Plagiarism Detection In Arabic Scripts Using Fuzzy Information Retrieval

Salha Mohammed Alzahrani [1], and Naomie Salim [2]
[1] Dept. of Computer Science, Faculty of CS & Info. Sys, Taif University, Hwiah 888, Taif, Saudi Arabia
[2] Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Johor, Malaysia
admin@u2learn.net, naomie@utm.com

Abstract—**The nature of Arabic language structure exposes the need for fuzzy or vague concept to reveal dishonest practices in Arabic documents. In this paper, we present a statement-based plagiarism detection approach in Arabic scripts using fuzzy-set IR model. The degree of similarity is calculated and compared to a threshold value to judge whether two statements are the same or different. Our corpus collection has been built in which all stopwords were removed and non-stop words were stemmed for typical Arabic IR. The corpora have 100 documents with 4367 statements in total. Five query documents with about 250 plagiarized statements were constructed and tested. Experimental results show that fuzzy-set IR successfully detected not only exact but also similar statements that have different *structure*. However, our Arabic fuzzy-set model approach does not handle the case of rewording with different *synonyms/antonyms*, a deficiency that will lead to future work of modeling the system using Arabic *thesaurus*.**

Keywords- fuzzy-set information retrieval; Arabic; plagiarism detection;

I. Introduction

Although plagiarism detection in Arabic natural language documents is important in schools and universities in Arab countries, yet there are no known techniques to detect plagiarism in Arabic scripts. Arabic is known as the richest human language in terms of word's constructions and diversity meanings, and hence plagiarism practices could be more complicated than simple copy and paste. Words could be cleverly changed to their synonyms and statements could be easily changed to another structure with the same meaning such as from active to passive. It is thus hypothesized that Arabic plagiarism best can be handled using fuzzy-set information retrieval to reveal more than the copy and paste plagiarism. This paper presents the work done to adapt fuzzy-set IR model for use with Arabic language to detect not only exact or accurate match but also to detect plagiarized statements based on the degree of membership between words.

To accomplish the work, we have built our corpus collection from Arabic Wikipedia and passed the collected html documents through a series of pre-processing including eliminating non-essential data from html files, removing stopwords and stemming. Then, we developed our proposed method and investigated its efficiency using several experiments on a set of query documents and the corpora, and finally derived conclusions and suggestions for future work.

This paper proceeds to present our results as follows. In Section 2, we discuss the related work involved in plagiarism detection on English scripts. In Section 3, we describe a typical Arabic IR and the work that has been done regarding stemming and stopwords removal in order to build our corpus collection. In Section 4, we discuss Arabic fuzzy-set IR approach for plagiarism detection. In Section 5, we include the experimental results of our approach. Finally, we give a concluding remark and suggestion for future work.

II. Related Work

To understand the problem domain of plagiarism auto-detection, we should look at what type of documents we deal with and what characteristics they have. The corpora in our study contain Arabic Natural Language (NL) documents so we will firstly discuss several documents' descriptors in natural languages before applying techniques that are based on fused descriptors.

A. Document Descriptors in NL

There are several schemes to characterize documents before applying a plagiarism detection method. Simple document descriptors include character-based [1], word-based [2], phrase-based [1, 3-5], statement-based [6-8], line-based such as diff command in Unix/Linux [6], paragraph-based [9], and structure-based [7, 10-12], in which a document is described as sections, subsections, subsubsections and so on. Different descriptors can be combined to assist in plagiarism detection. For example, a document can be divided into sections and subsections, each subsection into paragraphs, and each paragraph into statements. Statements can then be compared in detail to find plagiarism. Besides, character-based and word-based descriptors by themselves cannot be used to detect plagiarism but they can be used to build more sophisticated descriptors such as hash values [13] or suffix-trees [14, 15]. In this study, statement-based representation has been chosen since Arabic, just as English; can be easily segmented into statements using periods and that reduces computations complexity compared to other descriptors.

B. Plagiarism Detection Techniques in Natural Languages

Different document descriptors entail different techniques used for plagiarism detection. Techniques could

978-1-4244-2869-4/08/$25.00 ©2008 IEEE

be classified into four as follows. The first technique, fingerprint matching [1, 6] involves the process of scanning and examining all possible substrings (fingerprints) of documents generated on a descriptor-basis as discussed in the previous section. Although fingerprinting (the process of generating fingerprints) is time and space consuming, it is a useful measure for detecting exact match.

The second approach is clustering [7, 16] that uses specific words (or keywords) to find similar clusters between documents. Clustering by itself cannot be used to judge plagiarism but it might be used as a first level of detection to find similar documents that discuss the same subject. Clustering could be followed by another level of comparing similar documents in detail, i.e. on a sentence-per-sentence basis to highlight plagiarism.

Structure-related techniques or so-called Stylometry [10-12] focuses on the trend of structure that the overall document has. This technique has not gained popularity in the majority of plagiarism detection tools because it is hard to maintain the style or structure of natural languages compared to source code programming [12].

Previous techniques are not optimal in case of rewording and/or restructuring of statements, a deficiency that can be solved by using fuzzy IR model [17, 18]. The vagueness can be modeled using fuzzy framework as follows. Each word in a statement is associated with a fuzzy set, and hence this statement has a degree of membership with this fuzzy set. Thus, the membership is now a gradual notion ranging from 0 to 1, contrary to crisp Boolean logic. Reference [6] utilized fuzzy-set IR as a copy detection approach for web documents and the results show that it outperforms fingerprinting techniques in handling the cases of rewording and restructuring. For that reason, this paper uses the methodology of [6] to develop Arabic fuzzy-set IR approach for plagiarism detection.

III. Arabic Information Retrieval

Arabic IR is the science of searching and retrieving Arabic text, documents, web pages etc from various resources within homogenous or heterogeneous databases. Arabic nature exposes many complexities and difficulties in IR [19] because of the high inflection, momentous eloquence and diacritics influence. There are two steps essential to eliminate unnecessary data and thence accelerate the retrieval process; removing stopwords and stemming.

A. Removing Stop Words

Arabic has many words' constructions that build words not needed to judge plagiarism. Thus, stopwords removal algorithm [20] were run before applying the proposed technique. This avoids non-significant words from interfering during plagiarism detection. In addition, removing stopwords reduces the number of words and size of document by 30-50% as can be seen in Figure 1.

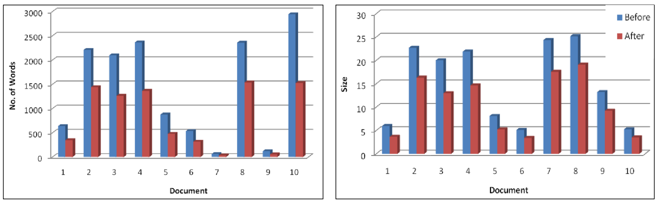

Figure 1. The effect of removing Arabic stopwords on the number of words and size for the first 10 documents

B. Stemming Arabic Words

Stemming is a technique aims to extract common affixes of various lengths from non-stopp words. Thus, words which are literally different but have a common stem may be abstracted by their stem [21]. As an example, the sentence S_i

*كعبة بناء موجود ارض مكة منطقة حجاز سعودي ربط ركن اسلم حج زيارة
كعبة طاف جزء فرض حج مسلم*

was obtained from its original sentence S

*الكعبة هي بناء موجود بأرض مكة بمنطقة الحجاز في السعودية يرتبط بأركان
الاسلام وخاصة الحج ، و يعد زيارة الكعبة والطواف حولها، جزءا من فريضة
الحج على كل مسلم.*

after applying stopwords removal and stemming algorithm. For this purpose, we used the algorithm in [22] to stem Arabic words wherein no root dictionary is needed.

C. Building Corpus Collection

The corpora of this study were collected from Arabic Wikipedia [23]. It consists of 100 documents chosen arbitrary about some general topics. Some non-essential data in the collected documents were removed to speed up processing. This includes removing HTML tags, eliminating empty lines, removing all lines that contain less than four words because they are meaningless to be plagiarized. The resulted documents have 4367 statements distributed evenly as shown in Figure 2. On the other hand, five query or test documents were used to verify the correctness and usefulness of our approach. Query documents (QDocs) were constructed manually with different degree of plagiarism from the corpus collection (CDocs) as follows.

i. QDoc1 is a duplicate of CDoc1.

ii. QDoc2 is closely related to CDoc2 wherein all sentences were included by with the order of paragraphs, statements and words changed.

iii. QDoc3 was constructed by adding unrelated statements to the CDoc3.

iv. QDoc4 is moderately related to CDoc4 but words were replaced with their synonyms and antonyms.

v. QDoc5 has exact and similar statements from more than one document in the corpus collection. This query is used to demonstrate the persistence, correctness and stability of our proposed approach.

IV. Arabic Fuzzy-Set IR Model

Two statements can be treated as the same although

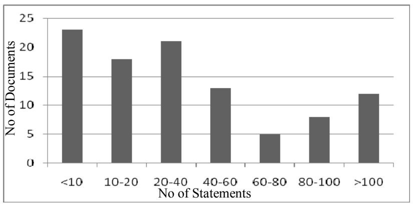

Figure 2. Statement Distribution in our Corpus Collection

they are semantically different based on the degree of similarity among words in both. Similarity between two statements has two cases: restructuring (i.e. changing the structure such as from active to passive) and rewording (replacing words with synonyms and antonyms). Fuzzy-set IR model [6, 17] can be used to judge similarity in both cases. This section describes the methodology used to adopt Arabic fuzzy-set IR model as in [6]. Table I exemplifies a pair of similar but restructured Arabic statements.

TABLE I. EXAMPLE SIMILAR ARABIC STATEMENTS

(S_i,S_j)	**Statements**
S_i	السيارة هي إحدى وسائل المواصلات
	Car is one of transportation means
S_j	بعض وسائل المواصلات تشمل السيارة والطائرة
	Some transportation means are cars and airplanes

To start with, we generate a different pairs of *Arabic terms* from both CDocs and QDocs. Samples of pairs are listed in Table II. Note that we use *"term"* here to refer to non-stop, stemmed words.

TABLE II. SAMPLE OF ARABIC TERMS PAIRS

Generated $<w_i,w_j>$ Pairs	
Arabic	***English***
<سيارة،عربة>	<car, automobile>
<سيارة،نقل>	<car, transportation>
<سيارة،طائرة>	<car, airplane>
<سيارة،منضدة>	<car, table>

a. English translation has been provided here to make Arabic words clearer to the reader but it does not interfere with programming

Secondly, we construct a *term-to-term correlation factor* ($c_{i,j}$) that defines the extent of similarity between $<w_i,w_j>$, using the equation

$$c_{i,j} = n_{i,j}/(n_i + n_j - n_{i,j}) \qquad (1)$$

where $n_{i,j}$ is the number of documents that has both *terms* and n_i, n_j if the number of documents that has w_i, w_j respectively. $c_{i,j}$ is one for synonyms and in the range [0-1] to measure how similar one word to another as shown in Table III.

TABLE III. TERM-TO-TERM CORRELATION FACTOR

$<w_i,w_j>$	**Correlation Factor**
<سيارة,عربة>	1
<سيارة،نقل>	0.5
<سيارة،طائرة>	0.1
<سيارة،منضدة>	0.001

Thirdly, we construct a *term-to-term correlation matrix* that consists of *term* pairs and their correlation factors as seen in Table IV.

TABLE IV. TERM-TO-TERM CORRELATION MATRIX

$<w_i, w_j>$	***سيارة***	***عربة***	***نقل***	***طائرة***	***منضدة***
سيارة	1	1	0.5	0.1	0.001
عربة	-	1	0.5	0.1	0.001
نقل	-	-	1	0.5	0.002
طائرة	-	-	-	1	0.003
منضدة	-	-	-	-	1

Fourthly, *term-to-sentence correlation factor* is computed as

$$\mu_{i,j} = 1 - \prod_{wk \in Sj}(1 - c_{i,k}) \qquad (2)$$

where w_k is one of the words in S_j and $c_{i,k}$ is the correlation factor between w_i and w_k. This is followed by *degree of similarity* between (S_i, S_j) as

$$\mathrm{Sim}(S_i, S_j) = (\mu_{w1,j} + \mu_{w2,j} + \ldots + \mu_{wn,j}) / n \qquad (3)$$

Then, we apply a threshold value[6] to find whether two statements should be treated as same or not. Finally, we compute the resemblance (degree of overlap) to measure how much of QDoc is contained in CDocs, as

$$R = (S_{QDoc} \cap S_{CDoc})/(S_{CDoc}), \quad 0 \leq R \leq 1 \qquad (4)$$

where $S_{QDoc} \cap S_{CDoc}$ is the number of similar statements in query document and corpus document, respectively.

To evaluate the retrieval process, precision and recall can be used [21]. They can be defined in our problem in terms of the number of statements detected as plagiarized and number of actual plagiarized statements. The following equations describe how to calculate precision and recall values

$$\text{Recall} = (\#\text{plagiarized} + \#\text{detected})/(\#\text{detected}) \qquad (5)$$

$$\text{Precision} = (\#\text{plagiarized} + \#\text{detected})/(\#\text{plagiarized}) \qquad (6)$$

V. EXPERIMENTAL RESULTS

Using the corpora and query documents, a *term-to-term correlation matrix* was constructed with around 4000 un-repeated Arabic terms. Our fuzzy-set IR model was tested using five test cases presented in Table V along with number of statements in corpus and query documents, number of plagiarized statements, number of detected statements by our approach, and finally the measure of resemblance R as explained in (4). As can be seen in case (i) and (iii), Arabic fuzzy-set IR performed very well in detecting duplicate statements although some unrelated statement were added in (iii). In case (ii), all statement in CDoc were included in QDoc with three different practices of plagiarism: copy and paste (duplicate), changing the structure of statements, or changing the words with synonyms and/or antonyms in some statements. The last two practices were also mixed. The measure of resemblance was 88% which indicates that Arabic fuzzy-set IR successfully highlights most of the plagiarized statements. The few statements not detected as plagiarism in this case is related to the third practice; replacing words by synonyms and/or antonyms. This can be significantly interpreted taking into account the last two cases.

In case (iv), only about 10% of plagiarized statements could be detected wherein 30 plagiarized statements were constructed by changing most of words with their synonyms

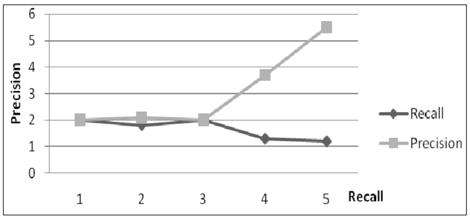

Figure 3. Precision & Recall

and antonyms. Our Arabic fuzzy-set IR detected only less than half of them. That is, the problem is not with restructuring but with rewording. This is because our *term-to-term correlation matrix* does not have enough pairs of *terms* with their synonyms and antonyms. Besides, we found that 25% of plagiarized statements detected in case (v) were mostly restructured but not reworded statements. Precision and recall in Figure 3 were calculated as illustrated in (5) and (6). As can be seen, the first three cases were optimal or near optimal since most of the statements were either duplicates or semantically the same but with different structure. In contrast, the effectiveness of our retrieval model measured in precision and recall in the last two cases was not encouraging, which gives a remarkable point for further enhancement of our model.

TABLE V. EXPERIMENTAL RESULTS

Case		**# Statements**			**Arabic Fuzzy-Set IR**		
		CDoc	***QDoc***	***Total***	***#Plagiarized***	***#Detected***	***R%***
i	duplicate	43	43	86	43	43	100
ii	closely related	67	67	134	67	59	88
iii	unrelated statement	69	73	142	69	69	100
iv	moderately related	80	80	160	30	11	14
v	any (restructuring & rewording)	All	50	4417	45	10	0.23

a. QDoc refers to Query Document, CDoc refers to corpus collection, and # indicates "number"

VI. CONCLUSION AND FUTURE WORK

We have presented the work done to construct Arabic fuzzy-set IR approach for the purpose of plagiarism detection. Based on the nature of Arabic which makes plagiarism easily done by restructuring and rewording statements rather than only copy and paste, it has been concluded that fuzzy-set IR is a suitable approach. Our Arabic fuzzy-set IR approach captures duplicates and similar statements with different structure perfectly. However, it does not handle cases of rewording with different synonyms/antonyms. This gap can be bridged by modeling fuzzy sets based on Arabic thesaurus. Future work also includes increasing the corpus collection with documents discussing the same subject, rather than varying topics, which can enhance the overall performance of the system. Besides, more test cases should be further included to increase the reliability of results in our approach.

REFERENCES

[1] Heintze, N. *Scalable document fingerprinting*. in *the Second USENIX Workshop on Electronic Commerce*. 1996.

[2] Shivakumar, N. and H. Garcia-Molina, *SCAM: A Copy Detection Mechanism for Digital Documents.* D-Lib Magazine 1995.

[3] *WCopyFind*. [cited 2008 April, 23]; Available from: http://www.plagiarism.phys.virginia.edu.

[4] Lyon, C., R. Barrett, and J. Malcolm, *Plagiarism is Easy, but also Easy To Detect.* Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification., 2006. **I**: p. 57-65.

[5] Bao, J., C. Lyon, and P. Lane, *Copy detection in Chinese documents using Ferret.* Language Resources and Evaluation, 2006. **40**(3): p. 357-365.

[6] Yerra, R. and Y.-K. Ng, *A Sentence-Based Copy Detection Approach for Web Documents*, in *Fuzzy Systems and Knowledge Discovery*. 2005. p. 557-570.

[7] Antonio, S., L. Hong Va, and W.H.L. Rynson, *CHECK: a document plagiarism detection system*, in *Proceedings of the 1997 ACM symposium on Applied computing*. 1997, ACM: San Jose, California, United States.

[8] Kang, N., A. Gelbukh, and S. Han, *PPChecker: Plagiarism Pattern Checker in Document Copy Detection*, in *Text, Speech and Dialogue*. 2006. p. 661-667.

[9] Sebastian, N. and P.W. Thomas, *SNITCH: a software tool for detecting cut and paste plagiarism*. 2006, ACM. p. 51-55.

[10] Stefan, G. and N. Stuart, *Tool support for plagiarism detection in text documents*, in *Proceedings of the 2005 ACM symposium on Applied computing*. 2005, ACM: Santa Fe, New Mexico.

[11] Meyer zu Eissen, S., B. Stein, and M. Kulig, *Plagiarism Detection Without Reference Collections*, in *Advances in Data Analysis*. 2007. p. 359-366.

[12] Byung-Ryul, A., K. Heon, and K. Moon-Hyun. *An Application of Detecting Plagiarism using Dynamic Incremental Comparison Method.* in *International Conference on Computational Intelligence and Security.* 2006.

[13] Sergey, B., et al., *Copy detection mechanisms for digital documents*. 1995, ACM. p. 398-409.

[14] Kriszti, M., Z. Arkdy, and S. Heinz, *Document overlap detection system for distributed digital libraries*, in *Proceedings of the fifth ACM conference on Digital libraries*. 2000, ACM: San Antonio, Texas, United States.

[15] Raphael, A.F., et al., *Signature extraction for overlap detection in documents*, in *Proceedings of the twenty-fifth Australasian conference on Computer science - Volume 4*. 2002, Australian Computer Society, Inc.: Melbourne, Victoria, Australia.

[16] Koberstein, J. and Y.-K. Ng, *Using Word Clusters to Detect Similar Web Documents*, in *Knowledge Science, Engineering and Management*. 2006. p. 215-228.

[17] Ogawa, Y., T. Morita, and K. Kobayashi, *A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method.* Fuzzy Sets and Systems, 1991. **39**: p. 163–179.

[18] Cross, V., *Fuzzy information retrieval.* Journal of Intelligent Information Systems, 1994. **3**(1): p. 29-56.

[19] Haddad, H., H.M. Harmain, and H. El-Katib, *Arabic Natural Language processing for Information Retrieval.* The UAEU 7th Annual Conference, Al Ain, UAE, 2006.

[20] Alzahrani, S.M. and N. Salim, *A modified algorithm for stemming and stopwords removal for effective Arabic IR*. 2008. unpublished.

[21] Salem, M., *Comparison and Fusion of Retrieval Schemes Based on Different Structures, Similarity Measures and Weighting Schemes*. 2006, Universiti Teknologi Malaysia.

[22] Taghva, K., R. Elkhoury, and J. Coombs. *Arabic stemming without a root dictionary*. in *International Conference on Information Technology: Coding and Computing.* 2005.

[23] *Arabic Wikipedia*. [cited 2008 August 24]; Available from: August 24, 2008