected 75 K papers which are indexed in three main topics/areas of computer science domain, which are: "*database/data mining*", "*artificial intelligence*" and "*networks*" (25 K papers for each topic). This set of 75 K papers associates with over 58,726 authors, and 2761 venues/journals. The topics of these papers are indexed by ACM digital library (ACM Digital Library, n.d.), as a trusted third-party resource. This set of computing domain topics are defined in the ACM Computing Classification System (CCS-2012)[3] (as illustrated in Fig. 11). From this set of 75 K papers, we apply the LDA topic model of BLEI (2012) to extract 10 topics (k = 10), with 20 keywords for each topic. (Figs. 12–19)

Authors and venues/journals labels. Similar to paper nodes, all the authors which are indexed in ACM digital library also contain set of labelled subject areas, following the CCS-2012 (Fig. 12). For a specific author, the labelled subject areas indicates interesting fields which this author has worked on. We only selected authors who works on three mentioned topics of this dataset. For the 7 K venues/journals nodes, we use the labels of Google Scholar Metric[4] (Fig. 13) as the trusted resource for labeling them. Like as the authors, we also select venues/journals belong to three topics, which are: "*database/data mining*", "*artificial intelligence*" and "*networks*".

4.1.2. DBIS dataset by Sun et al. usage (SUN-DBIS-Dataset)

This dataset is used by Sun et al. in previous works (Sun, et al., 2011; Sun, et al, 2009a, 2009b). The Sun et al. DBIS dataset[5] contains 72,902 papers and 60,694 authors which are associated with 464 venues/journals. From this set of 72,902 papers, we also collected their abstract content from Aminer dataset. Then, similar to previous DAC-Dataset, these abstract contents are applied LDA to extract 10 topics and 20 keywords for each topic.

4.1.3. 8-Area labelled dataset by Dong et al. (DONG-8AREA-Dataset)

This dataset is used by Dong et al. for testing the accuracy of Metapath2Vec model (Dong et al., 2017). This dataset[6] covers 8 topics computing knowledge domains which is constructed by Dong et al. based on the Google Scholar Metrics 2016. This dataset contains 133 venues and 246,678 authors which have been assigned into 8 computing domain areas, includes: "Computing Systems", "Theoretical Computer Science", "Computer Networks & Wireless Communication", "Computer Graphics", "Human Computer Interaction", "Computational Linguistics", "Computer Vision & Pattern Recognition", "Databases & Information Systems".

From Aminer dataset, we get the set of papers which is associated with given 246,678 authors and 133 venues in Dong et al. dataset. We have collected over 190 K papers based on Dong et al. dataset. Then, these 190 K papers are applied LDA to extract 10 topics and 20 keywords for each topic like previous two datasets.

4.2. Experimental setup and evaluation metric usage

For all approaches which are applied for heterogeneous networks, include: MethPath2Vec and W-Metapath2Vec models. We use four common meta-paths for evaluating the similarity between two types of nodes (author and venue/journal), which are (as described in Table 3):

Evaluation metrics for experimental results. For evaluating the output accuracy of models in resolving similarity searching task, we used three metrics:

- **Node similarity search evaluation metrics usage.** For evaluating the accuracy of node similarity task, we use nDCG (normalized Discounted Cumulative Gain) metric (Järvelin & Kekäläinen, 2002). The nDCG metric supports to evaluate the accuracy of output results within range of [0, 1] (the higher the better). This metric is used by Sun et al. for experiments with DBLP dataset in Sun et al. (2011). For each return node in the similarity search outputs, we score it based on the level of relevance. There are four levels of similarity score between two nodes which are shown in Table 4.

[1] DBLP dataset: http://dblp.uni-trier.de/.
[2] AMiner dataset: https://aminer.org/.
[3] ACM Computing Classification System: https://www.acm.org/publications/class-2012.
[4] Google Scholar Metric: https://scholar.google.com/citations?view_op=top_venues.
[5] Sun et al. DBIS dataset: http://web.cs.ucla.edu/~yzsun/.
[6] Dong et al. 8-area dataset: https://ericdongyx.github.io/metapath2vec/m2v.html.

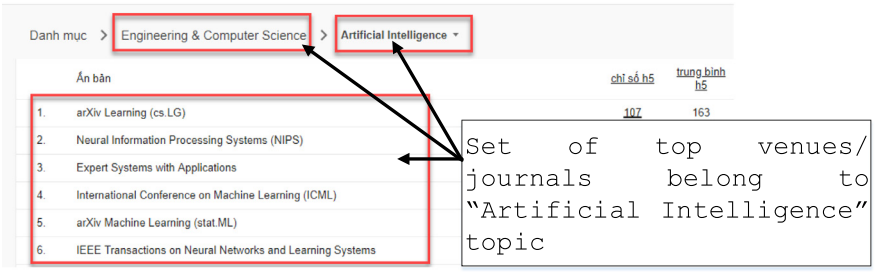

Fig. 13. Example of set of top venues/journals which belong to "Artificial Intelligence" topic which are indexed by Google scholar metric.

Table 3
Four common meta-paths in DBLP network.

Author similarity searching	These two following meta-paths are used for obtaining the set of top-k relevant authors: • **"A-P-V-P-A"** (author-paper-venue-paper-author): this meta-path indicates the semantic relations of two authors who often submit their works at the same venues/journals. • **"A-P-T-P-A"** (author-paper-topic-paper-author): this meta-path indicates the semantic relations of two authors who work on same subjects via their written papers.
Venue/ journal similarity searching	These two following meta-paths are used for obtaining set of top-k relevant venues/journals. • **"V-P-A-P-V"** (venue-paper-author-paper-venue): this meta-path presents for the semantic relations of two venues/journals which share the same set of authors, who often submit their papers to them. • **"V-P-T-P-V"** (venue-paper-topic-paper-venue): this meta-path is used to demonstrate the semantic relations of two venues/journals that have same set of covered topics via their published papers.

Table 4
The scores for level of node relevancy with the given node in query.

Score	Description
0	Non-relevant
1	Quite relevant
2	Closely relevant
3	Very/highly relevant

- **Node clustering & classification evaluation metrics usage**. For the output results of clustering task, we used the Purity, F-measure (average micro/macro F1) and NMI (Normalized Mutual Information) (Strehl & Ghosh, 2002) metrics. These metrics are used for evaluating the outputs of node clustering and classification tasks.

4.3. Experimental result and discussions

4.3.1. Model accuracy evaluation

In this part, we perform experiments with several approaches of information network embedding technique, include: DeepWalk, LINE, Node2Vec, MetaPath2Vec and W-MethPath2Vec on the same dataset. For each meta-path, we conducted the similarity searching by each technique, the output results contain three main type: top five (top-5), ten (top-10) and twenty (top-20) similar nodes for the specific node in the query. Each network embedding technique supports to produce training models contain set of feature vectors which are represented for each nodes of the given networks.

For both experiments on node similarity searching and clustering, we use the cosine similarity to determine the distance between nodes. Tables 5 and 6 show examples of top-10 relevant authors and venues/conferences searching with two meta-path A-P-V-PA and V-P-A-P-V, respectively.

4.3.1.1. Node similarity searching task. For the meta-path "A-P-V-P-A" (only used for model training while applying MetaPath2Vec and W-MethPath2Vec) and which is aimed for finding similar authors through their submitted papers in specific venues/journals, we select 10 most-active authors (has the highest number of published papers) as the source authors for finding similar authors within 58 K authors of experimental dataset. For 10 authors' output accuracy results which are evaluated by nDCG metric, we obtained the average results for each approach as the final output results. The final experiment results for this test-case are shown in Fig. 14 and Tables 7–9.

For the second meta-path "V-P-A-P-V" (only used for model training while applying MetaPath2Vec W-MethPath2Vec) which is used to find similar venues/journals which share the same set of common authors (via their paper's submissions), we also selected 10 most-active venues/conferences which have the highest number of published papers as the source venues. Similar to the first experiment, this averaged results are obtained as the final outputs for evaluating the performance of each embedding approach. The final experiment results for this test-case are shown in Fig. 15 and Tables 10–12.

4.3.1.2. Node clustering task. In this experiment, we test the performance of W-metapath2Vec in comparing with other approaches of network embedding techniques on solving network node clustering task. The cosine similarity and k-means clustering algorithm are applied to group the set of nodes into appropriate clusters. In this test, we only use DAC-Dataset and DONG-8AREA-Dataset, because only these two datasets have labels for author and venue/journal nodes. The label for each author and venue/journal is considered as the correct cluster which that node should belong to. Therefore, with k-means clustering algorithm, we assign number of clusters for DAC-Dataset is 3 ($k = 3$) and DONG-8AREA-Dataset is 8 ($k = 8$).

For evaluating the accuracy of the outputs, we use three metrics, include: Purity, F-measure and NMI. Tables 13 and 15 show the accuracy scores of authors clustering. Tables 14 and 16 show the accuracy scores of venues/journals clustering.

Table 5
Examples of top-10 relevant authors searching by W-MetaPath2Vec model with meta-path A-P-V-P-A on DAC-Dataset.

Philip S. Yu	Jiawei Han	Christos Faloutsos	Hari Balakrishnan
Jiawei Han	Philip S. Yu	Jian Pei	Yin Zhang
Ming-Syan Chen	Jian Pei	Ke Wang	Yunhao Liu
Jian Pei	Ming-Syan Chen	Hui Xiong	Mostafa H. Ammar
Elisa Bertino	Christos Faloutsos	Vipin Kumar	Scott Shenker
Hans-Peter Kriegel	Hans-Peter Kriegel	Eamonn J. Keogh	Balachander Krishnamurthy
Christos Faloutsos	Elisa Bertino	Jiawei Han	Daniela Rus
Ke Wang	Divesh Srivastava	Jianyong Wang	Bhaskaran Raman
Zheng Chen	Ke Wang	Huan Liu	B. R. Badrinath
Haixun Wang	H. V. Jagadish	Heikki Mannila	Boon Thau Loo
Jeffrey Xu Yu	Nick Koudas	Jieping Ye	Ion Stoica

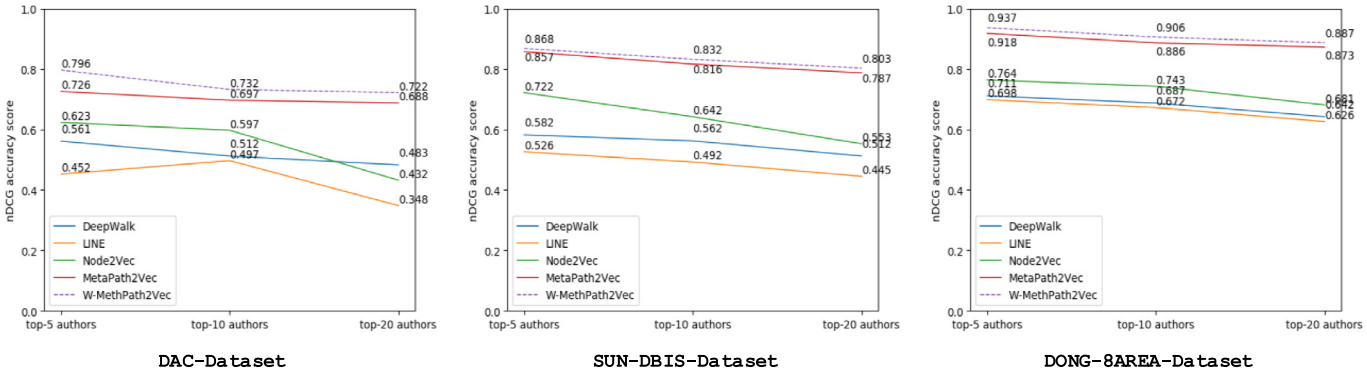

Fig. 14. Top-k similarity search accuracy for authors with different datasets.

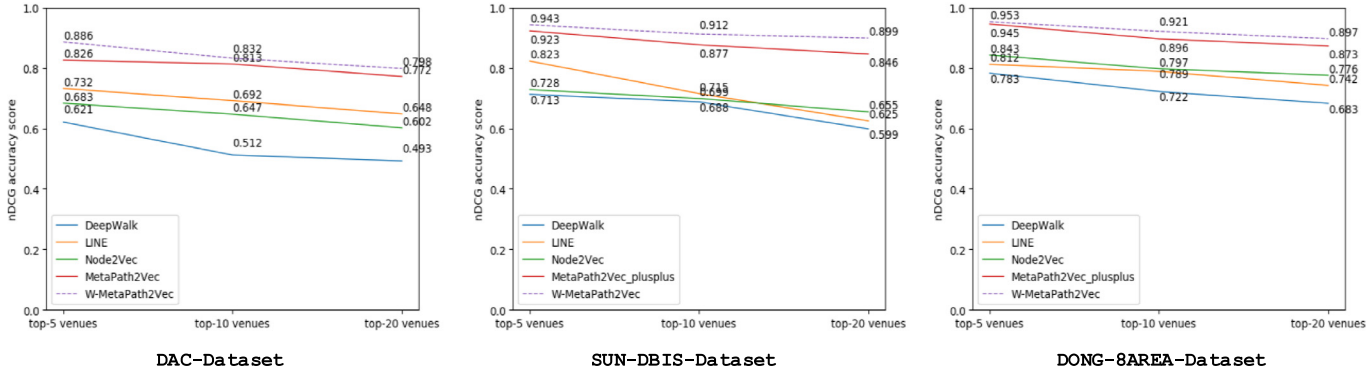

Fig. 15. Top-k similarity search accuracy for venues with different datasets.

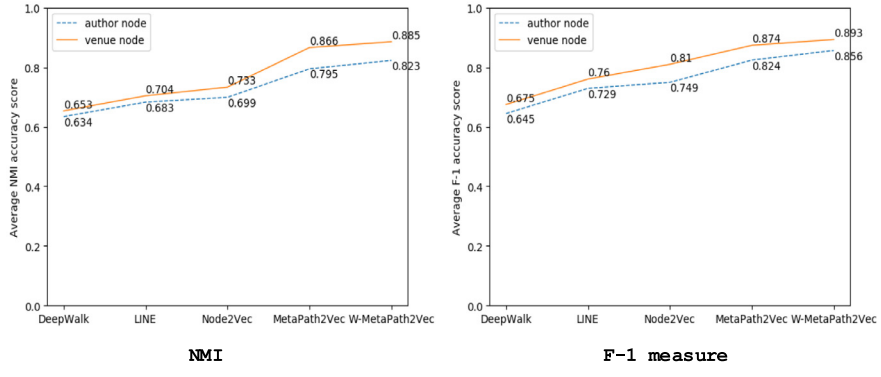

Fig. 16. Average F-measure and NMI accuracy scores for authors and venues clustering over 3 datasets with different network embedding approaches.

Table 6
Examples of top-10 relevant venues/conferences searching by W-MetaPath2Vec model with the meta-path V-P-A-P-V on DAC-Dataset.

KDD	PKDD	CVPR	ACL
ICDM	IJCAI	ICCV	NAACL
TKDE	SIGKDD	IJCV	COLING
CIKM	AAAI	SIGGRAPH	IJCAI
SIGMOD	ML	ICPR	EMNLP
WSDM	ICML	PAMI	SIGIR
VLDB	WSDM	ECCV	CIKM
WWW	PAKDD	TOG	AAAI
ICDE	DATAMINE	CVIU	WWW
ICML	KAIS	WACV	ICML
SIGKDD	CIKM	IJCAI	ICUIMC

Table 7
Top-k similarity search accuracy for authors on DAC-Dataset.

	nDCG Accuracy		
	Top-5	Top-10	Top-20
DeepWalk	0.56124	0.51213	0.48267
LINE	0.45223	0.39678	0.34812
Node2Vec	0.62341	0.59721	0.43218
MetaPath2Vec	0.72572	0.69672	0.68782
W-MethPath2Vec	**0.79628**	**0.73212**	**0.72213**

Table 8
Top-k similarity search accuracy for authors on SUN-DBIS-Dataset.

	nDCG Accuracy		
	Top-5	Top-10	Top-20
DeepWalk	0.58216	0.56217	0.51232
LINE	0.52627	0.49221	0.44521
Node2Vec	0.72162	0.64212	0.55272
MetaPath2Vec	0.85721	0.816271	0.78712
W-MetaPath2Vec	**0.86782**	**0.83212**	**0.80276**

Table 9
Top-k similarity search accuracy for authors on DONG-8AREA-Dataset.

	nDCG Accuracy		
	Top-5	Top-10	Top-20
DeepWalk	0.71123	0.68721	0.64223
LINE	0.69827	0.67229	0.62627
Node2Vec	0.76371	0.74281	0.68123
MetaPath2Vec	0.91772	0.88627	0.87261
W-MetaPath2Vec	**0.93672**	**0.90562**	**0.88671**

The experimental results (Fig. 16) show that the W-Metapath2Vec outperform other network embedding approaches for both authors and venues clustering tasks. For authors clustering test, for both DAC-Dataset and DONG-8AREA-Dataset the outputs show that W-Metapath2Vec model gains better performance around 4% more than Metapath2Vec model and over 12% than DeepWalk, LINE and Node2Vec models. The outperformance of Metapath2Vec in comparing with DeepWalk, LINE and Node2Vec models is quite clear, because these models are only applied for homogeneous networks only.

4.3.1.3. Multiple-class node classification task. In this section, we perform the node classification task for author and venue/journal nodes. In this test, we only use two datasets which are DAC-Dataset and DONG-8AREA-Dataset. The number of class in DAC-Dataset is three and DONG-8AREA-Dataset is eight for both author and venue/journal nodes. Detail information about the classes

Table 10
Top-k similarity search accuracy for venues on DAC-Dataset.

	nDCG Accuracy		
	Top-5	Top-10	Top-20
DeepWalk	0.62125	0.51213	0.49267
LINE	0.73222	0.69178	0.64812
Node2Vec	0.68343	0.64721	0.60218
MetaPath2Vec	0.82571	0.81323	0.77182
W-MethPath2Vec	**0.88624**	**0.83212**	**0.79823**

Table 11
Top-k similarity search accuracy for venues on SUN-DBIS-Dataset.

	nDCG Accuracy		
	Top-5	Top-10	Top-20
DeepWalk	0.71273	0.68778	0.59872
LINE	0.82271	0.71521	0.62512
Node2Vec	0.72826	0.69887	0.65479
MetaPath2Vec	0.92276	0.87652	0.84627
W-MethPath2Vec	**0.94272**	**0.91223**	**0.89862**

Table 12
Top-k similarity search accuracy for venues on DONG-8AREA-Dataset.

	nDCG Accuracy		
	Top-5	Top-10	Top-20
DeepWalk	0.78271	0.72212	0.68271
LINE	0.81221	0.78862	0.74221
Node2Vec	0.84271	0.797275	0.77582
MetaPath2Vec	0.94521	0.89627	0.87274
W-MethPath2Vec	**0.95286**	**0.92133**	**0.89675**

Table 13
Purity, F-measure and NMI accuracy scores for authors clustering task in DAC-Dataset ($k = 3$).

	Purity	F-measure	NMI
DeepWalk	0.662824	0.53291	0.58276
LINE	0.782472	0.69452	0.65218
Node2Vec	0.760217	0.69872	0.64128
MetaPath2Vec	0.844252	0.74892	0.70921
W-MethPath2Vec	**0.868736**	**0.75819**	**0.72416**

Table 14
Purity, F-measure and NMI accuracy scores for venues clustering task in DAC-Dataset ($k = 3$).

	Purity	F-measure	NMI
DeepWalk	0.712832	0.62872	0.63128
LINE	0.782472	0.76972	0.68213
Node2Vec	0.760217	0.79872	0.69321
MetaPath2Vec	0.916321	0.81232	0.83819
W-MethPath2Vec	**0.917368**	**0.84282**	**0.85416**

Table 15
Purity, F-measure and NMI accuracy scores for authors clustering task in DONG-8AREA-Dataset ($k = 8$).

	Purity	F-measure	NMI
DeepWalk	0.76526	0.72812	0.69281
LINE	0.79278	0.75628	0.71281
Node2Vec	0.84728	0.79281	0.75721
MetaPath2Vec	0.92682	0.89673	0.87827
W-MethPath2Vec	**0.93272**	**0.92378**	**0.91227**

Table 16
Purity, F-measure and NMI accuracy scores for venues clustering task in DONG-8AREA-Dataset ($k = 8$).

	Purity	F-measure	NMI
DeepWalk	0.76233	0.68276	0.64253
LINE	0.79682	0.75255	0.71254
Node2Vec	0.84378	0.81437	0.76425
MetaPath2Vec	0.93267	0.89672	0.88722
W-MethPath2Vec	**0.94468**	**0.91225**	**0.89443**

Table 17
Marco-average-F1 and Micro-average-F1 accuracy results for authors classification task in DAC-Dataset (number of class = 3).

	Marco-average-F1	Micro-average-F1
Metapath2Vec	0.92573	0.91756
W-Metapath2Vec	**0.93425**	**0.92886**

Table 18
Marco-average-F1 and Micro-average-F1 accuracy results for venues classification task in DAC-Dataset (number of class = 3).

	Marco-average-F1	Micro-average-F1
Metapath2Vec	0.96442	0.95882
W-Metapath2Vec	**0.97874**	**0.96478**

of each dataset have been mentioned in the Section 4.1.3. Both two dataset are splitted into two parts: testing and training sets. The training sets are used to feed the classification model. For all node classification experiments, we used the logistic regression alogrithm as the main classifier model. The trained classification models which are generated by logistic regression classifier are then used to predict the classes for the testing sets. The class prediction outputs are evaluated under the micro/macro F1 accuracy metrics in order to examine the accuracy of network embedding models.

Tables 17 and 18 show the average micro/macro F1 accuracy scores of author and venue/journal nodes classification in DAC-Dataset by applying 5-fold cross validation. For this dataset, the portion between training and testing sets is 70% and 30%, respectively. As shown in the experimental outputs, the W-Metapath2Vec model slightly outperforms the previous Metapath2Vec model for both author and venue classification tasks. Especially in author classification task, it gains better performance about 1.07% than Metpath2Vec model. For venue/journal node classification, the output accuracy scores for both two models are nearly similar, because the amount of venues/journals in this dataset is quite small, only 2761 venues/journals.

For the node classification experiments with DONG-8AREA-Dataset, we conduct the tests as the same testing scenario of Dong et al. (2017). We vary the size of the training set from 10% to 90% and the remaining nodes are used for testing. Tables 19 and 20 show the average micro/macro F1 accuracy scores for author and venue node classification task in DONG-8AREA-Dataset.

For author node classification task, the outputs of both Metapath2Vec and W-Metapath2Vec models are quite similar. For test cases with different size of input training set, the W-Metapath2Vec slightly outperforms Metapath2Vec in terms of both metrics. In venue node classification testing, the W-Metapath2Vec model significantly improve the accuracy especially with small training set cases (in 10% and 30%). In summary, our proposed W-Metapath2Vec model is proved to be better in solving heterogeneous network node embedding than previous Metapath2vec model (Fig. 17). In fact, due to the large number amount of author nodes, the predicting for authors' classes is relatively stable than the venues' classes when varying the train-test splits.

We also conducted several experiments for comparing our proposed W-Metapath2Vec model with previous homogeneous approaches, include: DeepWalk, LINE and Node2Vec. Tables 21 and 22 show the output accuracy in micro/macro F-1 terms for author and venue nodes classification task, respectively. The experimental results demonstrate better performance than DeepWalk, LINE (about 7%) and Node2Vec (about 19%) in author node classification.

4.3.2. Model scalability and time-consuming performance evaluation

In this section, we demonstrate the experiments on the scalability of our proposed W-Metapath2Vec model comparing with previous Metapath2Vec model. In this test, we use the same dataset with Dong et al. to test the performance of both approaches (DONG-8AREA-Dataset). The standalone W-Metpath2Vec and Metpath2Vec model are implemented in a single computer with Intel Xeon E3-1220 V5, 3.0 GHz, 8 MB Cache, (8 Cores) CPU with 32 Gb in memory. We vary the size of DONG-8AREA-Dataset from 20% to 100% and collect the execution time for completing the random walk process of each model. Both models' random walk mechanisms are setup with walk length is set to 100 ($l = 100$). Table 23 shows the execution times of Metapath2Vec and W-Metapath2Vec models in different size of the DONG-8AREA-Dataset (%).The experiments show that our proposed W-Metapath2Vec significantly outperforms the Metapath2Vec model in time-consuming performance for all dataset sizes. Especially, with the full DONG-8AREA-Dataset, the W-Metapath2Vec is faster than Metapath2Vec about 1.31 times (Fig. 18).

In the second part, we test the performance of W-Metapath2Vec model within the context of very large-scaled network. We use the full set of DBLP network with over 6 M nodes and more than 10 M relations. We setup the Apache Spark cluster with 5 nodes. Each node is a server with Intel Xeon E3-1220 V5, 3.0 GHz, 8 MB Cache, (8 Cores) CPU. For the full set DBLP network, with walk length is set to 100 ($l = 100$), we need approximate 445.47 min (nearly 7.5 h) to complete the process of random walk in standalone-based environment. We tested the distributed W-Metapath2Vec model which is implemented under Apache Spark GraphFrames with the standalone-based to demonstrate the scalability of our proposed model. The W-Metapath2Vec model are implemented under both standalone-based and distribution-based environment. We compare the execution time of both approach in different size of the DBLP dataset (Fig. 19). Table 24 shows the execution times of two approaches in different size of the DBLP dataset (%).

By using distributed W-Metapath2Vec model, the random walk processes for overall DBLP network costs only about 67.095 min. The distributed W-Metapath2Vec model supports to speed up the random walk processes nearly 6.6 times in comparing with the standalone-based model. Overall, the experimental results demonstrate that the proposed W-Metapath2Vec which is implemented under Apache Spark GraphFrames distributed environment can gain better performance for large-scale heterogeneous information networks with millions of nodes and relations.

5. Conclusion and future works

In this paper, we formally present studies related to problems of heterogeneous information network representation learning. There are remained challenges which are related to thorough evaluations on topic of text-based nodes in content-based HIN. Moreover, we are in the era of big data, it is necessary to develop network analysis model which is capable for handling large-scaled networks. To address these challenges, our works in this papers are focused on developing the W-metapath2Vec model. The proposed W-metapath2Vec model uses the topic-driven meta-path-based random walk mechanism. This novel random walk mech-

Table 19
Average micro/macro F1 accuracy results for venue nodes classification in DONG-8AREA-Dataset.

Training Set (%)		10%	30%	50%	70%	90%
Metapath2Vec	Micro-average-F1	0.62746	0.89261	0.93427	0.96241	0.97271
	Marco-average-F1	0.54351	0.88462	0.93678	0.94275	0.95203
W-Metapath2Vec	Micro-average-F1	0.64271	0.91267	0.94782	0.97821	0.98425
	Marco-average-F1	0.61257	0.90672	0.94017	0.95221	0.96214

Table 20
Average micro/macro F1 accuracy results for author nodes classification in DONG-8AREA-Dataset.

Training Set (%)		10%	30%	50%	70%	90%
Metapath2Vec	Micro-average-F1	0.92431	0.93456	0.93654	0.93851	0.93682
	Marco-average-F1	0.91262	0.92216	0.93147	0.93263	0.93244
W-Metapath2Vec	Micro-average-F1	0.93812	0.94281	0.94782	0.95213	0.95872
	Marco-average-F1	0.92728	0.93236	0.93877	0.94162	0.95062

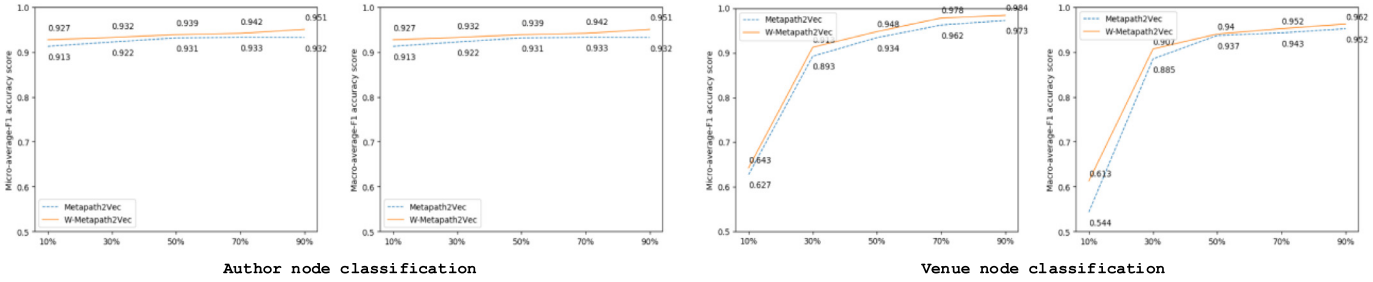

Fig. 17. Multiple-class node classification results in DONG-8AREA-Dataset.

Table 21
Comparative studies on different homogeneous network embedding approaches with proposed W-Metapath2Vec model for author nodes classification task.

	Marco-average-F1	Micro-average-F1
DeepWalk	0.672812	0.692823
Node2Vec	0.717281	0.743261
LINE	0.868212	0.882129
W-Metapath2Vec	0.938134	0.947926

Table 22
Comparative studies on different homogeneous network embedding approaches with proposed W-Metapath2Vec model for venue nodes classification task.

	Marco-average-F1	Micro-average-F1
DeepWalk	0.592812	0.612821
Node2Vec	0.561823	0.582712
LINE	0.827182	0.847281
W-Metapath2Vec	0.874762	0.893132

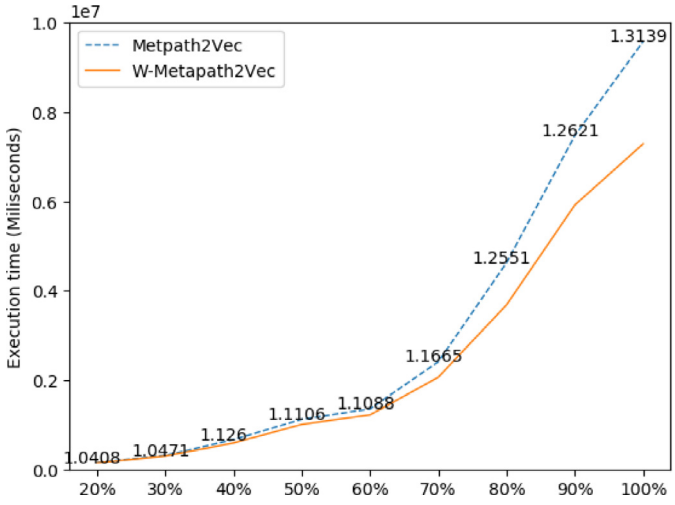

Fig. 18. Speed up rates in execution time between Metapath2Vec and W-Metapath2Vec models.

Table 23
Execution time of Metapath2Vec and W-Metapath2Vec models with different size of DONG-8AREA-Dataset.

Network Size (%)	Execution time (Milliseconds)		
	Metpath2Vec	W-Metapath2Vec	Speed-up rate (times)
20	168,204	161,614	1.0408
30	326,781	312,068	1.0471
40	682,031	605,703	1.1260
50	1131,781	1019,053	1.1106
60	1365,531	1231,532	1.1088
70	2418,203	2072,956	1.1665
80	4632,224	3690,703	1.2551
90	7469,574	5918,233	1.2621
100	9567,956	7282,030	1.3139

Table 24
Execution time of standalone-based and distribution-based W-Metapath2Vec model with different size of full DBLP dataset.

Network Size (%)	Execution time (Milliseconds)		
	Distributed W-Metapath2Vec	Standalone-based W-Metapath2Vec	Speed-up rate (times)
20	626,728	522,456	0.834
30	1072,816	928,271	0.865
40	1427,122	2422,812	1.698
50	2082,213	3276,212	1.573
60	2895,221	4526,129	1.563
70	3322,123	8291,823	2.496
80	3622,122	12,762,812	3.524
90	3822,421	19,272,931	5.042
100	4025,678	26,728,121	6.639

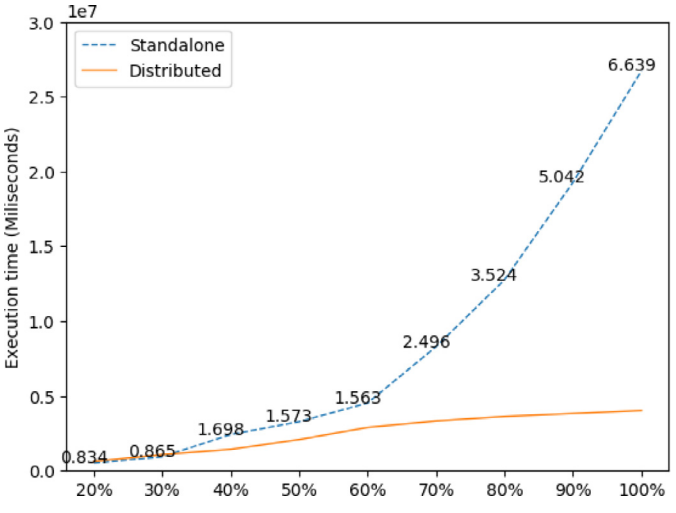

Fig. 19. Speed up rates in execution time between standalone-based and distribution-based W-Metapath2Vec.

anism allow to capture not only the semantic correlations but also topic similarity of nodes. Moreover, we also propose the approach of using distributed graph distributed processing mechanism of Apache Spark GraphFrames. The proposed distributed W-metapath2Vec is capable for handing massive networks with millions of nodes and relations. Through experiments, the W-MethPath2Vec model is proved to be able for improving heterogeneous network mining tasks as well as time-consuming performance.

Credit authorship contribution statement

Phu Pham: Conceptualization, Methodology, Software, Writing - review & editing, Data curation, Visualization. **Phuc Do:** Conceptualization, Methodology, Validation, Writing - original draft, Supervision, Funding acquisition.

Acknowledgement

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCMC) under the grant number B2017-26-02.

References

ACM Digital Library. Retrieved 03 16, 2018, from https://dl.acm.org/.

AMiner Dataset. Retrieved September 24, 2017, from https://aminer.org/data.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Chen, H., Yin, H., Wang, W., Wang, H., Nguyen, Q. V. H., & Li, X. (2018). PME: Projected metric embedding on heterogeneous networks for link prediction. In *Proceedings of the twenty-fourth ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1177–1186). ACM.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances In Neural Information Processing Systems*, 2172–2180.

DBLP - Computer Science Bibliography. Retrieved September 24, 2017, from http://dblp.uni-trier.de/.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Dong, Y., Chawla, N. V, & Swami, A. (2017). metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the twenty-third ACM SIGKDD international conference on knowledge discovery and data mining* ACM (pp. 135–144).

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the twenty-second ACM SIGKDD international conference on knowledge discovery and data mining* ACM (pp. 855–864).

Gupta, M., Kumar, P., & Bhasker, B. (2017). HeteClass: A Meta-path based framework for transductive classification of objects in heterogeneous information networks. *Expert Systems with Applications, 68*, 106–122.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 289–296). Morgan Kaufmann Publishers Inc..

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS), 20*(4), 422–446.

Jeh, G., & Widom, J. (2002). SimRank: A measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 538–543). ACM.

Jeh, G., & Widom, J. (2003). Scaling personalized web search. In *Proceedings of the twelfth international conference on world wide web* (pp. 271–279). ACM.

Michelson, M., & Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: A first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data* (pp. 73–80). ACM.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Mikolov, T., Chen, K., Corrado, G., amp; Dean, J. (2013b). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Pham, P., Do, P., & Ta, C. D. (2018). W-PathSim: Novel Approach of weighted similarity measure in content-based heterogeneous information networks by applying LDA topic modeling. In *Proceedings of the Asian conference on intelligent information and database systems* (pp. 539–549). Cham: Springer.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., & Raiko, T. (2015). Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 3546–3554.

Shi, C., Kong, X., Huang, Y., Philip, S. Y., & Wu, B. (2014). Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering, 26*(10), 2479–2492.

Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. (2017). A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering, 29*(1), 17–37.

Shi, C., Hu, B., Zhao, X., & Yu, P. (2018). Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering, 31*(2) https://ieeexplore.ieee.org/abstract/document/8355676.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research, 3*(Dec), 583–617.

Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., & Wu, T. (2009a). Rankclus: Integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the twelfth international conference on extending database technology: Advances in database technology* (pp. 565–576). ACM.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proceedings of the VLDB endowment* (pp. 992–1003).

Sun, Y., & Han, J. (2012). Mining heterogeneous information networks: Principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery, 3*(2), 1–159.

Sun, Y., & Han, J. (2013). Mining heterogeneous information networks: A structural analysis approach. In *Proceedings of the ACM SIGKDD explorations newsletter* (pp. 20–28).

Sun, Y., Yu, Y., & Han, J. (2009b). Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the fifteenth ACM SIGKDD international conference on knowledge discovery and data mining* ACM (pp. 797–806).

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). LINE: Large-scale information network embedding. In *Proceedings of the twenty-fourth international conference on world wide web. International world wide web conferences steering committee* (pp. 1067–1077).

Tang, J., Qu, M., & Mei, Q. (2015). PTE: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the twenty-first ACM SIGKDD international conference on knowledge discovery and data mining* ACM (pp. 165–1174).

Vahedian, Fatemeh, Burke, Robin&, & Mobasher, Bamshad (2017). Weighted random walk sampling for multi-relational recommendation. In *Proceedings of the twenty-fifth conference on user modeling, adaptation and personalization* ACM (pp. 230–237).

Xu, Z., Ru, L., Xiang, L., & Yang, Q. (2011). Discovering user interest on twitter with a modified author-topic model. In *Proceedings of the IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology-volume 01* IEEE Computer Society (pp. 422–429).

Zhang, M, Hu, H., He, Z., & Wang, W. (2015). Top-k similarity search in heterogeneous information networks with x-star network schema. *Expert Systems with Applications, 42*(2), 699–712.