Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/42/20230795

The application of NLP in information retrieval

Xurui Wang

Department of Computer Science, Dalian University of Technology, Dalian, Liaoning, China

wxr20030331@email.dlut.edu.cn

Abstract. The field of Natural Language Processing (NLP) has experienced impressive advancements and has found diverse applications. This paper presents a comprehensive review of the development of NLP in the field of information retrieval. It explores different stages of NLP techniques and methods, including keyword matching, rule-based approaches, statistical methods, and the utilization of machine learning and deep learning technologies. Furthermore, the paper provides detailed insights into the specific applications of NLP in domains such as academic information retrieval, medical information retrieval, travel information retrieval, and e-commerce information retrieval. It analyzes the current state of NLP applications in these domains, highlights their advantages, and discusses their associated limitations. Finally, the paper emphasizes the continuous advancement of the NLP field, with a particular focus on semantic understanding, personalized retrieval, and multimodal information retrieval, to better adapt to diverse data types and user requirements. The paper concludes by summarizing the main points discussed and providing future directions.

Keywords: NLP, Information Retrieval, Academic Information, Medical Information, Travel Information, E-commerce Information.

1. Introduction

The field of Natural Language Processing (NLP) has witnessed remarkable progress and diversified applications. The concept of NLP has evolved through various stages, from early keyword matching systems to contemporary deep learning techniques, revolutionizing how information is retrieved and understood. NLP has found applications in academic literature retrieval, medical knowledge extraction, travel information search, and e-commerce services. Understanding its evolution and applications is vital due to its potential to enhance efficiency and precision in diverse domains.

In the modern academic landscape, the exponential growth of scholarly literature necessitates efficient information retrieval. Academic databases and search engines now harness NLP to comprehend user queries, categorize documents, and provide relevant literature. This advancement significantly aids researchers in staying abreast of developments in their respective fields. Similarly, in the medical domain, NLP plays a pivotal role in constructing structured medical knowledge graphs and empowering question-answering systems, thereby improving healthcare outcomes.

Travel information retrieval benefits both travelers and service providers. NLP-driven systems offer personalized search and recommendations, voice-driven interactions, multilingual support, and sentiment analysis. This facilitates convenient trip planning for travelers and helps service providers cater to specific needs.

E-commerce information retrieval leverages NLP to provide personalized recommendations, sentiment analysis for brand management, user feedback analysis, and automated customer service. While enhancing user satisfaction, it also poses challenges regarding privacy and system accuracy. The motivation for this study stems from the evolving nature of NLP technology and its increasing relevance across domains. The research seeks to shed light on NLP's historical development, current applications, and future potential. By comprehensively understanding its evolution and applications, we can harness NLP's capabilities to advance information retrieval and improve user experiences.

This paper explores the evolution and applications of NLP in various fields of information retrieval. It aims to highlight the significance of NLP in addressing information overload, enhancing search efficiency, and facilitating personalized experiences for users across academic, medical, travel, and e-commerce domains.

2. Development of NLP in the Field of Information Retrieval

The evolution of NLP in the domain of information retrieval can be traced back several decades, encompassing various stages and significant milestones.

- ***Early Stages of Information Retrieval and Transition to Rule-Based Approaches (1950s-1990s)***

In the nascent years of information retrieval, systems primarily relied on a simple yet effective approach: keyword matching. During this period, users would input specific keywords, and the system would return documents that contained exact matches to those keywords. This approach, while straightforward, had its limitations. It excelled when keywords were precise and well-defined but struggled with more complex queries or when documents contained synonymous terms or related concepts. The primary drawback was its inability to grasp the context of the user's query, often resulting in irrelevant or incomplete results.

- **Rise of Statistical Methods and the Internet (2000s-2010s)**

The advent of the internet brought about a paradigm shift in information retrieval. The availability of vast amounts of text data led to the prominence of statistical methods. Bag-of-words models and vector space-based retrieval techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF), gained traction [1]. It discusses the use of TF-IDF (Term Frequency-Inverse Document Frequency) to identify relevant words for queries within a document corpus. TF-IDF calculates word values based on their frequency within a document and across the entire corpus. High TF-IDF values suggest a strong word-document relationship, indicating potential relevance in queries. The paper demonstrates that TF-IDF efficiently categorizes relevant words, improving query retrieval. Search engines like Google introduced algorithms like PageRank to refine result rankings. These statistical methods had the advantage of scalability and adaptability. They could accommodate larger datasets and adapt to variations in language and content. However, they still struggled with understanding the semantic nuances of language and context, often producing results that matched keywords but failed to capture the user's true intent.

- **Emergence of Machine Learning and Deep Learning (2010s-Present)**

The most recent phase has witnessed a transformative shift in information retrieval, driven by the rise of machine learning and deep learning techniques in NLP. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) revolutionized text feature extraction and sequence modeling [2]. After the release of the Transformer model, NLP entered a new era. Pre-trained language models like BERT and GPT, trained on extensive language corpora, provided powerful foundational models for various NLP tasks. These models excelled at understanding context and semantics, making them invaluable for query expansion, document retrieval, and question-answering systems. They significantly improved retrieval effectiveness by moving beyond keyword matching to semantic understanding.

- **Semantic Understanding and Personalized Retrieval and Multimodal Information Retrieval (Recent Years)**

In recent years, NLP technology has continued to advance, with a focus on deeper semantic understanding. Systems are increasingly capable of interpreting user queries beyond keywords, considering the intent and context behind the search. Moreover, personalized retrieval has emerged as a pivotal direction. Information retrieval systems now analyze users' historical behavior and interests to deliver more relevant search results. This personalization enhances the user experience, ensuring that the retrieved information aligns with individual preferences and needs. As the volume of multimedia data explodes, the importance of multimodal information retrieval has grown significantly. NLP, when combined with computer vision, audio processing, and other technologies, facilitates cross-modal information retrieval. For example, it enables the extraction of textual information from images or the retrieval of relevant content from speech data. This multimodal approach broadens the scope of information retrieval, making it more versatile and adaptable to the diverse data types encountered in the modern digital landscape.

These distinct phases in the evolution of information retrieval reflect not only technological advancements but also the evolving nature of user needs and expectations. While each phase has brought its own set of advantages and challenges, the progression toward deeper semantic understanding and personalization, coupled with multimodal capabilities, represents a promising future for information retrieval.

3. Information retrieval based on NLP

NLP has numerous applications in information retrieval. This paper will primarily focus on aspects such as Academic Information Retrieval, Medical Information Retrieval, Travel Information Retrieval, E-Commerce Information Retrieval.

3.1. Application of NLP in Academic Information Retrieval

The utilization of NLP in the realm of academic information retrieval can significantly enhance researchers' efficiency and the accuracy of information acquisition, aiding them in conducting scholarly research more effectively.

- **Literature Retrieval**

In today's scholarly databases and major search engines, the pervasive presence of NLP techniques is evident. Some classic algorithms and models include keyword-based retrieval algorithms, vector space model-based algorithms, and concept retrieval algorithms. Keyword-based retrieval algorithms are the simplest and traditional method that matches user-inputted keywords with keywords in the literature to retrieve relevant documents. Vector space model-based algorithms represent documents as vectors in a high-dimensional space and use techniques like term frequency-inverse document frequency (TF-IDF) to measure document importance and similarity for ranking and retrieval. Concept retrieval algorithms rely on knowledge graphs, ontologies, and semantic relationships to achieve more precise retrieval. For instance, WordNet is a classical lexical database that utilizes semantic relationships such as hypernyms, hyponyms, and synonyms to assist in literature retrieval. The widespread employment of language models such as BERT and GPT enables search tools to comprehend the context of text. Information extraction techniques swiftly and accurately extract key information from searched articles.

- **Literature Comparison**

Benefiting from the progress of NLP technology, tasks like named entity recognition have gradually become more accurate and efficient [3]. This study outlines the methodology employed by the Turku NLP group for the PharmaCoNER task, focusing on named entity recognition in Spanish biomedical texts. They employ both a CRF-based baseline approach and multilingual BERT for this task, resulting in impressive performance with an F-score of 88% on the development dataset and 87% on the test dataset when using BERT. It's worth noting that their approach involves the straightforward application of a cutting-edge multilingual model, without specific fine-tuning for either the language or the biomedical domain. In previous automated comparison systems, deviations in results often arose due to inaccurate analysis of the core content within texts. Utilizing contemporary NLP techniques, such as BERT, BILSTM networks, and CRF functions, facilitates the extraction of key textual information. document similarity-based algorithms, named entity recognition (NER) algorithms, and text model-based comparison algorithms. Document similarity-based algorithms compare the similarity between two documents using metrics like cosine similarity. These algorithms help researchers find related literature or detect duplicate documents. NER algorithms aim to identify named entities such as person names, locations, and organizations within literature. By recognizing and comparing named entities, researchers can determine the relevance between documents. Additionally, text model-based comparison algorithms leverage pre-trained language models like BERT and GPT to encode and represent literature. By calculating the similarity between documents, comparison and ranking can be performed. Further, similarity comparison algorithms like cosine similarity yield relatively precise measures of similarity between texts.

- **Literature Learning**

Learning from literature is an essential means of acquiring knowledge. Some classic algorithms and models include language models, topic modeling, and text classification algorithms. Language models such as GPT learn language patterns in literature by predicting the next word. These models can be used to generate text and answer questions. Topic modeling algorithms can identify and extract hidden topics or themes from literature, aiding researchers in understanding the structure and content of literature. Text classification algorithms classify and label literature. For example, literature can be classified based on domain, topic, or sentiment, facilitating better organization and management of documents. This research use NLP techniques to extract semantic relationships between medical concepts, like drugs and diseases [4]. The study combines NLP, UMLS ontology, and Support Vector Machines to accurately identify these relationships from medical texts, achieving impressive performance improvements, especially in identifying "cure" relationships with a 98.19% f-score. This work highlights NLP's significant role in enhancing knowledge acquisition and application in the medical field. The rise of large language models like GPT has made learning from literature effortless. We can interrogate these models to gain insights into advanced knowledge details.

- **Representative Examples**

In today's rapidly evolving academic landscape, the volume of scholarly literature has been growing exponentially, leading to an increasing demand for precise and efficient academic information retrieval. To address this challenge, various applications have emerged, including PubMed Central, ArXiv, Google Scholar, Semantic Scholar, among others. These platforms leverage NLP to enhance literature search and exploration, improve document categorization and tagging, comprehend user queries and provide relevant literature, as well as understand the semantic information within the documents. This enables them to better fulfil the task of information retrieval.

In the realm of Academic Information Retrieval, these techniques enhance information acquisition and research efficiency. The integration of advanced language models like BERT and GPT allows for a deeper understanding of scholarly articles, while precise information extraction from these articles is crucial. Challenges include the need for continuous adaptation to evolving language models, ensuring high precision in information extraction, and efficiently handling vast amounts of academic data.

3.2. Application of NLP in Medical Information Retrieval

When NLP techniques are applied in the field of medical information retrieval, they can significantly enhance the efficiency of utilizing medical data, improve clinical decision-making and patient care, and promote advancements in medical research.

- **Construction of Medical Knowledge Graphs**

One important application is the construction of medical knowledge graphs. By applying Named Entity Recognition (NER) algorithms such as Conditional Random Fields (CRF) and Recurrent Neural Networks (RNN), entities like diseases, medications, symptoms, and treatment methods can be identified. Additionally, relation extraction algorithms, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models, can analyze the semantic relationships between entities. By extracting relevant information from text using these algorithms, medical knowledge graphs can be constructed, supporting medical research, clinical decision-making, and patient care. Tianyong Hao, Likeng Liang, and others have summarized many previous methods and demonstrated the feasibility of using NLP to construct medical knowledge graphs [5].

- **Medical Question-Answering Systems**

Another significant application is medical question-answering systems, which utilize NLP techniques to provide immediate information and answer medical-related questions for patients and doctors. Text matching algorithms use similarity calculations, word vector models, or pretrained models like BERT to match user's natural language questions with medical knowledge bases. In knowledge graph-based reasoning algorithms, by processing entities, attributes, and relationships in medical knowledge graphs, logical reasoning or graph algorithms can be used to answer questions. Furthermore, information extraction algorithms can extract useful information from unstructured medical texts to obtain the required answers for specific questions [6]. In this paper, NLP is the foundation upon which the Semantic Question-Answering System "MEANS" is built, enabling it to provide precise and rapid answers in the medical domain. It empowers the extraction of meaningful insights from vast amounts of digital information, a necessity in the rapidly growing landscape of medical data.

In conclusion, the application of NLP techniques makes medical information retrieval more efficient, promoting advancements in the field of medicine. The construction of medical knowledge graphs and the development of medical question-answering systems have become critical tasks. Classic algorithms such as NER, relation extraction, text matching, and information extraction serve as the foundation for implementing these tasks. By applying these techniques, we can fully utilize the information from medical texts, improve the level of medical research and practice, and enhance the accuracy and efficiency of patient care and clinical decision-making. The application of NLP techniques in the medical field enhances the efficiency of medical information utilization, improves clinical decisions and patient care, and fosters advancements in medical research.

3.3. Application of NLP in Travel Information Retrieval

- **Travel Search and Recommendations:**

By leveraging NLP technology, systems can understand users' travel intentions and requirements. For instance, users can simply input phrases like "I want to go to a beach vacation destination" or "Find good restaurants nearby." The system, through analyzing user input, can provide personalized search and recommendation outcomes from a wealth of travel information. For example, the system can utilize collaborative filtering algorithms based on user preferences and historical data to recommend travel options that users may find interesting. It introduces an NLP-based approach to identify the multi-faceted characteristics of hotels, achieving promising results [7].

- **Voice Assistants and Intelligent Conversations:**

Travelers can engage in conversations with voice assistants, utilizing NLP technology to inquire about travel-related information. Travelers can directly ask questions like "What will the weather be like tomorrow?" or "What is the best mode of transportation to Paris?" The voice assistant can convert the voice inputs into text using speech recognition and synthesis techniques. By analyzing and comprehending the user's queries, the voice assistant can provide real-time responses and recommendations.

- **Translation and Multilingual Support:**

NLP translation technology helps travelers overcome language barriers. For instance, during their travels, travelers may need to understand and communicate destination-related information such as menus and signs. NLP translation technology can swiftly and accurately translate these pieces of information into languages familiar to the travelers, aiding their comprehension and adaptation to local culturesNLP techniques have been applied to address the challenges in translating African languages, including limited linguistic resources and diverse linguistic characteristics [8]. The MASAKHANE initiative, introduced in the paper, is a notable example of how NLP can be leveraged to bridge gaps in translation research and foster collaboration among researchers, even in regions with linguistic complexity and limited academic resources.

- **Sentiment Analysis and Review Evaluation:**

NLP technology can be employed to analyze reviews and evaluations from tourists on travel platforms, assisting service providers in understanding feedback sentiments. For example, sentiment analysis algorithms can assess the reviews, identifying positive and negative sentiments, and providing improvement suggestions to travel service providers (Figure 1). This enables travel service providers to gain insights into the quality of their services and take appropriate measures for improvement. An investigation is carried out into the application of sentiment analysis for the prediction of numerical ratings within a web-based travel diary platform [9]. This application enables users to document their experiences at various tourist attractions and provide written reviews. The textual reviews undergo a series of NLP processes, including part-of-speech (POS) tagging, rule-based phrase chunking, and dependency syntactic analysis, aimed at extracting opinion phrases from the original text, focusing on noun-adjective and noun-verb pairs. The study reports an impressive overall rating prediction accuracy of 82%.

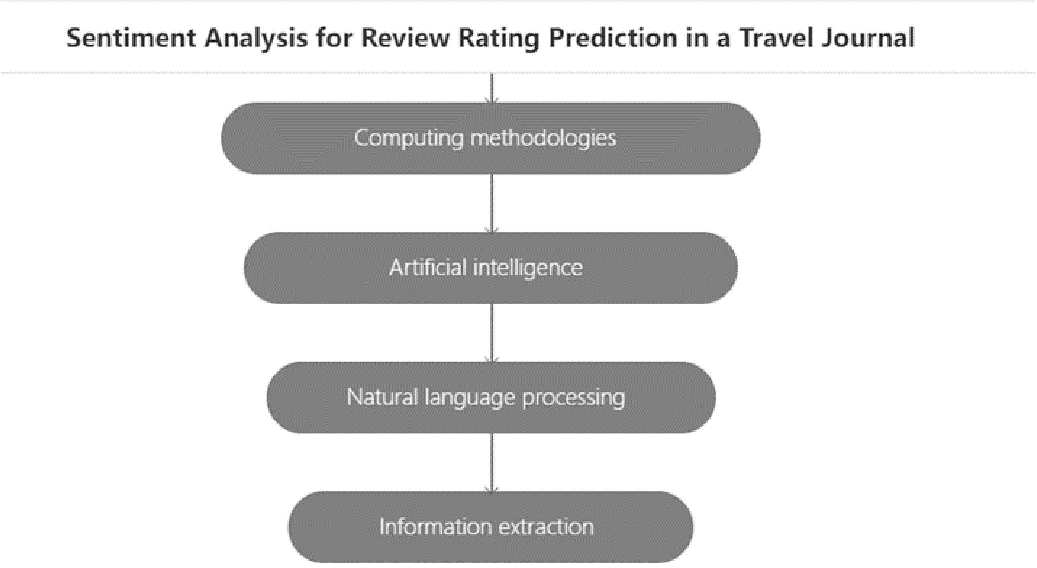
Sentiment Analysis for Review Rating Prediction in a Travel Journal
Computing methodologies
Artificial intelligence
Natural language processing
Information extraction

Figure 1. Sentiment Analysis for review rating [9].

In summary, the application of NLP technology offers travelers a more convenient and personalized travel experience. It facilitates easy travel destination search and discovery, provides local information, and offers functionalities such as multilingual support and sentiment analysis, enhancing the overall travel process and making it more enjoyable and seamless.

3.4. Application of NLP in E-Commerce Information Retrieval

NLP technology finds extensive application in the realm of e-commerce, offering numerous opportunities and advantages from personalized recommendations to sentiment analysis, customer service to market research.

- **Personalized Recommendations and Marketing**

By analyzing users' purchase history, search records, and behavior using NLP techniques, personalized product recommendations and coupons can be provided, enhancing both the conversion rate and user satisfaction. For example, collaborative filtering recommendation algorithms can analyze users' purchase history and preferences to find other users with similar interests and recommend products that these users have shown a preference for.

- **Sentiment Analysis and Brand Management**

By performing sentiment analysis on user comments and feedback, e-commerce platforms can understand the sentiment tendencies towards products and brands. This helps businesses make more precise marketing decisions and refine brand management strategies. Text classification algorithms can be trained to categorize user comments and feedback as positive, negative, or neutral, enabling quick understanding of users' attitudes and satisfaction levels towards products.

- **Analyzing User Feedback and Automated Customer Support**

By applying NLP techniques to analyze user comments and feedback on e-commerce platforms, businesses gain insights into user opinions and requirements. Integrating NLP into automated customer support systems allows for automatic handling of common inquiries, returns, and refunds, providing efficient customer support and enhancing user experience. This optimization contributes to improving products and services. Keyword extraction-based search algorithms can extract relevant keywords from product descriptions, comments, and other textual data, leading to more accurate matching of user search intent. In addition to the mentioned algorithms, there are other classic NLP algorithms used in e-commerce information retrieval, such as named entity recognition algorithms for identifying and extracting important information like product names, brands, and prices. Additionally, Syed Afeef Ahmed Shah and his colleagues present an innovative approach for the detection of e-commerce entities [10]. They employ a bidirectional LSTM model combined with a convolutional neural network (CNN) to effectively identify entities and entity groups associated with products available on the dark web. The proposed model is designed to capture intricate and comprehensive knowledge about these entities. Several experiments were conducted, comparing the model's performance with existing state-of-the-art techniques. The outcomes reveal outstanding results, achieving accuracy rates of 96.20% for the Dark Web dataset and 92.90% for the Conll-2003 dataset, surpassing the performance of other contemporary methods. By applying these algorithms and techniques to e-commerce, information retrieval, personalized recommendations, sentiment analysis, and automated customer support functionalities can be achieved, assisting businesses in better understanding user needs, optimizing products and services, and enhancing user experience.

E-Commerce Information Retrieval employs techniques like personalized product recommendations, sentiment analysis for brand management, user reviews, and automated customer service. Challenges include balancing personalized recommendations with user privacy concerns, addressing fluctuating customer sentiments, ensuring the accuracy of sentiment analysis for brand management, and the effective training and maintenance of automated customer service systems.

4. Conclusion

This paper has explored the evolution and applications of NLP in the domain of information retrieval. The historical development of NLP in this field, from keyword matching to the current era of deep learning and semantic understanding, has been discussed. Each phase brought its unique advantages and challenges, reflecting the evolving nature of user needs. NLP’s applications span across academic information retrieval, medical knowledge extraction, travel planning, and e-commerce services. These diverse applications showcase NLP’s adaptability in addressing various information retrieval tasks, from categorizing academic papers to delivering personalized recommendations.

Looking to the future, NLP in information retrieval holds great promise. Advancements in semantic understanding, multimodal integration, and personalization are expected. NLP will continue to play a vital role in meeting the challenges of the ever-expanding digital landscape, ensuring more efficient and tailored information retrieval experiences.

As NLP technology evolves and adapts, its potential to reshape how we access and interact with information is significant. With ongoing research and development, NLP is poised to define the future of information retrieval, offering enhanced convenience and relevance to users worldwide.

References

[1] Ramos J. Using tf-idf to determine word relevance in document queries, Proceedings of the first instructional conference on machine learning. 2003, 242(1): 29-48.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[3] Hakala K, Pyysalo S. Biomedical named entity recognition with multilingual BERT,oceedings of the 5th workshop on BioNLP open shared tasks. 2019: 56-61.

[4] Ben Abdessalem Karaa W, Alkhammash E H, Bchir A. Drug disease relation extraction from biomedical literature using NLP and machine learning, Mobile Information Systems, 2021: 1-10.

[5] Tianyong Hao, Zhengxing Huang, Likeng Liang, Heng Weng, Buzhou Tang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 21.10.2021.

[6] Abacha A B, Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. Information processing & management, 2015, 51(5): 570-594.

[7] Zhou X, Wang M, Li D. From stay to play–A travel planning tool based on crowdsourcing user-generated contents. Applied geography, 2017, 78: 1-11.

[8] Orife I, Kreutzer J, Sibanda B, et al. Masakhane--Machine Translation For Africa. arXiv preprint arXiv:2003.11529, 2020.

[9] Cuizon J C, Agravante C G. Sentiment analysis for review rating prediction in a travel journal, Proceedings of the 4th International Conference on NLP and Information Retrieval. 2020: 70-74.

[10] Shah S A A, Masood M A, Yasin A. Dark Web: E-Commerce Information Extraction Based on Name Entity Recognition Using Bidirectional-LSTM. IEEE Access, 2022, 10: 99633-99645.