Copy Detection Based On Winnowing Algorithm

Shuai Hao
Stevens Institute of Technology
shao8@stevens.edu

Yinghao Wang
Stevens Institute of Technology
ywang393@stevens.edu

Fan Luo
Stevens Institute of Technology
fluo4@stevens.edu

Dr.Dov Kruger
Stevens Institute of Technology
dkruger@stevens.edu

Abstract

Nowadays, most of documents are produced in digital format, in which they can be easily accessed and copied. Document copy detection is a very important tool for protecting the author's copyright.The Winnowing algorithm is a fingerprint based text similarity detection method. This paper makes a literature study of Winnowing, a fingerprinting algorithm for documents. The Winnowing selects fingerprints from hashes of k-grams, a contiguous substring of length k. We study different K-grams sizes and different winnowing windows size. We show the result and propose the improvement for this algorithm.

1 Introduction

Plagiarism detection, also known as duplicate or copy detection is to detect whether the contents of one file copy from one or more others. Plagiarism detection is an important method for academic misconduct, author identification and digital products intellectual property protection.Plagiarism detection has been widely discussed in recent years. Various approaches have been proposed such as the text-similarity calculation, structural-approaches, and the fingerprint. (Hasan et al., 2018) (Sutoyo et al., 2017) Winnowing algorithm is one method to pick up fingerprint and Copy detection is based on the winnowing algorithm. Therefore, we study the various conditions which K-gram size changes and winnowing window size changes. This paper have these following contributions:

- Show the whole progress of one copy detection method based on winnowing algorithm.
- Study the influence of K-gram size and Winnowing window size to the copy detection performance.
- Propose some improvements for this method.

2 Related Work

2.1 Copy detection

The research of text copy detection can be traced back to the code similarity detection which is used to prevent the large-scale program copying in 1970s. Natural language copy detection technology appeared in the 1990s. Till now, Some of the conventional approach to the detect of plagiarism indications including character-based, structural-based, cluster-based, syntax-based, cross-language based, and semantic-based. Furthermore, there are some character based algorithm such as:

- Fingerprint. Using parts of document to be checked to plagiarism. The checked parts will be proceed by certain mechanism (character-matching) to determine the fraud. There are two algorithms based on fingerprint which are fingerprint algorithm and winnowing algorithm
- String similarity. Using this approach overlapped document will be found using string-matching and sentence matching algorithm. This approach was developed and applied to COPS (copy protection system). In addition, detection system is also developed using Longest Common Subsequence (LCS), Levenshtein Distance and also Smith-Waterman
- Structural-based methods. While the two previous approaches above focused on exploiting lexical features of a document, structure-matching method focused on structural things such as header, section, paragraph, and reference. Tree-structure feature with POS tagging is also used in this study. Other approaches such as cluster-based, syntax-based,cross language-based and semantic-based also used by other research

K-values	# N=8	# N=9	N=10
K=3	62.50%	62.2%	64.4%
K=4	58.28%	57.41%	56.96%
K=5	56.60%	58.03%	55.44%
K=6	55.44%	55.89%	55.18%

Table 1: The result of different K values and different N values. The length of text is 50 words(L= 50)

K-values	# N=8	# N=9	N=10
K=3	63.70%	62.83%	63.18%
K=4	51.23%	50.51%	49.87%
K=5	44.23%	43.79%	42.81%
K=6	40.93%	40.75%	40.70%

Table 2: The result of different K values and different N values. The length of text is 1000 words(L= 1000)

2.2 winnowing

The Winnowing algorithm is a fingerprint based text similarity detection method, proposed by Schleimer et al in 2003. (Schleimer et al., 2003)The basic idea of Winnowing comes from the Karp-Rabin algorithm which using overlapping kgram and moving window for string matching. Winnowing chooses the minimize hash value in each window to compose the document fingerprint, and then compares documents' fingerprints using pairwise method to find the copied text. Winnowing is a lightweight and flexible similarity detection method, it is robust for sentence and text block rearrangement, and the influence of interference words can be effectively reduced through reasonable parameter setting.

3 Method

This section will introduce the copy detection method based on winnowing algorithm.

A Preprocessing Document Before a document is matched, it has to go through preprocessing steps. Preprocessing steps used in this study are: Case Folding and Filtering.

Case folding: All letters have to be case-insensitive. As for this study, all letters are changed to lowercase. Parsing process is needed so that textual contents of each paragraph are obtained. This process resulted a term in the article.

Filtering also called stop word removal. Important terms from previous process are decided. Unimportant terms are discarded so that stop list can be made.

B K-gram Pick Up K-gram is a series of substrings adjacent to the length of k . This method produces a substring sequence of k-grams, where k is the parameter chosen by the user. K-gram takes a substring of the character of a letter k of a word which is continuously read from the source text to the end of the document. This step will produce substrings to be processed in next steps.

C Hashing Process Hashing is the conversion of a series of characters into a value or code that becomes a marker of the sequence of characters. With this change, it creates a marker as an index to be used in retrieval or information retrieval. The function to generate this value is called the hash function, while the resulting value is called the hash value. The hash function creates "fingerprint" from various input data. The hash function will transpose the data to produce a fingerprint.In this step, Hashing algorithm will transfer every gram value to be hashed value.

D Winnowing Algorithm Process. The strategy adopted by the Winnowing algorithm is to select the smallest hash value in each window (obviously, two Windows may share the same minimum value). If there are multiple minimums, select the rightmost one. This strategy not only ensures that sufficient fingerprint information is selected, but also ensures that too large fingerprints are not generated. This step will choose the fingerprints from hashed values based on the windows size. Fingerprints present the document.

K-values	# N=8	# N=9	N=10
K=3	68.70%	66.88%	66.18%
K=4	44.14%	43.23%	42.88%
K=5	25.90%	9.41%	25.38%
K=6	15.79%	15.46%	15.13%

Table 3: The result of different K values and different N values. The length of text is over 10000 words.(L= 10000+)

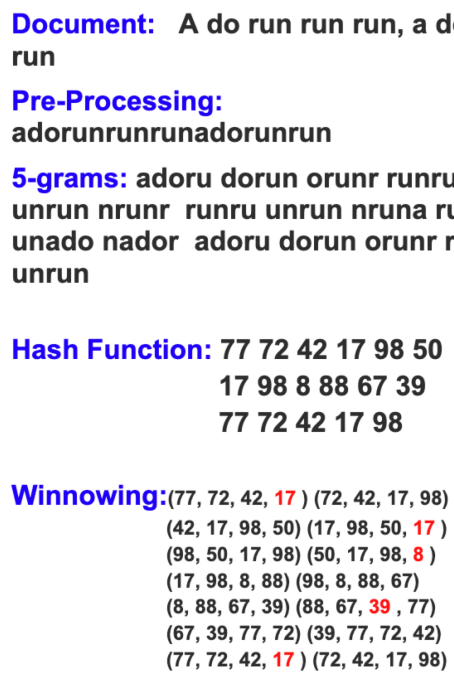

Figure 1: An example to show the process of producing fingerprint .The input of this method is the document which have different lengths. The output of this method is the fingerprint which can present this document.

D Similarity. The Jaccard similarity index (sometimes called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It's a measure of similarity for the two sets of data, with a range from 0 to 1. The higher the percentage, the more similar the two populations. Although it's easy to interpret, it is extremely sensitive to small samples sizes and may give erroneous results, especially with very small samples or data sets with missing observations. For two sets A and B define the Jaccard coefficient: $J(A, B) = |A \cap B|/|A \cup B|$.

4 Experiments and Results

4.1 Experiments Settings

k-grams will produce different grams for picking fingerprint and the winnowing algorithm will choose fingerprint based on the widows size. Therefore, we will select different K(gram size) and different N (window size) as the experiments item to see the performance of them. At the same time, we consider the different length texts. L means the number of words for every article. In our experiments, we test different K values(K= 3, k=4, K=5 and K= 6). We also test different N values(N= 8, N= 9 and N=10). Due to different lengths will have so different results, we also test different L values(L= 50, L= 1000 and L= 10000+).

4.2 Results

Table 1 shows the experiment results when L = 50.Table 2 shows the experiment results when L = 1000. Table 3 shows the experiment results when L = 10000+.The test results reflect that the length of the article has an impact on the test results. The similarity of articles will change with the change of k value.The value of n seems to have little effect on the results.

5 Improvements

About this copy detection, after we study it based on different conditions, we propose these following improvements:

- No current dataset: there is no current suitable datasets to test different copy detection models.
- Pick up dataset from internet randomly: When people test the copy detection methods, researchers always find dataset form internet randomly. This one will impact the reliability of the experiment.
- Do not know the exact similarity rate of two articles.

- Time complexity is O(n^2), which is not suitable for some large files.

References

Eric Ganiwijaya Hasan, Arya Wicaksana, and Seng Hansun. 2018. The implementation of winnowing algorithm for plagiarism detection in moodle-based e-learning. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 321–325. IEEE.

Saul Schleimer, Daniel S Wilkerson, and Alex Aiken. 2003. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 76–85.

Rhio Sutoyo, Insan Ramadhani, Angger Dwi Ardiatma, Sanditya Cakti Bavana, Harco Leslie Hendric Spits Warnars, Agung Trisetyarso, Bahtiar Saleh Abbas, and Wayan Suparta. 2017. Detecting documents plagiarism using winnowing algorithm and k-gram method. In *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pages 67–72. IEEE.