Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

Textual Document Restructuring using Frequent Terms and Named Entity

Wael M.S. Yafooz
Faculty of Computer and Information Technology,
Al-Madinah International University (MEDIU), 40100
Shah Alam, Selangor, Malaysia

***Abstract*— Data is mostly stored in digital format rather than hard copy because the former is safer, more secure, smaller in size, and faster to retrieve than the latter. With the increasing number of electronic documents to be organized for users to obtain knowledge and integrate information, document clustering has been applied by grouping textual documents based on their similarities. Many attempts have been made to perform textual document clustering with highly accurate results (i.e., close to nature classes) and high processing performance. However, such proposed techniques work in batch (or static) mode in which performance tend to be sacrificed with the use of all the terms in the document, at times resulting in overlapping or scalability issues. Few studies that focus on dynamic clustering also reported on performance issues. This paper contributes in the investigation of textual document clustering approaches and highlights the importance of using dynamic clustering in mining frequent terms with included named entity. This method is used to achieve high efficiency and high-quality data clustering. The method is also beneficial to be used in textual document clustering algorithms for many text domain applications.**

Keywords- Document Clustering, Frequent Term, Named Entity, Dynamic Textual Clustering

I. INTRODUCTION

Nowadays, digital work has increased at a high rate due to daily activities that depend on it. Digital work is stored in digital documents for easy and quick data retrieval. Digital documents are also more secure than hard copies [1]. With the massive amount of work covered by textual documents such as news articles and conference papers, they have become a basic source of knowledge [2]. However, users frequently encounter difficulties in obtaining knowledge from textual documents because of the massive amount of data they contain. Thus, textual document clustering (or text clustering) technique was introduced in the area of text mining. Textual document clustering, also known as text clustering, is the process of clustering or grouping of textual documents according to their content. The two important main goals in document clustering are achieving high performance or efficiency and obtaining highly accurate data clusters that are closed to their natural classes or textual document cluster quality. Traditional methods of textual document clustering can be categorized into partitional or hierarchical document clustering approaches [3-9]. Both approaches do not entirely achieve efficiency and data cluster quality. In the partitional approach, the clustering process is fast and is thus efficient. The accuracy of the data cluster, however, is not very good, especially when the textual document is large. Meanwhile, hierarchical document clustering produces much better data clusters than those of partitional methods. However, the approach is time consuming, rendering its performance as not much better than that of the partitional method.

Hierarchical document clustering outperforms partitional document clustering in the representation of textual document in multiple-level topics. Partitional approach represents documents in only one level topic. Generally, both methods have many issues that include the predefined number of clusters that the user must identify, requiring the user to have prior knowledge regarding the data [10]. The algorithms of both methods are also not scalable. Data become highly dimensional when the clustering involves a large collection of documents [11]. Finally, overlapping can occur when textual data belong to more than one textual data cluster.

Given these issues, many attempts had been made to come up with soultion by reducing terms [11-14] and by producing quality clusters [15-17]. Majority of the proposed methods are static or known as batch mode, in which textual documents are gathered prior to clustering. Studies that focus on dynamic and incremental methods are limited, especially those that aim to produce "good" quality data clusters. This paper extend of our work in [18, 19] investigates textual document clustering methods and highlights the importance and differences between the dynamic and static textual clustering. In addition, this paper shows the significance of using both frequent terms and named entities as a combined method for textual document clustering. Furthermore, this paper emphasizes on using the semantic (or synonym) of frequent terms can produce good quality data clusters.

The rest of this paper is organized as follows. In Section 2, related studies are discussed. Section 3 compares static and dynamic textual clustering algorithms. Section 4 presents similarity measures, frequent terms, and named entities. The results and discussion are presnted in section 5. Finally, the conclusion is presented in Section 6.

II. RELATED STUDIES

This section presents textual document clustering algorithms, which can be categorized into two main textual document approaches, namely, traditional and modern .

A. Traditional Textual Document Clustering approaches

In the traditional textual clustering approach, clustering algorithms are categorized into partitional and hierarchical.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

Partitional clustering algorithms are centroid-based clustering based on a center point and on a group surrounding data objects (data points). The most popular algorithm in this category is the k-mean algorithm [8]. The main idea of the k-mean algorithm is that a group of data objects is based on the distance between the center and the data objects (or data points). Many variants of the k-mean algorithm are introduced to overcome its weaknesses that affect the quality of data clusters.

In Hierarchical Clustering Algorithm (HCA), clusters are connected in the form of a hierarchy tree that is represented in nodes. The father node is the main node and the child nodes are the branches. Thus, the hierarchy tree provides meaningful clusters for textual documents. HCA has two types, namely, Agglomerative Hierarchical Clustering (AHC) and Divisive Hierarchical Clustering algorithms. AHC is the best traditional approach because it provides users with sufficient knowledge about the content of the textual document. An experiment by Steinbach et al. [3] shows that the Unweighted Pair Group Method with an Arithmetic Mean used in AHC is the distance measurement in producing quality clusters. AHC uses the top-down strategy, whereas Divisive Hierarchical Clustering algorithms use the bottom-up strategy. Generally, partitional approaches outperform hierarchical approaches in terms of efficiency, while hierarchical approaches outperform partitional approaches in terms of textual cluster quality [20]. Thus, the bisecting k-mean approach is a mix of the two produces results [3] that bisecting k-mean approach outperforms both hierarchical and partitional approaches.

The most popular textual-document clustering algorithm is the Scatter/Gather algorithm [7]. Scatter/Gather is used for searching or browsing system documents through text-based methods. The proposed system assists the ordinary user in obtaining relevant information by providing specific information. Then the system performs iterative clustering which is based on the two main processes, namely, the Buckshot and the Fractions systems. Normally, all traditional approaches suffer from many issues such as the predefined number of clusters, non-scalability, high dimensionality, and overlapping [10, 12, 21].

B. Modern Textual Document Clustering Approaches.

Traditional approaches suffer from high-dimensional data. Thus, modern textual document clustering approaches had been introduced and are categorized into three categories, namely, frequent-term [11, 12, 22-25], semantic-based [16, 17, 26] , and entity-based [13, 15, 27, 28] approaches. The Frequent Term-Based Clustering (FTC) approach was introduced by Beil et al. [11]. FTC begins by generating a set of frequent terms from the text document (or database) using the Apriori algorithm [29] then extended to Hierarchical Frequent Term-Based Clustering (HFTC) to represent data clusters in hierarchical view. However, HFTC is not scalable and is unsuccessful in the clustering process if the collection of documents is large. Thus, Fung et al. [12] introduced Frequent Itemset-based Hierarchical Clustering (FIHC) to overcome drawbacks of HFTC. FIHC reduces the number of frequent items by selecting frequent items that are greater than the minimum fraction of the document.

Chen et al. [23, 24] introduces a Fuzzy Frequent Item-set-Based Hierarchical Clustering (F^2IHC) in order to improve the clustering quality of FIHC. By using fuzzy association rules mining, it is easy to realize a relation and integrate linguistic terms [24]. F^2IHC consists of three stages, which are *Document Pre-processing*, *Candidate Clusters Extraction* and *Cluster Tree Construction*, as shown in Figure 1. Many attempts had been made based on frequent terms, such as by selecting only the Maximal Frequent Set [14, 30], the Sequence Frequent Term Set [1], and the Maximal Capturing Frequent Set [25].

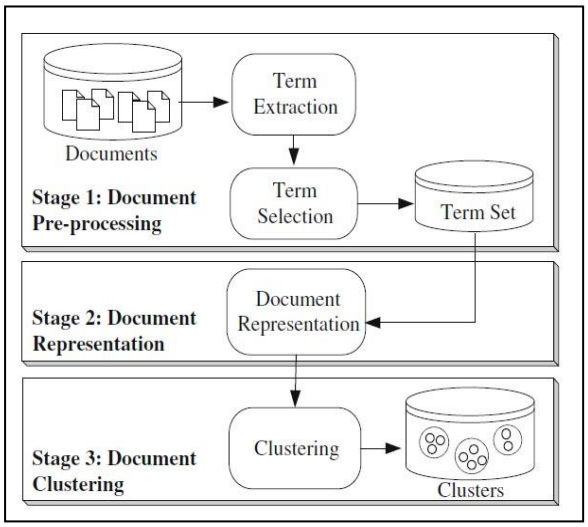

Figure 1: General View of clustering process [16] .

So far, all o the mentioned approaches treat textual data similar to a bag of words that are subsequently weighed according to Term Frequency. However, Shehata et al. [26] noted that frequency is insufficient for textual document clustering. In addition, such methods do not consider the semantic relationship between textual documents. Therefore, Shehata et al. [26] [17] propose Conceptual Term Frequency as a concept-similarity measure, focusing on the semantic structure of a sentence verb argument rather than based on word weight. However, a term (verb argument) can be significantly relevant to a sentence but more related to a document. Thus, in 2009, Concept-Based studies were enhanced using WordNet [31] to produce highly accurate textual document clusters. While focusing on the same concepts as [34], [36] developed the Frequent Concepts-based Document Clustering (FCDC) algorithm. FCDC can cluster a document based on concepts and a set of semantically related words using the WordNet database. In contrast, Huang et al. [32] introduced the Bag of Concepts to discover the conceptual relationship between two documents based on mapping using Wikipedia [33].

In the named entity-based approach, Montalvo et al. [15] noted that the content of news documents should answer six questions: Who, Where, What, Why, How, and When. Thus, answers to these questions should contain a name entity. A

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

method for multilingual document clustering was developed by Montalvo et al. [27] using cognate-named entities for bilingual clustering (English-Spanish) of online news articles with the application of a fuzzy logic set to determine the percentage of similarity between named entities and textual documents. Similar to the approach of [15], [28] proposed a document-clustering algorithm based on fuzzy rules. The proposed algorithm determined a semantic relation between entities to solve the disambiguation problem of similar entities by identifying features of named entities, which consist of three parts, namely, name, type, and identifier. The algorithm focused more on the geographical information of the named entities to show that the mentioned name belongs to a city, mountain, river, and so on.

Table 1:A comparison between the textual document clustering approaches

Approach	Traditional		Modern		
Method	Partitional	Hierarchical	Frequent Term	Semantic	Named Entity
Performance	Poor	Good	Very Good	Very Good	Very Good
accuracy	Good	Poor	Very Good	Very Good	Very Good
Terms	ALL	ALL	Frequent	Frequent	Named Entity
semantic	No	No	No	Yes	No
Literature	[3, 7]		[1, 4, 11, 15, 17, 25]		

Table 1 demonstrates the comparison between the textual document clustering approaches. The traditional approaches use all words in textual documents that affect the efficiency and quality of clusters. The efficiency and quality of clusters in modern approaches outperform the traditional approaches due the high dimensional of data reduced and semantic of words introduced.

III. STATIC AND DYNAMIC TEXTUAL DOCUMENTS CLUSTERING

Static textual documents clustering performs taskes in a batch mode. Batch mode is a method of gathering all documents for later clustering. This clustering method is time consuming and affects the performance of a system [34]. Suchh batch mode is suitable for documents that do not increase in number or those that only have a one-time process. however some textual documents continue to increase in number due to daily life activities, such as online news articles. We therefore dynamic clustering for such data is required.

Dynamic clustering for traditional approaches, such as partitional or hierarchical textual clustering, affect the performance of a system because of the following three main issues: predefined number of clusters, high dimensionality, and data cluster quality. Regarding high-dimensionality issues, traditional approaches use all terms in the document. The clustering process is thus repeated every time and is time consuming. Data cluster quality can also affect the performance during the clustering process because of the term used in textual documents. Moreover, the predefined number of clusters is another challenge because the user needs to have sufficient prior knowledge about the data to determine the number of clusters.

The modern approach of textual document clustering is much better than any of the traditional approaches because modern approaches do not require a user to enter the number of expected data clusters as an input parameter. In addition, modern approaches focus on reducing the number of terms or words that represent the textual document, which also reduces dimensionality of the data. Unfortunately, most modern approaches lack of dynamic clustering for frequent terms. This paper, shows the importance of using frequent terms and named entities as a clustering similarity measure. In addition, the semantic of terms are investigated along with the frequent terms that occur in textual documents.

IV. FREQUENT TERM AND NAMED ENTITY

In order to produce high-quality document clusters, a process of clustering or grouping of textual documents based on rules is formulated. A similarity measure rule is used to discover the relationship between two textual documents. In traditional approaches, this rule is known as the distance measure, which is the distance between data points. Many distance measures had been proposed, such as cosine similarity, Edlidean distance, Manhattan distance, and Maximum distance [35].

In modern approaches, the distance measure is called a similarity measure. In order to introduce a good similarity measure, we should know the structure of the textual document. The structure of textual documents is represented by several paragraphs. A paragraph consists of more than one sentence. A sentence consists of several words. Words can be repeated in textual documents. Repeated words are called frequent terms. Modern approaches focus on frequent term similarity measures [11, 12], maximal terms [14, 30], sequence frequent terms [1], Closed Interesting Itemsets [36], semantic of terms [16, 17], or named entities [15, 28]. All of these approaches do not consider the integration of the named entity and frequent terms. Overlapping is frequently word occurrence between data clusters. In addition, if the collection of textual document is large, the number of frequent terms will increase and the problem of high dimensionality will reappear. Moreover, named entities are usually found in the textual document, especially in news articles.

This research, we introduce a new similarity measure based on maximal frequency of frequent terms and on all named entities found in the textual document. A named entity is the name of a person, location, or organization. Thus, the use of frequent words or their maximal frequency according to the minimal support words and named entity is called a Clustering Frequent Set (CFS). CFS is useful in the textual documents clustering process. Obtaining the semantic of frequent terms from the WordNet database [31] will also improve the quality of data clusters.

V. RESULTS AND DISCUSSION

The dataset is a sample of the textual documents used to test the proposed approached, this dataset is the benchmark. The dataset of a classic dataset. Classic datasets contain abstracts

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

of papers and are classified into four classes, namely, CACM, CRAN, CISI, and MED. This dataset contain a general information about academic papers . Table 2 show

TABLE 2
Summary Of Dataset textual data structuring

Dataset	No of doc	No of classes	Max class size	Min class size	Avg. class size
Classic	7095	4	3203	1033	1774

Textual data structuring, linking and organizing processes link structured and unstructured textual data in textual databases. A user can find relationships between unstructured textual data by using such linking processes. The process of linking is achieved after structuring and organizing the unstructured textual data. The processes of converting textual files from unstructured to structured form by extracting important terms from the files. These terms are Named-Entity and Frequent-Term. For Named-Entity, all possible terms that represent the Named-Entity is considered important elements in extracting textual data. As for Frequent-Term which frequently appears in textual documents.

The experiments conducted to evaluate the textual document structuring and linking inside the textual database. Thus, the manner of structuring such data with the proposed approched evaluated by comparing the structuring approach of the information extraction technique. This evaluation performed based linking. linking is used to determine the relation between textual documents. In this experiment , the information extraction methods are used to compared its results of structuring textual unstructured data with TVSM model, due to it most common way of managing unstructured textual data.

TABLE V
Comparison of F-measure between IE methods and TVSM- Classic Dataset

	IE Minimum Support				TVSM Minimum Support			
	1	2	3	4	1	2	3	4
CACM	0.25	0.17	0.1	0.002	0.28	0.47	0.27	0.21
CISI	0.43	0.21	0.12	0.067	0.24	0.31	0.22	0.58
CRAN	0.037	0.15	0.005	0.062	0.58	0.18	0.37	0.37
MED	0.34	0.33	0.003 4	0.037	0.27	0.5	0.21	0.11

FIGURE 2 : Comparison of the IE -classic.

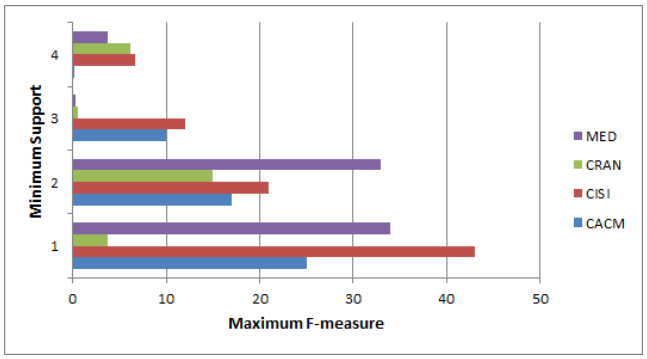

FIGURE 3 : Comparison of the proposed approch -classic.

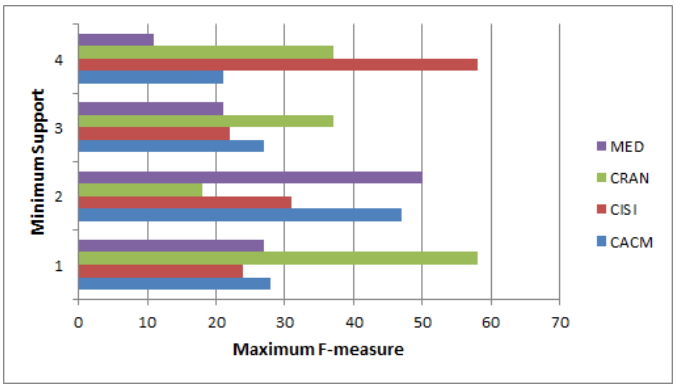

Figure 2 and Figure 3 shows the maximum F-measure of data clusters created by using IE and proposed approch , respectively. The best maximum F-measure can be realized when a minimum support is two and three words for both methods. The F-measure for a data cluster created using proposed approch is better than that created using IE

VI. CONCLUSION

In this paper, we presented the significance, importance, and difference between dynamic textual document clustering and batch (or static) mode. We also presented a clustering process based on the semantic of frequent terms and named entities. The proposed clustering process will improve the quality of data clusters that will provide users with knowledge about the content of a document as well as produce a close textual data cluster to natural classes. Additionally, dynamic textual clustering improve the performance of a system and of the clustering process. By contrast, traditional approaches encountered many issues during clustering such as high-dimensional data, a predefined number of clusters, non-scalability, overlapping, and poor quality clusters. These issues are partially resolved in modern approaches. However, quality and efficiency remain critical issues. Moreover, majority of studies focused on batch mode clustering, while dynamic textual document clustering was found significant for system efficiency.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 01, January-2017

ACKNOWLEDGMENT

This paper based on work support by Universiti Teknologi MARA (UiTM), Malaysia . The author would like to thanks UiTM.

REFERENCES

[1] Y. Li, et al., "Text document clustering based on frequent word meaning sequences," Data & Knowledge Engineering, vol. 64.1, pp. 381-404, 2008.
[2] W. M. S. Yafooz, et al., "Challenges and issues on online news management," Control System, Computing and Engineering (ICCSCE),IEEE International Conference on., 2011.
[3] M. Steinbach, et al., "A Comparison of Document Clustering Techniques," KDD workshop on text mining, vol. Vol. 400, 2000.
[4] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.
[5] Y. Zhao and G. Karypis, "Hierarchical Clustering Algorithms for Document Datasets," Data Mining and Knowledge Discovery, vol. 10, 2005.
[6] C. A. Bjorner Larsen "Fast and Effective Text Mining Using Linear-time Document Clustering," Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining., 1999.
[7] D. R. Cutting, et al., "Scatter/gather: A cluster-based approach to browsing large document collections," 15th annual international ACM SIGIR conference on Research and development in information retrieval., 1992.
[8] A. K. Jain, et al., "Data Clustering: A Review," ACM computing surveys (CSUR), vol. 31.3, pp. 264-323, 1999.
[9] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31.8, pp. 651-666, 2010.
[10] A. Sharma and R. Dhir, "A wordsets based document clustering algorithm for large datasets," Methods and Models in Computer Science, vol. ICM2CS . Proceeding of International Conference on. IEEE, 2009.
[11] F. Beil, et al., "Frequent Term-Based Text Clustering," Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. , 2002.
[12] B. C. M. Fung, et al., "Hierarchical Document Clustering Using Frequent Itemsets," Proceedings of the SIAM international conference on data mining, vol. 30. No. 5, 2003.
[13] T. H. Cao, et al., "Fuzzy named entity-based document clustering.," IEEE World Congress on Computational Intelligence. , 2008.
[14] E. Hernandez-Reyes, et al., "Document Clustering Based on Maximal Frequent Sequences," Advances in Natural Language Processing, vol. Springer Berlin Heidelberg, pp. 257-267, 2006.
[15] S. Montalvo, et al., "NESM: a Named Entity based Proximity Measure for Multilingual News Clustering," Procesamiento de Lenguaje Natural, vol. 48, pp. 81-88, 2012.
[16] C.-L. Chen, et al., "An integration of fuzzy association rules and WordNet for document clustering," Knowledge and information systems, vol. 28.3, pp. 687-708, 2011.
[17] S. Shehata, et al., "Efficient Concept-Based Mining Model for Enhancing Text Clustering," Knowledge and Data Engineering, IEEE Transactions on, 2010.
[18] W. M. S. Yafooz, et al., "Dynamic semantic textual document clustering using frequent terms and named entity," in System Engineering and Technology (ICSET), 2013 IEEE 3rd International Conference on, 2013, pp. 336-340.
[19] W. M. S. Yafooz, et al., "Future trends in managing extracted information," in Control System, Computing and Engineering (ICCSCE), 2013 IEEE International Conference on, 2013, pp. 279-283.
[20] D. B. Deshmukh and Y. Pandey, "A Review On Hierarchical Document Clustering," Journal of Data Mining and Knowledge Discovery, pp. 2229-6662, 2012.
[21] X. Liu and P. He, "A Study on Text Clustering Algorithms Based on Frequent Term Sets," Lecture Notes in Computer Science, vol. 3584, pp. 347-354, 2005.
[22] G. V. R. Kiran, et al., "frequent Itemset based Hierarchical Document Clustering using Wikipedia as External Knowledge," Knowledge-Based and Intelligent Information and Engineering Systems, vol. Springer Berlin Heidelberg, pp. 11-20, 2010.
[23] C.-L. Chen, et al., "Hierarchical Document Clustering Using Fuzzy Association Rule Mining " Innovative Computing Information and Control. 3rd International Conference on. IEEE, 2008.
[24] C.-L. Chen, et al., "Mining fuzzy frequent itemsets for hierarchical document clustering," Information processing & management vol. 46.2, pp. 193-211, 2010.
[25] W. Zhang, et al., "Text clustering using frequent itemsets," Knowledge-Based Systems, vol. 23.5, pp. 379-388, 2010.
[26] S. Shehata, et al., "Enhancing Text Clustering using Concept-based Mining Model," Data Mining. ICDM'06. Sixth International Conference on. IEEE, 2006.
[27] S. Montalvo, et al., "Bilingual News Clustering Using Named Entities and Fuzzy Similarity," Text, Speech and Dialogue. Springer Berlin Heidelberg, 2007.
[28] T. H. Cao, et al., "Data Mining: Foundations and Intelligent Paradigms.," Springer Berlin Heidelberg, pp. 267-287, 2012.
[29] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceedings of the 20th VLDB Conference Santiago, Chile, vol. Vol. 1215, 1994.
[30] C. Su, et al., "Text Clustering Approach Based on Maximal Frequent Term Sets," Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA, 2009.
[31] G. Miller, "WordNet: A Lexical Database for English," Communications of the ACM vol. 38.11, pp. 39-41, 1995.
[32] A. Huang, "Similarity Measures for Text Document Clustering," NZCSRSC 2008, Christchurch, New Zealand, 2008.
[33] D. Milne and I. H. Witten, "Learning to Link with Wikipedia," Proceedings of the 17th ACM conference on Information and knowledge management. ACM,, 2008.
[34] W. M. S. Yafooz, et al., "Towards automatic column-based data object clustering for multilingual databases," Control System, Computing and Engineering (ICCSCE), IEEE International Conference on. IEEE, 2011.
[35] A. Huang, "Similarity Measures for Text Document Clustering," NZCSRSC 2008 Christchurch, New Zealand, 2008.
[36] H. H. Malik, et al., "Hierarchical document clustering using local patterns," 2010.