Minimizing Legal Exposure of High-Tech Companies through Collaborative Filtering Methods

Bo Jin*
Dalian University of Technology
jinbo@dlut.edu.cn

Chao Che*
Dalian University
chechao@gmail.com

Kuifei Yu
Zhigu Tech
yukf@zhigu.com

Yue Qu
Dalian University of Technology
firestorm@mail.dlut.edu.cn

Li Guo
Dalian University of Technology
guoli@dlpu.edu.cn

Cuili Yao
Dalian University of Technology
yaocuili1984@dlut.edu.cn

ABSTRACT

Patent litigation not only covers legal and technical issues, it is also a key consideration for managers of high-technology (high-tech) companies when making strategic decisions. Patent litigation influences the market value of high-tech companies. However, this raises unique challenges. To this end, in this paper, we develop a novel recommendation framework to solve the problem of litigation risk prediction. We will introduce a specific type of patent-related litigation, that is, Section 337 investigations, which prohibit all acts of unfair competition, or any unfair trade practices, when exporting products to the United States. To build this recommendation framework, we collect and exploit a large amount of published information related to almost all Section 337 investigation cases. This study has two aims: (1) to predict the litigation risk in a specific industry category for high-tech companies and (2) to predict the litigation risk from competitors for high-tech companies. These aims can be achieved by mining historical investigation cases and related patents. Specifically, we propose two methods to meet the needs of both aims: a proximal slope one predictor and a time-aware predictor. Several factors are considered in the proposed methods, including the litigation risk if a company wants to enter a new market and the risk that a potential competitor would file a lawsuit against the new entrant. Comparative experiments using real-world data demonstrate that the proposed methods outperform several baselines with a significant margin.

CCS Concepts

•Social and professional topics → Patents; •Applied computing → Law; •Information systems → *Data mining; Recommender systems;*

*Corresponding Author

KDD '16, August 13–17, 2016, San Francisco, CA, USA.

 ISBN 978-1-4503-4232-2/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939708

Keywords

Patent Litigation; Section 337 Investigations; Collaborative Filtering; Proximal Slope One Predictor; Time-aware Predictor

1. INTRODUCTION

When an inventor, business, or other entity owns a patent and that patent is infringed, the patent owner has few alternatives other than to undertake patent litigation. Few entities wish to undertake patent litigation because it is usually lengthy and expensive. For example, Apple and Google spend more money on patent litigation and defensive patent acquisitions than they do on research and development. However, some firms try to *extort* money from others using patent litigation. These firms are known as patent trolls. Such firms collect patents solely for the purpose of litigation to extort money from small companies that cannot afford patent defense. In 2013, the president of the United States, Barack Obama introduced a proposal to crack down on patent trolls. His administration also pledged to protect innovators from frivolous litigation. A country's government rarely talks explicitly about detailed aspects of technology and intellectual property policy. Thus, it is clear that patent litigation is a huge challenge for high-technology (high-tech) companies in the United States.

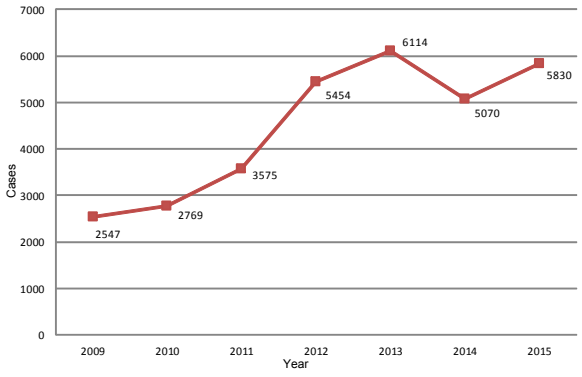


Figure 1: U.S. patent litigation cases filed, 2009–2015.

Patent litigation numbers have increased rapidly in recent years, as shown in Figure 1. This paper focuses on a specific type of litigation: Section 337 investigations. Section 337 investigations refer to investigations conducted according to Section 337 of the Tariff Act of 1930 and its related amendments by the United States International Trade Commission (USITC). Section 337 investigations prohibit all acts of unfair competition, or any unfair trade practices, when exporting products to the United States. If convicted under a Section 337 investigation, the exporter's products may be permanently excluded from the US market, which can result in considerable economic losses. Most Section 337 investigation cases concern patents. As with intellectual property rights, Section 337 investigations are frequently used for blocking purposes. Both parties in a Section 337 investigation are competitors; hence such investigations can be a means of blocking competitors by preventing their market expansion. Therefore, predicting the probability of its involvement in a Section 337 investigation is crucial for the development of a high-tech company.

Section 337 investigations consume precious resources, and the outcome can signify success or failure to a company. Hence, it is critical for high-tech companies to understand their legal exposure in their industry category and from their competitors to be successful in such patent litigations. Despite the increasing number of arguments and lawsuits, there are no reliable unbiased data mining methods on the patterns and trends of patent litigation in the United States[8] This paper aims to enable high-tech companies to develop data-driven intellectual property business strategies to succeed in their challenging environments. For example, the Chinese smartphone giant Xiaomi is being targeted by a U.S.-based patent troll. The troll company (Blue Spike LLS) filed a lawsuit against Xiaomi in a U.S. district court at the end of 2015. Thus, as Forbes reported, Xiaomi may have a major patent problem if it wants to enter the U.S. market. Xiaomi operates as an investment holding company and uses its subsidiaries for the development and transfer of its technologies and intellectual properties.

This paper has two aims: 1) to predict the litigation risk in a specific industry category for high-tech companies, and 2) to predict the litigation risk from competitors for the high-tech companies. To achieve these goals and to identify the risky industry categories and competitors for high-tech companies, we must consider several factors. We must consider the litigation risk if a company wants to enter a new market, and further consider the risk that a potential competitor would file a lawsuit against the new entrant. The risks must be formally estimated. These issues can be answered by mining historical investigation cases and related patents. Specifically, we develop two matrixes: 1) the company–category matrix predicts the litigation risk in a specific industry category for a high-tech company and 2) the complainant–respondent matrix predicts the litigation risk from the competitors for a high-tech company. We propose two methods to meet the needs of both aims: a proximal slope one predictor and a time-aware predictor. The proposed methods can work on different conditions, that is, companies previously engaged in lawsuits and companies never engaged in a lawsuit. Finally, we carry out extensive experiments on a real-world data set collected from the USITC. The experimental results clearly validate the effectiveness of the proposed methods. With the results of this paper, high-tech companies could adjust their strategies for development and explore space for market.

Overview. The paper is organized as follows. Section 2 describes the data used in this paper and introduces the trends of patent litigation in the United States. Section 3 formulates the problem of patent litigation prediction for high-tech companies, and introduces some preliminary matters in prediction. Section 4 and Section 5 describe the proposed methods, the proximal slope one predictor and time-aware predictor, respectively. Section 6 discusses the results of our study. Section 7 outlines some related works. Finally, Section 8 provides a conclusion and some suggestions for future research.

2. DATA DESCRIPTION AND ANALYSIS

2.1 Data Description

This paper uses several data sources. The first comes from USITC Section 337 investigation cases[1]. We analyze Section 337 case data (a sample is shown in Figure 2) for the period of 1982–2014. We find that 857 of the 906 cases in that period concern patent infringements, and 623 of these cases list the related patents. There are 1,654 patents, including design and utility patents. Several patents are used in multiple investigations.

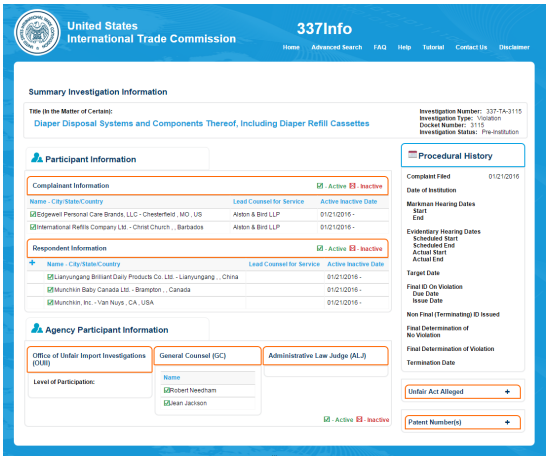


Figure 2: Section 337 investigation: sample HTML document.

We also use patent grants data from USPTO Bulk Downloads[2]. This public dataset is provided by the United States Patent and Trademark Office (USPTO). Each patent publication contains a group of structured and unstructured information. The USPTO updates the patent application publication each week. We first download the patent data package and retrieved the weekly data package files. We then import the patent data (in XML format) to our own local database. Finally, we retrieve the information from this local database using the ID of 1,654 patents as queries.

The final data source used is the industry category for the companies. It is difficult to identify the industry category of a company in relation to its patents. Each of the Section 337 investigation cases of patent infringement is associated

[1]https://www.usitc.gov/
[2]http://www.google.com/googlebooks/uspto-patents.html

with one or more patents, and each U.S. patent has at least one class number. In this study, we determine the industry category by the patents that a company holds. However, US patents have 1,088 classifications in total. Therefore, we reclassify the US patent data according to 56 industry categories with a mapping table.

2.2 Data Analysis

We analyze the Section 337 investigation cases by regions and the companies involved.

Figure 3 shows the top 10 Section 337 investigation regions from 1982–2014. The figure shows that the allegations relating to patent infringements vary across regions over time. In the 1980s and early 1990s, Japan was the most intensively investigated country, accounting for 43.15% of the total number of cases. Mainland China and Chinese Taiwan were also frequently investigated in the period 1982–2004. The Chinese mainland is currently the most frequent respondent in Section 337 investigations. Since 2004, Chinese products and companies have been involved in over 40% of Section 337 investigations, as the export volume of a region to the United States influences the number of patent litigations. When the trade volume of a country or region increases substantially with rapid economic development, the number of Section 337 investigations increases correspondingly. Japan, Chinese Taiwan, and Mainland China are all in the same situation.

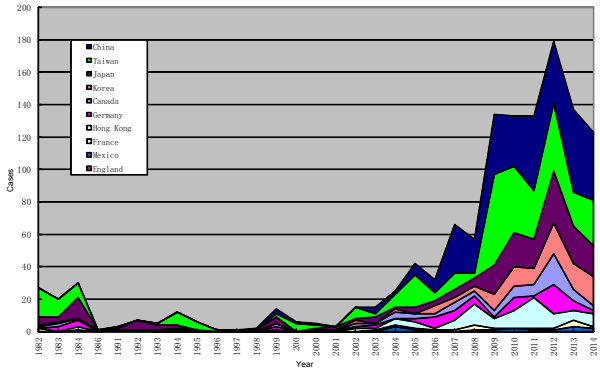


Figure 3: Top 10 respondent regions in Section 337 investigations.

Figure 4 shows the top 10 Section 337 investigation industry categories for 1982–2014. The figure shows that most of the cases are in material equipment domains such as building material processing equipment and assembly and material handling equipment, which are covered by a large number of patents. Electrical communication domains, including image and sound equipment and telecommunication, are also responsible for many Section 337 investigations.

Figure 5 shows the top 10 Section 337 investigation complainants for 1982–2014. We find that most of the complainants are large high-tech companies that hold core technologies and monopolistic advantages in their related domains. Such companies are motivated to retain their dominant position and to exclude competitors from the market.

Figure 6 shows the top 10 Section 337 investigation respondents for 1982–2014. We find that Section 337 investigations are associated with companies that experience explosive growth. The top four respondents are all smartphone companies. The telecommunication sector has undergone

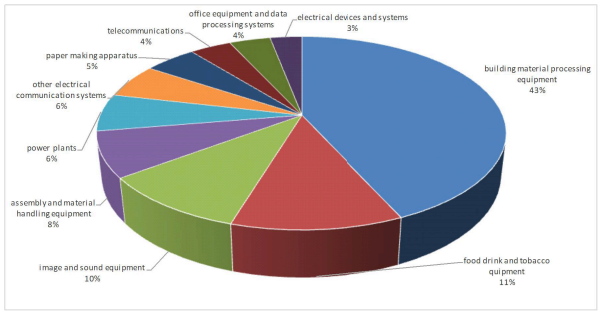


Figure 4: Top 10 industry categories in Section 337 investigations.

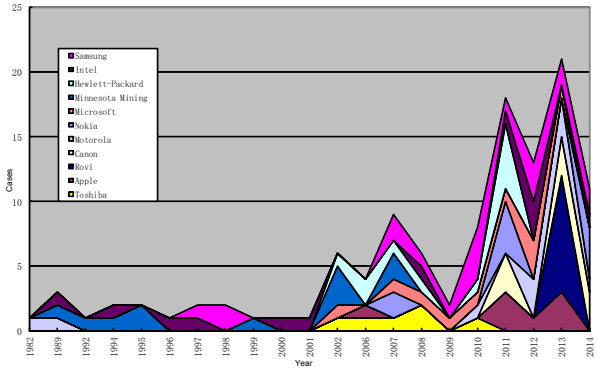


Figure 5: Top 10 complainants in Section 337 investigations.

the most investigations since 2009 because the rapid growth of mobile Internet has resulted in intense competition in the smartphone market. Hence, there has been an increase in lawsuits in this market.

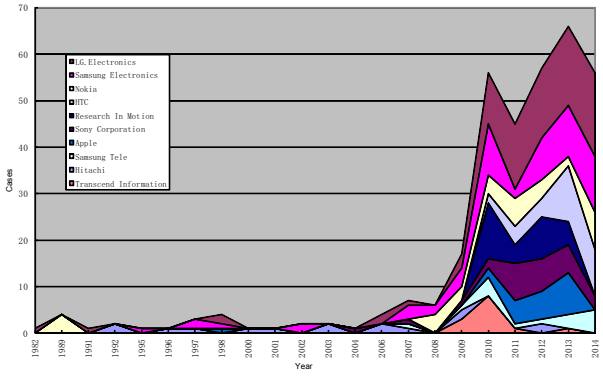


Figure 6: Top 10 respondents in Section 337 investigations.

3. PROBLEM FORMULATION

In this section, we introduce some preliminary matters, and then formally define the problem of minimizing the legal exposure of high-tech companies.

3.1 Preliminary Matters

A relatively small number of companies are involved in Section 337 investigations; however, the number of patents

is quite large. To reduce the number of calculations and enhance the precision, we analyze the data by patent category rather than by patent itself. Thus, individual patents are replaced by their corresponding patent classifications. As Section 337 investigations only protect the patents of United States, thus, we use the United States Patent Classification (USPC) system. The USPC system has 1,088 patent classes, making it difficult to identify the patent classification for a company. Thus, we further amalgamate the USPC classifications into industry categories. There are some works on patent classification mapping [13], we employ a common standards to map the USPC classifications into 56 industry categories.

To construct the company–category matrix, we need to first construct the company–classification matrix. The number of times that a company has undergone a Section 337 investigation determines its value in the matrix. After the normalization process, the company–classification value ranges between zero and five. Formula 1 is used to normalize the company-classification value as follows:

$$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}}\beta \quad (1)$$

where x_i indicates the actual number of patents in a patent classification related to Section 337 investigation cases that the company is involved in, x_i^* indicates the normalized value, x_{min} and x_{max} indicate the minimum and maximum number of patents related with Section 337 investigation cases, respectively. β is set to 5 for normalizing the value from 0 to 5.

The company–classification matrix is then transferred into the company–category matrix with a patent classification–industry category mapping table. The proposed methods are used to calculate the prediction values in this matrix.

A sample predicted results of the company–category matrix is shown in Table 1. Let us consider, for example, the 0.326 rating at the intersection between the row of Company ID 51 and column of Category 10. This rating indicates that there is a 0.326 probability that Adidas America Inc. (Company ID 51) may be investigated under the bleaching and dyeing industry category (Category 10).

In the same way we develop the complainant–respondent matrix. The total number of Section 337 investigation cases is not large enough, such that the complainant–respondent matrix is quite sparse. Take cases in 2014 as an example, for company pairs, there are only 564 non-zero values of the 2 million values in the matrix. For the full duration data set used in this paper, there are 3,370 non-zero values for company pairs in the matrix.

3.2 Recommendation Framework

Here, we convert the legal exposure prediction problem to a recommendation problem. We first try to recommend the riskiest industry category and the riskiest competitor for a high-tech company. We then formally define the problem of legal exposure recommendation with time-awareness.

DEFINITION 1. *We have company c, and a set of industry categories $A = a$, each of which contains a set of investigation probabilities p_{ai}. The goal of the legal exposure recommendation is to build an optimal ranked list of industry categories.*

DEFINITION 2. *We have company c, and a set of competitors $B = b$, each of which contains a set of investigation probabilities p_{bi}. The goal of the legal exposure recommendation is to build an optimal ranked list of competitors.*

The above problem statements raise two issues:

- How to mine data relating to the litigation risk of an industry category and how to produce a ranked list $\Lambda^{RiskA} = a|a \in c$ according to their risk scores $RiskA(a)$.
- How to mine data relating to the litigation risk of a competitor and how to produce a ranked list $\Lambda^{RiskB} = b|b \in c$ according to their risk scores $RiskB(b)$.

Although it is desirable to minimize legal exposure for high-tech companies, it is not an easy task to effectively discover and evaluate the litigation risks of both industry category and competitors. Patent litigation is a complex legal action and there are many factors to consider in a recommendation. Thus, it is difficult to answer how to efficiently manage litigation risk for high-tech companies. To this end, in this paper, we propose a novel recommendation framework to solve this problem.

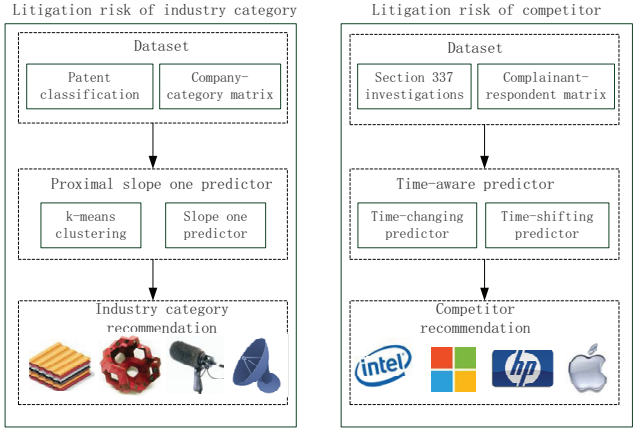


Figure 7: The recommendation framework.

Figure 7 shows the proposed recommendation framework, which consists of the two aims. *Litigation risk of industry category* predicts the legal exposure in a specific industry category for high-tech companies. *Litigation risk of competitor* predicts the litigation risk from competitors for high-tech companies.

4. PROXIMAL SLOPE ONE PREDICTOR

This section proposes a proximal slope one predictor as the recommendation algorithm to predict the litigation risk for high-tech companies in a specific industry category. In Section 337 investigations, the litigation procedures of various companies in an industry category are quite different. Slope One algorithm[9] is suitable to settle this problem. However, Slope One algorithm only improve the problem of item rating deviation without considering the difference between the items. Hence, the quality and performance of the recommender system are unsatisfactory for our problem. Because the investigations of companies other than the target company in different industry categories will affect the prediction accuracy, we therefore improved the Slope One

Table 1: Company–category matrix

Company ID	Category 1		Category 10		Category 23		Category 45		Category 56
1	0	...	0	...	0	...	0	...	0
...	···	...	···	...	···	...	···	...	···
51	···	...	0.326	...	0	...	0	...	···
...	···	...	···	...	···	...	···	...	···
411	0	...	0	...	0	...	0.217	...	0
...	···	...	···	...	···	...	···	...	···
1912	0	...	0	...	0	...	0.109	...	0
...	···	...	···	...	···	...	···	...	···
2338	0	...	0	...	0.109	...	0.217	...	0
...	···	...	···	...	···	...	···	...	···
2805	0	...	0	...	0	...	0	...	0

algorithm using a clustering method. In this paper, the k-means algorithm and the Slope One algorithm are combined to predict the litigation risk. The prediction value of the k-nearest neighbor vectors of the target company should be calculated using a sub-matrix of the company–classification matrix. This not only decreases the amount of calculations, but also relieves the problems caused by data sparseness. Specifically, we first establish patent classification vectors of the companies according to the historical patents related to Section 337 investigation cases. The k-means algorithm is employed to cluster the vectors into K clusterings. Then the k-nearest neighbor companies of the target company are screened and are used to construct the sub-matrix of the company–classification matrix. The prediction value in the sub-matrix is calculated using the Slope One algorithm. Finally, we transfer the company–classification matrix to the company–category matrix. We can predict the litigation risk of the target company based on the company–category matrix.

Suppose the patent classification vector of a company can be denoted as $X = \{x_i, i = 1, 2, \ldots, n\}$, where x_i is a patent classification and n is the total number of U.S. patent classifications. The m vectors X are then clustered into K clusterings as $C = \{c_k, k = 1, 2, \ldots, K\}$, where each c_k denotes a vector group of similar companies. Suppose μ_k is the clustering center of c_k, the distance between each patent classification vector of a company and the clustering center can be calculated as:

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_{ki}\|^2 \tag{2}$$

where μ_{ki} is the value of the i-th patent classification of μ_k.

For each company, we can obtain k-nearest neighbor patent classification vectors with the result of clusterings, and construct a sub-matrix A_{ij}. The Slope One algorithm is employed to predict the values R_{ij} in the sub-matrix. Basically, a Slope One algorithm is used as follows. Suppose x and y indicate the value of two patent classifications by company u. The two values conform to the linear relation $y = x + b$, where b is the parameter, and we obtain the estimated value $\hat{b}$ of parameter b. After substituting the known value x, we obtain the estimated rating $\hat{y}$ by company u. Formula 3 is used to calculate the value $dev_{j,i}$ between the target item $Item_j$ and item $Item_i$. Formula 4 is then used to obtain the prediction value $P(u)_j$:

$$dev_{j,i} = \sum_{u \in S_{j,i}(x)} \frac{u_j - u_i}{card(S_{j,i}(x))} \tag{3}$$

$$P(u)_j = \frac{\sum_{i \in R_j} (dev_{j,i} + u_i)}{card(R_j)} \tag{4}$$

where $S_{j,i}(x)$ indicates the set of companies who have corated items $Item_j$ and $Item_i$ and R_j indicates the set of items that have been rated by user u. The number of elements in a set S is $card(S)$.

As shown in Algorithm 1, in the proposed recommendation algorithm, using the company–classification matrix, $R_{m \times n}$ indicates the historical data of Section 337 investigations. Here m indicates the number of companies, n indicates the number of classifications, and R_{ij} indicates the normalized number of Section 337 investigation cases that company i filed according to the patent classification j.

Algorithm 1 Proximal Slope One Predictor

1: Construct the company–classification matrix A_{ij} with the k-means algorithm;
2: According to the classification i that has been rated by the target company u, select the nearest top k classifications with the classification i, and then produce the set K';
3: Use formula (3) to calculate the value $dev_{j,i}$ with other classifications rating for the target company u in a set K';
4: Use formula (4) to calculate the prediction rating $P(u)_j$, and then obtain the company–classification matrix $R'_{m \times n}$;
5: According to the company–classification matrix $R'_{m \times n}$, transfer into the company–category matrix;
6: Recommend the top r prediction categories for the target company.

5. TIME-AWARE PREDICTORS

This section proposes time-aware predictors as recommendation algorithms to predict the litigation risk for high-tech companies from their competitors. The proposed algorithm is based on a matrix factorization model that has been well studied in recent years. It is a popular model in the area of recommendation because it combines good scalability with predictive accuracy. Additionally, this model is very

flexible when modeling various temporal real-life situations. We improve the basic matrix factorization model [7] with a time-shifting baseline predictor. We combine the time-changing baseline predictor and the time-shifting baseline predictor to meet the needs of our problem. There are two different conditions, which are 1) a company pair that has no historical litigations, and 2) a company pair with historical litigations. The first condition is settled with the time-changing baseline predictor, and the second is settled with the time-shifting baseline predictor.

5.1 Static Baseline Predictor

As shown in Figures 5 and 6, the number of Section 337 investigation cases is time-dependent. Before we deal with temporal effects, we would like to establish the foundations of a static baseline predictor. A pure matrix factorization model could serve well in capturing the interaction between the users and items in a recommendation system. However, most of the observed values of litigation cases are due to effects independent of the interaction associated with either complainants or respondents. For example, typical Section 337 investigation data exhibit large complainant and respondent biases. That is, some companies initiate more investigations than other companies, and some companies face more investigations than others. We will encapsulate those effects, which do not involve complainant-respondent interactions, within the baseline predictors [7]. These baseline predictors tend to capture much of the observed values, in particular much of the temporal dynamics within the data. Hence, it is vital to model them accurately, which enables the better identification of the part of the values that truly represents complainant–respondent interactions and should be subject to factorization.

Now we discuss how to construct a static baseline predictor. Suppose each complainant u is associated with vector p_u and each respondent i is associated with vector q_i. The overall average number of investigation cases is denoted by μ. A baseline predictor for an unknown value r_{ui} is denoted by the observed deviations b_u and b_i of complainant u and respondent i, respectively, from the average as follows:

$$b_i = \frac{\sum_{u \in R(i)}(r_{ui} - u)}{\lambda_1 + |R(i)|} \tag{5}$$

$$b_u = \frac{\sum_{u \in R(u)}(r_{ui} - u - b_i)}{\lambda_2 + |R(u)|} \tag{6}$$

where λ_1 and λ_2 are regularization parameters which make the values of b_i and b_u shrink towards zero. Here $\lambda_1 = 10$, $\lambda_2 = 25$.

The static baseline predictor should then be integrated back into a factor model. The value is predicted by the rule:

$$\hat{r}_{ui} = \mu + b_i + b_u + q_i^T p_u \tag{7}$$

To determine the vectors p_u and q_i, we minimize the regularized squared error:

$$\min_{q^*,p^*,b^*} \sum_{(u,i)\in K} (r_{ui} - \mu - b_i - b_u - q_i^T p_u)^2 + \lambda_3(\|q_i\|^2 + \|p_u\|^2 + b_i^2 + b_u^2) \tag{8}$$

where the constant λ_3 controls the extent of regularization, as usually determined by cross validation. In this paper, minimization is performed by typical stochastic gradient descent. The parameters are updated as follows:

$$b_i = b_i + \gamma(\mathrm{e}_{ui} - \lambda_3 \cdot b_i) \tag{9}$$

$$b_u = b_u + \gamma(\mathrm{e}_{ui} - \lambda_3 \cdot b_u) \tag{10}$$

$$q_i = q_i + \gamma(\mathrm{e}_{ui} \cdot p_u - \lambda_4 \cdot p_i) \tag{11}$$

$$p_u = p_u + \gamma(\mathrm{e}_{ui} \cdot q_i - \lambda_4 \cdot p_u) \tag{12}$$

where e_{ui} is the prediction error as follows:

$$\mathrm{e}_{ui} = r_{ui} - (\mu + b_i + b_u + q_i^T \cdot p_u) \tag{13}$$

5.2 Time-changing Baseline Predictor

In this section, we discuss the following temporal effects which include in the baseline predictors: (1) complainant biases b_u change over time; (2) respondent biases b_i change over time; (3) complainant preferences p_u change over time; and (4) respondent preferences q_i change over time. Hence, in our models we would like to take the parameters as functions of time. For our problem of litigation risk prediction, it's no need to get the finest resolution. Thus, an adequate decision would be to split the biases into time-based bins. We employ a distinct bias for each time period related with a single bin. Such decision should strike a balance for the desire to achieve a finer resolution with enough ratings per bin. Considering the features of Section 337 litigation cases, we select 3 years as the bin size. We define the bias functions $b_i(t)$ of the respondent i, changing by year t, as follows:

$$b_i(t) = b_i + b_{i \cdot Bin(t)} \tag{14}$$

where $b_{i \cdot Bin(t)}$ is the bias of respondent i in those bins that contain year t.

In the same way, we define the bias functions $b_u(t)$ of the complainant u, changing by year t:

$$b_u(t) = b_u + \alpha \cdot dev_u(t) \tag{15}$$

where $dev_u(t)$ is the associated time deviation. If u files a lawsuit in the year t, $dev_u(t)$ is defined as:

$$dev_u(t) = sign(t - t_u) \cdot |t - t_u|^\beta \tag{16}$$

where $|t - t_u|$ measures the time distance between t and t_u. Furthermore, t_u is the average duration of Section 337 investigation cases of complainant u, which can be calculated as:

$$t_u = t_u(\lceil \frac{n_l - 1}{2} \rceil) \tag{17}$$

where n_l denotes the most recent Section 337 investigation case of complainant u.

Moreover, we note that the complainants' preferences change with time, leading to changes in the interactions between complainant and respondent. For example, if one company files a lawsuit against another company several times, the respondent may strike back. Suppose the feature vector of a company has k dimensions, the changing function of preferences is defined as:

$$p_{uk}(t) = p_{uk} + p_{uk \cdot Bin(t)} \tag{18}$$

where p_{uk} captures the stationary portion of the factor and $P_{uk \cdot Bin(t)}$ approximates a possible portion that changes in those bins containing year t.

Thus, the prediction rule is as follows:

$$\hat{r}_{ui} = \mu + b_i(t) + b_u(t) + q_i^T p_u(t) \tag{19}$$

where t indicates the year of the last lawsuit.

To determine the involved parameters $b_{i\cdot Bin(t)}$, α_u and α_{uk}, we should solve:

$$\begin{aligned} min \sum_{(u,i)\in K} & (r_{ui} - \mu - b_i(t) - b_u(t) - q_i^T p_u(t))^2 + \lambda_5(\|q_i\|^2 \\ & + \|p_u\|^2 + b_i^2 + b_{i\cdot Bin(t)}^2 + b_u^2 + \alpha_u^2 + p_{uk\cdot Bin(t)}^2) \end{aligned} \tag{20}$$

where the constant λ_5 controls the extent of regularization, as usually determined by cross validation. In this paper, minimization is performed by typical stochastic gradient descent. The parameters are updated as follows:

$$b_{i\cdot Bin(t)} = b_{i\cdot Bin(t)} + \gamma(\mathrm{e}_{ui} - \lambda_5 \cdot b_{i\cdot Bin(t)}) \tag{21}$$

$$\alpha_u = \alpha_u + \gamma(\mathrm{e}_{ui} \cdot dev_u(t) - \lambda_5 \cdot \alpha_u) \tag{22}$$

$$p_{uk\cdot Bin(t)} = p_{uk\cdot Bin(t)} + \gamma(\mathrm{e}_{ui} \cdot p_i - \lambda_5 \cdot p_{uk\cdot Bin(t)}) \tag{23}$$

5.3 Time-shifting Baseline Predictor

The time-changing baseline predictor only considers the bias of the complainant and respondent in the current time period. Practical application shows that litigation should be predicted based on the bias in the prior time period. Thus, the time-shift characteristics of bias are studied in this paper: $Shift(t)$ denotes the bias in a period of time, $b_{i\cdot Shift(t)}$ denotes the time-shift bias of the complainant, $b_{u\cdot Shift(t)}$ denotes the time-shift bias of the respondent, and $p_{u\cdot Shift(t)}$ denotes the time-shift bias of the complainant's suppressed feature vector. We then extend the bias functions denoted in the prior section as follows:

$$b_i(t) = b_i + b_{i\cdot Shift(t)} \tag{24}$$

$$b_u(t) = b_u + b_{u\cdot Shift(t)} \tag{25}$$

$$p_u(t) = p_u + p_{u\cdot Shift(t)} \tag{26}$$

Then the prediction rule is as follows:

$$\hat{r}_{ui} = \mu + b_i(t) + b_u(t) + q_i^T p_u(t) \tag{27}$$

We then split the data set into two groups: data of lawsuits before time t and data of lawsuits at time t. For the first data group, we employ a basic matrix algorithm to optimize the parameters b_u, b_i, q_i and p_u. For the second data group, to determine the involved parameters $p_{u\cdot Shift(t)}$, $b_{u\cdot Shift(t)}$, and $b_{i\cdot Shift(t)}$, we need to solve:

$$\begin{aligned} min \sum_{(u,i)\in K} & (r_{ui} - \mu - b_i(t) - b_u(t) - q_i^T p_u(t))^2 + \\ & \lambda_6(\|p_{u\cdot Shift(t)}\|^2 + b_{u\cdot Shift(t)}^2 + b_{i\cdot Shift(t)}^2) \end{aligned} \tag{28}$$

where the constant λ_6 controls the extent of regularization, as usually determined by cross validation. In this paper, minimization is performed by stochastic gradient descent. The parameters are updated as follows:

$$b_{i\cdot Shift(t)} = b_{i\cdot Shift(t)} + \gamma(\mathrm{e}_{ui} - \lambda_6 \cdot b_{i\cdot Shift(t)}) \tag{29}$$

$$b_{u\cdot Shift(t)} = b_{u\cdot Shift(t)} + \gamma(\mathrm{e}_{ui} - \lambda_6 \cdot b_{u\cdot Shift(t)}) \tag{30}$$

$$p_{u\cdot Shift(t)} = p_{u\cdot Shift(t)} + \gamma(\mathrm{e}_{ui} \cdot p_i - \lambda_7 \cdot p_{u\cdot Shift(t)}) \tag{31}$$

Furthermore, e_{ui} is the prediction error:

$$\mathrm{e}_{ui} = r_{ui} - \hat{r_{ui}} \tag{32}$$

where r_{ui} is the number of lawsuit filed by u against i in year t, $\hat{r}_{ui}$ is the predicted value based on the data before year t. As the exact value and predicted value are obtained from different durations, time-shifting trends are considered in our method.

As shown in Algorithm 2, in the proposed recommendation algorithm, the time-changing predictor trains the bias during a large range of time, and does not distinguish the conditions. Thus, it shows high accuracy for those companies without prior lawsuits. The time-shifting predictor trains the bias in the duration before year t. This indicates the shifting trend from the duration before year t and after year t. It provides high accuracy for those companies with prior lawsuits.

Algorithm 2 Time-aware Predictor

1: Constructs complainant–respondent matrix B_{ij}, and indicates the conditions of a company pair that has no historical litigation and a company pair that has historical litigations;
2: For the conditions of a company pair that has no historical litigation, use Formula 19 to calculate the predicted value $\hat{r}_{ui}$ and update the parameters using Formula 21, 22 and 23;
3: Repeat step (2) until the maximum number of iteration is reached or the changes in the RSME are less than the threshold;
4: Split the data set into D_t' and D_t according to year t, and optimize parameters b_u, b_i, q_i and p_u in data set D_t' before year t;
5: For the conditions of a company pair that has historical litigations, use Formula 27 to calculate the predicted value $\hat{r}_u i$ and update the parameters using Formulas 29, 30 and 31;
6: Repeat step (5) until the maximum number of iteration is reached or the changes in the RSME are less than the threshold;
7: Recommend the top r prediction competitors for the target company using the time-changing predictor and the time-shifting predictor for different conditions.

6. EXPERIMENTAL RESULTS

6.1 Evaluation Metrics

We use root-mean-square error (RMSE) to measure the prediction performance of the proposed algorithms. The smaller the RMSE values, the better the prediction accuracy of the algorithms. The RMSE formula are defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(p_i - q_i)^2}{N}} \tag{33}$$

where p_i is the prediction rating of the test set; q_i is the real rating of the test set; and N is the number of ratings of the test set.

6.2 Evaluation of Predictor for Industry Category

In the experiments, we employ k-fold cross-validation. The data is divided into k roughly equal-sized parts ($k = 5$ here). Of the k parts, a single part is retained as the validation data for testing the predictor, and the remaining $k - 1$ parts are used as training data. We use 1691 cases in the test set (out of 8,455 total cases in the data set). We then make recommendations for 2,806 companies, with an average of 2.41 observations per company for the entire training set.

We adopt three state-of-the-art baselines to evaluate the performances of our proximal slope one predictor. The first baseline is a traditional collaborative filtering algorithm, that is, an item-based collaborative filtering algorithm (IBCF). The second baseline is matrix factorization algorithm (MF), which is one of the most popular approaches to collaborative filtering. Moreover, we also use a popular slope one algorithm, that is, a bi-polar slope one algorithm, as the third baseline. All results reported below come from the same test set to facilitate comparison.

6.2.1 *Recommendation Performances*

We set up the evaluation as follows. First, we implemented our proximal slope one predictor and other baselines on the test data set. Specifically, by performing the proximal slope one predictor, we rate the company–category matrix. We then obtain the RMSE value by comparing the results with the original rating. We set the clustering parameters k as 20, and it remained invariant during the experiment.

Figures 8 show the results of each approach. In the figure we can see that our approach consistently outperforms other baselines and the improvement is significant. These results clearly validate the effectiveness of our proximal slope one predictor.

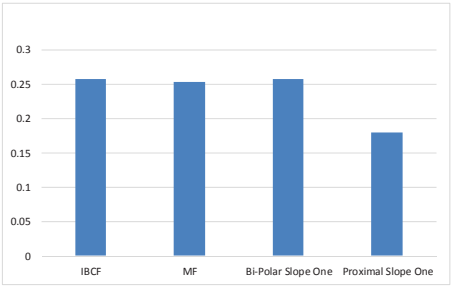


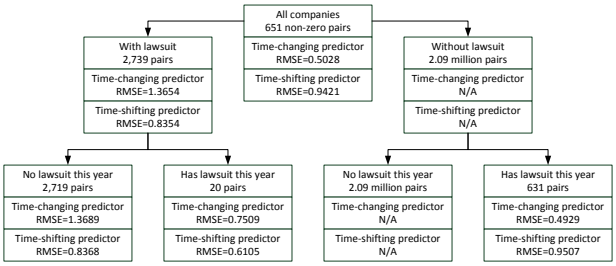
Figure 8: RMSEs of different approaches.

6.2.2 *Case Study*

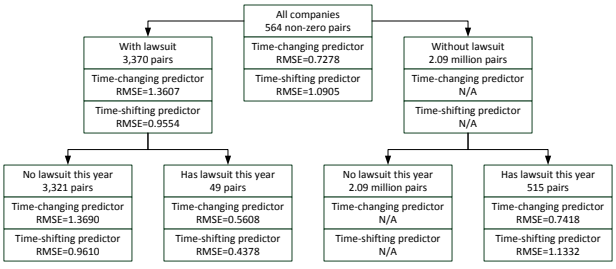
To further evaluate the performances of the different approaches, we randomly select 10 cases including 8 companies for case study. The results of the case study are shown in Table 2. Two companies (Company ID 200, 587) fall within two categories. The closer the difference between the predicted value and the rating, the better the prediction accuracy of the algorithm. According to the results, the prediction value of Proximal Slope One algorithm is closer to the original rating than that of the other three approaches. The extreme sparseness of the company-category matrix results that the parameters of MF could not be fully trained and optimized. Our approach is superior to matrix factorization algorithm due to no need of parameter training and optimizing. The IBCF and the bi-polar slope one algorithms use all of the data from the training data set as their prediction foundation, without clustering. The IBCF simply analyzes the similarity according to the investigation categories, and does not consider the degree of association between the categories. This means that the result is flawed. The bi-polar slope one algorithm analyzes the rating data that have been divided into two groups, according to the preference of the company regarding category. The disadvantage of this algorithm is that two totally different results will be produced when two ratings are close to the average value, but one is bigger than the average and the other is smaller. In the company–category matrix, there are a lot of data around the average: this leads to large errors when using the bi-polar slope one algorithm. Our approach only considers the data set that belongs to the same category as the target type. This minimizes the influence of unrelated categories, giving the highest prediction accuracy. However, in our approach, because of data sparseness, the nearest neighbor number of some types cannot achieve the number in the experimental setup. Thus, predicted partial results are not sufficiently accurate.

6.3 Evaluation of Predictor for Competitors

In these experiments, we divide the data set into training data and test data according to date. There are two experiments in this section: the first experiment used data from 2000 to 2012 as training data and data from 2013 as test data; the second experiment used data from 2000 to 2013 as training data and data from 2014 as test data. Figure 9 compares the results of the two proposed predictors. We can see that the results for the two years are similar.



(a) Test data set of 2013



(b) Test data set of 2014

Figure 9: The recommendation performances of different approaches.

Specifically, we set regularization parameters as $\lambda_3 = 0.005$, $\lambda_4 = 0.015$, $\lambda_5 = 0.005$, $\lambda_6 = 0.01$, and $\lambda_7 = 0.03$. The learning rate γ is set to 0.007 for static baseline predictor and time-changing baseline predictor, and 0.01 for time-shifting

Table 2: Prediction results for case study.

Company ID	Class	Category	Rating	IBCF	MF	Bi-polar Slope One	Proximal Slope One
71	258	20	0.109	0.277	0.084	0.249	0.109
200	452	56	0.217	0.168	0.061	0.249	0.217
200	602	53	0.217	0.228	0.085	0.249	0.217
336	11	17	0.109	0.113	0.011	0.249	0.109
587	229	18	0.109	0.129	0.039	0.249	0.109
587	471	17	0.109	0.132	0.027	0.249	0.109
846	448	48	0.109	0.191	0.076	0.203	0.117
1584	564	11	0.326	0.285	0.262	0.217	0.040
1874	526	9	0.217	0.243	0.048	0.249	0.217
2548	624	14	0.109	0.046	0.085	0.109	0.099

baseline predictor. For the setting of the number of iterations, if the number is smaller, the fitting degree of training is not high enough. However, if the number of iterations is larger, there might be overfitting. Thus, based on the results of the test, we set the number of iterations as 40.

As shown in Figure 9, the time-changing predictor performs better than the time-shifting predictor in the conditions of overall company pairs and company pairs with no historical lawsuits. In fact, it's extremely difficult to predict the legal exposure for a company without historical lawsuit data. It's even harder to predict the data in 2014 than in 2013, since there are much more new companies involved in Section 337 investigations in 2014. The rate of new companies (20) out of all companies (651) in 2013 is 0.0307, while the rate of new companies (49) out of all companies (564) in 2014 is 0.0869. On the other hand, the time-shifting predictor performed better in other conditions. This indicates that if a company pair has never been engaged in a lawsuit, and has a lawsuit in the test year, we can predict using the time-changing predictor. If a company pair has been engaged in a lawsuit previously, we can predict using the time-shifting predictor. It should be noted that the proposed approaches do have some limitations. If the company pair has been involved in a lawsuit previously, and has no lawsuit in the test year, it is difficult to predict using either of our approaches.

7. RELATED WORKS

Patent litigation covers both legal and technical issues, and is also a key concern of managers when making strategic decisions. Furthermore, patent litigation influences the market value of a company. Lerner's research[10] shows that patent litigation can decrease the market value of the respondent companies by 2%–3.1%. Such lawsuits are usually used to prevent imitation by other companies. The patentee prevents the infringement use of technology via litigation. The technology barriers constructed by litigation can transform into market entry barriers. Such barriers guarantee that the patentee retains its dominant position in the market. Lanjouw and Schankerman[8] argued there is a relationship between the characteristics of patents and the likelihood that they will become involved in infringement lawsuits. Chien[3] shows that whether a patent is going to be litigated depends on the economic value of the patent, the characteristics of the patent owner, and the owner's propensity to litigate. By studying all of the infringement claims for a sample of recently expired patents, Love[12] found considerable differences in litigation practices between practicing and non-practicing entities. Product-producing companies usually enforce their patents soon after issuance, and complete their enforcement activities well before their patent rights expire.

Existing research on patent recommendation have largely focused on method/approach, system/software, and patent derivatives (e.g., collaborator, maintenance). Most of the patent recommendation methods/approaches have been based on retrieval techniques and ranking methods. Bashir[1] proposed a query-expansion method using pseudo relevance feedback to increase the retrievability of patents. Cao[2] proposed a new user-friendly patent search paradigm and introduced three techniques (error correction, topic-based query suggestion, and query expansion) to improve the usability of patent searches. Oh[16] proposed a CV-PCR framework with a heterogeneous patent citation-bibliographic network that combines both patent citations (reflecting value relation) and bibliographic information (reflecting similarity relation). Based on this network, they proposed value-driven and context-guided features. Magdy[15] reported a comparison between simple and straightforward information retrieval techniques[14] and much more sophisticated techniques[11] in patent searches. Their experiments showed that the advanced search technique is statistically better only when no initial citations are provided.

Jie Tang[17] presented a topic-driven patent analysis and mining system (PatentMiner), which mainly focuses on mining heterogeneous patent networks. They proposed a dynamic probabilistic model to characterize the topical evolution of different objects in a heterogeneous network. The system uses patent summarization and technology trend visualization. Hasan[4]built a patent ranking software (Claim Originality Analysis) that rates a patent based on its value by measuring its age and the impact of the important phrases that appear in the "claims" section of the patent. Bo[5] designed an indicator to understand technology prospecting by studying the evolving distributions of technologies in high tech companies.

The research mentioned above largely focused on patent content. Other relevant studies have also been conducted such as partner recommendation and maintenance decisions. Wu[18] studied the problem of recommending patenting partners in enterprise social networks and proposed a ranking factor graph model to suggest co-invention relationships. Jin[6] proposed a systematic solution to analyze patents to recommend patent maintenance decisions. In their approach, the patents are modeled as a heterogeneous time-evolving information network and new patent features are

proposed to build models for a ranked prediction on whether to maintain or abandon a patent.

8. CONCLUSIONS

In this paper, we developed a recommendation framework for patent litigation risk prediction. The proposed framework clearly provides risky industry categories and risky competitors for high-tech companies when entering the US market. The design, analysis, and deployment of our recommendation framework are introduced in this paper. Specifically, we first acquired related information of Section 337 cases from USITC, and integrated it into a data set. We also presented the results of our statistical analysis, at both a regional and company level. Two methods of collaborative filtering, that is, proximal slope one predictor and time-aware predictor, are proposed in this paper. Finally, we evaluated our methods with extensive experiments on the real-world data set. The experimental results clearly validate the effectiveness of our methods.

Potentially, this study has many future research directions. First, it would be interesting to investigate new recommendation models to improve the accuracy of litigation risk prediction. Second, we plan to build a professional recommendation system to predict litigation risk for high-tech companies.

9. ACKNOWLEDGMENTS

The work was supported by the Natural Science Foundation of China (No. 61402068, 61425002) and Fundamental Research Funds for the Central Universities (No. DUT15QY04). It was also partially supported by a research grant from Zhigu Tech.

10. ADDITIONAL AUTHORS

Additional authors: Ruiyun Yu (Northeastern University, email: `yury@mail.neu.edu.cn`) and Qiang Zhang (Dalian University, email: `zhangq26@126.com`).

11. REFERENCES

[1] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *Advances in Information Retrieval*, pages 457–470. Springer, 2010.

[2] Y. Cao, J. Fan, and G. Li. A user-friendly patent search paradigm. *Knowledge and Data Engineering, IEEE Transactions on*, 25(6):1439–1443, 2013.

[3] C. V. Chien. Patent assertion and startup innovation. *New America Foundation, Open Technology Institute White Paper*, 2013.

[4] M. A. Hasan, W. S. Spangler, T. Griffin, and A. Alba. Coa: Finding novel patents through text analysis. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1175–1184. ACM, 2009.

[5] B. Jin, Y. Ge, H. Zhu, L. Guo, H. Xiong, and C. Zhang. Technology prospecting for high tech companies through patent mining. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 220–229. IEEE, 2014.

[6] 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 *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 280–289. IEEE, 2011.

[7] Y. Koren. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 447–456, New York, NY, USA, 2009. ACM.

[8] J. O. Lanjouw and M. Schankerman. Characteristics of patent litigation: a window on competition. *RAND journal of economics*, pages 129–151, 2001.

[9] D. Lemire and A. Maclachlan. Slope one predictors for online rating-based collaborative filtering. In *SDM*, volume 5, pages 1–5. SIAM, 2005.

[10] J. Lerner and A. Seru. The use and misuse of patent data: Issues for corporate finance and beyond. *Booth/Harvard Business School Working Paper*, 2015.

[11] P. Lopez, L. Romary, et al. Experiments with citation mining and key-term extraction for prior art search. In *CLEF 2010-Conference on Multilingual and Multimodal Information Access Evaluation*, 2010.

[12] B. J. Love. An empirical study of patent litigation timing: Could a patent term reduction decimate trolls without harming innovators? *University of Pennsylvania Law Review*, 161:1309, 2013.

[13] T. J. Lybbert and N. J. Zolas. Getting patents and economic data to speak to each other: An ąőalgorithmic links with probabilitiesąŕ approach for joint analyses of patenting and economic activity. *Research Policy*, 43(3):530–542, 2014.

[14] W. Magdy and G. J. Jones. Applying the kiss principle for the clef-ip 2010 prior art candidate patent search task. 2010.

[15] W. Magdy, P. Lopez, and G. J. Jones. Simple vs. sophisticated approaches for patent prior-art search. In *Advances in Information Retrieval*, pages 725–728. Springer, 2011.

[16] S. Oh, Z. Lei, W.-C. Lee, P. Mitra, and J. Yen. Cv-pcr: a context-guided value-driven framework for patent citation recommendation. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2291–2296. ACM, 2013.

[17] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, et al. Patentminer: topic-driven patent analysis and mining. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1366–1374. ACM, 2012.

[18] S. Wu, J. Sun, and J. Tang. Patent partner recommendation in enterprise social networks. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 43–52. ACM, 2013.