Text Reuse with ACL: (Upward) Trends

Parth Gupta and Paolo Rosso
Natural Language Engineering Lab - ELiRF
Department of Information Systems and Computation
Universidad Politécnica de Valencia, Spain
http://www.dsic.upv.es/grupos/nle
{pgupta,prosso}@dsic.upv.es

Abstract

With rapidly increasing community, a plethora of conferences related to Natural Language Processing and easy access to their proceedings make it essential to check the integrity and novelty of the new submissions. This study aims to investigate the trends of text reuse in the ACL submissions, if any. We carried a set of analyses on two spans of five years papers (the past and the present) of ACL using a publicly available text reuse detection application to notice the behaviour. In our study, we found some strong reuse cases which can be an indicator to establish a clear policy to handle text reuse for the upcoming editions of ACL. The results are anonymised.

1 Introduction

Text reuse refers to using the original text again in a different work. The text reuse in its most general form can be of two types: verbatim (quotations, definitions) and modified (paraphrasing, boilerplate text, translation). Although, the text reuse can be legal or illegal from a publishing authority perspective about the accreditation to the original author, more importantly it involves the ethical issues, especially in the scientific work.

There is a fuzzy line between the text reuse and the plagiarism and often this line is legislative. There are no straight-forward measures to declare a work as plagiarism and hence the publishing houses usually deploy their own rules and definitions to deal with plagiarism. For example, IEEE[1] and ACM[2] both consider the reuse as plagiarism in case of:

1. unaccredited reuse of text;
2. accredited large portion of text without proper delineation or quotes to the complete reused portion.

IEEE does not allow reusing large portion of own previous work, generally referred as self reuse or self plagiarism, without delineation, while ACM allows it provided the original source being explicitly cited.

With the advent of a large number of conferences and their publicly available proceedings, it is extremely easy to access the information on the desired topic to *refer* to and to *reuse*. Therefore, it becomes essential to check the authenticity and the novelty of the submitted text before the acceptance. It becomes nearly impossible for a human judge (reviewer) to discover the source of the submitted work, if any, unless the source is already known. Automatic plagiarism detection applications identify such potential sources for the submitted work and based on it a human judge can easily take the decision.

Unaccredited text reuse is often referred to as plagiarism and there has been abundant research about the same (Bouville, 2008; Loui, 2002; Maddox, 1995). Self plagiarism is another related issue, which is less known but not less unethical.

[1] http://www.ieee.org/publications_standards/publications/rights/ID_Plagiarism.html

[2] http://www.acm.org/publications/policies/plagiarism_policy

Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 76–82,
Jeju, Republic of Korea, 10 July 2012.

There has been limited research on the nature of self-plagiarism and its limit to the acceptability (Bretag and Mahmud, 2009; Collberg and Kobourov, 2005). In theory, the technologies to identify either of them do not differ at the core and there have been many approaches to it (Bendersky and Croft, 2009; Hoad and Zobel, 2003; Seo and Croft, 2008). The text reuse can also be present in the cross-language environment (Barrón-Cedeño et al., 2010; Potthast et al., 2011a). Since few years, PAN organises competitions at CLEF[3] (PAN@CLEF) on plagiarism detection (Potthast et al., 2010; Potthast et al., 2011b) and at FIRE[4] (PAN@FIRE) on cross-language text reuse (Barrón-Cedeño et al., 2011).

In the past, there has been an attempt to identify the plagiarism among the papers of ACL anthology in (HaCohen-Kerner et al., 2010), but it mainly aims to propose a new strategy to identify the plagiarism and uses the anthology as the corpus. In this study, we are concerned about the verbatim reuse and that too in large amount, only. We identify such strong text reuse cases in two spans of five years papers of ACL (conference and workshops) and analyse them to notice the trends in the past and the present based on their year of publication, paper type and the authorship. The detection method along with the subsection of the ACL anthology used are described in Section 2. Section 3 contains the details of the carried experiments and the analyses. Finally, in Section 4 we summarise the work with remarks.

2 Detection Method

The aim of this study is to investigate the trend of text reuse, and not proposing a new method. Looking at the importance of the replicability of the experiments, we use one of the publicly available tools to detect the text reuse. First we describe the best plagiarism detection system tested in (Potthast et al., 2010) and then explain how the tool we used works similarly. The partition of the ACL anthology used for the experiments is described in Section 2.1. The details of the system along with the detection method are presented in the Section 2.2.

[3] http://pan.webis.de/

[4] http://www.dsic.upv.es/grupos/nle/fire-workshop-clitr.html

Year	Long	Short	Workshop	Total
1993	47	0	68	115
1994	52	0	56	108
1995	56	0	15	71
1996	58	0	73	131
1997	73	0	232	305
2007	131	57	340	528
2008	119	68	363	550
2009	121	93	740	954
2010	160	70	772	1002
2011	164	128	783	1075

Table 1: The year-wise list of the number of accepted papers in ACL.

2.1 Data Partition

We crawled the long and short papers of the ACL conference and all the workshop papers from the ACL anthology of the years 1990-1997 and 2004-2011. We converted all the papers from the PDF format to plain text for processing using "pdftotext" utility available with "xpdf" package in linux[5]. The bibtex files available in the anthology are used for the author analysis. We investigate the trends over two span of five years (1993-97 and 2007-11) to depict the past and the present trends. The number of papers accepted for the mentioned categories in these years are listed in Table 1.

2.2 Reuse Identification

First, we describe how the best plagiarism detection system at PAN@CLEF 2010 works. Then we show that WCopyFind[6], the tool we used, works in a similar way.

2.2.1 State-of-the-art

The best system in PAN@CLEF 2010 edition was (Kasprzak and Brandejs, 2010). The overview of the system is as follows.

1. Preprocessing: The documents are processed to normalise the terms and word 5-gram chunks are made using MD5 hashing scheme.

[5] http://linux.die.net/man/1/pdftotext

[6] WCopyFind is freely available under GNU public license at http://plagiarism.bloomfieldmedia.com/z-wordpress/software/wcopyfind/. Version 4.1.1 is used.

2. Similarity: Inverted index of these chunks is created. Then for the given suspicious document, the source documents which contain at least 20 such chunks in common, are retrieved.

3. Annotation: The boundary of the exact fragments (cases) are annotated based on the position information of the common chunks. False positives are removed by neglecting the cases where the chunks are sparse (lay far from one another).

2.2.2 WCopyFind

For the identification of text reuse, we used an open source application WCopyFind. This system works very similarly to the approach explained in Sec. 2.2.1.[7] It handles the preprocessing by user defined variables as shown in Table 2 to tokenise the terms. Then it creates the word n-grams where n = *Shortest Phrase to Match* parameter and converts the chunks into 32-bit hash codes for similarity estimation. It outputs the reuse text portions among the documents in question explicitly as shown in Fig. 1. The system extends a wide variety of parameters with word and phrase-based similarity. We used the parameter values as depicted in Table 2. Most of the parameters are self-explanatory. We used word 6-grams for the identification because the value of n=6 is suggested by the developers of WCopyFind. Parameter "Fewest Matches to Report" interprets the number of words in the matching n-grams hence it is set to 500, which practically stands for ∼85 word 6-grams. There was a high overlap of the text among the papers in the "reference" section which can not be considered as reuse. To avoid this influence, we estimated the maximum words overlap of the reference section between two papers empirically, which turned out to be 200 words. Therefore, setting the threshold value to 500 words safely avoided high bibliographical similarity based false positives. In order to confirm the reliability of the threshold, we manually assessed 50 reported cases at random, in which 48 were actually cases of text reuse and only 2 were false positives.

Parameter	Value
Shortest Phrase to Match	6
Fewest Matches to Report	500
Ignore Punctuation	Yes
Ignore Outer Punctuation	Yes
Ignore Numbers	Yes
Ignore Letter Case	Yes
Skip Non-Words	Yes
Skip Long Words	No
Most Imperfections to Allow	0

Table 2: Parameters used of WCopyFind to identify the text reuse.

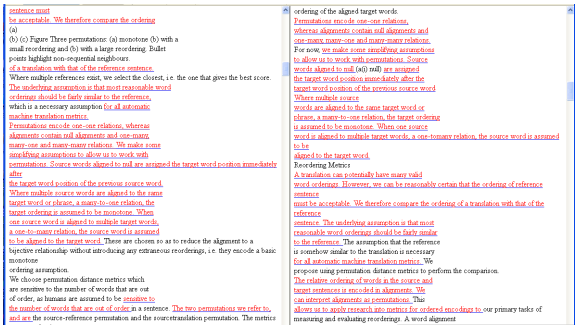

Figure 1: Screen-shot of the output of WCopyFind. The size is deliberately kept small to anonymise the case. Best viewed in color.

3 Experiments

We carried out a number of experiments to understand the nature and the trends of text reuse among the papers of ACL. These experiments were carried for papers over two spans of five years to notice the trends.

3.1 At present

In this category, we carry out the experiments on papers within the most recent five years.

I. Text reuse in the papers among the same year submissions This experiment aimed to identify the text reuse among the papers accepted in the same year. Each year, ACL welcomes the work in many different formats like long, short, demo, student session and workshop papers. This analysis reveals the same or highly similar text submitted in multiple formats.

Fig. 2 shows the number of reuse cases identified among the papers accepted in the same year.

[7] http://plagiarism.bloomfieldmedia.com/How_WCopyfind_and_Copyfind_Work.pdf

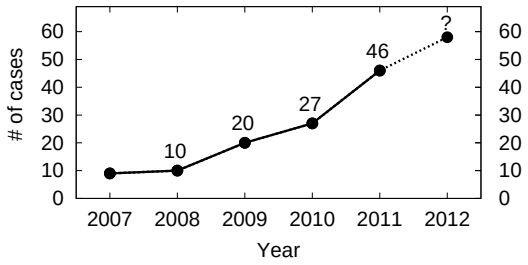

Figure 2: The text reuse cases identified among the papers of the same year submissions (span 2007-11).

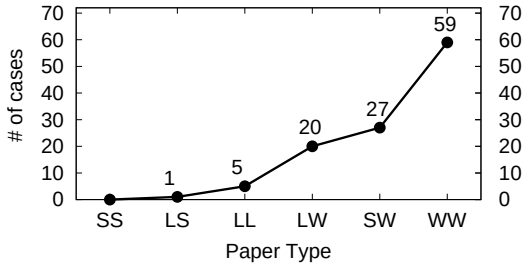

Figure 3: The text reuse cases based on the type of the papers involved. The 'L', 'S' and 'W' denote the long, short and workshop papers respectively. 'XY' refers to the cases of reuse involving one paper of type X and the other of type Y (span 2007-11).

We also analysed the types of the papers involved in these reuse cases. In the same year papers, it is difficult to decide the source and the target paper, because both are not published at the time of their review. Therefore, the number of cases based on the unordered pairs of the paper types involved in the reuse are shown in Fig. 3. It is noticeable from Fig. 2 and Table 1 that, although there is no big difference between the number of accepted papers in the last three years, the number of reuse cases are increasing rapidly. Moreover, Fig. 3 reveals that the chance of a workshop paper being involved in a reuse case with a long, short or another workshop paper is higher.

II. Text reuse in the papers from the previous year submissions This experiment aimed to depict the phenomenon of text reuse from an already published work, in this case, the ACL papers of the previous years. In this experimental setting, we considered the papers of a year 'X' as the target papers and the papers of the past three years from 'X' as the source papers. Fig. 4 depicts the reuse trend of

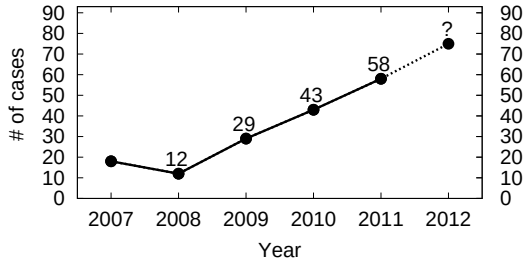

Figure 4: The text reuse cases in the papers of a year considering the papers of the past three years as the source (span 2007-11).

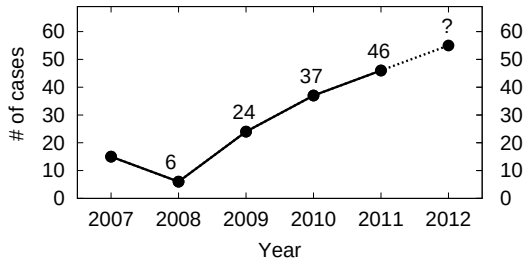

Figure 5: The text reuse cases in the papers of a year considering the papers of the immediate past year as the source (span 2007-11).

this nature over a span of five years.

We also carried a similar analysis considering only the immediate past year papers as the source. Fig. 5 presents the trend of such cases. It is noticeable from the Fig. 4 and 5 that the trend is upwards. Moreover, it is interesting to notice that the majority of the reuse cases involved the immediate past year papers as the source compared to the previous three year papers as the source.

We also analysed the trend of reuse based on the source and the target paper types and the findings are depicted in Fig. 6. Though the reuse cases involving the workshop papers are very high, there are noticeable amount of text reuse cases involving the papers where both of them (source and target) are of type long.

3.2 In retrospect

In this section we investigate the trends of text reuse in early 5 years papers i.e. papers from the span of years 1993-1997. Though the ACL Anthology contains papers from 1979, we chose this span because, for the consistency we wanted to include workshop

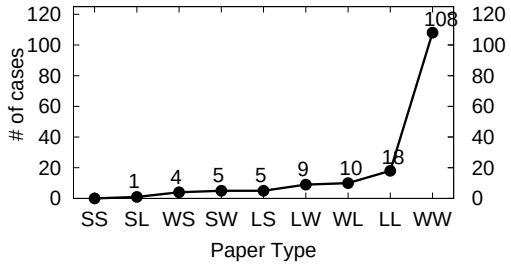

Figure 6: The text reuse trend based on the source and the target paper type. The 'L', 'S' and 'W' denote the long, short and workshop papers respectively. 'LS' refers to source is long paper and target is short paper, 'SL' refers to opposite and so on (span 2007-11).

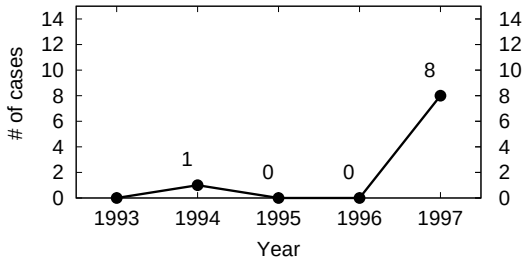

Figure 7: The text reuse cases identified among the papers of the same year submissions (span 1993-97).

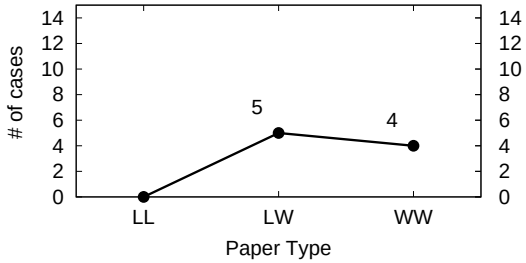

Figure 8: The text reuse cases based on the type of the papers involved. The 'L' and 'W' denote the long and workshop papers respectively. 'XY' refers to the cases of reuse involving one paper of type X and the other of type Y (span 1993-97).

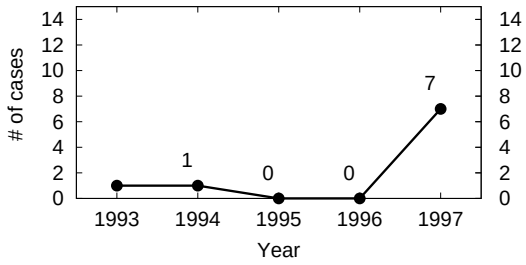

Figure 9: The text reuse cases in the papers of a year considering the papers of the past three years as the source (span 1993-97).

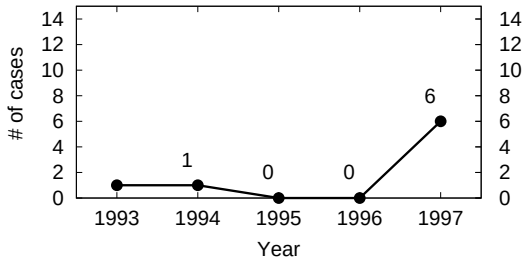

Figure 10: The text reuse cases in the papers of a year considering the papers of the immediate past year as the source (span 1993-97).

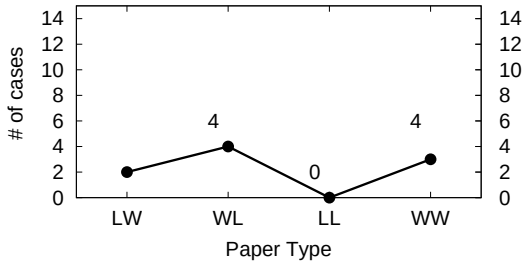

Figure 11: The text reuse trend based on the source and the target paper type. The 'L' and 'W' denote the long and workshop papers respectively. 'LW' refers to source is long paper and target is a workshop paper, 'WL' refers to opposite and so on (span 1993-97).

papers in the experiments, which only started in 1990. So our first test year became 1993 considering previous three years papers to it serving as the source.

Figs. 7, 8, 9, 10 and 11 show the behaviour in the past years for the experiments described in Section 3.1. These results are relatively low compared to the behaviour in the present. To better understand this, we present the number of text reuse cases in both the test spans as a relative frequency based on the total number of accepted papers in Table 3. It can be noticed from Table 3 that the reuse cases were quite a few in the past except the year 1997. Moreover, in the last five years the amount of text reuse cases have grown from 5.11% to 9.67%. In should also be noticed that in spite of these cases of text reuse,

a large partition of the accepted papers (more than 90%) still remains free from text reuse.

Year	Tot. Cases	Tot. Accepted	% Cases
1993	1	115	0.87
1994	2	108	1.85
1995	0	71	0
1996	0	131	0
1997	15	305	4.92
2007	27	528	5.11
2008	22	550	4.00
2009	49	954	5.14
2010	70	1002	6.99
2011	104	1075	9.67

Table 3: The relative frequency of text reuse cases over the years.

3.3 Author analysis of the reuse cases

Finally we analysed the authorship of these text reuse cases and categorised them as self and cross reuse. If the two papers involved in text reuse share at least one common author then it is considered as a case of self reuse otherwise is reffered as cross reuse. The number of the self and cross reuse cases in the last five year papers are reported in Table 4. The self reuse cases are much higher than the cross reuse cases.

We also analysed the frequency of a particular author being involved in the text reuse cases. This analysis is presented in Fig. 12. This phenomenon follows the Zipf's power law i.e. a small set of authors (635 out of 8855 = less than 10%) refer to the reported cases of reuse in the last five years. More interestingly, only 80 authors (roughly 1% of the total authors) are involved in more than 5 cases of text reuse.

Reuse Type	**No. of Cases**
Self	232
Cross	17
Total	249

Table 4: Authorship of the text reuse cases. "Self" denotes that at least one author is common in the papers involved and "Cross" denotes otherwise.

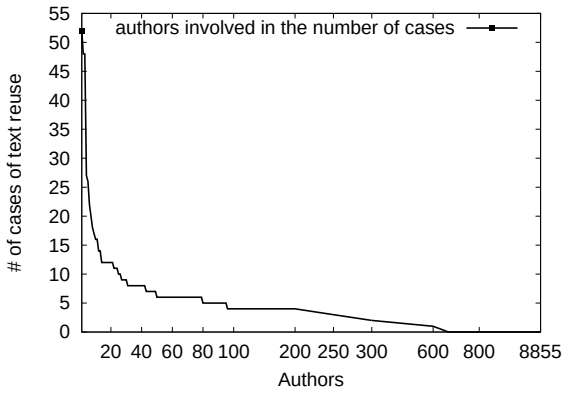

Figure 12: Involvement of an author in the number of text reuse cases.

4 Remarks

These cases are reported based on the verbatim copy of the text in the ACL proceedings only. We did not aim to detect any text reuse that is paraphrased, which in reality can not be neglected. The paraphrased cases of text reuse are even harder to detect, as remarked in (Stein et al., 2011): the state-of-the-art plagiarism detectors succeeded in detecting less than 30% of such plagiarised text fragments. Moreover, including the other major conferences and journals of the field, the number of reported cases may increase. The manual analysis revealed that, in some cases, the related work section is completely copied from another paper. There were many cases when two papers share a large portion of the text and differ mostly in the experiments and results section. This study revealed that self reuse is more prominent in the ACL papers compared to the cross reuse. The cross reuse could be a plagiarism case if the original authors are not acknowledged properly and explicitly. The ethicality and the acceptability of the self text reuse is arguable. Once more, the aim of this paper is not to judge the acceptability of the text reuse cases but to advocate the need of such systems to help in the review process. Text reuse in the same year submissions is also an eye opener because in such cases the text is novel but is used to publish in multiple formats and can stay unnoticed from the reviewers. In order to uphold the quality and the novelty of the work accepted in ACL, it is essential to implement a clear policy for text reuse and the technology to handle such reuse cases. We hope this work will help the ACL research commu-

nity to consider handling the text reuse for the upcoming editions.

Acknowledgment

This work has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, and by the Text-Enterprise 2.0 research project (TIN2009-13391-C04-03). We thank Rafael Banchs for his suggestions and ideas.

References

Alberto Barrón-Cedeño, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, Beijing, China, August 23-27.

Alberto Barrón-Cedeño, Paolo Rosso, Shobha Devi Lalitha, Paul Clough, and Mark Stevenson. 2011. Pan@fire: Overview of the cross-language !ndian text re-use detection competition. In *In Notebook Papers of FIRE 2011*, Mumbai, India, December 2-4.

Michael Bendersky and W. Bruce Croft. 2009. Finding text reuse on the web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 262–271, New York, NY, USA. ACM.

Mathieu Bouville. 2008. Plagiarism: Words and ideas. *Science and Engineering Ethics*, 14(3).

Tracey Bretag and Saadia Mahmud. 2009. Self-plagiarism or appropriate textual re-use? *Journal of Academic Ethics*, 7(3):193–205.

Christian Collberg and Stephen Kobourov. 2005. Self-plagiarism in computer science. *Commun. ACM*, 48(4):88–94, April.

Yaakov HaCohen-Kerner, Aharon Tayeb, and Natan Ben-Dror. 2010. Detection of simple plagiarism in computer science papers. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 421–429, Beijing, China.

Timothy C. Hoad and Justin Zobel. 2003. Methods for identifying versioned and plagiarized documents. *J. Am. Soc. Inf. Sci. Technol.*, 54(3):203–215, February.

Jan Kasprzak and Michal Brandejs. 2010. Improving the reliability of the plagiarism detection system - lab report for pan at clef 2010. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *CLEF (Notebook Papers/LABs/Workshops)*.

Michael C. Loui. 2002. Seven ways to plagiarize: handling real allegations of research misconduct. *Science and Engineering Ethics*, 8(4):529–539.

John Maddox. 1995. Plagiarism is worse than mere theft. *Nature*, 376(6543):721.

Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. 2010. Overview of the 2nd international competition on plagiarism detection. In *Notebook Papers of CLEF 2010 LABs and Workshops*, CLEF '10, Padua, Italy, September 22-23.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011a. Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1):45–62.

Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011b. Overview of the 3rd international competition on plagiarism detection. In *Notebook Papers of CLEF 2011 LABs and Workshops*, CLEF '11, Amsterdam, The Netherlands, September 19-22.

Jangwon Seo and W. Bruce Croft. 2008. Local text reuse detection. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 571–578, New York, NY, USA. ACM.

Benno Stein, Martin Potthast, Paolo Rosso, Alberto Barrón-Cedeño, Efstathios Stamatatos, and Moshe Koppel. 2011. Fourth international workshop on uncovering plagiarism, authorship, and social software misuse. *SIGIR Forum*, 45(1):45–48, May.