

Managing Knowledge in Light of its Evolution Process: An Empirical Study on Citation Network-based Patent Classification

Xin Li, Hsinchun Chen, Zhu Zhang, Jiexun Li, and Jay F. Nunamaker Jr.

Abstract:

Knowledge management is essential to modern organizations. Due to the information overload problem, managers are facing critical challenges in utilizing the data in organizations. Although several automated tools have been applied, previous applications often deem knowledge items independent and use solely contents, which may limit their analysis abilities. This study focuses on the process of knowledge evolution and proposes to incorporate this perspective into knowledge management tasks. Using a patent classification task as an example, we represent knowledge evolution processes with patent citations and introduce a labeled citation graph kernel to classify patents under a kernel-based machine learning framework. In the experimental study, our proposed approach shows more than 30 percent improvement in classification accuracy compared to traditional content-based methods. The approach can potentially affect the existing patent management procedures. Moreover, this research lends strong support to considering knowledge evolution processes in other knowledge management tasks.

Keywords: knowledge management, machine learning, classification, citation analysis, patent management, kernel-based method.

1. Introduction

The creation, transfer, and management of knowledge have attracted scholarly interest from different disciplines for years [21]. Recent information technology advances have led to increasingly large amounts of data, documents, and other types of knowledge and information. As a result, managers face more challenges in organizing and managing knowledge for future sharing and usage [12, 39]. Some typical knowledge management (KM) tasks include indexing and classifying patents for intellectual property protection and licensing, analyzing online news articles for decision support, managing technical documents for research and development, and maintaining employee profiles for team building.

To facilitate such KM tasks, automated tools such as classification, clustering, and visualization techniques have been widely adopted [49]. In documents and multimedia items, the textual and multimedia contents are often regarded as the major carrier of knowledge (i.e., explicit knowledge) for human cognition [50]. Most previous automated KM techniques treat knowledge items independently and process their contents alone for KM tasks.

However, knowledge items are not independent from each other. Ignoring relationships among them may limit the ability of the KM tools. Knowledge evolves after transfer and reuse during human collaboration [3]. Knowledge creation has been considered as a path-dependent evolution process [38], where innovation is created based on the recombination of prior knowledge elements [15]. For example, project reports can be written based on previous meeting memos; technical documents can be compiled based on older versions; and new patents are invented based on existing technologies. From this

perspective, the knowledge evolution processes may affect the newly created knowledge, as reflected in its content. Therefore, not only the knowledge content but also the knowledge evolution process should be taken into account in KM tasks.

The knowledge evolution process is often embedded in the relationships among individual documents, as the knowledge producers refer to related sources. For example, patents and scientific literature have citations to specify their intellectual basis; Webpages contain hyperlinks to related pages. As individual documents are composed into such “linked” documents, the links potentially represent the evolution history of knowledge. In this research we choose one type of linked document, patents, and conduct an empirical study to exploit the utility of knowledge evolution processes in KM tasks. Specifically, we focus on patent classification, which is both practically important to managers and theoretically representative to other KM tasks.

Patents contain a significant amount of knowledge on technical innovations. Patent management, at both the organization level and the society level, prompts the exchange of inventions [43] and reduces the duplication of research efforts [16]. In the past two decades, the advance of technology and the changes in patent policies have led to a surge in patent applications and publications, especially in high-tech fields [54]. As a result, patent processing time has been prolonged by more than 50% since 1994 [24] while the patent examiners' workload has been continuously increasing [28]. In patent management, classification plays a critical role, including assigning patent applications to examiners [48] and organizing patents based on patent classification schemes, e.g., the United States Patent Classification (USPC) system. The performance of patent classification affects the efficiency of patent examination and the effectiveness of patent search systems.

Most previous studies in patent classification focused on only content analysis and addressed the problem as a canonical text categorization problem [35, 44]. Although various features extracted from patent contents have been used and several machine learning algorithms have been applied [13], such approaches have not provided satisfactory performance [48]. On the other hand, patent citations have been considered a valid representation of knowledge diffusion and reuse in innovation creation [1, 46], in the sense that citing patents adopt knowledge elements from cited patents. The evolution processes of innovations can be represented as patent citation networks. The unsatisfactory performance of existing content analysis approaches and the explicit representation of the knowledge evolution process by patent citations make patent classification a good testbed to evaluate knowledge evolution processes' benefits to KM tasks.

Under a kernel-based machine learning framework, we explore different methods to model patent citation networks. We propose a novel model named labeled graph kernel, which shows a significant improvement in classification performance as compared with traditional content-based approaches. We also identify both the citation network structure and the features of cited patents as important factors in describing knowledge evolution processes for patent classification. This study shows the possibilities for further automating the patent examination process and the benefits of considering the knowledge evolution process in KM tasks.

The remainder of the paper is structured as follows. In Section 2, we review previous research on patent classification in the context of linked document classification. We also briefly review kernel-based methods on structured information. In Section 3, we describe

a kernel-based approach and propose several kernels that use patent citation networks and patent contents for classification. Section 4 reports our experiments on a nanotechnology-related patent testbed. Section 5 discusses the experimental results. Section 6 presents our conclusions and future directions.

2. Literature Review

As a common type of knowledge, linked documents such as patents, scientific literature, and Webpages are associated by links in the form of citations or hyperlinks. From a knowledge management perspective, the document content contains different forms of knowledge, while the links among them indicate the process of knowledge transfer and diffusion.

The classification of linked documents is of interest to both managers and scholars. Classification tools have been developed and adopted in patent management [48], Webpage management [9] and scientific literature management [19, 45, 49]. Among these tasks, patent classification has its unique challenges due to both its critical role in practice and its data characteristics [48]. Patent classification is usually conducted on a large number of categories (for example, the USPC has 450 first-level categories and 160,000 second-level categories). Many of these fine-grained classes have subtle semantic differences and usually have an uneven number of patent instances [30]. All these factors make patent classification difficult to address compared to other linked document classification tasks.

2.1 Classification of Linked Documents

We review previous patent classification studies in the context of linked document classification from two aspects: features, i.e., how the documents are represented, and

algorithms, i.e., how the documents are classified.

2.1.1 Feature Types

Previous studies on the classification of patents mainly consider the features in individual documents. Features related to the citations (links) between documents have also been used.

1) Features of individual documents:

Most previous research considered only the knowledge embedded in individual patents and extracted features from individual documents to represent patents. These features can be categorized into content features and metadata features. Content features are often considered good indicators of document subjects, which can be extracted at the word level (i.e., “bag-of-words”) or phrase level from different parts of the documents. In patent classification, previous studies examined the features extracted from patent title [32], abstract [14, 32, 35], claims [23], and full-text [29]. Features from patent title and abstract have been found to be more effective in patent classification.

The metadata, which usually describe the document’s author, institution, publication date, etc., may be highly correlated with its content and topic. In patent classification, Richter and MacFarlane have used a patent’s IPC category to help classify it into another classification scheme [42]. In Webpage classification, Yang et al. used Webpage headers to help label Webpages by industry sectors [56]. These studies demonstrated metadata’s effectiveness in improving classification performance.

2) Features of citations/links:

In machine learning literature, citations (links) indicate the close relationship between linked documents’ topics, methods, etc. From the knowledge creation perspective,

citations (links) indicate the inheritance or transfer of knowledge elements between linked documents [15]. In linked document classification, features can be defined on direct citations or the entire citation network (of directly and indirectly connected documents) by considering different levels of the knowledge evolution process.

The simplest way to take advantage of direct citations is to combine features of the neighboring (directly cited) documents and use them to describe the focal document. Studies in both patent classification [7] and Webpage classification [18, 40, 56] have shown that combining the neighbor documents' content features cannot significantly improve classification performance. However, it has been found that combining neighbor documents' classification category (metadata) features does yield improvement [7, 40].

Another method that utilizes direct citation information is to define features on linkage relationships. In Webpage classification research, hyperlinks have been represented as first-order logic clauses to build first-order rules describing the common characteristics of Webpages in the same category [9, 56]. Document similarity measures based on document in-links (co-citation similarity) [47], out-links (bibliographic coupling similarity) [27], or both in-links and out-links (Amsler similarity) [2] have been used with the K-nearest neighbor (KNN) algorithm and the Support Vector Machine (SVM) algorithm [6, 11, 25] in both Webpage and scientific literature classification studies. Although citation measures have been widely used in patent analysis studies to assess the impact of patents, inventors, and assignees [22, 37], few previous studies have taken advantage of linkage features.

While using direct citations only considers a single step of the knowledge transfer between citing and cited documents, using features extracted from the entire citation

network is a natural extension that gives a more complete picture of the knowledge evolution process. In recent studies on network topological analysis, researchers found that the networks of patents [34], Webpages [5], and scientific literature [41] are different from random networks. Their organized topological characteristics indicate rich information is contained in these networks. However, few studies have considered using features defined on patent citation networks to represent the knowledge transfer and innovation generation processes in patents and to address the patent classification problem.

2.1.2 Algorithm Types

The algorithms used in patent and other linked document classification can be categorized into feature-based methods and kernel-based methods.

1) Feature-based methods:

Feature-based methods are the major approach used in previous patent classification research. In feature-based methods, a data instance is represented by a feature vector, in which the features are explicitly constructed and selected based on domain knowledge or using automatic algorithms. In patent classification, KNN [53], Winnow [29] [30], Naïve Bayes, and probabilistic relational model (PRM) [52] have been widely applied on content features. Feature-based methods can utilize different types of information by incorporating different types of features in the feature vector. In previous research, content features and neighbor document features (direct citation features) have been used together with the Naïve Bayes algorithm in both patent and Webpage classification [7, 40].

2) Kernel-based methods:

Unlike feature-based methods, kernel-based methods do not require the explicit definition of feature vectors. A kernel-based method contains a kernel function and a kernel machine. The kernel function (or kernel) maps data instances from the input space \mathcal{X} to a feature space H (named reproducing kernel Hilbert space, RKHS) $\Phi(x) : \mathcal{X} \rightarrow H$, by defining a similarity measure between data instances $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathfrak{R} \quad (x, x') \rightarrow k(x, x')$. Although $\Phi(x)$ is not explicitly defined, for every pair of data instances the kernel function ensures that $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$. A kernel machine, such as SVM, is a learning algorithm which performs learning tasks in the feature space H [17]. Given limited types of kernel machines (with SVM being state-of-the-art), the performance of kernel-based learning is highly dependent on the selection and design of kernel functions [51].

In linked document classification, kernel-based methods have not been used as widely as feature-based methods. However, they have shown their potential in some recent studies. For example, Fall et al. compared the performances of KNN, Naïve Bayes, and Winnow with SVM on a linear kernel using content features and found that SVM with the linear kernel outperformed the other three feature-based methods [13, 14]. In Webpage classification, SVM has been used on kernels defined on linkage-based similarities and reported good performance [6].

In kernel-based methods, we can use well-established kernel composition rules to combine different types of information in a learning task [10, 25, 51]. In Webpage classification, Joachims et al. adopted such a kernel composition method to consider both direct citation information and content information [25].

2.2 Kernel-based Methods on Structure Information

Although feature-based methods have been widely used in classification problems, they are often criticized for requiring explicit feature extraction. It is also difficult to define and extract features from instances with complex structures. This may be one reason that citation networks have been used less in patent classification. Kernel-based methods provide an effective alternative to feature extraction for capturing such complex structure information.

In kernel-based methods different kernel functions have been designed to capture structure information [17]. Among these kernels, the convolution kernel [20] is one of the most widely used. For objects (data instances) containing a set of sub-objects, convolution kernels calculate the similarities between object pairs by conducting pairwise comparisons between the set of sub-objects they contain. As a special case of convolution kernels, graph kernels are designed for data instances whose sub-objects constitute a graph. The similarity between two graphs can be calculated by comparing the sub-structures in the graphs, such as nodes, paths, and sub-graphs. By representing graphs as random walk paths and conducting pairwise comparison of (matching) random walk paths, graph kernels have been successfully used to classify proteins according to their molecular (graph) structures [4, 26, 33].

Although previous studies showed the effectiveness of capturing structural information using graph kernels, most of these studies focus on the structure information of the sub-objects contained in data instances. In the patent classification problem, patent citation networks represent the structural information outside of data instances, i.e., the evolution processes of innovations. Few previous studies have made the effort to capture such context structure information for classification purposes.

2.3 Research Gaps and Research Questions

As an important knowledge management task, patent classification has been studied by a number of researchers. However, most previous studies isolated the knowledge contained in an innovation (patent) from its evolution process and employed only individual patent contents to address the classification problem. Even in the broader literature of linked document classification, use of the knowledge evolution process was limited to direct citations (one-step knowledge transfer). The structure of citation (linkage) networks has not been widely utilized.

We are interested in methods that will capture the structure of patent citation networks for the patent classification problem. We focus on the following two research questions in this research:

Q1. *Exploiting the evolution process*: Can the methods using citation networks outperform those using only direct citations? Will the features in the directly and indirectly cited patents be helpful for classifying the citing patent?

Q2. *Combining an innovation's intrinsic information with its evolution process*: Will combining citation information with patent contents improve patent classification performance compared with using citation or content information alone?

3. Research Design

To capture the structure of citation networks, we adopt a kernel-based approach, which also enables us to combine citation information with content information.

3.1 A Framework of Kernel-based Patent Classification

Figure 1 presents a general framework for addressing the patent classification problem using a kernel-based approach. 1) At the data acquisition and parsing stage, patent data

are retrieved and parsed into structured data. It should be noticed that both the patents of interest and their directly or indirectly cited patents need to be extracted. 2) At the kernel construction stage, the similarities between data instance pairs are pre-computed according to the kernel function designs. Different kernel functions can capture different information in patents and patent citation networks. 3) At the classifier learning stage, classifiers are learned based on the pre-computed kernel values using a kernel machine. In this research, we chose SVM as the kernel machine because of its reported good performance [13, 14, 25, 35]. 4) At the evaluation stage, testing data instances are provide to the classifiers for predictions. The classification performances of different classifiers are evaluated by comparing the predictions against the actual categories provide by experienced patent examiners.

[Figure 1. A framework of kernel-based patent classification]

In the proposed kernel-based framework, kernel functions define similarity measures between data instances and capture patterns in data instances. The kernel machine is in charge of building the classification models. The performance of kernel-based methods is highly dependant on the design of kernel functions [51]. The major problem (and contribution) of this research becomes designing appropriate kernel functions for patent classification.

3.2 Kernel Function Design

In light of the research gaps, we adopt and design several citation-related kernels that utilize patent citation and content information. Among these kernels, the labeled citation graph kernel is a novel kernel that captures more comprehensive information from the patent citation networks.

3.2.1 Using Citation Information

We considered two conditions in the design of citation-related kernels.

The scope of the cited documents: The different levels of citations represent the different steps of knowledge transfer. In addition to considering direct citations as an approximation for one-step knowledge transfer, we can extend the citation structure and consider multiple levels of cited documents, which represent a more complete picture of the knowledge evolution process.

The features of the cited documents: When modeling an innovation’s evolution process, we can choose to use or not use the cited documents’ features. Without considering features of cited documents, a patent’s cited patents are encoded only as identifiers. If cited documents’ features are considered, the semantics of knowledge elements in cited patents are used, which provide extra clues for understanding the focal innovation. In this study we consider the known classification categories of the cited patents as this type of feature, due to reported effectiveness in patent classification [7].

By combining these two conditions, we construct four kernels on patent citation information (see Table 1): bibliographic coupling kernel (K_{Bib}), labeled co-reference kernel (K_{Ref}), graph overlap kernel (K_{Ovr}), and labeled citation graph kernel (K_{Gra}).

[Table 1. Kernels for citation information]

a) Bibliographic coupling kernel:

The bibliographic coupling kernel (K_{Bib}) design adopted from [6] was initially used in the context of Webpage classification. It utilizes direct citations of patent documents without considering the cited documents’ features. In this kernel, a patent p is represented

by a set of patents it cites: $CV_p = \{q : p \text{ cites } q\}$. The similarity between two patents is defined as the number of their common citations divided by the total number of their citations:

$$K_Bib(p_1, p_2) = \frac{|CV_{p_1} \cap CV_{p_2}|}{|CV_{p_1} \cup CV_{p_2}|}$$

where p_1 and p_2 are two patents and CV_{p_1} and CV_{p_2} represent the two sets of patents they directly cited. In this kernel, the more common neighbors that two patents share, the more similar they are.

b) Labeled co-reference kernel:

We design a labeled co-reference kernel (K_Ref) to consider cited patents' features (classification category) while using only the direct citations. In this kernel, a patent p is represented as a classification category vector, $CC_p = (c_1, c_2, \dots, c_n)$, where the elements are the numbers of directly cited patents of p that belong to each classification category. The labeled co-reference kernel is defined as the normalized inner product of the classification category vectors:

$$K_Ref(p_1, p_2) = \frac{\langle CC_{p_1}, CC_{p_2} \rangle}{\sqrt{\langle CC_{p_1}, CC_{p_1} \rangle \cdot \langle CC_{p_2}, CC_{p_2} \rangle}}$$

where p_1 and p_2 are two patents and CC_{p_1} and CC_{p_2} represent their classification category vectors. In the labeled co-reference kernel, if two patents have similar citation patterns in different categories, they have relatively high similarity.

c) Graph overlap kernel:

Based on the idea of the bibliographic coupling kernel, we design a graph overlap kernel (K_Ovr) which considers more than one level of the cited patents in the patent citation network. In this kernel, a patent p is represented by the set of patents it directly or

indirectly cited: $GV_p = \{CV_p \subseteq GV_p; \text{ if } s \in GV_p \text{ and } s \text{ cites } t \text{ then } t \in GV_p\}$. The similarity of two patents is defined by the ratio of the overlap part of the two patent citation networks in the union of the two networks:

$$K_Ovr(p_1, p_2) = \frac{|GV_{p_1} \cap GV_{p_2}|}{|GV_{p_1} \cup GV_{p_2}|}$$

where $|GV_{p_1} \cap GV_{p_2}|$ is the number of common patents in the two citation networks, and $|GV_{p_1} \cup GV_{p_2}|$ is the total number of patents in the two networks. In the graph overlap kernel, the larger the overlap part of the two citation networks, the more similar the two patents are.

d) Labeled citation graph kernel:

Lastly, as our main contribution, we propose a labeled citation graph kernel (K_Gra) which considers both the network of cited documents and the cited documents' features. In this kernel, a patent p is associated with a labeled citation network,

$G_p := (GV_p, GE_p, GL_p)$, which contains the patents directly or indirectly cited by p : GV_p , and the citations between all patents in GV_p : $GE_p = \{(s, t) : \forall s, t \in GV_p \text{ and } s \text{ cites } t\}$. In this network, each node (patent) is labeled with its classification category:

$GL_p = \{label(q) : \forall q \in GV_p\}$. The similarity between two patents is measured by the similarity between the labeled citation networks associated with them.

[Figure 2. Random walk paths on a labeled citation network related to patent S]

In order to analyze patents using their associated labeled citation networks, the labeled citation graph kernel compares the random walk paths, starting from the focal patents on their associated labeled citation networks, and composes path similarities into the similarities of focal patents. This is different from previous graph kernel studies that

target analyzing graphs, which compare random walk paths starting from any nodes in the focal graphs [26]. The random walk paths are generated from the focal patents following patent citations (Figure 2). When a random walk is conducted, it follows a probability distribution and may jump from one patent to one of its neighbors (cited documents) or stop at the patent. From a knowledge diffusion perspective, the random walk paths represent the knowledge transfer paths (reversely) from prior innovations to focal patents. In this model, a longer random walk path has a lower probability of existence, indicating the less impact of older predecessors on new innovations. In the labeled graph kernel, each random walk path is represented as a sequence of labels (i.e., classification categories) of the nodes on the paths, which partially documents the knowledge elements related to this knowledge transfer path. The similarity of two paths is considered to be one if they share identical label sequences. Otherwise, it is considered to be zero for the sake of simplicity. The labeled citation graph kernel is defined as the sum of pairwise path similarity values, which are weighed according to the probabilities these random walk paths may exist. In other words, the kernel compares all knowledge transfer paths leading to each innovation to identify patents on similar topics. The algorithm to calculate the labeled citation graph kernel is summarized in Figure 3.

[Figure 3. The algorithm for labeled citation graph kernels]

3.2.2 Using Individual Documents' Content Information

The kernel that uses individual documents' content information represents the previous efforts that used content features to address the patent classification problem. In previous studies, features extracted from patent abstracts, claims, and descriptions have all been used. Patent abstracts have been reported to be slightly more informative than other

features in patent classification [32, 35]. The linear text kernel has been reported to have good classification performance [13, 14]. Therefore, we use the patent abstract to represent the entire patent content and choose the linear text kernel to capture patent content information. Such a setting works as a baseline to evaluate the performances of the citation-based kernels. In the linear text kernel, each patent p is represented by a term vector, $C_p = (t_1, t_2, \dots, t_m)$, where the elements are the number of occurrences of terms in the abstract. The linear text kernel (K_Txt) defines the similarity of two patents as the normalized inner product of the term vectors [25]:

$$K_Txt(p_1, p_2) = \langle C_{p_1}, C_{p_2} \rangle / \sqrt{\langle C_{p_1}, C_{p_1} \rangle \cdot \langle C_{p_2}, C_{p_2} \rangle}$$

where p_1 and p_2 represent two patents and C_{p_1} and C_{p_2} are their term vectors.

3.2.3 Using Both Content & Citation Information

Using kernel composition methods, it is easy to consolidate different types of information by combining multiple kernels. We use the simple addition operation to combine kernels that use citation information (K_Bib , K_Ref , K_Ovr and K_Gra) with the linear text kernel (K_Txt) into four composite kernels (K_Com_1 - K_Com_4). For any two kernel functions $K_1(p_1, p_2)$ and $K_2(p_1, p_2)$, the addition operation creates a new kernel by adding together corresponding kernel values: $K(p_1, p_2) = \lambda K_1(p_1, p_2) + (1 - \lambda) K_2(p_1, p_2)$. The addition operation on the two kernels implicitly combines the feature spaces defined by them. The parameter λ controls how much each kernel contributes to the composite kernel. This set of kernels represents the efforts that exploit both patent contents and the associated knowledge evolution process.

4. Experimental Study

4.1 Testbed

In order to examine the effectiveness of proposed kernel functions for patent classification, we conducted an experimental study on a nanotechnology-related patent dataset acquired from the USPTO. We chose USPTO patents because they have more complete citation information than patents from other patent offices (hence more reliable citation networks). We selected patents in a specific domain so as to restrict the size of the testbed without significantly reducing the difficulty of the patent classification task. Specifically, nanotechnology was selected due to its deep impact on a nation's technology advancement and its rapid development in patent publication in recent years, reflecting the characteristics of many high-tech domains.

We retrieved nanotechnology-related patents from the USPTO by keyword-searching in patent title, abstract, and claims, using a keyword list provided by domain experts [22]. The retrieved patents were parsed into structured data and stored in a relational database. We also retrieved the patents they directly or indirectly cited to reconstruct the citation network. Since the number of cited patents increases exponentially as the citation level increases, from a practical standpoint we retrieved only cited patents that are two steps away from the core set of patents. (The testbed may contain a patent's ancestors that are more than two steps away, if it cites the patents in the core set of patents.)

We split the testbed into a training set and a testing set following previous studies [30]. Given a specific date, patents published prior to that date were used as training data, while applications filed after that date were used as testing data. The patents under review on that day, which have been applied for but have not yet been issued, were not considered in either the training or testing dataset. In this research, the patents published

between 01/01/1999 and 12/31/2001 were used for training. The patent applications that were filed between 01/01/2002 and 12/31/2004 were used for testing. We used a patent's major USPC category as its classification label. To provide enough instances to train the classifier, we restricted the experiments to categories with more than 100 patents in the training dataset. After preprocessing, the training dataset contained 13,913 data instances and the testing dataset contained 4,358 data instances (see Table 2) which belong to 36 first-level USPTO categories. The number of instances in each category varied from 109 to 1,895 in the training data and from 15 to 705 in the testing data (Figure 4). The retrieved citation network of the training set contained 336,303 patents, and that of the testing set contained 227,833 patents. As there were overlap patents in these two citation networks, in total we collected 451,853 patents.

[Table 2. Number of data instances in the testing and training datasets]

[Figure 4. Patent distribution in USPC categories]

Our research testbed illustrates the challenges in patent classification discussed earlier. Produced by a multi-disciplinary research field, nanotechnology patents cover many USPC categories [22]. Some of these patents may have minor topical differences and are difficult to differentiate. In the dataset, the numbers of instances are uneven in different categories. This dataset contains 36 classification categories, which is comparable to previous patent classification studies.

4.2 Experimental Procedures

After creating the training and testing datasets, we calculated the kernel matrices that contain the kernel values between the patents in the datasets. To construct the linear text kernel matrix, we preprocessed the contents (abstracts) of the patents in the testbed using

the open source package “Rainbow” [36] for stemming, indexing, and feature selection (based on mutual information). To construct the kernel matrices of citation information, we used the extracted citation relations and classification categories of cited patents and pre-computed the kernel values according to their definitions. To construct the four composite kernels, we set λ as 0.5 and added the linear text kernel matrix to each of the four kernel matrices of citation information. We chose λ as 0.5 for consistency with past research [25], where individual documents’ content and citation information have equal effect on the final kernel matrix. It is worth knowing that parameter λ can be optimized by solving a semidefinite programming problem [31]. However, parameter optimization is out of the scope of the current research and will be considered in the future. After the pre-computation of kernel matrices, we used a well-known high-performance SVM package, “libSVM” [8], to build the classification models. We classify each patent to only one class, which is considered as its major classification category. The predictions on the testing dataset are used for evaluation.

4.3 Evaluation

For each of the data instances in the testing dataset, we compared the classifiers’ predictions with its actual classification category in the USPTO. We used standard classification performance metrics, accuracy, precision, recall, and F-measure, to evaluate the performance of different kernels with the SVM algorithm. These metrics have been widely used in information retrieval and machine learning studies.

Accuracy is usually used to assess the overall performance of a classifier at the instance level. For the instances in the testing set,

$$\text{Accuracy} = \frac{\text{number of all correctly identified instances}}{\text{total \# of instances}}.$$

Precision, recall, and F-measure are defined to evaluate the performance of a classifier on individual classes. For a class i , if TP_i is the number of correctly identified instances of class i , FP_i is the number of instances incorrectly assigned to class i , and FN_i is the number of instances which belong to class i and have been assigned to other classes by mistake, then

$$\text{Precision } P_i = TP_i / (TP_i + FP_i),$$

$$\text{Recall } R_i = TP_i / (TP_i + FN_i), \text{ and}$$

$$\text{F-measure } F_i = 2 \times P_i \times R_i / (P_i + R_i), \text{ which combines precision and recall.}$$

The micro-average (per-instance) value and macro-average (per-category) value of precision, recall, and F-measure can be used to compare the kernels' overall performances [44, 55]. Given that our experiments are designed as single-label classification, the micro-averaged precision, recall, and F-measure are equal to accuracy, which favors the categories with large numbers of instances by giving each instance the same weight. Thus, we report the macro-averaged precision, recall, and F-measure, which favor the categories with small numbers of instances since each category has the same weight.

4.4 Hypotheses

In correspondence with the research questions, we test two sets of hypotheses to examine the effects of using citation networks in patent classification. In these hypotheses, we adopt (a) accuracy and (b) F-measure (which combines the precision and recall) to gauge the instance-level and category-level performances of different settings.

H1.1a. Kernels that use the structures of patent citation networks will outperform those that use only direct citations on classification accuracy in patent classification.

H1.1b. Kernels that use the structures of patent citation networks will outperform those that use only direct citations on F-measure in patent classification.

H1.2a. Kernels that use classification categories as cited documents' features will outperform those that do not use any cited documents' features on classification accuracy in patent classification.

H1.2b. Kernels that use classification categories as cited documents' features will outperform those that do not use any cited documents' features on F-measure in patent classification.

H2.1a. Composite kernels of citation information and patent content will outperform the linear text kernel that uses patent contents on classification accuracy in patent classification.

H2.1b. Composite kernels of citation information and patent content will outperform the linear text kernel on patent contents on F-measure in patent classification.

H2.2a. Composite kernels of citation information and patent content will outperform kernels that use only citation information on classification accuracy in patent classification.

H2.2b. Composite kernels of citation information and patent content will outperform kernels that use only citation information on F-measure in patent classification.

We conducted single-sided pairwise *t*-tests to test these hypotheses. The *t*-test on accuracy was conducted at the instance level, in which the mean of every instance's correctness (0 or 1) is accuracy. The *t*-test on F-measure was conducted at the category level, in which the mean of every class's F-measure is the macro-averaged F-measure.

5. Results and Discussion

5.1 Overall Performances

Table 3 reports the performances achieved by the SVM classifiers with different kernels. We can observe that both the labeled citation graph kernel (K_{Gra}) and its composition with the linear text kernel (K_{Com_4}) have high accuracies, precisions, recalls, and F-measures. They achieve much better performances (31.12% and 32.29% absolute improvement in accuracy) than the baseline linear text kernel (K_{Txt}). Considering that the linear text kernel represents the performance of content analysis (using the knowledge embedded in patents) in previous research and applications, the two kernels show good potential to be used in real applications. Both kernels utilize the network structure of patent citations and the classification category features of cited documents, which account for their good performances.

[Table 3. Performances of different kernels]

In the experiments, the bibliographic coupling kernel (K_{Bib}) and the graph overlap kernel (K_{Ovr}) have low accuracy values (7.48% and 37.13%, respectively). This may be a direct result of their sparse kernel matrices. The designs of these two kernels compare patent citations according to exact match. Given the millions of patents existing in the world, the probability that two patents share the same references is very low. Thus, there is a high probability that the kernel values will be zero. In our experiments, the bibliographic coupling kernel has 99.88% zero values and the graph overlap kernel has 98.37% zero values. Compared with the linear text kernel whose matrix has 38.81% zero values, the two kernel matrices are too sparse to capture enough information to differentiate patents and build an effective classifier.

We also noticed that the bibliographic coupling kernel (K_{Bib}) and the graph overlap

kernel (K_{Ovr}) have much lower recalls (5.81% and 29.08%) than the other kernels, while most kernels have similar precision values (except the labeled citation graph kernel and its composition with the linear text kernel). Further inspection shows that the two kernels tend to assign most patents into certain classes by mistake. For example, the bibliographic coupling kernel assigns most instances into USPC category #435 (Chemistry: molecular biology and microbiology) with a low precision. The few instances left were assigned accurately, which lead to a high precision and a very low recall in most classes. For example, the bibliographic coupling kernel has 100% precision in assigning a couple of instances into some categories (e.g., 3 instances in USPC category #073, 3 instances in USPC category #106, and 1 instance in USPC category #252).

5.2 Hypotheses Testing

To further assess the factors that affect the performances of different kernels, we tested the hypotheses by conducting single-sided pairwise t -tests on accuracy and F-measure (Table 4). The pairwise t -tests on accuracy were conducted at the instance level ($n=4,358$); the pairwise t -tests on F-measure were conducted at the class level ($n=36$).

Statistical tests confirm that the kernels that use networks of patent citations significantly outperform the kernels that use only direct citations on both accuracy and F-measure (i.e., H1.1a and H1.1b are supported). Using citation networks explicates the relationship between the patents which do not share directly cited patents but share indirect ancestors. Such explications may provide more evidence when the classifiers try to categorize such patents into the same class. In addition, using citation networks differentiates the patents with similar directly cited patents more distinctly by inspecting

more levels of citations. Such detailed differentiation may enable the classifiers to categorize ambiguous patents into different classes more precisely.

[Table 4. Hypotheses testing for different kernels]

Statistical tests confirm that the kernels that use cited documents' classification category features significantly outperform those that do not use any cited documents' features on both accuracy and F-measure (i.e., H1.2a and H1.2b are supported). In previous research, it was found that employing neighbor documents' classification category information can improve the classification accuracy [7, 40], which is confirmed by our experiments. Our experiments further suggest that, when the entire citation network is considered, cited documents' features can still play an important role.

Statistical tests show that all four composite kernels significantly outperform the linear text kernel (K_Txt) on both classification accuracy and F-measure (i.e., H2.1a and H2.1b are supported). In the statistical test to compare composite kernels with the kernels using only citation information, although the labeled citation graph kernel and its composition with the linear text kernel do not have statistically significant differences in F-measures in the testing of H2.2b (p value ≈ 0.533), all other tests on accuracy and F-measure confirm a better performance when combining information (i.e., H2.2a is supported and H2.2b is partially supported). The statistical test results strongly suggest the complementary roles of patent citations and patent contents when used in patent classification tasks. In our experiments, the bibliographic coupling kernel (K_Bib) and the graph overlap kernel (K_Ovr) achieved only 7.48% and 37.13% accuracy, respectively. However, when they were combined with the linear text kernel, the classification performance improved significantly. This indicates that even though the

citation information may be sparse in patents and using it alone is not very helpful, combining citation and content information can still improve the performance for patent classification.

5.3 Individual Class's Performances

We also inspected the kernels' performances on all 36 classes. Figure 5 shows the F-measure each kernel achieved in each class. In general, the labeled citation graph kernel (K_{Gra}) and its composition with the linear text kernel (K_{Com_4}) have high performance in most of the 36 categories. However, the F-measures of the other seven kernels vary in the 36 categories, which may reduce their usability. We observe that the seven kernels' F-measures are relatively low in a similar group of categories. Table 5 provides some examples of these categories, which are difficult to classify and have a relatively small number of training instances. Our testbed includes other categories which share similar topics with these categories and have a larger number of training instances. The classifiers have a high probability of misclassifying patents belonging to these categories into other similar categories. For example, most of the instances in USPC category #216 (Etching a substrate: processes) were incorrectly assigned to category #438 (Semiconductor device manufacturing: process), which has 1,119 training instances. Many of the instances in category #264 (Plastic and nonmetallic article shaping or treating: processes) were assigned to category #428 (Stock material or miscellaneous articles), which has 774 training instances. Many of the instances in categories #422 (Chemical apparatus and process disinfecting, deodorizing, pre-serving, or sterilizing), #436 (Chemistry: analytical and immunological testing), #530 (Chemistry: natural resins or derivatives; peptides or proteins; lignins or reaction products thereof), and #536

(Organic compounds -- part of the class 532-570 series) were assigned to #435 (Chemistry: molecular biology and microbiology), which has 1,895 training instances. Even in these categories where most other kernels fail, the labeled citation graph kernel and its composition with the linear text kernel (K_{Com_4}) are highly accurate. By considering patent citation information, the two kernels have better differentiation abilities on the categories with very similar topics and uneven numbers of instances.

[Figure 5. The kernels' performances in different classes]

The performance of the labeled citation graph kernel (K_{Gra}) and its composition with the linear text kernel (K_{Com_4}) also changes slightly in different categories. In Figure 5, the labeled citation graph kernel (K_{Gra}) does not achieve a high performance in USPC category #435 (F-measure=54.71%). Although it is better than most of the other kernels in the same category, such a performance is not comparable to the performance it achieved in other categories (F-measures between 77.78% and 98.31%). USPC category #435 has the largest number of training instances in the dataset and a small number of testing instances. The patents in this category are on fundamental science topics or research tools, which were heavily cited by patents in all categories [34]. These characteristics may be the cause of the low performance of the labeled citation graph kernel and other kernels in USPC category #435. However, after combining it with the linear text kernel, the composite kernel (K_{Com_4}) achieves a high F-measure on USPC category #435 (81.25%). The composite kernel (K_{Com_4}) employs content features in addition to citations, which may help the classifiers differentiate the patents belonging to USPC category #435 from the others. Actually, the composite kernel (K_{Com_4}) achieves consistent good performance in all categories (F-measures between 74.42% and 96.73%,

average F-measure 87.96%, standard deviation 6.53%). Even the labeled citation graph kernel is highly accurate; considering patent contents (linear text kernel) ensures more consistent high performance for different categories.

[Table 5. Some of the categories which are difficult to classify]

6. Conclusions

Using patent classification as an example, this paper demonstrates that knowledge evolution processes can be useful in knowledge management tasks. In this research, we utilized patent citation networks, which encode evolution processes of innovations, for the classification of patent documents. Under a kernel-based framework, we designed different kernel functions to capture the information of citation networks and found that our proposed labeled citation graph kernel improved patent classification performance. Our research showed that the features of cited patents and the structure of patent citation networks, which together represent innovations' evolution history, can benefit the classification of focal patents. We also noticed that combining the information in citation networks with patent contents results in higher and more consistent performance.

In the practice of patent management, the significant performance improvement (>30% in accuracy) in our experiments indicates the good potential of using our approach in real-life patent examination applications. Previous content-based methods usually cannot match the performance of junior examiners with some basic general knowledge [30]. Our proposed approach can potentially alleviate human efforts in patent pre-classification and further expedite patent examination. Our research also lends support to a policy that requires inventors to file patent citations, since they often have the more complete knowledge about their innovation's evolution. The USPTO has adopted this policy,

which may need to be considered by other patent offices.

The effectiveness of our proposed approach implies a wider application area than patent classification. The proposed approach can be directly applied to classify other linked documents, such as Webpages and scientific literature. With appropriate adaptations it is also applicable to other knowledge codification and organization tasks such as building help desk systems, decision support systems, and knowledge repositories.

In the future, we will continue theorizing the role of knowledge evolution processes in KM and studying its applications in other types of KM tasks, such as the knowledge acquisition tasks in opinion mining and topic suggestion. For patent classification, we will extend this research to more realistic settings where multiple labels on hierarchical schemes are used to codify patents.

Acknowledgements

This research is supported by the NSF: IIS-0311652 “Intelligent Patent Analysis for Nanoscale Science and Engineering” and DMI-0533749 “NanoMap: Mapping Nanotechnology Development.” We thank the USPTO for making their data available for research purposes.

References

1. Almeida, P. and Kogut, B. Localization of knowledge and the mobility of engineers in regional networks. *Management Science*, 45, 7 (1999), 905-917.
2. Amsler, R. *Application of Citation-based Automatic Classification*. Austin, TX: University of Texas at Austin, Linguistics Research Center, 1972.

3. Bieber, M.; Engelbart, D.; Furuta, R.; Hiltz, S.R.; Noll, J.; Preece, J.; Stohr, E.A.; Turoff, M. and Van de Walle, B. Toward virtual community knowledge evolution. *Journal of Management Information Systems*, 18, 4 (2002), 11-35.
4. Borgwardt, K.M.; Ong, C.S.; Schonauer, S.; Vishwanathan, S.V.N.; Smola, A.J. and Kriegel, H.P. Protein function prediction via graph kernels. *Bioinformatics*, 21 (2005), I47-I56.
5. Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A. and Wiener, J. Graph structure in the Web. *Computer Networks-the International Journal of Computer and Telecommunications Networking*, 33, 1-6 (2000), 309-320.
6. Calado, P.; Cristo, M.; Goncalves, M.A.; de Moura, E.S.; Ribeiro-Neto, B. and Ziviani, N. Link-based similarity measures for the classification of Web documents. *Journal of the American Society for Information Science and Technology*, 57, 2 (2006), 208-221.
7. Chakrabarti, S.; Dom, B. and Indyk, P. Enhanced hypertext categorization using hyperlinks. *1998 ACM SIGMOD International Conference on Management of Data*. Seattle, WA, 1998, pp. 307-318.
8. Chang, C.-C. and Lin, C.-J. *LIBSVM: A Library for Support Vector Machines*. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2001.
9. Craven, M. and Slattery, S. Relational learning with statistical predicate invention: Better models for hypertext. *Machine Learning*, 43, 1-2 (2001), 97-119.

10. Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines (and Other Kernel-based Learning Methods)*. Cambridge: Cambridge University Press, 2000.
11. Cristo, M.; Calado, P.; de Moura, E.S.; Ziviani, N. and Ribeiro-Neto, B. Link information as a similarity measure in web classification. *International Symposium on String Processing and Information Retrieval*. 2003, pp. 43-55.
12. Dunford, R. Key challenges in the search for the effective management of knowledge in management consulting firms. *Journal of Knowledge Management*, 4, 4 (2000), 295-302.
13. Fall, C.J.; Torcsvari, A.; Benzineb, K. and Karetka, G. Automated categorization in the International patent classification. *ACM SIGIR Forum*, 37, 1 (2003), 10-25.
14. Fall, C.J.; Torcsvari, A.; Fievet, P. and Karetka, G. Automated categorization of German-language patent documents. *Expert Systems with Applications*, 26, 2 (2004), 269-277.
15. Fleming, L. Recombinant uncertainty in technological search. *Management Science*, 47, 1 (2001), 117-132.
16. Gallini, N.T. The economics of patents: Lessons from recent US patent reform. *Journal of Economic Perspectives*, 16, 2 (2002), 131-154.
17. Gartner, T. A survey of kernels for structured data. *ACM SIGKDD Explorations*, 5, 1 (2003), 49-58.
18. Ghani, R.; Slattery, S. and Yang, Y. Hypertext categorization using hyperlink patterns and meta data. *The 18th International Conference on Machine Learning*. 2001, pp. 178-185.

19. Ginsparg, P.; Houle, P.; Joachims, T. and Sul, J.H. Mapping subsets of scholarly information. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (2004), 5236-5240.
20. Haussler, D. *Convolution kernels on discrete structures*, Technical Report UCS-CRL-99-10. UC Santa Cruz, 1999.
21. Huang, M.H.; Wang, E.T.G. and Seidmann, A. Price mechanism for knowledge transfer: An integrative theory. *Journal of Management Information Systems*, 24, 3 (2007), 79-108.
22. Huang, Z.; Chen, H.; Yip, A.; Ng, G.; Guo, F.; Chen, Z.-K. and Roco, M.C. Longitudinal patent analysis for Nanoscale Science and Engineering: Country, institution and technology field. *Journal of Nanoparticle Research*, 5 (2003), 333-363.
23. Hull, D.; Ait-Mokhtar, S.; Chuat, M.; Eisele, A.; Gaussier, E.; Grefenstette, G.; Isabelle, P.; Samuelsson, C. and Segond, F. Language technologies and patent search and classification. *World Patent Information*, 21, 3 (2001), 265-268.
24. Hunt, R.M. You can patent that? Are patents on computer programs and business methods good for the economy? *Federal Reserve Bank of Philadelphia Business Review*, Q1 (2001), 5-15.
25. Joachims, T.; Cristianini, N. and Shawe-Taylor, J. Composite kernels for hypertext categorisation. *The 18th International Conference on Machine Learning*. 2001, pp. 250-257.
26. Kashima, H.; Tsuda, K. and Inokuchi, A. Marginalized kernels between labeled graphs. *The 20th International Conference on Machine Learning*. 2003, pp.

27. Kessler, M.M. Bibliographic coupling between scientific papers. *American Documentation*, 14, 1 (1963), 10-&.
28. King, J.L. Patent examination procedures and patent quality, In, Cohen, W.M. and Merrill, S.A. (eds.) *Patents in the Knowledge-Based Economy*. Washington, D.C.: National Academies Press, 2003, pp. 54-73.
29. Koster, C.H.A.; Seutter, M. and Beney, J. Multi-classification of patent applications with Winnow. *Perspectives of System Informatics*, 2890 (2003), 546-555.
30. Krier, M. and Zacca, F. Automatic categorisation applications at the European patent office. *World Patent Information*, 24, 3 (2002), 187-196.
31. Lanckriet, G.R.G.; De Bie, T.; Cristianini, N.; Jordan, M.I. and Noble, W.S. A statistical framework for genomic data fusion. *Bioinformatics*, 20, 16 (2004), 2626-2635.
32. Larkey, L.S. A patent search and classification system. *The 4th ACM Conference on Digital Libraries*. Berkeley, CA, 1999, pp. 79-87.
33. Le, S.Q.; Ho, T.B. and Phan, T.T.H. A novel graph-based similarity measure for 2D chemical structures. *Genome Informatics*, 14, 2 (2004), 82-91.
34. Li, X.; Chen, H.; Huang, Z. and Roco, M.C. Patent citation network in nanotechnology (1976-2004). *Journal of Nanoparticle Research*, 9, 3 (2007), 337-352.
35. Loh, H.T.; He, C. and Shen, L. Automatic classification of patent documents for TRIZ users. *World Patent Information*, 28, 1 (2006), 6-13.
36. McCallum, A.K. *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering*. <http://www.cs.cmu.edu/~mccallum/bow>, 1996.

37. Narin, F. Patent bibliometrics. *Scientometrics*, 30, 1 (1994), 147-155.
38. Nerkar, A. Old is gold? The value of temporal exploration in the creation of new knowledge. *Management Science*, 49, 2 (2003), 211-229.
39. Nidumolu, S.R.; Subramani, M. and Aldrich, A. Situated learning and the situated knowledge web: Exploring the ground beneath knowledge management. *Journal of Management Information Systems*, 18, 1 (2001), 115-150.
40. Oh, H.-J.; Myaeng, S.H. and Lee, M.-H. A practical hypertext categorization method using links and incrementally available class information. *The 23rd ACM International Conference on Research and Development in Information Retrieval*. 2000, pp. 264-271.
41. Redner, S. How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B*, 4, 2 (1998), 131-134.
42. Richter, G. and MacFarlane, A. The impact of metadata on the accuracy of automated patent classification. *World Patent Information*, 27, 1 (2005), 13-26.
43. Scherer, F.M. The economics of human gene patents. *Academic Medicine*, 77, 12 (2002), 1348-1367.
44. Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys*, 34, 1 (2002), 1-47.
45. Sinclair, G. and Webber, B. Classification from full text: A comparison of canonical sections of scientific papers. *The 20th International Conference on Computational Linguistics*. 2004, pp. 69-72.
46. Singh, J. Collaborative networks as determinants of knowledge diffusion patterns. *Management Science*, 51, 5 (2005), 756-770.

47. Small, H. Co-citation in scientific literature - New measure of relationship between 2 documents. *Current Contents*, 7 (1974), 7-10.
48. Smith, H. Automation of patent classification. *World Patent Information*, 24, 4 (2002), 269-271.
49. Spangler, S.; Kreulen, J.T. and Lessler, J. Generating and browsing multiple taxonomies over a document collection. *Journal of Management Information Systems*, 19, 4 (2003), 191-212.
50. Stenmark, D. Leveraging tacit organizational knowledge. *Journal of Management Information Systems*, 17, 3 (2000), 9-24.
51. Tan, Y. and Wang, J. A support vector machine with a hybrid kernel and minimal Vapnik-Chervonenkis dimension. *IEEE Transactions on Knowledge and Data Engineering*, 16, 4 (2004), 385-395.
52. Taskar, B.; Abbeel, P. and Koller, D. Discriminative probabilistic models of relational data. *The 18th Conference on Uncertainty in Artificial Intelligence*. 2002, pp. 485-492.
53. Teichert, T. and Mittermayer, M.-A. Text mining for technology monitoring. *IEEE International Engineering Management Conference 2002*. 2002, pp. 596-601.
54. USPTO. *U.S. Patent Statistics Chart Calendar Years 1963 - 2005*.
http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm, 2005.
55. Yang, Y. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1 (1999), 69-90.
56. Yang, Y.M.; Slattery, S. and Ghani, R. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 18, 2-3 (2002), 219-241.

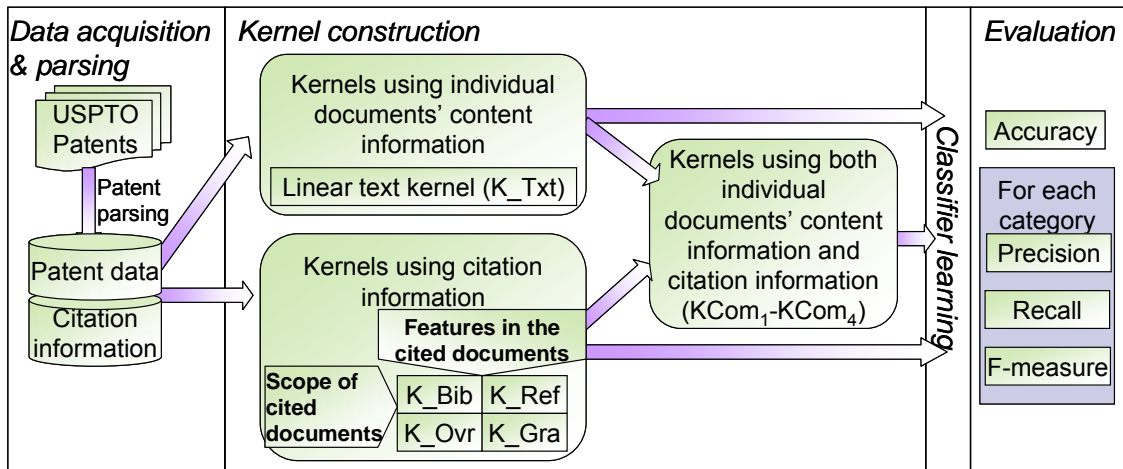


Figure 1. A framework of kernel-based patent classification

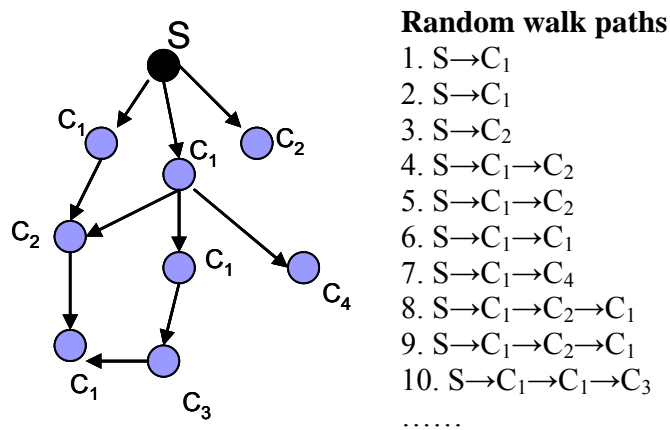


Figure 2. Random walk paths on a labeled citation network related to patent S

1. Random path generation

- 1) The random walk starts from the patent to be classified x_0 .
- 2) On node x_i the random walk has a probability of $p_q(x_i)$ to stop.
- 3) If the random walk does not stop, the random walk has equal probability to choose any of x_i 's neighbors (which is noted as x_{i+1}) to jump to. The probability is noted as $p_t(x_{i+1} | x_i)$.

- 4) Thus a random walk path $h = \{x_0, x_1, \dots, x_n\}$ has the probability

$$P(h | G) = p_t(x_1 | x_0) p_t(x_2 | x_1) \dots p_t(x_n | x_{n-1}) p_q(x_n) \text{ to exist.}$$

2. Kernel definition

The labeled citation graph kernel is defined as a convolution kernel

$$K_{Gra}(G_{p_1}, G_{p_2}) = \sum_h \sum_{h'} k(h, h') P(h | G_{p_1}) P(h' | G_{p_2})$$

For two random walk paths $h = \{x_0, x_1, \dots, x_n\}$ and $h' = \{x_0', x_1', \dots, x_m'\}$

$$\text{if } n \neq m, \quad k(h, h') = 0, \quad \text{else } k(h, h') = \prod_{i=1}^n \hat{k}(x_i, x_i') \quad \text{where } \hat{k}(x_i, x_i') = 1 \quad \text{iff}$$

$$\text{label}(x_i) = \text{label}(x_i').$$

Figure 3. The algorithm for labeled citation graph kernels

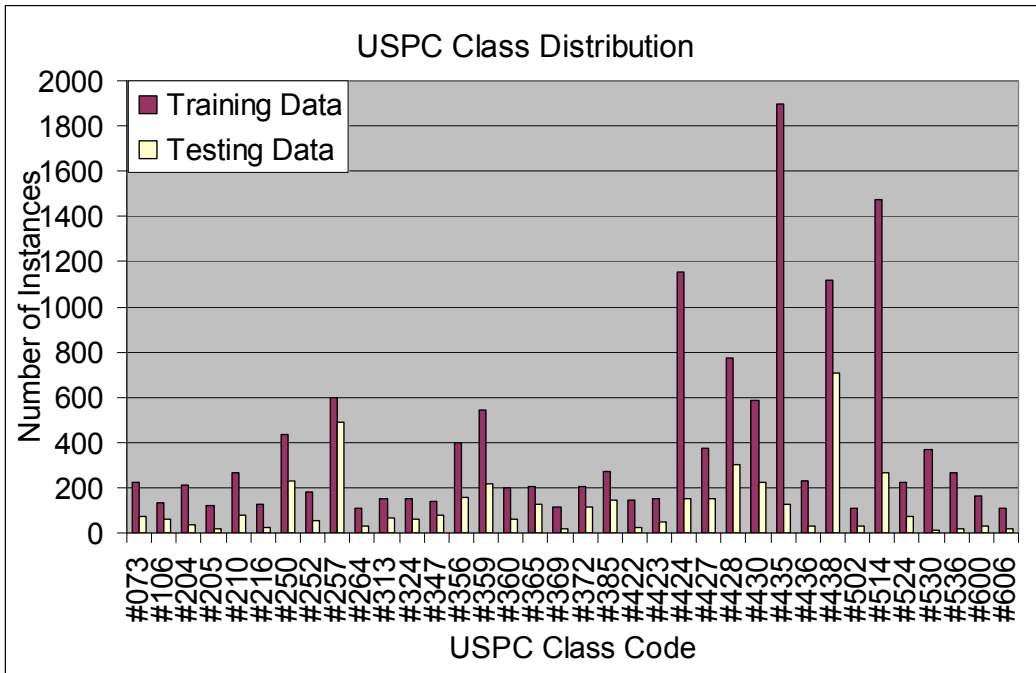


Figure 4. Patent distribution in USPC categories

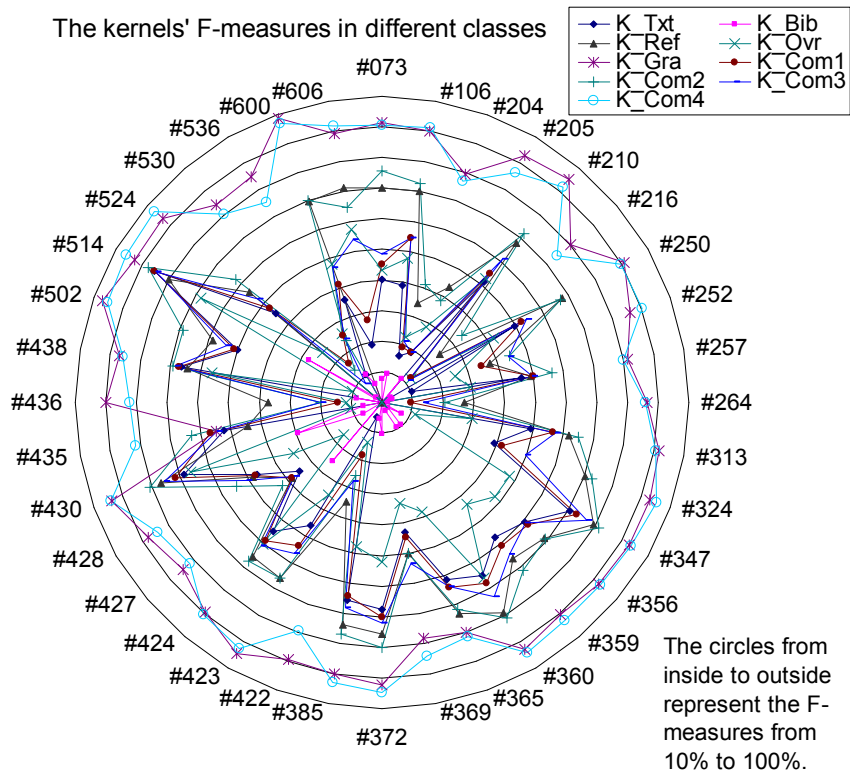


Figure 5. The kernels' performances in different classes

Table 1. Kernels for citation information

	No cited documents' features	Using cited documents' features
Direct citations	Bibliographic coupling kernel (K_{Bib})	Labeled co-reference kernel (K_{Ref})
Citation network	Graph overlap kernel (K_{Ovr})	Labeled citation graph kernel (K_{Gra})

Table 2. Number of data instances in the testing and training datasets

	<i># of patents</i>	<i># of categories</i>	<i># of patents in the citation network</i>	<i># of categories in the citation network</i>
Training	13,913	36	336,303	426
Testing	4,358	36	227,833	410

Table 3. Performances of different kernels

<i>Kernels</i>	<i>Accuracy</i>	<i>Averaged precision</i>	<i>Averaged recall</i>	<i>Averaged F-measure</i>
Bibliographic coupling kernel (K_{Bib})	7.48%	47.87%	5.81%	5.71%
Labeled co-reference kernel (K_{Ref})	61.50%	56.04%	56.82%	55.50%
Graph overlap kernel (K_{Ovr})	37.13%	53.32%	29.08%	34.91%
Labeled citation graph kernel (K_{Gra})	86.67%	89.09%	87.97%	88.04%
Composite kernel 1 (K_{Com_1})	57.82%	53.65%	44.24%	46.50%
Composite kernel 2 (K_{Com_2})	66.02%	59.43%	59.14%	58.78%
Composite kernel 3 (K_{Com_3})	59.64%	55.49%	47.56%	49.72%
Composite kernel 4 (K_{Com_4})	87.84%	89.43%	86.97%	87.96%
Linear Text Kernel (K_{Txt})	55.55%	51.65%	39.29%	40.83%

Table 4. Hypotheses testing for different kernels

<i>H1.1: p values</i>	<i>a) Pairwise t-test on accuracy</i>	<i>b) Pairwise t-test on F-measure</i>
$K_Bib < K_Ovr$	<0.001	<0.001
$K_Ref < K_Gra$	<0.001	<0.001

<i>H1.2: p values</i>	<i>a) Pairwise t-test on accuracy</i>	<i>b) Pairwise t-test on F-measure</i>
$K_Bib < K_Ref$	<0.001	<0.001
$K_Ovr < K_Gra$	<0.001	<0.001

<i>H2.1: p values</i>	<i>a) Pairwise t-test on accuracy</i>	<i>b) Pairwise t-test on F-measure</i>
$K_Txt < K_Com_1$	<0.001	<0.001
$K_Txt < K_Com_2$	<0.001	<0.001
$K_Txt < K_Com_3$	<0.001	<0.001
$K_Txt < K_Com_4$	<0.001	<0.001

<i>H2.2: p values</i>	<i>a) Pairwise t-test on accuracy</i>	<i>b) Pairwise t-test on F-measure</i>
$K_Bib < K_Com_1$	<0.001	<0.001
$K_Ref < K_Com_2$	<0.001	<0.005
$K_Ovr < K_Com_3$	<0.001	<0.001
$K_Gra < K_Com_4$	0.004	0.533

Table 5. Some of the categories that are difficult to classify

<i>USPC code</i>	<i>Category description</i>	<i># of training instances</i>	<i># of testing instances</i>
#216	Etching a substrate: processes	124	23
#264	Plastic and nonmetallic article shaping or treating: processes	111	33
#422	Chemical apparatus and process disinfecting, deodorizing, preserving, or sterilizing	143	23
#436	Chemistry: analytical and immunological testing	229	28
#530	Chemistry: natural resins or derivatives; peptides or proteins; lignins or reaction products thereof	367	15
#536	Organic compounds -- part of the class 532-570 series	265	18