SOM-BASED CLUSTERING OF TEXTUAL DOCUMENTS USING WORDNET

Abdelmalek AMINE[*,**], Zakaria ELBERRICHI[*], Michel SIMONET[***], Mimoun MALKI[*]

[*]Laboratoire EEDIS, Département d'informatique
Faculté de l'ingénieur. Université Djillali Liabes
Sidi bel Abbes – Algérie
[**]Département d'informatique, Centre Universitaire Taher Moulay de Saida-Algérie
amine_abd1@yahoo.fr, elberrichi@gmail.com, malki_m@yahoo.com

[***]Laboratoire TIMC-IMAG
INstitut de l'INgénierie et de l'INformation de santé (IN3S)
Faculté de médecine. Université Joseph Fourier
Grenoble - France
michel.simonet@imag.fr

ABSTRACT

The classification of textual documents has been the subject of many studies. Technologies like the web and numerical libraries facilitated the exponential growth of available documentation. The classification of textual documents is very important since it allows the users to effectively and quickly fly over and understand better the contents of large corpora. Most classification approaches use the supervised method of training, more suitable with small corpora and when human experts are available to generate the best classes of data for the training phase, which is not always feasible.
The non supervised classification or "clustering" methods make emerge latent (hidden) classes automatically with minimum human intervention, There are many, and the SOM (self Organized Maps) by kohonen is one the algorithms for non-supervised classification that gather a certain number of similar objects in groups without a priori knowledge. In this paper, we propose non supervised classification of textual document using the SOM of kohonen and based on a conceptual representation of the texts.

Keywords: *Non-supervised classification, descriptors, Similarity, self-organizing maps of Kohonen, Wordnet*

1. INTRODUCTION

With the abundance and the fast growth of the number of documents present in numerical form (Internet, numerical libraries, CD-ROM…), the categorization or automatic text classification became an important field of research.

The categorization or automatic text classification is the action to distribute by categories or classes a set of documents according to some common characteristics.
In English terminology the terms "categorization" or "classification" are used when it is about assigning a document to a class (classes being known in advance), in this case we are within the framework of supervised learning. And the term "clustering" (non supervised classification) when it is about the creation of classes or groups (clusters) of a certain number of similar objects without a priori knowledge, we are then within the framework of the non supervised learning.

Non supervised classification or "clustering" is automatic; it makes emerge latent (hidden) classes, not labeled. The classes are distinct and are to be discovered automatically. It is sometimes possible to fix their number.

A great number of methods of clustering were applied to the textual documents. In this article, we propose the method of the self-organizing maps of Kohonen for the classification of the textual documents based on the WordNet synsets as the terms for the textual documents representation.

Section II will introduce different manners to represent a text, explain similarity measurements and review the most known "clustering" algorithms, section III is devoted to the presentation of WordNet, in section IV, we describe the approach suggested with all its stages and the results, finally the section V will conclude the article.

2. STATE OF THE ART

To implement any method on textual document we initially, need to represent the documents [1], because there is currently no method of learning able to directly process not structured data (texts). In the second time it is necessary to choose a similarity measurement, and lastly to choose a clustering algorithm which one will develop starting from the descriptors and of metric chosen.

2.1. REPRESENTATION OF THE TEXTUAL DOCUMENTS

To implement any method classification it is initially necessary to transform the digitized texts in an economic and significant way so that they are analyzable. The vectorial 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 d_j will be transformed into a vector:

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

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

Table 1 : Matrix documents-terms

Docs	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 vectorial model is called the "bag of words" [2], [3], it consists in transforming the texts into vectors whose each component represents a word. This representation of the texts excludes any grammatical analysis and any concept of distance between the words and is destructive to the syntax of the texts while making them comprehensible for the machine.

• Another representation called "bag of sentences" carries out a selection of the sentences (sequences of words in the text, and not the lexeme "sentences" such as we usually understand it), by privileging those which are likely to carry an important meaning. Logically, such a representation must obtain better results that those obtained by the representation "bag of words", but the experiments [4] show that if the semantic qualities are preserved; the statistical qualities are largely degraded.

• Another method for the representation of the texts calls upon the techniques of lemmatization and stemming which consists in seeking the lexical roots for one [5] and replacing the verbs by their infinitive form and the nouns by their form in the singular [6] for the other in order to prevent that each inflection or form of a word is not regarded as a different descriptor and thus one more dimension.

• Another method of representation which has several advantages is the method based on the "n-gram" where a "n-gram" is a sequence of N consecutive characters. The whole set of the "n-gram" (generally N 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 [7]. The window is moved by a character at a time, the number of occurrences of different "n-gram" is then counted [8].

• The conceptual representation also called representation based on an ontology, is also based on the vectorial formalism to represent the documents but it remains basically different from the methods of representation presented before. The characteristic of this approach lies in the fact that the elements of the vector space are not here associated with the terms only but with the concepts also. This is possible by adding an additional phase, the phase of mapping terms into concepts.

There are various methods to calculate the weight w_{kj}, knowing that, for each term, it is possible to know on the one hand its frequency of appearance in the corpus but also the number of documents which contain this term. The majority of the approaches [1] are centered on a vectorial representation of the texts of the type TF-IDF.

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. 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 (product) in order to give the more strong weight as the term often appears in the document and seldom in the complete corpus.

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

Where Occ (t_k, d_j) is the number of occurrences of the tk term in the document d_j, Nbr_doc is the total number of documents of the corpus and Nbr_doc (t_k) is the number of documents of this set in which appears at least once the term t_k.

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

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

2.2. SIMILARITY MEASURE

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

- Cosinus distance:

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

- Euclidian distance:

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

- Manhattan distance:

$$\text{Manhattan}(\, d_i, d_j) = \sum_1^n \left| w_{ki} - w_k \right| \qquad (6)$$

2.3. ALGORITHMS FOR CLUSTERING OF TEXTUAL DOCUMENTS

Non supervised classification or "clustering" is one of the fundamental data mining techniques to cluster structured or unstructured data. Several methods were proposed, according to [10] and [11], 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 documents or starting from the whole set of all the documents or the corpus.

- If we start with the documents, each document is initially put in a class which contains only one. Then, the two most similar classes are amalgamated to form only one. This process is repeated until a certain condition of stop is satisfied. This method is called "agglomeration of similar groups" or ascending hierarchical "Clustering".

- On the other hand, if we start with the whole set of documents or the corpora, the method is called "division of dissimilar groups" or hierarchical "Clustering" going down. At the beginning of this process, there is only one class of all the documents. The class is divided into two subclasses at a time at the following iteration. The process continues until the condition of stop is satisfied. The similarity between two documents is based on the distance between documents.

• Partitioning Methods: Also called flat "clustering". The known methods are the K-medoid method, the dynamic clouds method and the K-means method or mobile centers. For the K-means method for example the number of classes is preset. A document is put in 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 is a question of grouping the objects as long as the vicinity density exceeds a certain limit. The groups or classes are dense areas separated by not very 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 proceeds in 2 stages:
- To choose a dense point randomly
- All the points which are attainable starting from this point, according to the threshold of density, form a group or a class.

• Grid based methods: It is a division the data space in multidimensional cells forming a grid (the data are represented like points in the grid) and grouping the close cells in term of distance. The classes are built by assembling the cells containing sufficient data (dense). Several levels of grids are used, with an increasingly high resolution.

• Models based Methods: One of the models based methods is the conceptual approach. In this approach we have a conceptual hierarchy inherent to the data where the concept is the couple (intention, extension) knowing that the intention 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 called also self-organizing maps (SOM) of Kohonen. It is a neuronal method interesting because ordering topologically the classes obtained in the form of a map, generally on a plan (two-dimensional).

3. WORDNET AND TEXTS CLASSIFICATION

WordNet [12] is an ontology of cross lexical references whose design was inspired by the current theories of human linguistic memory. The English names, verbs, adjectives and adverbs are organized in sets of synonyms (synsets), representing the subjacent lexical concept. Relations connect the sets of synonyms between them.

WordNet covers the large majority of the names, verbs, adjectives and adverbs of the English language. The last 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. The synsets are bound by the relations such as specific/generic or hypernym /hyponym (is-a), and meronym/holonym (part-whole).
We can define in a brief way the principal semantic relations supported by WordNet as follows [wikipedia]:

• Synonymy: the synset (synonym set), represents a set of words which are interchangeable in a specific context.

• Hypernymy: it is the generic term used to indicate a class including instances of more specific classes. A hypernym is the opposite of a hyponym. For example, *plant* is hypernymic to *flower* whereas *tulip* is hyponymic to *flower*

• Holonymy: (in Greek *holon* = whole and *onoma* = name) is a semantic relation. Holonymy defines the relationship between a term denoting the whole and a term denoting a part of, or a member of, the whole. That is, 'X' is a holonym of 'Y' if Ys are parts of Xs, or 'X' is a holonym of 'Y' if Ys are members of Xs. For example, 'tree' is a holonym of 'bark', of 'trunk' and of 'limb.'

• Holonymy is the opposite of meronymy.

• Méronymy (from the Greek words *meros* = part and *onoma* = name) is a semantic relation concept used in linguistics. A meronym denotes a constituent part of, or a member of something. That is, *X* is a meronym of *Y* if *X*s are parts of *Y*(s), or X is a meronym of *Y* if *X*s are members of *Y*(s). For example, 'finger' is a meronym of 'hand' because a finger is part of a hand. Meronymy is the opposite of holonymy.

• In addition to these relations, WordNet has some other relations which are however less used in practice. It is the case of the relation antonym to express the opposed

meanings for the synsets, or the relation troponym for the verbs. A word that denotes a manner of doing something *'march' is a troponym of 'walk'*.

The broad scale of WordNet and its free availability makes it used in many text classification methods and in information retrieval (IR) too. Some work in which the synsets of WordNet were used as index terms had very promising results.

4. APPROACH PROPOSED

The approach suggested is tested on a corpus obtained by mixing the first 50 documents of each the 22 categories of the Reuters21578 corpus. In a first phase, we eliminate from the corpus spaces, punctuations as well as the stop words.

4.1. A CONCEPTUAL APPROACH FOR DATA REPRESENTATION

We propose a representation which replaces the terms by their associated concepts in the Wordnet ontology. This representation requires two stages: the first being the "mapping" of the terms into concepts and the choice of the strategy of "merging", the second is the application of a strategy of disambiguation.

For the first stage, for the example indicated in the Figure 1, it is about mapping the two terms **government** and **politics** in the concept **government** and the frequencies of these two words will be added.
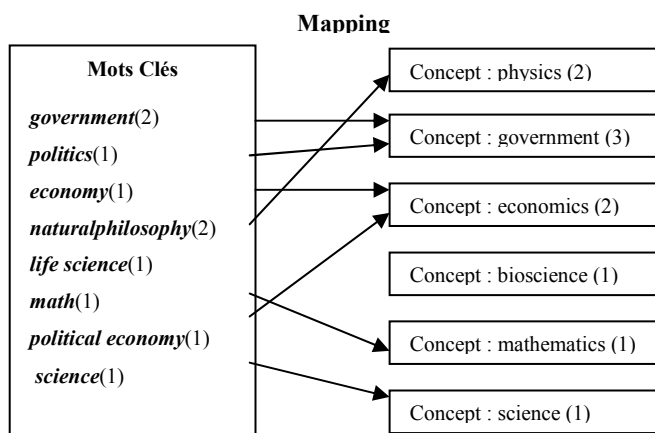
Mapping

Figure 1: Example of mapping words into concepts

Then, among the three possible strategies of "merging" offered by the conceptual approach "To add Concept" (adds terms and the found concepts), "To replace the terms by concepts" (replace all the terms by their found concepts in wordnet) and "Concept only" (use only the concepts), we opted for the last strategy "Concept only" where we replace the vector of the terms by the vector of the concepts by excluding all the terms from the new representation including the terms which do not appear Wordnet.

It is completely clear that assignments of the terms to the concepts in an ontology is ambiguous. For this reason to add or replace terms by concepts can cause a loss of information. Indeed, the choice of the more

appropriate concept to a term can influence the performances of a classification.

In our approach we used a simple method for the disambiguation: strategy called the "First concept". Wordnet gives for each term an ordered list of concepts according to certain criterion. This strategy of disambiguation consists in taking only the first concept of the list as the most suitable concept. The frequencies of concepts will thus be calculated as follows:

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

For the calculation of the weights (frequencies), we used the TFIDF function, knowing that the terms are synsets and the documents vectors are vectors of concepts which will be standardized.

4.2. SELF-ORGANIZING MAPS OF KOHONEN (SOM) FOR CLUSTERING OF TEXTUAL DOCUMENTS

The SOM is a non supervised learning method which is based on the principle of the competition according to an iterative process of updates [13] [14].

The Kohonen model or network proposed by Tuevo Kohonen [15] 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 affected which represents the center of the class.

• Of an output layer (or competition). The neurons of this layer enter in competition to be activated according to a distance chosen, only one neuron is activated (winner-takes-all neuron) following the competition.
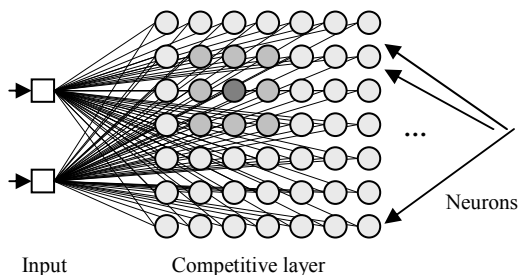

Figure 2: Kohonen network architecture

4.3. SIMILARITY MEASURE

We tested for various sizes of the map of Kohonen, and 4 similarity measurements: the cosine distance, the Euclidean distance, the squared Euclidean distance and the Manhattan distance.

4.4. EXPERIMENTAL RESULTS

It is necessary to specify here that our objective is to show that it is possible to extend the use of WordNet to the non supervised text classification.

For various sizes of the Map and each similarity measure quoted above we calculated the number of

classes, the time and the rate of learning. We obtained the following results:

Table 2: The Map size function of the classes number (for 4 similarity measures)

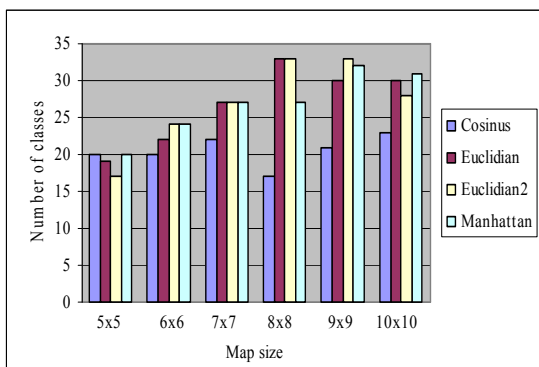| Map Size	Cosinus	Euclidian	Euclidian2	Manhattan
	Number of classes			
5x5	20	19	17	20
6x6	20	22	24	24
7x7	22	27	27	27
8x8	17	33	33	27
9x9	21	30	33	32
10x10	23	30	28	31

Figure 3: The Map size in function of the number of classes (for the 4 similarity measure)

Table 3: The Map size in function of the learning time (for the 4 similarity measures)

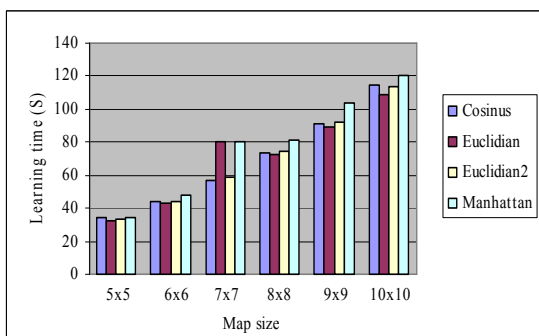| Map Size	Cosinus	Euclidian	Euclidian2	Manhattan
	Learning time (in S)			
5x5	34	32	33	34
6x6	44	43	44	48
7x7	57	80	59	80
8x8	73	72	74	81
9x9	91	89	92	104
10x10	115	109	114	120

Figure 4: The Map size in function of the learning time (for the 4 similarity measures)

Table 4: The Map size in function of the maximal learning rate (for the 4 similarity measures)

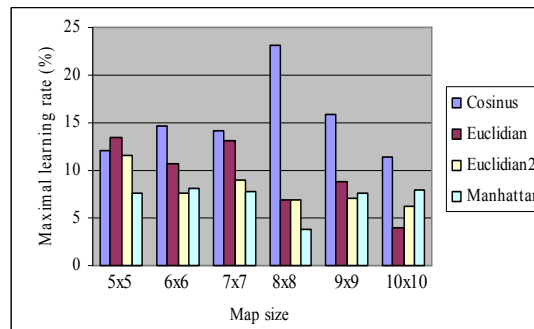| Map Size	Cosinus	Euclidian	Euclidian2	Manhattan
	The maximal learning rate (%)			
5x5	12	13,52	11,523	7,523
6x6	14,57	10,66	7,523	8,09
7x7	14,09	13,08	9,01	7,71
8x8	23,142	6,85	6,87	3,833
9x9	15,9	8,83	7,04	7,61
10x10	11,33	3,93	6,23	7,9

Figure 5: The Map size in function of the maximal learning rate (for the 4 similarity measures)

In the first, the second and the third case: the best performances in general are obtained with the cosine distance.

The evaluation of the relevance of the formed classes remains an open problem. The difficulty comes mainly from the fact that this evaluation is subjective by nature because there are often various possible relevant regroupings for the same data file.

Nevertheless, there are in general 4 principal criteria to evaluate a clustering of textual documents:

• The capacity to treat very large volumes of not structured data.

• The interpretation of the results: the system must offer various modes of visualization of the results. In our approach the map of Kohonen is a good example of visualization.

• Each group must be most homogeneous possible, and the groups must be the most different possible between them. For that you have to choose the most suitable similarity measure.

• A good representation unquestionably influences the clustering.

5. CONCLUSION
In this article we presented the concept of non supervised automatic text classification and its stages: the representation, the choice of metric and the choice of the method. It should be noted that the representation method is as important as the classification method because a good classification requires a good representation [4]. We proposed a new approach for the non supervised text classification based on the use of WordNet. The results obtained are encouraging. We project in a first time to use other strategies of

disambiguation and to see their influences on classification. In the second time, to use other conceptual approaches for multilingual texts classification using the SOM [16].

REFERENCES

[1] Sebastiani, F.: "Machine learning in automated text categorization". *ACM Computing Surveys*, 34(1), 1–47, 2002. Available from World Wide Web: http://faure.iei.pi.cnr.it/~fabrizio/Publications/ACMCS02.pdf

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

[3] Aas, K., Eikvil, L.: "Text categorization: a survey". *Technical report, Norwegian Computing Center*, 1999. Available from World Wide Web: http://www.nr.no/documents/samba/research_areas/BAMG/Publications/ tm_survey.ps

[4] 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. Available from World Wide Web: ftp://parcftp.xerox.com/pub/qca/papers/sigir95.ps.gz

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

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

[7] 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.

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

[9] 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, November 1987.

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

[11] 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.

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

[13] Amine, A.: "Les cartes auto–organisatrices de Kohonen pour la classification non supervisée de documents textuels : état de l'art". *In : Proceedings of Jetic'07*, Bechar Algeria, April 2007.

[14] Amine, A.: "Classification non supervisée de documents textuels : état de l'art". *In : Proceedings of COSI'07*, Oran Algeria, June 2007.

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

[16] 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.