

Article

An Integrated Human-in-the-Loop Patent Search Workflow: An Exploratory Case Study Combining Keyword, Similarity, Semantic, and Citation-Based Retrieval

Federico Rotini, Francesco Saverio Frillici and Lorenzo Fiorineschi *

Department of Industrial Engineering, University of Florence, Via di Santa Marta 3, 50139 Florence, Italy; federico.rotini@unifi.it (F.R.)

* Correspondence: lorenzo.fiorineschi@unifi.it

Featured Application

The workflow is intended as a structured procedure for preliminary prior-art exploration and technology-intelligence activities. The present single case does not establish its general effectiveness or equivalence to expert-led searching.

Abstract

Patent retrieval is important for innovation management, R&D planning, and intellectual property strategy, yet conventional searches are hindered by terminology variability, synonymy, broad patent scopes, and classification limitations. This study presents an integrated, iterative workflow combining keyword-based retrieval, semantic analysis, document-similarity searches, citation-based refinement, and consensus-based manual relevance assessment. The workflow is illustrated through a single exploratory case starting from the same reference patent used by Moeller and Moehrle. After correction of duplicate entries, the final dataset contains 49 distinct patent identifiers: 19 were assigned to the high-relevance category, 20 to the related category, and 10 to the peripheral-relevance category. Six of the ten documents listed in the reference study were also retrieved; this 60% document-level overlap is descriptive and must not be interpreted as recall because the published list is not an exhaustive ground truth. The case illustrates how complementary retrieval channels, set intersections, iterative seed updating, and human decisions can be organized in a transparent workflow. It does not demonstrate superiority, general effectiveness, or equivalence between non-domain experts and expert searchers.

Keywords: patent retrieval; prior-art search; semantic search; similarity search; keyword search; citation analysis; human-in-the-loop; innovation management

Academic Editor: Dong Hyun Jeong

Received: 12 September 2026

Revised: 9 October 2026

Accepted: 14 October 2026

Published: 15 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The growing complexity of technological innovation and the increasing volume of patent data have made patent retrieval a critical activity for innovation management, R&D planning, and intellectual property strategy. Patent searching is no longer limited to legal validation activities, but has become an essential component of technology intelligence and competitive analysis. However, traditional keyword-based patent retrieval approaches often suffer from limitations related to terminology variability, synonymy,

broad patent scopes, and domain-specific language. These issues can reduce retrieval precision and lead to incomplete identification of relevant prior art.

To address these limitations, research has explored semantic analysis, natural language processing, machine learning, network analysis, and hybrid retrieval strategies that combine multiple information sources [1–4]. Recent AI- and deep-learning-based work has focused mainly on algorithmic retrieval, large-scale processing, and quantitative evaluation [5–8]. The present study addresses a different level of the problem: the operational coordination of existing retrieval functions and human decisions within a traceable search procedure.

This paper presents an integrated and iterative patent retrieval workflow combining keyword-based search, semantic analysis, similarity search, and citation-based refinement. It does not introduce a new retrieval algorithm. Instead, it formalizes a human-in-the-loop sequence intended for exploratory patent searching and illustrates its application when the searchers do not possess specialized expertise in the target technological domain.

The methodological contribution lies in the organization and operational connection of existing techniques rather than in a new computational retrieval paradigm. More specifically, the workflow:

- coordinates keyword, semantic, similarity, and citation-based retrieval within one traceable procedure;
- uses intersections among independently generated result sets as an explicit mechanism for reducing the candidate space;
- updates similarity-search seeds iteratively by selecting documents assigned to the C1 category after manual consensus assessment; and
- makes the boundaries between automated operations and human decisions explicit, including the final citation-based refinement and the application of the procedure by searchers without specialized knowledge of the case domain.

The workflow is illustrated through a descriptive comparison with the single case reported by Moeller and Moehrl [9]. The comparison is used to document procedural behavior and document-level overlap; it is not treated as a validation of effectiveness, a recall study, or evidence of superiority over the reference method.

The remainder of the paper is organized as follows. Section 2 presents the literature background. Section 3 describes the proposed workflow, tool selection, and illustrative case-study design. Section 4 reports the case results. Section 5 discusses the descriptive comparison, limitations, and potential use. Section 6 presents the conclusions.

2. Literature Background

Patent searching is a multifaceted activity that can pursue different objectives and involve different costs and levels of depth [4–6,10,11]. This paper focuses on prior-art search, namely a specific form of state-of-the-art search aimed at identifying publications that are relevant to novelty and inventive-step assessment [12,13]. In patent examination and innovation management contexts, the search result should ideally be a concise set of documents representing the closest prior art, namely publications available before the filing date and highly relevant to the invention under analysis.

Traditional patent retrieval methods mainly rely on keyword-based searches and patent classification systems such as the International Patent Classification (IPC) and Cooperative Patent Classification (CPC). Keyword searches involve identifying relevant terms or phrases related to the invention of interest and querying patent databases using these terms. While simple and widely adopted, keyword-based retrieval can generate false positives and may miss relevant patents because of synonymy, terminology ambiguity, and

domain-specific language [7,14,15]. Classification searches provide a structured way to narrow the scope of retrieval, but they may overlook patents that straddle multiple domains or are classified under different categories. Classification systems may also struggle to keep pace with rapidly evolving technology landscapes, leading to incomplete or inconsistent coverage.

To overcome these limitations, semantic and NLP-based approaches have progressively emerged in patent analysis research. Natural language processing techniques allow extraction of contextual and conceptual information from patent descriptions, claims, and abstracts, enabling more advanced retrieval capabilities compared with purely lexical approaches [1,2,4]. Other studies explored ontology-based methods [3,16], SAO subject-action-object structures [17,18], thesaurus-based retrieval [19], machine-learning approaches for patent classification and trend identification [11,20], and network analysis for citation-based and technological trajectory analysis.

More recently, artificial intelligence and deep learning approaches have been applied to patent retrieval and prior-art analysis. These systems can identify latent semantic relationships and improve scalability in large patent databases [5,6,8]. However, fully automated AI-driven systems often present limitations related to interpretability, transparency, dependence on training datasets, and reproducibility across platforms. In practical industrial and innovation management contexts, explainability and user control remain important requirements.

Increasingly, researchers have explored hybrid approaches that combine the strengths of traditional retrieval methods with semantic, classification-based, or machine-learning techniques. By integrating keyword searches with semantic analysis or classification systems with computational tools, hybrid strategies can exploit the complementary nature of different retrieval channels. In parallel, integrated patent search platforms have provided researchers and practitioners with advanced analytics, similarity search, semantic expansion, visualization, and customizable search options. Appendix A reports an illustrative and non-exhaustive list of patent search platforms and tools.

An important contribution in this area was proposed by Moeller and Moehrle [9], who introduced a semiautomatic iterative patent-search method combining keyword searching and semantic similarity analysis. Their work established a useful operational reference for iterative near-search activities. It was selected here because it provides a published case, a common starting patent, and a documented result list against which a descriptive comparison can be made.

The present workflow builds on that foundation but formalizes a broader sequence of operations. Compared with the keyword and semantic-search combination described by Moeller and Moehrle [9], it adds an explicit parallel similarity-search channel, intersections among independently generated sets, iterative replacement of the similarity-search seed using C1 documents, and a final citation-based retrieval stage. Manual assessment is used as an operational decision gate between iterations. These are workflow-level refinements; no claim is made that the underlying proprietary retrieval algorithms are new or that the resulting procedure outperforms other hybrid approaches.

3. Materials and Methods

3.1. Proposed Methodology

The proposed methodology is a systematic and iterative process that combines keyword-based searches, semantic analysis, similarity-based techniques, and citation searching to organize the identification and assessment of potentially relevant patent documents. Each component addresses a different part of the retrieval process. Figure 1 and Table 1 show the sequence and provide a short description of each step.

In particular, the proposed workflow combines four complementary retrieval mechanisms:

- keyword-based retrieval,
- semantic retrieval,
- similarity-based retrieval,
- citation-based refinement.

The workflow is structured as an iterative human-in-the-loop process in which independently generated patent subsets are progressively intersected and manually assessed. The intersections are intended to reduce the number of candidates passed to subsequent stages; their use does not by itself guarantee higher retrieval performance.

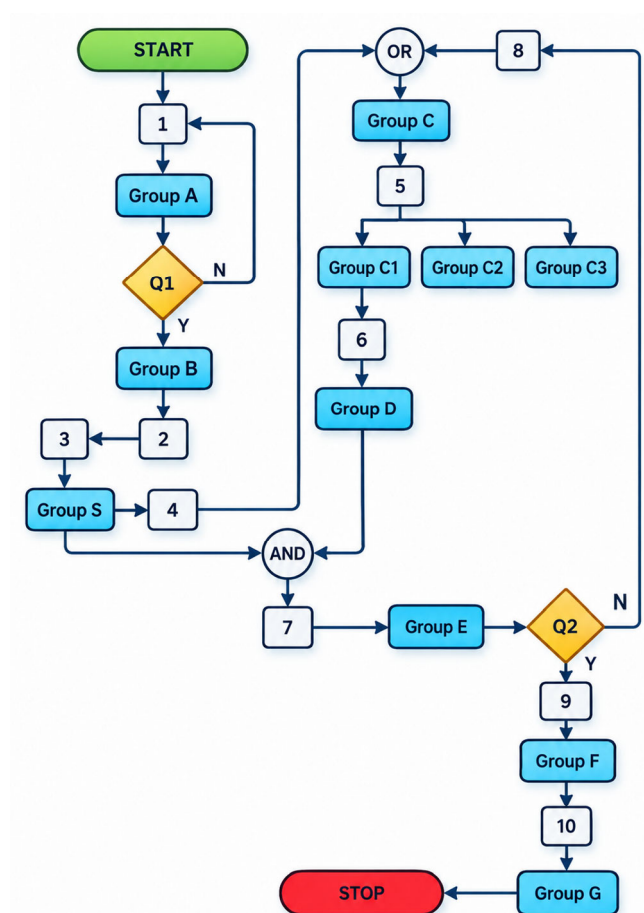


Figure 1. Schematic representation of the proposed procedure (see Table 1 for the step descriptions).

The process begins with an initial keyword search (Step 1 in Figure 1). Starting from a natural-language description of the idea, the operator identifies the functional problem, the elements considered distinctive, and terms judged representative of those elements. Boolean combinations, field restrictions, truncation, and patent classifications can then be used to broaden or narrow the query. Group A is reviewed to identify a potentially pertinent document, which becomes the seed for the similarity search in Step 2 and generates Group B. The choice of concepts and terms remains dependent on human judgement and on the terminology known to the operator.

Synonyms, multilingual variants, and domain-specific expressions are not normalized through a fixed automatic procedure in this workflow. They are handled through manual reformulation, truncation, classification codes, and the complementary semantic

channel. AI-assisted term extraction or multilingual query expansion could be incorporated, but such suggestions would require human verification because model-generated terms may drift from the intended inventive concept.

A semantic search is also performed from the natural-language description of the invention (Step 3). The platform processes this description and returns documents ranked according to its proprietary semantic model. This channel is intended to complement lexical retrieval by identifying documents that may use different wording, but the ranking logic and the treatment of semantic relationships are not fully accessible to the user.

Table 1. Description of the steps in the process shown in Figure 1.

Step	Description
1	Perform a first search by relying on subjective keywords. The obtained set of documents constitutes Group A.
Q1	Has at least one pertinent paper been found?
2	Perform a similarity search by referring to the most pertinent paper
3	Perform a semantic search by relying on the description of the invention provided by the inventor or the inventors. The obtained set of documents constitutes Group S.
4	Intersection of the results obtained at Steps 2 and 3. The obtained set of documents constitutes Group C.
5	Perform an in-depth analysis of each document in Group C, then subdivide the group into three subgroups based on the pertinence level.
6	Take the most pertinent document from Group C1 and perform a similarity search. The obtained set of documents constitutes Group D.
7	Intersection of Group S and Group D. The obtained set of documents constitutes Group E.
Q2	Is Group E empty?
8	Extend Group C by adding documents contained in Group E.
9	Search for citing and cited documents by referring to the most pertinent patents. The obtained set of documents constitutes Group F.
10	Analysis of the most pertinent documents contained in Group F, in order to find the final set of pertinent documents (Group G)

The output of the similarity search (Group B) is intersected with the semantic-search output (Group S) to form Group C (Step 4). Group C is therefore a smaller candidate set identified through both channels and passed to manual relevance assessment.

The process is then repeated using selected C1 documents as new similarity-search seeds. This iterative design exploits retrieval channels that operate from different inputs and allows the candidate set to be updated without treating the output of any single proprietary ranking algorithm as conclusive.

Manual assessment is performed at Step 5 using three ordinal relevance categories rather than a quantitative similarity score. Lower numerical category codes denote greater relevance: C1 identifies documents with high relevance, C2 identifies related documents, and C3 identifies documents with only peripheral relevance. The category determines whether a document can be used as a new search seed.

C1 documents address the same or a very close functional problem and present a substantially comparable solution or system architecture; they may be selected as seeds for the next iteration. C2 documents address a related problem or share core functional characteristics but are not used as seeds. C3 documents share only terminology, context, or peripheral technological elements and are retained for traceability without driving further iterations.

In Step 6, another similarity search is performed by considering the most pertinent document available in Group C1. The obtained set of documents constitutes Group D.

An intersection is performed between Group D and Group S, excluding those documents already available in Group C (Step 7). The obtained set of documents constitutes Group E. In Question Q2, the process verifies whether Group E is empty or not. If Group E is not empty, the documents are removed from the group and moved to Group C, which is then expanded (Step 8). Steps from 5 to 7 are repeated.

Once Group E is empty, documents cited in citing documents available in the last version of Group C1 are retrieved (Step 9 in Figure 1), in order to obtain Group F. Indeed, this step has been added since it is acknowledged that searching among citing patents potentially allows rapid improvement of the set of pertinent documents [21].

The most pertinent documents available in Group F are then identified (Step 10) and the final set of most pertinent documents is retrieved (Group G).

Each component of the proposed workflow was introduced to address a specific limitation commonly associated with patent retrieval activities.

- Keyword-based search was introduced to provide an initial high-precision retrieval channel relying on explicit technological terminology. Semantic retrieval was included to mitigate terminology variability, synonym ambiguity, and contextual differences frequently encountered in patent documentation.
- Similarity-based retrieval was adopted to exploit document-level technological relatedness starting from highly pertinent seed patents. This operation was intended to identify patents exhibiting structural or conceptual proximity even when explicit keyword overlap is limited.
- Intersection operations among independently generated patent subsets were introduced as a noise-reduction mechanism intended to retain documents simultaneously identified through multiple retrieval channels. The iterative refinement structure was designed to progressively converge toward increasingly relevant patents by recursively exploiting the most pertinent retrieved documents.
- Finally, citation analysis was included to recover technologically connected patents potentially missed through textual retrieval alone, exploiting the relational structure inherently present within patent citation networks.

Accordingly, the proposed methodology should not be interpreted as a fully automated retrieval algorithm, but rather as a structured human-in-the-loop workflow integrating complementary retrieval mechanisms.

3.2. Tool Selection

Among the tools listed in Appendix A, several offer features comparable to those used in the proposed methodology. Google Patents and PatSnap provide similarity-based document retrieval, while PatSnap and The Lens support semantic search via NLP-enhanced indexing. However, these tools differ in transparency and algorithmic control compared to Orbit®, which integrates semantic expansion, classification, and similarity in an iterative framework. Espacenet, although widely used, relies primarily on traditional keyword and classification search. As such, the choice of Orbit in this study is motivated by its unified support for semantic and similarity-based iterations, which are central to the proposed method.

Similarity and semantic-search functions are increasingly available across patent platforms, but their underlying ranking models are generally proprietary. Consequently, results may vary with the platform, database coverage, index updates, and software release. Manual assessment provides an explicit decision layer after automated retrieval, but it does not make the outputs platform-independent or guarantee reproducibility of the same result set.

The workflow separates automated retrieval and set operations from human decisions. The platform executes queries, semantic expansion, similarity ranking, set intersections, and citation retrieval. Human operators formulate the initial description and queries, choose seed documents, assign C1–C3 categories by consensus, and select documents after citation retrieval. Table A2 summarizes this allocation together with the output sizes recorded in the case study.

3.3. Illustrative Case Study

A single published case was used to illustrate the workflow and permit a descriptive comparison with an existing patent-search exercise [9]. The purpose was to document how the procedure operates from a common starting patent, not to validate its effectiveness or reliability.

Moeller and Moehrle [9] examined a semiautomatic iterative method and a keyword-search-emulating method. Their case provides a documented starting patent and a list of retrieved documents, making it suitable for observing convergences and divergences at the document level. The published list is not an exhaustive ground truth of all relevant patents.

The case study reported here therefore had two descriptive objectives:

- to determine which documents from the published list were also retrieved by the proposed workflow; and
- to document the additional candidates and their consensus-assigned relevance categories within the present workflow.

Both exercises started from patent US 7,996,252. The workflow can in principle also start from a natural-language description, but this possibility was not separately evaluated in the present case.

The search activities in the present case were conducted without specialized expertise in the technological field of the reference patent, whereas the original study involved familiarity with the subject matter. This difference is reported as a characteristic of the case only. It does not constitute an expert-versus-non-expert comparison and cannot establish equivalent performance across user groups. Both operators had previous experience with patent databases and patent-search procedures, although neither had specialist knowledge of the target technology.

For the keyword channel, the operators selected terms judged to represent the functional content of the invention. No systematic multilingual synonym normalization or AI-assisted query expansion was used. The semantic channel was generated by the search platform from the patent description and was combined with, rather than substituted for, the manually formulated queries.

The retrieved identifiers were compared descriptively with those listed by Moeller and Moehrle [9]. The analysis records overlap and divergence but does not interpret the overlap as recall or comparative retrieval performance.

Retrieved documents were assigned to the C1, C2, or C3 categories through joint assessment by two examiners. The classifications were produced directly through discussion and consensus; independent pre-consensus ratings were not collected.

Accordingly, the case provides an exploratory illustration of the workflow. It cannot establish general effectiveness, superiority over the reference method, inter-rater reliability, scalability, or applicability across technological domains.

4. Results

In this study, following Moeller and Moehrle [9], the patent US 7,996,252 has been used as a reference document to start the process. More specifically, it has been considered

as the primary foundation for the semantic search process in Step 3 (see Table 1 and Figure 1). This patent describes a global customer satisfaction system and was utilized to extract the following textual description of the invention, which guided the semantic search through the Orbit® software:

“A global customer satisfaction system includes point-of-sale customer feedback terminals for receiving in-store customer feedback, including satisfaction ratings and other customer feedback regarding participating vendors as the customer pays for purchases. The system also includes post-sale customer feedback terminals for receiving more detailed customer feedback regarding the vendors from members after the sales have been completed. The system also includes an aggregation and dissemination system for combining the ratings received from customers into global customer satisfaction ratings for participating vendors, and disseminating the customer satisfaction information through customer satisfaction-qualified directories, which are organized or searchable by geographical location, type of product or service offered by the participating vendors, and other factors.”

The description was submitted to the platform’s semantic-search function, which returned 2912 identifiers forming Group S. This was an automatically generated candidate set and was not read in full by the examiners; subsequent set operations reduced the number of documents passed to manual assessment.

In Step 1, the initial search was conducted using specific keywords directly related to the technology described in the reference patent. The query used was (see Appendix B for query codes and truncations):

(RATING+ AND FEEDBACK AND CUSTOMER SATISFACTION AND CUSTOMER FEEDBACK AND TERMINAL AND VENDOR)/TI/AB/CLSM/DESC/ODES/ICLM)

The query returned 75 identifiers, forming the first version of Group A. Additional keyword and classification searches were then used to modify the candidate set. The values reported below describe query outputs and set operations, not the number of documents read in full by the examiners.

The first additional search used the query:

((CUSTOMER FEEDBACK AND CUSTOMER SATISFACTION AND (IN STORE OR POINT OF SALE))/TI/AB/CLMS/DESC/ODES/ICLM/AND (G06Q-010/00)/IPC)

This returned 39 patents, which were then excluded from the initial set obtained in Step 1. Subsequently, another search was conducted with the query:

((CUSTOMER FEEDBACK AND CUSTOMER SATISFACTION AND RATING+)/TI/AB/CLMS/DESC/ODES/ICLM AND (G06Q-010/00)/IPC)

This returned 112 patents. These were then integrated with the existing set, but additional screening was applied to ensure no overlap with previously identified patents. Further, another query was executed to capture a broader range of patents related to customer feedback systems, with the following formulation:

((CUSTOMER FEEDBACK AND CUSTOMER SATISFACTION AND RATING+)/TI/AB/CLMS/DESC/ODES/ICLM AND (G06Q)/IPC)

This search returned 444 identifiers. After removing identifiers already present in the previously generated sets, 332 remained. These operations removed exact duplicate identifiers at this stage; patent-family members were not consolidated as single inventions.

To further narrow the focus, a final step of refinement was applied. The group of 332 patents was filtered by removing the documents returned from the earlier query that yielded 39 patents. Following this, another layer of screening eliminated the patents from the initial search (Step 1). This final refinement resulted in a set of 288 patents.

A pertinent document was selected from the keyword-derived candidates and used to initiate the similarity-search channel that forms Group B.

Step 2 involved a similarity-based search, which was conducted by selecting the most pertinent patent extracted from Group B, i.e., US 20130346155, for use as a reference document. This patent was considered the most similar to the reference patent, and a similarity search was performed using the “SIMILARITY SS 13 RANKED 1” query. This search generated a set of 1000 patents (limited by the search tool), which, while reflecting the technological proximity to the reference, were subject to further analysis and refinement.

The Step 2 and Step 3 outputs were intersected in Step 4. This operation produced 136 identifiers in Group C, a smaller candidate set that was then passed to manual assessment.

The documents were assigned to three ordinal relevance categories:

- C1 (coded 1)—high relevance and eligible for use as a new search seed;
- C2 (coded 2)—related and retained for documentation, but not used as a seed; and
- C3 (coded 3)—peripheral relevance and not used in subsequent iterations.

Two examiners assigned these categories through joint discussion and consensus. Because the ratings were not first produced independently, the study does not provide pre-consensus observations from which Cohen’s kappa, weighted kappa, or another inter-rater agreement statistic could validly be calculated.

C1 was assigned when a document addressed the same or a very close functional problem and presented a substantially comparable solution or system architecture. C2 was assigned when the problem or one or more core functions were related but the solution differed substantially. C3 was assigned when overlap was limited to terminology, application context, or peripheral technological elements. These are qualitative decision criteria, not similarity thresholds generated by the platform.

The three categories were retained because each corresponds to a distinct operational decision in the workflow. Adding a fourth category would imply a finer level of discrimination not supported by the consensus procedure. Percentile thresholds were not adopted because relevance is assessed against the inventive content rather than against the relative distribution of a result set; a fixed top percentage could classify documents as highly relevant even when the entire set is weakly related.

In Step 6, US 10380611, assigned to C1, was selected as the seed for a further similarity search. The platform returned a maximum of 1000 identifiers. Intersection with Group S in Step 7 produced 247 additional candidate identifiers in Group E.

The Group E candidates were added to the assessment process, and the sequence of category assignment, seed selection, similarity retrieval, and intersection was repeated until Group E was empty. The number of unique documents manually examined in each successive round was not recorded separately.

Step 9 retrieved citing and cited documents connected to selected C1 documents. This stage was intended to add a relational retrieval channel capable of identifying documents not returned through the textual and similarity-based channels.

In Step 10, the citation-derived candidates were manually examined and the final document-level set was assembled. The analysis was conducted at identifier level; members of the same patent family were not consolidated into a single record.

A consistency check of Appendix C identified three duplicated identifiers in the previously reported 52 rows. After retaining one occurrence of each identifier, the final dataset contains 49 distinct identifiers: 19 assigned to C1, 20 to C2, and 10 to C3 (Figure 2 and Appendix C). US20130346155 was retained in C1 because it is the document selected in Step 2 as the most pertinent similarity-search seed. Among the ten documents listed by Moeller and Moehrle [9], six were also retrieved: one was assigned to C1, three to C2, and two to C3 (Table 2).

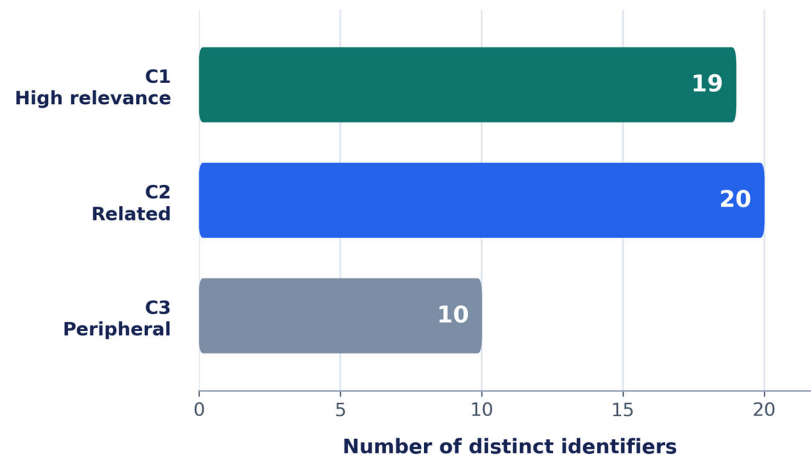


Figure 2. Distribution of the 49 distinct identifiers among the three consensus-assigned relevance categories.

Table 2. Descriptive document-level comparison with the ten patents reported by Moeller and Moehrle [9].

US Patents Identified in the Literature	Identified by the	Consensus-Assigned
Case Study	Workflow	Category
6,236,975	yes	3
6,960,988	yes	2
6,970,831	yes	3
7,376,570	yes	2
7,383,200	yes	2
7,698,162	no	-
7,778,862	no	-
6,108,665	no	-
7,818,203#	yes	1
5,995,941	no	-

5. Discussion

5.1. About the Obtained Results

The case shows that the workflow can be executed from the same starting patent used by Moeller and Moehrle [9] and can recover some of the documents in their published list. It does not show that the workflow achieves equivalent objectives or performance. The six-document overlap must be interpreted descriptively because the published list is not an exhaustive ground truth and the present classifications were produced within the same study.

The workflow also produced documents not contained in the published list, several of which were assigned to C1 through the consensus procedure. These internally assigned categories identify candidates considered closely related to the reference invention; without an independent ground truth or external expert assessment, they do not demonstrate that the workflow outperforms the reference method.

The operational rationale for combining retrieval channels is that a document must survive more than one form of filtering before entering selected parts of the manual assessment. Semantic search supplies a broad concept-based set, while keyword and similarity searches provide independently generated candidates for intersection. Citation retrieval adds a relational channel. In this case, these operations reduced large platform

outputs to smaller candidate sets, but no causal claim is made that the same configuration will improve retrieval quality in other domains.

The workflow should therefore be regarded as a transparent procedure for coordinating available retrieval functions and human judgement, not as a replacement for expert prior-art searching or as a competing AI retrieval algorithm. Its potential value lies in the explicit sequence of operations, decision points, and seed updates. Recent studies in computational chemistry likewise emphasize the value of combining machine learning with domain knowledge and human oversight, and of producing human-interpretable outputs [22,23]; however, these studies concern chemistry rather than patent retrieval and are cited here only as conceptual support.

Recent AI, deep-learning, and hybrid retrieval studies address semantic representation, scalability, and benchmark-based evaluation [5–8]. The present contribution is positioned at the workflow level rather than the algorithmic level. Because it uses one case and no exhaustive relevance set, it cannot be compared quantitatively with systems evaluated on larger benchmark datasets.

5.2. Descriptive Case Indicators

The case is described using dataset size, consensus-assigned categories, and overlap with the published list. These quantities characterize the observed output but are not information-retrieval performance metrics.

Six of the ten documents listed by Moeller and Moehrle [9] were also identified, corresponding to a 60% document-level overlap with that list. This value is not recall because the ten documents do not constitute an exhaustive set of all relevant patents. It also does not establish superiority or equivalence between the two procedures.

After exact duplicate identifiers were removed, the final set comprised 49 distinct identifiers: 19 C1 (38.8%), 20 C2 (40.8%), and 10 C3 (20.4%). Thus, 39 of 49 identifiers (79.6%) were assigned to C1 or C2 through the joint consensus assessment.

Additional C1 documents not present in the published list were observed, but their number is not used as evidence of comparative performance. The classification was internal to this study, patent families were not consolidated, and no independent ground truth or expert adjudication was available.

5.3. Limitations of the Work and Future Developments

The available records contain query-output and intersection sizes but not elapsed times or a complete count of unique documents manually reviewed during every iteration. Table A2 therefore reports only the recorded stage outputs and the human or automated responsibility at each stage. Retrospective time estimates or a workload comparison with the reference method would be unreliable and are not provided.

The relevance categories were assigned jointly by two examiners through discussion and consensus. Independent pre-consensus ratings were not collected, so inter-rater agreement cannot be quantified retrospectively. The criteria have been stated explicitly to make the decision process more transparent, but this does not remove subjectivity or establish classification reliability.

The initial description and keyword selection remain influential. The study did not apply systematic multilingual synonym mapping, terminology normalization, or AI-assisted query expansion. AI and advanced NLP tools could support term extraction, multilingual normalization, and synonym suggestions, but they may also introduce semantic drift, opaque expansions, or non-pertinent terms; any such output would require human review.

Orbit® was selected because it provided semantic retrieval, similarity ranking, and set operations in one environment. The exact search date and software/database release

were not preserved in the study records. Moreover, proprietary ranking algorithms and internal thresholds were not available to the operators. The queries, natural-language input, user-visible command, result limits, and recorded set sizes are reported, but repeating the workflow at another time or on another platform may not reproduce the same result set.

Future studies may compare the workflow with subject–action–object structures [2,17,18,24], ontology-based methods [3,16], thesaurus-based methods [19], AI-assisted retrieval [6,8,10], and deep-learning approaches [5]. Such studies would require multi-domain cases, independent ratings, expert assessment, timing records, and formal benchmark datasets.

The present results are therefore an exploratory, single-case illustration. They do not constitute validation of effectiveness, robustness, scalability, or comparative retrieval performance.

5.4. Expected Impact

The potential practical value of the workflow lies in making a complex search process explicit and traceable. The sequence may help researchers and practitioners organize preliminary patent exploration, document how candidate sets were generated, and identify where human judgement affects the outcome.

The workflow is intended for preliminary prior-art orientation and technology-intelligence activities. It is not a substitute for expert patent examination, a professional freedom-to-operate search, or a legal opinion. The present case does not establish that non-domain experts perform comparably to domain experts.

6. Conclusions

This study formalized an integrated human-in-the-loop patent-search workflow combining keyword retrieval, semantic search, similarity search, set intersections, iterative seed updating, and citation-based refinement. Its contribution is the operational design and connection of these elements, not a new retrieval algorithm.

The workflow makes explicit which operations are automated and which depend on human decisions. C1 documents act as new similarity-search seeds, while C2 and C3 documents are retained without driving subsequent iterations. This three-category structure links relevance assessment directly to an operational decision.

In the illustrative case, 49 distinct identifiers remained after correction of duplicate entries. Six of the ten documents in the published reference list were also retrieved. This overlap and the consensus-assigned categories describe the case outcome but do not constitute recall, proof of superiority, or independent confirmation of relevance.

The case also shows how intersections can reduce large automated outputs before selected sets are manually assessed. However, the contribution should be interpreted as a transparent search architecture whose performance remains to be established, rather than as evidence that the workflow improves retrieval quality.

The study is limited by its single case, the absence of an exhaustive ground truth, consensus rather than independent ratings, lack of timing and complete manual-workload records, identifier-level rather than family-level counting, and dependence on proprietary platform functions. It also does not compare expert and non-expert users.

Evaluation of effectiveness would require future multi-domain studies with prospectively collected independent ratings, domain-expert adjudication, timing data, family-level normalization, and formal retrieval metrics. These activities were outside the scope of the present exploratory case.

Author Contributions: Conceptualization, F.R.; methodology, F.S.F.; validation, F.R., F.S.F. and L.F.; formal analysis, F.S.F.; investigation, F.S.F. and F.R.; data curation, L.F.; visualization, L.F.; writing, L.F.; original draft preparation, L.F. and F.R.; and writing—review and editing, L.F. and F.R. The authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data supporting the reported results are contained within the article and its appendices.

Acknowledgments: During the preparation of this manuscript, the author used ChatGPT (OpenAI) for language editing, formatting, and assistance in revising the manuscript. The author reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A

Patent Search Platforms and Tools

Table A1. Illustrative and non-exhaustive list of patent search platforms and tools.

Platform Name	Description	Website
Google patents	A user-friendly platform providing access to a vast collection of global patent documents, offering search options, advanced filters, and citation analysis.	https://patents.google.com/
Espacenet	Operated by the European Patent Office (EPO), Espacenet offers access to over 120 million patent documents worldwide, with advanced search capabilities and patent family information.	https://world-wide.espacenet.com/
WIPO PatentScope	Maintained by the World Intellectual Property Organization (WIPO), PatentScope provides access to international patent applications, allowing users to search, browse, and analyze patent documents.	https://www.wipo.int/patentscope/en/
FreePatentsOnline	Offers a comprehensive database of US patents and patent applications, featuring advanced search options, document downloads, and alert services.	https://www.freepatentsonline.com/
Patent Lens	A free patent search tool providing access to worldwide patent documents, including full-text search and analysis capabilities.	https://www.lens.org/
PatentInspiration	Offers access to global patent data with advanced search functionalities, visualization tools, and analytics for patent landscape analysis.	https://www.patentinspiration.com/
InnovationQ	A comprehensive patent search and analytics platform providing access to global patent databases, with features for competitive intelligence and technology scouting.	https://www.innovationq.com/
PatSeer	A patent research platform with advanced search capabilities, analytics, and visualization tools for analyzing patent landscapes and identifying technology trends.	https://www.patseer.com/
Questel Orbit	Offers access to global patent databases, including advanced search functionalities, analytics, and IP management tools for patent professionals.	https://www.orbit.com/
PatBase	A patent search platform providing access to worldwide patent data, with advanced search options, analytics, and visualization tools for patent analysis and monitoring	https://www.patbase.com/

PatSnap	A comprehensive patent search and analytics platform offering access to global patent databases, with features for competitive intelligence, IP strategy, and innovation management.	https://www.patsnap.com/
PatentInspire	Provides access to global patent data, with advanced search capabilities, analytics, and visualization tools for technology scouting and patent portfolio analysis	https://patentinspire.com/
PatentHunter	A patent search tool with advanced search functionalities and analytics for analyzing patent landscapes and monitoring patent activity in specific technology areas.	https://neustelsoftware.com/patenthunter/ (accessed on 11 September 2026)
Patent Buddy	Offers access to US patent data with advanced search options, patent analytics, and portfolio management tools for patent professionals and attorneys.	https://www.patentbuddy.com/
Patent Search by USPTO	Provided by the United States Patent and Trademark Office (USPTO), this tool offers access to US patent and trademark databases, with search capabilities and resources for patent examination.	https://www.uspto.gov/patents/search (accessed on 11 September 2026)
PatentFiler	A patent search and analytics platform offering access to global patent databases, with features for patent monitoring, competitive analysis, and IP strategy development.	https://www.patentfiler.com/
LexisNexis PatentAdvisor	Provides access to US patent data with advanced search functionalities, analytics, and visualization tools for patent prosecution and portfolio management.	https://go.patentadvisor.com/
TotalPatent One	A patent research platform offering access to global patent databases, with search options, analytics, and visualization tools for patent analysis and competitive intelligence.	http://www.totalpatentone.com
PatentMiner	A patent search and analytics tool offering access to US patent data, with advanced search capabilities, analytics, and visualization tools for patent landscape analysis and technology scouting.	https://www.patentminer.com/
Patentcloud	Offers patent search and analysis tools, including global patent databases, AI-powered search capabilities, and IP analytics for patent professionals and organizations.	https://www.patentcloud.com/
PatentSight:	A patent analytics platform offering access to global patent databases, with AI-driven analytics, technology scoring, and competitive intelligence tools for patent strategy and decision-making.	https://www.patentsight.com/
Derwent Innovation	Provides access to global patent and scientific literature databases, with search options, analytics, and visualization tools for patent research, IP management, and competitive analysis.	https://www.derwentinnovation.com
Patent iNSIGHT Pro	A patent analytics platform offering access to global patent databases, with AI-driven analytics, technology scoring, and competitive intelligence tools for patent strategy and decision-making.	https://www.patentinsightpro.com/
InQuartik	Provides patent search and analytics solutions, including global patent databases, AI-powered search capabilities, and IP analytics tools for patent professionals and organizations.	https://www.inquartik.com/
PatentGate	A patent search and analysis platform offering access to global patent databases, with advanced search functionalities, analytics, and visualization tools for patent research and competitive intelligence	https://www.patentgate.com/
PatentServer	Provides access to global patent databases, with search options, analytics, and visualization tools for patent research and technology intelligence.	https://www.patentserver.io/ (accessed on 11 September 2026)

Appendix B

Orbit® Query Structure and Syntax

Orbit® was used to execute the keyword, semantic, similarity, set-operation, and citation-search stages. The queries and user-visible syntax available in the study records are reported below. Table A2 distinguishes platform operations from human decisions and lists the output sizes recorded during the case.

Table A2. Human–machine task allocation and recorded output sizes in the illustrative case study.

Step	Automated Operation	Human Decision	Recorded Output Size
1	Execute keyword and classification queries	Define concepts, terms, fields, and query revisions	75 initially; refinement outputs 39, 112, and 444
2	Rank documents by proprietary similarity function	Select the seed document	1000-result platform limit
3	Process the natural-language description semantically	Provide and verify the invention description	2912
4	Intersect Groups B and S	Review the resulting candidate set	136
5	None	Jointly assign C1, C2, or C3	136 in the first C set
6–8	Repeat similarity search and set intersection	Select C1 seed and assess new candidates	1000-result limit; 247 in the recorded Group E
9–10	Retrieve citing and cited identifiers	Select and classify citation-derived candidates	49 distinct identifiers in final table

Table A2 reports query or set-operation outputs, not a complete count of documents read in full. Elapsed times and unique manual-screening counts for every iteration were not recorded. The exact search date and Orbit® software/database release were also not preserved; therefore, exact result-level replication cannot be guaranteed.

Boolean Operators:

- **AND:** This operator was used to combine multiple search terms and ensure that only documents containing all the specified terms were retrieved. For example, (RATING+ AND FEEDBACK) would retrieve patents containing both “RATING” (and its variations) and “FEEDBACK”.
- **OR:** This operator is used to expand the search by including patents that contain at least one of the specified terms. For example, (IN STORE OR POINT OF SALE) retrieves patents containing either “IN STORE” or “POINT OF SALE”.
- **NOT:** This operator was used to exclude certain terms from the search. It helps narrow down results by omitting patents that contain unwanted terms.

Truncation Codes:

- **+ (Plus sign):** The plus sign is a truncation symbol used to account for variations of a word, such as different tenses, plural forms, or word stems. For instance, in the query (RATING+ AND FEEDBACK), the plus sign after “RATING” ensures that patents containing any form of the word “RATING” (e.g., “RATING”, “RATINGS”, etc.) are included in the search results. This allows for a broader search, ensuring that all potential relevant documents are retrieved without needing to specify each variation explicitly.

Field Codes:

- **TI:** This code specifies the title of the patent document. When included in a query, it restricts the search to patents where the specified terms appear in the title.
- **AB:** This code is used to search the abstract of a patent. It ensures that the search results include documents where the terms appear in the abstract of the patent.

- **CLSM:** This code targets the claims section of a patent. By using this field code, the search focuses on the legal claims and their specific details.
- **DESC:** This field code searches the description section of the patent, ensuring that the terms are included within the full description of the invention.
- **ODES:** This code pertains to the description of the object of the invention and ensures that the search targets this specific section of the patent.
- **ICLM:** This code is used to search for terms within the International Classification of Patents (IPC) section, which can provide insights into the patent's classification within specific technological domains.

Appendix C

Complete Set of Patent Identifiers Identified at the End of the Process

Table A3. Final set of 49 distinct patent identifiers and their consensus-assigned relevance categories.

Patent Identifier	Category Code
US9659299	2
US10380611	1
US20110251871	1
IN2991/CHE/2012	1
IN201741006328	3
IN0065/DEL/2013	2
CA2974427	2
US20170323355	2
US20170300943	1
US20130346155	1
US7996252	1
US20140344086	1
US8527307	1
US20080097769	1
US7281652	2
US20180218381	2
US8595061	1
US8473304	3
WO2009/029989	3
US20170300942	1
US9824323	2
US9489667	3
WO2021/003396	3
US10970732	2
WO2007/048125	1
US7383200	2
US4345315	2
GB2534342	3
US9418344	1
IN202231009022	3
KR10-1667552	2
US20220405781	1
WO2010/094092	2
RU2012115848	3

US7652558B2	2
US8335709	2
US20170098227	1
US20060259347	1
US20180189810	1
US20010037206	2
US20150170228	2
US20050149382	2
US20160253686	1
GB2456346	2
US7818203	1
US6960988	2
US6970831	3
US6236975	3
US7376570	2

References

1. Arts, S.; Hou, J.; Gomez, J.C. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Res. Policy* **2021**, *50*, 104144. <https://doi.org/10.1016/j.respol.2020.104144>.
2. Cascini, G.; Fantechi, A.; Spinicci, E. Natural language processing of patents and technical documentation. In *Document Analysis Systems VI*; Marinai, S., Dengel, A.R., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3163, pp. 508–520. <https://doi.org/10.1007/b100557>.
3. Dong, H.; Hussain, F.K.; Chang, E. A context-aware semantic similarity model for ontology environments. *Concurr. Comput. Pract. Exp.* **2011**, *23*, 505–524. <https://doi.org/10.1002/cpe.1652>.
4. Souili, A.; Cavallucci, D.; Rousselot, F. Natural Language Processing (NLP)—A Solution for Knowledge Extraction from Patent Unstructured Data. *Procedia Eng.* **2015**, *131*, 635–643. <https://doi.org/10.1016/j.proeng.2015.12.457>.
5. Krestel, R.; Chikkamath, R.; Hewel, C.; Risch, J. A survey on deep learning for patent analysis. *World Pat. Inf.* **2021**, *65*, 102035. <https://doi.org/10.1016/j.wpi.2021.102035>.
6. Setchi, R.; Spasić, I.; Morgan, J.; Harrison, C.; Corken, R. Artificial intelligence for patent prior art searching. *World Pat. Inf.* **2021**, *64*, 102021. <https://doi.org/10.1016/j.wpi.2021.102021>.
7. Ali, A.; Tufail, A.; De Silva, L.C.; Abas, P.E. Innovating Patent Retrieval: A Comprehensive Review of Techniques, Trends, and Challenges in Prior Art Searches. *Appl. Syst. Innov.* **2024**, *7*, 91. <https://doi.org/10.3390/asi7050091>.
8. Villa, A.M.; Wirz, M. A sequential patent search approach combining semantics and artificial intelligence to identify initial State-of-the-Art documents. *World Pat. Inf.* **2022**, *68*, 102096. <https://doi.org/10.1016/j.wpi.2022.102096>.
9. Moeller, A.; Moehrl, M.G. Completing keyword patent search with semantic patent search: Introducing a semiautomatic iterative method for patent near search based on semantic similarities. *Scientometrics* **2015**, *102*, 77–96. <https://doi.org/10.1007/s11192-014-1446-9>.
10. Alderucci, D.; Sicker, D. Applying Artificial Intelligence to the Patent System. *Technol. Innov.* **2019**, *20*, 415–425. <https://doi.org/10.21300/20.4.2019.415>.
11. Lee, C.; Kwon, O.; Kim, M.; Kwon, D. Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technol. Forecast. Soc. Change* **2018**, *127*, 291–303. <https://doi.org/10.1016/j.techfore.2017.10.002>.
12. Barbieri, M. Patent prior art searches: Basic principles and strategies. *Preprints* **2022**, 2022050054. <https://doi.org/10.20944/preprints202205.0054.v1>.
13. Clarke, N.S. The basics of patent searching. *World Pat. Inf.* **2018**, *54*, S4–S10. <https://doi.org/10.1016/j.wpi.2017.02.006>.
14. Kumar, R.; Tripathi, R.C.; Singh, V. Keyword Based Search and its Limitations in the Patent Document to Secure the Idea from its Infringement. *Procedia Comput. Sci.* **2016**, *78*, 439–446. <https://doi.org/10.1016/j.procs.2016.02.086>.
15. Montecchi, T.; Russo, D.; Liu, Y. Searching in Cooperative Patent Classification: Comparison between keyword and concept-based search. *Adv. Eng. Inform.* **2013**, *27*, 335–345. <https://doi.org/10.1016/j.aei.2013.02.002>.
16. Sánchez, D.; Batet, M.; Isern, D.; Valls, A. Ontology-based semantic similarity: A new feature-based approach. *Expert Syst. Appl.* **2012**, *39*, 7718–7728. <https://doi.org/10.1016/j.eswa.2012.01.082>.

17. Choi, S.; Yoon, J.; Kim, K.; Lee, J.Y.; Kim, C.H. SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics* **2011**, *88*, 863–883. <https://doi.org/10.1007/s11192-011-0420-z>.
18. Yoon, J.; Kim, K. Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics* **2011**, *88*, 213–228. <https://doi.org/10.1007/s11192-011-0383-0>.
19. Carpineto, C.; Romano, G. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.* **2012**, *44*, 1–50. <https://doi.org/10.1145/2071389.2071390>.
20. Fall, C.J.; Töröcsvári, A.; Benzineb, K.; Karetka, G. Automated categorization in the international patent classification. *ACM SIGIR Forum* **2003**, *37*, 10–25. <https://doi.org/10.1145/945546.945547>.
21. Couteau, O. Forward searching—A complement to keyword- and class-based patentability searches. *World Pat. Inf.* **2014**, *37*, 33–38. <https://doi.org/10.1016/j.wpi.2014.01.007>.
22. Ferraz-Caetano, J.; Teixeira, F.; Cordeiro, M.N.D.S. The Role of Knowledge Representation & Reasoning in Deciphering Chemical Complexity. *WIREs Comput. Mol. Sci.* **2026**, *16*, e70085. <https://doi.org/10.1002/wcms.70085>.
23. Wellawatte, G.P.; Schwaller, P. Human interpretable structure-property relationships in chemistry using explainable machine learning and large language models. *Commun. Chem.* **2025**, *8*, 11. <https://doi.org/10.1038/s42004-024-01393-y>.
24. Cascini, G.; Zini, M. Measuring patent similarity by comparing inventions functional trees. In *FIP 20th World Computer Congress, Proceedings of the Second Topical Session on Computer-Aided Innovation, WG 5.4/TC 5 Computer-Aided Innovation, Milano, Italy, 7–10 September 2008*; Springer: Boston, MA, USA, Volume 277, pp. 31–42. https://doi.org/10.1007/978-0-387-09697-1_3.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.