Classifying Search Engine Results as Potential Sources of Plagiarism

Kyle Williams[1], C. Lee Giles[1,2]
Information Sciences and Technology[1], Computer Science and Engineering[2]
The Pennsylvania State University, University Park, PA, 16802, USA

Introduction

In plagiarism detection the goal is usually to identify the similarities between a suspicious document and a set of documents that it might have been copied from. However, in some cases there is no set of documents to compare the suspicious document to and thus a candidate set needs to be identified. Once this candidate set has been identified, more accurate analysis can be performed.

Source Retrieval involves using a search engine to identify this candidate set. Queries from the suspicious document are submitted to the search engine and a decision is made about which search results are worth analyzing. In this poster, we present a competition winning approach to source retrieval that achieved the best overall performance in two plagiarism detection competitions (Williams et al., 2014a; Potthast et al., 2014) by using a supervised method to classify the search engine results as potential sources of plagiarism.

The Source Retrieval Problem

The source retrieval problem is formally described below.

Source Retrieval Problem: *Given a suspicious document and a search engine, use the search engine to retrieve candidate documents that may be sources of plagiarism while minimizing the number false positives retrieved.*

We solve this problem by automatically extracting queries from the suspicious document, submitting the queries to a search engine, using a machine learning approach to classify the search results returned by the search engine and then only retrieve the results classified as being potential sources of plagiarism while ignoring the others. This process is demonstrated in Figure 1.

Results Classifier

We train a Linear Discriminant Analysis classifier for search engine results. Each result returned by the search engine is classified as being a potential source of plagiarism based on features available at search time without ever retrieving the document. We consider features based on:

- The similarity between the suspicious document and the text returned for each search result, such as the title and snippet.
- The ranking of each search result.
- Information about the search result such as its domain, length, number of words, etc.

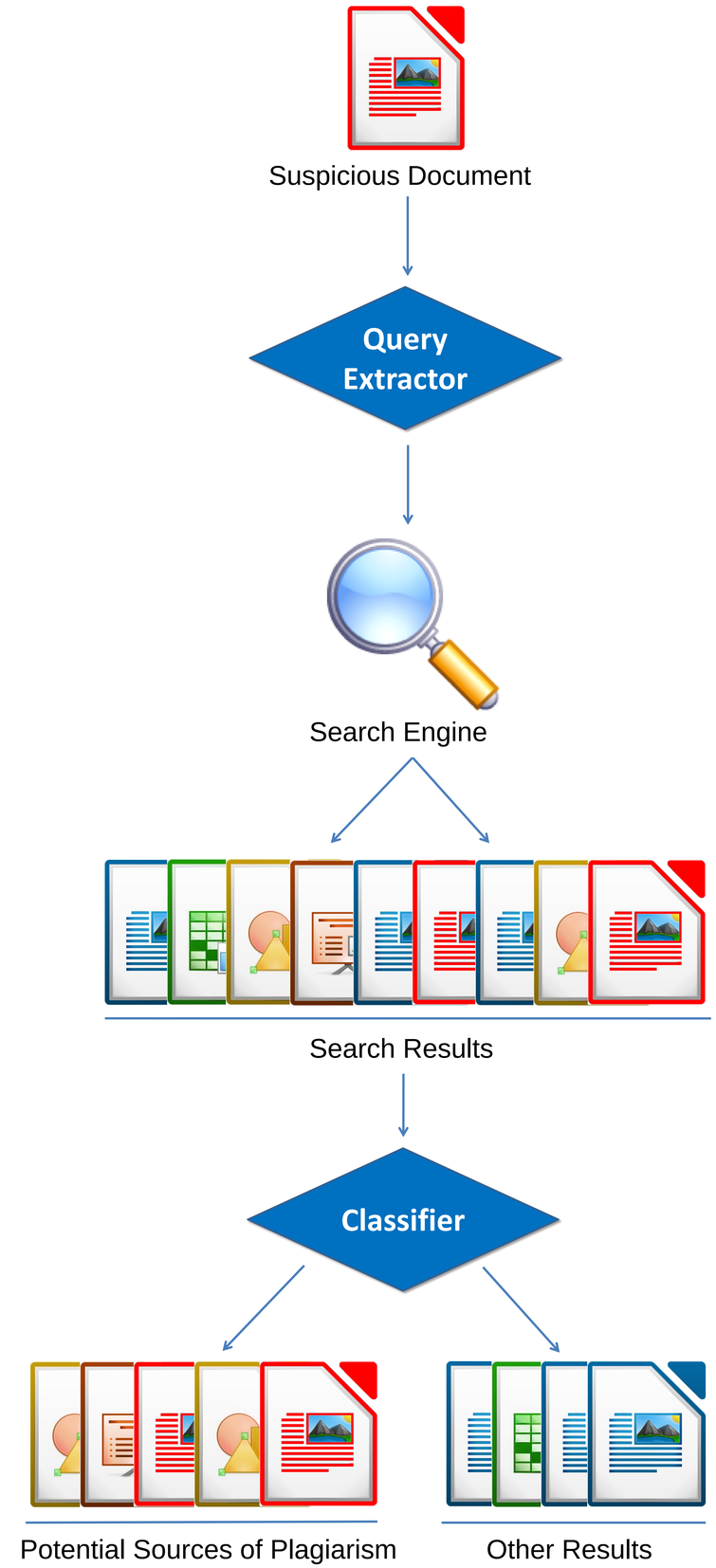

Figure 1: The Source Retrieval Strategy

Experiments

Table 1 shows the performance of our approach in the 2014 PAN Source Retrieval Task (Potthast et al., 2014). Our approach:

- Achieved the highest recall, meaning that it retrieved the highest number of sources of plagiarism
- Achieved the highest precision, meaning that it retrieved sources of plagiarism with the highest accuracy
- Achieved the best tradeoff between precision and recall (F1)
- Used a relatively high number of queries though with relatively few downloads.

Team	Precision	Recall	F1	Queries	Downloads
Williams	**0.57**	**0.48**	**0.47**	117.1	**14.4**
Zubarev	0.54	0.45	0.45	37.0	18.6
Prakash	0.38	0.51	0.39	60.0	38.8
Elizalde	0.40	0.39	0.34	54.5	33.2
Kong	0.08	**0.48**	0.12	83.5	207.1
Suchomel	0.08	0.40	0.11	**19.5**	237.3

Table 1: Performance in 2014 PAN Source Retrieval Task

Feature Analysis

We analyzed the features that are most important for search result classification (Williams et al., 2014b) and found that those based on the similarity between the suspicious document and the search results were the most important.

Interestingly, if a search result was from Wikipedia that was an important indicator that it could be a source of plagiarism.

Acknowledgments

We gratefully acknowledge support from the National Science Foundation and useful suggestions by Dr. Hung-Hsuan Chen. Icons by The Document Foundation used under CC-BY-SA 3.0 Unported license and Oxygen Project under LGPL.

References

1. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., & Stein, B. (2014) Overview of the 6th International Competition on Plagiarism Detection.
2. Williams, K., Chen, H., & Giles, C. L. (2014a) Supervised Ranking for Plagiarism Source Retrieval. CLEF 2014 Evaluation Labs and Workshop Working Notes Papers.
3. Williams, K., Wu, J., & Giles, C. L. (2014b) Classifying and Ranking Search Engine Results as Potential Sources of Plagiarism. In ACM Symposium on Document Engineering.