A Patent Search Strategy based on Machine Learning for the Emerging Field of Service Robotics

Vladimir Korzinov[*], Florian Kreuchauff

[*]*vladimir.korzinov@kit.edu*
Karlsruhe Institute of Technology (Germany)

Background and Motivation

High technologies are in the core focus of supra-national innovation policies. For being effective and efficient, these policies strongly rely on credible databases that display entire value creation chains, starting from research and development up to production and sales. Regarding emerging technologies, for which the latter are still marginal, tracking early development efforts becomes important. However, since these are not yet part of any official industry, patent or trademark classification systems, delineating boundaries to measure this early stage is a nontrivial task. Service robotics (SR) is such a technology. Its applications spread through a multiplicity of services including medical assistance, fully automated construction, delivery, inspection, maintenance, as well as cleaning of public places or even home entertainment.

This paper is aimed to present a methodology to automatically classify patents as concerning service or industrial robotics (IR). We introduce a synergy of a traditional technology identification process like keyword extraction and verification by an expert community with a machine learning algorithm. The result is a novel possibility to allocate patents avoiding an erratic lexical query approach and reducing the dependency on iterative input from third parties which is usually costly and time consuming.

Methodology

We start with a patent search strategy with a vision to extrapolate it to other lexical sources. All unstructured patent text data as well as related document meta data were extracted from the 'EPO Worldwide Patent Statistical Database' (PATSTAT), version April 2013.

Firstly, we extracted all patents that were either sorted in international patent class (IPC) B25J or contained a substring like '*robot**' in their respective title or abstract.[1] Hence, we established a set of documents describing *robotic devices*. Secondly, two independent academic expert groups with some 15 scientists, affiliated with the

- High Performance Humanoid Technologies (H2T) from the Institute for Anthropomatics and Robotics at KIT, Germany, and the
- Delft Center for Systems and Control / Robotics Institute at TU Delft, Netherlands,

took on the task to decide which of the patents belonged to SR and which belonged complementarily to IR. The above experts were specialized in humanoid robotics, computer science, and mechanical engineering. Their experience in the field of robotics varied between 1 and 15 years. We provided them with 250 full body versions of potential SR patents from all over the world, extracted with the primal subsample queries. All patents listed in PATSTAT disclose at least English titles and abstracts.

Thirdly, based upon a core set of 98 worldwide SR patent applications identified by independent experts we transformed the unstructured patent document text into structured data. This included several steps, namely (1) combining titles and abstracts in one body and splitting the resulting strings into single terms in normal lower cases, (2) removing stop words, (3) stemming, i.e. reducing inflected words to their stem,

[1] According to the USPTO, most of the manipulators classified in B25J are industrial robots. See http://www.uspto.gov/web/patents/classification/cpc/html/defB25J.html

(4) constructing n-grams of term combinations (up to 3 words in one), (5) deriving normalized word and n-gram frequencies for each document, and (6) adding IPC dummy variables to indicate class belongings.

Fourthly, due to the fact that only a small number of key words and n-grams are shared by majority of the patents, some of them were considered insignificant and were excluded from the data for the purpose of improvement of classifier performance because they contained too little information and introduced noise. The resulting matrix included 1206 variables and 250 observations/patents. Finally all of them were scaled to the interval [0, 1] since SVM is sensitive to the absolute values.

Finally, based on the structured data and knowing which patent is SR or IR we implement a support vector machine (SVM) (Cortes & Vapnik, 1995) using a procedure of k-fold cross-validation during our training process. In order to eliminate negative influence of the unbalanced dataset we introduced weights in our SVM proportionate to SR and IR classes. We also vary kernel functions and their constants in addition to the cross-validation parameter. Since there is no possibility to determine in advance which function should be used, we have chosen to start with polynomial, sigmoid and radial basis functions, which was motivated by their popularity (Burges, 1998) and availability within our software package.

First results

We exhaustively searched through all possible combinations of kernel functions and their constants looking for the best f1-score of the model. Our final model showed an 85% precision and 83% recall. It contained a radial basis function kernel with γ equal to 0.005 and C equal to 10. Table 1 shows a detailed classification report on our test set.

Table 1. Classification report showing the performance of SVM classifier.

Class	**Precision**	**Recall**	**f1-score**
Service Robotic	75 %	94 %	83 %
Industrial Robotic	93 %	74 %	82 %
Avg. / total	85 %	83 %	83 %

Therefore, the resulting model is able to find on average 83% of service/industrial robotics patents and classify them correctly with average probability of 85%.

Conclusion

The lack of clear definitions has so far impeded a comprehensive assessment of the economic importance of SR. Now that we have a method to overcome this issue, we are in position to answer questions with respect to SR's contribution to aggregate value creation, innovation and economic structures.

With a novel methodology for detecting early developments of an emerging technology in patent data we are able to classify new robotics patents without an input from expert community limiting the typical lexical bias towards preferred subfields when working with experts. The procedure offers strong portability to other data sources. Currently the work on classification of robotics publications is in progress. The proposed method could be further enhanced by introducing better kernel functions or second stage with other classifiers like genetic algorithms, k-nearest neighbor, neural networks or boosting.

References

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery,* **2**(2), 121–167. http://dx.doi.org/10.1023/A:1009715923555

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20, 273-297. http://dx.doi.org/10.1023/A:1022627411411