

# COA: Finding Novel Patents through Text Analysis\*

Mohammad Al Hasan<sup>1†</sup>, W. Scott Spangler<sup>2</sup>, Thomas Griffin<sup>2</sup>, and Alfredo Alba<sup>2</sup>

<sup>1</sup>Dept. of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, 12180

<sup>2</sup>IBM Almaden Research Center, San Jose, CA 95120

<sup>1</sup>alhasan@cs.rpi.edu, <sup>2</sup>{spangles@almaden, tdg@us, aalba@us}.ibm.com

## ABSTRACT

In recent years, the number of patents filed by the business enterprises in the technology industry are growing rapidly, thus providing unprecedented opportunities for knowledge discovery in patent data. One important task in this regard is to employ data mining techniques to rank patents in terms of their potential to earn money through licensing. Availability of such ranking can substantially reduce enterprise IP (Intellectual Property) management costs. Unfortunately, the existing software systems in the IP domain do not address this task directly. Through our research, we build a patent ranking software, named COA (Claim Originality Analysis) that rates a patent based on its value by measuring the *recency* and the *impact* of the important phrases that appear in the “claims” section of a patent. Experiments show 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.

## Categories and Subject Descriptors

H.2.8 [Database Management]: Database applications—*Data mining*

## General Terms

Algorithm, Design

\*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.

<sup>†</sup>Part of this work was done in the summer of 2007, when the first author was at IBM Almaden Research Center.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD’09, June 28–July 1, 2009, Paris, France

Copyright 2009 ACM 978-1-60558-495-9/09/06 ...\$5.00.

## Keywords

Document Ranking, Information Retrieval, Patent Processing, Patent Visualization

## 1. INTRODUCTION AND BACKGROUND

The business enterprises in the technology industry became increasingly patent savvy over the last two decades. It is caused by mainly two reasons: first, US courts have shown pro-patent stance in many patent-based litigations [12] and the second, the scope of patent-eligibility has been broadened to include software or business methods. As a result, patent portfolio of technology enterprises are growing at a much faster rate. According to the 2008 fiscal year report of United States Patent and Trademark Office (USPTO) [22], 496,762 patents were filed in the year of 2008, which is about 30% more than in year 2004 [23]. In the category of class 705 which includes patents related to business methods, USPTO awarded 13,037 patents until year 2006, of which only 958 (7.3%) were awarded before the year 1990.

As the patent portfolio grows, knowledge discovery in patent data becomes more valuable as it can save significant money for enterprises in their patent management cost. One important task in this regard is to rank the patents in terms of their value. It is beneficial in several ways. For example, while exploring new licensing revenue opportunities or searching for possible infringement [8, 14, 9], attorneys and IP analysts can prioritize their time in the patents that are at the top of the ranking list. This would increase the *return per dollar* for the money invested on highly-paid patent attorneys and agents. On the other hand, the set of patents that are at the tail of the ranking can be dropped to save a part of the maintenance fee (paid to the patent office). A value ranking of patents is also useful when enterprises make business decisions such as mergers, acquisitions, opening of new lines of business, etc.

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 art* while assessing the novelty of a pending patent application. An increased number of patent applications exacerbates 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 methods*, but, they comprise a large frac-

tion of patents issued in recent years [20, 1]. In fact, many granted patents in these areas received severe criticism as they did not seem to be novel [20]. Since they are composed as a sequence of business processes, comparison to *prior art* is difficult. Effective assessment techniques need to be invented that work well for these kinds of patents.

Existing practices of patent ranking are predominantly manual, accomplished by IP professionals and patent analysts. They mainly 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 the scope of legal language, etc. These metrics are significant as they are used in the formal process of patent litigation. But, there is another dimension in patent evaluation tasks—measuring its novelty and non-obviousness, for which IP professionals are not the most competent persons. Since they are not skillful on the core technology, they collaborate with SME (Subject Matter Expert) or the inventors to understand these aspects of a patent. Since attorneys’ and agents’ time is very expensive, an advanced software tool (be automated or semi-automated) is required to expedite the process.

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 value of that patent.

We validate the performance 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<sup>1</sup>. 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 art* 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<sup>2</sup> for back-end data store, together with **Lucene** [13] to index the textual phrases. **SOLR** [21] is used as the search server that communicates between COA Java application and Lucene. COA is integrated with the BIW (Business Insights Workbench) software [2]. 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 a 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 reveals 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 [12] 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 [9, 14, 11, 5]. In a recent work, Wang et. al. [24] 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 [15] 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, thus 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 this from USPTO web site [22].

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.

<sup>1</sup>patents that IBM continues to maintain

<sup>2</sup><http://www-01.ibm.com/software/data/db2/9/>

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 or 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. Robust claim structure and claim drafting are important issues 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.

## 2. CLAIM ORIGINALITY ANALYSIS (COA)

### 2.1 Definitions

**Patent Class:** US Patent office follows a classification system for organizing all US patent documents into relatively small collections based on common subject matter. Each subject matter division includes a major component called a class. Generally, a class delineates one technology from another. For example, class 706 is assigned for the technology, *Data Processing: Artificial Intelligence*. Classes may be further divided into subclasses to delineate processes, structural features and functional features of the subject matter. Patent office assigns at least one class label (also called mandatory class) to every patent it grants. Optionally, a patent may have multiple class labels. In COA, class-code of a patent plays an important role to compute the novelty of a patent from the patent text. COA uses only the mandatory class for a patent.

**Patent Novelty:** Novelty of a patent is its non-obviousness with respect to the existing *prior art*. According to the patent law, for a patent to be granted, it has to be novel. We attempt to quantify the novelty of a patent as a real number by comparing it (the subject patent) to other patents that are forerunners in the same or closely related technical domains. If the subject patent is one of the earliest patents in this comparison, it receives higher novelty score. On the other hand, if an earlier patent (or other work) is found that uses the invention (or a significant part of it) that is reported in the subject patent, the latter loses its value significantly.

The problem of quantifying novelty of a document has not been explored well in existing data mining and information retrieval researches. In these domains, researchers are generally interested to measure the *significance* of a document which is its relevance to a user-provided query. But, a significant document may not be novel at all. In COA, we quantify novelty of a patent from the patent text by analyzing the key phrases (defined later) that it uses. The earliness (defined later) of these key phrases is a good indicator of a patent’s novelty.

**Patent Impact:** Impact of a patent is its business importance in the technology market. It measures the influences that the subject patent has made on other related inventions

since it was published. We find the impact of a patent by analyzing the usages of its key phrases in other patents that are published later than it.

Measuring impact (also called *impact factor*) of a document is a well-studied research topic in information retrieval and bibliometrics. For instance, the HITS [10] algorithm assigns higher *authority score* for a document that has a high number of hyper-links pointing to it. In the academic research domain, citation statistics or bibliometrics are used to identify an influential (high-impact) 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 incomplete listing of the prior art. So, they do not measure the impact of a patent correctly.

**Key Phrases:** Key phrases<sup>3</sup> are selected phrases from the subject patent document. They can be a word, a multi-word term, a capitalized phrase, a noun phrase, etc. COA extracts key phrases from the *claims* section of a patent. Our experiments show that phrases appearing in the *claims* section are much more significant for patent evaluation. In COA a phrase is significant if it is essential to describe a patentable innovation. While in other information retrieval tasks, for example, in document categorization, a phrase can be significant if it is discriminatory among different classes of documents. Intuitive argument of using key phrases in patent ranking is based on the philosophy that the claims of a patent with a valuable innovation would use novel technical terms, phrases and keywords. For example, if a patent describes the innovation of the back-propagation algorithm as a neural network learning technique, the word *back-propagation* is most likely a novel term in that patent. A *dual-layer dvd* is another novel term in a patent that reports the invention of dual-layer recording of dvd disks.

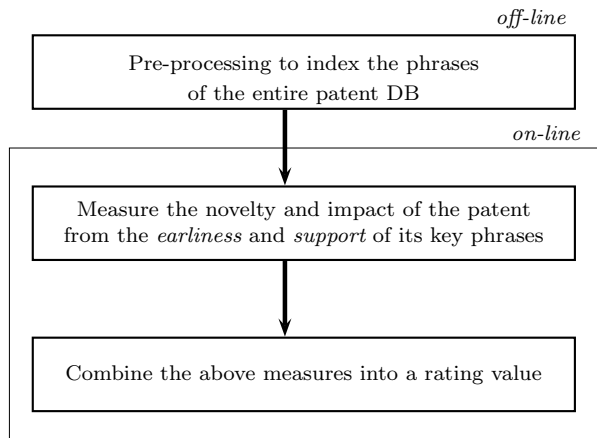
**Phrase Age:** In COA, the age of a phrase is defined with respect to a subject patent. To compute the age of a phrase, COA searches for that phrase in all patents that bear the same patent class as the class of the subject patent. If this search returns no earlier patent (with respect to the subject patent) then the age of the phrase is zero. But, if an earlier patent is found, then the time difference between their publish dates is the age of that phrase.

**Phrase earliness:** The *earliness* of a phrase is akin to its recency and is defined as the reciprocal of the age of that phrase. Thus earliness evaluates the degree of originality of a phrase. Usage of a high number of early phrases in a patent indicates that the patent is a forerunner (early) in its technological domain. By restricting the comparison within the patents belonging to the same class, COA ensures that the technical meaning of a phrase is comparable across different patents.

**Phrase Support:** The support of a phrase in a patent is the frequency of its usage by later patents in the same class. A high support phrase works favorably for a patent to achieve higher ranking. However, 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 by subtracting from it the average support values of a

---

<sup>3</sup>we may occasionally refer key phrase as phrase, if there is no confusion



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

collection of phrases that have similar age.

## 2.2 Methodologies

The steps of COA to obtain a patent’s value from its text is outlined in Figure 1. We discuss them as follows.

**Pre-processing Step:** This is an off-line step to extract a large set of phrases from the entire patent database and to index their occurrences in patent documents. Using this index, COA computes the earliness and the support of a phrase in nearly constant time. Thus, it (the index) provides real-time patent analysis capabilities with the provision of user interactions.

COA performs phrase indexing of one patent class at a time. To extract (and index) all possible key phrases from the patents in that class, it applies a simple  $n$ -gram method after removal of stop-words. In this method, all consecutive words up-to length three are considered as candidate phrases. Redundant phrases are removed by using a simple stemming algorithm. Thus, for each candidate phrase, COA indexes its occurrences in the claims section of all patents in the corresponding class. Our database holds all patents that were issued in the last twenty years. The off-line indexing process on this database generally takes few hours for each patent class.

The above method generates too many candidate phrases of which many are not useful. To filter them effectively, COA builds a background dictionary for each patent class that constitutes the phrases that are not critical for describing a novel invention; these phrases are too common and they constitute the upper tail of the phrase frequency distribution. Any phrase that appears with higher frequency than a frequency threshold is moved to the background dictionary. A suitable frequency threshold is determined for each class by analyzing the frequency histogram of the candidate phrases.

**Identifying Key Phrases:** This task is performed online for one or for a set of related patents that are evaluated together. This is the most important task since, in COA, the key phrases are the dimensions of patent evaluation metric. To obtain the key phrases, COA first finds the candidate phrases from the *claims* section of the subject patent. If any of these phrases appears in the appropriate background dictionary, the phrase is discarded. For each of the remain-

**Table 1: POS to extract phrases (NN=Noun, JJ=Adjective, VBG=Gerund)**

Rule	Example
NN NN	Product Information
JJ NN	Touch-sensitive Surface
VBG NN	Initiating Communications
NN NN NN	Global Information Network
JJ NN NN	Standard Message Format
VBG JJ NN	Monitoring Operational Status
VBG NN NN	Processing Product Orders

ing phrases, COA computes its contribution to the patent’s value as a product of the phrase’s support and earliness. Earliness is measured as the reciprocal of the term’s age. If  $T$  is a phrase, and  $\text{support}(T)$  and  $\text{age-in-days}(T)$  are its frequency and age, the term contribution is calculated as follows:

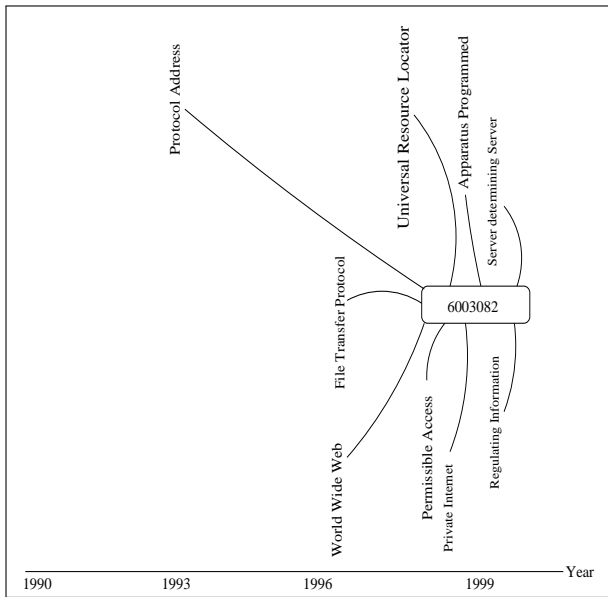
$$\text{Contribution}(T) = \max\left(\frac{\text{support}(T) - 2}{\text{age-in-days}(T) + 1}, 0\right)$$

In the above equation a one is added 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 below a threshold, COA discards the term. 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.

The above approach of finding key phrases is different than traditional text mining; for the latter, the phrases that appear around the lower tail of the distribution are generally discarded as they don’t have enough support to be significant for the text categorization or information retrieval task. However, COA does not discard those implicitly, as infrequently occurring terms may sometimes indicate novelty and in that case COA includes them in the set of key phrases.

An alternate way to extract key phrases is to use the POS (parts-of-speech) tagger. This is also popular is NLP based information retrieval. We employed rules on the POS-tagged claims to extract phrases. Table 1 shows the rules with an example for each. From our experiments on patent corpus, we found that this method, although it shows good precision, it often suffers from poor recall (misses a lot of important terms). Performance also depends on the quality of the POS tagger. So, POS tagger based techniques, although implemented, are not used in the final version of the system.

Depending on the patent, the above methods may generate too many or too few key phrases. So, COA allows 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. For the final set of key phrases, COA reports their occurrences and the support.



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

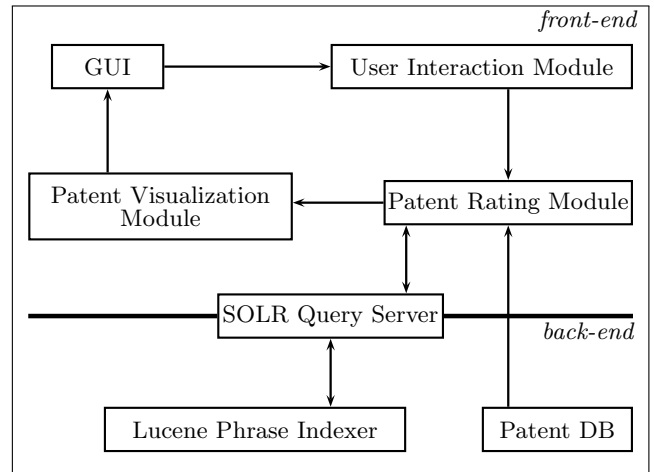
**Example:** Figure 2 shows a patent (number 6003082) with some of its key phrases. The horizontal line at the lowest part of the figure is the year axis that increases from left to right (from 1990-1999). The rectangular box denoting the patent 6003082 is vertically aligned with the year 1999 to indicate 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 is 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.

**Obtaining rating scores:** In COA, if a patent has many phrases with high earliness score and high support, the patent is considered valuable. However, to enable numerical comparison among different patents, we obtain two rating scores. The higher the rating score, the more valuable the patent is. The first score is just the linear sum of the individual contributions of each of the key phrases of a patent. The contribution value is computed the same way as discussed in section 2.2. The second score is just the size of the key phrase-set. 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 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.

### 3. COA:ARCHITECTURE

Figure 3 shows an architectural block diagram of COA. It denotes the different modules of the rating system in labeled rectangular boxes. It also shows the data flows among the modules by arrows. The diagram is partitioned into two



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

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 [13]. 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.

**SOLR query server:** We used SOLR [21] 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.

**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 pro-

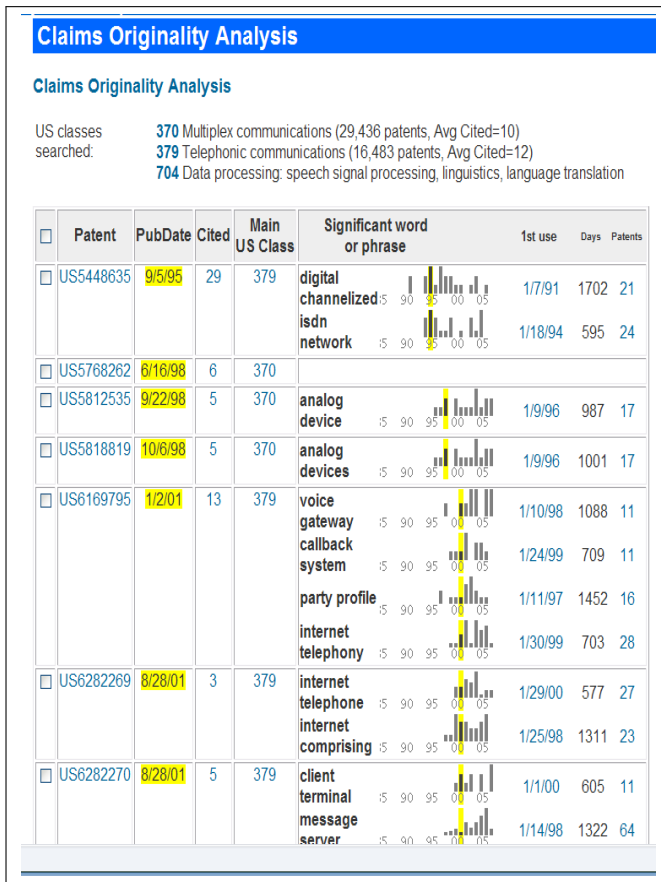


Figure 4: Patent Rating Table

duced 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, class-code, 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 595 days earlier than this patent. The support of the term is 24 patents, i.e. after the first use, the term has been subsequently used in 24 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

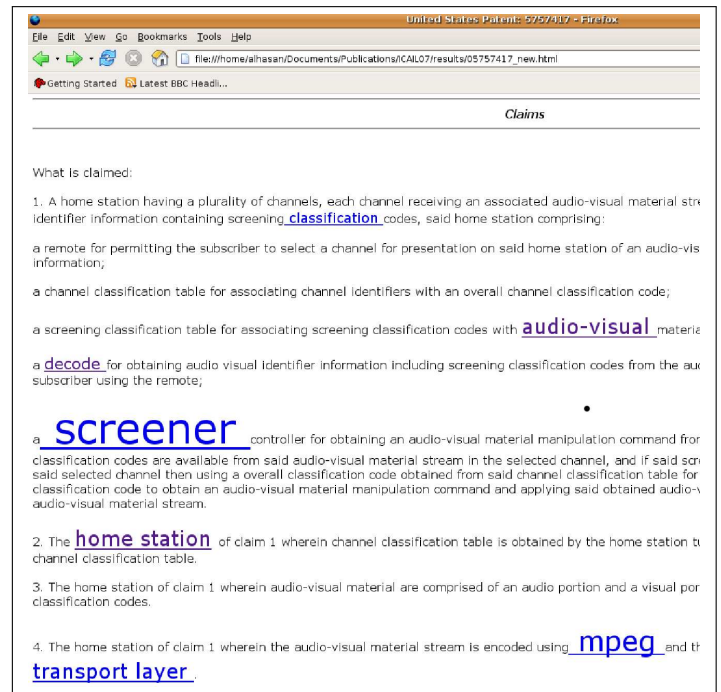


Figure 6: A screen-shot of the Claims Originality patent view. The terms that are the most original in the claims are highlighted by using larger fonts. Each term is also hyper-linked to the patent that used the term for the first time.

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. There are ample options to set different parameters to study the influences of different phrases on the rating value. For instance, user can edit the background dictionary to add or remove words or phrases from the dictionary. Contribution of a phrase towards a patent’s rating can be made void to investigate a suspicious rating. Default threshold for support and recency can be altered to modify the size of the key phrase list, etc.

## 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 [11] 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

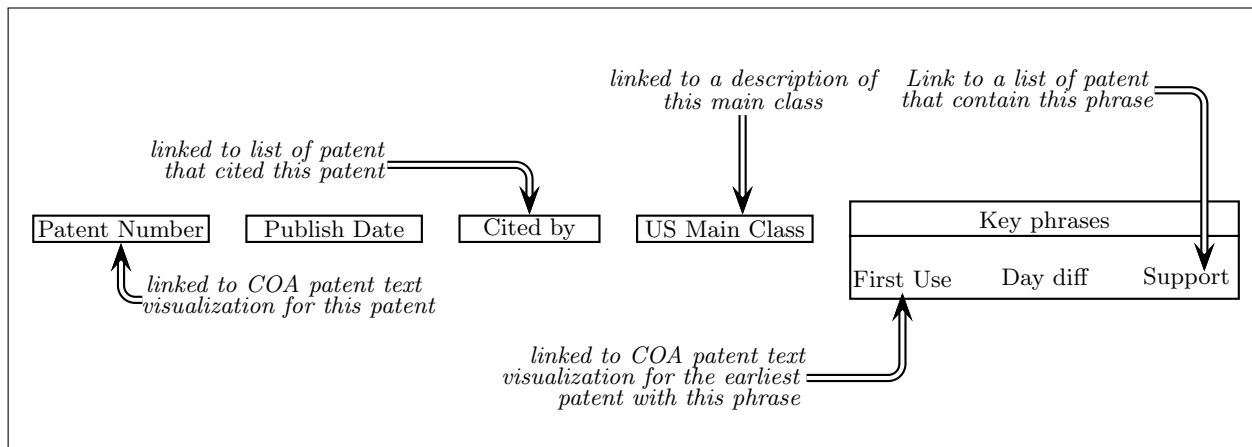


Figure 5: Hyper-links from different columns of patent rating table

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 [6], 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 scores with the status of a patent. We considered two sta-

Table 2: Comparison of active and lapsed patents

	COA (score1)	COA (score2)
<i>Current</i>	4.20	8.41
<i>Lapsed</i>	1.51	2.89

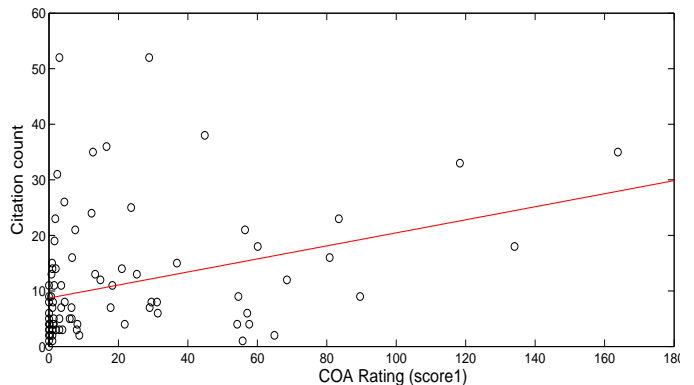


Figure 7: Citation count vs COA rating value

tus 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 2 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

**Table 3: Comparison of patents with attorney rating**

	Dataset 1		Dataset 2		Dataset 3	
	COA (score1)	COA (score2)	COA (score1)	COA (score2)	COA (score1)	COA (score2)
<i>Rating 1</i>	18.77	40.88	12.73	27.34	12.73	31.32
<i>Rating 3</i>	4.20	10.82	7.22	16.89	4.49	10.01

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 have been  $\frac{220 \times 683}{1544} = 97$ . Since, the obtained value of 126 was well above 97, these two variables can not be independent. A Chi-Square test, like Chi-Square-Distribution 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, 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 3. 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 SVM 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 4. 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 rank-

**Table 4: Classification accuracy using COA scores to classify rating-1 and rating-3 patents**

	Dataset 1	Dataset 2	Dataset 3
<i>Patent Citation</i>	59.75	54.84	51.82
<i>COA score1</i>	53.20	70.63	64.49
<i>COA score2</i>	59.76	69.79	60.15

ing 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

In recent years, works on patent data have got much attention in industrial domain. But, the majority of these works [4] have been web-based services that are targeted towards corporate clients. These works mostly provide patent data feed (patent texts, news, case updates, etc.) and, sometimes, an infrastructure for the clients to run queries on some structured fields (like, class-code, file histories, assignees name) of patent data. Few enterprises [3, 16] also provide web-based software tools that facilitate further analysis and visualization of the results obtained from these searches. 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 enterprise 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 hyper-link structure or citation information. Graph-based algorithms, like HITS [10] and Page-Rank [17] are the most successful in identifying the most useful document, especially in the domain of search engines. But, this approach is not well suited for our task, as the value of a document with respect to query word relevance may be very different from its value in terms of its originality index.

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. [18] 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 [19], the authors used a Gaussian mixture model of words in the text to model flows of ideas in documents. However, choosing the parameters for such a model is difficult as the intuitive meaning of a parameter value is difficult to comprehend by a patent analyst.

## 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 effectively. 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 patents 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.

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 enable 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 4). The latter uses weighted contribution of a term whereas the former just considers that all the terms have 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 existing system. Our ranking system is based only on the novelty of a patent. Although, it performs well for a pioneering 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. The user interface, user interac-

tion and patent visualization technique etc. can also evolve over 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 a 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 make it a very useful tool in practical patent evaluation jobs.

## 7. REFERENCES

- [1] J. Bessen, An Empirical Look at Software Patents, *J. of Econ. & Management Strategy* (2007), 16(1): 157-189
- [2] [www.almaden.ibm.com/asr/projects/biw/](http://www.almaden.ibm.com/asr/projects/biw/)
- [3] [www.delphion.com](http://www.delphion.com)
- [4] <http://www.freepatentsonline.com>
- [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] [www.ibm.com/ibm/licensing/patents/portfolio.shtml](http://www.ibm.com/ibm/licensing/patents/portfolio.shtml)
- [7] 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
- [8] K. Kasravi, M. Risov, Patent Mining - Discovery of Business Value from Patent Repositories, *Proc. of the 40th Intl. Conf. on System Science* (2007), Hawaii, US
- [9] H.J Knight, Patent Strategy for Researchers and Research Managers, *John Willey and Sons Ltd.*, 2001
- [10] J. Kleinberg, Authoritative sources in a hyperlinked environment, *Proc. of the Ninth Ann. ACM-SIAM Symp. Discrete Algorithms*, page 668-677., 1998
- [11] 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
- [12] J. O. Lanjouw, Economic Consequence of a Changing Litigation Environment: The case of Patents, *National Bureau of Economic Research*, W4835, 1994
- [13] <http://lucene.apache.org/java/docs/>
- [14] A. L. Miele, patent Strategy: The manager’s guide to profiting from patent portfolios, *Willey Intellectual Property Series*, 2001
- [15] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, Building a large annotated corpus of English: The PENN Treebank, *Computational Linguistics*, vol 19, 1993
- [16] [www.patentcafe.com](http://www.patentcafe.com)
- [17] L. Page, and S. Brin, The anatomy of a large-scale hypertextual Web search engine, *Proc. of the seventh international conference on World Wide Web*, pp107-117, 1998
- [18] B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims, 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
- [19] B. Shaparenko, and T. Joachims, Information Genealogy: Uncovering the Flow of Ideas in Non-Hyperlinked Document Databases, *Proc. of ACM SIGKDD Conference*, San Jose, CA, 2007
- [20] S. Shulman, Software Patents Tangle the Web, *Technology Review*, 2000
- [21] <http://lucene.apache.org/solr/>
- [22] <http://www.uspto.gov>
- [23] USPTO Performance and Accountability Report, 2008
- [24] B. Wang, M. Chu, J. Shyu, Patent value Measurement by Analytic Hierarchy Process, IAMOT (2006), Beijing, China