Tsinghua Science and Technology

Volume 18 | Issue 4 Article 2

2013

Finding Nuggets in Patent Portfolios: Core Patent Mining and Its Applications

Po Hu
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Minlie Huang
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Xiaoyan Zhu
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Recommended Citation

Po Hu, Minlie Huang, Xiaoyan Zhu. Finding Nuggets in Patent Portfolios: Core Patent Mining and Its Applications. Tsinghua Science and Technology 2013, 18(4): 339-352.

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 02/10 pp339-352
Volume 18, Number 4, August 2013

Finding Nuggets in Patent Portfolios: Core Patent Mining and Its Applications

Po Hu, Minlie Huang*, and Xiaoyan Zhu

Abstract: Patents are critically important for a company to protect its core business concepts and proprietary technologies. Effective patent mining in massive patent databases not only provides business enterprises with valuable insights to develop strategies for research and development, intellectual property management, and product marketing, but also helps patent offices to improve efficiency and optimize their patent examination processes. This paper describes the patent mining problem of automatically discovering core patents (i.e., novel and influential patents in a domain). In addition, the value of core patent mining is illustrated by revealing the potential competitive relationships among companies in their core patents. The work addresses the unique patent vocabulary usage which is not considered in traditional word-based statistical methods with a topic-based temporal mining approach that quantifies a patent's novelty and influence through topic activeness variations. Tests of this method on real-world patent portfolios show the effectiveness of this approach over state-of-the-art methods.

Key words: text mining; core patent; patent novelty; patent influence; company competitor

1 Introduction

"The distance between innovation and marketplace is shrinking."[1] Indeed, innovation is advancing more quickly from conception to deployment and has become the principal driver of world economic growth. Innovators and enterprises rely on *patents* to secure their core technologies and investment capital, and to bring new products and services to the global market without a flurry of third-party infringements. In 2011, the top 1400 largest global companies invested over 456 billion euros in their Research and Development (R&D) departments[2]. Without patent protection, R&D investments would be significantly reduced, limiting the possibility of technological advances and breakthroughs.

A company's patent portfolio forms the critical part of its Intellectual Property (IP). Effective patent portfolio management requires patent mining techniques to assess patent quality, identify technology gaps to invest or divest IP assets, and reveal a company's potential competitors. Among the various patent mining tasks, a key one is to discover core patents in a patent portfolio. A *core patent* is a patent in which the invention is novel and influential in a domain. *Core Patent Mining* (CPM) has both strategic and financial value in almost all aspects of the patent business. "If IP management and patent mining is not about the core, it is not worth doing."[3]

Firstly, CPM assists companies in patent licensing and investment. Patent licensing can bring huge profits to a company. According to Deloitte & Touche, Texas Instruments annually receives $500 million in patent royalties, more than its net income from manufacturing[4]. Patent investment can also bolster a company's IP assets and protect its R&D. In the latest twist of the smartphone patent wars, Google

• Po Hu, Minlie Huang, and Xiaoyan Zhu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: hup09@mails.tsinghua.edu.cn; {aihuang, zxy-dcs}@tsinghua.edu.cn.

∗ To whom correspondence should be addressed.

Manuscript received: 2013-05-27; accepted: 2013-06-17

spent $12.5 billion to acquire Motorola Mobility, aiming to use Motorola's patents to defend Android against competitor (Apple, Microsoft, Nokia, etc.) litigation. However, not all patents have the potential to sell licenses or deserve investment. In fact, only 5% of patents are licensed and 1% of patents actually generate revenue[3]. Manually identifying the prominent patents is like looking for a needle in a haystack. Automatic discovery of core patents can significantly promote a company's ability to explore opportunities related to patent licensing and investment.

Secondly, CPM helps a company maintain its patent portfolio. A patent grants its assignee a set of exclusive rights for a period of time (i.e., full term). During a patent's full term, maintenance fees must be paid periodically to keep the patent valid. For companies owning a large portfolio, the cost is too high for all patents to complete their full terms, so many patents are abandoned once they no longer have value. In the U.S., early-expired patents between 2007 and 2011 included over 30% of all issued patents[5]. Patent maintenance decisions are very important, as improper patent abandonments can eliminate a company's technology competitiveness. For large companies, making patent maintenance decisions is time-consuming and costly. CPM can help IP analysts make abandonment decisions by focusing on those patents with low technical value to balance between reducing maintenance fees and minimizing the risk of losing core patents.

Thirdly, CPM can identify and track competitors' activities. As one of the few real indicators of future product release, patents reveal research details long before the products hit the marketplace. From the discovered core patents, a company can identify its competitors and understand their R&D planning and technology strengths. By continuously monitoring the change of core patents and their owners, a company can spot competitive threats much sooner and adjust its strategies to survive in a fiercely competitive market.

Finally, CPM helps overburdened patent offices optimize their patent examination processes. The rapid growth of patenting activities has applied unprecedented pressure on patent offices like the United States Patent and Trademark Office (USPTO). In 2011, pending applications in USPTO had increased by 360% compared to 1991 and the average pending time had skyrocketed to 33.7 months per patent[5]. In addition, today's high-pace technology environment makes the patent examination process more complicated and error-prone than ever before. Examiners must collaborate with domain experts to comprehend emerging technologies and cross-disciplinary inventions, yet may still miss some important prior art in the patent review. As a result, numerous substandard applications are clogging limited examiners from issuing high-quality patents. Automatic ranking of a patent's technical value permits examiners to prioritize their time in examining those high-quality applications; thus, greatly promote the efficiency of patent offices.

This paper studies the problem of automatic discovery of core patents from patent texts. A topic-based temporal mining approach is given that quantifies the novelty and influence of a patent and ranks all patents by combining the novelty and influence scores. Top-ranked patents are selected as core patents in the domain. Unlike traditional word-based statistical methods, this approach considers the unique vocabulary usage in patent literature and can effectively discover core patents from massive patent databases. Finally, this study demonstrates the value of core patent mining by identifying competitors in the petroleum industry.

Despite the theoretical developments in information retrieval and data mining, advanced patent analysis tools are still in their infancy. Existing patent search engines (e.g., Google Patents[6]) and patent mining systems (e.g., Delphion[7], IPVision[8], and Aureka[9]) do not have core patent mining functions. Thus, this research is a pioneering effort towards this goal which aims to assist IP professionals to perform effective and scalable patent analysis work.

2 Related Work

This work is related to several lines of research, including patent quality assessment, patent prior art search, document novelty and influence analyses, and company competitor identification.

2.1 Patent quality assessment

Liu et al.[10] studied the problem of quantifying and predicting patent quality. They proposed a supervised graphical measurement model in which the patent quality was a latent variable. The value of the patent quality was estimated from correlated measurements, including court ruling decisions, patent citations, and some lexical features (e.g., number of claims, patent length, and text similarity between the abstract and claims). Jin et al.[11] worked on providing patent maintenance recommendations. Their method extracted a set of patent features, such as claim length, patent vocabulary size, citation count, and cohesion between abstract and description. A binary decision tree classifier was trained on historical patent maintenance decisions to predict whether a patent should be maintained. Finally, they proposed a network-based optimization process to refine the prediction results.

This work differs from these previous studies in that it is unsupervised and only utilizes the patent text as input. Court ruling decisions[10] and patent maintenance records[11] are only available for a limited number of patents. For example, only patents involved in Federal Circuit court cases have court ruling decisions and patents expired before 1995 have no maintenance information on the USPTO website. In addition, Liu et al.[10] and Jin et al.[11] focused more on the writing quality in the patents, while this study assesses the patent quality from the technical perspective (i.e., novelty and influence).

Several studies[12, 13] have studied patent quality in view of the economic value and legal strength of a patent as opposed to its intrinsic technical value to analyze the profit ability, protection breadth, and market demand of patents. Those factors cannot be obtained from the patent text and are outside the scope of patent mining and retrieval.

2.2 Patent prior art search

Patent prior art search (a.k.a. patent novelty search) aims to retrieve the patents that constitute the prior art of a given patent application to determine whether the retrieved patents invalidate the claims in the application.

Guo and Gomes[14] studied automatic patent ranking for prior art search. They proposed a method called SVM Patent Ranking (SVM_{PR}) which incorporates the differing importance of citations made by examiners and citations made by inventors into an optimization problem. SVM_{PR} ranks examiner cited patents higher than inventor cited patents, which are ranked higher than uncited patents. Xue and Croft[15] transformed a patent into an effective query for prior art search and focused on the effects of different sections of a patent. Azzopardi et al.[16] gave an international online survey of patent experts who had at least 10 years experience in patent retrieval. Their results indicated that patent novelty search is the most frequent type of search task and is routinely performed by patent experts. By contrast, 64.2% of the subjects rarely evaluated the patent writing quality. Their findings suggested that patent experts are much more focused on the novelty of patents than their writing quality.

There are also several workshops that aim to spark the development of patent prior art search, such as NTCIR[17], CLEF-IP[18], PaIR[19], and TREC-CHEM[20].

2.3 Document novelty and influence analysis

Hasan et al.[21] analyzed the patent novelty as assessed from patent claims. Their method obtained a set of keywords by extracting words and phrases from patent texts and removed frequently used keywords as domain stopwords. The score of a keyword in a patent is equal to the ratio of the keyword's support to its age. A patent's novelty score is then the sum of all its keyword scores. They implemented their method into the IBM patent mining system *SIMPLE*[22]. Shaparenko et al.[23] discovered important documents in a document collection. They first clustered the documents by their word bags. To determine the importance of a document, the most similar documents were examined by comparing their publishing dates to the subject document. A document was important if it had fewer similar documents published before it and had more similar documents published after it.

Unlike the word-based methods used in Hasan et al.[21] and Shaparenko et al.[23], the current approach is topic-based and addresses the unique vocabulary usage in the patent literature. This method eliminates the temporal bias in document novelty and influence analysis, which is not considered in previous works. The results are compared to Hasan et al.[21] and Shaparenko et al.[23] as baselines. The results show that this method is more effective.

Gerrish and Blei[24] analyzed the influence of scientific documents. They proposed a probabilistic model called the *Document Influence Model* which measured how past articles influence future articles by the changes in their thematic content. The present work shares their hypothesis that a document's influence is corroborated by how the language of its field changes after its publication. However, the present algorithm is much less complex than theirs, which requires complicated variational inferences. In addition, this algorithm integrates novelty and influence analyses into a unified framework, while they only analyzed the document influence.

Shaparenko and Joachims[25] also sought to identify influential documents. They proposed a language model based likelihood ratio test to determine how much influence a document has on another document. Unlike their pair-wise influence analysis, the present method assesses a patent's influence based on its impact on the technology developments in its field.

2.4 Company competitor identification

Bao et al.[26] proposed an algorithm called *CoMiner* to discover company competitors based on the co-occurrence of company names in web pages. Ma et al.[27] inferred competitor relationships from intercompany networks derived from company co-occurrence in online news articles. Those works are totally different from the present research since company co-occurrence in web pages and news articles may not necessarily represent competitor relationships. This method identifies competitors from the discovered core patents which shed light on the companies' R&D secrets that cannot be obtained from web pages and news articles.

3 Problem Formulation

Suppose the patent set in a domain is $\mathcal{D} = \{D_t | t = 1, \cdots, T\}$, where $D_t \subset \mathcal{D}$ contains the patents issued at year t. The patent novelty and patent influence are then defined as follows.

Definition 1 Patent novelty A patent $d \in D_t$ is novel if its ideas are not presented or little mentioned in its prior art, i.e., $\mathcal{D}_{\text{PA}}(d) = \{D_i | 1 \leqslant i < t\}$.

Definition 2 Patent influence A patent $d \in D_t$ is influential if its ideas are adopted or expanded by its follow-up work, i.e., $\mathcal{D}_{\text{FW}}(d) = \{D_i | t < i \leqslant T\}$.

This paper uses *latent topics* to represent the ideas in a patent and quantifies patent novelty and influence by topic activeness variations. Then, the novelty and influence scores are combined to rank the patents in $\mathcal{D}$ to select the top-ranked patents as core patents in the domain.

As scientific literature with legal significance and potential profits, patents have complex structures and special nomenclature, which differ from news articles, blogs, advertisements, and even research papers intended for non-profit use. The sophisticated patent language can pose significant challenges to patent mining and retrieval. This paper addresses the unique vocabulary usage in the patent literature. The semantic meanings of technical terms in patents are often inconsistent and indeterminate due to the following three reasons.

(1) **Lack of standard terminology for emerging technologies.** Before a new technology becomes mature, inventors use different terms to describe the same thing in their patents. For example, during the development of Digital Video Disc (DVD) system before the name "DVD" had been given, American companies such as Time Warner and IBM used "Optical Disk" in their patents, while Japanese companies such as Panasonic, Toshiba, and Hitachi used "Optical Disc" and "Optical Record Carrier"[28].

(2) **Heterogeneity in nomenclature.** Some technical terms, especially chemical and biological entities, have more than one reference name that are semantically identical to each other. For example, "Valium" (a.k.a. "Diazepam") has over 149 commonly accepted names and most of them are used in patents[22]. Another example is "speaker recognition" and "voice identification" in computer science. The choice of these aliases is mainly determined by the inventor's writing style.

(3) **Deliberate vocabulary inconsistency.** Inventors tend to use ambiguous words and expressions in their patents to pass the patent examination or extend the patent protection scope (e.g., "mobile unit" instead of "vehicle"[15]). In addition, some inventors create their own terminologies to hide certain technical details. For example, patent *WO2011109143* describes a method of producing hydrocarbon from oil shale where the inventors used the self-invented term "organic-rich rock" instead of the commonly used "oil shale" in the patent.

A combination of these factors greatly increases the difficulty of capturing the real contributions of a patent. Since 2008, the Federal Circuit has acknowledged that patent interpretation is not a rigid process in which only one answer applies. Instead, different judges may select different angles to generate multiple reasonable interpretations of a patent[29].

Classical word-based statistical methods, such as TF-IDF[21], word similarity[23], and word burstiness[30], are based on a homogeneous assumption that "imitation is the highest form of flattery". Those methods can cause serious word mismatch problems in the patent literature, since inventors may use different terms in different fields to describe the same technology. In addition, traditional TF-IDF methods can only discover the significant-frequent keywords, but will miss many *significant-rare* keywords which are low-frequency but highly informative and reliable. The patent search survey[16] gives another piece of evidence. 55.6% of patent experts rate "query expansion" as "very important" to formulate effective queries, while only 2.5% rate this as "unimportant". The result also indicates that classical word-based statistical methods are not suitable for patent mining and retrieval.

The unique vocabulary usage in the patent literature is addressed here by a topic-based temporal mining approach. Topics are used to depict the ideas in the patents and cluster synonyms and relevant keywords to avoid the word mismatch problem. In addition, the topic activeness trend is used to characterize the technology developments in a field with more effective results than traditional word statistics.

4 Methodology

The current method consists of three steps as follows.

Step 1 Preprocess the patent text and identify latent topics in the patent set of a domain using the topic model.

Step 2 Model the topic activeness trend as a Markov-Modulated Poisson Process (MMPP) and automatically remove noisy topics through topic activeness pattern matching.

Step 3 Calculate the patent novelty score and the influence score based on the topic activeness trends. Rank all the patents by their scores and select the top-ranked patents as core patents in the domain.

4.1 Latent topic identification

Patents are semi-structured documents consisting of several sections, including title, abstract, claims, background, and detailed descriptions. Each section has a specific purpose. For example, the abstract gives an overview of the invention, the background summarizes the related work, and the detailed descriptions explain the procedures of the invention in detail. The claim section is the heart of a patent as claims define the protection scope of the invention and their interpretations are the central issue in most patent disputes. This paper uses *title*, *abstract*, *claims*, and *detailed descriptions* as the most important sections to analyze patent novelty and influence.

All patents are transformed into lowercase with stopwords removed, followed by stemming and POS tagging. Only nouns and adjectives are maintained since most technical terms are nouns and adjectives. However, there are still some high-frequency but low-informative words (i.e., domain stopwords) in the processed patent text. This algorithm then automatically removes domain stopwords by noisy topic filtering as discussed in Section 4.2.

The model uses the distributed version of the Latent Dirichlet Allocation model[31], a highly scalable parallel-processing topic model, to efficiently discover the latent topics inside the patents in a domain. Suppose there are K topics $\{z_1, \cdots, z_K\}$ in the patent set $\mathcal{D} = \{D_1, \cdots, D_T\}$. Topic z_k is a probabilistic distribution of words in the word set V of $\mathcal{D}$; that is, z_k governs the multinomial word distribution $\{p(w|z_k)\}_{w\in V}$. A patent $d \in \mathcal{D}$ is a probabilistic distribution of topics, i.e., $\{p(z_k|d)\}_{k\in\{1,\cdots,K\}}$. The model is based on unigram words, yet some unigrams are too general to understand their meanings. To better represent the semantic meaning of a topic, the model also estimates the topic-phrase distributions based on the topic-word distributions. The model only considers bigram phrases to avoid the word sparseness problem. The probability of a bigram phrase $w_i w_j$ ($w_i \neq w_j$) generated by a topic z_k is estimated as follows.

$$p(w_i w_j|z_k) = p(w_i|z_k) \times p(w_j|z_k) \times p(w_i w_j) \quad (1)$$

where $p(w_i w_j)$ is the occurrence probability of the phrase $w_i w_j$ estimated as $p(w_i w_j) = \frac{\sum_{d\in\mathcal{D}} c(w_i w_j, d)}{\sum_{w_i', w_j' \in V} \sum_{d\in\mathcal{D}} c(w_i' w_j', d)}$, where $c(w_i w_j, d)$ is the count of the phrase $w_i w_j$ in d.

Finally, for each topic z_k, the model extracts a set of unigrams and bigrams with the highest probabilities under z_k (i.e., $p(w|z_k)$ and $p(w_i w_j|z_k)$) to be the topic signatures of z_k.

4.2 Topic activeness trend modeling

For each discovered topic, a discrete-time MMPP[32] is used to model its activeness trend as shown in Fig. 1. MMPP is a doubly stochastic Poisson process whose rate, λ, varies according to a latent Markov chain. The model is well-known in communication theory. This analysis assumes that the temporal behavior of a topic can be characterized by an MMPP model with respect to its activeness variations. The time span of $\mathcal{D}$ is divided into T time intervals (each lasts for a year) and a topic's activeness varies across different intervals. Suppose there are M levels of topic activeness and each level is represented by a hidden state in

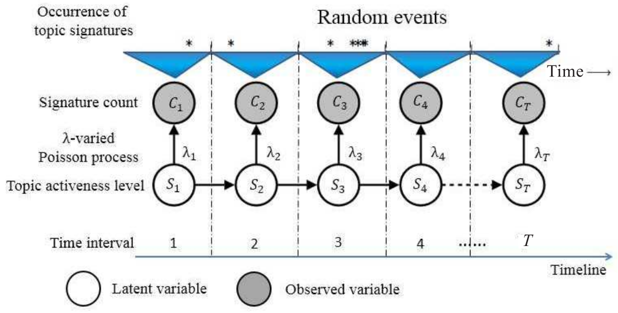


Fig. 1 Topic activeness modeling by MMPP.

ascending order as $S_i \in \{1, \cdots, M\},\ i = 1, \cdots, T$. The MMPP observations are the occurrence count of the topic's signatures in each interval with the emission probabilities set as Poisson distributions:

$$B[C_i][S_i] = \frac{{\lambda_i}^{C_i} \mathrm{e}^{-\lambda_i}}{C_i!} \tag{2}$$

where $B[C_i][S_i]$ is the emission probability for state S_i to generate C_i signatures of the topic in the i-th interval and λ_i is the expectation of the topic's signature count in the i-th interval. Equally divide the range of all observations $\{C_i | 1 \leqslant i \leqslant T\}$ into M bins with the initial value of λ_i set as the mean value of the observations in the bin holding C_i. Thus, λ_i can be viewed as the raw topic activeness as opposed to the topic activeness level S_i and higher topic activeness levels have higher probabilities of generating more signatures of the topic. The initial state probabilities and transition probabilities of the Markov chain and the rate parameters (i.e., $\lambda_i,\ i = 1, \cdots, T$) of the Poisson process are estimated via the EM algorithm[33] with complexity $O(M^2T)$. Finally, sort the values of $\{\lambda_i | 1 \leqslant i \leqslant T\}$ in ascending order and set the topic activeness level in the i-th interval equal to the rank of λ_i, i.e., $S_i = \text{Rank}(\lambda_i)$.

MMPP provides a uniform level of abstraction over the raw data stream with the same number of topic activeness levels for all topics in a domain. Thus, the topic activeness trends $\{S_i | 1 \leqslant i \leqslant T\}$ inferred by MMPP are comparable among different topics in a domain, while the raw topic activeness rates $\{\lambda_i | 1 \leqslant i \leqslant T\}$ are not.

Topic models have a common problem in that the number of topics must be determined in advance. Thus, some topics inevitably have little semantic meaning (i.e., noisy topics). Manually examining topic signatures to identify noisy topics is inefficient, tedious, and often requires domain knowledge. This algorithm uses an automatic noisy topic filtering method that assumes that noisy topics can be characterized by their activeness variations with two classes of noisy topics. The first class includes *trendless topics*, whose activeness trends have few variations along the timeline (e.g., the topic is seldom active). The second class is *jittering topics* where a topic jitters as its activeness fluctuates capriciously in adjacent time intervals (e.g., rises from inactive to bursty, then again drops to inactive). With three topic activeness levels (i.e., 1-inactive, 2-active, and 3-bursty), ten basic jittering patterns can be defined as shown in Table 1. More complex jittering patterns can be decomposed into the several basic jittering patterns. For example, the jittering pattern "2 -> 3 -> 1 -> 3 -> 2" consists of two basic patterns "2 -> 3 -> 1" and "1 -> 3 -> 2". If a topic jitters in too many time intervals, it is a jittering topic. All the trendless topics and jittering topics are removed as noisy topics.

The removal of noisy topics in a domain helps eliminate domain stopwords which are the signatures of noisy topics. This method avoids the drawbacks of traditional domain stopword removal methods. For example, some methods regard high-frequency or low-frequency words as domain stopwords, but this may eliminate important or novel technical terms as well. Other methods use words that occur frequently in different domains as domain stopwords, yet some keywords are interdisciplinary (e.g., "Super Computer" is used in both computer and biology domains).

4.3 Patent novelty and influence analysis

After removing the noisy topics in a domain, a patent's novelty and influence are quantified by analyzing the topic activeness variations along the timeline.

For each patent $d \in D_t$, the current method uses its dominant topics $Z_{\text{Dom}}(d) = \{z | p(z|d) > 10\%\}$ to depict the ideas in d. The novelty of d is evaluated by focusing on the activeness trends of the dominant topics in d's prior art $\{[S_{i,z}]|_{i=1}^{t-1}, z \in Z_{\text{Dom}}(d)\}$, where $S_{i,z}$ is the activeness level of topic z in the i-th interval. Low topic activeness in the prior art indicates that d is very

Table 1 Ten basic jittering patterns for the three topic activeness levels (1-inactive, 2-active, 3-bursty).

Number	Pattern	Number	Pattern
1	1 -> 2 -> 1	2	1 -> 3 -> 1
3	1 -> 3 -> 2	4	2 -> 1 -> 2
5	2 -> 1 -> 3	6	2 -> 3 -> 2
7	2 -> 3 -> 1	8	3 -> 1 -> 3
9	3 -> 2 -> 3	10	3 -> 1 -> 2

novel. The influence of d is evaluated by focusing on the topic trends in d's follow-up work $\{[S_{i,z}]|_{i=t+1}^{T}, z \in Z_{\text{Dom}}(d)\}$. High topic activeness in the follow-up work indicates that d is very influential. Figure 2 illustrates the analysis process in which d is highly novel and influential in topic 2, yet has low novelty in topic K and low influence in topic 1.

However, a temporal bias exists in the patent novelty and influence analyses because an old patent has fewer patents in its prior art and more patents published after it. Also, a new patent has fewer patents published after it, yet has more patents in its prior art. In an extreme case, the oldest patents have infinite novelty since they have no prior art at all, while the newest patents have zero influence since they have no follow-up work. Thus, old patents tend to be over-estimated in their novelty and influence, while new patents are under-estimated. We have noticed that core technologies (and also core patents) are time-sensitive and their values depend on technology developments. Thus, it is more appropriate to measure a patent's novelty and influence within a certain period of the topic activeness trends. This paper uses a *time decay factor* to restrict the scope of the topic trends and eliminate the temporal bias in patent novelty and influence analysis.

Consider two typical window functions used to determine the time decay factor, the *Rectangular window*, and the *Gaussian window*. Suppose Δt is the time difference between two time points and 2σ is the window size. The Rectangular window and the Gaussian window are defined as:

$$F(\Delta t) = \begin{cases} 1, & \text{if} \quad |\Delta t| \leqslant \sigma; \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

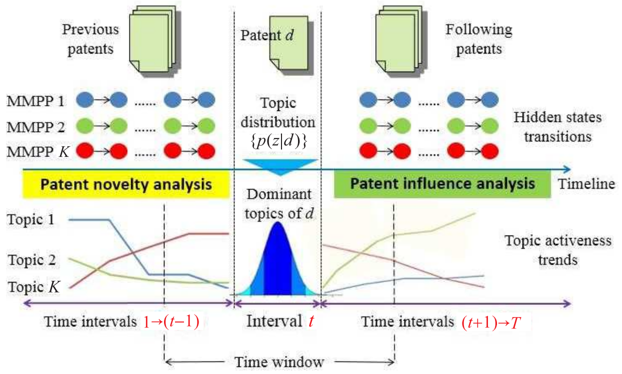


Fig. 2 Patent novelty and influence analysis through topic activeness trends.

$$F(\Delta t) = \mathrm{e}^{\frac{-\Delta t^2}{2\sigma^2}} \tag{4}$$

Figure 3 illustrates the two window functions. The Rectangular window only considers the values within the time window and has the two explicit thresholds $\pm\sigma$. The Gaussian window considers all values on the timeline with a decay factor based on Δt and σ, thus it has a smoothing effect on the topic trend normalization. For each topic z, the *normalized topic activeness level*, $S_{\text{Nor}}(i,z)$, in the i-th interval is defined as:

$$S_{\text{Nor}}(i,z) = F(\Delta t) \cdot S_{i,z} \tag{5}$$

where Δt is the time difference between the pending interval, t, and the neighboring interval, t', i.e., $\Delta t = t - t'$.

To quantify the novelty of patent d, determine the novelty score of topic $z \in Z_{\text{Dom}}(d)$ as follows:

$$\text{Novelty}(z) = \frac{t-1}{\sum_{i=1}^{t-1} S_{\text{Nor}}(i,z)} \tag{6}$$

The novelty score of d is the sum of its dominant topics' novelty scores, $\text{Novelty}(d) = \sum_{z \in Z_{\text{Dom}}(d)} \text{Novelty}(z)$.

To quantify the influence of d, determine the influence score of topic $z \in Z_{\text{Dom}}(d)$ as follows:

$$\text{Influence}(z) = \frac{\sum_{i=t+1}^{T} S_{\text{Nor}}(i,z)}{T-t} \tag{7}$$

The influence score of d is the sum of its dominant topics' influence scores, $\text{Influence}(d) = \sum_{z \in Z_{\text{Dom}}(d)} \text{Influence}(z)$.

Both $\text{Novelty}(z)$ and $\text{Influence}(z)$ are insensitive to insignificant and irregular topic activeness variations and can avoid overfitting of certain patterns of the topic activeness trends.

The score of d is the product of its novelty and influence scores, $\text{Score}(d) = \text{Novelty}(d) \cdot$

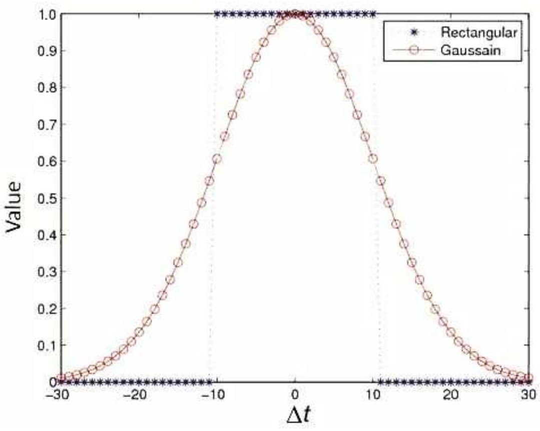


Fig. 3 Rectangular and Gaussian windows (σ= 10).

Influence(d). All patents are then ranked by their scores and the top-ranked patents are selected as core patents in the domain.

5 Tests

5.1 Dataset and domain definition

A dataset was constructed with patents from the petroleum industry. 108 large petroleum companies listed by Wikipedia[34] were selected. For each company, the infobox and history section of its web page on Wikipedia were downloaded to obtain the company's name alias, subsidiary companies, and acquired companies. Then, the patents assigned to the 108 companies and their subsidiaries and acquisitions from the USPTO patent database[35] were downloaded. Our dataset contains 82 648 U.S. patents spanning from 1976 to 2010.

A set of domains was then defined based on the United States Patent Classification (USPC) system, which is an authoritative classification standard adapted by USPTO. Each U.S. patent has a mandatory USPC class according to its technical subject. These classes are used here to classify the patents. A USPC class is defined as a domain if it has at least 800 patents in the dataset. The smaller classes of patents are not used in the tests, since domains with small numbers of patents are more suitable for manual analyses. To ensure the semantic consistency of the same technical term in a domain, one domain can only correspond to one USPC class, and vice versa. The only exceptions are the domains for "Organic compounds" and "Synthetic resins or natural rubbers", in which the corresponding classes are labeled as "one class series" in the USPC system. Table 2 shows the 21 domains that were defined.

Table 2 Domain definition and patent statistics. There are 65 846 patents in the 21 domains, accounting for 79.7% of the patents in the dataset.

USPC class	Domain definition by USPC	Patent count
520-528 Series	Synthetic resins and natural rubbers	12 963
532-570 Series	Organic compounds	9596
166	Wells	8179
175	Boring and penetrating	3929
502	Catalyst and solid sorbent	3780
208	Mineral oils	3776
585	Chemistry of hydrocarbon compounds	3201
423	Chemistry of inorganic compounds	1971
264	Plastic shaping or treating	1965
428	Stock material	1952
514	Drug	1914
508	Lubricant	1822
367	Acoustic wave systems	1818
073	Measuring and testing	1808
210	Liquid purification	1290
324	Electricity measurement	1110
252	Compositions	1029
702	Data measurement	1012
044	Fuel and related compositions	968
504	Plant protection	938
422	Chemical disinfection	825

5.2 Parameter setting and baseline methods

The parameters were set as follows. In the topic identification step, 50 topics were assigned to each domain with over 8 000 patents, 20 topics to each domain with 3 000 to 4 000 patents, and 10 topics to each domain with less than 2 000 patents. Excessive numbers of topics are not good for the algorithm, since they will produce less discrimination among topics and more noisy topics. The distributed LDA model was implemented in the MALLET toolkit[36] and run in parallel (3.0 GHz Dual-core Intel Xeon processors, 4 GB memory) to speed up the topic mining process. For each identified topic, 50 unigrams and 50 bigrams were selected with the highest probabilities in the topic to be the topic signatures. The topic activeness modeling step used three topic activeness levels (i.e., *inactive*, *active*, and *bursty*) in the MMPP model as a trade-off between performance and training complexity. The effectiveness of the MMPP model was illustrated by comparing to *equal-size binning* to model the topic activeness trends that uses neither a Poisson process nor a Markov chain. In equal-size binning, the range of a topic's signature counts is equally divided into three bins corresponding to the three topic activeness levels, with the topic activeness level in an interval determined by the bin holding its signature count. In the final step, topics whose activeness trends reach *inactive* or *bursty* only once, or which have jitters in over 50% of the intervals along the timeline, are removed as noisy topics. In addition, the effects of the Rectangular window and the Gaussian window in the topic activeness normalization were also evaluated.

Three word-based statistical methods from previous studies were implemented for comparison to the present approach. Baseline 1 (COA1)[21] removes stopwords from the patent text and any words whose document frequency exceeds 90% of the patent set. For each word w in patent d, the contribution of w is determined as:

$$\text{Contribution}(w) = \max\left(\frac{\text{support}(w) - 2}{\text{age}(w) + 1}, 0\right) \quad (8)$$

where age(w) is the time difference between the earliest year w occurs in the patent set and the issue year of d and support(w) is the number of follow-up patents that contain w. A word contributes to patent d only if its support exceeds a threshold (which was set to 2[21]). The score for d then equals the sum of the contributions of the words in d.

Baseline 2 (COA2)[21] uses the same procedures as in baseline 1 (COA1) with the only difference being that the patent score is equal to its word count after removal of the domain stopwords. Both COA1 and COA2 are used for patent evaluations in the IBM *SIMPLE*[22] system.

In baseline 3 (KeyPlayer)[23], each patent is represented as a TF-IDF vector of its words after removing the stopwords. Then for each patent d, the method finds the 50 most similar patents of d using the cosine similarity between the patents' TF-IDF vectors. The lead/lag index of d is then calculated as

$$\text{Index}(d) = \frac{\text{Follower}(d) - \text{Leader}(d)}{50} \quad (9)$$

where Follower(d) is the number of similar patents published after d and Leader(d) is the number of similar patents published before d. The lead/lag index ranges from -1 to 1 and is used as the score of d.

5.3 Evaluation metrics

Performance evaluations of core patent mining algorithms are extremely difficult, since there is no gold standard and manually labeling of each patent's novelty and influence is neither scalable nor consistent. This paper used two indicators to assess the discovered core patents in a domain. In addition, the human-written patent summaries from *Derwent Innovations* $Index^{SM}$ are used to show the effectiveness of this approach over the baselines.

The first indicator is *patent forward citation*, which is the citation count a patent received from subsequent patents. Before a patent application is issued, the examiners at USPTO are obligated to add patent citations beyond those provided by the inventors to ensure the quality of the citations. Thus, novel and influential patents (i.e., core patents) are more likely to receive more citations than the non-core patents in a domain.

Since new patents are less cited than old ones and may not contain citations to contemporary patents, patent forward citations tend to under-estimate the importance of new patents. To eliminate this bias, the discovered core patents were evaluated for each year, instead of for the whole timeline. All patents published in the same year were ranked by their scores and by their forward citation counts as the gold standard. Then, the Spearman correlation coefficient of the two patent rankings was calculated to assess the algorithm performance. In addition to the ranking evaluation, the 25% most cited patents were identified as the real core patents in the domain, with the 25% highest scored patents as the discovered core patents. The precision and the Mean Average Precision (MAP) of the two patent sets were calculated to complement the ranking correlation coefficient. Thus, new patents are only evaluated against their contemporary patents so the result is not dominated by old patents. Finally, the mean metric value of all the years' metric values is used to evaluate the algorithm.

The second indicator is *patent maintenance status*, which denotes if a patent is maintained or abandoned by its assignee. In the U.S., a patent can be kept valid for up to 20 years, with maintenance fees paid by the 4th (E1 stage), 8th (E2 stage), and 12th (E3 stage) years after the issue date of the patent. Due to the significant increase in the maintenance fees from the E1 stage to the E3 stage, assignees tend to abandon worthless patents as early as possible (e.g., at the E1 stage) and only those truly important patents are maintained at the E3 stage to complete their full terms. From 2007 to 2011, on average 90.6% of the patents were maintained in the E1 stage, 72.1% were maintained in the E2 stage, and 50.6% were maintained in the E3 stage[5]. Thus, only around half of the issued patents were truly important to their assignees.

A list of 958 523 patents abandoned between 1995 and 2011 were obtained from the official USPTO gazettes[37]. The time difference between the expiration and the issue year of a patent shows if a patent was abandoned at the E1, E2 or E3 stage. The results also show which patents were maintained throughout the entire 20-year period.

The gold standard assumes those patents abandoned at the E1 stage are *non-core patents*, which may be inappropriate for rapidly evolving industries such as the IT industry, in which even a core patent may become out-of-date and be abandoned during the E1 stage. The patents maintained throughout the 20 years in Fig. 4 are assumed to be *core patents*. Other patents, such as patents that were abandoned at the E2 or E3 stages, patents that expired before 1995, and newly issued patents, cannot be judged as core or non-core patents due to the complicated economic factors and lack of maintenance information. The dataset then had 9 527 core patents and 6 078 non-core patents in the 21 domains. The patents were ranked by their scores in each domain with the top 20%, 40%, 60%, and 80% of the patents assigned to be core patents to construct 4 cut-off levels. The False Positive Rate (FPR) and the True Positive Rate (TPR) were calculated for each cut-off level to draw the algorithm's Receiver Operating Characteristic (ROC) curve and Area Under the ROC Curve (AUC) to evaluate the algorithm performance.

5.4 Test results

The tests first compared the effects of different configurations of the approach, as discussed in Section 5.2. Table 3 shows the details of the configurations, including the topic activeness modeling (BR *vs.* MR), noisy topic filtering (MR *vs.* MRTF), and the time decay factor (MRTF *vs.* MGTF). The algorithm performance was also tested by varying the window size parameter σ to be 1, 5, 10, and 15 years. Figure 5 shows the average results on the 21 domains in the dataset.

The results in Fig. 5 show that for all four methods, a large window (σ of 10 or 15 years) gives better core patent mining performance than a small window (σ of 1 or 5 years). Small windows only consider the most recent topic trends and lose much information about technology developments in a domain. Thus, a large window is desirable to give sufficient information to estimate a patent's novelty and influence. A very large window (σ of 15 years) may harm the performance indicated by some metrics (e.g., precision and AUC in Fig. 5) since patents published long before or after the subject patent may not influence or be influenced by that patent, so a large window may introduce unnecessary topic activeness variations that harm the algorithm's performance. MR outperforms BR on all the metrics for large windows which shows that MMPP is more effective than equal-size binning because MMPP more accurately estimates the topic activeness trend through global model fitting, especially when the distribution of signature counts is very uneven. Besides, MMPP has a better smoothing effect through the transition probabilities when transforming the observed signature counts into the topic activeness levels. MRTF outperforms MR on all the metrics, especially on the patent ranking correlation as shown in Fig. 5c which shows that topic filtering effectively removes noisy topics and improves the algorithm's performance. On average, 36.5% of the discovered topics were identified as noisy topics in the 21 domains, with the noisy topic signatures successfully capturing many domain stopwords in the petroleum industry, such as "effective amount", "source device", and "weight percent". Finally, the Gaussian window (MGTF) outperforms the Rectangular window (MRTF) on all the metrics. Since the Rectangular window simply drops the non-recent topic information to eliminate the temporal bias in core patent mining, while Gaussian window maintains more useful information through the smoothing factor. Thus, the Gaussian window better reflects the time-sensitive characteristics of technology developments.

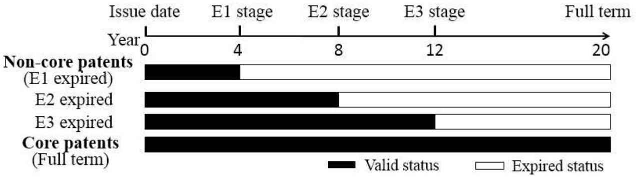


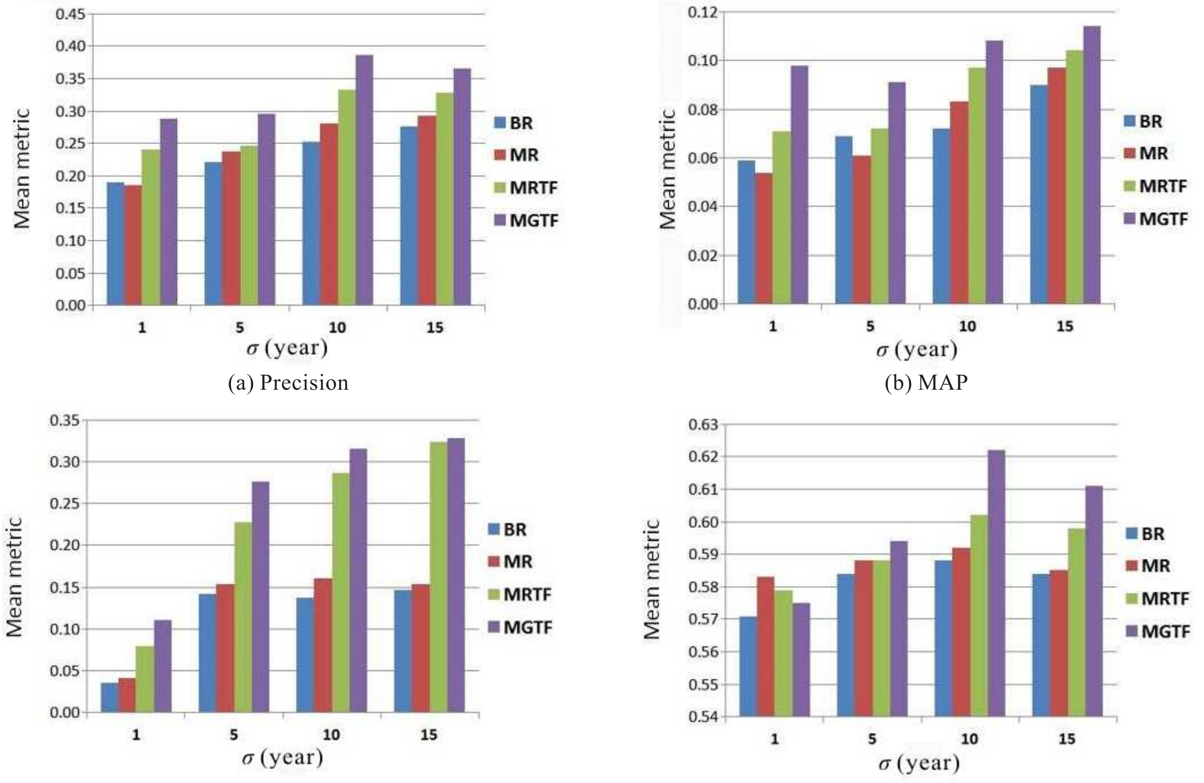
Fig. 4 U.S. patent maintenance illustration.

Table 3 Four configurations of our approach.

Name	Topic trend	Topic filtering	Time window
BR	Binning	No	Rectangular
MR	MMPP	No	Rectangular
MRTF	MMPP	Yes	Rectangular
MGTF	MMPP	Yes	Gaussian

Next, the current MGTF approach is compared against the performance of the three baseline models with large window sizes (σ of 10 and 15 years). For COA1, the term age and term support were calculated within the time window. For KeyPlayer, the 50 most similar patents were found within the time window.

Results in Tables 4 and 5 show that although baseline 1 (COA1) uses a more complex function, it performs worse than baseline 2 (COA2) on all the metrics, which agrees with previous results[21]. The unique patent vocabulary usage discussed in Section 3 can



(a) Precision

(b) MAP

(c) The Spearman coefficient

(d) AUC

Fig. 5 Results for the four configurations for the 21 domains. The details of the BR, MR, MRTF, and MGTF models are explained in Table 3.

Table 4 Results of the baseline models and the MGTF model (σ=10 years).

Method	Precision	MAP	Spearman	AUC
KeyPlayer	0.249	0.087	0.181	0.606
COA1	0.224	0.072	0.161	0.580
COA2	0.275	0.096	0.284	0.587
MGTF	**0.386**	**0.108**	**0.315**	**0.622**

Table 5 Results of the baseline models and the MGTF model (σ=15 years).

Method	Precision	MAP	Spearman	AUC
KeyPlayer	0.256	0.095	0.202	0.600
COA1	0.252	0.092	0.180	0.579
COA2	0.297	0.107	0.289	0.582
MGTF	**0.365**	**0.114**	**0.328**	**0.611**

create significant noise in the word statistics (e.g., TF-IDF); therefore quantifying the patent value by its term weights (COA1) will be more error-prone than by word counts (COA2). Baseline 2 (COA2) outperforms baseline 3 (KeyPlayer) on citation-based metrics (precision, MAP, and the Spearman coefficient), while KeyPlayer performs better on maintenance-based metric (AUC). This reflects the fact that patents with content similar to their prior art have less value in the business market and are more likely to be abandoned. The results also show that the topic-based temporal mining approach is statistically significantly better (p-value $<$ 0.005) than all three baselines on all the metrics. The core patents discovered by this method are then more likely to be cited by subsequent patents, and have longer maintenance lifespans than the non-core patents; thus, they are more valuable in the business market.

The third set of tests showed the advantages of this method over the baseline methods through human-written patent summaries in *Derwent Innovations Index*^{SM} (DII). Each patent in the DII database has a concise, high-quality summary (around 150 words) written by IP professionals which describes the novelty and advantages (impacts) of the invention. Due to the DII download restrictions, 150 patents were randomly selected from our dataset for comparison with the corresponding human-written summaries from the DII database. For each selected patent d, 50 signature words were extracted from each of the two most dominant topics in d (i.e., the topics with the highest probabilities $p(z|d)$, excluding noisy topics), to assemble a 100-word summary of the patent. 100 words were then extracted to the same patent that had the highest contributions calculated in baseline 1 (COA1). ROUGE-1[38] was used for each patent to compare the assembled summary to the human-written summary with stopwords removed. The current method achieved an average ROUGE-1 score of 0.223 and outperformed the baseline method (0.204). When writing a patent summary in DII, the domain experts choose the keywords from the technology background that most appropriately describe the invention rather than the terms used by the inventors. Compared to baseline 1 (COA1) which selects keywords in the subject patent based on term age and support, the current method captures the technical keywords that are closer to the expert choices for analyzing a patent's novelty and influence, which results in better core patent mining results.

5.5 Company competitor analysis

Core patents have significant value in many patent applications, such as company competitor analyses and key inventor identification. This section demonstrates the application value of core patent mining by revealing the company competitive relationships in the petroleum industry. The 3-level hierarchy of the petroleum industry defined by the American Petroleum Institute and shown in Fig. 6 was used, which covers 15 of the 21 domains in the dataset. The patent portfolio of each of the 108 companies was constructed by organizing the company's patents into the hierarchy.

The core patent mining results were used to discover a company's major competitors in its operating market. The scores of a company's patents in a domain were summed for each company and used to weight the domain in the company's patent portfolio. Then, a company's portfolio was compared to the other 107 company portfolios. If the weights of the same domain from two portfolios were comparable (i.e., one weight does not exceed the other by more than 100%), then the two companies are competitors in the domain. If two companies are competing in over half of the domains in an operating division, then the two companies are competing in the operating division. If two companies are competing in at least one operating division, then the two companies are *major competitors* to each other. A baseline method was also designed in which the weight of a domain was set to the company's patent count, instead of to the sum of the patent scores. Both methods were used to select the top three competitors for each company that have the highest number of competing domains with the company.

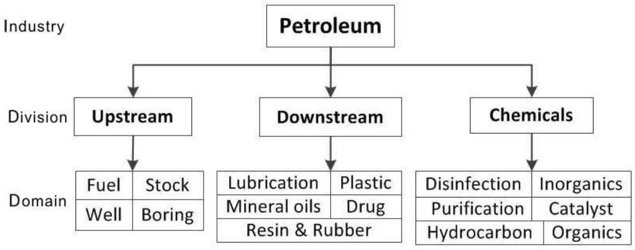


Fig. 6 3-level hierarchical structure in the petroleum industry. The upstream division (15 028 patents) is for *oilfield equipment & service*. The downstream division (22 440 patents) is for *major integrated oil & gas*. The chemicals division (20 663 patents) is for *chemical process & method*. The 3-level hierarchy covers 88.3% of the patents in the 21 domains.

The competitor information from Yahoo! Finance[39] was then used as the gold standard. For each of the 108 companies, experts selected three major competitors based on their business performance (such as market capacity and revenue) and also in which operating divisions they compete (e.g., ExxonMobil competes with Chevron on *major integrated oil & gas*). The method has an average competitor identification precision of 68.1% and outperforms the baseline method (55.4%) by 22.9%. The results show that core patent mining can identify a company's competitors and their technology strengths more accurately than simple patent counts.

6 Conclusions

Effective patent portfolio management can help a company capture potential licensing and investment opportunities and identify its competitors. This paper studies automatic core patent mining which is an important problem in patent portfolio management. The topic-based temporal mining approach more effectively discovered novel and influential patents in 21 domains in the petroleum industry than traditional word-based statistical methods. The patent mining techniques described here, if used wisely and in conjunction with manual patent analysis, can help companies find key information in patent portfolios.

Future work will extend this research from the petroleum industry to other industries with tests on larger patent datasets. The algorithm will be extended to jointly model topics and their activeness trends by incorporating dynamic topic models[40,41]. Future work will also explore more core patent mining applications, such as company competitor timeline analyses and domain-specific expert discovery. Finally, more reliable patent novelty and influence metrics will be developed to better evaluate the algorithm's performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61272227), the National Key Technology Research and Development Program (No. 20121857860), and the Tsinghua University Initiative Scientific Research Program (No. 20121088071). We thank the three anonymous reviewers for their valuable comments.

References

[1] D. Kappos, *Innovation Policy and the Economy*. Cambridge, MA, USA: National Bureau of Economic Research, 2010.

[2] Institute for Prospective Technological Studies, *The 2011 EU Industrial R&D Investment Scoreboard*. Brussels, Belgium: European Commission's Joint Research Centre, 2011.

[3] K. Edward, Patent mining in a changing world of technology and product development, *Intellectual Asset Management,* pp. 7-10, July/Aug. 2003.

[4] T. Keraan, *Extracting Maximum Value from Intellectual Assets*. New York City, USA: Deloitte & Touche, 2010.

[5] USPTO, Fiscal year 2011 performance and accountability report, http://www.uspto.gov/about/stratplan/ar/2011/index.jsp, 2011.

[6] Google Patents, http://www.google.com/patents, 2013.

[7] Delphion, http://www.delphion.com, 2013.

[8] IPVision, http://www.see-the-forest.com/G4/Main.act, 2013.

[9] Aureka, http://aureka.micropat.com, 2013.

[10] Y. Liu, P. Hseuh, R. Lawrence, S. Meliksetian, C. Perlich, and A. Veen, Latent graphical models for quantifying and predicting patent quality, in *Proc. of 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1145-1153.

[11] X. Jin, S. Spangler, Y. Chen, K. Cai, R. Ma, L. Zhang, X. Wu, and J. Han, Patent maintenance recommendation with patent information network model, in *Proc. of 11th IEEE International Conference on Data Mining*, 2011, pp. 280-289.

[12] B. Wang, M. Chu, and J. Z. Shyu, Patent value measurement by analytic hierarchy process, in *Proc. of 15th International Conference on Management of Technology*, 2006, pp. 1-12.

[13] R. J. Mann and M. Underweiser, A new look at patent quality: Relating patent prosecution to validity, *Journal of Empirical Legal Studies,* vol. 9, no. 1, pp. 1-32, 2012.

[14] Y. Guo and C. Gomes, Ranking structured documents: A large margin based approach for patent prior art search, in *Proc. of 21st International Joint Conference on Artificial Intelligence*, 2009, pp. 1058-1064.

[15] X. Xue and W. B. Croft, Transforming patents into prior-art queries, in *Proc. of 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, pp. 808-809.

[16] L. Azzopardi, W. Vanderbauwhede, and H. Joho, Search system requirements of patent analysts, in *Proc. of 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010, pp. 775-776.

[17] NTCIR, http://research.nii.ac.jp/ntcir/index-en.html, 2013.

[18] CLEF-IP, http://www.ir-facility.org/clef-ip, 2013.

[19] PaIR, http://www.ifs.tuwien.ac.at/pair2011/Site/PaIR11.html, 2013.

[20] TREC-CHEM, http://www.ir-facility.org/trec-chem, 2013.

[21] M. A. Hasan, W. S. Spangler, T. Griffin, and A. Alba, COA: Finding novel patents through text analysis, in *Proc. of 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 1175-1184.

[22] Y. Chen, S. Spangler, J. Kreulen, X. Wu, and L. Zhang, SIMPLE: A strategic information mining platform for licensing and execution, in *Proc. of 9th IEEE International Conference on Data Mining Workshops*, 2009, pp. 270-275.

[23] B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims, Identifying temporal patterns and key players in document collections, in *Proc. of 5th IEEE International Conference on Data Mining Workshops*, 2005, pp. 165-174.

[24] S. M. Gerrish and D. M. Blei, A language-based approach to measuring scholarly impact, in *Proc. of 27th International Conference on Machine Learning*, 2010, pp. 375-382.

[25] B. Shaparenko and T. Joachims, Information genealogy: Uncovering the flow of ideas in non-hyperlinked document databases, in *Proc. of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 619-628.

[26] S. Bao, R. Li, Y. Yu, and Y. Cao, Competitor mining with the web, *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 10, pp. 1297-1310, 2008.

[27] Z. Ma, G. Pant, and O. R. L. Sheng, Mining competitor relationships from online news: A network-based approach, *Electronic Commerce Research and Applications*, vol. 10, no. 4, pp. 418-427, 2011.

[28] Y. Li, L. Wang, and C. Hong, Extracting the significant-rare keywords for patent analysis, *Expert Systems with Applications*, vol. 36, no. 3, pp. 5200-5204, 2009.

[29] T. Chen, Patent claim construction: An appeal for Chevron deference, *Virginia Law Review*, vol. 94, no. 5, pp. 1165-1212, 2008.

[30] A. Kotov, C. Zhai, and R. Sproat, Mining named entities with temporally correlated bursts from multilingual web news streams, in *Proc. of 4th ACM International Conference on Web Search and Data Mining*, 2011, pp. 237-246.

[31] D. Newman, A. Asuncion, P. Smyth, and M. Welling, Distributed algorithms for topic models, *Journal of Machine Learning Research*, vol. 10, pp. 1801-1828, 2009.

[32] W. Fischer and K. Meier-Hellstern, The Markov-modulated Poisson process cookbook, *Performance Evaluation*, vol. 18, no. 2, pp. 149-171, 1993.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1-38, 1977.

[34] Large petroleum companies listed by Wikipedia, http://en.wikipedia.org/wiki/List_of_oil_exploration_and_production_companies and http://en.wikipedia.org/wiki/List_of_oilfield_service_companies, 2013.

[35] USPTO database, http://patft.uspto.gov, 2013.

[36] A. K. McCallum, MALLET: A machine learning for language toolkit, http://mallet.cs.umass.edu, 2002.

[37] The official USPTO gazettes, http://www.uspto.gov/news/og/index.jsp, 2013.

[38] C. Lin and E. Hovy, Automatic evaluation of summaries using N-gram co-occurrence statistics, in *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 71-78.

[39] Yahoo! Finance, http://finance.yahoo.com/q/co?s=MSFT, 2013.

[40] D. M. Blei and J. D. Lafferty, Dynamic topic models, in *Proc. of 23rd International Conference on Machine Learning*, 2006, pp. 113-120.

[41] X. Wang and A. McCallum, Topics over time: A non-Markov continuous-time model of topical trends, in *Proc. of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 424-433.

Po Hu is a currently PhD candidate in the Department of Computer Science and Technology, Tsinghua University. He received his BEng in computer science from Tsinghua University in 2009. Po has published and co-authored three research papers on ICDM, CIKM, and SIGKDD conferences. His research interests include textual temporal analysis, document retrieval, and topic modeling.

Minlie Huang is an associate professor in the Department of Computer Science and Technology, Tsinghua University. He received his BEng and PhD degree in computer science from Tsinghua University in 2000 and 2006, respectively. He has published and co-authored 15 research papers on ACL, COLING, IJCAI, ICDM, and AAAI conferences. His research interests include natural language processing, data mining, artificial intelligence, and machine learning.

Xiaoyan Zhu is currently a professor of Department of Computer Science and Technology, Tsinghua University. She received her BS degree from the University of Science and Technology of Beijing in 1982, and her PhD degree from the Nagoya Institute of Technology, Japan, in 1990. She is the editor for Journal of Chinese Information, and PC member for APBC 2007, CIKM 2008 and ECDM 2008. She has published and co-authored over 50 research papers on ACL, COLING, IJCAI, SIGKDD, ICDM, CIKM, and AAAI conferences, and KAIS, JCST and BMC Bioinformatics journals. Her research interests include pattern recognition, neural network, machine learning, natural language processing, and text mining.