Fuzzy Cross Language Plagiarism Detection Approach Based on Semantic Similarity and Hadoop MapReduce

H. Ezzikouri, M. Oukessou, M. Erritali and Y. Madani

Abstract Ranging from modifying texts into semantically equivalent up to translation and adopting ideas, without proper referencing to its originator, Cross Language Plagiarism can be of many different natures. Among the most common problems in any data processing system is reliable large-scale text comparison, especially in a fuzzy semantic based similarity due to the complexity of natural languages in particular Arabic, and the increasing number of publications which raise the rate of suspicious documents sources of plagiarism. CLPD is more complicated than monolingual plagiarism, it goes beyond copy+translate and paste, consequently the detecting process exposes the need for vague concept and fuzzy sets techniques in a big data environment to reveal dishonest practices of hidden plagiarism in Arabic documents translated from English or French sources. In this paper, we propose a fuzzy-semantic similarity for CLPD using WordNet taxonomy and three semantic approaches Wu and Palmer, Lin and Leacock-Chodorow for Arabic documents; the work has been parallelized using Apache Hadoop with HDFS file system and MapReduce programming model.

1 Introduction

Thanks to the rise of W3 and development of information technologies, information is within everyone's reach, people are able to generate more than you can imagine of data and information which lead to an explosive growth in the amount of data and the increase of originality and plagiarism issues. Big data is having a continuous exponential progress, intensified by the broadcast of digital information technologies, owing to this an already hard fuzzy task become even harder, Cross-Language Plagiarism detection (CLPD) consist of detecting plagiarism in documents from

H. Ezzikouri · M. Oukessou (✉) · M. Erritali · Y. Madani
Sultan Moulay Slimane University, BP 523 Beni Mellal, Morocco
e-mail: ouk_mohamed@yahoo.fr

H. Ezzikouri
e-mail: ezzikourihanane@gmail.com

S. Melliani and O. Castillo (eds.), *Recent Advances in Intuitionistic Fuzzy Logic Systems*, Studies in Fuzziness and Soft Computing 372,
https://doi.org/10.1007/978-3-030-02155-9_15

less-related languages such as English Arabic and French, in a massive amount of data seems to be impossible from the first sight even after applying some information retrieval systems that may generate thousands of candidate documents.

Cross-Language Plagiarism refers to the unacknowledged reuse of a text involving its translation from one natural language to another without proper referencing to the original source, its a sort of plagiarism idea, because texts are totally changed but ideas in the original texts remain unchanged; Such a change in the syntax and semantics of texts requires a deep and concentrated processing, then we confront two major factors, the management of large mass of data in all candidate documents and the number of operations required for this kind of plagiarism detection process. The nature of Cross-Language Plagiarism practices could be more complicated than simple copy translate and paste, in CLPD the languages from source and suspicious documents differ, thus the process exposes the need for a vague concept and fuzzy sets techniques to reveal dishonest practices in Arabic documents.

In this paper, we propose an detailed fuzzy semantic based similarity approach for analyzing and comparing texts in CLP cases using Big Data and WordNet lexical database (Miller, 1995) [1], to detect multilingual plagiarism in documents translated from English and French to Arabic. Arabic is known as one of the richest human languages in terms of words constructions and meanings diversity. We focus in our work on obfuscated plagiarism cases where texts are translated and rephrased from one language to another with no reference to the original source.

It's obvious that documents published in every field increases explosively, thus detecting plagiarism need a deep important treatment, therefore we need a large storage volume for storing all this data and also it is the problem of the required time to get results. To remedy this problem our proposal consists of parallelizing our method by working in a Big Data system with the Apache Hadoop using the HDFS (Hadoop Distributed File System) and the Hadoop MapReduce. Preliminary operations and text preprocessing are done to the documents such as tokenization, part-of-speech (POS) tagging, lemmatization and stop words removal, text segmentation (word 3-gram). The fuzzy semantic-based approach is based on the fact that words from two translated compared texts have in general, Strong fuzzy similarity words of the meaning from the second language

2 Related Work

The amount of generated data the frequency generated by on the web is incredible, produced the 'big data' term, defined by Gartner (Beyer and Douglas 2012) [2] as high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making and the recently added Veracity and Value based on the fact that accurate analysis could be affected by the quality of captured data. Most research works have used big data in the information retrieval phase [3, 4]. Zhang et al. [5] Presented a sequence-based method to detect the partial similarity of web pages using MapReduce based on sentence level

near-duplicate detection and sequence matching. Dwivedi et al. [6] introduced a SCAM (Standard Copy Analysis Mechanism) plagiarism detection algorithm, the proposed detection process is based on natural language processing by comparing documents and a modified Map-Reduce based SCAM algorithm for processing big data using Hadoop and detect plagiarism in big data.

3 Big Data a Solution for Fuzzy CLPD

Fuzzy set theory introduced by Lofti Zadeh in 1965 based on his mathematical theory of fuzzy sets, fuzzy set theory is a generalization of the theory of classical sets, and it permits the gradual assessment of the membership of elements in a set using a membership function valued in the real unit interval [0, 1]. Fuzzy set theory could be used in a wide range of domains especially for handling uncertain and imprecise data that linked with CLPD. In a cross language semantic based similarity detection process where words borders are not clear and the intersection of meanings of words are fuzzy, the fuzzy set theory seems to be the right way to treat such case. The huge masse of information and data generated from a voluminous corpuses, an efficient solution is the use of Big Data technologies to parallelize the fuzzy CLPD process in [7], the idea is to distribute and share it between several cluster, using Apache Hadoop framework with Hadoop distributed file system HDFS for distributing the storage of documents and also for storing results, and MapReduce programming model for the parallelisation and the development of our proposal.

CLP is a fuzzy complex process. Each word is associated with a fuzzy set that contains words with the same meaning with a similarity between 0 (for totally different) and 1 (for identic) (Fig. 1); thus fuzzy sets theory in CLPD looks to be an obvious way to solve the problem. Fuzzy set theory and CLPD turn up to be the perfect couple, however, the important number of operations and the running time implies to search for a solution to improve performance and results: is Big Data technologies.

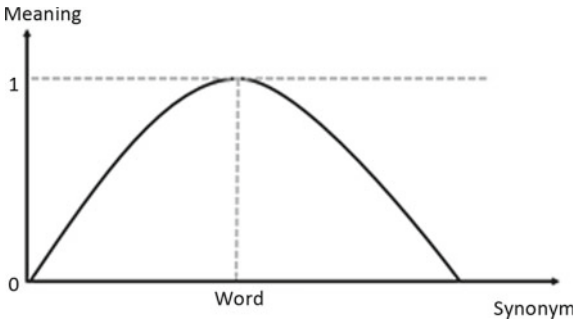

Fig. 1 A words fuzzy set synonyms

4 The Proposed Method

1. Preprocessing and segmentation

This paper is an extension and improvement of previous work by Ezzikouri et al. [7], where we present an intelligent multilingual plagiarism detection using Fuzzy-Semantic Similarity based methods (Wu and Palmer, Lin and Leacock-Chodorow) and Big data technologies (Hadoop, HDFS and MapReduce).

Input text and the collection corpus are from three distant languages Arabic-English and French, the creation of a suitable target data of each document is elementary, various text preprocessing methods based on NLP techniques are implemented (Fig. 2) and described in details in our previous work [7].

Semantic Similarity is defined here as the similarity between two concepts in a taxonomy(e.g. WordNet (Miller, 1995) [1]), where synonymous words are joined together to form synonyms sets called also synsets; Several similarity measures have been proposed in the last few years, LCH (Leacock and Chodorow, 1998), WuP (Wu and Palmer, 1994), RES (Resnik, 1995), LIN (Lin, 1998), LESK (Banerjee and Pedersen, 2003), and HSO (Hirst and St Onge, 1998) [8–12].

The proposed algorithm in this paper, is based on three semantic similarity approaches (Wu and Palmer, Lin and Leacock-Chodorow), that use WordNet to automatically evaluate semantic relations between words.

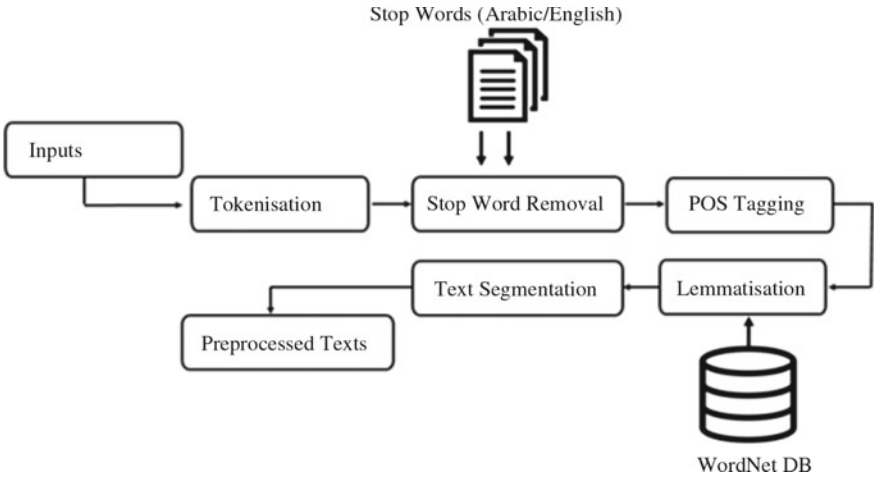

Fig. 2 Text preprocessing for CLPD [7]

2. Complexity

Leacock-Chodorow and Wu-Palmer are based on path length. They are based on counting the number of relation links between nodes in a taxonomy. Path length based measures are independent of corpus statistics and uninfluenced by sparse data. To compute the distance between two nodes in Wordnet, first we have to compute the shortest path between two nodes (length). This calculation process is equivalent to shortest path in a graph, i.e. the complexity could be estimated by $O(|V| + |E|)$ using a standard BFS approach where $|V|$ is the number of vertices and $|E|$ is the number of edges in the graph [13].

3. Fuzzification

Commonly two translated compared texts form a fuzzy similarity sets of words (differs in languages) sharing the same meaning class, based on that a Fuzzy semantic-based approach is obtained [4]. Many researches concentrate on text preprocessing methods especially Part-Of-Speech (POS) and its integration with fuzzy based methods for an efficient identification of similar documents [14].

Fuzzification is a fundamental operation in every fuzzy inference system, where relationship between inputs and linguistic variables is defined by a fuzzy membership function. In this paper we propose two semantic similarity approaches (Fuzzy-Lin and Fuzzy-Wup proposed in [7]) as fuzzy memberships function to fuzzify the relationship of word pairs (from text pairs), and compare results with Leacock-Chodorow (LCH).

The Lin and Wup fuzzy membership functions (expressed bellow Eq. 3) :

$$\mu_{1\,a_i\,b_j} = Lin(a_i, b_j) \tag{1}$$

$$\mu_{2\,a_i\,b_j} = Wup(a_i, b_j) \tag{2}$$

The membership degrees indicate to what extent an element belongs to a fuzzy set [15], i.e. for synonyms the value is 1, and 0 for dissimilar words.

To evaluate the similarity of two texts, a fuzzy inference system is needed. Fuzzy PROD operator is used for evaluating word relationship in first text with words in the second :

$$\begin{aligned} \mu_{a_1\,B} &= 1 - \prod_{\substack{b_j \in B \\ j \in [1,m]}} \left(1 - Wup(a_1, b_j)\right) \\ \mu_{a_n\,B} &= 1 - \prod_{\substack{b_j \in B \\ j \in [1,m]}} \left(1 - Wup(a_n, b_j)\right) \end{aligned} \tag{3}$$

The average sum is calculated by :

$$\mu_{A,B} = \left(\sum_{i=1}^{n} \mu_{a_i,B}\right) / n \tag{4}$$

4. Research methodology and Algorithm

Our main contribution in this article is fuzzifying cross language plagiarism detection using fuzzy semantic similarity approaches (fuzzy-WuP and fuzzy-Lin and LCH) in a parallel manner using Apache Hadoop (HDFS + MapReduce).

The idea is to store the inputs (the candidate documents (English, French) and the query document (Arabic text)) also results will be stored in HDFS for distributing the storage between several machines (Hadoop Cluster). Moreover, for the development of our proposal, MapReduce programming model was used (a parallel plagiarism's detection).

Inputs are from three different languages, an Arabic text and a corpus of potential candidates source of plagiarism in English and French. Some necessary steps must be done first like, text preprocessing, which contain several NLP processes (tokenization, stop words removal, post-tagging [16]) and word 3-grams (W3G) segmentation, this step is pivotal since Arabic has a complex morphology and one of the most difficult languages to treat. After that, then we prepare the inputs is the storage of them in HDFS (distributing the storage).

The resulting text (Arabic text) and every single text from the corpus (the English/French texts one by one) are used as inputs for the fuzzy inference system, then WuP Lin and LCH semantic similarity measures are modelled as membership functions. The output is a similarity score between the Arabic text and each input text from the corpus. Figure 3 shows the different steps of our work.

The MapReduce algorithm followed :

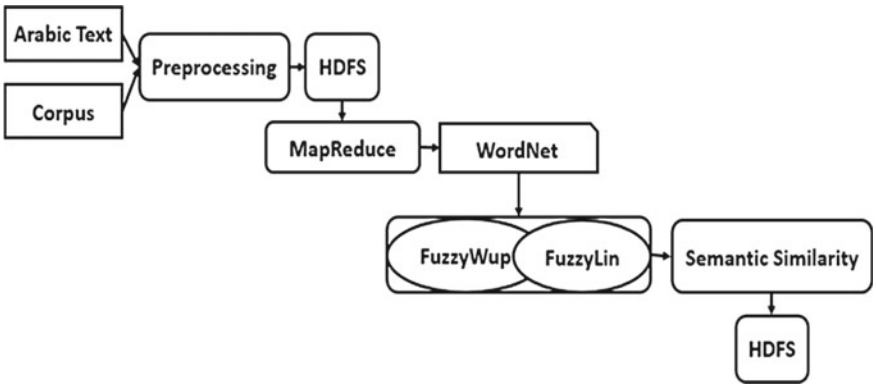

Fig. 3 Parallel Fuzzy CLPD system

Algorithm : MapReduce Programming Model for our proposed method

```
Inputs : Arabic Text , English text from the corpus
Require : Semantic Similarity between inputs
 S_Wup     0;
 S_Lin     0;
 S_LCH     0;
 C     0;
 AR      Text Preprocessing(Arabic Text);
 EN      Text Preprocessing(English Text);
 FR      Text Preprocessing(French Text);
 Segmentation1[]      W3G(AR);
 Segmentation2[]      W3G(EN);
 Segmentation3[]      W3G(FR);
 For all word In Segmentation1
    For all term In Segmentation2 and Segmentation3
       If word In WordNet
         Then word      WordNet(word);
       Else
         word      Translate(word);
       End If
       Fuzzy-Wup      1 - Wup(word, term);
       Fuzzy-Lin      1 - Lin(word, term);
       LCH      LCH(word, term);
       S_Wup      S_Wup + Fuzzy-Wup;
       S_Lin      S_Lin + Fuzzy-Lin;
       S_LCH      S_LCH + LCH;
       C      C + 1;
    End For
 End For
 Sim_Wup=S_Wup/C;
 Sim_Lin=S_Lin/C;
 Write(Arabic Text || English (French) text, Sim_Wup /Sim_Lin/LCH)
```

The inputs storage distributing and plagiarism detection parallelising is done by constructing a Hadoop cluster that contains five machines Hadoop Nodes, this cluster has a master machine and four slave machines. Each node is an Ubuntu 16.04 machine.

5 Experimental Results and Discussion

Testing corpus is built up from 600 English/French and Arabic documents from different sources (news, articles, tweets, and academic works). To detect cross language plagiarism, 200 are translated from English/French to Arabic (machine based) with no change, and 400 documents are translated and modified with a high percentage of obfuscated plagiarism (paraphrasing, back-translation, etc.).

Fuzzy semantic metrics (Fuzzy WuP, Lin and LCH) are implemented and results are compared. Results presented in Figs. 4 and 5 are some of the experimental tests that demonstrate that the Fuzzy Wu and Palmer have high performance than Fuzzy Lin.

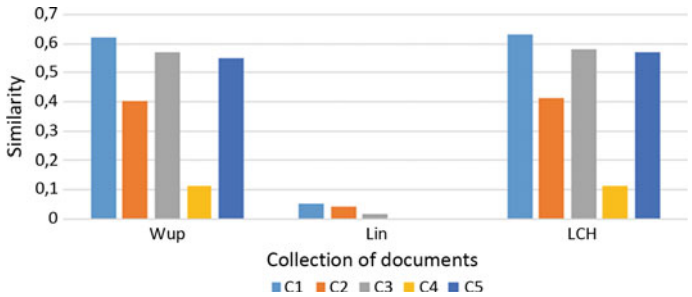

Fig. 4 Comparison of similarity for the proposed similarity measures

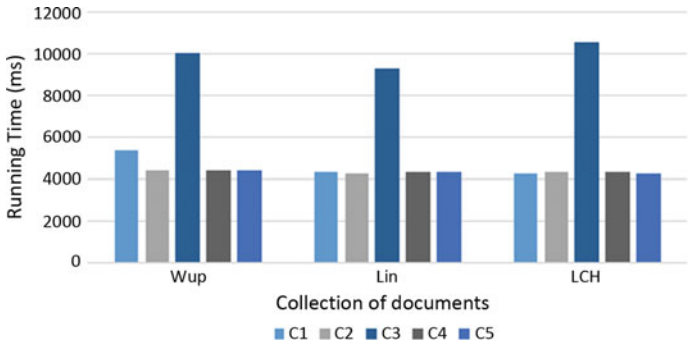

Fig. 5 Comparison of running time for the proposed similarity measures

6 Discussion

Based on the evaluation results and previous results in [8, 17, 18] and other ongoing works, Fuzzy-Lin gives good results in execution time but poor results for multilingual similarity and also it is not effective in detecting plagiarism in a big mass of data. Therefore Fuzzy-Lin is not suitable for detection multilingual plagiarism in high volumes of information. Fuzzy-WuP and LCH gives similar results in terms of similarity and execution time. For the execution time we notice a change in results of LCH for collections where documents are a little bit large. As said before our testing corpus contains several form of plagiarism ranging from simple translation to making serious changing in the text, which makes in addition to getting good results in detecting CLP, running time is also an important factor. For that we examined the two algorithms in terms of time with documents of different sizes (Fig. 6).

Fuzzy-WuP shows up to be the best semantic similarity measure for detection obfuscated plagiarism in a big data environment. Hence, no matter how the translated

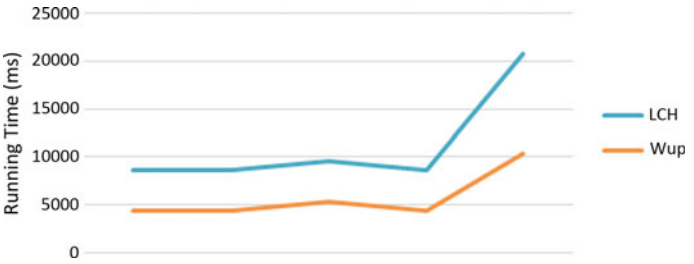

Fig. 6 Comparison of execution time by document size for WuP and LCH

plagiarized text is modified and the structure changed, such serious changing is the principal reason to involve fuzzy theory in plagiarism detection then similarity can still be detected.

The Parallel fuzzy-based Cross Language Plagiarism Detection using the lexical taxonomies WordNet in a big data environment presented in this paper achieves good results in comparison with some existing models and approaches namely fuzzy IR method in [19] such as word correlation factors obtained from large corpora that require allocation of disk space to save the word-to-word correlation factor tables. The processing time required to search for words and retrieve their correlation value is one of the main problems of former models and becomes a major issue for extending the use of parallel fuzzy-based CLPD approach, which have widely been reduced in our proposed model.

7 Conclusion

In this paper we have presented our Parallel fuzzy-based Cross Language Plagiarism Detection using WordNet and semantic similarity measures in a big data environment. Currently, most plagiarism detection tools are not suitable for detection a serious kind of plagiarism, where plagiarists are pushing to a high level using translation (human or machine based), paraphrasing, back-translation and a lots of manipulation to avoid to be caught with PD systems, obfuscated semantic plagiarism is a substantial issue and concern especially in academic works.

Different pre-processing methods based on NLP techniques were used (lemmatization, stop word removal and POS tagging), texts were segmented to 3-gram. Fuzzy semantic measures (Wu and Palmer, Lin and Leacock-Chodorow) were evaluated to judge the similarity in compared texts. Using a testing corpus of 600 handmade (rewording, paraphrasing, back-translation, idea adoption …etc.) and artificial plagiarism cases, the fuzzy plagiarism detecting method using two fuzzy semantic similarity approaches (fuzzy WuP and fuzzy Lin) in a parallel manner using Apache Hadoop (HDFS+MapReduce), hadoop was used for performance enhancement and time reducing, data were distributed across the cluster of machines (master and four slaves); processing time doesnt increase in comparison of the important number of

operations needed for such process. The results aid to conclude that the proposed model obtained reliable and significant performance with fuzzy Wu and Palmer similarity measure.

References

1. G.A. Miller, WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
2. M.A. Beyer, D. Laney, The importance of big data : a definition, Stamford CT Gart., 2014–2018 (2012)
3. M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar, N. Seliya, R. Wald, Muharemagic E., Deep learning applications and challenges in big data analytics. J. Big Data, **2** (2015)
4. B. Parhami, A Highly Parallel Computing System for Information Retrieval, in *Proceedings of the December 5–7, 1972, Fall Joint Computer Conference*, Part II (New York, NY, USA, 1972) pp. 681–690
5. Q. Zhang, Y. Zhang, H. Yu, X. Huang, Efficient partial-duplicate detection based on sequence matching, in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 675–682 (2010)
6. J. Dwivedi, A. Tiwary, Plagiarism detection on bigdata using modified map-reduced based SCAM algorithm. Int. Conference on Innovative Mechanisms Ind. Appl. (ICIMIA) **2017**, 608–610 (2017)
7. H. Ezzikouri, M. Erritali, M. Oukessou, Fuzzy-semantic similarity for automatic multilingual plagiarism detection. Int. J. Adv. Comput. Sci. Appl. **8**(9), 86–90 (2017)
8. C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification. WordNet Electron. Lex. Database **49**(2), 265–283 (1998)
9. Z. Wu, M. Palmer, Verbs semantics and lexical selection, in *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pp. 133–138 (1994)
10. P. Rensik, Using information content to evaluate semantic similarity, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453 (1995)
11. D. Lin, An information-theoretic definition of similarity, in Icml, **98**, 296–304 (1998)
12. G. Hirst, D. St-Onge, Lexical chains as representations of context for the detection and correction of malapropisms. WordNet Electron. Lex. Database **305**, 305–332 (1998)
13. T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms-Secund Edition. (McGraw-Hill, 2001)
14. D. Gupta, K. Vani, C.K. Singh, Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection, in *International Conference on Advances in Computing, Communications and Informatics* (ICACCI, 2014) pp. 2694–2699
15. N. Werro, *Fuzzy Classification of Online Customers* (Thesis University of Fribourg (Switzerland), Fuzzy Management Methods, 2008)
16. C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60 (2014)
17. H. Ezzikouri, M. Erritali, M. Oukessou, Semantic similarity/relatedness for cross language plagiarism detection. Indones. J. Electr. Eng. Comput. Sci. **1**(2), 371–374 (2016)
18. S. Alzahrani, N. Salim, Fuzzy semantic-based string similarity for extrinsic plagiarism detection. Braschler Harman **1176**, 1–8 (2010)
19. R. Yerra, Y.-K. Ng, A sentence-based copy detection approach for web documents, in *International Conference on Fuzzy Systems and Knowledge Discovery*. vol. 2005, pp. 557–570 (2005)