Scientometrics (2022) 127:1969–1993
https://doi.org/10.1007/s11192-022-04307-8

One-to-many comparative summarization for patents

Zheng Liu[1] · Jialing Zhang[1] · Tingting Qin[1] · Yanwen Qu[2] · Yun Li[1]

Received: 12 August 2021 / Accepted: 10 February 2022 / Published online: 2 March 2022

Abstract

Patents bring technology companies commercial values in modern business operations. However, companies have to bear the high cost of handling patent applications or infringement cases. A common yet expensive task among these jobs is to analyze relevant patent literature. Lengthy and technically complicated patents require a large number of human efforts. This paper focuses on automatically analyzing the similar contents between a patent and its relevant literature, relevant patents specifically, to help experts review the similarities among these patents. We formulate this as a one-to-many document comparison problem by generating a comparative summary of a given patent and its relevant patents. We extract essential technical features from semantic dependency trees based on sentences in claims and construct a multi-relational graph to model the relevance between features and patents. The key to generating the comparative summary is selecting comparative essential technical features, which we formulate as an optimization problem and solve by a fast greedy algorithm. Experiments on real-world datasets and case studies demonstrate the effectiveness and efficiency of the proposed methods.

Keywords Patent comparison · Essential technical feature · Semantic dependency tree · Feature-patent relevance graph

Introduction

Technology companies invest tremendous money on innovations to research and develop new and competitive technologies. Patents allow companies to enjoy the exclusive right of applying these technologies legally and serve as the moat of their revenues, leading to an increasing number of patent applications worldwide, especially in cutting-edge technologies such as 5G wireless telecommunications. Many companies have laid many human resources on patent-related intellectual property affairs, such as handling patent applications and infringement cases. These jobs take domain experts with professional backgrounds a massive number of working hours.

✉ Zheng Liu
zliu@njupt.edu.cn

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China

[2] School of Computer Information and Engineering, Jiangxi Normal University, Nanchang, China

However, the overwhelming volume of patents, the long and obscure sentences in patent documents, and the complex content of technical characteristics bring difficulties. Manually analyzing all patent documents is almost impossible. How to facilitate experts to save their efforts in the analysis process by utilizing computing technologies is an area with great potentials (Shalaby & Zadrozny, 2019; Zhang et al., 2018; Abbas et al., 2014; Lee et al., 2013).

In this paper, we focus on the problem of automatically comparative analysis of patent documents. Comparing similar contents is an essential task in patent analysis applications such as patentability determination and patent infringement cases. In patentability determination, patent comparative analysis can reveal the claims disclosed by other literature and help to verify the novelty and inventiveness of patent applications. In patent infringement cases, analogous characteristics discovered by patent comparative analysis can identify the infringed claims or contribute to the invalidation of patents employed in lawsuits.

Patent comparative analysis generates a concise and comparative summary by identifying similar contents among patent documents from various granularities, such as words, sentences, and topics. A comparative summary is not like the traditional document summarization in information retrieval, which intends to generate a summary covering the main topics (Souza et al., 2021; Wang et al., 2009; Gong & Liu, 2001). A small number of current research works focus on comparative summary (Zhang et al., 2015b; Wang et al., 2012) in terms of recapitulating the differences among documents by selecting the most discriminative sentences representing the document characteristics.

Because of the disparate objectives, there is no guarantee that applying the above methods directly in patent comparative analysis would yield summaries containing similar contents from different patents. Comparative summarization in information retrieval tends to select discriminative sentences from documents, while patent comparative analysis does exactly the opposite. Besides, most research efforts of comparative document analysis concentrate on one-to-one document comparison in a large text corpus and use words, phrases, or sentences as the characteristic features in the resulting summaries (Wang et al., 2012), which is not appropriate under the context of patents. We will see soon that essential technical features in patents are the keystones in the following of this paper.

We study the problem of one-to-many document analysis for patent comparison by leveraging essential technical features. Fig. 1 shows an example of a generated patent comparative summary. For a target patent and the comparative patents, the comparative summary

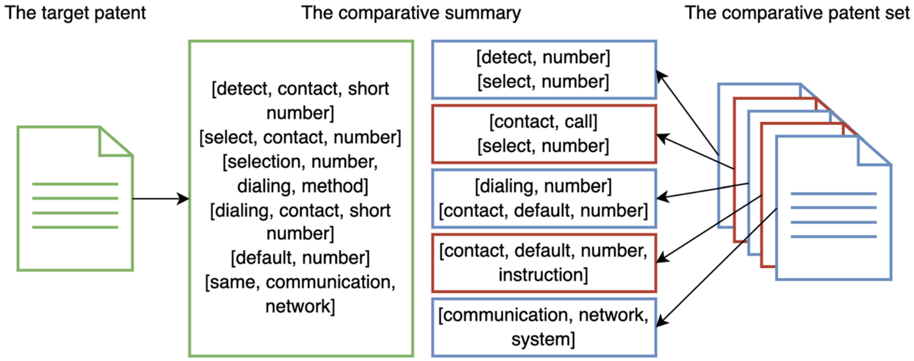
The target patent
The comparative summary
The comparative patent set
[detect, contact, short number]
[select, contact, number]
[selection, number, dialing, method]
[dialing, contact, short number]
[default, number]
[same, communication, network]
[detect, number]
[select, number]
[contact, call]
[select, number]
[dialing, number]
[contact, default, number]
[contact, default, number, instruction]
[communication, network, system]

Fig. 1 An example of patent comparative analysis

consists of the essential technical features from the target patent and a composite list of similar essential technical features from the comparative patents. In this paper, we propose a versatile framework (Section 3) for patent comparison concerning patent analysis applications in modern business operations:

- Experts pay more attention to essential technical features in patent analysis applications than other information in a patent, which are the fundamental components in evaluating the similarities between patents. We propose the definition of essential technical features as a composite of a ternary structure: Subject, Action, and Object (SAO). We extract technical features from the semantic dependency trees of sentences in patents (Section 4.1).
- We consider both the semantic similarity and the patent relevance similarity between essential technical features. The patent relevance of technical features is assessed based on a constructed feature-patent graph containing multiple relations in a semi-supervised learning manner (Section 4.2).
- We formulate the selection of essential technical feature pairs from both the target patent and its comparative patents as an optimization problem. Each technical feature in the target patent has a corresponding similar feature from as few comparative patents as possible.

We also report the results from extensive experiments to demonstrate the effectiveness and efficiency of the proposed framework (Section 5).

Related work

In this section, we concentrate on works that use data mining techniques to solve patent analysis problems and the existing comparative document summarization methods.

Patent mining

There are lots of research efforts put into the patent mining area (Tseng et al., 2007; Tang et al., 2012; Zhang et al., 2015; Krestel et al., 2021). Typical tasks include patent retrieval, patent classification, patent valuation, and patent visualization.

Patent retrieval is a subfield of information retrieval that focuses on developing technologies and methods to retrieve relevant patent documents efficiently. Helmers et al. (2019) proposed to use a full-text similarity search to get prior art results. They tested multiple approaches for feature representations and computing similarity between patents. However, they can only calculate the pairwise similarity using the cosine similarity instead of indicating the commonalities between two patent documents.

Patent classification based on the technical characteristics of their contents is an essential task in patent analysis applications. It enables the feasible search of documents about earlier disclosures similar to the invention for which a patent is applied and the tracking of technological trends in patent applications. Risch and Krestel (2019) presented domain-specific pre-trained word embeddings for the patent domain. They proposed a deep learning approach based on gated recurrent units for automatic patent classification built on the trained word embeddings.

Patent valuation is a typical process of assessing the economic value and quality of the patents, which can assist in the strategic decisions on the company's assets and facilitate the commercialization and transactions concerning intellectual property rights. Hu et al. (2012) proposed a topic-based temporal mining approach to quantify a patent's novelty and influence and automatically discover core patents.

Patent visualization, an application of information visualization, helps to clearly show the gist of patents and quickly identify the correlations between different patents. For more information about patent visualization, one can refer to (Federico et al. 2017), in which the interactive analysis and visualization approaches of patents and scientific articles are reviewed, ranging from exploration tools to sophisticated mining methods.

Comparative document summarization

Another research field tight with this paper is comparative document summarization. Comparative summarization plays an increasingly important role in the downstream tasks of the patent analysis. Traditional document comparison applies various methods to summarize the documents or patents. Erkan and Radev (2004) proposed an algorithm called LexPageRank to compute the sentence importance based on the concept of eigenvector centrality, which is extended in (Mihalcea & Tarau, 2005; Wan & Yang, 2008). Other classic methods include CRF-based summarization (Shen et al. 2007), sentence-based topic models (Wang et al., 2009), and ensemble methods (Wang & Li, 2010; Li & Ding, 2008).

Tseng et al. (2007) applied text mining techniques to segment patent texts and generate summaries for patent analysis. Tang et al. (2012) developed a novel topic-driven patent mining system. Instead of merely searching patent contents, they constructed a heterogeneous patent network for ranking and summarizing patents. Yang and Soo (2008) proposed combining domain knowledge ontologies and text parsing dependency trees to construct the concept map of patent claims, the foundation of patent comparison.

There are research works of comparative summarization that focus on finding the differences among documents. Huang et al. (2014) studied comparative news summarization to highlight the commonalities and differences between news. The evidence of comparativeness is cross-topic pairs of semantic-related concepts, and the evidence of representativeness is topic-related concepts. Such topic-based approaches are not suitable for patent comparison due to the complexity of the patent document. Wang et al. (2012) selected discriminative sentences in different document sets as a comparative summary. This method is not applicable in patent comparison analysis because we can not find valuable information from the discriminative sentences. Users need to evaluate the commonalities among patents by using the comparative summary. Shen and Li (2010) proposed a framework for multi-document summarization based on the minimum dominating set. The framework can accommodate four well-known summarization tasks: generic, query-focused, update, and comparative summarization.

Cascini and Zini (2008) calculated the similarities between two patents by comparing the invention functional tree. However, they only considered the global similarity instead of the local commonalities between the two patents. Zhang et al. (2015b) studied the problem of comparative patent analysis. They proposed extracting the significant parts of two patent documents and highlighting their relationship in terms of commonalities.

However, most existing comparative summarization methods for documents and patents are based on the discriminative words of two patents, which cannot capture their essential technical features and handle the case of one-to-many comparison. We focus on comparative patent summarization. The comparative summarization can provide strong evidence for patent analysis applications. Then experts can quickly determine whether previously granted patents or patent applications have disclosed the idea of a patent application.

The overall framework

Let d_0 denote a target patent, and $D = \left\{d_1, d_2, ..., d_n\right\}$ denote the comparative patent set. In this study, we assume that the target patent d_0 is comparable to the patents in the comparative patent set D, i.e., d_0 shares topics or technology characteristics with some patents in D. This assumption is reasonable because when patent experts do manual analysis, they search for the comparative patents by submitting related technology keywords first to patent search engines to narrow down the comparison range. Then they compare the essential technology features in these patents manually.

Figure 2 presents the overall framework for the one-to-many comparative patent summarization. There are three major steps:

1 The patent claims in $d \in D \cup \{d_0\}$ are parsed by the natural language processing (NLP) tools to build the sentence-level semantic dependency trees, then essential technical features, denoted as $T = \left\{t_1, t_2, ..., t_m\right\}$, are extracted from these dependency trees.
2 We construct a feature-patent relevance graph G which contains multiple relations from patent documents. The relevance $r(t, d)$ between a technical feature $t \in T$ and a patent $d \in (D \cup d_0)$ is learned in a semi-supervised learning manner.
3 We select the common and comparative essential technique features from the target patent and the comparative patents, respectively, and generate a concise summary based on these technical feature pairs, denoted as $S = \{\langle t, t' \rangle | t \in T_0, t' \in T_c\}$.

Such a concise summary can help experts quickly determine whether previously granted patents have disclosed the idea of a patent application or whether a product-related patent uses almost the same idea of a few patents. Table 1 shows the notations used in this article.

One-to-many comparative summarization for patents

The three steps introduced in the last section are (1) extracting the essential technical features from sentence-level semantic dependency trees, (2) constructing a feature-patent relevance graph for calculating the relevance between features and patent documents, and (3) generating a concise summary by selecting the common and comparative essential technique features from the target patent and the comparative patents. In this section, we will explain the details of each step.

Essential technical feature extraction

Essential technical features are the keystones of patent comparative analysis because they reveal the technical characteristics of patents. The Subject-Action-Object (SAO) structure can represent a variety of technical characteristics, i.e., a list of a subject (noun), an action

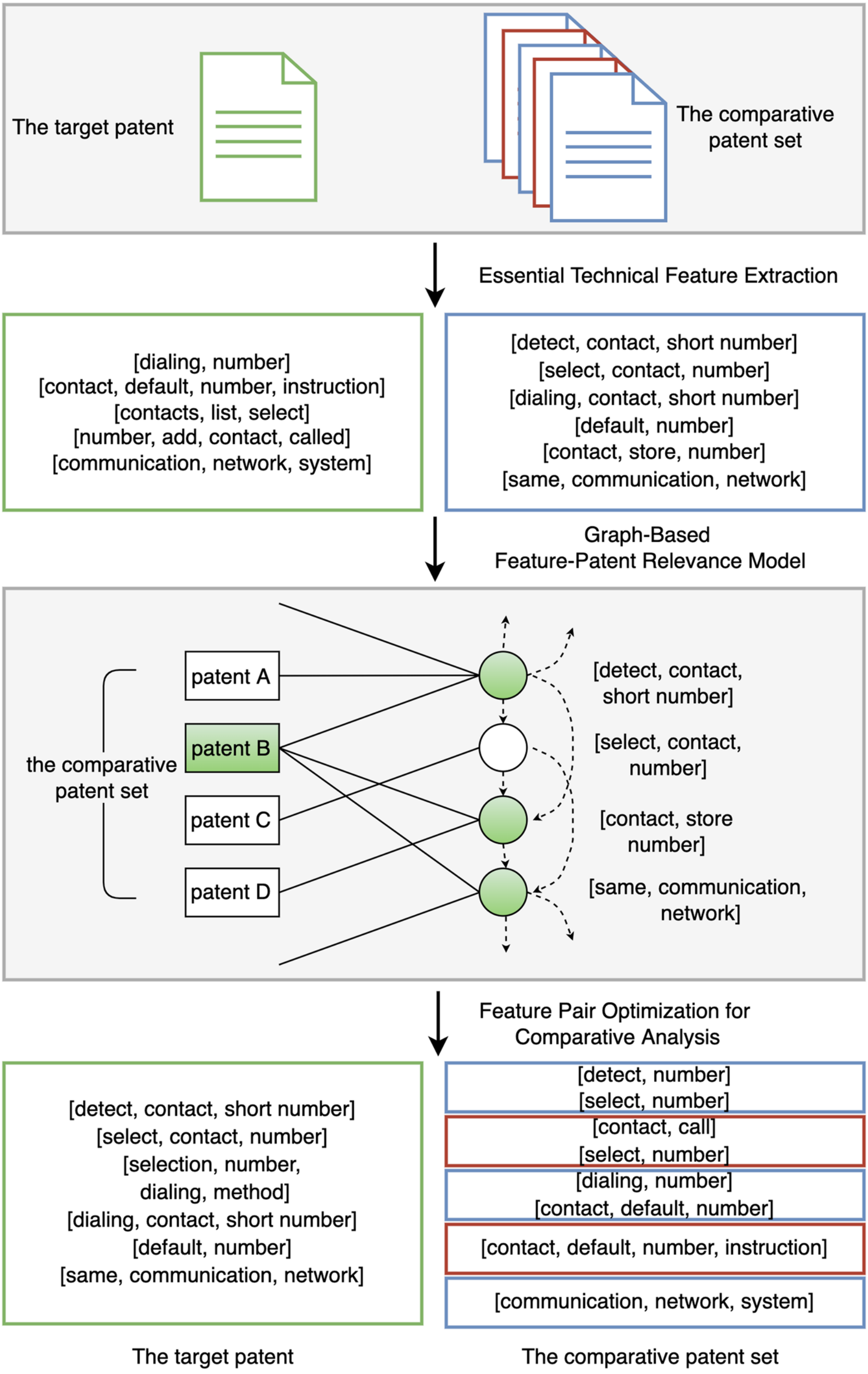
The target patent
The comparative patent set
Essential Technical Feature Extraction
[dialing, number]
[contact, default, number, instruction]
[contacts, list, select]
[number, add, contact, called]
[communication, network, system]
[detect, contact, short number]
[select, contact, number]
[dialing, contact, short number]
[default, number]
[contact, store, number]
[same, communication, network]
Graph-Based Feature-Patent Relevance Model
the comparative patent set
patent A
patent B
patent C
patent D
[detect, contact, short number]
[select, contact, number]
[contact, store number]
[same, communication, network]
Feature Pair Optimization for Comparative Analysis
[detect, contact, short number]
[select, contact, number]
[selection, number, dialing, method]
[dialing, contact, short number]
[default, number]
[same, communication, network]
[detect, number]
[select, number]
[contact, call]
[select, number]
[dialing, number]
[contact, default, number]
[contact, default, number, instruction]
[communication, network, system]
The target patent
The comparative patent set

Fig. 2 The overall framework for the one-to-many patent comparison summarization

Table 1 Notations

Notation	Description
d or d_i	A patent document
d_0	A target patent document
$D = \{d_i\}_{i=1}^{n}$	A comparative patent set
t or t'	An essential technical feature
$T = \{t_i\}_{i=1}^{m}$	A set of all essential technical features
T_0	A set of technical features from the target patent
T_c	A set of technical features from the comparative patents
$S = \{\langle t, t' \rangle \mid t \in T_0, t' \in T_s\}$	A set of similar technical feature pairs
$fsim(t, t')$	The similarity between two technical features t and t'.
$r(t, d)$	The relevance between feature t and patent document d.
Q^t	The relevance score vector associated with essential technical feature t.
G	A feature-patent graph containing multiple relations.
v_t	A vertex representing a technical feature t on G.
V_t	A vertex set of technical features.
v_d	A vertex representing a patent document d on G.
V_d	A vertex set of patent document

(verb), and an object (noun) (Choi et al., 2012). Subject or object might refer to a computer system component, and action might indicate a function executed by that component. For example, the SAO structure of [computer, retrieve, information] may represent a particular technical characteristic in a patent, where“computer”is the subject, “retrieve”is the action, and“information”is the object. In the following of this paper, we may use technical features or features to refer to essential technical features when there is no ambiguity.

However, simple SAO structures cannot represent complex essential technical features sometimes. There could be more than one word in subjects/objects, and adjectives or nouns usually modify them. We solve this issue by leveraging the semantic dependency trees of sentences. Figure 3 shows the semantic dependency tree of the sentence“A sensing device configured to detect an ambient light angle.”Many NLP tools can generate semantic dependency trees, and Stanford CoreNLP (Manning et al., 2014) is employed in this paper.

Each node in the tree corresponds to a word in the sentence. For node“2-n-device”in Fig. 3,“2” represents the index of the word“device”in the sentence, and“n” represents the part of speech of the word“device.”Each edge represents the semantic relationship between words in the sentence. For example,“nsubj”represents the subject-predicate relationship. Under the context of patent analysis, not all relationships defined in Stanford CoreNLP are necessary. The interesting relationships are v-n, n-v, n-n, n-a, n-n-a, n-v-n, v-v-n and v-n-n, where“v”means verb,“n”means noun and“a”means adjective. Table 2 shows descriptions and examples of the selected relationships.

We propose to define essential features as sub-trees in semantic dependency trees and extract them from patent claims according to the following two steps.

1 For each sentence in the claims in patents, we apply NLP tools to generate its semantic dependency tree. Then we remove the stop words and extract paths according to the relationships in Table 2.

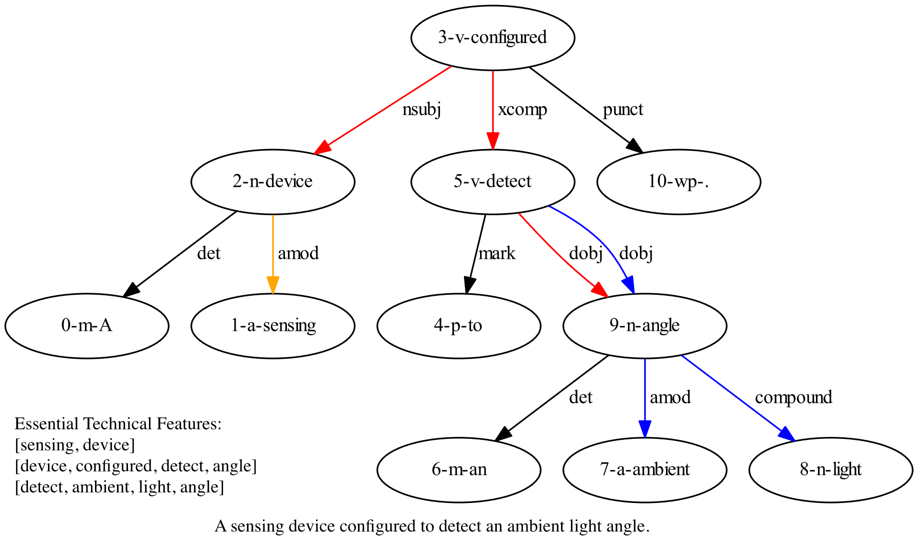

Fig. 3 An example of a semantic dependency tree and essential technical features

2 Each path is a composite of the noun, verb, and adjective nodes. We merge paths with the same root note to form sub-trees representing essential technical features, one feature for each tree.

In Fig. 3, there are three essential features, as indicated in the bottom left corner. We remove the structure of the essential features for simplicity.

Graph-based feature-patent relevance model

A straightforward solution for generating a concise summary consisting of common, similar, and comparative features is to employ essential technical features occurring in both patents to point out the commonality. However, this simple method ignores the semantic meanings of technical features, resulting in missing important pairwise commonalities (Mani & Bloedorn, 1997). Another way to derive the semantic connections between essential technical features is to apply clustering algorithms to find commonality clusters., but it is usually not easy to identify the number of clusters.

The connection between essential technical features lies in two aspects: semantic meanings and co-occurrences. It is easy to see that their semantic meaning should be similar when two essential technical features are comparable. Otherwise, the comparison is not helpful for patent analyzers. Furthermore, two semantically similar features are comparable if they are related to a common set of patents, that is, the resembling corpus-level co-occurrences. Suppose we can measure the relevance between a feature and patents, then for comparable features, their relevance distributions should be similar.

We design a graph-based feature-patent relevance model to capture the unified similarity in terms of the above two aspects. The relevance between features and patents is learned in a semi-supervised manner based on the constructed feature-patent graph containing multiple relations from the patent corpus. Figure 4 shows an example of feature-patent multi-relational

Table 2 Description of the relationship between each part of speech

Relation	Description	Example
n–v	It represents subject-action relationship. The first term is a noun and the second term is a verb	A sensing **device(n) configured(v)** to detect an ambient light angle
v–n	It represents action-object relationship. The first term is a verb and the second term is a noun	A sensing device configured to **detect(v)** an ambient light **angle(n)**
n–a	It represents adjective modification. The first term is a noun and the second term is an adjective	A **sensing(a) device(n)** configured to detect an **ambient(a)** light **angle(n)**
n–n	It represents noun modifications, and the two terms are both nouns. Sometimes nouns can be used as adjectives to modify nouns	A sensing device configured to detect an ambient **light(n) angle(n)**
v–v	It represents verb modification, and the two terms are both verbs. Sometimes a verb can be used as an auxiliary verb for another verb	A sensing device **configured(v)** to **detect(v)** an ambient light angle

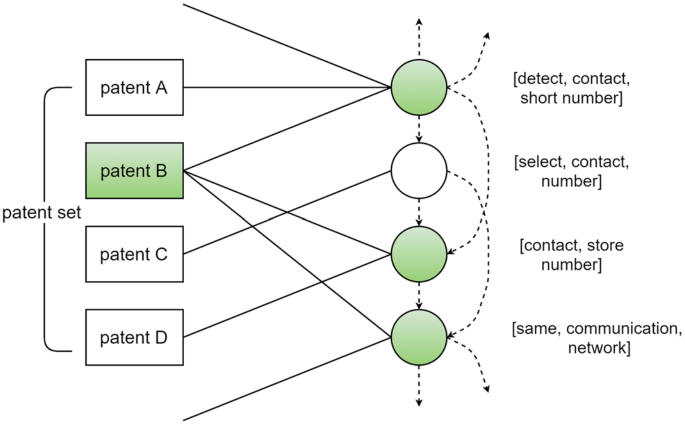
patent A
patent B
patent C
patent D
patent set
[detect, contact, short number]
[select, contact, number]
[contact, store number]
[same, communication, network]

Fig. 4 An example of feature-patent multi-relational graph

graph. Let $G = (V, E)$ denote the graph. The vertex set V is $V_t \cup V_d$, where V_t represents the set of essential technical features, and V_d represents the set of all patents. If a feature $t \in T$ occurs in a patent $d \in D \cup \{d_0\}$, these two vertices v_t and v_d are connected. The weight of edge (v_t, v_d) is defined as follows:

$$edgeweight(v_t, v_d) = \sum_{i}^{q} TF(t_i, d) * \log \frac{|D| + \epsilon}{DF(t_i, D) + \epsilon}, \tag{1}$$

where q is the number of words in t, $TF(t_i, d)$ is the frequency of the word t_i in patent d, and $DF(t_i, D)$ is the document frequency of the word t_i in patent set D. ϵ serves as a smoothing constant, which is 0.5 in the experiments.

To incorporate the semantic meaning of essential technical features, we apply Word2Vec (Mikolov et al. 2013) on the large patent dataset (see Section 5) to obtain the vector representations of words. At the same time, other embedding methods such as BERT (Devlin et al. 2019) are also applicable. Let $fsim(t, t')$ be the semantic similarity between essential technical feature t and t'. If $fsim(t, t')$ is large than a threshold λ, there is an edge between the corresponding vertices v_t and $v_{t'}$. The weight of edge $(v_t, v_{t'})$ is the semantic similarity of the corresponding features, i.e., $fsim(t, t')$.

We leverage semi-supervised learning on the feature-patent graph to model the relevance $r(t, d)$ between essential technical features and patents. Let W represent the adjacent matrix of the feature-patent graph with normalized weights. Let $\mathbf{r}$ denote the feature-patent relevance vector where $r_i = r(t_i, d_0)$, and let $\mathbf{f}$ denote the patent-patent relevance vector, where $f_j = f(d_j, d_0)$. All essential technical features in the target patent d_0 are labeled as positive. Then we can model the feature-patent relevance score propagation by combining a graph regularization term and a supervision term (Ren et al., 2017; Zhou et al., 2003), as shown in the following equation.

$$L_{d_0}(\mathbf{r}, \mathbf{f}) = \sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij} \left(\frac{r_i}{\sqrt{deg(t_i)}} - \frac{f_j}{\sqrt{deg(d_j)}} \right)^2 + \mu \sum_{j=1}^{n} \left(f_j - f_I\right)^2. \tag{2}$$

here m is the total number of essential features, n is $|D \cup \{d_0\}|$, and W_{ij} is the weight of edge (v_{t_i}, v_{d_j}). We use the degrees of vertex v_{t_i}, v_{t_j}, denoted as $deg(t_i), deg(t_j)$, to normalize W_{ij} in the first term. In the second term, f_I is the indicator to impose the positive label of d_0, i.e., $f_{d_0} = 1$ and $f_{\neg d_0} = 0$. A tuning parameter μ $(0 < \mu < 1)$ is used to control the strength of supervision from d on the score propagation.

Feature pair optimization for comparative summarization

The concise comparative summary consists of the essential technical features in the target patent and their corresponding similar features from the comparative patents, which should satisfy the following requirements: (1) a feature and its comparative feature should be semantic similar; (2) a feature and its comparative feature should have similar patent relevance distribution; (3) the comparative features should come from as few patents as possible.

The first and the second requirements have been introduced in Section 4.2. Now let us explain the third requirement. In patent analysis, patent examiners compare the essential technical features in the target patent with the features in prior arts to determine whether features in prior arts already disclose the features in the target patent. In an extreme case, if all features in the target patent are disclosed in one existing prior art, the target patent has no patentability. Suppose the features in the target patent are disclosed in a combination of prior arts. The examiner must consider whether these features in prior arts fit together (inter-connecting) in the instrumentality in the same way as those in the target patent. In this circumstance, more prior arts will reduce the possibility of similar instrumentality.

Moreover, if both essential technical features t and t' have a high relevance with a particular patent d, these two features, t and t', are more likely to be from the same patent. If all comparative essential technical feature pair has similar relevance score vectors, then the number of comparative patents should be small. In other words, the sum of all the distances between selected comparative essential technical features should be small.

Let Q^t denote the relevance score vector associated with essential technical feature t, i.e., $Q^t_i = r(t, d_i)$, where $d_i \in D \cup d_0$. The feature-patent relevance score $r(t, d)$ is obtained by minimizing Eq. 2. When selecting a technical feature pair $\langle t, t' \rangle$, we hope that the two features t, t' have a high semantic similarity and relevance similarity. High relevance similarity means if the relevance of technical feature t is very high in patent d, the relevance of technical feature t' is also very high in patent d.

We formulate selecting comparative essential technical feature pairs into an optimization problem. Let S be the set of pairs of essential technical features. For each pair $\langle t, t' \rangle \in S$, t is from the target patent d_0. t' is from a patent in the comparative patent set D. The objectives of the optimization problem corresponding to the requirements are (1) maximize the overall semantic similarity for feature pairs, (2) maximize the overall relevance distribution similarities for comparative feature pairs, (3) maximize the total relevance distribution similarities between features in the selected comparative features. Let O be the aggregated objective of the above ones, and then the optimization problem is as follows:

$$\begin{aligned} \max_S \; O &= \sum_{\langle t,t' \rangle \in S} \left(fsim(t,t') + \alpha Q^t \cdot Q^{t'} \right) + \beta \sum_{t_i,t_j \in T_s} Q^{t_i} \cdot Q^{t_j}, \\ s.t. \quad S &= \left\{ \langle t,t' \rangle | t \in T_0, t' \in T_s \right\}, T_s \subset T \; and \; |S| = K. \end{aligned} \tag{3}$$

In Eq. (3), we use the inner product to measure the similarity between relevance vectors. The first term in objective O aggregates the semantic similarity scores and the relevance similarity scores of the selected technical feature pairs. The tuning parameter α $(0 < \alpha < 1)$ controls the trade-off between the semantic similarity and the relevance similarity. The second term aggregates all the relevance similarity scores between technical features in order to reduce the total number of comparative patents. The tuning parameter β $(0 < \beta < 1)$ controls the trade-off between the feature similarity and the number of comparative patents.

Algorithm 1 Selecting comparative essential technical feature pairs

```
Input:
    $T_c$: all essential features extracted from patents $D$;
    $T_0$: essential features extracted from target patent $d_0$;
    $Q^t$: relevance score vector associated with essential technical feature $t$;
    $F$: similarity matrix between technical features;
    $K$: number of similar technical feature pairs;
Output:
    $S$: selected similar technical feature pairs;
 1: $S = \varnothing$;
 2: $T_s = \varnothing$;
 3: repeat
 4:     for each $t_i \in T_0$ do
 5:         for each $t_j \in (T_c - T_s)$ do
 6:             $o(t_i, t_j) = F[t_i, t_j] + \alpha Q^{t_i} \cdot Q^{t_j} + \beta \sum_{t_k \in T_s} Q^{t_j} \cdot Q^{t_k}$;
 7:         end for
 8:     end for
 9:     $t_i^*, t_j^* = \max_{t_i, t_j} o(t_i, t_j)$;
10:     $S = S \cup \{\langle t_i^*, t_j^* \rangle\}$;
11:     $T_0 = T_0 - \{t_i^*\}$;
12:     $T_s = T_s \cup \{t_j^*\}$;
13: until $|S| = K$
14: return $S$;
```

Considering the computation cost of the optimization problem, we propose a greedy algorithm to solve Eq. (3), which iteratively selects one essential technical feature at a time until all or K essential technical features in the target patent are selected. K is a pre-defined integer. Algorithm 1 presents the overall procedures, where T_c is the essential technical feature set extracted from the patent set D, and T_0 is the technical feature set extracted from the target patent d_0. Q^t is the relevance score vector associated with technical feature t. F is the semantic similarity matrix between technical features.

Lines 1 to 2 in Algorithm 1 are the initialization of variables. Line 4 selects a technical feature from the target patent feature set. Line 5 selects a technical feature from the comparative patent feature set. Line 6 calculates the objective o of each technical feature pair, which is selected based on lines 4,5. Then in line 9, we select the most similar technical feature to compose the feature pair. Lines 10 to 12 update corresponding variables. Loop from lines 3 to 13 stops until K technical feature pairs are selected.

Experimental evaluation

We report the evaluation results of the proposed framework in this section.

Experimental settings

Datasets

Comparative patent summarization is a new and novel application in patent analysis, and there is no standard benchmark patent dataset for evaluation so far. Our datasets used in the experiments come from two sources. One source is manual generation. According to real-world patentability or infringement analysis reports about several topics, a patent agent company provides data from this source. Another source is the patent search engines, where the search results are collected by submitting the search keywords from the first source. All patents we used in the experiments are Chinese patents. Part of these Chinese patents are PCT patents where PCT means the Patent Cooperation Treaty, so they have the corresponding English versions. Other patents are translated into English by Google Translate for presenting the experimental results. In addition, We also collected a large patent dataset, denoted as Patent-Large, to evaluate the efficiency and scalability of the proposed method. Table 3 shows the dataset statistics.

The number of comparative patents for each target patent is not equal. Each target patent has from 20 to 25 relevant patents in their corresponding comparative patent set. The average number of essential technical features of a semantic dependency tree is 6 (rounded to the nearest integer). The number of the essential technical features in each category in Table 3 is as follows: (1) the average number of features in a chemistry patent is 1356, (2) the average number of features in an electricity patent is 732, (3) the average number of features in a mechanical engineering patent is 336, and (4) the average number of technical features in a physics patent is 1212.

Compared methods

We selected both classical and state-of-the-art comparative summarization methods as the baselines:

- WordMatch (Mani & Bloedorn, 1997): It generates a common set of salient words if they occur in both two patent documents, where the top-N salient words are extracted based on TF-IDF scores.
- Word2Vec-Clus: It learns the embeddings for salient phrases from a patent corpus using the skip-gram model (Mikolov et al., 2013) and then clusters phrases using the X-means algorithm. The common set contains the phrases occurring in both patent documents and closest to each cluster centroid.
- PatentCom (Zhang et al., 2018): It is a state-of-the-art comparative summarization method for patents. The technical feature tree is generated by Steiner tree generation, and the phrase nodes of the feature trees serve as the result of the patent comparison.

Table 3 Dataset Statistics

Dataset	File size (MB)	Documents
Manual dataset	81.64	2891
Retrieval dataset	86.04	3437
Patent-Large	530.70	22039

We conduct experiments with the proposed method and its variants where different distance measures are employed to calculate the similarities between relevance vectors, including:

- ETFCom: It is the proposed method in this paper, where the inner product is used to measure the distance of relevance vectors of technical features, as shown in Eq. 3.
- ETFCom-Cos: It is a variation of the proposed method, and the difference is that in Eq. (3), cosine distance is used to measure the distance of relevance vectors of technical features.
- ETFCom-Euc: It is a variation of the proposed method, and the difference is that in Eq. (3), Euclidean distance is used to measure the distance of relevance vectors of technical features.
- ETFCom-KL: It is a variation of the proposed method, and the difference is that in Eq. (3), Kullback-Leibler divergence is used to measure the distance of relevance vectors of technical features. Since the Kullback-Leibler divergence is asymmetrical, we use the average of both directions.

Evaluation metrics

The purpose of the comparative summary is to save the human efforts of examiners and analyzers during patent analysis when they compare essential technical features from various patent documents, so do many other summarization methods. For this purpose, although we could invite patent experts to evaluate the utility of various methods subjectively, it is very complicated to design an evaluation metric to quantify the utility besides the high cost of human efforts and discrepant labels. So we employ an objective way to evaluate the performance.

In the paper, we use the essential technical feature pairs whose semantic similarity is larger than a threshold as the ground truth, and then we find out which algorithm can find these feature pairs. Let I denote the selected essential feature pairs in the generated concise comparative summary. Let H denote the composite list of essential features in I whose semantic similarity is larger than a threshold λ. λ is 0.7 in this paper. We employ precision, recall, and F1 score to measure the performance of various algorithms, which are popular evaluation metrics in retrieval tasks (Lupu et al. 2017). Precision (P) is calculated by $P = |I \cap H| / |I|$, Recall (R) is calculated by $R = |I \cap H| / |H|$, and F1 score is calculated by $F1 = 2 * P * R / (P + R)$. The reported numbers are the averages of multiple experiments.

Recall that the first requirement at the beginning of Section 4.3 states that comparative features should be semantic similar. This is a fundamental requirement for comparative feature pairs because feature pairs that are semantically dissimilar have little usage in the patent analysis problem focused on by this paper. On the other hand, we could also consider the second and third requirements in designing the evaluation during performance comparison. However, Considering the second and the third requirement in designing the evaluation would be unfair to other algorithms which do not consider such requirements in their computation process. With all the considerations, we think the current experimental setters are acceptable and fair for the performance comparison, although not perfect. We leave the standard evaluation framework as a potential future work.

Parameter settings

The number of salient words in WordMatch is 50. The number of clusters in Word2Vec-Clus is automatically decided by the X-means algorithm, whose range is [20, 50]. As for our proposed algorithm, we found empirically that the performance of the proposed method does not change dramatically across a wide choice of parameters. In the following experiments of one-to-many patent comparison and its variants, we set $\mu, \alpha, \beta = 100, 0.1, 0.1$.

Performance

The detailed evaluation results of all the compared methods are summarized in Table 4, which shows precision, recall, and F1 score of the four domains datasets, including Mechanical Engineering, Chemistry, Physics, and Electricity. Overall, our proposed method outperforms others on all test sets in terms of finding commonalities, validating the effectiveness of our patent technology generation. With the careful analysis of the results, we found that the precision of our method was significantly better than that of other methods and achieved the comparable recall resulting in a high F1 score and superior performance.

The recall of WordMatch is high since selected words by WordMatch occur in both two patents because the semantic similarity between two identical words is 1. However, its low precision dominates the F1 score and results in degraded performance. A similar situation that suffers from low precision occurs in the other two compared methods. It is hard to decide the appropriate cluster granularity for the clustering-based methods (e.g., Word2Vec-Clus). Therefore, many good phrases that are not close to the centroid are missed,

Table 4 Performance comparisons on patent dataset in terms of Precision, Recall and F1 score

Method	Mechanical engineering			Chemistry		
	P	R	F1	P	R	F1
WordMatch	0.183	**1.000**	0.309	0.204	**1.000**	0.339
Word2Vec-Clus	0.232	0.687	0.347	0.140	0.885	0.242
PatentCom	0.098	0.437	0.161	0.086	0.826	0.156
ETFCom-Cos	0.626	0.581	0.603	0.723	0.703	0.713
ETFCom-Euc	0.633	0.573	0.602	0.701	0.720	0.710
ETFCom-KL	0.626	0.563	0.593	0.723	0.711	0.717
ETFCom	**0.676**	0.585	**0.627**	**0.723**	0.796	**0.758**
Method	Physics			Electricity		
	P	R	F1	P	R	F1
WordMatch	0.077	**1.000**	0.142	0.049	**1.000**	0.094
Word2Vec-Clus	0.097	0.545	0.164	0.045	0.692	0.084
PatentCom	0.073	0.600	0.129	0.040	0.666	0.075
ETFCom-Cos	0.705	0.545	0.614	0.328	0.667	0.439
ETFCom-Euc	0.703	0.550	0.617	0.369	0.665	0.474
ETFCom-KL	0.427	0.599	0.499	0.386	0.655	0.485
ETFCom	**0.721**	0.559	**0.630**	**0.435**	0.667	**0.527**

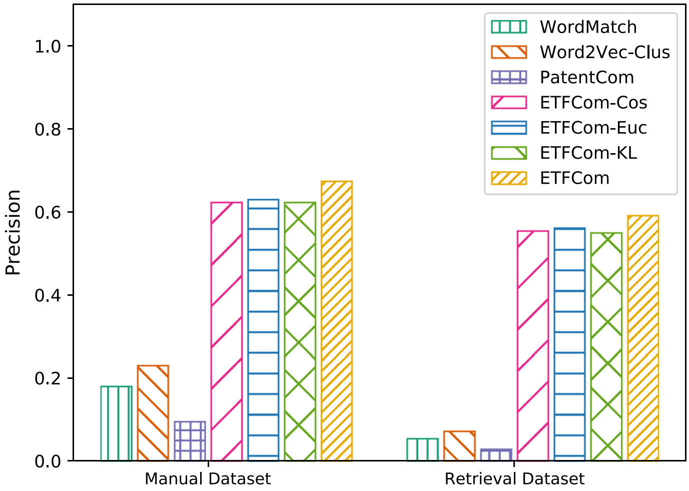

Fig. 5 Precision on manual dataset and retrieval dataset

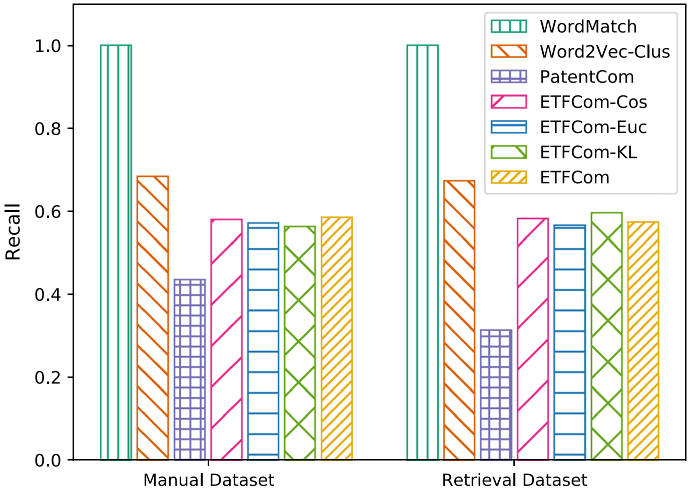

Fig. 6 Recall on manual dataset and retrieval dataset

which reduces precision results. PatentCom finds common terms simply by term overlap without considering semantic common words or phrases (same as WordMatch). Compared with ETFCom-Cos, ETFCom-Euc, and ETFCom-KL, ETFCom gains better performance from the inner product of relevance vectors. This proves the validity of using the inner product to measure the similarity of relevance vectors in this paper.

Figures 5, 6, 7 show the comparison results with different methods between manual dataset containing more related documents and retrieval dataset containing less related documents. ETFCom performs better than other methods in terms of Precision and F1 score on both kinds of document cases. As more semantic commonalities and subtle differences

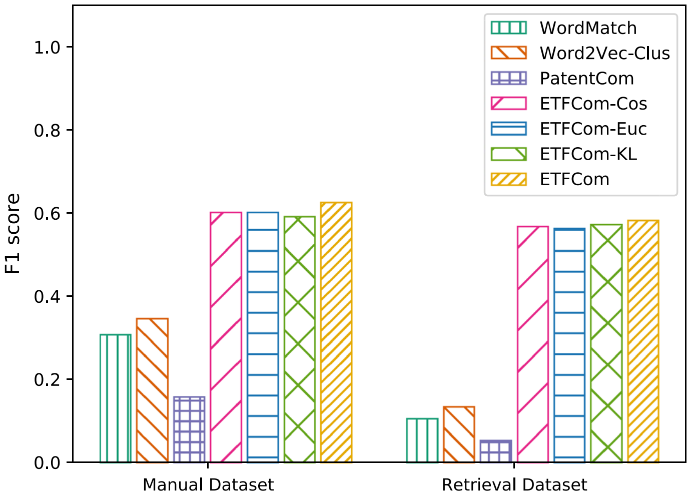

Fig. 7 F1 score on manual dataset and retrieval dataset

exist between a similar patent case generated manually, ETFCom gains more considerable improvement by optimizing the proposed measures and learning semantic relevance. It is also worth mentioning that our proposed methods ETFCom with different similarity measurements, perform more stably than other methods with leverage of the graph-based semantic relevance.

Then, we investigate the impact of the inner product on the patents' number. As shown in Fig. 8, the horizontal axis indicates the number of patents where technical features come from, and the vertical axis indicates the inner product of relevance scores between features.

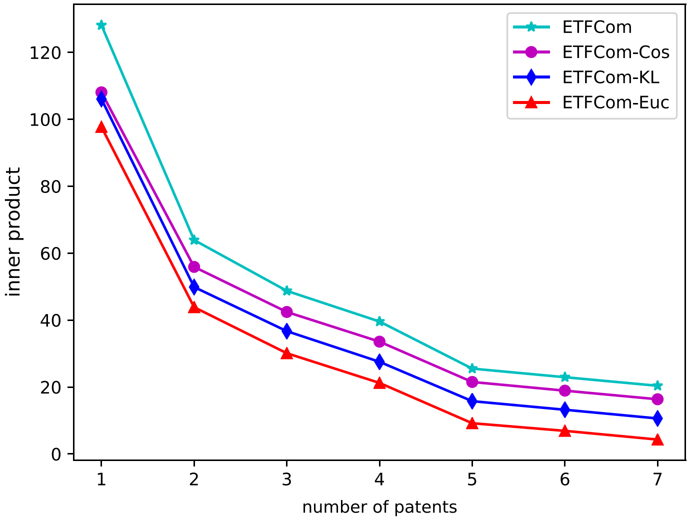

Fig. 8 The influence of the number of patents

We find that the inner product decreases as the number of patents increases. The underlying reason is that if two technical features have the same distribution of relevance scores, they have the same importance for each patent document. They may come from the same patent document. Thus, increasing the sum of the inner product of the feature's relevance scores can decrease the number of comparative patents.

Case study

We conduct a case study in this section. We collaborate with the patent department of a top ICT (Information and Communications Technology) company to evaluate the usage of the proposed framework. We report the results of one case. Table 5 shows the comparative summaries generated by the word-based method (Mani & Bloedorn, 1997) and sentence-based method (Shen & Li, 2010), where only the top-1 sentences are shown, as well as our proposed method. All patents are in Chinese, and as mentioned, we use their English versions for the presentation. The target patent application is CN103281426A, and the comparative patents include CN104010061B, CN102724347A, and CN101582944A. There are also a few other patents in the comparative patent set which are not in the generated summary. All these patents are related to the topic of the“telephone call method.”

Claims are the scope of the protection conferred in patents or the protection sought in patent applications. The generated comparative summaries here focus on claims in these patents. Table 5 presents the generated summaries by ETFCom, word-based methods(e.g., WordMatch (Mani & Bloedorn, 1997) and Word2Vec-Clus). The summaries generated by WordMatch and Word2Vec-Clus contain less valuable information than ETFCom since these word-based methods can only be used to generate the common word set of the compared patent documents. PatentCom (Zhang et al., 2018) generates a comparative summary of two comparable patents. However, the summary contains much fewer technical characteristics than the one generated by ETFCom. Having a more in-depth look at the features, we find that the summary generated by PatentCom is not concise enough and consists of more redundant nouns. In addition, the summary is mainly about the difference between the two patents, which is not the objective of patent analysis applications.

Table 5 shows essential technical features in the target patent claims, and their similar features from the comparative patents. We can see that mere three comparative patents can cover the technical features of the target patent. In other words, the existence of these three comparative patents significantly lowers the novelty and patentability of the target patent. When an expert or engineer tries to analyze these patents, he/she can focus on only three patents, which significantly saves the effort.

In addition, we also compare with a sentence-based method, i.e., DominatingSet (Shen & Li, 2010). The results are in Table 6. DominatingSet generates summaries of the discriminative sentences in each patent, and once again, it is not suitable for patent analysis applications. The sentence-based summarization techniques only provide distinction results for each patent. Our method provides cohesive and readable results compared with word-based methods, and it keeps the overall summary concise compared to sentence-based methods.

Scalability

To evaluate the efficiency of ETFCom, we intentionally add patents in the dataset to the comparative patent set of a target patent. We randomly add more patents from the manual

Table 5 Case study on summary of ETFCom and word-based methods

	Comparative patent set	
	Summary	Patent number
WordMatch	Technology, communication, selection, operation, steps, including, method, implementation, user, default, number, contact	CN102724347A
	Prompt, invention, implementation, technology, selection, inclusion, method, user, step, judgment, execution, dialing, dialing, contact, communication, number	CN104010061B
	Communication, judgment, prompt, network, including, automatic, invention, step, implementation, method, user, terminal, number, call, contact	CN101582944A
Word2Vec-Clus	Judge, each, plan, list, invention, reduce, settings, rights, communication, button, save, change, include	CN102724347A
	Network, quick, reception, random, trouble, customer, implementation, communication, setting, scheme, operation, mode, telephone	CN104010061B
	Communication, network, execution, interface, first, technology, customer, name, judgment, storage, plan	CN101582944A

	Target patent	Comparative patent set	
	Summary	Summary	Patent number
PatentCom	Terminal number, contact customer, short number	Module, number, information, contact, instruction	CN102724347A
	User, step, contact, customer, short number	List, information, name, user, step, contact, address book, number	CN104010061B
	Short number, contact, customer, phone	Primary contact, number, contact, unit, telephone, system	CN101582944A
ETFCom	[Detect, contact, short number]	[Detect, number]	CN102724347A
	[Select, contact, number]	[Select, number]	
	[Select, number, dialing, method]	[Select, number]	
	[Prompt, user, select, contact]	[User, select, contact, number]	
	[Default, number]	[Contact, default, number, instruction]	
	[Call, contact]	[Contact, call]	CN104010061B
	[Dialing, contact, short number]	[Dialing, number]	

Table 5 (continued)

	Target patent	Comparative patent set	
	Summary	Summary	Patent number
	[Contact, store, number]	[Number, add, contact, called]	
	[User, store, contact]	[User, store, contact]	
	[Same, communication, network]	[Communication, network, system]	CN101582944A

Table 6 Case study on summary of a sentence-based method.

	Comparative patent set	
	Summary	Patent number
DominatingSet	The invention discloses a method for automatically selecting a contact number dialing, comprising the following steps: (1) detecting whether the called contact has a short number, if yes, performing step (2), and if not, performing step (3); (2) dialing the short number of the contact; (3) detecting whether the contact has two or more numbers, if not, performing step (4), and if yes, performing step (5)	CN103281426A
	A method for displaying a contact number includes the following steps: receiving an instruction for a user to obtain a contact number; obtaining a contact name corresponding to the instruction and a number corresponding to the contact name	CN102724347A
	The present invention is applicable to the field of communication, and provides a dialing method and system. The method includes: storing at least one called contact name in the contact information of the first contact in advance, and the name of the called contact Correlating with the contact information of the calling contact in the address book; dialing the first contact; if the call fails, reading the stored in the contact information of the first contact Calling a contact person	CN104010061B
	The present invention is applicable to the field of mobile communication technologies, and provides a telephone dialing method, system, and mobile terminal, the method comprising the steps of: presetting at least one auxiliary contact associated with a primary contact; receiving a primary contact a phone number, calling the primary contact; when the call connection with the primary contact fails, receiving the phone number of the first secondary contact, calling the first secondary contact	CN101582944A

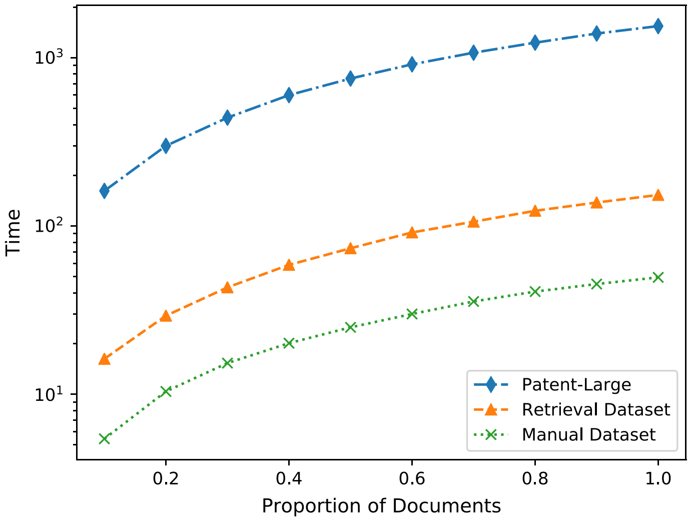

Fig. 9 Runtime analysis of Technology Extraction on the three datasets.(The sizes of technical features of Manual dataset, Retrieval dataset and Patent-Large are 41.70M, 45.95M and 285.40M, respectively)

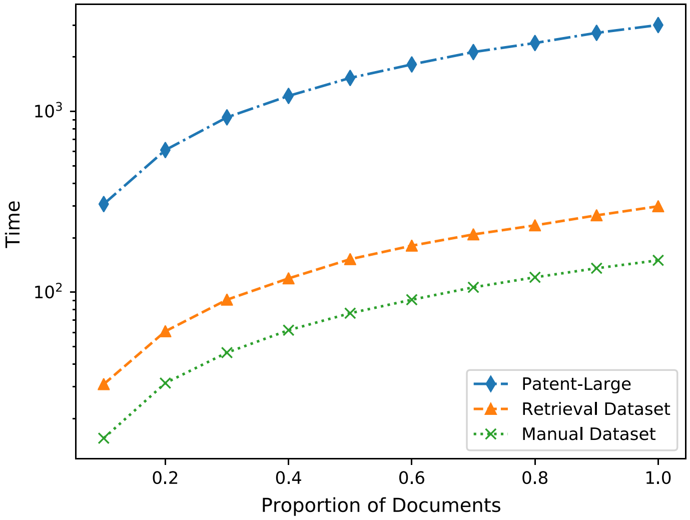

Fig. 10 Runtime analysis of Relevance Model on the three datasets.(The sizes of technical features of Manual dataset, Retrieval dataset and Patent-Large are 41.70M, 45.95M and 285.40M, respectively)

dataset, the retrieval dataset, and the large patent dataset to the comparative patent set of each target patent. To evaluate the efficiency of ETFCom, we randomly selected multiple target patents and measured the average runtime.

According to Section 4, ETFCom consists of three steps: extracting essential technical features, computing feature-patent relevance, and optimizing feature pair selection. Figures 9, 10, 11 present the runtime for the three steps, respectively, where the time axis is in log-scale for ease of interpretation. The number of added patents to the comparative patent

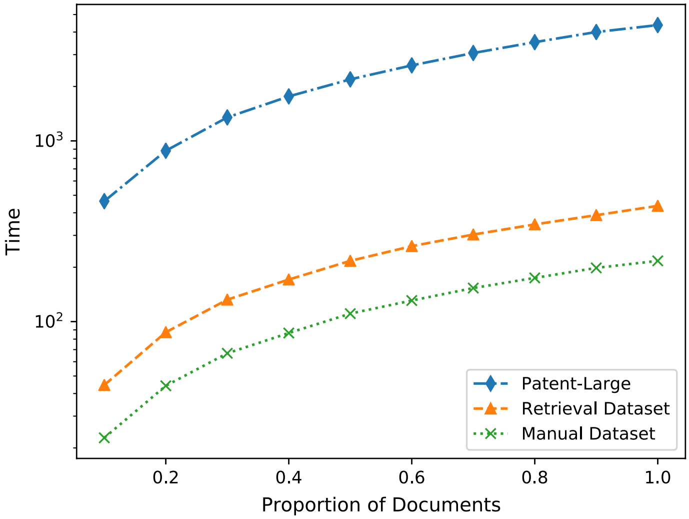

Fig. 11 Runtime analysis of Technology selection on the three datasets.(The sizes of technical features of Manual dataset, Retrieval dataset and Patent-Large are 41.70M, 45.95M and 285.40M, respectively)

set of each target patent is labeled as the proportion to the size of each data set. As we can see, the runtime of the proposed method scales linearly with the increment of the size of the corpus.

Conclusion

Comparative patent summarization is a powerful tool in many patent analysis applications. We proposed a framework for automatically generating comparative summaries by utilizing computing technologies to save expert efforts. The concise comparative summary focuses on essential technique features because they are the keystones when experts perform various tasks in patent analysis. The new definition of technical features surpasses pure SAO structures. In the optimization of selecting technical feature pairs, we considered the requirements from downstream patent analysis applications, especially that the features in the feature pairs should come from as few comparative patents as possible. Experimental results and the analysis in detail in the case study demonstrate the effectiveness of our proposed framework. As for the future direction, we plan to explore the performance quantification of the proposed comparative patent summarization and systematically study its impact in patent analysis in large technology companies.

Funding Funding was provided by Nanjing University of Posts and Telecommunications (Grant No. NY219084).

References

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information, 37*, 3–13.

Cascini, G., & Zini, M. (2008). Measuring patent similarity by comparing inventions functional trees. *IFIP International Federation for Information Processing, 277*, 31–42.

Choi, S., Kim, H., Yoon, J., Kim, K., & Lee, J. Y. (2012). An sao-based text-mining approach for technology roadmapping using patent information. *R & D Management, 43*(1), 52–74.

Devlin, J., Chang, MW., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota (Vol. 1, pp 4171–4186). https://doi.org/10.18653/v1/N19-1423.

Erkan, G., & Radev, D. R. (2004) LexPageRank: Prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Barcelona, Spain, pp. 365–371.

Federico, P., Heimerl, F., Koch, S., & Miksch, S. (2017). A survey on visual approaches for analyzing scientific literature and patents. *IEEE Transactions on Visualization and Computer Graphics, 23*(9), 2179–2198. https://doi.org/10.1109/TVCG.2016.2610422

Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, SIGIR'01, p 19–25

Helmers, L., Horn, F., Biegler, F., Oppermann, T., & Müller, K. R. (2019). Automating the search for a patent's prior art with a full text similarity search. *PLOS ONE, 14*(3), 1–17.

Hu, P., Huang, M., Xu, P., Li, W., Usadi, A. K., & Zhu, X. (2012). 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, Association for Computing Machinery, CIKM '12, pp. 1819–1823.

Huang, X., Wan, X., & Xiao, J. (2014). Comparative news summarization using concept-based optimization. *Knowledge & Information Systems, 38*(3), 691–716.

Krestel, R., Chikkamath, R., Hewel, C., & Risch, J. (2021). A survey on deep learning for patent analysis. *World Patent Information, 65*, 102035.

Lee, C., Song, B., & Park, Y. (2013). How to assess patent infringement risks: a semantic patent claim analysis using dependency relationships. *Technology Analysis & Strategic Management, 25*(1), 23–38.

Li, T., & Ding, C. (2008). Weighted consensus clustering. In *Proceedings of the 2008 SIAM International Conference on Data Mining, SIAM* (pp. 798–809).

Lupu, M., Mayer, K., Kando, N., & Trippe, A. J. (2017). *Current challenges in patent information retrieval*. Springer. https://doi.org/10.1007/978-3-662-53817-3

Mani, I., & Bloedorn, E. (1997). Multi-document summarization by graph search and matching. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI Press, AAAI'97/IAAI'97, pp. 622–628.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D (2014) The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations, pp 55–60

Mihalcea, R., Tarau, P (2005) A language independent algorithm for single and multiple document summarization. In: Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts, Asian Federation of Natural Language Processing

Mikolov, T., Sutskever, I., Chen, K., Corrado, GS., & Dean, J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119

Ren, X., Lv, Y., Wang, K., & Han, J (2017) Comparative document analysis for large text corpora. Association for Computing Machinery, New York, NY, USA, WSDM '17, p 325-334, 10.1145/3018661.3018690, https://doi.org/10.1145/3018661.3018690

Risch, J., & Krestel, R. (2019). Domain-specific word embeddings for patent classification. *Data Technologies and Applications, 53*(1), 108–122.

Shalaby, W., & Zadrozny, W. (2019). Patent retrieval: a literature review. *Knowledge and Information Systems, 61*(2), 631–660. https://doi.org/10.1007/s10115-018-1322-7

Shen C, & Li T (2010) Multi-document summarization via the minimum dominating set. In: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, USA, COLING'10, p 984–992

Shen, D., Sun, JT., Li, H., Yang, Q., & Chen, Z (2007) Document summarization using conditional random fields. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'07, p 2862–2867

Souza, C. M., Meireles, M. R. G., & Almeida, P. E. M. (2021). A comparative study of abstractive and extractive summarization techniques to label subgroups on patent dataset. *Scientometrics, 126*(1), 135–156. https://doi.org/10.1007/s11192-020-03732-x

Tang, J., Wang, B., Yang, Y., Hu, P., Zhao, Y., Yan, X., Gao, B., Huang, M., Xu, P., Li, W., et al (2012) Patentminer: Topic-driven patent analysis and mining. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, KDD'12, p 1366–1374, 10.1145/2339530.2339741

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Inf Process Manage, 43*(5), 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

Wan, X., & Yang, J (2008) Multi-document summarization using cluster-based link analysis. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, SIGIR'08, p 299–306, 10.1145/1390334.1390386

Wang, D., & Li, T (2010) Many are better than one: Improving multi-document summarization via weighted consensus. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, SIGIR'10, p 809–810, 10.1145/1835449.1835627

Wang, D., Zhu, S., Li, T., & Gong, Y (2009) Multi-document summarization using sentence-based topic models. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Association for Computational Linguistics, USA, ACLShort'09, p 297–300

Wang, D., Zhu, S., Li, T., & Gong, Y (2012) Comparative document summarization via discriminative sentence selection. ACM Trans Knowl Discov Data 6(3), 10.1145/2362383.2362386

Yang, SY., & Soo, VW (2008) Comparing the conceptual graphs extracted from patent claims. In: Proceedings of the 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (Sutc 2008), IEEE Computer Society, USA, SUTC'08, p 394–399, 10.1109/SUTC.2008.87

Zhang, L., Li, L., & Li, T. (2015). Patent mining: A survey. *SIGKDD Explor Newsl, 16*, 1–19.

Zhang, L., Li, L., Shen, C., & Li, T (2015b) Patentcom: A comparative view of patent document retrieval. In: Proceedings of the 2015 SIAM International Conference on Data Mining, SIAM, pp 163–171

Zhang, L., Liu, Z., Li, L., Shen, C., & Li, T. (2018). PatSearch: an integrated framework for patentability retrieval. *Knowledge and Information Systems, 57*(1), 135–158. https://doi.org/10.1007/s10115-017-1127-0

Zhou, D., Bousquet, O., Lal, TN., Weston, J.,&Schölkopf, B (2003) Learning with local and global consistency. In: Proceedings of the 16th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, NIPS'03, p 321–328