# 基于关键词的中文专利检索要素提取研究

**Research on** **Keyword-Based Element Extraction for Chinese Patent Retrieval**

**摘要：**专利检索是专利分析任务中的一个关键步骤，而检索要素是专利检索过程中构造检索式，进行精确查找的重要组成部分。随着每年专利申请提交数量的不断增加，人工提取检索要素、进行专利检索与审查的压力越来越大，这使得检索要素的自动提取问题亟待解决。在专利检索中，关键词是表达检索要素的手段之一，本文对中文专利申请文本中关键词检索要素的自动抽取问题进行了初步探究，使用不同的关键词抽取方法从中文专利申请文本中抽取检索要素。实验结果表明，关键词抽取方法能够在中文专利申请文本中有效地提取关键词作为检索要素用于检索，协助人工进行专利检索及审查。

**关键词：**专利检索；检索要素；关键词抽取

**Abstract:** Patent retrieval is a critical step in patent analysis, with retrievable elements playing a key role in constructing search queries and performing accurate searches. The growing number of patent applications each year has significantly increased the pressure on manual extraction of retrievable elements and patent examination, underscoring the need for automated solutions. Keywords serve as an effective means of expressing retrievable elements in patent retrieval. This study explores the automatic extraction of keyword-based retrievable elements from Chinese patent application texts. We employ various keyword extraction methods to identify retrievable elements within these texts. Our experiments have shown that these methods can effectively extract keywords as retrievable elements from Chinese patent applications, thereby facilitating more efficient manual patent searches and examinations.

**Keywords:** Patent retrieval; Retrievable elements; Keyword extraction

1 引言

专利权是知识产权最主要的组成部分之一，而专利是专利权产生的基础。专利记载了一项发明的技术内容，让其发明拥有者在一定时期内持有对这项发明的独享权，从而达到保护创新、鼓励创新的目的。随着知识保护意识的逐渐提高，专利的申请数量逐年增加，专利文献作为技术信息最集中、有效的载体，包含了世界上90%以上的最近研发成果[1]。对于企业或个人而言，从海量专利文献中了解行业技术现状，探寻技术布局与发展趋势，以便及时占领技术高地至关重要；对于专利机构来说，根据相关技术或文件查找最相近的现有技术，判断一项发明创造的专利性，从而构建合理的专利保护范围的难度也越来越大。尽管目的各有不同，但专利检索在这些过程中不可避免。

专利检索是信息检索的子领域，是几乎所有专利分析任务的关键一环[2]。专利检索作为发明专利申请实质审查程序中的一个关键步骤[3]，申请文本是这一步骤的主要依据。申请文本具有多模态、多语言、半结构化、元数据丰富等特点，冗长的文件内容以及模糊、抽象的行话与专业术语使得专利检索相较于一般文本来说，在实现自动化检索时更具有挑战性。而人工阅读申请文件、提炼检索要素、构造检索式并随时根据检索结果对前面的步骤进行动态调整的过程耗费大量的时间和精力，大大增加了时间成本，导致专利积压，延缓审查进度。借助自然语言处理技术实现某些环节的自动化处理并且可控，以算法协助专利检索来加速审查过程、降低审查难度是一个有价值的课题。

专利文本的特殊性使计算机难以如理解普通文本一般理解专利内容从而支持上层应用，因此针对专利的信息处理和规范化研究必不可少。目前，对于这方面的研究大多集中于专业术语抽取、特定领域内的命名实体识别以及语义关系抽取等，这些研究领域针对性强，抽取结果较为单一，并不能适用于专利审查过程中检索要素提取这一任务。针对这一问题，本文遵循专利检索的一般步骤，从“提取专利申请文本的检索要素”这一环节起进行探究，以检索要素的表达手段之一——关键词为切入点，并不限制申请文本的具体领域、关键词的种类等，比较多种无监督与有监督的关键词抽取方法应用于专利文本的效果，旨在探究不同抽取技术在多领域专利文本上的性能，探究计算机助力专利检索与审查的可行性，为后续的研究提供参考，为更多的工作提供启发。

1 Introduction

Patents are a fundamental component of intellectual property, serving as the foundation for patent rights. A patent documents the technical details of an invention, granting its owner exclusive rights for a certain period, thus protecting and promoting innovation. With the growing awareness of intellectual property protection, the number of patent applications has been increasing annually. Patent documents, as the most concentrated and effective carriers of technical information, encompass over 90% of the world's latest technological advancements[1]. For businesses and individuals, it is crucial to navigate through extensive patent ducuments to understand the current technological landscape, identify technological trends, and secure competitive advantages. For patent offices, the challenge is to find the closest prior art to assess an invention's patentability and define a reasonable scope of patent protection. Despite differing objectives, patent retrieval remains an essential process in these activities.

Patent retrieval, a subfield of information retrieval, stands as a pivotal component in nearly all patent analysis tasks[2]. As a crucial step in the substantive examination process for patent applications[3], patent retrieval relies heavily on the content of application texts. These texts possess characteristics such as multimodality, multilinguality, semi-structured formats, and rich metadata. The lengthy document content, coupled with the use of ambiguous, abstract jargon, and specialized terminology, renders patent retrieval more challenging compared to standard text retrieval when it comes to automation. The entirely manual process of patent examination consumes a significant amount of time and resources, substantially increasing time costs, leading to patent backlogs, and delaying the examination process. Leveraging Natural Language Processing (NLP) techniques to automate certain stages while maintaining control, thus using algorithms to assist in patent retrieval, presents a valuable avenue for accelerating examination processes and reducing examination complexity.

The specificity of patent texts poses challenges for computers to comprehend them as they would with ordinary text, thereby impeding support for higher-level applications. Hence, research into information processing and standardization for patents is indispensable. Currently, most research in this field focuses on tasks such as extracting specialized terminology, identifying named entities within specific domains, and extracting semantic relationships. However, these areas of research are highly specialized, yielding relatively homogeneous extraction results that may not be suitable for extracting retrievable elements during the patent examination process.

In response to this challenge, this study follows the general steps of patent retrieval, beginning with the exploration of "extracting retrievable elements". It takes keywords—one of the means of expressing retrievable elements—as the starting point without restricting the specific domain of the application text or the types of keywords. Multiple unsupervised and supervised keyword extraction methods are applied to patent texts to compare their effectiveness. The goal is to investigate the performance of different extraction techniques on multi-domain patent texts, explore the feasibility of computer-assisted patent retrieval and examination, provide reference for subsequent research, and inspire further work in this field.

2 专利基本信息

2 Preliminaries

## 2.1 专利申请文本

2.1 Patent Application Texts

专利文献是基于专利制度产生，记录有关发明创造信息的文献，包括一份发明专利从申请、公开、实质审查、授权乃至无效以及终止的全过程中产生的各种文本[4]。专利申请文本是重要的专利文献之一，是整个专利生命流程的起始文件，由申请人提交给专利局；也是专利审查过程中的重要依据文件，审查人员将申请文本的技术内容作为检索对象，通过检索现有技术对该专利的新颖性和创造性进行判断。

专利法对申请文本的构成有着明确要求，对于发明或者实用新型类型的专利，申请文本应该包括请求书、说明书、摘要和权利要求书等内容。申请文件扉页的著录项目中包含发明名称、分类号以及摘要等技术信息以及发明人、申请日期、法律状态等法律信息。权利要求书和说明书是记载发明内容且确定其保护范围的法律文件，根据专利法第六十四条第一款的规定，“发明或者实用新型专利权的保护范围以其权利要求的内容为准，说明书及附图可以用于解释权利要求的内容。”因此，权利要求是专利权人所要求保护的技术方案内容的具体体现，是判断一项技术是否具有新颖性和创造性的依据和审查重点，也是在检索时主要应当关注的内容。申请文本专业性强、半结构化以及冗长的特点使得将全文进行处理十分困难，在实验与分析时，以“权利要求”内容为主，必要时辅以“说明书”的内容合理且更具针对性。

Patent documents, stemming from the patent system, comprises documents detailing information about inventions. It encompasses various texts generated throughout the entire lifecycle of a patent, from application, publication, substantive examination, grant of patent rights, to invalidation and termination[4]. Among these, patent application texts stand out as pivotal documents, marking the inception of the patent lifecycle. Submitted by applicants to the patent office, they serve as crucial reference materials during the patent examination process. Examiners utilize the technical content as the basis for retrieval, assessing the novelty and inventiveness of the patent by comparing it with existing technologies.

The Patent Law sets clear requirements for the composition of application texts. For patents classified as inventions or utility models, the application text should include the request for granting a patent, specification, abstract, and claims. The cover page of the application document contains technical information such as title, kind codes, abstract, as well as legal information such as the inventors, application date, and legal status.

The claims and description are legal documents that record the invention's content and define its scope of protection. According to Article 64, Paragraph 1 of the Patent Law, "the scope of protection of an invention or utility model patent shall be determined by the content of its claims, with the description and drawings being used to interpret the content of the claims." Therefore, the claims represent the specific embodiment of the technical solution sought for protection by the patent holder. They serve as the basis for determining novelty and inventiveness and are the main focus during examination and retrieval. The highly specialized, semi-structured, and lengthy nature of application texts makes it challenging to process the entire document comprehensively. In experimental analysis, the primary focus is on the content of the claims, supplemented as necessary by the content of the description, for a more targeted approach.

## 2.2 检索要素

2.2 Retrievable elements

检索要素是体现技术方案基本构思的可检索要素[3]，可以被视为缩小集合范围的条件或者要素[5]。专利检索要素是一个抽象概念，并非仅限于某种词性或某种类型，需要使用具体的方式来表达，一般有两种表达手段：分类号与关键词。

专利在初步审查过程中会以国际专利分类表为依据对专利申请的技术主题进行分类，给定其一个或多个分类号。使用分类号进行检索的速度较快，但由于某些领域的专利分类号对技术的划分因比较上位而模糊，单独使用专利分类号进行检索会导致较大的噪声；对于跨领域的申请，当分类号并不准确时，也需要采用关键词进行初步检索[6]。尽管智能语义检索技术使得在用关键词难以表达的时候也可以直接对相关文本片段进行检索，但更多情况下，智能语义检索技术用于对检索结果进行排序，辅助于关键词和分类号检索。

因此，关键词作为检索要素之一在专利检索的精确匹配、保证检索结果可靠性的过程中扮演着不可或缺的角色。

Retrievable elements, as referenced in [3], represent the core conceptual aspects of a technical solution and serve as criteria or factors for narrowing down search parameters [5]. This concept is abstract and not restricted to a specific part of speech or category, necessitating specific expressions. Typically, there are two primary means of articulation: classification codes and keywords.

During the initial examination phase, patents undergo classification based on their technical subjects using the International Patent Classification (IPC) system, which assigns one or more classification codes to each patent application. While utilizing classification codes for retrieval offers speed advantages, the broad categorization of certain patent classification codes may introduce significant noise into search results. Moreover, in cases of interdisciplinary applications where classification codes may lack precision, the inclusion of keywords for initial retrieval becomes imperative [6]. Although intelligent semantic retrieval technology enables direct searches for relevant text fragments even in instances where keywords are insufficient, it is predominantly utilized for ranking search results and assisting keyword and classification code-based searches.

Thus, keywords, as one of the retrievable elements, play a critical role in ensuring the precision of patent retrieval and the reliability of search outcomes.

## 2.3专利检索

2.3 Patent retrieval

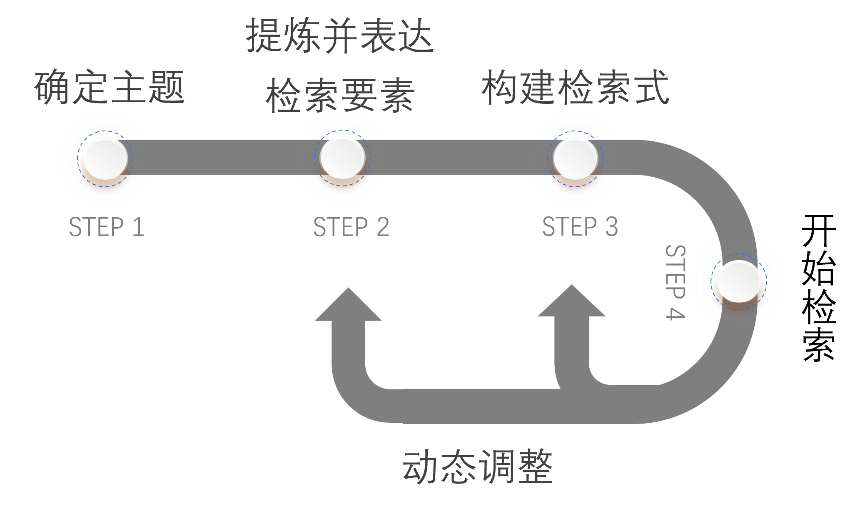
根据检索对象的不同，可以将专利检索分为面向新颖性和创造性评价的对比文件检索和面向专利分析的技术主题检索。前者注重于对现有技术的检索，例如常用于专利审查过程中的查新检索等；后者的目标是找到记载相同技术主题的目标文献，例如技术主题检索。两者目标的差异使得在检索过程中的一些具体细节上有所区别，但就整体的检索步骤而言大体是共通的，如图1所示：在专利检索过程中，首先根据独立权利要求中限定的技术方案理解其技术构思作为检索的主题；其次主要依据权利要求的技术方案，从技术领域、技术问题、技术手段或效果等方面提取检索要素，并将检索要素以一定的组合方式构建为检索式，在选定的专利数据库中进行检索。进行初步检索后，需要根据检索结果以及预期方向对检索策略进行调整，例如调整检索要素的表达或组合方式等。调整后重复前面的检索过程，直至达到检索终止条件。在第4节的实验部分，我们也将遵循基本流程进行实验。

图 1 专利检索的基本流程

Depending on the retrieval objectives, patent searches can be categorized into two types: prior art searches for evaluating novelty and inventiveness, and technical topic searches for patent analysis. The former focuses on retrieving existing technologies, commonly used in the patent examination process for novelty searches. The latter aims to find documents describing the same technical topics, such as in thematic searches. Although the goals differ, leading to variations in specific details during the search process, the overall steps are generally similar, as illustrated in Figure 1.

In a patent search, the process begins with understanding the technical concept delineated by the independent claims, which serves as the subject of the search. Next, based on the technical solution outlined in the claims, retrievable elements are extracted from aspects such as the technical field, technical problems, technical means, or effects. These retrievable elements are then combined into a search query, which is executed in a selected patent database. After an initial search, the search strategy may need to be adjusted based on the results and desired direction. This might involve modifying the expression or combination of retrievable elements. The process is repeated with these adjustments until the search termination conditions are met.

In Section 4, we will follow this basic workflow in our experiments.

3 相关工作

3 Related Work

## 3.1 现有专利信息提取研究及问题

3.1 Existing Research on Patent Information Extraction and Associated Challenges

专利信息的提取是进行专利检索与挖掘的基础。由于专利文本冗长、结构复杂、专业性强，不同领域的专利在术语表达、关系表述等方面都不尽相同，因此专利信息提取方面的研究主要集中于特定领域，如船舶建造领域等，亦或集中于特定类型，如术语、命名实体以及语义关系等[16]更具体，更有针对性的抽取工作，并基于这些研究基础进行后续的专利表示、检索、挖掘以及生成等任务。但在专利审查的检索中，专利检索要素的种类和表达繁多，既非特定词性，也非特定类别。综上，现有的专利信息提取技术并不能完全适用于检索要素的提取。在本文中，我们使用更一般的关键词抽取方法，在多领域的专利申请文本数据集上进行检索要素提取实验，比较不同种类提取方法、不同粒度提取结果分别在人工标注结果与检索引擎检索结果上的表现，探究计算机辅助检索要素提取的可行性。

Extraction of patent information for effective patent retrieval. The extraction of patent information is fundamental for patent retrieval and mining. Due to the lengthy, complex, and highly specialized nature of patent texts, and the variability in terminology and relational expressions across different fields, research in patent information extraction has primarily targeted specific domains, such as shipbuilding, or focused on specific types of data, like terminology, named entities, and semantic relationships [16]. These specialized studies have paved the way for further tasks such as patent representation, retrieval, mining, and generation.

However, in the context of patent examination and retrieval, the types and expressions of retrievable elements are diverse and not restricted to specific parts of speech or categories. As a result, existing techniques for patent information extraction are not entirely suitable for extracting these retrievable elements.

In this paper, we explore the feasibility of using more general keyword extraction methods for extracting retrievable elements from patent application texts across multiple domains. We conduct experiments to compare various extraction methods and the granularity of their results, evaluating their performance against gold standard and search engine results. This study aims to assess the potential of computer-assisted extraction of retrievable elements, providing insights for future research and practical applications in patent retrieval.

## 3.2 关键词提取方法

3.2 Keyword Extraction Methods

检索要素的关键词提取本质是关键词抽取任务，这一节我们将总结典型的关键词抽取方法，如表1所示。关键词提取的研究可以分为两类：无监督的方法和有监督的方法。无监督的方法主要可分为：基于统计的方法、基于图的方法、基于神经网络模型、预训练语言模型的嵌入式方法以及基于大语言模型的prompt方法。有监督的方法通常使用端到端的方式将关键词的选择与排序两个过程结合起来，同时针对这两个阶段进行优化。但有监督的方法需要大量的标注语料，难以推广到新的领域；随着语言模型的发展，模型参数量日渐增长导致模型的训练更加困难。因此，采用有监督的关键词提取方法需要权衡数据、算力以及训练时长等问题。

The essence of keyword-based retrieving elements lies in the task of keyword extraction. In this section, we will summarize typical keyword extraction methods, as shown in Table 1. The research on keyword extraction can be divided into two categories: unsupervised and supervised methods. Unsupervised Methods can be further divided into several types: statistics-based, graph-based, Neural Network-based, embedding-based models with Pre-trained Language Models and Prompt-based Methods with Large Language Models.

Supervised methods typically employ an end-to-end approach that integrates both the selection and ranking processes of keywords, optimizing both stages simultaneously. However, supervised methods require annotated corpora, making it difficult to generalize to new domains. Additionally, as language models evolve, the increasing number of parameters leads to more complex and resource-intensive training processes. Thus, the adoption of supervised keyword extraction methods involves balancing factors such as data availability, computational ability and training duration.

表格 1 关键词抽取方法 keyword extraction methods

|  |  |  |
| --- | --- | --- |
| Unsupervised | Statistics-based Methods | TF-IDF[7]、RAKE[8]、YAKE[9] |
| Graph-based Methods | TextRank[10]、TopicRank[11]、PositionRank[12] |
| Embedding-based Methods | Key2vec[13]、KeyBert[14]、PromptRank[15] |
| Prompt-based Methods | ChatGPT、ChatGLM[[1]](#footnote-1)、Qwen[[2]](#footnote-2) |
| Supervised | Pre-trained Language Models | ELMo、Bert、Transformer |
| Large Language Models | Efficient-tuning |
| Fine-tuning |

|  |  |  |
| --- | --- | --- |
| Unsupervised | 基于统计的方法 | TF-IDF[7]、RAKE[8]、YAKE[9] |
| 基于图的方法 | TextRank[10]、TopicRank[11]、PositionRank[12] |
| 基于嵌入的方法 | Key2vec[13]、KeyBert[14]、PromptRank[15] |
| 基于大预言模型的prompt方法 | ChatGPT、ChatGLM[[3]](#footnote-3)、Qwen[[4]](#footnote-4) |
| Supervised | 基于预训练语言模型 | ELMo、Bert、Transformer等架构 |
| 基于大语言模型 | Efficient-tuning |
| Fine-tuning |

4 基于关键词的中文专利检索要素提取实验

4 Experiments on Extracting Chinese Patent Retrievable elements Based on Keywords

为探究不同关键词提取方法在中文专利检索要素提取任务上的表现，我们分别选择3.2节中提到的每个分类中的一到两种方法，在中文专利申请文本数据集上进行实验，接下来依次介绍实验的数据集、评测指标、实验设置以及实验结果。

To investigate the effectiveness of various keyword extraction methods for extracting retrievable elements from Chinese patents, we selected one or two representative methods from each category mentioned in Section 3.2. We conducted experiments using a dataset of Chinese patent application texts. The following sections detail the dataset, evaluation metrics, experimental setup, and results.

## 4.1 数据集

4.1 Datasets

数据集包括测试集与检索数据集，测试集用于抽取检索要素，检索数据集用来评估检索要素的质量。我们使用不同的关键词抽取方法从测试集的内容中提取其检索要素，从两方面评价提取的检索要素的质量：一方面，使用F1值评价检索要素的质量；另一方面，将检索要素构造为检索式，使用检索引擎在检索数据集中检索相关文献，使用检索结果评价检索要素的质量。

The dataset is divided into two parts: a test set and a retrieval dataset. The test set is used to extract retrievable elements, while the retrieval dataset is used to evaluate the quality of these elements. We applied various keyword extraction methods to the test set to extract retrievable elements and evaluated their quality from two perspectives:

Quality of Extraction: This is measured using the F1-score, which assesses the accuracy and completeness of the extracted retrievable elements.

Retrieval Performance: The extracted retrievable elements are used to construct search queries, which are then input into a search engine to retrieve relevant documents from the retrieval dataset. The quality of the retrievable elements is evaluated based on the relevance of the search results.

### 4.1.1 检索数据集

4.1.1 Retrieval Dataset

为验证关键词提取方法在多领域的中文专利申请文本上的效果，我们构建了涵盖国际专利分类中A-H八个部类、超过25k篇的中文专利申请文本集作为检索实验的数据库（以下称为检索数据集），用于衡量检索要素提取的质量。我们统计了2014-2020年间的中文申请文本的分类号在国际专利分类中A-H八个部类中的占比，如表1所示。为方便实验进行，从六年的申请文本中按照其在国际专利分类中的自然分布随机抽取了25,696篇申请文本来模拟真实专利检索过程中使用的专利数据库。数据集中包含中文发明专利与中文实用新型专利两种，每一份申请文本中包含专利类型、申请日、申请号、分类号、发明人等著录项目以及标题、摘要、权利要求书、说明书及附图说明等发明内容。

To validate the effectiveness of keyword extraction methods across multiple domains of Chinese patent application texts, we constructed a database for retrieval experiments. This database comprises over 25,000 Chinese patent application texts covering the eight main sections (A-H) of the International Patent Classification (IPC). Henceforth referred to as the retrieval dataset, it serves as the basis for measuring the quality of retrievable element extraction.

We conducted a statistical analysis of the distribution of Chinese patent application texts across the eight main sections (A-H) of the International Patent Classification from 2014 to 2020. The proportions of these texts within each section are summarized in Table 2.

For experimental convenience, we randomly sampled 25,696 patent application texts from the six-year period according to their natural distribution. This sampling strategy aims to simulate the composition of a real patent database used in patent retrieval processes.

The dataset includes both Chinese invention patents and utility model patents. Each patent application text contains various metadata such as patent type, filing date, application number, kind code and inventors, as well as sections such as title, abstract, claims, description, and illustration descriptions, detailing the invention content.

表 2测试集八大部类占比 Distribution of the Eight Categories in the Test Set

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Categories | A | B | C | D | E | F | G | H | Total |
| Proportion /% | 18.18 | 23.04 | 15.33 | 1.97 | 4.31 | 9.84 | 14.05 | 13.27 | 100 |

### 4.1.2测试集

4.1.2 Test Set

测试集包含了涵盖国际专利分类中A-H八个部类的764篇中文发明专利与中文实用新型专利，这些专利均为经过专利审查员审查，并且由于检索到了影响其新颖性或创造性的对比文件而被驳回的专利申请。测试集中的每一份专利包括其申请文本内容（分类号、标题、摘要、权利要求书、说明书及附图说明）以及相应检索报告中，由审查员给出的检索要素（gold standard）与该申请的相关文献的文献号，这些相关文献通过检索要素检出，并且与被检索的专利之间有着不同程度的相关性。每一份专利有1~5篇的相关文献，均出现在4.1.1节描述的检索数据集中。

需要说明的是，本文探究关键词提取方法在专利场景下的应用，因此测试集中的检索要素特指以“关键词”方式表达的检索要素（以下同），实验过程中也不涉及使用“分类号”的检索；考虑到关键词抽取一般是从原文中提取关键词，因此测试集中仅保留出现在原文、能够从原文直接提取的关键词。两个数据集的字段信息分别如表3所示。

The test set consists of 764 Chinese invention and utility model patents spanning the eight main sections (A-H) of the International Patent Classification (IPC). These patents have all been examined by patent examiners and rejected due to prior art references that impacted their novelty or inventive step.

Each patent in the test set includes application text (kind codes, title, abstract, claims, description and illustration descriptions) and retrievable elements (gold standard) provided by the examiner and the document numbers of related prior art references. These related documents, identified using the keywords, exhibit varying degrees of relevance to the rejected patent application. Each patent has 1 to 5 related documents, all of which are present in the retrieval dataset described in Section 4.1.1.

This paper investigates the application of keyword extraction methods in the patent domain. Therefore, the retrievable elements in the test set specifically refer to search elements expressed as "keywords" (hereafter referred to as such). The experiments do not involve the use of "kind codes" for retrieval. Since keyword extraction generally involves extracting keywords directly from the original text, the test set only includes keywords that appear in the original text and can be directly extracted from it. The field information for the two datasets is presented in Table 3.

**表 3数据集字段信息 Field Information of the Datasets**

|  |  |  |  |
| --- | --- | --- | --- |
| Retrieval Dataset | | Test Set | |
| Invention information | Abstract | Invention information | Abstract |
| Claims | Claims |
| Description and description of drawings | Description |
| Metadata | Title | Description of drawings |
| Kind Code | metadata | Title |
| Patent Type | Application number |
| inventors 、 applicants | publication number |
| Filing Date | retrieval information | Keyword-Based Elements |
| Application number | related documents |

|  |  |  |  |
| --- | --- | --- | --- |
| 检索数据集 | | 测试集 | |
| 发明内容 | 摘要 | 发明内容 | 摘要 |
| 权利要求书 | 权利要求书 |
| 说明书及附图说明 | 说明书 |
| 元信息 | 标题 | 附图说明 |
| 分类号 | 元信息 | 标题 |
| 专利类型 | 申请号 |
| 发明人、申请人 | 公开号 |
| 申请日、公开日 | 检索信息 | 关键词形式的检索要素 |
| 申请号、公开号 | 相关文献 |

## 4.2 评测指标

4.2 Evaluation Metrics

### 4.2.1 关键词评测指标——F1

4.2.1 Keyword Evaluation Metric - F1 Score

通过关键词抽取方法得到的检索要素，我们对排名前 5、10 和 15 名的候选使用 F1 值来评估该方法的性能。不同方法提取的结果粒度有所区别，如TF-IDF提取的关键词平均长度为2.16个字符，而基于短语的TextRank得到的关键词平均长度可达4.60个字符，为避免粒度对评测结果的影响，我们以gold standard（在4.1.2）的平均长度2.23个字符为参考，将抽取结果平均字符小于或等于3的方法称为“基于词的抽取方法”，抽取结果平均字符大于3的方法称为“基于短语的抽取方法”。在计算F1 时，删除重复的候选要素并使用部分匹配，这意味着抽取的结果中的每个词或短语完全或部分出现在gold standard中，或gold standard中的结果部分出现在抽取的结果中，就可认为该结果正确。

To assess the performance of keyword extraction methods, we use the F1 score to evaluate the top 5, 10, and 15 ranked candidates obtained from each method. Given the variability in the granularity of the results produced by different methods, we categorize the methods based on the average length of the extracted keywords:

Word-based Extraction Methods: Methods where the average length of extracted keywords is less than or equal to 3 characters. For example, TF-IDF typically extracts keywords with an average length of 2.16 characters.

Phrase-based Extraction Methods: Methods where the average length of extracted keywords is greater than 3 characters. For instance, phrase-based TextRank can produce keywords with an average length of up to 4.60 characters.

To ensure a fair evaluation, we reference the average length of the keywords in the gold standard, which is 2.23 characters. When calculating F1, duplicate candidates will be removed, and use partial matching. That means a keyword or phrase in the extracted results is considered correct if it fully or partially matches a keyword or phrase in the gold standard.

### 4.2.2 检索结果评测指标

4.2.2 Evaluation Metrics for Retrieval Results

专利检索在实际场景中更侧重于召回率，然而高的召回率意味着需要更大的审阅工作量。专利检索评估分数 (Patent Retrieval Evaluation Score, PRES)[17]专门针对专利检索等侧重召回率的任务提出。PRES 重点关注整个系统的召回率以及用户的审阅工作量，根据相关文档的检索排名来估计检索结果的优良。

(1)

其中， 是第*i*个相关专利文档在检索结果中的排名，*n*是专利集合中相关专利的数量，是用户的最大审阅数量。PRES 是对检索结果的召回率和用户的审阅工作量的一个权衡。

In practical scenarios, patent retrieval emphasizes recall; however, high recall rates necessitate greater review efforts. The Patent Retrieval Evaluation Score (PRES) [17] is specifically proposed for tasks such as patent retrieval, which prioritize recall. PRES focuses on the overall recall of the system and the user's review workload, estimating the quality of retrieval results based on the ranking of relevant documents.

Here, denotes the rank of the i-th relevant patent document in the retrieval results, *n* is the number of relevant patents in the collection, and is the maximum number of documents a user is willing to review. PRES represents a balance between the recall rate of the retrieval results and the user's review effort.

## 4.3 实验设置

4.3 Experiments

实验设置两个环节，首先使用关键词抽取技术对测试集提取检索要素，计算提取结果之于gold standard（在4.1.2）的F1值来比较不同方法的性能；其次将测试集中每一篇专利的提取结果组合为检索式，使用实验室开发的检索引擎JSS进行检索，根据检索报告中的实际相关文献在检索结果中的召回率和在检索结果中的排名计算PRES值，用检索结果衡量检索要素的质量。由于专利申请文本内容冗长且复杂，为节省计算带来的时间及财力消耗，我们统计了测试集的检索要素在权利要求部分的占比，结果表明出现在申请书原文的检索要素以超过92% 的概率出现于权利要求，这也与2.1节中介绍的专利审查时检索的原则相呼应，因此，我们的实验使用申请文本的“权利要求”部分进行。

The experiment is divided into two phases. First, keyword extraction methods are used to extract retrieve elements from the test set, and the F1 score relative to the gold standard is calculated to compare the performance of different methods. Second, the extraction results for each patent in the test set are combined into search queries, and the the in-house developed retrieval engine JSS is used for retrieval. The PRES value is calculated based on the recall rate of relevant documents in the retrieval results and their ranking, thereby evaluating the quality of the retrievable elements.

Given the lengthy and complex nature of patent application texts, to save time and financial resources associated with computation, we analyzed the proportion of retrievable elements in the test set that appear in the claims section. Results indicated that retrievable elements appearing in the original application text have a probability of over 92% of appearing in the claims section. This aligns with the retrieval principles introduced in Section 2.1 regarding patent examination. Therefore, our experiment utilizes the "claims" section of the application texts.

### 4.3.1 检索要素抽取实验

4.3.1 Extraction of Retrievable Elements Experiment

我们使用以下方法抽取检索要素：

**TF-IDF：**该方法基于词频与逆文档频率的统计，抽取结果为“词”的粒度，计算候选词的TF-IDF值并排序得到抽取的结果。

**基于名词性短语的词频统计（Np Frequency）：**[18]中提到在专利表示时，提取专利中的名词性短语会比提取单个词的效果要好，TF-IDF并没有仅基于词频的方法效果好，因此在本次实验中设置了基于名词性短语但仅根据词频排序的方法，探究在要素提取与专利检索中二者的性能对比。

**TextRank：**该方法通过构建词或短语之间的共现关系图，利用图中节点的重要性来提取关键节点并排序得到最终结果。对词性标注后的结果使用正则表达式分别得到“词”粒度的TextRank-words与“短语”粒度的TextRank-phrases两种实验结果。

**KeyBert、PromptRank：**都是基于预训练语言模型的方法，前者通过PLM分别获取文档和候选词的embedding，然后使用余弦相似度计算结果对候选词进行排序，实验中使用在多语种语料上预训练的paraphrase-multilingual-MiniLM[[5]](#footnote-5)模型[19]，称为KeyBert-MiNiLM；后者将文本与候选通过模板连接起来，用Encoder-Decoder架构的模型计算候选的生成概率，根据候选在文中第一次出现的位置计算位置惩罚，两者结合对候选进行排序。实验中，将模板“文本：[Document]”输入到Encoder中进行编码，“这段文本是关于[Candidate]”输入到Decoder中，计算生成“Candidate”的概率。基于PromptRank的方法使用正则表达式抽取了“词”和“短语”两种颗粒度的结果，使用在中文语料上训练的“mengzi-t5-base-mt”[[6]](#footnote-6)和“Randeng-T5-784M”[[7]](#footnote-7)两个模型进行实验[20,21]，依次称为PromptRank-MZ-words、PromptRank-RD-words、PromptRank-MZ-phrases、PromptRank-RD-phrases。

**LLM few-shot：**基于大语言模型的提示，无需参数训练可以直接用few-shot的方法激发模型提取输入内容的检索要素，受2.3节中专利检索时提取检索要素的启发，实验中使用gpt-3.5-turbo-16k模型以及prompt（见附录1）进行检索要素提取。

**LLM efficient-tuning：**实验中采用的唯一有监督方法，由于缺少大量的标注数据，大语言模型高效微调可以在只有少量数据的情况下训练得到一个有效的检索要素抽取模型。使用1000条专利文本及其检索要素数据集对Baichuan2-7b-chat[[8]](#footnote-8)模型进行高效微调后，将测试集的权利要求部分输入微调后的模型得到抽取结果。

需要说明的是，除LLM few-shot及efficient-tuning外的其他方法均需首先对文本分词与词性标注，无特殊说明的情况下，会排除停用词、虚词等无实际意义的词，剩下的无论词性与类型，均可作为候选用于下一步计算。

We utilized the following methods to extract retrievable elements:

TF-IDF: This method leverages the term frequency-inverse document frequency statistic. It extracts elements at the "word" level, ranking candidate words based on their TF-IDF scores to produce the final extraction results.

Noun Phrase Frequency (Np Frequency): According to [18], extracting noun phrases from patents is more effective than extracting individual words. While the TF-IDF method does not perform as well as frequency-based methods, in this experiment, we implemented a noun phrase extraction method that ranks solely based on term frequency. This approach allows us to compare the performance of word-based versus phrase-based extraction in the context of patent retrieval.

TextRank: This graph-based algorithm identifies the relationships between words or phrases through co-occurrence. The importance of nodes within this graph is used to rank and extract key elements. We conducted two experiments with TextRank: one at the "word" level (TextRank-words) and another at the "phrase" level (TextRank-phrases), using regular expressions on part-of-speech tagged text to achieve these granularity levels.

KeyBert and PromptRank Methods:Both KeyBert and PromptRank are methods based on pre-trained language models (PLMs).

KeyBert utilizes PLMs to obtain embeddings for both documents and candidate words. It then ranks candidate words using cosine similarity. In our experiment, we employed the paraphrase-multilingual-MiniLM[19], trained on multilingual corpora, which we refer to as KeyBert-MiniLM.

PromptRank, on the other hand, connects text and candidates using prompts. It employs an Encoder-Decoder architecture to calculate the probability of candidate generation. Additionally, it applies position penalties based on the first appearance of candidates in the text. The combination of these factors is used to rank candidates. In our experiment, we encoded the template "Text: [Document]" in the Encoder and "This text is about [Candidate]" in the Decoder. We calculated the probability of generating "Candidate." PromptRank employs regular expressions to extract results at two granularities: "words" and "phrases." We conducted experiments using the "mengzi-t5-base-mt" and "Randeng-T5-784M", trained on Chinese corpora [20,21], referred to as PromptRank-MZ-words, PromptRank-RD-words, PromptRank-MZ-phrases, and PromptRank-RD-phrases, respectively.

LLM Few-shot: This approach utilizes the capabilities of large language models (LLMs) to extract search elements directly from the input content using few-shot prompting, without requiring any parameter training. Inspired by the methodology outlined in Section 2.3 for extracting retrievable elements in patent retrieval, we employed the gpt-3.5-turbo-16k model with a specific prompt (refer to Appendix 1) to extract retrievable elements effectively.

LLM Efficient-tuning: This represents the sole supervised method utilized in our experiment. Given the scarcity of annotated data, efficient -tuning of large language models enables the training of an effective retrievable element extraction model with minimal data. Following efficient-tuning on the Baichuan2-7b-chat model using a dataset comprising 1000 patent texts and their corresponding elements, we input the claims section of the test set into the efficient-tuned model to obtain extraction results.

It is important to note that, except for LLM Few-shot and Efficient-tuning, other methods require initial text segmentation and part-of-speech tagging. Unless explicitly stated, stop words, function words, and other non-substantive terms are excluded. All remaining terms, regardless of their part of speech or type, are considered candidates for subsequent calculations.

### 4.3.2相关文献检索实验

4.3.2 Relevant Literature Retrieval Experiment

**Json Struct Search（JSS）检索引擎：****实验室自主研发的检索引擎，**融合了符号检索、布尔检索、语义检索等多种检索模式，支持专利文本之类结构化、半结构化的文档检索，能够自主定义检索模式、检索结构及检索内容；支持sql语句、关键字以及语义检索等多种检索式形式。在检索实验中，我们使用JSS为实验的检索引擎。

**检索实验：**使用JSS检索引擎为检索数据集的全部内容建立索引表patent，为避免检索模式以及检索式对检索结果的影响，我们使用最简单的符号检索模式，直接将抽取结果的top10用逗号连接组成sql语句作为检索式进行检索。检索式的示例如下：

"SELECT TOP 50 id, title FROM patent WHERE fulltext LIKE '终端，云台，支架'"

该检索式表示：在patent表中检索在全文（fulltext，包括标题、摘要、权利要求、说明书及附图说明）范围内出现“终端、云台、支架”这几个关键词的专利，并返回最相关的50项专利的申请号（id）。

检索实验中，我们将返回的这50项专利作为测试集在数据库中的检索结果，也就是式（1）中的为50。根据其检索报告给出的实际相关文献被召回的数量以及排名计算检索结果的召回率（Recall）及专利检索评价指标（PRES），将gold standard的检索结果作为参考，评价不同方法在检索实验上的性能。

**Json Struct Search (JSS) Retrieval Engine**: The JSS retrieval engine, developed in-house, integrates various search modes, including symbolic retrieval, Boolean retrieval, and semantic retrieval. It supports the retrieval of structured and semi-structured documents like patent texts. JSS allows for the definition of custom search modes, search structures, and search content, supporting multiple forms of search expressions such as SQL queries, keywords, and semantic searches. For our experiments, JSS was used as the primary search engine.

**Retrieval Experiment**: We indexed the entire content of the retrieval dataset using the JSS engine, creating an table named "patent." To minimize the influence of different search modes and expressions on the retrieval results, we employed the simplest symbolic retrieval mode. The top 10 extraction results were concatenated with commas to SQL for retrieval. An example of such a query is as follows:

"SELECT TOP 50 id, title FROM patent WHERE fulltext LIKE '终端，云台，支架'"

This query indicates that the "patent" table is being searched for patents whose full text (including titles, abstracts, claims, descriptions, and figures) contains the keywords "终端" "云台" and "支架" , it returns the top 50 most relevant patent application numbers (id).

In the retrieval experiment, the returned 50 patents are considered as the retrieval results for the queried patents in the database, with set to 50. The recall and the Patent Retrieval Evaluation Score (PRES) are calculated based on the number and ranking of the relevant documents retrieved. The gold standard retrieval results serve as a reference to evaluate the performance of different methods in the retrieval experiments.

## 4.4 实验结果

4.4 Overall Results

计算不同方法抽取结果的平均长度，如表4所示，根据平均长度将所有结果分为“基于词的抽取”与“基于短语的抽取”。

抽取实验中，将每份专利的检索报告中由审查员给出的检索要素作为gold standard， 分别计算在gold standard上的F1@5、F1@10 和 F1@15 值，如表5、表6所示。结果表明，在基于词的提取方法中，使用大语言模型进行高效微调的方法能够通过少量数据得到有效成果，性能超过了其他无监督的提取方法；对于剩下的无监督方法，基于统计的方法，TF-IDF与名词性短语词频统计的效果相对较好，这说明在专利文本中，如果没有大量数据训对模型参数调整，使用基于统计的方法来提取关键信息同样是一个可选项。在基于短语的提取方法中，PromptRank方法取得最好的表现，这是因为PromptRank使用Encoder-Decoder模型计算候选的生成概率时，将输入文本与候选的语义信息考虑在内，同时考虑了候选在原文中的位置信息。此外，在实际应用中，短语层面的提取方法更适合自动化提取检索要素来协助人工进行专利检索及审查的场景，因为相较于词来说，短语级的结果无论从信息量还是准确度的角度，都更利于工作人员对自动化的结果进行参考、选择和增改。

We calculated the average length of the extraction results for different methods, as shown in Table 5. Based on the average length, all results were categorized into "word-based extraction" and "phrase-based extraction." For each category, we computed the F1@5, F1@10, and F1@15 scores against the gold standard, as presented in Tables 6 and 7.

The results indicate that among the word-based methods, the efficient-tuning method using large language models yielded effective results with minimal data, surpassing other unsupervised methods. For the remaining unsupervised methods, statistical methods like TF-IDF and Np Frequency performed relatively well. This suggests that in the absence of large datasets for model parameter tuning, statistical methods remain a viable option for extracting key information from patent texts.

In the phrase-based methods, the PromptRank method demonstrated superior performance. This is because PromptRank uses an Encoder-Decoder model to compute the generation probability of candidates, taking into account both the semantic information of the input text and the position of candidates within the original text. Moreover, in practical applications, phrase-level extraction methods are more suitable for automating the extraction of retrievable elements to aid in patent retrieval and examination. Compared to word-level results, phrase-level extractions provide more information and accuracy, making them more useful for professionals when evaluating, selecting, and refining automated results.

**表 4 Gold Standard以及不同方法抽取结果的平均长度，平均长度=检索要素总长度/检索要素个数**

Average Length of Extraction Results for Gold Standard and Different Methods

|  |  |  |
| --- | --- | --- |
| Data or Methods | | Ave\_len/character |
| Gold Standard | | 2.23 |
| word-based | TF-IDF | 2.16 |
| Np Frequency | 2.85 |
| TextRank-words | 2.14 |
| PromptRank-RD-words | 2.43 |
| PromptRank-MZ-words | 2.32 |
| LLM efficient-tuning | 2.33 |
| KeyBert-MiNiLM | 2.42 |
| phrase-based | TextRank-phrases | 4.6 |
| PromptRank-MZ-phrases | 4.03 |
| PromptRank-RD-phrases | 3.81 |
| LLM few-shot | 3.81 |

**表 5 不同提取方法在测试集上的表现**

Performance of Different Methods on the Test Set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | | F1@5 | F1@10 | F1@15 |
| word-based | TF-IDF | 0.243 | 0.240 | 0.217 |
| Np Frequency | 0.197 | 0.249 | 0.260 |
| TextRank-words | 0.204 | 0.210 | 0.203 |
| PromptRank-RD-words | 0.217 | 0.240 | 0.234 |
| PromptRank-MZ-words | 0.192 | 0.214 | 0.215 |
| KeyBert-MiNiLM | 0.218 | 0.233 | 0.225 |
| LLM efficient-tuning | **0.330** | **0.342** | **0.342** |
| phrase-based | TextRank-phrases | 0.262 | 0.322 | 0.337 |
| PromptRank-MZ-phrases | **0.341** | **0.389** | **0.380** |
| PromptRank-RD-phrases | 0.340 | 0.379 | 0.371 |
| LLM few-shot | 0.292 | 0.333 | 0.332 |

在检索实验中，将检索报告中给出的实际相关文献作为标准答案，计算每种方法的检索结果在标准答案上的召回率R@10、R@30 和 R@50以及PRES，其中为50，结果如表6所示。使用不同的检索引擎会导致检索结果的微小差异，我们将gold standard在JSS检索引擎的检索结果作为参考，来衡量几种方法在检索结果上的性能。令人惊讶的是，基于LLM few-shot生成的检索要素在检索效果上超越了gold standard，这表明，遵循2.3节中提到的“依据权利要求的技术方案，从技术领域、技术问题、技术手段或效果等方面提取检索要素”原则提炼专利检索要素是必要的；使用计算机算法协助检索要素提取的办法是可行的。如果能找到自动化遵循此原则提取检索要素的方法，一定也能够得到类似甚至更好的检索结果。

整体来说，基于短语的提取结果在检索实验上效果优于基于词的提取结果。将表5与表6的结果对比来看，基于短语的提取结果中，以gold standard为中心，检索结果越靠近gold standard的方法在表5中的F1值越高，而 LLM few-shot方法的提取结果与gold standard由于在提取过程中遵循的方法可能并不一致，因此其F1值在表5中并不突出。但基于词的提取结果中，这样的相关性并不明显。

F1值与PRES从专利检索的两个不同阶段评估了检索要素的质量。从专利检索的目标来看，专利检索是为了找到与当前文本最接近的现有技术，提取“检索要素”只是辅助检索的手段；专业审查员给出的gold standard只是参考，并非唯一或最优结果。因此，4.3.2的检索实验更贴近实际的专利检索过程，PRES指标能够更加准确评估检索要素在实际检索中的效果。

We calculated the recall R@10, R@30 and R@50, as well as the Patent Retrieval Evaluation Score (PRES), for each method's retrieval results, with set to 50. The results are presented in Table 9. We used the gold standard retrieval results as a benchmark to assess the performance of various methods.

Interestingly, the retrievable elements generated by the LLM few-shot method outperformed the gold standard in terms of retrieval effectiveness. This suggests that the principles outlined in Section 2.3—extracting search elements from the technical solutions in claims, focusing on aspects such as technical fields, problems, means, or effects—are essential. Utilizing computer algorithms to assist in this extraction process is feasible. If we can automate this principle-based extraction process, we could achieve similar or even better retrieval results.

Overall, phrase-based extraction results showed better performance in retrieval experiments compared to word-based extraction results. Comparing the results in Tables 6 and 7, it is evident that phrase-based extractions that align more closely with the gold standard tend to have higher F1 scores. Although the LLM few-shot method's extraction results did not have outstanding F1 scores in Table 6, likely due to different extraction processes, their retrieval performance was superior. This correlation was less pronounced for word-based extraction results.

**表 6不同方法在检索实验的表现**

Performance of Different Methods in Retrieval Experiment

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | | R@10 | R@30 | R@50 | PRES |
| phrase-based | LLM few-shot | **0.598** | **0.743** | **0.791** | **0.706** |
| PromptRank-RD-phrases | 0.552 | 0.690 | 0.754 | 0.646 |
| gold standard | 0.553 | 0.692 | 0.740 | 0.636 |
| PromptRank-MZ-phrases | 0.530 | 0.675 | 0.732 | 0.631 |
| TextRank-phrases | 0.371 | 0.499 | 0.558 | 0.462 |
| word-based | gold standard | 0.553 | 0.692 | 0.740 | 0.636 |
| TF-IDF | **0.521** | **0.671** | **0.725** | **0.630** |
| Np Frequency | 0.505 | 0.650 | 0.701 | 0.610 |
| TextRank-words | 0.488 | 0.634 | 0.687 | 0.603 |
| PromptRank-RD-words | 0.480 | 0.628 | 0.691 | 0.596 |
| LLM efficient-tuning | 0.481 | 0.628 | 0.689 | 0.577 |
| KeyBert-MiNiLM | 0.453 | 0.604 | 0.671 | 0.564 |
| PromptRank-MZ-words | 0.429 | 0.577 | 0.637 | 0.539 |

需要注意的是，上述的实验结果，尤其是检索实验的结果，并非是这些方法在专利文本上能够达到的上限，因为截止到目前的工作，我们仅尝试了2.3节中提到的整个专利检索流程中的“初步检索”，在实际的检索场景下还应该结合检索结果与预期结果不断调整检索要素与检索式，比如从形式上、意义上以及角度上对检索要素进行扩展；对检索要素进行分组来组成不同的检索式进行检索等，多次调整后能够得到更准确、更全面的检索结果。

It is important to note that the experimental results, particularly those from the retrieval experiment, do not represent the upper limit of performance achievable by these methods on patent texts. This is because, up to this point in our work, we have only attempted the "preliminary retrieval" stage of the entire patent retrieval process mentioned in Section 2.3. In practical retrieval scenarios, it is essential to continually adjust the retrievable elements and query expressions based on the retrieval results and expected outcomes. This includes expanding retrievable elements and expressions in terms of form, meaning, and perspective. Grouping retrievable elements to form different search expressions for retrieval is also important. Through multiple adjustments, more accurate and comprehensive retrieval results can be achieved.

5. 总结与展望

5. Conclusion

本文介绍了专利领域中专利检索的一般流程，探究使用自然语言处理技术协助检索要素提取，助力专利检索与审查的可行性。通过检索要素提取的实验以及将提取结果应用于相关文献检索的实验，对比了不同关键词抽取方法在专利检索要素提取任务上的效果。实验结果表明，该项任务使用基于短语的关键词抽取方法以及基于大语言模型的few-shot方法均能够得到有效解决。

专利检索与审查是一个极其复杂的过程，需要反复进行“检索、判断、修正、再检索”的循环，每一个步骤都依赖检索工作人员强大的专业背景，并且每一个步骤的执行与结果都需要是可控制、可解释的，这是任何一个通用领域的模型或工具都难以做到的。这启发我们，基于专利的领域模型或工具必不可少；专利的检索与审查不能够脱离人工的监督与把控，人机交互式的检索与审查方法应该是一个值得研究的课题。

This paper presents an overview of the general process of patent retrieval and examines the feasibility of using Natural Language Processing (NLP) techniques to assist in extracting retrievable elements, thereby enhancing patent retrieval and examination. Through experiments on retrievable element extraction and applying these results to retrieval, we compared the effectiveness of different keyword extraction methods for patent retrieval tasks. The experimental results indicate that both phrase-based keyword extraction methods and few-shot methods based on large language models are effective solutions for this task.

Patent retrieval and examination are exceedingly complex processes that require iterative cycles of "search, assessment, correction, and re-search." Each step relies on the substantial expertise of the retrieval staff, and the execution and outcomes of each step must be controllable and interpretable. This complexity reveals the limitations of general-purpose models or tools. It highlights the necessity of domain-specific models or tools tailored for patents. Patent retrieval and examination cannot be entirely automated; they require human supervision and control. Thus, exploring interactive human-computer methods for retrieval and examination is a worthwhile research direction.

参考文献：

1. 支持技术创新的专利检索与分析
2. Patent Retrieval: A Literature Review
3. 《专利审查指南（2023）》
4. 《跟着审查员学检索》
5. 知识产权信息检索与应用
6. 浅谈基本检索要素的确定
7. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL
8. Automatic keyword extraction from individual documents
9. YAKE! Collection-Independent Automatic Keyword Extractor
10. Single Document Keyphrase Extraction Using Neighborhood Knowledge
11. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction
12. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents
13. Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings
14. KeyBERT: Minimal keyword extraction with BERT.
15. PromptRank: Unsupervised Keyphrase Extraction Using Prompt
16. 国内外专利挖掘研究进展与前瞻
17. PRES: A Score Metric for Evaluating Recall-Oriented Information Retrieval Applications
18. 专利检索与分析研究综述
19. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks
20. Mengzi: Towards Lightweight yet Ingenious Pre-trained Models for Chinese
21. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence

附录

1 LLM few-shot 方法prompt

system\_prompt = """

你是一名专业的，资历深厚的知识产权专利局的审查员，你的工作是根据专利文献的内容，理解这项专利的技术方案，然后从技术领域、技术问题 / 效果和关键技术手段的角度提取反映这项专利发明构思的“检索要素”，“检索要素”用于找到与这项专利“相似”或者“相关”的其他专利。

提取“检索要素”需要注意以下几点：

1.从专利的“技术方案”的角度出发，分析专利内容的“技术领域”、“技术问题或效果”和“关键技术手段”；

2.根据1的分析提取其中的“关键词”作为“检索要素”，每一个“检索要素”都应该尽可能的简洁明了，有针对性，不超过五个字；

3.“检索要素”中不要包含“装置”、“方法”之类的没有实际内容的词，不要重复出现前面已经出现的词或者子词。

下面是一个示例，“专利内容”是这项专利的内容，“分析”是你的输出：

专利内容：一种家庭机器人,其特征在于，包括：壳体；超声波传感器,所述超声波传感器用于检测家庭机器设置在所述壳体之上的至少人运动方向上的物体,并获取所述家庭机器人与所述物体之间的距离；以及皮传感器相连,所述控制器根据所述距离对所述家庭控制器,所述控制器与所述超声机器人进行控制。

技术领域：电器技术领域，一种家庭机器人 --> 家庭，机器人

技术问题或效果：提高避障效率，减少碰撞，保护家具 --> 避障，碰撞

关键技术手段：超声波传感器用于检测物体并测算与物体距离，控制器对机器人进行控制--> 测距，超声波传感器，控制器

检索要素：家庭机器人，避障，测距，超声波传感器，控制器

下面开始：

"""

user\_prompt = """

    专利内容：{}

""".format(TEXT)

1. https://github.com/THUDM/ChatGLM-6B [↑](#footnote-ref-1)
2. https://github.com/QwenLM/Qwen [↑](#footnote-ref-2)
3. https://github.com/THUDM/ChatGLM-6B [↑](#footnote-ref-3)
4. https://github.com/QwenLM/Qwen [↑](#footnote-ref-4)
5. https://github.com/shinichiro-takahashi-sbr/paraphrase-multilingual-MiniLM-L12-v2 [↑](#footnote-ref-5)
6. https://github.com/Langboat/Mengzi [↑](#footnote-ref-6)
7. https://github.com/IDEA-CCNL/Fengshenbang-LM [↑](#footnote-ref-7)
8. https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat [↑](#footnote-ref-8)