Evaluation of Text Clustering Methods Using WordNet

Abdelmalek Amine[1,2], Zakaria Elberrichi[1], and Michel Simonet[3]
[1]EEDIS Laboratory, Department of Computer Science, UDL University, Algeria
[2]Department of Computer Science, UTMS University, Algeria
[3]IN3S, Joseph Fourier University, France

Abstract: *The increasing number of digitized texts presently available notably on the Web has developed an acute need in text mining techniques. Clustering systems are used more and more often in text mining, especially to analyze texts and to extract knowledge they contain. With the availability of the vast amount of clustering algorithms and techniques, it becomes highly confusing to a user to choose the algorithm that best suits its target dataset. Actually, it is very hard to define which algorithms work the best, since results depend considerably on the application and on the kinds of data at hand. In this paper, we propose, study and compare three text clustering methods: an ascending hierarchical clustering method, a SOM-based clustering method and an ant-based clustering method, all of these based on the synsets of WordNet as terms for the representation of textual documents. The effects of these methods are examined in several experiments using 3 similarity measurements: the cosine distance, the Euclidean distance and the manhattan distance. The reuters-21578 corpus is used for evaluation. The evaluation was done, by using the F-measure. The results obtained show that the SOM-based clustering method using the cosine distance provides the best results.*

Keywords: *Text clustering, similarity, WordNet, reuter-21578, and F-measure.*

Received February 24, 2009; accepted August 3, 2009

1. Introduction

With the great and rapidly growing number of documents available in digital form (Internet, library, CD-Rom…), the automatic classification of texts has become a significant research field. The automatic classification of texts is the action of distributing by categories or classes a set of documents according to some common characteristics. The terms "categorization" or "classification" are used when dealing with the assignation of a document to a class (with predefined classes). In this case we are within the framework of supervised learning. The term "clustering" (unsupervised classification) designates the creation of classes or groups (clusters) of a certain number of similar objects without prior knowledge; we are then within the framework of unsupervised learning.

Unsupervised classification or "clustering" is automatic and discover latent (hidden) unlabeled classes. The classes are isolated from one another and are to be discovered automatically. It is sometimes possible to fix their number. A great number of unsupervised classification methods have been applied to textual documents, however, the combinations between clustering methods and the representation of texts based on concepts was not extensively studied. In this paper, we study the clustering of textual documents, first with the ascending hierarchical clustering method, then with the Kohonen self-organizing Maps and last with the ant-based clustering method, all of these methods using WordNet synsets as terms for the representation of textual documents.

Section 2 will introduce different possible ways of representing a text, explain similarity measurements and will review the most known clustering algorithms. Section 3 is devoted to the description of the proposed approaches in all their stages, and in section 4 we evaluate and discuss the obtained results. Finally section 5 will conclude the article.

2. State of the Art

Implementing these methods initially consists in choosing a way of representing the documents [21], because there is currently no learning method able to directly process unstructured data (texts). Then, it is necessary to choose a similarity measurement, and lastly to choose an unsupervised classification algorithm which we will develop using the descriptors and the metric that have been chosen.

2.1. Representation of the Textual Documents

To implement any method of classification it is initially necessary to transform the digitized texts into an efficient and meaningful way so that they can be analyzed.

The space vector model is the most used approach to represent textual documents: we represent a text by a numerical vector obtained by counting the most relevant lexical elements present in the text. All document *dj* will be transformed into a vector:

$$d_j = (w_{1j}, w_{2j}, ..., w_{|T|j}) \quad (1)$$

where *T* is the whole set of terms (or descriptors) which appear at least once in the corpus (|T| is the size of the vocabulary), and w_{kj} represents the weight (frequency or importance) of the term t_k in the document *dj*.

Table 1. Document-term matrix.

Documents	Terms or Descriptors						
$\mathbf{d_1}$	w_{11}	w_{21}	w_{31}	...	w_{j1}	...	w_{n1}
$\mathbf{d_2}$	w_2	w_{22}	w_{32}	...	w_{j2}	...	w_{n2}
…	…	…	…	...	…	...	…
$\mathbf{d_m}$	w_{1m}	w_{2m}	w_{3m}	...	w_{jm}	...	w_{nm}

- The simplest representation of texts introduced within the framework of the vector space model is called "bag of words" [19, 1]; it consists in transforming texts into vectors where each component represents a word. This representation of texts excludes any grammatical analysis and any concept of distance between the words, and syntactically destructures texts by making them understandable to the machine.
- Another representation, called "bag of phrases", carries out a selection of sentences (sequences of words in the text, and not the lexeme "phrases" as we usually understand it), by favoring those which are likely to carry a significant meaning. Logically, such a representation must provide better results than those obtained by the "bag of words" representation. However, experiments [20] have shown that if semantic qualities are preserved, statistical qualities are much degraded.
- Another method for the representation of texts calls upon the techniques of lemmatization and stemming. Stemming consists in seeking the lexical root of a term [18] while lemmatization replaces a term by a conventional standard form, e.g., infinitive form for verbs and singular for nouns [12]. This prevents that each inflection or form of a word should be regarded as a different descriptor and consequently creates one more dimension.
- Another method of representation, which has several advantages, is based on "n-grams" (a "n-gram" is a sequence of n consecutive characters). The whole set of n-grams (n generally varies from 2 to 5) which can be generated for a given document is mainly the result of the displacement of a window of n characters along the text [15]. The window is moved by a character at a time and the number of occurrences of each n-gram is counted [5, 17].
- The concept-based representation, also called ontology-based representation, also uses the vector-space formalism to represent documents. The characteristic of this approach lies in the fact that the elements of the vector space are not associated with index terms only but with concepts, which is made possible by adding an additional stage to map terms into the concepts of ontology.

There are various methods to calculate the weight w_{kj} knowing that, for each term, it is possible to calculate not only its frequency in the corpus but also the number of documents which contain this term.

Most approaches [21] are centered on a vectorial representation of texts using the *TFxIDF* measure. The frequency *TF* of a term *T* in a corpus of textual documents corresponds to the number of occurrences of the term *T* in the corpus. The frequency *IDF* of a term *T* in a corpus of textual documents corresponds to the number of documents containing *T*. These two concepts are combined (by product) in order to assign a stronger weight to terms that appear often in a document and rarely in the complete corpus.

$$TF \times IDF(t_k, d_j) = Occ(t_k, d_j) \times Log \frac{Nb\_doc}{Nb\_doc(t_k)} \quad (2)$$

where $Occ(t_k, d_j)$ is the number of occurrences of the term t_k in the document d_j, *Nb_doc* is the total number of documents of the corpus and $Nb\_doc(t_K)$ is the number of documents of this unit in which the term t_k appears at least once.

There is another measurement of weighting called *TFC* similar to *TF×IDF* which corrects the lengths of the texts by a cosine standardization, to avoid giving more credit to the longest documents.

$$TFC(t_k, d_j) = \frac{TF \times IDF(t_k, d_j)}{\sqrt{\sum_{k=1}^{|T|} (TF \times IDF(t_k, d_j))^2}} \quad (3)$$

2.2. Similarity Measure

Typically, the similarity between documents is estimated by a function calculating the distance between the vectors of these documents: two close documents according to this distance are regarded as similar. Several measures of similarity have been proposed [9]. Among these measurements we can quote:

- The cosine the distance

$$Cos(d_i, d_j) = \frac{\sum_{t_k} [TF \times IDF(t_k, d_i)] \bullet [TF \times IDF(t_k, d_j)]}{\|d_i\|^2 \bullet \|d_j\|^2}$$

- The Euclidean distance

$$Euclidean(d_i, d_j) = \sqrt{\sum_{1}^{n} (w_{ki} - w_{kj})^2}$$

- The Manhattan distance

$$Manhattan(d_i, d_j) = \sum_{l}^{n} |w_{ki} - w_{kj}|$$

2.3. Algorithms for Clustering of Textual Documents

Unsupervised classification or "clustering" is one of the fundamental data mining techniques to cluster structured or unstructured data. Several methods have been proposed; according to [4, 23], these methods can be classified as follows:

- *Hierarchical methods:* these methods generate a hierarchical tree of classes called dendrogram. There are two ways of building the tree: starting from the document or starting from the set of all the documents or corpus. When starting with the documents, each document is initially put into a class of its own. Then, the two most similar classes are combined into one class. This process is repeated until a certain termination condition is satisfied. This method is called "agglomeration of similar groups" or "ascending hierarchical clustering". When starting with the whole set of documents (or corpora), the method is called "division of dissimilar groups" or "descending hierarchical clustering". At the beginning of this process, there is only one class, which contains all the documents. The class is divided into two subclasses at the following iteration. The process continues until the termination condition is satisfied. The similarity between two documents is based on the distance between the documents.
- *Partitioning methods:* these methods are also called flat "clustering". The most known methods are the method of K-medoids, the method of the dynamic clouds and the method of K-means or mobile centers. In the method of K-means, for example, the number of classes is preset. A document is put into a class if the distance between the vector of the document and the center of this class is the smallest in comparison with the distances between the vector and the centers of the other classes.
- *Density-based methods:* it consists in grouping the objects as long as the vicinity density exceeds a certain limit. The groups or classes are dense areas separated by sparsely dense areas. A point (document vector) is dense if the number of its neighbors exceeds a certain threshold and a point is close to another point if it is at a distance lower than a fixed value. The discovery of a group or class is made in two stages: choose a dense point randomly, and all the points which are attainable starting from this point, according to the density threshold, form a group or a class.
- *Grid-based methods:* it is a division of the data space into multidimensional cells forming a grid (points in the grid represent data items) and grouping close cells in terms of distance. Classes are built by assembling the cells containing enough data (dense). Several levels of grids are used, with an increasingly high resolution.
- *Model-based methods:* one of the model-based methods is the conceptual approach. In this approach we have a conceptual hierarchy inherent to the data where a concept is a couple (intension, extension) knowing that the intension is the maximal set of attributes common to the vectors and the extension is the maximal set of vectors sharing the attributes.

Another model-based method is the Kohonen networks method also called Self-Organizing Maps (SOM). It is an interesting neural method because it orders the obtained classes topologically in the form of a map, generally on a plan (i.e., two-dimensional).

Another model-based method is the Ant-based approach which is a biomimetic method inspired from the self-assembly behavior of real ants. Real ants can build complex structures by connecting themselves to each others.

3. Evaluation of Text Clustering Methods Using WordNet

Our experiments, we have developed 3 clustering methods based on WordNet for the representation of texts, that we evaluate and compare them: an ascending hierarchical clustering, a SOM-based clustering SOM and an ant-based clustering.

3.1. Corpus

The data used in our experiments come from the texts of the reuters-21578 corpus, which is a set of financial dispatches emitted during the year 1987 by the reuters agency in the English language and freely available on the web. This corpus is an update of the reuters-22173 corpus. This update was carried out in 1996. The texts of this corpus have a journalistic style. The characteristic of the corpus reuters-21578 is that each document is labeled with several classes. This corpus is often used as a basis for comparison between the various tools for documents classification.

We have used these texts in our experiments after having carried out some modifications in the pretreatment phase.

3.2. Configuration

Our algorithms were developed with Borland JBuilder version7 and a Windows XP platform, on a machine with a processor INTEL Pentium 4 (2,66 GHz) with 256 Mb RAM. We tested for each approach three similarity measurements: the cosine distance, the Euclidean distance and the Manhattan distance.

3.3. The Conceptual Approach for Documents Representation

3.3.1. Wordnet and the Classification of Texts

WordNet [16] is ontology of cross-lexical references whose design was inspired by the current theories of human linguistic memory. English names, verbs, adjectives and adverbs are organized in sets of synonyms (synsets), representing the underlying lexical concepts. Sets of synonyms are connected by relations. WordNet covers most names, verbs, adjectives and adverbs of the English language. The latest version of WordNet (2.1) is a vast network of 155000 words, organized in 117597 synsets. There is a rich set of 391.885 relations between the words and the synsets, and between the synsets themselves.

The basic semantic relation between the words in WordNet is synonymy. Synsets are linked by relations such as specific/generic or hypernym /hyponym (is-a), and meronym/holonym (part-whole). The principal semantic relations supported by WordNet is synonymy: the synset (synonym set), represents a set of words which are interchangeable in a specific context. WordNet is used in many text classification methods as well as in Information Retrieval (IR) because of its broad scale and free availability. Studies in which the synsets of WordNet were used as index terms have very promising results [6, 7, 14].

3.3.2. Representation of Documents Based on Wordnet

We propose a representation which replaces terms by their associated concepts in WordNet. In the pretreatment phase, we first convert uppercase characters into lowercase characters and then eliminate from texts 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 as shown in example Figure 1 is about mapping the two terms government and politic**s** 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"), we choose the strategy "concept only", where the vector of terms is replaced by the corresponding vector of concepts (excluding the terms which do not appear in WordNet).

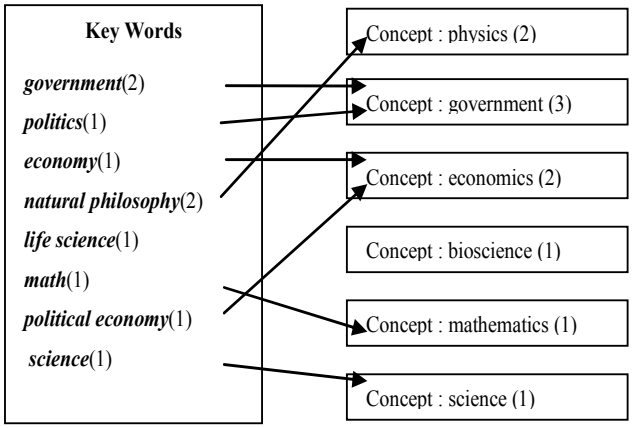


Figure 1. Example of mapping words in concepts.

It is clear that the assignment of terms to concepts in ontology can be ambiguous. For this reason adding or replacing terms by concepts can cause a loss of information. Indeed, the choice of the most appropriate concept for a term can influence the efficacy of the classification process.

In our approach we use a simple disambiguation method: the strategy of the "First concept". WordNet gives for each term a list of concepts ordered according to a certain criterion. This disambiguation strategy consists in taking only the first concept of the list as the most suitable concept. The frequency of a concept is then calculated as follows:

$$cf(d,c) = tf\{d \in t \mid first(ref_c(t)) = c\}\} \quad (4)$$

For the calculation of weights (frequencies), we use the *TFxIDF* function, knowing that the terms are synsets and the vectors of the documents are vectors of concepts which will be normalized

3.4. Ascending Hierarchical Clustering

This method consists of creating, for each step, a partition obtained by aggregating pairwise the closest elements. Then be designated by element both the objects to classify and the clustering (clusters) of objects generated by the algorithm. The result is a hierarchy of partitions in the form of a tree called dendrogram (containing *n-1* partitions) [22].The advantage of these trees is to give an idea of how many clusters exist in all the set of objects. Each cut of a tree provides a partition with fewer clusters that are cut above.

This method begins by identifying among the *n* objects, the 2 objects that are most similar compared to all the p variables specified. It will then merge these 2 objects to form a cluster. There are therefore at this level (*n-1*) cluster, one being composed of 2 previously grouped objects, the others containing a single object. The process continues by determining the 2 clusters which are most similar, and in bringing them together. This process is repeated until a single

cluster is obtained containing all objects. This process is based on 2 choices:

- The determination of the criterion of resemblance between objects (distance or similarity).
- The determination of the similarity criteria between clusters: called aggregation criteria.

Many aggregation criteria were proposed, the most known are: the single linkage, the average linkage, the complete linkage and the Ward's criteria. According to most references the criterion most commonly used is the Ward's criteria.

In practice, once the tree constructed, the user is not necessarily interested in the complete hierarchy but in a single partition obtained by cutting the tree generated by a horizontal line. It raises the problem of finding the best method to cut the tree in the right place and hence decide the proper number of clusters found. The most used approach to solve this problem is based on a simple principle: it is to cut the tree at level where the variation in the minimum distance between clusters is maximal, i.e., where the clusters are farthest.

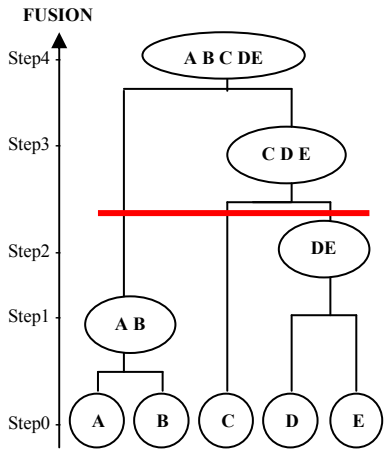


Figure 2. An example of ascending hierarchical clustering, the tree is cut by a horizontal line at level 3.

In our experiments, at the beginning of our algorithm, each document, characterized by *p* descriptors, is considered as a class:

- We calculate the square symmetric matrix of distance "documents x documents", with the cosine distance, then with the Euclidian distance and finally with the Manhattan distance.
- For each distance, with the average linkage, then with the complete linkage and finally with the Ward's criteria we do:
- Repeat: merge the two elements (documents or clusters) closest.
- Calculate a new distance matrix between clusters remaining,
- Until fusion of all clusters into one cluster in (*n-1*) steps.

With each distance and each aggregation criteria used, each obtained tree is cut at level where the variation in the minimum distance between clusters is maximal. We calculated learning rate, we obtained the results presented in the Table 2.

Table 2. Number of classes and learning rate according to the 3 similarity measurements with the ascending hierarchical clustering method.

	Cosine	Euclidean	Manhattan
Ward			
Number of Classes	19	22	16
Maximal Learning Rate (%)	12,85	13,65	10,76
Average Linkage			
Number of Classes	16	21	20
Maximal Learning Rate (%)	11,64	13,32	12,91
Complete Linkage			
Number of Classes	16	17	17
Maximal Learning Rate (%)	11,97	12,25	12,47

According to the number of classes and to the learning rate, the best values are obtained by the Euclidean distance and the Ward's criteria.

3.5. Self-Organizing Maps of Kohonen

Self-Organizing Maps (SOM) of Kohonen is an unsupervised learning method which is based on the principle of competition according to an iterative process of updates.

The Kohonen model or network proposed by Tuevo Kohonen [10] is a grid (map), generally two-dimensional, of *p* by *p* units (cells, nodes or neurons) *Ni*. It is made up of:

- An input layer: any object to be classified is represented by a multidimensional vector (the input vector). To each object a neuron is assigned, which represents the centre of the class.
- An output layer (or competition layer). The neurons of this layer enter in competition to be activated according to a chosen distance; only one neuron is activated (winner-takes-all neuron) following the competition.

The SOM Algorithm has been proposed and applied for a long time in the field of classification of textual documents. Many researchers are currently working on SOMs [11, 2, 3].

Each input vector is normalized. The initial weights (randomly generated) are also normalized. The map we used is two-dimensional and its size is 7x7.The initial neighbourhood is 8.

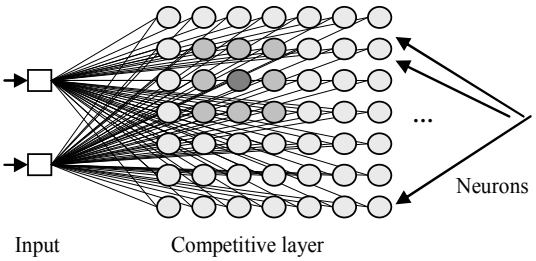


Figure 3. Kohonen network architecture.

In this approach too, we calculated learning rate, we obtained the results presented in the Table 3.

Table 3. Number of classes and learning rate according to the 3 similarity measurements with the SOM clustering method.

	Cosine	Euclidean	Manhattan
Number of Classes	22	27	27
Maximal Learning Rate (%)	14,09	13,08	7,71

According to the number of classes and to the learning rate, the best values are obtained by the cosine distance.

3.6. Ant-Based Clustering

The numerous abilities of ants have inspired researchers for more than ten years regarding designing new clustering algorithms [13, 8]. The model which has been studied the most is the way ants sort objects in their nest [13, 8]. These ants based algorithms may inherit from real ants interesting properties, such as the local/global optimization of the partitioning, the absence of need of a priori information on an initial partitioning or number of classes, or parallelism.

We used the algorithm proposed by Lumer and Faieta [13] where objects are initially distributed randomly on a 2D grid. Each ant is in a box of this grid and only reaches the objects in its neighbourhood (8 neighbours for example).

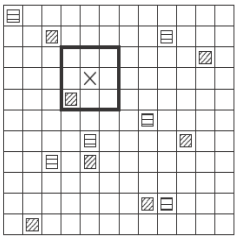
An object o_i on the grid is picked up with probability p_p heightened it is somewhat similar to neighbouring objects. In the same way, an object o_i carried by an ant is more easily deposited in a region with objects that are similar to him with probability p_d.

$$p_p(o_i) = \left(\frac{c_1}{c_1 + f(o_i)}\right)^2$$
$$p_d(o_i) = \begin{cases} 2f(o_i) & if \quad f(o_i) < c_2 \\ 1 & if \quad f(o_i) \geq c_2 s \end{cases} \quad (5)$$

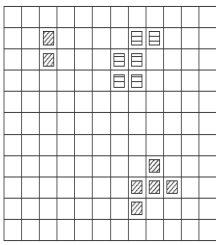
The local density function depends on the object oi and its position on the grid $r(o_i)$. It is calculated as follows:

$$f(o_i) = \begin{cases} \frac{1}{s^2} \sum_{o_j \in R_s(r(o_i))} 1 - \frac{d(o_i, o_j)}{\alpha} & if \quad f > 0 \\ 0 & else \end{cases} \quad (6)$$

$f(o_i)$ is then a measure of the average similarity of the object o_i with objects o_j in its neighbourhood. α is a scale factor determining the extent of dissimilarity between two objects is taken into account. The *LF* algorithm gives the steps of the method using A ants $\{a_1, \ldots, a_A\}$.



a. The objects are distributed randomly on the grid. Ants can pick up and deposit them in boxes where the density of similar objects is high.



b. The ant is represented by X and its perimeter detection by a thick line, objects are represented by squares with an interior ("invisible" to the ant) is the original class).

Figure 4. Principle of clustering by artificial ants according to the algorithm presented by [13].

LF Algorithm

1. *The N objects $o_1, \ldots, o_N$ are placed randomly on the grid G*
2. *for T = 1 to T_{max} do*
3. *for all $a_j \in \{a_1, \ldots, a_A\}$ do*
4. *if the ant a_j does not carry object and $r(o_i) = r(a_j)$ then*
5. *Calculate $f(o_i)$ and $p_p(o_i)$*
6. *The ant a_j collects object o_i according the probability $p_p(o_i)$*
7. *otherwise*
8. *if the ant a_j carries object o_i and box $r(a_j)$ is empty then*
9. *Calculate $f(o_i)$ and $p_d(o_i)$*
10. *The ant a_j deposit object o_i in the box $r(a_j)$ with probability $p_d(o_i)$*
11. *endif*
12. *endif*
13. *Move the ant a_j on a near box unoccupied by another ant*
14. *endfor*
15. *endfor*
16. *return the location of objects on the grid*

It will be necessary to choose the size of the grid, the number of artificial ants, the number of iterations and the vision of each ant. The size of the grid must be large enough to contain all documents. These are deposited randomly on the grid. A heuristic is used to select a systematic size of the grid it states that the size must be equal to 2 times the number of documents.

The parameters used in our application have the following values: $c_1 = 0,1$, $c_2 = 0,15$, $s = 3$, $\alpha = 0.5$, $T_{max} = 10^6$, a grid (50x50), each ant has perimeter detection (vision) (3x3) and the preferred number of ants is equal to ten (10).

Again in this approach, we calculated learning rate, we obtained the results presented in the Table 4.

Table 4. Number of classes and learning rate according to the 3 similarity measurements with the ant-based clustering method.

	Cosine	Euclidean	Manhattan
Number of classes	29	35	36
Maximal Learning Rate (%)	9,65	8,47	7,75

These results seem less good compared to clustering methods seen above. The best values for this method are obtained with the cosine distance.

4. Evaluation

The evaluation of the relevance of the classes formed remains an open problem. The difficulty mainly comes from the fact that this evaluation is subjective by nature because there are often various possible relevant groupings for the same data set. The four criteria most commonly used to evaluate an unsupervised classification of textual documents are:

- Ability to process very large volumes of unstructured data.
- Easy reading of results: the system must offer various modes of visualization of the results.
- The data must be as homogeneous as possible within each group, and the groups as distinct as possible. This amounts to choosing the best adapted similarity measure.
- A good representation unquestionably influences the clustering.

In our experiments, the clustering results of the different algorithms are evaluated and compared using the f-measure which make use of the known classes for each document.

This measure is based on two concepts: recall and precision:

$$Precision(i,k) = \frac{Nik}{Nk} \quad (7)$$

$$Recall(i,k) = \frac{Nik}{NCi} \quad (8)$$

where N is the total number of documents, i is the number of classes (predefined), K is the number of clusters in unsupervised classification, N_{Ci} is the number of documents of class i, N_K is the number of documents of cluster C_K, N_{ik} is the number of documents of class i in the cluster C_K. F-measure $F(P)$ is calculated as follows:

$$F(P) = \sum \frac{NCi}{N} Max_{k=1}^{K} \frac{(1+\beta) \times Recall(i,k) \times Precision(i,k)}{\beta \times Recall(i,k) + Precision(i,k)} \quad (9)$$

Typically $\beta = 1$. The partition P - considered as most relevant and which best corresponds to the awaited external solution - is that which maximizes the associated F-measure.

Table 5 gives the values of the f-measure obtained for each approach.

Table 5. Comparison of F-measure values obtained by the ascending hierarchical clustering method, the Kohonen self-organizing maps method and the Ant-based method (for the 3 similarity measurements).

	AHC			Ant	SOM
	Ward	Average	Compl		
Cosine	0.5653	0.5319	0.5409	0.4044	0.6250
Euclidean	0.6194	0.5925	0.5506	0.3645	0.2550
Manhattan	0.4934	0.5687	0.5582	0.3155	0,2495

The F-Measure reveals that best performances are obtained with the SOM-based method using the cosine distance. Ascending hierarchical clustering method gives satisfactory results especially with the Euclidean distance, but the ant-based method gives poor results especially with the Manhattan distance, this proves that the ant-based classical (LF) method perform very badly on this kind of data.

5. Conclusions

In this paper we have presented three text clustering methods and all its stages: representation of texts, choice of a metric and choice of the clustering algorithm. It should be noted that we used an original method for the representation of texts: a concept-based approach. The results we obtained during this work are satisfactory, given the complexity of data processed.

We realized that the choice of a similarity measure is important in the process of clustering. Indeed, two different measures can lead to two different results of clustering. The method of ascending hierarchical clustering has the advantage of providing flexibility regarding the level of granularity, a facility to handle any form of similarity or distance, applicability to any type of attribute and is easy to implement. Its drawback is that it is very costly in CPU from the moment where objects are compared pairwise at each step. Another disadvantage is the difficulty of choosing the right place to cut the dendrogram and for decide of the proper number of clusters found. The approach based on ant populations is certainly interesting, especially when viewing the results. The main criticisms concern the computation time relatively large ($T_{max} = 10^6$), the interpretation of results which becomes hazardous with the high dimensionality since the border between two groups of objects can be reduced to an empty box while these two groups may represent two very different objects; and the choice of parameters which is difficult and very complex. We finally conclude that the self-organizing maps of Kohonen method, in addition to its

simplicity to implement, has the advantage of representing large data without the need to establish a priori the number of clusters. It also has the advantage of ensuring that similar data will be projected onto nearby positions in an output representation space. It appears that the application of this simple algorithm, but effective, using the cosine distance, and representation based on concepts, provides the best results. But its disadvantages still, firstly, the difficulty of choosing the size of the neighbourhood and on the other hand, the convergence depends on the order in which entries are presented.

Finally, the choice of the clustering method should be done depending on the desired result and therefore the exploitation of this result.

References

[1] Aas K. and Eikvil L., "Text Categorization: A Survey," *Technical Report,* 1999.

[2] Amine A., Elberrichi Z., Simonet M., Bellatreche L., and Malki M., *SOM-Based Clustering of Textual Documents Using WordNet*, New Jersey Institute of Technology, 2008.

[3] Amine A., Elberrichi Z., Simonet M., Bellatreche L., and Malki M., *SOM pour la Classification Automatique Non Supervisée de Documents Textuels Basés sur WordNet*, Cépaduès-Éditions, 2008.

[4] Berkhin P., "Survey of Clustering Data Mining Techniques," *Technical Report*, 2002.

[5] Elberrichi Z., "Text Mining Using n-Grams," *in Proceedings of CIIA'06*," Algeria, pp. 15-16, 2006.

[6] Fukumoto F. and Suzuki Y., "Learning Lexical Representation for Text Categorization," *in Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*, USA, pp. 357-370, 2001.

[7] Gonzalo J., Verdejo F., Chugur I., and Cigarran J., "Indexing with WordNet Synsets Can Improve Text Retrieval," *in Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Spain, pp. 237-251, 1998.

[8] Goss S. and Deneubourg J., "Harvesting by a Group of Robots," *in Proceedings of the First European Conference on Arti_Cial Life*, France, pp. 195-204, 1991.

[9] Jones W. and Furnas G., "Pictures of Relevance: A Geometric Analysis of Similarity Measures," *Computer Journal of the American Society for Information Science*, vol. 38, no. 6, pp. 420-442, 1987.

[10] Kohonen T., "Self-Organized Formation of Topologically Correct Feature Maps," *Computer Journal of Biological Cybernetics*, vol. 43, no. 1, pp. 59-69, 1982.

[11] Kohonen T., Kaski S., Lagus K., Salojarvi J., Honkela J., Paatero V., and Saarela A., "Self Organization of a Massive Document Collection," *Computer Journal of IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 340-347, 2000.

[12] De Loupy C., "L'apport de Connaissances Linguistiques en Recherche Documentaire," *in Proceedings of Traitement Automatique des Langues Naturelles*, France, pp. 297-306, 2001.

[13] Lumer D. and Faieta B., "Diversity and Adaptation in Populations of Clustering Ants," *in Proceedings of the 3rd International Conference on Simulation of Adaptive Behavior*, USA, pp. 501-508, 1994.

[14] Mihalcea R. and Moldovan D., "Semantic Indexing Using WordNet Senses," *in Proceedings of ACL Workshop on IR and NLP*, USA, pp. 281-296, 2000.

[15] Miller E., Shen D., Liu J., Nicholas C., "Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System," *Computer Journal of Digital Information*, vol. 1, no. 5, pp. 257-265, 1999.

[16] Miller A., "WordNet: An On-Line Lexicaldatabase," *International Journal of Lexicography*, vol. 3, no. 4, pp. 265-277, 1990.

[17] Rahmoun A. and Elberrichi Z., "Experimenting N-Grams in Text Categorization," *International Arab Journal of Information Technology*, vol. 4, no. 4, pp. 377-385, 2007.

[18] Sahami M., "Using Machine Learning to Improve Information Access," *PhD Thesis*, Stanford University, 1999.

[19] Salton G. and McGill M., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[20] Schütze H., Hull D., and Pedersen A., "A Comparison of Classifiers and Document Representations for the Routing Problem," *in Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, New York, pp. 229-237, 1995.

[21] Sebastiani F., "Machine Learning in Automated Text Categorization," *Computer Journal of ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

[22] Sneath A. and Sokal R., *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, Freeman and Company, 1973.

[23] Wang Y., "Incorporating Semantic and Syntactic Information Into Document Representation for Document Clustering," *Guidelines for Dissertations and Thesis,* Faculty of Mississippi State University, 2002.

Abdelmalek Amine received an engineering degree in computer science from the Computer Science Department of Djillali Liabes University of Sidi-Belabbes-Algeria, the Magister diploma in computational science from the UTMS University of Saida-Algeria, and PhD from Djillali Liabes University.

Zakaria Elberrichi received his Master degree in computer science from the California State University, in addition to PGCert in higher education, and received his PhD in computer science from the university Djillali Liabes, Sidi-Belabbes, Algeria.

Michel Simonet received his PhD from Joseph Fourier University of Grenoble, France. He is the head of the knowledge base and database team of the TIMC laboratory at the Joseph Fourier University of Grenoble, France.