Using K-means Cluster Based Techniques in External Plagiarism Detection

Vani K
Department of Computer Science
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Bangalore, India
k_vani@blr.amrita.edu

Deepa Gupta
Department of Mathematics
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Bangalore, India
g_deepa@blr.amrita.edu

Abstract— **Text document categorization is one of the rapidly emerging research fields, where documents are identified, differentiated and classified manually or algorithmically. The paper focuses on application of automatic text document categorization in plagiarism detection domain. In today's world plagiarism has become a prime concern, especially in research and educational fields. This paper aims on the study and comparison of different methods of document categorization in external plagiarism detection. Here the primary focus is to explore the unsupervised document categorization/ clustering methods using different variations of K-means algorithm and compare it with the general N-gram based method and Vector Space Model based method. Finally the analysis and evaluation is done using data set from PAN-2013[1] and performance is compared based on precision, recall and efficiency in terms of time taken for algorithm execution.**

Keywords— Text Document Categorization; External Plagiarism; Candidate Retrieval; N-gram; Vector Space Model; K-means Clustering

1. INTRODUCTION

Automatic text document classification or categorization is the process of assigning documents to one or more specific categories algorithmically. There is no question concerning the commercial value of being able to classify documents automatically by content as it has myriad potential applications [1]. With the rapid development in technology, information overload has become a major problem and sorting out relevant documents and classifying them accurately has become a challenge. Document classification can be categorized as supervised, unsupervised and semi-supervised. In supervised document classification, some external feedback is available to provide the correct classification information. Unsupervised classification, also termed as document clustering performs classification without any reference to external information while in semi-supervised certain parts of document is labeled using external information [2].

Application areas of document classification are numerous, via language identification, spam filtering, genre classification, sentiment analysis, information retrieval (IR) and so on. Another such possible application area is plagiarism detection. Document classification methods are less explored in this domain. Rapid development of network technology, including large numbers of search engines, document repositories, translation software systems etc. not only provides people with the various knowledge acquisition channels, but also opens the door for text plagiarism. Plagiarism generally refers to the illegitimate use of someone else's information, text, ideas, etc. without proper reference to the original source of these data [3].

The type and degree of plagiarism ranges from the simplest to the more challenging and complex one's. External/extrinsic and Internal/intrinsic are the two main plagiarism detection methods. Extrinsic plagiarism evaluates plagiarism in accordance to one or more source documents while the intrinsic method analyze changes in the unique writing style of an author as an indicator for potential plagiarism[4]. This paper aligns on extrinsic plagiarism detection where the source documents are available in the form of some dataset/ the web/any online sources. In extrinsic plagiarism detection, five main tasks are generally performed [4]. Pre-processing is the initial step where the source and suspicious documents are subjected to certain refinements like stop-word removal, tokenization, lowercasing, sentence segmentation, punctuation removal etc. This helps to reduce the size of actual data by removing the irrelevant information with respect to the approach used. Next stage is candidate retrieval in which a subset of source document is retrieved that is globally similar to the suspicious document. In the third stage detailed comparison of each suspicious document against its candidate document set is done to retrieve the plagiarized fragments. Here different natural language processing (NLP) techniques like chunking, parsing, part of

The authors gratefully acknowledge Department of Science and Technology, Govt. of India (www.dst.gov.in), for sponsoring this research project, SERB/F/1511/2014-2015.

[1] http://pan.webis.de/

speech (POS) tagging etc can be adopted. Fourth stage, passage boundary detection deals with estimating the exact boundaries of plagiarized passages in both source and suspicious documents using certain criteria. Finally evaluation of the system is done using some standard measures, via PAN measures to rate its performance.

Text document classification can be applied in the initial phases of external plagiarism detection method to retrieve documents which are almost similar to the given suspicious document, i.e. candidate retrieval. This reduces the number of document comparisons in the detailed analysis stage and hence the complexity. The paper focuses on the comparative study and analysis of N-gram based methods, Vector Space Model (VSM) based methods and K-means clustering based methods when used in the candidate retrieval stage of external plagiarism detection. The methods are also compared based on NLP techniques like stemming, lemmatization and chunking.

2. RELATED WORKS

In external plagiarism detection, candidate retrieval can be viewed as an information retrieval task. Based on this concept, different candidate retrieval methods using IR models are proposed by eminent researchers. In this category, finger print or N-gram based method, hash-based method and vector space models have gained popularity. S Schleimer [5] proposes a finger print based method in which the document is divided into K-grams where 'K' is a user specified value. Then each K-gram is hashed and some of them are selected as finger prints of the document using a specific window size. The results are evaluated using Stanford Webbase. The method could detect matches of certain length only. William B. Cavnar & John M. Trenkle [6] presents an N-gram based text categorization method where initially documents are represented as N-gram profiles. Then using some distance measure the similarity between documents are computed and they are classified. It also calculates the N-gram frequency using Zipf's Law to take into account the top N-grams. For testing, the data from five newsgroups are used and it gives good results for classification of various news articles. Peter Nather [7] proposes a similar method in which 1-gram to 5-grams are used for experiments. The similarity measure of two N-gram profiles is taken as the sum of differences between the rank of the N-grams in one profile and the rank in the other profile. The method is evaluated using data from Project Gutenberg and it is found that the method performed well with short documents. With large volume and manipulations in content the method performance decreased.

In [8], 4 to 6 N-grams are formed to represent document profiles. Then the document ID is mapped to a set of hash values constructed. An inverted index, mapping the hash value to the sequence of document IDs is computed from it. Further similarity is calculated using Jaccard coefficient and the system is evaluated using PAN-09 dataset. An N-gram based candidate retrieval approach is proposed in [9] which uses shingling and Jaccard coefficient approach in candidate retrieval stage. Here after the required pre-processing N-gram profiles typically of N= 3 or 4 is made and then Jaccard similarity is calculated. A threshold of 0.1 is used to retrieve the candidate set of documents. The final system evaluation is done using PAN-10 dataset. Rajiv Yerra & Yiu –Kai Ng [10] performs a sentence-based copy detection method for web documents. Here the similarity between document sentences is calculated using the three least frequent 4-gram approaches where the rare tokens are taken into account. Fuzzy IR approach is also discussed, which outperformed the former one.

Another popular retrieval model is Vector Space Model (VSM) which is also used in candidate retrieval task. In [11] 16 grams are made which is then represented as vectors using tf-idf values and further document distance is measured using cosine similarity. Muhr.et.al [12] proposes a VSM based approach where initially vectorization of source and suspicious documents are done. Then the nearest documents corresponding to each suspicious document is found from the reference corpus using cosine similarity. Similar approach is carried in [13] and [14] for candidate retrieval and further the system is evaluated using PAN data sets. In VSM methods execution time increased rapidly with large datasets.

Self-Organizing Maps (SOM) is used in [15] for data organization and clustering. It uses a multi-layer tree structured SOM (MLSOM) where node data in different levels of a tree are processed in different layers of the MLSOM. Liping Jing et al.[16] presents a document clustering approach using K-means algorithm to cluster sparse data. It also automatically calculates the weights of keywords in each cluster to identify their importance. Experiments are conducted using datas from 20-Newsgroups collection and the results are promising. Different works are done using clustering techniques in IR and classification domain. But when it comes to the external plagiarism detection domain, the use of clustering techniques is too limited. Duo Zuo et al. [17] proposes an external plagiarism detection method which uses clustering for post processing. Here initially a pre-selecting step is done to narrow the scope of detection using the successive same fingerprint. Then all fragments between two documents are found and merged. Further clustering is done to reduce the impact of obfuscated text. The system is finally evaluated using PAN-10 data set. Thus from the literature survey conducted, it is observed that IR models, via N-gram based and VSM based methods are widely used in external plagiarism detection while clustering techniques are less explored. Unsupervised document classification method/ document clustering method have good potential to be applied in the candidate retrieval stage of external plagiarism detection. In this paper, document clustering using K-means algorithm is explored for candidate retrieval task. Different extensions of basic K-means using NLP techniques and N-grams are also discussed. The algorithms are evaluated using partial dataset from PAN-13 corpus and compared with the basic N-gram based methods and classical VSM based

method. The methods proposed and compared are discussed in the following sections.

3. METHODS PROPOSED AND COMPARED

This section describes the various methods proposed and compared in this paper. The following methods are discussed:

1. N-gram based method
2. Vector Space Model (VSM) Method
3. Cluster based method using K-means Algorithm
4. K-means with stemming
5. K-means with lemmatization
6. K-means with N-grams
7. K-means with chunking

Initially some basic pre-processing of the text document is done. This include tokenization, punctuation removal and stopword removal (Stopwords are the words without semantic meaning). A list of the 50 most frequent words of the English language provided by the British National Corpus[2] which includes about 90 million tokens are usually used. In the method which uses chunking, initially chunks are formed and then stopwords are removed, if any. First the traditional N-gram based method and VSM method is used for candidate retrieval. Then cluster based method is adopted using proposed K-means algorithm. Further different variations of K-means are also carried out and results are analyzed and compared.

3.1. N-gram based Method

Here initially the general pre-processing is carried out. Then the document is divided into N-grams or N-shingles. This refers to a sequence of consecutive words of size 'N', where 'N' is user specified. Both suspicious and source documents are converted to their N-gram profiles and similarity is calculated using Dice's coefficient. This is similar to Jaccard coefficient but it reduces the effect of shared terms between the documents. Let X_{suspng} and X_{srcng} be the suspicious and source N-gram profiles, then Dice's coefficient defined as:

$$Dice(X_{suspng}, X_{srcng}) = 2\frac{\left|X_{suspng} \cap X_{srcng}\right|}{\left|X_{suspng}\right| + \left|X_{srcng}\right|} \quad (1)$$

Consider the following English sentence (E1):

"The people left their countries and sailed with Gilbert."

After initial pre-processing, the tokens obtained (E1-tokens):

['people', 'left', 'countries', 'sailed', 'Gilbert']

After trigram formation, N=3 (E1-3-gram):

[['people', 'left', 'countries'], ['left', 'countries', 'sailed'], ['countries', 'sailed', 'Gilbert'], ['sailed', 'Gilbert']]

After the formation of both source and suspicious N-gram profiles, similarity is calculated using (1). Instead of using some threshold value to select candidate set, here the similarity between each source with all suspicious is calculated. Then the suspicious document which is having maximum Dice's coefficient measure is selected as the related document.

[2] http://www.natcorp.ox.ac.uk/

3.2. *Vector Space Model (VSM) Method*

Vector Space Model (VSM) is an algebraic model representing textual information as a vector. Here, after the required initial pre-processing, a dictionary of terms (vocabulary) is extracted from each source document which is compared against all the suspicious documents. VSM represents the importance of a word using term frequency-Inverse document frequency (tf-idf) metric. Inverse document frequency, idf(t) is then calculated which emphasizes that a term which is almost present in the entire corpus of documents is not good. Finally their product, i.e. tf-idf is calculated and the similarity between document vectors is calculated using cosine similarity. Cosine similarity between the two documents X_{susp} (suspicious) and X_{src} (source) is calculated as given below:

$$Cos(X_{susp}, X_{src}) = \frac{V(X_{susp}) \bullet V(X_{src})}{\left\|V(X_{susp})\right\| \left\|V(X_{src})\right\|} \quad (2)$$

Here $V(X_{susp})$ and $V(X_{src})$ represents the document vector representation of suspicious and source respectively. The numerator in (2) denotes the dot product of these vectors and the denominator denotes the product of their Euclidean norms. After calculating similarity using (2), the candidate documents are selected by the approach similar to that in N-gram method. Thus each source document is compared against all suspicious and the suspicious document with maximum cosine similarity is selected. It must be noted that there can be cases in which a source document is unrelated to any suspicious document, via when the source is the entire web. But the above non-thresholding approach is efficient in candidate retrieval task as here further detailed analysis is carried out to detect the actual plagiarized documents.

3.3. *Cluster based Method with K-means Algorithm-Proposed Method*

In this method clustering approach is used, where the similar documents are grouped together as a cluster. Here the algorithm used is K-means which is an efficient partition clustering technique [18]. In basic K-means clustering algorithm the two main parameters are the number of clusters (K) and the initial cluster centres/ centroids. Basic algorithm is given below:

1. Select K and the initial centroids, where 'K' no. of centroids is to be selected.
2. Assign each object to the group that has the closest centroid using some distance or similarity measure.

3. When all objects have been assigned, recalculate the positions of the 'K' centroids.

4. Repeat Steps 2 and 3 until the centroids no longer change.

Here the main problem is to decide the value of 'K' and the initial centroids, as these two parameters completely regulates the algorithm results. Considering these limitations, in the proposed approach these parameters are given fixed values. Here the concept used is that each suspicious document acts as the centroid. The source documents which are globally similar to suspicious documents are grouped to the cluster containing this suspicious document as the centroid. Thus 'K' is taken as the total no. of suspicious document, assuming that the suspicious corpus is given (PAN corpus). Using this concept, the basic K-means algorithm is modified for candidate retrieval task as given:

1. Set K = no.of suspicious documents.

2. Set initial K centroids = Each of the K suspicious documents.

3. Assign each source document to the cluster with closest centroid using cosine similarity.

The process is implemented using a Python package[3] that facilitates clustering and reduces the time taken to build document vectors and calculate their similarity. Then based on the similarity measure the source documents are grouped to their corresponding suspicious document. Thus each cluster corresponds to the candidate set of documents for a particular suspicious document. In the proposed algorithm, a single iteration is required as the centroids remain fixed. This makes it efficient in terms of time complexity.

3.4. Variations of Proposed K-means algorithm

The K-means candidate retrieval method discussed in Subsection 3.2. is used as the basic approach. Then certain extensions are brought into it and algorithm is evaluated. These methods are discussed in following sections.

3.4.1. K-means with Stemming(K-Stem)

In this method, the only change is that after tokenization of the document, stemming is carried out. Stemming is a heuristic process of removing the affixes from the words. Remaining steps are carried out similar to the basic approach.

3.4.2. K-means with Lemmatization(K-Lem)

This method uses lemmatization instead of stemming. Lemmatization produces the dictionary base forms of a word using vocabulary and morphology. It is closely related to stemming but stemming operates only on a single word at a time while lemmatization operates on the full text. It can thus discriminate between words that have different meanings depending on part of speech.

[3] http://www.clips.ua.ac.be/pattern

Table 1: Data Statistics

No. of Documents		
	Suspicious Document	***Source Document***
Set 1	39	205
Set 2	31	213
Set 3	35	209

3.4.3. K-means with N-gram(K-Ng)

Here K-means method is combined with N-gram based method. Instead of taking individual words N-grams are made and further proceedings is using the basic approach.

3.4.4. K-means with Chunking(K-Chk)

The method uses chunking to form grammatical phrases instead of dealing with unigrams. Initially a parse tree is constructed. Then noun phrases, verb phrases, adjective phrases and adverb phrases are extracted from it as these phrases contribute to the semantic meaning of a sentence. The algorithm is also evaluated using the pre-processing steps, stemming and lemmatization discussed in Subsections 3.3.1 and 3.3.2, i.e. K-Chk-Stem and K-Chk-Lem.

4. EXPERIMENTAL SETTINGS AND RESULT ANALYSIS

4.1. Data statistics

Algorithms are evaluated using three sets of documents taken from PAN-13 corpus [19]. Each set has suspicious and corresponding source document sets as shown in Table 1. Three sets of data used are:

- Set-1 : No Obfuscation
- Set-2 : Random Obfuscation
- Set-3 : Translation Obfuscation

No obfuscation set contains document pairs where the suspicious documents are exact copies of the source document. In random obfuscation the passages in document pairs are obfuscated with word shuffling, replacement with synonyms etc. Translation obfuscation set contains document pairs whose passages are run through a sequence of translations. Here the output of one translation becomes input to next translation and the final translation language is same as the original document language [19]. Due to the hardware limitations only some of the documents from these sets are used for the evaluation.

4.2. Evaluation

The methods discussed in Section 3 are evaluated using the measures, Recall (rec), Precision (prec), and Execution Time. In the scenario of candidate retrieval, recall is defined as the no. of relevant documents retrieved to the actual no. of relevant documents to be retrieved.

$$rec = \frac{\#\ of\ \ relevant\ \ \ documents\ \ \ retrieved}{Actual\ \ \#\ of\ \ relevant\ \ \ documents} \quad (3)$$

Precision is defined as the no. of relevant documents retrieved to the total no. of documents retrieved by the system.

$$prec = \frac{\#\ of\ \ relevant\ \ \ documents\ \ \ retrieved}{Total\ \ \#\ of\ \ \ documents\ \ \ retrieved\ \ \ by\ system} \quad (4)$$

4.3. *Results and Discussions*

Each set is evaluated using algorithms discussed in Section 3 and the results are compared. The data statistics given in Table 1 is used for the evaluation. The results obtained for each set in terms of precision, recall and execution time using the methods discussed in Section 3 is given in Fig.1, Fig.2 and Fig.3. Here recall and precision is shown in the primary axis and execution time in the secondary axis. Analyzing Fig. 1, it can be observed that N-gram based methods give good results for Set-1(No obfuscation set). It is also observed that as the 'N' value increases, both recall and precision degrades slightly. In this paper, only the basic classical VSM method is implemented and evaluated. Analyzing each of these plots, it can be noticed that VSM method gives a considerably reduced precision and recall with all the three sets. It can be also observed that this method requires high execution time compared to other approaches. VSM method is found to give good results when used along with approaches like ranking of documents, Latent Semantic Indexing (LSI) [20] etc. But time complexity still remains as a major limitation of this method especially when dealing with large datasets. From the plot in Fig.1, it is seen that for the proposed K-means method, recall and precision increases when compared to N-gram and VSM method. In K-means method the execution time decreases substantially. K-Chk and K-Ng(N=10) methods also give good precision and recall but with increased execution time. This is because formation of chunks in K-Chk method and N-grams in K-Ng method is time consuming. Thus K-means method outperforms the other two approaches when the documents are not obfuscated (Set-1). All the results show maximum 3 to 5 point difference only. This is because only small data set is taken for evaluation.

Fig.2. analyzes and compares the methods using Set-2 i.e. randomly obfuscated set. It can be observed that K-means method and its variations, via K-Chk and K-Ng (N=10) outperforms the N-gram based and VSM based methods. In N-gram based method N=3 and N=4 gives good results. Further it is found that K-means method takes considerably less execution time for Set-2 also. Fig 3. plots the evaluation of methods using Set-3 (translation obfuscation). The complexity of obfuscation is high in this set compared to other two sets. Analyzing the graph in Fig.3. and comparing the different methods, it can be noted that the proposed K-means clustering method surpasses N-gram method and VSM based methods when evaluated using Set- 3. In terms of precion and recall K-means and K-Ng(N=10) outperforms other methods but K-Ng(N=10) takes substantially large execution time.

From the comparisons and analysis made, it is observed that K-means method gives promising results when dealing with highly obfuscated data compared to the other two approaches. Further it can be concluded that in terms of execution time the proposed K-means algorithm gives performs efficiently with all the sets discussed in Subsection 4.1. Thus the proposed method fastens the candidate retrieval stage without compromising the accuracy of retrieval. Any algorithm's efficiency depends upon both accuracy and time, hence a trade-off between these two factors is to be considered.

5. CONCLUSION AND FUTURE SCOPE

The paper explores the application of automatic text document categorization in external plagiarism detection. Here an attempt is made to use the clustering techniques in candidate retrieval stage of external plagiarism detection.

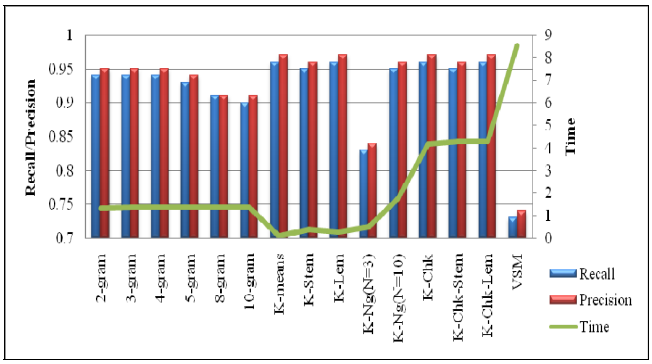

Fig.1. Recall , Precision & Execution Time of Set1

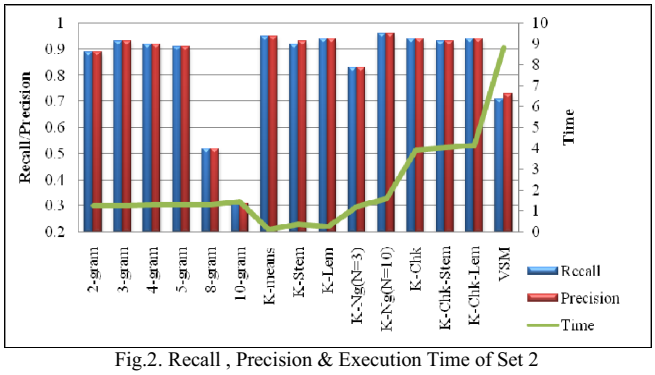

Fig.2. Recall , Precision & Execution Time of Set 2

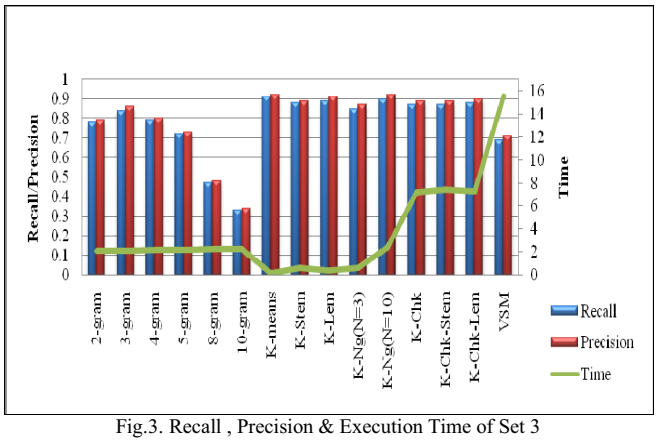

Fig.3. Recall , Precision & Execution Time of Set 3

Efficient candidate retrieval can help in reducing the number of document comparisons and hence the time complexity in the exhaustive local analysis stage of detection. Here N-gram based methods, Basic VSM approach and K-means cluster based methods are analyzed and compared. A new K-means approach for retrieval of candidate documents is proposed. The paper also throws light on certain variations that can be used to extend the proposed k-means algorithm. From the results and discussions in Section 4 it can be concluded that proposed K-means clustering algorithm surpasses both N-gram method and VSM method. Further the method gives promising results for documents that are obfuscated using methods like translation which is usually hard to detect. The K-means variations proposed using different NLP techniques via K-Ng (N=10), K-Stem, K-Lem, K-Chk, K-Chk-Stem, K-Chk-Lem also yield good recall and precision. In terms of execution time, which is a vital factor in any software system, the proposed method performs effectively. It speeds up the candidate document retrieval task by considerably reducing the execution time.

As future work, K-means approach with Word Net[4] can be used for better document comparisons, in more intelligent terms. The documents which are intelligently manipulated using synonyms can hence be retrieved. To improve the recall, Fuzzy-K-means algorithms can be employed which provide soft clustering. Multithreading and multiprocessing techniques can be used to improve the time efficiency of algorithms especially when dealing with large datasets.

6. REFERENCES

[1] Peter Jackson and Isabelle Moulinier,Natural Language Processing for Online Applications:Text Retrieval, Extraction , and Categorization,JAN 2002,pp.119-225.

[2] Arzucan Ozgur, "Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization",MSc.Bogazici University,2004.

[3] Ahmed Hamza Osman1, Naomie Salim and Albaraa Abuobieda, "Survey of Text Plagiarism Detection", Journal of Computer Engineering and Applications vol.1, June 2012.

[4] Salha M. Alzahrani, NaomieSalim, and Ajith Abraham, " Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods", IEEE transactions on systems, man, and cybernetics,vol.42, no.2, march 2012.

[5] S. Schleimer, D. Wilkerson, and A. Aiken, "Winnowing: Local algorithms for document fingerprinting," In *Proc. ACM SIGMOD Int. Conf.Manage. Data*, New York, 2003, pp. 76–85.

[6] William B. Cavnar , and John M. Trenkle, "N-Gram-Based Text Categorization", In Proc. of SDAIR-94, 3rd Annual Symposium On Document Analysis And Information Retrieval.

[7] Peter Nather,"N-Gram-Based Text Categorization",Thesis. Bratislva University,2005.

[8] J.Kasprzak,M.Brandejs, and M.K ripac, "Finding plagiarism by evaluating document similarities," In *Proc. SEPLN*, Donostia, Spain, pp. 24–28.

[9] S. Alzahrani, "Plagiarism auto-detection in arabic scripts using statement-based fingerprints matching and fuzzy-set information retrieval approaches," M.Sc. thesis, Univ. Technol. Malaysia, Johor Bahru, 2008.

[10] R. Yerra and Y.-K. Ng, "A sentence-based copy detection approach for web documents," in *Fuzzy System and Knowledge Discovery*, 2005, pp. 557–570.

[11] C. Grozea, C. Gehl, and M. Popescu, "ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection," in *Proc. SEPLN*, Donostia, Spain, 2012, pp. 10–18.

[12] M. Zechner, M. Muhr, R. Kern, and M. Granitzer, "External and intrinsicplagiarism detection using vector space models," in *Proc. SEPLN*, Donostia, Spain, pp. 47–55.

[13] Asif Ekbal, Sriparna Saha and Gaurav Choudhary, " Plagiarism Detection in Text using Vector Space Model, In Proc. of 12th International Conference on Hybrid Intelligent Systems(HIS), pp.366-371, Pune, 2012.

[14] Rasia Naseem and Sheena Kurian, " Extrinsic Plagiarism Detection in Text Combining VSM and Fuzzy Semantic Similarity Scheme", Journal of Advanced Computing, Engineering and application(IJACEA),vol.2, December 2013.

[15] T. W. S. Chow, and M. K. M. Rahman, "Multilayer SOM with treestructured data for efficient document retrieval and plagiarism detection,"*IEEE Trans. Neural Netw.*, vol. 20, no. 9, pp. 1385–1402, Sep. 2009.

[16] D. Zou, W.Long, and Z. Ling, "A cluster-based plagiarism detection method - Lab report for PAN at CLEF 2010", In Proc. of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse. Padua,Italy, 2010.

[17] Liping Jing, Michael K.Ng,Jun Xu, and Joshua Zhexue Huang, "Sub-space Clustering of text documentswith feature weighting K-means algorithm", In Proc. of the 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining,pp.802-812,Berlin,2005.

[18] Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification,2nd ed.,Wiley:2000.

[19] Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann(et al.),"Overview of 5th International Competition on Plgiarism Detection, CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September,Valencia, Spain. ISBN 978-88-904810-3-1. ISSN 2038-4963. 2013.

[20] Ch. Aswani Kumar, and S. Srinivas, "On the Performance of Latent Semantic Indexing-based Information Retrieval," Journal of Computing and Information Technology - CIT 17,2009, pp.259–264.

[4] http://wordnet.princeton.edu/