Exploring TF-IDF encoding for comprehensive industry partners' selection

Karl Trela karl.trela@imw.fraunhofer.de
Fraunhofer IMW, Leipzig
Yuri Campbell yuri.cassio.campbell.borges@imw.fraunhofer.de
Fraunhofer IMW, Leipzig
Friedrich Dornbusch friedrich.dornbusch@imw.fraunhofer.de
Fraunhofer IMW, Leipzig

Abstract:

We propose a novel method for public research organizations in the search for industry partners, which differs from existing approaches with regard to its range. While most approaches are mine information only from companies with R&D and/or patent filling activities, our approach is able to consider a vast spectrum of potential partners by using widely available data, as text content and firm attributes from commercial databases. We show that the TF-IDF information encoding technique together with an indirect technological fit estimation can reliably identify promising industry partners from a wide range of companies. We test the performance of the approach using data on cooperations of over sixty research institutes belonging to a large research organization in Europe. Using the proposed indicators in a classifier model turns out to be a powerful tool for predicting cooperation activity, especially when combined with some economic indicators like turnover, age and number of managers hired by the company.

Keywords:

Partner selection, information retrieval, TF-IDF, university-industry cooperation, collaborative research, knowledge transfer and technology transfer

Introduction and aim:

Public research organizations (PROs) often face the problem of searching for industry partners for publicly funded projects and other knowledge transfer activities like licensing or contracted research. Merely the vast number of small and medium enterprises (SMEs) in many markets makes it very difficult to find the right match with regard to technological or scientific fit. In turn, the performance of such collaboration activities is known to be highly dependent on partner selection procedure (Li et al., 2008). Therefore, although arduous, it is a task of fundamental importance for public research organizations and enterprises alike.

Many recent studies helped to understand which determinants empirically define promising industry partners and proposed various approaches to identify them. The methods range from simple regressions (Cardamone and Pupo 2015; Giunta et al., 2016; Roigas et al., 2018), over index-based approaches (Geum et al.,2013) to soft computing methods, like Bayesian Network Analysis (Lee et al., 2016) and Semantic Analysis (Park et al., 2014; Huang et al., 2016). Most of these works however rely either on information about the firm's R&D activity or on patent and publications data. This is problematic for two reasons. First, the availability of R&D and patent data is dependent on the sector. For example, many information technology (IT) companies rarely file patents or consider their work classic R&D, but they still intensively cooperate with universities and research institutes. Second, the majority of small and medium sized companies, while being valuable potential collaboration partners for applied research, do not report their R&D activities and do not have the resources or capabilities to file patents. SMEs however represent the majority of companies and are both, politically and economically relevant. Moreover, they are precisely the economic actors that could profit the most from such transfer of knowledge, because on average they would rarely achieve such innovative technological activity independently. It is therefore clear that these approaches are not comprehensive enough, since they exclude many potential partners from certain sectors and also below certain size.

Our goal is to develop a method that is able to identify the right industry partners among a wide range of companies. In order to do so, we use sources of information that are abundant for all sectors and independent of size.

Research methodology:

We propose a comprehensive method for cooperation partner selection based on information retrieval built upon text analysis and encoding techniques. Different from existing approaches, we use data from readily available sources that cover all relevant entities in a nearly uniform way. Moreover, another key difference on our approach is that in order to make a match prediction we use an indirect approach. Instead of relying on direct estimation of technological compatibility, via experts' knowledge and/or quantitative portfolio fit, *e.g.* via bibliometric data, which is a technique widely used (Geum et al., 2013). In our approach, a match between a company and a PRO is estimated based on previous cooperation history of a PRO and the similarity from this company to the PRO's previous partners. This similarity measure relies on firm's business operations description obtained from a comprehensive database of companies. For this end, the operations description is encoded using TF-IDF technique. Subsequently, the firm's capability to participate or pay for contracted research is evaluated with the usual economic indicators, like size, turnover and number of employees, commonly available in commercial business information databases. As a proof-of-concept, we evaluate the performance of our method applied to the matching problem of one publicly funded research organization, which granted us access to the list of its previous cooperation partners. Together with usual economic indicators and with the estimated technological similarity among previous partners, a random sample of the same size is appended to this list. This forms a balanced dataset, which is used to train a classifier model. Such model in turn identifies good cooperation partners for each of the PRO's independent research units. The performance of our model on unseen data is evaluated in a 10-fold cross-validation fashion.

Results:

Our work not only draws attention to the broader issue of partner selection among public research organizations and SMEs, which are most in need of innovation assistance. It also provides the proof of concept of a new partner selection methodology, which is conceived to cover a considerably broader spectrum of companies. In detail, our preliminary results show that:

(1) TF-IDF information encoding technique proves to be a useful way to explore the structure of the underlying code context. In doing so, we apply TF-IDF beyond its original scope and analyze tokens with different relational structure. This extends the applicability of TF-IDF encoding beyond the mere analysis of text to other codes, as industry classification systems.

(2) This together with our indirect approach in order to model cooperation activity is an efficient predictor for cooperation compatibility, with mean ROC AUC score of 0.90 with standard deviation 0.06, among all analyzed research units.

(3) As expected, adding structured data on economic indicators from commercial databases further improves the overall classifier performance of the classifier.

Final Remarks:

The preliminary results show the viability of our technique in this broader scope as well as a promising performance of the classifier on unseen data. The scalability of the proposed method however needs further investigation, especially with concern of new and prominent low dimensional encoding techniques, as word2vec embedding (Mikolov et al., 2013).

References:

Cardamone, P., Pupo, V. and Ricotta, F. (2015) 'University Technology Transfer and Manufacturing Innovation: the Case of Italy.' Review of Policy Research, 32 (3): 297-322.

Geum, Y., Lee, S., Yoon, B. and Park, Y. (2013) 'Identifying and Evaluating Strategic Partners for Collaborative R&D: Index-based Approach Using Patents and Publications.' Technovation, 33 (6-7): 211-224.

Giunta, A., Pericoli, F. M., and Pierucci, E. (2016) 'University–Industry collaboration in the biopharmaceuticals: the Italian case.' The Journal of Technology Transfer, 41 (4): 818-840.

Lee, K., Park, I. and Yoon, B. (2016) 'An Approach for R&D Partner Selection in Alliances between Large Companies, and Small and Medium Enterprises (SMEs): Application of Bayesian Network and Patent Analysis.' Sustainability, 8 (2): 117.

Li, D., Eden, L., Hitt, M. A. and Duane, R. (2008) 'Friends, Acquaintances, or Strangers? Partner Selection in R&D Alliances.' The Academy of Management Journal, 51 (2): 315-334.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient Estimation of Word Representations in Vector Space.' arXiv:1301.3781.

Park, I., Jeong, Y., Byungun, Y. and Mortara, L. (2014) 'Exploring Potential R&D Collaboration Partners Through Patent Analysis Based on Bibliographic Coupling and Latent Semantic Analysis.' Technology Analysis & Strategic Management, 27 (7): 759-781.

Huang, L., Shang, L., Wang, K., Porter, A. L. and Zhang, Y. (2016) 'Identifying Targets for Technology Mergers and Acquisitions Using Patent Information and Semantic Analysis.' In: Daim T., Chiavetta D., Porter A., Saritas O. (eds) *Anticipating Future Innovation Pathways Through Large Data Analysis* (pp. 173-186). Springer, Cham.