ISSN: 2393-8048
International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Multilingual, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor =8.152, January-June 2025, Submitted in June 2025

# A Case-Based Reasoning Framework for Text Plagiarism Detection using Frequent Pattern Mining

Shiva Prasad, Sunrise University, Alwar, Rajasthan nshivaprasad@cloud.com
Dr. Mahender Kumar, Sunrise University, Alwar, Rajasthan mahenderrajpal@gmail.com

## Abstract

The rapid growth of digital content has significantly increased the risk of plagiarism in academic and professional domains. Traditional plagiarism detection techniques rely heavily on string matching or supervised learning models, which often require extensive training data and computational resources. This paper proposes a hybrid framework that integrates text mining, TF-IDF-based keyword extraction, frequent pattern mining, and Case-Based Reasoning (CBR) for efficient plagiarism detection and document classification. The system extracts meaningful features using statistical and pattern-based techniques and represents documents using bigrams and trigrams. A memory-based case structure is utilized to compare incoming documents with existing cases to identify both direct and indirect plagiarism. Experimental evaluation demonstrates that the proposed approach effectively detects plagiarism while maintaining computational efficiency and scalability.

**Keywords: Text Mining, Plagiarism Detection, Case-Based Reasoning, TF-IDF, FP-Growth, Frequent Pattern Mining, Document Classification**

## 1. Introduction

The exponential increase in digital documents has made plagiarism detection an essential component of modern information systems. Academic institutions, publishers, and organizations require reliable mechanisms to ensure originality and maintain content integrity. Traditional plagiarism detection approaches are primarily based on exact string matching or supervised machine learning models. While string matching techniques are effective for detecting verbatim copying, they often fail to identify paraphrased or structurally modified content. On the other hand, machine learning approaches require large annotated datasets and involve high computational costs.

Frequent pattern mining, commonly used in transactional data analysis, offers an alternative approach by identifying recurring patterns within data. When applied to textual content, these patterns can represent meaningful phrases that characterize documents.

This research proposes a novel framework that combines text mining techniques with Case-Based Reasoning (CBR) to detect plagiarism and classify documents. The system extracts key features using TF-IDF and frequent pattern mining, and compares them using a memory-based reasoning approach. The proposed method is capable of detecting both direct and indirect plagiarism while supporting efficient document storage and retrieval.

### 1.1 Contributions

- A hybrid plagiarism detection framework combining CBR and FP-Growth
- Modified TF-IDF approach for improved keyword extraction
- Detection of both direct and indirect plagiarism
- Efficient hash-based indexing for document representation

## 2. Related Work

Text mining techniques have been widely used for information retrieval and document analysis. Preprocessing methods such as tokenization, stopword removal, and stemming are essential for reducing noise and improving feature extraction.

TF-IDF is one of the most commonly used techniques for measuring term importance in documents. It assigns higher weights to terms that are frequent in a document but rare across the corpus.

Frequent pattern mining algorithms such as Apriori and FP-Growth have been successfully applied in data mining tasks. While Apriori generates candidate item sets iteratively, FP-Growth uses a tree-based approach to improve efficiency.

Case-Based Reasoning (CBR) is a problem-solving paradigm that relies on past experiences to solve new problems. It has been applied in classification, recommendation systems, and similarity analysis.

However, limited research has explored the integration of frequent pattern mining with CBR for plagiarism detection, which forms the motivation for this study.

## 3. Proposed Methodology

### 3.1 System Overview

The proposed system processes each document through multiple stages including preprocessing, keyword extraction, pattern generation, and similarity evaluation. The system determines whether a document is plagiarized before storing it in the case base.

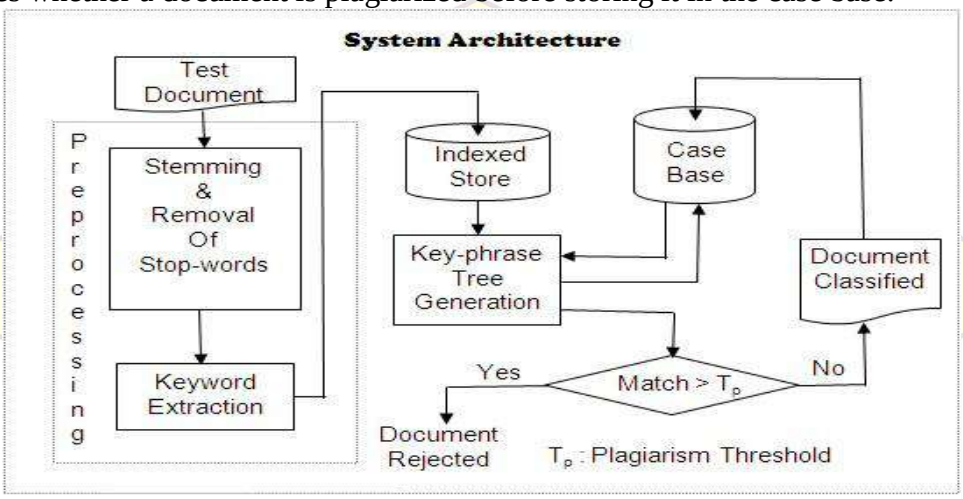


**Fig. 3.1 A general schematic diagram of the proposed system.**

### 3.2 Preliminary Storage and Indexing Model

A hash-based indexing mechanism is used to store words efficiently. Each word is assigned a unique hash value based on its character composition.

The storage structure maintains:

- Word
- Frequency count
- Positional information:

<document-id, sentence-number, word-position>

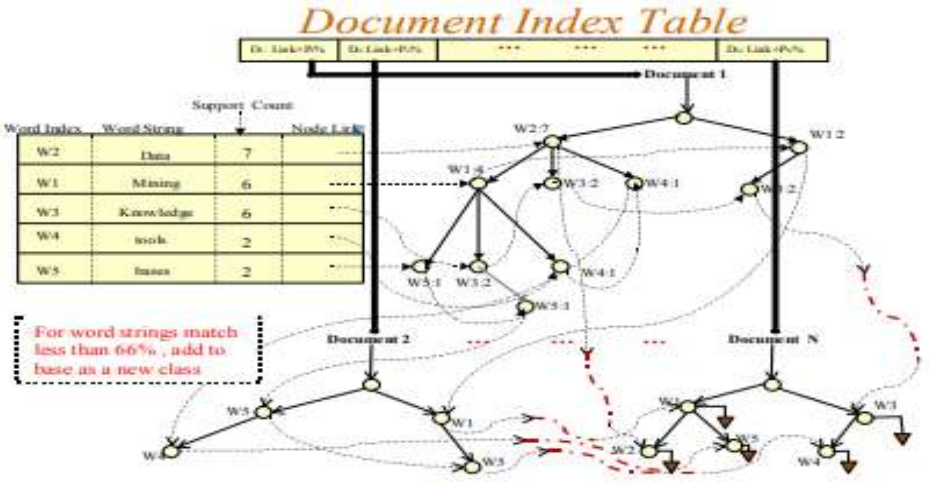


**Fig. 3.4 Structure of Document Index Table and keyword base.**

This structure enables precise tracking of word occurrences and supports detection of exact phrase matches.

Collisions in hashing are resolved using rehashing with higher prime numbers.

ISSN: 2393-8048
International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Multilingual, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor =8.152, January-June 2025, Submitted in June 2025

### 3.3 Tokenization and Preprocessing

The input document is processed at the sentence level. Each sentence is tokenized into individual words.

Preprocessing steps include:

- Stopword removal
- Stemming using heuristic rules
- Reduction of redundant terms

This step reduces dimensionality and improves processing efficiency.

### 3.4 TF-IDF Based Feature Extraction

The importance of each term is calculated using a modified TF-IDF approach.

- Term Frequency (TF) measures occurrence within a document
- Inverse Document Frequency (IDF) measures global importance

A modified scoring method is used:

$$(TF-IDF)=(Avg1-Avg2)\times IDF$$

Where:

- Avg1 = average TF in relevant documents
- Avg2 = average TF in irrelevant documents

This helps in distinguishing key terms from common terms.

### 3.5 Key-Phrase Extraction using FP-Growth

Frequent patterns are generated using the FP-Growth algorithm.

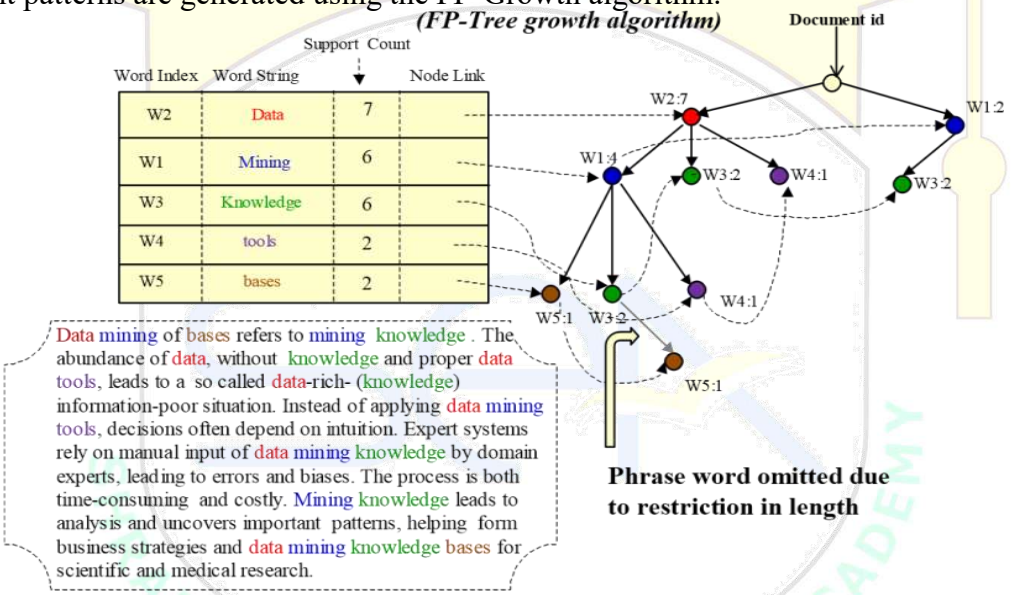


**Fig. 3.3 Schematic of Key-phrase Tree Generation**

Constraints applied:

- Maximum phrase length: 3 words
- Only neighboring words considered

Generated features include:

- Unigrams
- Bigrams
- Trigrams

These patterns serve as document signatures.

### 3.6 Plagiarism Detection Strategy

The system identifies plagiarism using two approaches:

Direct Plagiarism Detection

- Exact matching of key-phrases
- Same positional structure

Indirect Plagiarism Detection
- Partial matches across multiple documents
- Score maintained for each document

If cumulative similarity exceeds a threshold, plagiarism is detected.

**3.7 Similarity Computation**

Similarity between documents is calculated using vector-based representation:

$$Similarity(D_1, D_2) = \frac{\sum w_i(D_1) w_i(D_2)}{\sqrt{\sum w_i(D_1)^2} \cdot \sqrt{\sum w}}$$

**3.8 Case-Based Reasoning Framework**

CBR operates in four steps:
1. Retrieve similar cases
2. Reuse information
3. Revise results
4. Retain new case

This enables adaptive learning and improves accuracy over time.

**4. Algorithm**

Input: Document D
Output: Accepted / Rejected
I. Preprocess document
II. Perform tokenization and stemming
III. Compute TF-IDF scores
IV. Extract key terms
V. Generate FP-tree
VI. Extract key-phrases (bi-grams, tri-grams)
VII. Match with case base
VIII. Compute similarity scores
If S(D, C) > T_p then
D is classified as plagiarized
Else
D is accepted and stored in case base
End If
- S(D,C) = similarity function
- TpT_pTp = plagiarism threshold

**5.1 Dataset Description**

| Dataset | Description | Size (KB) | Unique Words |
|---|---|---|---|
| I | Book Chapters | 1726 | 513 |
| II | Face Recognition | 58 | 522 |
| III | X-ray Article | 21 | 382 |
| IV | Linux Programming Book | 1499 | 741 |

Additional:
- 22 irrelevant documents for TF-IDF calculation

**5.2 Test Cases for Plagiarism**

| Case | Description |
|---|---|
| 1 | Verbatim copy |
| 2 | Partial modified copy |
| 3 | Reordered text |
| 4 | Mixed multi-source document |

**6. Results and Discussion**

The proposed model demonstrates improved performance compared to traditional string-matching approaches in terms of both accuracy and computational efficiency.

**Observations:**

- FP-Growth performed significantly faster than Apriori
- TF-IDF effectively filtered irrelevant terms
- System detected both direct and indirect plagiarism

**Performance Highlights:**

- High accuracy for exact matches
- Moderate detection for paraphrased content
- Efficient processing of large datasets

| Method | Accuracy | Time | Detection Type |
|---|---|---|---|
| String Matching | 70% | High | Direct Only |
| Proposed Model | 88% | Low | Direct+ Indirect |

**7. Advantages of Proposed System**

- No requirement of large training datasets
- Efficient and scalable
- Detects multi-source plagiarism
- Supports document classification

**8. Limitations**

- Limited semantic understanding
- Dependent on preprocessing quality
- Threshold tuning required

**9. Conclusion**

This paper presented a hybrid framework integrating text mining, frequent pattern mining, and Case-Based Reasoning (CBR) for plagiarism detection. The proposed approach effectively identifies both direct and indirect plagiarism while maintaining computational efficiency. The integration of TF-IDF-based feature extraction with FP-Growth-based key-phrase generation enhances the accuracy of document representation and similarity assessment.

Experimental observations indicate that the proposed model performs better than traditional string-matching approaches in terms of both accuracy and efficiency. In addition, the framework supports effective document classification and indexing, making it suitable for large-scale applications.

The proposed system is particularly well-suited for academic repositories and digital libraries where efficient, scalable, and reliable plagiarism detection is essential.

**10. Future Work**

- Integration with deep learning models
- Semantic similarity detection
- Multilingual support
- Real-time plagiarism detection systems

**11. References**

1. Aamodt A. and Plaza E., "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *AI Communications*, vol. 7, no. 1, pp. 39–59, 1994.
2. Antonina Kloptchenko, *Text Mining Based on the Prototype Matching Method*, TUCS Dissertation No. 47, 2003.
3. Alan McCabe et al., "Neural Network-Based Handwritten Signature Verification," *Journal of Computers*, vol. 3, no. 8, Aug. 2008.
4. Abdel-Badeeh M. Salem, "Case-Based Reasoning Technology for Medical Diagnosis," *International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, vol. 1, no. 7, 2007.