Separating Precision from Recall, and the Use of Machine Learning Methods for the Generation of High Value Patent Landscape Collections

Managing Director, Patinformatics, LLC **Anthony Trippe**

Anthony (Tony) Trippe is Managing Director of Patinformatics, LLC. Patinformatics is an advisory firm specializing in patent analytics and landscaping to support decision making for technology based businesses. In addition to operating Patinformatics, Mr. Trippe is also an Adjunct Professor of IP Management and Markets at Illinois Institute of Technology teaching a course on patent analysis, and landscapes for strategic decision making. Tony is an influential thought leader in the patent analysis space, recently wrote the Guidelines for the Preparation of Patent Landscape Reports for the World Intellectual Property Organization, has been named one of the Top 300 IP Strategists by IAM Magazine, and is the author of the popular Patinformatics blog.

1 INTRODUCTION

It can be argued that the single most important step in generating a patent landscape report (PLR) is the creation of an appropriate data collection for performing the corresponding analysis. The consequences of creating a sub-optimal data collection goes by many names, but most often the acronym used to describe this situation is GIGO — "garbage in, garbage out". In the case of patent landscape reports the data most frequently used is collected by performing a patent search. The patent information retrieval must be fit for purpose, or the analyst runs the risk of producing results that are irrelevant at best, and incorrect and misleading at worst. These poor results are the garbage out that GIGO refers to. Understanding the consequences of building a sub-standard patent collection for the generation of a PLR, how does an analyst measure the quality of a patent search, and perhaps more importantly how do they optimize their approach to building a collection that will produce the most relevant results.

2 Measuring Patent Information Retrieval Effectiveness

In the data science world, information retrieval, or searching effectiveness is traditionally described in terms of two measures, recall and precision. These items are defined as:

- *Recall – how much of the useful information has my search retrieved?*
- *Precision – how much of the information that I have retrieved is useful?*

Stated another way, recall is an estimate of the probability that a relevant document will be retrieved in response to a query and precision is an estimate of the probability that a retrieved document will be relevant. Figure 1 provides a demonstration of these concepts[1]:

Thinking about the issues in searching during the preparation of a PLR, information retrieval methods usually look at precision and recall simultaneously and measure their methods by how search techniques stack up against both elements. Even though this is the case,

[1] https://www.tutorialspoint.com/data_mining/dm_mining_text_data.htm

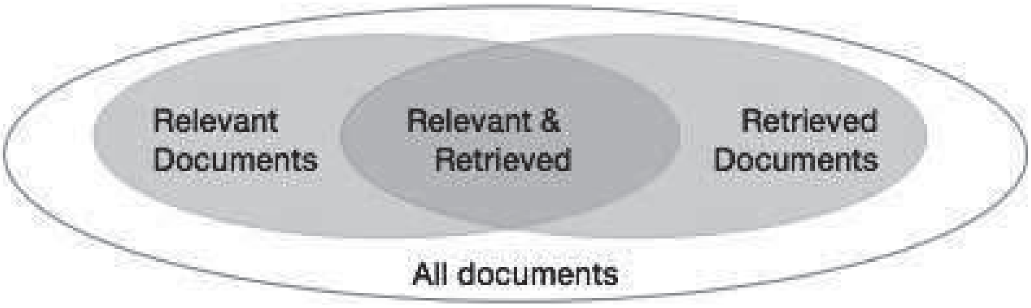

Figure 1: A Venn Diagram of Recall vs. Precision

precision and recall are normally opposed to one another such that with an increase in recall there is usually a subsequent drop in the level of precision. Generally speaking, as PLR searches are often designed to maximize recall, the precision of the results can suffer since more off-topic references may get included in the collection. This can produce the "garbage in" scenario described above if something isn't done to also improve the precision of the collection.

Having said this though, it is also important to keep in mind that there is a point of diminishing returns when it comes to high precision. High recall is always desirable since it is important to have as much of the relevant information available for analysis as possible when building a PLR. A good rule of thumb is for analysts to try and achieve at least 90% recall in their collection. One way to measure this is to intentionally broaden a search by adding additional search terms, or classification codes to the query. If the majority of the new records discovered were not in the previous iteration of the search, and most of the new records are not relevant then the analyst can begin to feel that the search has achieved the desired level of recall.

3 Separating Precision and Recall for Patent Information Retrieval

Coming back to the topic of high precision keep in mind that when statistical analysis is performed on large, or macro-level sets only major trends, or items that appear frequently are going to be seen. Precision, in this instance, can suffer to some degree with these types of searches, since minor occurrences within these sets will not be seen in the larger context. This can often be evaluated by examining several of the significant trends to ensure that they are coming from reasonably precise references. If this is the case then it is generally acceptable to sacrifice some precision for the sake of recall. From personal experience a greater than 80% precision is required to avoid the GIGO scenario.

As previously stated though, in order to get to 90% recall there is a high likelihood that precision will suffer, and could in fact fall below the 80% threshold. Since this is the case how does an analyst get to the point where they can achieve the desired thresholds with regards to both measures. In generating collections for PLRs, it might be more productive to begin by creating sets using methods that produce high recall, and then look to increase the precision separately once the initial, high recall collection is built. This approach is contrary to traditional information retrieval methods that try to maximize both simultaneously.

4 Maximizing Recall

Of the two tasks associated with optimizing recall, and precision, maximizing recall is by far the easiest. Almost anyone can search with a couple of intentionally broad patent classification codes by simply truncating them to four-digits. For instance, searching with A01H (New Plants or Processes for Obtaining Them; Plant Reproduction by Tissue Culture Techniques), or C12N (Biochemistry — Microorganisms or Enzymes) will generate a pretty comprehensive collection of documents associated with plant breeding. The issue in this case is that this particular search will generate more than 2 million worldwide records. Keeping in mind that no competent patent searcher would ever suggest building a query like this it is still illustrative of what could be done to try to achieve nearly absolute recall. This would be the ultimate example of "garbage in, garbage out" since much of the output associated with an analysis of this collection would probably have very little to do with plant breeding, assuming it would even be possible to produce results based on the analysis of this many records.

In a similar fashion, an obscenely broad search can be accomplished using keywords, for example, a search of plant "and" breeding in the full-text of worldwide patents would produce over 300,000 records. While this is certainly less than over 2 million this approach suffers from similar issues as the query that involved overly broad use of patent classification codes.

Perhaps more constructively, one of the more useful ways of increasing recall without producing collections that are overwhelmingly broad is to use multiple searches that incorporate a variety of search tools. For example, listed below are several search methods that are used to query patent collections:

- ●Reasonably defined keyword hedges — including the use of synonyms and proximity operators as well as broader Boolean operators
- ●Narrow patent classification codes — at the group/subgroup/class/subclass level some codes are specific enough that they produce precise results by themselves
- ●Citation analysis — backward citations in particular are often used to identify relevant references
- ●Semantic analysis — using techniques such as natural language processing relevant records can be found that weren't retrieved using keywords and Boolean

To further improve recall these methods can be used in combination with one another, for instance:

- ●Broader patent classification codes can be qualified with broader keyword hedges
- ●Citation, and Semantic analyses can be filtered by using broad patent classification codes

All of these methods often produce additional results beyond the ones generated when the more specific versions of them alone were used. When all of these approaches are combined a high recall collection is normally produced, and can be verified by comparing the difference between these results with an overly broad search and checking to see if any of the unique references associated with the overly broad search are in fact relevant.

5 Maximizing Precision with Machine Learning Methods

Even using the methods above to produce a high recall, but reasonably sized collection there is still a likelihood that the patent records will not meet the 80% precision threshold that is desirable for a high-quality landscape data set. Many of these collections will also be quite large, often numbering in the thousands, or tens of thousands of patent families, so the traditional method of manually reviewing even titles, or enhanced titles for relevance would be a large, tedious chore. So now that recall has been maximized how does an analyst approach the issue of precision, especially with larger data collections?

One potential solution for increasing precision in a patent data set is to turn to the field of machine learning for organizing, and prioritizing documents. These methods have quickly become some of the most polarizing tasks associated with patent analytics. While these approaches have caught on, and are used in many industries, the adoption in the patent information space has been sporadic. Having said this, the tide may be turning with regards to implementation of machine learning methods for patent information retrieval. At the most recent PIUG Annual Meeting held in May of 2017[2] a significant percentage of the presentations covered machine learning, and many of the vendors introduced products that incorporated it in one form, or another. So what exactly is machine learning?

Consulting Wikipedia the following definition is found[3]:

"Machine learning, a branch of Statistical Learning, is about the construction and study of systems that can learn from data. For example, a machine learning system could be trained on email messages to learn to distinguish between spam and non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders."

There are many machine learning methods that can be applied to patent information retrieval, and analytics including text clustering, and a more involved interpretation of clustering to produce spatial concept maps. For the purposes of improving precision though, classification is the method that is likely to be of the highest value. Classification is usually accomplished with a supervised, machine learning method that uses "learning sets" to identify key attributes of documents in a category. The "learning sets" are small sub-collections, one for each category, generated by the analyst, who decides which test documents should appear in each class. New documents are compared to the learning collections, and assigned to a class based on their similarity to the documents that have already been assigned to the category.

When it comes to classification the most frequently applied supervised machine learning methods are Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs):

- ●Artificial Neural Networks[4] – In computer science and related fields, artificial neural networks are models inspired by animal central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition. They are usually

[2] https://www.piug.org/an17program

[3] https://en.wikipedia.org/wiki/Machine_learning

[4] https://en.wikipedia.org/wiki/Artificial_neural_network

presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network.

- Support Vector Machines[5] – supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories; an SVM training algorithm builds a model that assigns new examples into one category or the other.

6 Binary Classification Using a Support Vector Machine to Increase Precision

The methods used for automatic classification have been around for some time, and have been used by patent offices, publishers and database producers, in association with patent information, but there have not been many commercial tools providing classification capabilities to analysts, and patent information retrieval specialists. This is changing however, and in the not too distant future patent information professionals will have a variety of machine learning based classification systems available to them. Before launching into an example of how classification can assist with the identification, and prioritization of relevant references, within large patent document sets, let's look at some details of the task itself.

The Wikipedia entry on Statistical Classification provides the following description of this method[6]:

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category.

Classification can be thought of as two separate problems – binary classification and multiclass classification. In binary classification, a better understood task, only two classes are involved, whereas multiclass classification involves assigning an object to one of several classes. Since many classification methods have been developed specifically for binary classification, multiclass classification often requires the combined use of multiple binary classifiers.

When it comes to increasing precision in a patent data collection binary classification is the method that should be used. For this article, a support vector machine (SVM) implementation of binary classification will be used for the task. Referring back to the Wikipedia reference on SVMs the motivation behind the method, and an illustration are provided:

Suppose some given data points each belong to one of two classes, and the goal is

[5] https://en.wikipedia.org/wiki/Support_vector_machine

[6] https://en.wikipedia.org/wiki/Statistical_classification

to decide which class a new data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector (a list of p items), and we want to know whether we can separate such points with a (p − 1) -dimensional hyperplane. This is called a linear classifier. There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes.

A theoretical example of this can be seen in Figure 2:

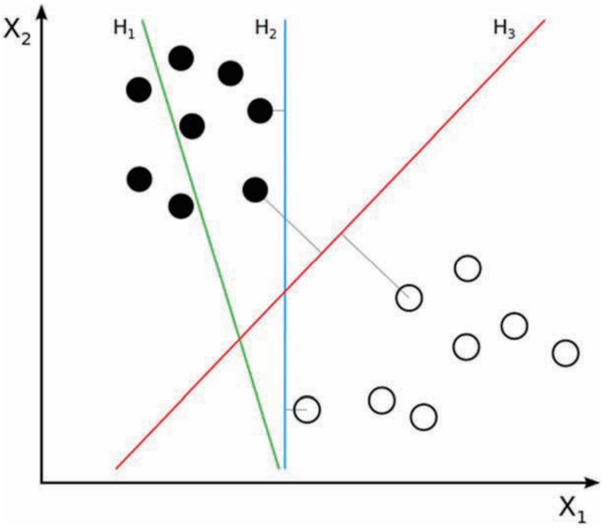

Figure 2: Illustration of separating hyperplanes in Support Vector Machines

Binary classification provides a means for categorizing large collections of patent documents into the references that are likely to be of highest interest to the information professional, and those that are likely not related, but were still retrieved in a high recall search. The training set, in this case will be made up of references that would be found in a high precision version of the collection. In training the classifier, examples of patent records that are not relevant also need to be identified, so the classifier can establish a hyperplane that will distinguish between the two categories.

7 An Example, Putting these Ideas into Practice

Several years ago, the author developed an interest in wearable fitness monitors and began using this field as an example when exploring machine learning methods and the problem of recall, and precision in patent data collections. Two of the major companies working in the space at the time were Aliphcom (doing business as Jawbone) and Nike. Both organizations sell other products, and have extensive patent portfolios, which cover their fitness monitors, as well as many additional items. A binary classifier, using a SVM can help identify the patents associated with personal fitness monitors in the midst of many other patents from these companies.

Searching worldwide, several hundred patent documents are assigned to Aliphcom. Of these, more than 100 are associated with their personal fitness band, based on a previous analysis conducted using a manual method of classification. Ten of these documents were used to represent the positive examples in the training set. The Aliphcom portfolio also contains patent documents associated with Bluetooth headsets and speakers. Ten documents associated with these items were identified as the negative examples.

The first step in preparing to build a classifier is to decide on the sections of text that will be used to create the individual document vectors. In this case, the source titles, and abstracts were used, but in other circumstances enhanced titles, and abstracts, or source claims could be used as well. All potential family members were imported in this case, but frankly, under normal usage, it' s probably a good idea to put the documents through some type of family reduction before performing a classification task. The source titles and abstracts can be used for a binary classification since the user is simply trying to separate relevant documents from the remainder of the collection.

The next step involves training the classifier, and as discussed above, ten positive and ten negative examples were initially chosen to accomplish this. Depending on the application used to generate the SVM there will be various methods for selecting the documents for the training set. Once the initial classification is accomplished most systems will present the analyst with a set of at least ten middle-of-the-road documents that the classifier is having difficulty classifying. The analyst is asked to look at these, and manually classify them before a second training round takes place with the new training set.

After only three training rounds a classifier was created that successfully classified all but one of the Aliphcom documents correctly into those covering the personal fitness monitor compared with the remainder of the company's products. The one document, and its equivalent family members were new, recently published, and dealt with a new application of the product line. All and all, with minimal effort, a result with greater than 95% precision was achieved. In this particular example, working with a few hundred records where about 25% are relevant to start with is a fairly simple test. A real challenge would be whether a SVM classifier could be used to classify larger, more highly diversified portfolios.

To test this, 11,126 worldwide patent documents from Nike were submitted to the SVM for classification using the final classifier generated from the classification effort on the Aliphcom patents. The initial use of the Aliphcom classifier did not produce particularly good results. This was to be expected since the language used in the Aliphcom records was different from the text used in the Nike portfolio. Ultimately, after additional training, a classifier will need to be able to handle this eventuality since patent collections are rarely homogeneous with regards to language, especially between different companies. Having looked at patents associated with the Nike personal fitness monitor using traditional review methods, many of the known documents did not score well with the Aliphcom training data. This situation was remedied by retraining the classifier, as was done with the first Aliphcom classifier. After three generations of training, the classifier had successful scored ~85% of the Nike documents accurately. It still scored some of the originally discovered documents poorly, but frankly, many of these were associated more with a sensor system that was similar but distinctly different than the personal fitness monitor. Interestingly, the classifier in this case identified several Nike families that were not discovered using the original, traditional search. Combing the methods, in this case, would have led to a more comprehensive result when studying the Nike fitness monitor filings.

Finally, and representing an even bigger challenge for a computer assisted method of classification 43,612 US, and WO documents in International Patent Classification class A61B005, the class under which the majority of the relevant documents analyzed to this point were assigned, were classified using the Aliphcom and Nike classifiers.

This is an enormous number of extraordinarily diverse documents, and a very tall order for a machine learning method. A61B005 is the IPC class for measuring for diagnostic purposes, and it includes MRI, and blood glucose monitoring as well as the fitness devices being investigated. The language used in the titles and abstracts, of these documents, can be very different than what was used in the Nike and Aliphcom documents.

The first attempt at classification produced very poor results with a few documents receiving a high score and a handful that received reasonable scores, but were clearly off-topic. Following the previous pattern, retraining of the initial model was performed. Due to the size and diversity of the collection, this process was repeated five times before a reasonable outcome was produced. Using the fifth-generation classifier, 620 documents received a score of 50, or better. The titles of these were studied, and from a preliminary examination it appeared as if ~80% of these were on topic.

8 Conclusions

Taking a large number of documents from a diverse IPC collection is the ultimate test for a machine learning method, and in this example the SVM performed reasonably. In real world situations, it is recommended that family equivalents be removed and document collections created that are not quite so broad. Alternatively, additional retraining sessions will help bring more relevant references over the relevance threshold.

While patent landscape searches were the focus of this particular article the machine learning methods discussed can also be applied to other patent searches as well. Regardless of whether the patent information professional is interested in Patentability, Freedom-to-Operate, or Validity it is often the case that results are presented in reverse chronologic order, as opposed to relevancy order. Even when a relevancy rank is provided it is sometimes difficult to determine what criteria was used to generate the ranking. With the use of a SVM the analyst is in control of choosing the training set, evaluating the results, and deciding how many retraining sessions are required to generate a sub-collection which is highly likely to be related to source patents of interest. In these circumstances, a classifier could save significant time in reviewing patent records.

Information retrieval systems typically look at precision and recall simultaneously and measure their results by how techniques stack up against both elements. When it comes to patent searching it might be more productive to separate these functions so that they can be maximized independently. It has been demonstrated that precision and recall do not follow the same linear path when producing a patent search. Since this is the case it might be more productive to begin with methods that produce high recall exclusive of precision. Once this is accomplished the results can be ranked using different methods to improve precision and manage the way the results are shared with the analyst. It will likely be the case that different approaches will be used to provide higher recall than those that can be employed to share records with higher precision. Instead of expecting a single method to do both it would be useful to the patent searching, and analysis community if the process was done stepwise to maximize the value to the analyst.

高付加価値特許ランドスケープ収集を作成するための適合率／再現率の分離と機械学習法の利用（抄録）

Anthony Trippe

1．はじめに

特許ランドスケープ報告書（PLR）を作成する際の重要事項はその分析に相応しい特許集合を準備することである。最適化の前の集合状態を示す頭字語は GIGO（ゴミ入れ、ゴミ出し）と呼ぶ。その収集は検索によって集められる。その収集は目的に相応しいものでなければならない。さもないと、分析結果がせいぜい不適切で、最悪の場合にはミスリードを犯すリスクが伴う。このような不適切な結果をゴミ出しと言う。PLR を産む准最適化された収集を構築するためには、その特許検索の品質を如何に測るかが重要だ。

2．特許情報検索の有効性の測定

データ科学では、伝統的に再現率と適合率の 2 つの測定因子で測られてきた。それぞれの定義と説明は、図 1 を参照されたい。一般的には、再現率を上げるとその結果として適合率が下がってくる。一般的にいって PLR は再現率を最大にしようとするから garbage in シナリオになる。良い例は、再現率が少なくとも 90% 以上の時に履行される。だから検索用語を追加し、分類コードを追加することが多い。個人的な経験から判断すると、GIGO シナリオを回避するには適合率 80% 以上が求められる。

3．特許情報検索のために再現率と適合率を分離する

重要な語句が頻繁に出現する大量データを扱う場合には再現率が犠牲になっても傾向が解るが、重要な語句の頻度が低い場合には要注意である。経験から言うと適合率が 80 以上では GIGO シナリオに陥るのを避けることが必須である。再現率を 90%以上にすると再現率が悪くなるのが一般的である。良い PLR を得るには先ず高い再現率を求め、次に適合率を高めたい。しかし、この方法は伝統的な再現率と適合率を同時に高めようというやり方とは反対のやり方である。

4．再現率を最大化する

必要な分類コードを使えば、ある事例では、200 万件が直ぐに集められ再現率も高い。まさに、“garbage in, garbage out”である。全文検索によりキーワードキーだけで検索すると 30 万件がヒットする。200 万件よりかなり少ないので良くない。再現率を高く保ちながら適合率を下げない工夫として各種の検索ツールを組み合わせるのが良い。その例を下記に示す。

●広義のブーリアン演算だけでなく同義語を含む近接演算を利用する。
●適切な狭い特許分類コードを使う。
●後方引用を含む引用被引用分析を利用する。
●キーワードやブーリアン検索で得られなかった自然言語処理の概念検索を利用する。
●広義の特許分類を広義のキーワードで組み合わせる。
●引用および概念分析の結果を広義の特許分類を使用してフィルタリングを行う。

5．機械学習法を用いて適合率を最大化する

Wikipedia には各種の用語が定義されている。

1）機械学習
2）人工ニューラルネットワーク
3）サポートベクターマシン（SVM）
4）統計的分類
5）SVM による二値化識別

6．適合率を増加させるための SVM を利用した二値化識別（重要な主張ポイント）

二値化識別は大量の特許データを最も関心が高く可能性が高い特許文献と、検索はされたが関連性が高くない可能性がある特許文献に識別する手段を提供する。二つのカテゴリーに区別する識別超平面を描くことができる。（図 2 参照）（回帰問題として解く。）

7．当アイデアを実践する事例研究

数年前にウェアラブルフィットネスモニターへの関心が高まり、機械学習の方法と、再現率の問題と特許

データ収集の適合率の問題を検討した。Aliphcom 社とNike 社の 2 社のデータを使って SVM を利用して二値化識別を行った事例が役立った。

最初に、Aliphcom の少数の事例のタイトル、要約、請求項を対象にして識別を行った。次に、良い該当事例を 10 個抽出し、該当から外れた事例を 10 個抽出した。この最初の方の処理段階で二値化識別が難しく、中間領域に混在する場合も発生するが、その場合にはトレーニングをやり直した。この二値化識別のトレーニングを三回も繰り返すと、適合率が 95%より大きくするこができるようになった。

そこで、Aliphcom 社の教師データを使って、Nike 社の 11,126 件の特許のカテゴリー化を試みた。三回のトレーニングの後には、85%程度の識別成功率がえられた。

別の A61B005 分類の 43,612 件の US 特許と WO 文献でも検討を行った。三回のトレーニングでも識別がうまくできない場合には 5 回目の見直しを行う五世代識別を行った結果、Score が 50 以上の 620 件の特許文献を抽出することができた。

8. 結論

ここで議論された機械学習をこの記事の焦点の特許ランドスケープ検索に応用した。他の特許検索にも適用が可能である。特許専門家が特許性調査、FTO 調査、特許有効性調査に興味を持つか持たないかに関係なく、この手法の結果は時系列とは関係なく提示されることが多い。関連性ランクが提供されている場合でも、ランク付けを生成するためにどの基準が使用されたのかが解らない場合が多い。SVM を利用することにより、分析者はトレーニングセットの選択、その結果の評価、および関心あるソース特許に関連する可能性が高いサブコレクションの生成に必要な再トレーニングセッションの数を決めなければならない。大量の特許文献のレビューをする際には識別法はかなりの時間節約になる。

特許情報検索システムは、通常、適合率と再現率の両方を同時に探求する。しかし、特許検索ではそれらが独立して稼働させるために分離して扱う方がより生産的である場合もあることを実証している。だから、先ず高い再現率を生成する手法から始める方が生産的である。これを達成してから、異なる方法を活用して適合率を向上させ、分析者と結果を共有する方法を管理して、結果をランク付けすることができる。再現率と適合率を同時に高める方法を期待するのでなく、分離して検討し、分析者の価値を最大化するために段階的プロセスを行うことだ。それは、特許検索や分析コミュニティにも有益なものである。