# Inherent Document Value

Patentics.com White Paper

Jenny Qiu

## Introduction

As the cost of data storage becomes cheaper and cheaper, the amount of data that we keep and store increases but the cost of human effort to process and understand the data still remains the same. For each document that requires a human to read and evaluate, the approximate man-hour-dollar amount can be as much as $10 per document, which starts to add up when the document count is in the millions and multiple readers are cycling through the same document. And not to mention the sheer quantity of documents would take a team of qualified analysts several days, if not weeks, to digest and process into manageable chunks of ingestible data. This timeframe is often unacceptable when we need results within hours or even minutes.

Patentics has developed an unparalleled and powerful solution to this problem of filtering out valuable documents from not as valued ones, relevant from the irrelevant. Within a matter of minutes, data-heavy documents like patents can be separated into groups of valuable and non-valuable document groups. The method is simple: using the document set, compare each document in the first set with all of the documents from the second set, ignoring duplicates. The results are then ranked and scored and only documents in the second set with a high relevance score will be accepted. This partitioning process is illustrated in figure 1 below.

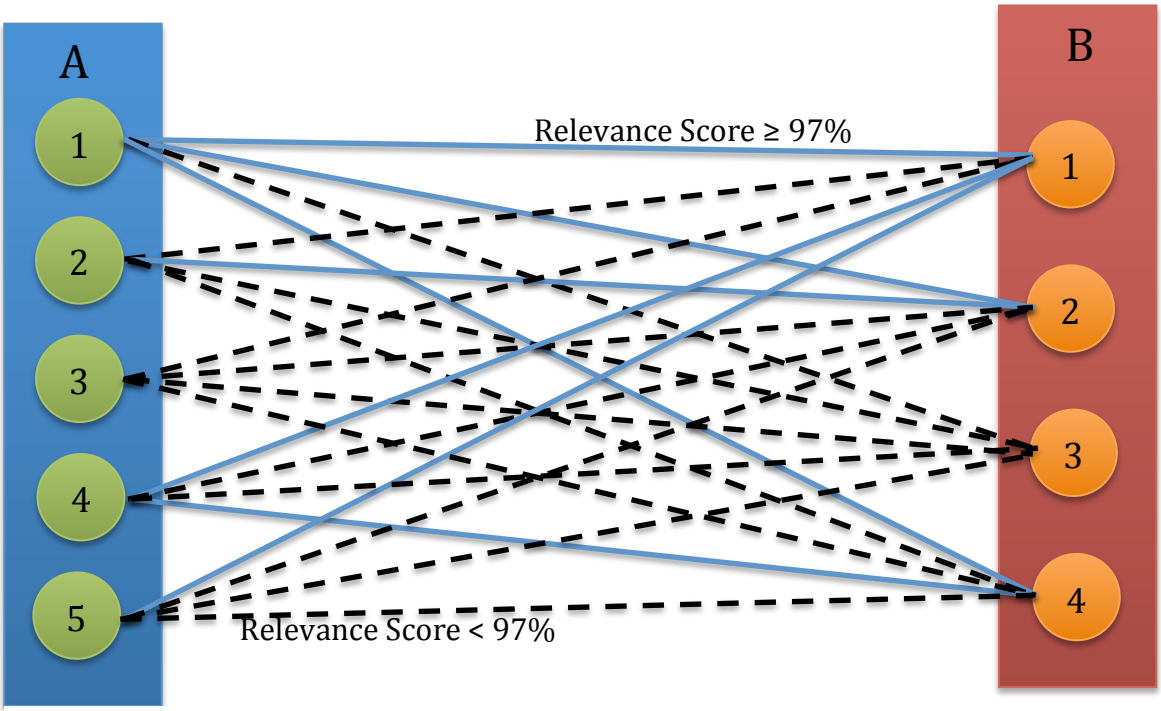


Figure 1

As illustrated in figure 1, top documents in A (1, 2, 4, 5) are aligned with top documents in B (1, 2, 4) because the second set contains documents that have a relevance score greater than or equal to 97% with any document in the first set. The relevance score can be set to any number, the higher the score, the smaller and more relevant the result set will be.

What is unique about Patentics is its ability to calculate the relevance score. Each document can be represented in space by a vector, defined by the words and terms contained in the document. To determine the similarity between any two documents, the proprietary algorithm calculates the proximity of the two vectors by computing the inner product, or the angle between the vectors, which is a mathematical representation of the content and concept similarity of the documents. What distinguishes Patentics' version of the algorithm from other versions is the accuracy and speed of calculation for large full-text documents.

## Citations Made

Traditionally, scientific journals and other publications rely on the number of citations that articles receive as a proxy for the relative importance. A similar concept is employed by Google's PageRank algorithm, which determines the relative importance of a webpage based on the number and quality of hyperlinks that refer to it. However, what if a document lacks explicit citations, such as intelligence briefs or news articles? In this instance, there is no existing way to determine an article's value in comparison to other articles because there are no external factors that indicate its effect on others. Patentics' algorithm is able to analyze a document's semantics to create a map of implicit citations across multiple documents.

To test our hypothesis, we've taken a group of patent documents that have citations. US granted patents are unique in that they contain a list of references but our algorithm is not trained to use these references to calculate importance. As a result, we can use them after the computation to determine how well Patentics was able to select the good documents from the bad as a benchmark.

We've chosen four USPC classifications; the following table lists a description for each.

| | |
|---|---|
| **386** | MOTION VIDEO SIGNAL PROCESSING FOR RECORDING OR REPRODUCING |
| **369** | DYNAMIC INFORMATION STORAGE OR RETRIEVAL |
| **375** | PULSE OR DIGITAL COMMUNICATIONS |
| **382** | IMAGE ANALYSIS |

Table 2

Using these classifications, we limit our full-text patent document set into four datasets for testing. For each set, each document in the set will be run against all other documents in the same set using our relevance algorithm to find ones that have at least 97% relevancy in common, with the percentage adjustable. We include another dataset for consideration: the inverse of our calculated set, which we consider as non-valuable patents for comparison purposes.

The following is a graph that shows the average number of patents that have cited a patent in each group.

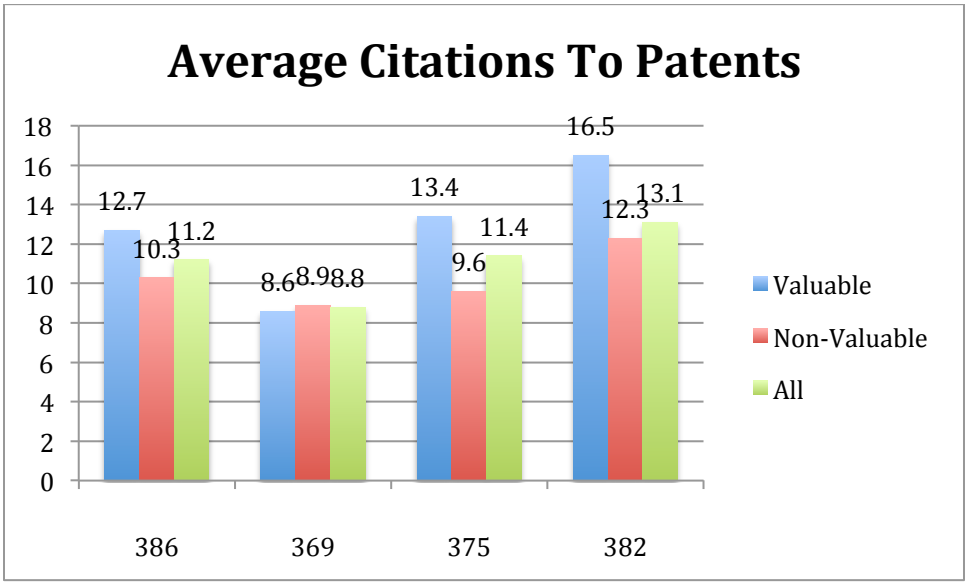


Figure 3

Overall, the patents contained in the valuable sets have more patents referencing them on average, and therefore more important, than those patents in the non-valuable sets.

Using citation data alone to evaluate our system's computational quality is not sufficient since citation analysis poses some limitations such as patent data truncation due to time lag and changes in citation rates over time. We should then look at measurements unique to information retrieval, precision and recall.

## Patent Pools

In the case of a human conducting the document evaluation, the human will be able to select the good documents from the not so good ones, with high precision. However, for him to have good recall where he must be comprehensive in his analysis and find all documents that may be of value to him, it will be difficult or impossible if the document set contains more than thousands. On the other hand, the advantage of using a computer to calculate and separate documents into groups is that it is able to exhaustively process each and every document to determine the

quality. When this same computer is able to also perform with high precision in addition to high recall, this becomes a powerful and efficient tool to process and analyze large amounts of data that would otherwise be impossible to do with manpower alone.

To gather these metrics, we can still use the same method as described previously to limit and procure our set of top documents and then compare against established patent pools covering similar technological areas as a measuring stick for "valuableness".

Patent pools represent groups of related patents by technology and concept that have been submitted through extensive expert review and if these pools are well governed, comprise of essential patents that jointly establish a common technological standard with no viable alternatives. Any company wishing to implement any portion of the technology must obtain a license from the pool. As a result, there is an inherent quality to these patents that other groups of patents may not have.

By comparing the overlap of patent documents between our calculated set and the patent pool set, we can measure the effectiveness of the Patentics engine in determining a document's relative value. A control, or reference, set can be obtained by selecting a patent at random for each patent in the pool set that has the same classification and same application date. These patents should then represent the average state of the technology at the same time the patent from the pool was written.

We have selected a few well-known patent pools: DVD6C and MPEG LA, to use in our evaluations as well as the previously listed classifications: 386, 369, 375, and 382. The following table 4 provides more detail into the patent pools.

| | |
|---|---|
| **DVD6C** | Patents required to produce DVD discs, players, drives, recorders, decoders, and encoders. |
| **MPEG LA** | Patents required for use of the MPEG-2, MPEG-4 Visual (Part 2), IEEE 1394, VC-1, ATSC, and AVC/H.264 standards. |

Table 4

With the appropriate combinations of patent pool and classification document sets, we have counted the overlap and calculated the precision and recall metrics for each data set combination.

## Patent Pool Precision

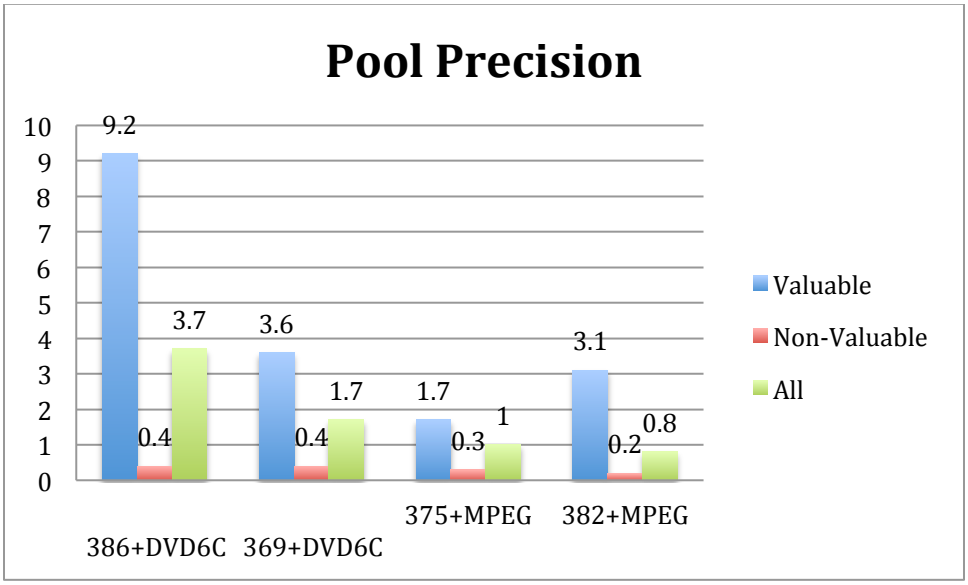


Figure 5

In figure 5 above, the blue bars represent the precision percentage of the calculated valuable set for specific classifications with a corresponding pool set and the red bars represent the non-valuable set (the green bars represent the entire document set for a single classification). Most importantly, the overwhelming majority of patents in the pools were also found to be in the valuable sets.

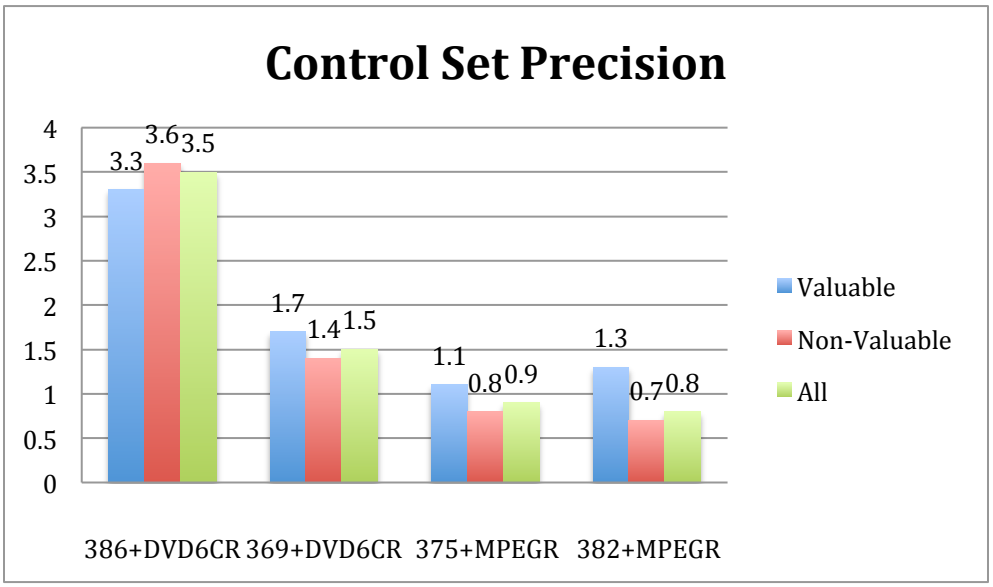


Figure 6

In comparison, when we use the control set, which contains a mixture of various patent quality, to measure the precision of the calculated valuable and non-valuable sets as shown in figure 6, the patent overlap is less biased to any particular set.

Using a reference set as a control variable allows us to show that using patent pools does not unfairly manipulate the data and that these trends apply universally to any other groups of patents.

## Patent Pool Recall

Using precision to measure the effectiveness of the evaluation is only one side of the coin. We should also look at the recall factor to determine how effective Patentics is able to filter out as many valuable patent documents that are available in the full document set.

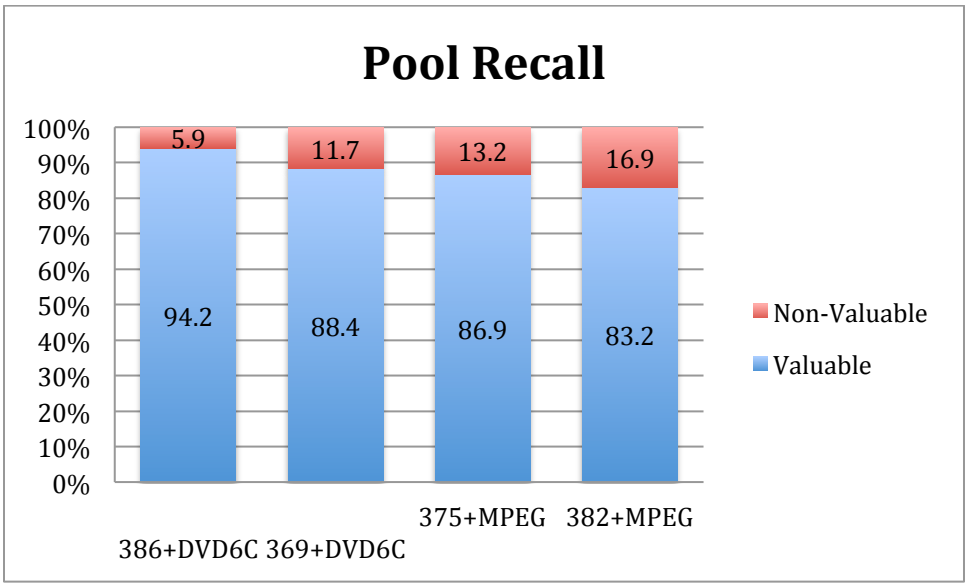


Figure 7

In figure 7 above, we can see that close to 90% of the patent pool documents are returned in the valuable set. This clearly demonstrates that in terms of exhaustiveness, Patentics is able to find a large majority of valuable documents and mark them as such.

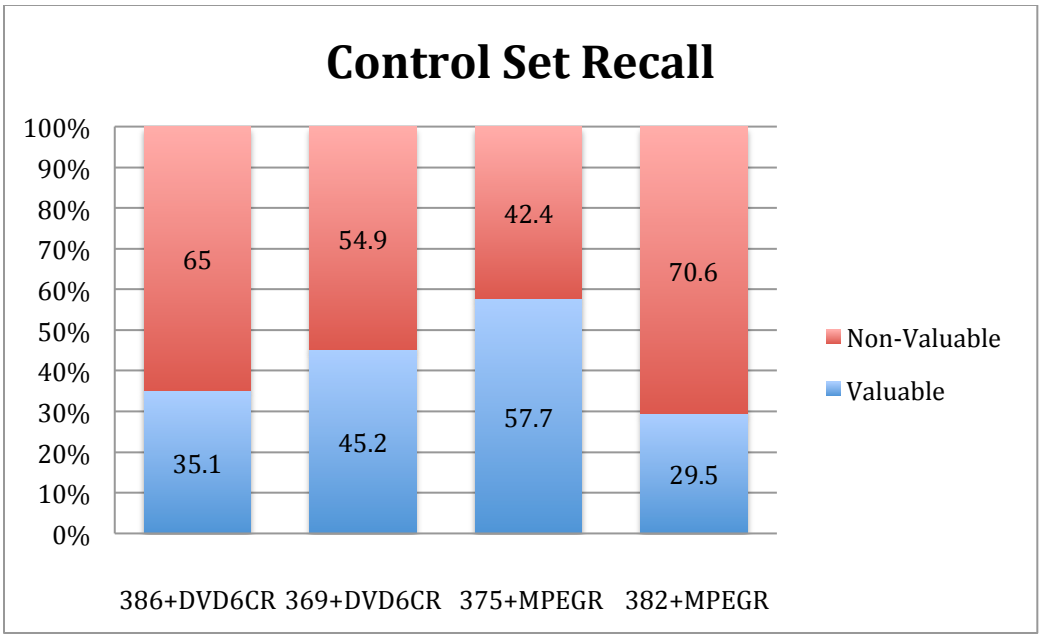


Figure 8

If we look at the recall metric for the control set in figure 8 above, we can see that the non-valuable sets contain a majority of the control set. This shows that Patentics is also able to do the inverse by distinguishing a non-valuable, or average, document from a more valuable document during its evaluation.

## Patent Pool Family Patents

Another interesting metric to look at is the number of family patents on average to which each patent in the calculated sets are linked, as shown in figure 9 below.

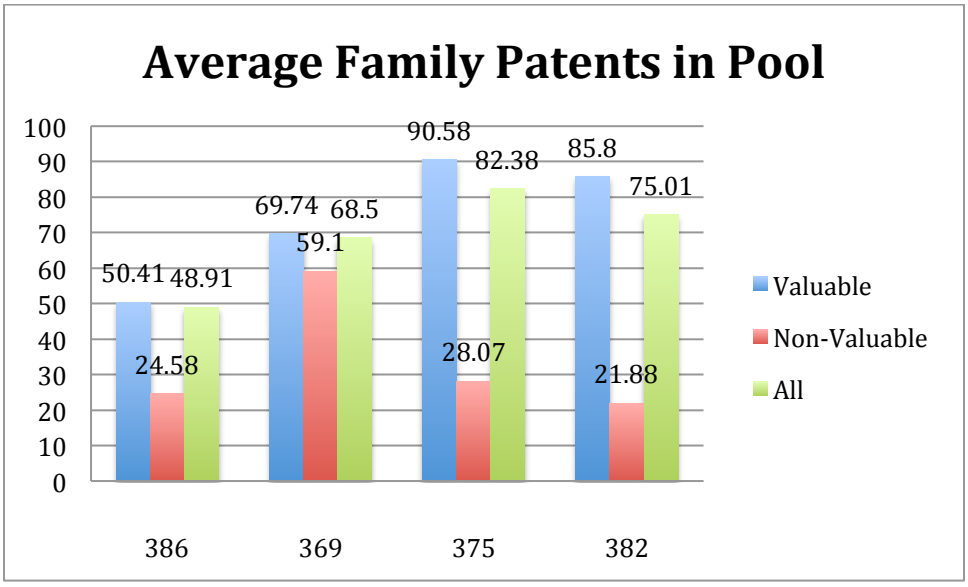


Figure 9

A family patent is the same patent that has been applied for in different countries in possibly a different language. The more family patents there are, the more the company values the patent, enough for them to expend additional resources to cover the invention in other countries. As such, patents with a high family patent count should be of higher value. In the figure above, the valuable patents that overlap with the patent pool have significantly higher family patent averages in comparison to the non-valuable sets. What is more telling is how the average family patent metrics trend remains consistent when applied to the overlap between the control patent sets, as shown in figure 10 below.

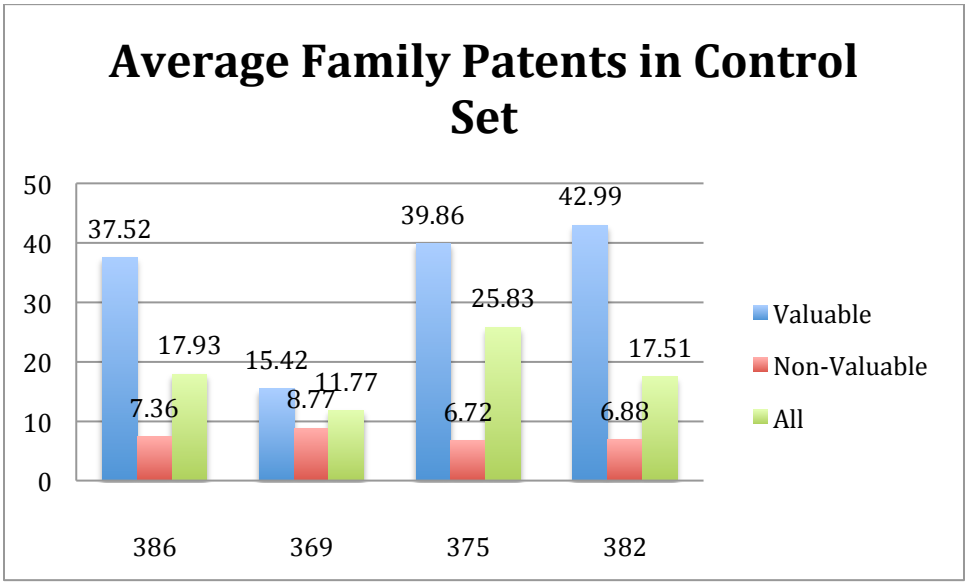


Figure 10

In comparing the results with the control set patents, the trend is shown to be universal and not just applicable to patent pools: the patents that Patentics has calculated as having more value still display a higher family patent average than the non-valuable patents. This shows that the algorithm is able to discern an inherent value that a company places on a patent, regardless of the actual patent quality. This is an important ability to have since quality and value can be subjective to a human, though not to an impartial computer.

## Conclusion

In the above tables, we have shown that Patentics is able to divide a document set into valuable and non-valuable sections with high precision and recall. In conducting our tests, we have chosen to use patent pools to validate our machine judgment capability because we believe our algorithm to be adept and strong enough to withstand a test against standards that are powered by humans. This provides us with an opportunity to solve real world problems with real world data that is more complex than any previous judging standards could provide.

Challenging other machine processes and engines in computing value and relevance is not enough; we are now pushing to match human logic and intellect.

There are many interesting applications for this technology. As an example, some companies apply for more patents than are put into practice so when it comes time to allocating resources to defending and prosecuting valuable patents, they need to evaluate which intellectual property rights to protect, specifically ones that generate the most revenue. Without expending too much effort, they can then search for valid patent implementations in their sales and marketing documents that align with their intellectual property. There is currently no workable technology service that can solve this problem; Patentics can and we have demonstrated how well it is able to find these documents with barely an instruction or human intervention.