Identification of Plagiarism using Syntactic and Semantic Filters

Vijay Sundar Ram, R[1], Efstathios Stamatatos[2], and Sobha Lalitha Devi[1]

[1]AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India

[2]Dept. of Information and Communication Systems Eng. University of the Aegean, 83200–Karlovassi, Greece

sundar@au-kbc.org, stamatatos@aegean.gr, sobha@au-kbc.org

Abstract. We present a work on detection of manual paraphrasing in documents in comparison with a set of source documents. Manual paraphrasing is a realistic type of plagiarism, where the obfuscation is introduced manually in documents. We have used PAN-PC-10 data set to develop and evaluate our algorithm. The proposed approach consists of two steps, namely, identification of probable plagiarized passages using dice similarity measure and filtering the obtained passages using syntactic rules and lexical semantic features extracted from obfuscation patterns. The algorithm works at sentence level. The results are encouraging in difficult cases of plagiarism that most of the existing approaches fail to detect.

Keywords: Manual paraphrasing, Syntactic rules and Lexical Semantics, Plagiarism detection

1 Introduction

Manual paraphrasing concerns the transformation of an original text, so that the resulted text has the same meaning as the original, but with significant differences in wording and phrasing. Detecting manual paraphrasing in plagiarism cases is challenging, since the similarity of the original text with the re-written text is purposefully hidden. The exponential increase of unstructured data on the web provides multiple sources for plagiarism. The need for accurate plagiarism detection is vital to ensure originality of text with applications in areas such as publishing, journalism, patent verification, academics etc. There are many commercial plagiarism detection tools such as Turnitin and prototype systems such as COPS (Copy Protection System), SCAM (Stanford Copy Analysis Mechanism), and MOSS (Measure of Software similarity) [6]. Yet identification of manual paraphrasing is a challenge to available plagiarism detection tools since most of them are only able to detect easy, copy-and-paste cases. The detection of plagiarism when paraphrasing is used requires under-

standing of re-ordering of phrases governed by syntactic rules and use of synonym words governed by lexical semantics.

Several researchers have studied plagiarism detection for the past few decades. Plagiarism detection gained more focus and geared up with the yearly automatic plagiarism detection competition started in 2009 in the framework of PAN evaluation campaigns. In these competitions, various text re-use issues such as external and intrinsic plagiarism detection as well as cross-lingual plagiarism detection [10]. The competition started in 2009 (PAN 2009) with a small corpus. In 2010 (PAN 2010), it evolved with huge set of suspicious and source data (PAN-PC-10) and continued similarly in the year 2011 (PAN 2011). The overview report of PAN 2010 shows, 18 participants participated in the competition. They have followed similar steps in the detection of plagiarized passages, namely, candidate (or source) retrieval, detailed analysis (or text alignment) and post processing.

In the candidate retrieval step, most of the participants reduced the search space by removing the source document which does not have significant similarities with the suspicious documents. In this step, the used techniques are based on such as finger prints and position of data, finger prints with threshold are used. Information retrieval, document fingerprinting, string kernel matrices, word and character n-grams,. In the detailed analysis step some of the techniques used were winnowing fingerprinting, FastDoCode technique, cosine similarity, and jaccard similarity. Post processing techniques were focused on filtering the detected passages using chunk ratio with a predefined threshold and n-gram match, using different similarity measures [10].

In PAN-2011, the participants used approaches similar to those used in PAN-2010. They have improved the efficiency and the processing time of their approaches in comparison with systems in PAN-2010 [10]. In PAN-2012 and PAN-2013, the competition was remodeled with smaller datasets to focus on individual steps. Evaluation was done at different steps [8, 9]. We will look in detail the various plagiarism detection approaches using semantic, syntactic and structural information, focused on identifying manual paraphrasing.

Palkovskii et al. [7] have used a semantic similarity measure to detect the plagiarized passages. Semantic similarity measure used hyponym taxonomy in WordNet for calculating the path length similarity measure. Uzner et al. [15] used a low-level syntactic structure to show linguistic similarities along with the similarity measure based on tf-idf weighted keywords. Here they have used Levin's verb classes. A fuzzy semantic string similarity algorithm for filtering the plagiarized passages was used by Alzahrani et al [1]. The authors had mentioned that their approach didn't work for all levels especially for higher obfuscation, which includes manual paraphrasing. Chong and Specia [3] used a lexical generalization approach using WordNet to generalize the words and performed n-gram based similarity measure, namely overlap co-efficient, to identify the plagiarized passages. Stamatatos [14] used structural information for identifying the plagiarized passages. He used the stop-words to identify the structure of sentence. This approach varied from most of the other approaches where the content words (nouns, verb, adjective, and adverb) were considered important and the stop-words were removed as these words occur frequently.

In our work, we try to identify the manual paraphrasing using lexical semantics and syntactic rules. The paper is organized as follows: In the next section, we describe our approach, where we elaborate on identification of probable plagiarized passages and filtering the irrelevant passages with lexical semantics and syntactic rules. In the third section, we demonstrate our experiment and describe our results. The paper ends with the concluding section.

2 Our Approach

We present an approach for identifying the manual paraphrasing. Our approach works at sentence level. We detect the plagiarized passages in a two step approach. In the first step, we try to find all probable plagiarized passages using dice similarity measure. In the second step, we identify the correct passages from the probable plagiarized passages obtained in the first level using lexical semantics and syntactic rule based filters. The approach is described in detail in the following sections.

2.1 Retrieval of Probable Plagiarized Passages

In this step, we try to identify all probable plagiarized passages from a given suspicious document in comparison with the source documents. Here we identify the plagiarized sentences in the suspicious documents using dice similarity measure and group the sentences into plagiarized passages using heuristic rules. We start by preprocessing the suspicious and source documents with a tokeniser and sentence splitter. We remove the function words, connectives, pronouns from the text, making the sentences incoherent.

Identification of Similar Sentences.
We have collected the sentences with common bigram words from the suspicious and the source documents and performed similarity comparison using dice similarity measure. Dice similarity (Q_s) is defined as follows,

$$Q_s = 2C/(A+B)$$

where A and B are the number of words in sentence A and B, respectively, and C is the number of words shared by the two sentence; Q_S is the quotient of similarity and ranges from 0 to 1 [4].

We have used dice similarity measure instead of cosine similarity measure with tf-idf as weights. The cosine similarity measure provides a smaller score, if a sentence has many common words and a rarely occurred word.
Consider the following sentences,

1. *This was written by E.F. in 1627 and printed exactly as the original.*
2. *Written by E.F. in the year 1627, and printed verbatim from the original.*

The dice similarity score between sentence 1 and 2 is 0.769, whereas the cosine similarity score is 0.462. The cosine similarity score is smaller as the term ‘verbatim’ has very high inverse-document frequency value and the denominator value in cosine similarity becomes big.

The steps followed in this task are described in the algorithm below.

1. Sentences with common bigram words in the suspicious and source documents are collected.

2. Between the pair of sentences from the suspicious and source documents having bigram words, dice similarity comparison is performed. Comparison is done between the following set of suspicious and source pair sentences.

   (a) One suspicious and source sentence having a common word bigram.

   (b) Two consecutive suspicious sentences and one source sentence having a common word bigram.

   (c) One suspicious sentence and two consecutive source sentences having a common word bigram.

3. Those suspicious-source pairs with similarity greater than a predefined threshold *t* are collected to form plagiarized passages.

Formation of Probable Plagiarized Passages.
In this step, we try to group the suspicious-source sentence pairs which are greater than the predefined threshold *t* into plagiarized passages. We group the suspicious-source sentence pairs with another pair which have neighboring sentence in both suspicious document and source document into a plagiarized passage. Here we also consider the next neighboring sentence to form a passage. We describe the steps in detail in the algorithm given below.

Passage Forming Algorithm.

1. For each of the suspicious–source document pair which have suspicious-source sentence pairs having dice score greater than the predefined threshold *t*, do steps 2-7.

2. For each pair from the probable pairs of suspicious-source sentences collected using dice similarity, do step 3-6.

3. Compare this pair with the rest of the probable pairs, consider the pair has suspicious sentence x and source sentence y.

4. If the current suspicious-source sentence pair has another suspicious-source sentence pair which is consecutive to the current pair i.e., (x+1, y+1), group these suspicious sentences and source sentences to form a new passage, (x, x+1; y, y+1).

5. If a suspicious-source pair has a consecutive suspicious sentence (x+1) and source sentence which is the sentence following the consecutive sentence y+2, to

the current pair, group the suspicious sentences and source sentences to form a new passage (x, x+1; y, y+1, y+2).

6. If a suspicious-source pair has a suspicious sentence which is the sentence following the consecutive sentence x+2 and consecutive source sentence (y+1), to the current pair, group the suspicious sentences and source sentences to form a new passage (x,x+1,x+2; y, y+1).

7. Repeat steps 3 to 7 for the set of new passages till there are no pairs to group.

By performing the steps in the algorithm above, we form passages having from one sentence to n sentences. Here n is determined based on the clustering of neighboring sentences having similarity score greater than the threshold t.

2.2 Filtering of Plagiarized Passages based on Syntactic Rules and Lexical Semantics

The methods mentioned in previous work for filtering out the irrelevant suspicious-source passages were using similarity measures such as jaccard, cosine similarity or dice similarity with a given threshold. In manual paraphrasing, these similarity measures fail to filter out the irrelevant passages without drastically affecting the true positive passages, as these plagiarized passages are well rewritten. To filter out the irrelevant passages, we manually analysed the manual paraphrased passages in PAN-PC-10 training data and identified a set of patterns of changes performed by people, while plagiarizing a text. From these patterns we came up with a set of syntactic rules and lexical semantic features. These lexical semantics and syntactic rules are helpful in identifying the manual paraphrased passages. These rules obtained by analysing the training corpus are then used in identifying the correct passages in the test corpus.

The set of identified patterns and rules to handle them are presented below in this section. As the rules are based on syntactic information, all probable passages are preprocessed with a POS tagger [2] and a text chunker [13]. Pleonastic ‘it’ is also identified in the passages using a CRFs engine [5].

Synonym Substitution.

While plagiarizing a sentence, people tend to substitute nouns, verbs, adjectives and adverbs by its equivalent synonym words. Verbs are also substituted by combinations of verbs and prepositional phrases. For example 'kill' can be substituted by 'put to death'. Phrases are also replaced by its semantically-equivalent phrases. By observing the data, this type of synonym substitution covers 75.71% of changes.

Consider the following sentences:

3.a *This question is linked closely to the often-debated issue of the Pointed Style's beginnings.*

3.b *This Query is, of course, intimately connected with the much-disputed question of the origin of the Pointed Style itself.*

Here, 3.a is the suspicious sentence and 3.b is the source sentence. By comparing sentences 3.a and 3.b, we see that the noun ‘query’ is substituted with 'question', ‘origin’ is substituted with ‘beginning’. The noun phrase ‘much-disputed question’ is substituted with ‘often-debated issue’ and the verb phrase ‘intimately connected’ is substituted with ‘linked closely’.

We handle this synonym substitution using Rule 1.

Rule1:

a. Check for common words in the preprocessed suspicious and source sentence.

b. if common words exists then do the following steps:

1. The words in the suspicious sentences which are not common with the source sentences are examined and their synonyms are obtained from the WordNet, taking into account the POS category of the word. The synset words obtained are matched with words in the source text and the matched synonyms replace initial words in the suspicious sentence. While performing comparison with the source sentence, synset words are matched with words in the source sentence in the position corresponding to the suspicious sentence.

2. Similarly, we obtain equivalent phrases from the phrase dictionary, which is built using the WordNet synonym dictionary, for the phrases in the suspicious sentence and match with the equivalent phrases in the source sentence. If the equivalent phrase exists in the source sentence, then in the suspicious sentence the phrase is replaced by its equivalent phrase.

Re-ordering of Phrases.

When plagiarizing a sentence, people tend to re-order prepositional phrases in sentences by moving it to the end of the sentence or from the end of the sentence to the start of the sentence. Moreover, the position of adverbs are re-ordered within verb phrases, the position of adjectives and nouns are re-ordered within noun phrases and possessive nouns are changed into prepositional phrase following its head noun and vive-versa. This re-ordering of phrases cover 10% of the introduced changes. The re-ordering of phrases is explained with the example sentences below.

4.a *After the treaty of Tilsit, Emporer Alexander took control of Finland.*

4.b *The Emperor Alexander possessed himself of Finland after the treaty of Tilsit.*

Sentence 4.a is the suspicious sentence and 4.b is the source sentence. In the sentence 4.a and 4.b, the prepositional phrase, ‘after the treaty of Tilsit’ which occurs in the end of the source sentence is moved to the start of the suspicious sentence.

5.a *I saw a little of the Palmyra's inner life.*

5.b *I saw something of the inner life of Palmyra.*

Here sentences 5.a and 5.b are suspicious and source sentence respectively. In the source sentence, a noun phrase followed by a prepositional phrase “the inner life of Palmyra” is replaced in the suspicious sentence by a possessive noun following a head noun “Palmyra's inner life”.

This re-ordering of phrases is handled by Rule 2.

Rule 2 .

1. If the suspicious and source sentences have common prepositional phrases, check its position.

   (a) If the position of the prepositional phrase varies then re-order the phrase exactly as it has occurred in the source sentence.

2. If either the suspicious or the source sentence has a possessive noun phrase and the other sentence has the head noun of a possessive noun phrase with prepositional phrase, then normalize possessive noun phrase as it has occurred in the source sentence.

3. If a noun phrase exists in suspicious and the source sentences with common words but different ordering, then re-order the adjectives and nouns as it has occurred in the source sentence.

4. If a verb phrase exists in suspicious and the source sentences with common words but different ordering, then re-order the adverb and verb as it has occurred in the source sentence.

Introduction of Pleonastic 'it'.

If the source sentence starts with a phrase such as 'by pure luck', in the plagiarized sentence it may be re-written with pleonastic 'it' as subject 'It was pure luck'. This type of introduction of pleonastic 'it' is 2.14% of the total changes.

This is explained with sentence 6.a and 6.b, where 6.a is the suspicious sentence and 6.b is source sentence.

6.a *It was pure luck that I made that shot.*

6.b *By good luck I hit.*

In the above sentence 6.a and 6.b., the source sentence starts with a prepositional phrase and does not have a subject and in the suspicious sentence a null subject 'it' is introduced. This is handled by Rule 3.

Rule 3:

a. Check if occurrences of 'it' in the suspicious or source sentence is marked as pleonastic 'it'.

b. If the pleonastic 'it' occurs in the suspicious sentence and not in the source sentence, then remove 'it'.

Antecedent Replacement.

The pronouns in source sentences may be replaced by its antecedent while plagiarizing the sentence. This covers 2.85% of the total changes.

Consider the following sentences:

7.a *Madame Omar was a lovely German woman who's husband kept her all but locked away in the harim.*

7.b *She was a charming German lady ; but her husband kept her secluded in the harim like a Moslem woman.*

Here the sentence 7.a is the suspicious sentence and 7.b is the source sentence. In the suspicious sentence, the pronoun 'she' is replaced by 'Madame Omar'. The suspicious sentence also has adjective (charming -> lovely), noun (lady -> woman) and verb (secluded -> locked) synonym substitutions and an adjunct drop (like a Moslem woman). The antecedent replacement is handled by Rule 4.

Rule 4:

If in a given pair of suspicious and source sentences a noun phrase exists in the place of a pronoun and if the context of the noun phrase and the pronoun are same, then replace the pronoun with the noun phrase.

Other Observed Patterns.

In this section we list the set of extracted patterns which we have not handled in this study. These cover 9.26% of the changes.

Rewriting the sentence completely.

When plagiarizing the sentence, people assimilate the sense conveyed in the sentences and they re-write in a completely different manner.

Consider the example 8.a and 8.b. Sentence 8.a is plagiarized from sentence 8.b. Here the sentence 8.a and 8.b does not have any structural similarity.

8.a *Even if the man is skilled enough to hook the fish, he could catch the fish from the rough sea merely because of his luck.*

8.b *No mortal skill could have killed that fish.*

Addition / Reduction of the descriptions in Sentences.

While plagiarizing a sentence, people tend to add more or reduce the descriptions in the plagiarized sentence, as in example sentences 9.a and 9.b.

9.a *The proof that a wire may be stretched to long that the current will no longer have enough strength to bring forward at the station to which the despatch is made known.*

9.b *It is evident, therefore, that the wire may be continued to such a length that the current will no longer have sufficient intensity to produce at the station to which the despatch is transmitted those effects by which the language of the despatch is signified.*

Here sentence 9.a is the suspicious sentence and 9.b is the source sentence. In the suspicious sentence, there is a reduction in the description.

The Proposed Algorithm.

Using the lexical semantics and syntactic rules mentioned in the previous section, we try to filter out the irrelevant passages from the retrieved probable passages. The steps involved in this process are described in detail in the algorithm given below.

1. For each sentence in the suspicious passage do step 2.

2. Compare the suspicious sentence with all sentences in the source passage.

(a) Compare suspicious sentence and source sentence, check if the suspicious sentence has prepositional phrase re-ordering, possessive noun re-ordering, adverb re-ordering or adjective – noun re-ordering, then apply Rule 2.

(b) To correct the synonym substitutions in the suspicious sentence, apply Rule 1.

(c) If the suspicious sentence has pleonastic 'it', then apply Rule 3.

(d) Compare the suspicious and the source sentence, if there is a pronoun in a sentence and noun phrase in another sentence in the same position, then apply Rule 4.

(e) Similarity between suspicious and source sentences after applying the above rules is measured based on similarity of syntactic features and their positions such as noun phrase, verb phrase, proportional phrase, and their order of occurrence.

3. Similar suspicious and source sentences are identified.

4. If there exists a set of similar sentences in the suspicious and source passage, then this plagiarized passage is considered as probable plagiarized passage.

3 Experiments and Results

We have used PAN-PC-10 test data set for evaluating our algorithm for detecting manual paraphrasing [12]. This dataset has 15,925 suspicious documents and 11,147 source documents. In the suspicious dataset 50% of the documents do not have plagiarized passages. In the suspicious documents with plagiarism, the plagiarized passages are classified into three major types: passages with artificial obfuscation, passages where the obfuscation was introduced by translated passages, and passages with manually paraphrased obfuscation. In this evaluation, we have considered suspicious document with only manual paraphrasing, which counts to 397 suspicious documents. We have tested our algorithm with 397 suspicious documents having manual paraphrasing and 11,147 source documents. The evaluation is done using the performance metrics used in PAN-2010 evaluation campaign [11]. First, macro-average precision and recall are calculated at the passage level. In addition, granularity measures the ability of the the plagiarism detection algorithm to detect a plagiarized passage as a whole or as several pieces. Then, precision, recall and granularity are combined in the overall performance score called Plagdet.

In table 1, we have presented the scores obtained by the top 9 out of 18 participants in PAN2010 competition for these 397 suspicious documents with manual paraphrasing along with the scores obtained by our approach. The scores in table 1 are generated using the runs of each participants and the evaluation program provided at the PAN-2010 website.

Table 1. Scores of PAN 10 participants on suspicious documents with manual paraphrasing

Participant	Plagdet Score	Recall	Precision	Granularity
Our Approach	*0.4804*	*0.3914*	*0.8234*	*1.15*
Muhr	0.387	0.244	0.938	1.0
Grozea	0.322	0.198	0.944	1.0
Zou	0.290	0.172	0.948	1.0
Oberreuter	0.283	0.178	0.691	1.0
Kasprzak	0.231	0.131	0.954	1.0
Torrejon	0.230	0.133	0.836	1.0
Palkovskii	0.115	0.062	0.938	1.02
Sobha	0.061	0.031	0.889	1.0
Gottron	0.016	0.009	0.809	1.25

The low performance scores in table 1 show the need of an effective algorithm for detecting manual paraphrasing, which is the more realistic scenario in plagiarism detection. Our algorithm substantially improves recall while precision remains relatively high. Identification of probable plagiarized passages with a low similarity score helps in getting better recall.

Table 2. Scores of our approach after step 1 (without filtering)

	PlagdetScore	Recall	Precision	Granularity
Without lexical semantics and syntactic filtering	0.1954	0.4314	0.1553	1.25

The syntactic rule and lexical semantics filtering of the probable plagiarized passages helps in getting a high precision without largely disturbing the recall. This is shown from table 2. Before filtering the probable plagiarized passages using lexical semantics and syntactic rules, the precision is low. The filtering helps in removing the irrelevant passages without drastically disturbing the recall.

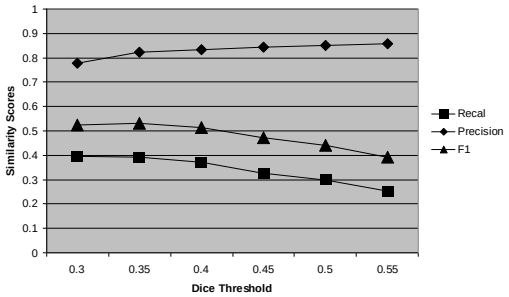

Fig. 1. Performance with varying similarity thresholds.

Figure 1 shows the performance of our approach, in terms of recall, precision and F1 for varying threshold values used in retrieving probable sentences. Choosing the dice similarity threshold as 0.35 enhances the recall by extracted most of the probable plagiarized passages. We empirically derived and used this threshold in the reported experiments. In this study, we have not handled completely rewritten sentences and

sentences with additional description. This affects the recall slightly, as the filtering algorithm filters out the complete rewritten sentences, though the passages may have many common words, which are identified in the first step. Improving the phrase to word and phrase to phrase dictionaries will help in improving the precision as well as the recall.

Figure 2 shows precision and recall while using different features in filtering the probable passages. The results show an increase in recall when both lexical semantics and syntactic rules are used in filtering the probable passages. When one of the features either lexical semantics or syntactic rules is used in filtering there is a drastic reduction in recall as compared to the recall achieved when both features are used together in filtering. We also observe that the use of lexical semantic features in filtering gives greater recall than filtering using syntactic rules.

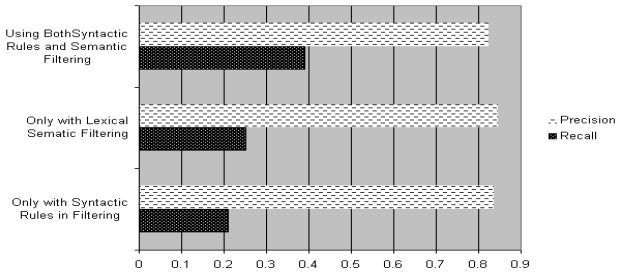

Fig. 2. Comparison of performance with different filtering features

4 Conclusion and Future Work

We have presented a two step approach to detect simulated plagiarized passages. In the first step, we extract the probable plagiarized passages using dice similarity measure from the sentences in the suspicious and source documents having common word bigrams. We maintain the similarity score between the suspicious and source sentence as low as 0.35 to find most of the probable plagiarized passages, which helps in getting better recall. In the second step, we have used lexical semantics and syntactic rules for filtering the relevant plagiarized passages, which helps in achieving high precision without largely harming the recall. Normalising the words using synonym dictionaries at the initial stage will help in boosting the recall drastically. We have not handled sentences which are completely rewritten. We are planning to take up the above mentioned tasks as our future work.

Reference

1. Alzahrani, S., and Salim, N.: Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection: Lab Report for PAN at CLEF 2010. In: Notebook Papers of Labs and Workshops CLEF'10, Padua, Italy, (2010)

2. Brill, E.: Some Advances in transformation Based Part of Speech Tagging. In Proceedings of the Twelfth International Conference on Artificial Intelligence (AAAI-94), Seattle, WA. (1994)
3. Chong, M. and Specia. L.: Lexical Generalisation for Word-level Matching in Plagiarism Detection. In: Recent Advances in Natural Language Processing, pp 704–709, Hissar, Bulgaria, (2011)
4. Dice, Lee R.: Measures of the Amount of Ecologic Association Between Species. In: Ecology 26 (3): 297–302 (1945)
5. Lalitha Devi, S., Vijay Sundar Ram and Pattabhi RK Rao.: Resolution of Pronominal Anaphors using Linear and Tree CRFs. In. 8th DAARC, Faro, Portugal, (2011)
6. Pakinee Aimmanee.Automatic Plaiarism Detection Using Word-Sentence Based S-gram. In: Chiang Mai Journal of Science, Vol. 38 (Special Issue), pp. 1-7 (2011)
7. Palkovskii, Y., Belov, A., Muzyka, I.: Using WordNet-based Semantic Similarity Measurement in External Plagiarism Detection - Notebook for PAN at CLEF (2011)
8. Potthast M., Hagen M., Gollub T., Tippmann M., Kiesel J., Rosso P., Stamatatos E., Stein B.: Overview of the 5th International Competition on Plagiarism Detection. In: Forner P., Navigli R., Tufis D. (Eds.), Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013, Valencia, Spain, September 23-26 (2013)
9. Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P. and Stein, B.: Overview of the 4th International Competition on Plagiarism Detection. In: Forner, P., Karlgren, J. and Womser-Hacker, C (Eds), CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, September 2012. (2012)
10. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B. and Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: Petras, V., Forner, P. and Paul D. Clough, (Eds), Notebook Papers of CLEF 11 Labs and Workshops (2011)
11. Potthast M., Barrón-Cedeño A., Stein B., Rosso P.: An Evaluation Framework for Plagiarism Detection. In: Proc. of the 23rd Int. Conf. on Computational Linguistics, COLING-2010, Beijing, China, August 23-27, pp. 997-1005 (2010)
12. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B. and Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In Braschler, M., Harman, D. and Pianta, E. (Eds), Notebook Papers of CLEF 10 Labs and Workshops, September 2010. (2010)
13. Ngai, G., Florian, R.: Transformation-Based Learning in the Fast Lane. In: NAACL'2001, Pittsburgh, PA, pp. 40-47 (2001)
14. Stamatatos, E.: Plagiarism Detection Using Stopword n-grams. Journal of the American Society for Information Science and Technology, 62(12), pp. 2512-2527, Wiley, (2011)
15. Uzuner, O., and Katz, B., and Nahnsen, T.: Using Syntactic Information to Identify Plagiarism. In: 2nd Workshop on Building Educational Applications using NLP (2005)