Exploration of Fuzzy C Means Clustering Algorithm in External Plagiarism Detection System

N. Riya Ravi, K. Vani and Deepa Gupta

Abstract With the advent of World Wide Web, plagiarism has become a prime issue in field of academia. A plagiarized document may contain content from a number of sources available on the web and it is beyond any individual to detect such plagiarism manually. This paper focuses on the exploration of soft clustering, via, Fuzzy C Means algorithm in the candidate retrieval stage of external plagiarism detection task. Partial data sets from PAN 2013 corpus is used for the evaluation of the system and the results are compared with existing approaches, via, N-gram and K Means Clustering. The performance of the systems is measured using the standard measures, precision and recall and comparison is done.

1 Introduction

"Plagiarism is the reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source" [1]. Plagiarism corresponds to copying the work or idea of another author and presenting it as own without acknowledging the original work. It can be a literal copy of the material or part of it from another source. However, the author may modify some portions of it by inserting or deleting some parts of text. It also includes cases where the author just copies the idea from another, writing it in his own words by paraphrasing or summarizing the original work, which is termed as intelligent plagiarism [2]. Plagiarism is of serious concern as the works of many authors are readily available to anyone through the high availability of the World Wide Web (WWW). Many tools and techniques have been built and analyzed over the years

N.R. Ravi(✉) · K. Vani · D. Gupta
Department of Computer Science, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Bangalore, India
e-mail: riya.sanjesh@gmail.com, {k_vani,g_deepa}@blr.amrita.edu

S. Berretti et al. (eds.), *Intelligent Systems Technologies and Applications*,
Advances in Intelligent Systems and Computing 384,
DOI: 10.1007/978-3-319-23036-8_11

but still lacks proper detection efficiency. Plagiarism detection techniques can be Intrinsic or Extrinsic. Former detection techniques involve detecting plagiarized passages within a document using the author's style of writing. Latter detection techniques on the other hand compare multiple existing documents with suspected documents and try to figure out the plagiarized passages [2]. Some of the External detection techniques are String matching, Vector Space Model (VSM), Finger printing etc. In general, the major steps followed in an Extrinsic plagiarism detection system (PDS) are pre-processing, candidate document identification, detailed analysis and postprocessing. Pre-processing is the initial step and its purpose is to discard all irrelevant information from the source and suspicious documents .Candidate document identification is an important step , where the candidate source documents for each suspicious document is picked out and this further reduces the complexity of the final processing. Taking the whole set of the source documents for a detailed plagiarism detection is too cumbersome and a time consuming process [3]. Most of the currently available PDS depends on N-gram based approaches which are inherently slow. Thus finding naïve and efficient futuristic method or technology is highly in need. Machine learning (ML) approaches are explored in detail in the text classification domain with great success. Such ML techniques can be experimented with in various stages of the plagiarism detection systems as well.

In this paper the main focus is given to the candidate retrieval stage of extrinsic PDS. Usually the complete PDS evaluation is done and efficiency is measured rather than evaluating candidate retrieval task separately. But in practical scenario, the incorporation of this stage is important to aid the final exhaustive analysis of an extrinsic PDS. Here ML approaches, via, clustering techniques are used for candidate retrieval task. A Fuzzy C-Means (FCM) based clustering approach is used and compared with the existing candidate retrieval approaches, via, N-gram based and K Means clustering approach.

2 Literature Survey

The concept of Plagiarism was first introduced in the field of Arts in the first century but the word was adopted in English language in around 1620. In the twentieth century plagiarism was getting noticed as a big menace in Academia and Journalism [4]. Serious actions were taken against the people involved in the act. As pointed out by Alzahrani et.al [2], the first code plagiarism detection tools came in 1970s but it was not before 1990s we saw first tool to detect extrinsic plagiarism detection in natural language documents. Alzahrani et.al [2] describes the taxonomy of Plagiarism and plagiarism detection types, via, Extrinsic and Intrinsic. Some of the popular tools available for plagiarism detection are COPS, SCAM, CHECK and Turnitin which are discussed by Zdenek Ceaska [3]. These tools mainly focuses on copy & paste detection in the extrinsic plagiarism. Documents were represented as set of words/sentences and compared. This made the detection slow. To circumvent this problem, feature reduction was introduced

to reduce the detection time. The features used to represent the document are reduced with one of the Feature reduction techniques like TF-IDF, Information Gain, Chi^2 [2]. The other techniques used to reduce the complexity are the Stop words removal, lemma etc. To reduce the number of the sources against which a given suspicious document is compared could be further reduced using one of the Candidate retrieval technique like Fingerprint, Hash based, Vector Space Model, Latent Semantic Indexing [2] etc. Other document representations and comparison techniques were developed to improve the detection. One of the ways was to represent the document as a graph [5]. Alzahrani and Salim introduced a semantic based plagiarism detection technique which used fuzzy semantic-based string similarity. Their process can be subdivided into four stages. In the first stage text is tokenized, stop words removed and stemmed. Second stage involved identifying a set of candidate documents for each suspicious document using Jaccard coefficient and shingling algorithm. Detailed comparison is carried out next between the suspicious document and the corresponding candidate documents. In this stage a fuzzy similarity is calculated. If the similarity is more than a threshold then the sentences are marked as similar. In the end a post processing is carried out where the consecutive sentences are combined to form a paragraph. Asif Ekbal et. al. [7] proposed a Vector Space Model (VSM) to overcome the problems which arises due to matching of strings. In this case the document is represented using vectors and similarity is calculated based on the cosine similarity measure. The downside of VSM model is failure to detect disguised plagiarism. It also involves lot of effort in computation.RasiaNaseemet.al. [8] utilizes VSM approach for candidate retrieval task and uses a fuzzy semantic similarity measure for detailed analysis. A further improvement on this technique was done by Ahmed Hamza et.al. [9] where Semantic Role Labeling (SRL) is used. Although this technique gives good detection, it takes lot of time and have scalability issues. Alberto and Paolo [10] proposes a basic N-gram approach for locating the plagiarized fragments. Word n-gram documents are compared with each other to detect the similarity. This approach is time consuming and did not consider any intelligent aspects, via, synonyms. N-gram method is the base for most of the detection techniques like character based, syntactic based, vector based, semantic based etc. These techniques have been traditionally slow and thus the plagiarism detection tools that utilizes these techniques are also slow in nature. Most of the PDS employ string matching approach only mainly in candidate retrieval stage, which means that such systems cannot detect intelligent plagiarism effectively.

Computational Intelligence or ML techniques are not much explored in plagiarism detection domain but are popular in document classification domains. Document classification deals with classification of text documents into multiple categories. In the field of text classification, ML techniques such as Naïve Bayes, Support Vector Machine (SVM) etc. along with clustering techniques have been explored in detail. Chanzing et. al. [11] focus on feature selection using Information Gain along with SVM for classification. K Nearest Neighbour (KNN), Naïve Bayes and Term Graph techniques are well compared by Vishwanath et. al. [12]. Even

Genetic algorithm technique have been employed in [13] for document classification but with not much of success. KNN gives good accuracy but with high time complexity. K Means, KNN and the combination and variations of them are used for text classification applications widely [14,15,16,17]. But these ML techniques are less explored in plagiarism domain. Most of the papers in plagiarism detection domain talks about the exhaustive analysis stage giving less importance to the pioneer candidate retrieval stage. A recent paper by Vani and Gupta [18] explores the ML technique, via, K Means clustering based technique specifically for the candidate retrieval stage. K Means clustering creates K non-overlapping clusters of the document features and thus produce hard boundaries.

To overcome this problem, the proposed method uses a soft clustering approach, via, Fuzzy C Means (FCM) clustering which creates K clusters which can be overlapping. Thus the objective of this paper is the exploration of FCM clustering algorithm in the candidate retrieval step of Extrinsic Plagiarism Detection and the comparative study with existing candidate retrieval approaches.

Further the method is explored using NLP techniques, via, stemming, lemmatization and chunking. These algorithms are evaluated on PAN 2013 data set and compared.

3 Existing Candidate Retrieval Approaches

Traditional N-gram based approach [10] and K Means [18] based clustering approach, the two existing algorithms for the candidate retrieval stage are explained in detail in Sub-Sections 3.1 and 3.2 respectively.

3.1 N-gram Approach

N-gram is a contiguous sequence of N items from a given sequence of text or speech. This contiguous sequence of N items can be words or character. Given a suspicious document and a set of source documents, the objective is to find out the relevant candidate source documents using N-gram method. Here, N-grams are generated from both suspicious and source. Similarity between each pair of document is calculated based on the Dice coefficient measure as in (1).

$$C(d_{sus}, d_{src}) = \frac{2|N(d_{sus}) \cap N(d_{src})|}{|N(d_{sus}) \cup N(d_{src})|} \quad (1)$$

Here N(*) is the set of N-grams in (*), C is the similarity measure, d_{sus} and d_{src} are the suspicious and source document respectively. Now for any pair of suspicious and source documents, if the computed similarity value is more than a specified threshold, then the source document is added to the candidate set of the particular suspicious document.

3.2 K Means Approach

In the basic K Means algorithm, the main problem is to decide the value of K and the initial centroids, which is important. This is well tackled by K Vani and Gupta D [18], where the value of K is fixed as the number of suspicious documents, and the initial centroids are considered as the individual suspicious documents itself. Here each centroid is fixed as a suspicious document, as the clusters/candidate set with respect to each suspicious document has to be formed. Initially Stop Words, which are the semantically irrelevant words, via, pre-position, conjunction etc are removed from the documents. Then the documents, d_{src} and d_{sus} is represented using VSM model with TF-IDF representation using Equation (2) to form vectors V(d_{sus}) and V(d_{src}).

$$wt, d = tft, d * \log \frac{|D|}{|\{d' \in D; t \in d'\}|} \tag{2}$$

Here $tf_{t,\,d}$ is term frequency of term t in document d and $\log \frac{|D|}{|\{d' \in D; t \in d'\}|}$ is inverse document frequency where numerator defines the total number of documents in the document set and denominator is the number of documents containing the term t. Then the similarity between document vectors is computed using cosine measure using Equation (3).

$$Cos(d_{sus}, d_{src}) = \frac{V(d_{sus}) \cdot V(d_{src})}{||V(d_{sus})|| \, ||V(d_{src})||} \tag{3}$$

In Equation (2), Cos (*) is the cosine similarity and the dot product of document vectors are represented in the numerator and the denominator computes the Euclidean norms of these document vectors

4 Proposed Approach

The proposed Fuzzy C Means algorithm is explained in Algorithm 1 and its variations with NLP techniques such as lemmatization, chunking and stemming are discussed in Sub-Section 4.1. Here initially, fuzzy membership is calculated for each d_{src} with all d_{sus} using Equation (4). In Equation (4), $\mu ij$is the degree of membership of xiin the cluster j, xi is the i^{th} data point and $cj$is the centroid.

Here m is the fuzziness factor and C is cluster centre. $d_{src}$with a μij greater than a particular threshold is considered as candidates for the given suspicious document, d_{sus}.

$$\mu_{ij} = \frac{1}{\Sigma_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{4}$$

4.1 Proposed Algorithm

Algorithm 1. Algorithm based on Fuzzy C Means

Input:

D_{sus} = Set of suspicious documents

D_{src} = Set of source documents

Output:

Candidate sets (*Cand*)

Begin

Convert all $D_{sus}$and D_{src} into TF-IDF form using Equation(2)

Set $K = \#\ D_{sus}$

Initialize each centroid c_j as each *xj* in D_{sus}

Repeat for each x_j in D_{sus}

Compute μ_{ij} for each $x_i$in D_{src} using Equation (4)

Normalize μ_{ij} between 0 and 1

If $\mu_{ij} > \alpha$, then Add x_i to the $Cand(x_j)$

Repeat for each x_i in $Cand(x_j)$: Compute

$dis = 1 - Cos(x_i, x_j)$

If $dis > \theta$, then Remove *xi* from $Cand(x_j)$

End

Here, α is the threshold for selecting the initial candidate documents with respect to a particular suspicious document. Further a cut off θ is applied for each candidate set formed, which helps in pruning out the highly dissimilar documents from the candidate sets.

4.2 Fuzzy C Means Algorithm with Different NLP Techniques

This Section describes the NLP techniques which are incorporated with the proposed FCM algorithm. The NLP techniques, via, Stemming **(FCM-stem)**, lemmatization **(FCM-lem)** and Chunking **(FCM-chk)** are used to explore its impact on the proposed algorithm. Stemming refers to a heuristic process that removes the ends of words that often includes the derivational affixes. Lemmatization is the process of converting the word tokens into their dictionary base forms. In chunking, the document is segmented into sub constituents, such as noun phrases, verb phrases, prepositional phrases etc. With chunking, stop words are not removed. Two other combinations of the above NLP techniques are also used, via, **Chunking with Lemma** (FCM-chklem) and **Chunking with Stem** (FCM-chk stem).The examples demonstrating the NLP techniques used are given in Fig.1.

English Sent: "This Document describes strategies carried out by companies for their agricultural chemicals."
Tokenization :['This', 'Document', 'describes', 'strategies', 'carried', 'out', 'by', 'U.S.', 'companies', 'for', 'their', 'agricultural', 'chemicals', '.']
After Stop Word Removal: ['Document', 'describes', 'strategies', 'carried', 'U.S.', 'companies', 'their', 'agricultural', 'chemicals']
Stemming: ['docu', 'describ', 'strateg', 'carr', 'compan', 'agricultur', 'chemic']
Lemmatization: ['Document', 'describes', 'strategy', 'carried', 'company', 'agricultural', 'chemical']
Chunking:['This Document', 'strategies', 'describes', 'carried', 'companies', 'their agricultural chemicals']

Fig. 1 Example of NLP techniques used with proposed algorithm

5 Experimental Settings and Result Evaluation

The existing approaches and the proposed algorithm discussed in Section 3 and 4 respectively are evaluated on PAN-13 partial data set. Then the algorithm efficiency is measured using the standard IR measures, via, Recall and Precision.

5.1 Data Set

The data set in PAN 2013 is divided based on their complexity and the statistics used is given in Table.1

Table 1 Data Statistics

	Suspicious Documents	**Source Documents**
Set 1	39	205
Set 2	31	213
Set 3	35	209

Here, No Obfuscation set (Set 1) consists of document pairs where the suspicious document contains exact copies of passages in the source document. In random Obfuscation (Set2) the document is manipulated via word shuffling, adding, deleting and replacing words or short phrases, synonym replacements etc. In translation obfuscation (Set 3), the given text is run through a series of translations, so that the output of one translation forms the input of next one while the last language in the sequence is the original language of the text [19].

5.2 Evaluation Metrics

The performance of each algorithm is measured using the metrics, via, Recall and precision as given in Equation (4) and (5) and comparison is done.

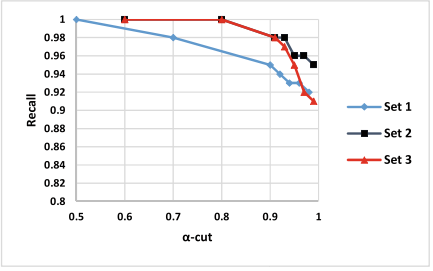
$$prec = \frac{|D_{ret} \cap D_{exp}|}{|D_{ret}|} \tag{5}$$

$$rec = \frac{|D_{ret} \cap D_{exp}|}{|D_{exp}|} \quad (6)$$

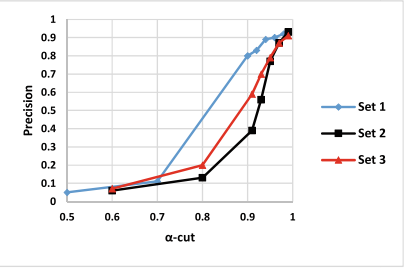
Here D_{ret}denote the set of documents that are retrieved by the system and D_{exp} is the actual expected documents. Precision measures the ratio of correctly retrieved cases to the total documents retrieved by the system while recall measures the ratio of correctly retrieved cases to the actual expected documents.

5.3 Results Analysis

The proposed and existing candidate retrieval algorithms are evaluated on the data sets mentioned in Table 1 and performance is evaluated. The proposed algorithm results are then compared with the existing K Means and N-grams methods and performance is measured using Equation (4) and (5). The proposed FCM algorithm efficiency is determined mainly by two parameters, via, α and θ as discussed in Sub-Section 4.1. The algorithm is tuned with different values of these two parameters to obtain the optimal recall and precision. Initially, the proposed FCM algorithm is evaluated with various α values on the three data sets discussed in SubSection 5.1. The recall and precision based on different αvalues is plotted in Fig.2 (a) and (b) respectively. Due to the variation in complexity of sets, the performance exhibited by different α values on these sets also differs. After conducting many evaluations, the α value for each set is determined based on the precision and recall presented using the specific α value. From Fig.2 (a) it can be observed that with Set-1, a 100 % recall is obtained with α= 0.5 and for Set-2 and Set-3 from 0.6 to 0.9 the recall is high. Now considering both precision and recall, the α value selected for each set is given in Table. 2. In a similar way, the

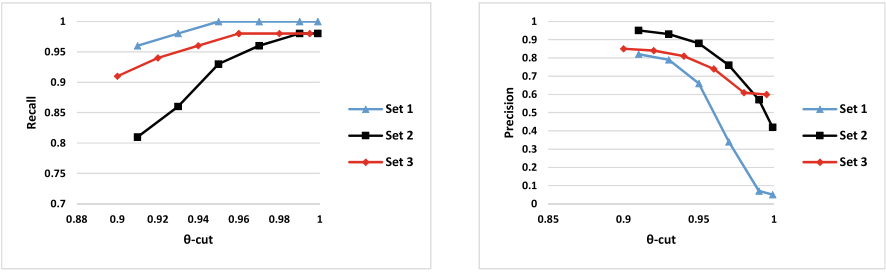


(a)



(b)

Fig. 2 (a) Recall results with different α values of FCM-basic; (b) Precision results with different α values of FCM-basic



(a) (b)

Fig. 3 (a).Recall results with different θ values of FCM-basic; (b) Precision results with different θ values of FCM-basic

Table 2 θ and α values for each set

	α	θ
Set 1	0.5	0.95
Set 2	0.91	0.98
Set 3	0.91	0.98

algorithm is evaluated with multiple θ values and the performance of the system is plotted in Fig.3 (a) and (b). The θ value that presents maximum recall and a high precision is then considered for each Set. The θ value finally selected for each set is given in Table.2.

Fig.4 and 5 plots the proposed FCM algorithm and its variations with different NLP techniques using the α and θ values as given in Table.2. Further these algorithms are compared with the existing candidate retrieval approaches, via, N-gram and K Means approach. In N-gram evaluation, N = 3 is employed with a threshold of 0.05 and the basic K Means candidate retrieval algorithm is also evaluated with the data sets. It can be noted that for Set-1 with no obfuscations, the basic FCM algorithm exhibits high recall but precision reduces. Proposed algorithm incorporated with lemmatization and stemming exhibits good performance with respect to recall. With respect to precision, FCM-chk and FCM-chklem outperforms the other NLP variations incorporated with proposed FCM. Compared to the existing approaches, it can be noted that the proposed algorithm and its variations presents a considerable improvement in recall which is the main focus of candidate stage. With Set-2 the performance variation is almost the same as in Set-1, while in Set-3 with respect to both measures FCM-chk outperforms the other FCM based variations.

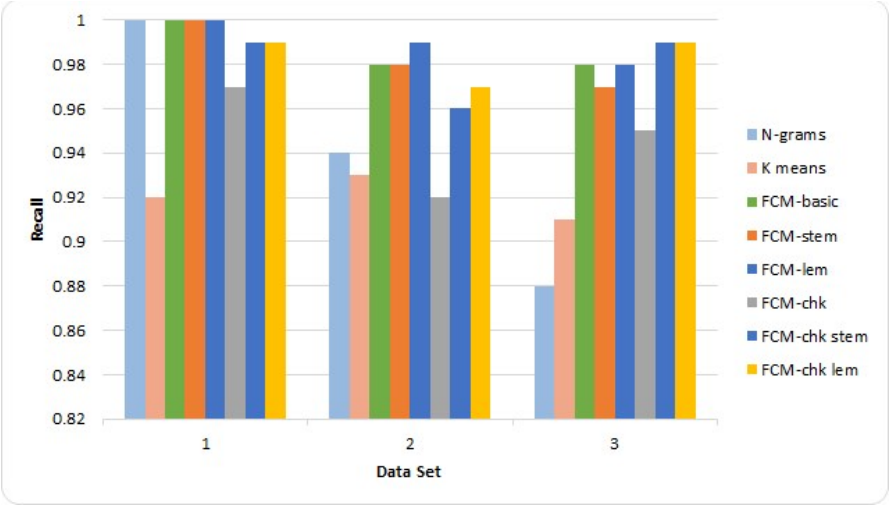


Fig. 4 Fuzzy C Means Recall values with various NLP techniques

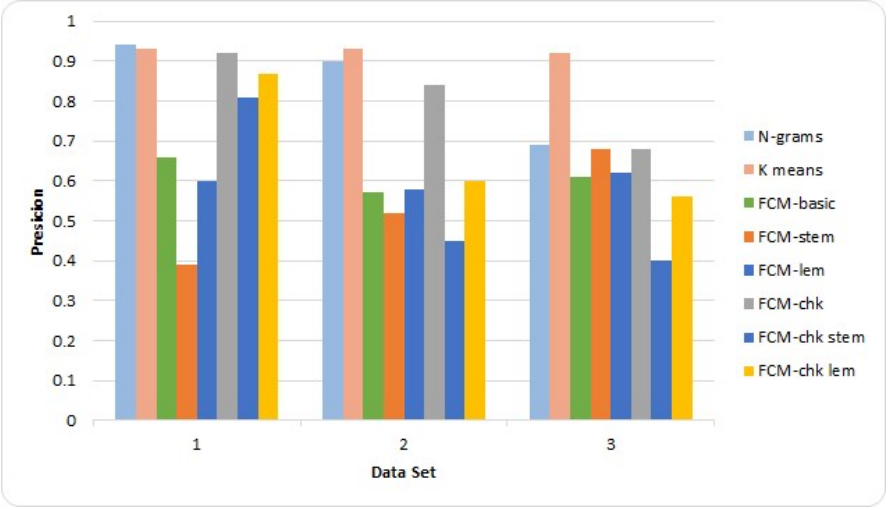


Fig. 5 Fuzzy C Means Precision values with various NLP techniques

The proposed method presents an improvement of 5% in Set-2 and 11% in Set-3 with respect to the traditional N-gram approach and compared to K Means approach an improvement of 8% in Set-1, 6% in Set-2 and 8% in Set-3. Thus compared to N-gram and K Means candidate retrieval approaches, the proposed FCM algorithm and its variations exhibit an increased performance with respect to recall, which is the main focus in candidate retrieval task.

6 Conclusions and Future Work

The proposed work focuses on exploration of clustering techniques in candidate retrieval task of extrinsic PDS. The soft clustering approach with Fuzzy C Means algorithm is employed here and FCM algorithm is tuned for the candidate retrieval task. The proposed approach exhibits a recall value of 100% in Set 1 and around 98% in Sets 2 and 3 which is considerably higher than the existing methods. Within FCM based approaches, in Set 2, FCM-lem presents about 99% recall and with Set-3 FCM-chk stem and FCM-chklem provides a recall of about 98%. K Means exhibits a good precision but with lower recall values compared to other approaches because it utilize 'hard' clustering approach. This is where Fuzzy C Means approach gains ground because of its 'soft' clustering technique. With the main focus in candidate retrieval being higher recall, it can be concluded from the analysis and discussions that the proposed algorithm and its variations exhibits a considerable improvement in recall compared to existing approaches discussed. In future, more efficient filtering techniques can be employed with the focus on improvement of precision values. For efficient comparison of performance, the algorithms will be evaluated on larger data sets.

References

1. 'A Plagiarism FAQ' IEEE. http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html
2. Alzahrani, S.M., Salim, N., Abraham, A.: Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews 42(2) (2012)
3. Ceska, Z.: The future of copy detection techniques. In: Proceedings of the First Young Researchers Conference on Applied Sciences, pp. 5–10 (2007)
4. Freedman, J.: The Ombudsman as Go-Between. In: The Fourth Annual Report of the Office of the Ombudsman (1974–1975)
5. Osman,A H., Salim, N., Bin Wahlan, S., Hentabli, H., Ali, A.M.: Conceptual similarity and graph-based method for plagiarism detection. Journal of Theoretical and Applied Information Technology **32**(2), 135–145 (2011)
6. Alzahrani, S., Salim, N.: Fuzzy Semantic-based String Similarity for Extrinsic Plagiarism Detection. CLEF (Notebook Papers/LABs/Workshops) (2010)
7. Ekbal, A., Saha, S., Choudhary, G.: Plagiarism Detection in Text using Vector Space Model. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews 42(2) (2012)
8. Naseem, R., Kurian, S.: Extrinsic Plagiarism Detection in Text Combining Vector Space Model and Fuzzy Semantic Similarity Scheme. International Journal of Advanced Computing, Engineering and Application (IJACEA) 2(6) (2013) ISSN: 2319–281X
9. Osmana, A.H., Salima, N., Binwahlan, M.S., Alteeb, R., Abuobiedaa, A.: An improved plagiarism detection scheme based on semantic role labeling. In: Applied Soft Computing 12 (2012)

10. Barrón-Cedeño, A., Rosso, P.: On Automatic Plagiarism Detection Based on n-Grams Comparison. In: ECIR Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval, pp. 696–700 (2009)
11. Shang, C., Li, M., Feng, S., Jiang, Q., Fan, J.: Feature selection via maximizing global information gain for text classification. KnowledgeBased Systems **54**, 298–309 (2013)
12. Bijalwan, V., Kumar, V., Kumari, P., Pascual, J.: KNN based Machine Learning Approach for Text and Document Mining. International Journal of Database Theory and Application **7**(1), 61–70 (2014)
13. Uysal, A.K., Gunal, S.: Text classification using genetic algorithm oriented latent semantic features. Expert Syst. Appl. **41**, 5938–5947 (2014)
14. Buana, P.W., Jannet, S., Putra, I.: Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News. International Journal of Computer Applications 50(11) (2012)
15. Šilić, A., Moens, M.-F., Žmak, L., Bašić, B.D.: Comparing Document Classification Schemes Using K-Means Clustering. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part I. LNCS (LNAI), vol. 5177, pp. 615–624. Springer, Heidelberg (2008)
16. Pappuswamy, U., Bhembe, D., Jordan, P.W., VanLehn, K.: A Supervised Clustering Method for Text Classification. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 704–714. Springer, Heidelberg (2005)
17. Miao, Y., Kešelj, V., Milios, E.: Document Clustering using Character N-gram: A Comparitive Evaluation with Term-based and Word-based Clustering. In: ACM International Conference on Information and Knowledge Management, pp. 357–358 (2005)
18. Vani, K., Gupta, D.: Using K-means Cluster based Techniques in External Plagiarism Detection. In: Proceedings of International Conference on Contemporary Computing and Informatics (IC3I), pp. 27–29 (2014)
19. Potthast, M., Hagen, M., Gollub, T., Tippmann, M.: Overview of 5th International Competition on Plagiarism Detection. In: CLEF 2013 Evaluation Labs and Workshop–Working Notes Papers, pp. 23–26 (2014) , ISBN 978-88-904810-3-1, ISSN 2038-4963