Documents Similarity Detection Based on Bloom Filter

Final Presentation

Sujit Maharjan
071/MSCS/668

Overview

1. Introduction
2. Problem Definition
3. Objectives
4. Algorithms
5. Results
6. Conclusions
7. Limitations
8. Future Enhancements
9. References

Introduction

- Internet is huge
- Google indexed 200TB of data in 2014 (approx. 0.004% of total internet)
- Many of the documents are duplicate websites, blogs, articles and reports
- People want to find whether the document is copied from other source or not
- There are various method of matching documents and finding similarity
- One of the method is using Bloom filter data structure.

Problem Definition

- Document matching is very simple process to do.
- Compare each word or sentence from one doc to another.
- Is it possible to apply this method when we have to compare all the documents in internet ?
- This method will have complexity of O(n^2).
- So, we need better algorithm to compare documents efficiently
- In this regard, bloom filter can represent whole document with fixed number of bits and can search the whole document within constant time.

Objectives

- To build application to detect similarity between two documents
- Effectively use bloom filter for similarity measure

Algorithms

Rabin Fingerprinting

Given an n-bit message $m_0, \ldots\ldots, m_{n-1}$, we view it as a polynomial of degree n-1 over the finite field.

$$f(x) = m_0 + m_1 x + \ldots\ldots + m_{n-1} x^{n-1}$$

We then pick a randon irreducible polynomial p(x) of degree k over finite field and we define the fingerprint of the message m to be the remainder r(x) after division of f(x) by p(x) over finite field.

W-shingling

In natural language processing a w-shingling is a set of unique "shingles" that can be used to gauge the similarity of two documents. The w denotes the number of tokens in each shingle in the set.

The document, "a rose is a rose is a rose" can be tokenized as follows:

{(a,rose,is,a),(rose,is,a,rose),(is,a,rose,is),(is,a,rose,is),(a,rose,is,a),(rose,is,a,rose)}=

{(a,rose,is,a),(rose,is,a,rose),(is,a,rose,is)}

Bloom Filter

- A set U is represented by an array of m bits
- Each u (uEU) is hashed using k independent hash functions hk.
- Each hk maps to one bit in the array.
- For set membership check, Bloom filter may yield a false positive.
- For n = |U|, bloom filter size m
- The optimum value of k that minimizes the false positivity p^k where p = probability that a given bit is set is k = (m/n)ln2

Jaccard Similarity

Consider two sets A = {0,1,2,5,6} and B = {0,2,3,5,6,9}.

The Jaccard Similarity is defined

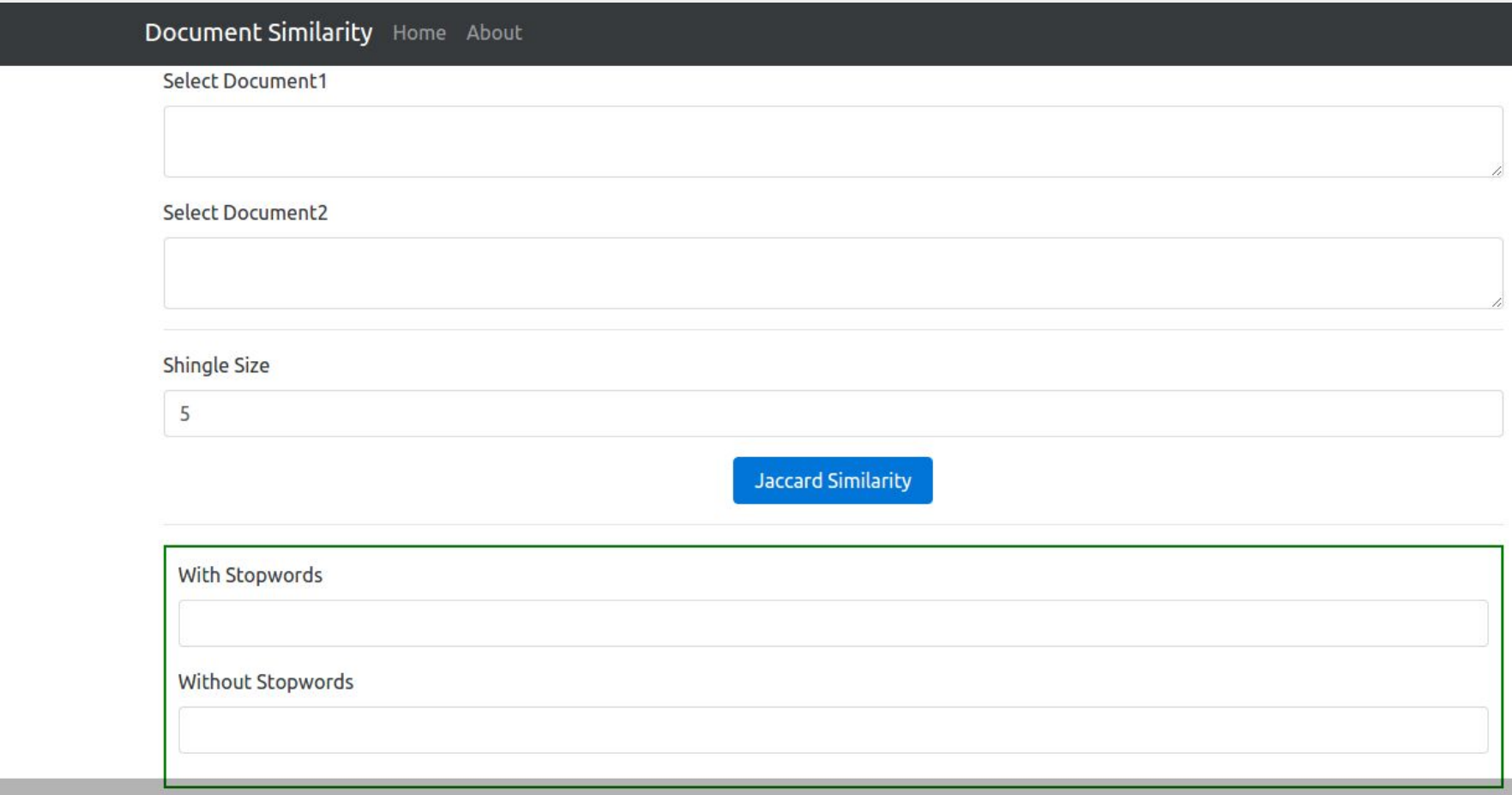
JS(A,B) = $|A \cap B| / |AUB| = |\{0,2,5,6\}| / |\{0,1,2,3,5,6,9\}| = 4/7$

More, given a set A, the cardinality of A denoted |A| counts how many elemtns are in A. The intersection between two sets A and B is denoted $A \cap B$ and reveals all items which are in both sets.The union between sets A and B is denoted AUB and reveals all the items which are in either set.

Tools used

Results

Landing Page



Documents Input

Select Document1

We all want to see our kids going towards success which is only possible through the good and proper education. Every parent tells their kids from childhood about the importance of education in the life and all the advantages of education to make their mind towards better study in the future. Make your kids and children habitual of writing essays, participate in debates and discussion and many more skill enhancing activities in the schools or at home using such simple essays. We are here to help you all parents in making your kid's better future by providing simple essay on importance of education. Following importance of education essay are easily worded and given under various words limit

Select Document2

Better education is very necessary for all to go ahead in the life and get success. It develops confidence and helps building personality of a person. School education plays a great role in everyone's life. The whole education has been divided into three divisions such as the primary education, secondary education and Higher Secondary education. All the divisions of education have their own importance and benefits. Primary education prepares the base which helps throughout the life, secondary education prepares the path for further study and higher secondary education prepares the ultimate path of the future and whole life. Our good or bad education decides that which type of person we would in the future.
Importance of Education Essay 2 (150 words)

Shingle Size

5

Jaccard Similarity

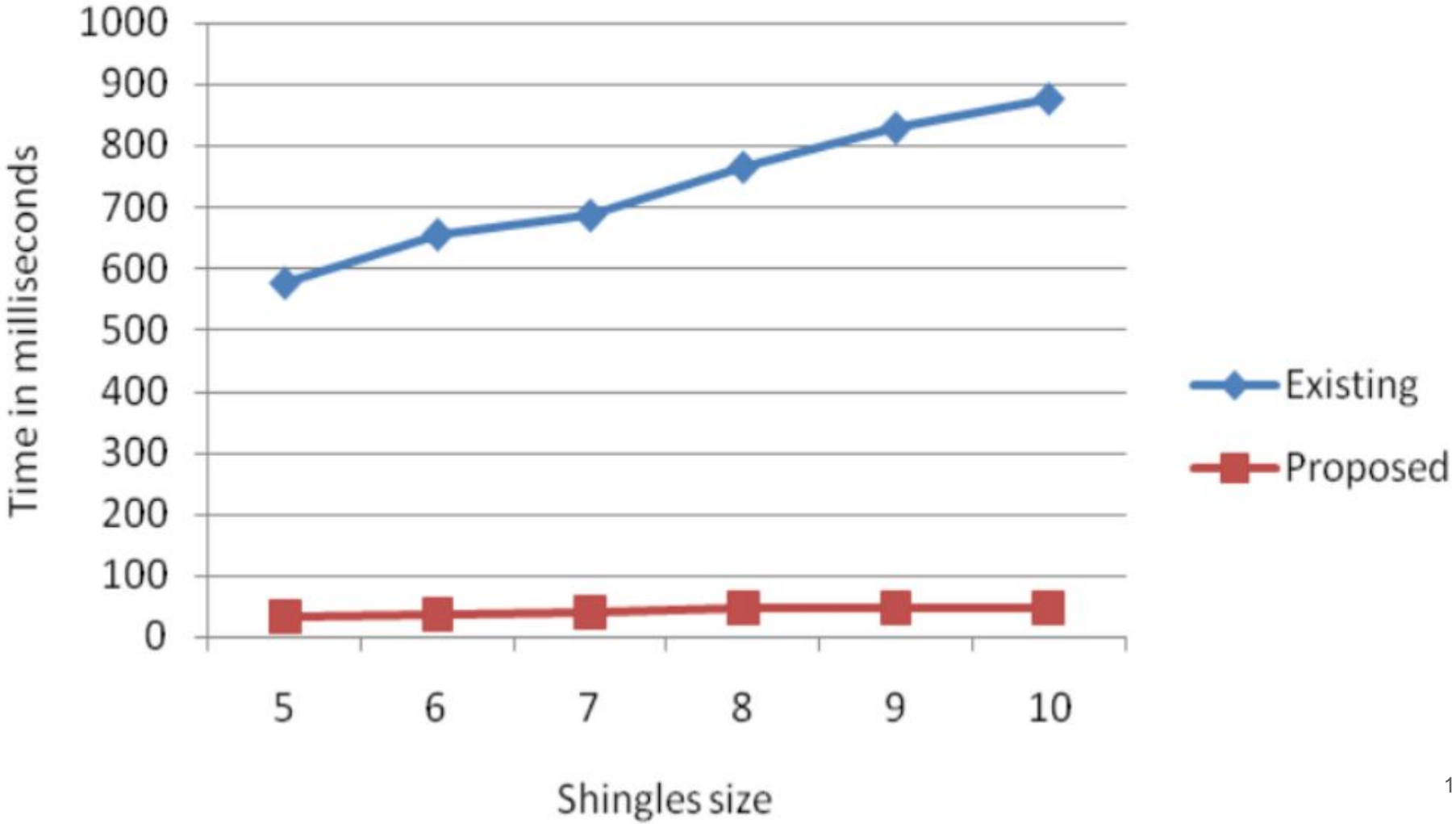
Similarity output

With Stopwords

0.657601977750309

Without Stopwords

0.6706586826347305


1000
900
800
700
600
500
400
300
200
100
0
Time in milliseconds
5
6
7
8
9
10
Shingles size
Existing
Proposed

Conclusions

- The shingles is computed and stored in bloom filter
- The searching of shingle is linearly dependent on the file size in traditional method
- The searching time using bloom filter is constant and reduces searching complexity for large sized files

Limitations

- The computation of fixed sized shingles is computationally expensive for huge file size.
-
- Fixed sized bloom filter make the compuation more expensive for small sized files.

Future Enhancements

- Optimize the hashing function so as to compute the bloom filter in less time
- Make bloom filter size dynamic.
- A parallel processing environment can be build such that shingling of documents can be distributed in parallel and multiple files can be selected for similarity search.
- This work can be extended for finding semantic similarity in text document i.e to find the similarity between text files which are semantically similar

References

- Navendu Jain, Mike Dahlin, and Renu Tewari, "Using Bloom Filters to Refine Web Search," in Eighth International Workshop on the Web and Databases, Baltimore, 2005.
- Anna Huang, "Similarity Measures for Text Document Clustering".
- Sachindra Singh Chauhan, "Similarity Search using Locality Sensitive Hashing and
- Bloom Filter," THAPAR UNIVERSITY, PATIALA, Masters Thesis 2014.
- A. Broder et al., "Min-wise independent permutations," Proc. Theory of computing,1998.
- Burton H. Bloom, "Space / Time Trade-offs in Hash Coding with Allowable Errors," Communications of the ACM, vol. 13, no. 7, July 1970.
- Chow Kok Kent and Naomie Salim, "Features Based Text Similarity Detection,"
- JOURNAL OF COMPUTING, vol. 2, no. 1, January 2010.