Automatic Textual Patent Similarity from the Search Report at the European Patent Office

Pablo Fonseca[*], Isaac Yrigoyen[†], Jacques Wainer[‡]

September 20, 2018

1 Introduction

When an application is filled at the European Patent Office, an examiner would search for its prior art in order to find other potential conflicting patents. As it is well known, a patent should be verified to be novel before it is granted. The Search Report is issued before the first patent publication after this and would contain citations to the conflicting documents along a code for each citation, the so-called citation category. Our hypothesis is that these citation categories do carry valuable information on relative similarity among patents and might inform a pairwise textual similarity function through weak supervision. Finally, such similarity measure might be fit for the task of prior-art search: in a retrieval setup or in patent mapping (a 2D distance-preserving visualization commonly used).

2 Methods

We use two data sources: PATSTAT, a relational database issued by WIPO twice a year (which has patent metadata as well as textual features such as the title and abstract) and also the Open Patent Services, a web service provided by the European Patent Office that allows to retrieve other sections of the patents such as the claims and description. We used a Metric Learning approach [1] to obtain a supervised pairwise similarity, specially we implement OASIS [2] that would produce a pairwise distance function $d_W(x, y) = x^T W y$, where W is a matrix that is learn in order to minimize a constraint provided as triplets (p, p^+, p^-), where a patent p should have a closer distance to a more similar patent p^+ (the possitive sample) than to p^- (the negative sample). We use citation categories from the search report to build the triplet samples. In order to build a lower dimensional feature description of the text we use topic

[*] Montreal Institute for Learning Algorithms, University of Montreal, Montreal, Canada

[†] Department of Engineering, Pontifical Catholic University of Peru, Lima 32, Peru

[‡] Institute of Computing, University of Campinas, Campinas, Brazil

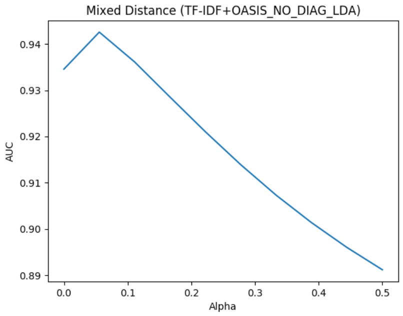

Figure 1: AUC for $\alpha*$OASIS$+(1-\alpha)$*TFIDF for IPC A47

modelling trained on the patent class analyzed. Finally, in order to simulate a prior-art search scenario, train and test sets are separated on a temporal axis. A positive pair is, i.e. citing-cited relationship, and a negative pair is randomly matched with another patent given that there is no cited-citing relationship between these patents.

3 Results and Conclusions

The best distance we manage to produce was a linear combination of TF-IDF Cosine Similarity + OASIS on the Topic Distribution. However, OASIS was forced to learn outside the W matrix diagonal. The results are measured in terms of the ROC-AUC on patent pairs (positive/negative). In figure 1 we see the improvement of the method, which is modest, yet seem to proof that there is information in the citation categories, extending the observation for citations in [3].

References

[1] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[2] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(Mar):1109–1135, 2010.

[3] Lixin Chen. Do patent citations indicate knowledge linkage? the evidence from text similarities between patents and their citations. *Journal of Informetrics*, 11(1):63–79, 2017.