External Plagiarism Detection
Lab Report for PAN at CLEF 2010

Sobha Lalitha Devi, Pattabhi R K Rao, Vijay Sundar Ram and A Akilandeswari

AU-KBC Research Centre, MIT Campus of Anna University Chennai
sobha@au-kbc.org

Abstract. Here we describe our algorithm for detecting external plagiarism in PAN-10 competition. The algorithm has two steps 1. Identification of similar documents and the plagiarized section for a suspicious document with the source documents using Vector Space Model (VSM) and cosine similarity measure and 2. Identify the plagiarized area in the suspicious document using Chunk ratio.

1 Introduction

Plagiarism is defined as stealing or imitation of the language of another author and the representation of them as one's own original work. Here we work on external plagiarism where a set of suspicious documents are given along with set of source documents from where the text is copied in the suspicious documents. We have referred the following works: external plagiarism detection which compares different similarity measures Hoad and Zoble [3], using hashing or fingerprinting Brin, Davis, and Garcia-Molina [2], Ferret System based on trigrams [4], and Mixed-length comparisons [1]. In our work we differ from these approaches, by taking a moving window of 4 word sequence and use Chunk ratio R for identifying plagiarized passages.

2 Our Methodology, Evaluation and Conclusion

Our Algorithm has two steps 1) Document Filtering and 2) Identification of plagiarized passages. In step one all the suspicious documents are compared with the source documents. The documents are represented as vector of terms and a term is a sequence of four (4) words, called chunk. The chunk is defined as set of four consecutive words, where the last three words in the preceding sequence is considered as the first three words in the following sequence. For example, the chunk is w1w2w3w4, w2w3w4w5, w3w4w5w6 etc. The weights of the chunk in the vector are the term frequency and inverse document frequency (tf-idf). In Similarity identifier we compare each suspicious document with all the source documents. The PAN-10 test collection has 15925 suspicious and 11148 source documents and comparison is

of the order 1.77 x 10^8. For similarity we used the cosine similarity. The pairs of suspicious and source documents, for which the similarity score obtained is greater than the threshold of 0.005 is taken for step 2. The threshold was based on the development corpus. In step 2, in identification of plagiarized passages, we take all the pairs of suspicious and source documents above the threshold and identify the area where the plagiarism is done. We mark the line numbers in suspicious and source documents where the chunks have occurred. The consecutive lines where the chunks have occurred are grouped together. The difference between the lines n and n+1 is kept at less than or equal to 10. In a pair of suspicious and source documents, we get several such groups. To identify which groups of lines to consider as plagiarized we calculate a ratio, which we term it as "chunk ratio (R)" and the formula is R = C^2/(cosine score). C = frequency of commonly occurring chunks in suspicious and source documents.

Implementation: We consider the plagiarized area for which the R is greater than 0.65. The comparisons are of the order 1.77 x 10^8. The documents are split into 5 parts with 2000 documents and computed the inverted index. It is stored as hashes data structure of Perl. The five parts were run parallel in 6 different machines and the configuration of machines used was Pentium 4 with 2 GB RAM and 800 FSB and one with Core2duo with 2 GB RAM and 1033 FSB and the task was finished in 38 Hrs.

Evaluation and Conclusion: We have obtained the best precision of 0.9561. The recall obtained is 0.2868. The low recall can be attributed to setting the chunk ratio score to greater than or equal to 0.65. While working on the development set we found that taking 'R' below 0.65 was reducing the precision below 0.9. We obtained overall score of 0.4378. In the first step of document filtering we obtained an accuracy of 88.76%. The system has good precision. In our work we differ from previous approaches, by defining a chunk as moving window of 4 word sequence and we use a new measure called Chunk ratio R, for identifying plagiarized passages. The preprocessing of the documents is not done.

References

1. Barr´on-Cede˜no, Alberto and Paolo Rosso: On Automatic Plagiarism Detection based on n-grams Comparison. In: Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soul´e- Dupuy (eds.) ECIR 2009, LNCS vol. 5478, pp. 696–700. Springer (2009)

2. Brin, S., Davis J., Garcia-Molina H.: Copy detection mechanisms for digital documents. In ACM International Conference on Management of Data SIGMOD (1995)

3. Hoad, Timothy C. and Justin Zobel: Methods for identifying versioned and plagiarized documents. Journal of the American Society for Information Science and Technology (JASIST) 54(3), 203–215 (2003)

4. Lyon, Caroline, Ruth Barrett and James Malcolm: A theoretical basis to the automated detection of copying between texts, and its practical implementation in the ferret plagiarism and collusion detector. In: Plagiarism: Prevention, Practice and Policies Conference, June (2004)