Automated Patent Landscaping

Aaron Abood
Google, Inc.
aabood@google.com

Dave Feltenberger
Google, Inc.
feltenberger@google.com

ABSTRACT

Patent landscaping is the process of finding patents related to a particular topic. It is important for companies, investors, governments, and academics seeking to gauge innovation and assess risk. However, there is no broadly recognized best approach to landscaping. Frequently, patent landscaping is a bespoke human-driven process that relies heavily on complex queries over bibliographic patent databases.

In this paper, we present Automated Patent Landscaping, an approach that jointly leverages human domain expertise, heuristics based on patent metadata, and machine learning to generate high-quality patent landscapes with minimal effort.

In particular, this paper describes a flexible automated methodology to construct a patent landscape for a topic based on an initial seed set of patents. This approach takes human-selected seed patents that are representative of a topic, such as *operating systems*, and uses structure inherent in patent data such as references and class codes to "expand" the seed set to a set of "probably-related" patents and anti-seed "probably-unrelated" patents. The expanded set of patents is then pruned with a semi-supervised machine learning model trained on seed and anti-seed patents. This removes patents from the expanded set that are unrelated to the topic and ensures a comprehensive and accurate landscape.

CCS Concepts

• **Computing / technology policy → Intellectual Property → Patents • Machine learning → Learning paradigms → Supervised learning → Supervised learning by classification • Machine learning → Machine learning approaches → Neural networks • Machine learning → Machine learning approaches → Learning linear models → Perceptron model • Applied computing → Law, social and behavioral sciences → Law • Applied computing → Document management and text processing → Document metadata.**

Keywords

Patent landscape; classification; text analytics; semi-supervised machine learning

1. INTRODUCTION

At the height of the smartphone wars, it was rumored that the smartphone was covered by 250,000 patents [13]. Or was it 314,390 [14]? Which is right, and how does one even arrive at such numbers? This also leads to follow on questions such as: who owns the patents? when were the patents filed? where are the inventors from? what fraction of the patents have been litigated?

These questions are often answered by a technique called patent landscaping. Patent landscaping is important to a number of audiences, including (1) companies that desire to assess risk posed by other patent holders and understand their own relative strength, (2) academics and governments that seek to gauge the level of R&D investment and innovation in particular fields [18], and (3) investors looking to value companies and assess risk [3].

Patent landscaping is the exercise of identifying all the patents that are relevant to a topic. This is discussed in detail in the guidelines prepared by the World Intellectual Property Office (WIPO) [18]. It is a challenging task that can involve substantial time and expense. In particular, patent landscaping is made difficult by issues such as: over/under inclusion, needing to fully understand the associated technical/product landscape, scalability/data limitations of commercial tools, and proper search methodology. Building patent landscapes is often seen as more of an art than a science with analysts constructing elaborate queries over bibliographic patent databases and assembling multiple lists of patents for inclusion in the resulting landscape. This can make reproducing and updating landscapes difficult. Further, methodologies may be inconsistent or difficult to explain.

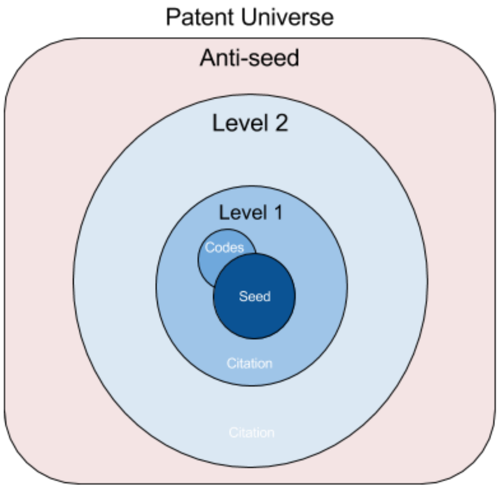

Figure 1 – Expansion levels

This paper describes a semi-supervised machine learning approach for automated patent landscaping. The approach starts with a small human curated seed set of patents narrowly related to the topic of interest. The seed set is expanded by citations (forward and backward) and class codes to identify candidate patents for inclusion in the landscape (Levels 1 and 2 in Figure 1). Patents not included in the expansion are referred to as the anti-seed. A machine learning model is then applied to prune the candidate patents to create a subset of the candidate patents that are relevant to the topic. The model is trained using the seed (positive examples) and a sample of patents from the anti-seed (negative examples). The final landscape is the pruned subset of candidate patents and the seed set. Alternatively, the model could be applied to the full patent corpus to increase coverage missed by the expansion to Levels 1 and 2.

2. PATENT DATA

A basic understanding of patent data and its traditional application to patent landscaping is useful for understanding the automated approach described in this paper. Those who are familiar with patent data can skip to the next section, AUTOMATED LANDSCAPING.

Patent documents and their associated data are somewhat unique in that they include multiple text segments as well as a significant array of structured metadata. Such metadata includes class codes, citations, inventors, assignees, and family relationships. The following highlights some of the key types of patent data as they relate to patent landscaping. Note that each type has significant limitations when used in patent landscaping.

2.1 Text

Patents contain multiple text segments, such as a title, abstract, and detailed description (body of the patent). Additionally, patents include claims. The claims define the scope of the property right and must be supported by the detailed description.

A common traditional technique for constructing landscapes is to perform boolean keyword searches on some or all of the text segments. However, this can be difficult when words have multiple meanings, when multiple words can describe the same subject, or simply because of spelling variations. Further, constructing such queries can often involve constructing long boolean queries with multi-word phrases and require that the user understand all aspects of the topic.

2.2 Class Codes

Class codes are topical labels applied by a patent office to classify patents. The primary class code regimes are the US class codes (exclusive to the USPTO) and the Cooperative Patent Class (CPC) codes (applied worldwide). Both are large hierarchical taxonomies with thousands of nested labels. Patents are typically assigned to multiple class codes.

A common technique for constructing landscapes is to select all the patents from one or more class codes. However, identifying the desired class codes can be difficult and the underlying patents sometimes deviate from the description of the class codes. Additionally, the codes, either alone or in combination, may not line up well with the topic for which the user wants to construct a landscape. Even if they do, selecting the right combinations of codes requires the user to have a nuanced understanding of the topic.

2.3 Citations

Like scholarly publications, many patents have citations. During prosecution (the process for obtaining a patent), the examiner as well as the applicant may cite publications that are relevant to the patent's claims in evaluating whether the claims are novel and nonobvious. Frequently, the cited publications are patent publications. A patent publication is a published version of a patent or patent application produced by a patent office. A given patent may publish multiple times in slightly different forms. The most common example of this is the publication of a patent at the application stage and later publishing again as an issued patent.

When building landscapes, analysts sometimes will expand an initial list to include cited patents. However, this can lead to noise as examiners and applicants often cite patents that are not especially relevant.

2.4 Family

One somewhat unique aspect of patents is that they have family relationships. This is often expressed as a family ID. A patent family includes different publications of the same patent/patent application as well as other patents/patent applications that have a priority relationship. One common instance where a priority relationship occurs is where the applicant files the same application in multiple countries (patent rights are country specific). Another is where the applicant files multiple applications in the same country that share a detailed description, but have different claims. Typically, all family members relate to the same technical topic.

It is common to include all family members in a landscape. However, this is typically just a final step to increase coverage for the landscape.

3. AUTOMATED LANDSCAPING

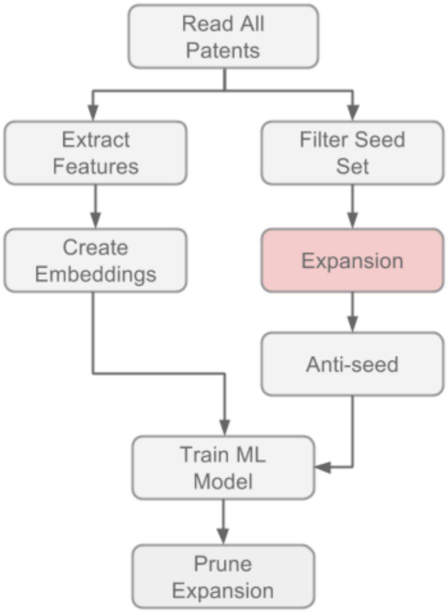

Figure 2 – landscape construction flow

A new automated approach for generating patent landscapes is presented below. This approach greatly mitigates many of the limitations discussed above by leveraging human insights, patent metadata, and machine learning.

This approach takes a human curated set of seed patents and expands it in order to populate a patent landscape. A human curated seed set is a sound starting point as it provides human insight as to the contours of the landscape. The seed set is then expanded using citations and class codes.

The initial results are often over-inclusive, but this is mitigated by pruning out less relevant patents using a machine learning model. The model provides a form of double verification. It is trained using the seed set as positive examples and a random subset of patents not included in the expansion (anti-seed) as negative examples.

The machine learning model can also be applied to the entire patent corpus, not just the expanded patents. This will lead to greater recall, but will also likely result in a drop in precision as there is no longer double verification outside the expansion. Additionally, this may require significantly more computing resources.

Figure 2 shows a high-level flow of the process. Importantly, the Expansion will be discussed later in the Expansion section.

3.1 Seed Patents

The seed set is the basis for the landscape so its composition is very important. Errors in the seed will propagate throughout the resulting landscape and become magnified. It is important that the seed set be accurate as well as representative. For example, if one wanted to build a landscape for mobile phones, it would be important to include patents relating to screens, antennas, housings, etc.

An advantage to using a seed set is that it minimizes human effort. It frees the user from having to distill search features. No need to identify keywords or class codes. It can also mitigate human review, by allowing the user to simply focus on a narrow subset of the patents to be included in the landscape. Seed sets can often be sourced from pre-existing lists that a user created for another purpose. Further, a user may not need to fully understand all the technical nuances of the topic as much of this is extractable from the seed set. For example, a user building a *speech recognition* landscape would not need to know about phonemes or hidden Markov models as these common features can be inferred from the patents in the seed set.

Depending on the topic, the size of the seed set may vary. For narrower topics, such as *email*, a seed set of a few hundred patents may be sufficient. For broader topics, such as *processors*, a couple thousand patents may be required. Note that these estimates are merely based on informal experimentation taking into account subjective evaluations of patents in the resulting landscapes, the amount of expansion, and the accuracy of the models all in relation to the seed set size. Additional research would be worthwhile here.

3.2 Expansion

To identify candidate patents to include in the landscape, the seed set is expanded based on structured patent metadata. This approach uses class codes and citations, though one could also consider using other types of metadata (e.g., inventors and assignees).

3.2.1 Family Citations

A particularly robust technique for citation expansion is to use bidirectional (i.e., forward and backward) family citations. This captures any patent where that patent, or its family members, cite to or are cited by a patent in the seed set or a family member of a patent in the seed set. This increases the size of the expansion, while sacrificing little accuracy as family members all tend to involve the same technical topic. An efficient way to compute the expansion is to build a family citation graph, where the nodes are families and edges are citation relationships. Then one can simply query the graph for all families within one degree of a family in the seed set.

3.2.2 Class Codes

Class codes are another way to expand the seed set. This can be accomplished by identifying highly relevant class codes and expanding the seed set to include all patents that have those class codes. Highly relevant class codes can be identified by evaluating the distribution of class codes in the seed set relative to the distribution of class codes in all patents. For example, selecting a class code where (1) it occurs in at least 5% of the patents in the seed set, and (2) the ratio of patents in the seed set having the class code is 50 times higher than the ratio of all patents having the class code. This second condition is important because some class codes could be very prevalent in the seed set, but not very relevant. These very prevalent class codes tend to be very general and can span multiple topics (e.g., CPC code G06 “computing; calculating; counting” [7]). Expanding on such class codes could bring in hundreds of thousands of patents, most of which would not be especially relevant.

3.2.3 Combining Approaches

Expansion by citation and expansion by class code are not mutually exclusive and performing both together is beneficial. Expanding by citations is good because the expansion generally reflects all the main technical aspects present in the seed set. However, patent families with few to no citations may not be identified and some citations may be to irrelevant patents. Expanding by class codes is good because all patents have at least one class code and the application of the codes is fairly accurate. However, many technical aspects may not have a specific class code, so expanding on class codes alone will leave gaps. Thus a combination of class codes and citations is preferred.

3.2.4 Running at Scale

We run our automated expansion and machine learning process (discussed in section 3.4) against all US-issued patents since 1980, encompassing approximately 15 million patents and published applications. The data is taken from an internal corpus and is approximately 40 terabytes in size. Essentially three pipelines are run: first, for expansion, second for feature extraction and model training, and thirdly to classify patents. These pipelines are scaled through Map Reduce [4] and an internal implementation of Google’s Cloud Dataflow.

3.3 Types of Expansions

The citation and class code expansion described above can be combined in multiple ways to construct landscapes. The following describes two examples: one with relatively narrow relevance and one with relatively broad relevance. Note that both these approaches use expansion both to identify the anti-seed, but also to pre-filter the results of the final landscape. Both the broad and narrow landscapes use the same sequence of expansions. The specific sequence used here is enumerated in pseudo-code as follows:

```
Seed Citations =
    FamilyOf(ExpandByRef(FamilyOf(SeedSet)))
Codes =
    ExpandByCode(HighlyRelevantCodes(
        SeedSet.CPC, FullCorpus.CPC))
Level 1 = Seed Citations ∪ Codes
Level 2 = FamilyOf(ExpandByRef(Level 1))
Anti-Seed = Sample(FullCorpus - Level 2, n)
```

In step 1, the seed is expanded by family citation. In step 2, the seed is expanded by highly relevant class codes. In step 3, the citation and class code expansions are combined resulting in Level 1 of the landscape. In step 4, Level 1 is expanded by family citation to produce Level 2 of the landscape. Finally, in step 5, the anti-seed is created by sampling patents outside Level 2 (Level 1 and the seed are a subset of Level 2). The anti-seed is used as negative examples in a later machine learning pruning process described in section 3.4. In this research, we used between 10,000 and 50,000 patents for the anti-seed set. Deeper, more broad topics tended to benefit more from larger anti-seed sets.

3.3.1 Narrow

In this narrow landscape example, only Level 1 is retained. The expanded patents in Level 1 are pruned using the machine learning model to ensure precision. This arrangement is well suited for constructing relatively narrow landscapes. Figure 3 gives the intuition of the narrow landscape. Note that the Level 2 expansion described above is still performed to select the anti-seed set.

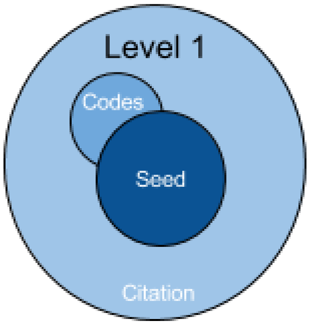

Figure 3 – narrow expansion

3.3.2 Broad

In this broad example, the patents from both Level 1 and Level 2 are retained. The expanded patents from Levels 1 and 2 are pruned using the machine learning model. The broad landscape is shown in Figure 4. This broad landscape example is more useful where a user desires to capture tangential topics as well as the primary topic. For example, in a GPS device landscape, a user may want to capture general touchscreen patents as touchscreens are a popular feature of GPS devices.

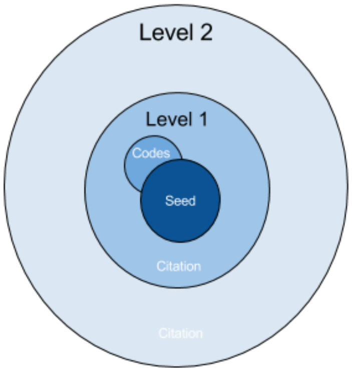

Figure 4 – broad expansion

3.4 Machine Learning Based Pruning

Both the class code and citation expansions will be over-inclusive. The results can be filtered using a supervised machine learning model to exclude non-relevant patents, which is covered in sections below. First, however, we look at some existing machine learning work used to automate patent classification.

3.4.1 Existing Machine Learning Approaches

Most available literature on machine learning based approaches to patent classification focus on fully supervised classification, where both positive and negative training samples are available. Many approaches use the existing hierarchical International Patent Class (IPC) and its replacement (CPC) schemas. These approaches tend to have relatively low accuracy due to the sheer volume of labels, or they focus on a small subset of labels. There are many thousands of classes across the IPC and CPC schemas, which vary greatly in distribution, making for a challenging classification task. For example, C.J. Fall, et. al. [8] focus on a small subset of IPC labels a using bag-of-words approach and achieve between 30-79% accuracy, but conclude that the results are not sufficiently strong to rely on in real-world scenarios. Other interesting approaches incorporate time-based information to increase robustness [12] with accuracy above 80%, or using information retrieval techniques to classify patents by finding similar patents already classified into the IPC and then use those patents' classes [16]. Additional methods using bag-of-words models have been described as well, e.g. for routing requests to examiners at the European patent office [9] and to assist with patent searching in a manner similar to medical search on PubMed [6], with relatively good results (approaching 90% accuracy), but again against the IPC classes. Most of the existing automated approaches do not consider structured metadata outside of class codes, with some notable exceptions [11].

We are unaware of any techniques that combine 1) semi-supervised techniques to auto-generate portions of the training data, 2) arbitrary labels where the user controls the domain, and 3) running at scale across the entire patent corpus. We were unable to find approaches that do any two of the three, in fact.

3.4.2 Training Data - Seed & Anti-seed

Good training data is essential for any supervised machine learning model. The seed set provides positive examples; however, negative examples are needed as well. One can think the of negative examples as the anti-seed.

An efficient way to create the anti-seed is to sample from patents not included in the expansion. If the expansion is done correctly, only a very small fraction patents not included in the expansion should be relevant to the topic of the landscape. Sampling from these patents provides a large amount of reasonably accurate and representative negative training examples. The best part is that it requires no human effort.

In this research, positive (seed) and negative (anti-seed) data reside at differing poles, with little representative training data that reside in "the middle". This is by design, since the goal is to prune data that isn't very similar to the seed set, however, it is also a limitation of the semi-supervised approach since the sampling technique has built-in bias.

3.4.3 Machine Learning Methodologies

Here, one could use a wide variety of supervised machine learning methodology from simple regressions to deep neural networks. The following describes two different methodologies that were applied with good success.

3.4.3.1 Ensemble of Shallow Neural Networks using SVD Embeddings

One methodology that produced good results was an ensemble of shallow neural networks using singular value decomposition (SVD) embedding [10]. This involves constructing SVD embeddings for each patent data type and training multiple neural networks using the embeddings as inputs. The patent data types that proved to be most useful were family citations, text (title plus abstract and claims), class codes, and inventor citations. The outputs of the neural networks are weighted using a logistic regression model. The results for this approach (below) reflect a single machine implementation with very small SVD embeddings (64 values per data type). Increasing the embedding size should increase accuracy.

Constructing the SVD embeddings is fairly straightforward. For each patent data type construct a sparse matrix of patents by features then apply SVD to compress the sparse feature space into fewer dense features. For family citations the matrix is a binary patent by patent matrix with the positive cells representing citation relationships between patent families. This is very similar to how one would apply SVD embedding for network analysis [17]. For text, use a patent by term frequency matrix. This is essentially a LSA approach [5]. For class codes use a binary patent by class code matrix. For inventor citations use a binary patent by inventor matrix with the positive cells representing a citation relationship between a patent and an inventor. Given the size of the matrices (millions by tens of thousands), a SVD approximation algorithm may be preferred, such as Lanczos bi-diagonalization [1].

The SVD embeddings are then used as inputs to the ensemble of neural networks. A different neural network is trained for each patent data type, as well as a composite neural network that is trained on all of the embeddings concatenated together. The hyperparameters for each neural network were: (1) an input layer that is the same size as the embeddings, (2) a single hidden layer, and (3) an output layer with a single node.

Once all the neural networks are trained, a logistic regression model is used to weigh their outputs. Using the ensemble increased accuracy and helped prevent overfitting making the machine learning more robust. This is especially important for an automated approach that seeks to reduce human effort by avoiding tuning. Interestingly, the weights assigned to each of the patent type specific networks can vary significantly from landscape to landscape. For example, landscapes that involved topics having well defined class codes had higher weights on the class code network. Conversely, landscapes that did not have well-defined class codes had higher weights on citations and examiner citations.

3.4.3.2 Perceptrons using Random Feature Projection Embeddings

A second machine learning approach applied during research was the Perceptron [15] algorithm and embeddings using Random Feature Projection (RFP) [2]. When dealing with data at a large scale, such as the universe of patents, sometimes it's preferable to sacrifice a small amount of accuracy for large increases in efficiency. In addition to more traditional methods such as TF/IDF, a very fast way to reduce a high-dimensionality bag-of-word feature space -- in this case, hundreds of millions of tokens across tens of millions of instances -- is by using RFP. Random feature projection is very fast and depending on the dimensionality of the resulting feature vector, can provide increased efficiency of machine learning algorithms at little-to-no loss of accuracy.

As with the shallow neural net approach described above, many types of features were used. In particular, the following proved useful:

- n-grams of size 1 and 2 for abstract, description, claims, and references' titles
- bigrams occurring at least twice across a patent's description *and* claims
- the CPC classifications of the patent

Notably absent in this approach, however, was the family citation and LSA of the text. Instead, bag of word text features and CPC classifications were combined into a single matrix. RFP was then applied on the full matrix, reducing the dimensionality to only 5,000 columns. Larger feature vectors, e.g. 10,000 or 15,000 columns, proved to have negligible increase in accuracy while increasing model size and slowing down the training and inference process.

Finally, while many algorithms were experimented with on the RFP data embeddings, including support vectors and logistic regression, the perceptron proved the most accurate on RFP embeddings by a minimum of 5% F1 score across most experimental landscapes.

4. RESULTS

As discussed above, most automated classification research focuses around automating classification into the existing IPC or CPC classification hierarchies [6; 9; 12; 16].

The current research was performed on an internal Google dataset containing all patents since 1980 that are English-language and US-issued, which is approximately 15 million patents and published applications. However, this automated technique should apply to non-US patents as most countries make use of citations and class codes.

4.1.1 Analysis Methodology

The machine learning analysis is primarily done by evaluating F1 scores [19]. Reported here are the F1 scores and additional details for three example landscapes. Further, a plot of 60 landscapes illustrating the broad applicability of the methodology.

In addition to the standard machine learning analysis, internal patent domain experts reviewed the resulting landscapes for validity. This was somewhat subjective and involved reviewing representative portions of the results, comparing overlap to traditionally generated landscapes, and analysis of aggregate statistics (e.g., assignee distributions).

4.1.2 Machine Learning Results

Figure 5 shows the landscapes across the top columns, and in subsequent rows are the F1 scores for the machine learning process, the number of patents in the seed set, the number in the narrow expansion, the number in the narrow landscape after pruning, and similarly for the broad landscape.

The F1 score was calculated by training the machine learning models on the seed and anti-seed sets, running k-folds cross-validation (k=10), and retaining the maximum F1 score. Note that the scores reflect the F1 score for the *minority class*, that is, the patents in the seed set. What this tells us is that, while the training set is imbalanced (as heavily as 1:10 in extreme cases), by calculating the F1 for the minority class we show the classifier isn't always choosing the majority class to "cheat." The majority (negative) class also scores strongly, but results are not presented as it is not important to the task of pruning with high precision. We also vary the randomization seed for cross-validation splits between runs to help avoid overfitting.

While the results are very promising, they may overstate the actual accuracy of the model as there are likely to be few borderline positive/negative examples in the training data. In the current context, this is acceptable, as the goal is to prune patents from the expansion that are not clearly related to the seed topic.

Topic	*browser*	*operating system*	*machine learning*
RFP F1 Score	.954	.882	.929
SNN F1 Score	.930	.797	.895
# Patents in Seed Set	1094	1083	1238
# in Level 1 Expansion	72,895	44,779	46,570
# in Narrow Landscape (RFP)	49,363	30,319	28,947
# in Level 2 Expansion	511,173	505,176	527,203
# in Broad Landscape (RFP)	92,362	118,675	86,432

Figure 5 – seed and expansion counts for three sample landscapes

As shown in Figure 5, patents identified in the Level 1 expansion are much less likely to be pruned by the machine learning model than patents identified in the Level 2 expansion. For the *browser* topic 68% (49,363/72,895) of patents identified in Level 1 are retained whereas only 10% (92,362-49,363)/(511,173-72,895) of patents identified first in Level 2 are retained. This, especially in light of the high F1 scores, supports the notion that the citation and class code expansions are a reasonably good way to identify potentially relevant patents and strong negatives (anti-seed). This also suggests the model is dealing with close positives and negatives fairly well, at least in aggregate. We see similar shapes of expansion and pruning across other landscapes not detailed here.

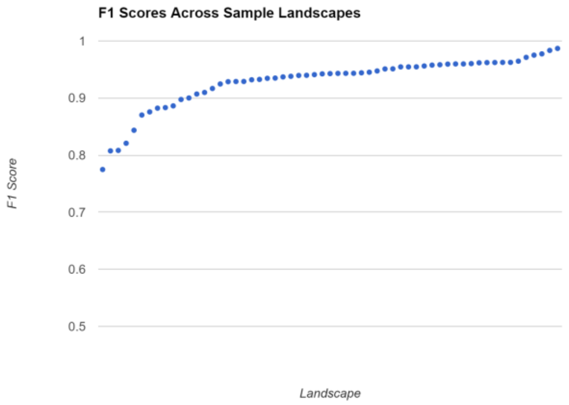

Figure 6 – distribution of F1 scores for some sample patent landscapes produced using RFP & Perceptron approach

The plot from Figure 6 below shows the distribution of F1 scores across a range of 60 topics in ascending order by F1 score. Details of each topic are omitted, however this demonstrates that the Automated Patent Landscaping technique can be applied to numerous topics with success. Most landscapes achieved an F1 score above .90. There were some with poorer scores, of course. For example, the lowest score was for a landscape about "car infotainment." In that model, the classifier frequently confused car-related patents that had no infotainment components as being part of the topic. Were this landscape to be used in a real world scenario, more tuning of the model and seed set would likely be necessary.

4.1.3 *Patent Landscape Results*

The resulting patents in both the broad and narrow sets have two sets of validation (expansion and machine learning), as described previously. The results are in line with internal expert evaluations that found the resulting landscapes to be both accurate and reasonably comprehensive using the methodology described above in section 4.1.1.

As demonstrated in the results and expert evaluation, this automated patent landscaping approach provides a repeatable, accurate, and scalable way to generate patent landscapes with minimal human effort. This has significant implications for companies, governments, and investors that benefit from patent landscape analysis, but are discouraged because of cost and effort. Further this approach, especially if widely adopted, makes patent landscapes more compelling to external audiences as they can focus on the results rather than questioning the underlying methodology.

It should be noted that the landscaping approach described here could be applicable to other domains that have similar metadata. Two that come to mind, in particular, are scholarly articles and legal opinions.

5. REFERENCES

[1] BAGLAMA, J. and REICHEL, L., 2005. AUGMENTED IMPLICITLY RESTARTED LANCZOS BIDIAGONALIZATION METHODS. *Society for Industrial and Applied Mathematics*, 19-42.

[2] BLUM, A., 2005. Random projection, margins, kernels, and feature-selection. In *Proceedings of the 2005 international conference on Subspace, Latent Structure and Feature Selection* Springer-Verlag, Pittsburgh, 52-68.

[3] COCKBURN, I.M. and MACGARVIE, M.J., 2009. Patents, Thickets and the Financing of Early-Stage Firms: Evidence from the Software Industry. *Journal of Economics and Management Strategy, vol 18, issue 3*, 729-773.

[4] DEAN, J. and GHEMAWAT, S., 2004. MapReduce: simplified data processing on large clusters. In *Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6 (OSDI'04)*, Berkeley.

[5] DEERWESTER, S. and AL, E., 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 391-.

[6] EISINGER, D., TSATSARONIS, G., BUNDSCHUS, M., WIENEKE, U., and SCHROEDER, M., 2013. Automated Patent Categorization and Guided Patent Search using IPC as Inspired by MeSH and PubMed. In *Journal of Biomedical Semantics*.

[7] EPO and USPTO, 2016. Cooperative Patent Classification. http://worldwide.espacenet.com/classification?locale=en_EP#!/CPC=G06.

[8] FALL, C.J., TÖRCSVÁRI, A., BENZINEB, K., and KARETKA, G., 2003. Automated categorization in the international patent classification. In *SIGIR Forum 37*, 10-25.

[9] KRIER, M. and ZACCÀ, F., 2002. Automatic categorisation applications at the European patent office. *World Patent Information 24*, 3, 187-196.

[10] LATHAUWER, D., LIEVEN, and OTHERS, 2000. A MULTILINEAR SINGULAR VALUE DECOMPOSITION. *Society for Industrial and Applied Mathematics*, 1253-1278.

[11] LI, X., CHEN, H., ZHANG, Z., and LI, J., 2007. Automatic patent classification using citation network information: an experimental study in nanotechnology. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, New York.

[12] MA, C., LU, B.-L., and UTIYAMA, M., 2009. Incorporating Prior Knowledge into Task Decomposition for Large-Scale Patent Classification. In *Proceedings of the 6th International Symposium on Neural Networks: Advances in Neural Networks - Part II*, Berlin.

[13] O'CONNOR, D., 2012. ONE IN SIX ACTIVE U.S. PATENTS PERTAIN TO THE SMARTPHONE. In *Disruptive Competition Project*. http://www.project-disco.org/intellectual-property/one-in-six-active-u-s-patents-pertain-to-the-smartphone/.

[14] REIDENBERG, J.R. and AL, E., 2015. *Patents and Small Participants in the Smartphone Industry*. Center on Law and Information Policy at Fordham Law School.

[15] ROSENBLATT, F., 1958. THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION. *Psychological Review*, 386-408.

[16] SHIMANO, T. and YUKAWA, T., 2008. An automated research paper classification method for the IPC system with the concept base. In *Proceedings of the NTCIR-7 Workshop Meeting*, Tokyo.

[17] SKILLICORN, D.B., 2006. Social Network Analysis via Matrix Decompositions. In *Popp, Robert L.; Yen, John* John Wiley & Sons, Hoboken, 367-392.

[18] TRIPPE, A., 2015. *Guidelines for Preparing Patent Landscape Reports.* World Intellectual Property Organization.

[19] VAN RIJSBERGEN, C.J., 1979. Evaluation. In *Information Retrieval* Butterworths, 133-134. http://www.dcs.gla.ac.uk/Keith/Preface.html.