Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, Volume 22, No. 1, ISSN 1726-9679
ISBN 978-3-901509-83-4, Editor B. Katalinic, Published by DAAAM International, Vienna, Austria, EU, 2011

Make Harmony between Technology and Nature, and Your Mind will Fly Free as a Bird

Annals & Proceedings of DAAAM International 2011

PLAGIARISM DETECTION IN A MULTILINGUAL ENVIRONMENT

TRIFAN, I[onut]

Abstract: *The purpose of this paper is to research an automatic similarity and plagiarism detection system in a multi-lingual environment. It was developed a new system that is using Winnowing fingerprint extraction method and overlapping word-5-grams algorithm to obtain the fingerprint of the documents. A language guesser tool and an online dictionary are used to identify the language and to translate the documents to be used in cross-lingual similarity algorithms.*

Key words: *similarity detection systems, fingerprint systems, cross-lingual similarity, plagiarism detection*

1. INTRODUCTION

The concept of similarity has become more important over the years, algorithms for calculating the similarity being used by the search engines to be able to detect duplicate pages, by indexing and clustering systems, by plagiarism detection systems and many other systems.

The concept of similarity varies depending on the context of the data analyzed. Inverse Document Frequency (comparing the unusual frequency terms) is the most used algorithm for text blocks and space vector approach for structured data.

There are two types of similarity and plagiarism analysis: intrinsic and extrinsic. Intrinsic plagiarism method is not using any reference collection and it tries to determine passages by analyzing style changes within the document. Intrinsic plagiarism detection is related to authorfingerprint . Extrinsic similarity method compares the document and tries to find substrings that have been copied from the collection.

2. SIMILARITY DEDECTION SYSTEMS

Each system is using its own definition and level for a document to be considered similar. Manber (1994) proposes a definition for the similarity files: „We say that two files are similar if they contain a significant number of common substrings that are not too small".

2.1 Online Systems

Online detection systems are searching parts of documents on the Internet. Because the Internet is one of the most important data sources, its importance should not be underestimated. Online systems are considered to be very similar to the search engines, which are focused on speed and quantity and not optimized for quality.

To create an online system means that the system will use a large database collection. For example, the online system Turnitin database is having over 14 billion web pages and 150 million articles indexed. Because of this constraint, most online detection systems are using existing search engines.

Online systems are not able to use all the advantages, but time consuming document comparison routines used by offline similarity detection systems. Most of the time, the programmers have to solve technical problems that are not related to a similarity detection algorithms, like organizing large-scale documents collections.

2.2 Offline Systems

Offline detection systems are using models where all information is available within the documents in a collection. For example, for offline systems the collection may consist of archived student papers.

The best method is to combine the online and offline systems. First step is to search and choose a number of suspected documents using an online search engine. Next step is to analyze the similarity between the documents using specific offline algorithms.

The existence of several similarity detection systems raises the question of a classification. A classification of these types of systems is proposed by Mozgovoy (2006), the most important categories are: fingerprint based, string matching based and tree-matching-based systems.

New systems are use artificial intelligence to create automatic systems. Artificial neural networks possess remarkable capacity to learn, to adapt, to generalize, and it can solve nonlinear problems using neural networks and neuro fuzzy. (Engels et al., 2007; Ceska, 2008; Trifan 2010).

3. FINGERPRINT SYSTEMS

One of the best fingerprint systems was developed by Narayanan Shivakumar and Hector Garcia-Molina., using as a starting point signature-based techniques implemented by the SIF and the COPS. SCAM is designed and optimized to detect the similarity and plagiarism of documents. SCAM uses information retrieval techniques to implement a word based system.

The authors of SCAMP have preferred to develop a detection system that is using a words based similarity algorithm rather than phrases based system, primarily due to the fact that the separation of words is much easier than the separation of the phrases (one of the COPS system problems). Second, is more likely to find common sequences of words rather than common phrases, such as partial fragments.

4. SYSTEM OVERVIEW

The system is using Winnowing fingerprint extraction method. The algorithm of hashing selection is one of the main advantages of this method, the trade-of between fingerprint length and the smallest detected string has theoretical guarantees. The similarity detection algorithm consists of several main steps.

4.1 Cross-lingual similarity

First step consists from translating Non-English documents using word alignment algorithm. This algorithm is using pairs of words that are used as translation candidates. The system is trying to find the best 3 translation candidate for each word.

The untranslatable words are not replaced. Also, a language guesser tool is used to identify the language of each document. After all words are replaced, the new document is used for the second step.

4.2 External similarity

The pre-selecting method is used to find potential similar documents and to reduce the number of candidate documents. The system is using a search engine to find, select best 100 online candidate documents and include them in the offline documents collection.

4.3 Retrieval of Candidate Documents

The pre-selecting method is used to find the best matching documents. To reduce the number of candidate documents the system selects only documents having the fingerprints similarity greater than a given threshold.

4.4 Fingerprint

The preprocessing step is to remove all non all non alphanumeric characters and to convert all characters to lower case. Each document is split into sequences of at least three consecutive alphanumeric characters (words). The position of each word is saved.

The overlapping word-5-grams algorithm is used, then sorting the words in it and calculate the MD5 hash to obtain the fingerprint of the documents. The most signficant 3 2 bits are saved to be used. Using the Winnowing fingerprint method, six fingerprints are used as a window and the smallest one is selected as a sample fingerprint of the window. The system is using the inverted index method to examine the candidate documents. The documents having less than 30 common fingerprints are ignored.

5. RESULTS

5.1 Variables

To evaluate the performance of the similarity detection systems, there are implemented the variables recall, precision, granularity and overall (Potthast et al., 2009):

$$recall = \frac{1}{|R|}\sum_{i=1}^{|R|}\frac{|\hat{r}_i|}{|r_i|} \tag{1}$$

$$precision = \frac{1}{|P|}\sum_{i=1}^{|P|}\frac{|\hat{p}_i|}{|p_i|} \tag{2}$$

$$granularity = log_2(1 + \frac{1}{|R_p|}\sum_{i=1}^{|R_p|}|r_i \cap P|) \tag{3}$$

$$overall = \frac{F}{granularity} \tag{4}$$

where *r* is a similar passage and |*R*| is the set of all similar blocks, *p* is the detected similar passage using the algorithm and |*P*| the set of all detected similar passages, $|R_p|$ is the subset of *R* for |*P*|, |*r*| is the length of the passage *r*, |*p*| is the length of the detected similar passage *p* and F is the harmonic mean of (*recall, precision*).

5.2 Data

To test the system it was used a collection of 900 documents.

Language/ Obfuscation	**None**	**Low**	**High**	**Translated**
English	150	150	50	50
French	50	50	30	20
Italian	50	50	30	20
Romanian	50	50	50	50

Tab. 1. The collection of documents

5.3 Results

The system ran on a Dual-route Intel Xeon 4 cores 5500 series processor and 16Gb memory.

The results obtained by running the test documents are illustrated in the next tables

Language/ Obfuscation	None	Low	High	Translated
English	0.9141	0.8942	0.6734	0.9482
French	0.8284	0.8021	0.5935	0.8563
Italian	0.7525	0.7511	0.5512	0.8121
Romanian	0.7736	0.7523	0.5014	0.8692

Tab. 2. Recall results

Language/ Obfuscation	None	Low	High	Translated
English	0.9141	0.6320	0.4612	0.6829
French	0.6393	0.6221	0.4221	0.6343
Italian	0.5832	0.5294	0.3953	0.6222
Romanian	0.5928	0.5382	0.3982	0.6492

Tab. 3. Score results

6. CONCLUSION

In the last years many similarity and plagiarism detection systems have been developed, but none of them are able to work with the documents written in different languages. This system is using a language guesser tool and an online dictionary to translate the documents into English. The results are showing that the score results are lower for the Non-English documents.

One possible improvement is to use more than one online dictionary to translate the documents and keep up to 5 different translations for each word.

The system is showing that the score for high obfuscation document was lower then the rest. There are several possible improvements. First is using N-gram profile distance algorithm and using an intrinsic detector for the external detector. A new improvement is by using neural networks and neuro fuzzy approach.

7. REFERENCES

Ceska Z., Toman, M, Jezek K. (2009). Multilingual Plagiarism Detection. Artificial Intelligence: Methodology, Systems, and Applications, *Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications*, pp. 83-92

Engels, S., Lakshmanan, V., Craig, M. (2007). Plagiarism detection using feature-based neural networks. *SIGCSE*, pp. 34-38

Mozgovoy, M. (2006). Desktop Tools for Offline Plagiarism Detection in Computer Programs. *Informatics in Education*, pp. 97-112

Pereira, R. C., Moreira, V. P., Galante, R. (2010). A New Approach for Cross-Language Plagiarism Analysis. *Proceedings of the CLEF 2010 Conference on Multilingual and Multimodal Information Access Evaluation*, 2010

Potthast, M, Stein, B, Eiselt, A. (2009). Overview of the 1st International Competition on Plagiarism Detection, *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, 2010, pp 1-9

Trifan, A.L. (2010). Financial time series forecasting using neural networks: A case study of the Bucharest Stock Exchange, *Annals of DAAAM for 2010 & Proceedings of the 21st International DAAAM Symposium*, pp. 1381- 1383, ISBN 978-3-901509-73-5, ISSN 1726-9679