Link Prediction using Textual and Graphical Features

Xin HUANG
Student ID:116260910040
Leaderbroad name:sebastian
Email: hx_mz@qq.com

***Abstract*—In this report, we focus on the problem of link prediction. Given the information of papers and their existing citation networks, we are supposed to predict the existance of citing behavoirs of paper pairs in the test set. In our proposal, we extract the features both from the textual information of papers as well as the graphical information of the citation networks and then train a neural network classifier 5 times with these features we select the predictions appearing in at least twice to make the final prediction. The f_1 score of our prediction reaches around 0.9765.**

I. Introduction

Link prediction is a task with multidisciplinary applications including bioinformatics, social networks and online stores. For example, given users' profiles such as interests, career and their social network statue (i.e. who they follow in the facebook), we may be able to infer the possible relations among two users. Another important application is the ciataion prediction in the scholarly networks. In the citation prediction, we are supposed to predict whether the citations exist between papers, these predictions can be applied in the recommendation of scientific articles as well as discovering the milestone papers in a specific scientific area.

In this data challage, we mainly focus on link prediction on the scholarly data. The dataset contains citation networks of 27770 documents with the information of title, abstract, authors and publish time. The intuition of the feature extraction is to simulate the way the scholar do when adding the citations to their academic works. Generally, when an author sellect the documents that he wants to cite, there are multiple factors to consider: the correspondance of textual content, the statue of cited documents in the citation network, etc. In the following sections, we will give a concrete description about the features that we select.

II. Selected features

The features that we selected can be divided into three categories, the textual features which mainly focus the level of word and topic; the network features about the importance of papers as well as their interaction activities with other papers; the third category mainly focuses on the the features about the authors, publish year, etc.

A. Textual informations

For an author, when he cites some documents, there is a high probability that he cite the related papers to his own writing. The correspondance of papers mainly lays on the topic level. If two papers describe the similar topic, these two papers are likely to link each other via citation. Another important sign indicating the citation is the common words in their titles or their abstract. A concrete example is the papers contains the words ”topic model”, citations occurred frequently among them since, all these paper describe the topic modelling.

1) Topic level features extraction: In order to extract features in the topic level, we adopted two main strategies, probabilistic latent semantic analysis and the latent semantic analysis.

The first method for extract the topic feature is to use Probabilistic Latent Semantic Analysis. For a document, it contain a distribution of topic, and the word. The intuition of PLSA is to simulate the process of writing a scientific article. When we writa a document, we firstly select a topic following the document-topic distribition and then selece the words with the topic-word distribution. Figure 1 shows the intuition of latent dirichlet allocation and its graphical representation is shown in the Figure 2.

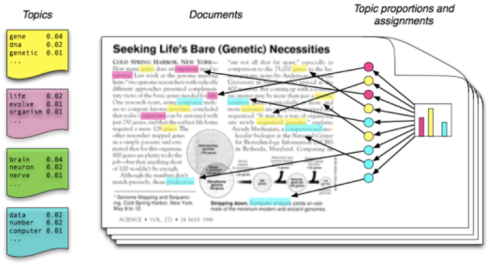

Fig. 1. Topic model

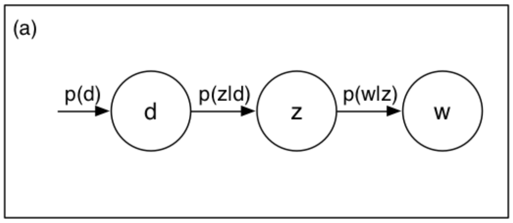

Fig. 2. Graphical representation of PLSA

According to [1], the method of non-negative matrix decomposition has the close link with the probablistic latent semantic analysis[2]. The goal of non-negative matrix fatorizations is to find two non-negative matrices (W, H) whose product approximates the original non-negative matrix X. Thus, we can implement the PLSA model using non-negative matrix fatorizations.

Another method is to use the LSA, which mainly concern the singular value decomposition of the correated maxtix of the tf-idf representation of documents.

2) Graph of word feature extraction: While the topic modelling based method is build on the bag of word assumption and thus pay no attention to the context of words in the document, we thus utilise the feature named *tw-idf* to leverage the context sensitive features of features.The *tw-idf* is built on the graph of word model[3], which aims at create a graph which reflect not only the occurrance of important words but also the co-occurrance of phases. Words, in this model, forms a graph and the *tw-idf* is computed by the formula 1

$$tw - idf(t, d, D) = degree(node_t) \times idf(t, D) \qquad (1)$$

3) Word level feature extraction: We may observed that among the similar scientific, authors tend to utilize the same words, especially the terminologies. For example, in the following five articles discribing ”topic model”:

- *Correlated topic models*
- *Hierarchical topic models and the nested chinese restaurant process*
- *Dynamic topic models*
- *Joint latent topic models for text and citations*

They all indlude the words of ”Topic model” and among them, the newer documents cite the older documents. This phenomenon gives us a inspiration to select the common words the title and the abstract as features. Moreover, we also take into consideration of the common used words therefore we also vectorize the titles and abstract in forms of *tf-idf* vector.

B. Network feature extraction

Citation network directly reflects the activities of citation, therefore the network feature is crucial in making link prediction in citation network. In this aspect we mainly select six following features:

- *Common neighbors*
- *Node importance*
- *Community*
- *Nodes’ clustering factors*

1) Common neighbors: When we write a paper, we usually cite the papers which are refered by papers in our citation list since these papers form a chain of citations and thus related to each other in the aspect of topic. For example, if we want to write the paper about the *topic evolution* we may refer paper *Dynamic topic models*. Moreover, we may refer to the paper *Latent Dirichlet Allocation* which put forward the lda topic model. Based on this phenomenon, we compute the number of common neighbour between each pairs to predict as the feature.

2) Node importance: In the academic field, we usually cite the milestone paper in the area, thus, the important nodes in the citation neworks are likely to be cited. To leverage the important papers in the citation network, we adopt the PageRank algorithm[4]. The output of the Pagerank is a vector reflecting the invariant distribution of the nodes. The larger number in the vector, the more important the node is.

3) Community: Community detection is to find the best partition of graph, with each part tends to be a dense network. Therefore, the community of the node is considered as an important feature, if two nodes belongs to the same community, the probability that one paper cite the other paper is high.

4) Nodes’ clustering factors: The feature that we choose regarding the clustering factors of nodes mainly contains the core value of nodes and the clustering value of the node.The clustering coefficient is calculated by formular *1*.

$$c = \frac{3 \times \#triangle}{\#connected\ truplelets\ of\ vertices} \qquad (2)$$

C. Author and Time Features

In the research areas works done by a single person are not independent. For most of researchers or Ph.d candidtates, their publications mostly come from the projects that they participate in. In these case, the papers with the same authors tend to cite each other due to the correlation in the topical level. Another important issue of paper citation is the temporal gap, since researchers may be more interested in the recenly published works, since they represent the most advanced technologies. For the reasons mentioned above we extract the number of common author and the time difference between papers as the features.

III. Classifier selection

Since the data is labelled, we may utilised the supervised learning method to train the model and give the predictions. Generally speaking, we have tried five different classifiers:

- *Support vector machine*
- *Neural Network*
- *K nearest neighbour classifier*
- *Random Forest*

A. Support vector machine

Support vector machine (SVM) is a supervised learning methods mainly built on the principle of maximizing the margin. Besides, we can also use the kernal function to project the non-linear seperable data to a novel hyperplan. The advantges of Support vector machine is that it can effectively in high dimensional spaces.

B. K nearest neighbour classifier

The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label based on the number of samples in each classes. There are two types of hyperparameters, the first one is the number of samples k, the other one is distance, which can, in general, be any metric measure: standard Euclidean distance is the most common choice.

C. Random Forest

A random forest is a ensemble classifier that aims at fitting a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement using the method of bootstrap.

D. Neural Network

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function $f(\cdot) : R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features $X = x_1, x_2, ..., x_m$ and a target y, the advantage of neural network is that it can learn a non-linear function approximator for classification.

IV. Programming pipeline

The program is written in python. It has three main parts: data pre-processing, feature extraction and the model learning. For the simplicity of progeamming, we adopt the packet *scikit learn*[1]. To extract the network feature and the community feature, we utilised the packet *networkx*[2] and *community*[3]. Last but not least, we utlisized the *keras* for construct the neural network. For the detail implementation, please refer the source code.

V. Result

All together, there are 19 features given a paper:

- Topic representation selected by PLSA
- Lsa representation of abstract
- Lsa representation of topic
- Tf-idf representation of abstract
- Tf-idf representation of title
- Tw-idf representation of title
- Tw-idf representation of abstract
- Number of common words in abstract
- Number of common words in title
- Community of papers
- Pagerank score of cited paper
- Pagerank score of citing paper
- Clustering score of cited paper
- Clustering score of citing paper
- Degree of cited paper
- Degree of citing paper
- Number of common author
- Time gap

We have tested our results on the the five classifiers that we choose. For knn classifier, we set k to be 5; for supprt vector machine, we utlisize linear kernal; for random forest, the number of tree is set to be 300 with the maximun depth of 6. For neural network, we construct a three-layer network.

TABLE I
Topic parameter

Model	Number of Topics
PLSA	50
LSA(abstract)	200
LSA(title)	100

TABLE II
Performance

Model	F_1 score
K nearest neighbour	0.9359
SVM	0.9642
Random forest	0.9737
Neural Network	0.9760
Hybrid Neural Network	**0.9765**

Since the neural network runs with random initialzation, we run the model five times and admit the prediction that appears at least 2 time (namely Hybrid neural network). Table I demonstrate the hyper-paramer that we choose in the feature selection, for the sliding window of tw-idf, we set it to three. We have tested classifiers and the f_1 score are used to evaluate the performance of classifiers. From the table II we can concluded that the hybrid neural network classifier works the best.

VI. Conclusion

In this data challenge, we have studied differents methods to leverage the feature from the data and extracted 18 features for the link prediction. We have also utilised different classifier to predict the existance of citations in the test set. We eventually adopted the hybrid neural network classifiers which gives a *f1* score of 0.9765.

Acknowledgment

We would like to thank Prof. Michalis Vazirgiannis for giving us course of data mining, and thank Dr. Christos Giatsidis for the help in the lab session.

References

[1] E. Gaussier and C. Goutte, "Relation between plsa and nmf and implications," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '05. New York, NY, USA: ACM, 2005, pp. 601–602. [Online]. Available: http://doi.acm.org/10.1145/1076034.1076148

[2] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[3] F. Rousseau and M. Vazirgiannis, "Graph-of-word and tw-idf: new approach to ad hoc ir," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 2013, pp. 59–68.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[1] http://scikit-learn.org/

[2] http://networkx.github.io/

[3] http://perso.crans.org/aynaud/communities/