PatentCom: A Comparative View of Patent Document Retrieval

Longhui Zhang, Lei Li, Chao Shen, and Tao Li

School of Computing and Information Sciences, Florida International University

Email: {lzhan015, lli003, cshen001, taoli}@cs.fiu.edu

Abstract

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 paper, 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. Extensive quantitative analysis and case studies on real world patent documents demonstrate the effectiveness of our proposed approach.

Keywords: Patent Retrieval; Patentability Search; Patent Infringement; Comparative Summarization

1 Introduction

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 [24].

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. [1]. 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 the results is not a trivial task. For instance, the task of determining patentability involves analyzing prior patent documents 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 and legal 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 [1, 4, 17], 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 [5, 14]. However,

most summarization methods cannot provide comparative information. Recently, comparative summarization [21], 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 [6, 21] 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 regarding their differences;
- We conduct extensive evaluation on a collection of US patent documents, and the results demonstrate the effectiveness of our proposed approach.

The rest of the paper is organized as follows. In §2 we discuss existing patent retrieval solutions that provide refined search results for the analysts. In §3 we formulate the problem, and explore possible solutions that provide comparative summaries. In §4 we present our graph-based comparative summarization approach, `PatentCom`. Empirical evaluation is conducted and reported in §5. Finally, §6 concludes our work.

2 Related Work

Patent retrieval is essentially different from searching general web documents due to the characteristics of patent documents and special requirements of patent retrieval tasks [20]. In the last decade, a lot of research work has been published in the domain of patent retrieval, e.g., generating search queries [9, 22], expanding queries [4, 11], technology evolution analysis [26], key patent discovery [25], etc. In this paper, we focus on analyzing retrieved patent documents to improve the readability, which has not been well explored in the community of patent retrieval. In the following, we highlight the previous research that are most relevant to our work.

A patent document is often of rich content, consisting of descriptions, embodiments, claims, etc. The lexical content, as well as the structure of a patent document, is often the obstacle that makes it difficult to read. To ease the understanding of patent documents, Shinmori et al. [17] utilize nature language processing methods to reduce the structural complexity. Sheremetyeva [16] proposes similar approach to capture both the structure and lexical content of claims from US patent documents. Although they achieve a promising performance for improving the readability of patent document, human efforts are not significantly reduced for comparing given patent documents.

Another direction of refining search results is to use summarization techniques to represent original patent documents. In [20], Tseng et al. utilize an extractive summarization method that selects sentences based on occurrence of keywords, title words, and clue words contained in the document. Trappey et al. [19] employ a clustering-based approach that combines the ontological concepts and vector space models. The ontology captures the general concepts of patents in a given domain. Then, the proposed methodology extracts, clusters, and integrates the content of a patent document to derive a summary and a tree diagram of key terms. These approaches might be able to capture the major information of a patent; however, they are not suitable to highlight the differences of two patent documents.

Our work is orthogonal to the aforementioned approaches. By presenting the comparative information, we are able to provide strong evidence for patent analysts of the difference between patent documents. Based on such evidence, patent analysts can quickly determine whether the idea of a patent application has been disclosed by previously granted patents, or whether a product-related patent documents uses almost the same idea of another patent, etc.

3 Problem Statement and Possible Solutions

In this section, we first formally define the problem under the setting of summarization, and then explore possible solutions to this problem.

3.1 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.

3.2 Existing Solutions Recently, a list of approaches have been reported to tackle the problem of comparative summarization [6, 8, 12, 18, 21]. 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.

3.2.1 Selection via Difference The extraction-based summarization process generally involves selecting sentences from documents [14]. 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 [21], 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

$$\mathcal{H}(Y|X_S) \stackrel{\text{def}}{=} -\mathbf{E}_{p(Y,X_S)}(\ln p(Y|X_S)), \tag{3.1}$$

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 [21] 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.

3.2.2 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 [6], 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. [6] 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

$$\lambda \sum_{j=1}^{|C_1|} \sum_{k=1}^{|C_2|} u_{jk} \cdot op_{jk} + (1-\lambda) \sum_{i=1}^{2} \sum_{j=1}^{|C_i|} w_{ij} \cdot op_{ij}, \tag{3.2}$$

which is a linear combination of the representativeness and the comparative importance. The first term in Eq.(3.2) 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 [13]). The optimization problem of Eq.(3.2) can be solved using linear programming, as indicated in [6].

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 Our Approach: 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 1 presents an overview of our proposed approach. It contains 4 major modules, described as follows.

1. *Selecting Discriminative Features* (§4.1): 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* (§4.2): 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): 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* (§4.4): 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.

4.1 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 [23].

4.2 Feature Graph Construction The discriminative features from §4.1 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].

4.3 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

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

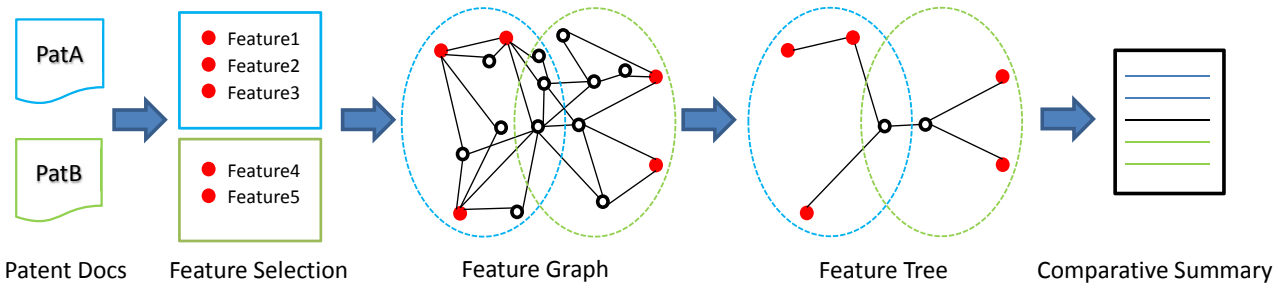

Figure 1: An overview of PatentCom.

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 §4.2) and a subset of vertices S (the discriminative features in §4.1), 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 4.1. *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 [7]. As suggested by [2], 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.

To solve this problem, we employ 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 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.

4.4 Comparative Summarization Generation

The Steiner tree obtained from §4.3 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 4.2. *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[7], can be easily reduced to the SGCP problem. Please refer to the appendix for the reduction. 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 [3] 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:

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

5 Empirical Evaluation

5.1 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.

5.2 Experimental Setup To evaluate the quality of the generated summaries by automatic methods, we use ROUGE [10] 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.

5.3 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 §3 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.

5.3.1 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 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 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

5.3.2 Comparison with Existing Solutions For comparison purpose, we implement the following document summarization methods: (1) Minimal Dominate Set Model (MDSM) [15], which selects the most representative sentences from each patent document; (2) Discriminative Sentence Selection Model (DSSM) [21], which extracts comparative sentences via the method introduced in § 3.2.1, 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) [6], which considers cross-topic concept pairs as comparative evidences, and topic-related concepts as representative evidences, as introduced in § 3.2.2.

Table 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 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 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.

5.4 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 (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. 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.

6 Conclusion

In this paper, 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.

APPENDIX

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

Table 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 **cone-shape presieve**, which separates the **dirt** particles entrained.

Table 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.

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. □

ACKNOWLEDGMENT

The work was supported in part by the National Science Foundation under grants DBI-0850203, CNS-1126619, and IIS-1213026, the U.S. Department of Homeland Security under Award Number 2010-ST-06200039, the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001, and the Army Research Office under grants W911NF-10-1-0366 and W911NF-12-1-0431.

References

[1] D. Alberts, C. B. Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, and D. DeMarco. Introduction to patent searching. In *Current challenges in patent information retrieval*, pages 3–43. 2011.

[2] M. Charikar, C. Chekuri, T.-y. Cheung, Z. Dai, A. Goel, S. Guha, and M. Li. Approximation algorithms for directed steiner problems. *Journal of Algorithms*, 33(1):73–91, 1999.

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

[4] A. Fujii. Enhancing patent retrieval by citation analysis. In *SIGIR*, pages 793–794. ACM, 2007.

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

[6] X. Huang, X. Wan, and J. Xiao. Comparative news summarization using linear programming. In *ACL-HLT*, pages 648–653. ACL, 2011.

[7] R. M. Karp. *Reducibility among combinatorial problems*. Springer, 1972.

[8] H. D. Kim and C. Zhai. Generating comparative summaries of contradictory opinions in text. In *CIKM*, pages 385–394. ACM, 2009.

[9] Y. Kim, J. Seo, and W. B. Croft. Automatic boolean query suggestion for professional search. In *SIGIR*, pages 825–834. ACM, 2011.

[10] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL-HLT*, pages 71–78. ACL, 2003.

[11] P. Mahdabi, L. Andersson, M. Keikha, and F. Crestani. Automatic refinement of patent queries using concept importance predictors. In *SIGIR*, pages 505–514. ACM, 2012.

[12] C. Pasupathi, B. Ramachandran, and S. Karunakaran. Selection based comparative summarization of search results using concept based segmentation. In *Trends in Network and Communications*, pages 655–664. 2011.

[13] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In *NAACL-HLT*, pages 38–41. ACL, 2004.

[14] G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207, 1997.

[15] C. Shen and T. Li. Multi-document summarization via the minimum dominating set. In *Computational Linguistics*, pages 984–992. ACL, 2010.

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

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

[18] R. Sipos and T. Joachims. Generating comparative summaries from reviews. In *CIKM*, pages 1853–1856. ACM, 2013.

[19] A. J. Trappey, C. V. Trappey, and C.-Y. Wu. Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering*, 18(1):71–94, 2009.

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

[21] D. Wang, S. Zhu, T. Li, and Y. Gong. Comparative document summarization via discriminative sentence selection. *TKDD*, 6(3):12, 2012.

[22] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In *SIGIR*, pages 808–809. 2009.

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

[24] L. Zhang, L. Li, and T. Li. Patent mining: A survey. *SIGKDD explorations*, 2015. to appear.

[25] L. Zhang, L. Li, T. Li, and D. Wang. Patentdom: Analyzing patent relationships on multi-view patent graphs. In *CIKM*, pages 1369–1378. ACM, 2014.

[26] L. Zhang, L. Li, T. Li, and Q. Zhang. Patentline: analyzing technology evolution on multi-view patent graphs. In *SIGIR*, pages 1095–1098. ACM, 2014.