Detection of Simple Plagiarism in Computer Science Papers

Yaakov HaCohen-Kerner
Department of Computer Science, Jerusalem College of Technology (Machon Lev)
kerner@jct.ac.il

Aharon Tayeb
Department of Computer Science, Jerusalem College of Technology (Machon Lev)
aharontayeb@gmail.com

Natan Ben-Dror
Department of Computer Science, Jerusalem College of Technology (Machon Lev)
bd.natan@gmail.com

Abstract

Plagiarism is the use of the language and thoughts of another work and the representation of them as one's own original work. Various levels of plagiarism exist in many domains in general and in academic papers in particular. Therefore, diverse efforts are taken to automatically identify plagiarism. In this research, we developed software capable of simple plagiarism detection. We have built a corpus (C) containing 10,100 academic papers in computer science written in English and two test sets including papers that were randomly chosen from C. A widespread variety of baseline methods has been developed to identify identical or similar papers. Several methods are novel. The experimental results and their analysis show interesting findings. Some of the novel methods are among the best predictive methods.

1 Introduction

In light of the explosion in the number of available documents, fast and accurate searching for plagiarism is becoming more needed. Identification of identical and similar documents is becoming very important.

Plagiarism is the use of the language and thoughts of another work and the representation of them as one's own original work (Wikipedia, 2010; Library and Information Services, 2010). Plagiarism can be committed by "recycling" other's work as well as by one’s own work (self-plagiarism).

Various levels of plagiarism exist in many domains in general and in academic papers in particular. In addition to the ethical problem, plagiarism in Academics can be illegal if copyright of the previous publication has been transferred to another entity.

It is important to mention, that in many cases similar papers are different versions of the same work, e.g., a technical report, a poster paper, a conference paper, a journal paper and a Ph. D. dissertation.

To avoid any kind of plagiarism, all sources which were used in the completion of a work/research must be mentioned (Library and Information Services, 2010).

Over the last decade, various softwares have been built to automatically identify plagiarism (e.g., Collberg et al. (2005), Sorokina et al. (2006), and Keuskamp and Sliuzas (2007)).

In this research, we developed such a system. This system is planned to deal with simple kinds of plagiarism, e.g., copying of sentences or part of sentences. We have built a corpus that contains academic papers in computer science written in English. Most of the papers are related to the domain research of Natural Language Processing (NLP) and are from the last ten years.

The remainder of this paper is organized as follows: Section 2 gives a background regarding plagiarism. Section 3 overviews researches and systems dealing with detection of plagiarism. Section 4 describes five groups of baseline methods, which have been implemented by us to detect plagiarism. Section 5 presents the experiments that have been performed and their analysis. Section 6 gives an illustrative example. Section 7 concludes and proposes future directions for research.

2 Plagiarism

Plagiarism is defined in the 1995 Random House Compact Unabridged Dictionary 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."

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 421–429,
Beijing, August 2010

Self-plagiarism is the reuse of significant, identical, or nearly identical parts of one’s own work without citing the original work. In addition to the ethical issue, this phenomenon can be illegal if copyright of the previous work has been transferred to another entity. Usually, self-plagiarism is considered to be a serious ethical problem in cases where a publication needs to contain an important portion of a new material, such as in academic papers (Wikipedia, 2010).

On the other hand, it is common for researchers to rephrase and republish their research, tailoring it for different academic journals and conference articles, to disseminate their research to the widest possible interested public. However, these researchers must include in each publication a meaningful or an important portion of a new material (Wikipedia, 2010).

There are various classifications for levels of plagiarism. For instance, IEEE (2010) categorized plagiarism into five levels, or degrees, of misconduct, ranging from the most serious (Level One) to the least serious (Level Five):

Level One: The uncredited verbatim copying of a full paper, or the verbatim copying of a major portion (greater than half of the original paper)

Level Two: The uncredited verbatim copying of a large portion (less than half of the original paper).

Level Three: The uncredited verbatim copying of individual elements (e.g., paragraphs, sentences, figures).

Level Four: The uncredited improper paraphrasing of pages or paragraphs.

Level Five: The credited verbatim copying of a major portion of a paper without clear delineation (e.g., quotes or indents).

Loui (2002) handled eight allegations of plagiarism related to students' works. Collberg et al. (2005) proposes eight ranks of plagiarism.

3 Related Research

There are two main attitudes concerning discovery of similar documents: ranking and fingerprinting. Ranking methods are derived from information retrieval (IR) and are widely used in IR systems and Internet search engines. Known ranking methods are the cosine measure, the inner product, and the normalized inner product. Hoad and Zobel (2003) extended the ranking family by defining identity measures, designed for identification of co-derivative documents.

Fingerprinting aims to compare between two documents based on their fingerprints. Fingerprint methods have been used by many previous researches, e.g., Manber (1994). Heintze (1996), Lyo et al. (2001), Hoad and Zobel (2003), and Shivakumar and Garcia-Molina (1996).

3.1 Full Fingerprinting

Given a document, a full fingerprint of the document consists of the set of all the possible sequential substrings of length α in words (a definition that is based on characters is also possible). There are N–α+1 such substrings, where N is the length of the document in words. This fingerprinting selects overlapping sub-strings. For instance, if α is 3, this method selects the 3-word phrases that begin at position 0; 1; 2; etc. The size of α is known as the fingerprint granularity. This variable can have a significant impact of the accuracy of fingerprinting (Shivakumar and Garcia-Molina, 1996).

Comparing a document X to a document Y where X's size is |X| and if n is the number of substrings common to both documents then n/|X| is the measure of how much of X is contained in Y.

3.2 Selective Fingerprinting

To decrease the size of a full fingerprint, there are various versions of selective fingerprints.

The simplest kind of selective fingerprinting is the "All substrings selection" described in Hoad and Zobel (2003). This fingerprinting is similar to the full fingerprinting, but it does not select overlapping sub-strings. Rather, it selects all non-overlapping substrings of size α (in words) from the document. For example, if α is 3, this strategy selects the 3-word phrases that begin at position 0; 3; 6; 9; etc.

Heintze (1996) performed various experiments using a fixed number of fingerprints independent of the size of the document and a fixed number of substrings of size α (in characters). The best results were achieved by 1,000 fingerprints with α=50. Another possibility is to work with a fixed proportion of the substrings, so that the size of the selective fingerprint is proportional to the size of the document. The main disadvantage of this possibility is space consumption.

Hoad and Zobel (2003) suggested many additional general types of selective fingerprinting, e.g., positional, frequency-based, and structure-based.

3.3 Additional Similarity Measures

SymmetricSimilarity

Monostori1 et al. (2002) defined a measure called SymmetricSimilarity as follows:

$SS(X, Y) = |d(X) \cap d(Y)| / |d(X) + d(Y)|$

where X and Y are the two compared documents, d(X) and d(Y) are the number of the fingerprints of X and Y, respectively, and $|d(X) \cap d(Y)|$ is the number of the common fingerprints.

S2 and S3

Bernstein and Zobel (2004) defined several additional similarity measures, such as S2 and S3:

$S2(X, Y) = |d(X) \cap d(Y)| / \min(|d(X)|, |d(Y)|)$

$S3(X, Y) = |d(X) \cap d(Y)| / (d(X) + d(Y))/2)$

where $\min(|d(X)|, |d(Y)|)$ is the minimal number of the fingerprints of X and Y, respectively, and d(X) + d(Y) is the average number of the fingerprints of X and Y.

Rarest-in-document

The Rarest-in-Document method is one of the frequency-based methods defined by Hoad and Zobel (2003). This method chooses the substrings that produce the rarest substrings with length of k words in the document. This means that all of the substrings must be calculated and sorted according to their frequency in the document, and then the rarest of them are selected. The intuition is that sub-strings, which are less common, are more effective discriminators when comparing documents for similarity.

Anchor methods

Hoad and Zobel (2003) defined anchor methods. These methods are based on specific, predefined strings (called anchors), in the text of the document. The anchors are chosen to be common enough that there is at least one in almost every document, but not so common that the fingerprint becomes very large (Manber, 1994).

Various anchors were used by Hoad and Zobel. The anchors were randomly selected, but extremely common strings such as "th" and "it" were rejected. The 35 2-character anchor method detects all of the documents that were considered as similar by a human expert.

Additional experiments have been applied to identify the optimal size of an anchor. Manber (1994) used 50-character anchors in a collection of over 20,000 "readme" documents, identifying 3,620 sets of identical files and 2,810 sets of similar files. Shivakumar and Garcia-Molina (1996) achieved the best results with one-sentence anchors and Heintze (1996) achieved the best results with 1000-character anchors.

4 Baseline Detection Methods

To find whether there is a plagiarism, novel and old baseline methods have been implemented. These methods can be divided into five groups: full fingerprint methods, selective fingerprint methods, anchor methods, word comparison methods, and combinations of methods.

Full fingerprint methods

All the full fingerprint methods are defined for overlapping substrings of length k in words from the beginning of the document.

1. FF(k) - Full Fingerprints of length k
2. SSF(k) - SymmetricSimilarity for Full fingerprints of length k
3. S2F(k) - S2 for Full fingerprints of length k
4. S3F(k) - S3 for Full fingerprints of length k
5. RDF(k) - Rarest-in-Document for Full fingerprints of length k
6. CA - Compare between the abstracts of the two documents using FF(3)

Selective Fingerprint methods

In this research, all the selective fingerprint methods are selective by the sense of non-overlapping substrings of length k in words from the beginning of the document.

7. SF(k) - Selective Fingerprints of length k

8. SSS(k) - SymmetricSimilarity for Selective fingerprints of length k
9. S2S(k) - S2 for Selective fingerprints of length k
10. S3S(k) - S3 for Selective fingerprints of length k
11. RDS(k) - Rarest-in-Document for Selective fingerprints of length k

Anchor methods

We decided to work with seventy (N=70) 3-character anchors. Based on these anchors we have defined the following methods:

12. AFW - Anchor First Words - First 3-charcters from each one of the first N words in the tested document
13. AFS - Anchor First Sentences - First 3-charcters from each one of the first N sentences in the tested document
14. AF - most Frequent Anchors - N most frequent 3-charcter prefixes in the tested document
15. AR - Rarest Anchors - N rarest frequent 3-charcter prefixes in the tested document
16. ALW - Anchor Last Words - First 3-charcters from each one of the last N words in the tested document
17. ALS - Anchor Last Sentences - First 3-charcters from each one of the last N sentences in the tested document Word comparisons
18. CR - CompareReferences. This method compares between the titles of the papers included in the references section of the two examined papers.

Combinations of methods

19. CARA- CompareAbstractReferencesAverage. This method returns the average value of CA and CR.
20. CARM - CompareAbstractReferencesMin. This method returns the minimal value between CA and CR.

As mentioned above, Hoad and Zobel (2003) defined anchor methods based on the first/last N sentences/words/3-charcter prefixes in the tested document. As shown in Table 1 and in its analysis, the anchor methods are not successful, probably because they use a small portion of data. Therefore, we decided to implement methods defined for the following portions of the paper: the first third (*first*), the middle third (*middle*), and the last third (*end*) of the paper according to the number of the words in the discussed paper. All the *first*, *middle* and *end* methods use FF(3). These methods were combined with CA or CR. CA was not combined with the *first* methods because the abstract is included in the first part of the paper. CR was not combined with the *last* methods because the references are included in the end part of the paper.

21. CAMA- CompareAbstractMiddleAve. This method returns the average value of CA and FF(3) computed for the middle parts of the two examined papers.
22. CAMM - CompareAbstractMiddleMin. This method returns the minimal value between CA and FF(3) computed for the middle parts of the two examined papers.
23. CAEA - CompareAbstractEndAverage. This method returns the average value of CA and FF(3) computed for the end parts of the two examined papers.
24. CAEM - CompareAbstractEndMin. This method returns the minimal value between CA and FF(3) computed for the end parts of the two examined papers.
25. CRFA - CompareReferencesFirstAverage. This method returns the average value of CR and FF(3) computed for the first parts of the two examined papers.
26. CRFM - CompareReferencesFirstMin. This method returns the minimal value between between CR and FF(3) computed for the first parts of the two examined papers.
27. CRMA - CompareReferencesMiddleAverage. This method returns the average value of CR and FF(3) computed for the middle parts of the two examined papers.
28. CRMM - CompareReferencesMiddleMin. This method returns the minimal value between CR and FF(3) computed for the middle parts of the two examined papers.

To the best of our knowledge, we are the first to implement methods that compare special and important sections in academic papers: abstract and references: CA and CR, and combinations of them. In addition, we implemented new methods defined for the three thirds: the first (F) third, the middle (M) third, and the last (E) third of the paper. These methods were combined with CA and CR in various variants. All in total, we have defined 12 new baseline methods.

5 Experimental Results

5.1 Dataset

As mentioned above, the examined dataset includes 10,100 academic papers in computer science. Most of the papers are related to NLP and are from the last ten years. Most of the papers were downloaded from http://www.aclweb.org/anthology/.

These documents include 52,909,234 words that are contained in 3,722,766 sentences. Each document includes in average 5,262 words. The minimum and maximum number of words in a document are 28,758 and 305, respectively.

The original PDF files were downloaded using IDM - Internet Download Manager (http://www.internetdownloadmanager.com/). Then we convert them to TXT files using ghostscript (http://pages.cs.wisc.edu/~ghost/). Many PDF files were not papers and many others were converted to gibberish files. Therefore, the examined corpus contains only 10,100 papers.

5.2 Experiment I

Table 1 presents the results of the 38 implemented methods regarding the corpus of 10,100 documents. The test set includes 100 papers that were randomly chosen from the examined dataset. For each tested document, all the other 10,099 documents were compared using the various baseline methods.

The IDN, VHS, HS, MS columns present the number of the document pairs that found as identical, very high similar, high similar, and medium similar to the 100 tested documents, respectively. The IDN, VHS, HS, MS levels were granted to document pairs that got the following similarity values: 100%, [80%, 100%), [60%, 80%), and [40%, 60%), respectively. However, similar pair of papers is not always a case of plagiarism, e.g., in case where one paper cites the second one.

The first left column indicates a simple ordinal number. The second left column indicates the serial number of the baseline method (Section 4) and the number in parentheses indicates the number of the chosen words (3 or 4) to be included in each substring.

On the one hand, the anchor methods (# 12-17) tried on the interval of 70-500 anchors report on relatively high numbers of suspicious document pairs, especially at the MS level. According to our expert, these high numbers are rather exaggerated. The reason might be that such fix numbers of anchors are not suitable for detection of similar papers in various degrees of similarity.

#	#(k)	Method	IDN	VHS	HS	MS
1	1(3)	FF(3)	9	0	2	1
2	1(4)	FF(4)	9	0	1	1
3	2(3)	SSF(3)	0	0	0	9
4	2(4)	SSF(4)	0	0	0	9
5	3(3)	S2F(3)	9	0	2	2
6	3(4)	S2F(4)	9	0	1	1
7	4(3)	S3F(3)	0	0	9	0
8	4(4)	S3F(4)	0	0	9	0
9	5(3)	RDF(3)	1	5	1	3
10	5(4)	RDF(4)	1	6	0	3
11	6	CA	9	0	1	0
12	7(3)	SF(3)	9	0	0	1
13	7(4)	SF(4)	9	0	0	1
14	8(3)	SSS(3)	0	0	0	9
15	8(4)	SSS(4)	0	0	0	9
16	9(3)	S2S(3)	9	0	0	1
17	9(4)	S2S(4)	9	0	0	1
18	10(3)	S3S(3)	0	0	9	0
19	10(4)	S3S(4)	0	0	9	0
20	11(3)	RDS(3)	0	0	0	1
21	11(4)	RDS(4)	0	0	0	0
22	12	AFW	4	6	18	2772
23	13	AFS	6	3	10	708
24	14	AF	6	4	4	313
25	15	AR	4	6	19	2789
26	16	ALW	4	6	9	500
27	17	ALS	4	5	12	704
28	18	CR	9	0	1	3
29	19	CARA	8	2	1	0
30	20	CARM	8	0	2	0
31	21	CAMA	9	0	1	0
32	22	CAMM	9	0	0	1
33	23	CAEA	9	0	1	0
34	24	CAEM	9	0	0	1
35	25	CRFA	8	0	3	0
36	26	CRFM	8	0	2	0
37	27	CRMA	8	0	3	0
38	28	CRMM	8	0	1	1

Table 1. Results of the 38 implemented methods for 100 tested papers.

On the other hand, the SSF(k), S3F(k), S3S(k), RDF(k), and RDS(k) methods report on relatively very low numbers of suspicious document pairs. According to our expert, these numbers are too low. The reason for this finding might be that these methods are quite stringent for detection of similar document pairs.

The full fingerprint methods: FF(k), S2F(k) and the selective fingerprint methods SF(k), and S2S(k) present very similar results, which are reasonable according to our expert. Most of these methods report on 9 IDN, 0 VHS, 0-2 HS, and 1-2 MS document pairs. The full fingerprint methods report on slightly more HS and MS document pairs. According to our expert, these methods are regarded as the best.

Our novel methods: CA and CR also report on 9 IDN, 0 VHS, one HS, and 0 or 3 MS document pairs, respectively. The sum (10-13) of the IDN, VHS, HS and MS document pairs found by the best full and selective fingerprint methods mentioned in the last paragraph is the same sum of the IDN, VHS, HS and MS document pairs found by the CA and CR methods. That is, the CA and CR are very close in their quailty to the best methods. However, the CA and the CR have a clear advantage on the other methods. They check a rather small portion of the papers, and therfore their run time is much more smaller.

On the one hand, CR seems to be better than CA (and even the best selective fingerprint methods SF(k), and S2S(k)) because it reports on more MS document pairs, which means that CR is closer in its quality to the best full fingerprint methods. On the other hand, according to our expert CA is better than CR, since CR has more detection failures.

The combinations of CA and/or CR and/or the methods defined for the three thirds of the papers report on results that are less or equal from the viewpoint of their quality to CA or CR.

Several general conclusions can be drawn from the experimental results as follows:

(1) There are 9 documents (in the examined corpus) that are identical to one of the 100 tested papers. According to our expert, each one of these documents is IDN to a different paper from the 100 tested papers. This means that at least 9% of our random tested papers have IDN files in a corpus that contains 10, 099 files (for each test file).

(2) Several papers that have been found as IDN might be legal copies. For example: (a) by mistake, the same paper might be stored twice at the same conference website or (b) the paper, which is stored in its conference website might also be stored at its author's website.

(3) All the methods that run with two possible values of k (3 or 4 words) present similar results for the two values of k.

(4) FF(3) found as better than FF(4). FF(3) discovers 9 IDN papers, 2 HS papers, and 1 MS paper. These results were approved by a human expert. FF(4) missed one paper. One HS paper identified by FF(3) was identified as MS by FF(4) and one MS paper identified by FF(3) was identified as less than MS by FF(4). Moreover, also for other methods, variants with K=3 were better or equal to those with K=4. The main reason for these findings might be that the variants with K=4 check less substrings because the checks are done for each sentence. Substrings that end at the sequential sentence are not checked. Therefore, it is likely that additional equal substrings from the checked papers are not identified.

(5) S2F(3) discovers one more MS paper compared to FF(3). According to the human expert, the similarity measure of this paper should be less than MS. Therefore, we decided to select FF(3) as the best method.

(6) FF(3)'s run time is very high since it works on overlapping substrings for the whole papers.

(7) Our two novel methods: CA and CR are among the best methods for identification of various levels of plagiarism. As mentioned before, CA was found as a better predictor.

5.3 Selection of Methods and Experiment II

Sixteen methods out of the thirty-eight methods presented in Table 1, were selected for additional experiments. All the methods with k=4, the anchor methods, SSF, S3F, S3S, RDF, and RDS methods were omitted, due to their faulty results (as explained above). The remaining 16 methods (with k=3) are: FF, S2F, S2F, SF, S2S and all our 12 baseline methods: CA, and CR- CRMM.

Table 2 presents the results of these methods regarding the corpus of 10,100 documents. Since we selected less than half of the original methods we allow ourselves to test 1,000 documents instead of 100.

#	Method	IDN	VHS	HS	MS	Time d:h:m
1	FF	38	0	11	5	1:3:57.3
2	S2F	41	1	10	18	32:00.0
3	SF	37	1	1	6	31:12.2
4	S2	38	1	1	14	20:10.8
5	CA	38	1	11	5	09:16.7
6	CR	41	2	11	67	05:57.7
7	CARA	33	2	1	21	31:53.4
8	CARM	30	4	1	5	33:40.1
9	CAMA	38	0	5	6	11:26.5
10	CAMM	38	0	3	4	10:09.8
11	CAEA	38	0	6	7	10:42.1
12	CAEM	38	0	3	4	12:35.3
13	CRFA	32	1	3	25	54:20.7
14	CRFM	30	3	3	6	54:10.0
15	CRMA	33	2	3	25	58:52.2
16	CRMM	30	2	2	5	54:17.7

Table 2. Results of the 16 selected methods for 1,000 tested papers.

Again, according to our expert, FF has been found as the best predictive method. Surprisingly, CA achieved the second best results with one additional VHS paper. 11 HS documents and 5 MS documents have been identified by CA as by FF. The meaning of this finding is that the abstracts in almost all the simple similar documents were not significantly changed. That is, the authors of the non-IDN documents did not invest enough to change their abstracts.

CR indentified 41 documents as identical. The reason for this is probably because 3 additional papers have the same reference section as in 3 other tested papers, although these 3 document pairs are different in other sections. Furthermore, CR reports on relatively high number of suspicious document pairs, especially at the MS level. The meaning of this finding is that the references in many document pairs are not significantly different although these documents have larger differences in other sections. Consequently, combinations with CA achieved better results than combinations with CR.

A very important finding is that the run time of FF was very expensive (one day, 3 hours and 57.3 minutes) compared to the run time of CA (9 hours and 16.7 minutes). In other words, CA achieved almost the same results as FF but more efficiently.

5.4 An Error Analysis

The selected methods presented in Table 2 were analyzed according to the results of FF. Table 3 shows the distributions of false true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), regarding the 10,099 retrieved documents for the 1,000 tested document.

The false positive rate is the proportion in percents of positive test results (i.e., a plagiarism was identified by a baseline function) that are really negative values (i.e., the truth is that there is no plagiarism). The false negative rate is the proportion of negative test results that are really positive values.

#	Method	TP	FP	TN	FN
1	FF	0.534	0	99.465	0
2	S2F	0.524	0.168	99.296	0.010
3	SF	0.425	0.019	99.445	0.108
4	S2	0.435	0.099	99.366	0.099
5	CA	0.534	0.010	99.455	0
6	CR	0.534	0.663	98.801	0
7	CARA	0.386	0.178	99.287	0.148
8	CARM	0.356	0.039	99.425	0.178
9	CAMA	0.475	0	99.465	0.059
10	CAMM	0.445	0	99.465	0.089
11	CAEA	0.485	0.020	99.445	0.049
12	CAEM	0.445	0	99.465	0.089
13	CRFA	0.396	0.207	99.257	0.138
14	CRFM	0.376	0.039	99.425	0.158
15	CRMA	0.405	0.217	99.247	0.128
16	CRMM	0.366	0.020	99.445	0.168

Table 3. Distributions of the various possible statistical results.

FF is the only method that detects all cases of simple plagiarism. According to FF, there are 0.534% true positives. That is, 54 papers out of 10,099 are suspected as plagiarized versions of

54 papers of the 1,000 tested papers. This finding fits the results of FF(3) in Table 2, where there are 38 IDN, 11 HS, and 5 MS.

CA, the second best method has 0% false positives, and 0.01% false negatives, which means that CA identified one suspected plagiarized version that is really a non-plagiarized document. This finding is presented in Table 2, where CA identified 55 suspected plagiarized documents, one more than FF.

CR has 0% false positives, and 0.663% false negatives, which means that CR identified 67 suspected plagiarized versions that are really non-plagiarized documents. This finding is presented in Table 2, where CR identified 121 suspected plagiarized documents, 67 more than FF.

6 Illustrative Example

Due to space limitations, we briefly present an illustrative example of comparison between a couple of papers found as HS (High Similar) according to FF(3), the best detection method. However, this is not a case of plagiarism, since the longer paper cited the shorter one as needed and there are differences in the submission length and quality.

The tested paper (Snider and Diab, 2006A) contains 4 pages and it was published on June 06. The retrieved paper (Snider and Diab, 2006B) contains 8 pages and it was published a month later. The title of the tested paper is identical to the first eight words of the title of the retrieved paper. The authors of both papers are the same and their names appear in the same order. Most of the abstracts are the same. One of the main differences is the report of other results (probably updated results).

A relatively big portion of the beginning of the Introduction section in both papers is identical. Very similar sentences are found at the beginning of different sections (Section 2 in the 4-page paper and Section 3 in the the 8-page paper).

Many sentences or phrases from the rest of the papers are identical and some are very similar (e.g., addition of 'The' before "verbs are classified" in the abstract of the retrieved paper.

It is important to point that the authors in their 8-page paper wrote "This paper is an extension of our previous work in Snider and Diab (2006)". This sentence together with the detailed reference prove that the authors cite their previous work as required.

Concerning the references in both papers, at the first glance we found many differences between the two papers. The short paper contains only 7 references while the larger paper contains 14 references. However, a second closer look identifies that 5 out of the 7 references in the shorter paper are found in the reference section of the larger paper. Indeed, regarding the reference sections we did not find HS; but we have to remember that the larger paper include 8 pages twice than the shorter paper and therfore, more references could be included.

7 Conclusions and Future Work

To the best of our knowledge, we are the first to implement the CA and CR methods that compare two basic and important sections in academic papers: the abstract and references, respectively. In addition, we defined combinations of them. Furthermore, we implemented methods defined for the three thirds of the paper. These methods were combined with CA or CR in various variants. All in total, we have defined 12 new baseline methods.

Especially CA and also CR are among the best methods for identification of various levels of plagiarism. In contrast to the best full and selective fingerprint methods, CA and CR check a rather small portion of the papers, and therefore, their run time is much more smaller.

The success of CA and CR teaches us that most documents that are suspected as simple plagiarized papers include abstracts and references, which have not been significantly changed compared to other documents or vice versa.

There is a continuous need for automatic detection of plagiarism due to web influences, and advanced and more complex levels of plagiarism. Therefore, some possible future directions for research are: (1) Developing new kinds of selective fingerprint methods and new combinations of methods to improve detection, (2) Applying this research to larger and/or other corpora, and (3) Dealing with complex kinds of plagiarism, e.g., the use of synonyms, paraphrases, and transpositions of active sentences to passive sentences and vice versa.

References

Bernstein, Y., and Zobel, J., 2004. A Scalable System for Identifying Co-Derivative Documents. In *Proceedings of 11th International Conference on String Processing and Information Retrieval (SPIRE)*, vol. 3246, pp. 55-67.

Bretag, T., and Carapiet, S., 2007. A Preliminary Study to Identify the Extent of Self Plagiarism in Australian Academic Research. *Plagiary*, 2(5), pp. 1-12.

Collberg, C., Kobourov, S., Louie, J., and Slattery, T., 2005. Self-Plagiarism in Computer Science. *Communications of the ACM*, 48(4), pp. 88-94.

Heintze, N., 1996. Scalable Document Fingerprinting. In *Proceedings of the USENIX Workshop on Electronic Commerce*, Oakland California.

Hoad, T. C., and Zobel, J., 2003. Methods for Identifying Versioned and Plagiarised Documents. *Journal of the American Society for Information Science and Technology*, Vol 54(3), pp. 203-215.

IEEE, 2010. Introduction to the Guidelines for Handling Plagiarism Complaints. http://www.ieee.org/publications_standards/publications/rights/plagiarism.html.

Keuskamp, D., and Sliuzas, R., 2007. Plagiarism Prevention or Detection? The Contribution of Text-Matching Software to Education about Academic Integrity. *Journal of Academic Language and Learning*, Vol 1(1), pp. 91-99.

Library and Information Services, 2010. Cyprus University of Technology in Scopus, http://www.cut.ac.cy/library/english/services/references_en.html#plagiarism.

Loui, M. C., 2002. Seven Ways to Plagiarize: Handling Real Allegations of Research Misconduct. *Science and Engineering Ethics*, 8, pp. 529-539.

Lyon, C., Malcolm, J., and Dickerson, B., 2001. Detecting Short Passages of Similar Text in Large Document Collections. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 118-125.

Manber, U., 1994. Finding Similar Files in a Large File System, In *Proceedings of the USENIX Technical Conference*, pp. 1-10.

Monostori1, K., Finkel, R., Zaslavsky, A., Hodasz, G., and Patke, M., 2002. Comparison of Overlap Detection Techniques. In *Proceedings of the 2002 International Conference on Computational Science*, Lecture Notes in Computer Science, vol 2329, pp. 51-60.

Shivakumar, N., and Garcia-Molina, H., 1996. Building a Scalable and Accurate Copy Detection Mechanism. In *Proceedings of the International Conference on Digital Libraries*, pp. 160-168.

Snider, N., and Diab, M., JUNE 2006A. Unsupervised Induction of Modern Standard Arabic Verb Classes. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 153- 156, June 2006.

Snider, N., and Diab, M., JULY 2006B. Unsupervised Induction of Modern Standard Arabic Verb Classes Using Syntactic Frames and LSA. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 795- 802.

Sorokina, D., Gehrke, J., Warner, S., Ginsparg, P., 2006. Plagiarism Detection in arXiv. In *Proceedings of Sixth International Conference on Data Mining (ICDM)*, pp. 1070-1075.

Wikipedia, 2010. Plagiarism. http://en.wikipedia.org/wiki/Plagiarism.

Witten, I. H., Moffat, A., and Bell, T. C., 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, second edition.