Representation of Textual Documents by the Approach Wordnet and N-grams for the Unsupervised Classification (Clustering) with 2D Cellular Automata: A Comparative Study

HAMOU Reda Mohamed (Corresponding author)
University Dr Tahar MOULAY of Saïda, EEDIS Lab, Computer Science Department, Algeria
E-mail: hamoureda@yahoo.fr

Ahmed LEHIRECHE
University Djillali Liabes of Sidi Bel Abbes, EEDIS Lab, Computer Science Department, Algeria
E-mail: elhir@yahoo.com.

LOKBANI Ahmed Chaouki
University Dr Tahar MOULAY of Saïda, EEDIS Lab, Computer Science Department, Algeria
E-mail: ahchlokbani@yahoo.fr

RAHMANI Mohamed
University Dr Tahar MOULAY of Saïda, EEDIS Lab, Computer Science Department, Algeria
E-mail: rahmanimed@yahoo.fr

Abstract

In this article we present a 2D cellular automaton (Class_AC) to solve a problem of text mining in the case of unsupervised classification (clustering). Before to experiment the cellular automaton, we vectorized our data indexing textual documents from the database REUTERS 21,578 by Wordnet approach and the representation of text documents by the method n-grams. Our work is to make a comparative study of two approaches to representation that is the conceptual approach (Wordnet) and the n-grams. Section 1 gives an introduction on the biomimétisme and text mining, Section 2 presents representation of texts based on Wordnet approach and the n grams, Section 3 describes the cellular automaton for clustering, Section 4 shows the experimentation and comparison results and finally Section 5 gives a conclusion and perspectives.

Keywords: Data classification, Cellular Automata, Biomimetic methods, Data mining, Clustering and segmentation, Unsupervised classification

1. Introduction

Biomimétic in a literary sense is the imitation of life. Biology has always been a source of inspiration for researchers in different fields. These have found an almost ideal in the observation of natural phenomena and their adaptation to solve problems. Among these models are the genetic algorithms, ant colonies, swarms particles, and clouds of flying insects [Nicolas Monmarché 2003] and of course cellular automata that we will detail in the next section. The first approaches mentioned methods are widely recognized and studied but cellular automata against methods are rarely used and in particular in the field of unsupervised classification. It has been our motivation for the use of this method in this field. This method is known by scientific community as a tool for implementing machinery and other (A **cellular automaton** (CA) is primarily a formal machine). We consider cellular automaton as a biomimetic method.

At all times, many researchers have been inspired by the model of the nature. Opportunistic man has always dreamed of flying like a bird or swim like a fish...

Leonardo de Vinci (1452 - 1519) was a universal genius. As both an artist, philosopher, scientist and it was also the first true biomimétique researcher. After studying the flight of birds, he constructed in 1505 of theft of equipment, helicopters and parachutes. Unfortunately, the society of the time was not yet ready and has prevented his ideas to be transformed into real products.

Since the 50th, biomimétic has been steadily progressing and is a major issue for current research.

Biomimétic is a scientific practice that tries to imitate or draw inspiration from natural systems. Examples of this area, we find among other things: imitating skin fishes used in cars or vehicles aerodynamics, or the algorithm inspired from ant colonies to find the shortest path in a graph…

The Text mining is the combination of techniques and methods for the automatic processing of textual data in natural language. It is a multidimensional analysis of textual data, which aims to analyze and discover knowledge and connections from the available documents. In text mining similarities are used to produce synthetic representations of large collection of documents. Text mining includes a series of steps to go documents to text, text to number, number to the analysis and analysis to decision-making.

State of the art

To implement classification methods, because there is currently no method of learning that can directly represent unstructured data (text), we should choice a mode of representation of documents [Sebastiani, F 2002]. Also, it is necessary to choose a similarity measure and an algorithm for unsupervised classification.

a- Textual representation

A document (text) di is represented by a numerical vector as follows: di = (V_{1i}, V_{2i}, ..., $V_{|T|i}$)

Where T is the set of terms (or descriptors) that appear at least once in the corpus.

(|T| is the vocabulary size), and V_{ki} is the weight (or frequency).

The simplest representation of textual documents is called a "bag of words" representation [Aas, K.1999], it is to transform texts in vectors where each element represents a word. This text representation excludes any form of grammatical analysis and any notion of distance between words.

Another representation, called "bag of phrases", provides a selection of phrases (sequences of words in the texts, not the lexeme "phrases"), by favouring those who are likely to carry a meaning. Logically, such a representation should provide better results than those obtained by the representation "bag of words".

Another method is based on stemming; it is to seek the root of a lexical term [Sahami M 1999], for example, the infinitive forms of the singular for verbs and nouns.

Another representation, which has several advantages (mainly, this method treats the textual regardless of the language), is based on "n-grams" (a "n-gram is a sequence of n consecutive characters).

There are different methods to compute the weight Vki knowing that for each term, it is possible to compute not only its frequency in the corpus, but also the number of documents containing that term.

Most approaches [Sebastiani, F 2002] focus on the vector representation of text using the measure TF ∗ IDF. TF represents "Term Frequency": the number of occurrences of the term in the corpus. IDF represents the number of documents containing the term.

These two concepts are combined (by product) to assign a higher weight to terms that often appear in a document and rarely in the entire corpus.

b. Similarity measure

Several measures of similarity between documents have been proposed in literature in particular is the Euclidean distance, Manhattan and Cosinus that we detail in Section 3.

c. Unsupervised classifications algorithm

The principle of unsupervised classification (clustering) is to group texts that seem similar (having common affinities) in the same class. The texts in different classes have different affinities.

The unsupervised classification of a set of documents is a highly combinatorial problem. Indeed, the number of possible partitions $P_{n\,k}$, n documents into k classes is given by the Stirling number of second species:

$$P_{n,k} = \frac{1}{k!} \sum_{k} C_k^i (-1)^{k-i} i^n$$

The methods of unsupervised classification can be divided into two families: the family of methods of hierarchical classification and methods of non-hierarchical classifications (Figure 1).

Unsupervised classification or clustering is a fundamental technique in data mining (structured or unstructured). Several methods have been proposed:

- Hierarchical classification: tree of classes.
- Ascending hierarchical classification: Successive Agglomeration.
- Descending hierarchical classification: Successive divisions.
- Flat Classification : algorithm for k-means: Partition

Some work in the field of classification:

- An overview of biomimetic algorithms for classification carried out in Informatics' Laboratory in University of Tours, Ecole Polytechnique of the University of Tours [AZZAG H 2004].
- Classification provided by cellular automata [AZZAG H 2005].
- Visual search and classification data cloud of flying insects. [Nicolas Monmarché 2003].
- Competition colony of ants for supervised learning: CompetAnts. [VER 2005].
- Unsupervised Classification contextualized [Laurent Candillier 2004].
- SOM for unsupervised Automatic Classification of Textual based WordNet [Amine et al., 2008].

Mainly these are the most seen in the inspiration of our work.

2. Representation of texts

2.1 Representation of texts based on the n grams

An n-gram is a sequence of n consecutive characters in a document, all n-grams (usually n = (2, 3, 4.5)) is the result obtained by moving a window of n boxes on the body text.

This movement is made by steps of one character and each step we take a picture. All of these pictures gives the set of all n-grams of the document. Then there are the frequencies of n-grams found.

For example, to generate all 5-grams in the sentence: "The girl eats the apple":
we get:

The-gi, he-gi, e-gir, -girl, girl-,irl-e, rl-ea, l-eat, -eats, eats-, ats-t, ts-th, s-the, -the-, the-a, he-ap, e-app, -appl, apple.

Note. A dash (-) represents the space between words in text

Since its creation by (Shannon) in 1948, the concept of n-grams has been widely used in several areas such as the identification of speech and information retrieval: representation of textual document by the method of n-grams has many advantages In fact, the n-grams capture the knowledge of most frequent words of each language which facilitates identification of language and the method of n-grams is independent of language, while the systems based on words For example, are dependent on language.

Another advantage of the representation of texts with the n-grams is that this method is tolerant to spelling mistakes, for example, when a document is scanned using an OCR, the OCR is often imperfect.

For example, the word "character" can be read as "claracter. A system based on the words hardly recognize the word "character" or even its root; But, a system based on the n-grams will be able to take into account other n-grams as " aract", "racte" etc., with n = 5, some retrieval systems based on n-grams have retained their performance with the deformation rate of 30%, a rate at which any system based on words can not function properly.

Finally, with the use of n-grams for the representation of textual documents is not required to pre-treatment language, that is to say, the application of techniques stemming or elimination of "stopwords", sequences of n-grams do not improve performance.

For example, if a document contains many words from the same root, the frequencies of n-grams corresponding increase without requiring any prior language processing.

In the phrase "the fisherman fishing" the 5-grams and the corresponding representative vector are as Figure 2.

2.2 Representation of texts based on Wordnet Approch

We used in our experiments the Reuters 21578 corpus, which represents a database of 21578 text documents of news information in English. Thus, before the clustering phase, text documents must be indexed i.e. vectorized without loss of semantics. The first step in indexing is the pre-treatment; it is to remove any symbol that does not correspond to a letter of the alphabet (points, commas, hyphens, numbers, etc.). This operation is motivated by the fact that these characters are not related to the contents of documents and does not change the meaning if they are omitted, and therefore they can be neglected. The second step is called stopping which corresponds to the deletion of all words that are too frequent (they do not have to distinguish between documents) or play a purely functional role in the construction of sentences (articles, prepositions, etc...). This operation is motivated by the fact that these characters are not related to the contents of documents and does not change the meaning if they are omitted, and therefore they can be neglected. The stopping corresponds to the deletion of all words that are too frequent (they do not have to distinguish between documents) or play a purely functional role in the construction of sentences (articles, prepositions, etc...). The result of stopping is that the number of words in the collection, what is called the mass of words is reduced by an average of 50%. To eliminate the words, known as stop-words, are harvested in the stop-list which usually between 300 and 400 elements. Then comes the step of stemming which is to replace every word in the document root e.g. national, nationality and nationalization are replaced by their root "national" and conjugated verbs by their infinitives. The stemming does not affect the mass of words, but reduced by 30% in the average size of the document. We used the algorithm PORTER to address this step. Then comes the lemmatization using WordNet approach; WordNet is lexical database. WordNet is commonly referenced as a lexicon. The words in WordNet are represented by their canonical form called lemma. This step is used to prepare the next step, which is crucial to indexing vectorization (scanning). Lemmatization replaces each word of the document by its SYNSET (set of synonyms in the lexicon).

The vectorization is realised by the method TF-IDF (Term Frequency / Inverse Document Frequency) which is derived from an algorithm for information retrieval. The basic idea is to represent documents by vectors and measure the closeness between documents by the angle between vectors, this angle is assumed to represent a semantics distance. The idea is to encode each word of the bag by a scalar (number) called tf-idf to give a mathematical aspect to text documents.

$$tfidf = tf(i, j).idf(i) = tf(i, j).\log\left(\frac{N}{N_i}\right)$$

Where:

• tf(i,j) is the term frequency: the frequency of term t_i in document d_j

• idf(i) is the inverse document frequency: the logarithm of the ratio between the number N of documents in the corpus and N_i the number of documents containing the term t_i.

A document corpus di after vectorization is:

$d_i = (x_1, x_2, \dots \dots, x_m)$ where m is the number of word of i^{th} bag of word and x_j is the tf-idf.

This indexing scheme gives more weight to words that appear with high frequency in some documents. The underlying idea is that these words help to discriminate between texts with different subject. The tf-idf has two fundamental limitations: The first is that longer documents are typically rather strong weight because they contain more words, so "the term frequencies" tend to be higher. The second is that the dependence of the "term frequency" is too important. If a word appears twice in a document d_j, it does not necessarily mean that it has two times more important than in a document d_k where it appears only once.

In our study we standardized frequencies TFxIDF by the following formula:

$$\boldsymbol{TF\times IDF\ (t_k, d) = \frac{n_i}{\sum_i n_i} * \log\left(\frac{N}{DF}\right)}$$

The weighting of TF × IDF has the following effects:

1) Importance of each word in the standard text

2) A word that appears in all documents is not important for differentiation of texts

3) Relevance words globally uncommon but common in certain documents.

Encoding TF × IDF does not correct the lengths of texts, to this end, the coding TFC is similar to that of TF × IDF, but it fixes the lengths of the texts by cosine normalization, in order not to encourage the longest.

$$TFC\left(t_k, d\right) = \frac{TF \times IDF\left(t_k, d\right)}{\sqrt{\sum^{|r|}\left(TF \times IDF\left(t_s, d\right)\right)^2}}$$

3. The cellular automaton for clustering

The cellular automaton we propose is a network of cells in a 2D space and belongs to the family (k, r) where k is the number of possible states of a cell i.e. the cardinal of all states and r is the environment of the cell i.e. r is the radius of the neighbourhood.

This automaton has 4 possible states (k = 4) and the radius of neighbourhood is a single cell (r = 1). In other words, the neighbourhood used is the vicinity of Moore (8 neighbouring cells around the cell itself) slightly modified.

Thus a cell of the automaton is dead, alive, alone or contains a data that all states of the automaton is (dead, alive, isolated, Active).

A dead cell will contain the value 0, a living cell will contain the value -1, an isolated cell will contain the value -2 and an active cell contain data (number of the document corpus).

We used these values and especially the value of the living cell (-1) to make the difference between a living cell containing the value 1 and a cell containing a data (number of the document) 1. Thus a cell will contain a value of {-2, -1, 0, 1, 2... N} where N is the number of the last document of the corpus used.

If N is the number of documents to classify the size of the 2D grid cell is m∗m with m = 2 ∗ (Int (sqrt (N)) + 1) where 2 is an empirical coefficient used for the organization of spatial class in the grid.

Example: to classify 150 documents in the corpus REUTER 21,578 we must have a grid of 13 x 13 to represent the 150 documents and a 26 x 26 grid to represent the different classes of 150 documents in the grid spacing (Figure. 3).

3.1 The Neighbourhood

The neighbourhood used in the cellular automaton that we propose is an hybrid containing neighbourhood near which is the Moore neighbourhood of radius 1 containing 8 cells around the cell itself and two neighbourhoods of radius 1 arising from the fact that the grid is planar. Since the grid is the planar ends of the four neighbourhood contains only three (3) and the neighboring cells surrounding a cell (i, j) belonging to the perimeter of the grid (without the ends) is the set of five (5) cells surrounding the cell (i, j) of radius 1. (Figure 4)

3.2 The transition function of automaton Class_AC

Rule 1:

If the cell $C_{i,j}$ dies when then cell $C_{i,j}$ ← data
Neighbourhood

Rule 2:

If cell $C_{i,j}$ Living Then Check Neighbourhood
If neighbourhood contains at least
one activate cell ←
Then
$C_{i,j}$ similar data

Rule 3:

If a cell is isolated then unchanged (Isolated remains)

3.3 The similarity matrix

We experiment the classification using three type of similarity distance: the Euclidean distance, Manhattan and cosine.

3.3.1 The Euclidian distance

Distances between vectors Ti and Tj in multidimensional space are:

$$D(T_i, T_j) = \sqrt{\sum_k (x_k(T_i) - x_k(T_j))^2}$$

3.3.2 The Manhattan distance

Distances between vectors Ti and Tj in multidimensional space are:

$$D(T_i, T_j) = \sum_k |(x_k(T_i) - x_k(T_j))|$$

3.3.3 The Cosine:

Distances between vectors Ti and T_j in multidimensional space are

$$Cos(T_i, T_j) = \frac{T_i . T_j}{||T_i|| . ||T_j||}$$

Where Ti, Tj represents the scalar product of the vectors Ti and Tj

| | Ti | | and | | Tj | | Represent respectively the standards of Ti and Tj.

The similarity matrix is a symmetric matrix of dimension N ∗ N, where N is the number of documents to classify diagonal equal to zero for the Euclidean distance and Manhattan and diagonal equal to 1 for the cosine distance. The indices represent the indexes in the corpus of the documents to classify.

3.4 Description of the algorithm Class_AC

```
- Indexing documents corpus to classify.
- Vectorize each text document corpus by the TF-IDF method.
- Compute the similarity matrix from the vectors found: $sim (i, j) = D (d_i, d_j)$.
- Initialize all the cells of the automaton to the state 'dead' (status = 0).
- Repeat (for each t)
- For each cell of the automaton do
        If cell is dead then
                cell becomes Active
                Neighbourhood cell becomes Living
        End If
        If cell is alive then
             Check neighbourhood
               If neighbourhood contains at least 1 cell
                 then
                 Cell becomes active (similar data)
                 Neighbourhood cell becomes living
                 Otherwise
                 Neighbourhood cell becomes isolated
               End If
           If cell is isolated then cells isolated (Unchanged)
- End For
- Until End of data.
```

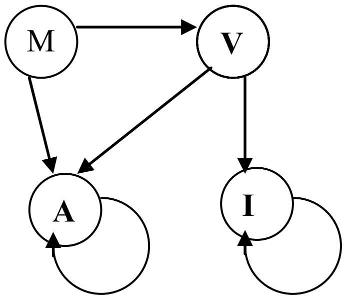
Class_AC Algorithm

At each iteration of the algorithm, the cells will change their status under the transition rules defined by the cellular automaton that will seek to consolidate similar statements to the active cells (containing the documents). The classification is recovered (The data may appear multiple times in the grid).

3.5 Simulation of the cellular automata

- At time t, the cell is dead.
- At time t+1, the cell is active (contains data) and the neighbourhood is alive.
- At time t, the cell is alive.
- At time t+1, after checking the neighbourhood, the cell will become active after consulting the similarity matrix and the index will contain similar (if this similarity is less than or equal to a predefined threshold).

These sequences are iterated until we obtain a cellular automaton with state activated, isolated and dead; data clusters are separated by cells in isolated state. (Stopping criterion: when we have finished with all documents)



M: cell death
V: living cell
A: active cell
I: isolated cell.

Transitions of a cell

4. Experimentations

After testing the algorithm Class_AC on the outcome of the Reuters 21578 corpus, we obtained the following results in terms of number of classes and purity of the clusters for both methods of representing text documents in this case the conceptual representation based on WordNet and representation by N-grams.:

Regarding the purity of a cluster we used a threshold of similarity is the distance between two documents. If this distance is less than or equal to the threshold then the documents are similar. Cosine to the distance threshold is compared to the value | 1 - cos (Vi, Vj) |.

In terms of purity of the cluster, and error rate in classification we used two measures of assessment in this case the entropy and f-measure. Both measures are based on two concepts: recall and precision defined as follows:

$$precision(i,k) = \frac{N_{i,k}}{N_k}$$

$$recall(i,k) = \frac{N_{i,k}}{N_{c_i}}$$

where N is the total number of documents, i is the number of classes (predefined), K is the number of clusters unsupervised classification, N_{Ci} is the number of documents of class i, N_k is the number of documents to cluster C_k , N_{ik} is the number of documents of class i in cluster Ck. The entropy and f-measure are calculated on a partition P as follows:

$$F(p) = \sum \frac{N_{C_i}}{N} \max_{k=1}^{K} \frac{(1+\beta)\times recall(i,k)\times precision(i,k)}{\beta \times recall(i,k) + precision(i,k)} \quad (1)$$

$$E(p) = \sum_{k=1}^{K} \frac{N_k}{N} \times (-\sum precision(i,k) \times \log(precision(i,k))) \quad (2)$$

In general β = 1

The partition P corresponds to the expected solution is one that maximizes the F-measure or minimizes the associated entropy. (In our study P is the partition that corresponds to the class of results of classification by the SOM method for the number of documents associated).

4.1 Definition of the threshold

Threshold 1: For the Euclidean distance and Manhattan and after normalization of the matrix of similarity (distance in [0,1]) we allowed an error rate of 10% (threshold 1 = 0.1) and the distance cosine we have tolerated 20%.

Threshold 2: A 15% threshold (threshold 2 = 0.15) for the Euclidean and Manhattan distances against threshold 2 = 0.25 (25%) for the cosine distance.

These threshold values were chosen after testing of the classification by cellular automata.

4.2 Results

The experiment started with 50 documents and the results have only proved that cellular automata were able to make the unsupervised classification of text documents (Clustering of text) because they regrouped effectively similar documents. Then we performed the experiment with 150 documents (first 50 texts of three documents REUTERS 21,578) and we have achieved concrete results from the previous, which led us to increase the size of the corpus documents in 350 to take a decision on the quality of the classification after evaluation by the entropy and F-measure.

In this study and the representation of text documents by the method of n-grams we have chosen n = 5, i e the method of 5-grams.

We experienced our cellular automaton on the Reuters 21578 corpus, we proceeded to the extraction 350 texts we have indexed and then calculated the similarity matrix. After testing we have achieved the results grouped in tables and figures below.

For each representation of documents and in terms of results we obtained different classes by the three distances used by varying the threshold of similarity (Table 1, 2, 3, 4, 5, 6). The classes are found in a grouping of similar documents in a way guided by the threshold.

4.3 Interpretation

From Figure 5 we see that there is almost no difference for the two representations of documents in terms of number of classes and seeing figures 6 and 7 we note that the method of n-grams has an influence on the approach wordnet because for the f-measure, the graph of the n-gram is superior to that of wordnet and the same thing for the graph entropy n-grams is inferior to that of wordnet therefore may conclude that for cosine distance, the representation of n-grams is much better than wordnet.

Regarding the Euclidean distance and from Figure 8 the number of classes varies roughly the same for both methods of representation. According to Figure 9 which shows the rating curve of the f-measure, the graph approach wordnet is well above the graph representation of the n-grams but from Figure 10 which shows the rating curve of the entropy, the graph representation of the n-grams is much less than the graph approach wordnet. So it is difficult to decide on the best representation of documents with the Euclidean distance because based on the f-measure can be said that the approach wordnet is better than n-grams as the maximum f-measure is achieved in the approach wordnet and based on the entropy we can say that the representation by the n-gram is better than the approach wordnet as the minimum entropy is achieved in the method of n-grams. In conclusion for this distance, we can opt for the textual representation of the WordNet approach because we can rely on the assessment by the F-measure itself because it is based on two concepts namely recall and precision but the entropy is based only on precision.

Regarding the Manhattan distance from Figure 11 we see a total difference of variation curve for both methods of representation. From Figures 12 and 13 we note that the curves behave the same as those observed for the Euclidean distance. So what has been concluded for the Euclidean distance is valid on the Manhattan distance.

The colored boxes in the tables above represent the best Threshold of classification because of the choice of the minimum entropy (gray boxes) or the maximum F-measure (blue boxes). The best values of the entropy or the F-measure are highlighting in yellow in the table above. Regarding the choice of the best classification we opted for the F-measure because it is based on two concepts (recall and precision) as shown in formula (1).

In conclusion we can say that to a good classification should be a good representation of textual documents and a good choice of similarity metric. To make a good representation must choose the right approach. To choose the right approach and based on the results of our study, we can choose the representation of text documents by the methods of n-grams if the distance chosen is the cosine distance, and opt for the representation of textual documents by wordnet approach if the distance chosen is the Euclidean distance or Manhattan distance. It recapitulates in the following table.

In terms of time, the convergence of the algorithm is very fast (less than 1 second) as indicated in the table above for the Euclidean distance, Manhattan and the Cosine distance. Therefore what was said in the literature on cellular automata is observed in our study. We noticed that the execution time increases with the number of documents to classify.

5. Conclusion and perspectives

In conclusion, we proposed a first algorithm for unsupervised classification (clustering) using cellular automata and represented by two methods (approach Wordnet and N-grams). After testing we have proved that this algorithm can solve a text mining problem .i.e. the clustering.

The transition function used in our automaton is changed by forming groups (cluster) similar to a certain threshold meadows. The methods of indexing text documents such as TF-IDF ,n-grams approach and Wordnet approach helped us to mummeries documents so that the use of cellular automaton on digital vectors is possible. So passage of documents to text, text to number, number to the analysis by cellular automata and analysis to decision making on the classification have been the subject of this study in this article.

As future work; we are studying 3D cellular automaton and its use for visual viewing and navigation classes in 3D space. These algorithms will also be tested for other types of data such as images and multimedia data in general.

References

A .Lehireche, A .Rahmoun. (2006). on line learning: evolving in real time a neural Net Controller of 3D-robot-arm. Track and Evolve, 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06),, Dubai/Sharjah, UAE. 2006.

A. Amine, Z. Elberrichi, M. Simonet, L. Bellatreche and M. Malki. (2008). SOM pour la Classification Automatique Non supervisée de Documents Textuels basés sur Wordnet. *Extraction et gestion des connaissances (EGC'2008)* – INRIA-Sophia Antipolis -France-. Volume 1. Revue des Nouvelles Technologies de l'Information RNTI-E-11 Cépaduès-Éditions 2008. ISSN : 1764-1667.

A.Lehireche, A Rahmoun. (2004). Evolving in Real Time a Neural Net Controller of Robot-Arm: Track and Evolve.Lithuanian Academy of Sciences, INFORMATICA International Journal ,Vol. 15, No. 1, 63-76, 2004, © IMI, LT.

Aas, K., Eikvil, L. (1999). Text categorization: a survey. *Technical report, Norwegian Computing Center*, 1999.

Agata Kramm. (2005). AUTOMATES CELLULAIRES, Mémoire de maîtrise d'informatique, Universit´e Paris VIII, Septembre 2005

Alessandro Vinciarelli, Indexation de Documents Manuscrits Offline

APTÉ C., DAMERAU F., WEISS S. (1994). Automated learning decision rules for text categorization, ACM Transactions on Information Systems, vol. 12, no 3, 1994, pp. 233-251.

AZZAG H., PICAROUGNE F., GUINOT C., VENTURINI G. (2004). *Un survol des algorithmes biomimétiques pour la classification*. Classification Et Fouille de Donnée, pages 13-24, RNTI-C-1, Cépaduès. 2004.

BURGES C. (1998). A tutorial on Support Vector Machines for pattern recognition, Data Mining and Knowledge Discovery, vol. 2, no 2, 1998, pp. 121-167.

Classification de données par automate cellulaire, H. Azzag, F. Picarougne, C. Guinot, G. Venturini, *Université François-Rabelais de Tours, Laboratoire d'Informatique (EA 2101)*

Efficient and effective clustering methods for spatial data mining. In J. BOCCA, M. JARKE & C. ZANIOLO, Eds., *20th Int. Conf. on Very Large Data Bases*, p. 144–155, Santiago, Chile : Morgan Kaufmann Publishers.

GANGULY N., SIKDAR B. K., DEUTSCH A., CANRIGHT G., CHAUDHURI P. P. (2003). *A Survey on Cellular Automata*. Technical Report Centre for High Performance Computing, Dresden University of Technology, December 2003.

H. Azzag, F. Picarougne, C. Guinot, G. Venturini. (2005). VRMiner: a tool for multimedia databases mining with virtual reality. Processing and Managing Complex Data for Decision Support (2005). J. Darmont and O. Boussaid, Editors.

J. Hansohm. (2000). Two-mode clustering with genetic algorithms. In Classification, Automation, and New Media: Proceedings of the 24th Annual Conference of the Gesellschaft Fur Klassifikation E.V., pages 87–94, 2000.

Laurent Candillier, Isabelle Tellier, Fabien Torre. Tuareg: Classification non supervisée contextualisée - Université Charles de Gaulle - Lille 3 France.

LUMER E., FAIETA B. (1994). *Diversity and adaption in populations of clustering ants.* In Proceedings of the Third International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3, pages 501-508. MIT Press, Cambridge, MA, 1994.

Nicolas Monmarché, Christiane Guinot,Gilles Venturini, Fouille visuelle et classification de données par nuage d'insectes volants, *Laboratoire d'Informatique de l'Université de Tours, École Polytechnique de l'Université de Tours - Département Informatique.*

R.M .Hamou, A .Lehireche. (2009). La classification non supervisée (clustering) de documents textuels par les automates cellulaires, International Conference on Information Technology and its Applications (CIIA-09), Saïda/Algeria 2009. CEUR Workshop proceedings Volume 547/45 Edition 2009. ISSN : 1613-0073

R.M .Hamou, A .Lehireche. Proposal of cellular automaton for solving text-mining problem:Unsupervised classification (clustering) based on wordnet. Proceedings of the second International Conference on web and Information Technology (ICWIT'09), Kerkennah Island, Sfax, Tunisia IHE Edition. Pages 63-75. ISBN : 978-9973-868-23-7

Sahami, M. (1999). Using Machine Learning to Improve Information Access. PhD thesis, Computer Science Department, Stanford University, 1999.

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

VON NEUMANN J. (1966). *Theory of Self Reproducing Automata.*, University of Illinois Press, Urbana Champaign, Illinois, 1966.

Table 1. Result of Cosine distance with Wordnet Approach

	# Classe	Time (ms)	F-measure F(p)	Entropy E(p)
1	**159**	**3,354**	**50%**	**21%**
2	**71**	**3,521**	**40%**	**27%**
3	**52**	**3,362**	**37%**	**30%**
4	**45**	**2,245**	**36%**	**15%**
5	**37**	**2,860**	**15%**	**14%**
6	**24**	**1,802**	**25%**	**36%**
7	**22**	**1,600**	**19%**	**36%**
8	**20**	**1,564**	**15%**	**37%**
9	**15**	**1,486**	**10%**	**39%**
10	**17**	**1,480**	**9,1%**	**39%**
11	**14**	**1,384**	**6,2%**	**40%**
12	**10**	**0,320**	**6,3%**	**42%**

Table 2. Result of Cosine distance with N-grams Approach

	# Class	Time (ms)	F-measure F(p)	Entropy E(p)
1	**230**	**393**	**0.35**	**0.11**
2	**157**	**386**	**0.40**	**0.63**
3	**101**	**386**	**0.34**	**0.53**
4	**56**	**328**	**0.37**	**0.48**
5	**53**	**344**	**0.29**	**0.29**
6	**35**	**282**	**0.26**	**0.19**
7	**30**	**297**	**0.25**	**0.18**
8	**24**	**298**	**0.23**	**0.16**
9	**21**	**301**	**0.22**	**0.14**
10	**18**	**300**	**0.21**	**0.15**
11	**17**	**299**	**0.20**	**0.14**
12	**17**	**301**	**0.22**	**0.14**

Table 3. Result of Euclidian distance with Wordnet Approach

	# Classe	Time (ms)	F-measure F(p)	Entropy E(p)
1	**142**	**0,389**	**46%**	**37%**
2	**125**	**0,845**	**45%**	**36%**
3	**112**	**0,785**	**43%**	**35%**
4	**90**	**0,765**	**42%**	**34%**
5	**75**	**0,471**	**44%**	**27%**
6	**64**	**0,548**	**45%**	**29%**
7	**56**	**0,678**	**42%**	**26%**
8	**48**	**0,104**	**41%**	**30%**
9	**43**	**0,786**	**51%**	**40%**
10	**37**	**0,387**	**45%**	**21%**
11	**34**	**0,946**	**14%**	**23%**
12	**32**	**0,812**	**25%**	**8%**

Table 4. Result of Euclidian distance with N-grams Approach

	# Class	Time (ms)	F-measure F(p)	Entropy E(p)
1	175	266	0.20	0.026
2	170	250	0.19	0.034
3	159	266	0.20	0.040
4	143	250	0.21	0.049
5	126	250	0.23	0.048
6	108	250	0.24	0.080
7	96	242	0.25	0.076
8	73	241	0.25	0.12
9	42	243	0.27	0.38
10	24	236	0.22	0.39
11	175	266	0.20	0.026
12	170	250	0.19	0.034

Table 5. Result of Manhattan distance with Wordnet Approach

	# Classe	Time (ms)	F-measure F(p)	Entropy E(p)
1	142	0,389	37%	46%
2	125	0,845	36%	45%
3	112	0,785	35%	43%
4	90	0,765	34%	42%
5	75	0,471	27%	44%
6	64	0,548	29%	45%
7	56	0,678	26%	42%
8	48	0,104	30%	41%
9	43	0,786	40%	51%
10	37	0,387	21%	45%
11	34	0,946	23%	14%
12	32	0,812	8%	25%

Table 6. Result of Manhattan distance with N-grams Approach

	# Class	Time (ms)	F-measure F(p)	Entropy E(p)
1	5	188	0,16	0,80
2	7	188	0,19	0,83
3	36	219	0,18	0,71
4	81	188	0,18	0,18
5	97	187	0,18	0,09
6	104	187	0,18	0,10
7	114	188	0,21	0,079
8	118	188	0,31	0,058
9	149	187	0,31	0,004
10	152	192	0,19	0,03
11	153	191	0,17	0,06
12	157	192	0,18	0,12

Table 7. Choice of approach and distance

	Wordnet Approach	N-Grams method
Cosine distance		✓
Euclidian distance	✓	
Manhattan distance	✓	

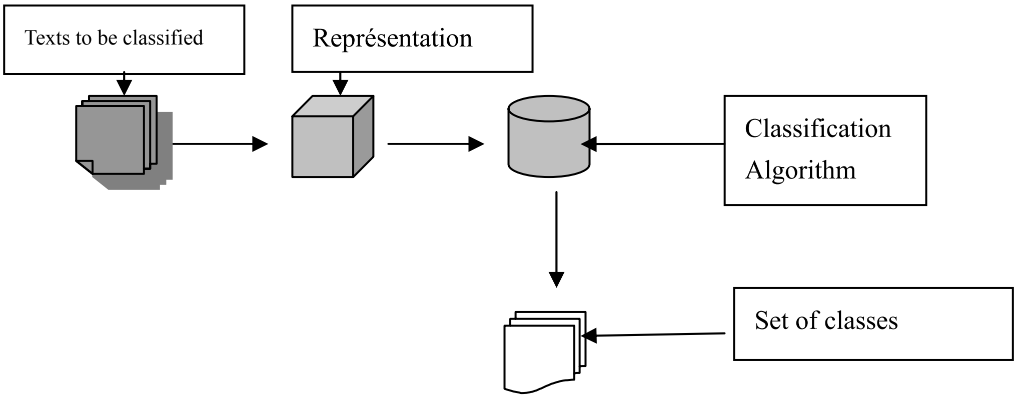


Figure 1. Unsupervised automatic classification (clustering)

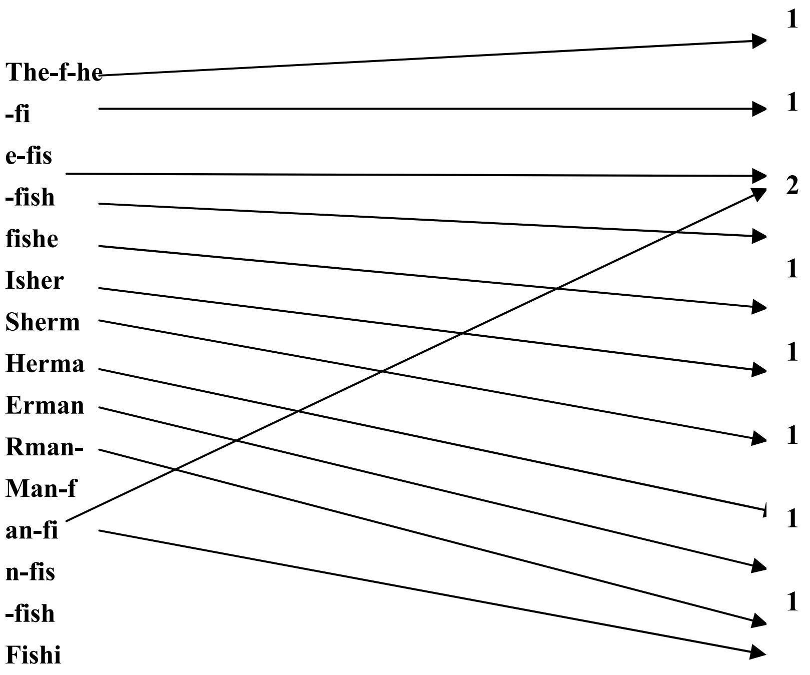
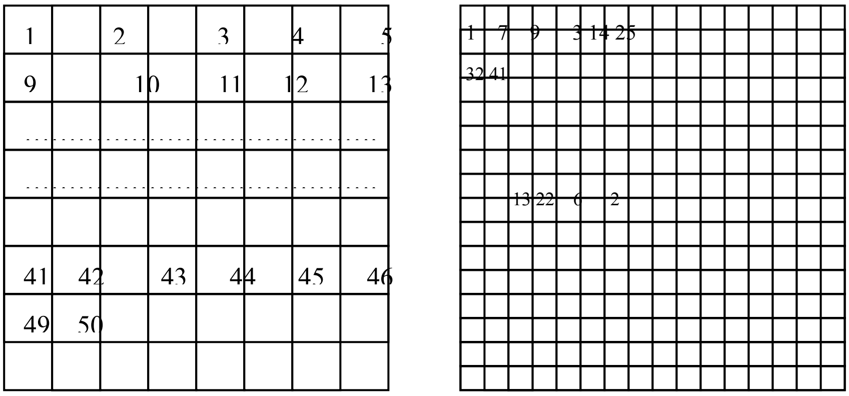


Figure 2. The 5-grams corresponding to the phrase "the fisherman fishing" with their representative vector.



13 x 13 26 x 26

Figure 3. Example of grid for 150 documents

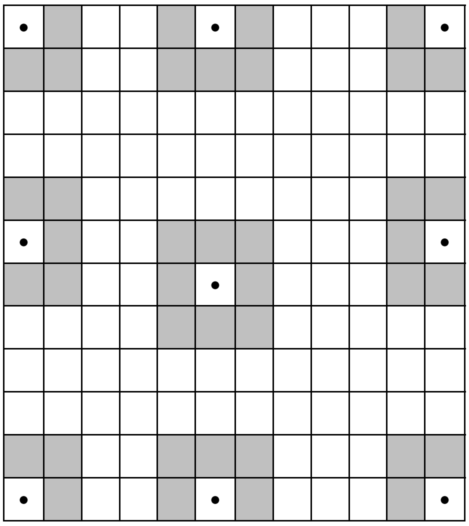


Figure 4. Neighbourhood

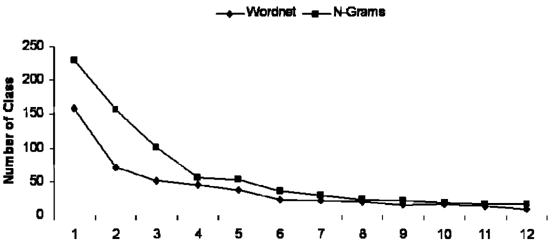


Figure 5. Number of Class – Cosine distance

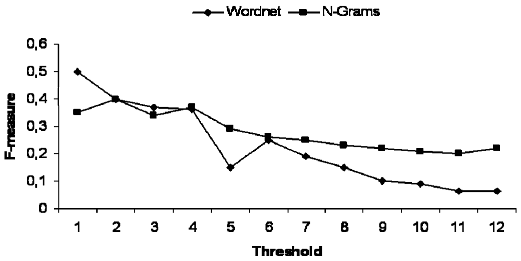


Figure 6. F-measure – Cosine distance

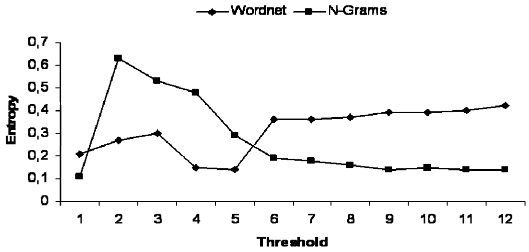


Figure 7. Entropy – Cosine distance

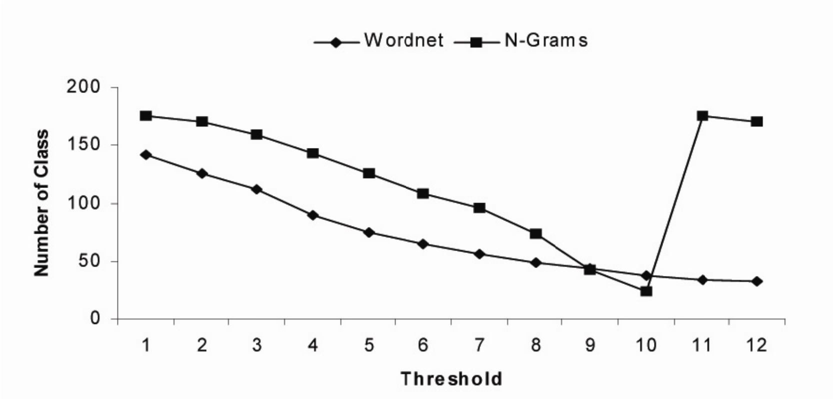


Figure 8. Number of Class – Euclidian distance

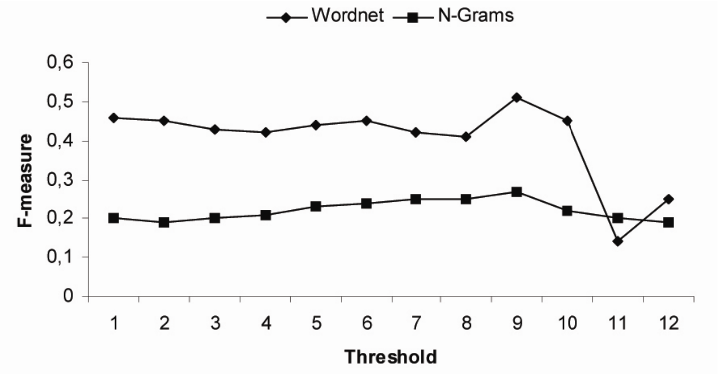


Figure 9. F-measure – Euclidian distance

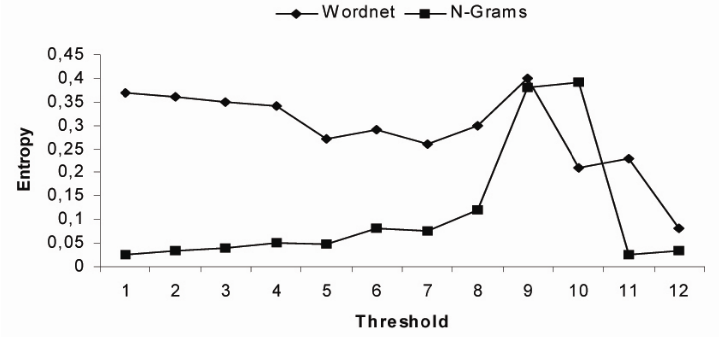


Figure 10. Entropy – Euclidian distance

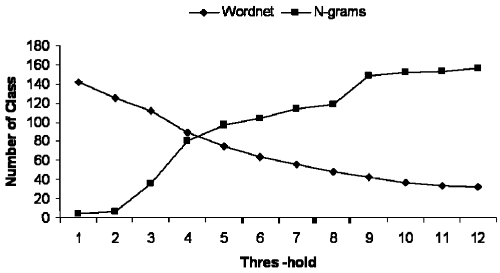


Figure 11. Number of Class – Manhattan distance

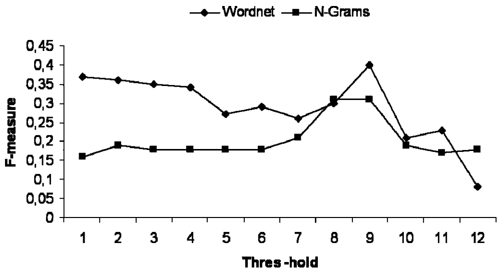


Figure 12. F-measure – Manhattan distance

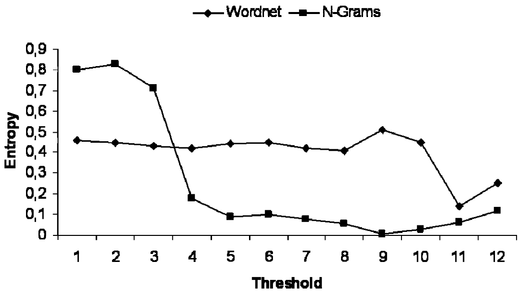


Figure 13. Entropy – Manhattan distance