Homepage Augmentation by Predicting Links in Heterogenous Networks

Jianming Lv
South China University of Technology
jmlv@scut.edu.cn

Jiajie Zhong
South China University of Technology
cszhongjiajie@mail.scut.edu.cn

Weihang Chen
South China University of Technology
csscut@mail.scut.edu.cn

Qinzhe Xiao
South China University of Technology
qinzhexiao@gmail.com

Zhenguo Yang
Guangdong University of Technology
yzgcityu@gmail.com

Qing Li
City University of Hong Kong
qing.li@cityu.edu.hk

ABSTRACT

Scholars' homepages are important places to show personal research interest and academic achievement through the Web. However, according to our observation, only a small portion of scholars update their publications and related events on their homepages in time. In this paper, we propose a homepage augmentation technique, which automatically shows the newest academic events related to a scholar on his/her homepage. Specifically, we model the relations between homepages and the events collected from the Web as a complex heterogenous network, and propose an Embedding-based Heterogenous random Walk algorithm, namely EHWalk, to predict the links between homepages and events. Compared with existing embedding-based link prediction algorithms, EHWalk supports more efficient modeling of complex heterogenous relations in a dynamically changing network, which helps link the massive new updated events to homepages precisely and efficiently. Comprehensive experiments on a real-world dataset are conducted and the results show that our algorithm can achieve both good effectiveness and efficiency for real-world deployment.

KEYWORDS

Homepage Augmentation, Heterogenous Networks, Embedding

ACM Reference Format:
Jianming Lv, Jiajie Zhong, Weihang Chen, Qinzhe Xiao, Zhenguo Yang, and Qing Li. 2018. Homepage Augmentation by Predicting Links in Heterogenous Networks. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269249

1 INTRODUCTION

When people want to know a scholar, they get used to browse his/her homepage at first. Thus, personal research homepages are very important for scholars to show their research interest and build cooperative relationship. However, according to our observation, only a small portion of scholars update their publications and related events on their homepages in time. In the 5,998 scholars' homepages collected from the computer science departments of 52 high-ranking Chinese universities, only 25% of the homepages contain publication list, and only 5% of them contain latest publications after 2016. This may cause a seriously adverse impact on understanding the up-to-date advance from each others. In this paper, we propose a homepage augmentation technique that automatically shows the new academic events related to a scholar on his/her homepage by inserting a short Javascript code snippet, as shown in Fig. 1. The events can be the newly published papers, academic activities or new courses, which are collected from DBLP, social networks and web search engines. Without loss of generality, we only represent the procedure to associate homepages with the paper records collected from DBLP here. The proposed model can be easily extended to integrate with other events from the Web.

Figure 1: Example of the homepage augmentation system.

The key challenge of deploying such system is the name ambiguity problem while associating a scholar with an event collected from the Web. Specifically, in the case of linking DBLP papers, because precise affiliations of authors are usually missing in the records of DBLP, it is not trivial to distinguish the papers written by different authors sharing a same name. Recently, some name disambiguation algorithms based on clustering mechanism [3, 7, 9] are proposed to partition the papers into groups, each of which is expected to contain the papers written by one single author. These algorithms focus on clustering of papers, rather than associating papers with a given homepage. Furthermore, most of the existing solutions are conducted based on simple direct relations among papers and authors, and neglect some deep semantic information such as the topics of papers and the content of authors' profiles.

To address those challenges above, we model the relations between homepages and DBLP papers as a dynamically evolving complex heterogenous network, which contains papers, homepages, journals/conferences and words. Furthermore, we propose an

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269249

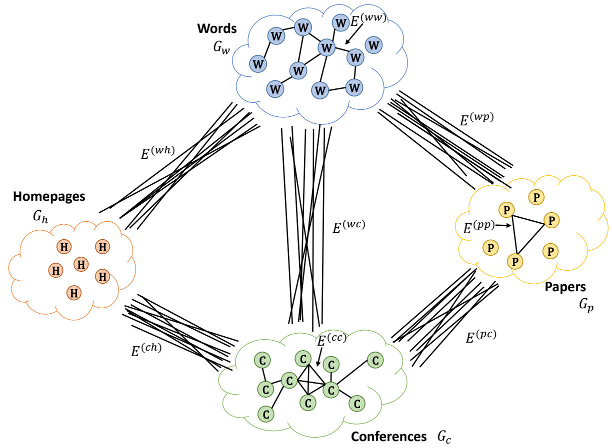

Figure 2: The heterogenous network G.

Embedding-based Heterogenous random Walk algorithm, namely EHWalk, to predict the links between homepages and papers. Compared with existing embedding-based link prediction algorithms[2, 5, 6, 8], which aim to model short relation paths in a static network, EHWalk supports more efficient modeling of complex heterogenous relations in a dynamically changing network. This helps link massive new updated events to homepages precisely and efficiently. Comprehensive experiments on real-world datasets show that our algorithm can achieve both high precision and efficiency.

2 HETEROGENOUS NETWORK MODELING

To leverage the rich semantic relations between the entities such as papers, journals/conferences, words and homepages, we model the relations as an undirected heterogenous networks (as shown in Fig. 2): $G =< V, E >$,where $V = P \cup C \cup W \cup H$, and $E = E^{(pp)} \cup E^{(cc)} \cup E^{(ww)} \cup E^{(wp)} \cup E^{(wc)} \cup E^{(wh)} \cup E^{(ch)} \cup E^{(pc)}$. Here P is the set of papers, C is the set of journals/conferences (for abbreviation, we use 'conferences' to represent journals or conferences in the following sections), W is the set of words, and H is the set of homepages. An edge in $E^{(pp)}$ indicates that two papers share an author, and the weight of the edge between two papers P_i and P_j is denoted as $E^{(pp)}_{i,j}$, which is set as:

$$E^{(pp)}_{i,j} = \begin{cases} \text{\#coauthors(i,j)} & \text{if } \#coauthors(i,j) > 1 \\ 0 & \text{if } \#coauthors(i,j) \leq 1 \end{cases}$$

, where $\#coauthors(i, j)$ means the number of coauthors between the two papers. $E^{(cc)}$ indicates the citation relations between conferences, and $E^{(cc)}_{i,j}$ is set as the number of cross-citations between two conferences. $E^{(ww)}$ indicates the co-occurrence relations between words, and $E^{(ww)}_{i,j}$ is set as the co-occurrence frequency of word W_i and W_j in those papers. $E^{(wp)}$ indicates the relations between words and papers, and $E^{(wp)}_{i,j}$ is set as the frequency of word W_i appearing in paper P_j. $E^{(wc)}$ indicates the relations between words and conferences, and $E^{(wc)}_{i,j}$ is set as the frequency of word W_i appearing in the papers from conference C_j. $E^{(wh)}$ indicates the relations between words and homepages, and $E^{(wh)}_{i,j}$ is set as the frequency of word W_i appearing in homepage H_j. $E^{(ch)}$ indicates the relations between conferences and homepages, and $E^{(ch)}_{i,j}$ is set as the frequency of conference C_i appearing in homepage H_j. $E^{(pc)}$ indicates the publication relations between papers and conferences, and $E^{(pc)}_{i,j}$ is set to 1 if paper P_i is published in conference C_j.

The network G contains totally 4 homogeneous sub-networks: the conference sub-network $G_c =< C, E^{(cc)} >$, the word sub-network $G_w =< W, E^{(ww)} >$, the paper sub-network $G_p =< P, E^{(pp)} >$ and the homepage sub-network $G_h =< H, \phi >$ (no internal links between homepages). Specifically, the paper sub-network G_p is a dynamically evolving network which keeps growing while new papers are published. The heterogenous network is a powerful tool to model not only the simple direct relation such as $E^{(pp)}$, but also the complex high-order relation such as 'paper $\rightarrow$ word 1 $\rightarrow$ word 2 $\rightarrow$ journal $\rightarrow$ homepage', which implies that a paper contains word 1, word 1 is related to word 2, word 2 is related to a topic in a journal, and the journal is appearing at a scholar's homepage.

Given the heterogenous network revealing academic relations, the key procedure to solve the homepage augmentation problem is to predict the links between homepage nodes and paper nodes. Different from traditional embedding-based link prediction algorithms[2, 6] in static heterogenous networks, the dynamically evolving property of network G prefers a more efficient way to link each new added paper node to a homepage. This motivates us to propose EHWalk, which can flexibly infer the likelihood that a new coming paper P_i is written by the scholar of homepage H_j. The detail of EHWalk will be introduced in the following sections.

3 EMBEDDING-BASED HETEROGENOUS RANDOM WALK

In this section, we propose the embedding-based heterogenous random walk solution, EHWalk, which contains the following three key steps. Firstly, we propose a novel heterogenous random walk algorithm to measure the relations between the nodes in the heterogenous network G (section 3.1). Secondly, we enhance G with hidden relations to construct an embedding heterogenous network to solve the sparsity problem (section 3.2). Finally, we apply the heterogenous random walk over the embedding network to determine the probability to link papers and homepages (section 3.3).

3.1 Heterogenous Random Walk

To capture the hidden relations between the nodes of heterogenous network G, we propose the heterogenous random walk algorithm to calculate the transition probability from each node to others. Specifically, starting from each node V_i, we initialize λ random walkers to visit other nodes through the edges for ℓ hops. In consideration of different types of edges and nodes in the heterogenous network, the probability $Prob_{i,j}$ of a walker at V_i to select each next hop V_j is determined according to both weight and type of edges as follows:

$$Prob_{i,j} = \frac{1}{N_i} \cdot E_{i,j} \cdot \Big(\sum_{V_k \odot V_j} E_{i,k} \Big)^{-1} \qquad (2)$$

Here N_i is the number of different types of V_i's neighbors. $V_k \odot V_j$ means that V_k belongs to the homogeneous sub-network (G_p,G_w,G_c or G_h) containing V_j. Eq. (2) means that the neighbors of V_i are firstly grouped according to their types, and the probability to walk into each group is equal to $\frac{1}{N_i}$. The transition probability is then further normalized in each group as $\frac{E_{i,j}}{\sum_{V_k \odot V_j} E_{i,k}}$. This type-specific normalization can capture the heterogenous edge information better than traditional unified normalization in random walk can do.

The nodes visited by each walker form a node sequence, in which every adjacent $t+1$ nodes compose a t-step transition sub-sequence $\{V_x, V_{x+1}, ..., V_{x+t}\}$, from which we can achieve the t-step transition matrix T, where $T_{i,j}$ indicates the number of sub-sequences start from V_i and contain V_j. After applying row normalization to T, we get the t-step transition probability matrix M, where $M_{i,j}$ indicates the probability of a walk from V_i reaches V_j within t steps, and reflects the strength of the relation between V_i and V_j.

3.2 Embedding Heterogenous Network

Given any paper node V_i, we can use the transition probability $M_{i,j}$ to measure the relation between V_i and any homepage node V_j. However, due to the sparsity of G, directly using the transition probability based on G may not reveal the latent complex dependencies between nodes. Particularly, for the long dependencies between homepages and papers across multiple heterogenous sub-networks (G_w and G_c), the sparsity problem can be further amplified.

In order to solve the sparsity problem, we propose an embedding-based topology augmentation technique to capture the hidden relations in G_w and G_c. Firstly, we assign each node V_i in G_w with an embedding vector $\vec{v}_i$ by maximizing the following objective:

$$\sum_{V_i, V_j \in W}[M_{i,j} log(\sigma(\vec{v}_i \cdot \vec{v}_j)) + \beta \mathbb{E}_{V_k \sim P_w(V)}(1-M_{i,k}) log(\sigma(-\vec{v}_i \cdot \vec{v}_k))] \quad (3)$$

Here $\sigma(.)$ denotes the sigmoid function. M is the t-step transition probability matrix obtained by the heterogenous random walk. By maximizing the first part of Eq. (3), the word nodes with higher transition probability are assigned with closer embedding vectors. The second part of Eq. (3) takes the negative sampling technique like [4] to measure the relation between V_i and another node V_k randomly sampled from G_w according to the degree distribution $P_w(V)$ (the probability to select V_k is proportional to its degree). By maximizing the second part, the embedding vectors of word nodes with lower transition probability can be further dispersed. β is a hyperparameter to tune the importance of each portion.

Compared with the traditional homogenous network based embedding algorithms[5, 8], our heterogenous random walk based embedding is more flexible to capture complex long dependency across multiple heterogenous sub-networks by using the t-step transition probability matrix M to constrain embedding vectors.

After achieving the embedding vectors in sub-network G_w, we can obtain an embedding network of the words as: $\widehat{G}_w =< W, \widehat{E}^{(ww)} >$, where the weight of edge $\widehat{E}^{(ww)}_{i,j}$ is set as $\sigma(\vec{v}^{(w)}_i \cdot \vec{v}^{(w)}_j)$. The weight of each edge reveals the hidden relations between the nodes in the embedding space. To reduce the cost of storage and computation, we erase the edges with small weights. Specifically, for each node V_i in $\widehat{G}_w$, we rank its adjacent edges according to their weights and reserve the top $k = \max\{D(V_i), \theta\}$ edges. Here $D(V_i)$ is the degree of node V_i in the original network G_w, and θ is a positive constant to define the minimum degree of $\widehat{G}_w$.

In a similar way, we can achieve the embedding network of conferences: $\widehat{G}_c =< C, \widehat{E}^{(cc)} >$ to embed the latent relations between conferences. Based on the embedding networks $\widehat{G}_w$ and $\widehat{G}_c$, we can achieve a new embedding heterogenous networks:

$$\widehat{G} =< V, \widehat{E} > \quad (4)$$

, where $E = E^{(pp)} \cup \widehat{E}^{(cc)} \cup \widehat{E}^{(ww)} \cup E^{(wp)} \cup E^{(wc)} \cup E^{(wh)} \cup E^{(ch)} \cup E^{(pc)}\}$. Beyond the direct relations revealed in the original heterogenous network G, $\widehat{G}$ further integrates with the hidden relations between words and those between conferences.

3.3 Random Walk in Embedding Network

After achieving the embedding network $\widehat{G}$, we can perform the heterogenous random walk on $\widehat{G}$ to get a new t-step transition probability matrix $\widehat{M}$. Given any paper node V_x and homepage node V_y, $\widehat{M}_{x,y}$ indicates the relation between V_x and V_y. Thus, we can link paper V_x to the homepage with the strongest relation to it.

Furthermore, we extend this link prediction algorithm to support efficient linking of new coming papers. Different from traditional embedding-based link prediction algorithms for static networks, which require rebuilding the embedding space for a new network, EHWalk supports very light-weighted incremental updating. For any new paper V_x, we initialize $\widehat{\lambda}$ random walkers starting from V_x in the embedding network $\widehat{G}$, and follow the transition probability defined in Eq. (2) to perform heterogenous random walk for t hops. For each walk, we can achieve a walking sequence $\{V_x, ...\}$, which embeds the complex relations across heterogenous networks. By counting the frequency of each homepage V_y appearing in the walking sequences, the t-step transition probability from V_x to V_y can be easily calculated as:

$$P_t(V_x \to V_y) = \frac{\mathbb{N}_x(V_y)}{\sum_{V_i \in H} \mathbb{N}_x(V_i)} \quad (5)$$

Here $\mathbb{N}_x(V_i)$ indicates the frequency of V_i appearing in the walking sequences started from V_x. The higher $P_t(V_x \to V_y)$ indicates the closer relation between V_x and V_y.

4 EXPERIMENTS

4.1 Experimental Settings

The dataset we use is the 'DBLP-Citation-Network-V10' corpus from AMiner[1], which contains 3,079,007 papers and is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. To evaluate the precision of link prediction between homepages and papers, we also crawl 58 homepages from the computer science departments of 29 Chinese universities. We use the scholar's name of each homepage to search the dataset for candidate papers containing the name, and then rank the papers according to the link probability defined in Eq. (5). We also construct the ground-truth dataset by verifying the authorship manually and filtering out the wrong mapping in the search results. In the following experiments, the number of random walkers λ is set as 10, the length of walking sequence ℓ is 80 and the length of sub-sequence t is 10. β of Eq. (3) is 0.002 while computing the embedding vectors in G_c, and set as 0.0001 in G_w. θ for generating embedding networks is set as 100.

We compare EHWalk with state-of-art embedding-based link prediction algorithms, including DeepWalk[5], LINE[8], ESim[6] and HIN2Vec[2]. Both DeepWalk[5] and LINE[8] are based on homogeneous networks, so they are deployed here neglecting the different types of nodes and edges. ESim[6] and HIN2Vec[2] are recently proposed embedding methods for heterogenous networks, which are based on the predefined meta-paths of a network. We also compare EHWalk with several widely used academic search engines such as Google Scholar, Bing and Aminer[1], which maintain auto-generated scholars' homepages with publication lists.

Table 1: The AUC scores of different algorithms.

Method	Average AUC Score
DeepWalk[5]	0.8673
LINE (unweighted)[8]	0.7768
LINE (weighted)[8]	0.7853
ESim [6]	0.7153
HIN2Vec[2]	0.7239
EHWalk (original network)	0.8780
EHWalk (embedding network)	**0.9098**

Table 2: The AUC scores of different algorithms on the homepages associating with different number of papers.

#Papers	1∼4	5∼10	11∼20	21∼50	Over 50
DeepWalk[5]	0.9001	0.9526	0.7930	0.8346	0.8011
LINE (unweighted)[8]	0.7723	0.8765	0.7738	0.6924	0.7652
LINE (weighted)[8]	0.7581	0.8990	07615	0.6760	0.7871
ESim[6]	0.7302	0.8281	0.6640	0.6352	0.6362
HIN2Vec[2]	0.7844	0.8733	0.6138	0.6117	0.7127
EHWalk (original network)	0.9608	**0.9560**	0.8356	0.8158	0.8010
EHWalk (embedding network)	**0.9641**	0.9527	**0.8743**	**0.8820**	**0.8150**

4.2 Performance Comparison

We compare the performance of our methods with the state-of-art embedding-based link prediction algorithms. In particular, LINE[8] can process both weighted graph and unweighted graph, so we carry out two experiments with LINE[8]. Furthermore, to validate the effectiveness of constructing the embedding heterogenous networks (Eq. (4)) in EHWalk, we also implement a simplified version of EHWalk, which performs heterogenous random walk on the original networks G instead of $\widehat{G}$. For each homepage, all papers are ranked according to their linking probability. The commonly used AUC (area under the curve) score, which measures the general predictive power of binary classifiers, is employed to evaluate the performance. The average AUC score over 58 randomly selected homepages are shown in Table 1. It shows that EHWalk obviously outperforms other algorithms, and the embedding network improves the performance of EHWalk significantly.

To evaluate the performance on processing different kinds of homepages, we further divide the homepages into 5 groups according to the ground-truth number of the papers associated with each homepage. In all cases as shown in Table 2, EHWalk achieves the best results, especially for the homepages with less publications. This validates the power of EHWalk to address the sparsity problem.

We also compare the performance of EHWalk with DBLP, and several widely used academic search engines such as Google Scholar, Bing and Aminer[1], which maintain the auto-generated scholars' homepages with publication lists. Given any homepage, the average F_1 score is used to measure the accuracy of linked papers. Table 3 shows a typical examples of linking results. '-' in the table means that the search engine does not generate any homepage for the scholar. It shows clearly that EHWalk performs quite steadily and can achieve relatively high performance in most test cases.

4.3 Efficiency of Linking

Different from traditional embedding-based link prediction algorithms for static networks[2, 5, 6, 8], which require rebuilding the embedding vectors for a new network (usually need hours for large networks), EHWalk supports very light-weighted incremental updating. For each new paper, it takes less than 20ms to determine which homepage it should be mapped to on the server with 3.50GHz Intel Xeon E5-1650 v3 CPU and 128G memory. This makes EHWalk quite suitable to perform real-time homepage augmentation.

Table 3: Randomly chosen examples to show the F_1 scores in different systems to associate homepages with papers.

Author	AMiner	Bing	Google Scholar	DBLP	EHWalk
Peng Lu - BUPT	-	-	-	0.3656	**0.7182**
Jing Zhang - ECUST	-	-	-	0.0750	**0.7403**
Jing Xu - NKU	-	-	-	0.1360	**0.7806**
Li Jiang - SJTU	-	-	**1.0**	0.2820	0.9863
Yu Long - SJTU	-	-	-	**1.0**	**1.0**
Dongmo Zhang - SJTU	-	0.0	-	0.1081	**1.0**
Yong Xiang - THU	0.2629	-	-	0.5600	**0.8983**
Wei Xue - THU	0.8930	0.7881	**0.9333**	0.7007	0.8792
Peng Song - USTC	0.5999	0.5497	-	0.4842	**0.7803**
Min Peng - WHU	0.4545	-	-	0.5797	**0.8948**
Chong Zhao - XMU	-	-	-	0.5714	**1.0**

5 CONCLUSIONS

In this paper, we propose a homepage augmentation technique to associate a scholar's homepage with an auto-updated list of events (e.g. paper publications). In particular, we model the relations between homepages and events as a complex heterogenous network, and propose an embedding-based heterogenous random walk algorithm, EHWalk, to predict links between homepages and events. EHWalk can not only capture complex heterogenous relations, but also support efficient linking in a dynamically evolving network. Experiments show that EHWalk outperforms state-of-art embedding-based prediction algorithms and some popular academic search engines in most test cases. EHWalk is also efficient enough for real-time deployment of homepage augmentation. In the future, we will further extend the EHWalk algorithm to support effective linking of multi-source and multi-modal events from the Web.

6 ACKNOWLEDGEMENT

The work was supported by the grants from NSFC (No.61876065, 61703109), Science and Technology Program of Guangdong Province, China (No. 2016A010101012), and CAS Key Lab of Network Data Science and Technology, China (No.CASNDST201703).

REFERENCES

[1] AMiner. Citation Network Dataset. "https://www.aminer.cn/citation".
[2] Tao-Yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning. In *CIKM 2017*. 1797–1806.
[3] Xin Jin and Jiawei Han. 2017. *K*-Way Spectral Clustering. In *Encyclopedia of Machine Learning and Data Mining*.
[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS 2013*. 3111–3119.
[5] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: online learning of social representations. In *KDD 2014*. 701–710.
[6] Jingbo Shang, Meng Qu, Jialu Liu, Lance M. Kaplan, Jiawei Han, and Jian Peng. 2016. Meta-Path Guided Embedding for Similarity Search in Large-Scale Heterogeneous Information Networks. *CoRR* abs/1610.09769 (2016).
[7] Jie Tang, Alvis Cheuk M. Fong, Bo Wang, and Jing Zhang. 2012. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. *TKDE* 24, 6 (2012), 975–987.
[8] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *WWW 2015*. 1067–1077.
[9] Baichuan Zhang and Mohammad Al Hasan. 2017. Name Disambiguation in Anonymized Graphs using Network Embedding. In *CIKM 2017*. 1239–1248.