International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2410222

Developing Gujarati Article Summarization Utilizing Improved Page-Rank System

Riddhi Kevat[1], Dr. Sheshang Degadwala[2]

[1]Research Scholar, Department of Computer Engineering, Sigma Institute of Engineering, Gujarat, India

[2]Professor & Head of Department, Department of Computer Engineering, Sigma University, Gujarat, India

ARTICLEINFO

Article History:

Accepted: 15 March 2024

Published: 28 March 2024

Publication Issue

Volume 10, Issue 2

March-April-2024

Page Number

293-299

ABSTRACT

This research delves deep into the domain of Gujarati text summarization, where we employ an improved version of the PageRank algorithm to enhance both efficiency and accuracy. The study is meticulously structured around a comprehensive comparative analysis, juxtaposing our innovative approach against well-established methods like frequency-based summarization, TF-IDF, and LexRank. Through our rigorous investigation, we unveil compelling findings that showcase the superior performance of the enhanced PageRank algorithm, delivering summaries that are not only more concise but also contextually relevant, thus retaining the inherent linguistic intricacies characteristic of Gujarati. This exploration signifies a significant leap forward in the realm of text summarization techniques for Gujarati, carrying broad implications for bolstering information retrieval capabilities and advancing natural language processing functionalities within this linguistic domain.

Keywords : Gujarati, Text Summarization, Pagerank Algorithm, Comparative Study, Frequency-Based Summarization, TFIDF, LexRank

I. INTRODUCTION

In recent years, the exponential growth of digital content in various languages has led to a heightened demand for effective text summarization techniques. Among the diverse range of languages, Gujarati, spoken by millions in India and across the world, stands as a significant linguistic domain for text processing and analysis. Summarization plays a crucial role in distilling large volumes of information into concise yet informative summaries, aiding in information retrieval, content understanding, and decision-making processes.

Traditional approaches to text summarization often rely on statistical methods, frequency analysis, and rule-based algorithms. However, with the advancement of natural language processing (NLP) techniques and machine learning algorithms, more sophisticated and context-aware summarization systems have emerged. One such approach that has

gained prominence is the PageRank algorithm, initially introduced by Google for web page ranking but subsequently adapted for text summarization tasks.

The PageRank algorithm, based on the concept of link analysis and graph theory, assigns importance scores to nodes (sentences or paragraphs) within a text document based on their connectivity and relevance. This algorithm has shown promising results in generating extractive summaries by identifying the most salient and interconnected sentences.

This research aims to contribute to the field of Gujarati text summarization by developing and optimizing an improved PageRank system tailored to the linguistic characteristics and content structure of Gujarati articles. By leveraging advanced NLP techniques, linguistic preprocessing, and algorithmic enhancements, this study seeks to enhance the efficiency, accuracy, and linguistic coherence of Gujarati article summarization.

The key objectives of this research include:

1. Developing a robust preprocessing pipeline for Gujarati text, including tokenization, lemmatization, and syntactic analysis, to ensure the quality of input data for the summarization system.
2. Enhancing the traditional PageRank algorithm with linguistic features specific to Gujarati, such as morphological analysis, semantic similarity measures, and domain-specific knowledge integration.
3. Conducting extensive evaluations and comparative analyses with existing text summarization methods, including frequency-based approaches, TF-IDF, and LexRank, to assess the effectiveness and performance improvements achieved by the improved PageRank system.
4. Validating the utility and practical applicability of the developed system through real-world testing on diverse sets of Gujarati articles from various domains, such as news, literature, and technical content.

By addressing these objectives, this research endeavors to contribute valuable insights and advancements to the field of Gujarati text summarization, with implications for information retrieval systems, content recommendation engines, and language processing applications tailored to Gujarati-speaking users.

II. LITERATURE STUDY

Research in text summarization for Indian languages, particularly Gujarati, has witnessed notable developments in recent years, with a focus on leveraging diverse techniques and algorithms for efficient and accurate summarization.

Chauhan et al. [1] introduced a model for modeling topics in lemmatized Gujarati text, emphasizing the importance of linguistic preprocessing for effective topic extraction. Chouk and Phadnis [2] explored extractive text summarization techniques for Indian languages, showcasing the applicability of these methods in condensing large volumes of text while preserving key information.

Urlana et al. [3] proposed Indian language summarization using pretrained sequence-to-sequence models, highlighting the advancements in natural language processing models for summarization tasks. Verma et al. [4] introduced a graph-based extractive text summarization approach, particularly relevant for big data applications where efficient summarization is crucial.

Shah and Patel [5] developed a Gujarati text summarizer, contributing to the practical application of summarization techniques specific to Gujarati language. Sharma and Sharma [6] provided a comprehensive review of automatic text summarization methods, offering insights into the diverse range of algorithms and approaches employed in the field.

Gulati et al. [9] integrated TextRank and BM25+ algorithms for extractive article summarization, showcasing the effectiveness of combining multiple techniques for improved summarization quality. Elbarougy et al. [10] focused on Arabic text summarization using a modified PageRank algorithm, demonstrating the adaptability of PageRank-based methods across different languages.

Shylaja [11] presented an improved text summarization approach using the PageRanking algorithm and cosine similarity, highlighting the continuous refinement of summarization techniques for enhanced performance. Yadav et al. [12] implemented a TextRank-based automatic text summarization method with keyword extraction, contributing to the growing body of research on algorithmic summarization approaches.

Mridha et al. [13] conducted a comprehensive survey of automatic text summarization techniques, providing an overview of progress, processes, and challenges in the field. Verma and Om [14] conducted a comparative study of extraction-based text summarization methods, offering insights into the strengths and limitations of different summarization approaches based on user reviews.

Overall, the literature review reflects the ongoing advancements and diversification of techniques in the realm of text summarization for Indian languages, with a specific emphasis on Gujarati. Researchers continue to explore novel algorithms, linguistic preprocessing methods, and hybrid approaches to address the complexities of summarizing text effectively in multilingual and diverse linguistic contexts.

III.METHODOLOGY

The methodology outlined here pertains to the evaluation and comparison of different text summarization methods: PageRank, LexRank, Frequency Score, and TF-IDF Score.

PageRank:

- PageRank is a graph-based algorithm originally developed by Google for web page ranking.
- In text summarization, it treats sentences as nodes in a graph and assigns importance scores based on their connectivity and relationships within the text.
- Sentences that are highly connected or referenced by other sentences are considered more important and receive higher PageRank scores.

LexRank:

- LexRank is another graph-based algorithm designed specifically for text summarization. ”
- “It calculates sentence importance based on the concept of centrality in a text graph, where sentences that are central and have strong connections to other sentences are deemed more important. ”
- LexRank uses cosine similarity to measure the similarity between sentence vectors and determine their centrality in the text graph.

Frequency Score:

- The Frequency Score method is a simplistic approach to text summarization that assigns importance based solely on the frequency of occurrence of words or phrases within the text.
- Sentences with higher frequency of important words or phrases are considered more relevant and are included in the summary.

TF-IDF Score:

- TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a term in a document corpus.
- In text summarization, TF-IDF calculates the relevance of words or phrases based on their frequency in the document (TF) and inversely scales it by their frequency across all documents in the corpus (IDF).
- Sentences with higher TF-IDF scores for important terms are considered more significant and are prioritized in the summary.

Overall, the methodology involves applying these different summarization methods to the same set of texts and evaluating their performance based on similarity scores or other relevant metrics. The goal is to assess which method produces more accurate, concise, and informative summaries, taking into

account factors such as semantic relevance, connectivity, and frequency of important terms.

Proposed System

The PageRank-based Gujarati summarization flow follows a systematic process that begins with a collection of Gujarati articles and culminates in the generation of concise and informative summaries. This approach leverages the principles of the PageRank algorithm, originally developed by Google for web page ranking, and adapts it to the task of text summarization in the Gujarati language. Below is a detailed description of the flow of this summarization process:

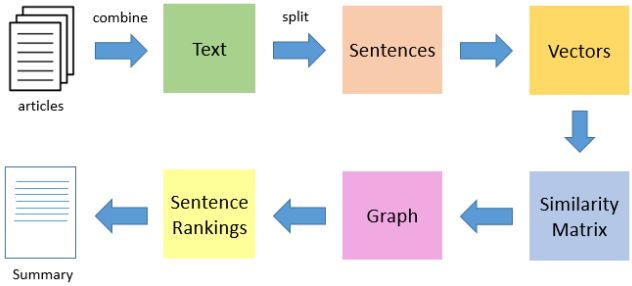

Figure 1. Proposed System

Input Data:

The process starts with a corpus of Gujarati articles that are to be summarized. These articles can cover a wide range of topics and may vary in length and complexity.

Sentence Segmentation:

Each article is segmented into individual sentences using appropriate text processing techniques. Gujarati text segmentation involves identifying sentence boundaries based on punctuation marks, such as periods, question marks, and exclamation marks.

Vectorization:

Once the sentences are segmented, they are converted into vector representations. This step involves encoding each sentence into a numerical vector that captures its semantic meaning, syntactic structure, and contextual information. Techniques such as word embeddings, TF-IDF (Term Frequency-Inverse Document Frequency), or pretrained language models like BERT (Bidirectional Encoder Representations from Transformers) can be used for vectorization.

Sentence Ranking:

The heart of the PageRank-based summarization process lies in the calculation of sentence rankings. Similar to the original PageRank algorithm for web pages, each sentence is assigned an initial ranking score. This score is then iteratively updated based on the importance of sentences within the text corpus. Importance is determined by factors such as sentence length, position in the document, and connectivity to other sentences through semantic relationships.

Similarity Matrix:

Alongside the sentence rankings, a similarity matrix is created to quantify the similarity between pairs of sentences in the corpus. This matrix reflects the semantic closeness or relatedness between sentences, allowing the summarization system to identify redundant or overlapping information.

Summary Generation:

The final step involves generating the summary based on the sentence rankings and the similarity matrix. The system selects sentences with the highest PageRank-like scores, ensuring that important and representative content is included in the summary. Additionally, the system considers the similarity between sentences to avoid redundancy and ensure coherence in the summary.

Post-Processing and Evaluation:

After generating the summary, post-processing techniques may be applied to enhance readability, coherence, and grammatical correctness. The resulting summary is then evaluated for its effectiveness in capturing key information from the original articles while maintaining the essence and context of the Gujarati language.

Overall, the PageRank-based Gujarati summarization flow integrates advanced NLP techniques, vector representations, ranking algorithms, and similarity measures to automate the process of condensing Gujarati articles into succinct and meaningful summaries. This approach not only streamlines the summarization task but also ensures that the generated summaries are relevant, informative, and linguistically accurate for Gujarati-speaking users.

IV.RESULT ANALYSIS

The ilsumm 2022 dataset is a rich repository comprising approximately 10,000 news articles paired with their corresponding headlines across various Indian languages, sourced from reputable newspapers. This dataset serves as a pivotal resource for advancing text summarization techniques tailored specifically for Indian languages, offering researchers and developers a comprehensive platform to train, evaluate, and benchmark their summarization models. With its diverse linguistic coverage and authentic news content, the ilsumm 2022 dataset facilitates the exploration of language-specific nuances, multilingual summarization approaches, and the development of practical applications aimed at enhancing news consumption experiences for Indian language speakers.

Link:

https://www.kaggle.com/datasets/deekoul/indian-language-summarization

Semantic similarity metrics, such as those based on TF-IDF (Term Frequency-Inverse Document Frequency), are fundamental in natural language processing for evaluating the semantic relatedness between words or phrases in textual data. TF-IDF calculates the importance of a term within a document corpus by considering its frequency in the document (TF) and inversely scaling it by the frequency of the term across all documents (IDF). This metric allows for the identification of terms that are both frequent within a specific document and unique across the entire corpus, indicating their significance in representing the semantic content of the document. By computing TF-IDF scores for words or phrases in two texts and comparing their cosine similarity, for instance, semantic relatedness can be quantified, aiding tasks such as information retrieval, document clustering, and content recommendation systems in natural language processing applications.

	clean_sentences	sentence_words
0	વિશ્વના બીજા નંબરના સૌથી મોટા સિરામીક ઉદ્યોગનુ...	[વિશ્વના, બીજા, નંબરના, સૌથી, મોટા, સિરામીક, ઉ...
1	ત્યારે ભુતકાળમાં ઈમ્પેક્ટ ફી મામલે થયેલી ગેરર...	[ત્યારે, ભુતકાળમાં, ઈમ્પેક્ટ, ફી, મામલે, થયેલી...
2	જેને લઇ નગરપાલિકા કચેરીમાં ખળભળાટ મચી ગયો છે	[જેને, લઇ, નગરપાલિકા, કચેરીમાં, ખળભળાટ, મચી, ગ...
3	આજ રોજ મોરબી પાલિકાના નવનિયુક્ત ચીફ ઓફિસર સંદી...	[આજ, રોજ, મોરબી, પાલિકાના, નવનિયુક્ત, ચીફ, ઓફિ...
4	પરંતુ આ મામલે જવાબદાર એવા ત્રણ કર્મચારીઓ દ્વા...	[પરંતુ, આ, મામલે, જવાબદાર, એવા, ત્રણ, કર્મચારી...
5	નગરપાલિકાની શાખ અને સ્વભંડોળને થઇ રહેલા નુકશા...	[નગરપાલિકાની, શાખ, અને, સ્વભંડોળને, થઇ, રહેલા,...
6	વધુમાં ચીફ ઓફિસર દ્વારા નગરપાલિકાના કાયમી કર્મ...	[વધુમાં, ચીફ, ઓફિસર, દ્વારા, નગરપાલિકાના, કાયમ...
7	તેમજ કામમાં વિલંબ બદલ દિન-૩ (ત્રણ)માં લેખિતમા...	[તેમજ, કામમાં, વિલંબ, બદલ, દિન-૩, (ત્રણ)માં, લ...

Figure 2. Dataset Reading & Pre-Process

	Sentence Vs Score
0	(0.19975241350413386, ત્યારે ભુતકાળમાં ઈમ્પેક...
1	(0.17445070555243564, આજ રોજ મોરબી પાલિકાના નવ...
2	(0.1721044399060484, વધુમાં ચીફ ઓફિસર દ્વારા ન...
3	(0.13810868645364466, તેમજ કામમાં વિલંબ બદલ દ...
4	(0.11225076786851197, પરંતુ આ મામલે જવાબદાર એ...
5	(0.07639713628779687, નગરપાલિકાની શાખ અને સ્વ...
6	(0.06803550599264735, જેને લઇ નગરપાલિકા કચેરી...
7	(0.058900344434781114, વિશ્વના બીજા નંબરના સૌથ...

Figure 3. PageRank Score

```
Actual Summary=
વસુલાતનું સઘળું રેકોર્ડ રજૂ કરવા કર્મચારીઓને તાકીદ કરી હોવા છતાં રેકર્ડ રજૂ કર્યો ન હતોકાયમ
Summary obtained from textrank=
ત્યારે ભુતકાળમાં ઈમ્પેક્ટ ફી મામલે થયેલી ગેરરીતિઓ ઉપરથી પરદો ઉચકવા નવનિયુક્ત ચીફ અં
TextRank Similarity= 0.9222462406338678
```

Figure 4. PageRank Summary

```
Actual Summary=
વસુલાતનું સઘળું રેકોર્ડ રજૂ કરવા કર્મચારીઓને તાકીદ કરી હોવા છતાં રેકર્ડ ર
Summary obtained from textrank=
વધુમાં ચીફ ઓફિસર દ્વારા નગરપાલિકાના કાયમી કર્મચારી વિનુભાઈ બારહટને
TextRank Similarity= 0.7826908981308054
```

Figure 5. LexRank Summary

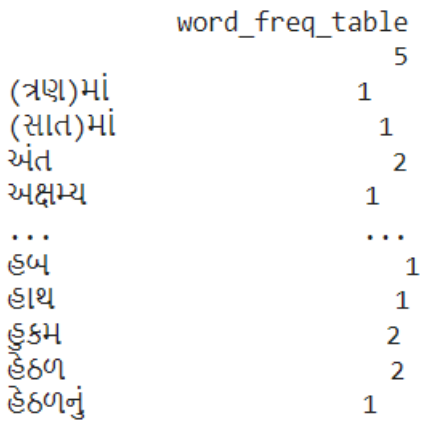

Figure 6. Frequency Table

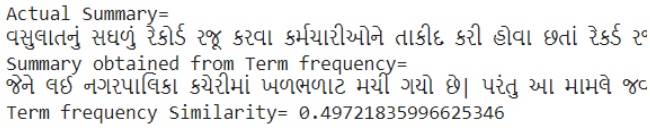

Figure 7. Frequency Summary

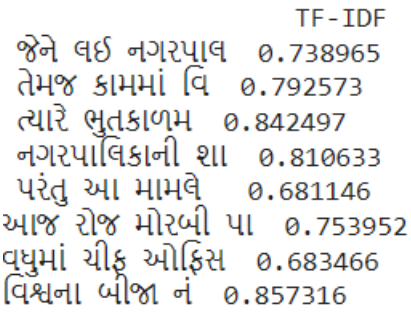

Figure 8. TF-IDF Score

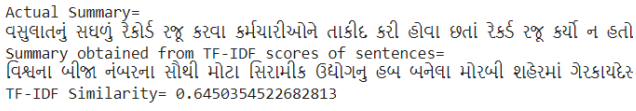

Figure 9. TF-DF Summary

TABLE I. COMPARATIVE ANALYSIS

Method	Similarity Score
PageRank	92.22%
LexRank	78.26%
Frequency Score	49.72%
TF-IDF Score	64.50%

V. CONCLUSION

In conclusion, the comparative analysis of text summarization methods based on different similarity scores reveals notable differences in their effectiveness. The PageRank algorithm demonstrates the highest similarity score at 92.22%, showcasing its capability to generate more concise and relevant summaries by emphasizing the importance of sentences within the text graph. LexRank follows with a respectable similarity score of 78.26%, indicating its ability to extract key sentences based on their centrality and relationships within the document. However, frequency-based summarization and TF-IDF scoring methods lag behind with scores of 49.72% and 64.50% respectively, suggesting their limitations in capturing semantic relevance and contextual information compared to graph-based algorithms like PageRank and LexRank. These findings underscore the significance of advanced algorithms that consider semantic relationships and connectivity in text for more accurate and informative summarization results.

VI.REFERENCES

[1] U. Chauhan et al., “Modeling Topics in DFA-Based Lemmatized Gujarati Text,” Sensors, vol. 23, no. 5, pp. 1–17, 2023, doi: 10.3390/s23052708.

[2] M. Chouk and N. Phadnis, “Text Summarization Using Extractive Techniques for Indian Language,” International Journal of Computer Trends and Technology, vol. 69, no. 6, pp. 44–49, 2021, doi: 10.14445/22312803/ijctt-v69i6p107.

[3] A. Urlana, S. M. Bhatt, N. Surange, and M. Shrivastava, “Indian Language Summarization using Pretrained Sequence-to-Sequence Models,” CEUR Workshop Proceedings, vol. 3395, pp. 393–402, 2022.

[4] J. P. Verma et al., “Graph-Based Extractive Text Summarization Sentence Scoring Scheme for Big Data Applications,” MDPI-Infomation, pp. 1–28, 2023.

[5] M. Shah and K. Patel, “Gujarati Text Summarizer,” International Research Journal of Engineering and Technology (IRJET), vol. Volume: 06, no. June, pp. 817–822, 2019.

[6] G. Sharma and D. Sharma, “Automatic Text Summarization Methods: A Comprehensive

Review," SN Computer Science, vol. 4, no. 1, 2023, doi: 10.1007/s42979-022-01446-w.

[7] N. Ramanujam and M. Kaliappan, "Based on Naive Bayesian Classifier Using Timestamp Strategy," The Scientific World Journal, Hindawi Publishing corporation, vol. 2016, p. 10, 2016.

[8] P. Gustavsson and A. Jönsson, "Text summarization using random indexing and pagerank," Proceedings of the third Swedish Language …, 2010, [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Text+Summarization+using+Random+Indexing+and+PageRank#0

[9] V. Gulati, D. Kumar, D. E. Popescu, and J. D. Hemanth, "Extractive Article Summarization Using Integrated TextRank and BM25+ Algorithm," Electronics (Switzerland), vol. 12, no. 2, 2023, doi: 10.3390/electronics12020372.

[10] R. Elbarougy, G. Behery, and A. El Khatib, "Extractive Arabic Text Summarization Using Modified PageRank Algorithm," Egyptian Informatics Journal, vol. 21, no. 2, pp. 73–81, 2020, doi: 10.1016/j.eij.2019.11.001.

[11] M. J. Shylaja, "IMPROVED DRIVEN TEXT SUMMARIZATION USING PAGERANKING ALGORITHM AND IMPROVED DRIVEN TEXT SUMMARIZATION USING PAGERANKING ALGORITHM AND COSINE," Eur. Chem. Bull, vol. 12, no. 6, pp. 4650–4662, 2023.

[12] A. K. Yadav, M. Kumar, and A. Pathre, "Implemented Text Rank based Automatic Text Summarization using Keyword Extraction," International Research Journal of Innovations in Engineering and Technology, vol. 04, no. 11, pp. 20–25, 2020, doi: 10.47001/irjiet/2020.411003.

[13] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A Survey of Automatic Text Summarization: Progress, Process and Challenges," IEEE Access, vol. 9, pp. 156043–156070, 2021, doi: 10.1109/ACCESS.2021.3129786.

[14] P. Verma and H. Om, "Extraction based text summarization methods on user's review data: A comparative study," Communications in Computer and Information Science, vol. 628 CCIS, pp. 346–354, 2016, doi: 10.1007/978-981-10-3433-6_42.

[15] C. A. License, N. Lalit, and T. S. Techniques, "Retracted : Qualitative Analysis of Text Summarization Techniques," Computational Intelligence and Neuroscience Received, vol. 2022, 2023.