RJ10402 (A0701-011) January 17, 2007
Computer Science

IBM Research Report

Assessing Patent Value through Advanced Text Analytics

Mohammad Al Hasan
Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180

W. Scott Spangler
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

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

Assessing Patent Value through Advanced Text Analysis*

Mohammad Al Hasan†
Rensselaer Polytechnic Institute
110, 8th Street
Troy, NY 12180
alhasan@cs.rpi.edu

W. Scott Spangler
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
spangles@almaden.ibm.com

ABSTRACT

Patent, as an intellectual property, got tremendous attention lately from the technological companies, who are filing more and more patents, evidently to realize a rich portfolio. It serves both the defensive and the license revenue generating purposes. But, along with that, the portfolio management is also becoming difficult. One of the major challenges in this regard is to identify a small set of patents that are highly innovative and hence, are valuable in the technology market in terms of licensability. Again, a large fraction of recent patents are software or business method kinds, for which the novelty or the innovation is difficult to assess. Hence, an automated or semi-automated software system is required that can employ a ranking mechanism for patents. Unfortunately, no such system exists. The existing patent software systems, mostly web-based, provide the following services: patent data feeds, structured or unstructured search platform on patents, portfolio analysis, like, comparison among different assignees patent strength, or patent visualization, like patent citation graph, etc. These services are helpful for prior search or analyzing assignees market strength; yet, not capable to provide any insight to compare the novelty among a set of patents. Therefore, identifying patents that have high license potential, is still, predominately, a manual, laborious and time-consuming process. In this research, we proposed a patent ranking method that is very suitable for ranking software or business-method kind of patents. It adopts information retrieval methodologies that use text from the patent claim sections to rank the patents based on their novelty. Moreover, it provides user interaction provisions in all critical steps of the ranking to fine tune the rank results. This method also employs innovative visualization tools to assist the users in understanding the salient features of a patent. We implemented the proposed method to build a patent ranking tool, named COA (Claim Originality Analysis) and subsequently, used it in analyzing IBM's patent portfolio. Our experiments and analyses show that COA is very effective in identifying innovative patents in a very short time and effort.

Keywords

document ranking, information retrieval, patent processing, patent visualization

1. INTRODUCTION AND BACKGROUND

In current technology market, the value of patent as an intellectual property is tremendous. Besides protecting the invention, it also provides the inventor an opportunity to generate revenue by means of licensing the invention. In computer industry, Big research companies, like IBM, Motorola, Sony, Intel, Mitsubishi and etc. earn more than hundreds of million dollars yearly just from patent licensing revenue and this trend is going upward. Moreover, a big patent portfolio gives a company the competitive edge in the technology market, especially, in making business deals, like merging, acquisition or even in marketing new products. Therefore, companies are filing an increased number of patents in each year. From the 2006 fiscal year report of USPTO [28], 443652 patents are filed in the year 2006, which is about 10% more than that of previous year [19].

However, as the patent portfolio of a company gets larger and larger, it becomes increasingly difficult for it to manage it. Firstly, The company needs to pay a maintenance fee to the patent office for each patent in its portfolio. But, many patents in the portfolio may become obsolete, due to numerous reasons, like–a change in the trend of the technology market, invention of alternative technology, a change in the company's growth plan, and etc. Hence, it employs strategist to carry out a periodic assessment of each of the patents in it's portfolio to make the decision whether to keep the patent or not. Secondly, high competitiveness in the market, also forces a company to carefully search the competitor's product line with respect to the patents in its portfolio, to identify possible infringement; that, sometimes, opens a new source of revenue through licensing, if infringement can be claimed. However, each such search is very through, time consuming and predominately a manual process, since it requires considering all the elements that influence the legal sustain-ability of an infringement case. Hence,

*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 summer 2006, when the first author was a research intern at IBM Almaden Research Center

the search is usually localized to a small subset of very fundamental patents in the related field and that also requires a ranking of the patents. Besides assessing a single patent, sometimes a company assess a group of patents or even the entire patent portfolio of another company. For instance, if it want to make a decision to merge with a company or to acquire it, or to make an offer to license some of the patents of that company, a study of their patent portfolio helps in assessing their market stronghold. Furthermore, if a company like to make a decision to enter in a new business, it analyzes the existing patents in that business area to get an answer to the question: "How crowded is that technological field?". Answer to this question, helps the management to make a better decision regarding their investment. In summary, effectively maintaining large patent portfolio is very crucial, especially, for large companies and the major success in this regard, depends on the ability to correctly evaluate a patent from the prospect of its licensability. At present, companies spend millions of dollars in intellectual property management, mostly through the employment of IP (Intellectual Property) lawyers and patent analysts, who evaluate each and every patent and order those by the prospect of license values. But, such analyses is time consuming and very subjective. Hence, to expedite the process, advanced software tool is required that is automated or semi-automated. One of the key functionalities of such tools is to assess a patent's license value. Our research is a pioneering effort towards achieving this goal.

We developed a software system, named, "Claims Originality Analysis (COA)", to assess a patent by evaluating the originality of its invention. COA is fundamentally different from any concurrent patent analysis tool. It uses a information retrieval approach, where a patent is considered to be 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. For each word in the claim section of all patents, COA builds an index to store a record. inside it, it stores the following information—the id and the publish time of the patent that used the word for the first time, the count of the subsequent usages of the word in other patents, and etc. To rank a patent, COA first extracts all the words in its claim section and then by using the above index, retrieves all the records corresponding to these words. Then, it presents this result in a patent rating table, from which a patent's value can be estimated. Currently, COA is being used in-house by the IP (Intellectual Property) department of IBM. IBM's rich patent portfolio provides us an ideal testbed to evaluate the effectiveness of the ranking approach that COA adopts. So far, the experiences of using COA in evaluating patent is very rewarding and we find that the ranking criteria and approach of COA is very effective to evaluate software patents, or patents on business processes, which, otherwise, is very difficult to analyze.

COA is designed in a modular fashion, hence can be integrated with any text analysis software. Currently it is housed in the framework of an advanced text analysis engine. Thus, it exploits the existing data analysis techniques, like clustering, trend analysis, word analysis, scatter plot visualization, and etc. that the host engine provides. Furthermore, it provides the above tool for patent ranking. We itemize the features that COA provides as below:

- It rates a patent from the innovativeness perspective, by using techniques from information retrieval domain. To do so, it uses only the claim texts of a patent.
- 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 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.
- The system can run independently or can be incorporated with a text analysis tools because of its modular and effective design.

We claim the following contributions.

1. We proposed a method for a patent's ranking that is simple, efficient and practical. It has numerous usage in all steps of the patent processing, like prior search, portfolio evaluation, etc. Beside patent, this method can also be used to rank other technical documents.
2. We implemented our proposed method in a software tool. Experiences of its usage by our IP lawyers show that it can save substantial time and effort in patent analysis. To the best of our knowledge, it is the first tool of such kind.
3. We identified different steps of a patent rating system, where human knowledge can be injected and then, incorporated those steps in our system.
4. We designed novel methods for patent visualization that are very effective in patent analysis tasks.

The rest of the document is organized as follows. Section 2 discusses the related works. Section 3 explains different criteria on which a patent can be evaluated. It also discusses the challenges in analyzing patent documents. Section 4 provides a short description on the structure of patent documents. Section 5 describes our approach with the architectural framework and the implementation of the proposed system. The next section presents the results and future directions. Section 7 concludes with a discussion.

2. RELATED WORKS

In recent years, works on patent data got much attention in industrial domain. But, majority of these works [25, 26, 27] are 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 field, like class-code, file histories, assignees name, references and sometimes also,

on unstructured field, like patent claims, description of invention, prior work etc. Few companies [12, 14] also provide web-based software tools that facilitate further analysis of the results obtained from these search. They typically use clustering or summarizing techniques to find interesting patterns in patent data and then apply effective visualization techniques to display those. Now, Works in these kinds can only help in understanding the global picture of a collection of patents, such as, to discover 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 hypergraph structure or citation information. Graph-based algorithms, like HITS [11] and PageRank [9] are most successful in identifying the most useful document, especially in the domain of search engine. 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 [12] use the reference and citation information in patent data to form forward and backward reference graph, which is very useful, specially for the prior search in the patent domain.

Our work assess patent's value from the patent text and we did not find any prior work that build software tool to do such assessment. Very recently, Shaparenko et. al. [10] proposed a method for identifying influential papers and authors from a collection of research papers that solely uses the text. Conceptually, this work is similar to ours, since finding influential papers is homologous to finding valuable patents. But, patent documents are very different from the research papers in their style and structure, hence their algorithm may not be directly applicable in this domain. Again, due to the very subjective nature of patent evaluation, expert opinion and user friendly visualization are immensely required for any patent assessment work and their work does not has the provision of employing such.

There are some less technical research works, that tried to identify factors that are influential in ranking a patent. One of the best among these is the work by Wang et. al [2]. In this research, the authors elaborate different metrics that can be used to evaluate patent and assign numerical weights on each of those metrics. The weight represents the relative importance of the corresponding metric. Note that, the work is more inclined towards technology management and no direct study was made using real patents or the patent text.

Besides patent assessment, there are few researches that solve other problems in the patent domain. Tseng. et. al. [1] 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. Sheremetyeva et. al. [4, 5] has two distinct works that use statistical NLP (Natural Language Processing) and rule based technique to parse patent claim section. In one work, the authors decompose a patent claim to different elements to present it in an understandable manner; while, in other, they combine different claim elements to automatically generate the patent claim texts. But, due to the enormous complexity in understanding natural language, and specially, in understanding the patent text, their works are still in a rudimentary phase. Nevertheless, these are very important works, since, they attempted to understand the overwhelmingly long, complex and peculiar sentence structure of the patent documents. Such understanding is very useful with respect to patent ranking also, since robustness of claim structure, understandability of claims, etc. are key properties of good patents. However, our current work did not explore this avenue. To learn more about other works related to patent, interested readers can read the papers in the ACL workshop on Patent Corpus Processing (2003).

Our system uses patent text to find its value, hence, it borrows several ideas from the domain of text mining. Specially, the idea of interactive text mining [18, 22, 17] was very useful. However, text mining itself is a very active research domain with numerous researchers working on this topic. Interested reader can read the following textbooks [20, 21] and the references therein to cover the broad details.

3. PATENT ASSESSMENT CHALLENGES

Accurately assessing a patent's license value is a difficult task for a human, let alone, for an automated software system. It is so, because numerous criteria affect the license value of a patent. There are many literature [30, 23] that outlined many of these considerations. However, Wang et. al. in [2] broke those in three different categories: (1) Patent Strategic Value, (2) Patent Protection Value, and (3) Patent Application Value. The first category determines the innovativeness 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 assess the property that a patent protects through its claims section. The last category–Patent Application Value, mainly considers the breadth of the patent's applications in the relevant industry.

Our ranking method is akin to evaluating the patent's strategic value. So, we concentrate on measuring the novelty and impact of a patent. Since, a patent is about a new innovation, it must contribute something novel on top of the existing prior knowledge and we aim to extract that part, from the patent text. However, evaluating the exact contribution of an invention from the legalistic textual description in extremely hard, and we do not claim to solve it, either. But, we solve a rather simplified problem, where we view an invention as a vector of technical terms or words. Thus, our perception of a patent text is just a collection of keywords without any linguistic structure. This enables us to use the techniques from the domain of text mining only and to avoid the complex linguistic based technique.

We systematically redefine the ranking problem, in its narrow scope that considers only the strategic value. Under this criteria, the analyst needs to find out the current market value of the innovative substances of the patent and its impact on the technology market. There are, indeed three distinct steps in above process: (1) Identifying innovative substances (2) Finding their current market value, and (3) finding their continuing impact on the market in future.

Later, we describe how our ranking method solve each of these steps.

Now, we like to explain briefly, why we did not consider other ranking criteria. Firstly, it is not the case that those criteria are less important. Rather, in some situations they are more important than the one that we considered. But, they are much more difficult to evaluate. For instance, to evaluate the patent's protection value, the analyst needs to find patent claim elements, their scope and the robustness of the claim language in protecting those claim elements. These processes require techniques from information extraction [13] to, first, identify the claim elements and then, advanced techniques from Natural Language Processing (NLP), to distinguish bad claims from the good claims. Unfortunately, technique of NLP are very much corpus-driven, and no tagged corpus for patent language exists yet. Note that, all popular text corpus [3, 15] are newspaper or literature text based and performs very poorly in patent document for its stand-out peculiarity and intricacy.

Finally, estimating the patent's application value is the hardest, as analysts need to find it's application market; if it has a very broad market, then it gets a higher ranking and vice versa. It is very difficult to evaluate this by a software system. We give an example to explain this. Consider, a patent on some memory chip design, which covers a part of industry standard of the memory chip. This patent is very valuable since its market segment is very broad. Manufacturers who want to make memories, need to license the patent, since they want to conform to the industry standard so that their device is compatible to other co-operating devices, like processor, mother board etc. Now, this information is not accessible from the patent text or neither can it be inferred from there, so a software system surely will fail in ranking this patent close to its actual value. A patent's application value can also vary subject to its owner. For instance, in software domain, a patent is valuable only when it is considered collectively within the context of other supporting patents. Collectively, they cover a broad area of the technical domain; thereby, do not allow a manufacturer to build a relevant product without infringing one or more of the patents in that set. Whereas, with only one patent, a roundabout route can easily be obtained to manufacture the product without infringing that particular patent. So, value of this patent is different to different parties. Any companies that are building a portfolio in that technical area, would like to have that patent to make their portfolio bullet-proof. On the other hand, other companies don't want to have one such patent, as it does not gain them anything. A software tool has no way to consider these cases. There are numerous other examples, which justifies our approach to attack the ranking problem only from the patent's novelty, at least, to start with.

4. STRUCTURE OF A PATENT DOCUMENT

Patent text is very different from the ordinary newspaper text and text analytic tools that analyze 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. Reader can get more information from US Patent and Trademark office (USPTO) web site [28] or books on IP law [29].

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 the title, abstract, or description; one must read the claim section for it. Claim describes the invention, by listing its constituent part (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 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. Wang's [2] patent evaluation criteria— patent protection value, indeed concentrate on the claim section to identify the intellectual property elements that are protected through the patent. Identifying such elements is the most important step of claim analysis. Moreover, claim drafting is an important issue as well, since, choices of words (that are more specific), using poor language and etc. can generate claims that have very narrow scope and henceforth, can not sustain the infringement attack.

5. OUR APPROACH: PATENT RATING ON EARLINESS

Our approach is based on a very simple and well known principle which states that the patents that are early in the technology cycle are innovative, hence have higher values. These patents, most likely, innovate some breakthrough techniques that have been extended by later patents. We mentioned in earlier section that their rating can be obtained through the following three steps. Here, we describe each step elaborately.

Identifying Innovative Substances Each new technique usually introduces its technical jargon and keywords, and also, frequently uses a set of terms that are essential to describe the technical essence of the innovation. So, that set of qualified keywords/terms can represent the innovative substances of a patent. For example, if a patent innovates the back-propagation as a neural network based learning technique, the set {**neural network, back-propagation, supervised, weight, neuron, weight vector, epoch**} can be a potential set of frequently used terms that represent the innovative substance of that patent.

Finding term's market value All frequently used term do not have the same weight while evaluating a patent. Moreover, a term's value may differs depending on the followings: in which patent the term appears, to which class the patent

belongs, and etc. So, we adopts a generic approach of term evaluation that does not depends on the patent or its class-code. According to this approach, a term has high market value if it is introduced in the patent literature lately. Like, in the previous example, if the above patent uses the term "back-propagation" for the first time in the patent literature, this term is novel during the time the patent is published. Such term gets higher weight in patent evaluation. Our method extract all such terms from the patent's claim section. They are named as *innovation set*. For each term in the innovation set of a patent, we calculate the time difference between this patent's publish time and an earlier patent's publish time, where the earlier patent had used the term for the first time. So, we also name our approach as "ranking on the earliness of a patent" with the consideration that, if a patent uses lot of new terms, it is an early patent in that technology domain and should get higher rating.

Finding a term's future impact on the market Now, whether the patent has impact on the technology field, depends on the usage of the innovation therein in later inventions. This can be roughly estimated by finding the number of patent that used the terms in *innovation set*, later. We called it the *support* of the term. Note that, the support of a term 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.

Now, a patent's rating is depended on the above three measures: (1) Terms in the innovation set, (2) Their importance as measured by the time difference, and (3) their impact on the market as measured by the support. So, we present the analyst a table with all these information, that we named as *patent rating table*. Obviously, a lower value in the time difference and a higher value in the support for the terms in a patent, rates a patent highly.

Selecting the right set of terms in the innovation set is difficult. So, we use a user defined time-window method to solve it. From all the important terms in the claim text, we consider only those that have appeared first in patents that are published within the given time window. The value of time difference from the patent rating table can be used as a guideline to choose the time window length. A length of zero considers only those terms that are used for the first time in that particular patent. Selecting a higher value for the time window length allows more terms to enter into the innovation set.

Our approach is conceptually similar to the approach proposed by Shaparenko et. al. [10]. To identify the value of a document, d, they first represent all the documents by term vectors in TFIDF format. Then, they use the cosine product to find the similarity between two documents. In this way they find the k nearest neighbors of the document, d. Then, they find two values, k^d_{later} and $k^d_{earlier}$, representing the number of documents out of k neighbors, that has later and earlier time-stamps, respectively, than this document. Now, document d's rank can be obtained as below: $R^d_{raw} = \frac{k_{later} - k_{earlier}}{k}$. The higher the value of R_{raw}, the more influential the document is. However, to avoid the edge effects, the rank value is normalized by subtracting the average rank of all documents bearing same time stamps.

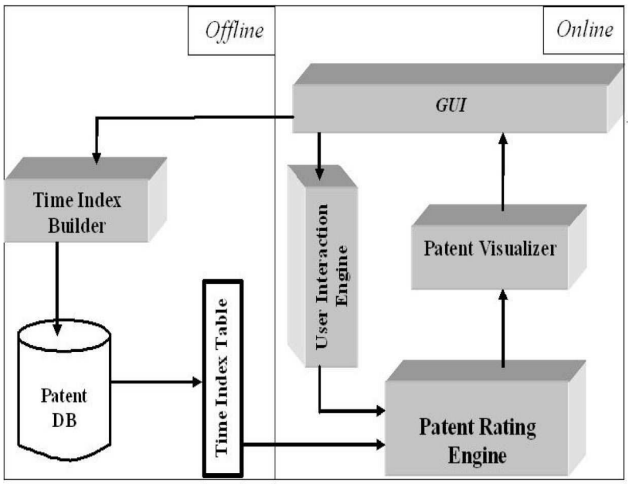

Figure 1: Different Architectural Component of the Patent Rating System

The final equation that they used to obtain the score is:

$$R^d_{scaled} = \frac{1}{k}(R^d_{raw} - \sum_{d_i : time(d_i) = time(d)} R^{d_i}_{raw})$$

Now, note that a document's k_{later} is analogous to the *support* of its *innovation set* and it's $k_{earlier}$ is inversely proportional to the size of it's *innovation set*. Thus these two methods are comparable. But, this approach gives a numerical score for each document's value. However, we learnt from the patent attorneys that an exact numerical score may not be very meaningful for patent data, since they need to know what are the terms in this document that are responsible to obtain a high or low score. Our approach makes that information available to the users and hence, provide options for subsequent user interactions.

5.1 Architecture of the proposed System

The patent rating system has the following modules

- Module to index the earliest usage of a term (offline)
- Rating Module
- User Interaction Module
- Visualization Module

Figure 1 shows the different modules in a bock diagram.

Indexing terms with the earliest-use time This is an offline step that indexes all the important terms or words in the patent literature to store their earliest appearance time. The index also stores the id of the corresponding patent, where an word appeared. Moreover, it stores the support of the term. Since, patents in different class-codes usually have very disjoint sets of terms, we index the patents in each class-code separately, i.e., there is one index file for each patent class-code. Before that, we build a background dictionary to contain stop words, and other terms which are very common to the patent literature and hence, are not important to rate a patent. Entire indexing step is summarized below and a flow chart of the steps is given in figure 2.

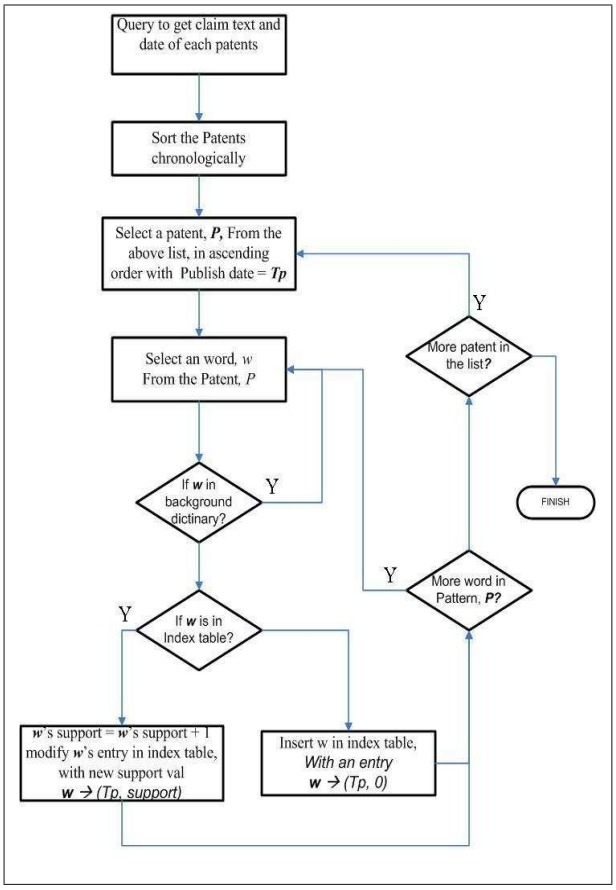

Figure 2: Patent word Indexing Flow Chart

1. From the entire patent data-set, we build a background dictionary by considering all the words that appears in more that 90% of the patent documents.
2. For each patent class-code, we build a separate index-file. to store the id and publish time of a patent (having that class-code) that used a term for the first time. We do so for all the terms, if it does not appear in the background dictionary.
3. To build the index-file efficiently, we first sort all the patents chronologically. Then for each file according to the above order, we insert all its terms in a hash-table to map the term to the patent-id and publish time, provided that, the term is not in the background dictionary or is not already inserted in the hash-table. For the later case, we increase its support count. The support count of a term is increased only once per patent. We also use some stemming algorithm to consider the synonymy effect of the terms.
4. After, all the patent files are considered, the index file is saved in the local hard drive. The support value for a term is also normalized for different patents based on that patent's publish date.

Patent Rating module This is the online module that an analyst uses to measure a patent's novelty. In the first step, the rating module identifies the innovation set by automatically extracting important claim words or terms from the patent text. Then, an interface similar to table 1 is presented to the user, by using the default set of application parameters. We call it, *patent rating table*. It displays, in each row, one innovation term, the time of the earliest patent that used this term, the time difference between these two patent and the term's support. For instance, from table 1 we notice that, while ranking the patent bearing number 06181781, one of the innovation term is *applet*, it was first used in a patent published in 6/22/1999, which is 588 days earlier than this patent. The support of the term is 23, i.e. after the first use, the term has been subsequently used in 23 distinct patents. Other innovation terms are placed in subsequent rows, similarly. Sometimes, an analyst do a collective assessment for a set of patent in a portfolio. Our system also provide that option. In table 1, we show such an analysis, where the analyst, first searched all the patent related to the term "voice mail" and then performed a collective rating similar to the one above. Here, the table displays the innovation terms and other data for all the patents in the collection. The table row can be sorted on different column values, as required. To get a quick reference to the earliest patent for an innovation term, each of the earliest date on the table (i.e. the fourth column) is hyperlinked to the actual earliest patent. So, by clicking those dates an analyst can access the patent text of that earliest patent. This is very helpful while analyzing a patent for noveltiness.

User Interaction Module The user interaction module allows an user to alter the default setting of the patent rating

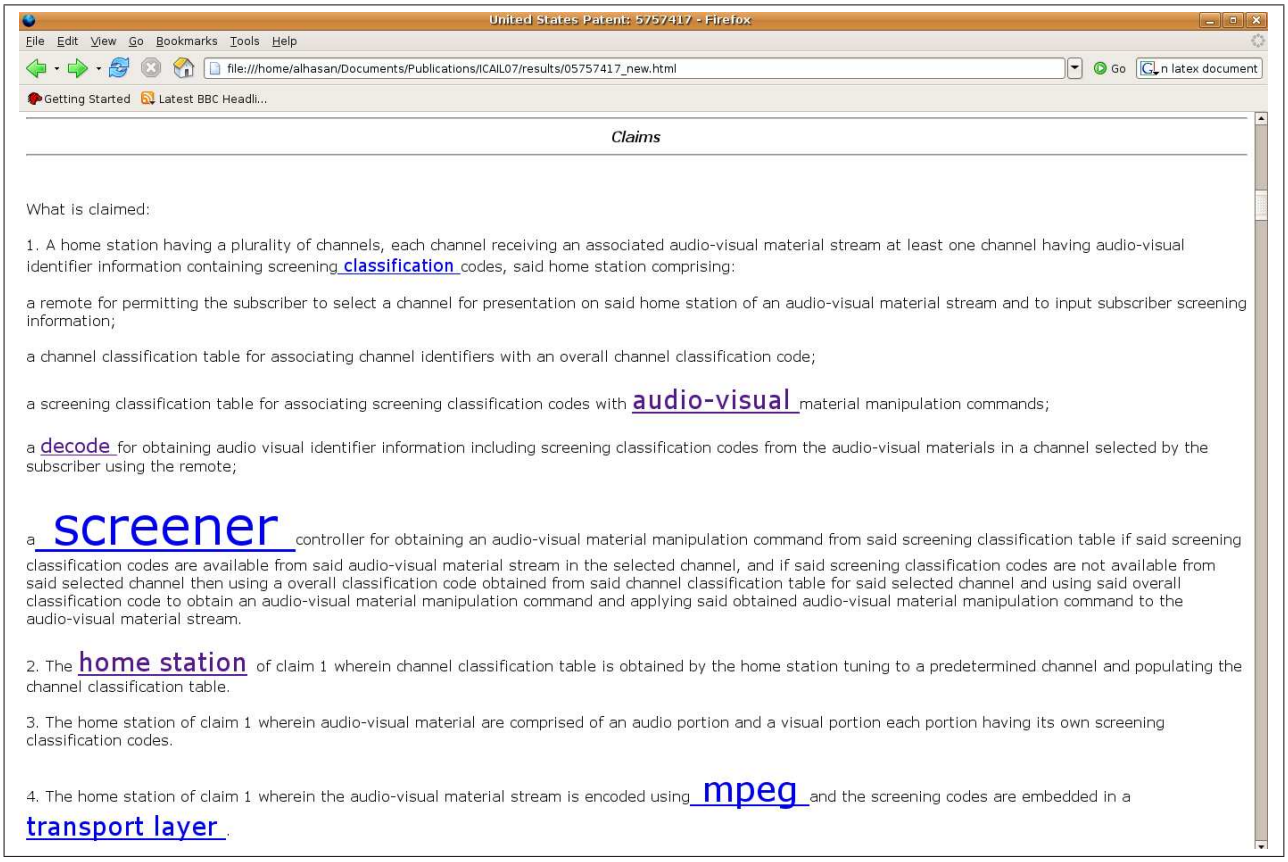

Figure 3: A screenshot of Patent Visualization In a browser window. The terms that are the most original in the claims are highlighted by using larger fonts. The font size also represents the degree of originality. Each terms is also hyperlinked to the patent that used the term for the first time.

Patent ID	Date	Words	First Used	Difference (in Days)	Support
06963637	11/8/05	*machine-accessible*	5/6/03	917	12
06181781	1/30/01	*java*	8/31/99	518	38
		applet	6/22/99	588	23
		www	1/30/01	0	16
		hyperlink	8/25/98	889	30
06775651	8/10/04	*unsupervised*	8/19/97	2548	27
		text-independent	6/30/98	2233	19
		non-enrolled	11/14/02	635	4
06219407	4/17/01	*spotting*	2/17/98	1155	9
		trained	8/2/94	2450	42
07003083	2/21/06	-	-	-	-
07079632	7/18/06	*browse*	3/17/98	3045	246

Table 1: Patent Claims Word Analysis using patent rating table

modules. Many different setting parameters can be altered. This is important because flexibility is the key for an efficient analysis. Different domains of the technology have different measure of prior works. Again, the sizes of their innovation sets also vary significantly. So, user needs to adjust the setting parameter according to the technological domain and the result-set configuration. In this system, User can add an word/term in the innovation set or can delete the same from it. User can also change the threshold of the time interval for which the innovation set will be considered. For instance, for a four years of threshold, all the words that appeared first in a patents within the previous 4 years of this patent's publish time, will be considered in the innovation set. Again, an innovation set can be selected by the minimum support count. Say, if a minimum support count is set to 20, a term shall not be considered in the innovation set, unless, it has at least 20 distinct usages after the first use.

Visualization Module Analyzing patent is a difficult task that requires substantial domain knowledge and it is hardly possible for an IP lawyer to be knowledgeable in many different domains, in which the client company files patents. So, if the important terms of a domain is highlighted, that can helps the lawyer in reviewing the patent's novelty instantly. In our patent ranking tool, we provide the patent analysts such options through the patent visualization module. This module actually displays the patent text, but in a creative manner. The text is formatted in html file and obviously, browser in the host machine are invoked to display them. Since, in patent, the words in the innovation set are considered to be the most important factor for its rating, those words are highlighted with different color in this visualization. Furthermore, the font size of the word is varied according to the novelty of the word; i.e., if a word is considered to be part of the innovation set, then it is displayed in a different color and its size is inversely proportional to the value of the date difference column in the patent rating table. Moreover, a hyperlink is also attached to the word that is linked to the patent, where this word appeared the first. Figure 3 shows an exemplary patent in a browsers window of the host machine.

6. RESULTS AND FUTURE WORKS

IBM has a large patent portfolio which currently has more than 40,000 patents [16], in more than ten different classes. In a broad categorization, they roughly fall under any of the following categories:micro-electronics, server, display, storage, network computing and software. In recent years, number of patents in the software and the business process categories are increasing rapidly. Most of these patents are comparably harder to assess, since, for software or business process, it is hard to disassociate the invention from the prior art. We used Claim Originality Analysis to analyze patents in three different portfolios related to software technology or business proceess that IBM was considering for divestiture. In each case, the results of the analysis were shared with the IP attorneys in the form of HTML reports showing for each patent, which words were considered to be most original in the claims of that patent at the time it was published. The IP attorneys found these results to be intriguing, and it helped them to focus their discussions on the most salient features of the portfolios. While there is still more work to do in proving the validity of these results in assessing patent value, the IP attorneys felt this to be a promising direction to pursue.

This is an ongoing research and hence, has substantial rooms for improvement. The improvement can be of 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 on the earliness of a patent, only. Although, it performs excellently for a pioneer effort, it is far from a perfect system. 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 claim's linguistic simplicity, claim's unambiguousness, claims 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 parameter. For instance, the "support" of a term, currently, just count 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 can also evolves over the time from the suggestions of the current user.

7. DISCUSSION AND CONCLUSIONS

At this point, since, the user is informed about the approach of our patent ranking, we like to reemphasize the usefulness of such a system. First and foremost, we provide the ana-

lysts, a system, where the analyst can both learn and rank. For instance, the idea of innovation set is simple, but is immensely valuable. This give the analysts a set of keyword on which that technology is evolving. Now, if the analyst is not well informed about that technology domain, he still can: (1) Get a quick hand-on knowledge on those keywords, (2) run a prior search on those keyword just by following the hyperlinks on those words, (3) Get a feeling of the novelty and impact of those words from the patent rating table, and (4) finally, based on his learning of the patent, he can change the default parameter of the patent ranking system to get a better result. From the experience to out IP teams, this was extremely helpful in expedite the patent evaluation. Secondly, In recent days, software or business process patents had 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 art [24]. 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 art or related work. Moreover, it provides the examiner a visual cue about the dominant keyword of that technology field, what is very helpful to obtain fast domain knowledge.

To summarize, we build 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, 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.

8. ACKNOWLEDGMENT

We like to thank Dr. Ying Chen and Dr. Jeffrey Kreulen of IBM Almaden Research center for numerous discussions and suggestions regarding this research work.

9. REFERENCES

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

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

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

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

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

[6] G. Karypis, E. Han, and V. Kumar, *CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling*, IEEE Computer: Special Issue on Data Analysis and Mining, 32(8),1999

[7] F. Beil, M. Ester, and X. Xu, *Frequent Term-Based Text Clustering*, SIGKDD, 2002

[8] H. Hacigumus, J. Rhodes and W.Scott Spangler, J. Kruelen, *BISON: Providing Business Information analysis as a Service*, International Conference of Extending Database Technology, Munich, 2006

[9] 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

[10] 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.

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

[12] `www.delphion.com`

[13] A. Doan, R. Ramakrishnan, and S. Vaithyanathan, *Managing information extraction: state of the art and research directions*, In Proceedings of SIGMOD, 2006.

[14] `www.patentcafe.com`

[15] `http://www.natcorp.ox.ac.uk/`, British National Corpus.

[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] S. Spangler, J. Kreulen, and J. Lessler, *Modeling Docuemnt Taxonomies*, IBM Research Report, RJ10288, 2003.

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

[20] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, The Morgan Kaufmann publications, 1999

[21] S. Chakrabarti, *Mining the Web: Analysis of Hypertext and Semi Structured Data*, The Morgan Kaufmann publications, 2003

[22] S. Spangler and J. Kreulen, *Interactive methods for Taxonomy Editing and Validation*, IBM Research Report, RJ10300, 2003

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

[24] 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.

[25] `http://www.bustpatents.com`

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

[27] `http://www.uspat.com/uspat/index.shtml`

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

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

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