Intelligent Patent Retrieval using TF-IDF

Yen Le YTL2008

Biraj Chowdhury BC3052

Isaiah Levy IL1016

Arnav Kanwal AJK8685

Abstract

The acceleration of technological innovation and the subsequent expansion of intellectual property patents demand a sophisticated mechanism to efficiently navigate the vast expanse of IP documentation. In this paper, we aim to address these challenges associated with patent searching by developing a system to rank and return relevant patents based on search queries. The primary objective of this system is to streamline the process of retrieving existing intellectual property (IP). We standardized the steps of data pre-processing using dimension reduction and word stemming, and experimented with Latent Dirichlet Allocation (LDA) and Query Extension to create an enhanced TF-IDF system for IP retrieval on widely-used patent databases. Unlike other NLP tasks, information retrieval requires sensitivity to both relevance and ranking, thus leading us to evaluate our system with average precision, mean average precision (MAP), and normalized discounted cumulative gain (NDCG). Our findings indicate that incorporating synonyms of words using WordNet significantly boosted model efficiency and our best-performing models achieved a MAP score of 61.6% and average NDCG of 87.4%. Our study showcases an efficient and user-friendly TF-IDF system in retrieving intellectual property data based on a widely distributed set of search queries.

1. Introduction

In the realm of patent research and analysis, the effectiveness of query search tools is paramount. The existing systems, notably the United States Patent and Trademark Office's (USPTO) search tool and Google Patents, present significant challenges in terms of complexity, user-friendliness, and relevance of search results. These limitations highlight a critical gap in the field of patent search and retrieval, underscoring the need for a more efficient and intuitive tool. Our project addresses this need by developing an advanced query searching tool that leverages popular techniques in Natural Language Processing (NLP) and similar information retrieval tasks.

The primary motivation behind our project stems from the complexities and inadequacies inherent in current patent search tools. The USPTO's search tool, while comprehensive, suffers from a steep learning curve due to its intricate interface and lack of clear documentation. This complexity often deters users, especially those without extensive legal or technical background, from effectively utilizing the system. On the other hand, Google Patents, despite its user-friendly interface, falls short in delivering the most relevant patent information. These shortcomings in existing platforms hinder the efficient discovery and analysis of patent information, a critical component for innovation and research.

To bridge this gap, our project introduces an innovative approach to patent query searching. We focus on enhancing the accuracy and relevance of search results through sophisticated text preprocessing and query expansion techniques. Our system employs dimensionality reduction and word stemming as part of its preprocessing strategy, essential for handling the vast and varied nature of patent data. A notable innovation in our approach is the implementation of query expansion. By incorporating synonyms into search queries, enhanced through part-of-speech (POS) tagging to filter unrelated synonyms, our tool significantly improves upon traditional term frequency-inverse document frequency (TF-IDF) models.

The backbone of our search algorithm is a combination of TF-IDF along with manual scoring and vast evaluation metrics. This method ensures that the search results are not only relevant but also contextually aligned with the user's query. Our system has been meticulously developed and tested using the Harvard USPTO dataset, a comprehensive repository of patent data. We have processed and stored this dataset in MongoDB, ensuring efficient data retrieval and management.

In summary, our project aims to revolutionize the patent search experience by introducing a tool that is not only powerful in its search capabilities but also intuitive and user-friendly. By addressing the shortcomings of existing systems, our tool stands to significantly benefit researchers, legal professionals, and inventors in navigating the vast expanse of patent information with ease and precision.

2. Related Works

2.1 Evaluation Metrics for Information Retrieval Tasks

Filip Radlinski and Nick Craswell (2010) discuss the sensitivity of popular information retrieval metrics and helped us guide the decision making process for the evaluation methods of our system. Unlike other traditional measures used for other NLP tasks like precision, recall, and f-score, information retrieval tasks require methods that evaluate performance of the system and take into account the ranking and relevance of documents. In their paper, they compare the performance and insight gained of metrics such as Precision at cutoff k, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG). To do this, they created an experiment in which they interleaved five real web-based search and ranking functions. From their results, they discovered a high correlation between NDCG and the amount of manually annotated data as these scores rely on human-assigned relevance scores. On queryset sizes under a thousand, it was hard for NDCG to pick up minute differences in ranking quality, however, still recommend its usage in addition to MAP-scores as an extremely important tool for evaluating relevance and order of a search function's retrieval results.

2.2 Patent Retrieval and Document Classification using Enhanced TF-IDF

Lee, Jung-Hoon, et al. (2020) provide a foundational methodology in "Human-centric Computing and Information Sciences," which our Intellectual Property Tracker system aims to incorporate for improved patent retrieval. Their work introduces a system that constructs a keyword dictionary using user inputs and LDA-extracted topics, ensuring thematic alignment with search queries. They further enhance document classification by coupling TF-IDF with K-means clustering, a technique our project will adopt to categorize patents efficiently. By integrating these strategies, we anticipate a heightened precision in our search results, providing users with patents that are relevant both contextually and thematically. Emulating their evaluative methods, we aim to achieve a performance standard that is both robust and comparative to established systems.

2.3 Patent Retrieval and Document Classification using Hybrid Adaptable Machine Learning

Terryn, Ayla Rigouts, et al. (2021) provide a machine learning approach to computational linguistics problems in their paper "Hybrid Adaptable Machine Learning approach to Extract Terminology." Since our task includes providing key-words to search through and return the best-matching patents, this paper presents us with a possible algorithm for terminology extraction which achieves relatively high average precision results. The paper walks through their process of picking a large public dataset to train and test their algorithm on and explains how they used a mixture of features from traditional POS-tags and TF-IDF scores to more complex statistical features based on term-likelihood and more. After narrowing down the top 30 most important features from this list, they implemented a simple Random Forest Classifier (type of supervised machine learning model) as their algorithm for terminology extraction. The paper's conclusions state how their results would be incredibly useful for domain-specific applications such as intellectual property patents which may contain very specific and unusual vocabulary and terminology that is not commonly present in other forms of text. They also provide a detailed error analysis including potential areas of improvement and how they calculated their measurements such as precision, recall, and f1-scores. This paper provides insightful information for when we need to do our own error analysis and provides a helpful foundation for discovering an algorithm that can help us with our task for document classification of patents using terminology.

3. Methodology

Inspired by Lee, Jung-Hoon et al. (2020), we aimed to refine our TF-IDF model for patent retrieval. We build an enhanced TF-IDF model using various Natural Language Processing methodology to process and analyze patent texts. For each patent, we create a comprehensive text amalgamation, selecting five main attributes of our data, namely 'abstract', 'claims', 'background', 'summary', and 'description' fields. This selection of data ensures a holistic representation of each patent's content.

We conduct data pre-processing through a series of steps, including lowercasing, punctuation removal, tokenization, stop word removal, and stemming. These steps are fundamental in reducing the dimensionality of our text data, thereby enhancing the focus on relevant linguistic features.

Post-preprocessing, we constructed a bag-of-words model, which simplifies textual representation while preserving essential information. This model facilitates the creation of a structured corpus and dictionary, which is pivotal for our TF-IDF (Term Frequency-Inverse Document Frequency) modeling.

A significant enhancement in our model is the implementation of query expansion. Using the synset (synonym set) structure from WordNet, we enrich user queries with contextually appropriate synonyms. This expansion successfully ensures the semantic integrity of the original query, thus enhancing the retrieval process without diluting query intent.

Our model's effectiveness is quantitatively evaluated using metrics such as average precision, MAP-scores, and normalized discounted cumulative gain. These metrics together determine the model's capability in accurately retrieving and ranking relevant patents from a vast corpus. In conclusion, our methodology is structured with data selection, text preprocessing, structured TF-IDF modeling, and enhanced query expansion, all aimed at creating a state-of-the-art patent retrieval system that is both efficient and accurate.

4. Data

4.1 Data sources

4.1.1 Harvard USPTO Dataset (HUPD)

For this project, we utilized the Harvard USPTO dataset (HUPD), released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license. This dataset is a comprehensive corpus of over 4.5 million English-language utility patent applications filed with the USPTO from January 2004 to December 2018. The dataset comprises inventor-submitted versions of patent applications, rather than the final versions of granted patents. Key features included in the dataset dictionary are patent number, decision, title, abstract, claims, background, summary, description, CPC label, IPC label, filing data, patent issue date, date published, and examiner ID. For our patent retrieval purposes, we selectively extracted five main features, mentioned above, from this dataset.

4.1.2 Google Patents Dataset

In addition to the HUPD dataset, we also utilized a secondary dataset for model testing, derived from Google Patents. This dataset was constructed by performing Google searches for patent IDs related to various queries using SerpAPI, and then scraping patent details from Google Patents URLs with BeautifulSoup. To ensure compliance with web scraping best practices, we implemented a rate limit with a 1-second delay between requests. This dataset, comprising approximately 600 patent documents obtained from about 15 results for each of the 40 queries, includes details like patent name, number, date produced, and content. This content was processed and stored in a JSON format, encompassing the abstract, description, title, and summary of each patent.

4.2 Data Exploration and Data Ultilization

4.2.1 Development with the Harvard USPTO Dataset

The Harvard USPTO dataset (HUPD) served as the primary resource for developing our patent search model. This dataset's extensive collection of over 4.5 million patent documents provided a rich and varied base for our model to learn from. By training on this dataset, our model was exposed to a wide range of patent application types and linguistic styles, enabling it to develop a robust understanding of patent language and structure.

4.2.2 Validation with the Google Patents Dataset

After developing our model on the HUPD dataset, we moved to the testing phase using the Google Patents dataset. This dataset, constructed through web scraping, consisted of around 600 patent documents derived from various Google Patents search results. Unlike the training phase, where the focus was on learning patterns and features of patent documents, the testing phase aimed at evaluating the model's performance in real-world scenarios.

We preserved the original ranking of each patent result as well as the initial query it was associated with in the Google Patents dataset. This approach allowed us to generate accurate testing metrics, assessing the model's effectiveness in retrieving relevant patents and maintaining the fidelity of search results.

5. System Overview

5.1 System Development

5.1.1 Data Acquisition and Database Integration

The system initiates by either fetching patent data from a JSON file or directly interfacing with a MongoDB database. This dual-source approach ensures flexibility and scalability in data handling. For our purpose, we chose MongoDB for its efficiency in dealing with large datasets. This is crucial for storing and retrieving patent texts. The database setup includes connectivity testing and dynamic collection creation, ensuring a robust data management framework.

5.1.2 Advanced Text Preprocessing

Each patent text undergoes a series of thorough text preprocessing steps. After lowercasing and punctuation removal, tokenization is performed to break down the text into individual tokens.

A critical preprocessing step is the use of the Porter Stemmer algorithm, which simplifies words to their base form. This is done to reduce data complexity and enhance focus on key lexical elements. We also remove stopwords to filter out common, less significant words; ultimately, streamlining the dataset for more effective analysis.

5.1.3 Corpus Construction and Semantic Processing

The system employs the Gensim library to construct a bag-of-words corpus from the preprocessed text. This step transforms the textual data into a structured format, crucial in preparing the data for model building.

A TF-IDF model is then generated from this corpus. This model quantifies the importance of words within the corpus, balancing their frequency across documents and the entire dataset.

5.1.4 Query Expansion and Synonym Incorporation

An innovative aspect of our system is query expansion. Queries are enriched with synonyms identified through WordNet, leveraging part-of-speech tagging to ensure contextual relevance. This process significantly enhances the query's scope without compromising its semantic integrity.

Expanded queries are preprocessed identically to the patent texts, maintaining consistency in the system's treatment of textual data.

5.1.5 Similarity Analysis and Patent Retrieval

The system converts queries into the TF-IDF space and employs cosine similarity to gauge their relevance against the patent corpus. This similarity computation is key to identifying patents most pertinent to the query.

The top-ranking patents, determined by their similarity scores, are earmarked for each query, enabling precise and relevant retrieval.

5.1.6 Data Structuring and Output Generation

The retrieval results are organized into a structured JSON format, retaining patent names and their corresponding relevance scores. This structured output is crucial for downstream analysis and user interpretation.

5.1.7 System Robustness and Scalability

The system's architecture allows for efficient processing of large datasets and adaptation to varying data sources; thus, fulfilling scalability. This is achieved through modular design and flexible data handling mechanisms.

MongoDB integration also ensures robust data storage and retrieval capabilities, crucial for managing the vast and varied nature of patent data.

5.2 Implementation of Baseline & Test Models

A plain TF-IDF model was tested to evaluate the performance of the baseline model. The baseline model was trained using the Harvard HUPD dataset. We then test the model on the acquired data through web scraping Google Patents API. The subsequent tests include various ways to potentially improve the performance of our system, including adding LDA in topic modeling and expanding our query using synonyms words from the WordNet.

6. Evaluation and Results

In this section, we discuss the criteria used for creating and manually annotating an answer key, the reasoning behind the evaluation metrics we chose to use, and the results from our system's output.

6.1 Manual Scoring

For the evaluation process, we wanted to create a human annotated answer key for validating and testing our system. While we used the retrieved results from the Harvard USPTO and Google Patent's search systems as our benchmark in training, manually grading a subset of such data serves for creating a better source of ground truth when evaluating an IR system. We took a subset of 200 patents which were roughly equally distributed among 10 random search queries for manual evaluation. For scoring, we created a list of relevant documents per query and assigned each document a relevance score from 0-100. We also ranked each list in sorted order by their relevance which allowed us to preserve the importance of rank when evaluating our system. Two independent scorers evaluated patents based on well-defined criteria: alignment with key elements of the query, addressing the intent of the query, and the depth and scope of coverage. A third evaluator then reviewed these scores to ensure consistency and resolve any discrepancies.

Criteria:

- Alignment with Key Elements of Query (50 points)
  - 0-10 Points: The patent has minimal or no mention of key elements.
  - 11-30 Points: The patent mentions some key elements but does not focus on them.
  - 31-50 Points: The patent directly addresses most or all key elements of the query.
- Addressing the Intent of the Query (30 points)
  - 0-10 Points: The patent does not address the underlying intent or need of the query.
  - 11-20 Points: The patent partially addresses the intent or need but not in a comprehensive manner.
  - 21-30 Points: The patent fully addresses the intent or need outlined in the query.
- Depth and Scope of Coverage (20 points)
  - 0-6 Points: The patent provides a superficial or tangential connection to the query's subject.
  - 7-13 Points: The patent covers the subject to a moderate degree but lacks depth.
  - 14-20 Points: The patent thoroughly explores the subject and is closely related to the query.

6.2 Average Precision and MAP-Score

Metrics like precision and recall are single value metrics based on the whole list of documents, but for retrieval tasks like ours, it is also important to consider the order in which the documents were returned thus leading us to use average precision as one of our scoring criteria. We calculated average precision with using the following formula

$$\frac{\sum_{k=1}^{n} P(k) \times rel(k)}{\text{total number of relevant documents}}$$

where k is the rank in sequence of returned documents, n the total number of retrieved documents, $P(k)$ the precision at cutoff k, and $rel(k)$ a binary indicator function equal to 1 if the kth document is relevant or 0 if not. The mean of these scores across all queries results in a MAP-score which is indicative of our system's ability to generalize across any dataset.

6.3 Normalized Discounted Cumulative Gain

The next set of metrics we chose were NDCG which utilizes the graded relevance scores in our manually annotated answer key. In contrast to average precision, it takes into account the relative position of documents returned rather than just placing emphasis on the top ranked documents. It is calculated as such

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}$$
$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

where i is the position of the current document and rel_i the relevance score of that document. Similarly, the ideal discounted cumulative gain ($IDCG_p$) is calculated like DCG_p but using the documents and their respective scores in order per the answer key. The quotient of DCG and IDCG provides a normalized score for assessing performance. We can also calculate a mean of NDCG scores across all queries to measure our model's ability to generalize and return patents in order of relevance.

6.4 Results

Figure 1: Patent Retrieval Scores on Different System Versions

	Retrieval System		
	TF-IDF (base)	TF-IDF + LDA	TF-IDF + Query Expansion
MAP-Score	0.531	0.438	0.616
Mean NDCG	0.761	0.672	0.874

Per the results of figure 1 above, we found that in addition to the various word pre-processing techniques we used, a well refined TF-IDF system with query expansion allowed our searches to be more robust resulting in both the highest and MAP and mean NDCG scores.

Recall that normalized discounted cumulative gain as a metric allows us to evaluate the overall quality of returned documents by our system. In all such versions of the system, we found that it was easy to achieve a high mean NDCG score when averaging NDCG across all queries that we tested. A potential limitation of the high scores we see here could just be because of the limited size of our test dataset. If we could manually annotate a larger set of patents we could potentially see a wider variety of different documents retrieved by our system, thus lowering the overall relevance and NDCG scores. Our system did utilize highly advanced text pre-processing techniques and created a well defined corpus of words from which to build our system on top of. A final mean NDCG score of 87.4% is indicative of our system as a whole in returning the most relevant documents.

When experimenting with LDA, we saw the most various in our MAP-score used these results to conclude that our system would perform better without. By using query-expansion, we achieved a final MAP-score of 61.6%, showing our system's ability to retrieve relevant documents at the top of the ranked lists for all queries. Recall that MAP scores are evaluated from a range of 0 to 1, and our score signifies the effectiveness of our system in delivering the top results as the first set of results that a user will see.

7. Future Works and Concerns

The main challenge with patent search was the vast and varied nature of patent documentation, along with intricate domain languages and large volumes of documents. Our model, enhanced TD-IDF along with Latent Dirichlet Allocation (LDA) and Query Extension, was designed with an aim to mitigate these challenges. Our methodologies, as explained above, revealed the inherent difficulties in providing definitive assessments of model accuracy in such a complex domain.

Our findings indicate that LDA did not significantly improve the performance of our system. However, the implementation of Query Extension markedly enhanced the relevance of the retrieved results. This success underscores the value of synonym incorporation and advanced query processing techniques in improving search effectiveness. Aside from that, our research also highlights a critical limitation: the scope of available data. The Harvard USPTO dataset represents a limited perspective of the global patent landscape. Similarly, the Google Patents dataset also failed to include the diversity in patent documents, due to our specific search queries and scraping methodologies.

We posit that with access to a broader variety of data and global patent databases, our model could potentially yield more relevant and accurate results. This hypothesis stems from the understanding that diversity of training and testing data play a crucial role in the refinement and success of NLP models.

In future works, we aim to expand our dataset to include a wider range of global patent documentation. This will be done by expanding our training data to new sources to provide a more comprehensive view of the patent landscape. We will also aim to expand our testing data by increasing the number of self-annotated answer keys to best fit our purpose of testing and evaluation. These steps will ultimately enhance the model's accuracy and relevance by incorporating diverse training and testing data.

Inspired by Terryn, Ayla Rigouts, et al. (2021) , we also aim to explore hybrid Machine Learning and Natural Language Processing models in our venture to create a more efficient and accurate intellectual property tracker system. With regards to this, we considered the effect of transfer learning, by utilizing our existing pretrained enhanced TF-IDF with query expansion models to extract semantic features and perform further learning based on these features. We can then train the network with our extended data to enhance the model's performance.

8. Conclusion

This paper explores the complexities of patent searching, a task that has become increasingly critical with the rapid growth of technological innovation and intellectual property documentation. We have tackled the challenge of streamlining the process of classifying and retrieving intellectual property patents through an enhanced TF-IDF system with Query Expansion, leveraging the potential of Natural Language Processing techniques.

In terms of broader and more practical applications, our study contributes to the existing research aiming to streamline the process of patent retrieval. Through our evaluation, our model demonstrated promising capabilities in improving efficiency and accuracy of patent searches. Future endeavors in this field should aim to expand the dataset horizons, incorporate more nuanced NLP/ML techniques, and continuously refine models to adapt to the evolving nature of patent documentation. Our study contributes to the ongoing efforts in enhancing intellectual property research and opens avenues for more advanced, comprehensive models in the realm of information retrieval.

References

Helmers, L., Horn, F., Biegler, F., Oppermann, T., Müller, K. R. (2019). ”Automating the search for a patent’s prior art with a full text similarity search.” PloS one, 14(3), e0212103. https://doi.org/10.1371/journal.pone.0212103

Kaur, Bipanjyot, and Gourav Bathla. ”Document classification using various classification algorithms: a survey.” Int J Fut Revol Comput Sci Commun Eng. 2018. [PDF] academia.edu

Kavitha, Modepalli and P. Prabhavathy, ”A review on machine learning techniques for text classification,” 2021 4th International Conference on Computing and Communications Technologies (ICCCT), Chennai, India, 2021, pp. 605-610, https://ieeexplore.ieee.org/abstract/document/9711858

Krestel, Ralf, et al. ”A survey on deep learning for patent analysis.” World Patent Information. 2021. https://www.sciencedirect.com/science/article/pii/S017221902100017X

Lee, Jung-Hoon, et al. ”Paper classification systems based on TF-IDF and LDA schemes.” Human-centric Computing and Information Sciences. 2020. [HTML] springer.com

Radlinski, Filip, and Nick Craswell. ”Comparing the sensitivity of information retrieval metrics.” Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010. https://dl.acm.org/doi/abs/10.1145/1835449.1835560

Terryn, Ayla Rigouts, et al. ”Hybrid Adaptable Machine Learning Approach to Extract Terminology.” International Journal of Theoretical and Applied Issues in Specialized Communication,Vol 27, Issue 2, Oct 2021, p. 254 - 293, https://www.jbe-platform.com/content/journals/10.1075/term.20017.rig

Wang, Meiyun, et al. ”Discovering new applications: Cross-domain exploration of patent documents using causal extraction and similarity analysis.” World Patent Information 75. 2023. [HTML] sciencedirect.com

Yanagihori, Kyoko, Koji Tanaka, and Kazuhiko Tsuda. ”Improvement of terminology extraction method for specific patent search.” Procedia Computer Science. 2014. [PDF] sciencedirect.com