Patent Mining: A Survey

Longhui Zhang
lzhan015@cs.fiu.edu

Lei Li
lli003@cs.fiu.edu

Tao Li
taoli@cs.fiu.edu

School of Computing and Information Sciences
Florida International University
Miami, FL 33199

ABSTRACT

Patent documents are important intellectual resources of protecting interests of individuals, organizations and companies. Different from general web documents, patent documents have a well-defined format including frontpage, description, claims, and figures. However, they are lengthy and rich in technical terms, which requires enormous human efforts for analysis. Hence, a new research area, called patent mining, emerges in recent years, aiming to assist patent analysts in investigating, processing, and analyzing patent documents. Despite the recent advances in patent mining, it is still far from being well explored in research communities. To help patent analysts and interested readers obtain a big picture of patent mining, we thus provide a systematic summary of existing research efforts along this direction. In this survey, we first present an overview of the technical trend in patent mining. We then investigate multiple research questions related to patent documents, including patent retrieval, patent classification, and patent visualization, and provide summaries and highlights for each question by delving into the corresponding research efforts.

Keywords

Patent Mining; Patent Information Retrieval; Patent Classification; Patent Visualization; Patent Valuation; Cross-Language Patent Mining; Patent Application

1. INTRODUCTION

Patent application is one of the key aspects of protecting intellectual properties. In the past decades, with the advanced development of various techniques in different application domains, a myriad of patent documents are filed and be approved. They serve as one of the important intellectual property components for individuals, organizations and companies. These patent documents are open to public and made available by various authorities in a lot of countries or regions around the world. For example, World Intellectual Property Organization (WIPO)[1] reported 1.98 million total patent applications filed worldwide in 2010.

Patent documents have great research values, beneficial to the industry, business, law, and policy-making communities. If patent documents are carefully analyzed, important technical details and relations can be revealed, leading business trends can be illustrated, novel industrial solutions can be inspired, and consequently vital investment decisions can be made [15]. Thus, it is imperative to carefully analyze patent documents for evaluating and maintaining patent values. In recent years, patent analysis has been recognized as an important task at the government level. Public patent authorities[2] in United States, United Kingdom, China and Japan have invested various resources to improve the performances of creating valuable patent analysis results for various patent analysis tasks.

However, patent analysis is a non-trivial task, which often requires tremendous amount of human efforts. In general, it is necessary for patent analysts to have a certain degree of expertise in different research domains, including information retrieval, data mining, domain-specific technologies, and business intelligence. In reality, it is difficult to find and train such analysts to match those multi-disciplinary requirements within a relatively short period of time. Another challenge of patent analysis is that patent documents are often lengthy, and full of technical and legal terminologies. Even for domain experts, it may also require a lot of time to read and analyze a single patent document. Therefore, patent mining plays an important role in automatically processing and analyzing patent documents [106; 127].

A patent document often contains dozens of items that can be grouped into two categories: (1) structured items, which are uniform in semantics and format (such as patent number, inventor, filing date, issued date, and assignees); and (2) unstructured items, which consist of text content in different length (including claims, abstracts, and descriptions of the invention.). Given such a well-defined structure, patent documents are considerably different from general web documents (e.g., web pages), most of which contain unstructured data, involving free texts, links, tags, images, and videos. Hence, the analysis of patent documents might be different from the one for web documents in terms of the format and various application-wise purposes.

In this survey, we comprehensively investigate multiple critical research questions in the domain of patent mining, including (1) how to effectively retrieve patent documents based on user-defined queries (See Section 3)? (2) how to efficiently perform patent classification for high-quality maintenance (See Section 4)? (3) how to informatively represent patent documents to users (See Section 5)? (4) how to explore and evaluate the potential benefit of patent documents (See Section 6)? and (5) how to effectively deal with cross-language patent documents (See Section 7)? For

[1] http://www.wipo.int/ipstats/en/general_info.html.

[2] http://www.wipo.int/directory/en/urls.jsp.

Table 1: Representative patent mining tasks and approaches.

Tasks	Techniques	References
Patent Retrieval (See Section 3)	Query Generation	[6; 7; 10; 17; 50; 76; 78; 79; 104; 108; 114; 118; 119]
	Query Expansion	[2; 9; 25; 28; 29; 30; 31; 34; 40; 43; 52; 68] [69; 72; 73; 74; 78; 83; 96; 98; 99; 107; 114]
Patent Classification (See Section 4)	Using Different Resources	[4; 33; 49; 56; 58; 59; 66; 86; 101]
	Using Different Classifier	[13; 19; 23; 24; 33; 103; 116]
Patent Visualization (See Section 5)	Structured Data Visualization	[42; 93; 97; 120; 122; 123]
	Unstructured Text Visualization	[5; 39; 61; 105; 124]
	Hybrid Visualization	[16; 51; 63; 80; 94; 97; 121; 124]
Patent Valuation (See Section 6)	Unsupervised Exploration	[3; 21; 45; 46; 60; 67; 81; 84; 109; 111]
	Supervised Evaluation	[21; 41; 46; 67; 85; 109]
Cross-Language Mining (See Section 7)	Machine Translation	[18; 26; 27; 32; 48; 53; 70; 77]
	Semantic Correspondence	[54; 62; 64; 47; 102; 110]

each question, we first identify several critical research challenges, and then discuss different research efforts and various techniques used for addressing these challenges. Table 1 summarizes different patent mining tasks, including patent retrieval, patent classification, patent visualization, patent exploration, and cross-language patent mining. Up-to-date references/lists related to patent mining can be found at http://users.cis.fiu.edu/∼lzhan015/patmining.html. In the following sections, we will briefly introduce the existing solutions to each task based on the techniques being utilized.

The rest of the paper is organized as follows. In§ 2, we provide an introduction to patent documents by describing patent document structures, patent classification systems, and various patent mining tasks. Section 3 presents a summary of research efforts for addressing patent retrieval, especially, patent search. In Section 4, we investigate how patent documents can be automatically classified into different predefined categories. In Section 5, we explore how patent documents can be represented to analysts in a way that the core ideas of patents can be clearly illustrated and the correlations of different documents can be easily identified. In Section 6, we show that the quality of a patent document can be automatically evaluated based on some predefined measurements that help companies decide which patent is more important and should be further maintained for effective property protection. In Section 7, we present different techniques for cross-language patent mining, including approaches to solving machine translation and semantic correspondence. Section 8 discusses existing free and commercial patent mining systems that provide various functionalities to allow patent analysts to perform different patent mining tasks. Finally, Section 9 concludes our survey and discusses emerging research- and application-wise challenges in the domain of patent mining.

2. BACKGROUND

In this section, we first provide a brief overview of patent documents and their structure, and then describe the current patent classification systems, followed by introducing the tasks in the entire process of patent application.

2.1 The Structure of Patent Documents

According to World Intellectual Property Organization[3], the definition of a patent is: "*patents are legal documents issued by a government that grants a set of rights of exclusivity and protection to the owner of an invention. The right of exclusivity allows the patent owner to exclude others from making, using, selling, offering for sale, or importing the patented invention during the patent term, typically period from the earliest filing date, and in the country or countries where patent protection exists.*" Based upon the understanding of the definition, patent documents are one of the key components that serve to protect the intellectual properties of patent owners. Note that patents and inventions are two different yet interleaved concepts: patents are legal documents, whereas inventions are the content of patents. Different countries or regions may have their own patent laws and regulations, but in general there are two common types of patent documents: utility patents and design patents. Utility patents describe technical solutions related to a product, a process, or a useful improvement, etc., whereas design patents often represent original designs related to the specifications of a product. In practice, due to the distinct properties of these two types of patents, the structure of patent document may vary slightly; however, a typical patent document often contains several requisite sections, including a front page, detailed specifications, claims, declaration, and/or a list of drawings to illustrate the idea of the solution.

Figure 1 shows an example of the front page of a patent document. In general, a `frontpage` contains four parts, described as follows:

1. `Announcement`, which includes Authority Name (e.g. United States Patent), Patent No., and Date of Patent (i.e., patent publication date).;

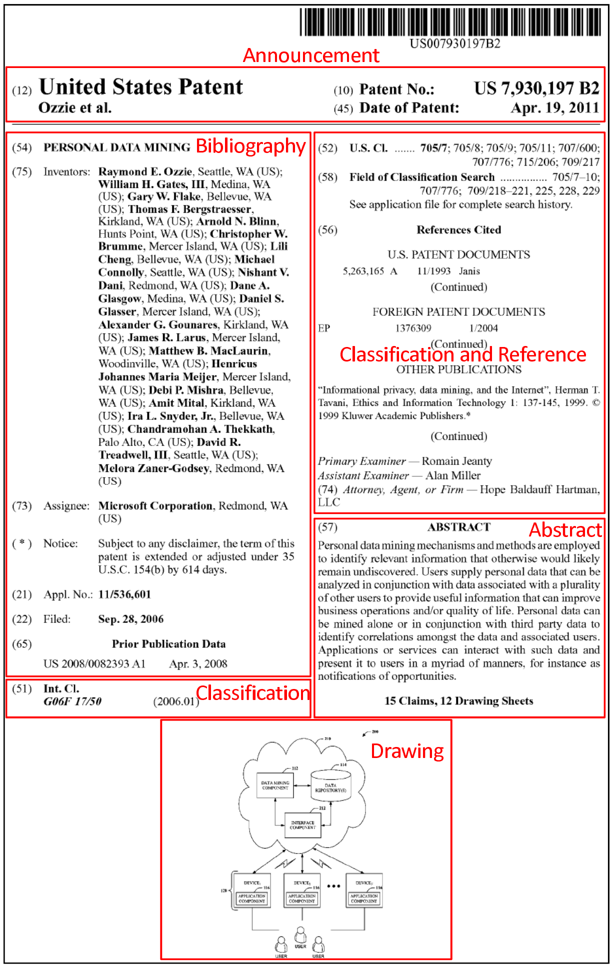
2. `Bibliography`, which often includes Title, Inventors, Assignee, Application No., and Date of filing.;

3. `Classification` and `Reference`, which include International Patent Classification Code, Region-based Classification Code (e.g., United State Classification Code), and/or other patent classification categories, along with references assigned by the examiner;

4. `Abstract`, which may contain a short description of the invention and sometimes a drawing that is the most representative one in terms of illustrating the general idea of the invention.

Beside the front page, a patent document contains detailed description of the solution, claims, and/or a list of draw-

[3] http://www.wipo.int.



US007930197B2

(12) **United States Patent**
Ozzie et al.

(10) **Patent No.:** **US 7,930,197 B2**
(45) **Date of Patent:** **Apr. 19, 2011**

(54) **PERSONAL DATA MINING**

(75) Inventors: **Raymond E. Ozzie**, Seattle, WA (US); **William H. Gates, III**, Medina, WA (US); **Gary W. Flake**, Bellevue, WA (US); **Thomas F. Bergstraesser**, Kirkland, WA (US); **Arnold N. Blinn**, Hunts Point, WA (US); **Christopher W. Brumme**, Mercer Island, WA (US); **Lili Cheng**, Bellevue, WA (US); **Michael Connolly**, Seattle, WA (US); **Nishant V. Dani**, Redmond, WA (US); **Dane A. Glasgow**, Medina, WA (US); **Daniel S. Glasser**, Mercer Island, WA (US); **Alexander G. Gounares**, Kirkland, WA (US); **James R. Larus**, Mercer Island, WA (US); **Matthew B. MacLaurin**, Woodinville, WA (US); **Henricus Johannes Maria Meijer**, Mercer Island, WA (US); **Debi P. Mishra**, Bellevue, WA (US); **Amit Mital**, Kirkland, WA (US); **Ira L. Snyder, Jr.**, Bellevue, WA (US); **Chandramohan A. Thekkath**, Palo Alto, CA (US); **David R. Treadwell, III**, Seattle, WA (US); **Melora Zaner-Godsey**, Redmond, WA (US)

(73) Assignee: **Microsoft Corporation**, Redmond, WA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 614 days.

(21) Appl. No.: **11/536,601**

(22) Filed: **Sep. 28, 2006**

(65) **Prior Publication Data**

US 2008/0082393 A1 Apr. 3, 2008

(51) **Int. Cl.**
G06F 17/50 (2006.01)

(52) **U.S. Cl.** **705/7**; 705/8; 705/9; 705/11; 707/600; 707/776; 715/206; 709/217

(58) **Field of Classification Search** 705/7–10; 707/776; 709/218–221, 225, 228, 229
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,263,165 A 11/1993 Janis

(Continued)

FOREIGN PATENT DOCUMENTS

EP 1376309 1/2004

(Continued)

OTHER PUBLICATIONS

"Informational privacy, data mining, and the Internet", Herman T. Tavani, Ethics and Information Technology 1: 137-145, 1999. © 1999 Kluwer Academic Publishers.*

(Continued)

Primary Examiner — Romain Jeanty
Assistant Examiner — Alan Miller
(74) *Attorney, Agent, or Firm* — Hope Baldauff Hartman, LLC

(57) **ABSTRACT**

Personal data mining mechanisms and methods are employed to identify relevant information that otherwise would likely remain undiscovered. Users supply personal data that can be analyzed in conjunction with data associated with a plurality of other users to provide useful information that can improve business operations and/or quality of life. Personal data can be mined alone or in conjunction with third party data to identify correlations amongst the data and associated users. Applications or services can interact with such data and present it to users in a myriad of manners, for instance as notifications of opportunities.

15 Claims, 12 Drawing Sheets

Figure 1: Front page of a patent document.

ings. The description section, in general, depicts the background and summary of the invention, brief description of the drawings, and detailed description of preferred embodiments. The claim section is the primary component of a patent document, which defines the scope of protection conveyed by the invention. It often contains two types of claims: (1) the independent claim which stands on itself; and (2) the dependent claims which refer to its antecedent claim.

A patent document is often lengthy, compared with other types of documents, e.g., web pages. Although the structure of a patent document is well-defined, a myriad of obscure and ambiguous text snippets are often involved, and various technical terms are often used in the content, which render the analysis of patent document more difficult.

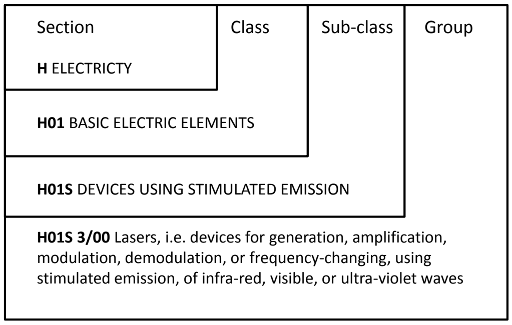
2.2 Patent Classification Criteria

Before the publication of patent applications, one or more classification codes are often assigned to patent documents based on their textual contents for the purpose of efficient management and retrieval. Different patent authorities may maintain their own classification hierarchies, such as the United States Patent Classification (USPC) in the United States, the International Patent Classification (IPC) for the World Intellectual Property Organization, and the Derwent classification system fixed by Thomson Reuters. In the following, we will introduce the classification taxonomies of IPC and USPC in more details.

2.2.1 IPC Taxonomy

IPC was established in 1971 based on Patent Cooperation Treaty [22]. This hierarchical patent classification system categorizes patents to different technological groups. There are over 100 countries using IPC system to classify their national patent applications. Specifically, the IPC category taxonomy contains 8 sections, 120 classes, 630 subclasses, 7,200 main groups and approximately 70,000 sub-groups. A typical IPC category contains a class label and a piece of text description to indicate the specific category content.

In IPC, all technological fields are first grouped into 8 sections represented by one of the capital letters from A to H[4], including (A) "Human necessities"; (B) "Performing operations, transporting"; (C) "Chemistry, metallurgy"; (D) "Textiles, paper"; (E) "Fixed constructions"; (F) "Mechanical engineering, lighting, heating, weapons, blasting"; (G) "Physics"; and (H) "Electricity". Then, within each section, the technological fields are regrouped into classes as the second level of the IPC taxonomy. Each class consists of one or more subclasses, which are treated as the third level of the taxonomy. Finally, each subclass is further divided into subdivisions referred to as "groups". As an illustrative example, Figure 2 describes the class label "H01S 3/00" and its ancestors.



Section
Class
Sub-class
Group

H ELECTRICTY

H01 BASIC ELECTRIC ELEMENTS

H01S DEVICES USING STIMULATED EMISSION

H01S 3/00 Lasers, i.e. devices for generation, amplification, modulation, demodulation, or frequency-changing, using stimulated emission, of infra-red, visible, or ultra-violet waves

Figure 2: An example of IPC.

2.2.2 USPC Taxonomy

The USPC system was developed in 1836, which is the first patent taxonomy established in the world [88]. In USPC, the patent categories are organized as a two-level taxonomy, i.e., class and subclass. Each class has a designated class number, and includes a descriptive title, class schedule, and definitions. Then each class is subdivided into a number of subclasses. A subclass has a number, a title, an indent level indicated by one or more dots, a definition, a hierarchical relationship to other subclasses in a class, and relationships to other subclasses in other classes. A subclass is the smallest searchable group of patents in USPC.

2.3 Tasks in Patent Analysis and Investigation

Based upon the filing status of a patent document, a patent mining system can be decomposed into two modules: (1) *Pre-filing* module, in which the patent documents are carefully examined to ensure the non-infringement; and (2) *Post-*

[4] http://www.wipo.int/classifications/ipc/en.

filing module, in which patent documents are maintained and analyzed. The general architecture of a patent mining system is depicted in Figure 3.

During the *pre-filing* process, or say, the application process, there are two major tasks:

1. Classifying the patent application into multiple predefined categories (e.g., IPC and USPC). This task aims to not only restrict the searching scope, but also ease the maintenance of patent applications/documents.

2. Searching all relevance patent documents from patent databases and non-patent documents from online resources. The primary goal of this task is to examine the infringement/patentability, and assigning a list of appropriate references for better understanding the idea of the patent application.

Currently in most intellectual property authorities and/or patent law firms, these two tasks are often being conducted manually. In practice, these two tasks, especially the latter one, may require specific domain expertise and a huge amount of time/human efforts.

The major focus of the *post-filing* process is to maintain and analyze patent documents in order to provide fully functional support to various types of enterprises. For example, a company plans to develop a new product. Prior to the design/implementation of this product, it is essential to determine what related products have already been produced and patented. Therefore, a typical task is to perform a comprehensive investigation towards the related domain/products by virtue of patent search. By doing this, the company is able to obtain an overview of the general technologies applied in the corresponding domain, as well as the technical details of relevant products. In general, in the process of *post-filing*, besides the task of patent search, three additional tasks are often involved:

1. Patent visualization, which aims to represent patent documents to help patent analysts easily understand the core idea of patents;

2. Patent valuation, which explores patent documents in different ways to evaluate their value, potential, impact, etc.;

3. Cross-language mining, which localizes patent information from patent documents that are described by multiple languages.

However, due to the large volume of patent files and diverse writing styles of patent applications, these processes are time-consuming, and often require a lot of human efforts for patent reading and analysis. The ultimate goal of these efforts is to provide automatic tools to ease the procedure of patent analysis. In the following sections, we will introduce the existing academic/industrial efforts in designing patent mining algorithms and building patent mining applications using the architecture shown in Figure 3.

3. PATENT RETRIEVAL

Patent retrieval is a subdomain of information retrieval, in which the basic elements to search are patent documents. Due to the characteristics of patent documents and special requirements of patent retrieval, patent search is quite different from searching general web documents. For example, queries in patent search are generally much longer and more complex than the ones in web search.

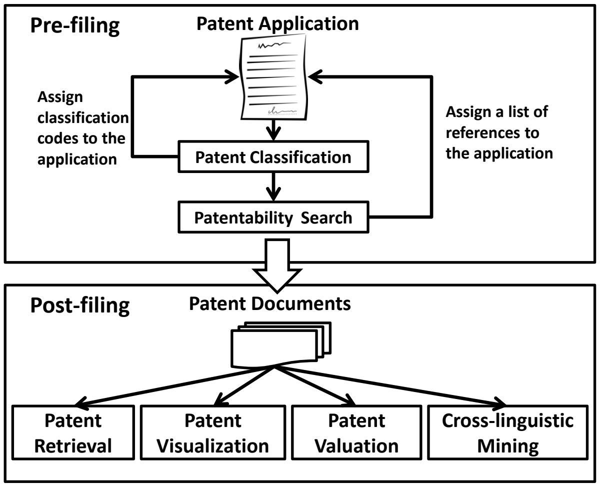


Figure 3: The architecture of a patent mining system.

With the domain-specific requirement of patent retrieval, patent search has gained great attention in the last decade in both academia and industry. Currently, there are numerous benchmark collections of patent documents available in information retrieval community, and several workshops and symposiums on patent retrieval have been organized, including NTCIR[5], CLEF[6] and TREC[7]. In 2003, the third NTCIR workshop [44] firstly provided benchmark collections of patent documents for enhancing research on patent information processing. They assigned the "Patent Retrieval Task" to explore the effect of retrieving patent documents in real-world applications. The recent advancement in patent search is driven by the "Intellectual Property" task initialized by CLEF [87]. Several teams participated in the prior-art search task of the CLEF-IP 2010 and proposed approaches to reduce the number of returned patent documents by extracting a set of key terms and expanding queries for broader coverage.

Table 2: Challenges in patent retrieval.

Challenges	Reasons
Low Readability	People may use rhetorical structures and ambiguous terms to defend their invention in order to obtain broader protection.
Lengthy Query	People often use the whole patent document as a query to perform searching.
High Recall	Missing one strongly relevant document in patent retrieval is unacceptable because of the tremendous cost of patent lawsuit.

Despite the recent advances, the task of patent retrieval remains challenging from multiple perspectives. We summa-

[5] http://research.nii.ac.jp/ntcir/index-en.html.
[6] http://ifs.tuwien.ac.at/∼clef-ip.
[7] http://trec.nist.gov.

rize several challenges related to patent retrieval as listed in Table 2. In the following, we first introduce various types of patent search tasks in Section 3.1, and then discuss existing solutions/approaches to the aforementioned challenges. A summary of patent retrieval techniques is depicted in Figure 4. Specifically, in Section 3.2 we discuss how to improve the readability of patent documents; in Section 3.3 we introduce existing methods that assist patent examiners in generating query keywords; and in Section 3.4 we describe the techniques to expand the query keyword set.

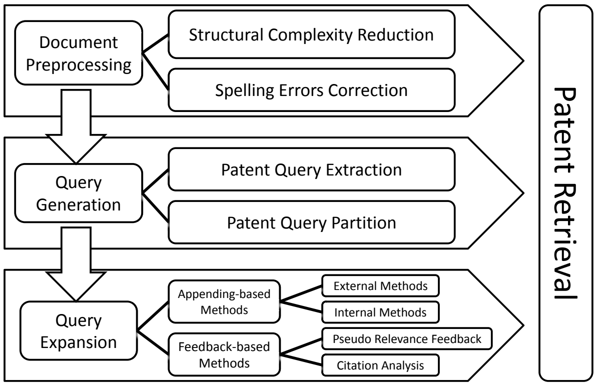


Figure 4: A summary of patent retrieval techniques.

3.1 Patent Search and a Typical Scenario

In practice, there are five representative patent search tasks listed as follows:

- *Prior-Art Search*, which aims at understanding the state-of-the-art of a general topic or a targeted technology. It is often referred to as patent landscaping or technology survey. The scope of this task mainly focuses on all the available publications[8] worldwide.

- *Patentability Search*, which tries to retrieve relevant documents worldwide that have been published prior to the application date, and may disclose the core concept in the invention. This task is often performed before/after patent application.

- *Invalidity Search*, which searches the available publications that invalidate a published patent document. This task is usually performed after a patent is granted.

- *Infringement Search*, which retrieves valid patent publications that are infringed by a given product or patent document. In general, the search operates on the claim section of the available patent documents.

- *Legal Status Search*, which determines whether an invention has freedom to make, use, and sell; that is, whether the granted patent has lapsed or not.

In Figure 5, we provide an overview of the procedure to perform patent search tasks. As depicted, it contains 4 major steps:

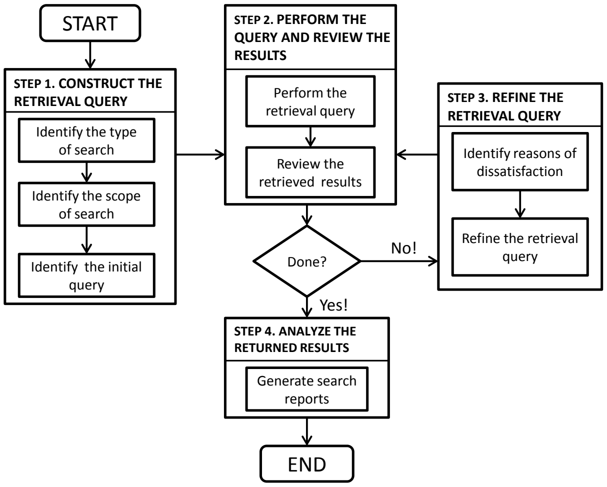

START
STEP 1. CONSTRUCT THE RETRIEVAL QUERY
Identify the type of search
Identify the scope of search
Identify the initial query
STEP 2. PERFORM THE QUERY AND REVIEW THE RESULTS
Perform the retrieval query
Review the retrieved results
Done?
No!
Yes!
STEP 3. REFINE THE RETRIEVAL QUERY
Identify reasons of dissatisfaction
Refine the retrieval query
STEP 4. ANALYZE THE RETURNED RESULTS
Generate search reports
END

Figure 5: A typical procedure of patent search.

Step 1 Construct the retrieval query:
An initial action is to determine the type of patent search task (as aforementioned) based on the purpose of patent retrieval. Then, the search scope can be identified accordingly. For example, patentability search is to retrieve relevant documents that are published prior to the filing/application date, and therefore the scope of patentability search contains all the available documents worldwide. Finally, we need to construct the initial retrieval query based on the user's information need, as well as the type of the task. For example, in the task of invalidity search, both the core invention and the classification code of the patent document need to be identified.

Step 2 Perform the query and review the results:
Queries are executed in the scope of the task identified in **Step 1**, and relevant documents are returned to the user. Then the user will review the returned results to determine whether the documents are desired. If so, go to **Step 4**; otherwise, go to **Step 3**.

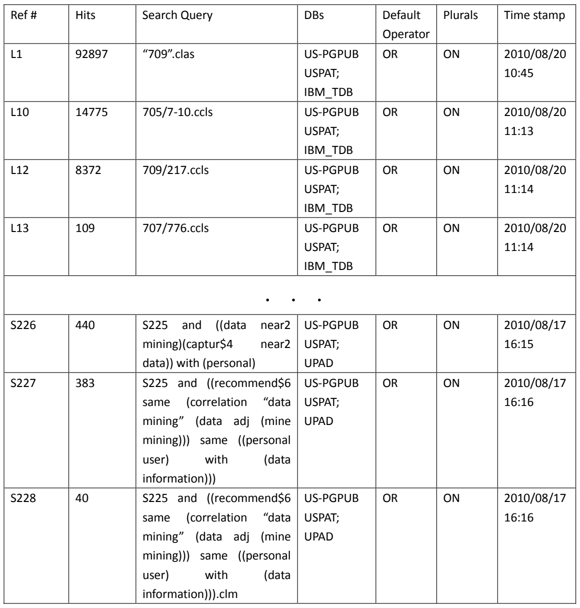
Step 3 Refine the retrieval query:
If the returned results in **Step 2** are not satisfactory (e.g., too many documents, too few results, or many irrelevant results), we need to refine search queries in order to improve the search results. For example, we can put more constrains (hyponyms) in the query if we want to reduce the number of returned documents, or remove several constrains (hyponyms) if we get too few results, or replace the query with new keywords if the results are irrelevant.

Step 4 Analyze the returned results:
After a user reviews each returned document, he/she will write a search report based on the search task in accordance with the patent law and regulation. The search report, in general, consists of: (1) a summary of the invention; (2) classification codes; (3) databases or retrieval tools used for search; (4) relevant documents; (5) query logs; and (6) retrieval conclusions.

We take patentability search as an illustrative example to further explain the search procedure. Suppose a patent ex-

[8] Here the publications are public literatures, including patent documents and scientific papers.

aminer tries to perform the patentability search for a patent application related to "Personal Data Mining". In Step 1, he/she will read the application file and extract keywords such as "data mining", "capture data", and "correlation connection link", and generate the search query based on these keywords. Then he/she will perform the search query within a series of patent databases, such as USPAT and IBM_TDB, and iteratively refine the query according to the search results in Step 2 and 3. Finally, he/she will read all 40 "hits" (the returned documents) to find a list of relevant documents and write a search report in Step 4. Figure 6 shows a query log of this example[9].



Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time stamp
L1	92897	"709".clas	US-PGPUB USPAT; IBM_TDB	OR	ON	2010/08/20 10:45
L10	14775	705/7-10.ccls	US-PGPUB USPAT; IBM_TDB	OR	ON	2010/08/20 11:13
L12	8372	709/217.ccls	US-PGPUB USPAT; IBM_TDB	OR	ON	2010/08/20 11:14
L13	109	707/776.ccls	US-PGPUB USPAT; IBM_TDB	OR	ON	2010/08/20 11:14
. . .						
S226	440	S225 and ((data near2 mining)(captur$4 near2 data)) with (personal)	US-PGPUB USPAT; UPAD	OR	ON	2010/08/17 16:15
S227	383	S225 and ((recommend$6 same (correlation "data mining" (data adj (mine mining))) same ((personal user) with (data information)))	US-PGPUB USPAT; UPAD	OR	ON	2010/08/17 16:16
S228	40	S225 and ((recommend$6 same (correlation "data mining" (data adj (mine mining))) same ((personal user) with (data information))).clm	US-PGPUB USPAT; UPAD	OR	ON	2010/08/17 16:16

Figure 6: A sample query log of patent search.

3.2 Patent Document Preprocessing

In Section 2.1, we have introduced the typical structure of patent documents. Besides the structured content in the front page, a patent document, in practice, often contains a large amount of unstructured textual information. In order to ensure the patentability of patent documents and maximize the scope of the protection, patent attorneys or inventors, in general, use complex sentences with domain-specific words to describe the invention, which renders patent documents difficult to understand or read, even for domain experts. This phenomenon is more common in the claims, which is the most important part of a patent document, as claims often define the implementation of essential components of the patent invention. In order to help users quickly grasp the core idea of a patent document, and consequently improve the efficiency of patent retrieval, it is imperative to refine the readability of patent documents.

A patent document often involves complex structure and/or lexicon. To ease the understanding of patent document, researchers usually try to reduce both *structural complexity* and *lexical complexity* using techniques of information retrieval, data mining, natural language processing, etc. For example, in [91], Shinmori et al. utilize nature language processing methods to reduce the structural complexity. They predefine six relationships (procedure, component, elaboration, feature, precondition, composition) to capture the structure information of Japanese patent claims. In addition, they use cue-phrase-based approaches to extract both cue phrase tokens and morpheme tokens, and then employ them to create a structure tree to represent the first independent claim. Their experimental results on NTCIR3 patent data collection indicate that the proposed tree-based approach can achieve better performance in terms of accuracy. In contrast, Sheremetyeva [90] proposes the similar approach to capture both the structure and lexical content of claims from US patent documents. The author decomposes the long claim sentences into short segments, and then analyzes the dependence relations among them. After that, a tree-basd representation is provided to capture both content and structure information of claims, and consequently the readability of the patent documents is improved.

Besides the complexity, patent documents often contain some spelling errors. Stein et al. [92] indicate that many patents from USPTO contain the spelling errors, e.g., "Samsung Inc" may be written as "Sumsung Inc". Such errors may increase the inconsistency of the patent corpus and hence may deteriorate the readability of patent documents. Thus, they provide an error detection approach to identify the spelling errors in the field of patent assignee (e.g., company name). The experiments have shown that both precision and recall can be improved after they correct the spell errors.

3.3 Patent Query Generation

In general, users may specify only several keywords in adhoc web search. Most web-based search systems have the restriction on the length of the input query, e.g., the maximum number of query keywords in Google search engine is 32. One possible reason is that the retrieval response time of search engines increases along with the length of the input. Comparatively in patent retrieval systems, a patent query often consists of tens or even hundreds of keywords on average. A common practice of generating such a query is to manually extract representative terms from original patent documents or add additional technological terms. This is often achieved by patent examiners, which requires a tremendous amount of time and human efforts. Also, patent examiners are expected to have strong technological background in order to provide a concise yet precise query. To assist patent examiners in generating patent queries, a lot of research work has been proposed in the last decade. In general, there are two automatic ways to produce a patent query, i.e., *query extraction* and *query partition*.

3.3.1 Query Extraction

Query extraction aims to extract representative information from an invention that describes the core idea of the invention. The simplest way of query extraction is to extract the abstract which is the summary of the invention given by the patent applicant, or the independent claims which define the scope of the protection. However, the extracted information based on abstracts or claims may not be suitable to form the patent query. The reason is straightforward: applicants often describe the abstract/claim without enough technical details in order to decrease the retrievability of their patent, and the terms in the abstract/claims often contain

[9] http://portal.uspto.gov/pair/PublicPair.

obscure meaning (e.g., "comprises" means "consists at least of") [106].

To alleviate this issue, Konishi [55] tries to expand the query by selecting terms from the explanative sentences in the description. As mentioned in Section 2, the description section of a patent document consists of the detailed information of the invention. Additional efforts along this direction involve [76; 119] that extract query terms from different sections of a patent document to automatically transform a patent file into a query. In [119], different weights are assigned to terms from different sections of patents. Their experiments on a USPTO patent collection indicate that using the terms from the description section can produce high-quality queries, and using the term frequency weighting scheme can achieve superior retrieval performance. In [76], a patent query is constructed by selecting the most representative terms from each section based on both log-likelihood weighting model and parsimonious language model [38]. While the authors only consider 4 sections, including title, abstract, description and claims, they draw the same conclusion that extracting terms from the description section of a patent document is the best way to generate queries. Mahdabi et al. [73] further propose to utilize the international patent code as an additional indicator to facilitate automatic query generation from the description section of patents.

In addition to extracting query terms from a single section [73; 76; 119], Konishi [55] exploits the combination of queries from multiple sections to build a query. The intuition is that the terms extracted from a single section is more cohesive from the ones from different sections, whereas the terms of multiple sections can help emphasize the differences between sections. Therefore, the generated queries from single sections can be treated as subqueries for searching patent documents. The experiments [55] demonstrate that the best retrieval performance could be achieved by combining the extracted terms from the abstract, claims, and description sections.

However, the aforementioned approaches require to assign weights to terms from different sections. In most cases, the weights of terms are difficult to obtain, and hence have to be heuristically assigned. To further improve the retrieval, Xue and Croft consider to employ additional features, including patent structural features, retrieval-score features, and the combinations of these features to construct a "learning-to-rank" model [118]. Their experiments on a USPTO patent collection demonstrate that the combination of terms and noun-phrases from the summary field can achieve the best retrieval performance.

3.3.2 Query Partition

An alternative way for query generation is to automatically partition the query document into multiple subtopics, and generate keywords based on each subtopic. Along this direction, several partition-based approaches have been proposed to improve the quality of patent queries. For example, Takaki et al. [95] partition the original query document into multiple subtopics, and then builds sub-queries to retrieval similar documents for each subtopic. A entropy-based "relevance score" of each subtopic is defined to determine relevance documents. However, this method involves extracting terms from the query document for each subtopic element, and hence the time complexity will increase along with the number of subtopics. Borgonovi et al. [11] present a similar approach to segment original query into subtopics. Instead of extracting terms form subtopics, they treat subtopics as sub-queries, and directly use them to execute the search and merge results obtained from each sub-query as the final result. Another approach [10] splits the original query document into multiple sentences, and then treats each sentence as an individual query to perform search. The top k relevant documents of each sub-query are merged as the final retrieval result. The empirical evaluation demonstrates that this approach is able to achieve reasonable retrieval performance, and also can significantly improve the running time compared with other baselines.

3.4 Patent Query Expansion

Patent search, as a recall-orientated search task, does not allow missing relevant patent documents due to the highly commercial value of patents and high costs of processing a patent application or patent infringement. Thus, it is important to retrieve all possible relevant documents rather than finding only a small subset of relevant patents from the top ranked results. To this end, a common practice is to enrich the query keywords in order to improve the keyword coverage, which is often referred to as *query expansion*. Recently, many query expansion techniques have been introduced in the field of patent search to improve the effectiveness of the retrieval. As discussed in [69; 78], the methods for tackling this problem can be categorized into two major groups: (1) *appending-based methods*, which either introduce similar terms or synonyms from patent document or external resources, or extract new terms from patent document to expand or reformulate a query; and (2) *feedback-based methods*, which modify the query based on the retrieved results, e.g. using pseudo relevance feedback or citation analysis.

3.4.1 Appending-Based Methods

Appending-based methods try to append additional terms to the original keyword set. In practice, the additional terms can be extracted from either the query document or the external resources, e.g., Wordnet and Wikipedia. Based on the information sources utilized by query expansion, this type of methods can be further decomposed into two groups: (1) methods that employ the query document as the expansion basis; and (2) methods that use external resources to expand the query.

Internal methods: This type of techniques exploits the query patent document itself as the resource to expand the original keyword set. The general process is to extract relevant or new terms that represent the major idea of the invention. A lot of query expansion approaches fall into this group. For example, Konishi [55] expands query terms by virtue of the "explanative sentences" extracted from the description section of the query patent, where the explanative sentences are obtained based on the longest common substring with respect to the original keyword set. In addition, several approaches [69; 99] use multi-language translation models to create a patent-related synonyms set (SynSet) from a CLEP-IP patent collection, and expand the original query based on SynSet. Parvaz et al. [73] introduce various features that can be used to estimate the importance of the noun-phrase queries. In their method, important noun-phrase queries are selected to reformulate original keyword set. These approaches are able to improve the retrieval performance; however, the improvement purely based on the extraction paradigm is quite marginal.

To further enhance the retrieval capability, semantic relations, e.g., the keyword dependencies, between query keywords are often explored. For example, Krishnan et al. [57] propose an approach to identifying the extracted treatment and causal relationships from medical patent documents. In [83], linguistic clues and word relations are exploited to identify important terms in patent documents. Based on the extracted relations between problems and solutions, the original query is reformulated. The evaluation shows that by considering the semantic relations of keywords, the retrieval performance can be improved to a great extent.

External methods: This type of techniques aims to utilize external resources, e.g., WordNet and Wikipedia, to expand original queries. WordNet is a large lexical database of English that groups different terms into sets of cognitive synonyms. It is often employed by researchers from the information retrieval community to enhance retrieval effectiveness. Recently, WordNet has been used to facilitate the process of query expansion in patent retrieval. For instance, Magdy and Jones [69] build a keyword-based synonym set with extracted synonyms and hyponyms from WordNet, and utilize this synonym set to improve the retrieval performance. However, in some cases it cannot obtain reasonable results due to the deficiency of contextual information. To solve this problem, Al-Shboul and Myaeng [2] introduce another external resource, i.e., Wikipedia, to capture the contextual information, i.e., the category dependencies. Based on the category information of Wikipedia, another query candidate set is generated. Finally, the WordNet-based synonym set and the Wikipedia-based candidate set are integrated to refine the original query.

Besides the public resources available online, the domain-specific ontology is another reliable resource that can be utilized to expand the keyword set. For example, Mukherjea et al. [82] apply Unified Medical Language System as an ontology to facilitate keyword-based patent query expansion in biomedical domain, and the result can be refined based on the semantic relations defined by the ontology. Another useful resource is the patent classification information that defines the general topic/scope of patent documents [1; 35]. Mahdabi et al. [75] treat patent classification information as domain knowledge to facilitate query expansion. Based on the international patent classification information, a conceptual lexicon is created and serves as a candidate pool to expand the keyword set. To further improve the effectiveness of patent retrieval, the proximity information of patent documents is exploited to restrict the boundary of query expansion. Recently, Tannebaum et al. [99; 100] introduce the query logs as expert knowledge to improve query expansion. Based on the analysis of query logs, they extract the frequent patterns of query terms and treat them as rules to expand the original keyword set.

3.4.2 Feedback-Based Methods

The idea of relevance feedback [89] is to employ user feedbacks to improve the search result in the process of information retrieval. However in practice, it is often difficult to obtain direct user feedbacks on the relevance of the retrieved documents, especially in patent retrieval. Hence, researchers usually exploit indirect evidence rather than explicit feedback of the search result. Generally, there are two types of approaches to acquire indirected relevant feedback: *pseudo relevance feedback* and *citation analysis*.

Pseudo relevance feedback: Pseudo relevance feedback (Pseudo-RF) [117], also known as blind relevance feedback, is a standard retrieval technique that regards the top k ranked documents from an initial retrieval as relevant documents. It automates the manual process of relevance feedback so that the user gets improved retrieval performance without an extended interaction [78]. Pseudo-RF has been extensively explored in the area of patent retrieval. Several related approaches have been proposed to employ Pseudo-RF to facilitate the retrieval performance of patent search. In NTCIR3, Kazuaki [52] exploits two relevance feedback models, including the Rocchio [89] model and Taylor expansion based model, and then extends relevance feedback methods to pseudo relevance feedback methods by assuming the top-ranked k documents as relevant documents. In NTCIR4 [43] and NTCIR5 [96], several participants attempt to utilize different Pseudo-RF approaches to improve the retrieval effectiveness. However, existing studies indicate that Pseudo-RF based approaches perform relatively poor on patent retrieval tasks, as it suffers from the problem of topic drift due to the ambiguity and synonymity of terms [71]. To alleviate the negative effect of topic drift, Bashir and Rauber [8] provide a clustering-based approach to determine whether a document is relevant or irrelevant. Based upon the intra-cluster similarity, they select top ranked documents as relevant feedback from top ranked clusters. Recently, Mahdabi et al. [74] utilize a regression model to predict the relevance of a returned document combined with a set of features (e.g. IPC clarity and query clarity). Their experiments demonstrate the superiority of the proposed method over the standard pseudo relevance feedback method. Based on this approach, in [73], they introduce an additional key-phrase extraction method by calculating phrase importance scores to further improve the performance.

Citation analysis: There are two types of citations assigned to patent documents: applicant-assigned citations and examiner-assigned citations. The first type of citations are produced by patent applicants, and often appear in the specification of patent applications in a way similar to the case that research papers are cited. Comparatively, citations assigned by patent examiners are often obtained based on the results from patentability search of the patent application, and hence might be more accurate because of the authority of the examiners.

Citations are good indicators of relevance among patent documents, and thus are often utilized to improve the search results. For example, Fuji [25] considers the cited documents as relevance feedback to expand the original query. Based on the empirical evaluation, the retrieval performance can be significantly improved by virtue of patents citation information. In CLEF 2009 IP track, Magdy et al. [68] propose to automatically extract the applicant-assigned citations from patent documents, and utilize these cited documents to facilitate patent retrieval. They further improve the citation feedback method by introducing additional terminological resources such as Wikipedia [72].

4. PATENT CLASSIFICATION

Patent classification is an important task in the process of patent application, as it provides functionalities to enable flexible management and maintenance of patent documents. However in recent years, the number of patent documents is rapidly increasing worldwide, which increases the demand for powerful patent mining systems to automatically categorize patents. The primary goal of such systems is to replace the time-consuming and labor-intensive manual categorization, and hence to offer patent analysts an efficient way to manage patent documents.

Since 1960, automatic classification has been identified as an interesting problem in text mining and natural language processing. Nowadays, in the field of text classification, researchers have devised many excellent algorithms to address this problem. However, as we previously described, it is still a non-trivial task in the domain of patent mining due to the complexity of patent documents and patent classification criteria. There are several challenges during the process of patent classification, including (1) patent documents often involve the sophisticated structures, verbose pages, and rhetorical descriptions, which renders automatic classification ineffective as it is difficult to extract useful features; (2) the hierarchical structure of the patent classification schema is quite complex, e.g. there are approximately 72,000 subgroups in the bottom level of IPC taxonomy; and (3) the huge volume of patent documents, as well as the increasing variety of patent topics, exacerbates the difficulty of automatic patent classification.

To overcome these challenges, researchers have put a lot of efforts in designing effective classification systems in the past decades. The major focus along this research direction includes (1) utilizing different types of information to perform classification; and (2) testing the performance of different classification algorithms on patent documents.

4.1 On Using Different Resources

The bag-of-words (BOW) model is often employed to represent unstructured text document. In the domain of patent document classification, the BOW representation has been widely explored. For example, Larkey [58] proposes a patent classification system in which terms and phrases are selected to represent patent documents, weighted by the frequency and structural information. Based on the vector space model, KNN (K-Nearest Neighbors) and Naïve Bayes classification models are employed to categorize US patent documents. The experiments indicate that the performance of KNN-based classifier is better than that of Naïve Bayes in the task of patent classification. After that, Koster et al. [56] propose a new approach which employs the Winnow algorithm [33] to classify patent applications. The BOW-based model is utilized to represent patent documents. Based on their experiment result, they state that the accuracy of using full-text documents is much better than that of abstracts.

The popularity of the BOW-based representation is originated from its simplicity. However, it is often difficult to convey the relationships among terms by using the BOW-based model. To address this issue, Kim et al. [49] propose a new approach to facilitate patent classification by introducing the semantic structural information. They predefine six semantic tags, including technological field, purpose, method, claim, explanation and example. Given a patent document, they convert it to the new representation based on these semantic tags. They then calculate the similarity based on both the term frequency and the semantic tag. Finally, KNN-based model is exploited to automatically classify the Japanese patent documents. The proposed approach achieves 74% improvement over the prior approaches in Japanese patent classification.

It has been widely recognized that patent classification is difficult due to the complexly structure and professional criteria of the current patent classification schema. Hence, beside exploiting the existing patent classification schema to categorize patent documents, some researchers explore the possibility of using other types of taxonomies to fulfill this task. For example, in [86], Pesenhofer et al. exploit a new taxonomy generated from Wikipedia to categorize patent documents. Cong et al. [66] design a TRIZ-based patent classification system in which TRIZ [4] is a widely used technical problem solving theory. These systems provide flexible functionalities to allow users to search relevant patent documents based on the applied taxonomy.

4.2 On Using Different Classifiers

Following the aforementioned efforts, researchers are also interested in exploring what types of classification algorithm can help improve the classification accuracy. For example, Fall et al [23; 24] compare the performance of different classification algorithms in categorizing patent documents, including Naïve Bayes, Support Vector Machine (SVM), KNN, and Winnow. Besides, they also compare the effect of utilizing different parts of patent documents, such as titles, claims, and the first 300 words of the description. Their experiments have shown that SVM achieves the best performance for class-level patent document categorization, and it is the best way to use the first 300 words of the description for representing patent documents.

As mentioned in Section 2, the IPC classification system is a five-level classification schema which contains more than 70,000 sub-groups in the bottom level. The fine-grained class label information renders patent classification more difficult. To alleviate this problem, Chen et al. [19] present a hybrid categorization system that contains three steps. Firstly, they train an SVM classifier to categorize patent documents to different sub-classes; they then train another SVM classifier to separate the documents to the bottom level of IPC; finally, they exploit KNN classification algorithms to assign the classification code to the given patent document based on the selected candidates. In their experiments, they compare various approaches employed in the sub-group level patent classification and show that their approach achieves the best performance.

Besides the traditional classification models, hierarchical approaches have also been explored, given the fact that the patent classification schema can naturally be represented as a taxonomy, as described in Section 2. For example, in [13], Cai and Hofmann present a novel hierarchical classification method that generalizes SVM. In their method, structured discriminant functions are used to mirror the class hierarchy. All the parameters are learned jointly by optimizing a common objective function with respect to a regularized upper bound on the empirical loss. The experiments on the WIPO-alpha patent collection demonstrate the effectiveness of their method. Another hierarchical model involves [103], in which the taxonomy information is integrated into an online classifier. The results on the WIPO-alpha and Espace A/B patent collections show that the method outperforms other state-of-the-art approaches significantly.

5. PATENT VISUALIZATION

The complex structure of patent documents often prevents the analysts from quickly understanding the core idea of patents. To resolve this issue, it would be helpful to visualize patent documents in a way that the gist of patents can be clearly shown to the analysts, and the correlations between different patents can be easily identified. This is often referred to as *patent visualization*, an application of information visualization.

As introduced in Section 1, a patent document contains dozens of items for analysis, which can be grouped into two categories:

- *structured data*, including patent number, filing date, issued date, and assignees, which can be utilized to generate a patent graph by employing data mining techniques;
- *unstructured text*, consisting of textual content of patent documents, such as abstract, descriptions of the invention, and major claims, which can be used to generate a patent map by employing text mining techniques.

In the following, we will discuss how patent documents can be visualized using these two types of data, as well as the integration of them.

5.1 Using Structured Data

For the purpose of analysis, structured data in patent documents are often represented as graphs. The primary resource used for constructing graphs is the citation information among different patents. By analyzing the citation graph, it is easy to discover interesting patterns with respect to particular patent documents. An example of patent citation graphs is illustrated in Figure 7a. Along this direction, several research work has been published, in which graphs are used to model patent citations. For example, in [42], Huang et al. create a patent citation graph of high-tech electronic companies in Taiwan between 1998 and 2000, where each point denotes an assignee, and the link between two points represents the relationship between them. They categorize the companies into 6 major groups, and apply graph analysis to show the similarity and distinction between different groups.

Citation analysis has been the most frequently adopted tool in visualizing the relationships of patent documents. However in some cases, it is difficult to capture the big picture of all the patent documents purely using a citation graph, as citations are insufficient to grasp the inner relations among patents. To alleviate this issue, Yoon and Park propose a network-based patent analysis method, in which the overall relationship among patents is represented as a visual network [123]. In addition, the proposed method takes more diverse keywords into account and produces more meaningful indices, which enable deeper analysis of patent documents. Tang et al. [97] further extend this idea by constructing a patent heterogeneous network, which involves a dynamic probabilistic model to characterize the topical evolution of patent documents within the network.

5.2 Using Unstructured Text

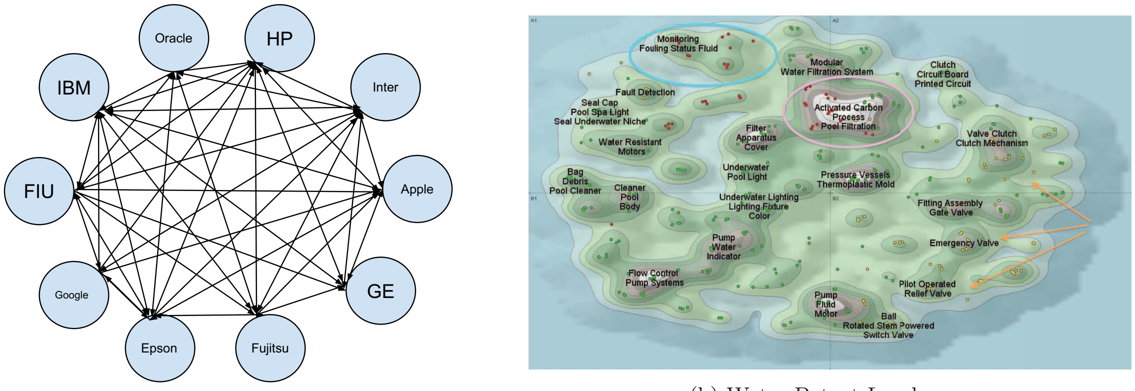
Unstructured text in patent documents provides rich information of the core ideas of patents, and therefore it becomes the primary resource for patent analysts to perform content analysis. Compared with the citation analysis, the content-based patent map has considerable advantages in latent information extraction and global technology visualization. It can also help reduce the burden of domain knowledge dependance. In the last decade, several visualization approaches have been proposed to explore the underlying patterns of patent documents and present them to users. For example, in [124], Yoon et al. present three types of patent maps, including technology vacuum map, claim point map, and technology portfolio map, all of which are generated from the unstructured text of patent documents. Figure 7b shows a patent landscape map. Similarly, Atsushi et al. [5] propose a technology portfolio map generated using the concept-based vector space model. In their model, they apply single value decomposition on the word co-occurrence matrix to obtain the word-concept matrix, and then exploit the concept-based vector to represent patent documents. To generate the patent landscape map, they employ the hierarchical clustering method based on the calculated document-concept matrix. More recently, Lee et al. [61] present an approach to generating the technology vacuum map based on patent keyword vectors. They employ principal component analysis to reduce the space of keyword features to make suitable for use on a two-dimensional map, and then identify the "technology vacuum areas" as the blank zones with sparse density and large size in the map.

5.3 Integrating Structured and Unstructured Data for Visualization

Unstructured text is useful for analyzing the core ideas of patents, and structure data provide evidences on the correlations of different patent documents. These two types of information are often integrated together for the purpose of visualization. As a representative work, Kim et al. [51] propose a novel visualization method based on both structured and unstructured data. Specifically, they first collect keywords from patent documents under a specific technology domain, and represent patent documents using keyword-based vectors. They then perform clustering on patent documents to generate k clusters. With the clustering result, they form a semantic network of keywords, and then build up a patent map by rearranging each keyword node according to its earliest filing date and frequency in patent documents. Their approach not only describes the general picture of the targeted technology domain, but also presents the evolutionary process of the corresponding techniques. In addition, natural language prossing is utilized to facilitate patent map generation [125]. Compared with the traditional technology vacuum map purely built on patent content, this approach integrates bibliographic information of patent documents, such as assignee and file date, to construct the patent maps. The generated patent map is able to assist experts in understanding technological competition trends in the process of formulating R&D strategies.

6. PATENT VALUATION

Patent documents are the core of many technology organizations and companies. To support decision making, it is imperative to assess the quality of patent documents for further actions. In practice, a common process of evaluating the importance/quality of patent documents is called *patent valuation*, which aims to assist internal decision making for patent protection strategies. For example, companies may



(a) Patent Assignee
Citation Graph (Source:NodeXL)

(b) Water Patent Landscape
Map (Source:CleanTech)

Figure 7: Representative examples of patent visualization.

create a collection of related patents, called *patent portfolio* [113], to form a "super-patent" in order to increase the coverage of protection. In this case, a critical question is how to explore and evaluate the potential benefit of patent documents so as to select the most important ones. To tackle this issue, researchers often resort to two types of approaches: *unsupervised exploration* and *supervised evaluation*. In the following, we discuss existing research publications related to patent valuation from these two perspectives.

6.1 Unsupervised Exploration

Unsupervised exploration on the importance of patent documents is often oriented towards two aspects: *influence power* and *technical strength*. The former relies on the linkage between patent documents, e.g., citations, whereas the latter mainly focus on the content analysis.

Influence power: The first work of using citations to evaluate the influence power of patent documents involves [20]. In this work, a citation graph is constructed, where each node indicates a patent document, and nodes link to others based on their citation relations. The case study of semi-synthetic penicillin demonstrates the effectiveness of using citation counts in assessing the influence power of patents. In [3], Albert et al. further extend the idea of using citation counts, and prove the correctness of citation analysis to evaluate patent documents. In addition, two related techniques are proposed, including the bibliographic coupling that indicates two patent documents share one or more citation, and co-citation analysis that indicates two patent documents have been cited by one or more patent documents. Based on these two techniques, Huang et al. [42] integrate the bibliographic coupling analysis and multidimensional scaling to assess the importance of patent documents. Further, ranking-based approaches can also be applied to the process of patent valuation. For example, Fujii [25] proposes the use of PageRank [12] to calculate citation-based score for patent documents.

Technical strength: Unlike approaches that rely on the analysis of the influence power of patent documents, some research publications focus on the analysis of the technical strength of inventions, which is relevant to the content of patents. For instance, Hasan et al. [36] define the technical strength as claim originality, and exploit text mining approaches to analysis the novelty of patent documents. They use NLP techniques to extract the key phrases from the claims section of patent documents, and then calculate the originality score based on the extracted key phrases. This valuation method has been adopted by IBM, and is applied to various patent valuation scenarios; however, the term-based approaches suffer the problem of term ambiguity, which may deteriorate the rationality of the scores in some cases. To alleviate this issue, Hu et al. [41] exploit the topic model to represent the concept of the patents instead of using words or phrases. In additional, they state that traditional patent valuation approaches cannot handle the case that the novelty of patents evolves over time, i.e., the novelty may decrease along time. Therefore, they exploit the time decay factor to capture the evolution of patent novelty. The experiment indicates that their proposed approach achieves the improvement compared with the baselines.

6.2 Supervised Evaluation

The aforementioned approaches define the importance of patent documents from either content or citation links. In essence, they are unsupervised methods as the goal is to extract meaningful patterns to assess the value of patents purely based on the patent itself. In practice, besides these two types of resources, some other information may also be available to exploit. Some researchers introduce other types of patent related records, such as patent examination results [37], patent maintenance decisions [46], and court judgments [65], to generate predicated models to evaluate patent documents. For example, Hido et al. [37] create a learning model to estimate the patentability of patent applications from the historical Japan patent examination data, and then use the model to predict the examination decision for new patent applications. They define the patentability prediction problem as a binary classification problem (reject or approval). In order to obtain an accuracy classifier, they exploit four types of features, including patent document structure, term frequency, syntactic complexity, and word age [36]. From their experiments, they demonstrate the superiority of the proposed method in estimating the examination decision. Jin et al. [46] construct a heterogeneous information network from patent documents corpus, in which nodes could be inventors, classification codes, or patent documents and edges could denote the classification similarity, the citation relation or inventor cooperation, etc. Based on this heterogeneous network, they define interesting features, such as meta features, novelty features, and writing quality features, to created a patent quality model that is able to predict the value of patents and give the maintenance decision suggestion. Liu et al. [65] propose a graphical model that discovers the valid patents which have highly probability to achieve the victory during the patent litigation process. Based on the patent citation count and court judgments, they define a latent variable to estimate the quality of patent documents. They further incorporate various quality-related features, e.g., citation quality, complexity, reported coverage, and claim originality, to improve the probabilistic model. The experiments indicate that their approach achieves promising performance for predicting court decisions.

7. CROSS-LANGUAGE PATENT MINING

Patent documents are quite sensitive to regions, i.e., patents from different regions might be described by different languages. However in reality, patent analysts prefer to receive localized patent information, even if they are described by multiple languages. For example, a patent document is written by English, but an analyst from Spain expects that this patent can be translated to Spanish for better understanding. In addition, international patent documents are required to be written by the language accepted worldwide, which is often referred to as patent globalization. In such cases, cross-language patent mining is needed to support patent localization/globalization.

In the current stage of cross-language patent mining, the primary task is cross-language information retrieval, which enables us to retrieve information from other languages using a query written in the language that we are familiar with. In general, a cross-language patent retrieval system can be constructed using two techniques: *machine translation* and *semantic correspondence*. In the following, we describe the details of these two techniques and discuss existing research efforts on this direction.

7.1 Using Machine Translation

A well-known technique to address cross-language retrieval is machine translation. By translating a query to the desired language, the problem can be reduced to a monolingual information retrieval task that various approaches can be employed. Popular machine translation systems, such as Google Translate[10], Bing Translator[11], and Cross Language[12], have been widely exploited in tackling the problem of cross-language patent retrieval [18; 48; 70; 77]. The NTCIR Workshop holds a machine translation track to encourage researchers to practice the cross-lingual patent retrieval task [27]. In [77], Makita et al. present a multilingual patent retrieval system based on the method proposed in [26], which employs a probabilistic model to reduce the ambiguity of query translation. As indicated in the report of NTCIR9 Patent Machine Translation task [32], several participants propose word-based and phrase-based translation approaches by exploiting Moses [53], an open source toolkit for statistical machine translation. Their experiments demonstrate that lexicon-based approaches are able to achieve acceptable performance; however, the domain-specific terms and structural sentences of patent documents are difficult to translate. Hence, it is imperative to explore the syntactic structure of patents when performing patent document translation.

7.2 Using Semantic Correspondence

An alternative way of building a cross-language patent search engine is to explore the semantic correspondence among languages. The basic idea is to first construct the semantic relations of a pair of languages, and then interpret the query to another language. In [64], Littman et al. present a novel approach which creates a cross-language space by exploiting latent semantic indexing(LSI) in cross-language information retrieval domain. Base on the research of [64], Li et al. [62] propose a new approach to retrieve patent documents in the Japanese-English collection. They introduce the method of kernel canonical correlation analysis [110] to build a cross-language sematic space from Japanese-English patent documents. The empirical evaluation shows that the proposed method achieves significant improvement over the state-of-the-art. However, it may require a lot of efforts to build a cross-language semantic space, and also the performance of this type of approaches is restricted by the quality of the semantic space.

8. APPLICATIONS

Patent mining aims to assist patent analysts in efficiently and effectively managing huge volume of patent documents. It is essentially an application-driven area that has been extensively explored in both academia and industry. There are a lot of online patent mining systems, either with free access or having commercial purposes. Table 3 lists several representative systems that provide flexible functionalities of patent retrieval and patent analysis (Part of the content is obtained from *Intellogist*[13]).

Patent mining systems, e.g., *Google Patent*[14], *Baidu Patent*[15] and *FreePatentOnine*[16], provide free access and basic retrieval functionalities and are very easy to use for the majority. In addition, a list of patent authorities, e.g., USPTO[17], EPO[18], WIPO[19], provide advanced search functions to allow professional users to input more complex patent queries for high-recall retrieval. These authority-based systems usually require more human efforts and domain expertise.

Some leading companies, e.g., Thomson Reuters, Questel, and Lexisnesxis, offer commercial patent mining systems. Compared with the systems with free access, commercial systems provide more advanced features to assist analysts in retrieval and processing patent documents. These commercial systems often have:

- Widespread scope. Most commercial systems not only cover patent data from multiple authorities, but also

[10] http://translate.google.com.

[11] http://www.bing.com/translator.

[12] http://www.crosslanguage.co.jp.

[13] http://www.intellogist.com.

[14] https://www.google.com/?tbm=pts.

[15] http://zhuanli.baidu.com.

[16] http://www.freepatentsonline.com.

[17] http://www.uspto.gov.

[18] http://www.epo.org.

[19] http://www.wipo.int.

Table 3: Comparison among different patent mining systems.

Systems	Thomson Innovation	Orbit	Total Patent	ProQuest	PatFT	Espacenet	Patent Scope	Google Patent	Free Patents Online
Owner	Thomson Reuters	Questel	LexisNexis	Quest	USPTO	EPO	WIPO	Google	Free Patents Online
Data Coverage(Number of authorities)	8	21	32	3	1	2	1	6	3
Legal Status Data	Yes	Yes	Yes	Yes	No	No	No	No	No
Non-Patent Sources	Yes	Yes	Yes	Yes	No	Yes	No	No	No
Legal Status Data	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Quick Search	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Advanced Search	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Keyword Term Highlighting	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Personalize Result	Yes	Yes	Yes	No	Yes	No	No	No	Yes
Keep Queries History	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes
Queries Combination	Yes	Yes	Yes	Yes	No	No	No	No	No
Bulk Documents Download	Yes	Yes	Yes	Yes	No	Yes	No	No	No
Warning Mechanism	Yes	Yes	Yes	No	No	No	No	No	No
Statistical Analysis	Yes	Yes	Yes	Yes	No	No	Yes	No	No
Patents Graphs	Yes	Yes	Yes	Yes	No	No	No	No	No
Keyword Analysis	Yes	Yes	Yes	Yes	No	No	No	No	No
Advanced Analysis	Yes	Yes	Yes	Yes	No	No	No	No	No

integrate other types of resources. For example, Thomson Reuters includes science and business articles, Questel combines news and blogs, and Lexisnesxis considers law cases. These resources are complementary to patent documents and are able to enhance the analysis power of the systems.

- Cutting-edge analysis. Commercial systems often provide patent analysis functionalities, by which more meaningful and understandable results can be obtained. For example, Thosmson Innovation provides a function called *Themescape* that identifies common themes within the search results by analyzing the concept clusters and then vividly presents them to users.

- Export functionality. Compared with free patent retrieval systems that do not allow people to export the search results, most commercial systems provide customized export functions that enable users to select and save different types of information.

Recently, several patent mining systems have been proposed in academia, most of which are constructed by utilizing the available online resources. For example, *PatentSearcher* [40] leverages the domain semantics to improve the quality of discovery and ranking. The system uses more patent fields, such as abstract, claims, descriptions and images, to retrieve and rank patents. *PatentLight* [14] is an extension of *PatentSearcher*, which categorizes the search results by virtue of the tags of the XML-structure, and ranks the results by considering flexible constraints on both structure and content. Another representative system is called *PatentMiner* [97], which studies the problem of dynamic topic modeling of patent documents and provides the topic-level competition analysis. Such analysis can help patent analysts identify the existing or potential competitors in the same topic. Further, there are some mining systems focusing on patent image search. For instance, *PATExpert* [115] presents a semantic multimedia content representation for patent documents based on semantic web technologies. *PatMedia* [112] provides patent image retrieval functionalities in content-based manner. The visual similarity is realized by comparing visual descriptors extracted from patent images.

9. CONCLUDING REMARKS

In this survey, we comprehensively investigated several technical issues in the field of patent mining, including patent search, patent categorization, patent visualization, and patent evaluation. For each issue, we summarize the corresponding technical challenges exposed in real-world applications, and explore different solutions to them from existing publications. We also introduce various patent mining systems, and discuss how the techniques are applied to these systems for efficient and effective patent mining. In summary, this survey provides an overview on existing patent mining techniques, and also sheds light on specific application tasks related to patent mining.

With the increasing volume of patent documents, a lot of application-oriented issues are emerging in the domain of patent mining. In the following, we identify a list of challenges in this domain with respect to several mining tasks.

- *Figure-Based Patent Search* introduces patent drawings as additional information to facilitate traditional patent search tasks, as technical figures are able to vividly demonstrate the core idea of invention in some domains, especially in electronics and mechanisms. The similarity between technical figures may help improve the accuracy of patent search.

- *Product-Based Patent Search*: In general, a product may be associated with multiple patents. For example, "iPhone" contains a list of key components, such as touchscreen, frame, adapter, and operating systems. What are the patents related to each component? We call this case as product-based patent search, which provides the component-level patent search results for a product.

- *Patent Infringement Analysis* aims to decide whether two patent documents are similar or one is covered by another. In general, the analysts have to manually read through lengthy patent documents to determine the equivalence/coverage. It is necessary to automate this process, or at least to provide concise summaries to ease the understanding.

- *Large-Scale Patent Retrieval* aims to alleviate the scalability issue of patent search engines. Due to the large volume of patent documents, the performance of traditional patent retrieval systems cannot meed the expectation of patent analysts. To resolve this problem, patent documents need to be carefully processed and indexed.

- *Multi-Label Hierarchical Patent Classification* denotes the process of automatically categorizing patent documents into the pre-defined classification taxonomies [13], e.g., IPC or USPC. This is a crucial step in patent document management and maintenance. However, existing approaches to solving this problem cannot efficiently handle large classification taxonomies.

- *Technique Evolution Analysis* involves generating a technology evolution tree for a given topic or a classification code related to granted patents [126]. It is a representative application of patent visualization, which enables us to effectively understand technological progress, comprehend the evolution of technologies and grab the emergence of new technologies.

- *Detecting Potential Collaborators/Competitors*: When a company would like to design a new product, a problem usually encountered by the company is who to collaborate with. Identifying potential collaborators is helpful to reduce the cost, as well as to accelerate the process of the product. In addition, the company needs to acquire features of similar products by the competitors.

- *Cross-Domain Patent Recommendation*: Online news services give people opportunities to quickly grasp the trending techniques in industry by reading technical news articles. However, tech news articles often contain a list of uncommon terms that cannot be easily understood by the audience, and consequently hinder news readers' reading experience. Therefore, it would be helpful to present patent summaries to news readers for better understanding of tech news.

Some challenges, such as the scalability and classification issues, are imperative to solve in order to assist patent analysts in efficiently and effectively performing patent analysis tasks. Other challenges can stimulate the emergence of new types of patent-oriented applications, such as evolutionary analysis and drawing-based retrieval. Even though it is impossible to describe all algorithms and applications in detail for patent mining, we believe that the ideas and challenges discussed in this survey should give readers a big picture of this field and several interesting directions for future studies.

10. ACKNOWLEDGMENTS

The work was supported in part by the National Science Foundation under grants DBI-0850203, HRD-0833093, CNS-1126619, and IIS-1213026, the U.S. Department of Homeland Security under grant Award Number 2010-ST-06200039, and Army Research Office under grant number W911NF-1010366 and W911NF-12-1-0431.

11. REFERENCES

[1] S. Adams. Comparing the ipc and the us classification systems for the patent searcher. *World Patent Information*, 23(1):15–23, 2001.

[2] B. Al-Shboul and S. Myaeng. Query phrase expansion using wikipedia in patent class search. *Information Retrieval Technology*, pages 115–126, 2011.

[3] M. Albert, D. Avery, F. Narin, and P. McAllister. Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20(3):251–259, 1991.

[4] G. S. Altšuller. *The innovation algorithm: TRIZ, systematic innovation and technical creativity.* Technical Innovation Center, Inc., 1999.

[5] H. Atsushi and T. YUKAWA. Patent map generation using concept-based vector space model. *working notes of NTCIR-4, Tokyo*, pages 2–4, 2004.

[6] L. Azzopardi, W. Vanderbauwhede, and H. Joho. Search system requirements of patent analysts. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 775–776. ACM, 2010.

[7] R. Bache and L. Azzopardi. Improving access to large patent corpora. *Transactions on large-scale data-and knowledge-centered systems II*, pages 103–121, 2010.

[8] S. Bashir and A. Rauber. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1863–1866. ACM, 2009.

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

[10] S. Bhatia, B. He, Q. He, and S. Spangler. A scalable approach for performing proximal search for verbose patent search queries. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2603–2606. ACM, 2012.

[11] F. Borgonovi. Divided we stand, united we fall: Religious pluralism, giving, and volunteering. *American Sociological Review*, 73(1):105–128, 2008.

[12] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[13] L. Cai and T. Hofmann. Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 78–87. ACM, 2004.

[14] S. Calegari, E. Panzeri, and G. Pasi. Patentlight: a patent search application. In *Proceedings of the 4th Information Interaction in Context Symposium*, pages 242–245. ACM, 2012.

[15] R. S. Campbell. Patent trends as a technological forecasting tool. *World Patent Information*, 5(3):137–143, 1983.

[16] M. Carrier. A roadmap to the smartphone patent wars and frand licensing. *Antitrust Chronicle*, 4, 2012.

[17] S. Cetintas and L. Si. Effective query generation and postprocessing strategies for prior art patent search. *Journal of the American Society for Information Science and Technology*, 63(3):512–527, 2012.

[18] M. Chechev, M. Gonzàlez, L. Màrquez, and C. España-Bonet. The patents retrieval prototype in the molto project. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 231–234. ACM, 2012.

[19] Y. Chen and Y. Chang. A three-phase method for patent classification. *Information Processing & Management*, 2012.

[20] P. Ellis, G. Hepburn, and C. Oppenhein. Studies on patent citation networks. *Journal of Documentation*, 34(1):12–20, 1978.

[21] P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik, P. Volf, and L. Zalányi. Prediction of emerging technologies based on analysis of the us patent citation network. *Scientometrics*, pages 1–18, 2012.

[22] J. Erstling and I. Boutillon. Patent cooperation treaty: At the center of the international patent system. *Wm. Mitchell L. Rev.*, 32:1583–1600, 2005.

[23] C. Fall, A. Torcsvari, K. Benzineb, and G. Karetka. Automated categorization in the international patent classification. In *ACM SIGIR Forum*, volume 37, pages 10–25, 2003.

[24] C. Fall, A. Törcsvári, P. Fievet, and G. Karetka. Automated categorization of german-language patent documents. *Expert Systems with Applications*, 26(2):269–277, 2004.

[25] A. Fujii. Enhancing patent retrieval by citation analysis. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 793–794. ACM, 2007.

[26] A. Fujii and T. Ishikawa. Japanese/english cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4):389–420, 2001.

[27] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Evaluating effects of machine translation accuracy on cross-lingual patent retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 674–675. ACM, 2009.

[28] S. Fujita. Technology survey and invalidity search: A comparative study of different tasks for japanese patent document retrieval. *Information processing & management*, 43(5):1154–1172, 2007.

[29] D. Ganguly, J. Leveling, and G. Jones. United we fall, divided we stand: A study of query segmentation and prf for patent prior art search. In *Proceedings of the 4th workshop on Patent information retrieval*, pages 13–18. ACM, 2011.

[30] D. Ganguly, J. Leveling, W. Magdy, and G. Jones. Patent query reduction using pseudo relevance feedback. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1953–1956. ACM, 2011.

[31] J. Gobeill, A. Gaudinat, P. Ruch, E. Pasche, D. Teodoro, and D. Vishnyakova. Bitem site report for trec chemistry 2010: Impact of citations feeback for patent prior art search and chemical compounds expansion for ad hoc retrieval. In *TREC*, 2010.

[32] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou. Overview of the patent machine translation task at the ntcir-9 workshop. In *Proceedings of NTCIR*, volume 9, pages 559–578, 2011.

[33] A. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43:173–210, 2001.

[34] H. Gurulingappa, B. Müller, R. Klinger, H.-T. Mevissen, M. Hofmann-Apitius, C. M. Friedrich, and J. Fluck. Prior art search in chemistry patents based on semantic concepts and co-citation analysis. In *TREC*, 2010.

[35] C. G. Harris, R. Arens, and P. Srinivasan. Comparison of ipc and uspc classification systems in patent prior art searches. In *Proceedings of the 3rd international workshop on Patent Information Retrieval*, pages 27–32. ACM, 2010.

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

[37] S. Hido, S. Suzuki, R. Nishiyama, T. Imamichi, R. Takahashi, T. Nasukawa, T. Idé, Y. Kanehira, R. Yohda, T. Ueno, et al. Modeling patent quality: A system for large-scale patentability analysis using text mining. *JIP*, 20(3):655–666, 2012.

[38] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2004.

[39] Q. Honghua and Y. Xiang. Research on a method for building up a patent map based on k-means clustering algorithm. *Science Research Management*, 2:1–9, 2009.

[40] V. Hristidis, E. Ruiz, A. Hernández, F. Farfán, and R. Varadarajan. Patentssearcher: a novel portal to search and explore patents. In *Proceedings of the 3rd international workshop on Patent information retrieval*, pages 33–38. ACM, 2010.

[41] P. Hu, M. Huang, P. Xu, W. Li, A. K. Usadi, and X. Zhu. Finding nuggets in ip portfolios: core patent mining through textual temporal analysis. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1819–1823. ACM, 2012.

[42] M. Huang, L. Chiang, and D. Chen. Constructing a patent citation map using bibliographic coupling: A study of taiwan's high-tech companies. *Scientometrics*, 58(3):489–506, 2003.

[43] H. Itoh. Ntcir-4 patent retrieval experiments at ricoh. In *NTCIR-4*, 2004.

[44] M. Iwayama, A. Fujii, N. Kando, and A. Takano. Overview of patent retrieval task at ntcir-3. In *Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20*, pages 24–32. Association for Computational Linguistics, 2003.

[45] A. Jaffe and M. Trajtenberg. *Patents, citations, and innovations: A window on the knowledge economy.* MIT press, 2005.

[46] X. Jin, S. Spangler, Y. Chen, K. Cai, R. Ma, L. Zhang, X. Wu, and J. Han. Patent maintenance recommendation with patent information network model. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 280–289. IEEE, 2011.

[47] Y. Jin. A hybrid-strategy method combining semantic analysis with rule-based mt for patent machine translation. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on*, pages 1–4. IEEE, 2010.

[48] C. Jochim, C. Lioma, H. Schütze, S. Koch, and T. Ertl. Preliminary study into query translation for patent retrieval. In *Proceedings of the 3rd international workshop on Patent information retrieval*, pages 57–66. ACM, 2010.

[49] J. Kim and K. Choi. Patent document categorization based on semantic structural information. *Information processing & management*, 43(5):1200–1215, 2007.

[50] Y. Kim, J. Seo, and W. Croft. Automatic boolean query suggestion for professional search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 825–834. ACM, 2011.

[51] Y. Kim, J. Suh, and S. Park. Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, 34(3):1804–1812, 2008.

[52] K. Kishida. Experiment on pseudo relevance feedback method using taylor formula at ntcir-3 patent retrieval task. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, NII, Tokyo. http://research. nii. ac. jp/ntcir*, 2003.

[53] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[54] S. Kondo, M. Komachi, Y. Matsumoto, K. Sudoh, K. Duh, and H. Tsukada. Learning of linear ordering problems and its application to je patent translation in ntcir-9 patentmt. In *Proceedings of NTCIR*, volume 9, pages 641–645, 2011.

[55] K. Konishi. Query terms extraction from patent document for invalidity search. In *Proceedings of NTCIR*, volume 5, 2005.

[56] C. Koster, M. Seutter, and J. Beney. Multi-classification of patent applications with winnow. In *Perspectives of System Informatics*, pages 111–125, 2003.

[57] A. Krishnan, A. F. Cardenas, and D. Springer. Search for patents using treatment and causal relationships. In *Proceedings of the 3rd international workshop on Patent information retrieval*, pages 1–10. ACM, 2010.

[58] L. Larkey. *Some issues in the automatic classification of US patents*. Massachusetts univ amherst Department of computer Science, 1997.

[59] L. Larkey. A patent search and classification system. In *International Conference on Digital Libraries: Proceedings of the fourth ACM conference on Digital libraries*, volume 11, pages 179–187, 1999.

[60] C. Lee, Y. Cho, H. Seol, and Y. Park. A stochastic patent citation analysis approach to assessing future technological impacts. *Technological Forecasting and Social Change*, 79(1):16–29, 2012.

[61] S. Lee, B. Yoon, and Y. Park. An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, 29(6):481–497, 2009.

[62] Y. Li and J. Shawe-Taylor. Advanced learning algorithms for cross-language patent retrieval and classification. *Information processing & management*, 43(5):1183–1199, 2007.

[63] Y.-R. Li. The technological roadmap of cisco's business ecosystem. *Technovation*, 29(5):379–386, 2009.

[64] M. Littman, S. Dumais, T. Landauer, et al. Automatic cross-language information retrieval using latent semantic indexing. *Cross-language information retrieval*, pages 51–62, 1998.

[65] Y. Liu, P. Hseuh, R. Lawrence, S. Meliksetian, C. Perlich, and A. Veen. Latent graphical models for quantifying and predicting patent quality. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1145–1153. ACM, 2011.

[66] H. T. Loh, C. He, and L. Shen. Automatic classification of patent documents for triz users. *World Patent Information*, 28(1):6–13, 2006.

[67] M. Lupu, F. Piroi, and A. Hanbury. Aspects and analysis of patent test collections. In *Proceedings of the 3rd international workshop on Patent information retrieval*, pages 17–22. ACM, 2010.

[68] W. Magdy and G. Jones. Applying the kiss principle for the clef-ip 2010 prior art candidate patent search task. In *Proceedings of the CLEF-2010 Conferences and Labs of the Evaluation Forum*, 2010.

[69] W. Magdy and G. Jones. A study on query expansion methods for patent retrieval. In *Proceedings of the 4th workshop on Patent information retrieval*, pages 19–24. ACM, 2011.

[70] W. Magdy and G. J. Jones. An efficient method for using machine translation technologies in cross-language patent search. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1925–1928. ACM, 2011.

[71] W. Magdy, J. Leveling, and G. J. Jones. Exploring structured documents and query formulation techniques for patent retrieval. *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pages 410–417, 2010.

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

[73] P. Mahdabi, L. Andersson, M. Keikha, and F. Crestani. Automatic refinement of patent queries using concept importance predictors. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 505–514. ACM, 2012.

[74] P. Mahdabi and F. Crestani. Learning-based pseudo-relevance feedback for patent retrieval. *Multidisciplinary Information Retrieval*, pages 1–11, 2012.

[75] P. Mahdabi, S. Gerani, J. X. Huang, and F. Crestani. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 113–122. ACM, 2013.

[76] P. Mahdabi, M. Keikha, S. Gerani, M. Landoni, and F. Crestani. Building queries for prior-art search. *Multidisciplinary Information Retrieval*, pages 3–15, 2011.

[77] M. Makita, S. Higuchi, A. Fujii, and T. Ishikawa. A system for japanese/english/korean multilingual patent retrieval. *Proceedings of Machine Translation Summit IX(online at http://www. amtaweb. org/summit/MTSummit/papers. html)*, 2003.

[78] C. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[79] E. Meij, W. Weerkamp, and M. de Rijke. A query model based on normalized log-likelihood. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1903–1906. ACM, 2009.

[80] H.-C. Meng. Innovation cluster as the national competitiveness tool in the innovation driven economy. *International Journal of Foresight and Innovation Policy*, 2(1):104–116, 2005.

[81] A. Messeni Petruzzelli, D. Rotolo, and V. Albino. Determinants of patent citations in biotechnology: An analysis of patent influence across the industrial and organizational boundaries. *Technological Forecasting and Social Change*, 2014.

[82] S. Mukherjea and B. Bamba. Biopatentminer: an information retrieval system for biomedical patents. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 1066–1077. VLDB Endowment, 2004.

[83] K.-L. Nguyen and S.-H. Myaeng. Query enhancement for patent prior-art-search based on keyterm dependency relations and semantic tags. *Multidisciplinary Information Retrieval*, pages 28–42, 2012.

[84] S. Oh, Z. Lei, P. Mitra, and J. Yen. Evaluating and ranking patents using weighted citations. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 281–284. ACM, 2012.

[85] K. OuYang and C. Weng. A new comprehensive patent analysis approach for new product design in mechanical engineering. *Technological Forecasting and Social Change*, 78(7):1183–1199, 2011.

[86] A. Pesenhofer, S. Edler, H. Berger, and M. Dittenbach. Towards a patent taxonomy integration and interaction framework. In *Proceedings of the 1st ACM workshop on Patent information retrieval*, pages 19–24. ACM, 2008.

[87] F. Piroi and J. Tait. Clef-ip 2010: Retrieval experiments in the intellectual property domain. In *Proceedings of the CLEF-2010 Conferences and Labs of the Evaluation Forum*, 2010.

[88] I. J. Rotkin, K. J. Dood, and M. A. Thexton. *A history of patent classification in the United States Patent and Trademark Office*. Patent Documentation Society, 1999.

[89] G. Salton. *The SMART retrieval system – experiments in automatic document processing*. Prentice-Hall, Inc., 1971.

[90] S. Sheremetyeva. Natural language analysis of patent claims. In *Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20*, pages 66–73. Association for Computational Linguistics, 2003.

[91] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama. Patent claim processing for readability: structure analysis and term explanation. In *Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20*, pages 56–65. Association for Computational Linguistics, 2003.

[92] B. Stein, D. Hoppe, and T. Gollub. The impact of spelling errors on patent search. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 570–579. Association for Computational Linguistics, 2012.

[93] C. Sternitzke, A. Bartkowski, and R. Schramm. Visualizing patent statistics by means of social network analysis tools. *World Patent Information*, 30(2):115–131, 2008.

[94] J. H. Suh and S. C. Park. Service-oriented technology roadmap (sotrm) using patent map for r&d strategy of service industry. *Expert Systems with Applications*, 36(3):6754–6772, 2009.

[95] T. Takaki, A. Fujii, and T. Ishikawa. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 399–405. ACM, 2004.

[96] H. Takeuchi, N. Uramoto, and K. Takeda. Experiments on patent retrieval at ntcir-5 workshop. In *NTCIR-5*, 2005.

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

[98] W. Tannebaum and A. Rauber. Acquiring lexical knowledge from query logs for query expansion in patent searching. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pages 336–338. IEEE, 2012.

[99] W. Tannebaum and A. Rauber. Analyzing query logs of uspto examiners to identify useful query terms in patent documents for query expansion in patent searching: a preliminary study. *Multidisciplinary Information Retrieval*, pages 127–136, 2012.

[100] W. Tannebaum and A. Rauber. Mining query logs of uspto patent examiners. *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 136–142, 2013.

[101] D. Teodoro, J. Gobeill, E. Pasche, D. Vishnyakova, P. Ruch, and C. Lovis. Automatic prior art searching and patent encoding at clef-ip'10. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[102] E. Terumasa. Rule based machine translation combined with statistical post editor for japanese to english patent translation. In *Proceedings of the MT Summit XI Workshop on Patent Translation*, volume 11, pages 13–18, 2007.

[103] D. Tikk, G. Biró, and A. Törcsvári. A hierarchical online classifier for patent categorization. *Emerging Technologies of Text Mining: Techniques and Applications. Idea Group Inc*, 2007.

[104] A. J. Trappey, C. Y. Fan, C. Trappey, Y.-L. Lin, and C.-Y. Wu. Intelligent recommendation methodology and system for patent search. In *Computer Supported Cooperative Work in Design (CSCWD), 2012 IEEE 16th International Conference on*, pages 172–178. IEEE, 2012.

[105] Y. Tseng et al. Text mining for patent map analysis. In *Proceedings of IACIS Pacific 2005 Conference*, pages 1109–1116, 2005.

[106] Y. Tseng, C. Lin, and Y. Lin. Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216–1247, 2007.

[107] Y. Tseng, C. Tsai, and D. Juang. Invalidity search for uspto patent documents using different patent surrogates. In *Proceedings of NTCIR-6 Workshop*, 2007.

[108] Y. Tseng and Y. Wu. A study of search tactics for patentability search: a case study on patent engineers. In *Proceedings of the 1st ACM workshop on Patent information retrieval*, pages 33–36. ACM, 2008.

[109] N. Van Zeebroeck. The puzzle of patent value indicators. *Economics of Innovation and New Technology*, 20(1):33–62, 2011.

[110] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. *Advances in neural information processing systems*, 15:1473–1480, 2003.

[111] I. Von Wartburg, T. Teichert, and K. Rost. Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, 34(10):1591–1607, 2005.

[112] S. Vrochidis, A. Moumtzidou, G. Ypma, and I. Kompatsiaris. Patmedia: augmenting patent search with content-based image retrieval. *Multidisciplinary Information Retrieval*, pages 109–112, 2012.

[113] R. P. Wagner and G. Parchomovsky. Patent portfolios. *U of Penn. Law School, Public Law Working Paper*, 56:04–16, 2005.

[114] J. Wang and D. W. Oard. Combining bidirectional translation and synonymy for cross-language information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 202–209. ACM, 2006.

[115] L. Wanner, S. Brügmann, B. Diallo, M. Giereth, Y. Kompatsiaris, E. Pianta, G. Rao, P. Schoester, and V. Zervaki. Patexpert: Semantic processing of patent documentation. In *SAMT (Posters and Demos)*, 2006.

[116] T. Xiao, F. Cao, T. Li, G. Song, K. Zhou, J. Zhu, and H. Wang. Knn and re-ranking models for english patent mining at ntcir-7. In *Proceedings of NTCIR-7 Workshop Meeting*, 2008.

[117] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11. ACM, 1996.

[118] X. Xue and W. Croft. Automatic query generation for patent search. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 2037–2040. ACM, 2009.

[119] X. Xue and W. Croft. Transforming patents into prior-art queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 808–809. ACM, 2009.

[120] Y. Yang, L. Akers, T. Klose, and C. Barcelon Yang. Text mining and visualization tools–impressions of emerging capabilities. *World Patent Information*, 30(4):280–293, 2008.

[121] Y. Y. Yang, L. Akers, C. B. Yang, T. Klose, and S. Pavlek. Enhancing patent landscape analysis with visualization output. *World Patent Information*, 32(3):203–220, 2010.

[122] T. Yeap, G. Loo, and S. Pang. Computational patent mapping: intelligent agents for nanotechnology. In *MEMS, NANO and Smart Systems, 2003. Proceedings. International Conference on*, pages 274–278. IEEE, 2003.

[123] B. Yoon and Y. Park. A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, 15(1):37–50, 2004.

[124] B.-U. Yoon, C.-B. Yoon, and Y.-T. Park. On the development and application of a self–organizing feature map–based patent map. *R&D Management*, 32(4):291–300, 2002.

[125] J. Yoon, H. Park, and K. Kim. Identifying technological competition trends for r&d planning using dynamic patent maps: Sao-based content analysis. *Scientometrics*, 94(1):313–331, 2013.

[126] L. Zhang, L. Li, T. Li, and Q. Zhang. Patentline: Analyzing technology evolution on multi-view patent graphs. In *Proceedings of the 37th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2014.

[127] L. Zhang and T. Li. Data mining applications in patent analysis. In *Data mining where theory meets practice*, pages 392–416. Xiamen University Press, 2013.

APPENDIX

A. BENCHMARK DATA

- *NTCIR Patent Data*[20]: This data set is provided by NII Testbeds and Community for Information access Research. The data set contains 3,496,252 unexamined Japanese patent applications and 1,315,470 grant patents of United States from 1993 to 2002. It is used to evaluate techniques related to patent mining such as patent retrieval, patent classification, and cross-language mining.
- *WIPO Patent Data*[21]: This patent collection is created by Fall, et al [23; 24], which aims to provide benchmark data for automatic patent classification. The data set contains about 75,000 patent applications in English (called WIPO-alpha) and 110,000 patent applications in German (called WIPO-de) from 1998 to 2002. Each patent application file consists of bibliographic data, abstract, claims, and description.
- *MAREC Patent Data*[22]: MAtrixware REsearch Collection (MAREC) is a standard patent data collection provided by Information Retrieval Facility for research purpose. It consists of over 19 million patent application and grated patents (1976-2008) from multiple authorities in 19 languages, the majority being English, German and French. MAREC has a wide usage in different areas such as patent information processing, patent retrieval, and patent translation.
- *ESPACE EP Patent Data*[23]: ESPACE EP is created by EPO, and consists of two sets of patent documents (EP-A and EP-B). Both patent collections contain bibliographic data, full text and embedded facsimile images of European patent documents from 1978 to 2006. The difference is that EP-A are patent applications, whereas EP-B are granted patents. These two patent collections are often used to carry out state-of-the-art searches on EP documents.

B. GLOSSARY

WIPO	World Intellectual Property Organization
USPTO	United States Patent and Trademark Office
EPO	European Patent Office
PCT	Patent Cooperation Treaty
IPC	International Patent Classification
USPC	United States Patent Classification
NTCIR	NII Testbeds and Community for Information access Research
CLEF	Conference and Labs of the Evaluation Forum
TREC	Text REtrieval Conference
USPAT	U.S. Patent Document Copies
IBM_TDB	IBM Technical Disclosure Bulletin
TRIZ	Theory of Inventive Problem Solving
IRF	Information Retrieval Facility

[20] http://research.nii.ac.jp/ntcir/permission/ntcir-8/perm-en-PATMN.html.

[21] http://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/index.html.

[22] http://www.ir-facility.org/prototypes/marec.

[23] http://www.epo.org/searching/subscription/ep.html.