Santa Clara High Technology Law Journal

Volume 36 | Issue 4 Article 3

5-13-2020

THE PROMISE OF MACHINE LEARNING FOR PATENT LANDSCAPING

Toole, Andrew A.

Pairolero, Nicholas A.

Forman, James Q.

Giczy, Alexander V.

Recommended Citation

Toole, Andrew A.; Pairolero, Nicholas A.; Forman, James Q.; and Giczy, Alexander V., *THE PROMISE OF MACHINE LEARNING FOR PATENT LANDSCAPING*, 36 Santa Clara High Tech. L.J. 433 ().
Available at: https://digitalcommons.law.scu.edu/chtlj/vol36/iss4/3

THE PROMISE OF MACHINE LEARNING FOR PATENT LANDSCAPING

Andrew A. Toole
Chief Economist, U.S. Patent and Trademark Office

Nicholas A. Pairolero
Economist, U.S. Patent and Trademark Office

James Q. Forman
Data Scientist, Google LLC[1]

Alexander V. Giczy
Data Scientist, U.S. Patent and Trademark Office (Addx Corporation)

ABSTRACT

Patent landscaping involves the identification of patents in a specific technology area to understand the business, economic, and policy implications of technological change. Traditionally, patent landscapes were constructed using keyword and classification queries, a labor-intensive process that produced results limited to the scope of the query. In this paper, we discuss the advantages and disadvantages of using machine learning to produce patent landscapes. Machine learning leverages traditional queries to construct the data necessary to train machine learning models, and the models allow the resultant landscapes to extend more broadly into areas of technology not expected a priori. The models, however, are "black boxes" that limit transparency regarding their underlying reasoning. To illustrate these points, we summarize two landscapes we recently conducted, one in mineral mining and another in artificial intelligence.

DISCLAIMER: The views expressed are those of the individual authors and do not necessarily reflect the official positions of the Office of the Chief Economist or the U.S. Patent and Trademark Office.

[1] Formerly at the U.S. Patent and Trademark Office (Addx Corporation).

ACKNOWLEDGEMENTS: We wish to thank Kakali Chaki, Jesse Frumkin, Christian Hannon, Anne Thomas Homescu, David Landrith, Steve Melnick, David B. Orange, Christyann Pulliam, and Matthew Such, current and former employees at the U.S. Patent and Trademark Office for their assistance regarding the machine learning approach used in the Artificial Intelligence ("AI") Landscape.

INTRODUCTION

Patent landscaping identifies patents in a specific technology area to understand the business, economic, and policy implications of technological change. It has traditionally been a time consuming and complex process relying on the careful construction of queries to identify relevant patents (Trippe 2015; Abood and Feltenberger 2018). Recent machine learning advances promise to reduce these costs by automating landscaping while providing scalability and accuracy (Abood and Feltenberger 2018). This paper provides an overview of how machine learning overcomes shortcomings of traditional approaches and clarifies these points by describing two studies conducted by the U.S. Patent and Trademark Office (USPTO).

I. TRADITIONAL APPROACH

Several traditional methods exist to search for patents: (1) keywords against patent text, (2) classification classes, and (3) citations. These queries may be narrow or broad, and allow for precise control over results. This leads to high transparency in the resulting landscape. There are, however, several shortcomings. Queries may become very complex with keyword synonyms explicitly stated. Since words and concepts change over time (e.g., "horseless carriages" are now "automobiles"), a specific query may become less effective over time. Word context matters (e.g., oil "extraction" versus dental "extraction"), and the applicant may be their own lexicographer.[2] Patent classification schema are dynamic: classes are created for new technologies or to reduce the scope of large existing classes. Finally, citations are subject to truncation error and may be influenced by many factors (Lerner and Seru 2017)[3]. All these considerations lead to increasingly complex queries. Table 1 displays a query from a WIPO (2019) landscape to illustrate.

[2] U.S. PATENT AND TRADEMARK OFFICE MANUAL OF PATENT EXAMINATION PROCEDURE (MPEP) § 2111.01 IV (2018).

[3] *See also* Kunh et al., *Patent Citations Reexamined* (2019).

Table 1 - Sample Text Query for Artificial Intelligence

(((ARTIFIC+ OR COMPUTATION+) 1W INTELLIGEN+) OR (NEURAL 1W NETWORK+) OR NEURAL_NETWORK+ OR NEURAL_NETWORK+ OR (BAYES+ 1W NETWORK+) OR BAYESIAN-NETWORK+ OR BAYESIAN_NETWORK+ OR (CHATBOT?) OR (DATA 1W MINING+) OR (DECISION 1W MODEL?) OR (DEEP 1W LEARNING+) OR DEEP-LEARNING+ OR DEEP_LEARNING+ OR (GENETIC 1W ALGORITHM?) OR ((INDUCTIVE 1W LOGIC) 1D PROGRAMM+) OR (MACHINE 1W LEARNING+) OR MACHINE_LEARNING+ OR MACHINE-LEARNING+ OR ((NATURAL 1D LANGUAGE) 1W (GENERATION OR PROCESSING)) OR (REINFORCEMENT 1W LEARNING) OR (SUPERVISED 1W (LEARNING+ OR TRAINING)) OR SUPERVISED-LEARNING+ OR SUPERVISED_LEARNING+ OR (SWARM 1W INTELLIGEN+) OR SWARM-INTELLIGEN+ OR SWARM_INTELLIGEN+ OR (UNSUPERVISED 1W (LEARNING+ OR TRAINING)) OR UNSUPERVISED-LEARNING+ OR UNSUPERVISED_LEARNING+ OR (SEMI-SUPERVISED 1W (LEARNING+ OR TRAINING)) OR SEMI-SUPERVISED-LEARNING OR SEMI_SUPERVISED_LEARNING+OR CONNECTIONIS# OR (EXPERT 1W SYSTEM?) OR (FUZZY 1W LOGIC?) OR TRANSFER-LEARNING OR TRANSFER_LEARNING OR (TRANSFER 1W LEARNING) OR (LEARNING 3W ALGORITHM?) OR (LEARNING 1W MODEL?) OR (SUPPORT VECTOR MACHINE?) OR (RANDOM FOREST?) OR (DECISION TREE?) OR (GRADIENT TREE BOOSTING) OR (XGBOOST) OR ADABOOST OR RANKBOOST OR (LOGISTIC REGRESSION) OR (STOCHASTIC GRADIENT DESCENT) OR (MULTILAYER PERCEPTRON?) OR (LATENT SEMANTIC ANALYSIS) OR (LATENT DIRICHLET ALLOCATION) OR (MULTI-AGENT SYSTEM?) OR (HIDDEN MARKOV MODEL?))/BI/OBJ/CLM

Source: WIPO Technology Trends 2019 Artificial Intelligence, Data collection and method and clustering scheme: Background paper, 23.

This approach is essentially trial and error – defining a query, examining results, refining the query – and may become very time consuming. In the end, the results mirror *a priori* expectations about where the technology is and what language is used to describe it.

II. MACHINE LEARNING APPROACH

Patent landscaping is a classification problem: does a patent document belong in the technology of interest or not? Models classify patent documents by learning from a set of pre-classified documents belonging to the technology of interest (the "seed" set) and not (the "anti-seed" set). Traditional queries build the seed set; the anti-seed set is trickier. Abood and Feltenberger (2018) solve this problem by expanding from the seed set using families, citations, or classifications, and randomly sampling outside this expansion (presumed unlikely to contain the technology of interest) for the anti-seed set.[4] Several models may be used, e.g., support vector machines

[4] Aaron Abood & Dave Feltenberger, *Automated patent landscaping*, ARTIF. INTELL. L., 103, 109-114 (2018).

(SVM) and neural networks (Abood and Feltenberger 2018; Alderucci 2019). Inputs commonly include patent text (or a subset thereof) and may be augmented by classification and citations. Text must be encoded.[5] Model output is typically the probability that a given document is in the technology of interest.

One advantage of this approach is the results are not constrained to the seed queries, enabling the landscape to better capture diffusion across technology. However, the seed and anti-seed must be representative, with the seed set covering all significant aspects of the target technology or the model will not detect these aspects, and borderline cases (i.e. patents that are more challenging to classify) should be included in training. One disadvantage is a lack of transparency, particularly with more complex models. Finally, if traditional approaches are overly narrow then machine learning runs the risk being overly broad, classifying documents *a posteriori* for reasons that are not entirely clear.

III. Examples

A. Mineral Mining

This project explored the safety and health impact of U.S. mineral mining patents (Toole et al. 2019). A mineral mining patent landscape was a necessary starting point. After receiving a dataset of 92,000 patents generated using a set of queries, it became evident the dataset contained non-relevant documents; e.g., data mining and landmines. Manual filtering was impractical, so we employed a machine learning approach. For the seed set we matched patent assignees to known mining companies and extracted their patents, and for the anti-seed set to known oil/gas and non-mining companies.[6] We selected an SVM model. Only 50% of the original 92,000 patents were classified as relevant to mineral mining. We further used queries and a neural network to identify safety and health-related patents. Machine learning, in combination with traditional query methods, allowed us to complete our analysis with a high degree of confidence.

[5] Text encoding or “embedding” runs from “bag of word” counts to algorithms capturing word context, e.g., *word2vec* (Mikolov et al., *Distributed Representations of Words and Phrases and their Compositionality* (2013)) and BERT (Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (2019)). Wikipedia, “Bag-of-words model”, https://en.wikipedia.org/wiki/Bag-of-words_model (last accessed March 2, 2020).

[6] Known non-mining companies included assignees for some of the questionable patents mentioned above, such as data mining.

B. Artificial Intelligence

In the second project, we developed a patent landscape for U.S. AI patents[7] using the approach of Abood and Feltenberger (2018). Since a consensus definition of AI does not exist (Russell and Norvig 2009), we defined eight AI categories (Figure 1).

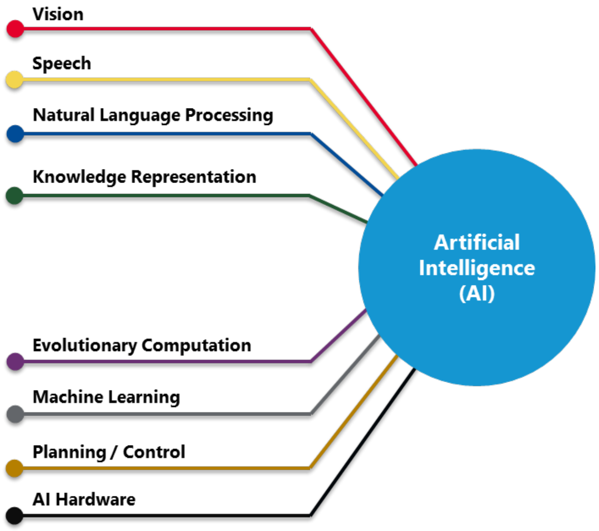


Figure 1 - USPTO Artificial Intelligence Patent Landscape AI Categories

Source: USPTO analysis

We trained a neural network for each category, with seed sets drawn from narrowly defined traditional search queries.[8] The models included patent abstract text, claims text, and citations as inputs. This analysis resulted in 1.3M of 11.7M patent documents (10.8%) categorized in at least one of the eight AI categories.

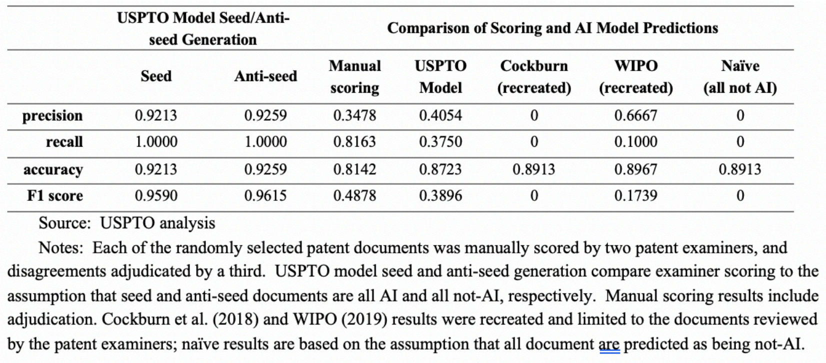
Additionally, we manually scored 800 randomly selected documents *ex post facto* using experienced patent examiners, enabling us to review and compare results across methodologies (Table 2). The review shows that our seed and anti-seed sets were not perfect, although this may be due to interpretation differences across examiners, highlighting difficulties in defining AI. Of the different methods, the evaluation examiners outperform

[7] A USPTO Office of the Chief Economist IP Data Highlights report is anticipated mid-2020; *see* www.uspto.gov/economics.

[8] The AI categories are not mutually exclusive, so a single document may be in several models.

based on F1 scores,[9] and accuracy is comparable across all.[10] Interestingly, the traditional approach used in Cockburn et al. (2018) did not identify any AI documents in our sample, illustrating the limitations of overly narrow queries). Our neural network model achieved higher recall than WIPO's (2019) traditional query approach, and our higher F1 score indicates our method did not adversely sacrifice precision.

Table 2 - AI Landscape Model Comparisons



	USPTO Model Seed/Anti-seed Generation		Comparison of Scoring and AI Model Predictions				
	Seed	**Anti-seed**	**Manual scoring**	**USPTO Model**	**Cockburn (recreated)**	**WIPO (recreated)**	**Naïve (all not AI)**
precision	0.9213	0.9259	0.3478	0.4054	0	0.6667	0
recall	1.0000	1.0000	0.8163	0.3750	0	0.1000	0
accuracy	0.9213	0.9259	0.8142	0.8723	0.8913	0.8967	0.8913
F1 score	0.9590	0.9615	0.4878	0.3896	0	0.1739	0

Source: USPTO analysis

Notes: Each of the randomly selected patent documents was manually scored by two patent examiners, and disagreements adjudicated by a third. USPTO model seed and anti-seed generation compare examiner scoring to the assumption that seed and anti-seed documents are all AI and all not-AI, respectively. Manual scoring results include adjudication. Cockburn et al. (2018) and WIPO (2019) results were recreated and limited to the documents reviewed by the patent examiners; naïve results are based on the assumption that all document are predicted as being not-AI.

CONCLUSION

Both traditional queries and machine learning are beneficial in patent landscaping. In our mineral mining study, the query returned results that were too broad, and we pruned this set down by using machine learning. In our AI study, we used a narrow query to build training data (seed and anti-seed sets). The machine learning classifier then accurately identified a landscape beyond patents obtained through traditional approaches. Seed and anti-seed generation is crucial to machine learning, as is rigorous evaluation. Manual review outperforms any traditional or machine learning approach but is too costly to scale to large document sets. The promise of machine learning is not to replace traditional query approaches but to allow the landscape to extend beyond preconceived notions of where, and what constitutes the technology. This greater representation allows for better decision-making by business leaders and policy-makers.

[9] The F1 score combines recall and precision. Recall, or the true positive rate, measures the likelihood the model predicts a positive when then the document is positive. Precision, or the positive predictive value, is the likelihood of the document being positive when the model predicts it to be positive. Accuracy includes true positives and true negatives. (See Wikipedia, "F1 score," https://en.wikipedia.org/wiki/F1_score).

[10] Accuracy is less relevant here than recall, precision, and F1 since we are trying to identify a rare class. In fact, one can do well with accuracy by guessing that all documents are not AI.