THESIS INFORMATION

Thesis Title	**Researches on heterogeneous information networks mining model and applications**
Major	Information Technology
Major Code	62-48-02-01
PhD Student	Phạm Thế Anh Phú
Supervisor	Assoc. Prof. Dr. Đỗ Phúc
Training Place	University of Information Technology, Vietnam National University - Ho Chi Minh City

1. ABSTRACT

The entire content of this thesis explicitly presents research issues which are related to the heterogeneous information network analysis and mining domain. Our studies in this thesis include contents regarding with literature reviews as well as proposed improvements on the similarity search problem within the Heterogeneous Information Network (HIN). The core contributions of this thesis mainly focus on the enhancements of integrating between the content-based similarity and graph-based topological features, in the form of meta-paths, between nodes in a given HIN. This combination enables to improve the performance of similarity measurement problem in HIN. Besides that, in this thesis, we also proposed enhancing directions for the link prediction problem within the context of content-enriched heterogeneous information network. Our achievements in this thesis are presented in 4 papers which are published in prestigious SCIE indexed journals, including: Expert System with Applications (1), International Journal of Machine Learning and Cybernetics (1), Intelligent Data Analysis (2). In the overall, the stated research issues as well as enhancement proposals in our thesis are generally structured into 3 main problems, as the following:

- **Problem 01**: In the first problem, our research mainly concentrating on building a theoretical background and problem formulation of discovering the distribution of latent topics over the content-based heterogeneous information network (a.k.a. Content-based HIN, or: C-HIN). In fact, the combination between the content-based similarity in the topic with the structural topological features between network nodes enable to improve the similarity search performance within the C-HINs. To modelling the proposed assumptions and ideas in this problem, we proposed a novel meta-path-based similarity measure model, called as: **W-PathSim** algorithm. Our proposed W-PathSim model enables to measure the

similarity between nodes in C-HIN by jointly evaluating the content-based and meta-path-based relational structure factors.

- **Problem 2**: majorly inheriting from the previous achievements with the proposed W-PathSim model, in the second problem, we mainly focus on enhancing the performance of network representation learning problem within the context of C-HIN. In this problem, we formulate the traditional network embedding model as joint content-based and structure-based learning problem for C-HIN. To deal with the network representation learning in C-HIN, we proposed two novel models: **W-MetaPath2Vec** and **W-MetaGraph2Vec**.
- **Problem 3**: From the achievements which are obtained in the two previous problems (1 & 2), we apply the integrated content-based and structure-based similarity evaluation paradigms in handling the link prediction problem with C-HIN. The main content of the third problem is concentrated on the proposal of the **W-MMP2Vec** algorithm.

2. OUR ACHIEVEMENTS IN THIS THESIS

By formally formulate the existing challenges regarding with the problems in C-HIN analysis and mining as well as proposed corresponding enhancements for each challenge, we have achieved the notable results which are categorized into two aspects:

- **In terms of academic and scientific aspect**: Our proposals in combining content-based and structure-based evaluation in network node similarity measurement within the context of C-HIN provide a new direction for content-enrich heterogeneous network analysis and mining domain. Our proposed models in this thesis enable to improve the performance of similarity search on the information network as well as provide meaningful similarity search results in HIN in which the similar weight between nodes are under evaluated in both content-based and structure-based relevancies. Besides the novel proposal in content-enhanced similarity measure on the C-HIN, our works also contain the integration of content-based and structure-based evaluation upon the heterogeneous network representation problem. These content-enriched network embedding models are utilized to effectively support the node transformation in C-HIN into embedding vector which are later used for different network analysis and mining tasks like as similarity search, clustering or link prediction. Thus, our research and proposals in this thesis have played as a strong baseline for solving the similarity search problem in HIN with the integration of auxiliary information like as content and topics between network nodes.

- **In terms of practical application aspect**: for the real-world application aspect, our proposed models proposed in thesis will directly support to build realistic applications which are related to the scientific collaboration recommendation problem. Our proposed similarity search via heterogeneous network embedding approach explicitly help to identify relevant authors as well as predicting possible collaborative links between authors for conducting recommendation. Moreover, our proposed models are also generally applied for handling other networked data analysis and mining problems on different types of information networks such as social network and e-commerce platforms.

3. FUTURE DIRECTIONS

In general, our proposed approaches in this present thesis are mainly designed upon the single-layered neural network training with a single hidden layer. Therefore, they can’t reach the maximum accuracy performance in similarity search as well as link prediction problems. Therefore, among possible improvement directions of our thesis is replacing the single-layered training mechanism with other advanced multi-layer neural network architectures within the deep learning domain. The substitution of advanced deep neural architectures in our proposed models might enable to significantly increase the accuracy performance of our proposed algorithms in primitive tasks of heterogeneous network analysis and mining. In more specifics, the multi-layer Graph Convolutional Network (GCN) neural network architecture is one of most recent and potential improvement direction.

In addition, our proposed models in the thesis are mainly designed upon the graph-based random walk sampling mechanism to model the proximity information between network nodes which is considered as local relevant structure of a given network. Therefore, they might be unable to effectively preserve the global structure of the given heterogeneous information network. Among potential future improvements, focusing on preserving the joint local and global structure information of the given HIN is a promising direction for significantly improving the accuracy performance of proposed models in this thesis.

Supervisor

Assoc. Prof. Dr. ĐỖ PHÚC
Date: 05/10/2021

PhD Student

PHẠM THẾ ANH PHÚ
Date: 05/10/2021