

DATA CHALLENGE: PREDICTION OF MISSING LINKS IN A CITATION NETWORK

Léo Boulouc ^(1, 2)

Cyril Gaudetroy ^(1, 2)

Ariel Shemtov ^(1, 2)

Thai-Chau Truong ^(1, 3)

⁽¹⁾ ENS Cachan, 61 Avenue du Président Wilson, 94230 Cachan

⁽²⁾ ENSTA ParisTech, 828 Boulevard des Maréchaux, 91120 Palaiseau

⁽³⁾ Télécom ParisTech, 46 Rue Barrault, 75013 Paris

ABSTRACT

This report presents our work on the data challenge for the “Advanced Learning on Text and Graph Data” (ALTeGraD) the purpose of which was to predict missing links in a citation network.

1. INTRODUCTION

The citation network that we consider here is made of research papers at its nodes, and links between nodes where one of them cites the other. The information available for each paper in the network is the following: title, authors (some with affiliation), year published, publisher and the abstract.

2. FEATURE ENGINEERING

For our learning strategy, we compute features to describe the relationship between each pair of nodes in the citation network.

2.1. Shared citations

The quantity used to describe the shared citations is the number of common neighbours in the network. In other words, the number of distinct papers that are linked to both nodes of the considered pair. As described in section 3, this appears to be the most important feature of all.

2.2. Shared relevant words

The strategy used here is a bag-of-words model: a matrix with lines corresponding to documents and columns to words (ignoring stop-words) is created. The values in this matrix stand for how many times each word is present each document. In order to introduce a notion of word weight as first proposed in [1], each column is then normalized so that it sums to one. Now, to compute the similarity between two documents, we used the intersection kernel, which computes the sum of the element-wise minimum between two line vectors.

2.3. Publication dates

Although it appears to have less impact on the result, we used the difference between the years of publication for each pair. The fact that this feature doesn’t have much impact comes from the fact that the network that we consider is non-oriented.

2.4. Shared authors

The number of shared authors between two articles is a very significant indicator, although shared authors don’t happen in many pairs.

3. MODEL TUNING AND COMPARISON

The strategy that we chose is a random forest. The advantages of this method is that it provides feedback on which features perform best, and also it tends to avoid overfitting. The model is trained on the training set, and then used on the test set to predict whether there exists a link or not.

4. REFERENCES

- [1] Karen Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.