

# Supplementary Materials for “Innovating Patent Retrieval: A Comprehensive Review of Techniques, Trends, and Challenges in Prior Art Searches”

Amna Ali <sup>1,\*</sup>, Ali Tufail <sup>2</sup>, Liyanage Chandratilak De Silva <sup>2</sup> and Pg Emeroylariffion Abas <sup>1,\*</sup>

<sup>1</sup> Faculty of Integrated Technologies, Universiti Brunei Darussalam, Gadong BE1410, Brunei

<sup>2</sup> School of Digital Science, Universiti Brunei Darussalam, Gadong BE1410, Brunei;  
liyanage.silva@ubd.edu.bn (L.C.D.S.)

\* Correspondence: aamna.hamed@gmail.com (A.A.); emeroylariffion.abas@ubd.edu.bn (P.E.A.)

## S1. Introduction

This Supplementary Materials document accompanies the review article titled “Innovating Patent Retrieval: A Comprehensive Review of Techniques, Trends, and Challenges in Prior Art Searches.” It is designed to provide readers with a deeper understanding of the detailed aspects of the study that could not be fully covered in the main article due to space constraints. This document includes additional data, extended analyses, and further methodological details intended to enhance the comprehensiveness of the research findings and support further scholarly inquiry.

### Contents of this Supplementary Document:

1. S1. Introduction: An overview of the supplementary materials and guidance on how to navigate and use the additional content provided.
2. S2. Supplementary Materials to Section 2 Background: This section includes extended content that supports the background information discussed in the main manuscript. It provides additional context, data, and historical perspectives that underpin the research questions addressed in the study.
3. S3. Supplementary Materials to Section 3 Review Methodology:
  - a. S3.1 Supplementary Materials to RQ2: Detailed explanations and formulas for the evaluation metrics used in assessing the effectiveness of patent retrieval methods.
  - b. S3.2 Supplementary Materials to RQ3: Additional tables and descriptions that provide deeper insights into the methodologies used for patent retrieval tasks during specific NTCIR events.
  - c. S3.3 Supplementary Materials to RQ4: Expanded content on the use of NLP techniques in patent retrieval, illustrated through figures and detailed descriptions.
  - d. S3.4 Supplementary Materials to RQ5: A comprehensive breakdown of the components of a patent document, aiding in understanding the structure and functionality of patents in the context of retrieval processes.

Each section of these supplementary materials is designed to be self-contained, offering extensive background and context to enrich the reader’s understanding of the topics discussed in the main manuscript. By providing these detailed supplementary materials, the document aims to serve as a valuable resource for researchers, academics, and practitioners interested in the nuances of patent retrieval processes and the efficacy of various techniques explored throughout the study.

## S2. Supplementary Materials to Section 2 Background.

This section of the Supplementary Materials provides additional data and explanations that expand upon and support the background information discussed in Section 2 of the main manuscript. The supplementary materials in this section help to deepen the reader's understanding of the historical context, foundational theories, or key concepts that frame the research questions addressed in the study.

### *S2.1 Patent Retrieval Tasks*

The State of the Art (SOA) search is commonly performed by an inventor or researcher at the pre-R&D stage to seek a grasp of the latest developments within a specific technological domain. This comprehensive inquiry not only aids in understanding current technological advancements but also in identifying potential areas for innovation. The outcomes of an SOA search typically culminate in a detailed patent landscape report, providing an extensive overview of the domain, and highlighting existing technologies, key players, and potential gaps in the market, thereby guiding inventors or researchers toward the most promising directions for development or research.

Once a specific invention concept is established, a pre-filing patentability search is performed to ensure the invention's novelty and non-obviousness before filing a patent application. Insights from the SOA search help tailor the pre-filing search to the invention's context, enhancing the effectiveness of identifying prior art that could affect the patentability of the new invention. The pre-filing patentability search aims to ensure that the invention meets the criteria for patentability before significant resources are invested in the R&D and patent application process. This search can influence the decision on whether to proceed with the patent application and may also guide modifications to the invention or its application strategy to avoid infringing on existing patents.

Following a decision to proceed with a patent application based on a Pre-Filing Patentability Search, detailed and formal Patentability Search are undertaken by examiners. This step involves an exhaustive review of existing patent documents to confirm novelty, inventive step, and industrial applicability—key criteria for patentability of the proposed invention. The objective is to ensure that the invention stands unique against the backdrop of existing patents and adheres to all patentability requirements, laying a foundational step towards securing a patent grant. A series of iterative reviews between examiners and inventors normally occur, involving detailed reports and responses to address concerns over patentability. This is aimed at refining the application to meet patentability criteria and can lead to either the granting of the patent or abandonment of the application if objections cannot be resolved.

Even before the grant of a patent, the inventor can already commercialize it as a product or process, potentially incorporating various Intellectual Property (IP) elements. Conducting a Freedom To Operate (FTO) search before product launch ensures there is no infringement on active patents, clearing legal pathways for market entry. The FTO analysis is a preventive measure to mitigate the risk of patent infringement litigation, ensuring that the commercial exploitation of the patent respects existing Intellectual Properties (IPs) including other active patents. Additionally, a proactive Infringement Search may also be performed during the product development stage to identify potential patent violations early. While FTO assesses commercialization risks in targeted markets, Infringement Searches are broader, focusing on avoiding legal issues during development. If potential infringements are identified, measures can be taken to legally leverage the infringed patent, such as negotiating a licensing agreement or patent acquisition. Alternatively, an Invalidity Search can be initiated to challenge the validity of the infringed patent, seeking prior art that could render the patent invalid. This strategy is crucial for ensuring both innovation and compliance, addressing commercialization risks and legal obligations efficiently.

After commercialization, the new product or process, often an amalgamation of various Intellectual Properties (IPs), may attract scrutiny from competitors. These competi-

tors may conduct Infringement Searches to check for potential IP violations, possibly leading to infringement charges. In response, the originating company may also perform an Infringement Search to evaluate the legitimacy of these claims. Depending on the outcome, strategies such as negotiating licensing agreements, acquiring the questioned patents, or commencing a counter Invalidity Search to contest the validity of the claims are viable options. These proactive measures are vital for navigating through potential legal intricacies and ensuring the commercial venture adheres to IP laws, safeguarding the innovation's unique contributions and its composite IP framework.

Indeed, conducting a Patent Portfolio Search is essential for assessing the comprehensive value and strategic alignment of a patent collection throughout the innovation lifecycle. This analysis aids in making informed decisions regarding the management, protection, and capitalization of IP assets. Such a search is pivotal for companies aiming to delineate their competitive advantage, pinpoint deficiencies in IP coverage, or identify avenues for innovation and growth, thereby ensuring a well-rounded and forward-thinking approach to intellectual property strategy.

### S2.1 Existing Works on Retrieval Tasks

Table S1 offers a detailed analysis of the shortcomings identified in previous literature survey studies within the field of patent retrieval. By documenting these limitations, the table provides a clear justification for the necessity of this current review, highlighting the gaps and challenges that have persisted in the domain. This comprehensive summary underscores the importance of our study, demonstrating the ongoing need for advancements and fresh perspectives in patent retrieval research. It ensures that researchers recognize areas that are ripe for further investigation and innovative approaches, thereby setting the stage for this review's contributions to the field. Table S1 not only supports the background information provided in Section 2 of the main manuscript but also elaborates on the motivations behind our systematic review by identifying critical areas where previous research has fallen short.

**Table S1.** Comprehensive Summary of Existing Works on Retrieval Tasks in Patent and Information Retrieval Research.

Study	Deficiencies	References
Highlights numerous research efforts targeted at refining existing information retrieval strategies or applying standard procedures in various stages of patent retrieval	Lacks detailed sectioning of the information gathered from the articles under review Not enough graphical and tabular representations of articles based on different aspects that are crucial to patent retrieval Discusses patent retrieval in general and doesn't specifically talk about patent prior art retrieval.	[12]
Highlights the need for patent domain-specific modifications in the retrieval systems and also suggests a need to create interactive search tools compatible with the practices and requirements of patent domain experts	Lacks to discuss the latest state-of-the-art Natural Language Processing (NLP) approaches Broadly mentions patent retrieval, rather than in detail review of patent prior art retrieval	[8]
Examines the way information retrieval research has influenced and altered patent search strategies till 2013	Lacks the coverage of the most recent advancements in patent retrieval techniques because it was conducted more than ten years ago	[20]
Examines the use of deep learning in patent analysis by summarizing state-of-the-art approaches and categorizing 40 research publications	Sole focus is on deep learning methodologies for patent analysis, hence the lack of complete comparison with conventional or hybrid approaches	[16]

	Broadly mentions patent retrieval, rather than in detail review of patent prior art retrieval	
Examines the existing Natural Language Processing (NLP) methodologies for summarizing, simplifying, and generating patent texts, whilst acknowledging the unique challenges posed by patents in the research and development process	Does not directly address the patent retrieval process	[21]
Probes the current landscape of patent analysis, its various tasks and the prevalent tools and methodologies for efficient patent analysis, as well as the limitations of the existing tools.	Briefly touches on patent retrieval tasks and predominantly it centers around various aspects of patent analysis	[22]
Performs a comparative analysis of various pre-application prior art search strategies, including partial application search and query reformulation approaches from the perspective of inventors determining the patentability of their ideas before filing a full application	Sole emphasis has been on query reformulation, with no thorough investigation of alternative patent retrieval techniques	[23]
Employs bibliometric and keyword-based network analysis to map out the evolution of patent analysis and patent mining	Discuss patent retrieval as a single component of a larger investigation within the subject of patent mining	[24]
A comprehensive survey of the latest trends in data mining relevant to patents, to enrich the understanding of the patent analysts on the landscape of data mining	Discuss patent retrieval as a single component of a larger investigation within the subject of patent mining	[25]

### S3. Supplementary Materials to Section 3 Review Methodology

This section of the Supplementary Materials provides detailed information that complements and supports the review methodology discussed in the main manuscript. These supplementary materials offer a deeper insight into the methodologies, data sources, and analytical techniques used in our study, specifically addressing the research questions posed. By providing this additional context and detail, we aim to enhance the transparency, reproducibility, and comprehensiveness of our research findings.

#### S3.1 Supplementary Materials to RQ2

This section provides detailed explanations and formulas for the key evaluation metrics used to assess the effectiveness of patent retrieval methods discussed in response to Research Question 2 (RQ2). These metrics are critical for understanding both the efficiency and accuracy of the retrieval techniques employed in the study.

When undertaking a patent retrieval task with a specific patent, whether for a new application of an innovative idea, the objective is to identify patents within a database that are pertinent to the referenced patent. This process entails a detailed examination of the patent's claims, descriptions, and technical sphere to formulate search strategies capable of pinpointing similar inventions. The goal is to discover patents with shared technological, functional, or inventive traits, thereby providing a holistic view of the related prior art. Such an extensive search identifies not only exact matches but also patents with sufficient similarities to be deemed relevant, enriching the understanding of the patent environment surrounding the new invention. This method strives to optimize the identification of relevant patents (true positives, TP) and minimize the misidentification of irrelevant patents as relevant (false positives, FP) or the oversight of pertinent patents (false

negatives, FN), ensuring the collection's most relevant patents are precisely retrieved. Recall, Precision, F1-score, Mean Average Precision (MAP), and accuracy are performance measures utilized in the literature for patent retrieval tasks.

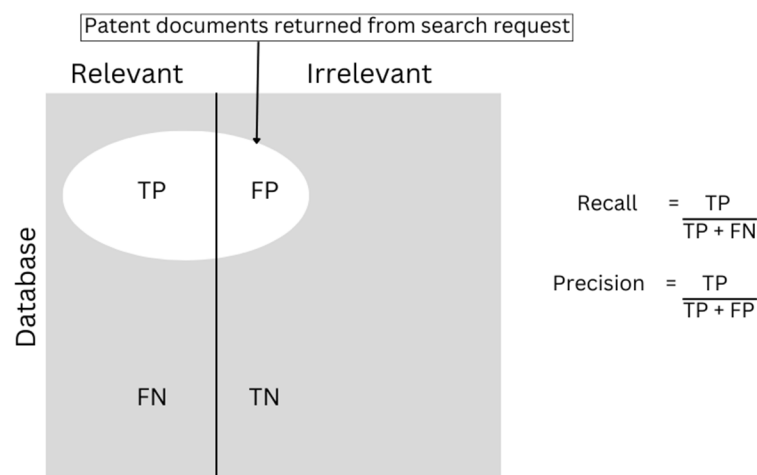
**Recall:** This measures the fraction of relevant patents that were successfully identified as relevant, highlighting the comprehensiveness of the method in identifying relevant patent documents in the collection. High recall indicates that the method is effective at capturing a greater percentage of the relevant information [100].

$$Recall = \frac{TP}{TP + FN} \quad (S1)$$

**Precision:** This measures the proportion of identified patents that are truly relevant, emphasizing the accuracy of the retrieval. High precision indicates that the method returns more relevant patent documents in the results [100].

$$Precision = \frac{TP}{TP + FP} \quad (S2)$$

The precision-recall trade-off and the method's ability to precisely identify relevant patent documents while reducing the presence of irrelevant patent documents in response to a prior art search request is depicted in Figure S1. It shows a share of relevant patent documents and irrelevant patent documents in a collection of patents in relation to a specific prior art search request. The circled region shows the true positives (relevant patents correctly identified) and false positives (irrelevant patents mistakenly retrieved) returned by the search request.



**Figure S1.** Recall and Precision.

**F1-score:** This metric combines both Precision and Recall into a single metric, providing a balanced view of the overall effectiveness of the retrieval task by considering both precision and completeness of the search results. Its effectiveness is beneficial when there is an imbalance between relevant and irrelevant documents [101], which is expected when doing a patent retrieval task.

$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (S3)$$

**Average Precision:** This measures the quality of a retrieval system's ranked results for a given query by taking into account the precision at various recall levels [102]. Unlike Precision, AP considers each item in the ranked list.

$$AP = \frac{1}{n} \sum_{k=1}^n P(k) \times rel(k) \quad (S4)$$

where  $k$  iterates over positions where relevant items are retrieved,  $n$  denotes the total number of retrieved patents for a particular query,  $P(k)$  denotes “Precision at position  $k$ ” in the ranked list of top  $k$  retrieved patents, and  $rel(k)$  denotes “Relevance at position  $k$ ”, and serves as an indicator function showing whether the patent at position  $k$  in the ranked list is relevant (1) or not relevant (0) to the given query.

**MAP (Mean Average Precision):** This is an extension of Average Precision (AP) metric that evaluates the overall effectiveness of a patent retrieval system across several queries. [117].

$$MAP = \left( \frac{1}{Q} \right) * \sum (AP(q)) \quad (S5)$$

where  $Q$  represents the total number of queries and  $AP(q)$  represents the Average Precision for query  $q$ .

**Accuracy:** This quantifies the ability to correctly identify relevant patents while effectively excluding irrelevant ones. It measures the success of the method in delivering precise results by balancing true positives (TP) and true negatives (TN) against the total number of patents examined. Achieving high accuracy is crucial, as it ensures that the search results closely match the user's needs, reducing the time and resources spent on sifting through non-relevant patents and enhancing the efficiency of the patent search process.

$$Accuracy = \frac{(TP + TN)}{TP + FP + FN + TN} \quad (S6)$$

However, accuracy can be misleading in evaluating the performance of a patent retrieval system due to its inability to effectively handle imbalanced data distribution. In common scenarios where the majority class (e.g., non-relevant documents) significantly outweighs the minority class (e.g., relevant documents) in terms case numbers, a system can still attain high accuracy by simply classifying all instances as belonging to the majority class, despite being unable to find relevant cases. This deceptively high accuracy ignores the main objective of retrieving relevant data. More detailed metrics such as recall and precision focus specifically on the retrieval of relevant documents, making them more suitable for assessing the effectiveness of a system in information retrieval [104].

### S3.2 Supplementary Materials to RQ3

This sub-section contains supplementary materials that provide further details on the methodologies and results discussed in relation to Research Question 3 (RQ3). Each table offers an in-depth look at specific patent retrieval tasks and datasets, shedding light on the varied approaches and tools used in the research. These tables ensure a comprehensive understanding of the technical aspects and data sources relevant to the patent retrieval studies.

#### S3.2.1 Table S2: Variations of the CLEF-IP Dataset

Table S2 provides a comprehensive overview of the variations in the CLEF-IP dataset from 2009 to 2013, as utilized in various patent retrieval tasks. Each iteration of the CLEF-IP dataset was tailored to meet specific research needs, ranging from prior art searches to classification and image-based tasks. This table details the unique components and focus of each year's dataset, including the size of the corpus and topic pools, the range of documents included, and the specific challenges addressed. It also outlines the criticism and modifications each dataset version received to better align with the evolving needs of the patent retrieval community.

By summarizing these variations, Table S2 not only serves as a valuable resource for researchers looking to understand the historical context and development of the CLEF-IP datasets but also aids in selecting the appropriate dataset version for future studies based on past applications and their outcomes. This detailed breakdown enhances the transparency of the dataset selection process in the studies cited throughout the main document and supports a deeper understanding of the methodological choices made in the field of patent retrieval.

**Table S2.** CLEF-IP Corpus Variants.

	Details	Corpus pool/Topic pool	References
<b>CLEF-IP 2009</b>	<ul style="list-style-type: none"> <li>Specifically customized for prior art search tasks</li> <li>Received criticism for selecting topics from granted patent documents instead of patent application documents (a practice that is opposite to the typical process in patent searches), which may have impacted the relevance of the documents collected</li> </ul>	<p>Corpus pool:</p> <ul style="list-style-type: none"> <li>Comprises around 2 million documents related to around 1 million unique patents, published between 1985 and 2000</li> </ul> <p>Topic pool:</p> <ul style="list-style-type: none"> <li>Comprises of around 0.7 million documents related to around 0.5 million distinct patents that were published between 2001 and 2006, with around 500-10,000 topics extracted from granted patents</li> </ul>	[123], [8]
<b>CLEF-IP 2010:</b>	Tailored for the prior art candidates search (PAC) and classification (Cls) tasks, encompassing over 3.5 million patent documents	<p>Corpus pool:</p> <ul style="list-style-type: none"> <li>Include around 2.6 million documents related to 1.9 million distinct patents that were published before 2002</li> </ul> <p>Topic pool:</p> <ul style="list-style-type: none"> <li>Include around 0.8 million documents related to 0.6 million distinct patents, that were published between 2002 and 2009.</li> <li>The topic pool of the CLEF-IP 2010 Collection contains two</li> </ul>	[124], [8]

		sets of documents; one larger set with 2000 topics, and the other smaller set with 500 topics, for the prior art task	
<b>CLEF-IP 2011</b>	<ul style="list-style-type: none"> <li>Created for four tasks: prior art search, patent classification, image-based prior art search, and image classification</li> <li>Moreover, 290,880 images (occupying 5.4 GB) associated with the patents in three IPC subclasses were added.</li> <li>For the prior art tasks, CLEF-IP 2011 includes 3,973 topics obtained from patent application documents as a separate repository</li> </ul>	Similar to the CLEF-IP 2010 dataset, the corpus pool and topic pool remained the same	[125], [8]
<b>CLEF-IP 2012</b>	<ul style="list-style-type: none"> <li>Designed for three tasks: passage retrieval starting from claims, chemical structure recognition, and flowchart recognition.</li> <li>The image data added in CLEF-IP 2011 was not included this year.</li> <li>The Passage Retrieval Starting From Claims task in Clef-IP 2012 aims at finding both relevant passages and documents based on the claims taken from patents published after 2001, with a total of 156 training and test topics manually generated, and with relevance assessments derived from highly relevant citations as reported by technically-skilled examiners in the search reports of the specific patents</li> </ul>	<ul style="list-style-type: none"> <li>Leverage the corpus and topic pools of CLEF-IP 2010</li> <li>Includes patent documents from the World Intellectual Property Organization (WIPO) and the European Patent Office (EPO), totaling more than 1.5 million patents up to 2002.</li> </ul>	[126],[8]
<b>CLEF-IP 2013</b>	The dataset was collated as a testing dataset for two specific tasks; passage retrieval from claims and structure recognition from patent images.	<ul style="list-style-type: none"> <li>Similar to the CLEF-IP 2012 Collection, this collection utilized the CLEF-IP 2010 dataset's topic pool and corpus pool and maintained the same objective behind the passage retrieval task.</li> <li>For this task, a total of 148 topics were selected from 69 patent applications published after 2002</li> </ul>	[126],[8]

S3.2.2 Table S3: Patent Retrieval Tasks in NTCIR-3 to NTCIR-6

Table S3 provides a comprehensive summary of the patent retrieval tasks conducted as part of the NTCIR (NII Test Collection for Information Retrieval Research) workshops



from NTCIR-3 through NTCIR-6. Initiated in 1997 by the Japanese National Institute of Informatics, NTCIR workshops promote research in information retrieval and related fields, with particular emphasis on cross-language information retrieval (CLIR). The table delineates the main tasks, datasets used, and the specific details of the retrieval challenges and datasets for each iteration of the workshop. These include the types of documents involved, the scope and structure of the search topics, and the criteria for relevance judgments. The information in this table is crucial for understanding the evolution and complexity of patent retrieval tasks over these sessions, providing detailed insights into the methodologies and data used in these internationally recognized research efforts.

**Table S3.** Detailed Overview of Patent Retrieval Tasks from NTCIR-3 to NTCIR-6.

NTCIR PR Tasks	Main task	Dataset	References
NTCIR-3 PATENT	Technology survey problem	<b>Documents:</b> Japanese patent applications from 1998-1999 (about 17GB). JAPIO patent abstracts from 1995-1999. Patent Abstracts of Japan (PAJ) (English translations for JAPIO patent abstracts) from 1995-1999  <b>Task data:</b> Search Topics: 30 search topics with associated newspaper articles (translated from Japanese to Traditional Chinese, Simplified Chinese, Korean, and English), and their corresponding relevance judgements	[8],[127]
NTCIR-4 PATENT	Patent map generation Invalidity search	<b>Documents</b> Unexamined Japanese patent applications published between 1993-1997 along with English translations of the abstract Number of documents: 3,496,252  <b>Task data:</b> Search Topic: Each of the 34 search topics corresponds to a claim from a denied patent application that was declared void by already-existing prior art Relevance Judgements: Individual patents (or in combination with other patents) that can invalidate or void a topic claim Specific relevant passages that invalidate a claim are tagged with the relevance judgements	[8],[128]
NTCIR-5 PATENT	Document Retrieval (Invalidity Search) Passage Retrieval	<b>Documents:</b> Unexamined Japanese patent applications published between (1993-2002) along with English translations of the abstract Number of documents 3,496,252	[8],[129]

		<b>Task data:</b> <b>Document Retrieval Subtask</b> Search Topics: 1223 search topics, with 34 topics reused from NTCIR-4 Relevance Judgement: For 34 search topics from NTCIR-4, relevance judgement is also the same as in NTCIR-4 Citations from examiners in the Japanese Patent Office were used for the remaining 1,189 topics	
		<b>Passage Retrieval Subtask</b> Search Topics: 356 search topics Relevance Judgement: Relevant passages were determined based on specific criteria	
<b>NTCIR-6 PATENT</b>	Japanese retrieval (invalidity search) English retrieval.	<b>Documents:</b> Publication of unexamined patent applications from the Japanese Patent Office (JPO) in the years 1993-2002, along with English translations of the abstract Number of documents: 3,496,252 Patent grant data published from the United States Patent and Trademark Office (USPTO) between 1993-2002, Number of documents: 1,315,470 <b>Task data:</b> <b>Japanese Retrieval Subtask:</b> Search topics: 2,908 Search topics are claims in Japanese patent applications, and the aim is to retrieve patents that can invalidate the claims Relevance judgments: 4 levels  <b>English Retrieval Subtask:</b> Search topics: 3,221 Search topics are claims in USPTO patent grants, and the objective is to retrieve patents relevant to the claims Relevance judgments: 3 levels	[8],[130].

S3.2.3 Table S4: Overview of Chemical Patent Tasks in TREC-CHEM

Table S4 provides a detailed overview of the patent retrieval tasks in TREC-CHEM 2009, TREC-CHEM 2010, and TREC-CHEM 2011 [8], as part of the Text Retrieval Conference (TREC). TREC, initiated in 1992 and sponsored by the National Institute of Standards and Technology (NIST) and the United States Department of Defense, serves as a prominent forum for evaluating information retrieval methods. The TREC-CHEM track specifically addresses the challenges of chemical patent retrieval, aiming to advance research in handling chemical datasets.

This table outlines the tasks, topics, relevance judgments, and search corpus details for each year of the TREC-CHEM track. It includes specifics such as the number of topics,

the nature of the relevance judgments, and the composition of the search corpus, which comprises millions of chemical patents and scientific articles. These details are crucial for understanding the scope and depth of chemical patent retrieval challenges addressed during these sessions, providing valuable insights into the methodologies and data used in this specialized area of research.

**Table S4.** Chemical Patent retrieval tasks related [30] in TREC-CHEM (2009-2011).

	Tasks	Topics	Relevance Judgment	Search Corpus
<b>TREC-CHEM 2009</b> [131]	Technology survey Prior art search	188 topics for the technology survey task 1,000 patents for the prior art search.	Obtained from experts and chemistry graduate students for the technology survey Collected from citations of topic patents and their family members for the prior art search.	Approximately 1.2 million chemical patents were filed until 2007 at EPO, USPTO, and WIPO, along with 59,000 scientific articles.
<b>TREC-CHEM 2010</b> [132]	Same as TREC-CHEM 2009 (technology survey and prior art search)	30 topics for the technology survey task	Shaped in the same way as TREC-CHEM 2009	About 1.3 million chemical patents and 177,000 scientific articles
<b>TREC-CHEM 2011</b> [133]	Same as previous CHEM-TREC (Technology Survey and Prior art search) Chemical image recognition task	Biomedical and pharmaceutical patents	Same procedure as previous TREC-CHEM tracks	

### S3.2.4 Table S5: USPTO Research Datasets

Table S5 presents a detailed overview of the research datasets provided by the United States Patent and Trademark Office (USPTO) [140]. The USPTO, as a pivotal government office, grants patents and registers trademarks in the United States, offering essential tools for inventors, researchers, and corporations to access comprehensive patent-related data. This table categorizes the various datasets available, describing their contents, the type of information they include, and the time brackets they cover.

Each dataset listed in Table S5, such as the Patent Examination Research Dataset (PatEx) and the Patent Litigation Docket Reports Data, serves distinct purposes ranging from offering elaborate information on patent applications and assignments to providing in-depth details on trademark filings and litigation. These datasets are crucial for conducting thorough research in intellectual property law and technology, providing a rich source of data for academic and commercial research. The table not only helps in understanding the scope of each dataset but also illustrates the breadth of data available for diverse research needs.

**Table S5.** Description of USPTO Research Datasets and Their Applications.

Dataset	Explanation	Time Bracket
---------	-------------	--------------

<b>Patent Examination Research Dataset (PatEx) [140]</b>	Elaborate information on over 13 million publicly viewable patent applications filed with the USPTO, including more than 1 million Patent Cooperation Treaty (PCT) applications	Spans from the year 2014 to the latest release in 2022
<b>Patent Assignment Dataset [140]</b>	Extensive insights on patent assignments and other transactions recorded at the USPTO	since 1970
<b>Trademark Case Files Dataset [140]</b>	In-depth details on millions of trademark applications filed with or registrations issued by the USPTO	since 1870
<b>Trademark Assignment Dataset [140]</b>	Extensive information on trademark assignments and other transactions recorded at the USPTO	since 1952
<b>PatentsView [140]</b>	A highly flexible API, Search and download query builder Bulk download Visualization interface for exploring and analyzing	40 years of patent data
<b>Artificial Intelligence Patent Dataset [140]</b>	Comprises data files identifying artificial intelligence-related patents and pre-grant publications	1976 - 2020
<b>Patent Litigation Docket Reports Data [140]</b>	Thorough patent litigation data on 81,350 unique district court cases	1963-2016
<b>Static datasets [140]</b>	Office Action Research Dataset for Patents Patent Claims Research Dataset Cancer Moonshot Patent Data Historical Patent Data Files	

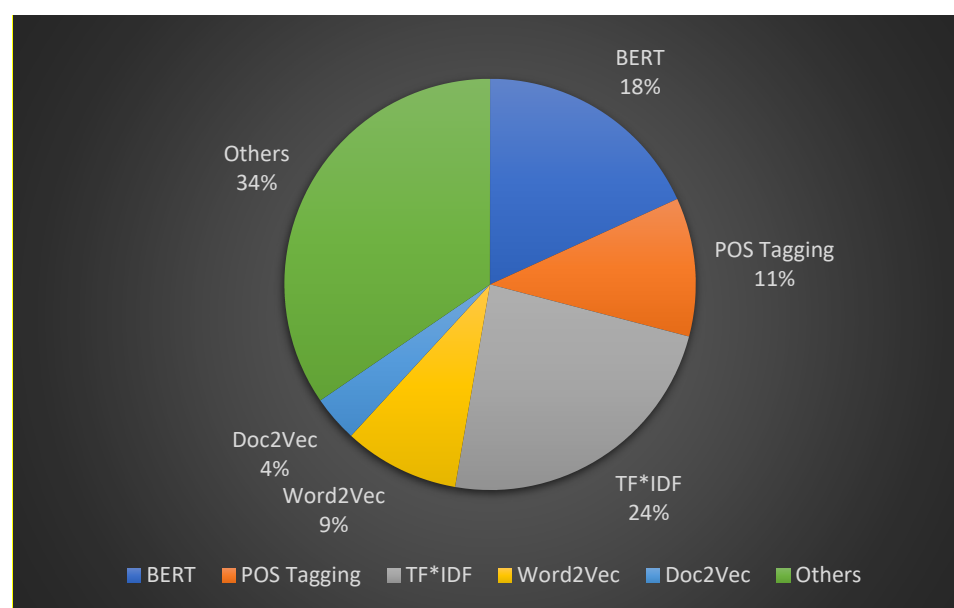
### S3.3 Supplementary Materials to RQ4

This section provides supplementary information supporting Research Question 4 (RQ4), which focuses on the use of Natural Language Processing (NLP) techniques in patent retrieval research. Detailed here are specific NLP models employed and their efficacy in processing and understanding patent data as demonstrated through various studies.

#### S3.3.1 Figure S2: Utilization of NLP Models in Surveyed Research

Figure S2 summarizes the overall NLP-based techniques reported to have been used in various surveyed research. The largest share, 24%, is attributed to the relatively simple NLP-based technique Term Frequency-Inverse Document Frequency (TF\*IDF). TF-IDF has been frequently used due to its efficacy in capturing the relevance and importance of terms, scalability suited for processing large document collections, interpretability allowing simple to grasp and interpretation of findings, and flexibility adaptable to the incorporation of additional features. It is widely utilized as it offers an effective framework for gauging the significance of terms within documents.

Part-of-speech (POS) tagging has been used in 11% of the studies, while Bidirectional Encoder Representations from Transformers (BERT), which is considered a highly advanced NLP technique, has been utilized in 18% of the studies. POS tagging, frequently employed to provide insights into the grammatical structure of text, enhances text comprehension and feature extraction. However, it does not grasp the context of the text. To fill this gap in contextual understanding, BERT was introduced a few years ago, with the ability to capture rich contextual information within the text. Its state-of-the-art performance makes it a popular choice among researchers.



**Figure S2.** Widely utilized NLP models.

### S3.3.2 Table S6: Natural Language Processing Techniques Used in Selected Studies.

Table S6 presents a detailed explanation of various Natural Language Processing (NLP) techniques as they are applied in the context of information retrieval and text analysis in the selected studies. Each entry in the table describes a specific NLP technique, offering insights into its application, underlying principles, and the typical tasks it is used for within the realm of patent data and broader textual analysis.

This table serves as an essential resource for readers seeking to understand the diverse range of NLP tools and methods employed in contemporary research. It covers a variety of techniques from foundational models like TF\*IDF and basic language models to advanced neural network approaches such as BERT and Gated Recurrent Units (GRUs). The explanations include details on the operational mechanisms of these techniques and their practical applications in handling complex language data, thereby enriching the reader's comprehension of the technical nuances involved in NLP-driven research.

**Table S6.** Brief explanation of NLP techniques utilized in the selected studies.

NLP Technique	Explanation
Kullback-Leibler Divergence	Kullback-Leibler (KL) divergence, in the context of Natural Language Processing (NLP), is a metric that measures the degree to which one probability distribution deviates from an expected distribution. It is used for tasks such as language modeling, text classification, and information retrieval [142].
LM (Dirichlet smoothing, and	Smoothing is a strategy used to address the issue of zero probabilities in language modelling for unseen words or n-grams by revising estimated probabilities and maintaining non-zero probabilities for unobserved scenarios [143].

<b>Jelinek-Mercer smoothing)</b>	<p>LM-Dirichlet smoothing: Dirichlet smoothing or Bayesian smoothing, adjusts word probabilities to smooth out the extreme probabilities by utilizing prior distribution (estimated from the entire corpus) based on observed frequencies [143].</p> <p>LM-Jelinek-Mercer smoothing: Jelinek-Mercer smoothing adjusts the probabilities of words in language models by utilizing both foreground distribution (estimated from the observed data) and background distributions (often derived from the entire corpus) by taking a linear combination of their probabilities [143].</p> <p>The Jelinek-Mercer smoothing method outperforms the Dirichlet method for long queries whereas Dirichlet smoothing outperforms the Jelinek-Mercer smoothing for short queries [144].</p>
<b>Positional Language Model</b>	A positional language model takes into account the position of words (positional information) within a document when computing their probabilities, to improve understanding of the text [145].
<b>Approximate nearest-neighbor techniques</b>	Approximate nearest-neighbor (ANN) techniques provide effective ways to find documents or data points that are most comparable (closest and similar) to a particular query or reference point. It includes techniques such as hashing, locality-sensitive hashing (LSH), or tree-based algorithms like k-d trees [146].
<b>BERT</b>	<p>Designed by Google and pre-trained (large text corpora, such as Wikipedia and Book-corpus) language model for the bidirectional contextual understanding of text [147]. To meet different requirements and use cases, BERT comes in a number of variations that differ in terms of size, computational efficiency, task-specific fine-tuning, or domain adaption:</p> <p><b>ColBERT:</b> Contextualized Language Model for Passage Re-ranking [148].</p> <p><b>DistilBERT:</b> A lighter version of BERT with similar performance [149].</p> <p><b>ParaBERT:</b> Enhanced BERT model for paraphrase identification [150].</p> <p><b>SBERT:</b> Sentence-BERT for semantic similarity between sentences [151].</p> <p><b>TinyBERT:</b> A smaller version of BERT designed for resource-constrained environments [152].</p> <p>PLI (Modelling Paragraph Level Interactions): Specialized language model that determine the semantic relationships at the paragraph-level and then infers the relevance between two cases by combining paragraph-level interactions [153].</p>
<b>BiLSTM</b>	BiLSTM (Bidirectional Long Short-Term Memory) is a sequence model in natural language processing that can process input data in both forward and backward directions to capture bidirectional context information and enhance sequence understanding [154].
<b>Bigram Language Model</b>	Bigram Language Model predicts the likelihood of a word given its preceding word. It ruminates pairs of adjacent words (bi-grams) in text data to estimate the likelihood of word sequences (finding a particular word following another) [155].
<b>Conditional Random Fields (CRFs)</b>	CRFs (Conditional Random Fields) is a discriminative probabilistic model, used to segment and label sequential data. They take into consideration contextual information obtained from surrounding labels in a sequence to discover dependencies and generate more precise predictions. This method is frequently utilized in natural language processing tasks such as named entity recognition, part-of-speech tagging, and sequence labelling [156].
<b>Doc2Vec</b>	Doc2Vec (Document-to-Vector), also referred to as Paragraph Vector is a Natural Language Processing technique to represent the entire documents or text segments

	(such as paragraphs, sentences, or whole articles) as fixed-length vectors. Doc2Vec has two primary variants: Distributed Bag of Words (DBOW), which focuses on word distribution within documents, and Distributed Memory (DM), which takes into account word order as well as document context [157], [158].
<b>Gated Recurrent Units (GRUs)</b>	Gated Recurrent Units (GRUs) are a type of recurrent neural network (RNN) architecture frequently used in Natural Language Processing (NLP). It is designed to solve a long-standing problem (also known as vanishing gradient problem) which limits RNNs ability to retain information from earlier time steps in long sequences. It uses a gating method to regulate the flow of information through the network [159].
<b>LDA</b>	Latent Dirichlet Allocation (LDA) is a commonly used topic modelling approach in Natural Language Processing (NLP) to find underlying topics within a set of documents or texts. It produces a set of topics, each characterized by a probability distribution over words, for a particular collection of documents [160].
<b>POS Tagging</b>	Part-of-speech (POS) tagging is a key task in natural language processing (NLP) to grasp the syntactic structure of sentences within the given text by assigning grammatical tags (such as nouns, verbs, adjectives, etc.) to words in a sentence [161].
<b>Semantic Trees</b>	Semantic trees (also known as syntax trees or parse trees) are visual aids that show how sentences are syntactically constructed. They highlight the grammatical relationships between words or phrases within sentences (like subject, object, verb, and others). In these trees (hieratical structures) nodes represent the words and edges represent the relationships between them [162].
<b>Skip-gram model</b>	The skip-gram model attempts to predict the surrounding words (context words) for a given target word (central word). It generates vector representations for words in a given text that reflect their syntactic and semantic relationships. However, Skip-gram captures limited contextual information because of a fixed-size window around each target word in larger contexts [163].
<b>TF*IDF</b>	TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic used in Natural Language Processing (NLP) to quantify the significance of a word in a document. It incorporates two metrics: <ul style="list-style-type: none"> <li>Term Frequency (TF): It indicates the frequency with which a term occurs in a document [164].</li> <li>Inverse Document Frequency (IDF): It indicates the frequency of the term across all documents in the corpus [164].</li> </ul>
<b>Tools</b>	<p>Stanford CoreNLP: A collection of Natural Language Processing (NLP) tools built by Stanford University, providing several NLP functionalities including part-of-speech tagging, named entity recognition, and sentiment analysis.</p> <p>OpenNLP: An open-source Natural Language Processing (NLP) library that offers tools for tasks like tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, language detection and co-reference resolution [165].</p> <p>TreeTagger: A software tool for annotating text using part-of-speech tagging, lemmatization, and named entity recognition in text analysis [166].</p>
<b>Topic Modelling</b>	Topic modelling is used to find generic topics or concepts in a group of documents by examining the density of words or terms. It looks for hidden semantic structures

	in text data to improve comprehension, organization, and summarization of large document collections [167].
<b>Unigram language model</b>	Unigram language model (or bag-of-words model) considers each document as a histogram of word occurrences and disregards word order and context, as a result, fails to capture semantic relationships within the document [168].
<b>Word2vec</b>	Word2vec (Word to vector) generates word embeddings (vector representations of a particular word) to capture semantic and syntactic relationships between words in a given text corpus [169].

### S3.4 Supplementary Materials to RQ5

This section supplements Research Question 5 (RQ5) by providing detailed insights into the structural elements of patent documents. The understanding of these components is crucial for researchers and practitioners in the field of patent retrieval and analysis, as it affects how information is accessed and interpreted.

#### S3.4.1 Table S7: Components of a Patent Document

Table S7 provides a comprehensive breakdown of the various sections of a patent document, outlining their specific functions and the type of information each contains. Each component of the patent document, from the title and abstract to the claims and drawings, plays a crucial role in conveying essential information about the patent. This table categorizes these components into structured and unstructured information, helping to clarify how each part contributes to the overall utility and understanding of the patent. It serves as an invaluable resource for researchers and practitioners in the fields of intellectual property and legal studies, offering insights into the standard structure and expected content of patents.

**Table S7.** Detailed Functions and Nature of Each Component in a Patent Document.

Component of a Patent Document		Functions	Nature
<b>Title</b>		Compact one-liner synopsis of the patented invention	Unstructured Information (Open text)
<b>Abstract</b>		Give a summary of the invention	Unstructured Information (Open text)
<b>Bibliographic data</b>	<b>Inventor(s)</b>	Individual(s) credited with inventing the technology	Structured Information (Follows a stringent format)
	<b>Patent Number</b>	Unique identifier assigned to the patent	
	<b>Assignee(s)</b>	Individual(s) that owns the patent rights	
	<b>Filing Date</b>	Date when the patent application was filed	
	<b>Publication Date</b>	Date when the patent application was published and made available for the public	
	<b>Issue Date</b>	Date when the patent was granted	
	<b>Patent Classification Code</b>	Classification codes (CPC) are the categories assigned based on the subject matter	
<b>Others</b>		Include applicant, representative, field of search etc.	
<b>Description</b>		Detailed explanation of the invention's background, functioning, and advantages	Unstructured Information (Open text)
<b>Claims</b>		Sets the boundaries of patent protection by identifying the distinctive features of the invention	Unstructured Information (Open text)



---

	Consists of one or more numbered claims, written in a specific format, each presenting a distinct feature of the invention
<b>Drawings</b>	If applicable, drawings are provided along with the description to illustrate the invention and its various components
<b>References</b>	Citations to prior art (related patents) included to provide context and establish the uniqueness of the invention

---