Evaluation and Comparison of Concept Based and N-Grams Based Text Clustering Using SOM

ABDELMALEK AMINE[1,2], ZAKARIA ELBERRICHI[1], MICHEL SIMONET[3], MIMOUN MALKI[1]

[1] EEDIS Laboratory, Department of computer science, UDL University, Sidi Belabbes – Algeria
[2] Department of computer science, University Center Taher Moulay, Saida - Algeria
{amine_abd1, elberrichi, malki_m}@univ-sba.dz
[3] TIMC-IMAG Laboratory
IN3S, Joseph Fourier University, Grenoble - France
michel.simonet@imag.fr

Abstract. 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 and a fundamental task in document processing. This paper deals with unsupervised classification of textual documents also called text clustering using Self-Organizing Maps of Kohonen in two new situations: a conceptual representation of texts and a representation based on n-grams, instead of a representation based on words. The effects of these combinations are examined in several experiments using 4 measurements of similarity. The Reuters-21578 corpus is used for evaluation. The evaluation was done by using the F-measure and the entropy.

Keywords: Text clustering, Self-Organizing Maps of Kohonen, n-grams, concept, similarity, Reuters21578.

(Received November 09, 2007 / Accepted February 02, 2008)

1. Introduction

With the great and rapidly growing number of documents available in digital form (Internet, library, CD-Rom…), the categorization or automatic classification of texts has become a significant research field.

The categorization or 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. In this paper, we first study the method of Kohonen self-organizing maps for the classification of textual documents based on n-grams representation. The same method using WordNet synsets as terms for the representation of textual documents is then studied and compared with the former approach.

Section II will introduce different possible ways of representing a text, explain similarity measurements and will review the best known clustering algorithms. Section III is devoted to the presentation of the model of Kohonen, and in section IV we describe the proposed approaches in all their stages. Finally section V will conclude the article.

2. State of the Art

Implementing these methods initially consists in choosing a way of representing the documents [20], 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 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:

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{|T|j}) \tag{1}$$

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						
d_1	w_{11}	w_{21}	w_{31}	…	w_{j1}	…	w_{n1}
d_2	w_2	w_{22}	w_{32}	…	w_{j2}	…	w_{n2}
…	…	…	…	…	…	…	…
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” [18], [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 favouring 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 [19] 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 [17] while lemmatization replaces a term by a conventional standard form, e.g., infinitive form for verbs and singular for nouns [11]. This prevents that each inflection or form of a word should be regarded as a different descriptor and consequently create 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 [13]. The window is moved by a character at a time and the number of occurrences of each n-gram is counted [5],[16].
- The conceptual 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 an 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 [20] 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.

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

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.

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

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 [8]. Among these measurements we can quote:

- The cosine distance:

$$Cos\left(d_i, d_j\right) = \frac{\sum_{t_k} \left[TF \times IDF\left(t_k, d_i\right)\right] \bullet \left\lfloor TF \times IDF\left(t_k, d_j\right)\right]}{\left\|d_i\right\|^2 \bullet \left\|d_j\right\|^2} \qquad (4)$$

- The Euclidean distance:

$$Euclidean(d_i, d_j) = \sqrt{\sum_{1}^{n} (\ w_{ki} - w_{kj})^2} \qquad (5)$$

- The Manhattan distance:

$$Manhattan(\ d_i, d_j) = \sum_{1}^{n} \left| w_{ki} - w_{k} \right| \qquad (6)$$

2.3. Algorithms for the 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] and [21], 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,
- 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).

3. Self-Organizing Maps of Kohonen (SOM) for the Clustering of Textual Documents

SOM (Self-Organizing Maps of Kohonen) is an unsupervised learning method which is based on the principle of competition according to an iterative process of updates [2], [3].

The Kohonen model or network proposed by Tuevo Kohonen [9] 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 center 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.

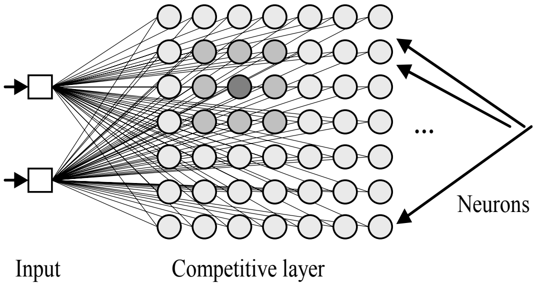

Figure 1. Kohonen network architecture

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 [10].

However, the combination between SOMs and the conceptual representation of texts on the one hand, SOMs and representation based on the n-grams on the other hand were not extensively studied. In the following section, we shall try out these combinations, evaluate and compare them.

4. Experiments

4.1. 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.

4.2. Approach Based on N-grams for Document representation

In this approach and at the first stage, we eliminate from the texts in an automatic way the punctuation marks such as: point, comma, semicolon, exclamation and question marks… etc., because these characters have no influence on the classification results.

Then, we count the frequencies of the n-grams found. For a given document, the set of n-grams (in general *n = {2, 3, 4, 5}*) is the result obtained by moving a window of *dimension n* throughout the text. We move this window one character at a time and at each step we take a "snapshot". Together, these snapshots constitute the set of n-grams associated with the document.

For example, to generate all the 5-grams in the sentence: "*the_fisherman_fish*", we obtain:

[the_f=1, he_fi=1, e_fis=1, _fish=2, fishe=1, isher= 1, sherm = 1, herma= 1, erman = 1, rman_ = 1, man_f=1, an_fi=1, n_fis=1]

The character _ is used to represent a blank character.

With the use of n-grams for the representation of textual documents, we do not need to make a linguistic pretreatment, i.e., we do not need to apply lemmatization and stemming techniques or to eliminate the stop words. This method offers other advantages such as capturing the roots of the most frequent words, operating independently from the languages, and being tolerant of errors due to spelling mistakes and the scanning of documents.

To calculate the weight (frequency) of each n-gram in a text, we use *TFxIDF* function. Each document will be thus represented by its standardized vector of n-grams.

4.3. The Conceptual Approach for Document Representation

4.3.1. Wordnet and the Classification of Texts

WordNet [14] is an 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], [12].

4.3.2. Representation of Documents Based on Wordnet

In this approach, we propose a representation which replaces terms by their associated concepts in Wordnet. In the pretreatment phase, we eliminate from texts punctuation marks and stop words such as: *are*, *that*, *what*, *do*.

This representation requires two more stages: 1) the "mapping" of terms into concepts and the choice of the "merging" strategy, and 2) the application of a disambiguation strategy.

The first stage (see example 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"), 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).

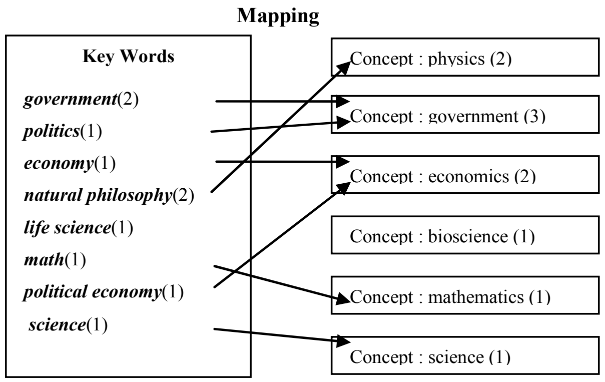

Figure 2. Example of mapping words in concepts

It is clear that the assignment of terms to concepts in an 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:

$$cf(d,c) = tf\{d, \{t \in T \mid \ first(ref_c(t)) = c\}\} \qquad (7)$$

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.

4.4. Configuration

Our system was developed with Borland JBuilder version7 and a Windows XP platform, on a machine with a processor INTEL Pentium 4 (2,66 GHz) with 256 Mo RAM. We used a 7x7 Kohonen map and we tested for each approach four similarity measurements: the cosine distance, the Euclidean distance, the Squared Euclidean distance and the Manhattan distance.

4.5. Results

It is necessary to state here that our objective is to demonstrate that it is possible to extend the use of WordNet to unsupervised text classification and to assess the efficiency of this approach when we add a lexical dimension compared with a statistical approach such as that based on n-grams.

For each similarity measurement quoted above, for the method based on n-grams and for the method based on Wordnet, we have calculated the number of classes, the time and the rate of training. We have obtained the following results (as shown in: Table.2, Table.3, Figure.3, Figure.4, Figure.5 and Figure.6):

Table 2: Number of classes, learning time and learning rate according to 4 measurements of similarity (for the n-grams, N = {2, 3, 4,5})

	Cosine	Euclidean	Euclidean2	Manhattan
n	**Number of classes**			
2	36	28	26	17
3	33	26	24	19
4	15	29	32	34
5	29	30	24	28
	Learning time (in S)			
2	62	40	55	70
3	94	68	98	106
4	83	64	86	92
5	53	37	49	52
	Maximal learning rate (%)			
2	5,81	5,81	13,71	26,1
3	5,62	9,71	16,09	12,09
4	17,05	7,14	8,47	7,71
5	6,95	5,61	5,61	5,9

Table 3: Number of classes, learning time and learning rate according to 4 measurements of similarity for the approach based on Wordnet

	Cosine	Euclidean	Euclidean2	Manhattan
Number of classes	22	27	27	27
Learning Time (in S)	57	80	59	80
Maximal learning Rate (%)	14,09	13,08	9,01	7,71

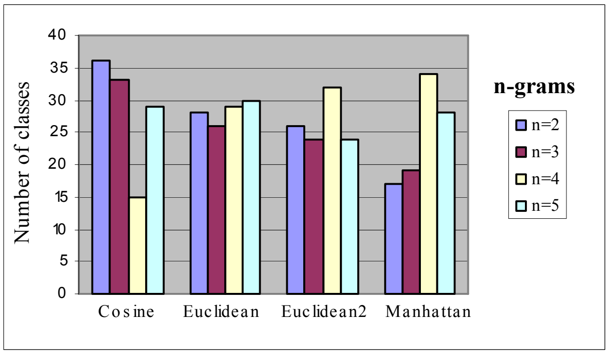

Figure 3: Number of classes according to 4 measurements of similarity (for the n-grams, N = {2, 3, 4, 5})

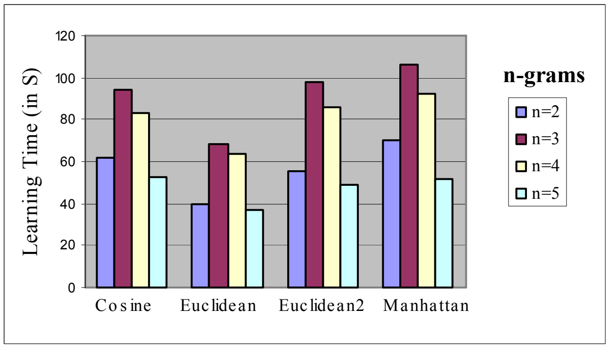

Figure 4: Learning time according to 4 measurements of similarity (for the n-grams, N = {2, 3, 4, 5})

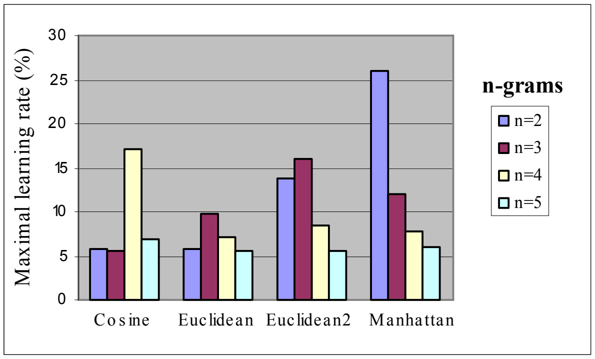

Figure 5: Learning rate according to 4 measurements of similarity (for the n-grams, N = {2, 3, 4, 5})

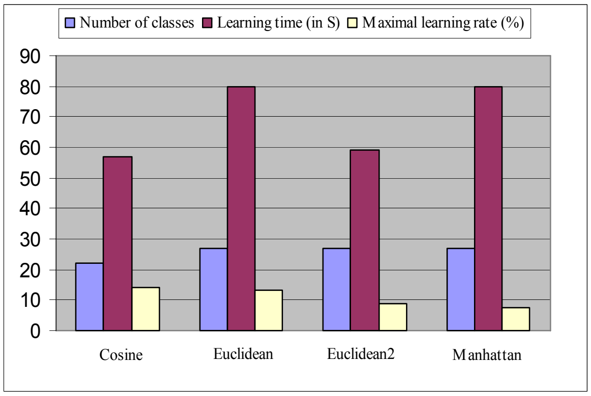

Figure 6: Number of classes, learning time and learning rate according to 4 measurements of similarity for the approach based on Wordnet.

It should be noted that in spite of the good results obtained by the n-gram method, particularly for n=3 and n=4, the results obtained by the conceptual method are better, especially when using the cosine distance.

4.6. 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. In our approach the Kohonen map is a good example of visualization,
- 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.

If we wish to evaluate the quality of unsupervised classification with respect to the known classes for each document, two measurements of external quality are classically used: F-measure and entropy. These two measurements are based on two concepts: recall and precision:

$$Pr\ ecision\ (i,k\ ) = \frac{Nik}{Nk} \qquad (7)$$

$$Re\ call\ (\ i,k\ ) = \frac{Nik}{NCi} \qquad (8)$$

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) and Entropy are calculated as follows:

$$F(P) = \sum \frac{NCi}{N} Max_{k=1}^{K} \frac{(1+\beta)\times Re call(i,k)\times Pr ecision(i,k)}{\beta\times Re call(i,k)+Pr ecision(i,k)} \qquad (9)$$

$$E(P) = \sum_{k=1}^{K} \frac{Nk}{N} \times \left(-\sum_{i} Pr ecision(i,k) \times log\, Pr ecision(i,k) \right) \qquad (10)$$

Typically β = 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 or minimizes the associated entropy.

Table 4: Comparison of the F-measure values obtained by the n-grams and Wordnet-based approaches (for 4 measurements of similarity)

	Method				
Distance	Wordnet	n-grams			
		2	***3***	***4***	***5***
Cosine	**0.6250**	0.2304	0.2370	0.4690	0.2348
Euclidean	0.2550	0.1807	0.2510	0.2180	0.1780
Euclidean2	0.3607	0.2466	0.3027	0.2735	0.2203
Manhattan	0.2495	0.2738	0.3274	0.2120	0.1970

Table 5: Comparison of the values of entropy obtained by the n-grams and Wordnet-based approaches (for 4 measurements of similarity)

	Method				
Distance	Wordnet	n-grams			
		2	***3***	***4***	***5***
Cosine	**0.3750**	0.7412	0.7049	0.4305	0.7210
Euclidean	0.7213	0.7956	0.7453	0.7683	0.8083
Euclidean2	0.6216	0.7297	0.6436	0.6928	0.7565
Manhattan	0.7368	0.7025	0.6650	0.7843	0.7673

These results corroborate the conclusions drawn from this study. The best performance for the n-grams method is obtained with the cosine distance for n=4; we note that, for this approach, performance improves by increasing the value of n and start degrading from n = 5. We confirm previous works [16] which showed that the 4-gram approach produces better results than 3-gram where fewer features are generated.

The best performance in general is obtained by the conceptual approach (Wordnet) and the cosine distance.

Our experiments on actual textual data sets demonstrate that clustering with concepts yields a better performance.

5. Conclusion

In this paper we have presented the concept of unsupervised automatic classification of texts and all its stages: representation, choice of a metric and choice of the method. It should be noted that the choice of the methods of representation deserves as much consideration as the methods of classification; this is because a good classification requires a good representation [19].

We have proposed two new approaches for the SOM-based unsupervised classification of texts, one based on the use of WordNet and the other on the use of n-grams. The results obtained show that in spite of the good results obtained by the n-grams method, adding lexical knowledge in the representation makes it possible to build a better classification. This is a particularly interesting path for possible future work.

We first plan to use other strategies for disambiguation and analyze their influence on classification. Then, we plan to use other conceptual approaches of syntactic references, aiming at the classification of multilingual texts by SOM-based methods [15].

References

[1] Aas, K., Eikvil, L.: Text categorization: a survey. *Technical report, Norwegian Computing Center*, 1999.

[2] Amine, A., Elberrichi, Z., Bellatreche, L., Simonet, M., Malki, M.: Concept-based Clustering of Textual Documents Using SOM. *In: proceedings of the 6th ACS/IEEE International Conference on Computer Systems and Applications AICCSA-08,* 2008.

[3] Amine, A., Elberrichi, Z., Simonet, M., Malki, M.: SOM pour la Classification Automatique Non supervisée de Documents Textuels basés sur Wordnet. *EGC'2008,* INRIA Sophia Antipolis, France, 2008.

[4] Berkhin, P.: *Survey of Clustering Data Mining Techniques*. Accrue Software, San Jose CA, 2002.

[5] Elberrichi, Z.: Text mining using n-grams. *In: Proceedings of CIIA'06*, Saida, Algeria, 2006.

[6] Fukumoto, F., Suzuki, Y.: Learning lexical representation for text categorization. *In: Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, 2001.

[7] Gonzalo, J., Verdejo, F., Chugur, I., 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*, 1998.

[8] Jones, W. and Furnas, G.: Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6):420-442, 1987.

[9] Kohonen, T.: *Self-organized formation of topologically correct feature maps*. Biological Cybernetics, 43, 59-69, 1982.

[10] Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., Saarela, A.: *Self organization of a massive document collection*. IEEE Transactions on Neural Networks. 11(3), May, 2000.

[11] de Loupy, C: L'apport de connaissances linguistiques en recherche documentaire. *In: TALN'01*, 2001.

[12] Mihalcea, R., Moldovan, D.: Semantic indexing using WordNet senses. *In: Proceedings of ACL Workshop on IR and NLP*, 2000.

[13] Miller, E., Shen, D., Liu, J., Nicholas, C: Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System. *Journal of Digital Information*, 1(5), 1999.

[14] Miller, G.A.: Wordnet: An on-line lexicaldatabase. In Special. *Issue of International Journal of Lexicography*, Vol 3, No.4, Chongqing, China, 1990.

[15] Pham, M.H., Bernhard, D., Diallo, G., Messai, R., Simonet, M.: SOM-based Clustering of Multilingual Documents Using Ontology. *In: Data Mining with Ontologies: Implementations, Findings and Idea Frameworks*. Nigro, H.O., Císaro, S.G., Xodo, D. (ed.), Group Inc, 2007.

[16] Rahmoun, A., Elberrichi, Z.: Experimenting N-Grams in Text Categorization. *Issue of International Arab Journal of Information Technology*, Vol 4, N°.4, pp. 377-385, 2007.

[17] Sahami, M.: Using Machine Learning to Improve Information Access. PhD thesis, Computer Science Department, Stanford University, 1999.

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

[19] Schütze, H., Hull, D. A., Pedersen, J.O.: A comparison of classifiers and document representations for the routing problem. *In: Fox, E. A., Ingwersen, P., and Fidel, R., editors, Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, Seattle, US. ACM Press, New York, 229–237, 1995.

[20] Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47, 2002.

[21] Wang, Y.: Incorporating semantic and syntactic information into document representation for document clustering. A dissertation submitted to the Faculty of Mississippi State University August, 2002.