Agent for Documents Clustering using Semantic-based Model and Fuzzy

Khaled M. Fouad
Computer Science Dept., College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia (KSA).

Moataz O. Hassan
Computer Engineering Dep., Faculty of Engineering Cairo University, Egypt.

ABSTRACT
Text clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing large sets of documents into a small number of meaningful clusters. Many fuzzy clustering algorithms, such as K-means, deal with documents as bag of words. The bag of words representation method used for these clustering is often unsatisfactory because it ignores the semantic of words.

The proposed agent exploits WordNet ontology to create low dimensional feature vector which allows us to develop an efficient clustering algorithm. A new semantic-based model, that represents documents based on semantic concepts of words, is proposed. The proposed approach aims at increasing the performance of information retrieval process by enhancing the document clustering. The accuracy and the speed of clustering have been examined before and after combining ontology with Vector Space Model (VSM). Experimental results demonstrate that using semantic-based model and fuzzy clustering enhances the clustering quality of sets of documents.

General Terms
Document Clustering, Fuzzy Clustering, K-Means.

Keywords
Document Clustering, Semantic Text Representation, Agent, WordNet.

1. INTRODUCTION
Nowadays, the internet has become the largest data repository, facing the problem of information overload. In the same time, more and more people use the World Wide Web as their main source of information. The existence of an abundance of information, in combination with the dynamic and heterogeneous nature of the Web, makes information retrieval a tedious process for the average user. Search engines, meta-search engines and Web Directories have been developed in order to help the users quickly and easily satisfy their required information [1]. This has led to the requirement for the development of new techniques to assist users effectively navigate, trace and organize the available web documents, with the ultimate goal of finding those best matching their needs. One of the techniques that can play an important role to achieve this objective is document clustering.

Clustering is the process of grouping/dividing a set of objects into subsets (called clusters) so that the objects are similar to one another within the cluster and are dissimilar to objects in other clusters regarding some selected features of these objects. Clustering is a method of unsupervised classification. It is a common technique of statistical data analysis used in many fields and applications such as biology, geology, medicine, market research, educational research, social network analysis, image segmentation, and data mining. The process of clustering typically involves the following steps: (1) text representation (optionally feature extraction and /or selection), (2) definition of distance/similarity measure, (3) clustering or grouping, and (4) data abstraction or labeling (optional) [2].

Text representation is the step of selecting features to represent text that will be clustered. Feature selection is a process of identifying the most effective subset of the original features to be used in clustering. Feature extraction is the process of using linear or non-linear transformations on original features to generate projected features to be used in clustering [3].

The effective clustering algorithm must have the following features: (1) the ability to detect clusters with various shapes and different distributions; (2) the capability of finding clusters with considerably different sizes; (3) the ability to work when outliers are present; (4) no or few parameters needed as input; and (5) scalability to both the size and the dimensionality of data [4].

In this paper, the proposed agent presents semantic-based model to enhance the standard k-means algorithm using WordNet ontology [5] by enriching the text representation used in the clustering process. The model leads to create low dimensional feature vector which allows developing an efficient clustering algorithm. The major challenge is to use the background knowledge in the similarity measure to acquire the concepts for each term in the feature vector from WordNet ontology. Through experiments, the performance of the proposed system is analyzed in terms of accuracy and speed of clustering.

2. RELATED WORK
Drakshayani and Prasad [6] proposed a new model for text document representation. The proposed model follows parsing, preprocessing and assignment of semantic weights to document phrases to reflect the semantic similarity between phrases and k-means clustering algorithm. They evaluated the proposed model using 5 different datasets in terms of F-Measure, Entropy, and Purity for K-Means clustering algorithm. The results demonstrated a performance improvement compared to the traditional vector space model and latent semantic indexing model. More NLP techniques may be included to enhance the performance of the text document clustering.

Luo, Li and Chung [7] have proposed three different methods of using the neighbors and link in the k-means and bisecting k-means algorithms for document clustering. Comparing with the local information given by the cosine function, the link function provides the global view in evaluating the closeness between two documents by using the neighbor documents.

They enhanced the k-means and bisecting k-means algorithms by using the ranks of documents for the selection of initial centroids, by using the linear combination of the cosine and link functions as a new similarity measure between a document and a centroid, and by selecting a cluster to split based on the neighbors of the centroids. All these algorithms are compared with the original k-means and bisecting k-means on real-life data sets. Their experimental results showed that the clustering accuracy of k-means and bisecting k-means is improved by adopting the new methods individually and also in combinations.

Shah and Mahajan [8] presented the survey of various clustering techniques based on semantics. All these techniques are described in brief. They presented the comparison of these techniques in tabular format with various parameters like the semantic approach applied by author, datasets used, evaluation parameters applied, limitations and future work; which would be very easy for quick interpretation. This survey is very useful for researchers in this area.

In [9, 12] authors proposed a semantic text document clustering approach based on the WordNet lexical categories and Self Organizing Map (SOM) neural network. The proposed approach generates documents vectors using the lexical category mapping of WordNet after preprocessing the input documents. They applied three different clustering algorithms, SOM neural network, k-means, and bisecting k-means to the generated documents vectors. The output clusters in each case are evaluated using silhouette coefficient measure to test the performance of the proposed approach. The results showed that SOM neural network achieves higher clustering quality than other two clustering algorithms k-means, and bisecting k-means. Also, the results showed that by using WordNet lexical categories in the feature extraction process for text documents improves the overall clustering quality.

Thangamani and Thangaraj [10] introduced the method of building ontologies into unsupervised text learning in order to consider the text semantics in the preview of linguistics. The fuzzy document clustering uses the sub space-clustering model. The relevant attributes are used for the comparison process. The semantic analysis is used to reduce the vector size. The relevancy is also improved by the semantic analysis. Their system can be enhanced with multi domain ontology to analyze documents with any domain. This also applied to distribute clustering on web document and in XML document.

Fodeh, Punch and Tan [11] presented a methodology for clustering using core semantic features. Their analysis showed that clustering using terms identified by WordNet as nouns often produce results that are comparable to those using Word Sense Disambiguation (WSD). Furthermore, the polysemous and synonymous nouns play an important role in clustering, even though their disambiguation does not necessarily lead to significant improvement in cluster purity. They also showed that it is possible to select a subset of the semantic features that are useful for clustering. They introduced an "unsupervised" information gain measure to determine whether a "disambiguated" noun should be used as a feature in clustering. Their experimental results showed that the core semantic features were sufficient to not only substantially reduce the dimensionality of the feature set, but also maintain or possibly improve clustering using all nouns.

3. BASIC CONCEPT

3.1 Agent System

Agent technology is a new algorithm model, which is highly intelligent, easy to construct distributed system and having strong reusability. The concept of agent and technology has appeared in the development of distributed applied system and shown its remarkable effectiveness. From some research about agent and developing work in the aspect of distributed application, the meaning of the concept and technology of agent is shown in [13].

- Agent technology can improve the application of Internet such as the agent which develops "finding person with information";
- Agent technology can improve the application of parallel projects, such as the manager of agent technology developing work. It can make the workflow and programming known to each workstation, and initiatively guide each workstation according to the workflow and programming, handle and estimate the reports of work condition of each workstation, and manage centrally all kinds of data, and so on.
- Agent technology can be used to develop the distributed interactive simulation system. For example, it can connect the simulator of flight training and several workstations in the computer network, and realize many agents imitating airplanes in workstations to form interactive aviation simulation system together with simulator.

3.2 Fuzzy K-Means

Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. Therefore, it embraces various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis.

The simplest and most commonly used algorithm, employing a squared error criterion is the K-means algorithm [14]. This algorithm partitions the data into K clusters (C_1, C_2, …, C_K), represented by their centers or means. The center of each cluster is calculated as the mean of all the instances belonging to that cluster. The algorithm [14] starts with an initial set of cluster centers, chosen at random or according to some heuristic procedure.

In each iteration, each instance is assigned to its nearest cluster center according to the euclidean distance between the two. Then the cluster centers are re-calculated. The center of each cluster is calculated as the mean of all the instances belonging to that cluster as found in equation 1:

$$\mu_{k=\frac{1}{N_k}\sum_{q=1}^{N_k} x_q} \quad \text{.....(1)}$$

Where N_k is the number of instances belonging to cluster k and μ_k is the mean of the cluster k.

A number of convergence conditions are possible. For example, the search may stop when the partitioning error is not reduced by the relocation of the centers. This indicates that the present partition is locally optimal. Other stopping criteria can be used also such as exceeding a pre-defined number of iterations. Figure 1 presents the pseudo-code [14] of the K-means algorithm.

Input: S (instance set), K (number of cluster)
Output: clusters
1: Initialize K cluster centers.
2: While termination condition is not satisfied do
3: Assign instances to the closest cluster center.
4: Update cluster centers based on the assignment.
5: End While

Figure 1: K-means Algorithm

3.3 WordNet

WordNet [5, 15] has been used in several capacities to improve the performance of information retrieval (IR) systems. WordNet can be used to solve the research problems in information retrieval.

To overcome the weaknesses of term-based representation that is found in the conventional IR approaches, an ontology-based representation has been recently proposed [16], which exploits the hierarchical is-a relation among concepts, i.e., the meanings of words. For example, to describe with a term-based representation documents containing the three words: "animal", "dog", and "cat" a vector of three elements is needed; with an ontology-based representation, since "animal" subsumes both "dog" and "cat", it is possible to use a vector with only two elements, related to the "dog" and "cat" concepts, that can also implicitly contain the information given by the presence of the "animal" concept. Moreover, by defining an ontology base, which is a set of independent concepts that covers the whole ontology, an ontology-based representation allows the system to use fixed-size document vectors, consisting of one component per base concept.

In the text representation, the terms are replaced by their associated concepts in WordNet [17]. In the pretreatment phase, it firstly convert uppercase characters into lowercase characters and then eliminate from text punctuation marks and stop words such as: are, that, what, do. This representation requires two more stages: a) the "mapping" of terms into concepts and the choice of the "merging" strategy, and b) the application of a disambiguation strategy. The first stage is shown in example, as found in figure 2, is about mapping the two terms government and politics into the concept government (the frequencies of these two terms are thus cumulated). Then, among the three "merging" strategies offered by the conceptual approach ("To add concept", "To replace terms by concepts" and "concept only"), the strategy "concept only" can be chosen, where the vector of terms is replaced by the corresponding vector of concepts (excluding the terms which do not appear in WordNet).

Voorhees [18] suggested that WordNet can be used in IR for query expansion. Query expansion is considered to be one of the techniques that can be used to improve the retrieval performance of short queries. Most of the indexing and retrieval methods are based on statistical methods; short queries posed challenges to this model due to the limited amount of information that can be gathered during its processing. In expanding the query, Voorhees suggested using of synonyms, hypernyms, hyponyms, and their combinations. The results showed that using of synonyms, hypernyms, and hyponyms are significant in the retrieval performance for short queries, but little improvement when they are applied to the long query.

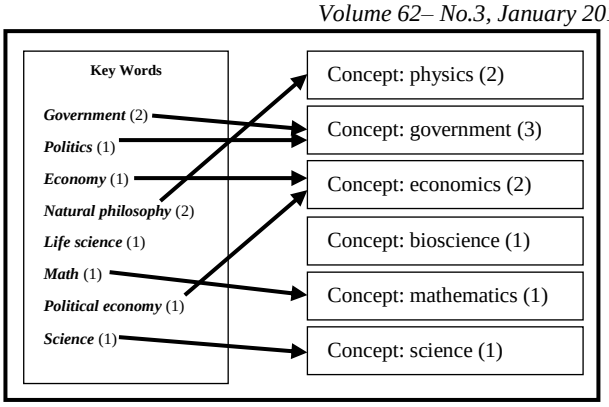

Figure 2: Example of mapping words in concepts

3.4 Clustering Evaluation

The clusters produced by the alternative techniques are compared with the gold standard by means of a proper similarity measure, such as the overlap (number of common elements divided by the total number of elements) [19, 20] or Jaccard similarity coefficient [7] is another similarity measure and, for document clustering, it can be defined as the ratio between the number of common terms in two documents and the number of terms in the union of two documents..

The goal of clustering is attaining high intra-cluster similarity and low inter-cluster similarity. The parameters, true positive, true negative, false positive and false negative are calculated. Two similar documents in the same cluster are true positive (TP) and two similar documents in different clusters are true negative (TN). The decision of assigning two dissimilar documents to the same cluster is false positive (FP) and two similar documents to different clusters is false negative (FN). Precision (P) and recall (R) are two popular measures for clustering. The higher precision is obtained at the low recall value. F-measure (F) is used to evaluate the performance measure of the model. F-measure combines precision and recall, the two most widely used measure. The precision and recall are given in Equation 2. F-measure is evaluated using the Equation 3. A higher F-measure indicates better performance [21, 22].

$$P = \frac{TP}{TP \times FP} \qquad R = \frac{TP}{TP \times FN} \quad \ldots\ldots\ldots\ldots (2)$$

$$F = \frac{2 \times P \times R}{P + R} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

4. THE PROPOSED ARCHITECTURE

The proposed approach utilizes the semantic relationship between words to create concepts in the semantic-based model such as semantic vector space model (SVSM) [23, 24]. It exploits the WordNet ontology in turn to create low dimensional feature vector which allows us to develop an efficient clustering algorithm. A new semantic-based model that analyzes documents based on their meaning is proposed. The proposed model analyzes terms and their corresponding synonyms and/or hypernyms in the documents. The proposed agent aims at increasing the performance of IR process by enhancing the document clustering. The accuracy of clustering has been computed before and after combining ontology with vector space model (VSM) [25]. Experimental results demonstrate that the newly developed semantic-based model enhances the clustering quality of sets of documents substantially. The proposed architecture is shown in figure 3.

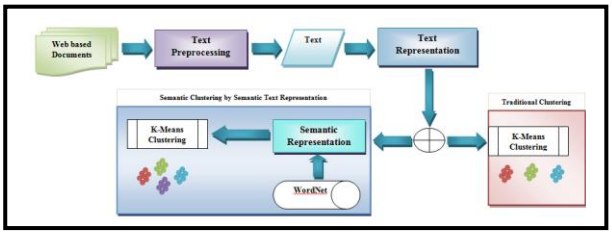

Figure 3: The proposed architecture

4.1 Text Preprocessing

The most widely accepted document representation model in text classification is probably vector space model [25]. VSM is adapted in the proposed system to achieve effective representations of documents. The documents must be preprocessed before the text representation. The main procedures of preprocessing are shown in figure 4.

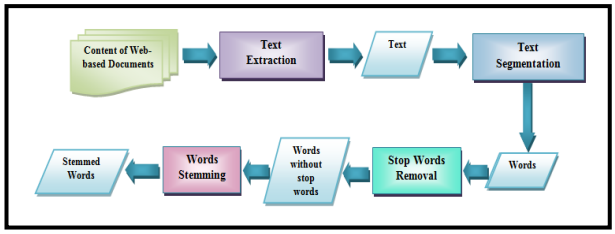

Figure 4: Main steps for text preprocessing

4.1.1 Text Extraction

The first step of in preprocessing process is extracting textual data from the web pages. Then convert each page into individual text document to apply text preprocessing techniques on it. This step is applied on input Web documents dataset by scanning the web pages and categorizing the HTML tags in each page. Then exclude the tags that contain no textual information like formatting tags and imaging tags. Then extract the textual data from other tags (like paragraphs, hyperlinks, and metadata tags) and store it into individual text documents as input for next steps. To extract the text from Web documents, we used the open source high-performance .NET C# module that was created to parse HTML [26] for links, indexing and other purposes.

4.1.2 Stop Words Removal

Stop words, i.e. words thought not to convey any meaning, are removed from the text. In this work, the proposed approach uses a static list of stop words about all tokens. This process removes all words that are not nouns, verbs or adjectives. For example, stop words removal process will remove all the words like: he, all, his, from, is, an, of, your, and so on. Removing these words will save spaces for storing document contents and reduce time taken during the search process.

4.1.3 Words Stemming

The stem is the common root-form of the words with the same meaning appear in various morphological forms (e.g. player, played, plays from stem play). In the proposed approach, we used the morphology function [27] provided with WordNet [28] that is used for stemming process. Stemming will find the stems of the output terms to enhance term frequency counting process This process will output all the stems of extracted terms [12, 29].

4.2 Text Representation

4.2.1 Document Representation

VSM [25] is adapted in the proposed system to achieve effective representations of documents. Each document is identified by n-dimensional feature vector where each dimension corresponds to a distinct term. Each term in a given document vector has an associated weight.

The weight is a function of the term frequency, collection frequency and normalization factors. Different weighting approaches may be applied by varying this function. Hence, a document j is represented by the document vector d_j :

$$d_j = (w_{1j}, w_{2j}, w_{nj}) \quad \text{.....(4)}$$

Where, w_{nj} is the weight of the kth term in the document j.
The term frequency reflects the importance of term k within a particular document j. The weighting factor may be global or local. The global weighting factor clarify the importance of a term k within the entire collection of documents, whereas a local weighting factor considers the given document only.
The document keywords were extracted by using a term-frequency and inverse-document-frequency (*tf-idf*) calculation [30], which is a well-established technique in information retrieval. The weight of term k in document j is represented as:

$$w_{kj} = tf_{kj} \times (log_2^n - log_2^{df_k} + 1) \quad \text{.....(5)}$$

Where: tf_{kj} = the term k frequency in document j, df_k = number of documents in which term k occurs, n = total number of documents in collection.

The main purpose of this step is to extract the item in the document, then get term frequency that reflects the importance of term. Finally get the weight of terms in the selected document. The output of this step is the weight of terms in selected document.

4.2.2 Words Mapping into Concepts using WordNet

The purpose of this step is identifing WordNet concepts that correspond to the document words. Concept identification [31] is based on the overlap of the local context of the analyzed word with every corresponding WordNet entry. The words mapping into concepts algorithm is given in figure 5.

Input: (B_w) Bag of words (W_i) in document *D* that was gotten from Words Stemming phase.
Output: Set of all WordNet concepts belonging to terms (words) in document *D*.
Procedure:
// (C_w) is the count of words in the bag, and ($Cont_i$) the context of the word in the document, it is the sentence in document D that contains the word occurrence being analyzed.
Do While $i <= C_w$
  Get WordNet entries C_i set ($CSet_i$) that is containing the word W_i , where $C_i \in CSet_i$.
  Save W_i and its C_i in database table.
EndDo
Rank concepts C_i in $CSet_i$ where $|C_1| > |C_2| > |C_3| \ldots > |C_n|$ *// | | denotes the concept length, in terms of the number of words in the corresponding terms. $CSet_{ri}$ is the ranked concepts set.*
FOR each C_i in $CSet_{ri}$
  Get common words between $Cont_i$ and representative term of C_i , which is the intersection $Cint = \cap (Cont_i, C_i)$.
  If $|Cint| < |C_i|$ then
    The concept-sense C_i is not within the context $Cont_i$.
  EndIf
  If $|Cint| = |C_i|$ then
    The concept-sense C_i is within the context $Cont_i$.
    Add C_i to the set of possible senses associated with the document.
  EndIf
EndFor

Figure 5: Words Mapping into Concepts using WordNet

4.2.3 Weight of Concept Computation

The concepts in the documents are identified as a set of terms that have identified or synonym relationships, i.e., synsets in the WordNet ontology. Then, the concept frequencies Cf_c are calculated based on term frequency tf_{tm} [32] as found in equation 6.

$$Cf_c = \sum_{t_m \in r(c)} tf_{tm} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Where r(c) is the set of different terms that belongs to concept C .

Note that WordNet returns an ordered list of synsets based on a term. The ordering is supposed to reflect how common it is that the term is related to the concept in standard English language. More common term meanings are listed before less common ones. Using the first synset as the identified concept for a term can improve the clustering performance more than that of using all the synsets to calculate concept frequencies.

Hypernyms of concepts can represent such concepts up to a certain level of generality. The concept frequencies are updated as found in equation 7.

$$hf_c = \sum_{b \in H(c,r)} Cf_b \ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

where $H(c, r)$ is the set of concepts C_H , which are all the concepts within r levels of hypernym concepts of c. In WordNet, it is obtained by gathering all the synsets that are hypernym concepts of synset c within r levels. In particular, $H(c, \infty)$ returns all the hypernym concepts of c and $H(c, 0)$ returns just c. The weight of each concept c in document d is computed as follows:

$$wh_c = hf_c \times idf_c \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

where idf_c is the inverted document frequency of concept c by counting how many documents in which concept c appears as the weight of each term t in the document d.

5. EXPERIMENTAL EVALUATION

In this section, the experiment is performed to gauge many aspects of the proposed system by comparing the proposed clustering method using semantic-based model and k-means clustering using VSM. This is accomplished by using two clustering evaluation metric; f-measure and refrence ovelap, to compute the clustering efficiency of set of documents. In the experiment, a speed of clustering is compared for the two methods. The input ducuments in the implemented agent is collected by Web crawler that collects the documents, which are from Web and about computer science domain. The number of collected documents is about 3000 documents and the agent saves it in the system repasitory. Figure 6 depicts a extracted text of documents that are saved in the system repasitory.

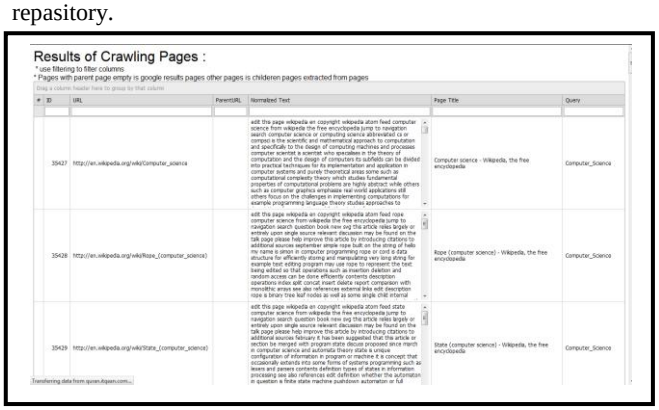

Figure 6: Sample of collected documents from Web

After the agent collects the Web-based documents, it can perform the clustering approach based on semantic-based model and k-means.

The evaluation of proposed system depends on performing the clustering to extract two clusters of documents or three cluster or four clusters or five clusters using and VSM and the proposed semantic-based model. Figure 7 shows the main form of the clustering in the proposed system.

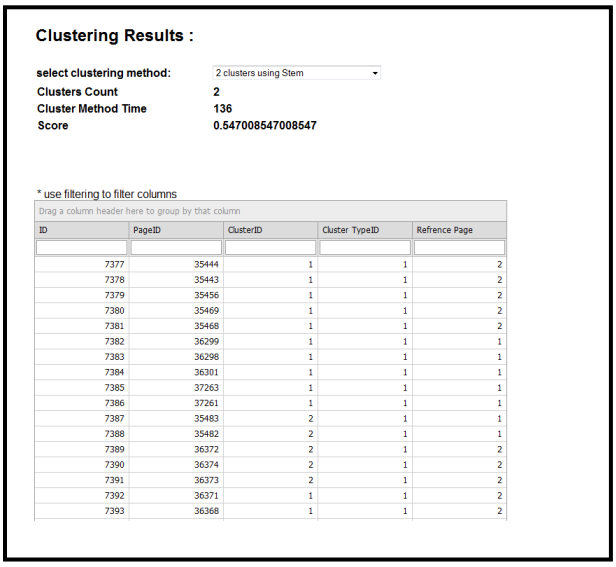

Figure 7: Sample of clustering documents

Table 1 shows the difference between the ouput clusters that used VSM and the output clusters that used the semantic-based model.

Table 1: Difference between the clusters using VSM and the proposed Semantic-based Model for two clusters

Using VSM			Using Semantic-Based Model		
DocID	Cluster ID	Reference Cluster	DocID	Cluster ID	Reference Cluster
35444	1	2	37263	1	1
35443	1	2	37261	2	1
35456	1	2	35470	2	1
35469	1	2	35479	2	1
35468	1	2	35477	1	1
36299	1	1	36374	1	2
36298	1	1	36371	1	2
36301	1	1	36298	2	1
37263	1	1	36301	2	1
37261	1	1	36299	1	1
35483	2	1	35469	2	2
35482	2	1	35468	2	2
36372	2	2	36368	2	2
36374	2	2	36373	2	2
36373	2	2	36372	2	2
36371	1	2	35483	2	1
36368	1	2	35482	2	1
35470	2	1	35443	2	2
35479	2	1	35456	2	2
35477	2	1	35444	2	2

Figure 8 shows the difference between the clustering that is used VSM and clustering that is used semantic-based model by reference overlap metric.

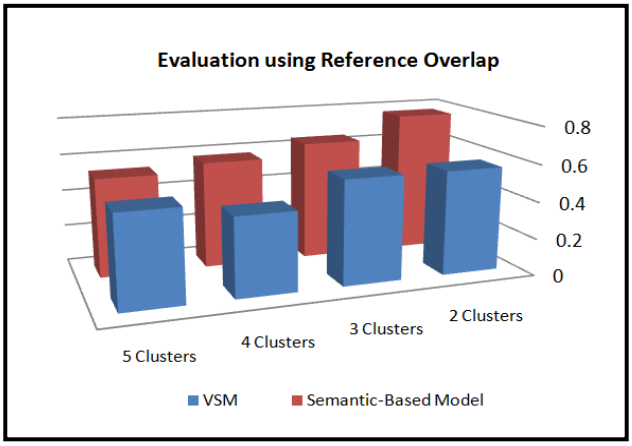

Figure 8: Difference between the clustering using VSM and clustering using semantic-based model by reference overlap metric

Figure 9 shows the difference between the clustering that is used VSM and clustering that is used semantic-based model by f-measure.

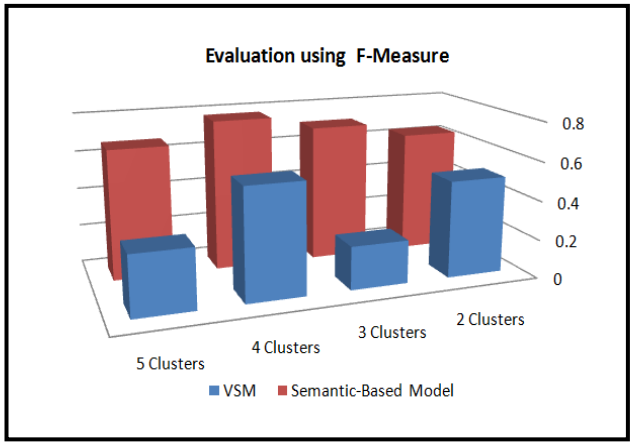

Figure 9: Difference between the clustering using VSM and clustering using semantic-based model by F-Measure

Figure 10 shows that the clustering that is used semantic-based model is faster than the clustering that is used traditional VSM, because the process time of clustering that is used semantic-based model is shorter than the clustering that is used traditional VSM.

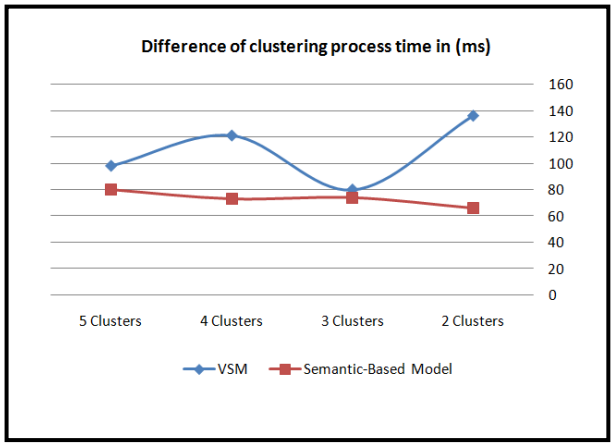

Figure 10: Difference of clustering process time using two methods

6. CONCLUSION

Because of the volume of information continues to increase, there is growing interest in helping people better find, filter and manage these resources. Text clustering, which is the process of grouping documents having similar properties based on semantic and statistical content, is an important component in many IR and management tasks. In the present work, the proposed agent to document clustering was considered. Also the methods and performance of the proposed agent were analyzed.

The proposed agent aims at providing qualitative improvement over traditional VSM by using semantic-based model based on WordNet ontology.

The proposed semantic-based model framework provides improved performance and makes a clustering be efficient. It also overcomes the problems existing in the VSM commonly used for clustering. The clustering result based on semantic-based model has higher efficiency values and faster than those based on the traditional VSM.

7. REFERENCES

[1] Oikonomakou, N. & Vazirgiannis, M. (2010). A Review of Web Document Clustering Approaches. In: Data Mining and Knowledge Discovery Handbook, 2nd edition. DOI 10.1007/978-0-387-09823-4_48, Springer Science+Business Media.

[2] TONG, T. (2010). Semantic frameworks for document and ontology clustering. A dissertation in Computer Science and Computer Networking Presented to the Faculty of the University of Missouri–Kansas City..

[3] Viswanth, p., Patra, b. & Babu, v. (2009). Some Efficient and Fast Approaches to Document Clustering. In: Handbook of Research on Text and Web Mining Technologies, 181-188 pp, DOI: 10.4018/978-1-59904-990-8.ch011. IGI Global.

[4] Zhao, Y., Cao, L., Zhang, H. & Zhang, C. (2009). Data Clustering. In: Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends. 562-572 pp. DOI: 10.4018/978-1-60566-242-8.ch060. IGI Global.

[5] Fellbaum, C. (2010). WordNet. Theory and Applications of Ontology: Computer Applications, 231, PP: 231-243, Springer Science+Business Media B.V.

[6] Drakshayani, B. & Prasad, E. (2012). Text Document Clustering based on Semantics. International Journal of Computer Applications (0975 – 8887). Vol. 45– No.4.

[7] Luo, C., Li, Y. & Chung, S. (2009). Text document clustering based on neighbors. Data & Knowledge Engineering 68 (2009) 1271–1288. Elsevier B.V.

[8] Shah, N & Mahajan, S. (2012). Semantic based Document Clustering: A Detailed Review. International Journal of Computer Applications (0975 – 8887). Vol. 52– No.5.

[9] Gharib, T., Fouad, M., Mashat, A. & Bidawi, I. (2012). Self Organizing Map -based Document Clustering Using WordNet Ontologies, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2.

[10] Thangamani, M. & Thangaraj.P. (2010). Ontology Based Fuzzy Document Clustering Scheme. Modern Applied Science. Vol. 4, No. 7.

[11] Fodeh, S., Punch, B. & Tan P. (2011). On ontology-driven document clustering using core semantic features. Knowl Inf Syst (2011) 28:395–421. DOI 10.1007/s10115-010-0370-4. Springer-Verlag London Limited.

[12] Gharib, T., Fouad, M. & Aref, M. (2010). Fuzzy Document Clustering Approach using WordNet Lexical Categories. In: Advanced Techniques in Computing Sciences and Software Engineering. DOI 10.1007/978-90-481-3660-5, Springer Science+Business Media.

[13] Georgakarakou, C. E., & Economides, A. A. (2008). Software Agent Technology: An Overview. Software Applications: Concepts, Methodologies, Tools, and Applications. 128-151 pp. IGI Global.

[14] M. Oded, R. Lior. (2010). A survey of Clustering Algorithms. In: Data Mining and Knowledge Discovery Handbook Second Edition. DOI 10.1007/978-0-387-09823-4_14. Springer Science+Business Media.

[15] Maria, I. & Loke, S. (2010). The Impact of Ontology on the Performance of Information Retrieval: A Case of WordNet, In G. I. Alkhatib, D. C. Rine, Web Engineering Advancements and Trends: Building New Dimensions of Information Technology, DOI: 10.4018/978-1-60566-719-5.ch002, 24-37.

[16] Pereira, d. C., Tettamanzi, C. (2006). A.G.B.: An ontology-based method for user model acquisition. In: Ma, Z. (ed.) Soft computing in ontologies and semantic Web. Studies in fuzziness and soft computing, pp. 211–227. Springer, Heidelberg.

[17] Amine, A., Elberrichi, Z. & Simonet, M. (2010). Evaluation of Text Clustering Methods Using WordNet, The International Arab Journal of Information Technology, (7) 4.

[18] Voorhees, E. (1994). Query Expansion Using Lexical-Semantic Relations. The 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (Dublin Ireland, 1994), 61. ACM.

[19] Koschke, R. & Eisenbarth, T. (2000). A Framework for Experimental Evaluation of Clustering Techniques. 0-7695-0656-9/00, IEEE.

[20] Tonella, P., Ricca, F., Pianta, E. & Girardi, C. (2003). Evaluation Methods for Web Application Clustering. Proceedings of the Fifth IEEE International Workshop on Web Site Evolution (WSE'03). 0-7695-2016-2/03, IEEE.

[21] Sridevi, U. (2011). An Ontology Based Model for Document Clustering. International Journal of Intelligent Information Technologies (IJIIT). Vol.7 (3), PP: 54-69. DOI: 10.4018/jiit.2011070105.

[22] Punitha, S. & Punithavalli, M. (2012). Performance Evaluation of Semantic Based and Ontology Based Text Document Clustering Techniques. Procedia Engineering 30 (2012) 100 – 106. Elsevier Ltd.

[23] Liu, G. (1994). The Semantic Vector Space Model (SVSM) A Text Representation and Searching Technique. 1060-3425/94, IEEE.

[24] Zhao, L., Jianguo, D. (2010). An Efficient Semantic VSM based Email Categorization Method. International Conference on Computer Application and System Modeling (ICCASM 2010), 978-1-4244-7237-6, IEEE

[25] Liu, Y. (2009). On Document Representation and Term Weights in Text Classification. In: Handbook of Research on Text and Web Mining Technologies. PP: 1-22. DOI: 10.4018/978-1-59904-990-8.ch001. IGI Global.

[26] Majestic-12: Projects : C# HTML parser (.NET). http://www.majestic12.co.uk/projects/html_parser.php.

[27] wordnetdotnet - Revision 262. http://wordnetdotnet.googlecode.com/svn/trunk/Projects/Thanh/ .

[28] Bai, R., Wang, X. & Liao, J. (2010). Extract Semantic Information from WordNet to Improve Text Classification Performance. AST/UCMA/ISA/ACN 2010, LNCS 6059, pp. 409–420. Springer-Verlag Berlin Heidelberg.

[29] Tarek, G., Fouad, M. & Aref, M. (2008). Web Document Clustering Approach using WordNet Lexical Categories and Fuzzy Clustering. Proceedings of International Workshop on Data Mining and Artificial Intelligence (DMAI' 08), 24 December, 2008, Khulna, Bangladesh. 1-4244-2136-7/08, IEEE.

[30] Jones, K. (2004). A Statistical Interpretation of Term Specificity and its Application to Retrieval. Journal of Documentation, 60 (5), p.493-502.

[31] B. Fatiha, B. Mohand, T. Lynda, D. Mariam. (2010). Using WordNet for Concept-Based Document Indexing in Information Retrieval, SEMAPRO: The Fourth International Conference on Advances in Semantic Processing, Pages: 151 to 157, IARIA.

[32] Dragoni, M., Pereira, C. & Tettamanzi, A. (2010). An Ontological Representation of Documents and Queries for Information Retrieval Systems, IEA/AIE 2010, Part II, LNAI 6097, pp. 555–564, Springer-Verlag Berlin Heidelberg.