Putting Ourselves in SME's Shoes: Automatic Detection of Plagiarism by the WCopyFind tool

Enrique Vallés Balaguer
Private Competitor
enriquevallesbalaguer@gmail.com

Abstract: Thanks in part, to the large amount of information circulating today on the Internet, unfortunately, the plagiarism has become a very common practice, up to become one of the biggest problems of today's society. One of the most affected sectors by the plagiarism are small and medium entreprises (SME's), which are daily victims from their competitors. Finding a system able to detect plagiarism in texts, has become a major goal for the interests of SME's, which are forced to solve the problem through the tools available on the web. In this paper we analyze the results obtained in the PAN'09 competition with the WCopyFind tool.

Keywords: Plagiarism detection, WCopyFind

1 Introduction

Internet is one of the greatest advances in history in the area of communication. Thanks to (the) Internet, you can have immediate access to information, regardless of the distances. However, the easy access to information, has increased the number of plagiarism cases.

Within the business area must be emphasized the importance of the automatic plagiarism detection for SME's. For SME's is vital to know if their proposals, products, ideas, etc, have been plagiarized by competitors. To solve this problem, the companies have mainly to rely on the software available on the web. In this paper, we attempt using the software WCopyFind (Dreher, 2007) developed in the University of Virginia.

2 Plagiarism detection for SME's

SME's build web pages to enter information about themselves, advertise their products, etc..., to approach (to) the consumer. But the information on the web is also visible for the competitors. When a company launches a new tool, this is discovered by competitors within a few hours or days. But, there are companies that use this information to copy. The automatic plagiarism detection aims to try to find an automated approach that is able to locate fragments of texts suspects of plagiarism.

Currently the automatic plagiarism detection is divided into two different branches. By one side, is the *external plagiarism analysis*, which requires a set of original sources from which seeking possible plagiarized fragments in suspicious texts. Within this branch, there are methods developed with the intention to locate fragments suspected of plagiarism through search strategies.

Given the large amount of information available at present, comparing a suspected document with all the available ones is a virtually unmanageable task. Therefore, emerged the *intrinsic plagiarism analysis*, tries to rely on the suspected document. Its intention is to capture the style and the complexity of a document with the aim of finding unusual fragments that are candidates to be instances of plagiarism (Barrón-Cedeño and Rosso, 2009).

2.1 WCopyFind

WCopyFind[1] is a software developed in 2004 by Bloomfield at the University of Virginia. To detect suspicious fragments of plagiarism, WCopyfind conducts a search through the comparison of n-grams.

Since WCopyfind works with n-grams, language is not important and matches are

[1] http://plagiarism.phys.virginia.edu/

Stein, Rosso, Stamatatos, Koppel, Agirre (Eds.): PAN'09, pp. 34-35, 2009.

n-gram	Precision	Recall
4	2.05 %	66.34 %
5	11.34 %	59.08 %
6	17.85 %	57.06 %

Table 1: Training Phase

readily identified from the candidate documents submitted for analysis (Dreher, 2007).

3 Corpus

The PAN'09 corpus which refers to the External Plagiarism Analysis task, consists mainly of documents in English, in which you can find any type of plagiarism.

There are a total of 7,214 suspicious documents, which may contain plagiarized fragments from one or more original documents or do not contain any plagiarized fragment at all. On the other hand, the number of original documents that constitute the corpus is 7,215.

4 Results

Due to the fact that the WCopyFind tool allows the user to select the size of the n-grams, before carrying out the analysis on the competition corpus, we have made several experiments with training corpus to find the appropriate size of the n-grams. Table 1 shows the results for each one of the experiments. We can highlight several interesting points. By one side it is noteworthy that contrary to other language engineering tasks, we must stress that the obtained precision is smaller than the obtained recall.

Another interesting fact observed in Table 1 is that, how much smaller size of n-grams is, the smaller is the precision. However, it happens all the contrary to the measure of recall, that is, the smaller the n-grams, the greater is the recall. This is because, the smaller are the n-grams, the greater is the possibility of finding similar fragments in plagiarized documents. In (Barrón-Cedeño and Rosso, 2009), the authors analyzed this fact, and they showed that the probability of finding common n-grams in different documents decreases as n increases.

Finally, we have taken the decision that the best size for the n-grams was *hexagrams*, because there is no great loss with to respect of recall and it has the best result in precision.

Software	Precision	Recall
WCopyFind	1.36 %	45.86 %

Table 2: Final results obtained

Table 2 shows the results that we have obtained. From the results, we can noting that the results are not good, especially in terms of precision which is very low.

5 Conclusions and further work

Unlike most areas of the language engineering, in the automatic detection of plagiarism, the precision is lower than the recall. This is because it is very likely to find similar fragments between two documents, although these are not plagiarized fragments. For a future work, it would be interesting search for a automated approach to reduce the space of search before conducting the search based on the comparison between n-grams. In (Barrón-Cedeño, Rosso, and Benedí, 2009), the author proposed the reduction of the space of search on the basis of the Kullback-Leibler distance.

In this paper we tried to put ourselves in a SME's shoes and in its need of detecting cases of plagiarism of its marketing campaign on the web. The idea was to investigate to what extent this could be done using the plagiarism detection software which is available on the web. The poor results we obtained with WCopyFind tool, highlight the need to develop at-hoc plagiarism detection methods for SME's.

References

Barrón-Cedeño, A. and P. Rosso. 2009. On automatic plagiarism detection based on n-grams comparisons. *Proc. European Conference on Information Retrieval, ECIR-2009*, pages 696–700.

Barrón-Cedeño, A., P. Rosso, and J.M. Benedí. 2009. Reducing the plagiarism detection search space on the basis of the Kullback-Leibler Distance. *Proc. 10th Int. Conf. on Comput. Ling. and Intelligent Text Processing, CICLing-2009, Springer-Verlag, LNCS(5449)*, pages 523–534.

Dreher, H. 2007. Automatic conceptual analysis for plagiarism detection. *Journal of Issues in Informing Science and Information Technology 4*, pages 601–614.