Electronic Preprint for Journal of Information Processing Vol.20 No.3

Regular Paper

Modeling Patent Quality: A System for Large-scale Patentability Analysis using Text Mining

Shohei Hido[1,†1,a)] Shoko Suzuki[1] Risa Nishiyama[1] Takashi Imamichi[1]
Rikiya Takahashi[1] Tetsuya Nasukawa[1] Tsuyoshi Idé[1,b)] Yusuke Kanehira[2]
Rinju Yohda[2] Takeshi Ueno[2] Akira Tajima[1,†2] Toshiya Watanabe[3]

Received: September 16, 2011, Accepted: February 3, 2012

Abstract: Current patent systems face a serious problem of declining quality of patents as the larger number of applications make it difficult for patent officers to spend enough time for evaluating each application. For building a better patent system, it is necessary to define a public consensus on the quality of patent applications in a quantitative way. In this article, we tackle the problem of assessing the quality of patent applications based on machine learning and text mining techniques. For each patent application, our tool automatically computes a score called patentability, which indicates how likely it is that the application will be approved by the patent office. We employ a new statistical prediction model to estimate examination results (approval or rejection) based on a large data set including 0.3 million patent applications. The model computes the patentability score based on a set of feature variables including the text contents of the specification documents. Experimental results showed that our model outperforms a conventional method which uses only the structural properties of the documents. Since users can access the estimated result through a Web-browser-based GUI, this system allows both patent examiners and applicants to quickly detect weak applications and to find their specific flaws.

Keywords: patent quality index, patentability, document classification

1. Introduction

Automatic classification of text documents has been one of the biggest challenges in natural language processing for decades [2], [15], [21], [22]. Distinguishing good and bad documents is relevant for various types of real-world situations such as finding useful Web pages or reviewing research papers. If computers were able to look into the documents at the semantic level, that would support or at least assist humans in judging such documents. As a first step, in this article we address the problem of evaluating the quality of patent documents, a typical task for patent examiners, by using text mining and machine learning techniques.

Patent applications which are examined and approved in patent systems have a key role in the industry of each country. Industries cannot grow or thrive without patenting their important inventions and legally preventing them from being unfairly used by competitors. A good patent system properly protects the rights of inventors, prohibits infringements, and promotes fair competition. On the other hand, if the patent office grants a flawed patent, it can hinder future progress and business development in that technology field. Therefore, maintaining the fairness and quality of the granted patents is one of the primary responsibilities of the patent office. However, patent systems in many countries are facing two major and related problems: substantial examination delays and the declining quality of the granted patents. Since the number of filed patent applications is steadily increasing, it is sometimes difficult for the officers to allocate sufficient time to properly and fully evaluate each of the applications. According to the 2011 annual report [5] by WIPO (World Intellectual Property Organization), the total number of patent applications in the world was almost doubled in 15 years from 1995 to 2010 as shown in **Fig. 1**. **Figure 2** shows the changes in the total numbers of the patent application in the top 3 countries, U.S, China, and Japan. Though the Japanese patent applications were recently decreased due to its economical depression, those of the U.S. and China were increased more than twice and 17 times, respectively in the last 10 years. Average tendency time (the difference between application and grant date) also tends to increase in many countries. In fact, the average tendency time increased by over 100% from 2000 to 2009 in the U.S. which has the largest number of submitted patent applications.

In response to the crises of the current patent systems, most of the patent offices have taken parallel actions to accelerate the examination process and to improve the quality of patents [13], [16]. For those purposes, inventors, attorneys, and examiners need to reach a consensus on the quality of patents. A widely-used criterion to measure the quality is the legal validity of the granted

1 Analytics & Intelligence, IBM Research - Tokyo, Yamato, Kanagawa 242–8502, Japan
2 IP Law Department, IBM Japan, Chuo, Tokyo 103–8510, Japan
3 Research Center for Advanced Science and Technology, The University of Tokyo, Meguro, Tokyo 153–8904, Japan
†1 Presently with Preferred Infrastructure, Inc.
†2 Presently with Yahoo! Japan
a) hido.jp@gmail.com
b) goodidea@jp.ibm.com

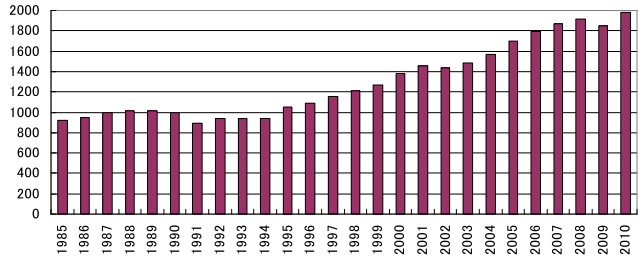

Fig. 1 Changes in the total numbers of the patent application in the world from 1985 to 2010. The unit for the vertical axis is 1,000 applications. There is a long increasing trend from 1995 to 2010.

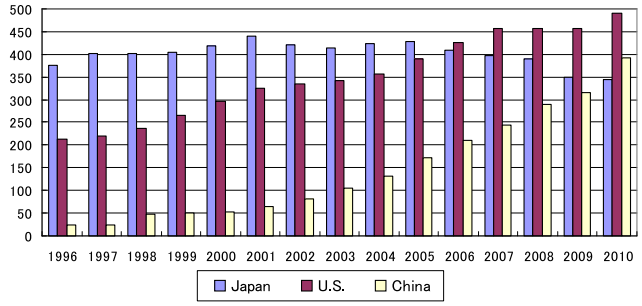

Fig. 2 Changes in the total numbers of the patent application in the top 3 countries, U.S, China, and Japan. The unit for the vertical axis is 1,000 applications. The increasing speed in China is very high.

patents, i.e., the likelihood that a patent will be upheld as valid if an invalidation trial is held. If a patent was improperly approved in the original examination process in spite of its legal invalidity, it will be invalidated in the trial. Though the validity can be estimated by manually studying the actual invalidation decisions, this is slow and expensive. In addition, while the case studies might reveal some qualitative characteristics of the valid patents, they do not provide any objective metrics for the quality of patents. Such a quantitative criterion is sometimes called a *Patent Quality Index*. Researchers and practitioners have both paid attention to them for many years [8], [9]. Some previous studies tackled this problem of defining a metric for validity with predictive modeling [10], [14], but the results lack generality since their experiments used at most a thousand cases of invalidation decisions, a small number compared to the total number of patent applications.

In this article, we introduce an alternative metric of patent quality named *patentability*, which represents the likelihood that an application will be approved by the patent office. The main contributions of this work are summarized as follows:

- The patentability score is formalized based on the specification documents of patent applications and their examination results.
- New sets of features called *word age* and *syntactic complexity* are introduced for evaluating the text of patent applications.
- Prediction models to compute the patentability score are obtained by using a supervised classification method.
- The prediction effectiveness of the models is shown by an evaluation using more than 0.3 million Japanese applications.
- A new GUI-based patentability analysis system allows users to interactively visualize and analyze the patentability of patent applications.
- A group of intellectual property experts at Japanese technology companies obtained analysis results by comparing some sets of applications on the analysis system.

These contributions cannot be made without having patent attorneys and intellectual property experts in the members. To the best of our knowledge, this is the first work that addresses the assessment of the patentability of patent applications using predictive modeling techniques. Evaluating the quality of patent application in advance of their examinations can be helpful for both examiners and applicants to make the patent process smoother, faster, and more reliable.

2. Document Classification and Related Work

In this section, we briefly introduce the prior art on classification of text documents and the related studies on the quantitative evaluation of the quality of patent documents.

Automatic classification or categorization of text documents has been one of the biggest challenges in natural language processing since the 1960s [2]. Since documents are increasingly stored in digital forms, document classification covers many real-world applications as summarized in a survey [21], such as automatic indexing, text filtering, and document organization. The improvements in this area mainly come from applying new textual features which effectively represent the contexts of text documents, such as TF-IDF (term frequency - inverse document frequency) [19] and n-gram [22]. Following this line of research, we cast the problem of evaluating the patent quality into a binary classification problem on patent documents.

The general difficulties in assessing the quality of text documents come from the lack of supervised label information about the quality and the vague format of the contents. There have been several projects aiming at estimating the quality of Web documents to guide Web users away from incorrect information. Alexander and Tate pointed out that manual lists of good Web pages can cover only a small fraction of the documents on the ever-growing Web [1]. Rieh compared the relationships between various types of characteristic factors of the Web sites and questionnaire results about their information quality and perceived authority [18]. In particular, medical researchers have paid close attention to the quality of medical information on the Internet, since users' misunderstanding based on incorrect information on the Web can potentially cause fatal medical accidents. In a survey, Eysenbach et al. summarized several small empirical studies that manually evaluate medical webpages and reported that some operational quality criteria are needed since the study results and conclusions widely vary due to the differences in the methodologies and data sets [6]. This means that qualitative evaluations are of limited effectiveness for assessing the quality of large numbers of documents. Since the design of webpages and the format of text content vary, it is hard to define a unified criterion to assess their quality. Since no one person can read and examine the contextual quality of thousands of websites, there must not be enough Web documents which are labeled based on a consistent criterion. In contrast, the specification documents of patent applications have a standard format. All of the patent applications will

have examination results given by the patent examiners that we can regard as class labels for the patent, as an application should be a good one if it is approved. Therefore, the patentability prediction from the specification documents is a valuable example of large-scale classification problem of text documents.

For assessing the quality of patents, there are also many studies on the legal validity of the granted patents. Sampat et al. examined the relationships of the number of citations and their types in the U.S. patents since the completeness of the references to prior art is one of the most important criteria for valid patents [20]. Nagata et al. introduced a predictive modeling approach which builds a logistic regression model to predict the invalidation trial result based on the characteristic features of the patents. The feature set for patent applications includes various types of statistical characteristics of specification documents such as the number of characters, the number of citations, and the number of independent claims. This method was extended by Kashima et al. by introducing other features including TF-IDF and n-grams, and the experiments showed that these textual features work well for predicting the validity [10]. However, since there are only a limited number of actual trials that were used for training the validity models, they lack generality. Though our method also follows their approach, our objective is different since we predict the patentability of each patent application to estimate how likely it is to be approved, rather than the validity of the granted patents. In addition, we introduce new types of features, the syntactic complexity and word age.

There were also some attempts to use text mining techniques to help examiners and applicants in analyzing large numbers of patent documents. Markellos et al. developed a system to apply clustering algorithms to a patent database and examine the detailed characteristics of the patents in each cluster [12]. Tseng et al. pointed out that text mining techniques are very useful in helping the domain experts in many tasks related to patent analysis, such as creating a patent map or prior art search [23]. In the following paper, Tseng and Wu also discussed how the patent engineers can efficiently perform patentability searches, a type of patent search that aims to determine whether an invention meets the requirements for being granted as a patent [24].

Some studies also have evaluated the quality of scientific papers. Yogatama et al. recently addressed the problem of predicting the number of citations of scientific articles [26]. Similarly to our approach, they also used textual features including n-grams in addition to the baseline bibliographic features. However, the motivation of predicting the impact of published scientific papers is not the same with ours, which is to estimate the patentability of patent applications, since the definitions of the quality of papers and patents are different. Note that the used data sets were both from the same fields (economics and natural language processing) and relatively small so that a small subset of domain-specific textual features would work well. In contrast, our problem is more challenging since the objective is to build a more general model for patent quality that covers broader area of technology.

3. Patent Quality Assessment

In this section we review the examination process for patent applications, discuss how to define and evaluate the patent quality, and formalize the patentability prediction problem [*1].

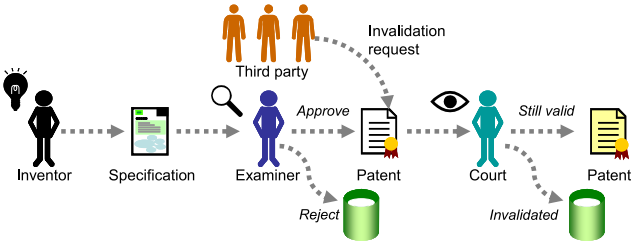

Fig. 3 An overview of the patent examination process in Japan.

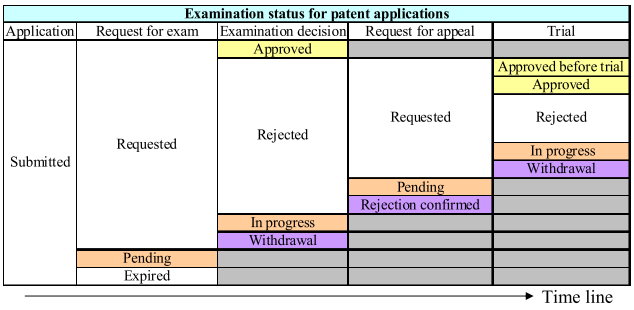

Fig. 4 The status of the patent applications during the examination process.

3.1 Examination Process in Japan

The process of patent examination begins when an inventor submits an application for a patent to the patent office. **Figure 3** shows an overview of the process after the submission. The application contains a specification document that represents the key ideas of the invention. After the inventor requests an examination, an examiner at the Japan Patent Office (JPO) reads the document, evaluates the novelty and non-obviousness of the invention compared to the prior art, and decides whether or not to approve the application to grant the patent. Note that there may be a long series of extensive communications, repeated discussions, and various procedures between the inventor and the examiner before the final decision. In fact, it takes 25 months on average for the patent office to take the first action after the submission of an application. It also takes more than four years on average to finally grant a patent. **Figure 4** shows the status of applications during the examination process. The time line flows from left to right. After the application is submitted, the applicants will make a request for examination. Next a patent officer starts examination and communication with the attorneys and applicants for revising the application documents. For simplicity we omit the details of such prosecution history in Fig. 4. After that, each application will be given a final decision, approval or rejection. After the rejection, the applicants can still make a request for appeal, second chance for having the application evaluated and granted. Even after being granted as a patent, a third party may request an invalidation trial arguing that the patent should not have been granted. To invalidate a patent, the plaintiff must prove that flaws in the original application were overlooked in the examination process. Obviously, invalidation trials are requested for only a small frac-

[*1] Note that we focus on the current patent system in Japan and a data set of Japanese patents. However, we believe that our approach will also work for patent databases in other countries.

tion of the granted patents. In fact, there were only an average of 142 requests for invalidation trials submitted to the Japanese intellectual property high court each year during 2006 to 2009 [25]. In comparison, the total number of patent applications in Japan was 348,596 in 2009 [4].

3.2 Patent Quality

In this section we briefly introduce the patent quality evaluation problem.

The patent systems in many countries are facing a problem of long delays in conducting patent applications. Since the management of intellectual property is also becoming critically important for many industries as a source of competitive advantage, the number of patent applications is dramatically increasing everywhere [5]. In fact, Fig. 1 shows that the total number of the patent applications in the world increased almost twice from 1995 to 2010. In particular, China showed a 17 times increase in the number of applications from 1996 to 2010 as shown in Fig. 2. This growth is causing substantial delays in the examination processes in many countries including the U.S. Patent and Trademark Office are being forced to take actions to shorten the processes for examinations from submission to decision. At the same time, these accelerated examination processes for large numbers of applications must be handled without degrading the quality of the granted patents. If a national patent office carelessly grants a large number of unjustifiable or overly broad patents because of hasty validations by pressured examiners, the flawed patents can actually hinder future progress and business development in that field of technology, since competitors cannot go into the same area of business without violating the excessively broad patents. Even if they have a solid basis to invalidate the unfair patents, the legal processes are slow and expensive.

In the requirements for a "quality patent," the U.S. Patent and Trademark Office emphasizes the complete examination confirmed by the prosecution history and the fact that the scope of protection is defined properly and clearly. Patent offices require better patent systems that allow only high quality patent to be granted through a prompt and fair examination process. Meanwhile, the applicants are also responsible for submitting examiner-friendly applications that are well-written and well-organized. All of the related parties agree on the necessity for a consensus on the quality of patents and tangible criteria. In addition, academic communities studying intellectual property have also extensively focused on the definition of the quality of patents.

The primary approach for analyzing the quality of patents is case-based study of actual court trials. These analyses can give qualitative explanations of the legal validity. However, such case studies impose a heavy workload for the extensive investigation of the prosecution history and of the process of each trial. Hence, it is difficult to make a thorough survey of thousands of invalidation lawsuits.

In this article, we focus on another metric called *patentability* described in the next section. Predicting the patentability of a patent application is equivalent to estimating the possibility that it will be approved. Based on the specification document and following communication with the applicants, patent officer classifies each patent application into one of the two groups, approval group or rejection group. We regard this examination process as a binary classification problem. Approval and rejection are binary class labels and feature variables can be extracted from the specification documents and so on. The decision of patent officer gives the true class label of each patent application. Then the task is to build a classifier which resembles the examination results for patent applications. By using probabilistic classifiers such as naive bayes or logistic regression, we can estimate the probability of being approved for each application as patentability score. In addition, if the classifier model is interpretable as each feature variable is explicitly given a weight, we can estimate the importance of the feature variables for the decisions by evaluating which one has stronger impact on the prediction. Our approach is based only on the quantitative analysis of the patent application information which is publicly available. By collecting the electronic records and the prosecution histories for a number of patent applications and by analyzing them with computational text processing and statistical techniques, it is possible to capture the characteristics of solid patents and weak ones, and their differences.

Note that the main contribution of this article relates to the readability and clarity of the specification documents which is only a part of the complete patent quality. However, we believe that this work can be the first step towards assessing the quality of patent application in a quantitative way.

3.3 Patentability Label

Consulting with patent attorneys, we quantitatively defined patentability as a metric that can be computed as an output of prediction models.

Based on the set of patent applications and their specification documents published by the Japan Patent Office (JPO), we made a training data set of applications with class labels to build patentability models as a supervised classification problem. First we collected a subset of Japanese applications submitted during ten years from 1989 to 1998. We define the class label for an application as +1 if it was approved, and 0 otherwise. All of the approved applications had high patentability with a score of 1.0, while the rejected ones had no patentability with scores of 0.0. Note that we could not assign labels to a subset of the applications such as those that are still under examination. In addition, we also omit the applications for which examinations have not yet been requested by their applicants [*2]. Therefore, we used only the subset of the applications for which the final decisions have been made as our training data set. However, there are no formal records about the current status of each application in the complicated examination process. Then we refine class labels depending on the records of the intermediate actions by the JPO. We assign the class label +1 to the two types of applications approved at the decision or at the appeal trial, since they must have high patentability. Figure 4 also shows that there are three types of approval in the examination process, approval at decision, approval before appeal trial, and approval in trial. In contrast, we

[*2] The request for examination is never made for almost half of the applications in Japan.

give the class label 0 to the applications for which the decision for rejection has been made or for which the time limit for appeal has passed. Figure 4 shows that there are also three types of confirmed rejections including withdrawal. All of the other applications in review or in pending were discarded since their final states have not been decided yet. In our experiments, we derived the class labels for 10% of the applications chosen randomly. Finally we had collected roughly 0.3 million applications with the class labels. It is generally enough data for learning classification models that have good generalization ability and that do not cause over-fitting. We did not use the newer applications submitted after 1998, since there must be fewer applications for which we could define the class labels because of the long examination process.

In contrast, we used all of the Japanese applications submitted from 1993 to 2007 as validation data set for computing the patentability scores and presenting the results of our analysis system.

3.4 Patentability Prediction Problem

In this subsection, we formalize the classification problem that we solved for deriving patentability models to compute the patentability scores. Let x be a data sample that corresponds to a patent application. Each x is a d-dimensional vector that consists of a set of feature values extracted from its specification document. The class label of an application is represented as $y = \{0, +1\}$. We set $y = +1$ when the corresponding application is approved and $y = 0$ otherwise.

Assume that we are given two kinds of data sets, the training data set D^{tr} and the test data set D^{te} as follows.

$$D^{tr} = \{(x_1^{tr}, y_1^{tr}), (x_2^{tr}, y_2^{tr}), \ldots, (x_n^{tr}, y_n^{tr})\}$$
$$D^{te} = \{(x_1^{te}, y_1^{te}), (x_2^{te}, y_2^{te}), \ldots, (x_m^{te}, y_m^{te})\}$$

Note that we do not know the class labels of the test samples (y_i^{te}). The purpose is to learn a classification model based on D^{tr} and predict the unknown labels y_i^{te} in D^{te}.

In this case, D^{tr} represents a set of feature vectors $\{x_i^{tr}\}$ related to the existing patent applications with their examination results $\{y_i^{tr}\}$. We formally define the patentability prediction problem as follows.

Problem 3.1 Given a training data set D^{tr}, learn a classifier $f(\cdot)$ which minimizes the summation error E:

$$E = \sum_{i=1}^{n} (y_i^{tr} - f(x_i^{tr}))^2.$$

Then the classifier $f(\cdot)$ can predict the class label, i.e., the examination result of x_1^{te} as $f(x_1^{te})$. $y_1^{te} = f(x_1^{te})$ means the classification is correct.

If the classifier $f(\cdot)$ can also estimate the posterior probability that the sample x_i belongs to the class label $+1$ as $p(x_i) = p(y_i = +1|x_i)$, then $p(\cdot)$ can be regarded as a probability model that produces a confidence value. In the problem of assessing the quality of patent applications, we can assume that the value of $p(x_i)$ is the likelihood that the application x_i will be approved, which is the patentability score. The more accurate the classifier $f(\cdot)$ becomes, the better the patentability score $p(x_i)$ assesses the quality of patent applications. Therefore, our goal is to build an accurate classifier and to derive a good patentability model by using an effective representation of x_i and a powerful classification algorithm.

Table 1 Structural property features.

Number of characters in title
Number of characters in specification
Number of sheets of drawings
Number of claims
Number of independent claims
Depth of claims tree
Whether IPC includes 'A'
Number of combinations of IPC
Number of inventors
Number of cited references
Number of positive expressions

4. Methodology

In this section we describe what kinds of feature values to use, how to train the patentability models based on the logistic regression, and the results of our performance studies.

4.1 Feature Set

We use four kinds of feature sets to represent the characteristics of the specification documents for each patent application. In the following, we describe each of them.

4.1.1 Structural Properties

The structural properties consist of various types of values computed based on the statistical characteristics of the specification document. We use a subset of the features used in a previous work done by Nagata et al. [14]. We show some of the structural property features in **Table 1**. Most of them are the counts of various kind of properties. For example, we count the number of characters in the title or in the body text of the specification. *IPC* is the acronym for the International Patent Classification, which represents the technology domain of the invention with a taxonomic code such as "F16C1/00." An application can be assigned more than one IPC code. The number of positive expression is the only textual property in this feature set, which counts the number of the occurrences of a pre-defined set of positive expressions such as "can" or "enable."

Note that we omit some of the features defined in the original paper [14] such as "the number of applications that the examiner has examined," since they require manual work to determine the values for each application which is expensive. Therefore, we focus only on the easy-to-compute subsets of the features.

Since these kinds of feature sets have been widely-used for the qualitative evaluation of general documents and for machine learning methods on textual data, we also use our structural property features as a baseline method in the experiments. In the following, we regard the other features as options to be added after the feature values of the structural properties.

4.1.2 Feature Set

TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency) is definitely the most commonly-used feature set for textual data in natural language processing [19]. The concept is to represent the

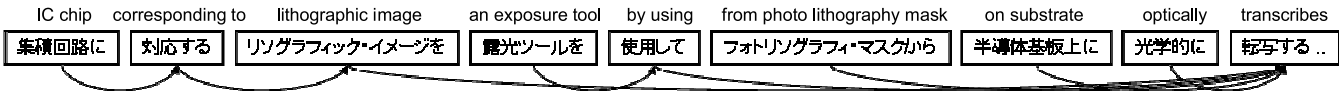

Fig. 5 An example of the dependency structure for the syntactic complexity feature.

Table 2 Syntactic complexity features.

Maximum depth of dependencies in a sentence
Maximum number of terms in a phrase
Maximum number of phrases in a sentence
Number of sentences in the main part
Number of terms in the main part
Number of phrases in the main part

document as a bag-of-words and compute the TF-IDF value of each word s for the i-th document as

$$\mathrm{tfidf}(s,i) = \mathrm{tf}(s,i) \cdot \log(\mathrm{idf}(s)),$$

where $\mathrm{tf}(s,i)$ represents the frequency of the word s in the i-th document, and $\mathrm{df}(s)$ denotes the inverted frequency of s among all of the documents.

Though TF-IDF is known to be very powerful in representing documents with fixed-length vectors, the computational cost increases linearly with respect to the number of words used. Since it is unrealistic to include tens of thousands of words in the features when the number of documents is also large, one has to prune the set of the words. In this article, we only use 2,000 words for which the document frequency df(s) is in the top 2,000.

4.1.3 Syntactic Complexity

Syntactic complexity represents the complexity of the structure of the sentences. The underlying assumption is that when experienced patent attorneys write specification documents, the syntax of the sentences tends to be more complex than general documents such as news articles. Since a specification document should be logical and precise to ensure that there is no misunderstanding, even if it reduces the clarity of the expressions in a general sense. On the other hand, non-experts are expected to use simpler expressions in their specifications. Therefore, if we calculate some numerical values related to the complexity, they can be used as informative features.

The syntactic complexity can be computed based on the dependencies between terms and phrases in the sentences. First we extract the dependency structures from the specification documents by using a commercial text mining software. **Figure 5** shows a small example of such dependency structures. Next we compute the values of the syntactic complexity features listed in **Table 2**. In particular, the maximum depth of the dependencies in a sentence matters because most of the claims in granted patents contain a lot of complex compound sentences. More sophisticated features could be defined, but the current six features are straightforward to use.

4.1.4 Word Age

The previous three sets of features are all based only on the individual specification documents. However, their quality assessment cannot be completed without evaluating the originality compared to prior art. When inventors submit an application to the patent office, they prepare specification documents that describe in detail the central technical problem, the weaknesses of prior art, and the proposed technique to solve the problem. In particular, the *novelty* and *non-obviousness* are the key parts of the requirements of the patent laws.

To simulate the examiners' evaluation of the novelty and non-obviousness of an application, the word age features aim at measuring the ages of the words included in a specification. Since each word has its first occurrence in a document in the data set, we can compute its age based on the duration from the earliest appearance.

First we represent each specification document as a bag-of-words. Then we remove the rare words for which the document frequencies are quite low. Next we find the first occurrences of the words. Later we calculate the word age for each word in each specification. Since the raw values of the word ages are uninformative in characterizing the applications, we aggregate them into the monthly groups to make a histogram over 300 months. To cancel out the effect of small fluctuations in the word ages of the same-generation words, we make the histogram smooth with a sliding window by calculating the average density of word ages for a small range in the histogram (15 months). Finally, we obtain 20-dimensional features for the word age.

Hasan et al. also proposed a method called *COA* (Claim Originality Analysis), which is similar to the word-age metric, to measure the novelty and impact of the patents [7]. However, the motivation is different from ours since COA also calculates the support for the words (meaning how often the words were used in subsequent patents), divided by their ages, to estimate how significant a patent WAS with respect to the later patents. The word age in our method simply aims at evaluating how original an application is compared to prior art since we cannot access the future patent applications in evaluating an under-review application.

4.2 Learning Prediction Models

Given a training data set D^{tr} with the set of features defined in Section 4.1.2, our goal is to learn a classifier $f(\cdot)$ with the prediction model $p(\cdot)$ by solving Problem 3.1.

Based on the previous studies [10], [14], we also use the L2-regularized logistic regression model and maximize the objective function by using the conjugate gradient descent method [11]. In the logistic regression algorithm, the probability prediction model is derived as follows,

$$p(y=+1|x,w) = \frac{1}{1+\exp(-w^T x - b)},$$
$$p(y=0|x,w) = 1 - p(y=+1|x,w) = \frac{\exp(-w^T x - b)}{1+\exp(-w^T x - b)},$$

where w denotes the weight coefficient of the features and b is an offset. Then the output of the classifier is defined as

$$f(x) = \begin{cases} +1 & \text{if } p(y=+1|x,w) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

Note that 0.5 is just an example of the threshold parameter which can be changed for maximizing prediction accuracy. However, since our goal is to make a good metric which reflects the quality of the patent application rather than predicting its actual class label by choosing the optimal threshold, we use the probability outputs of logistic regression models ($p(y = +1|x, w)$) as patentability scores.

The advantage of logistic regression model is that it belongs to linear models so that the trained model can be interpreted to examine which feature has a significant impact on the prediction. This is a very important capability in real-world applications of machine learning, since users tend to prefer interpretable models to black-box models. Therefore, our Patentability Analysis System allows users to understand *why* a patentability score is high (or low).

4.3 Performance Evaluation

In this subsection, we evaluate the predictive performance of the patentability model and the correlation between the patentability scores and the examination results.

We used 0.3 million applications in this experiment. This data set includes about 10% of the Japanese patent applications submitted from 1989 to 1998 which were randomly chosen and given patentability labels (approved or rejected) as described in Section 3.3. We evaluated the predictive performance on this data set by 10-fold cross validation. First we randomly separated the data set into ten subgroups. Next, for each subgroup, we built a patentability prediction model based on logistic regression by using the other nine subgroups as a training data set. Then we computed the value of a performance metric by using the targeted group as a validation data set. We repeated this operation for ten times for the subgroups and obtained the averaged value which represents the predictive performance of the approach.

We use the AUC (Area Under ROC Curves) value which is a widely-used metric to measure the performance of scoring [3]. ROC is the acronym for Receiver Operating Characteristic. Intuitively, an AUC value for a scoring model corresponds to the likelihood that for a randomly selected pair of approved and rejected applications, a higher score is assigned to the approved ones. A higher AUC means a more accurate patentability score, and the maximum AUC value is 1.0. Here we compare four models, the baseline model with only the structural property features, and the extended models with the other three sets of features, respectively. The AUC values are computed with 10-fold cross validation and **Fig. 6** shows the results, from left to right. The AUC value of the baseline model is 0.594. All of the three extended models had higher AUC values. In particular, the word age-based model achieved the highest AUC, 0.607. Note that 0.607 is generally not a good value in terms of the prediction accuracy of classifiers. However, the purpose of the patentability model is not to perfectly predict the examination results. If we could build such model, it could replace the patent examiners and provide automated examination of patent applications, which is obviously unrealistic with current technologies. Thus, raising the AUC value above 0.600 by adding more features is still a meaningful improvement for the patentability model. We believe that our future work will improve on these results.

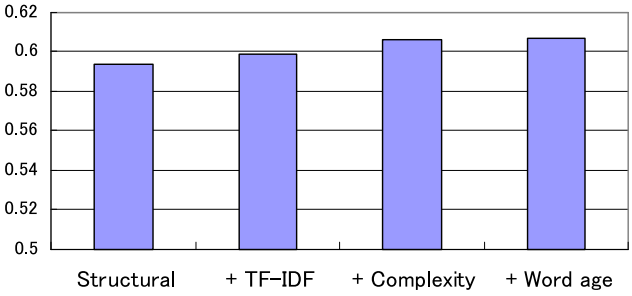

Fig. 6 The vertical axis represents the AUC values, averaged with 10-fold cross validation.

From these models, we will use the third model with the syntactic complexity features as the primary model (M01) in our Patentability Analysis System. We also tested a combination of IPC-specific models. The aggregated model (M05) uses the IPC-specific models depending on the IPC codes for each application. However, M05 had lower AUC values since averaging multiple scores is problematic for AUC due to the different distributions of the scores for each IPC-specific model.

5. Patentability Analysis System

In this section, we introduce a GUI-based service called the *Patentability Analysis System* that was developed for predicting the patentability scores for existing patent applications and analyzing the results. This system is already in service, being used by intellectual property experts and they conducted some of the case studies introduced in Section 6.

5.1 Architecture

This system is designed for intellectual property experts. The objective is to provide them insights into the quality of patent applications by visualizing the quantitative evaluations of the applications based on the patentability models. For example, when a company would like to know how likely their application will be granted, its patentability score is an indicator of the overall quality of the application. In addition, by comparing the feature values defined in Section 4.1.2 with those of the granted and rejected applications, the most important features (both strength and weakness) can be identified. This functionality is implemented in the *Application view*. A user can input the number of any existing patent applications submitted in Japan between 1993 and 2007 of which size is about 5 million. Based on the set of feature values of the application as defined in Section 4.1.2, the system computes the score by applying the patentability model in Section 4.2 and the result appears on the display. In addition, we also developed a *Group comparison view* that enables users to compare the histograms of the patentability scores between paired sets of applications to examine the differences between them.

Figure 7 is an overview of the architecture and the user interface of the Patentability Analysis System. All of the required information including the feature values and the precomputed patentability scores for each applications are stored in the databases on the background server. The input from users and the visualization of the results are handled via Web browsers. In

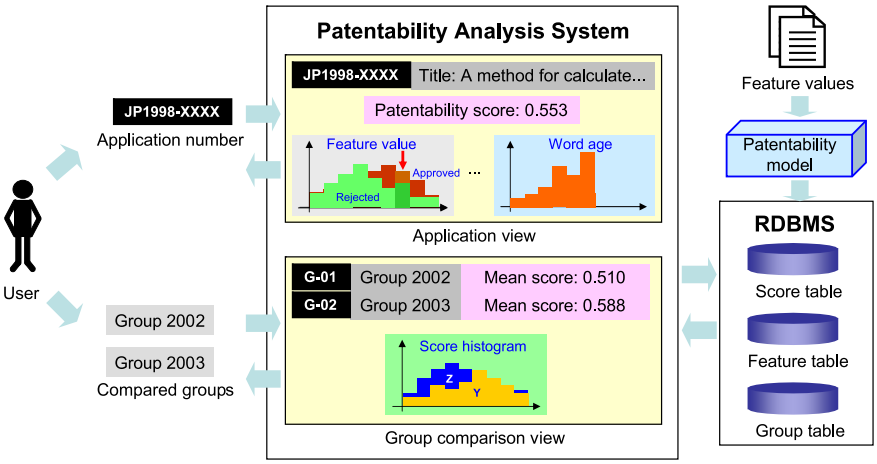

Fig. 7 The architecture of the Patentability Analysis System.

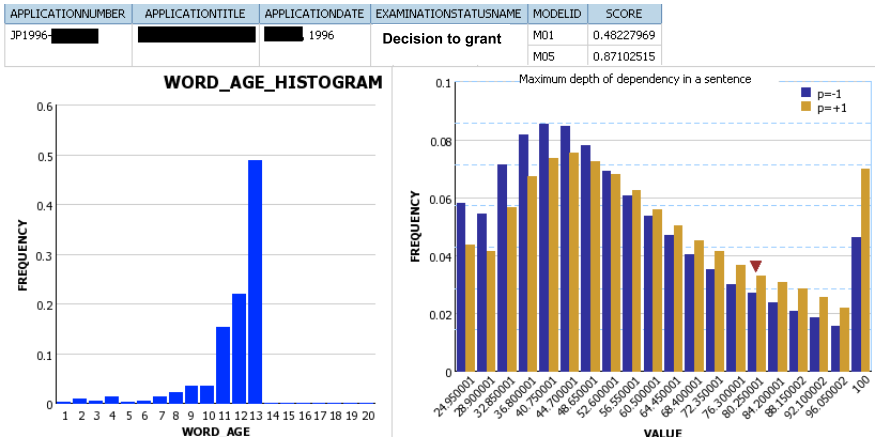

Fig. 8 An example of the Application view: granted patent.

this system, we used a standard relational database system and a commercial visualization program is deployed for the GUI interface.

Next we describe the details of the two views.

5.2 Application View

First we describe the Application view. **Figure 8** shows an example of this view presenting the output of applying the patentability model to a patent application. The application ID number, title and application date of the application are shown from the first to third columns of the top table. Since this is not public work yet, we black out these information except for the application year though this is a result for an actual application. In the fourth column, the examination result is shown as "decision to grant," which means that this application has been granted as a patent. In the fifth and sixth columns, the patentability scores based on two different models (M01 and M05) are shown. The M05 model aggregates the outputs of multiple IPC-aware models relying on the same set of features. In this case, while the M01 model gives a low score of approximately 0.48, the M05 model has a much higher score of about 0.87. The histogram on the bottom left is for the word age of this application. The bars on the right show the frequencies of older words compared to those on the left. Since the ages of frequent words are calculated depending on the submission date of the earliest application in the used data set, the frequency tends to become higher for older sets of words. The figure at the bottom right is for the histograms of the features values of the granted and rejected applications, respectively. Blue bars belong to the set of rejected applications, and yellow bars are those of granted applications. Though the system guarantees these histograms for all of the features, we only show one example for the maximum depth of the dependencies, which is a member of the feature set of syntactic complexity. Note that each neighboring pair of blue and yellow bars represent the frequencies for the same range of feature values. Comparing the two histograms reveals how much impact the feature has on the patentability score. Since the distributions are different here, the patentability score can vary depending on the value of this feature due to the large weight given to this feature in the trained patentability models. In contrast, if the distributions of the bars are almost the same on both histograms, that feature has little effect on patentability. The red triangle (▽) located above the bars represents the value of the feature for this particular application. At the point the triangle indicates, the yellow bar is higher than the blue one. Thus we see that this feature contributes to increasing the patentability score of this application. In other words, the specification document of this application has longer dependency which in general means higher quality. In the same way, users can analyze how and why the patentability score of a target application is high (or low) compared to other applications. **Figure 9**

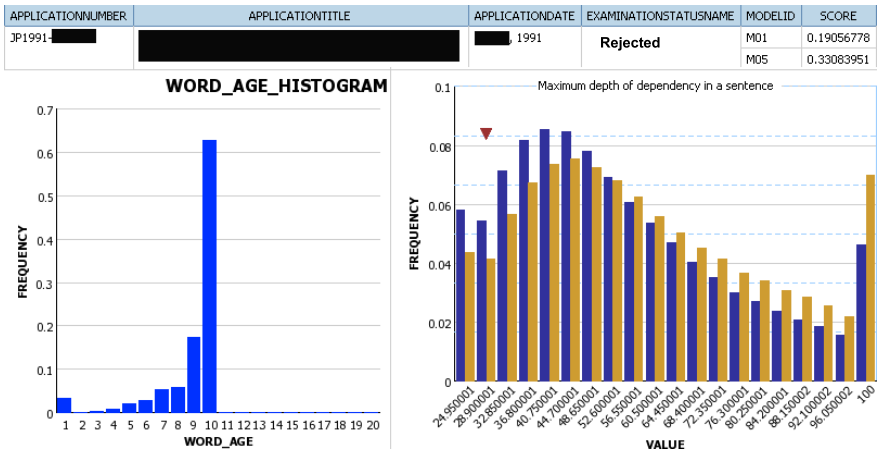

Fig. 9 An example of the Application view: rejected application.

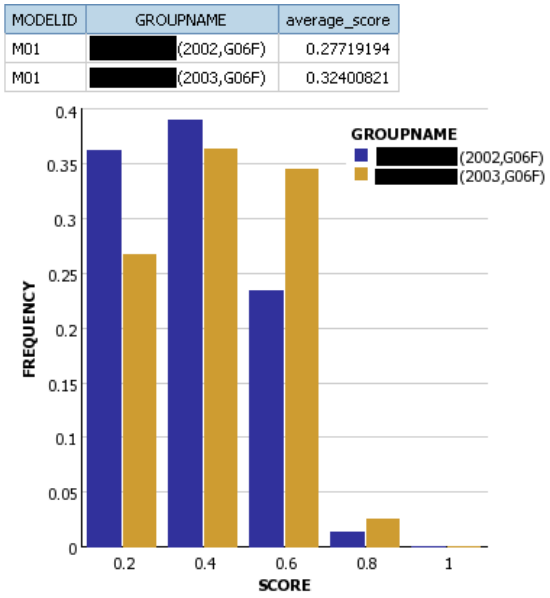

Fig. 10 An example of the Group comparison view: comparison between a company's applications filed in 2000 and 2003.

shows the same output for a rejected application, showing that the patentability scores are both much lower than those for the granted patent in Fig. 8. In addition, the feature histogram at the bottom-right shows that the maximum dependencies of this application is smaller than average, and this feature contributes to lowering the patentability score.

5.3 Group Comparison View

Next we introduce another view for comparing multiple groups of applications. **Figure 10** shows an example of the Group comparison view. In this case we define two groups of applications whose IPC code is G06F, which corresponds to "electric digital data processing," as filed by a major IT company in 2002 and 2003. We show the average patentability score for 2002 and 2003 based on the model M01 in the top table. This indicates that the company filed better quality applications in 2003 compared to 2002. We can actually see the difference in the histograms shown below in which the yellow and blue bars represent the scores for the groups 2002 and 2003. The horizontal axis corresponds to the patentability scores, and the vertical axis represents the proportion of the applications belonging to each bar. The figures under the bars denotes the maximum score for each bar. These histograms show that in the low score range the group 2002 has a higher frequency, while the group 2003 has more applications with scores larger than 0.4. This result suggests that there were some changes in the company's IP strategy between 2002 and 2003. In a similar way, users can analyze and obtain insight on the group view by comparing groups of applications as defined from different perspectives.

6. Case Studies

In this section, we provide some results of case studies conducted by intellectual property experts using the Patentability Analysis System described in Section 5. These results were also published in an article of a magazine for patent experts [17].

6.1 Comparisons of Law Offices

Based on a set of patent applications from a company, the users compared the performance of five law firms whose attorneys are in charge of the patent applications. The total number of cases was 143. **Table 3** shows the grant ratios and the average patentability scores of the approved rejected applications. The names of firms are anonymized. We can see that for the top three firms (A, B, and C) with higher grant ratio, the patentability scores for the approved applications are also higher. In contrast, the average scores are both low for the approved and rejected applications handled Firm D and E, and there is no clear difference between them. Interestingly, Firm C had the highest score for the approved applications (0.640). However, the grant ratio of Firm C was only 66.7% and the patentability scores for the rejected applications was also low (0.588). This suggests that Firm C might have more than two attorneys working on patents, one who is more experienced and one who is a relative novice. These results show that there seems to be a correlation between the quality of the patent applications, the grant ratio, and the patentability score.

6.2 Comparison of Examiner

In the next case, we assume that a user tries to examine the difference between the patent examiners who are supposed to use similar criteria to make their decisions. The total number of ap-

Table 3 The relationship between the average patentability scores after the final decisions are made in the examinations for five law firms.

Firm	Grant ratio	Score (approve)	Score (reject)
A	82.1%	0.614	0.598
B	81.1%	0.629	0.614
C	66.7%	0.640	0.588
D	60.0%	0.585	0.579
E	40.6%	0.568	0.562

Table 4 The relationship between the average patentability scores after the final decisions are made in the examinations for five patent examiners.

Examiner	Approval ratio	Score (approve)	Score (reject)
A	75.0%	0.590	0.588
B	73.1%	0.618	0.624
C	68.4%	0.611	0.583
D	65.6%	0.623	0.556
E	52.4%	0.620	0.571

Table 5 The relationships between the average patentability scores after the trial decisions are made for the same member of examiners.

Examiner	Grant ratio	Score (grant)	Score (deny)
A	81.3%	0.591	0.582
B	92.3%	0.625	0.558
C	78.9%	0.609	0.576
D	90.6%	0.604	0.565
E	75.0%	0.612	0.549

plications was 133.

In **Table 4**, the results are shown for five anonymous examiners, with their average scores and the approval ratios at the stage where the final decision of the examination has been made so that an inventor can still request a trial. For Examiners A and B with the higher approval ratios, the differences between the scores of the approved and rejected applications is unclear. In contrast, for the rest of the patent examiners, the scores for the approved applications are substantially higher. This might indicate that the final decision depends on the examiners in charge. **Table 5** shows the corresponding results for the same examiners after the trial decisions. The differences between the granted and denied patents are much clearer. This may be because only one examiner makes the final decision at the examination, so the result might be biased depending on the examiner. In contrast, since another expert is also involved in the trial, the trial decisions tend to be more neutral and well-correlated with the patentability scores.

Note that though these case studies are based on a small data set, the results support generalizing the patentability models to help the intellectual property experts, allowing them to compare the sets of patent applications that were processed under different conditions.

7. Conclusion

In this article, we studied how to evaluate the quality of patent applications automatically by using text mining and machine learning techniques. First we introduced a new metric of patent quality named patentability. By regarding the approval estimation of patent applications as a supervised document classification problem, we devised a prediction model that computes the likelihood that an application will be approved as patentability score. We used over 0.3 million patent applications submitted to the Japan Patent Office in ten years as a training data set. The feature values for the patent applications were computed based on text mining techniques combined with domain-specific knowledge.

Experiments showed that the proposed prediction model achieved a higher accuracy in predicting the examination results than conventional methods and that the actual grant ratio of the applications was reflected in the computed patentability scores. We also developed a GUI-based visualization tool with which users can see and analyze the patentability of any Japanese patent application filed from 1993 to 2007. This work is a part of our patent quality project done as a collaboration of computer science researchers, patent attorneys, and intellectual property experts working in the industry.

We believe that this work can be the first step towards assessing the quality of patent application in a quantitative way.

Acknowledgments We would like to thank Prof. Hisashi Kashima and Dr. Kentaro Nagata for providing their materials. We also appreciate Dr. Koichi Takeda for his useful comments.

References

[1] Alexander, J.E. and Tate, M.A.: *Web wisdom: How to evaluate and create information quality on the Web*, Lawrence Erlbaum Associates Inc., Hillsdale, NJ, USA, 1st edition (1999).

[2] Borko, H. and Bernick, M.: Automatic Document Classification, *J. ACM*, Vol.10, pp.151–162 (online), DOI: http://doi.acm.org/10.1145/321160.321165 (1963).

[3] Bradley, A.P.: The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms, *Pattern Recogn.*, Vol.30, No.7, pp.1145–1159 (1997).

[4] Economics and Statistics Division of World Intellectual Property Organization: World Intellectual Property Indicators 2010 (2010).

[5] Economics and Statistics Division of World Intellectual Property Organization: World Intellectual Property Indicators 2011 (2011).

[6] Eysenbach, G., Powell, J., Kuss, O. and Sa, E.-R.: Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web, *Journal of the American Medical Association*, Vol.287, No.20, pp.2691–2700 (2002).

[7] Hasan, M.A., Spangler, W.S., Griffin, T. and Alba, A.: COA: Finding novel patents through text analysis, *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'09*), pp.1175–1184, ACM (online), DOI: http://doi.acm.org/10.1145/1557019.1557146 (2009).

[8] IBM Corporation: Introduces Initiatives for Improved Patent Quality (2006).

[9] Kappos, D.J.: Request for Comments on Enhancement in the Quality of Patents, *US Federal Register*, Vol.74, No.235, pp.65093–65100 (online) (2009), available from ⟨http://www.federalregister.gov/articles/2009/12/09/E9-29328/request-for-comments-on-enhancement-in-the-quality-of-patents⟩.

[10] Kashima, H., Hido, S., Tsuboi, Y., Tajima, A., Ueno, T., Shibata, N., Sakata, I. and Watanabe, T.: Predictive Modeling of Patent Quality by Using Text Mining, *Proc. 19th International Conference on Management of Technology* (*IAMOT'09*) (2009).

[11] Koh, K., Kim, S.-J. and Boyd, S.: An Interior-Point Method for Large-Scale l1-Regularized Logistic Regression, *Journal of Machine Learning Research*, Vol.8, pp.1519–1555 (online) (2007), available from ⟨http://portal.acm.org/citation.cfm?id=1314498.1314550⟩.

[12] Markellos, K., Perdikuri, K., Markellou, P., Sirmakessis, S., Mayritsakis, G. and Tsakalidis, A.: Knowledge discovery in patent databases, *Proc. 11th International Conference on Information and Knowledge Management* (*CIKM'02*), pp.672–674, ACM (online), DOI: 10.1145/584792.584915 (2002).

[13] Ministry of Economy, Trade and Industry in Japan: Advanced Measures for Accelerating Reform toward Innovation (2007).

[14] Nagata, K., Shima, M., Ono, N., Kuboyama, T. and Watanabe, T.: Empirical Analysis of Japan Patent Quality, *Proc. 18th International Conference on Management of Technology* (*IAMOT'08*) (2008).

[15] Papka, R. and Allan, J.: Document classification using multiword features, *Proc. 7th ACM International Conference on Information and Knowledge Management* (*CIKM'98*), pp.124–131, ACM (online), DOI: http://doi.acm.org/10.1145/288627.288648 (1998).

[16] U.S. Patent and Trademark Office: Strategic Goal 1: Optimize Patent

Quality and Timeliness (2009).
[17] Patent Committee Group 1 - Japan Intellectual Property Association: Strategic Use of Objective Quality Metrics of Patents/Patent Applications, *Intellectual Property Management*, Vol.61, pp.1–17 (2011). (to be published).
[18] Rieh, S.Y.: Judgement of information quality and cognitive authority in the Web, *J. Am. Soc. Inf. Sci. Technol.*, Vol.53, pp.145–161 (online), DOI: 10.1002/asi.10017.abs (2002).
[19] Salton, G.: *Automatic text processing: The transformation, analysis, and retrieval of information by computer*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1989).
[20] Sampat, B.N.: Determinants of Patent Quality: An Empirical Analysis (2005).
[21] Sebastiani, F.: Machine learning in automated text categorization, *ACM Computing Surveys* (*CSUR*), Vol.34, pp.1–47 (online), DOI: http://doi.acm.org/10.1145/505282.505283 (2002).
[22] Shen, D., Sun, J.-T., Yang, Q. and Chen, Z.: Text classification improved through multigram models, *Proc. 15th ACM International Conference on Information and Knowledge Management* (*CIKM'06*), pp.672–681, ACM (online), DOI: http://doi.acm.org/10.1145/1183614.1183710 (2006).
[23] Tseng, Y.-H., Lin, C.-J. and Lin, Y.-I.: Text mining techniques for patent analysis, *Inf. Process. Manage.*, Vol.43, No.5, pp.1216–1247 (online), DOI: http://dx.doi.org/10.1016/j.ipm.2006.11.011 (2007).
[24] Tseng, Y.-H. and Wu, Y.-J.: A study of search tactics for patentability search, *Proc. 1st ACM Workshop on Patent Information Retrieval* (*PaIR'08*), pp.33–36, ACM (online), DOI: 10.1145/1458572.1458581 (2008).
[25] Yamazaki, J.: Intellectual Property High Court of Japan Important Patent Case Decisions in 2009, LES Pan-European Conference (2010).
[26] Yogatama, D., Heilman, M., O'Connor, B., Dyer, C., Routledge, B.R. and Smith, N.A.: Predicting a Scientific Community's Response to an Article, *Proc. 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., Association for Computational Linguistics, pp.594–604 (online) (2011), available from ⟨http://www.aclweb.org/anthology/D11-1055⟩.

Shohei Hido received his M.Info. degree from Kyoto University in 2006. He has been working in IBM Research - Tokyo since 2006. His research interest includes knowledge discovery and data mining on sensor data and semi-structured data.

Shoko Suzuki is a researcher at IBM Research - Tokyo. She received her Ph.D. in Physics from the University of Tokyo in 2004. Her research interests include data mining, machine learning, text mining and statistics.

Risa Nishiyama received her M.Sc. degree from University of Edinburgh in 2005 and has been working in IBM Research - Tokyo since 2006. Her research interest includes information extraction and analysis methods of patent disclosures, research papers and other types of technical documents. She is a member of IPSJ, JSAI, and the Association for Natural Language Processing.

Takashi Imamichi is a researcher in IBM Research - Tokyo. He belongs to the Analytics and Optimization group. He received his M.Info. and Ph.D. degrees from Kyoto University in 2006 and 2009, respectively. His research interests include cutting and packing problems and metaheuristics. Dr. Imamichi is a member of the Operations Research Society of Japan and the Scheduling Society of Japan.

Rikiya Takahashi received his M.S. degree from the University of Tokyo in 2004. He has been working in IBM Research - Tokyo since 2004, and now is a staff researcher there. He has broad research interests in machine learning and data mining, and performs basic and application studies especially in nonparametric methods.

Tetsuya Nasukawa is a Senior Technical Staff Member at IBM Research - Tokyo. He received his Ph.D. degree from Waseda University in 1998 for his work on natural language processing. He has been working for IBM Research since 1989. His research interests include natural language understanding, text mining, sentiment analysis, and conversation mining.

Tsuyoshi Idé is a Senior Researcher and the manager of the Analytics & Optimization group in IBM Research - Tokyo. He received his M.Sc. and Ph.D. degrees in theoretical physics from the University of Tokyo, in 1997 and 2000, respectively. In 2000, he joined IBM Research as a researcher in display technology. Since 2003, he has been working on data mining research. His current research interest includes knowledge discovery techniques from time-series data.

Yusuke Kanehira received his B.Eng. degree from Kobe University in 1996, and LL.M. degree from Temple University School of Law in 2012, respectively. He has been working in IP Law department of IBM Japan since 2001, and now is a Senior Patent Attorney. He has been engaging in the various IP law practice areas including patent, copyright, IP transaction and IP policy matters.

Rinju Yohda received her M.S. degree from Keio University in 1987. She has been working in Intellectual Property Law Department of IBM Japan since 1994 and now is a counsel whose responsibilities include any Intellectual Property Law services at Research and Development of IBM Japan. She has also been working as an Associate Corporate Licensing Staff.

Takeshi Ueno received his B.Eng. degree from the University of Tokyo. He became a registered Japanese Patent Attorney in 1996, and passed United States Patent and Trademark Office Patent Bar Exam in 2000. He is Senior Counsel, Intellectual Property Law Department, IBM Japan. In this position, he has overall responsibility for all IP matters and operations in Japan. He is Vice President of Japan Intellectual Property Association.

Akira Tajima received his M.Eng. degree from the University of Tokyo in 1992, and Ph.D. degree in 2000. He worked with IBM Research - Tokyo more than 10 years, and also has 4 years experience with A.T. Kearney. He now works with Yahoo! Japan.

Toshiya Watanabe graduated from Tokyo Institute of Technology with Ph.D. degree in 1994. He has experienced R&D and Executive Manager for TOTO LTD from 1985 to 2001. He has been invited to a Professor of the University of Tokyo from 2001. He is currently a Professor of Research Center for Advanced Science and Technology of the University of Tokyo. He is also a Professor of Department of Technology Management for Innovation, the graduate school of Engineering, the University of Tokyo. He has served as a Professional Committee Member of Council for Science and Technology Policy, a Professional Committee Member of Intellectual Property Strategy Headquarters and other many committee members of Ministry of Education, Culture, Sports, Science and Technology, Ministry of Economy, Trade and Industry and Cabinet Office. He is one of founders of intellectual property association of Japan and serves as a director of the association. He has authored seven books and more than 100 academic papers. He has been invited to many international conferences worldwide.