Information Processing and Management 000 (2016) 1–17

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Rapid detection of similar peer-reviewed scientific papers via constant number of randomized fingerprints

Yaakov HaCohen-Kerner*, Aharon Tayeb

Department of Computer Science, Jerusalem College of Technology – Lev Academic Center, 21 Havaad Haleumi St., P.O.B. 16031, 9116001 Jerusalem, Israel

ARTICLE INFO

Article history:
Received 28 September 2014
Revised 22 June 2016
Accepted 27 June 2016
Available online xxx

Keywords:
Fingerprinting
Heuristic methods
Plagiarism detection
Similar peer-reviewed scientific papers

ABSTRACT

This research is concerned with the detection of similar academic papers. Given a tested paper from a given corpus of 10,099 peer-reviewed scientific papers, a two-stage process was activated. During the first stage, most of the papers were filtered out using a fast filter method. In the second stage, in order to detect similar papers we applied 23 heuristic variants derived from 3 novel prototype methods using various parameter settings. The three novel prototype methods are: CT-TR – Constant Number of randomized T fingerprints, compared to each one-third of R (first/middle/last) fingerprints, CT-AR: Constant Number of randomized T fingerprints, compared to all R fingerprints, and CDT-AR: Constant Number of divided randomized T fingerprints compared, to all R fingerprints. Results achieved by the new methods are superior to those of previous heuristic methods, which were approximations of the "Full Fingerprint" (FF) method, currently considered the best heuristic method. The order of this new methods' run-time, $\Theta(n)$, is far more efficient than the order of the FF method run-time, $\Theta(n^2)$ (after removing short documents from the corpus).

1. Introduction

Plagiarism is defined as the "use or close imitation of the language and thoughts of another author and the representation of them as one's own original work".[1] "Since antiquity, writers and artists have borrowed words, images, and ideas from predecessors without attribution" (Loui, 2002). Nowadays, with the proliferation of online text, plagiarism has become simpler and more sophisticated. Inter alia, plagiarism is noticeable in academia (i.e., reports, papers, theses and dissertations) and presswork (i.e., Internet news, newspapers, and blogs) (Clough, 2003). Plagiarism, particularly self-plagiarism, is considerably evident in the recycling of academic papers. This phenomenon appears to be constantly developing. It is necessary, therefore, to develop quick, efficient and complex methods to discover similarities between existing articles and any tested article.

Two main general approaches exist for the detection of similar documents: Global Similarity Assessment methods and Local Similarity Assessment methods (Meuschke & Gipp, 2013). Global Similarity Assessment methods analyze features of longer text segments such as paragraphs, sections and full papers. Global methods can be divided into term occurrence analysis, citation-based, and stylometry methods. Local methods analyze pairs of relatively small text segments such as character

* Corresponding author. Fax: (972) 2 6751046.
E-mail addresses: kerner@jct.ac.il (Y. HaCohen-Kerner), aharontayeb@gmail.com (A. Tayeb).

[1] The 1995 Random House Compact Unabridged Dictionary.

http://dx.doi.org/10.1016/j.ipm.2016.06.007
0306-4573/

n-grams, word n-grams and sentences (Stein & Meyer zu Eissen, 2006). In this research, our focus is Local Similarity Assessment methods.

Fingerprinting is the most popular of Local Similarity Assessment methods. A Fingerprint is a sequential substring. A document can be represented by segmenting it into a set of substrings and selecting a subset of substrings. A full fingerprint (FF) of a document is a set of all its possible α length sequential substrings in characters/words. There are $N-\alpha+1$ such substrings, with N representing the length of the document in characters/words. An FF contains overlapping sub-strings. Given a tested paper (T) and a retrieved paper (R), where T's size is $|T|$ and n is the number of fingerprints common to T and R, the similarity measure between T and R can be calculated by $n/|T|$, which measures the amount of T contained in R.

Fingerprinting methods can be classified according to the Chunking Unit of the Fingerprint. A Fingerprint can either be a character n-gram (Butakov & Scherbinin, 2009; Heintze, 1996), a word n-gram (HaCohen-Kerner, Tayeb, & Ben-Dror, 2010; Hoad & Zobel, 2003), a sentence (Barrón-Cedeño & Rosso, 2009; Kang, Gelbukh, & Han, 2006), or a combined method such as sentence and word n-gram (Sorokina, Gehrke, Warner, & Ginsparg, 2006).

Various Selective Fingerprints (SFs) require fewer fingerprints than the FF method to represent a document, which reduces the run-time for computing the similarity measure of the documents. The most trivial type of SF is the "All Substrings Selection" described by Hoad and Zobel (2003). This method contains all non-overlapping substrings of length α in characters/words from the document. Additional types of SF methods are as follows: Hoad and Zobel (2003) suggest positional, frequency-based, and structure-based methods; Monostori, Finkel, Zaslavsky, Hodász, and Pataki (2002) propose a Symmetric Similarity (SS) measure; Bernstein and Zobel (2004) suggest other similarity measures, such as S2 and S3, which are based on the minimal number of T and R Fingerprints and the average number of T and R Fingerprints, respectively. Implementation and evaluation of these SF methods as well as other variations of SF methods are presented in HaCohen-Kerner et al. (2010).

In this research, we are interested in discovering recycled articles (i.e. articles that are similar to previous articles (from the level of Medium Similarity (MS), which corresponds to 40–60% common Fingerprints between the two papers) using fast and efficient Heuristic Fingerprinting method(s).

This paper is organized as follows: Section 2 provides background information concerning plagiarism detection. Section 3 describes methods for detecting similar papers introduced in relevant previous Fingerprints studies. Section 4 introduces the filtering and detection stages of the proposed methodology including the Heuristic methods that aim to detect similar papers. Section 5 presents the examined corpus, the experiments that have been performed and their analysis. Section 6 discusses the statistical significance of the error results. Section 7 details two illustrative examples. Section 8 concludes and proposes future directions for research.

2. Plagiarism detection

A wide range of research has been carried out regarding plagiarism detection in general and the detection of similar papers in particular. During the last two decades, many methods and systems were developed to automatically identify plagiarism. Several surveys have been held regarding automatic plagiarism detection in text. Maurer, Kappe, and Zaka (2006) discuss the general setting of textual plagiarism, present various algorithms and report on results of plagiarism detection software. Kumar and Govindarajulu (2009) classify and review works related to detection of duplicate and near duplicate general documents and web documents with web crawling. Meuschke and Gipp (2013) present an overview and classification of methods and systems aiding in detection of academic plagiarism. Test cases for plagiarism detection software are presented by Weber-Wulff (2010).

Another task related task plagiarism detection is the detection of near-duplicate web pages. This sort of detection can be performed using various techniques such as Shingling (Brin, Davis, & Garcia-Molina, 1995; Broder, 1993; Chowdhury et al., 2002) and Simhash (Charikar, 2002; Manku, Jain, & Das Sarma, 2007). These techniques were either developed or used by various web search engines. A "shingle" is the hash-value of a word n-gram. A document is represented by a set of shingles that represent a set of features of a document. "Simhash" is a dimensionality reduction technique, which maps high-dimensional vectors to small-sized fingerprints. This technique requires less space than the shingling technique. Henzinger (2006) compared shingling and simhash algorithms on a set of 1.6B distinct web pages. The results reveal that both algorithms did not achieve results of adequate precision for near-duplicate pairs on the same website. Henzinger presents a combined algorithm, which achieves more precise results with the same corpora. Additional advanced methods proposed for the discovery of near-duplicate web-pages are Spot Signatures (Theobald, Siddharth, & Paepcke, 2008), which are consist of stopwords with short chains of adjacent content terms, and "super shingles" (Broder, 2000) in order to reduce the complexity of shingling when processing large corpora.

Lyon, Malcolm, and Dickerson (2001) propose a method for detection of short passages of similar text based on the characteristic distribution of word trigrams and a set of theoretic principles. Barrón-Cedeño and Rosso (2009) investigate various methods for identifying plagiarized sentences that have been modified by rewording, insertion or deletion. Their experiments with the METER corpus show that the use of word bigrams and trigrams yields the best results. Bigrams achieve higher recall results while trigrams achieve higher precision results. Lyon, Barrett, and Malcolm (2004) refer to a set of trigram words as the typical representation of any text document. Trigram words were also the chosen compared string of Bao, Lyon, and Lane (2006) for their FERRET copy detector system in student course work in Chienese. HaCohen-Kerner et al. (2010) found that trigram words enable better detection of similar papers than quadgram words.

Please cite this article as: Y. HaCohen-Kerner, A. Tayeb, Rapid detection of similar peer-reviewed scientific papers via constant number of randomized fingerprints, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.06.007