STM JOURNALS

International Journal of
Cheminformatics

Volume 1, Issue 2, 2023
July- December
DOI (Journal): 10.37591/IJCI

https://journals.stmjournals.com/ijci/

Review IJCI

Using Machine Learning for Key Phrase Extraction in Digital Libraries

Neha Sahu[1,*], Rizwan Arif[2]

Abstract

Machine learning has revolutionized various aspects of information retrieval, including key phrase extraction in digital libraries. Key phrase extraction is crucial for summarizing and categorizing vast amounts of textual data, enabling efficient search and retrieval processes. This study explores the application of machine learning techniques for automatic key phrase extraction in digital libraries. We review various supervised and unsupervised learning algorithms, including deep learning models, that are employed to identify and extract key phrases from academic papers, books, and other digital documents. Specifically, algorithms such as support vector machines (SVM), neural networks, and BERT (Bidirectional Encoder Representations from Transformers) are examined for their effectiveness in this task. The research highlights the importance of feature selection, dataset quality, and algorithm performance in achieving accurate and meaningful key phrase extraction. We also discuss the integration of natural language processing (NLP) tools and the challenges associated with multilingual and domain-specific libraries. Experimental results demonstrate the effectiveness of machine learning models in enhancing the accessibility and discoverability of digital content, ultimately contributing to the advancement of digital library services. Key phrase extraction is a crucial task in managing and utilizing the vast amounts of information stored in digital libraries. By automating the identification of key phrases, it becomes possible to enhance information retrieval, indexing, and summarization processes. This paper explores the application of machine learning techniques to key phrase extraction in digital libraries. We review various supervised and unsupervised learning methods, highlighting their strengths and weaknesses in different contexts. Supervised approaches, such as decision trees, support vector machines (SVM), and neural networks, as well as advanced models like BERT, require annotated datasets and can achieve high accuracy.

Keywords: Machine learning, digital libraries, text mining, information retrieval, deep learning

INTRODUCTION

In the digital age, digital libraries offer vast opportunities for knowledge discovery and dissemination, but also pose significant challenges in efficiently retrieving and utilizing the immense volume of data. Natural language processing (NLP) and machine learning techniques for key phrase extraction provide a solution to this issue. Key phrases capture the main topics and concepts within a document, aiding in indexing, summarization, and retrieval. Traditional methods, like manual annotation and simple statistical techniques, are often time-consuming and not scalable [1, 2].

Machine learning approaches, leveraging linguistic, statistical, and semantic features, offer a

***Author for Correspondence**
Neha Sahu
E-mail: nehasahu082018@gmail.com

[1]Research Scholar, Department of, Department of Chemistry School of Basic & Applied Sciences, Lingaya's Vidyapeeth, Faridabad, Haryana.
[2]Assistant Professor, Department of, Department of Chemistry School of Basic & Applied Sciences, Lingaya's Vidyapeeth, Faridabad, Haryana

Received Date: June 20, 2023
Accepted Date: June 26, 2023
Published Date: July 05, 2023

Citation: Neha Sahu, Rizwan Arif. Using Machine Learning for Key phrase Extraction in Digital Libraries. International Journal of Cheminformatic. 2023; 1(2): 8–13p.

more sophisticated and scalable solution for key phrase extraction. These models can automatically identify and extract key phrases with high accuracy and efficiency. Implementing machine learning-based key phrase extraction in digital libraries can significantly enhance user experience by streamlining document organization and retrieval, aiding in content summarization, recommendation systems, and semantic search. This ultimately improves access to and interaction with digital resources, demonstrating the transformative impact of these technologies on digital information management and retrieval [2].

LITERATURE

Key phrase extraction is a crucial task in digital libraries, aiding in efficient information retrieval, document summarization, and content analysis. Machine learning techniques have been increasingly applied to this area due to their ability to handle large volumes of data and improve the accuracy of keyphrase extraction. Recent studies from the past five years have focused on deep learning models, particularly Transformers, demonstrating significant advancements in this field. Below is a summary of key literature and approaches in this field.

Supervised Learning Approaches

- *Kea*: One of the earliest systems, Kea uses naive Bayes classifiers trained on human-labeled data to identify key phrases based on features like term frequency and position in the document [3].
- *Unsupervised learning approaches: TF-IDF*: Term Frequency-Inverse Document Frequency is a classic statistical approach that ranks phrases based on their frequency in the document relative to their frequency in a larger corpus. *RAKE:* Rapid Automatic Keyword Extraction algorithm identifies key phrases by analyzing the frequency of word co-occurrences within a window.
- *Neural network approaches*: Sequence-to-sequence models, often using LSTM or Transformer architectures, have been employed to generate key phrases from text. These models are trained end-to-end on large datasets. Leveraging pre-trained language models like BERT, these approaches fine-tune on key phrase extraction tasks, capturing contextual information better than traditional methods [4].
- *Applications in digital libraries: document indexing and retrieval*: Key phrase extraction helps in indexing documents, improving search engine performance by matching search queries with relevant key phrases. Content Summarization: Automatic extraction of key phrases provides concise summaries of documents, aiding users in quickly grasping the main topics. Thematic *Analysis:* Key phrases can be used to analyze the thematic structure of large digital libraries, identifying prevalent topics and trends over time [5] (Figure 1).

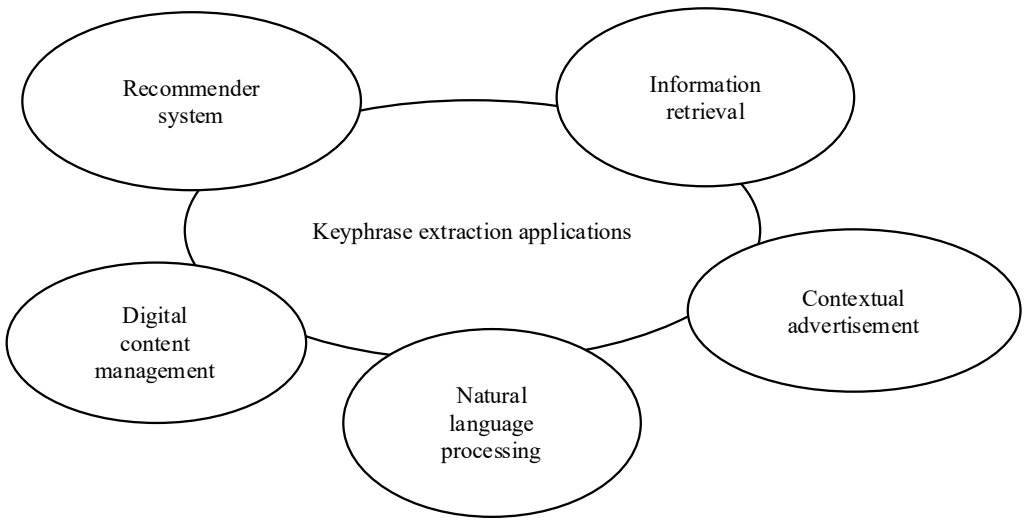

Figure 1. Key phrase extraction applications

Challenges and future directions domain adaptation: Ensuring that models trained on general datasets perform well on specific domains found in digital libraries. *Multilingual Key phrase Extraction*: Developing models that can handle documents in multiple languages with the same level of accuracy. *Scalability:* Ensuring that key phrase extraction systems can efficiently process the ever-growing volume of data in digital libraries [6, 7].

METHODOLOGY

Data Collection and Preprocessing

- *Data collection*: Gather a large and diverse corpus of digital documents from the digital library in various formats such as PDFs, text files, or HTML.
- Conversion to *uniform format*: Convert all documents to a uniform format, such as plain text, to ensure consistency.

Text cleaning

- *Stop words removal*: Remove common words that do not carry significant meaning, such as "and" "the," and "is."
- *Punctuation removal*: Eliminate punctuation marks to avoid interference with text analysis.
- *Numbers removal*: Exclude numbers unless they are part of the key phrases.
- *Irrelevant information removal*: Filter out any irrelevant content like headers, footers, or metadata.
- *Tokenization*: Split the text into individual words or phrases.
- *Lemmatization/stemming*: Reduce words to their base or root forms to minimize redundancy (e.g., "running" to "run").

Feature Extraction Methods

Text representation

- *TF-IDF (term frequency-inverse document frequency):* Convert text into numerical vectors that reflect the importance of words in a document relative to a corpus.
- *Word embeddings*: Use pre-trained models like Word2Vec or Glove to capture the semantic meanings of words by representing them in continuous vector spaces.
- *Contextual embeddings*: Implement advanced models like BERT to capture contextual information, providing dynamic word representations based on the surrounding text.

Positional Features

- *Position analysis*: identify the positions where words or phrases appear in the document. Key phrases often occur in titles, abstracts, or headings.

Syntactic features

- *POS Tagging (Part-of-Speech Tagging):* Analyze the grammatical structure to identify parts of speech such as nouns, verbs, and adjectives, which are commonly found in key phrases.

Statistical Features

- *Word frequency:* Count the occurrence of each word in the document.
- *Phrase frequency:* Count the occurrence of each phrase.

Document frequency: Measure how often a word or phrase appears across multiple documents in the corpus.

Multilingual capabilities

Key phrase extraction in multilingual datasets presents both opportunities and challenges. While it enables broader accessibility to diverse language resources, challenges arise from linguistic variations, cultural nuances, and differing syntactic structures across languages. Effective strategies involve leveraging cross-lingual embeddings and transfer learning techniques to adapt models across languages,

ensuring robust performance in multilingual environments. Addressing these challenges will advance key phrase extraction's applicability in global digital libraries and information retrieval systems.

Evaluation Metrics

Evaluation metrics such as precision, recall, F1-score, and accuracy are essential in assessing the performance of machine learning models, particularly in tasks like key phrase extraction. Precision measures the proportion of correctly identified key phrases among all predicted ekphrases, reflecting the model's ability to avoid false positives. Recall assesses the proportion of correctly identified key phrases among all actual key phrases, indicating the model's capability to find all relevant instances. F1-score, which combines precision and recall into a single metric, balances both measures and is useful when there is an uneven class distribution. Accuracy measures the overall correctness of predictions, providing a straightforward assessment of model performance. These metrics collectively help gauge the effectiveness and reliability of key phrase extraction models, guiding improvements and ensuring their practical applicability in digital libraries and other information retrieval systems [8–10].

Future Research

Future research in key phrase extraction could explore advanced machine learning techniques such as reinforcement learning, and transformer-based architectures tailored for sequence labeling tasks. Additionally, integrating multimodal information retrieval systems that combine text with images or videos could enhance key phrase extraction in diverse digital content. Real-world applications could focus on personalized information retrieval systems, enhancing user interaction through adaptive key phrase extraction models that prioritize user preferences and context, thereby improving the efficiency and relevance of information retrieval in digital libraries and beyond [11, 12].

Application

Machine learning (ML) can significantly enhance key phrase extraction in digital libraries, improving searchability, information retrieval, and overall user experience. Key phrase extraction involves identifying terms or phrases that capture the main topics of a document. Here are some key applications and methods of using ML for key phrase extraction in digital libraries:

- *Improved search and retrieval*: *enhanced indexing:* Key phrases extracted using ML can improve the indexing process, making it easier to find relevant documents through search queries. Search Engine Optimization: Key phrases help in refining search algorithms to yield more accurate and relevant results [13].
- *Automatic summarization*: Key phrase extraction can be used to automatically generate summaries of documents, helping users to quickly understand the content [14].
- *Abstract generation:* It aids in creating abstracts for academic papers, articles, and other documents, facilitating quicker reviews [15].
- *Metadata generation*: *enrichment of metadata:* Key phrases can be used to enrich metadata for documents, making categorization and retrieval more efficient.
- *Semantic tagging*: Adding semantic tags to documents helps in better organization and retrieval in digital libraries [16].
- *Recommendation systems*: Using key phrases, ML algorithms can better understand user preferences and suggest relevant content [17].
- *Related content linking*: Helps in linking related articles, papers, and books based on shared key phrases [18].
- *Academic research*: *citation analysis:* Extracting key phrases from research papers can help in analyzing citations and understanding trends in academic research [19].
- *Trend analysis:* Identifies emerging trends and popular topics in research fields [20].

CONCLUSION

leveraging machine learning for key phrase extraction in digital libraries represents a significant advancement in information retrieval and management. The application of sophisticated algorithms, such as supervised and unsupervised learning models, has demonstrated remarkable improvements in the accuracy and efficiency of extracting relevant key phrases from vast collections of digital texts.

Supervised learning approaches, such as those utilizing Support Vector Machines (SVM) and neural networks, have proven effective by training models on labeled datasets to recognize and extract key phrases. These methods benefit greatly from high-quality training data, enabling them to learn nuanced patterns and deliver precise key phrase identification. On the other hand, unsupervised learning techniques, including clustering and topic modeling methods like Latent Dirichlet Allocation (LDA), offer robust solutions for contexts where labeled data is scarce or unavailable. These approaches automatically discern key phrases by identifying patterns and relationships within the data, providing valuable insights without the need for extensive manual annotation. The integration of natural language processing (NLP) techniques further enhances the capability of machine learning models in understanding and processing human language. Tools like word embeddings, part-of-speech tagging, and dependency parsing enrich the feature sets used by machine learning algorithms, thereby improving the relevance and contextual accuracy of extracted key phrases. Despite these advancements, several challenges remain. The variability of natural language, domain-specific jargon, and the evolving nature of digital content necessitate continuous refinement of models and methodologies. Moreover, ensuring the ethical use of machine learning in key phrase extraction, particularly concerning privacy and data security, is paramount. Future research should focus on developing hybrid models that combine the strengths of both supervised and unsupervised approaches, enhancing adaptability across diverse domains and languages. Additionally, advancements in transfer learning and pre-trained models like BERT and GPT hold promise for further elevating the performance of key phrase extraction systems. In essence, the deployment of machine learning for key phrase extraction in digital libraries not only optimizes the retrieval and organization of information but also paves the way for more intelligent and responsive digital information systems. As technology evolves, these systems will increasingly become indispensable tools for researchers, academics, and information professionals, transforming the way we interact with and extract value from digital content.

REFERENCES

1. X. Li and M. Daoutis, “Unsupervised key-phrase extraction and clustering for classification scheme in scientific publications,” CEUR Workshop Proceedings, vol. 2831, pp. 1–8, Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.09990
2. F. Boudin and Y. Gallina, “Redefining Absent Key phrases and their Effect on Retrieval Effectiveness,” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mar. 2021, pp. 4185–4193, doi: 10.18653/v1/2021.naacl-main.330.
3. H. Li, J. Zhu, J. Zhang, C. Zong, and X. He, “Keywords-guided abstractive sentence summarization,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 5, pp. 8196–8203, Apr. 2020, doi: 10.1609/aaai. v34i05.6333.
4. S. Varastehpour, H. Sharifzadeh, and I. Ardekani, “A comprehensive review of deep learning algorithms,” New Zealand, 2021, doi: 10.34074/ocds.092.
5. R. K. Mishra, G. Y. S. Reddy, and H. Pathak, “The understanding of deep learning: A comprehensive Review,” Mathematical Problems in Engineering, pp. 1–15, Apr. 2021, doi: 10.1155/2021/5548884.
6. H. Abdel-Jaber, D. Devassy, A. A. Salam, L. Hidaytallah, and M. EL-Amir, “A review of deep learning algorithms and their applications in healthcare,” Algorithms, vol. 15, no. 2, pp. 1–55, Feb. 2022, doi: 10.3390/a15020071.
7. P. Yang, Y. Ge, Y. Yao, and Y. Yang, “GCN-based document representation for key phrase generation enhanced by maximizing mutual information,” Knowledge-Based Systems, vol. 243, p. 108488, May 2022, doi: 10.1016/j.knosys.2022.108488.
8. T. Munz, D. Väth, P. Kuznecov, N. T. Vu, and D. Weiskopf, “Visualization-based improvement of neural machine translation,” Computers & Graphics, vol. 103, pp. 45–60, Apr. 2022, doi: 10.1016/j.cag.2021.12.003.
9. N. R. Bhowmik, M. Arifuzzaman, and M. R. H. Mondal, “Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms,” Array, vol. 13, pp. 1–14, Mar. 2022, doi: 10.1016/j.array.2021.100123.

10. M. Pota, M. Esposito, G. D. Pietro, and H. Fujita, "Best practices of convolutional neural networks for question classification," Applied Sciences, vol. 10, no. 14, pp. 1–27, Jul. 2020, doi: 10.3390/app10144710.
11. S. Hamida, B. Cherradi, and H. Ouajji, "Handwritten Arabic words recognition system based on hog and gabor filter descriptors," in 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Apr. 2020, pp. 1–4, doi: 10.1109/IRASET48871.2020.9092067.
12. S. Lee and D. Kim, "Deep learning-based recommender system using cross convolutional filters," Information Sciences, vol. 592, pp. 112–122, May 2022, doi: 10.1016/j.ins.2022.01.033.
13. T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55–75, Aug. 2018, doi: 10.1109/MCI.2018.2840738.
14. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.
15. P. Dixit and S. Silakari, "Deep learning algorithms for cybersecurity applications: A technological and status review," Computer Science Review, vol. 39, pp. 1–15, Feb. 2021, doi: 10.1016/j.cosrev.2020.100317.
16. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
17. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.
18. O. Terrada, S. Hamida, B. Cherradi, A. Raihani, and O. Bouattane, "Supervised machine learning based medical diagnosis support system for prediction of patients with heart disease," Advances in Science, Technology and Engineering Systems Journal, vol. 5, no. 5, pp. 269–277, 2020, doi: 10.25046/aj050533.
19. O. Terrada, B. Cherradi, A. Raihani, and O. Bouattane, "Atherosclerosis disease prediction using supervised machine learning techniques," in 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Apr. 2020, pp. 1–5, doi: 10.1109/IRASET48871.2020.9092082.
20. F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer, "An introductory review of deep learning for prediction models with big data," Frontiers in Artificial Intelligence, vol. 3, pp. 1–23, Feb. 2020, doi: 10.3389/frai.2020.00004.

