Finding Nuggets in IP Portfolios: Core Patent Mining through Textual Temporal Analysis

Po Hu
State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University
hup09@mails.tsinghua.edu.cn

Minlie Huang
Department of Computer Science and Technology
Tsinghua University
Beijing 100084, China
aihuang@tsinghua.edu.cn

Peng Xu
ExxonMobil Research and Engineering Company
Annandale, New Jersey, U.S.A
peng.xu@exxonmobil.com

Weichang Li
ExxonMobil Research and Engineering Company
Annandale, New Jersey, U.S.A
weichang.li@exxonmobil.com

Adam K. Usadi
ExxonMobil Research and Engineering Company
Annandale, New Jersey, U.S.A
adam.k.usadi@exxonmobil.com

Xiaoyan Zhu
Department of Computer Science and Technology
Tsinghua University
zxy-dcs@tsinghua.edu.cn

ABSTRACT

Patents are critical for a company to protect its core technologies. Effective patent mining in massive patent databases can provide companies with valuable insights to develop strategies for IP management and marketing. In this paper, we study a novel patent mining problem of automatically discovering core patents (i.e., patents with high novelty and influence in a domain). We address the unique patent vocabulary usage problem, which is not considered in traditional word-based statistical methods, and propose a topic-based temporal mining approach to quantify a patent's novelty and influence. Comprehensive experimental results on real-world patent portfolios show the effectiveness of our method.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*

Keywords

Core patent mining, patent novelty, patent influence, textual temporal analysis

1. INTRODUCTION

Effective patent portfolio management requires the assessment of patent quality, identification of technology gaps, and analysis of competitors' patent activities. This current work focuses on identification of core patents (patents with high novelty and influence). Automated Core Patent Mining (CPM) can play an important role in patent portfolio management and R&D strategy development.

CPM improves the efficiency of patent portfolio management. For a company with large number of patents, it can be time consuming and costly to manually screen patents to identify ones with licensing potential and/or patents with low values that could be dropped. By automatically evaluating a patent's novelty and influence, CPM methods help IP analysts to focus on a smaller set of patents that require manual analysis. In this way, the overall time on patent analysis is reduced, resulting in more efficient portfolio management.

CPM can also be used to track competitors' activities. Generally speaking, patent work reveals research details before a product hits the marketplace. By continuously monitoring the change of core patents and their owners, a company can spot competitive threats, and adjusts its R&D strategy accordingly.

In this paper, we propose a topic-based temporal mining approach that quantifies the novelty and influence of a patent, and ranks all patents by combining the novelty score and influence score. The top-ranked patents are selected as core patents in the domain. Unlike traditional word-based statistical methods, our approach considers the unique vocabulary usage in patent literature, and can effectively discover core patents from massive patent databases.

The rest of the paper is organized as follows. We survey related work in Section 2. Section 3 formulates the task of core patent mining and illustrates the problem challenges.

Our methodology is described in Section 4. Section 5 analyzes the experimental results, and Section 6 concludes our work.

2. RELATED WORK

Our work is mainly related to patent quality assessment, and document novelty and influence analysis.

2.1 Patent Quality Assessment

Liu et al. [4] study the problem of predicting patent quality. They propose a measurement model in which patent quality is a latent variable. Patent quality can be estimated from correlated measurements, including court ruling decisions and some lexical features. Jin et al. [3] work on automatic patent maintenance recommendation. Their method extracts a set of patent lexical features. Based on the features, a classifier is trained on historical patent maintenance decisions, and predicts whether a patent should be maintained or not. Finally, they propose a network-based optimization process to refine the prediction results.

Our work is different from the above. Our method only utilizes patent text as input. Court ruling decisions used in [4] and patent maintenance records used in [3] are only available for a small number of patents. In addition, [4] and [3] are more focused on the writing quality of patents, while we assess patent quality from the technical perspective.

2.2 Document Novelty and Influence Analysis

Hasan et al. [2] analyze patent novelty from patent claims. Their method obtains a set of keywords from the claim section. The score of a keyword in a patent equals to the ratio of the keyword's support to its age (i.e., the time difference between the word's first appearance in patents and the issue year of the subject patent). A patent's novelty score is the sum of all its keywords' scores.

Shaparenko et al. [6] discover important documents in a document collection. The documents are clustered by their word bags. A document is important if it has fewer similar documents published before it, and has more similar documents published after it.

Unlike the above work, our approach is topic-based, and addresses the unique vocabulary usage in patent literature. The experimental results show that our approach is more effective than the methods in [2] and [6].

3. PROBLEM FORMULATION

Suppose the patent set in a domain is $\mathcal{D} = \{D_t | t = 1, ..., T\}$, where $D_t \subset \mathcal{D}$ contains the patents issued at year t. We define patent novelty and patent influence 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., $\mathcal{D}_{PA}(d) = \{D_i | 1 \leq i < t\}$.*

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., $\mathcal{D}_{FW}(d) = \{D_i | t < i \leq T\}$.*

In this paper, we use *latent topics* to represent the ideas in a patent, and quantify patent novelty and influence by topic activeness variations. Then we combine the novelty score and influence score to rank the patents in $\mathcal{D}$, and select the top-ranked patents as core patents in the domain.

As a scientific literature with legal significance and potential profits, patents have complex structures and special nomenclature. The sophisticated patent language can pose significant challenges to patent mining. We have noted that the semantic meanings of technical terms in patents are often inconsistent due to the following two 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.

2. **Heterogeneity in nomenclature.** Some technical terms, especially those chemical and biological entities, have more than one reference names that are semantically identical to each other. Choice of the aliases is mainly determined by the inventor's writing style.

A combination of these factors greatly raises the difficulty of capturing the real contributions of a patent. Classical word-based statistical methods, such as TF-IDF [2] and word similarity [6], are based on a homogeneous assumption. Those methods can cause serious word mismatch problem in 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, yet miss many *significant-rare* keywords which are low-frequency but highly informative.

To address the unique vocabulary usage in patent literature, we propose a novel topic-based temporal mining approach. We use topics to depict the ideas in patents, which can 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, which outperforms the traditional word statistics.

4. METHODOLOGY

Our method consists of three steps. Firstly, identify latent topics in the patents. Secondly, model topic activeness trend, and remove noisy topics. Thirdly, quantify patent novelty and influence, and rank patents by their scores.

4.1 Latent Topic Identification

Patents are semi-structured documents containing several sections. Each section has a specific purpose. For example, abstract gives an overview of the invention, background summarizes related works, and detailed description explains the invention in details. The claim section is the heart of a patent as claims define the protection scope of the invention. In this paper, we use *title*, *abstract*, *claims* and *detailed description* to analyze patent novelty and influence. All patents are transformed into lowercase with stopwords removed.

We utilize the distributed version of Latent Dirichlet Allocation model (AD-LDA) [5] to efficiently discover the latent topics inside the patents in a domain. Suppose there are K topics $\{z_1, ..., z_K\}$ in the patent set $\mathcal{D} = \{D_1, ..., D_T\}$. A topic z_k is a probabilistic distribution of words in the word set V of $\mathcal{D}$, i.e., $\{p(w|z_k)\}_{w \in V}$. A patent $d \in \mathcal{D}$ is a probabilistic distribution of topics, i.e., $\{p(z_k|d)\}_{k \in \{1,...,K\}}$. To better represent the semantic meaning of a topic, we also estimate the topic-bigram distributions. The probability of a bigram phrase $w_i w_j$ ($w_i \neq w_j$) generated by a topic z_k is

Table 1: 10 Basic Jittering Patterns for 3 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

estimated as follows.

$$p(w_iw_j|z_k) = p(w_i|z_k) * p(w_j|z_k) * p(w_iw_j) \qquad (1)$$

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

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

4.2 Topic Activeness Trend Modeling

For each discovered topic, we use a Markov-Modulated Poisson Process (MMPP) [1] to model its activeness trend. The time span of $\mathcal{D}$ is divided into T time intervals, and each interval lasts for a year. Suppose there are M levels of topic activeness, and each level is represented by a hidden state in ascending order $S_i \in \{1,...,M\}, i = 1,...,T$. MMPP's observation is the occurrence count of the topic's signatures in each interval, and the emission probabilities are set as Poisson distribution:

$$B[C_i][S_i] = \frac{{\lambda_i}^{C_i} e^{-\lambda_i}}{C_i!} \qquad (2)$$

where $B[C_i][S_i]$ is the emission probability for state S_i to generate C_i signatures of the topic in the *ith* interval; λ_i is the expectation of the topic's signature count in the *ith* interval. We equally divide the range of all observations $\{C_i|1 \leq i \leq T\}$ into M bins, and the initial value of λ_i is set as the mean value of the observations in the bin that C_i belongs to. The initial state probabilities and transition probabilities of the Markov chain, and also the rate parameters of the Poisson process are estimated via EM algorithm. Finally, we sort the value of $\{\lambda_i|1 \leq i \leq T\}$ in ascending order, and the topic activeness level in the *ith* interval equals to the rank of λ_i, i.e., $S_i = Rank(\lambda_i)$.

Topic models have a common problem: the number of topics must be determined in advance. It is inevitable that some topics have little semantic meanings (i.e., noisy topics). In this paper, we propose an automatic noisy topic filtering method. We observe that noisy topics can be characterized by their activeness variations, and we have summarized two classes of noisy topics. The first class is *trendless topics*, whose activeness trends have few variations along the timeline. The second class is *jittering topics*. A topic jitters when its activeness fluctuates capriciously in adjacent time intervals. Suppose there are 3 topic activeness levels, we can define 10 basic jittering patterns as shown in Table 1. More complex jittering patterns can be decomposed into several basic jittering patterns. If a topic jitters in too many time intervals, we consider it as a jittering topic. All the trendless topics and jittering topics are removed as noisy topics.

4.3 Patent Novelty and Influence Analysis

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

For each patent $d \in D_t$, our method uses its dominant topics $Z_{Dom}(d) = \{z|p(z|d) > 10\%\}$ to depict the ideas in d. To evaluate the novelty of d, we focus on the activeness trends of dominant topics in d's prior art $\{[S_{i,z}]|_{i=1}^{t-1}, z \in Z_{Dom}(d)\}$, where $S_{i,z}$ is the activeness level of topic z in the *ith* interval. Low topic activeness in the prior art indicates that d is very novel. To evaluate the influence of d, we focus on the topic trends in d's follow-up work $\{[S_{i,z}]|_{i=t+1}^{T}, z \in Z_{Dom}(d)\}$. High topic activeness in the follow-up work indicates that d is very influential.

However, temporal bias exists in patent novelty and influence analysis. On one hand, an old patent has fewer patents in its prior art, and more patents published after it. On the other hand, a new patent has fewer patents published after it, yet has more patents in its prior art. Thus, old patents tend to be over-estimated in their novelty and influence, while new patents are under-estimated. We have noticed that core patents are time-sensitive, and their values depend on certain aspects of the technology developments. Thus, it is more appropriate to measure a patent's novelty and influence within a certain period of the topic activeness trends. In this paper, we use *time decay factor* to restrict the scope of the topic trends, and eliminate the temporal bias in patent novelty and influence analysis.

We consider two typical window functions to determine the time decay factor, namely *Rectangular window* and *Gaussian window*. Suppose Δt is the time difference between two time points, and 2σ is the window size. Rectangular window and Gaussian window are defined in (3) and (4) respectively.

$$F(\Delta t) = \begin{cases} 1 & if \quad |\Delta t| \leq \sigma \\ 0 & otherwise \end{cases} \qquad (3)$$

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

For each topic z, the *normalized topic activeness level* $S_{Nor}(i,z)$ in the *ith* interval is defined as:

$$S_{Nor}(i,z) = F(\Delta t) * S_{i,z} \qquad (5)$$

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

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

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

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

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

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

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

The score of d is the product of its novelty score and influence score, $Score(d) = Novelty(d) * Influence(d)$. We rank all patents by their scores, and the top-ranked patents are selected as core patents in the domain.

5. EXPERIMENT

5.1 Dataset and Domain Definition

We construct our dataset with patents from the petroleum industry. Firstly, we select 108 large petroleum companies listed by Wikipedia. Then we download the patents assigned to the 108 companies from USPTO patent database. Our dataset contains 82,648 U.S. patents from 1976 to 2010.

Secondly, we define a set of domains 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, and we use this class to classify the patents. A USPC class is defined as a domain, if it has at least 800 patents in our dataset. The remaining classes with their patents are not used in our experiments, since a domain with too few patents is more suitable for manual analysis. As a result, we have defined 21 domains in the petroleum industry with 65,846 U.S. patents, accounting for 79.7% patents in our dataset.

5.2 Parameter Setting and Baseline Methods

The parameters in our method are set as follows. In the topic identification step, we empirically set the topic-patent density to be 200 patents per topic in each domain. We select 50 unigrams and 50 bigrams with the highest probabilities under a topic to be the topic's signatures. In the topic activeness modeling step, we set 3 topic activeness levels in the MMPP model to trade-off between performance and training complexity. To show the effectiveness of MMPP, we also use *equal-size binning* to model the topic activeness trend, in which the range of a topic's signature counts is equally divided into 3 bins, and the topic activeness level in an interval depends on the bin that its signature count belongs to. A topic whose activeness trend reaches *inactive* or *bursty* only once, or jitters in over 50% of all intervals, is removed as a noisy topic. We also test the effects of Rectangular window and Gaussian window for topic activeness normalization.

We have also implemented three word-based statistical methods to compare to our approach. Baseline 1 (COA1) is the algorithm used in [2]. In this method, stopwords are removed from the patent text. Those words whose document frequency exceeds 90% in the patent set are also removed as domain stopwords. For each word w in a patent d, the contribution of w is determined as follows:

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

where $age(w)$ is the time difference between the earliest year w occurs in the patent set and the issue year of d; $support(w)$ is the number of follow-up patents that contain w. The score of d is the sum of the contributions of the words in d.

Baseline 2 (COA2) takes the same procedures as in baseline 1 (COA1). The difference is that the score of a patent equals to its word count after removing domain stopwords.

Baseline 3 (KeyPlayer) is adapted from [6]. Each patent is represented as a TF-IDF vector of its words after removing 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. Finally, the score of d is calculated as follows:

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

where $Leader(d)$ and $Follower(d)$ is the number of the similar patents published before and after d respectively.

Table 2: Four Configurations of Our Approach

Name	Topic Trend	**T**opic **F**iltering	Time Window
BR	**B**inning	No	**R**ectangular
MR	**M**MPP	No	**R**ectangular
MRTF	**M**MPP	Yes	**R**ectangular
MGTF	**M**MPP	Yes	**G**aussian

5.3 Evaluation Metrics

In this section, we utilize two indicators to help assess the discovered core patents in a domain.

The first indicator is *patent forward citation*, which is the citation count a patent receives from its follow-up work. In this paper, we assume that novel and influential patents (i.e., core patents) are more likely to receive more citations than those non-core patents in a domain.

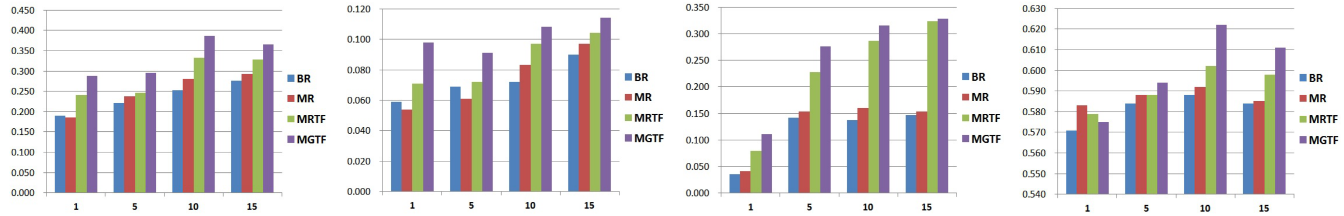
Since new patents are less cited than old ones, patent forward citation tends to under-estimate the importance of new patents. To eliminate this bias, we evaluate the discovered core patents within each year, instead of on the whole timeline. All patents published in the same year are ranked by their scores, and also by forward citations as the gold standard. Then the Spearman correlation coefficient of the two patent rankings is calculated to evaluate the algorithm. In addition, we also assign the 25% most cited patents as the real core patents in the domain, while the 25% highest scored patents as the discovered core patents. Precision and mean average precision (MAP) of the two patent sets are calculated to complement the ranking correlation coefficient.

The second indicator is *patent maintenance status*. In the U.S., a patent can be kept valid for up to 20 years, and maintenance fees must be 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 E1 stage to E3 stage, assignees tend to abandon their worthless patents as early as possible (e.g., at E1 stage), and only those truly important patents are maintained at E3 stage to complete their full terms.

We have obtained the patent abandonment information between 1995 and 2011 from USPTO. To construct the gold standard, we regard those patents expired at E1 stage as *non-core patents*, and those patents maintained throughout the 20 years as *core patents*. In our dataset, there are 9527 core patents and 6078 non-core patents in the 21 domains. For each domain, the patents are ranked by their scores, and the top 20%, 40%, 60% and 80% patents are set as core patents respectively to construct 4 cut-off levels. False positive rate (FPR) and true positive rate (TPR) are calculated at each cut-off level to draw the algorithm's ROC (Receiver Operating Characteristic) curve, and AUC (Area Under the ROC Curve) is used to evaluate the algorithm's performance.

5.4 Experimental Results

We first compare the effects of different configurations of our approach, as discussed in Section 5.2. Table 2 shows the



(a) Precision (b) MAP (c) Spearman coefficient (d) AUC

Figure 1: Experimental results of four configurations in our approach. The X label is σ.

Table 3: Experimental Results of Baselines (σ=10)

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

details of the configurations, including topic activeness modeling (BR *vs.* MR), noisy topics filtering (MR *vs.* MRTF), and time decay factor (MRTF *vs.* MGTF). We also test the algorithm's performance by varying the window size parameter σ to be 1, 5, 10 and 15 years respectively. Figure 1 shows the average results on the 21 domains in our dataset.

From Figure 1, we can draw the following conclusions. Firstly, for all four methods, large window size can bring better performance than small window size. Small window only considers the most recent part of the topic trends, and loses a lot of information of technology developments in a domain. Thus, large window size is desirable to acquire sufficient information to estimate a patent's novelty and influence. Besides, we also notice that too large window size (σ as 15) may deteriorate the performance by some metrics (e.g., precision and AUC in Figure 1). Since some patents published long before or after the subject patent may not influence or be influenced by that patent, too large window size may introduce unnecessary topic activeness variations that harm the algorithm's performance. Secondly, MR outperforms BR on all metrics for large window size. This is because MMPP can estimate the topic activeness trend more accurately through global model fitting, and has better smoothing effect than equal-size binning. Thirdly, MRTF outperforms MR on all metrics, showing that topic filtering can effectively remove noisy topics and improves the algorithm's performance. Finally, Gaussian window (MGTF) outperforms Rectangular window (MRTF) on all metrics. To eliminate the temporal bias in core patent mining, Rectangular window simply drops the non-recent topic information away, while Gaussian window maintains more useful information through the smoothing factor. Thus, Gaussian window can better reflect the time-sensitive characteristic of technology developments.

Next, we compare the performance of the baselines against our approach (MGTF) with large window size (σ as 10). Table 3 shows the experimental results, and we can draw the following conclusions. Firstly, although baseline 1 (COA1) uses more complex function, it performs worse than baseline 2 (COA2) on all metrics, which agrees with the experimental results in [2]. The unique patent vocabulary usage discussed in Section 3 can bring significant noises in word statistics, thus quantifying the patent value by term weights (COA1) would be more error-prone than word counts (COA2). Secondly, baseline 2 (COA2) outperforms baseline 3 (KeyPlayer) on citation-based metrics (precision, MAP and Spearman coefficient), while baseline 3 (KeyPlayer) performs better on maintenance-based metric (AUC). This reflects the fact that those patents similar to their prior art in content have less values in the business market, thus are more likely to be abandoned. Finally, MGTF outperforms all baselines on all metrics. The core patents discovered by our method are more likely to be cited by their follow-up work, and have longer maintenance lifespan than those non-core patents.

6. CONCLUSION

In this paper, we study automatic core patent mining, which is an important problem in patent portfolio management. Compared to traditional word-based statistical methods, our topic-based temporal mining approach can effectively discover novel and influential patents, thus can help companies develop better IP strategy and improve competitiveness.

7. ACKNOWLEDGMENTS

This work is supported by ExxonMobil Research and Engineering Company. We would like to thank the three anonymous reviewers for their valuable comments.

8. REFERENCES

[1] W. Fischer and K. Meier-Hellstern. The markov-modulated poisson process cookbook. *Performance Evaluation*, 18(2):149–171, 1993.

[2] M. A. Hasan, W. S. Spangler, T. Griffin, and A. Alba. Coa: Finding novel patents through text analysis. In *ACM SIGKDD Conference Proceedings*, 2009.

[3] 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 *IEEE ICDM Conference Proceedings*, 2011.

[4] Y. Liu, P. yun Hseuh, R. Lawrence, S. Meliksetian, C. Perlich, and A. Veen. Latent graphical models for quantifying and predicting patent quality. In *ACM SIGKDD Conference Proceedings*, 2011.

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

[6] B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims. Identifying temporal patterns and key players in document collections. In *IEEE ICDM Workshop Proceedings*, 2005.