Link Prediction in Dynamic Heterogeneous Social Networks

Anonymous Author(s)

ABSTRACT

Most real-world social networks are heterogeneous and relations between different entities have different semantic meanings. Therefore techniques for link prediction in homogeneous networks can not be directly applied on heterogeneous ones. On the other hand, recent works that investigate link prediction in heterogeneous networks, do not consider the dynamics of social networks and ignore the timestamps associated to the relations. We propose a technique, called `LinkPredict`, that predict links between two nodes of given types and a target relation in a dynamic heterogeneous.

KEYWORDS

Link prediction, social network, heterogeneous networks, network topology, meta path-based relationship.

ACM Reference format:
Anonymous Author(s). 2016. Link Prediction in Dynamic Heterogeneous Social Networks. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 3 pages.
DOI: 10.475/123_4

1 INTRODUCTION

The goal of link prediction in a social network graph [2] is to estimate the likelihood of the relationship between two nodes in future, based on the observed network. Recommending such future links have multiple applications such as friendship, item, or ad suggestions, network completion, or predicting protein-protein interactions.

Traditional link prediction techniques consider social networks to be homogeneous, i.e., graphs with only one type of edges and nodes, however, most real-world social networks (e.g. Twitter, Facebook, DBLP) are heterogeneous, i.e., have multiple relation and node types. For example, in a bibliographic social network there are different types of nodes such as authors, papers, and venues, and edges such as write, cite and publish. There are limited number of works that focused on this problem. For example, the probabilistic latent tensor factorization model... Recent works, such as [6], investigated this problem. However, such techniques do not consider the dynamics of social networks and ignore the timestamps associated to the relations.

In this work we study the problem of temporal and heterogeneous link prediction, that can be stated as follows: *Given a dynamic heterogeneous social network graph G (network with different types of nodes and links, attached with timestamps), how can we predict the future graph structure?*

1.1 Motivation

The link prediction problem for homogeneous networks have been studied in the past. However most real social networks are heterogeneous and relations between different entities have different semantic meanings. Thus techniques for homogeneous networks can not be directly applied on heterogeneous ones. Recent works, such as [4, 6], investigated this problem. However, such techniques do not consider the dynamics of social networks and ignore the timestamps associated to the relations. This is important as incorporating temporal changes helps in more accurate prediction (e.g. [11]). To the best of our knowledge, the problem of link prediction for dynamic (temporal) heterogeneous networks was not studied before.

1.2 Contributions

The main contributions of our work include:

- We present a technique, called `LinkPredict`, that predict links between two nodes of given types and a target relation;
- An evaluation of the accuracy and performance of the proposed algorithm on real social network data.

2 PROBLEM STATEMENT

Definition 2.1 (Dynamic heterogeneous social network). A dynamic heterogeneous social network is defined as a directed graph $G = (V, E)$, where V and E are set of nodes and edges of different types, and edges have timestamps.

Example 2.2. The DBLP bibliographic network[1] is a dynamic heterogeneous social network, containing different types of nodes such as papers, authors, topics, and publication venues, with publication links associated with date. Twitter social network is another example with nodes of types posted tweets, users, topics, and hashtags and time window associated with these tweets.

Link prediction problem. Given A dynamic heterogeneous social network G at time t and a target relation R, we aim to predict links of type R in G at time $t + 1$.

In order to better understand different types of nodes and their relation in a network, the concept of *network schema* [6] is used. The network schema is a meta structure graph that summarizes a heterogeneous social network and is formally defined as bellow.

Definition 2.3 (Network schema). The network schema $S_G = (\mathcal{A}, \mathcal{R})$ is a directed meta graph for a heterogeneous network $G = (V, E)$, where $\mathcal{A}$ is the set of node types in V and $\mathcal{R}$ is the set of relation types in E.

Example 2.4. Figure 1 shows the network schema for DBLP bibliographic network, where $\mathcal{A} = Authors, Paper, Venues, Topics$.

[1] http://dblp.uni-trier.de/db/

Conference'17, Washington, DC, USA
© 2016 ACM. 123-4567-24-567/08/06...$15.00
DOI: 10.475/123_4

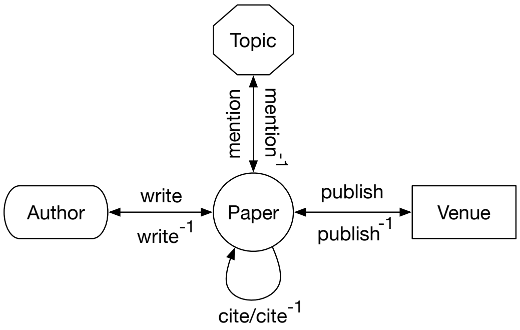

Figure 1: Network schema for DBLP network.

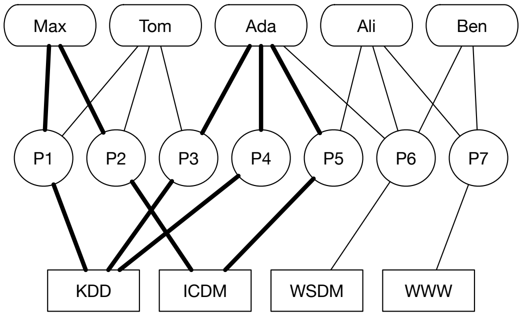

Figure 2: An example of *A–P–V–P–A* meta paths between two authors Max and Ada.

In this paper, we refer to different types of nodes for DBLP bibliographic network using abbreviations P for papers, A for authors, T for topic, and V for venues.

2.1 Meta path-based topology

Similar to the notion of network schema that provides a meta structure for the network, a *meta path* [6] provides a meta structure for paths between different nodes in the network.

Definition 2.5 (Meta path). A meta path $\mathcal{P}$ is a path in the network schema graph $S_G = (\mathcal{A}, \mathcal{R})$, denoted in the form of $P = A_1 \xrightarrow{R_1} A1 \ldots \xrightarrow{R_n} A_{n+1}$, as a sequence of relations between node types, which defines a new composite relation between its starting type and ending type.

Example 2.6. In the DBLP network example, the co-author relation can be described with the meta path $A \xrightarrow{write} P \xrightarrow{write^{-1}} A$ or in short form *A–P–A*. Paths in thick solid lines in Figure 2 correspond to *A–P–V–P–A* meta paths between Max and Ada.

Each meta path indicates a different semantic for a path connecting two nodes, and defines a unique topology representing a special relation.

3 LINK PREDICTION APPROACH

Overall approach: Given a dynamic heterogeneous social network graph $G = (V, E)$, and number of graph snapshots t, we first decompose G to a sequence of t heterogeneous graphs $G_1, .., G_t$ with respect to associated timestamps. We then use the idea of meta path [6] on a given network schema and a target relation type R (e.g. relation between two authors) to generate an augmented reduced graph (Definition 3.1) G_i^R from G_i based on R. We finally leverage the technique in [11] to predict G_{t+1}^R given $G_1^R, ..., G_t^R$, by inferring the temporal latent space representation for nodes at time $t + 1$.

Definition 3.1 (Augmented reduced graph). For a given heterogeneous graph $G = (V, E)$ and a target relation $R(a, b)$ between nodes of type a and b, an *augmented reduced graph* $G^R = (V^R, E^R)$ is a graph, where $V^R \subseteq V$ and nodes in V^R are of type a and b, and edges in E^R are of type R.

Example 3.2. An augmented reduced graph for the network in Figure 2 and $R(Author, Author)$ is a co-authorship graph, where nodes are of type Author and edges represent co-authorship relationship such as publishing in the same venue as in meta path *A–P–V–P–A*.

3.1 Algorithm

Algorithm 1 gets as an input a dynamic heterogeneous graph G, number of graph snapshots t, a meta path P, and latent space dimension k. Note that the set of all possible meta paths of length less than a given amount, can be simply produced by traversing the network schema S (for instance a BFS traversal) from a node with type a to the one with type b defined in the given target relation $R(a, b)$. We assume the meta path P is set by the user.

Algorithm. 1 Generate Predicted Graph

```
Input: A dynamic heterogeneous graph $G$, number of graph snapshots $t$, a meta path $P$, latent space dimension $k$
Output: The predicted graph $G^R$ at time $t + 1$ based on the given target relation $R$
1:  $\{G_1, .., G_t\} \leftarrow DecomposeGraph(G, t)$
2:  for each graph $G_i$ do
3:      Let $a$ and $b$ be the node types of beginning and end of $P$
4:      for each node $x \in V_i$ of type $a$ in $G_i$ do
5:          follow $P$ to reach a node $y$ of type $b$ in $G_i$
6:          $w_{xy} \leftarrow MeasureSimilarity(x, y)$
7:          add edge $(x, y)$ with weight $w_{xy}$ to augmented reduced graph $G_i^R$
8:      end for
9:  end for
10: Infer temporal latent spaces $Z_1, .., Z_t$ by optimizing Eq. 1
11: $G_{t+1}^R \leftarrow Z_t Z_t^T$
12: return $G_{t+1}^R$
```

The algorithm first decomposes G into a sequence of graphs $\{G_1, .., G_t\}$ (line 1) by considering the associated timestamps on edges. Next from each graph snapshot G_i, a corresponding augmented reduced graph G_i^R is generated (lines 2-9) for which nodes are of type a and b (beginning and end of meta path P) and edges have weight between 0 and 1, based on a similarity measure. For example *PathSim* [6], path count, or random walk are measures for the relation of two nodes given link type and meta path, such as authorship.

Once the sequence of augmented reduced graphs $\{G_1^R, ..., G_t^R\}$ are generated, we apply the matrix factorization with the block-coordinate gradient descent (BCGD) technique presented in [11] to find a k-dimensional latent space representation at each timestamp Z_τ that minimizes the quadratic loss with temporal regularization

$$\begin{aligned}&\underset{Z_1,..,Z_t}{\operatorname{argmin}} \sum_{\tau=1}^{t} \left\| G_\tau^R - Z_\tau B Z_\tau^T \right\|_F^2 + \lambda \sum_{\tau=1}^{t} \sum_u (1 - Z_\tau(u) Z_{\tau-1}(U)^T)\\ &\text{subject to} : \forall u, \tau, Z_\tau \geq 0, Z_\tau(U) Z_\tau(U)^T = 1\end{aligned} \tag{1}$$

where λ is a regularization parameter, and $(1 - Z_\tau(u) Z_{\tau-1}(U)^T)$ penalizes node u for suddenly changing its latent position. Zhu at al. [11] assume that the probability of a link between nodes depends only on their latent positions. The intuition behind their model is that nodes move smoothly in the latent space over time, and it is less likely to have large moves [3, 10]. Thus, given a sequence of graphs $G = (G_1, ..., G_t)$, we need Z_{t+1} to predict future graph G_{t+1}. They assume Z_{t+1} can be approximated by $Z_1, ..., Z_t$.

Following the temporal latent space inference technique in [11], given a sequence of graphs $\{G_1, ..., G_t\}$, we first infer $Z_1, ..., Z_t$ based on $G_1, ..., G_t$, and then use $Z_t Z_t^T$ to predict G_{t+1}.

The model assumes that user interactions are more likely to occur between similar users in a latent space representation.

3.2 Implementation

We use the implementation[2] of temporal latent space inference for a sequence of dynamic graph snapshots [11].

4 EXPERIMENTS

4.1 Dataset

The *aminer* citation dataset[3] V8 (2016-07-14) is extracted from DBLP, ACM, and other sources. It contains 3,272,991 papers and 8,466,859 citation relationships. Each paper is associated with abstract, authors, year, venue, and title.

The ml-latest-small[4] dataset describes 5-star rating and free-text tagging activity from MovieLens[5], a movie recommendation service. It contains 100004 ratings and 1296 tag applications across 9125 movies. These data were created by 671 users between January 09, 1995 and October 16, 2016. This dataset was generated on October 17, 2016.

4.2 Baseline methods

Considering the effect of time-wise data decomposition. What if we shorten timespans of each G_t? The extreme is having only one graph or having it for each year. Can we find a trade-off?

Heterogeneous non-temporal - Collection of PathSim (PathCount, NormalPCount, RandomWalk, Symmetric random walk) Homogeneous non-temporal (Katz, Jaccard) Homogeneous temporal (Katz, Jaccard)

1. We use prediction error to evaluate the inference accuracy. Given the training graph G1, . . . , Gt, prediction error is defined as... Therefore, a smaller prediction error indicates better inference accuracy.
2. For link prediction accuracy, we use Area Under Curves (both Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves), termed as AUCROC and AUCPR.
3. NDCG
4. Trade-off analysis (time overhead)

5 RELATED WORK

[11] [6] [5] [1] [8] [7] [4] [9] [2]

6 CONCLUSIONS

TBA.

REFERENCES

[1] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li. Meta structure: Computing relevance in large heterogeneous information networks. In *KDD'16*, 2016.

[2] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

[3] P. Sarkar and A. W. Moore. Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 7(2):31–40, 2005.

[4] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 121–128, July 2011.

[5] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla. When will it happen?: Relationship prediction in heterogeneous information networks. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 663–672, New York, NY, USA, 2012. ACM.

[6] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proceedings of the VLDB Endowment 4 (11)*, pages 992–1003. VLDB Endowment, 2011.

[7] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):11, 2013.

[8] C. Wang, Y. Sun, Y. Song, J. Han, Y. Song, L. Wang, and M. Zhang. Relsim: Relation similarity search in schema-rich heterogeneous information networks. 2016.

[9] Y. Yang, N. Chawla, Y. Sun, and J. Hani. Predicting links in multi-relational and heterogeneous networks. In *2012 IEEE 12th International Conference on Data Mining*, pages 755–764, Dec 2012.

[10] J. Zhang, C. Wang, J. Wang, and J. X. Yu. Inferring continuous dynamic social influence and personal preference for temporal behavior prediction. *Proceedings of the VLDB Endowment*, 8(3):269–280, 2014.

[11] L. Zhu, D. Guo, J. Yin, G. V. Steeg, and A. Galstyan. Scalable temporal latent space inference for link prediction in dynamic social networks. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 28(10):2765–2777, Oct 2016.

[2] https://github.com/linhongseba/Temporal-Network-Embedding

[3] https://aminer.org/citation

[4] http://files.grouplens.org/datasets/movielens/ml-latest-small-README.html

[5] https://movielens.org/