**RJ10428 (A0803-004) March 4, 2008**
**Computer Science**

# IBM Research Report

## COA: Finding Novel Patents through Text Analysis

**Mohammad Al Hasan**
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180

**W. Scott Spangler, Thomas D. Griffin, Alfredo Alba**
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

**Research Division**
**Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich**

# COA: Finding Novel Patents through Text Analysis*

Mohammad Al Hasan[1]†, W. Scott Spangler[2], Thomas D. Griffin[2], and Alfredo Alba[2]
[1]Dept. of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, 12180
[2]IBM Almaden Research Center, San Jose, CA 95120
[1]alhasan@cs.rpi.edu,[2]{spangles@almaden, tdg@us, aalba@us}.ibm.com

## ABSTRACT

In the last two decades, the value of patents have increased enormously. As a result, companies are showing higher propensity to patent their invention and to strengthen their patent portfolio. However, large portfolios are difficult to manage. One major objective in portfolio management is to rank the patents in terms of their values. Since intellectual property (IP) attorneys' time is very expensive, an automated or semi-automated software system that expedites and assists the ranking process would be of great value. The existing software systems, targeted at IP professionals, mostly provide web-based services, data feed, advanced patent search interfaces etc. These are very helpful to commend a *prior art* search or to obtain answers to basic patent related inquiries but are not adequate to assess the value of a patent. Through our research, we build a patent ranking software, named COA (Claim Originality Analysis) that rates a patent based on its novelty. It computes novelty by measuring the *impact* and the *recency* of the important phrases that appear in the "claims" section of a patent. In our experiments, we found that COA produces meaningful ranking when comparing it with other indirect patent evaluation metrics— citation count, patent status, and attorney's rating. In real-life settings, this tool was used by beta-testers in the IBM IP department. Lawyers found it very useful in patent rating, specifically, in highlighting potentially valuable patents in a patent cluster. In this article, we describe the ranking techniques and system architecture of COA. We also present the results that validate its effectiveness.

*This material is based upon work funded in whole or in part by International Business Machines (IBM) and any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of IBM.

†Part of this work was done in the summer of 2006 and 2007, when the first author was a research intern at IBM Almaden Research Center

## Keywords

document ranking, information retrieval, patent processing, patent visualization

## 1. INTRODUCTION AND BACKGROUND

In the last two decades, the value of patents as intellectual property have increased tremendously. US legal systems have also shown pro-patent stance in many recent patent-based litigations [2]. So, now-a-days companies are much more aggressive in patenting their inventions and in building a substantial patent portfolio. From the 2006 fiscal year report of USPTO [23], 443652 patents were filed in the year of 2006, which is about 10% more than that of previous year [18].

Patents protect the invention; they also provide the inventor an opportunity to generate revenue by means of licensing them. In the technology industry, the research-driven companies like IBM, Sony, Intel, Mitsubishi and etc. earn excess of hundreds of millions of dollars yearly just from patent licensing revenue and this trend is rising. Moreover, a comprehensive patent portfolio gives a company the competitive edge in the market, especially when handling business transactions like mergers, acquisitions or even emerging products marketing.

As the patent portfolio of a company grows, it becomes increasingly difficult to manage. Firstly, the company needs to pay maintenance fees to the patent office for each patent in its portfolio. But, many patents in the portfolio may become obsolete, due to numerous reasons, like— industry and technology trend changes, availability of alternative technology, changes in the company's growth plan, strategy, etc. Hence, it needs to find a value ranking of the patents in its portfolio to optimize this cost. Secondly, companies need to identify the fundamental patents in their portfolio to optimize the legal cost while searching for possible infringement (by third parties) or exploring new licensing revenue opportunities (or both) [3, 19, 25]. Moreover, business activities such as mergers, acquisitions, opening of new lines of business, etc. require identifying fundamental patents of the partner companies and competitors in the desired business area to effectively evaluate the business prospects. In present days, these tasks are accomplished mostly by IP professionals and patent analysts.

However, the evaluation of patents by patent attorneys is more direct. They measure the legal strength of the claims of a patent; for instance, how broad the claims are, on what product(s) they "read on", how they interpret in the scope of legal linguistic, and etc. All these are very important as they

are used in the formal processes of patent litigation. But, there is another dimension in patent evaluation—its novelty and non-obviousness. Lawyers are not the most competent people to measure this effectively, since they are not skillful on the core technology. They usually depend on experts (or the inventors) in this regard. Since attorneys' and agents' time is very expensive, an advanced software tool (be automated or semi-automated) is required to expedite the process.

Identifying a patent's novelty is also significant for public sector agencies like USPTO since novelty is a prerequisite for a patent to be granted. But today's high-pace technological invention environment can easily defeat anyone in his endeavor towards being well-informed regarding the state-of-the-art of a technical field. As a result, there is always a chance to miss some significant *prior arts* while assessing the novelty of a pending patent application. An increased number of patent applications acerbates the situation. Therefore, human evaluation needs to be complemented by effective software based application in this task.

Novelty assessment is more difficult for patents on *software or business method*, but, they comprise a large fraction of patents issued in recent years [4, 28]. In fact, many granted patents in these areas received severe criticism as they did not seem to be novel [4]. Since they are composed as a sequence of business processes, comparison to *prior arts* is difficult. Effective assessment techniques need to be invented that work well for these kinds of patents.

We develop a software system, named, "Claims Originality Analysis (COA)", to address the problems described above. It assesses a patent by evaluating the originality of its invention. COA is fundamentally different from any concurrent patent analysis tool. It uses an information retrieval approach, where a patent is considered valuable, if the invention presented in the patent is novel and also, is subsequently used or expanded by later patents. This knowledge is gleaned from the patent text, specifically, from the text composing the patent claims. From the "claims" section of a patent, we first identify a set of phrases (single word or multi-word) that retain the key ideas of the patent. For every phrase, we then find the earliest patent that had used that phrase. We also track the usages of that phrase by later patents. Finally, we feed these information into a ranking function to obtain a numeric value that denotes the novelty of that patent.

We validate the performances of COA with the patents in IBM patent portfolio. We find that the patents with high COA rating are mostly those that have positive status [1]. It is also observed that these patents have been cited more often compared to other patents that have low COA rating. COA's ranking criteria is particularly useful for the patents on software and business methods for which the analysis of novelty is difficult and ambiguous. Besides portfolio evaluation, COA features can also be useful to identify *prior arts* when evaluating the merit of a new invention. This method is also general enough to be used in ranking other technical documents.

COA is developed as a Java application. It uses **DB2** databases for back-end data store, together with **Lucene** to index the textual phrases. **SOLR** is used as the search server that communicates between COA Java application and Lucene. COA is integrated with the BIW (Business Insights Workbench) software [29, 17]. It provides the following features:

- It rates a patent from the novelty perspective, by using techniques from information retrieval domain. The texts of patent claims are used for this purpose.
- The system is mostly automatic. However, expert opinion from human is indispensable for any patent analysis tool. Hence, our system provides the option to incorporate human knowledge in all different aspects of the system.
- It provides innovative ways to visualize a patent that reveal inherent information of a patent's rank status. From this, an analyst is informed about the reason why a patent is ranked high or low. That facilitates the option for further adjustment of the ranking criteria.

## 1.1 Patent assessment Challenges

Accurately assessing a patent's license value is a difficult task for an expert, let alone, for an automated software system. It not only depends on the patent; but sometimes depends on assignee, assignee's patent portfolio, and on other complex economic factors. Some economic research [2] suggested that the true values of patents are not revealed until such rights are held valid by the courts. However, there are many research efforts to outline the major criteria to assess the value of a patent [25, 19, 1, 5]. In a recent work, Wang et. al. [7] summarized those in three broad categories: (1) Patent Strategic Value, (2) Patent Protection Value, and (3) Patent Application Value. The first category determines the novelty of the invention and its impact on the technology market in near future. The second category evaluates patents from its protection value, i.e. it mostly assesses the property that a patent protects through its claims. The last category–Patent Application Value, mainly considers the breadth of the patent's applications in the relevant industry.

Our ranking method is limited to evaluating the patent's strategic value; that sums to measuring the novelty and impact of a patent. Though other aspects of evaluation are equally important, we found that they are too difficult to handle by a software system. For instance, to evaluate the protection value of a patent, the analyst needs to find its claim elements and their scope, the strength of the claim language to protect the claim elements and other legal measures of the patent claims. These tasks require software systems with the ability to understand the claim language semantically. Unfortunately, current techniques of NLP (Natural Language Processing) are not adequate for this purpose. They are usually trained on newspaper based corpus [8] and perform very poorly for a patent document. Finally, estimating the patent's application value is completely outside the scope of a data mining domain, and is more appropriate topic for industrial economics and market strategy research.

## 1.2 Structure of a Patent Document

Patent text is very different from the ordinary newspaper text and text analytic tools that analyze a patent, need to be aware of its structure to achieve high performance. In this section, we provide a brief overview of the important sections of a patent document. Readers can get more information on

[1] patents that IBM continues to maintain

this from US Patent and Trademark office (USPTO) web site [23] or books on IP law [24].

Every patent has a section, titled, "Description of the Invention". It includes a brief abstract of the invention followed by a longer description. The description must detail the best way of making and using the invention that the inventor is aware of, at the time of the patent application. It also includes relevant figures and flow-charts of the invention described in the patent.

Then usually comes the "Claims" section, where claims are listed with a numeric label to each of them. They are the most significant part of the patent as they define those aspects of the invention that are protected by the patent. Note that, it is not possible to determine what is protected by the patent from its title, abstract, or description; one must read the claims. Claim describes the invention, by listing its constituent parts (in case the invention is a device of apparatus) or by listing its method sequences (for business process or software-based invention). The most important concept in understanding a claim is whether the claim *reads on* something. A claim reads on a physical object or on a process when all the elements of the claim are component of that object or process. For instance, if a hypothetical claim begins as follow: "A device $X$ comprising $A$, $B$ and $C$ ...", then this claim reads on all devices which are of type $X$ and have $A$, $B$ and $C$. Robust claim structure is an important property for a good patent. Moreover, claim drafting is an important issue as well, since, choices of words (that are more specific), using poor language, etc. can generate claims that have very narrow scope and henceforth can diminish the value of a patent.

### 1.3 Patent Classes

Every patent is assigned a class label based on its subject matter. A class generally delineates one technology from another. Classes may be further divided into subclasses to delineate processes, structural features and functional features of the subject matter. In COA, class-code of a patent plays an important role. While comparing the novelty of a patent from the patent text, we generally restrict the search in patents belonging to the same class.

## 2. COA: METHODOLOGIES

We measure a patent's novelty by two factors: its earliness and impact. Earliness factor concerns about the fact whether the patent is one of the earliest patents in terms of its technical matter or in its application domain. It thus attempts to rank a patent in comparison to other patents that are forerunners in the same or closely related technical domains. On the other hand, the impact measures the influences that the patent makes on the other related inventions over the course of time since it was born. Among these two, "earliness" has not been used much in the information retrieval or data mining research. However, for patent documents, earliness is indeed more important than impact because of the *prior arts* rule. For a patent $X$, its value diminishes enormously if an earlier work is found that uses the invention (or a significant part of it) that is described in $X$. The latter factor, i.e. the impact has been extensively used in document ranking. For instance, the web page ranking algorithms [11, 13] that are used in search engines assign a higher rank for documents with many hyper-links pointing to it. In the academic research domain, citation statistics or bibliometrics are used to identify an influential document. In the above cases, the number of hyper-links or citations measure the impact that the document has made since it was created. Unfortunately for patents, citations can be poorly drafted by the additions of many auxiliary citations by the patent examiners [21] or sometimes by intentional omissions by some inventors. So, alternate data need to be sought to measure the novelty of a patent.

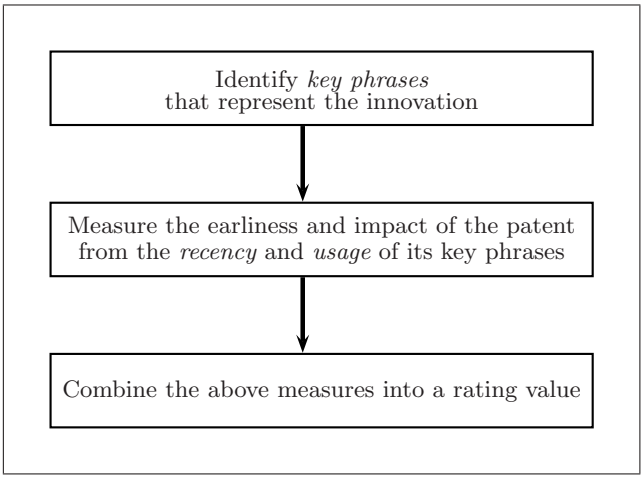


**Figure 1: Steps to obtain COA rating of a patent from its novelty**

COA uses the text data in the patent to measure its novelty. Since a patent is about a new innovation, it should have a solid contribution on top of the existing *prior arts* and the patent text should reflect this contribution to a considerable extent. Since the claim section describes the novelty of the invention in the format of formal claims, we use only the text of the claims. It reduces the noise by not considering the terms or phrases that are only tangentially related to the main contribution of a patent. The option to use the entire text is also available, if required. To enable us to use the text mining techniques, we represent the claim section of a patent as a vector of technical phrases composed of a set of consecutive words upto length three. By discarding the complex linguistic structure of the patent document, we avoid NLP based techniques that usually perform poorly for patent documents. There are three distinct steps in our process, as shown in figure 1. We describe each of these step next.

### 2.1 Identifying Key Phrases

A new innovation usually comes with its own technical terms, phrases and keywords. They are also frequently used in the claim section to describe the patented invention that is protected through the claims. For example, if a patent innovates the back-propagation as a neural network based learning technique, the set {**neural network, back-propagation, hidden layer, weight, neuron, weight vector**} can be a potential set of frequently used terms that represent the key phrases of the invention. So, finding this set is the first step in measuring the novelty of the patent. To distinguish the key phrases from other commonly occurring phrases, we built (off-line) a background dictionary for every patent-class. It constitutes the phrases that appeared frequently in many different patents of that class and thus can be dis-

carded as background noise. The rest of the jobs are done online for a given patent whose rating is to be determined.

We use a simple $n$-gram method to extract phrases from the claim text. In this method, we first remove the stop-words from the claim-text. A simple stemming algorithm is also used to discard redundant phrases. Then, we construct all consecutive words upto length three. Each such word is a prospective phrase and its contribution to the patent's value is computed as follows:

$$Contribution(T) = max\left(\frac{support(T)-2}{age-in-days(T)+1}, 0\right)$$

In the above equation, $T$ is a phrase, and **support**($T$) and **age-in-days**($T$) are its frequency and age, respectively. The function **support** and **age-in-days** are discussed elaborately in the next subsection. We added a one to the denominator to avoid an infinity value for the contribution. A value of two is subtracted from the support in the numerator to ensure that the support value is at least 3, otherwise the contribution of that term to the patent's value is zero. Now, if the contribution value of a term falls bellow a threshold, the term is discarded. The threshold value is fixed for a class-code and is computed empirically by taking random samples of patents from that class. The set of terms whose scores pass the threshold value constitutes the set of key phrases for that patent.

An alternate way to extract key phrases is to use the POS (parts-of-speech) tagger. This is popular is traditional information retrieval domain. We found that such methodologies, although select good terms, miss a lot of important terms. So, POS tagger based techniques, although implemented, are not used in the final version of the system.

Depending on the patent, the above method may generate too many or too few key phrases. So, we allow a user to define a time-window, which is used as follows. Only the key phrases that first appeared in some patents published within the given time-window are considered. A zero length time-window considers only those key phrases that are used for the first time in the patent that we are evaluating. Selecting a higher value for the window length allows more terms to enter into the key phrase list.

## 2.2 The earliness and impact of a patent through phrases

The *earliness* of a patent can be measured by finding the recency of the key phrases in its innovation. Thus, if a patent is early in some technology domain, it uses lot of novel phrases and most likely has significant contributions beyond the available *prior arts*. Like, for the previous example, if a patent uses the phrase "back-propagation" for the first time in the patent literature, it most likely had invented the back-propagation algorithm for training a neural network. The age of a phrase is the inverse of its recency and is defined as the time difference between this patent's (the one that we are evaluating) publish time and any earlier patent's publish time that used the corresponding key phrase for the first time. The earlier patent's class-code needs to be the same as this patent's class-code so that the technical meanings of the phrase are comparable across different patents. If the time difference decreases, the age increases and subsequently, the recency of that key phrase also increases.

However, adopting numerous new phrases does not necessarily imply that the innovation is useful. So, we also compute the frequency of a phrase by counting its usage by other patents in the same class. We call it the *support* of a phrase. Many phrases with high recency and high support imply that the invention described by the patent is novel and also useful. Note that, the support of a phrase also depends on the patent's publish date. A patent that is published very recently may be very innovative although its support value is small. So, we normalize the support value appropriately to consider this fact.

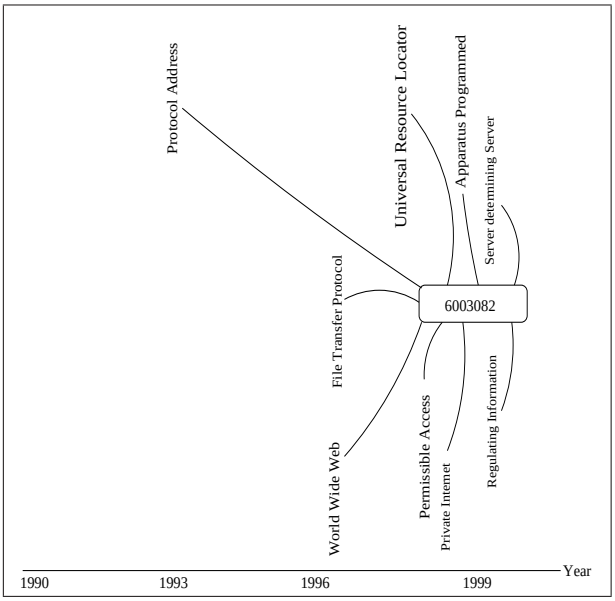


**Figure 2: A novel patent with some of its key phrases**

Figure 2 shows a patent (number 6003082) with some of its key phrases. The horizontal line at the lowest part of the figure is the year axis that increases from left to right (from 1990-1999). The rectangular box denoting the patent 6003082 is vertically aligned with the year 1999 to indicate that the patent was published in that year. The key phrases are similarly aligned along the year line when they were first used by some other same-class patents. Font size of a phrase are roughly proportional to its support. From the figure we can see that many of the phrases that are used by this patent are very recent, mostly within the three years time window; some of those phrases were also first used by this patent. So in COA the ranking of this patent will be high.

## 2.3 Obtaining rating scores

We obtain rating scores that enable us to compare the novelty of a patent with respect to others. A higher rating score generally infers that the patent is novel. Rating scores are always positive. Two simple scores are computed. The first score is just the linear sum of the individual contributions of each of the important phrases of a patent. The contribution value is computed the same way as discussed in section 2.1. The second score is just the count of the important phrases. These scores are called **COA score1** and **COA score2** respectively, in the later sections.

A rating value obtained above is just a way to obtain a fast estimate of a patent's novelty in comparison to other related patents. What is more important than the value is to find the justification of an obtained value. For instance, if we

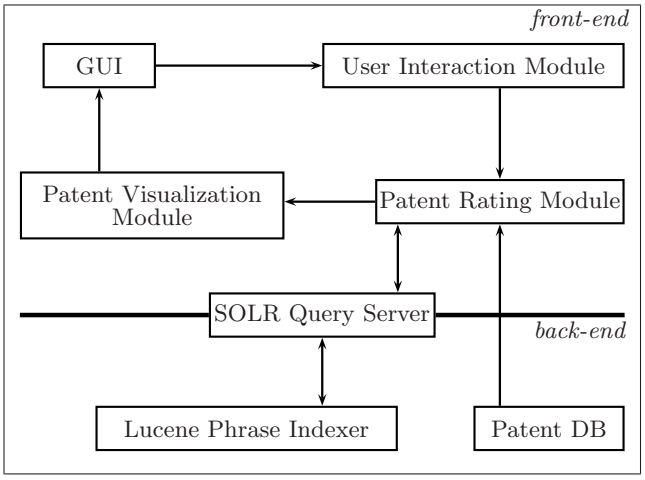


**Figure 3: Different Architectural components of COA**

find that a patent has very high value, we should also notice that there exist some phrases that have very high support and high recency. So, generally a patent rating table like that in figure 4 is presented to the analyst. We will discuss more about the rating table in later sections.

## 3. COA:ARCHITECTURE

The patent rating system has the following modules

- Database and Lucene Text Indexer
- SOLR query server
- Rating Module
- User Interaction Module
- Visualization Module

Figure 3 shows an architectural block diagram of COA. It denotes the different modules in labeled rectangular boxes. It also shows the data flows among the modules by arrows. The diagram is partitioned into two parts: front-end and back-end. The back-end manages the data and the front end implements the application logic and the user interface.

### *Database and Lucene Text Indexer.*

A DB2 database and Lucene search engine comprise the Back-end of the COA application. Entire patent data (both structured and text fields) is stored in the database. Patent number, title, assignee name, inventors name, publication date, filing date, cited-by, references etc. are some of the structured fields. The text information, like description and claims, are stored as CLOBS (Character large objects). To facilitate search in these text fields we used Lucene [30]. We built a Lucene index on patent class-code and publish date. Using this index, the recency and support of any phrase can be obtained instantly. The Front end Java application provides a phrase and a class-code value, and the Lucene search engine returns all the patents that have that class-code and that contain the given phrase in the claim section. Lucene also takes care of stop-word removal and stemming with its built-in functionalities. Besides these, the COA back-end also has background dictionaries as flat text files that contain frequently occurring words of different patent classes. While finding key phrases, these words are discarded and hence do not contribute to the patent rating.

### *SOLR query server.*

We used SOLR [31] query server to mediate between front end of COA and the Lucene index, see Figure 3. The Front end application builds a query in the SOLR query language and sends it to SOLR, which communicates with Lucene, prepares the result in XML format and sends the result back to the front-end application.

### *Patent Rating module.*

This module implements the rating algorithm that we described in the section 2. It accepts a patent number and optionally accepts a list of parameter values. It communicates with the back-end to retrieve the claim-text and the class-code of the patent. Then the key phrases are extracted by using the $n$-gram method. For each of the key phrases, this module builds a SOLR query to search the phrase in the patents of the same class. Once the result is achieved, an XML parser is used to parse the result and calculate the recency and support of all the key phrases. Then it computes the rating value using the linear rating equation. The rating value, key phrases and other information are sent to the visualization module to prepare the presentation.

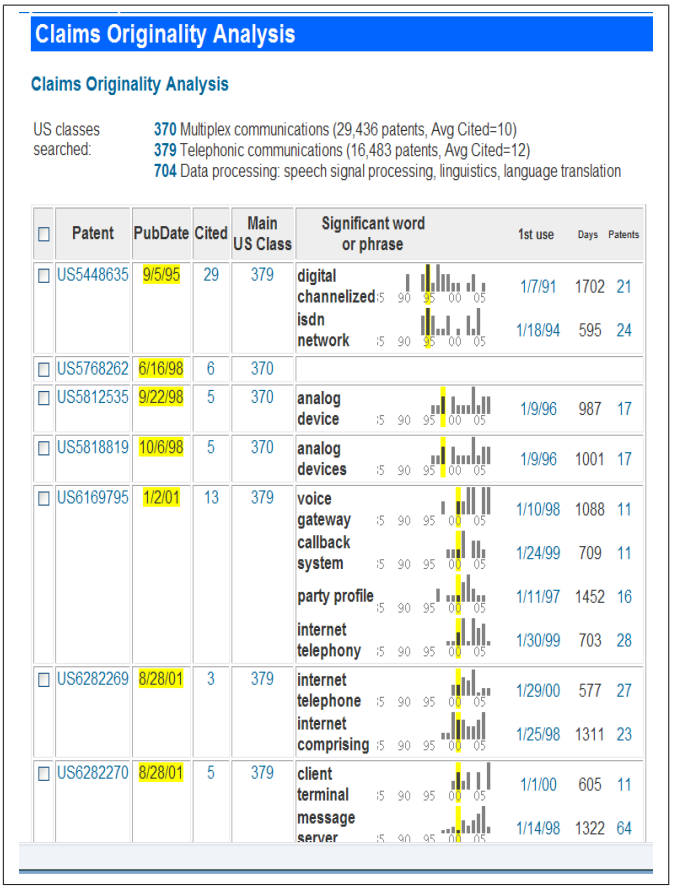


**Figure 4: Patent Rating Table**

### *Visualization Module.*

This module displays the result in a user friendly manner

that can help a patent analyst to efficiently evaluate a patent . We consulted with patent analysts to identify their evaluation methodologies and produced the visualization tools that would assist them the most. A patent table is generated which is displayed in a browser window, as shown in figure 4. Since patent analysis is generally performed over a set of related patents, the rating table is designed to display a summary analysis of all the patents in the given set. The rating result of each patent is listed in one row of the table. The columns contain patent number, publish date, classcode, citation count, key phrases, and the rating value. For each of the key phrases, we also show the first use date, the day difference (inverse of recency) and the support value for that phrase. For instance, from table 4 we notice that, while ranking patent 5448635, one of the important phrases is *isdn network*, it was first used in a patent published in 1/18/1994, which is 1702 days earlier than this patent. The support of the term is 21 patents, i.e. after the first use, the term has been subsequently used in 21 distinct patents.

The rating table also provides effective navigation capability by hyper-linking the objects of different columns to other relevant objects. For instance, the "first use" date of any phrase is hyper-linked to the text of the patent that used that phrase for the first time. So, an analyst can quickly get the context in which the phrase was actually used in the earlier patent. In Figure 5 we describe the different hyper-links that were used with different columns of the rating table.

When displaying the text of a patent in the browser, we highlight the key phrases in different colors. Furthermore, the font size of those words are varied according to the recency of the word; i.e., the size is inversely proportional to the value of the date difference column in the patent rating table. Moreover, the phrase is linked to the patent, where this word appeared first. Figure 6 shows an sample patent in a browser window of the client machine.

### *User Interaction Module.*

The user interaction module allows a user to alter the default setting of different modules of COA. In the following, we describe a few important interactions.

**Edit Background dictionary** The user can view the terms that are listed in the background dictionary. These terms are ignored by the COA rating module when rating a patent. The user interaction module provides the user the option to add(remove) words and phrases from the dictionary. These changes can be made permanent or can just be used for the current session.

**Edit phrase's contribution in rating** Contribution of a phrase towards a patent's rating can be made void from the patent rating table by clicking on it and disabling that phrase. The process can be undone as well. This is helpful to investigate a suspicious rating value for a patent.

**Thresholding support/recency** Different domains of technology have different measures of prior works; and, the sizes of their key phrases also vary significantly. The user interaction tool allows the user to change the threshold of the time interval for which the key phrases will be considered. For instance, for a four year threshold, all the phrases that appeared first in patents within the previous four years of this patent's publish time will be considered in the key phrase set. The key phrase set can also be filtered by setting a minimum support count. For example, if the minimum support count is set to 20, a term shall not be considered in the key phrase unless it has been used by at least 20 other patents after the first use.

## 4. COA: RESULTS

In this section, we present some numerical results to validate the effectiveness of COA rating. Generally, such validation is difficult for patent data, as no gold standard exists and the business value is not publicly available. For recent patents, the business values are uncertain. For relatively aged patents, economic scientists have found some indirect measures that are somewhat correlated with actual monetary values. Patent citation count [5] and patent status [1] are two of those. The former is the number of citations that a patent receives from any other patents. The latter (patent status) denotes whether the patent is still maintained by its assignee through regular payments of the license fees. We also had at our disposal, IBM confidential attorney ratings of many patents that IBM owns. The effectiveness of COA is established by comparing its scores with these alternate quality measurements. Note that, the correlation of these measures with the actual patent value is considerably noisy and does not hold for many patents, but they provide a viable option for us to cross-validate the result that we obtain from COA.

Most of our experiments were performed on IBM patent portfolios. IBM owns more than 40,000 patents [16], in more than twenty different classes; from which, we picked a set of different portfolios related to software technology or business process. Typically, one such portfolio contains 50-100 patents. For each patent in these portfolios, COA scores are computed and recorded. We also collected the data related to patent status, citation count and attorney rating for each of these patents. Results are discussed in the following paragraphs. The actual patent numbers are not shown in any of the results, because this information is confidential.

Figure 7 shows a scatter plot of 95 patents from one of IBM portfolios. Each small circle denotes a patent and its $x$ and $y$ co-ordinate values represent its COA rating (score1) and the citation count respectively. For instance, the circle at position (119,33) represents that this patent has COA score 119 and it has cited by 33 other patents. From the figure a positive correlation between these two metrics is evident, although it is quite noisy. There are some patents for which we have high citation but low COA rating. The same behavior of noisy correlation between patent citation and its value was also observed in previous research [5]. In this figure, we also show the linear regression line for these scatter points. A positive slope of this regression line confirms the existence of positive correlation. The computed Pearson correlation value is 0.33. The p-value for testing the hypothesis, "there is no correlation" against the alternative, "there is a positive correlation" is .0005. So, the null hypothesis can be rejected since the probability that such a correlation in the data will be seen (assuming that they are uncorrelated) is only .0005. The scatter plot of the same dataset that compares COA rating (score2) with patent citation is similar, hence not shown. However, the Pearson correlation value between **COA score2** and citation is 0.42 with a p-value of .0004 for this dataset, which is better than that of COA score1. In fact, for most of the dataset, score2 shows stronger correlation with citation count compare to that of score1.

Our second criterion to validate COA is to compare COA

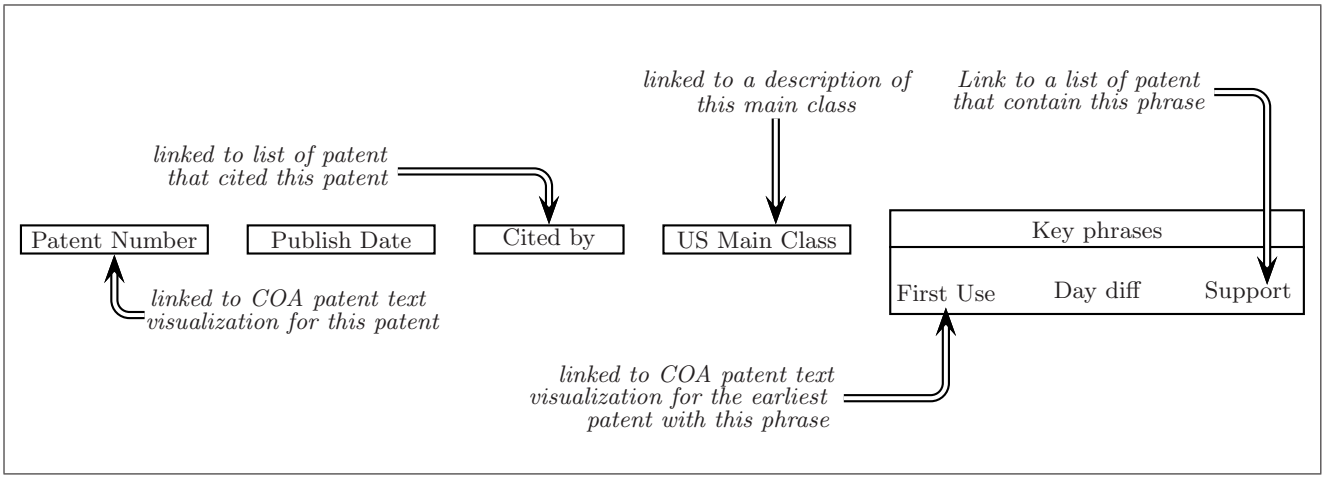


**Figure 5: Hyper-links from different columns of patent rating table**

**Table 2: Comparison of patents with attorey rating**

| | Dataset 1 | | Dataset 2 | | Dataset 3 | |
|---|---|---|---|---|---|---|
| | **COA (score1)** | **COA (score2)** | **COA (score1)** | **COA (score2)** | **COA (score1)** | **COA (score2)** |
| *Rating 1* | 18.77 | 40.88 | 12.73 | 27.34 | 12.73 | 31.32 |
| *Rating 3* | 4.20 | 10.82 | 7.22 | 16.89 | 4.49 | 10.01 |

**Table 1: Comparison of active and lapsed patents**

| | **COA (score1)** | **COA (score2)** |
|---|---|---|
| *Current* | 4.20 | 8.41 |
| *Lapsed* | 1.51 | 2.89 |

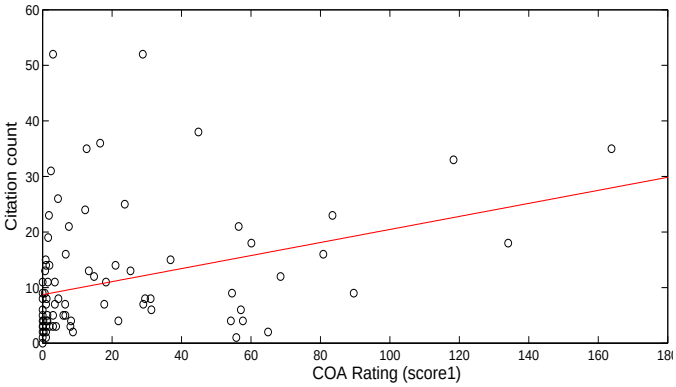


**Figure 7: Citation count vs COA rating value**

scores with the status of a patent. We considered two status values: *current* and *lapsed*. Obviously, a patent with current status is more valuable. But, for a lapsed patent it is not necessarily true that it is not valuable; because, it can happen that a patent had lapsed only for the latest part of its lifespan, but it was a valuable patent in the earlier lifespan. So, a meaningful comparison should consider a set of patents that have approximately similar age. We took a set of total 1544 patents whose numbers start with "61". Since, patent numbers are assigned in the order of acceptance, these set of patents had comparable ages. Also, they were relatively new patents, so any patent that had lapsed in this set did not finish its full term (i.e. not expired by age), rather, the assignee discontinued to pay the license fee for it. Out of 1544 patents in our experiment-list, 220 had expired. Table 1 compares the current and lapsed patents in terms of COA score1 and COA score2. It is easy to see that for both the COA scores, current patents have higher average values than the lapsed patents. To further establish the significance of COA scores in relation to the patent status, we performed the following statistical significance test. We partitioned the above set of patents in two different sets, this time based on COA scores. Set 1 consisted of patents with any of the COA score equal to zero and the set 2 consisted of the remaining patents. Sizes of these sets were 683 and 861, respectively. We expected to see more expired patents in set 1, as patents in this set had low COA scores. It was found that 126 members in this set were expired patents. With an assumption that COA scores and patent status are independent, the above value should had been $\frac{220*683}{1544} = 97$. Since, the obtained value of 126 was well above 97, these two variables can not be independent. A Chi-Square test, like $Chi - Squared(1544, 220, 683, 126)$ yielded a value of 0.0026%, which suggested that with independence assumption, obtaining a value as high as 126 were only .0026% probable. Similar test with fisher probability (which is more exact) was only 0.0019%. Hence, there exists a significant relationship between having at least non-zero COA scores and not being a lapsed patent.

IBM IP department also has its own rating system that rates all patents into three different classes: 1(excellent), 2(good), and 3(not-so-good)–based on their merit. This rating is done for every new patent by the internal IP attorneys as soon as the patent is filed to the patent office. Since,

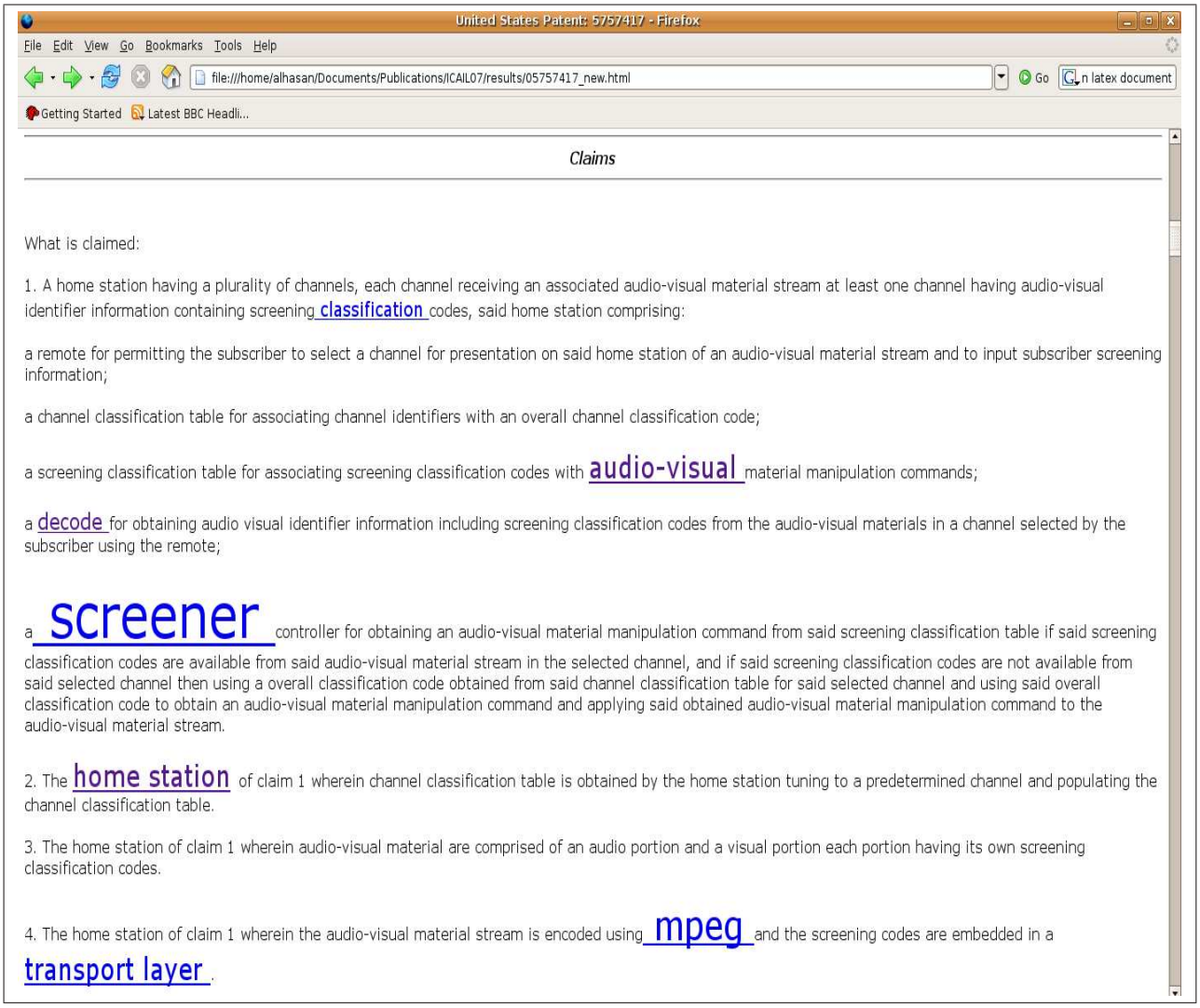


**Figure 6: A screen-shot of the Claims Originality patent view. The terms that are the most original in the claims are highlighted by using larger fonts. The font size also represents the degree of originality. Each term is also hyper-linked to the patent that used the term for the first time.**

the evaluation is made even before the patent is granted, it does not reflect the actual monetary benefit that was earned through the patent, rather it evaluates the novelty of the patent in comparison to the *prior arts*. So, this evaluation is complementary with respect to the evaluation that is based on the status of a patent since the status to some extent, depends on the actual income earned by a patent. We collected the attorney rating values of three different sets of patents, whose numbers start with "61", "62" and "63"; from there, we built three datasets by considering patents with rank 1 and 3 only. For each of these datasets, COA scores were computed and is shown in Table 2. In all the datasets, rating-1 patents rate highly over the rating-3 patents in both the COA scores.

To further evaluate the COA score1 and COA score2, we used them independently as a feature in a unit-feature supervised classification task that classifies patents into rating-1 and rating-3. An SM linear classifier with default parameter setting was used. A balanced dataset (equal number of rating-1 and rating-3 patents) was used for this job, so the baseline accuracy was 50%. The results are shown in table 3. We can see that, both COA scores1 and COA score2 achieved around 70% accuracy on dataset 2. On dataset 3, the accuracy is around 65% and 60%, respectively. The accuracy on dataset 1 is relatively lower, but well above the baseline. To compare COA scores with patent citation metric, patent citation was also used as a classification feature to perform the same job . Compared to both the COA scores, citation data performed poorly in this classification task (see row one in the table). It validates that the COA scores are better predictors in predicting a human based patent ranking score.

In practice, COA has been used in a preliminary fashion by beta-testers in the IBM IP licensing department. So far they have found it to be useful tool for highlighting potentially valuable patents in a patent cluster. It also helped them to identify the concepts in a patent that constitutes the most salient features.

## 5. RELATED WORK

**Table 3: Classification accuracy using COA scores to classify rating-1 and rating-3 patents**

| | Dataset 1 | Dataset 2 | Dataset 3 |
|---|---|---|---|
| *Patent Citation* | 59.75 | 54.84 | 51.82 |
| *COA score1* | 53.20 | 70.63 | 64.49 |
| *COA score2* | 59.76 | 69.79 | 60.15 |

In recent years, works on patent data have got much attention in industrial domain. But, the majority of these works [22] have been web-based services that are targeted towards corporate clients. These works mostly provide patent data feed (patent text, news, case update, etc.) and, sometimes, an infrastructure for the clients to run queries on patent data. Usually, these queries are on structured fields like class-code, file histories, assignees name, references and sometimes also on unstructured fields, like patent claims, description of invention, prior work etc. Few companies [14, 15] also provide web-based software tools that facilitate further analysis of the results obtained from these searches. They typically use clustering or summarizing techniques to find interesting patterns in patent data and then apply effective visualization techniques to display those. Works of these kinds can help in understanding the global picture of a collection of patents, such as discovering the trend of the innovation, to identify the industry leader in some technology, to identify the technology focus of some company and so on. But, they are not applicable in assessing an individual patent in terms of its value.

Finding a document's value is a well-studied research area, in the domain of information retrieval and text mining where majority of techniques use meta-data information, like hyperlink structure or citation information. Graph-based algorithms, like HITS [13] and Page-Rank [11] are the most successful in identifying the most useful document, especially in the domain of search engines. But, this approach is not well explored in patent data, most likely because of the poor quality in their reference and citation information. Nonetheless, some software tools [14] use the reference and citation information in patent data to form forward and backward reference graphs, which are very useful, specially for the *prior arts* search in the patent domain.

Our work assesses a patent's value from the patent text and we did not find any prior work on this. Recently, Shaparenko et. al. [12] proposed a method for identifying influential papers and authors from a collection of research papers that solely uses the text. They represent a document $d$ by a term vector in a TFIDF format and compute the nearest neighbor documents of $d$. The nearest neighbors are partitioned in two sets, $\mathcal{N}_{earlier}$ and $\mathcal{N}_{later}$, depending on whether they were published before or after $d$. The size of the first set is subtracted from the size of the second set and is used to evaluate the novelty. Intuitively, this approach is similar to our approach. The larger size of $\mathcal{N}_{later}$ corresponds to larger support of the key phrases in our approach and the smaller size of $\mathcal{N}_{earlier}$ corresponds to more novel phrases. But, our approach finds the specific phrases that contribute to the rating and provide options for subsequent user interactions. In another recent work [26], the authors used Gaussian mixture model of words in the text to model flows of ideas in documents. However, such a modeling is not appropriate when the relation is very noisy and secondly, choosing the parameter for the model is difficult as the intuitive meaning of the parameter value is difficult to comprehend by a patent analyst.

There are some excellent researches [7, 5, 1] in the domain of economics and management, that tried to identify factors that are influential in ranking a patent. Most of these are based on survey data and testing those data to verify some hypothesis. They do not offer any direct method for patent assessment.

There are few researches that solve other related problems in the patent domain. Tseng. et. al. [6] use patent text mining to understand the distribution of words and terms in different patent documents, which is useful for automating the patent categorization task. Yoon et. al. [27] built a text-mining based patent network which also uses patent text to identify the technology trend. Sheremetyeva et. al. [9, 10] have two distinct works that use statistical NLP (Natural Language Processing) and rule based technique to parse patent claim section. To learn more about other works related to patents, interested readers can read the papers in the ACL workshop on Patent Corpus Processing (2003).

## 6. DISCUSSION AND CONCLUSION

Patent ranking is a challenging task with numerous factors that determine its value. Hence, it would be too optimistic to expect a perfect ranking just by focusing on the text of the patents. But, if we just consider the novelty factor, COA works excellent. It produces a rating value that satisfactorily agrees with other indirect rating criteria. However, from the experiences of its users, its main appeal is not the rating value, rather the usability, flexibility, and versatility that it offers, when rating a patent. First and foremost, we provide the analysts, a system where the analyst can both learn and rank. For instance, the key phrases that we display retain valuable information and COA offers numerous other options to use those phrases in the patent ranking task; like, (1) to run a prior search on those keywords just by following the hyper-links on those phrases, (2) to get a measure on the novelty and impact of those phrases from the patent rating table, (3) to study the distribution of the phrases in earlier and later patents, (4) to analyze the co-occurrence behavior of a set of key phrases to model an innovation concept, and (5) finally, to change the default setting to one that is appropriate for a particular class of patent based on the output of the above analysis tasks. From the experience of our IP teams, this was extremely helpful in expediting the patent evaluation.

In recent days, software or business process patents have received some criticism regarding their quality or importance. The main reason behind that is the inability of the patent examiner to understand the technicality of the patent or their failure to search the *prior arts* [20, 4]. Our term indexing approach is very useful there, as it capture the systematic flow of knowledge evolution in the patent literature over the time. Such indexing is very helpful in finding the *prior arts* or related work. Moreover, it provides the examiner a visual cue about the dominant keywords of that technology field; thus, it helps him(her) to obtain domain knowledge instantly.

One final remark regarding COA ranking is that it uses the simplest statistic measures(like the average) to obtain different scores and parameters, like threshold, which enables COA to achieve very good generalization abilities over different classes of patents. Intuitively the huge contrasts

among patents in different classes make the ranking task similar to learning in a very noisy dataset. So, any complex criteria suffers from over-fitting, and hence, does not perform well. An instant example is the better performance of COA score2 over COA score1 in the classification task (see table 3). The latter uses weighted contribution of a term whereas the former just considers that all the term has an weight value 1. Here, although COA score1 uses more complex function, it performs worse compare to COA score2.

This is an ongoing research and hence, has substantial room for improvement. The improvements can be made in two distinct arenas. One is in the ranking technique and the other is in the improvement of the current system. Our ranking system is based only on the novelty of a patent. Although, it performs well for a pioneer effort, it is far from perfect. Specifically, "claim robustness analysis" is another compelling criteria that IP attorneys think can add significant value to the current system. We like to maneuver this approach by understanding a claim's linguistic simplicity, unambiguousness, generality etc., by using some form of statistical NLP techniques. Regarding the current system, the major improvement is to streamline the definition of different ranking parameters. For instance, the "support" of a term, currently, just counts the number of its usage in subsequent patents. But, one important modification could be to understand the distribution of the the term's usage over the time interval instead of just looking at the raw count. The user interface, user interaction and patent visualization technique etc. can also evolve over the time from the suggestions of the current users.

To summarize, we built a patent evaluation system that considers the earliness and impact of the claim words to measure the novelty of a patent. By indexing the words in the patent literature for its earliest occurrence, it can present a patent rating table which is very helpful in defining patent's value in a very fast and efficient manner. Moreover, the user friendly manner of visualization and ample user interaction options in the entire system makes it a very useful tool in practical patent evaluation jobs.

## 7. REFERENCES


[1] J. O. Lanjouw, A. Pakes, and J. Putnam, *How to count Patent and Value Intellectual Property: The Use of Patent Renewal and Application Data*, The Journal of Industrial Economics, 46(4):405-432, 1998

[2] J. O. Lanjouw, *Economic Consequence of a Changing Litigation Environment: The case of Patents*, National Bureau of Economic Research, W4835, 1994.

[3] K. Kasravi, M. Risov, *Patent Mining - Discovery of Business Value from Patent Repositories*, Proc. of the 40th International Conference on System Science (2007), Hawaii, US

[4] S. Shulman, *Software Patents Tangle the Web*, Technology Review (2000), pp68-72,74,76

[5] D. Harhoff, F. Narin, F. Scherer, and K. Vopel, *Citation Frequency and the Value of Patented Inventions*, The Review of Economics and Statistics, 81(3):511-515

[6] y. Tseng, Y. Wang, D. Juang, C. Lin, *Text Mining for Patent map Analysis*, IACIS Pacific Conference (2005), Taipei, Taiwan

[7] B. Wang, M. Chu, J. Shyu, *Patent value Measurement by Analytic Hierarchy Process*, IAMOT (2006), Beijing, China

[8] M. Marcus, B. Santorini, and M. A. Marcinkiewicz,*Building a large annotated corpus of English:The PENN Treebank*, Computational Linguistics, vol 19, 1993.

[9] S. Sheremetyeva, *Natural Language Analysis of Patent Claims*, ACL Workshop on Patent Corpus Processing (2003), Sapporo, Spain

[10] S. Sheremetyeva, *Generating Patent Claim from Interactive Input*, 8th International Workshop of Natural Language Generation (1996), Herstmonceux, England.

[11] L. Page, and S. Brin, *The anatomy of a large-scale hypertextual Web search engine*, Proceedings of the seventh international conference on World Wide Web, pp107 117, 1998

[12] B. Shaparenko, R. Caruana, J. Gehrke, and T. Jocachims, *Identifying Temporal Patterns and Key Players in Document Collection*, In Proceedings of the IEEE ICDM Workshop on Temporal Data Mining, Houston, TX, (2005), pp164 174.

[13] J. Kleinberg, Authoritative sources in a hyperlinked environment, Proc. of the Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pp668 677., 1998

[14] `www.delphion.com`

[15] `www.patentcafe.com`

[16] `http://www.ibm.com/ibm/licensing/patents/portfolio.shtml`

[17] W. Cody, J. Kreulen, V. Krishna, W. S. Spangler, *The integration of Business intelligence and Knowledge Management*, IBM System Journal, 41(4), 2002

[18] *USPTO Performance and Accountability Report, 2006*

[19] A. L. Miele, *patent Strategy: The manager's guide to profiting from patent portfolios*, Willey Intellectual Property Series, 2001.

[20] A. B. Jaffe, and J. Lerner, *Innovation and its discontents: How our broken patent system is endangering innovation and progress, and what to do about it*, Princeton University Press, 2004.

[21] J. Alcacer, M. Gittelman, *How to I know what you know? Patent examiners and the generation of patent citations*, Stern School of Business (working paper).

[22] `http://www.freepatentsonline.com`

[23] `http://www.uspto.gov`

[24] L. Hollaar, *Legal Protection of Digital Information*, BNA Books, 2002

[25] H.J Knight, *Patent Strategy for Researchers and Research Managers*, John Willey and Sons Ltd., 2001.

[26] B. Shaparenko, and T. Joachims, *Information Genealogy: Uncovering the Flow of Ideas in Non-Hyperlinked Document Databases*, Proceedings of ACM SIGKDD Conference, San Jose, CA, 2007

[27] *A text-mining-based patent network:Analytical tool for high-technology trend*, The journal of High Technology Management Research, 15(1), 2004, pp37-50

[28] *An Empirical Look at Software Patents*, Journal of Economics & Management Strategy (2007), 16(1), pp157-189

[29] http://www.almaden.ibm.com/asr/projects/biw/

[30] http://lucene.apache.org/java/docs/

[31] http://lucene.apache.org/solr/