4-1-2016

An Integrated Framework for Patent Analysis and Mining

longhui zhang
lzhan015@cs.fiu.edu

DOI: 10.25148/etd.FIDC000278

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

AN INTEGRATED FRAMEWORK FOR PATENT ANALYSIS AND MINING

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Longhui Zhang

2016

To: Interim Dean Ranu Jung
College of Engineering and Computing

This dissertation, written by Longhui Zhang, and entitled An Integrated Framework for Patent Analysis and Mining, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Sundaraja Sitharama Iyengar

Jainendra K Navlakha

Ning Xie

Debra VanderMeer

Tao Li, Major Professor

Date of Defense: April 1, 2016

The dissertation of Longhui Zhang is approved.

Interim Dean Ranu Jung
College of Engineering and Computing

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2016

DEDICATION

To my love and family.

ACKNOWLEDGMENTS

So many people have helped me come to this stage of my Ph.D. study. First and foremost, my greatest debt is to my advisor, Tao Li, an excellent researcher in the community of Data Mining. Thanks to Dr. Li confidence in me I was able to explore research directions with freedom, while getting suggestion and feedback from him at critical times.

I am grateful to my doctoral thesis committee, Sundaraja Sitharama Iyengar, Jainendra K Navlakha, Ning Xie, and Debra VanderMeer for your helpful advices and valuable feedback on my research, and my dissertation.

I would like to thank the members of the KDRG group in FIU, Li Zhen, Lei Li, Jingxuan Li, Chao Shen, Liang Tang, Hongtai Li, Wubai Zhou, Chunqiu Zeng, and Wei Xue for sharing their knowledge and their passion with me in the area of Data Mining and Information Retrieval. In particular, I would like to thank Lei Li and Chao Shen, for giving me their insights, suggestions, feedback and for many interesting discussions about my research.

My deep thanks to all the faculty, staff, and students at Florida International University who helped me in ways large and small. My special thanks goes to the staff members of the Office Support of the school of computing and information sciences, especially, Olga Carbonell for providing help and warm support during my PhD time.

Finally, I am grateful to my wife, Dan Xi, and my son, Daniel Zhang, for their generous support, who encouraged me and stood by me through thick and thin. I am thankful to my parents for their sincere support and continuous encouragement. Thank you.

ABSTRACT OF THE DISSERTATION

AN INTEGRATED FRAMEWORK FOR PATENT ANALYSIS AND MINING

by

Longhui Zhang

Florida International University, 2016

Miami, Florida

Professor Tao Li, Major Professor

Patent documents are important intellectual resources of protecting interests of individuals, organizations and companies. These patent documents have great research values, beneficial to the industry, business, law, and policy-making communities. Patent mining aims at assisting patent analysts in investigating, processing, and analyzing patent documents, which has attracted increasing interest in academia and industry. However, despite recent advances of patent mining, several critical issues in current patent mining systems have not been well explored in previous studies.

These issues include: 1) the query retrieval problem that assists patent analysts finding all relevant patent documents for a given patent application; 2) the patent documents comparative summarization problem that facilitates patent analysts in quickly reviewing any given patent document pairs; and 3) the key patent documents discovery problem that helps patent analysts to quickly grasp the linkage between different technologies in order to better understand the technical trend from a collection of patent documents.

This dissertation follows the stream of research that covers the aforementioned issues of existing patent analysis and mining systems. In this work, we delve into three interleaved aspects of patent mining techniques, including (1) PatSearch, a framework of automatically generating the search query from a given patent application and retrieving relevant patents to user; (2) PatCom, a framework of investigating the relationship in terms of commonality and difference between patent documents pairs,

and (3) PatDom, a framework of integrating multiple types of patent information to identify important patents from a large volume of patent documents.

In summary, the increasing amount and textual complexity of patent repository lead to a series of challenges that are not well addressed in the current generation systems. My work proposed reasonable solutions to these challenges and provided insights on how to address these challenges using a simple yet effective integrated patent mining framework.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Background

Patent documents are important intellectual resources that can help protect interests of individuals, organizations and companies. In the past decades, with the advanced development of various techniques in different application domains, a myriad of patent documents are filed and approved. They serve as one of the important intellectual property components for individuals, organizations and companies. These patent documents are open to the public and made available by various authorities in a lot of countries or regions around the world. For example, World Intellectual Property Organization (WIPO) [wip11] 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 [ZLL15]. 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 [Cam83]. Thus, it is imperative to carefully analyze patent documents for evaluating and maintaining patent values. 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 [TLL07, ZL13].

Patent mining aims at assisting patent analysts in investigating, processing, and analyzing patent document. Patent mining has attracted increasing interest in academia and industry [TLL07]. Recently, patent mining has been widely explored by a lot of researchers from different perspectives. These research activities mainly focus on the specific tasks in the domain of patent analysis, which include (1) effectively retrieving patent documents based on user-defined queries [AVJ10, BA10]; (2) efficiently performing patent classification for high-quality maintenance [Alt99, GLS01]; (3) informatively representing patent documents to users [Car12, KSP08]; (4) exploring the potential benefit of patent documents [AANM91, ÉMS+12] and (5) effectively dealing with cross-language patent documents [CGMEB12, FI01]. However, despite recent advances of patent mining, several critical issues in current patent mining systems have not been well explored in previous studies. These issues include:

1. Patent Retrieval: Patent retrieval is a subdomain of information retrieval, in which the basic elements to search are patents. Due to the characteristics of patents and special requirements of patent retrieval, patent search is significantly 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. Because of the tremendous cost of patent prosecution and litigation, it would be beneficial to patent retrieval if these patent queries are comprehensively understood.

2. Analysis of Single Patent Document: Patents are one of the major carriers for technology documentation. In-depth analysis of patent documents enables uncovering important technical details and relations, which can provide valuable information to develop strategies for R&D. However, patent document is often lengthy, and full of

technical and legal terminologies. Even for domain experts, it may require a huge amount of time to read and analyze a single patent document.

3. Analysis of Multiple Patent Documents: Analyzing large volume of patent documents can help us effectively understand technological progress, comprehend the evolution of technologies, and capture the emergence of new technologies. However, analysis of multiple patent documents is a non-trivial task, as there might be a lot of underlying relations among multiple documents, which requires a huge 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. Hence, automatic approaches for assisting patent analysts in the patent processing and analyzing are in high demand.

1.2 Contribution

My dissertation follows the stream of research that covers the aforementioned issues of existing patent analysis and mining systems. In this work, I delve into three interleaved aspects of patent mining techniques, including query generation for improving the performance of patent retrieval, patent summarization for understanding both commonality and difference between patent pairs, and key patent mining from a large volume of patent documents. In particular, the contributions of my dissertation are summarized as follows.

1.2.1 A unified framework for Patent Retrieval

The first contribution is a unified framework for patent retrieval, where the user submits the entire patent document as the query. Given a patent document, our

framework will automatically extract representative yet distinguishable terms to generate a search query. In order to alleviate the issues of ambiguity and topic drifting, a novel query expansion approach is proposed, which combines content proximity with topic relevance. Our framework aims to help users retrieve relevant patent documents as many as possible, and provide enough information to assist patent analysts in making the patentability decision. Specifically, the system has the following significant merits:

- *Automatic keywords extraction*: Based on the analysis of patent documents, our framework is able to automatically extract important yet distinguishable keywords from a given patent document, which integrates special characters of patent documents (e.g. patent classification code and patent structure).
- *Relevant keywords expansion*: Based on the knowledge base and term thesaurus, our framework is capable of expanding a list of keywords related to a given query term. The expansion is achieved by combining the content proximity with topic relevance.
- *Result filtering with topic*: Based on the expanded search query, our framework is able to retrieve relevant patent documents. The result is achieved by finding all potential relevant patent documents and then filtering them within the corresponding topics.

1.2.2 A comprehensive framework for Patents Comparison

The second contribution is a novel and comprehensive framework to model and compare given patent documents, which utilizes graph-based techniques to connect the dots among various aspects of the two patent documents on a term co-occurrence graph. When analyzing the retrieved patents for different retrieval tasks, our approach can serve as automatic baseline, and consequently allow the analysts to quickly

go through the results. To the best of our knowledge, our work is the first journey towards reducing human efforts of comparing patent documents by leveraging comparative summarization techniques. In summary, the contributions of our work are three-fold:

- We formulate the problem of comparing patent documents as a comparative summarization problem, and explore different means to solve this problem;
- We utilize a graph-based method to highlight the commonalities and differences between patents, and meanwhile show the relationship between the patents regarding their differences;
- We conduct extensive evaluation on a collection of US patent documents, and the results demonstrate the effectiveness of our proposed approach.

1.2.3 Discovering Key Patent on Multi-View Patent Graphs

The third contribution is a unified framework of discovering dominant patent documents, in which multiple types of patent-related information are employed, including the content and citation relations of patent documents. The input to the system is a topic or a classification code relevant to a specific technical field. The system first retrieves all the patent documents related to the topic/code from a patent database. We then construct a multi-view patent graph in which patent content, citation relations and temporal orders are integrated. We model the problem of identifying key patents as a minimum-cost dominating set problem, and select key patents using an approximation algorithm. We further discover a list of patent-related problems based on the identified key patents. These problems can be resolved by considering the temporal order of patent documents and connecting the dots between the key patents through graph-based algorithms.

To the best of our knowledge, our work is the first journey towards unifying the process of understanding the linkage between different technologies in the domain of patent analysis, by considering both document content and citation relations of patents. The contributions of our work along this direction are three-fold:

- We present a unified framework to identify dominating technologies on a multi-view patent graph that synthesizes both patent content and citation relations.
- We apply the proposed framework to multiple patent-related analysis problems that aim to discover the linkage of patents, including:
  - `PatentLine`, i.e., to outline the technology evolution of a particular domain;
  - `PatentTrace`, i.e., to trace a given technique to previous related technologies;
  - `PatentLink`, i.e., to discover the technical connection of two given patent documents.
- We conduct extensive empirical evaluation on a collection of US patent documents, and the results demonstrate the efficacy of the framework.

In summary, the increasing amount and textual complexity of patent repository lead to a series of challenges that are not well addressed in the current generation of patent mining systems. My work proposed reasonable solutions to these challenges and provided insights on how to address these challenges using a simple yet effective integrated patent mining framework.

1.3 Organization of this Dissertation

To assist the understanding and reading this dissertation, an outline of the material presented in this dissertation is given as follows. In Chapter 2, we will investigate

multiple critical research questions in the domain of patent mining and briefly introduce the existing solutions to each task based on the techniques being utilized. In Chapter 3, we will study the problem of leveraging text mining techniques, especially the query expansion techniques, to conduct the query reformulation task for improving patent retrieval performance of the current system. Then in Chapter 4, we will explore comparative summarization methods in addressing the problem of comparing patent documents. Moveover, a novel comparative summarization approach is proposed, which utilizes graph-based techniques to connect the dots among various aspects of the two patent documents on a term co-occurrence graph. Afterwards, in Chapter 5, we study the problem of mining dominating technologies from a large collection of patent documents. Finally, we will conclude my research in Chapter 6.

CHAPTER 2

Preliminaries and Related Work

Patent documents are important intellectual resources of protecting interests of individuals, organizations and companies. Different from general web documents (e.g., web pages), 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 chapter, we present an overview of the technical trend in patent mining. The rest of the chapter is organized as follows. In § 2.1, we give a brief introduction of several technical research questions in the domain of patent mining, including patent search, patent categorization, patent visualization, and patent evaluation. In§ 2.2, we provide an introduction to patent documents by describing patent document structures, patent classification systems, and various patent mining tasks. Section 2.3 presents a summary of research efforts for addressing patent retrieval, especially, patent search. In Section 2.4, we investigate how patent documents can be automatically classified into different predefined categories. In Section 2.5, 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 2.6, 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 2.7, we present different techniques for

cross-language patent mining, including approaches to solving machine translation and semantic correspondence. Section 2.8 discusses existing free and commercial patent mining systems that provide various functionalities to allow patent analysts to perform different patent mining tasks. Finally, Section 2.9 concludes this chapter and discusses emerging research- and application-wise challenges in the domain of patent mining.

2.1 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 [Cam83]. 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

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

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

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 [TLL07, ZL13].

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 chapter, 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 2.3)? (2) how to efficiently perform patent classification for high-quality maintenance (See Section 2.4)? (3) how to informatively represent patent documents to users (See Section 2.5)? (4) how to explore and evaluate the potential benefit of patent documents (See Sec-

Table 2.1: Representative patent mining tasks and approaches.

Tasks	Techniques	References
Patent Retrieval (See Section 2.3)	Query Generation	[AVJ10, BA10, BHHS12, CS12, KSC11, MKG+11, MRS08, MWdR09, TFT+12, TW08, WO06, XC09a, XC09b]
	Query Expansion	[ASM11, BR10, Fuj07a, Fuj07b, GLJ11, GLMJ11, GGR+10, GMK+10, HRH+10, Ito04, Kis03, MJ10] [MJ11a, MLJ11, MAKC12, MC12, MRS08, NM12, TUT05, TR12a, TR12b, TTJ07, WO06]
Patent Classification (See Section 2.4)	Using Different Resources	[Alt99, GLS01, KC07, KSB03, Lar97, Lar99, LHS06, PEBD08, TGP+10]
	Using Different Classifier	[CH04, CC12, FTBK03, FTFK04, GLS01, TBT07, XCL+08]
Patent Visualization (See Section 2.5)	Structured Data Visualization	[HCC03, SBS08, TWY+12, YAKBY08, YLP03, YP04]
	Unstructured Text Visualization	[AY04, HX09, LYP09, Tse05, YYP02]
	Hybrid Visualization	[Car12, KSP08, Li09, Men05, SP09, TWY+12, YAY+10, YYP02]
Patent Valuation (See Section 2.6)	Unsupervised Exploration	[AANM91, ÉMS+12, JT05, JSC+11, LCSP12, LPH10, MPRA14, OLMY12, VZ11, VWTR05]
	Supervised Evaluation	[ÉMS+12, HHX+12, JSC+11, LPH10, OW11, VZ11]
Cross-Language Mining (See Section 2.7)	Machine Translation	[CGMEB12, FI01, FUYU09, GLC+11, JLS+10, KHB+07, MJ11b, MHFI03]
	Semantic Correspondence	[KKM+11, LST07, LDL+98, Jin10, Ter07, VSTC03]

tion 2.6)? and (5) how to effectively deal with cross-language patent documents (See Section 2.7)? For each question, we first identify several critical research challenges, and then discuss different research efforts and various techniques used for addressing these challenges. Table 2.1 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.

2.2 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.

2.2.1 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*

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

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 2.1 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

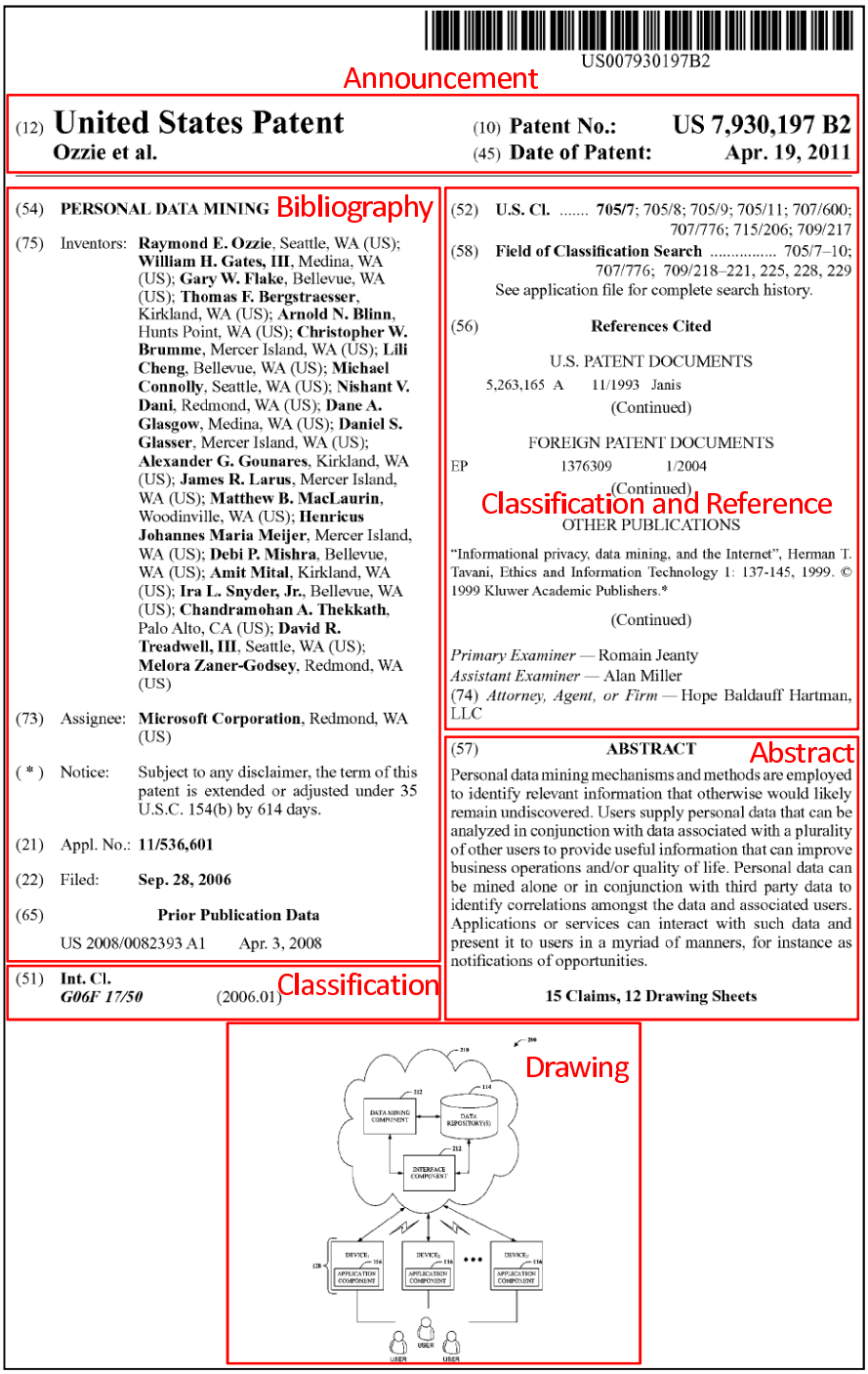


Figure 2.1: Front page of a patent document.

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 drawings. 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.

2.2.2 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.

IPC Taxonomy

IPC was established in 1971 based on Patent Cooperation Treaty [PT74]. 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 2.2 describes the class label "H01S 3/00" and its ancestors.

USPC Taxonomy

The USPC system was developed in 1836, which is the first patent taxonomy established in the world [RDT99]. In USPC, the patent categories are organized as a

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

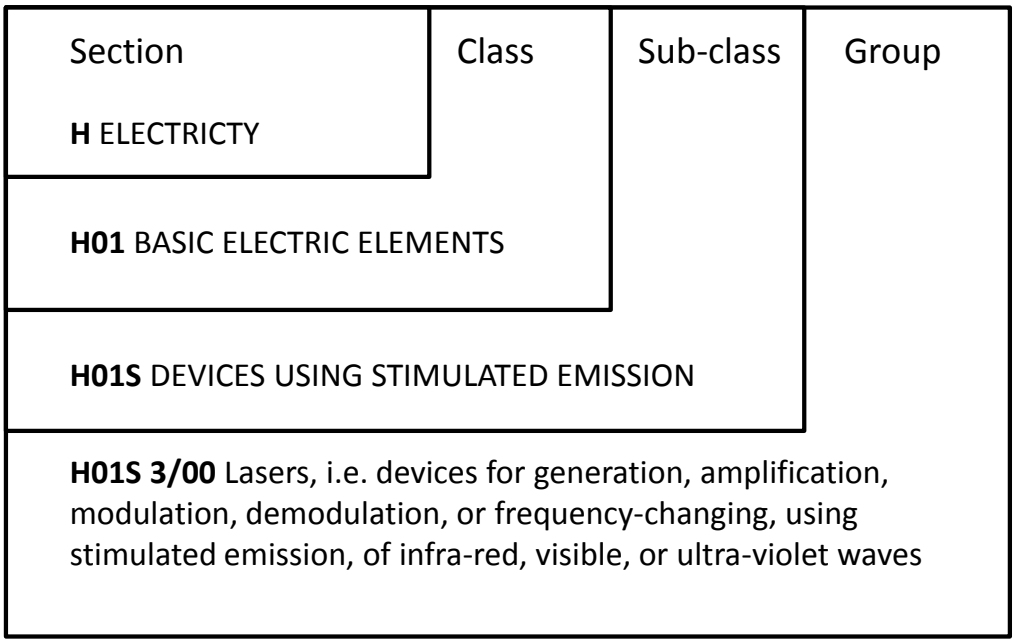


Figure 2.2: An example of IPC.

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.

2.2.3 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-filing* module, in which patent documents are maintained and analyzed. The general architecture of a patent mining system is depicted in Figure 2.3.

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

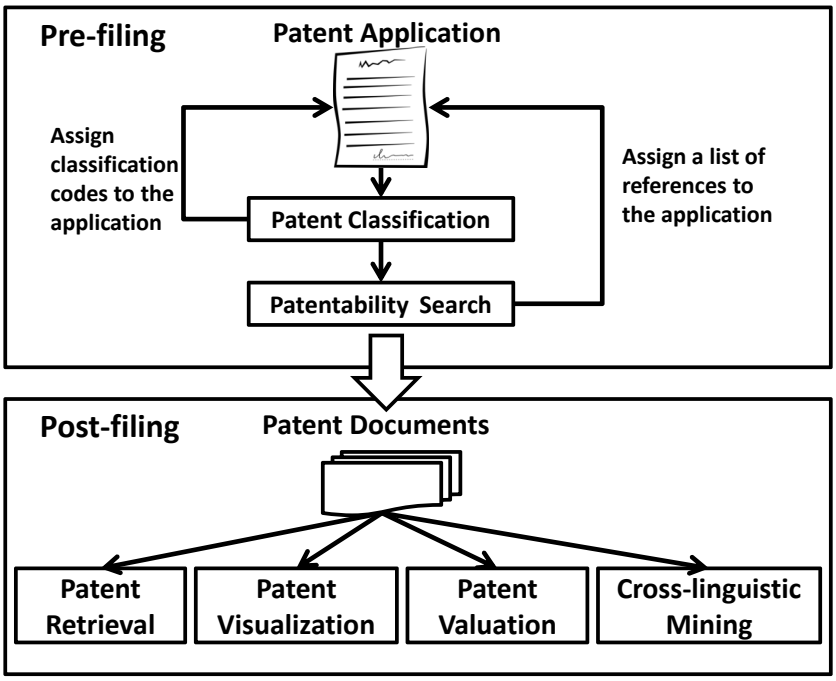


Figure 2.3: The architecture of a patent mining system.

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 2.3.

2.3 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.

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 [IFKT03] 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 [PT10]. 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.

Despite the recent advances, the task of patent retrieval remains challenging from multiple perspectives. We summarize several challenges related to patent retrieval as listed in Table 2.2. In the following, we first introduce various types of patent search tasks in Section 2.3.1, and then discuss existing solutions/approaches to the

[5]http://research.nii.ac.jp/ntcir/index-en.html.

[6]http://ifs.tuwien.ac.at/~clef-ip.

[7]http://trec.nist.gov.

Table 2.2: 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.

aforementioned challenges. A summary of patent retrieval techniques is depicted in Figure 2.4. Specifically, in Section 2.3.2 we discuss how to improve the readability of patent documents; in Section 2.3.3 we introduce existing methods that assist patent examiners in generating query keywords; and in Section 2.3.4 we describe the techniques to expand the query keyword set.

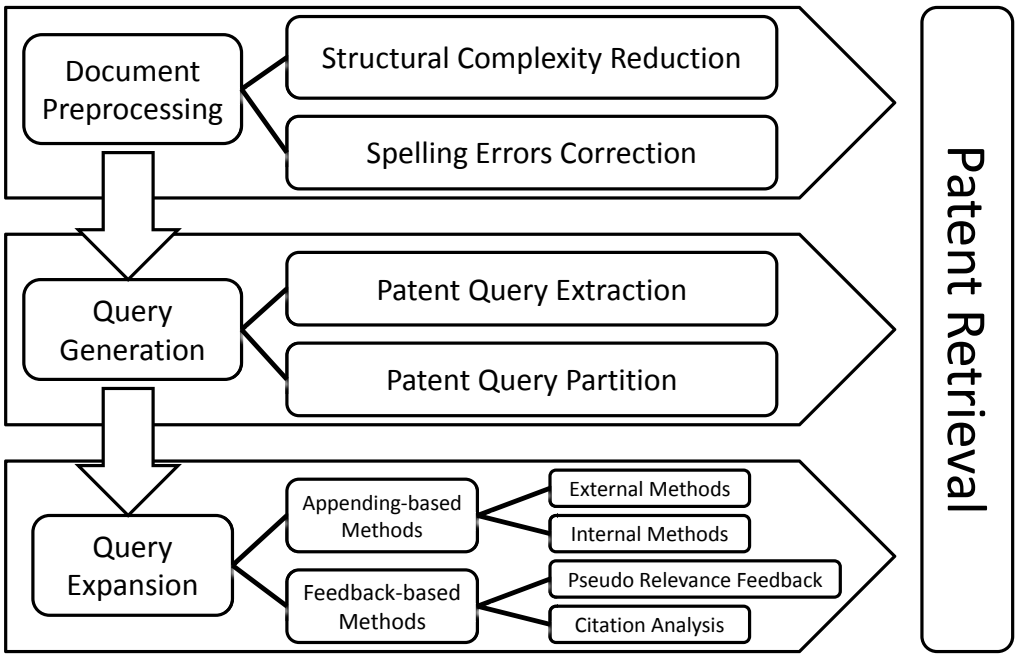


Figure 2.4: A summary of patent retrieval techniques.

2.3.1 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 2.5, we provide an overview of the procedure to perform patent search tasks. As depicted, it contains 4 major steps:

Step 1 Construct the retrieval query:

An initial action is to determine the type of patent search task (as aforemen-

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

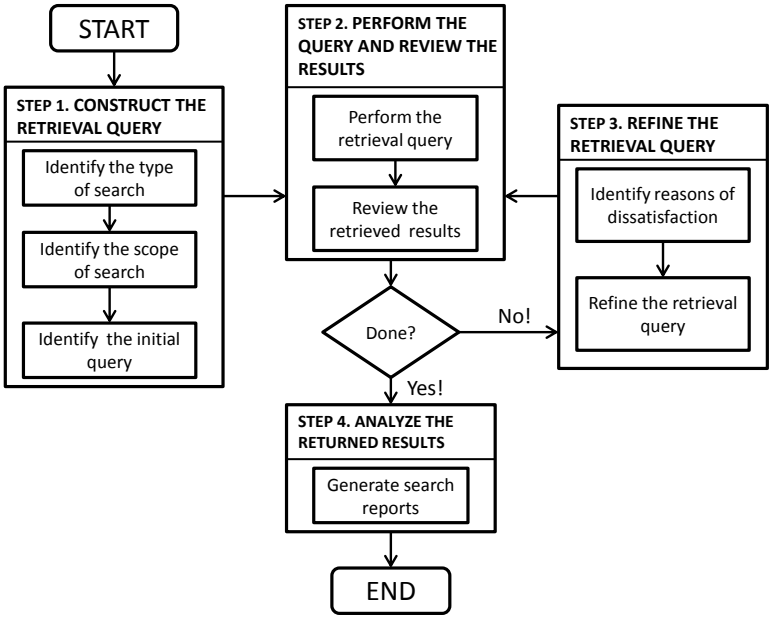


Figure 2.5: A typical procedure of patent search.

tioned) 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 examiner 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 2.6 shows a query log of this example[9].

2.3.2 Patent Document Preprocessing

In Section 2.2.1, we have introduced the typical structure of patent documents. Besides the structured content in the front page, a patent document, in practice, often

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

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 2.6: A sample query log of patent search.

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 [SOMI03], Shinmori et al. utilize nature language processing methods to reduce the structural complexity. They predefine six relationships (procedure, component, elaboration, etc) 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 [She03] 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 is improved.

Besides the complexity, patent documents often contain some spelling errors. Stein et al. [SHG12] 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.

2.3.3 Patent Query Generation

In general, users may specify only several keywords in ad-hoc 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*.

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 obscure meaning (e.g., "comprises" means "consists at least of") [TLL07].

To alleviate this issue, Konishi [Kon05] tries to expand the query by selecting terms from the explanative sentences in the description. As mentioned in Section 2.2, the description section of a patent document consists of the detailed information of the invention.

Additional efforts along this direction involve [MKG+11, XC09b] that extract query terms from different sections of a patent document to automatically transform a patent file into a query. In [XC09b], 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 [MKG+11], 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 [HRZ04]. 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. [MAKC12] further propose to utilize the interna-

tional 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 [MAKC12, MKG+11, XC09b], Konishi [Kon05] 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 [Kon05] 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 [XC09a]. 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.

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. [TFI04] 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. [Bor08] 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 [BHHS12] 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.

2.3.4 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 [MJ11a, MRS08], 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.

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 [Kon05] 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 [MJ11a, TR12b] 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. [MAKC12] 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. [KCS10] propose an approach to identifying the extracted treatment and causal relationships from medical patent documents.

In [NM12], 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 [MJ11a] 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 [ASM11] 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 public resource that can be utilized to expand the keyword set. For example, Mukherjea et al. [MB04] apply the Unified Medical Language System as an ontology to facilitate the 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 [Ada01, HAS10]. Mahdabi et al. [MGHC13] 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. [TR12b, TR13] 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.

Feedback-Based Methods

The idea of relevance feedback [Sal71] 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) [XC96], 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 [MRS08]. 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 [Kis03] exploits two relevance feedback

models, including the Rocchio [Sal71] 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 [Ito04] and NTCIR5 [TUT05], 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 [MLJ10].

To alleviate the negative effect of topic drift, Bashir and Rauber [BR09] 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. [MC12] 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 [MAKC12], 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 [Fuj07a] 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. [MJ10] 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 [MLJ11].

2.4 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 sub-groups

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.

2.4.1 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 [Lar97] 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. [KSB03] propose a new approach which employs the Winnow algorithm [GLS01] 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. [KC07] 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 [PEBD08], Pesenhofer et al. exploit a new taxonomy generated from Wikipedia to categorize patent documents. Cong et al. [LHS06] design a TRIZ-based patent classification system in which TRIZ [Alt99] 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.

2.4.2 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 [FTBK03, FTFK04] 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 2.2, 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. [CC12] 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.

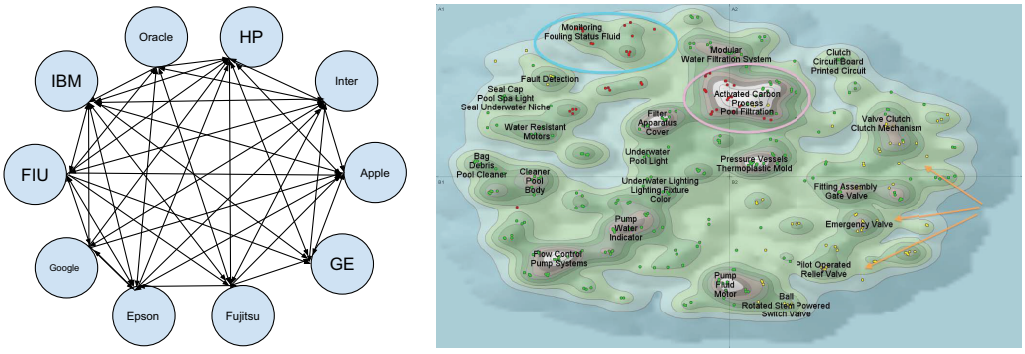
Besides the traditional classification models, hierarchical approaches have also been explored, given the fact that the patent classification schema can naturally be represented as a taxonomy, as described in Section 2.2. For example, in [CH04], Cai and Hofmann present a novel hierarchical classification 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 [TBT07], 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.

2.5 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 2.1, 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.



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

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

Figure 2.7: Representative examples of patent visualization.

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.

2.5.1 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 2.7a. Along this direction, several research work has been published, in which graphs are used to model patent citations. For example, in [HCC03], 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 [YP04]. 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. [TWY+12] 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.

2.5.2 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 [YYP02], 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 2.7b shows a patent landscape map. Similarly, Atsushi et al. [AY04] 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. [LYP09] 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.

2.5.3 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. [KSP08] 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 [YPK13]. 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.

2.6 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 create a collection of related patents, called *patent portfolio* [WP05], 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.

2.6.1 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 [EHO78]. 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 [AANM91], 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. [HCC03] 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 [Fuj07a] proposes the use of PageRank [BP98] 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. [HSGA09] 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. [HHX+12] 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.

2.6.2 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 [HSN+12], patent maintenance decisions [JSC+11], and court judgments [LHL+11], to generate predicated models to evaluate patent documents.

For example, Hido et al. [HSN+12] 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 [HSGA09]. From their experiments, they demonstrate the superiority of the proposed method in estimating the examination decision. Jin et al. [JSC+11] 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. [LHL+11] 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.

2.7 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.

2.7.1 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

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

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

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

cross-language patent retrieval [CGMEB12, JLS+10, MJ11b, MHFI03]. The NTCIR Workshop holds a machine translation track to encourage researchers to practice the cross-lingual patent retrieval task [FUYU09].

In [MHFI03], Makita et al. present a multilingual patent retrieval system based on the method proposed in [FI01], which employs a probabilistic model to reduce the ambiguity of query translation. As indicated in the report of NTCIR9 Patent Machine Translation task [GLC+11], several participants propose word-based and phrase-based translation approaches by exploiting Moses [KHB+07], 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.

2.7.2 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 [LDL+98], 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 [LDL+98], Li et al. [LST07] propose a new approach to retrieve patent documents in the Japanese-English collection. They introduce the method of kernel canonical correlation analysis [VSTC03] 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.

2.8 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 2.3 lists several representative systems that provide flexible functionalities of patent retrieval and patent analysis (Part of the content is obtained from *Intellogist*[13]).

Table 2.3: 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

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

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 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*

[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.

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* [HRH+10] 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* [CPP12] 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* [TWY+12], 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* [WBD+06] presents a semantic multimedia content representation for patent documents based on semantic web technologies. *PatMedia* [VMYK12] provides patent image retrieval functionalities in content-based manner. The visual similarity is realized by comparing visual descriptors extracted from patent images.

2.9 Chapter Summary

In this chapter, 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 [CH04], 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 [ZLLZ14]. 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.

CHAPTER 3

PatSearch: An Integrated Framework for Patent Document Retrieval

Patent retrieval primarily focuses on searching relevant legal documents with respect to a given query. Processes of patent retrieval may differ significantly, depending on the purposes of specific retrieval tasks. Given a patent application, it is challenging to determine the patentability, i.e., to decide whether a similar invention has been published. Therefore, it would be helpful to use the patent document as the query, which could reduce the labor cost and time consuming. However, it is not a trivial task to find all relevant prior art using the entire patent document as a query, as such a query is composed thousands of terms which cannot represent a focused information need.

To this end, in this chapter, we propose a unified system, name `PatSearch`, that automatically transforms the query patent into a reasonable and effective search query. It firstly extracts reprehensive yet distinguishable terms from a given patent application to generate a search query, and then expands the query by combining content proximity with topic relevance. Finally, a list of relevant patent documents have been retrieved that provide enough information to assist patent analysts in making the patentability decision. An empirical evaluation of real-world patents collection provides interesting insights and demonstrates the effectiveness of our system.

3.1 Motivation

From Chapter 2, we know that patent documents are an important type of intellectual resources that helps protect interests of companies. Different from general web documents (e.g., web pages), patent documents have a well-defined format, and they are often lengthy and rich in technical terms, which may require many human efforts

for analysis. Therefore, patent retrieval, as a new research area, emerges in recent years, aiming to assist patent analysts in retrieving, processing and analyzing patent documents [ZLL15].

In practice, patent retrieval tasks may differ from each other in terms of the retrieval purpose. Typical patent retrieval tasks include *prior-art search* (understanding the state-of-the-art of a targeted technology), *patentability search* (retrieving relevant patent documents to check if similar ideas exist), *infringement search* (examining if a product infringes a valid patent or not), etc. [AYFD+11]. Due to the great commercial value of patents and significant costs of processing a patent application or a patent infringement case, these tasks share a common requirement, i.e., to provide full coverage with respect to the query document as much as possible.

The high quality of the search query is the cornerstone of patent retrieval; however, it is not a trivial task to find/form such a query. 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 read or understand. 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. A common practice of generating the expected query is to manually extract representative terms from original patent documents or add additional technological terms by domain expects, which requires a tremendous amount of time and human efforts. Hence, it is imperative to automate this process and assist the analysts to find more relevant patent documents. As an example, Xue et al. [XC09b] extract query terms from the summary field of a patent document, and rely on the term frequency to automatically transform a patent file into a query.

On the other hand, patentability retrieval, 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 [MJ11a, MRS08]. However, despite recent advances of query expansion technique, several critical issues in current patent search systems have not been well explored in previous studies. For example, the expansion of query terms may result in topic drifting, i.e., the topics of the query may change/shift to an unintended direction after query expansion. Another critical issue is the ambiguity of search query, i.e., a single term may have multiple meanings with respect to specific context.

3.2 System Overview

To overcome the aforementioned issues, we proposed a unified framework, name `PatSearch`, where the user submits the entire patent application as the query. Given a patent application, `PatSearch` will automatically extract representative yet distinguishable terms to generate an search query. In order to alleviate the issues of ambiguity and topic drifting, a novel query expansion approach is proposed, which combines content proximity with topic relevance. `PatSearch` aims to help users retrieval relevant patent documents as many as possible, and provide enough information to assist patent analysts in making the patentability decision. Specifically, the framework has the following significant merits:

- *Automatic keywords extraction*: Based on the analysis of patent documents, `PatSearch` is able to automatically extract important yet distinguishable key-

words from a given patent application, which integrates special characters of patent documents(e.g. patent classification code, patent structure).

- *Relevant keywords expansion*: Based on the knowledge base and term thesaurus, `PatSearch` is capable of expanding a list of keywords related to a given query term. The expansion is achieved by combining the content proximity with topic relevance.
- *Results filtering with Topic*: Based on the expanded search query, `PatSearch` is able to retrieve relevant patent documents. The result is achieved by finding all potential relevant patent documents and then filtering them within the corresponding topics.

Figure 3.1 shows an overview of the framework architecture of `PatSearch`. Specifically, it contains two major modules: (1) an offline module that performs patent term analysis to create a domain-related keyword thesaurus, and patent topic analysis to generate the knowledge base; and (2) an online module that automatically generates and expands search query from a given patent application and retrieves relevant patent documents from the patent repository for the search query.

3.2.1 Offline Analysis

The offline module contains two submodules: patent term analysis and patent topic analysis. We first collect patent documents from USPTO[1] in multiple domains based on the classification of patents[2]. To enable our analysis, for each patent document, we extract its title, abstract, claims, and description. The textual content are pre-

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

[2]http://www.uspto.gov/patents/resources/classification.

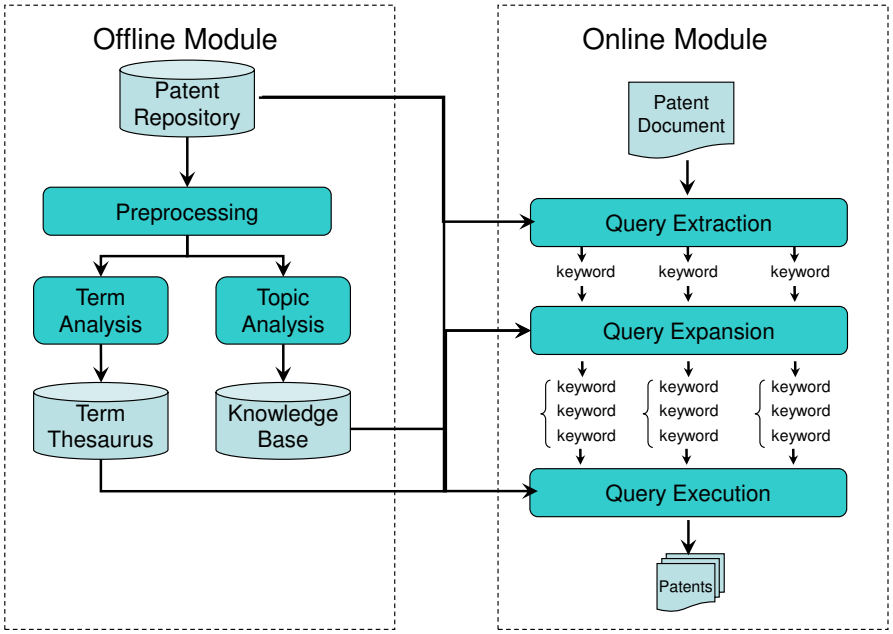


Figure 3.1: Framework architecture of PatSearch.

processed using Mallet [McC02], including stopwords removal, tokenizing, stemming, postagging, etc.

Patent Term Analysis

Patent, as a legal document, has complex structures and technical content that cause significant challenges for the retrieval system. 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, 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 analyze the technical terms and create a domain-related thesaurus.

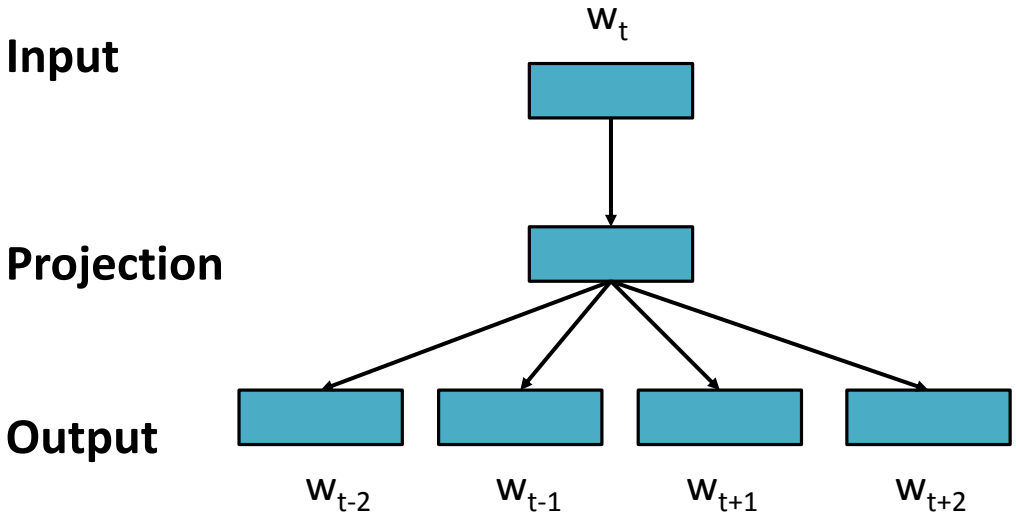


Figure 3.2: The neural network architecture of the skip-gram model

In stead of using the bag-of-word representation model, in `PatSearch`, we employ the skip-gram model [MSC+13], a novel word-embedding procedure for learning high quality vector representations of words from large amount of data, to build the keyword thesaurus. Rather then involving dense matrix multiplication, this model learns a vector representation for each word using a language model obtained from building a neural network. Figure 3.2 shows the a neural network architecture of the skip-gram model, that consists of an input layer, a projection layer, and an output layer to predict nearby words. Given a sequence that contains words $w_1, w_2, ..., w_T$, the objective is to maximize the average log probability in a corpus:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c} logp(w_{t+j} \mid w_t) \tag{3.1}$$

where c is the size context window (in the experiment, we set window size c=2) and define $p(w_{t+j}|w_t)$ using the hierarchical softmax. The speed of training the skip-gram model could be billions of word per hour using standard computer, because the simply of architecture and the use of negative sampling. After model trained, the keyword thesaurus is able to find the proximal term given a term in terms of the word vector representation, e.g., vec(handoff)$\approx$ vec(handover).

Patent Topic Analysis

In some cases, a keyword might represent multiple meanings. For example, a "chip" may present a "computer chip" or a "potato chip", and there are corresponding patent documents with respect to these two meanings. If we retrieve patent documents purely based on keyword search, the results might not be reasonable due to the ambiguity of the keywords.

Figure 3.3: Topic Model Analysis.

To resolve this issue, we try to discover the underlining topic that occur in the collection of documents. Figure 3.3 show a topic model technique for analyzing the abstract topic from a patent document. Moreover, we perform patent retrieval based on the derived topics. Specifically in `PatSearch`, we use Latent Dirichlet Allocation (LDA [BNJ03]) model to extract topics from patent documents. Formally, we treat each patent abstract as a document d, and assume that the generation of d is affected by a topic factor z, i.e., d is considered as a mixture of topics in the domain of patent documents. Each topic corresponds to a multinomial distribution over the vocabulary. The existence of the observed word w in d is considered to be drawn from a word distribution specific to topic z, i.e., $p(w|z)$. Similarly, a topic z is drawn from a document-specific topic distribution, i.e., $p(z|d)$. Based on the learned posterior probabilities, we are able to group the words contained in each patent abstract into semantic topics, and therefore treat these topics as a knowledge base for further usage.

3.2.2 Online Analysis

Online analysis includes three submodules: query extraction, query expansion, and query execution. Given a patent application, we first identify possible keywords to form an initial query, and then expand the search query based on term analysis and topic analysis. Once a query has been generated, the query execution module will retrieve a list of patent documents relevant to the original patent application to help patent analysts make the patentability decision.

Query Extraction

In general, users may specify only several keywords in ad-hoc 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.

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 obscure meaning [TLL07].

To alleviate this issue, in `PatSearch`, we try to automatically extract important yet distinguishable keywords from a given patent application. We evaluate the quality of a term in the patent application using:

$$\frac{1}{n_f}\sum_{f=1}^{n_f} f(w, q_f) \cdot log(1 + \frac{1}{f(w, D)}) \tag{3.2}$$

where f belongs to the fields of patent documents that are {title, abstract, description, claim}, $f(w, q_f)$ is the frequency of term w in the field f of patent application q, and $f(w, D)$ is the frequency of term w in the relevant patent document collection D (i.e. the patent documents that have at least a same International Patent Classification code with the given patent application). The intuition behind this formula is that a term t have a high average term frequency in all fields of the given patent application p is more likely to relevant to queries containing this term. Moreover, terms that are infrequent in the relevant patent documents have more discriminate power that is the better choice to selected for describing the information content.

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 [MJ11a, MRS08], 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.

Appending-based methods try to append additional terms to the original keyword set. In practice, the additional terms can be extracted from the external resources, e.g., WordNet and Wikipedia. WordNet is a large lexical database of English words 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 [MJ11a] 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. Hence, in`PatSearch`, instead of using the general-purpose word thesaurus, we expand query terms using domain-specific thesaurus obtained from the module of term analysis.

Another direction is to utilize relevance feedback based on the initial search result. The idea of relevance feedback [Sal71] 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. For example, in NTCIR workshop[3], several participants attempt to utilize pseudo-feedback approaches to im-

[3]http://research.nii.ac.jp/ntcir/index-en.html.

prove the retrieval effectiveness, which regards the top k ranked documents from an initial retrieval as relevant documents. However, existing studies indicate that pseudo-feedback based approaches perform relatively poor on patent retrieval tasks, as it suffers from the problem of topic drifting due to the ambiguity and synonymity of terms [MLJ10]. To alleviate the negative effect of topic drifting, we introduce a topic-based approach to determine whether a term is relevant or irrelevant.

To solve aforementioned issues, in `PatSearch`, we proposed a novel approach for expanding the give query, which combines the content proximity with topic relevance. The query expansion modular firstly finds the top-K closest terms (with K=10) from term thesaurus for each term t in the given query q to generate a potential expansion list L, and then it employs a topic-based approach to evaluate the topic relevance for each term l in the list L corresponding to the term t in the given query q for alleviating the problem of topic drifting. We calculate a relevance score (RS) with respect to the query term t with term l in the expansion list L as following:

$$RS(l,t) = \delta sim_{term}(l,t) + (1-\delta) sim_{topic}(l,t) \tag{3.3}$$

where $sim_{term}(l,t)$ is the similarity function based on cosine similarity of word embedding feature vectors $v_{term}(l)$ of and $v_{term}(t)$, and $sim_{topic}(l,t)$ is the similarity function in terms of the word topic vectors $v_{topic}(l)$ and $v_{topic}(t)$. $\delta \in [0,1]$ controls the relative importance of these two terms[4].

Query Execution

In query execution module, `PatSearch` is able to retrieve relevant patent documents given the expanded search query. The result is achieved by finding all potential relevant patent documents and then filtering them within the corresponding topics. We employ the language model with the latent topics smooth [WC06]. We compute

[4]In the experiment, we empirically set δ as 0.5

the similarity score of given query q for a document d as follow:

$$score(q,d) = \lambda score_{topic}(q,d) + (1-\lambda) score_{term}(q,d)$$

which is a linear combination of the topic similarity and the term similarity. The first term in Eq.(3.2.2) evaluates the similarity between query q and document d based on topic model, whereas the second term estimates the similarity in terms of the language model. $\lambda \in [0,1]$ controls the relative importance of these two terms[5]. The first term is calculated as follow:

$$score_{topic}(q,d) = \prod_{\text{t in q}} \sum_{z=1}^{N} p(t|z)p(z|d)$$

Here $p(z|d)$ and $p(t|z)$ are the posterior probabilities obtained in §3.2.1. For language model, we employ the Dirichlet smoothed language model:

$$score_{term}(q,d) = \prod_{w \in q} \frac{N}{N+500} P(w|d) + (1 - \frac{N}{N+500}) P(w|c)$$

where N is the number of tokens in document d, $P(w|d)$ is maximum likelihood estimation of word w in document d, and $P(w|c)$ is maximum likelihood estimation of word w in the collection c.

3.3 Experiment

In this section, we provide a comprehensive experimental evaluation to show the efficacy of our proposed framework `PatSearch`. We start with an introduction to the patent collection used in the experiment. To evaluate our proposed framework, we compare our method with other existing solutions. To enable our analysis, for each patent document, we extract its title, abstract, claims, and description. The tex-

[5]In the experiment, we set λ to 0.3 that suggested in [KS13]

tual content are preprocessed using Mallet, including stopwords removal, tokenizing, stemming, postagging, etc. The Lucene[6] toolkit is used for text indexing.

3.3.1 Data Collection

The data set used in our experiment is obtained from the United States Patent and Trademark Office [7], including 1,847,225 US granted patents, whose filing dates are ranging from 2001 to 2012. The statistics of the data are depicted in Table 3.1. To conduct the experiment, we extract the title, abstract, claim, and description, and preprocess the content using natural language processing technique, such as stop removal, tokenizing, and stemming. The number of token is more than 14B and size of vocabulary is more than 8M. We using word2vec to build the keywords repository, which fix the number of vector is 1000. We using Mallet to build the topic model among the patent collection, which set the number of topic is 1000.

In order to build a test query set, we randomly selected 100 patents that have at least 20 citations included in the patent collection. Note that there is no standard patent data set that provides the ground truth of relevant documents with respect to a patent application. Hence, for evaluation purpose, we consider the citation field of a patent as a substitute in terms of relevance judgements, which is the same strategy was also used in the NTCIR workshop series [TUT05]. These references are usually assigned by examiners during patent prosecution, but it is quite common in practice that truly relevant patents are not cited. Although the strategy of using citations as relevance judgements has a number of limitations, the same setting affects all of the algorithms of patent retrieval. Therefore, it provides a reasonable insight for comparing and evaluating algorithms in the patent retrieval. We discard these citations to

[6]http://lucene.apache.org/

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

Table 3.1: The statistics of patent data

Number of patent	1.8M
Number of tokens	14B
Size of vocabulary	8M
Number of query	100
Average num of relevance	23
Number of topic	1000
Number of vector	1000

non-US patents and non-patent literature, and also do not include references to us patent that are not covered in data collection.

3.3.2 Evaluation Methodology

To evaluate our proposed framework, we implement two existing method for query expansion:

- *WordNet [MJ11a]*: It employs WordNet to extract the synonyms and hyponyms for each term in the search query. 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.

- *PRF [XC96]*: Pseudo relevance feedback(PRF),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. After an initial run of a given query q_0, it use the Rocchio [Sal71] algorithm to generate a modified query q_m.

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j$$

  where q_0 define the original query vector, and D_r and D_{nr} are the set of relevant and nonrelevant documents, respectively. We set the weights variable $\alpha = 1$,

$\beta = 0.75$, $\gamma = 0.15$, and consider the top-20 retrieved documents as relevant document and others as nonrelevant documents.

and three retrieval approaches as baseline methods:

- *VSM [SWY75]*: In Vector Space Model(VSM), documents and queries are represented as weighted vectors in a multi-dimensional space. Here we set weights to be term frequencyCinverse document frequency (TF-IDF) value, which is well-used in information retrieval and text mining. VSM ranking the document d for query q is based on cosine similarity score of the vector v_d and v_d:

$$scored = \frac{v_q \cdot v_d}{|v_q||v_d|}$$

  Where $v_q \cdot v_d$ is the dot product of the weighted vectors, and $|v_q|$ and $|v_d|$ are their Euclidean norms.

- *BM25 [RWB+96]*: BM25 is a ranking function based on probability model, it ranks the document d for query q is computed as:

$$score(d) = \sum_{w \in q} idf(w) \cdot \frac{tf(w)(k_1 + 1)}{tf(w) + k_1(1 - b + b * \frac{|C|}{avgdl})}$$

  where $idf(w)$ is inverse document frequency of term w in collection, $tf(w)$ is tern frequency of term w in document d, $|C|$ is length of collection, and $avgdl$ is average document length for all documents in the collection. We set parameter k_1, b to 1.5, 0.75, respectively.

- *LM [ZL01]*: It employs the Dirichlet smoothed language model:

$$score(d) = \prod_{w \in q} \frac{N}{N + \lambda} P(w|d) + (1 - \frac{N}{N + \lambda}) P(w|c)$$

  where $P(w|d)$ is maximum likehood estimate of word w in document d, and $P(w|c)$ is maximum likehood estimation of word w in the collection c, and N is number of tokens in document d. We set the smoothing parameter λ to 500.

For evaluation purpose, we use Recall, Precision, F1 score, Mean Average Precision (MAP) to compare the performance on the top-100 retrieved relevant patents retrieved by our method with baseline methods for all test patents, since Joho [JAV10] conducts a survey on patent users to show that the patent examiners are willing to review the top 100 patents.

- *Recall [AYFD+11]*: It is the ratio of the number of retrieved relevant patents to all the relevant patents.

$$Recall = \frac{|\text{relevant items retrieve}|}{|\text{relevant items}|} \tag{3.4}$$

- *F1 score [AYFD+11]*: It is a measure that trades off between precision and recall, which is the evenly weighted of them.

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3.5}$$

- *Mean Average Precision (MAP) [AYFD+11]*: It is the mean of average precision for all test patents.

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{3.6}$$

  where R_{jk} is the set of ranked retrieval results from the top of retrieved list to item k in the list, and the set of relevant document for query $q_1 \in Q$ is $/p_1$, p_2,...,$p_{m_i}/$. If a relevant document is not occurred in retrieval list, the precision value is taken to be 0.

3.3.3 Result and Analysis

Query Extraction Performance

In `PatSearch`, we extract top-30 important terms to form an initial search query from the given patent query based on our query extraction module. To evaluate our query

extraction approach, we compare the generated initial search query with baseline methods [XC09b], including using title(TLE), abstract(ABS), claim(CLM), and the entire patent document(ALL) as search query, respectively. The results are reported in Table 3.2. As depicted in the table, our generated search query achieves the best

Table 3.2: Performance for query extraction

Method	Recall	MAP	F1 score
TLE	0.153	0.044	0.054
ABS	0.185	0.052	0.066
CLM	0.169	0.047	0.06
ALL	0.215	0.058	0.077
PatSearch without expansion	**0.254***	**0.06***	**0.09***

performance compared with other baseline in terms of recall, MAP, and F1-score. Especially for the recall, it significantly outperforms other methods. This is very valuable as the retrieval in the domain of patent analysis is a recall-based task. It is extremely important to have a higher recall in order to reduce the human efforts as well as to lower the risk of missing important documents. The reason is straightforward: applicants often describe the abstract/claim without enough technical details in order to decrease the retrievability of their patents, and the terms in the abstract/claims often contain obscure meaning.

Query Expansion Performance

In patent retrieval, it is important to retrieve all possible relevant documents rather than finding only a small subset of relevant patents from the top ranked results. In `PatSearch`, given the generated search query, our query expansion module selects top-3 relevant terms for expansion based on the combination of content proximity and topic relevance. For comparison, we compute the recall, F1-score, and MAP with baseline query expansion methods including expansion using wordNet and pseudo relevance feedback. The results are reported in Table 3.3

Table 3.3: Performance for query expansion

Method	Recall	MAP	F1 score
wordNet	0.293	0.063	0.104
PRF	0.248	0.058	0.088
PatSearch without expansion	0.254	0.06	0.09
PatSearch	**0.385***	**0.082***	**0.137***

As depicted in the table, after query generation, the retrieval performance has been improved significantly. Compared with other baseline methods, our approach achieves the best performance. Based on the analysis, we observe that the query expansion within wordNet slightly improves the retrieval performance. However, in some cases it cannot obtain reasonable results due to the deficiency of contextual information. The expansion using pseudo relevance feedback performs relatively poor on patent retrieval tasks, as it suffers from the problem of topic drifting, i.e., the topics of the query may change/shift to an unintended direction after query expansion, due to the ambiguity and synonymity of terms.

Query Execution Performance

After query expansion, in `PatSearch`, we retrieve relevant results with topic based filtering. In order to demonstrate the efficacy of our framework, we compare with different retrieval models. The results of our experiments are shown in Table 3.4. It

Table 3.4: Performance for query execution

	Recall	MAP	F1 score
VSM	0.276	0.063	0.099
BM25	0.308	0.07	0.11
LM	0.339	0.072	0.121
PatSearch	**0.385***	**0.082***	**0.137***

is clear that the combination of language model within topic constrain outperforms the individual methods in terms of the recall, MAP, and F1-score.

3.4 Chapter Conclusion

Patent retrieval primarily focuses on searching relevant legal documents with respect to a given query. Processes of patent retrieval may differ significantly, depending on the purposes of specific retrieval tasks. Given a patent application, it is challenging to determine its patentability, i.e., to decide whether a similar invention has been published. In general, it dose not allow missing any relevant documents due to high costs of patent prosecution and lawsuit. In this paper, we explore automatic strategies that reformulate search query of given query documents to assist the analysts in easily finding all possibly relevant documents. To this end, we proposed a unified system that combines content proximity with topic relevance to expand the original search query. Empirical evaluation on a collection of patent documents provides interesting insights, and demonstrates the effectiveness of our approach.

CHAPTER 4

PatentCom: A Comparative View of Patent Document Retrieval

Patent document retrieval, as a recall-orientated search task, does not allow missing relevant patent documents due to the great commercial value of patents and significant costs of processing a patent application or patent infringement case. Thus, it is important to retrieve all possible relevant documents rather than only a small subset of patents from the top ranked results. However, patents are often lengthy and rich in technical terms, and it often requires enormous human efforts to compare a given document with retrieved results.

In this Chapter, we formulate the problem of comparing patent documents as a comparative summarization problem, and explore automatic strategies that generate comparative summaries to assist patent analysts in quickly reviewing any given patent document pairs. To this end, we present a novel approach, named `PatentCom`, which first extracts discriminative terms from each patent document, and then connects the dots on a term co-occurrence graph. In this way, we are able to comprehensively extract the gists of the two patent documents being compared, and meanwhile highlight their relationship in terms of commonalities and differences.

The rest of this chapter is organized as follows. In §4.1 we indicate the motivation for generating comparative summaries to assist patent analysts in quickly reviewing any given patent document pairs. In §4.2 we formulate the problem, and explore possible solutions that provide comparative summaries. In §4.3 we present our graph-based comparative summarization approach, `PatentCom`. Empirical evaluation is conducted and reported in §4.4. Finally, §4.5 concludes our work.

4.1 Motivation

Patent documents are important intellectual resources of protecting interests of companies. Different from general web documents (e.g., web pages), patent documents have a well-defined format, and they are often lengthy and rich in technical terms, which may require many human efforts for analysis. Therefore, patent retrieval, as a new research area, emerges in recent years, aiming to assist patent analysts in retrieving, processing and analyzing patent documents [ZLL15].

In practice, patent retrieval tasks may differ from each other in terms of the retrieval purpose. Typical patent retrieval tasks involve *prior-art search* (understanding the state-of-the-art of a targeted technology), *patentability search* (retrieving relevant patent documents to check if similar ideas exist), *infringement search* (examining if a product infringes a valid patent or not), etc. [AYFD+11]. Due to the great commercial value of patents and significant costs of processing a patent application or patent infringement case, these tasks share a common requirement, i.e., to provide full coverage with respect to the query document as much as possible.

However, even for a few retrieved patent documents, analyzing results is not a trivial task. For instance, the task of determining patentability involves analyzing prior patents that possibly disclosed the target document. In this task, the analysts have to read through all the retrieved patent documents to determine whether the target document satisfied the patentability requirements. Nonetheless, patent documents are often lengthy, and full of technical terminologies. Even for domain experts, it may also require a huge amount of time to read and analyze a single patent document. Hence, it is imperative to automate this process and assist the analysts in reviewing the relationship between the query and the retrieved patents. Despite of some recent advancement in patent retrieval [AYFD+11,Fuj07a,SOMI03], this comparison process is still far from being well explored in research communities and industry.

In our work, we observe that typical patent retrieval tasks often require examining how similar/different two patent documents are in multiple aspects. To ease the process, it would be helpful if we can provide a comparative summary of the two patent documents being examined. To this end, we model the problem of comparing patent documents as a summarization problem, in which both commonalities and differences of documents are preferred. Traditional document summarization aims to generate a summary delivering the major information expressed in documents [GL01, SSMB97]. However, most summarization methods cannot provide comparative information. Recently, comparative summarization [WZLG12], as a special stream of summarization problems, has been proposed to summarize the differences between documents. We hence resort to this technique to address the problem of comparing patent documents.

Specifically, we first investigate available comparative summarization methods [HWX11, WZLG12] in addressing the comparison problem in patent domain. We find that although these methods can provide comparative summaries of patent documents, they fail to capture the linkage of aspects in original patent documents. To address this limitation, we propose a novel comparative summarization approach, named `PatentCom`, which utilizes graph-based techniques to connect the dots among various aspects of the two patent documents on a term co-occurrence graph. When analyzing the retrieved patents for different retrieval tasks, our approach can serve as automatic baseline, and consequently allow the analysts to quickly go through the results. To the best of our knowledge, our work is the first journey towards reducing human efforts of comparing patent documents by leveraging comparative summarization techniques. In summary, the contributions of our work are three-fold:

- We formulate the problem of comparing patent documents as a comparative summarization problem, and explore different means to solve this problem;
- We utilize a graph-based method to highlight the commonalities and differences between patents, and meanwhile show the relationship between the patents;

- We conduct extensive evaluation on a collection of US patent documents, and the results demonstrate the effectiveness of our proposed approach.

4.2 Problem Statement and Possible Solutions

The problem of comparing patent documents is a relatively new topic in the area of patent retrieval. In this section, we first formally define the problem under the setting of summarization, and then explore possible solutions to this problem.

4.2.1 Problem Formulation

Suppose there are two patents $\mathbf{d}^1$ and $\mathbf{d}^2$ for comparison. Each patent document is composed of a set of sentences, i.e., $\mathbf{d}^1 = \{s_1^1, s_2^1, \ldots, s_m^1\}$ and $\mathbf{d}^2 = \{s_1^2, s_2^2, \ldots, s_n^2\}$. The problem of comparing two patent documents is essentially a comparative summarization problem, i.e., to select a subset of sentences $\mathbf{s}^1 \subset \mathbf{d}^1$ and $\mathbf{s}^2 \subset \mathbf{d}^2$ with an identical summary length L, to accurately discriminate the two documents. The generated comparative summaries $\mathbf{s}^1$ and $\mathbf{s}^2$ can represent the general comparison of the major topic in $\mathbf{d}^1$ and $\mathbf{d}^2$, respectively. They can also be decomposed into several sections, each of which focuses on a specific aspect. For analysis purpose, the summaries should have not only acceptable quality, i.e., to be representative to the corresponding patent, but also wide coverage with less redundant information.

In general, a comparison identifies the commonalities or differences among objects. Therefore, a comparative summary should convey representative information in the documents, and contain as many comparative evidences as possible. Specifically, given two documents, the comparative summarization problem is to generate a short summary for each document to deliver the differences of these documents by extracting the most discriminative sentences in each document. This problem is related to the traditional document summarization problem as both of them try to extract

sentences from documents to form a summary. However, traditional document summarization aims to cover the majority of information among document collections, whereas comparative summarization is to find differences.

4.2.2 Existing Solutions

Recently, a list of approaches have been reported to tackle the problem of comparative summarization [HWX11, KZ09, PRK11, SJ13, WZLG12]. These approaches can mainly be categorized into two types of strategies: (1) considering only the differences between documents; and (2) focusing on both commonalities and differences of documents. In the following, we investigate these two strategies in more details.

Selection via Difference

The extraction-based summarization process generally involves selecting sentences from documents [SSMB97]. To this end, one strategy of comparative summarization is to select sentences that describe the notable difference of the two documents without considering their commonality.

A representative work in this direction involves [WZLG12], in which the selection is modeled as an optimization problem that tries to minimize the conditional entropy of the sentence membership given the selected sentence set. Let Y denote the membership identity variable of sentences, X be the entire sentence set, and X_S be the selected sentence set for comparative summary. Then the prediction capability of Y given X_S can be measured by the conditional entropy, defined as

$$\mathcal{H}(Y|X_S) \stackrel{\text{def}}{=} -\mathbf{E}_{p(Y,X_S)}(\ln p(Y|X_S)), \tag{4.1}$$

where $\mathbf{E}_{p(\cdot)}$ is the expectation given the distribution p, e.g., the joint distribution of Y and X_S. The comparative summarization problem can then be modeled as an optimization problem, i.e., $\arg\min_S \mathcal{H}(Y|X_S)$, that is, to find the most discriminative

sentences. This optimization problem can then be solved using a greedy strategy (please refer to [WZLG12] for more details).

This type of comparative summarization techniques might be suitable for general purpose. However in practice, the sentence-document matrix is quite sparse; directly selecting sentences may not be a good choice. In addition, the analysts often expect to obtain not only the differences between patent documents, but also the evidences of what aspects on which the patents are different from each other, i.e., the common yet different information. Hence, comparison between patent documents should be originated from a more fine-grained level, rather than only describing the differences.

Selection via Commonality & Difference

Another paradigm for comparative summarization considers both commonalities and differences of documents when selecting representative sentences. Typically, two patent documents are related to each other, i.e., they share some common aspects; nevertheless, their focus on these aspects might be different. Based on this observation, several methods have been reported to generate comparative summaries. One representative work involves [HWX11], which considers semantic-related cross-topic concept pairs as comparative evidences, and topic-related concepts as representative evidences.

In more details, let $C_i = \{c_{ij}\}$ be the set of concepts in document $d_i, i = 1, 2$. Each concept has a weight $w_{ij} \in \mathbb{R}$, indicating the representativeness of the concept, and a binary factor $op_{ij} \in \{0, 1\}$ indicating whether c_{ij} is presented in the summary. [HWX11] considers the cross-document concept pair $< c_{1j}, c_{2k} >$, which has a weight $u_{jk} \in \mathbb{R}$ indicating the comparative importance as well as a binary factor $op_{jk} \in \{0, 1\}$. Then the quality of a comparative summary is evaluated using

$$\lambda \sum_{j=1}^{|C_1|} \sum_{k=1}^{|C_2|} u_{jk} \cdot op_{jk} + (1 - \lambda) \sum_{i=1}^{2} \sum_{j=1}^{|C_i|} w_{ij} \cdot op_{ij}, \tag{4.2}$$

which is a linear combination of the representativeness and the comparative importance. The first term in Eq.(4.2) evaluates the cross-document comparativeness in terms of the concepts presented in the summary, whereas the second term estimates the representativeness of the concepts. $\lambda \in [0, 1]$ controls the relative importance of these two terms. w_{ij} is calculated as the term frequency, whereas u_{jk} is computed as the averaged term frequency if the corresponding two terms are semantically relevant (using WordNet [PPM04]). The optimization problem of Eq.(4.2) can be solved using linear programming, as indicated in [HWX11].

This type of comparative summarization methods relies on external resources, e.g., WordNet, to extract semantically relevant concepts from documents. However in the domain of patent retrieval, the terms in a patent document are often used from a legal perspective. It is difficult to extract meaningful concept pairs from such documents by utilizing general thesaurus. In addition, the generated summaries of this method are presented as a list of sentence pairs without indicating the relevance cross different pairs. Consequently, the readability of the summaries might be deteriorated.

4.3 Our Approach: PatentCom

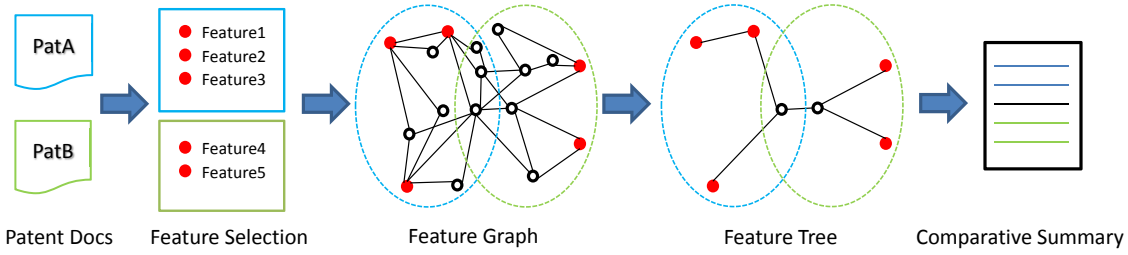


Figure 4.1: An overview of PatentCom.

To address the limitations of the aforementioned tentative solutions, we propose a novel approach, named `PatentCom`, in which graph-based methods are utilized to tackle the comparative summarization problem. Figure 5.1 presents an overview of our proposed approach. It contains 4 major modules, described as follows.

1. *Selecting Discriminative Features* (§4.3.1): Given two patents, we treat each document as a class, and perform feature selection to extract discriminative terms (i.e., nouns).

2. *Constructing Feature Graph* (§4.3.2): We construct an undirected feature graph using the feature co-occurrence information in the original patent documents, and map the discriminative features onto the graph.

3. *Extracting Representative Tree* (§4.3.3): Based on the discriminative features, we extract common information of two patents on the feature graph. The discriminative and common features are represented as a tree-based structure.

4. *Generating Comparative Summaries* (§4.3.4): We select sentences from the two patent documents by using the connected dots on the generated feature tree. The resulted summary covers both commonalities and differences of patents.

4.3.1 Discriminative Feature Selection

Patent documents often differ from each other on specific aspects. For instance, technical patents often utilize different techniques in their inventions. Hence, as the first step, we try to extract discriminative terms, i.e., nouns, from patent documents. These terms can be regarded as aspects that distinguish the two patents being compared. We therefore treat each patent document as a class, and nouns/noun phrases as features, and model the problem as a feature selection problem.

Formally, suppose we have t feature variables from the two patent documents, denoted by $\{x_i | x_i \in F\}$, where F is the full feature index set, having $|F| = t$. We have the class variable, $C = \{c_1, c_2\}$. The problem of feature selection is to select a subset of features, $S \subset F$, to accurately predict the target class variable C. There are various ways to perform feature selection, e.g., information theory based methods (such as information gain and mutual information), and statistical methods (such as

χ^2 statistics). In our work, we use χ^2 statistics as the feature selection method as it has been successfully applied to the field of text mining [YP97].

4.3.2 Feature Graph Construction

The discriminative features from §4.3.1 are able to describe the differences between patents. However, a comparative summary of two patent documents should include both different and common aspects. To obtain the common aspects and link them to the differences, we resort to graph-based approaches.

Particularly in our work, we construct an undirected graph $\mathbb{G}$ to represent two patent documents, where $\mathbb{G} = (V, E)$. $\mathbb{G}$ contains a set of vertices (i.e., features) V, where each vertex represents the nouns/noun phrases in patent documents. Two vertices connect to each other only if they co-occur in the same sentence. In order to link two vertices, we consider both their co-occurrence and their corresponding frequencies in each document. Specifically, we define a linkage score of two vertices v_1 and v_2 in a single document A as

$$w_A(v_1, v_2) = 2\frac{|\{(v_1, v_2)|v_1 \in A, v_2 \in A\}|}{|\{v_1|v_1 \in A\}| \times |\{v_2|v_2 \in A\}|}, \tag{4.3}$$

where $|\{v_1|v_1 \in A\}|$ and $|\{v_2|v_2 \in A\}|$ denote the frequencies of v_1 and v_2 in document A, respectively. $|\{(v_1, v_2)|v_1 \in A, v_2 \in A\}|$ represents the number of times that v_1 and v_2 appear in the same sentence of A. $w_A(v_1, v_2)$ essentially models the co-occurring probability of v_1 and v_2 in A. Given two patent documents A and B, we connect v_1 and v_2 if their averaged linkage score on both A and B exceeds a predefined threshold τ[1].

[1]In the experiment, we empirically set τ as 0.1.

4.3.3 Feature Tree Extraction

The discriminative features obtained from feature selection are capable of representing the difference of patent documents. However, there might be some gaps among these features, that is, they may not be well connected in the feature graph. In order to provide a fluent structure of comparative summary, we have to discover the relationship among discriminative features. This can possibly be achieved by connecting the discriminative vertices and the vertices shared by two patent documents. Also, for presentation purpose, the generated summary should be as dense and informative as possible, i.e., to include the minimum number of features and convey the major commonalities/differences.

In our problem setting, we expect that the identified features can be connected in a meaningful way, we hence formulate it as the minimum Steiner tree problem. Given a graph $\mathbb{G}$ (the feature graph in §4.3.2) and a subset of vertices S (the discriminative features in §4.3.1), a Steiner tree of $\mathbb{G}$ is similar to minimum spanning tree, defined as the subtree of $\mathbb{G}$ that contains S with the minimum number of edges. **Definition** Given a graph $\mathbb{G} = (V, E)$, a vertex set $S \subset V$ (terminals) and a vertex $v_0 \in S$ from which every vertex of S is reachable in $\mathbb{G}$, the problem of minimum Steiner tree (MST) is to find the subtree of $\mathbb{G}$ rooted at v_0 that subsumes S with minimum number of edges.

The problem of MST, is known as an NP-hard problem [Kar72]. As suggested by [CCC+99], a reasonable approximation can be achieved by finding the shortest path from the root to each terminal and then combining the paths, with the approximation ratio of $O(\log^2 k)$, where k is the number of terminals. The approximation algorithm is described in Algorithm 1.

The algorithm employs a recursive way to generate the Steiner tree T. It takes a level parameter $i \geq 1$. When $i = 1$, the algorithm tries to find the k terminals

Algorithm 1 $Steiner_i(\mathbb{G}, S, v_0, k)$ for a undirected graph

Require: $\mathbb{G} = (V, E)$: an undirected features graph; S: terminal set; $v_0 \in S$: root of the Steiner tree; k: target size of terminals to be covered
Ensure: **T**: a Steiner tree rooted at v_0 covering at least k terminals

$T \leftarrow \emptyset$
while $k > 0$ **do**
    $T_{opt} \leftarrow \emptyset$;
    $cost(T_{opt}) \leftarrow \infty$
    for v, $(v_0, v) \in E_{ct}$, and k', $1 \leq k' \leq k$ **do**
        $T' \leftarrow Steiner_{i-1}(\mathbb{G}, S, v, k') \cup \{(v_0, v)\}$
        if $(cost(T_{opt}) > cost(T'))$ **then**
            $T_{opt} \leftarrow T'$
        end if
    end for
    $T \leftarrow T \cup T_{opt}$; $k \leftarrow k - |S \cap V(T_{opt})|$;
    $S \leftarrow S \setminus V(T_{opt})$
end while
return T

which are the closest to the root v_0 and connect them to v_0 using shortest paths. As each vertex in the feature graph can reach to any other vertices, we hence randomly choose v_0 from the terminal set. As $i > 1$, the algorithm repeatedly finds a vertex v adjacent to the input root of the i-th function and a number k' such that the cost of the updated tree is the least among all tree of this form. Here the cost of a tree is calculated as the number of edges in the tree. After obtaining the expected path, we update the corresponding Steiner tree, the target size k and the terminal set S.

The generated Steiner tree of the feature graph gives us an elegant representation of patent comparison, which describes the transitions among all the other discriminative features, connected by the common features shared by two patents. Once the Steiner tree is generated, we can easily obtain a concise feature-based comparative summary of given patent documents.

4.3.4 Comparative Summarization Generation

The Steiner tree obtained from §4.3.3 provides us the basis to generate comparative summaries of two patent documents. Our goal is to select the minimum set of sentences from the original documents, by which the features in the Steiner tree can be fully covered. Each sentence can be represented as a subgraph of the entire feature graph, whereas the Steiner tree can also be regarded as a subgraph. Hence, the problem is to select the minimum set of subgraphs that cover the Steiner tree. Formally, we define the union of two graphs $G_a = (V_a, E_a)$ and $G_b = (V_b, E_b)$ as the union of their vertex and edge sets, i.e., $G_a \cup G_b = (V_a \cup V_b, E_a \cup E_b)$.

We denote each sentence as $G_i = (V_i, E_i)$, which is a subgraph of $\mathbb{G}(V, \mathbf{w}_v, E, \mathbf{w}_e)$. We then formulate the problem of generating comparative summaries as the problem of finding the smallest subset of subgraphs whose union covers the Steiner tree. **Definition** Given a graph $\mathbb{G} = (V, E)$, a set of subgraphs S, and a Steiner tree T of $\mathbb{G}$, the subgraph cover problem (SGCP) is to find a minimum subgraph set $C \subset S$, whose union, $\mathbb{U} = (V_U, E_U)$, covers all the vertices and edges in T.

The SGCP problem is closely related to the set cover problem. The set cover problem (SCP), which is known as an NP-hard problem [Kar72], can be easily reduced to the SGCP problem. Reduction to SGCP Problem: Given a universe U, a set of elements $\{1, 2, ..., m\}$, and a family S of subset of U. We generate a fully connected graph $\mathbb{G} = (V, E)$ for each subset, where nodes are elements of subset and every pair of nodes has a edge. This construction can be done in polynomial time in the size of set cover instance.

Assume the universe U has a cover C with length k, where C is a smallest subfamily $C \subset S$ of sets whose union is U. Based on set cover C, we generate a set S of a fully connected graph G_i, where the vertex set of G_i is the same with C_i. Suppose we have a graph $T = (V_T, E_T)$, the vertex set V_T equals the union of C. It

is straightforward that the set S is the cover of T, because T is a subgraph of union of S and there is not smaller set of subgraph to cover all the vertex in T.

For the reverse direction, assume that $T = (V_T, E_T)$ has a subgraph cover S with length k. Let us only consider the vertex part of S, we can get a set C of k sets whose union equals V_T, the universe. This set will cover the universe, and thus the subgraph cover in $\mathbb{G}$ is a set cover in U. $\square$

The greedy algorithm for the set cover problem chooses sets according to one rule: choose the set that contains the largest number of uncovered elements at each iteration. It has been shown [Chv79] that this algorithm gets an approximation ratio of $H(s)$, where s is the size of the set to be covered, $H(m)$ is the m-th harmonic number:

$$H(m) = \sum_{j=1}^{m} \frac{1}{j} \leq \ln m + 1$$

4.4 Empirical Evaluation

4.4.1 Real World Data Set

Comparative patent document summarization is a novel application in patent retrieval, and hence there is no benchmark patent dataset for evaluation. In the experiment, a patent comparative summarization data set is provided by a patent agent company according to the real-world patentability or infringement analysis reports. The data set is composed of 300 pairs of US patents related to various topics, including "DOMESTIC PLUMBIN", "OPTICS DEVICE OR ARRANGEMENT", "INFORMATION STORAGE", under the administration of USPTO (http://www.uspto.gov). For each comparable patent pair, manual summaries are provided by three patent attorneys as the references.

4.4.2 Experimental Setup

To evaluate the quality of the generated summaries by automatic methods, we use ROUGE [LH03] as the metric, which has been widely used in document summarization evaluation. Given a system generated summary and a set of reference summaries, ROUGE measures the summary quality based on the unit overlap counting. In the experiment, for each summarization method, we calculate the averaged scores of ROUGE-1, ROUGE-2, ROUGE-W and ROUGE-SU over 300 pairs of patent documents.

For evaluation purpose, we perform preprocessing on patent documents, including stopwords removal, tokenization, stemming, etc. To emphasize the technical difference, we extract noun terms and phrases for each sentence in the documents. In practice, the number of features could vary depending on the size of the documents. For simplicity, we choose the top 20 discriminative features using χ^2 statistics for each patent document pair.

4.4.3 Results and Discussion

In the experiments, we start by using the features from different sections of patents to generate summaries. We then compare `PatentCom` with several baselines introduced in §4.2 from both quantitative and qualitative perspectives. Finally, we present an illustrative case study of using `PatentCom` to determine patentability. The results have been assessed and validated by patent analysts.

Summarization using Different Sections

A typical patent document often contains multiple sections, including summary of the invention, description of the preferred embodiments, claims, etc. Some sections may describe the invention in more details, whereas others may represent the idea

using abstractive terms. To evaluate how important of each section in delivering the comparative information, we generate the comparative summaries from different sections of patent documents, e.g., claims (CLM), embodiments (EMB), the summary of the invention (SUM), the combinations of these three sections and the entire patent document (ALL).

In Table 4.1, we report the averaged ROUGE scores of PatentCom for the summaries generated from different sections of patent pairs. **Bold** indicates the corresponding result is statistically significant. We observe that the best score is achieved by the summaries generated from combination of embodiment section and claim, because the claim section is the core part of the entire patent document and the embodiment of a patent document describes how the invention can be made and practiced in details, that contains sufficient resources to generate a comparative summary. Besides, it is not enough consider them separately, because claim is generally full of legal or domain-specific terminologies, and embodiment contains detail information without significance.

Table 4.1: Comparison of using different sections.

Sections	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-SU
CLM	0.5424	0.3306	0.1552	0.2230
SUM	0.4831	0.2642	0.1139	0.1961
EMB	0.4477	0.2317	0.0972	0.1460
CLM+SUM	0.5938	0.4174	0.2037	0.2887
CLM+EMB	**0.6078**	**0.4623**	**0.2244**	**0.3113**
EMB+SUM	0.4988	0.3007	0.1270	0.2171
ALL	0.6053	0.4593	0.2226	0.3093

Comparison with Existing Solutions

For comparison purpose, we implement the following document summarization methods: (1) Minimal Dominate Set Model (MDSM) [SL10], which selects the most representative sentences from each patent document; (2) Discriminative Sentence Selection

Model (DSSM) [WZLG12], which extracts comparative sentences via the method introduced in § 4.2.2, that is, to select the most discriminative sentences for describing the unique characteristics of each document; and (3) Comparative Summarization via Linear Programming Model (CSLPM) [HWX11], which considers cross-topic concept pairs as comparative evidences, and topic-related concepts as representative evidences, as introduced in § 4.2.2.

Table 4.2 shows the comparison results of different summarization methods, which are averaged ROUGE scores over 300 pairs of patent documents. We observe that (1) `PatentCom` achieves the best performance in terms of all the ROUGE scores by considering both commonalities and differences between two patent documents; (2) The performance of DSSM is not comparable with the other two methods, indicating that only considering the difference of the patent pair is not sufficient for this task, since such difference may not be significant or comparable; and (3) MDSM has similar ROUGE-1 with CSLPM, since MDSM selects importance sentences for each patent so that the summaries by MDSM contain frequent words used in patents, and may have significant overlap with reference summaries based on unigram. However, MDSM performs poorly on ROUEG-2, ROUGE-W and ROUGE-SU, as it does not match the purpose of this task.

Table 4.2: Comparison of different models.

Models	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-SU
MDSM	0.5210	0.3099	0.1499	0.2886
DSSM	0.4604	0.2645	0.1148	0.1583
CSLPM	0.5309	0.4066	0.2118	0.3015
PatentCom	**0.6053**	**0.4593**	**0.2226**	**0.3093**

To further illustrate the efficacy of comparative summarization approaches for the problem studied in our work, we conduct a case study of two comparable patent documents, US689,296,4 (as US689) and US775,796,9 (as US775). Both patents are related to the topic of "jet regulator", which distributes the incoming water flow into

individual jets. The difference between the two patents is that US775 provides an extra component called "deflection projection" which is used to keep the water jets away from aeration openings.

The comparative summaries generated by different methods are shown in Table 4.3. The result of MDSM misleads us to believe the major difference between two patents is that US775 contains a new "jet fractionating device" for dispersing water flow. However, US689 mentions "jet splitting device" which has similar functionality as "jet fractionating device"; on this aspect, both patents are similar. The reason here is straightforward: traditional summarization methods like MDSM try to capture the major information of the document, without considering whether the concepts are semantically identical. The differences identified by DSSM are trivial and the summaries are not comparable, and hence we cannot rely on this to decide whether US775 infringes US689 or not. From the summary by `PatentCom`, we observe that US775 contains "a deflection projection" and "cone-shape presieve", which are not described in US689. The reason why CSLPM misses "cone-shape presieve" is straightforward: "dirt" is a relatively low-frequency feature, which is difficult to find without considering the relationship between common and discriminative features. Such summaries provide informative information to patent analysts in a sense that there is a low probability that US775 infringes US689.

4.4.4 An Illustrative Case Study for Determining Patentability

Our proposed comparative summarization approach can serve as the basis of different patent retrieval tasks. As an example, we choose the task of determining patentability of a patent document to evaluate the efficacy of our proposed method, `PatentCom`. We conduct a real-world case study between a patent application US2013,0301,299

Table 4.3: Sample summaries by MDSM, DSSM, CSLPM, and PatentCom.

Patent	MDSM	DSSM
US689	A **jet regulator** comprising a **jet regulator housing** having an interior in which a jet regulation device is provided that has **passage openings**...Thereby, the projections on the support ring of the **insertable components** can be formed out of an un-deformed section of the metal sheet.	Metallic **insertable parts** can also be manufactured in small numbers especially economically...The **insertable components** of the **jet regulator** according to the invention can be manufactured in a simple manner using simple conventional manufacturing methods.
US775	A **jet regulator** comprising a **jet fractionating device** for dispersing an incoming water flow into a multitude of individual jets...Additionally the **jet regulator** may also be embodied as an aerated jet regulator with its jet regulator housing being provided at its exterior perimeter with at least one separate **aerating opening**.	the circular **deflecting projection** at its side facing away from the **aeration openings** in the flow direction is provided with an angled deflection surface...At the interior circumference of the housing, in the flow direction downstream in reference to the **aeration openings**, a **deflecting projection**is provided.

Patent	CSLPM	PatentCom
US689	The fluid stream that flows into the **jet regulator** is divided into a number of individual jets in the **jet splitting device**, which is designed as a **perforated plate**...A ventilated jet regulator has **ventilation openings** at the peripheral cover of its **jet regulator housing**.	A **jet regulator** comprising a **jet regulator housing** having an interior ... A ventilated jet regulator has **ventilation openings** at the peripheral cover of its **jet regulator housing**. In order to keep **dirt** particles out of the interior of the housing..., an **intake filter** is placed.
US775	A **jet regulator** has a **jet fractionating device** comprised of a **perforated plate**, which distributes the incoming water jet into a multitude of individual jets...At the interior circumference of the housing, in the flow direction downstream in reference to the **aeration openings**, a **deflecting projection** is provided.	A **jet regulator** comprising a **jet fractionating device** for dispersing an incoming water flow...in the flow direction downstream in reference to the **aeration openings**, a **deflecting projection** is provided...at the incoming side, are essentially provided upstream with a **coneshape presieve**, which separates the **dirt** particles entrained.

Table 4.4: Sample comparative summary for patentability analysis.

Patent	US253	US299
	The formation of the **molded pattern** on the **mold base** by the use of the positive-type **photosensitive heat-resistant resin** comprises the steps of coating the **mold base** with the positive-type photosensitive heat-resistant resin to form the photoresist film on its surface, pre-heating the photoresist film so as to harden slightly, exposing the applied photoresist film to light via the positive-type **pattern film** for forming the **optical pattern**.	Claim 1. A fabricating method of grid points on a **light guiding plate**, comprising following steps of: S1, forming a layer of **photosensitive material** on a mold for the light guiding plate; and S2, performing **photolithography** on the photosensitive material in order to form **grid points** on the light guiding plate. Claim 2. The method according to claim 1, wherein the photosensitive material is a **photosensitive resist**.
Patent	US520	US299
	a development step in which the **photosensitive heat-resistant resin layer** 12 exposed is developed; a rinsing step in which the portions removed by the development are rinsed away; and a baking step in which the pattern formed by the development is baked at a high temperature to cure the photosensitive heat-resistant resin and form a raised or depressed **pattern**...	Claim 5. The method according to claim 2, wherein the step of S2 further comprises following steps of: S21 using a film formed with **grid points** arrangement **pattern** as a mask, S22 sequentially performing exposing and developing process on the photosensitive resist in order to form a grid points pattern on the photosensitive resin, and S23 curing the photosensitive resist and removing **residual solvent** and moisture.

(US299) and the combination of US7,094,520 (as US520) and US6,663,253 (as US253). Both patents are related to the topic of "optical panel", which distributes the incoming light form light source over the entire upper face of the panel.

The comparative summaries generated by `PatentCom` are shown in Table 4.4. From the selected comparative summarizes, we observe that the combination of US520 and US253 disclose similar process for producing an optical panel molding die, which is described as light guild panel in US299. Such summaries provide informative information to patent analysts that there is a high probability that US520 and US253 will affect the patentability of US299.

4.5 Chapter Conclusion

In this chapter, we study the problem of comparing patent documents, which refers to examining the equivalence or coverage of two patent documents. We formulate this

problem as a comparative summarization problem, and propose a novel automatic comparative summarization approach, named `PatentCom`, to generate representative yet comparative summaries for given patent document pair. The generated summary is able to assist patent analysts in quickly understanding the relationship of two patents, and hence can help reduce the cost of different patent retrieval tasks. Extensive empirical evaluation on a collection of US patent documents demonstrates the effectiveness of our proposed approach. From the experiments we notice that features from different sections of patent documents may affect the performance of the summarization. For future work, we plan to consider the domain characteristics of patent documents, e.g., by assigning weights to different sections of a patent when selecting discriminative features.

CHAPTER 5

PatentDom: Analyzing Patent Relationships on Multi-View Patent Graphs

The fast growth of technologies has driven the advancement of our society. It is often necessary to quickly grasp the linkage between different technologies in order to better understand the technical trend. The availability of huge volumes of granted patent documents provides a reasonable basis for analyzing the relationships between technologies. In this paper, we propose a unified framework, named `PatentDom`, to identify important patents related to key techniques from a large number of patent documents. The framework integrates different types of patent information, including patent content, citations of patents, and temporal relations, and provides a concise yet comprehensive technology summary. The identified key patents enable a variety of patent-related analytical applications, e.g., outlining the technology evolution of a particular domain, tracing a given technique to prior technologies, and mining the technical connection of two given patent documents.

The rest of the paper is organized as follows. Section 5.1 presents the motivation of the key patent discovery. In Section 5.2, we formalize the problem and describe the algorithmic details of our proposed framework. We then present several potential patent-related applications and the corresponding solutions in Section 5.3. Empirical evaluation of our framework is reported in Section 5.4. Finally Section 5.5 concludes the paper.

5.1 Motivation

Technological innovation is becoming one of the important factors that stimulate the development of our society. Granted patents, as the major carrier for technology

documentation, have great potential to provide valuable insights of technologies. Analyzing patent documents enables us to effectively understand technological progress, comprehend the evolution of technologies and capture the emergence of new technologies [BM02, DRMG06].

One representative application of patent analysis involves that enterprises evaluate the prior art or technology evolution of a specific technical field in the development of new products. To conduct such an analysis, a key step is to identify important patents from a large number of related patent documents, where these patents can represent dominating technologies in the corresponding technical field [MB01]. In addition, for a technology company who maintains a large number of patents, it is often time-consuming and costly to manually examine these patents to identify the important ones for further maintenance. Automatic discovery of key patents from patent collections is able to help improve the efficiency and reduce the cost of patent portfolio management. Further, connecting the dots between the identified key patents enables a variety of patent analysis tasks.

In this chapter, we study the problem of mining dominating technologies from a large collection of patent documents. Previous research efforts [HSGA09, HHX+12, SCGJ05] tackle this problem via clustering or topic-based mining, where the key patents are essentially identified through content analysis. However, as a scientific means of technology documentation with legal significance and potential economic values, a patent document often has complex structures and special terminologies. The sophisticated patent language poses great challenges to automatic patent analysis, and hence it is difficult to identify key patents purely based on patent content.

In the domain of patent analysis, patent documents are often explicitly organized using citation links [HCP+09]. The citation relations among patents documents provide good indicators for the importance of patents. Representative work involves [WCL10, WC07], which utilizes the co-citation relations of patent documents

to identify key patents. However, citations among patent documents are usually sparse, which may result in the technology gap, and consequently hinder the comprehension of dominating technologies.

To address the aforementioned issues, in our work, we explore the possibility of integrating both patent content and citation relations in identifying key patents. To this end, we propose a unified framework, named `PatentDom`, in which multiple types of patent-related information are employed, including the content and citation relations of patent documents. The input to the system is a topic or a classification code relevant to a specific technical field. The system first retrieves all the patent documents related to the topic/code from a patent database. We then construct a multi-view patent graph in which patent content, citation relations and temporal orders are integrated. We model the problem of identifying key patents as a minimum-cost dominating set problem, and select key patents using an approximation algorithm. We further discover a list of patent-related problems based on the identified key patents. These problems can be resolved by considering the temporal order of patent documents and connecting the dots between the key patents through graph-based algorithms.

To the best of our knowledge, our work is the first journey towards unifying the process of understanding the linkage between different technologies in the domain of patent analysis, by considering both document content and citation relations of patents. In summary, the contributions of our work are three-fold:

- We present a unified framework to identify dominating technologies on a multi-view patent graph that synthesizes both patent content and citation relations.
- We apply the proposed framework to multiple patent-related analysis problems that aim to discover the linkage of patents, including:

  - `PatentLine`, i.e., to outline the technology evolution of a particular domain;
  - `PatentTrace`, i.e., to trace a given technique to previous related technologies;
  - `PatentLink`, i.e., to discover the technical connection of two given patent documents.

- We conduct extensive empirical evaluation on a collection of US patent documents, and the results demonstrate the efficacy of the framework.

5.2 Identifying Dominating Patents

In the domain of patent analysis, it makes more sense to restrict the scope to a particular technical field. Hence, given a classification code related to a specific technical field, we initially retrieve all available patent documents under the code from a patent database. The problem of identifying key patents can be defined as follows:

Problem 1. *Given a collection of granted patents $D = \{d_1, d_2, \ldots, d_n\}$, extract a subset of patents $P \subseteq D$, where $P = \{p_1, p_2, \ldots, p_m\}$ and each p_i denotes a key patent that can represent the dominating technology within the patent collection.*

Problem 1 gives us a generic definition of key patents, which can be used to describe the general problems of key patent discovery. In some cases, patent analysts expect to obtain important patents with respect to specific queries, e.g., a set of query patents. Then Problem 1 can be redefined as follows:

Problem 2. *Given a collection of granted patents $D = \{d_1, d_2, \ldots, d_n\}$ and a set of query patents $Q = \{q_1, q_2, \ldots, q_k\}$, extract a subset of patents $P \subseteq D$, where*

$P = \{p_1, p_2, \ldots, p_m\}$ *and each patent* p_i *is able to represent the dominating technology related to the query set* Q*.*

To address the aforementioned problems, we propose a unified framework, named `PatentDom`, which employs the minimum dominating set of a patent graph to represent the key patents. Specifically, we first construct a multi-view patent graph using the information of patent content, citation relations and temporal orders of patent documents, and then identify dominating/influential patents from the graph. Taking PROBLEM 1 as an example, we can assume that the extracted key patents should represent all the patent documents (i.e., every patent in the collection should be relevant to the extracted patents in terms of technologies). In other words, these key patents serve as a brief summary of the entire patent collection. Meanwhile, the number of these patents should be as small as possible. Such a summary of the patent collection under the above assumption is exactly the minimum dominating set of the patent graph. We hence model the problems as a minimum-cost dominating set problem, where the cost can be defined using different types of information, depending on the problem being solved. The framework is described in Figure 5.1.

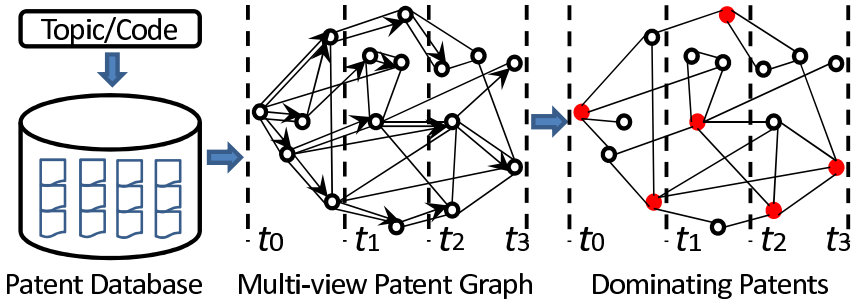


Figure 5.1: An overview of the PatDom.

5.2.1 Constructing Multi-View Patent Graph

As introduced in Section 5.1, the patent data consists of multiple types of information including patent content, citation relationship, and temporal order, that shape the

relations among patent documents. We use a multi-view graph $\mathbb{G}$ to represent these relations, where $\mathbb{G} = (V, \mathbf{w}_v, E_s, \mathbf{w}_s, E_{ct}, \mathbf{w}_{ct})$.

$\mathbb{G}$ contains a set of nodes/vertices (patent documents) V, where each node $v \in V$ is associated with a cost value w_v and a timestamp t. In our problem setting, the cost w_v can be defined using the information of patent content and/or citation relations. For example, to address PROBLEM 1, the cost can be calculated as the inverse of the total number of citations of the corresponding patent document, as we expect the selected patent is more influential than others. When selecting dominating nodes, the total cost of selected nodes should be minimized.

In addition, the vertices are connected by two types of edges: E_s and E_{ct}. E_s contains undirected edges, where each edge connects two patent vertices and the edge weight w_s denotes the content proximity of connected vertices. For patent documents, it is often difficult to calculate the similarity/proximity, as there are a lot of domain-specific and ambiguous terms, and different patents may have their own writing styles. To this end, we extract the most significant section of patents, i.e., `claims`, since this section defines the major invention of patents and often has relatively stable writing structures. We employ "bag-of-words" representation and the cosine measure for proximity computation. Two vertices are linked if and only if the content proximity is greater than a predefined threshold δ. In our proposed framework, E_s is used for dominating patent selection.

Another set of edges, E_{ct}, are directed edges, where each edge represents either the citation linkage between two vertices, or the temporal order of two vertices. Two vertices form a temporal link if and only if they do not have a citation link and their respective timestamp difference falls into a predefined time range $[\tau_1, \tau_2]$. For simplicity, we assign a unit value 1 to the weight of edges E_{ct}, i.e., $w_{ct} = 1$. E_{ct} serves to connect the selected dominating patents for specific patent applications. For example, to outline the technology evolution of a particular technical field, we

can employ E_{ct} to generate an evolution tree of dominating patents. Details can be found in Section 5.3 for different applications.

5.2.2 Identifying Dominating/Influential Patents

Our goal is to detect the patent documents with representative power, or say, dominating/influential patents. To this end, we define the problem on the undirected part, i.e., $(V, \mathbf{w}_v, E_s, \mathbf{w}_s)$, of the multi-view graph introduced in Section 5.2.1. Specifically, given the graph $\mathbb{G}$, a *dominating* set of $\mathbb{G}$ is a subset S of vertices with the following property: each vertex $v \in V$ is either in the dominating set S, or is adjacent to some vertices in S. Note that in $\mathbb{G}$, each vertex has a cost with respect to specific applications. The problem of finding a set of dominating patent documents can be formulated as the minimum-cost dominating set problem [CHL$^+$03, SL10].

Problem 3. *Given a graph $\mathbb{G} = (V, \mathbf{w}_v, E_s, \mathbf{w}_s)$ and a budget L, the problem of minimum-cost dominating set (MCDS) is to find a dominating set S, with size L, of vertices in $\mathbb{G}$ whose total vertex cost is the minimum.*

The MCDS problem is closely related to the problem of minimum dominating set (MDS). The vertex cover problem, which is known as an NP-hard problem, can be reduced to the MDS problem.

Reduction. Given a connected graph $\mathbb{G} = (V, E)$, we replace each edge of $\mathbb{G}$ by a triangle to create another graph $\mathbb{G}' = (V', E')$. In $\mathbb{G}'$, $V' = V \cup V_e$ where $V_e = \{v_{e_i} | e_i \in E\}$, and $E' = E \cup E_e$ where $E_e = \{(v_{e_i}, v_k), (v_{e_i}, v_l) | e_i = (v_k, v_l) \in E\}$. Such a transformation can be viewed as subdividing each edge (u, v) by the addition of a vertex, and adding an edge directly from u to v.

Assume $\mathbb{G}$ has a vertex cover S with size K, then S forms a dominating set in $\mathbb{G}'$. As each vertex v has at least one edge (v, u), and u must be in the cover if v is not. Since v is adjacent to u, then v has a neighbor in S.

For the reverse direction, assume that $\mathbb{G}'$ has a dominating set S' with size K, which only contains vertices from the vertex set V. If v_{e_i} is selected in S', then we can replace it by either v_k or v_l, without increasing the size of S'. We now claim that S' forms a vertex cover. For each edge e_i, v_{e_i} must have a neighbor (either v_k or v_l) in S'. This neighbor will cover the edge e_i, and thus the dominating set in $\mathbb{G}'$ is a vertex cover in $\mathbb{G}$. □

It has been shown that no algorithm can achieve an approximation factor better than $c \log |V|$ for some $c > 0$ [Kan92]. However, we can obtain a greedy approximation for MCDS, as shown in Algorithm 2. Starting from an empty set, if the current subset of vertices is not the dominating set, a new vertex with the minimum averaged cost (with respect to its neighbor size) and not adjacent to any vertex in the current set will be added. In other words, the cost of the new vertex can be evenly shared by its neighbors. Such a greedy algorithm provides a factor of $1 + \log |V|$ approximation of MCDS [RS97].

Algorithm 2 Approximation of MCDS.

Require: $\mathbb{G} = (V, \mathbf{w}_v, E_s, \mathbf{w}_s)$: undirected patent graph; L: predefined threshold of dominating patents
Ensure: **T**: a minimum-cost dominating set S
1: $S \leftarrow \emptyset$; $T \leftarrow \emptyset$
2: **while** $|S| < L$ **do**
3: **for** $v \in V - S$ **do**
4: $s(v) = |\{v' | (v', v) \in E_s\} \setminus T|$
5: **end for**
6: $v^* = \arg\min_v \frac{cost(v)}{s(v)}$
7: $S = S \cup \{v^*\}$; $T = T \cup \{v' | (v', v^*) \in E_s\}$
8: **end while**
9: **return** T

By Algorithm 2, we can obtain a set of dominating patents related to the specific technical field, with the limit of a predefined dominator number L. Note that in Algorithm 2, $cost(v)$ represents the value of $w(v)$, i.e., the cost of the vertex v. It

may be related to the citation relations as indicated in PROBLEM 1, or relevant to the query set as indicated in PROBLEM 2.

5.3 Potential Applications

The identified dominating patents from Section 5.2.2 enable a list of patent-related applications. In this section, we will discuss these applications from the perspective of connecting the dots between dominating patents.

5.3.1 Generating Tree-Based PatentLine

The first application is named as `PatentLine`, aiming to discover the technology evolution tree of a particular technical field. This problem has recently attracted increasing interest in the information retrieval community. Most existing approaches focus on identifying evolutionary topics in scientific literatures [BEG09, BEZG09] by making use of vector space model or LDA-alike topic models. Some recent work further tries to analyze the roles of linkage analysis (e.g., the co-authorship [ZJZG06] or citation analysis [HCP+09]) in topic detection and evolution. However, these existing methods cannot be simply applied to our problem setting of generating an evolutionary tree of patents. In addition, the characteristics of patent domain (e.g., lengthy and ambiguous description, full of technical terms) render these methods ineffective in generating patent evolution tree.

The dominating patents obtained from dominating set approximation are capable of representing the rest of patents in the graph in terms of content proximity and citation influence. Note that when utilizing Algorithm 2 to identify dominating patents, the cost of a vertex, i.e., $cost(v)$, is defined as the inverse of the total number of citations of the corresponding patent document, as we expect the selected patent is more influential than others. However, there might be some technical gaps among

these patents, that is, they may not be well connected. In order to provide a fluent structure of patent documents, e.g., a patentline, we have to find ways to link them together. Also, for presentation purpose, the generated structure of patent documents should be as dense and informative as possible, i.e., to include the minimum number of patents or have the maximum influence over other options.

To tackle this problem, we utilize the directed part, i.e., $(V, \mathbf{w}_v, E_{ct}, \mathbf{w}_{ct})$, of the multi-view graph introduced in Section 5.2.1. The procedure is depicted in Figure 5.2.

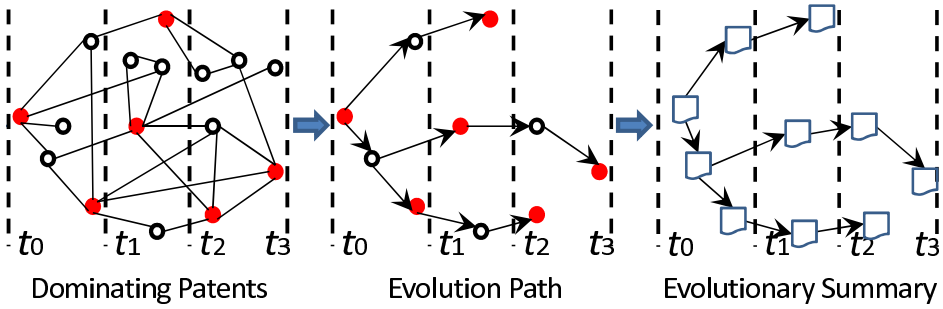


Figure 5.2: The procedure of PatentLine.

We formulate the problem as the minimum-cost Steiner tree problem. Given a graph $\mathbb{G}$ and a subset of vertices S, a Steiner tree of $\mathbb{G}$ is similar to minimum spanning tree, defined as the subtree of $\mathbb{G}$ that contains S with the minimum total cost. In our problem setting, the total cost is defined as the sum of vertex cost of the entire Steiner tree.

Problem 4. *Given a graph $\mathbb{G} = (V, \mathbf{w}_v, E_{ct}, \mathbf{w}_{ct})$, a vertex set $S \subset V$ (terminals) and a vertex $v_0 \in S$ from which every vertex of S is reachable in $\mathbb{G}$, the problem of minimum-cost Steiner tree (MCST) is to find the subtree of $\mathbb{G}$ rooted at v_0 that subsumes S with minimum total vertex cost.*

The problem of MCST, a directed version of the Steiner tree problem, is known as an NP-hard problem [Kar72]. As suggested by [CCC+98], a reasonable approximation can be achieved by finding the shortest path from the root to each terminal and then

Algorithm 3 $Steiner_i(\mathbb{G}, S, v_0, k)$ for a directed graph

Require: $\mathbb{G} = (V, \mathbf{w}_v, E_{ct}, \mathbf{w}_{ct})$: directed patent graph; S: terminal set; $v_0 \in S$: root of the Steiner tree; k: target size of terminals to be covered
Ensure: **T**: a Steiner tree rooted at v_0 covering at least k terminals

$T \leftarrow \emptyset$
while $k > 0$ **do**
  $T_{opt} \leftarrow \emptyset$;
  $cost(T_{opt}) \leftarrow \infty$
  for v, $(v_0, v) \in E_{ct}$, and k', $1 \leq k' \leq k$ **do**
    $T' \leftarrow Steiner_{i-1}(\mathbb{G}, S, v, k') \cup \{(v_0, v)\}$
    if $(cost(T_{opt}) > cost(T'))$ **then**
      $T_{opt} \leftarrow T'$
    end if
  end for
  $T \leftarrow T \cup T_{opt}$; $k \leftarrow k - |S \cap V(T_{opt})|$;
  $S \leftarrow S \setminus V(T_{opt})$
end while
return T

combining the paths, with the approximation ratio of $O(\log^2 k)$, where k is the number of terminals. The approximation algorithm is described in Algorithm 3.

The algorithm employs a recursive way to generate the Steiner tree T. It takes a level parameter $i \geq 1$. When $i = 1$, $Steiner_1$ is simple to describe, i.e., to find the k terminals which are the closest to the root v_0 and connect them to v_0 using shortest paths. As $i > 1$, $Steiner_i$ repeatedly finds a vertex v adjacent to the input root of the i-th function and a number k' such that the cost of the updated tree is the least among all the trees of this form. After obtaining the expected path, we update the corresponding Steiner tree, the target size k and the terminal set S.

The generated Steiner tree of the patent graph gives us an elegant representation of patent evolution, which describes the transitions from the root patent to all the other dominating patents. Once the Steiner tree is generated, we can easily obtain a concise summary for each patent in the tree by applying document summarization techniques.

5.3.2 Tracing Technologies To Ancestors

The second application is called `PatentTrace`, which aims to trace a given patent document back to its ancestors to examine what techniques that the given patent utilizes. This problem is relatively new in the domain of patent analysis. One major issue of modern patent documents is the growing complexity of the involved tasks, i.e., a single patent may contain a list of procedures and involve a lot of technologies. For such inventions, one may often need multiple research teams to develop different processes, and various inventions may be interlinked. Hence, to ease the understanding of patent analysts, it is imperative to identify key techniques related to the patent being investigated, and represent them in an informative manner.

To tackle this problem, we rely on the identified dominating patents based on the framework of `PatentDom`. The procedure of `PatentTrace` is described in Figure 5.3.

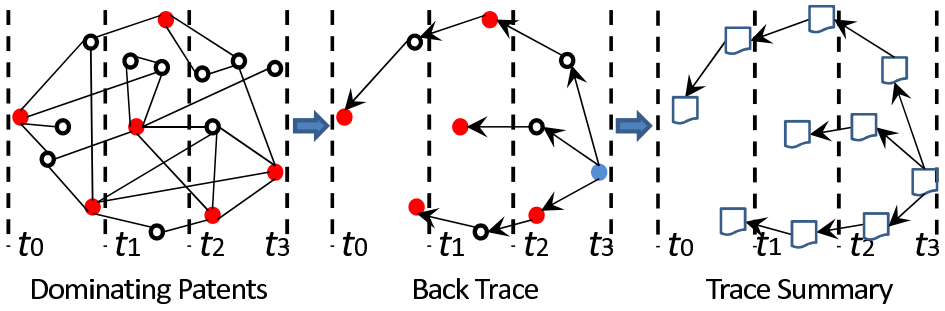


Figure 5.3: The procedure of PatentTrace.

Given a patent document as a query q, we first utilize Algorithm 2 to discover dominating patents based on the undirected part of the multi-view patent graph introduced in Section 5.2.1. Here we expect that the dominating patents are not only relevant to the query patent, but also reflect the important technologies. Hence in Algorithm 2, the cost of a vertex, i.e., $cost(v)$, should be defined in a way different from the one introduced in Section 5.3.1. To this end, we consider both content and

citation relations of patent documents, and define $cost(v)$ as

$$cost(v) = \frac{1 - sim(v, q)}{citation(v)}, \tag{5.1}$$

where the numerator denotes the content distance between the query patent q and the node v, and the denominator represents the citation count of the patent v. The similarity between patents is calculated using the content from `claims`, as indicated in Section 5.2.1. By Eq.(5.1), we expect to select the patents with content similar to the query patent, as well as with more citations to represent its influential power.

After identifying a list of dominating patents related to the given query, the next step is to connect these patents in order to provide a fluent trace from the query back to its ancestors. Some of the identified key patents may have a timestamp later than the one of the query patent, and hence they cannot be included in the final trace. To this end, we employ the directed part of the multi-view patent graph. Starting from the query node, we iteratively reverse the directed edges, and remove the nodes later than the query node, as well as the edges with opposite directions. The resulted subgraph $\mathbb{G}^*$ serves as the basis for trace generation.

Similar to `PatentLine`, we formulate the problem of tracing a patent to its ancestors as the minimum-cost Steiner tree problem. We then utilize Algorithm 3 to form the trace. The input is slightly different from the one in Section 5.3.1. v_0, as the root of the Steiner tree, is the query patent q. The terminal set S contains the dominators that are reachable from v_0 in the subgraph $\mathbb{G}^*$. The generated Steiner tree presents an informative representation of patent trace, which vividly describes the related ancestor technologies with respect to the query patent. Similar to `PatentLine`, we can generate a concise summary for each patent in the tree by applying document summarization techniques.

5.3.3 Discovering Technical Connections

The third application is named as `PatentLink`, aiming at discovering the potential relations between two patent documents. Given two patents p_1 and p_2 from different time periods, where p_1 is published earlier than p_2, they may not connect directly through citation relations. However, it is possible that p_2 is an implicit extension of p_1 in terms of technologies, or an application of the techniques described in p_1. Such latent connections are valuable for companies to design the corresponding product strategy. To the best of our knowledge, this problem has not yet attracted any research attention in the domain of patent analysis.

To address this problem, we first utilize the framework of `PatentDom` to identify dominating patents. Given a query set $Q = \{p_1, p_2\}$, we discover the dominating patents relevant to Q using Algorithm 2. The calculation of vertex cost is similar to Eq.(5.1). The only difference is the similarity score, which is computed as the averaged similarity between the vertex and the query patents.

The key patents are able to help connect the two query patents. However in the multi-view patent graph, multiple paths may exist between the given query patents. The challenge here is how to identify important paths in order to depict the strong connection between queries. In other words, how to find the nodes that are the center-piece, and have direct or indirect connections to all the query nodes? To this end, we employ the so-called center-piece subgraph [TF06, TFK07] and apply it to the direct part of the multi-view graph. We expand the query set Q by adding all the dominators falling in between the time period from p_1 and p_2. By doing this, the generated center-piece subgraph is able to show how the two patents are connected through leading technologies. The procedure is shown in Figure 5.4.

The algorithm `CEPS` described in [TFK07] for generating center-piece subgraph involves three steps: (1) calculating individual goodness score for a single node with

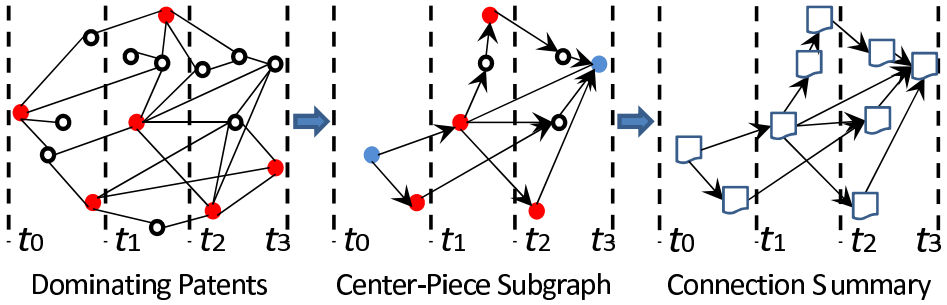


Figure 5.4: The procedure of PatentLink.

respect to each query node; (2) combining individual scores to obtain the goodness score for a single node with respect to the query set; and (3) extracting a connection subgraph maximizing the goodness criteria. The individual goodness score can be calculated using random walk with restart. Given a query p_i, a random particle starts from p_i, and then iteratively transmits to its neighborhood with the probability proportional to the edge weight between them, and also at each step, it has some probability c to return to the node p_i. Let $\mathbf{R}$ be the matrix containing the probability that the particle will finally stay at node p_i, then the matrix form of random walk calculation can be represented as

$$\mathbf{R}^T = c\mathbf{R}^T \times \mathbf{W} + (1-c)\mathbf{E},$$

where $\mathbf{E} = [\vec{e_i}](i = 1, \ldots, |Q|)$, and each $\vec{e_i}$ is the unit query vector with all zeros except one at row p_i. Notice that in our problem setting, we expand the query set by including appropriate dominators, and hence the corresponding dominators' entries are 1. $\mathbf{W}$ is the normalized adjacency matrix. Detailed procedure can be found in [TFK07].

5.4 Empirical Evaluation

In this section, we provide a comprehensive experimental evaluation to show the efficacy of our proposed framework `PatentDom`. We start with an introduction to

the patent collection used in the experiment. To validate the proposed framework, we compare our method with other existing solutions of identifying key patents. We further present several case studies to show the efficacy of the approaches for different applications.

5.4.1 Patent Data

The data set used in our experiment is provided by State Intellectual Property Office of the P.R.C (SIPO)[1], containing 16,518 US granted patents under the section G (physics), whose filing dates are ranging from 2001 to 2012. It covers three subdomains, including patents related to data processing system (G06Q 10/00), photomechanical production (G03F 7/00), and optical operation (G02F 1/00). The statistics of the data are depicted in Table 5.1. Under each patent code, there are a list of major patent groups, and each group contains at least 250 patents. Note that there is no standard patent data set that provides the ground truth of important patent documents with respect to a domain. Hence, for evaluation purpose, we ask patent analysts to manually select at least 20 key patents for each patent group as the ground truth.

Table 5.1: The description of patent data.

Domain Code	Groups	# of Patents	Average
G02F 1/00	17	11,218	660
G03F 7/00	6	2,922	487
G06Q 10/00	5	2,378	476

To conduct the experiment, we extract the title, claims, citations and publishing timestamp of each patent document, and preprocess the content using natural language processing techniques, such as removing stop words, tokenizing, and stemming.

[1]http://english.sipo.gov.cn.

The content of each patent is represented as a term vector, and the content proximity of patents is calculated using the cosine similarity for the purpose of similarity calculation. The citation relations are restricted in the patent collection.

5.4.2 Evaluation on PatentDom

In `PatentDom`, to construct the multi-view patent graph, we empirically set the content proximity threshold δ as 0.2, and the time range as 3 months. To evaluate our proposed framework, we implement three existing methods of identifying key patents as the baselines:

- `COA` [HSGA09]: It rates a patent based on its value by measuring the recency and impact of important phrases that appear in the `claims`. The score of a word w in a patent d is determined as follows:

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

  where $age(w)$ defines the recency of w, which is the time difference between the year w first occurs in the patent collection and the issue year of d; $support(w)$ is the number of follow-up patents that contain w. The score of d is the sum of scores of all the words in d. This method is based on the content and temporal information of patent documents.

- `PageRank` [Fuj07a]: It employs PageRank to rank patent documents, where the probability of accessing a patent is treated as the citation-based score for each document. This method is purely based on the citation relations of patents.

- `CorePatent` [HHX+12]: It aims to address the unique patent vocabulary usage problem by using a topic-based temporal mining approach to quantify a patent's novelty and influence. It initially identifies latent topics using an LDA-alike model [NASW09], and then examines the activeness of topics and removes noisy

Table 5.2: Comparison with existing methods. (Bold indicates the best performance. * indicates the statistical significance at $p < 0.01$.)

Methods	top@10			top@30			top@50		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
COA	0.11	0.056	0.07	0.092	0.138	0.11	0.086	0.215	0.123
PageRank	0.106	0.053	0.07	0.1	0.15	0.12	0.112	0.28	0.16
CorePatent	0.188	0.094	0.125	0.192	0.288	0.231	0.192	0.48	0.274
PatentDom	**0.194**	**0.097**	**0.129**	**0.22***	**0.33***	**0.263***	**0.212***	**0.53***	**0.3***

topics. Finally it quantifies patent novelty and influence, and ranks patents by their scores. This method utilizes both content and temporal information of patents.

The problem of identifying key patents is essentially a retrieval problem. For each method and each patent group, we rank and select top@10, top@30, top@50 patent documents based on its corresponding ranking criterion, and compare the results with the ground truth provided by patent analysts. For comparison, we compute the averaged precision, recall and F1-score of the entire 28 patent groups. The results are reported in Table 5.2.

As depicted in the table, our proposed framework, PatentDom, achieves the best performance compared with other baselines in terms of the precision, recall and F1-score. Especially for the recall, it significantly outperforms other methods. This is very valuable as the retrieval in the domain of patent analysis is a recall-based task. It is extremely important to have a higher recall in order to reduce the human efforts as well as to lower the risk of missing important patent documents.

We further examine the details of the results by investigating the content as well as citations of patent documents. Based on the analysis, we observe that PatentDom presents important patents of different time periods, and these patents are able to cover the dominating technologies in the corresponding domain without too much interlinking. Compared with PatentDom, the baselines provide partial or unreasonable key patents:

- Most patents in the results of `COA` fall into the earlier time periods, i.e., it only identifies key patents in the early years. However, there might be some patents serving as a connection link between the preceding and the following technologies, which are also important. `COA` fails to capture these patents, and hence its performance is comparatively worse than other baselines.
- `PageRank` only identifies important patents in the early and middle stages, due to the property of the PageRank algorithm. However in practice, technologies often evolve over time, and hence in recent stages we may have emerging technologies used by a lot of companies, which are also important in some sense.
- `CorePatent` discovers important patents from the topic-oriented perspective, and the results generated by this method are important in terms of the content. However, it fails to consider the citation relations of patent documents. Because of this, the identified key patents often center on several major technology companies, e.g., FujiFilm Corporation presents a lot of patents in photomechanical production. Nonetheless, these patents are usually related to each other with much more redundancy. This is the reason for which the performance of `CorePatent` is comparable to ours when the number of retrieved key patents is small, but is getting worse with more key patents.

5.4.3 Case Studies of Different Applications

Validating the efficacy of our proposed solutions to the three applications is a subjective process, as it is difficult to obtain annotated ground truth. We hence resort to case studies on the collected patent data. All the cases used in this section are reviewed by domain experts and are confirmed to be effective.

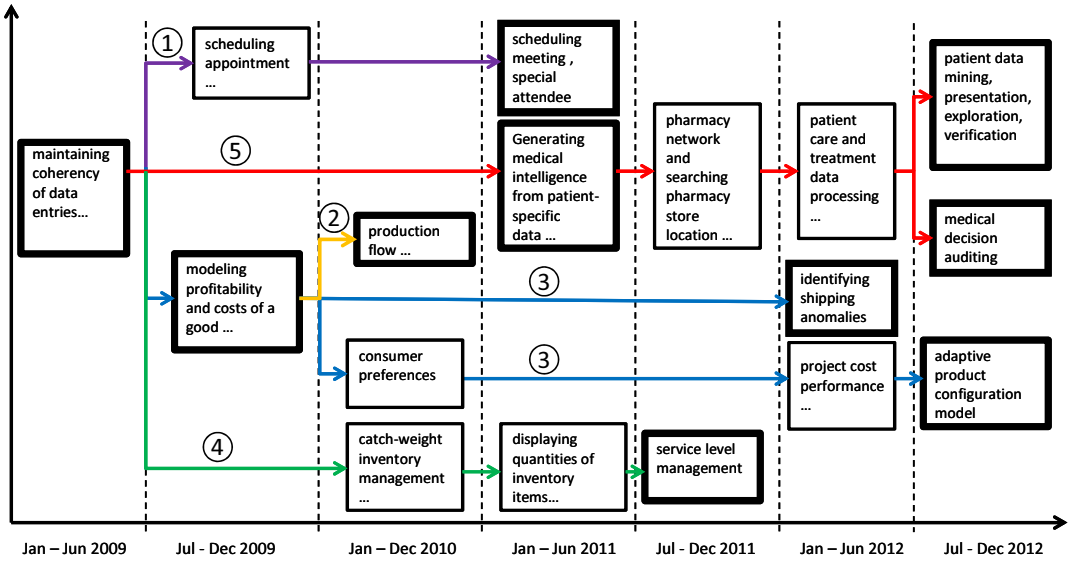


Figure 5.5: A case study of PatentLine.

A Case Study on `PatentLine`

`PatentLine` presents a way to explore the technology evolution of a specific technical field. To evaluate the efficacy of `PatentLine`, we perform a case study on a collection of patent data. The major international classification code of the patent data is "G06Q 10/00", representing the topic of "data processing systems or processes for administration and management of an organization, enterprise or employees". This code includes 5 sub-domains, and their descriptions are shown in Table 5.3.

Table 5.3: The description of patent classification.

Code	Description	# of Patents
G06Q 10/02	Reservations, e.g., meetings	288
G06Q 10/04	Forecasting or optimization	341
G06Q 10/06	Workflow management	404
G06Q 10/08	Inventory management	534
G06Q 10/10	Office automation	811

We run Algorithm 2 (limiting the number of dominators to be 10) and Algorithm 3 on the generated multi-view patent graph, and the resulted Steiner tree is demonstrated in Figure 5.5, organized by the temporal order of patents. For representation purpose, we only list the keywords that are contained in the title of patents. The

bold rectangles denote the dominators identified by Algorithm 2. The X-axis describes the publishing dates of the patents. As observed in Figure 5.5, "Management" in "G06Q 10/00" starts from manipulating data, as described in the first dominator, and then can be decomposed into several subtopics. The line labeled as **1** mainly describes meeting scheduling, which is related to "G06Q 10/02". The lines of **2** and **3** include production workflows and optimizing project, etc., which correspond to "G06Q 10/06" and "G06Q 10/04", respectively. The path labeled as **4** depicts some techniques of inventory and service management, which is relevant to "G06Q 10/08". These three evolution paths give us a general understanding of how technologies evolve with respect to the corresponding categories. These results have been reviewed and assessed by domain experts.

One interesting phenomenon in Figure 5.5 is the path of **5**, which describes the technologies of health care management, such as medical intelligence, patient treatment, etc. From Table 5.3 we cannot find a mapping between this topic and the available codes. We further check the detailed assignments of classification codes to the patents along this line, and find that besides "G06Q 10/00", the patents are all assigned to the code "G06Q 50/00", which includes the classification of health care and patient record management. It somehow indicates that "G06Q 50/00" is more suitable to these patents rather than "G06Q 10/00". The analysts may be able to obtain more insights by using our proposed framework.

A Case Study on `PatentTrace`

`PatentTrace` formalizes the problem of tracing back a given technology/patent. The purpose is to trace a given patent document back to its ancestors to investigate what techniques that the given patent utilizes. To validate the proposed solution for this problem, we use the patent data under the international classification code of "G02F 1/1335", which represents the structural association of optical devices, e.g., polarisers

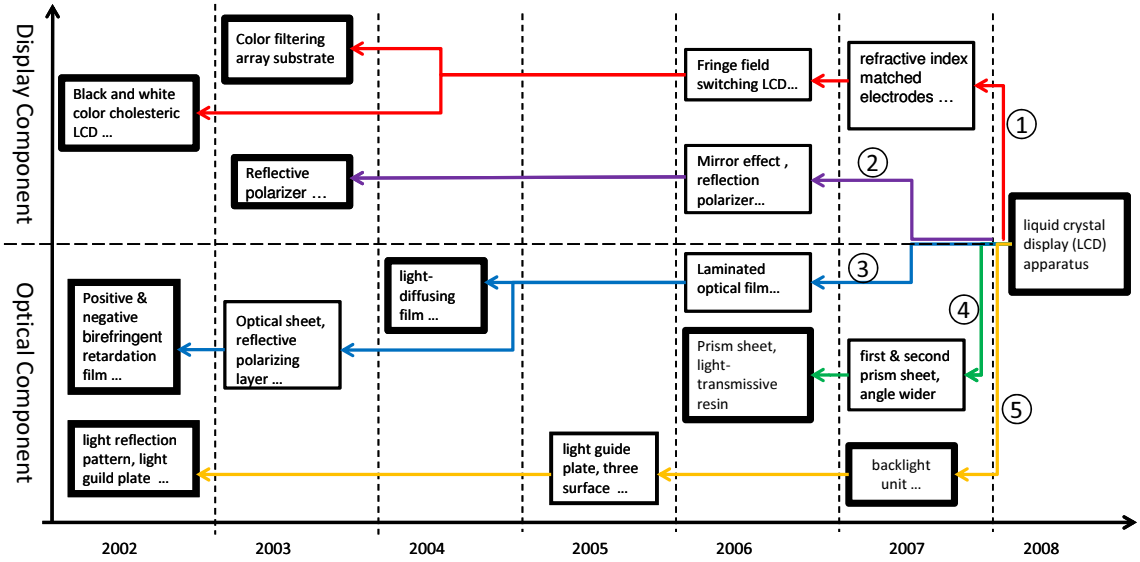


Figure 5.6: A case study of PatentTrace.

and reflectors. The data contains 3,080 patent documents. The query patent used in this case study is US8269915, which is related to a type of liquid crystal display apparatus (LCD), and was filed in 2008. Our goal is to examine what techniques are adopted in this product and how these techniques evolve to the product.

We treat US8269915 as a query and run the query-focused version of Algorithm 2 (limiting the number of dominators to be 20). We then run Algorithm 3 on the generated multi-view patent graph. The resulted back tracing Steiner tree is demonstrated in Figure 5.6. Similar to the case study of `PatentLine`, we only list the keywords of the title of patents for each patent document. The **bold** rectangles denote the dominators identified by query-focused MCDS. The X-axis represents the filing dates of patents.

This type of LCD contains two major components, i.e., the display and optical components. Our proposed solution to `PatentTrace` has successfully identified these two components (as depicted in Figure 5.6). For the display component, it involves polarized lighting plate (as indicated in the line of **2**) and color filtering array (described by the line of **1**). For the optical component, it consists of three major devices,

i.e., optical film (**3**), prism sheet (**4**), and back-light unit (**5**). The figure outlines the major constituent parts of LCD, and describes how related techniques evolve to the corresponding components. For example, as indicated by line **3**, the function of the optical film was originally fulfilled by birefringent retardation film, and then changed to reflective optical sheet, and finally laminated optical film. These results have been validated by patent analysts.

A Case Study on `PatentLink`

In practice, the linkage between two technologies is often achieved by technology evolution or technology application. The goal of `PatentLink` is to discover the details of evolution or application, in which the identified key patents serve to the ties that bind the technologies together. This would be very helpful for patent analysts to effectively understand the linkage between technologies.

To validate the efficacy of our solution to `PatentLink`, we present a case study on a collection of patent documents under the international classification code of "G03F 7/00", which represents the photomechanical production of textured or patterned surfaces. This data set contains 2,922 granted patents. We try to find the linkage between the patents US7771916 and US8053172. The former describes a polymerizable composition, which was filed in 2004; the latter proposes a method of forming a photoresist pattern using the photoresist composition, which was filed in 2008. The polymeriable composition is not directly used in the latter patent.

The experimental setup is similar to the one of `PatentTrace`. The resulted centerpiece subgraph is depicted in Figure 5.4. There are 4 dominators falling in between the filing time period of the two query patents. With the help of patent analysts, we can identify several interesting paths that reflect the technology evolution/application. For example, the path of the dotted line indicates how the technique of polymerizable

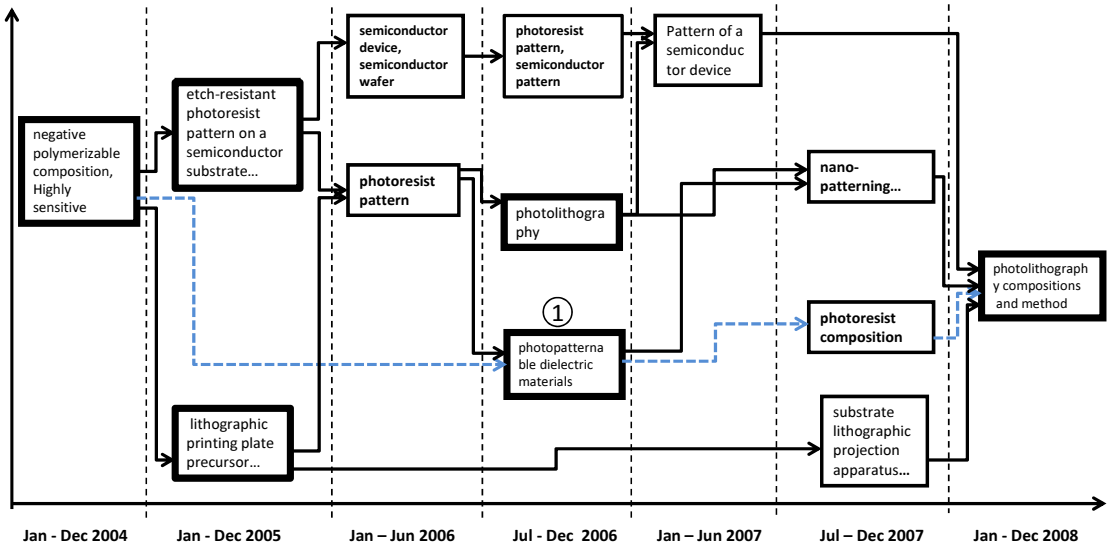


Figure 5.7: A case study of PatentLink.

composition evolves to the one of photresist composition, connected by the technique of photolithography in **1**.

5.5 Chapter Conclusion and Future Work

In this chapter, we study the problem of identifying dominating technologies using granted patent documents. Based on the analysis of domain characteristics of patents, we propose a unified framework, called `PatentDom`, to detect key patents from a large number of patents in a structural way. We formulate the problem as the minimum-cost dominating set problem, and employ graph-based optimization approaches to solve this problem. We further present potential applications of the proposed framework, including outlining the technology evolution of a particular domain (`PatentLine`), tracing a given technique to prior technologies (`PatentTrace`), and mining the technical connection of two given patent documents (`PatentLink`). Simple yet effective graph-based approaches are proposed based on the identified key patents as well as the requirements of the corresponding applications. Extensive empirical evaluation and

case studies on a collection of US patents demonstrate the efficacy and effectiveness of our proposed framework.

In our proposed framework, the cost of a vertex (patent) is defined based on the content and citation counts of the corresponding patent. It is interesting to extend it using external resources, such as patent examination results [HSN+12], patent maintenance decisions [JSC+11], and court judgments [LHL+11]. These resources explicitly indicate the relative importance of the patents, and hence are helpful to refine the definition of the cost. Further, to construct the multi-view patent graph, we utilize the content from `claims` to calculate the similarity. Due to the complex structure of patent documents as well as the diverse writing styles, the similarity may not represent the actual proximity between patents. We plan to explore semantic methods to improve the rationality of the edge weight in the undirected part of the graph.

The three applications introduced in Section 5.3 are all exploratory studies. In the domain of patent analysis, these applications are able to help patent analysts quickly identify the expected information without too much human effort, and make the corresponding decisions. It is worthy to provide quantitative measures to evaluate the generated results based on the requirement of the applications. In addition, we also plan to discover more applications/problems that can be solved using the dominating patents identified by `PatentDom`. Further, to ease the understanding, an interesting direction is to explore ways of visualizing the generated tree/graph based structures of patent documents.

CHAPTER 6

Conclusion

Patent mining aims at assisting patent analysts in investigating, processing, and analyzing patent document, which has attracted increasing interest in academia and industry. However, despite recent advances of patent mining, several critical issues in current patent mining systems have not been well explored in previous studies. These issues include: 1) the query retrieval problem that assists patent analysts finding all the relevant patent documents for a given patent application; 2) the patent documents comparative summarization problem that facilitates patent analysts in quickly reviewing any given patent document pairs; and 3) the key patent discovery problem that helps patent analysts to quickly grasp the linkage between different technologies in order to better understand the technical trend from a collection of patent documents.

For the issue of patent retrieval, a unified framework, named `PatSearch` is proposed, where the user submits the entire patent document as the query. Given a patent document, our framework will automatically extract representative yet distinguishable terms to generate a search query. In order to alleviate the issues of ambiguity and topic drifting, a novel query expansion approach is proposed, which combines content proximity with topic relevance. Our framework aims to help users retrieve relevant patent documents as many as possible, and provide enough information to assist patent analysts in making the patentability decision. The experimental evaluation demonstrates the effectiveness and efficacy of the proposed solution.

For the issue of patent documents comparison, we proposed a novel and comprehensive framework to model and compare given patent documents, named `PatCom`, which utilizes graph-based techniques to connect the dots among various aspects of the two patent documents on a term co-occurrence graph. When analyzing the retrieved patents for different retrieval tasks, our approach can serve as automatic baseline, and

consequently allow the analysts to quickly go through the results. To the best of our knowledge, our work is the first journey towards reducing human efforts of comparing patent documents by leveraging comparative summarization techniques. Extensive quantitative analysis and case studies on real world patent documents demonstrate the effectiveness of our proposed approach.

For the issue of key patents discovering, we proposed a unified framework of discovering dominant patent documents, named `PatDom`, in which multiple types of patent-related information are employed, including the content and citation relations of patent documents. The input to the system is a topic or a classification code relevant to a specific technical field. The system first retrieves all the patent documents related to the topic/code from a patent database. We then construct a multi-view patent graph in which patent content, citation relations and temporal orders are integrated. We model the problem of identifying key patents as a minimum-cost dominating set problem, and select key patents using an approximation algorithm. We further discover a list of patent-related problems based on the identified key patents. These problems can be resolved by considering the temporal order of patent documents and connecting the dots between the key patents through graph-based algorithms. To the best of our knowledge, our work is the first journey towards unifying the process of understanding the linkage between different technologies in the domain of patent analysis, by considering both document content and citation relations of patents. Empirical analysis and extensive case studies on a collection of US patent documents demonstrate the efficacy of our proposed framework.

In summary, this dissertation attempts to data mining techniques to resolve the patent mining issues in different aspects. As far as we know, this dissertation is the one of the earliest attempts that solves such issues from the analytic perspective instead from the system perspective.

Based on these initial exploration, we also found several limitation of the proposed works and there are some promising extensions can be done in the future. In our current method of key patents discovering, the cost of a vertex (patent) is defined based on the content and citation counts of the corresponding patent. It is interesting to extend it using external resources, such as patent examination results, patent maintenance decisions, and court judgments. These resources explicitly indicate the relative importance of the patents, and hence are helpful to refine the definition of the cost. Further, to construct the multi-view patent graph, we utilize the content from `claims` to calculate the similarity. Due to the complex structure of patent documents as well as the diverse writing styles, the similarity may not represent the actual proximity between patents. We plan to explore semantic methods to improve the rationality of the edge weight in the undirected part of the graph.

BIBLIOGRAPHY

[AANM91] M.B. 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.

[Ada01] Stephen Adams. Comparing the ipc and the us classification systems for the patent searcher. *World Patent Information*, 23(1):15–23, 2001.

[Alt99] Genrich S Altšuller. *The innovation algorithm: TRIZ, systematic innovation and technical creativity*. Technical Innovation Center, Inc., 1999.

[ASM11] B. Al-Shboul and S.H. Myaeng. Query phrase expansion using wikipedia in patent class search. *Information Retrieval Technology*, pages 115–126, 2011.

[AVJ10] 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.

[AY04] H.U. Atsushi and T. YUKAWA. Patent map generation using concept-based vector space model. *working notes of NTCIR-4, Tokyo*, pages 2–4, 2004.

[AYFD+11] Doreen Alberts, Cynthia Barcelon Yang, Denise Fobare-DePonio, Ken Koubek, Suzanne Robins, Matthew Rodgers, Edlyn Simmons, and Dominic DeMarco. Introduction to patent searching. In *Current challenges in patent information retrieval*, pages 3–43. Springer, 2011.

[BA10] Richard Bache and Leif Azzopardi. Improving access to large patent corpora. In *Transactions on large-scale data-and knowledge-centered systems II*, pages 103–121. Springer, 2010.

[BEG09] Levent Bolelli, Şeyda Ertekin, and C Lee Giles. Topic and trend detection in text collections using latent dirichlet allocation. In *Advances in Information Retrieval*, pages 776–780. 2009.

[BEZG09] Levent Bolelli, Seyda Ertekin, Ding Zhou, and C Lee Giles. Finding topic trends in digital libraries. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 69–72. ACM, 2009.

[BHHS12] Sumit Bhatia, Bin He, Qi He, and Scott 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.

[BM02] Anthony F Breitzman and Mary Ellen Mogee. The many applications of patent analysis. *Journal of Information Science*, 28(3):187–205, 2002.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[Bor08] Francesca Borgonovi. Divided we stand, united we fall: Religious pluralism, giving, and volunteering. *American Sociological Review*, 73(1):105–128, 2008.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[BR09] Shariq Bashir and Andreas 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.

[BR10] Shariq Bashir and Andreas Rauber. Improving retrievability of patents in prior-art search. In *Advances in Information Retrieval*, pages 457–470. Springer, 2010.

[Cam83] Richard S Campbell. Patent trends as a technological forecasting tool. *World Patent Information*, 5(3):137–143, 1983.

[Car12] Michael Carrier. A roadmap to the smartphone patent wars and frand licensing. *Antitrust Chronicle*, 4, 2012.

[CC12] Y.L. Chen and Y.C. Chang. A three-phase method for patent classification. *Information Processing & Management*, 2012.

[CCC+98] Moses Charikar, Chandra Chekuri, To-yat Cheung, Zuo Dai, Ashish Goel, Sudipto Guha, and Ming Li. Approximation algorithms for directed steiner problems. In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 192–200. ACM, 1998.

[CCC+99] Moses Charikar, Chandra Chekuri, To-yat Cheung, Zuo Dai, Ashish Goel, Sudipto Guha, and Ming Li. Approximation algorithms for directed steiner problems. *Journal of Algorithms*, 33(1):73–91, 1999.

[CGMEB12] Milen Chechev, Meritxell Gonzàlez, Lluís Màrquez, and Cristina 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.

[CH04] 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.

[CHL+03] Xiuzhen Cheng, Xiao Huang, Deying Li, Weili Wu, and Ding-Zhu Du. A polynomial-time approximation scheme for the minimum-connected dominating set in ad hoc wireless networks. *Networks*, 42(4):202–208, 2003.

[Chv79] Vasek Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235, 1979.

[CPP12] 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.

[CS12] Suleyman Cetintas and Luo 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.

[DRMG06] Tugrul U Daim, Guillermo Rueda, Hilary Martin, and Pisek Gerdsri. Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, 73(8):981–1012, 2006.

[EHO78] P Ellis, G Hepburn, and C Oppenhein. Studies on patent citation networks. *Journal of Documentation*, 34(1):12–20, 1978.

[ÉMS+12] 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.

[FI01] 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.

[FTBK03] C.J. 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.

[FTFK04] CJ Fall, A. Törcsvári, P. Fievet, and G. Karetka. Automated categorization of german-language patent documents. *Expert Systems with Applications*, 26(2):269–277, 2004.

[Fuj07a] Atsushi 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.

[Fuj07b] 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.

[FUYU09] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito 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.

[GGR+10] Julien Gobeill, Arnaud Gaudinat, Patrick Ruch, Emilie Pasche, Douglas Teodoro, and Dina 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.

[GL01] Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR*, pages 19–25. ACM, 2001.

[GLC+11] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K Tsou. Overview of the patent machine translation task at the ntcir-9 workshop. In *Proceedings of NTCIR*, volume 9, pages 559–578, 2011.

[GLJ11] D. Ganguly, J. Leveling, and G.J.F. 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.

[GLMJ11] D. Ganguly, J. Leveling, W. Magdy, and G.J.F. 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.

[GLS01] A.J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43:173–210, 2001.

[GMK+10] Harsha Gurulingappa, Bernd Müller, Roman Klinger, Heinz-Theodor Mevissen, Martin Hofmann-Apitius, Christoph M Friedrich, and Juliane Fluck. Prior art search in chemistry patents based on semantic concepts and co-citation analysis. In *TREC*, 2010.

[HAS10] Christopher G Harris, Robert Arens, and Padmini 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.

[HCC03] M.H. Huang, L.Y. Chiang, and D.Z. Chen. Constructing a patent citation map using bibliographic coupling: A study of taiwan's high-tech companies. *Scientometrics*, 58(3):489–506, 2003.

[HCP+09] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 957–966. ACM, 2009.

[HHX+12] Po Hu, Minlie Huang, Peng Xu, Weichang Li, Adam K Usadi, and Xiaoyan 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.

[HRH+10] 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.

[HRZ04] Djoerd Hiemstra, Stephen Robertson, and Hugo 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.

[HSGA09] Mohammad Al Hasan, W Scott Spangler, Thomas Griffin, and Alfredo 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.

[HSN+12] Shohei Hido, Shoko Suzuki, Risa Nishiyama, Takashi Imamichi, Rikiya Takahashi, Tetsuya Nasukawa, Tsuyoshi Idé, Yusuke Kanehira, Rinju Yohda, Takeshi Ueno, et al. Modeling patent quality: A system for large-scale patentability analysis using text mining. *Journal of Information Processing*, 20(3):655–666, 2012.

[HWX11] Xiaojiang Huang, Xiaojun Wan, and Jianguo Xiao. Comparative news summarization using linear programming. In *ACL-HLT*, pages 648–653. ACL, 2011.

[HX09] Qiu Honghua and Yu Xiang. Research on a method for building up a patent map based on k-means clustering algorithm [j]. *Science Research Management*, 2:009, 2009.

[IFKT03] 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.

[Ito04] Hideo Itoh. Ntcir-4 patent retrieval experiments at ricoh. In *NTCIR-4*, 2004.

[JAV10] Hideo Joho, Leif A Azzopardi, and Wim Vanderbauwhede. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Proceedings of the third symposium on Information interaction in context*, pages 13–24. ACM, 2010.

[Jin10] Yaohong 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.

[JLS+10] Charles Jochim, Christina Lioma, Hinrich Schütze, Steffen Koch, and Thomas 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.

[JSC+11] Xin Jin, Scott Spangler, Ying Chen, Keke Cai, Rui Ma, Li Zhang, Xian Wu, and Jiawei Han. Patent maintenance recommendation with patent information network model. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 280–289. IEEE, 2011.

[JT05] A.B. Jaffe and M. Trajtenberg. *Patents, citations, and innovations: A window on the knowledge economy*. MIT press, 2005.

[Kan92] Viggo Kann. *On the approximability of NP-complete optimization problems*. PhD thesis, Royal Institute of Technology Stockholm, 1992.

[Kar72] Richard M Karp. *Reducibility among combinatorial problems*. 1972.

[KC07] J.H. Kim and K.S. Choi. Patent document categorization based on semantic structural information. *Information processing & management*, 43(5):1200–1215, 2007.

[KCS10] Ashwathi Krishnan, Alfonso F Cardenas, and Derek 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.

[KHB+07] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard 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.

[Kis03] 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*. Citeseer, 2003.

[KKM+11] Shuhei Kondo, Mamoru Komachi, Yuji Matsumoto, Katsuhito Sudoh, Kevin Duh, and Hajime Tsukada. Learning of linear ordering prob-

lems and its application to je patent translation in ntcir-9 patentmt. In *Proceedings of NTCIR*, volume 9, pages 641–645, 2011.

[Kon05] Kazuya Konishi. Query terms extraction from patent document for invalidity search. In *Proc. of NTCIR*, volume 5, 2005.

[KS13] Ralf Krestel and Padhraic Smyth. Recommending patents based on latent topics. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 395–398. ACM, 2013.

[KSB03] C. Koster, M. Seutter, and J. Beney. Multi-classification of patent applications with winnow. In *Perspectives of System Informatics*, pages 111–125. Springer, 2003.

[KSC11] Youngho Kim, Jangwon Seo, and W Bruce Croft. Automatic boolean query suggestion for professional search. In *SIGIR*, pages 825–834. ACM, 2011.

[KSP08] Y.G. Kim, J.H. Suh, and S.C. Park. Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, 34(3):1804–1812, 2008.

[KZ09] Hyun Duk Kim and ChengXiang Zhai. Generating comparative summaries of contradictory opinions in text. In *CIKM*, pages 385–394. ACM, 2009.

[Lar97] L. Larkey. *Some issues in the automatic classification of US patents*. Massachusetts univ amherst Department of computer Science, 1997.

[Lar99] L.S. 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.

[LCSP12] Changyong Lee, Yangrae Cho, Hyeonju Seol, and Yongtae Park. A stochastic patent citation analysis approach to assessing future technological impacts. *Technological Forecasting and Social Change*, 79(1):16–29, 2012.

[LDL+98] M.L. Littman, S.T. Dumais, T.K. Landauer, et al. Automatic cross-language information retrieval using latent semantic indexing. *Cross-language information retrieval*, pages 51–62, 1998.

[LH03] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL-HLT*, pages 71–78. ACL, 2003.

[LHL+11] 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.

[LHS06] Han Tong Loh, Cong He, and Lixiang Shen. Automatic classification of patent documents for triz users. *World Patent Information*, 28(1):6–13, 2006.

[Li09] Yan-Ru Li. The technological roadmap of cisco's business ecosystem. *Technovation*, 29(5):379–386, 2009.

[LPH10] 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.

[LST07] Y. Li and J. Shawe-Taylor. Advanced learning algorithms for cross-language patent retrieval and classification. *Information processing & management*, 43(5):1183–1199, 2007.

[LYP09] 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.

[MAKC12] Parvaz Mahdabi, Linda Andersson, Mostafa Keikha, and Fabio Crestani. Automatic refinement of patent queries using concept importance predictors. In *SIGIR*, pages 505–514. ACM, 2012.

[MB01] Jacques Michel and Bernd Bettels. Patent citation analysis. a closer look at the basic input data from patent search reports. *Scientometrics*, 51(1):185–201, 2001.

[MB04] Sougata Mukherjea and Bhuvan 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.

[MC12] Parvaz Mahdabi and Fabio Crestani. Learning-based pseudo-relevance feedback for patent retrieval. In *Multidisciplinary Information Retrieval*, pages 1–11. Springer, 2012.

[McC02] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. 2002.

[Men05] Hsien-Chun 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.

[MGHC13] Parvaz Mahdabi, Shima Gerani, Jimmy Xiangji Huang, and Fabio 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.

[MHFI03] Mitsuharu Makita, Shigeto Higuchi, Atsushi Fujii, and Tetsuya 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.

[MJ10] W. Magdy and G.J.F. Jones. Applying the kiss principle for the clef-ip 2010 prior art candidate patent search task. 2010.

[MJ11a] W. Magdy and G.J.F. 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.

[MJ11b] Walid Magdy and Gareth JF 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.

[MKG+11] Parvaz Mahdabi, Mostafa Keikha, Shima Gerani, Monica Landoni, and Fabio Crestani. Building queries for prior-art search. In *Multidisciplinary Information Retrieval*, pages 3–15. Springer, 2011.

[MLJ10] W. Magdy, J. Leveling, and G. Jones. Exploring structured documents and query formulation techniques for patent retrieval. *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pages 410–417, 2010.

[MLJ11] W. Magdy, P. Lopez, and G. Jones. Simple vs. sophisticated approaches for patent prior-art search. *Advances in Information Retrieval*, pages 725–728, 2011.

[MPRA14] Antonio Messeni Petruzzelli, Daniele Rotolo, and Vito Albino. Determinants of patent citations in biotechnology: An analysis of patent influence across the industrial and organizational boundaries. *Technological Forecasting and Social Change*, 2014.

[MRS08] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[MWdR09] 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.

[NASW09] David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *The Journal of Machine Learning Research*, 10:1801–1828, 2009.

[NM12] Khanh-Ly Nguyen and Sung-Hyon Myaeng. Query enhancement for patent prior-art-search based on keyterm dependency relations and semantic tags. In *Multidisciplinary Information Retrieval*, pages 28–42. Springer, 2012.

[OLMY12] 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.

[OW11] K. OuYang and C.S. Weng. A new comprehensive patent analysis approach for new product design in mechanical engineering. *Technological Forecasting and Social Change*, 78(7):1183–1199, 2011.

[PEBD08] 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.

[PPM04] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In *NAACL-HLT*, pages 38–41. ACL, 2004.

[PRK11] Chitra Pasupathi, Baskaran Ramachandran, and Sarukesi Karunakaran. Selection based comparative summarization of search results using concept based segmentation. In *Trends in Network and Communications*, pages 655–664. 2011.

[PT74] Basic Patent Cooperation Treaty PCT and IIV TCO. Patent cooperation treaty. 1974.

[PT10] F. Piroi and J. Tait. Clef-ip 2010: Retrieval experiments in the intellectual property domain. In *Proc. of CLEF*, 2010.

[RDT99] Irving J Rotkin, Kendall J Dood, and Matthew A Thexton. *A history of patent classification in the United States Patent and Trademark Office*. Patent Documentation Society, 1999.

[RS97] Ran Raz and Shmuel Safra. A sub-constant error-probability low-degree test, and a sub-constant error-probability pcp characterization of np. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 475–484. ACM, 1997.

[RWB+96] Stephen E Robertson, Steve Walker, MM Beaulieu, Mike Gatford, and Alison Payne. Okapi at trec-4. In *Proceedings of the fourth text retrieval conference*, volume 500, pages 73–97, 1996.

[Sal71] Gerard Salton. The smart retrieval systemexperiments in automatic document processing. 1971.

[SBS08] Christian Sternitzke, Adam Bartkowski, and Reinhard Schramm. Visualizing patent statistics by means of social network analysis tools. *World Patent Information*, 30(2):115–131, 2008.

[SCGJ05] Benyah Shaparenko, Rich Caruana, Johannes Gehrke, and Thorsten Joachims. Identifying temporal patterns and key players in document collections. In *Proceedings of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM-05)*, pages 165–174, 2005.

[She03] Svetlana 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.

[SHG12] Benno Stein, Dennis Hoppe, and Tim 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.

[SJ13] Ruben Sipos and Thorsten Joachims. Generating comparative summaries from reviews. In *CIKM*, pages 1853–1856. ACM, 2013.

[SL10] Chao Shen and Tao Li. Multi-document summarization via the minimum dominating set. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 984–992. Association for Computational Linguistics, 2010.

[SOMI03] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto 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. ACL, 2003.

[SP09] Jong Hwan Suh and Sang Chan 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.

[SSMB97] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207, 1997.

[SWY75] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[TBT07] 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.

[Ter07] Ehara 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.

[TF06] Hanghang Tong and Christos Faloutsos. Center-piece subgraphs: problem definition and fast solutions. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 404–413. ACM, 2006.

[TFI04] Toru Takaki, Atsushi Fujii, and Tetsuya 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.

[TFK07] Hanghang Tong, Christos Faloutsos, and Yehuda Koren. Fast direction-aware proximity for graph mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 747–756. ACM, 2007.

[TFT+12] Amy JC Trappey, Chin Yuan Fan, CV Trappey, Yi-Liang Lin, and Chun-Yi 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.

[TGP+10] Douglas Teodoro, Julien Gobeill, Emilie Pasche, Dina Vishnyakova, Patrick Ruch, and Christian Lovis. Automatic prior art searching and patent encoding at clef-ip'10. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[TLL07] Yuen-Hsien Tseng, Chi-Jen Lin, and Yu-I Lin. Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216–1247, 2007.

[TR12a] Wolfgang Tannebaum and Andreas 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.

[TR12b] Wolfgang Tannebaum and Andreas Rauber. Analyzing query logs of uspto examiners to identify useful query terms in patent documents for query expansion in patent searching: a preliminary study. In *Multidisciplinary Information Retrieval*, pages 127–136. Springer, 2012.

[TR13] Wolfgang Tannebaum and Andreas Rauber. Mining query logs of uspto patent examiners. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 136–142. Springer, 2013.

[Tse05] Y.H. Tseng. Text mining for patent map analysis. *Catalyst*, 5424054(5780101):6333016, 2005.

[TTJ07] Y.H. Tseng, C.Y. Tsai, and D.W. Juang. Invalidity search for uspto patent documents using different patent surrogates. In *Proceedings of NTCIR-6 Workshop*, 2007.

[TUT05] Hironori Takeuchi, Naohiko Uramoto, and Koichi Takeda. Experiments on patent retrieval at ntcir-5 workshop. In *NTCIR-5*, 2005.

[TW08] Y.H. Tseng and Y.J. 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.

[TWY+12] 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.

[VMYK12] 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.

[VSTC03] 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.

[VWTR05] I. Von Wartburg, T. Teichert, and K. Rost. Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, 34(10):1591–1607, 2005.

[VZ11] N. Van Zeebroeck. The puzzle of patent value indicators. *Economics of Innovation and New Technology*, 20(1):33–62, 2011.

[WBD+06] Leo Wanner, Sören Brügmann, Boubacar Diallo, Mark Giereth, Yiannis Kompatsiaris, Emanuele Pianta, Gautam Rao, Pia Schoester, and Vasiliki Zervaki. Patexpert: Semantic processing of patent documentation. In *SAMT (Posters and Demos)*, 2006.

[WC06] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006.

[WC07] Shiao-Chun Wu and Hung-Yi Chen. Recognizing the core technology capabilities for companies through patent co-citations. In *Industrial Engineering and Engineering Management, 2007 IEEE International Conference on*, pages 2081–2085. IEEE, 2007.

[WCL10] Hsiao-Chun Wu, Hung-Yi Chen, and Kung-Yen Lee. Unveiling the core technology structure for companies through patent information. *Technological forecasting and social change*, 77(7):1167–1178, 2010.

[wip11] Intellectual property statistics. 2011.

[WO06] Jianqiang Wang and Douglas 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.

[WP05] R Polk Wagner and Gideon Parchomovsky. Patent portfolios. *U of Penn. Law School, Public Law Working Paper*, 56:04–16, 2005.

[WZLG12] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. Comparative document summarization via discriminative sentence selection. *TKDD*, 6(3):12, 2012.

[XC96] Jinxi Xu and W Bruce 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.

[XC09a] X. Xue and W.B. Croft. Automatic query generation for patent search. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 2037–2040. ACM, 2009.

[XC09b] Xiaoibng Xue and W Bruce Croft. Transforming patents into prior-art queries. In *SIGIR*, pages 808–809. ACM, 2009.

[XCL+08] 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.

[YAKBY08] YunYun Yang, Lucy Akers, Thomas Klose, and Cynthia Barcelon Yang. Text mining and visualization tools–impressions of emerging capabilities. *World Patent Information*, 30(4):280–293, 2008.

[YAY+10] Yun Yun Yang, Lucy Akers, Cynthia Barcelon Yang, Thomas Klose, and Shelley Pavlek. Enhancing patent landscape analysis with visualization output. *World Patent Information*, 32(3):203–220, 2010.

[YLP03] T. Yeap, G.H. 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.

[YP97] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.

[YP04] Byungun Yoon and Yongtae 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.

[YPK13] Janghyeok Yoon, Hyunseok Park, and Kwangsoo Kim. Identifying technological competition trends for r&d planning using dynamic patent maps: Sao-based content analysis. *Scientometrics*, 94(1):313–331, 2013.

[YYP02] Byung-Un Yoon, Chang-Byung Yoon, and Yong-Tae Park. On the development and application of a self–organizing feature map–based patent map. *R&D Management*, 32(4):291–300, 2002.

[ZJZG06] Ding Zhou, Xiang Ji, Hongyuan Zha, and C Lee Giles. Topic evolution and social interactions: how authors effect research. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 248–257. ACM, 2006.

[ZL01] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001.

[ZL13] Longhui Zhang and Tao Li. Data mining applications in patent analysis. In *Data mining where theory meets practice*, pages 392–416. Xiamen University Press, 2013.

[ZLL15] Longhui Zhang, Lei Li, and Tao Li. Patent mining: A survey. *ACM SIGKDD Explorations Newsletter*, 16(2):1–19, 2015.

[ZLLZ14] Longhui Zhang, Lei Li, Tao Li, and Qi Zhang. Patentline: analyzing technology evolution on multi-view patent graphs. In *SIGIR*, pages 1095–1098. ACM, 2014.

VITA

LONGHUI ZHANG

2012 – Now	Ph.D., Computer Science Florida International University Miami, Florida, U.S.A.
2012	M.S., Computer Science Florida International University Miami, Florida, U.S.A.
2009	B.E., Computer Science University of Shanghai for Science and Technology Shanghai, China

PUBLICATIONS & PRESENTATIONS

- Longhui Zhang and Tao Li. "Data mining application in patent mining". In Data Mining Where Theory Meets Practice. Xiamen University Press, 2013, Page 38-64, ISBN 9787561542941.
- Longhui Zhang, Lei Li, Tao Li, and Qi Zhang. "Patentline: Analyzing technology evolution on multi-view patent graphs". In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 1095-1098. ACM, 2014.
- Longhui Zhang, Lei Li, and Tao Li. "Patent mining: A survey". ACM SIGKDD Explorations Newsletter 16, no. 2 (2015): 1-19.
- Longhui Zhang, Lei Li, Chao Shen, and Tao Li. "PatentCom: A Comparative View of Patent Document Retrieval". SDM, 2015.
- Longhui Zhang, Lei Li, Tao Li, and Dingding Wang. "Patentdom: Analyzing patent relationships on multi-view patent graphs". In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1369-1378. ACM, 2014.
- Lei Li, Chao Shen, Long Wang, Li Zheng, Yexi Jiang, Liang Tang, Hongtai Li, Longhui Zhang and Chunqiu Zeng. "iMiner: Mining Inventory Data for Intelligent Management". ACM Conference on Information and Knowledge Management (CIKM), Pages 2057 - 2059, 2014.
- An Integrated Framework for Patent Analysis and Mining, Doctoral Student Forum, SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, May, 2015.
- PatentCom: A Comparative View of Patent Document Retrieval, SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April, 2015.

- Patentdom: Analyzing patent relationships on multi-view patent graphs, Graduate Student Association Scholarly Forum, FIU, Miami, FL, April, 2015.
- pReader: Knowledge-based Recommendation Framework for technique articles understanding, School of Computing and Information Sciences Seminar Presentation, FIU, Miami, FL, Feb, 2014.