Centre for Technology Management working paper series

ISSN 2058-8887

No. 1

January 2019

Intellectual Property Analytics Decision Support Tool (IPDST) for Early Stage Technology Decision Making

https://doi.org/10.17863/CAM.35544

Leonidas Aristodemou (CTM University of Cambridge, The Alan Turing Institute) *
Frank Tietze (CTM, University of Cambridge)
Alexandra Brintrup (CTM, University of Cambridge)
Sam Deeble (CTM, University of Cambridge

* Please contact the corresponding author for feedback:
la324@cam.ac.uk ; laristodemou@turing.ac.uk

An earlier version of this paper has been presented to the R&D Conference 2018, in Milan, Italy

UNIVERSITY OF CAMBRIDGE

IfM Centre for Technology Management

An earlier version of this paper was presented in:
R&D Management Conference 2018 *"R&D Designing Innovation: Transformational Challenges for Organizations and Society"*,
Trach 18: Big Data Analytics for R&D Management, June, 30th -July, 4th, 2018, Milan, Italy

Intellectual Property Analytics Decision Support Tool (IPDST) for Early Stage Technology Decision Making

Leonidas Aristodemou[1], Frank Tietze[1], Alexandra Brintrup[2], Sam Deeble[1]

[1] Centre for Technology Management (CTM), Institute for Manufacturing (IfM), Department of Engineering, University of Cambridge

[2] Distributed Information and Automation Lab (DIAL), Institute for Manufacturing (IfM), Department of Engineering, University of Cambridge

Increased data availability presents an opportunity for better decision making, to introduce the next generation of innovative and disruptive technologies. In this paper, we aim to complement early stage technology strategic decision making with Intellectual Property Analytics (IPA). We follow a system design approach, where we propose an intellectual property analytics decision support tool (IPDST), which makes use of machine learning to analyse patent data, predicting technological impact. Firstly, we extract and operationalise features from 274,609 randomly selected patents from the United States Patent and Trademark Office (USPTO). Secondly, we employ a multi-layer perceptron artificial neural network to capture the nonlinear relationships between the input features and the output feature of number of citations. We assess the performance of the model with 61% accuracy and propose future research directions.

Keywords: **Intellectual Property Analytics, Machine Learning, Deep Learning, Early stage technology, Strategic Decision Making,**

1. Introduction

Forecasting high technological impact technologies are of great interest to a wide range of stakeholders especially at the early stage of technology projects. The existing literature has shown that patent citation information is useful for measuring the economic and technological value and impact of a technology (Carpenter, Narin and Woolf, 1981; Narin, Albert and Smith, 1992; Narin, 2006; Leonidas Aristodemou and Tietze, 2018).

We propose a machine learning approach to forecast the technological impact of an early stage technology project, using multiple patent indicators that can be defined immediately after the relevant patents have been drafted. Economic and innovation literature has presented a wide range of patent indicators that may be indicative of the future citation count of patents and that further the relevant technology's economic impact (Narin, 2006; Squicciarini, Dernis and Criscuolo, 2013; Leonidas Aristodemou and Tietze, 2018). This approach is mainly based on Intellectual Property Analytics, which is the data science of analysing large amount of intellectual property information, to discover relationships, trends and patterns in the data for decision making. It is a multidisciplinary approach that makes use of mathematics, statistics, computer programming, and operations research to gain valuable knowledge from data, to support decision making rooted in the business context (L. Aristodemou and Tietze, 2018).

Firstly, a dataset of 274,609 sampled patent are extracted from the United States Patent and Trademark Office (USPTO) database. Secondly, a multilayer perceptron (MLP) neural network, is deployed to capture the complex nonlinear relationships between six input and one output feature in a time period of interest. Our proposed method provides a proof of concept and further work is required to build a full comprehensive intellectual property decision support tool (IPDST), which will complement strategic decision making for early stage technology projects (Aristodemou and Tietze, 2017; Aristodemou *et al.*, 2017). The paper is organised as follows: section 2 presents the methodology, section 3 presents the results, with section 4 concluding the paper and outlining next steps.

2. Methodology

We follow a system design methodology: firstly, data is sourced from the United States Patents and Trademark Office (USPTO). The dataset is sampled and stored locally with the same structure as the original datasets, such that all further processes are compatible with the full dataset. Collation is performed in Python to produce a single table containing all 274,609 sampled patents and associated information. Secondly, we deploy a multi-layer perceptron (MLP) using the Keras API, with classification, to predict technological impact (Basheer and Hajmeer, 2000; Schmidhuber, 2015).

2.1 Measures of technological impact

The use of patent quality or technological value indicators has been the subject of research for a number of years. There are a number of indicators proposed in the literature to assess patent quality and subsequently technological value (Harhoff *et al.*, 2007; Squicciarini, Dernis and Criscuolo, 2013). One such indicator is the number of forward citations, as dominant theory suggests that the number of citations a patent receives is correlated to the technological and commercial importance of that patent, and the invention described therein (Leonidas Aristodemou and Tietze, 2018).

2.2 Feature extraction from text

Word embeddings is a method by which words are transformed to vectors using their semantic meaning. Google offer an advancement on traditional methods by using the word2vec (Mikolov *et al.*, 2013). This allows words with similar meanings to be both clustered together and share relationship. Word embeddings can be applied to larger corpora of text by taking the centroids of the individual vectors to generate a document embedding (Le and Mikolov, 2014). Advancements on this method include combining these embeddings with traditional frequency information such as TF-IDF to improve the representation of short texts (De Boom *et al.*, 2016), or adopting a paragraph vector within an embedding model (Dai, Olah and Le, 2015).

2.3 Data extraction

The USPTO provides patent data via Google BigQuery. The dataset is constrained to utility patents and grant dates between 1997 and 2016 inclusive (Marco, Sarnoff and DeGrazia, 2016; Marco and Tesfayesus, 2017). Of the 3,920,108 patents identified, 274,609 are randomly sampled to provide a representative dataset. The percentage of patents against grant date and IPC section are plotted in the full dataset and the sampled dataset to demonstrate the appropriateness of random sampling, which is shown in Figure 1. Once sampled, the datasets

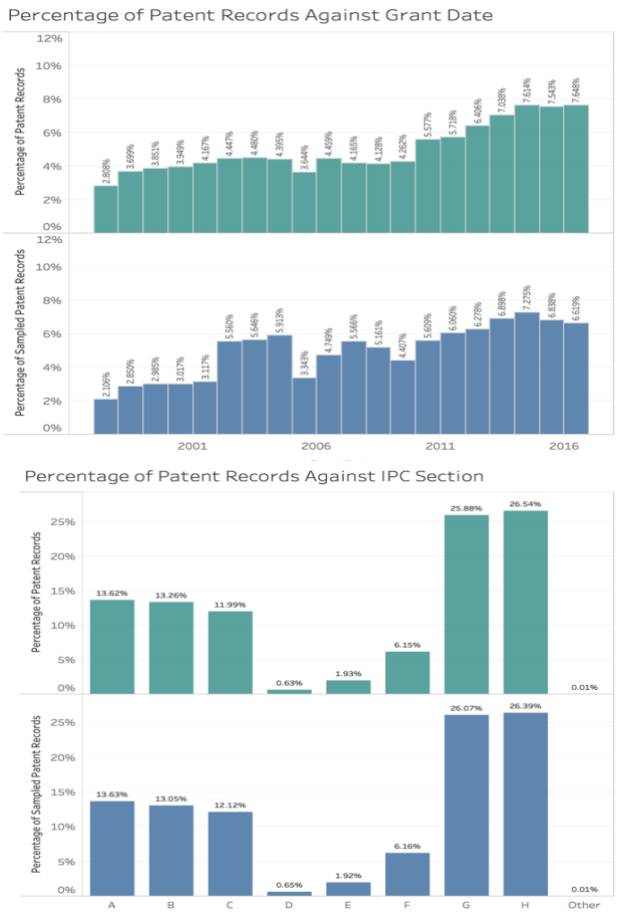

Figure 1 Full dataset (green) vs. sampled dataset (blue)

were combined into a single data frame with one row per patent.

2.4 Operationalisation of features

Feature selection and operationalisation is the process of selecting features and translating them as inputs to a machine learning model. The input and output features in this paper have been selected through a feature selection method for continuous and discrete features, using a mutual information test; and a variance-to-mean ratio test (Manning, Raghavan and Schütze, 2008). An example is shown in **Error! Reference source not found.**, but for the purpose of simplicity, this method is not discussed. The input features deployed in this study are the patent abstract, number of dependent claims, primary IPC classification, number of backwards citations, originality index (based on IPC) and radicalness index (based on IPC). The output feature is the number of forward citations within 4 years since patent publication.

Table 1 Feature selection analysis

Output Feature	Continuous or Discrete Input Feature	Complete Data Points	Mutual Information	Variance/ Mean
citation_t4	num_independent_claims	189128	0.0199	0.1516
	num_dependent_claims	189129	0.0183	0.1516
	num_invention_ipc_classes	93966	0.0064	0.1392
	num_additional_ipc_classes	93966	0	0.1392
	num_us_back_citations	170524	0.0170	0.1514
	originality_index	169186	0.0149	0.1519
	radicalness_index	159414	0.0154	0.1535

We make use of the word2vec model to transform the patent abstract into a vectorised numerical form (Mikolov *et al.*, 2013; Le and Mikolov, 2014). The advantage of word2vec is twofold: firstly, words with similar meanings are mapped to similar vectors; and secondly, the vector distance between word pairs represents relationships which are transferable to other word pairs. Stop words (which are commonly removed from text during natural language processing such as 'the', 'and' and 'it) are not taken out of the list and words are only excluded if not included in the vocabulary of the model (Mikolov *et al.*, 2013; Le and Mikolov, 2014; Jurafsky and Martin, 2016).

Categorical features are represented using a one-hot encoding scheme. The number of forward citations is a discrete numerical feature and is transformed into categorical features by 'binning' into four classes 'L1', 'L2', 'L3', 'L4'. The classes are ordered such that 'L1' indicates a value indicator of the highest category while 'L4' indicates the lowest category as shown in Table 2 (Lee *et al.*, 2018). Standardisation is the process of scaling each input feature such that it has a mean of zero and standard deviation of one. The purpose of this process is to ensure the cost function is equal in gradient in all directions and reduce the variation in space so as to improve the rate of convergence (LeCun *et al.*, 2012).

Table 2 Binning scheme for categorisation of output feature

Output Feature	**Binning Scheme**			
	L4	**L3**	**L2**	**L1**
Forward Citations (t4)	0-1	2-9	10-19	20+

2.5 Model architecture

Machine learning models are constructed using the Keras API, with the main architecture implemented is the multi-layer perceptron (Schmidhuber, 2015). The input features are standardised to a unit variance and zero mean, which allows for initialising the network weights around zero. Intuitively, the model learns the simpler linear relationships before the more difficult non-linear ones. We employ a grid-search method, which is an extensive search for the best parameters given a range of possibilities. The grid search creates a model for each combination of hidden layer depth and epochs, checks the model performance (measured with accuracy) using k-fold cross validation, and returns the best parameters (Zhang,

Patuwo and Hu, 1998). Assessing the technological impact of patents is equivalent to classifying all the patents into four classes according to the expected number of forward citations of patents. We make use of the accuracy measure to evaluate our model, which is defined as the number of correct prediction over the total number of predictions, using 90% of the dataset for training and 10% for testing (Kim and Lee, 2017; Lee *et al.*, 2018).

3. Results

The grid is applied to optimise the structure and training time of the MLP with a single hidden layer. The best performing model was found to have a hidden layer depth of 10 nodes and trained for 200 epochs.

Figure 4 shows a cross-tabulation of the primary IPC section against number of citations after four years. The cross-tabulation is colour coded with the percentage of samples of each IPC section and citation category (allowing comparison between categories with differing numbers of patents). Absolute values are included numerically in the cells. The figure indicates that sections C and D have the lowest number of average citations, and so the primary IPC class provides partial information for the classification problem.

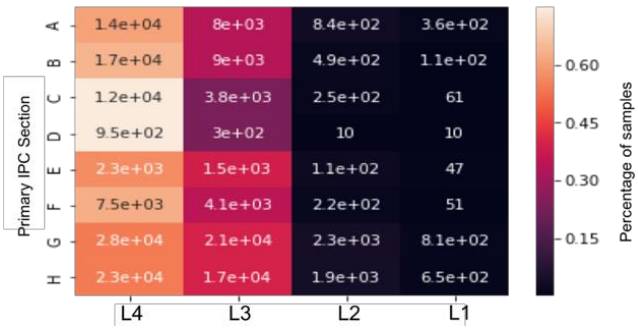

Figure 4 Cross tabulation of the primary IPC section against number of citation after four years

To visualise the patent abstracts, Figure 5 shows a t-SNE plot of the document vector, with colour coding given by the citation categories. This plot shows a minor clustering of L3 citations in the upper left corner however the low dimensional plot fails to capture any further, more complex relationships

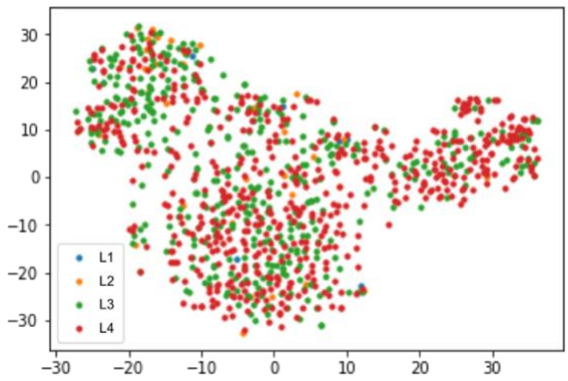

Figure 4 t-SNE Plot: sampled abstract embeddings, coded by number of citations within four years

Figure 6 shows the accuracy of the training set (blue) and the validation set (orange) against the number of epoch for this model. The model reaches a validation accuracy of over 61%, at around 40 to 60 epochs, however the training set accuracy and validation set accuracy begin to slightly diverge as training continues. This demonstrates the advantage of plotting the two accuracies together as it can be seen that it is likely the model has begun to slightly over-fit the training features. To reduce the risk of overfitting, we use dropout and regularisation.

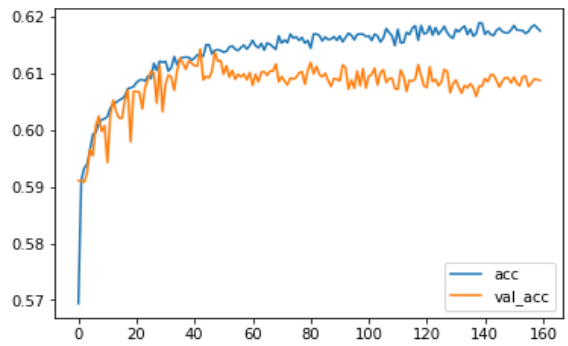

Figure 4 Accuracy plot of the model (blue line = training set; orange line = testing set)

4. Conclusion

In this paper, we complement early stage technology strategic decision making with intellectual property analytics, where we predict technological impact, defined as the number of forward citations in the fourth year since publication, at the technology development process. We follow a system design approach, where we propose an intellectual property analytics decision support (IPDST) methodology. The model has an accuracy of 61%.

There are a number of limitations with this research, which were mainly due to computing capability issues. However, this paper provides a good overview and proof of concept of using machine learning models in predicting technological impact and value. We intend to expand the number of input and output features for the model and assess a variety of models in predicting technological value.

Acknowledgements

The authors would like to thank the Engineering and Physical Sciences Research Council (EPSRC) for funding this research, under the EPSRC DTP award. In addition, we would like to thank Dr. Robert Phaal, and the Strategic Technology and Innovation Management (STIM) consortium 2018.

References

Aristodemou, L. *et al.* (2017) 'Exploring the Future of Patent Analytics: A Technology Roadmapping Approach', in *R&D Management Conference 2017, Leuven, Belgium*, pp. 1–9. doi: 10.17863/CAM.13967.
Aristodemou, L. and Tietze, F. (2017) *Exploring the Future of Patent Analytics*. Cambridge, UK.
Aristodemou, L. and Tietze, F. (2018) 'Citations as a measure of technological impact: A review of forward citation-based measures', *World Patent Information*. doi: 10.1016/j.wpi.2018.05.001.
Aristodemou, L. and Tietze, F. (2018) 'The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data', *World Patent Information*, 55. doi: 10.1016/j.wpi.2018.07.002.
Basheer, I. . and Hajmeer, M. (2000) 'Artificial neural networks: fundamentals, computing, design, and application', *Journal of Microbiological Methods*, 43, pp. 3–31. doi: 10.1016/S0167-7012(00)00201-3.
De Boom, C. *et al.* (2016) 'Representation learning for very short texts using weighted word embedding aggregation', pp. 1–8. doi: 10.1016/j.patrec.2016.06.012.
Carpenter, M. P., Narin, F. and Woolf, P. (1981) 'Citation rates to technologically important patents', *World Patent Information*, 3(4), pp. 160–163. doi: 10.1016/0172-2190(81)90098-3.
Dai, A. M., Olah, C. and Le, Q. V. (2015) 'Document Embedding with Paragraph Vectors', pp. 1–8. Available at: http://arxiv.org/abs/1507.07998.
Harhoff, D. *et al.* (2007) *The strategic use of patents and its implications for enterprise and competition policies*, *European Commission Ref. Ares(2014)78204 - 15/01/2014*.
Jurafsky, D. and Martin, J. H. (2016) 'Semantics with Dense Vectors', *Speech and Language Processing, 3rd edition*.
Kim, J. and Lee, S. (2017) 'Forecasting and identifying multi-technology convergence based on patent data: the case of IT and BT industries in 2020', *Scientometrics*. Springer Netherlands, 111(1), pp. 47–65. doi: 10.1007/s11192-017-2275-4.
Le, Q. V. and Mikolov, T. (2014) 'Distributed Representations of Sentences and Documents', 32. doi: 10.1145/2740908.2742760.
LeCun, Y. A. *et al.* (2012) 'Efficient backprop', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7700 LECTU, pp. 9–48. doi: 10.1007/978-3-642-35289-8-3.
Lee, C. *et al.* (2018) 'Early identification of emerging technologies: A machine learning approach using multiple patent indicators', *Technological Forecasting and Social Change*. Elsevier, 127(April 2017), pp. 291–303. doi: 10.1016/j.techfore.2017.10.002.
Manning, C. D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval Introduction*, *Computational Linguistics*. doi: 10.1162/coli.2009.35.2.307.
Marco, A. C., Sarnoff, J. D. and DeGrazia, C. A. (2016) 'Patent Claims and Patent Scope', p. 53. doi: 10.2139/ssrn.2844964.
Marco, A. C. and Tesfayesus, A. (2017) 'Patent Litigation Data from US District Court Electronic Records (1963-2015)', *SSRN Electronic Journal*, pp. 1–40. doi: 10.2139/ssrn.2942295.
Mikolov, T. *et al.* (2013) 'Efficient Estimation of Word Representations in Vector Space', pp. 1–12. doi: 10.1162/153244303322533223.
Narin, F. (2006) 'Assessing Technological Competencies', in *From Knowledge Management to Strategic Competence*. doi: doi:10.1142/9781860948138_0008.
Narin, F., Albert, M. B. and Smith, V. M. (1992) 'Technology indicators in strategic planning', *Science and Public Policy*, 19(6), pp. 369–381. doi: 10.1093/spp/19.6.369.
Schmidhuber, J. (2015) 'Deep Learning in neural networks: An overview', *Neural Networks*, 61, pp. 85–117. doi: 10.1016/j.neunet.2014.09.003.
Squicciarini, M., Dernis, H. and Criscuolo, C. (2013) 'Measuring Patent Quality: Indicators of Technological and Economic Value', *OECD Science, Technology and Industry Working Papers*, (03), p. 70. doi: 10.1787/5k4522wkw1r8-en.
Zhang, G., Patuwo, E. B. and Hu, M. Y. (1998) 'Forecasting with artificial neural networks: the state of the art',