Text Enrichment-Based Enhanced Patent Mining using Clustering Techniques

Caner Aksoy caner.aksoy@stu.bahcesehir.edu.tr Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey

C. Okan Sakar okan.sakar@eng.bau.edu.tr Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey

Sercan Ozcan sercan.ozcan@port.ac.uk Portsmouth Business School, University of Portsmouth, Portsmouth, United Kingdom

Abstract

While the pace of technological developments is at a level where it is difficult for the end users to stay up-to-date with changes, it is even more challenging for most Information Technology (IT) companies as it is a matter of survival in the industry. IT sector is a rapidly developing, knowledge intensive field where tangible assets are far less of a concern. As IT field requires lower resource-intensive investments, it is very tempting for new entrants and existing companies to be in the field. However, if the technological roadmap and product/service investment decisions of a firm is not right, the invested time, effort and resources may be wasted and, lead to poor performance and possibly bankruptcy at the end.

In order to follow the rapid developments in a high-tech field and to be leading firm in a cutting-edge technology, either R&D intensive (i.e. IBM) or innovation efficient approach (i.e. Apple) is essential. Therefore, there is an increasing effort in analysing various kinds of text data gathered from diverse data sources, such as patent sources, scientific publications and, social network platforms to explore the promising technology fields and potential applications of the cutting-edge technologies. These efforts require the need of applying many text processing techniques on the raw text data such as text segmentation, summary extraction, feature selection, term association, clustering, taxonomy generation, topic identification, keyword identification, information mapping, co-word analysis and domain analysis (Tseng et al. 2007).

Besides the technique, it is of great importance to look at the data that best represents the actual information because one fact does not apply for all domains. For patents, one can look at many pieces of information such as the title itself, abstract, main text or even citations (Rodriguez et al. 2016). There are still many other possibilities to understand, examine and identify different sections of patent documents where valuable knowledge can be extracted. Patent examination itself is also a valuable process and requires specific know-how. For example, publication or social media is another type of data where academicians follow analysis on similar matters where co-occurrence, citation, or co-ownership/co-authorship are used.

In this study, we aim to investigate, to the best of our knowledge for the first time, the usefulness of text enrichment approach in clustering of patent documents considering relevant studies such as patent mining, tech mining and scientometrics. This approach has been successfully used in other fields with different enrichment methods (Gharib et al. 2012, Faatz et al. 2002). Text enrichment technique is a decisive solution for sparse datasets. Considering the sparsity of the co-occurrence matrix that is constituted from the abstracts and titles of the patent documents, using text enrichment we aim to obtain better clustering that represent the fields in the relevant topic. For this purpose, first a cloud computing related patent dataset of approximately 60.000 is obtained using advanced retrieval methods (Ozcan et al., 2017). After

the data retrieval process, the first step of the methodology followed in this paper is data cleaning and reduction which involves stop-word removal, lemmatization, and filter feature selection processes. Then, the co-occurrence matrix of the remaining words representing the similarities between the word pairs is generated. After this step, we apply two different approaches for comparison. The first approach is based on applying principal component analysis (PCA) to the original co-occurrence matrix and then feeding the obtained reduced set to various clustering algorithms. The literature shows that k-means and density-based clustering algorithms are the most popular ones for similar analysis, so we have included both with expectation-maximization algorithm to illustrate results with different algorithms. The second approach is, in parallel to the mail goal of the study, is to enrich the content of the co-occurrence matrix, apply PCA to the enriched matrix and then feed it to the clustering algorithms. The results are evaluated with sum-of-squared error and silhouette coefficient metrics. The resulting clusters are also represented by top-N keywords with the highest term frequencies in the corresponding cluster and analysed in terms of the complementary and semantic relations among the keywords in the same cluster.

The knowledge base used to enrich the co-occurrence matrix is the Wikipedia English corpus, approximately 6 GB of raw data, which is trained by continuous bag-of-words and the skip-gram models. Wikipedia data has been successfully used for text enrichment in the literature (Hu et al., 2009, Ran et al., 2015). The results showed that text enrichment technique improves the evaluation metrics of all clustering algorithms by 6% up to 18% when compared to the results obtained with the initial matrix. This proved to be closely related with the initial sparsity of the matrix as the scores were getting higher when the term frequency threshold is increased. We observe that the resulting clusters have a better distribution of words which form meaningful cloud computing related phrases. Another observation is that the patents retrieved by using the keywords explored using the enriched matrix returns more recent studies in these areas. We should also note that both with original and enriched matrix, expectation maximization outperformed the other clustering methods in terms of the cluster evaluation metrics used in this study and the keywords that fall into specific clusters are semantically or logically related representing the applications areas and sub-categories of cloud computing.

References

Faatz, A., Steinmetz, R. Ontology Enrichment with Texts from the WWW. Proceedings of the 13th European Conference on Machine Learning, Helsinki, Finland (2002).

Gharib, Tarek F., Nagwa Lotfy Badr, Shaimaa Haridy and Ajith Abraham. Enriching Ontology Concepts Based on Texts from WWW and Corpus. J. UCS 18 (2012): 2234-2251.

Jun, S. (2012). A clustering method of highly dimensional patent data using Bayesian approach. International Journal of Computer Science Issues, 9(1), 7-11.

Ozcan S., Alp B. & Sakar C. O. (2017). A Patent-mining Study Focusing on Technological Trends and Diffusions: A Case of Cloud Computing. R&D Management Conference, Leuven, Belgium, July 1-5, 2017.

Rodriguez et al. (2016). Patent Clustering and Outlier Ranking Methodologies for Attributed Patent Citation Networks for Technology Opportunity Discovery. IEEE Transactions on Engineering Management, vol. 63, no. 4, pp. 426-437, Nov. 2016.

Shanie, T., Suprijadi, J., & Zulhanif. (2017, March). Text grouping in patent analysis using adaptive K-means clustering algorithm. In AIP Conference Proceedings (Vol. 1827, No. 1, p. 020041). AIP Publishing.

Sharma, A. (2012). A survey on different text clustering techniques for patent analysis. International Journal of Engineering, 1(9).

Tseng Y., Lin C. & Lin Y. (2007). Text mining techniques for patent analysis. Information Processing & Management Volume 43, Issue 5, September 2007, Pages 1216-1247.

Hu, X., Zhang, X., Lu, C., Park, E. K., & Zhou, X. (2009, June). Exploiting Wikipedia as external knowledge for document clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 389-396). ACM.

Ran, C., Shen, W., Wang, J., & Zhu, X. (2015, November). Domain-specific knowledge base enrichment using Wikipedia tables. In Data Mining (ICDM), 2015 IEEE International Conference on (pp. 349-358). IEEE.