A Topical Graph Kernel for Link Prediction in Labeled Graphs

Snigdha Chaturvedi SNIGDHAC@CS.UMD.EDU
Hal Daumé III HAL@CS.UMD.EDU
Taesun Moon TSMOON@UMIACS.UMD.EDU
University of Maryland, College Park, MD, USA

Shashank Srivastava SHSRIVA@GMAIL.COM
Tower Research Capital, Gurgaon, India

Abstract

This paper proposes a solution to the problem of link prediction in labeled graphs with additional text information associated with the nodes. By fitting a topic model on the text corpus and some processing, we compute the topics of interest to a node. We propose a walk based graph kernel which incorporates the node's interest and thus represents structural as well as textual information. We then make predictions about the existence of unseen links using a kernelized SVM. Our experiments with an author citation network shows that our method is effective and significantly outperforms a network-oriented approach.

1. Introduction

Link prediction is the problem of predicting future links within some graph structure (e.g. a social network or a citation network) given some snapshot of this graph at a preceding moment. Many approaches to the problem consider only the structural features intrinsic to the network itself (Liben-Nowell & Kleinberg, 2007), such as friendship links or citation links, and ignore the properties of the nodes themselves which could be social agents or scientific articles. Alternative approaches to link prediction incorporate intrinsic properties of nodes to both improve prediction performance and provide greater insight into the network. The node properties provide an alternative view into the network that complements approaches that subsist solely on the edges.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

In this paper, we propose a graph kernel which we call the *topical graph kernel* to solve the task of link prediction in labeled graphs with additional text information associated with the nodes. Our kernel subsumes the idea of walk-based kernels (Li & Chen, 2009), and enriches them with textual data from nodes by using a topic model to represent node interest. We then make predictions about the existence of unseen links using a kernelized SVM. Our experiments on a citation graph of authors of scientific papers shows that our method significantly outperforms a network-oriented approach.

2. Related Work

The link prediction problem has been addressed by several researchers. There are node neighborhood based methods which work under the assumption that two nodes are likely to form link if their sets of neighbors have a large overlap. These techniques employ common neighbors (Newman, 2001; Kossinets, 2006), Adamic-Adar Index (Adamic & Adar, 2003), preferential attachment (Newman, 2001), etc. to measure neighborhood similarity. There are other ensemble-of-all-paths based approaches which refine the notion of shortest path distance by considering an ensemble of all paths (Katz, 1953; Jeh & Widom, 2002).

Recently there have been attempts to enhance text models with information from graphs. Mei et al. (2008) address the problem of learning topic models on labeled graphs. They propose to regularize a topic model with a regularizer based on the graph structure. However, their approach does not directly apply to link prediction. Nallapati et al. (2008) and Chang & Blei (2010) propose solutions to the citation prediction problem using topic models to model the parameters which determine if a given paper would cite another given document. However their method is limited to

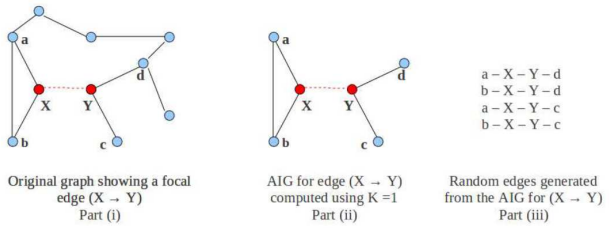

Figure 1. AIG construction and random walks generation for a given link

modeling existence of individual links and does not incorporate the graphical topology of the neighborhood.

Our approach is most closely related to that proposed by Li & Chen (2009) which uses graph kernels to predict links in recommendation systems. They experiment with a bipartite graph of users and items and a link exists between a user-item pair if the user buys the item. For making predictions, they compare the neighborhood of a given user-item pair with other pairs using a graph kernel. Their method is solely dependent on graph structure and doesn't utilize any textual information about nodes.

3. Methodology

This paper proposes a method to predict unseen links in a labeled graph. In the context of the author citation graphs, the nodes represent authors and there is a link between two nodes whenever one of them cites the other. In social networks, a link can represent the 'follower' or 'friend' relation between two people. Each node has text documents associated with it in the form of papers authored or blogs or tweets posted.

We model this problem as a learning task where we use a kernerlized classifier (an SVM) to learn from historical data and make predictions about future links. Our graph kernel, the *topical graph kernel*, compares two links by quantifying the topological similarity and node interests. Node interests are modeled using topics learnt from a topic model (Blei et al., 2003) on the text documents. The rest of this section describes this approach in detail.

3.1. AIG Construction

This is the first step of the approach. Computation of the graph kernel is based on a comparison of the subgraph centered on the link of interest $X - Y$ where X and Y are nodes incident to the link. For this purpose, Li & Chen define an AIG to be an induced sub-graph consisting of the two nodes X and Y and neighbors of these nodes accessible within n hops from either node. The AIG also contains all the links between these neighboring nodes. The link which the AIG centers on is the *focal link* and the two nodes which are incident to this link are *focal nodes*. While training, since the truth is known, the focal link $X - Y$ will actually exist whereas during testing, one will have to construct the AIG assuming that the focal link exists. In the testing phase the classifier predicts how probable it is for this hypothetical focal link to exist. An example of an AIG induced from a bigger graph is shown in Part (ii) of Figure 1.

3.2. Graph Kernel Computation

This is the second step of the approach which quantifies the similarity between two focal links by computing the graph kernel of their respective AIGs.

Given an AIG of a focal link $X - Y$ the preliminary step is to decompose the AIG into random walks. Only those random walks are significant which include both the focal nodes X and Y. An example of generating random walks is shown in Part (iii) of Figure 1.

Once both the AIGs have been decomposed into sets of random walks, the graph kernel, $K(AIG_i, AIG_j)$, can be computed as:

$$K(AIG_i, AIG_j) = \frac{\sum_{h_i \in AIG_i} \sum_{h_j \in AIG_j} K_h(h_i, h_j)}{(N_i + N_j)/2}$$

where

h_i = random walk from the AIG_i
h_j = random walk from the AIG_j
N_i = Number of random walks from AIG_i
N_j = Number of random walks from AIG_j
$K_h(h_i, h_j)$ = Random walk Kernel representing the similarity between two random walks h_i and h_j

Given two random walks h_i and h_j related to focal links $X^{h_i} - Y^{h_i}$ and $X^{h_j} - Y^{h_j}$ respectively, the random walk kernel $K_h(h_i, h_j)$ is defined as:

$$\begin{aligned} K_h(h_i, h_j) &= K_n(n_1^{h_i}, n_1^{h_j}) \times K_l(n_1^{h_i} - n_2^{h_i}, n_1^{h_j} - n_2^{h_j}) \\ &\times K_n(n_2^{h_i}, n_2^{h_j}) \times \cdots \times K_n(X^{h_i}, X^{h_j}) \times K_n(Y^{h_i}, Y^{h_j}) \\ &\times \cdots \times K_n(n_{t-1}^{h_i}, n_{t-1}^{h_j}) \times K_l(n_{t-1}^{h_i} - n_t^{h_i}, n_{t-1}^{h_j} - n_t^{h_j}) \\ &\times K_n(n_t^{h_i}, n_t^{h_j}) \end{aligned}$$

where the random walks h_i and h_j are aligned as

$$\begin{aligned} h_i &= n_1^{h_i} - n_2^{h_i} - \ldots - X^{h_i} - Y^{h_i} - \ldots - n_l^{h_i} \\ h_j &= n_1^{h_j} - n_2^{h_j} - \ldots - X^{h_j} - Y^{h_j} - \ldots - n_l^{h_j} \end{aligned}$$

and

K_l = Kernel defining similarity between two two links
K_n = Kernel defining similarity between two nodes

For our task, the link kernel K_l for all link pairs is simply taken as 1. On the other hand, the node kernel K_n could be defined in several ways given the complexity of associated text. We first define our baseline node kernel which does not incorporate textual content. We then propose our alternative—the labeled graph kernel—based on topic models which allows us to incorporate textual content associated with individual nodes in the node kernels.

Baseline Method Li & Chen assumes that non-equal nodes share the same similarity value. On this assumption a binary node kernel is defined below assuming that a non-focal node is similar only to itself. Focal nodes are, however, similar to all nodes of the graph to allow predictions about new nodes.

$$K_n(n_i, n_j) = \begin{cases} 1 & \text{if } n_i = n_j \\ 1 & \text{if } n_i \text{ or } n_j \text{ is a focal node} \\ 0 & \text{otherwise} \end{cases}$$

The intuition for the definition goes as follows: suppose $a - X - Y - b$ is a random walk. We could learn from this walk that if X links to a and Y links to b, then X links to Y. To make prediction about a new link $(X' - Y')$ with a random walk as $a - X' - Y' - b$, we could say that the link $X' - Y'$ has a high probability of existence based on our recent learning.

Topical graph kernel (TGK) The above node kernel makes predictions based only on the structural cues. We propose to extend the node kernel to include a node's interests in various topics across the corpus.

We learn a topic model (Blei et al., 2003) over the complete text corpus. A topic model is a generative mixture model of latent variables (called "topics") over a corpus where each document is represented as a sparse mixture of K topics. It is a way of reducing complex text to a smaller topical space to facilitate exploration or manipulation of large corpora. The model parameters are learned by fitting the model to the observed data. Letting $P(k|d)$ the probability of topic k in document d, the interest of node n in k is computed as:

$$P(k|n) = \sum_d P(k|d)P(d|n)$$

For simplicity, we assume that all the documents associated with a node/author are equally probable and so $P(d|n)$ is one divided by the number documents n has authored. We then represent a node as a K dimensional vector of interests in each of these topics:

$$\vec{n} = \langle P(k_1|n), P(k_2|n), P(k_3|n) \dots P(k_K|n) \rangle$$

Finally, the node kernel is defined as:

$$K_n(n_i, n_j) = \frac{\vec{n}_i \cdot \vec{n}_j}{\|\vec{n}_i\| \|\vec{n}_j\|}$$

All of these kernels meet the semi-positive definiteness property (Li & Chen, 2009). The K_n and K_l satisfy this condition and since the kernel of a random walk is a product of these kernels, it is a well formed kernel.

3.3. Model Training and Prediction

The graph kernel described above can be used to compute kernel values between all pairs of links present in the training data. A kernelized learning algorithm can then be trained on this data. Since we train on only positive examples of links present, we use a one-class SVM for training.

At test time, given a test focal link, it is assumed that the link exists and the AIG is constructed and decomposed into random walks. The trained SVM is then used on this test link to predict its existence.

4. Empirical Evaluation

4.1. Dataset

For our experiments, we use an author citation graph built from the ACL anthology data (Radev et al., 2009). Each node of the graph is an author and there exists a link between author i and author j if i ever cited j in at least one paper. We pruned the graph to exclude authors who published less than 5 papers.

The training graph consists of citation information from papers published between 2006 to 2010 and contained about 136000 links. The test graph consists of author citation links based on publications from 2011. It contains about 26000 links.

4.2. Experiments and Results

The computational overhead of computing the graph kernel can get prohibitive for real-life graphs. In our experiments, training was performed only on a random sample of L links randomly chosen from the training set. However, in order to incorporate information from the whole graph, AIGs for the training links were built using the complete graph.

Our approach was tested on two separate test sets of size 200 each: Test Set1 and Test Set2, each containing both positive and negative links. A positive link is one that is randomly sampled from the set of links in the test graph whereas a negative link is a link which doesn't appear in the complete time period of 2006-2011. For a fair evaluation we ensure that the model

Method Name	5-fold CV on train set	F-measure	
		Test Set1	Test Set2
Baseline	30.66%	66.40%	60.49%
TGK	42.92%	76.70%	74.34%

Table 1. Comparison of TGK with the baseline (Li & Chen, 2009) ($L = 500$ and $K = 500$)

is not tested on links that were seen during training. Also, in order to test the sensitivity of the model towards skewness of the test sets, ratio of positive to negative links in the two test sets, Test Set1 and Test Set2, was kept different: 3:1 and 1:1 respectively.

Table 1 compares the topical graph kernel (TGK) with the baseline using F-measure of the positive links. It shows that TGK significantly outperforms the baseline method during 5 fold cross validation on the train set and on the two test sets. Also, the model's performance on the two sets was comparable, indicating robustness of the model towards class bias at test time.

We also study the effect of training set size L and number of topics K on the performance of TGK on one of the test sets. Part (a) of Figure 2 shows a learning curve of the F-measure as the size of the training set is increased. While both methods benefit from an increase in train set size, we see that TGK constantly outperforms the baseline method. Also, for the smallest training set of size 100, the baseline method yields an F-measure of around 20% while the proposed approach still leads to reasonable accuracy (about 53%). Similarly, part (b) of Figure 2 shows that the performance of TGK increases with increase in number of topics. This happens because a higher K strengthens the representative power of the node vector and is better at modeling node interests.

5. Conclusion

In this paper we have presented the topical graph kernel to predict links in rich labeled graphs using a graph kernel. The kernel, based on random walks to capture structural cues, was enhanced using node similarity as computed from the text documents associated with the nodes. For computing similarity, an LDA model was used to compute a node's interests in various topics and nodes were compared using cosine similarity. Our experiments with an author citation network demonstrated the usefulness of the enriched kernel.

With the availability of rich labeled graphs such as social networks, approaches which work for multi-faceted data are increasingly desirable. Another characteristic of such networks is their continuously evolving nature. Future work could focus on improving the kernel

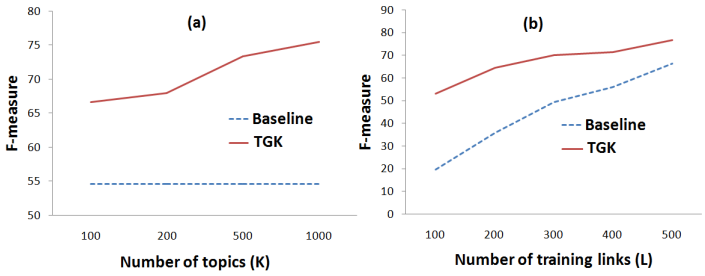

Figure 2. Performance of model with increasing train set size and number of topics

to incorporate the dynamic nature of such graphs by modeling the age or dynamics of individual links.

References

Adamic, Lada A. and Adar, Eytan. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

Chang, Jonathan and Blei, David M. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150, 2010.

Jeh, Glen and Widom, Jennifer. Simrank: a measure of structural-context similarity. In *SIGKDD*, pp. 538–543, 2002.

Katz, Leo. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39– 43, 1953.

Kossinets, G. Effects of missing data in social networks. *Social Networks*, 28(3):247–268, 2006.

Li, Xin and Chen, Hsinchun. Recommendation as link prediction: a graph kernel-based machine learning approach. In *JCDL*, pp. 213–216, 2009.

Liben-Nowell, David and Kleinberg, Jon. The link-prediction problem for social networks. *JASIST*, 58(7): 1019–1031, 2007.

Mei, Qiaozhu, Cai, Deng, Zhang, Duo, and Zhai, ChengXiang. Topic modeling with network regularization. In *WWW*, pp. 101–110, 2008.

Nallapati, Ramesh M., Ahmed, Amr, Xing, Eric P., and Cohen, William W. Joint latent topic models for text and citations. In *SIGKDD*, pp. 542–550, 2008.

Newman, Mark E. Clustering and preferential attachment in growing networks. In *Physical Review E*, volume 64, 2001.

Radev, Dragomir R., Muthukrishnan, Pradeep, and Qazvinian, Vahed. The ACL Anthology Network corpus. In *NLPIR4DL*, pp. 54–61, 2009.