International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

Unsupervised Document Summarization Using Graph-based Methods

Akshata Upadhye

Data Scientist

Abstract: ***Document summarization plays a vital role in Natural Language Processing (NLP), which is used to condense extensive textual content into concise versions while preserving the essential information. This paper offers a comprehensive survey of unsupervised document summarization techniques, with a particular emphasis on graph-based methods. We explore the foundational principles of graph-based summarization and illustrate how documents are represented as nodes in a graph which is utilized for extraction of crucial sentences based on their centrality and connectivity within the graph structure. Furthermore, we also conduct a comparative analysis between graph-based methods and traditional approaches like TF-IDF and Latent Dirichlet Allocation (LDA) to identify their respective strengths and limitations. Through this exploration, we aim to provide insights into the efficacy of graph-based techniques in document summarization and guide future research endeavors in this domain.***

Keywords: Document summarization, Graph-based methods, TF-IDF, Latent Dirichlet Allocation (LDA), Natural Language Processing

1. Introduction

The substantial increase of digital content in recent years has resulted in an unprecedented growth in the volume of unstructured information that overwhelms the users and requires efficient methods for document summarization. Traditional supervised approaches to document summarization require large amounts of labeled data for training which poses a significant challenge in terms of cost and scalability. In contrast, unsupervised methods present an lucrative alternative, as they do not require labeled data and can automatically extract key information from the text documents.

This survey focuses on unsupervised document summarization techniques, with a particular emphasis on graph-based approaches. Graph-based methods offer a promising avenue for summarization by representing documents as nodes in a graph and leveraging the relationships between the nodes to identify the most relevant content. Unlike supervised methods that rely on annotated data, graph-based techniques utilize the inherent structural patterns and semantic connections within the document corpus to generate concise summaries.

By focusing on graph-based approaches, this survey aims to provide a comprehensive overview of the principles, method- ologies, and applications of unsupervised document summarization. In this paper the fundamental concepts behind graph- based summarization, including graph construction, centrality and connectivity measures, and summarization algorithms are explored. Additionally, the survey presents a comparison of the performance of graph-based methods with traditional super- vised techniques to highlight their advantages and limitations. This survey, intends to offer insights into the state-of-the-art in unsupervised document summarization and identify potential avenues for future research. By understanding the capabilities and challenges of graph-based approaches, researchers and practitioners can make informed decisions when selecting summarization techniques for various applications across do- mains such as information retrieval, document clustering, and content recommendation.

Graph-Based Document Representation

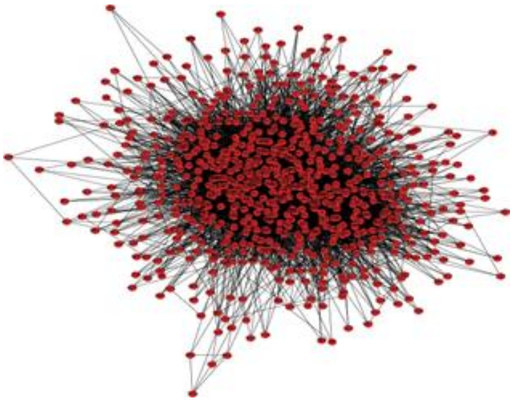

Figure 1: Graph-based Document Representation

Graph-based document representation methods have emerged as powerful tools for document summarization, offering a versatile framework for capturing semantic relationships between documents. In these approaches, documents are represented as nodes in a graph, with edges between nodes encoding the semantic connections between documents [1] as shown in 1. This graph-based representation facilitates the extraction of important sentences by analyzing the centrality and connectivity of nodes within the graph.

a) Graph Construction Techniques

- One of the critical aspects of graph-based document summarization is the construction of the underlying graph structure for the documents. Various techniques can be utilized to create this graph, by capturing different aspects of document semantics.
- Similarity Graphs: In similarity-based graph construction, documents are connected based on their semantic similarity. Some of the commonly used similarity measures include cosine similarity, Jaccard similarity, and Euclidean distance. Documents with higher semantic similarity are linked by edges in the graph, forming clusters of related content [2].

- Topic Graphs: Another approach is to construct a topic graph, where nodes represent topics or themes extracted from the document collection. Edges between nodes indicate the presence of shared terms or concepts across topics. Topic graphs provide a higher-level abstraction of document semantics and can facilitate the identification of overarching themes within the corpus [3].
- Hybrid Approaches: Some methods combine multiple sources of information, such as textual similarity and topic modeling, to construct hybrid graphs. These approaches aim to leverage complementary aspects of document semantics to enhance summarization performance.

b) *Application in Document Summarization*

- Graph-based document representation offers a flexible framework for summarization, capable of capturing complex semantic relationships between the documents. By leveraging graph-based techniques, summarization algorithms can identify important sentences or clusters of related content within the document corpus. These approaches enable the generation of concise yet informative summaries that preserve the key information contained in the original documents.

2. Centrality and Connectivity Measures

In graph-based document summarization, centrality and connectivity measures serve as fundamental concepts used for identifying important nodes and extracting relevant informa- tion. These measures are used to quantify the significance of nodes within the graph structure and their roles in facilitating connections between different parts of the document corpus.

a) *Centrality Measures*

- Centrality measures evaluate the importance of a node in a graph based on its connections to other nodes. These measures help identify nodes that play central roles in the network and are often indicative of key documents or sentences within the document collection [4]. Common centrality measures utilized in document summarization include:
- Degree Centrality: Degree centrality is a simple yet effective measure that quantifies the number of edges incident to a node [5]. Nodes with higher degree centrality are those that have a greater number of connections to other nodes. In the context of document summarization, nodes with high degree centrality typically correspond to sentences that are highly connected to other sentences in terms of semantic similarity.
- PageRank: PageRank is a widely used centrality measure that evaluates the importance of a node based on the structure of the entire graph [6]. Originally developed by Google to rank web pages, PageRank assigns higher scores to nodes that are connected to by other high- ranking nodes. In document summarization, PageRank can identify sentences that are authoritative or central to the overall semantic structure of the corpus.
- Betweenness Centrality: Betweenness centrality quantifies the extent to which a node lies on the shortest paths between other nodes in the graph [7]. Nodes with high betweenness centrality act as bridges or bottlenecks within the network, facilitating communication between disparate parts of the graph. In document summarization, nodes with high betweenness centrality may represent sentences that bridge different topics or themes within the corpus.

b) *Connectivity Measures*

- Connectivity measures evaluate the importance of a node based on its role in connecting different parts of the graph. These measures are particularly relevant for identifying nodes that serve as bridges between distinct clusters or communities within the corpus. Common connectivity measures include:
- Closeness Centrality: Closeness centrality quantifies how close a node is to all other nodes in the graph, taking into account the length of the shortest paths between them. Nodes with high closeness centrality are those that are geographically close to other nodes, making them efficient communicators within the network. In document summarization, nodes with high closeness centrality may represent sentences that are centrally located in terms of semantic similarity to other sentences.

c) *Role in Document Summarization*

- Centrality and connectivity measures play a crucial role in identifying the most relevant information for summarization by highlighting the key documents or sentences within the document corpus. By leveraging these measures, summarization algorithms can extract important content that captures the underlying semantic structure and themes of the document col- lection. Centrality measures identify documents or sentences that are central to the semantic network, while connectivity measures identify documents that bridge different parts of the corpus, facilitating the creation of concise and informative summaries.

3. Comparative Analysis

In this section, we conduct a comparative analysis between graph-based summarization methods and traditional techniques such as TF-IDF and LDA. TF-IDF (Term Frequency- Inverse Document Frequency) measures the importance of a word in a document relative to the entire collection of documents, while LDA (Latent Dirichlet Allocation) represents documents as mixtures of topics. We evaluate the performance of each approach in terms of summarization quality, scalability, and interpretability.

a) *TF-IDF*

- TF-IDF is a widely used statistical measure that evaluates the importance of a term in a document relative to a collection of documents [8]. It computes a weight for each term based on its frequency in the document (TF) and its rarity across the document collection (IDF). TF-IDF is commonly used in extractive summarization approaches, where sentences containing important terms are selected for inclusion in the summary.
- Strengths: TF-IDF is simple to implement and computationally efficient, making it suitable for large

document collections. It effectively identifies important terms that are discriminative for a specific document while discounting common terms that appear frequently across the entire collection.
- Weaknesses: TF-IDF relies solely on term frequencies and does not capture the semantic relationships between terms or documents. Hence, it may overlook important context or nuances present in the document corpus, leading to suboptimal summarization quality.

b) *Latent Dirichlet Allocation (LDA)*

- LDA is a probabilistic topic model that represents documents as mixtures of topics [9]. It assumes that each document is generated from a mixture of underlying topics, and each topic is characterized by a distribution over terms. LDA is commonly used in topic modeling and has been adapted for document summarization by selecting representative sentences from the dominant topics in each document.
- Strengths: LDA captures the latent thematic structure of the document corpus, enabling the identification of coherent topics and themes. It provides a probabilistic framework for summarization, allowing for uncertainty quantification and model interpretation.
- Weaknesses: LDA requires careful tuning of hyperparameters and may be sensitive to the choice of the number of topics. It assumes a bag-of-words representation of documents, neglecting the structural information present in the text. Additionally, LDA-based summarization approaches may struggle with documents containing diverse or overlapping topics.

c) *Graph-based Summarization Methods*

- Graph-based summarization methods leverage the semantic relationships between documents to identify important sentences for summarization. These methods represent documents as nodes in a graph, where edges capture semantic connections between documents. Centrality and connectivity measures are then used to identify the most important information within the graph structure.
- Strengths: Graph-based methods capture the complex semantic relationships between documents, enabling the extraction of coherent and informative summaries. There- fore these methods can handle documents with diverse topics or overlapping themes by leveraging the global structure of the document corpus.
- Weaknesses: Graph-based summarization methods may require substantial computational resources for graph construction and analysis, particularly for large document collections. They may also be sensitive to the choice of graph construction parameters and similarity measures, leading to variability in summarization performance.

4. Applications and Challenges

Graph-based methods for document summarization offer a versatile framework with a wide range of applications across various domains. However, several challenges persist in the development and deployment of these techniques, affecting their effectiveness and scalability.

a) *Applications*

- Graph-based methods for document summarization find applications in several domains, including:
- Information Retrieval: Graph-based summarization techniques can enhance information retrieval systems by providing users with concise and informative summaries of large document collections. By leveraging semantic relationships between documents, these methods enable more accurate and efficient retrieval of relevant information.
- Document Clustering: Graph-based summarization can facilitate document clustering tasks by identifying clusters of related documents based on relevant information and their semantic similarity. By summarizing each document, these methods provide a comprehensive representation of the main topics within the document collection, aiding in cluster interpretation and analysis.
- Content Recommendation: Graph-based summarization techniques can be employed in content recommendation systems to generate personalized summaries of documents or articles tailored to the preferences and interests of individual users. By analyzing the semantic relation- ships between documents and users interaction patterns, these methods can deliver relevant and engaging content recommendations.

b) *Challenges*

- Despite their potential applications, graph-based methods for document summarization face several challenges:
- Graph Construction Techniques: Selecting appropriate graph construction techniques is crucial for the effective- ness of graph-based summarization. Different approaches, such as similarity graphs and topic graphs, capture different aspects of document semantics. However, determining the optimal graph construction method for a given document corpus remains a challenge, as it depends on factors such as the nature of the documents and the desired summarization objectives.
- Centrality and Connectivity Measures: Designing effective centrality and connectivity measures is essential for identifying important information within the graph structure. While centrality measures such as degree centrality and PageRank are commonly used, their applicability to document summarization tasks may vary depending on the characteristics of the document corpus. Similarly, connectivity measures need to capture the nuances of semantic relationships between documents accurately.
- Integration of External Knowledge Sources: Incorporating external knowledge sources, such as domain-specific ontologies or knowledge graphs, can enrich the semantic representation of documents and improve summarization quality. However, integrating external knowledge sources into graph-based summarization frameworks poses challenges in terms of data integration, knowledge representation, and computational complexity.

Addressing these challenges requires interdisciplinary research efforts spanning natural language processing, machine learning, and network science. By overcoming these challenges, graph-based methods for document summarization can be utilized to their full potential and contribute to a variety of applications in information management and knowledge discovery.

5. Future Directions

While significant strides have been made in the field of graph-based document summarization, there are several promising avenues for future research that require exploration. These potential directions encompass the development of advanced graph models, the integration of multi-modal information, and the investigation of hybrid approaches combining graph-based and deep learning neural methods.

a) Advanced Graph Models

One promising direction for future research is the development of more sophisticated graph models that are tailored specifically for document summarization tasks. These models could incorporate rich representations of document semantics, capturing not only semantic relationships between documents but also the contextual information, temporal dynamics, and user preferences. Advanced graph models could leverage techniques from network science, graph theory, and deep learning to extract detailed insights from document collections and generate more informative summaries.

b) Integration of Multi-Modal Information

Another area of research interest is the integration of multi-modal information into graph-based summarization frame-works. In addition to textual content, documents often contain other modalities such as images, audio, and video. Integrating these diverse modalities into graph-based models could enable more comprehensive summarization, capturing a broader range of content and providing richer insights. Multi-modal graph- based summarization approaches could leverage techniques from computer vision, audio processing, and multimedia analysis to extract complementary information from heterogeneous document sources.

c) Hybrid Approaches

Hybrid approaches that combine the strengths of graph-based and neural methods represent an intriguing direction for future research. By integrating graph-based representations with neural architectures, researchers can leverage the expressive power of deep learning models while retaining the interpretability and structure-awareness of graph-based tech- niques. Hybrid approaches could encompass techniques such as graph neural networks (GNNs) [10], which operate directly on graph structures by integrating graph-based representations into neural architectures for document summarization tasks.

d) Evaluation and Benchmarking

In addition to algorithmic developments, future research should focus on standardizing evaluation metrics and bench- mark datasets for graph-based document summarization. Ro- bust evaluation frameworks are essential for objectively com- paring the performance of different summarization techniques and assessing their generalization capabilities across diverse document corpora. Benchmark datasets that capture various aspects of document semantics, including topic diversity, document length, and linguistic complexity, are crucial for fostering progress in the field and facilitating reproducible research.

e) Ethical and Societal Implications

Finally, future research in graph-based document summarization should also consider the ethical and societal im- plications of summarization technologies. As summarization systems become increasingly pervasive, it is essential to ad- dress concerns related to privacy, bias, and the responsible use of automated summarization algorithms. Researchers and practitioners must work collaboratively to develop transparent, accountable, and ethically responsible summarization systems that benefit society while minimizing potential harms.

By addressing these future directions, researchers can advance the graph-based document summarization techniques and unlock new possibilities for enhancing information management, knowledge discovery, and content understanding in this digital age.

6. Conclusion

In this survey, we have dived into the field of graph-based methods for unsupervised document summarization, offering an insightful overview of their principles, methodologies, and applications. These methods present a promising approach to addressing the challenges posed by the explosive growth of digital content, enabling the efficient summarization of large document collections by leveraging the inherent semantic relationships between documents. Through our comparative analysis, we have highlighted the strengths and limitations of graph-based techniques in contrast to traditional approaches such as TF-IDF and LDA. Graph-based methods excel in capturing complex semantic structures and identifying relevant information within document corpora, making them well-suited for a wide range of applications in information retrieval, document clustering, and content recommendation. However, despite their promise, graph-based methods for document summarization face several challenges. Addressing these challenges requires interdisciplinary research efforts and the ex- ploration of novel algorithmic techniques which will help in advancing the field of graph-based document summarization, for enhancing information management, knowledge discovery, and content understanding in this digital age.

References

[1] Sonawane, Sheetal S., and Parag A. Kulkarni. ”Graph based repre- sentation and analysis of text document: A survey of techniques.” International Journal of Computer Applications 96, no. 19 (2014).

[2] Jin, Wei, and Rohini K. Srihari. ”Graph-based text representation and knowledge discovery.” In Proceedings of the 2007 ACM symposium on

Applied computing, pp. 807-811. 2007.

[3] Ma, Qiang, Akiyo Nadamoto, and Katsumi Tanaka. ”Complementary information retrieval for cross-media news content.” In Proceedings of the 2nd ACM international workshop on Multimedia databases, pp. 45-54. 2004.

[4] Bonacich, Phillip. ”Power and centrality: A family of measures.” Amer- ican journal of sociology 92, no. 5 (1987): 1170-1182.

[5] Zhang, Junlong, and Yu Luo. ”Degree centrality, betweenness centrality, and closeness centrality in social network.” In 2017 2nd interna- tional conference on modelling, simulation and applied mathematics (MSAM2017), pp. 300-303. Atlantis press, 2017.

[6] Brin, Sergey, and Lawrence Page. ”The anatomy of a large-scale hypertextual web search engine.” Computer networks and ISDN systems 30, no. 1-7 (1998): 107-117.

[7] Barthelemy, Marc. ”Betweenness centrality in large complex networks.” The European physical journal B 38, no. 2 (2004): 163-168.

[8] Ramos, Juan. ”Using tf-idf to determine word relevance in document queries.” In Proceedings of the first instructional conference on machine learning, vol. 242, no. 1, pp. 29-48. 2003.

[9] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ”Latent dirichlet allocation.” Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[10] Xu, Keyulu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. ”How powerful are graph neural networks?.” arXiv preprint arXiv:1810.00826 (2018).