titulo

Remigio Hurtado[1]

Universidad Politécnica Salesiana, Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador.
rhurtadoo@ups.edu.ec
ups.edu.ec

Abstract.

Keywords:

1 Introduction

This paper presents a comprehensive methodology for the development of survey recommendation systems using Natural Language Processing (NLP) and Machine Learning (ML) techniques. The approach ensures systematic analysis and generation of knowledge representations from scientific literature.

2 Methodology

Our methodology is designed to ensure a comprehensive and systematic analysis of scientific literature. The methodology used in this study builds on our previous work, including a methodology that leverages machine learning and natural language processing techniques [?], and a novel method for predicting the importance of scientific articles on topics of interest using natural language processing and recurrent neural networks [?]. The process is structured into three phases: data preparation, topic modeling, and the generation and integration of knowledge representations. Each phase is essential for transforming raw text data into meaningful insights, and the detailed parameters and algorithm are explained below. The table **??** describes the parameters for understanding the overall process and algorithm. The high-level process is presented in Fig. **??**, and the detailed algorithm is outlined in Table **??**.

In the data preparation phase, we focus on extracting and cleaning text from scientific documents using natural language processing (NLP) techniques. This phase involves several steps to ensure that the text data is ready for analysis. First, text is extracted from the documents, and non-alphabetic characters that do not add value to the analysis are removed. Next, the text is converted to lowercase, and stopwords (common words that do not contribute much meaning) are removed. We then apply lemmatization, which transforms words to their base form (e.g., "running" becomes "run"). Each document is tokenized (split into individual words or terms), and n-grams (combinations of words) are identified to find common terms. We generate a unified set of common terms, denoted as TE, which includes both the terms extracted from the documents and basic terms relevant to any field of study, such as [Fundamentals, Evaluation of Solutions, Trends]. Finally, each document is vectorized with respect to TE, resulting in the Document-Term Matrix (DTM).

The DTM is a crucial component for topic modeling. It is a matrix where the rows represent the documents in the corpus, and the columns represent the terms (words or n-grams) extracted from the corpus. Each cell in the matrix contains a value indicating the presence or frequency of a term in a document. This structured representation of the text data allows us to apply machine learning techniques to uncover hidden patterns.

In the topic modeling phase, we use Latent Dirichlet Allocation (LDA), a popular machine learning technique for identifying topics within a set of documents. By applying LDA to the DTM, we transform the matrix into a space of topics. Specifically, LDA provides us with two key matrices: the Topic-Term Matrix (β) and the Document-Topic Matrix (θ). The Topic-Term Matrix (β) indicates the probability that a term is associated with a specific topic, while the Document-Topic Matrix (θ) indicates the probability that a document belongs to a specific topic.

To enhance this approach, we integrate predefined topic representations and refine document-topic assignments. In addition to extracting topics purely from the document-term matrix (DTM), we incorporate a set of predefined topics, namely [Fundamentals, Evaluation of Solutions, Trends], which are represented using a predefined set of keywords generated by a language generation model. This ensures that the model captures both the inherent structure of the dataset and domain-relevant themes. The predefined topics are processed through a keyword extraction function, which selects the most relevant words associated with each topic based on the provided corpus.

To construct a more robust topic representation, we create predefined topic matrices that integrate predefined topic vectors with the most relevant keywords extracted for each topic. These matrices are then normalized and combined with the LDA-generated topic distribution, ensuring a refined alignment of document-topic assignments. By doing so, we mitigate the limitations of purely unsupervised topic modeling, which might generate topics that lack semantic clarity.

Finally, to improve interpretability, we assign meaningful names to the discovered topics by combining the highest-probability terms from the topic-term matrix (β). Using a language generation model, we generate concise and descriptive topic names, ensuring that each topic is easily understandable. This process results in a refined set of topics (T), the most relevant terms (K), and a topic-term graph (G_{tk}) that illustrates relationships between topics and key terms. This methodology enhances the effectiveness of topic modeling by integrating both machine learning and domain-specific knowledge, leading to a more structured and meaningful representation of the analyzed corpus.

The final phase involves the generation and integration of knowledge representations, which include summaries, keywords, and interactive visualizations. For each topic t in T, we identify the N most relevant documents—those with the highest probabilities in θ. This set, D_t, represents the documents most closely related to each topic. We then generate summaries of these documents, each with a maximum of W_b words, using a language generation model. These summaries include references to the most relevant documents, which are added to the set R_d if they are not already included. We also integrate all the summaries and generate J suggested keywords using a language generation model. Additionally, we create interactive graphs from the G_{tk} graph, showcasing nodes and relationships between the topics and terms, and highlighting significant connections.We also integrate all the summaries and generate J suggested keywords using a language generation model. Additionally, the topic titles were improved based on the generated summaries using a language generation model, ensuring that the final topic names more accurately reflect the summarized content.

This methodology provides a clear, structured approach to analyzing scientific literature, leveraging advanced NLP and machine learning techniques to generate useful and comprehensible knowledge representations. This is the methodology used in the development of this study, which presents the fundamentals, solution evaluation techniques, trends, and other topics of interest within this field of study. This structured approach ensures a thorough review and synthesis of the current state of knowledge, providing valuable insights and a solid foundation for future research.

Table 1. Description of parameters of the proposed methodology

Parameter	Description
T	Set of topics to be generated with the LDA model.
$\#T$	Cardinality of T. That is, the number of topics (dimensions).
K	Set of common terms with the highest probability in topic t used to label that topic.
$\#K$	Cardinality of K.
W_t	Maximum number of words for combining the common terms of topic t into a new topic name.
N	Number of the most relevant documents to generate a summary of topic t.
W_b	Maximum number of words to generate a summary of the list of N most related documents to topic t.
J	Number of suggested keywords to be generated as knowledge representation.
W_k	Number of keywords selected by language generation model from the LDA-generated words to construct the predefined matrix.

Figures/method.png

Fig. 1. Methodology for the generation of knowledge representations

General Algorithm

Input: Field of study, Scientific articles collected from virtual libraries, $\#T$, $\#K$, W_t, N, W_b, J, W_k

Phase 1: Data Preparation with Natural Language Processing (NLP)

a. Extract of text from documents.
b. Remove non-alphabetic characters that do not add value to the analysis.
c. Convert to lowercase and remove stopwords.
d. Apply lemmatization to transform words to their base form.
e. In each document, tokenize and identify n-grams to identify common terms (words or n-grams).
f. Unified generation of common terms of all documents. Where TE is the unified set of common terms.
g. Add in TE the basic terms for any field of study, such as: [Fundamentals, Evaluation of Solutions, Trends].
h. Vectorize each document with respect to TE and generate Document-Term Matrix (DTM).
Output: For each document [title, original text, common terms], and Document-Term Matrix (DTM)

Phase 2: Topic Modeling whit Machine Learning

Input: Document-Term Matrix (DTM)
a. Apply Latent Dirichlet Allocation (LDA) to transform the DTM Matrix into a space of $\#T$ topics (dimensions).
b. Obtain the Topic-Term Matrix (β) that indicates the probability that a term is generated by a specific topic.
c. Obtain the Document-Topic Matrix (θ) that indicates the probability that a document belongs to a specific topic.
d. Generate predefined topic representations using language generation model by sending the LDA-generated words and selecting W_k keywords for each predefined topic.
e. Construct predefined matrices by combining predefined topic vectors with the W_k selected keywords.
f. Normalize and combine the predefined matrices with the LDA-generated topic distribution to refine document-topic assignments, resulting in the new Matrix (β).
g. For each unknown t topic in β, assign a name or label to the t topic by combining the $\#K$ highest probability common terms in β associated with that t topic. This generates:
The T Set with the $\#T$ most relevant topics.
The K Set with the $\#K$ most relevant terms.
The G_{tk} Graph of the relationships between the most relevant topics (T) and the most relevant terms (K).
h. For each topic t in T, modify topic t by combining the common terms of t into a new topic name with at most W_t words using a language generation model.
Output: Relevant Topics (T), Topic-Term Matrix (β), Document-Topic Matrix (θ)

Phase 3: Generation and Integration of Knowledge Representations

Input: Relevant Topics T, Document-Topic Matrix (θ)

3.1: Knowledge representations through summaries and keywords
a. For each t topic in T, obtain its N most relevant documents, i.e., those with the highest probabilities in θ. Thus, D_t represents the set of documents most related to each topic t.
b. For each topic t in T, and from D_t generate a summary of the list of documents most related to that topic t with at most W_b words using a language generation model.
c. Incorporate into the summary the text citing references to the most relevant documents. Add these references to the set R_d, which will contain all cited references, including new references if they have not been previously included.
d. Integrate all the summaries and from them generate J suggested keywords using a language generation model.
e. Improve topic titles with a language generation model.

3.2: Knowledge representations through knowledge visualizations with interactive graphs
a. From the G_{tk} graph, generate an interactive graph with nodes and relationships between the topics T and the K most relevant terms.
b. Highlight the connections between the topics T and the K most relevant terms.

3.3: Integration of Knowledge Representations

Output: Knowledge Representations

Fig. 2. General algorithm of the methodology incorporating natural language processing, machine learning techniques and language generation models

3 Conclusion

- Our work introduces an innovative approach to scientific literature analysis by combining natural language processing (NLP) and machine learning. One of its main contributions is the improvement of topic modeling by integrating Latent Dirichlet Allocation (LDA) with predefined topic representations generated by a language model.
- This combination captures both the inherent structure of the data and domain-specific knowledge, improving the coherence and interpretability of the generated topics. Additionally, we incorporate predefined topic matrices and normalize them with the LDA-generated distributions, ensuring a more precise and contextually aligned topic assignment.
- Another key contribution of our study is the automated generation of summaries and keywords based on the identified topics. For each topic, we select the most relevant documents and generate concise summaries using language generation models, facilitating the synthesis of key information.
- Additionally, we create an interactive graph that illustrates the relationships between topics and their most relevant terms. This visualization enables an intuitive exploration of the extracted knowledge structure, enhancing the understanding of connections between the analyzed concepts.
- Finally, our project optimizes knowledge generation through a three-phase structure: data preparation, topic modeling, and the generation of knowledge representations. By integrating advanced NLP and machine learning techniques, our work establishes a solid foundation for future research in the automation of scientific literature analysis.

This paper presents a robust methodology for creating survey recommendation systems. By integrating NLP and ML techniques, we ensure a systematic and comprehensive analysis, leading to high-quality knowledge representations and user-friendly reports.

References

1. Hurtado, Remigio; Picón, Cristian; Muñoz, Arantxa; Hurtado, Juan. "Survey of Intent-Based Networks and a Methodology Based on Machine Learning and Natural Language Processing." In Proceedings of Eighth International Congress on Information and Communication Technology. Springer Nature Singapore, Singapore, 2024.
2. Park, Keunheung; Kim, Jinmi; Lee, Jiwoong. "Visual Field Prediction using Recurrent Neural Network," *Scientific Reports*, vol. 9, no. 1, p. 8385, 2019, https://doi.org/10.1038/s41598-019-44852-6.
3. Xu, M.; Du, J.; Guan, Z.; Xue, Z.; Kou, F.; Shi, L.; Xu, X.; Li, A. "A Multi-RNN Research Topic Prediction Model Based on Spatial Attention and Semantic Consistency-Based Scientific Influence Modeling," *Comput Intell Neurosci*, vol. 2021, 2021, p. 1766743, doi: 10.1155/2021/1766743.
4. Kreutz, Christin; Schenkel, Ralf. "Scientific Paper Recommendation Systems: a Literature Review of recent Publications," 2022/01/03.

5. Hurtado, R., et al. "Survey of Intent-Based Networks and a Methodology Based on Machine Learning and Natural Language Processing." International Congress on Information and Communication Technology. Singapore: Springer Nature Singapore, 2023.
6. Lopez, A., Dutan, D., Hurtado, R. "A New Method for Predicting the Importance of Scientific Articles on Topics of Interest Using Natural Language Processing and Recurrent Neural Networks." In: Yang, X.S., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Ninth International Congress on Information and Communication Technology. ICICT 2024 2024. Lecture Notes in Networks and Systems, vol 1013. Springer, Singapore. https://doi.org/10.1007/978-981-97-3559-4_50.