INFORMATION MANAGEMENT IN GLOBAL ENVIRONMENTS: Swarm Intelligence in multilingual economic document repositories

artigo de revisão

*Angel Cobo Ortega**
*Rocío Rocha Blanco***
*Adolfo Alberto Vanti ****

ABSTRACT: The information is a strategic resource of first order for organizations, so it is essential to have methodologies and tools that allow them to properly manage information and extract knowledge from it. Organizations also need knowledge generation strategies using unstructured textual information from different sources and in different languages. This paper presents two bio-inspired approaches to clustering multilingual document collections in a particular field (economics and business). This problem is quite significant and necessary to organize the huge volume of information managed within organisations in a global context characterised by the intensive use of Information and Communication Technologies. The proposed clustering algorithms take inspiration from the behaviour of real ant colonies and can be applied to identify groups of related multilingual documents in the field of economics and business. In order to obtain a language independent vector representation, several linguistic resources and tools are used. The performance of the algorithms is analysed using a corpus of 250 documents in Spanish and English from different functional areas of the enterprise, and experimental results are presented. The results demonstrate the usefulness and effectiveness of the algorithms as clustering technique.

Keywords: clustering, ant-based algorithms, multilingual documents, text mining

* Doutor em Ciências Matemáticas. Universidad de Cantabria, Espanha. Universidad de Cantabria, Espanha.
Email: angel.cobo@unican.es

** Doutora em Informática pela Universidade de Cantabria, Espanha. Universidad de Cantabria, Espanha.
Email: eliana.rocha@unican.es

*** Doutor em Direção de Empresas pela Universidade de Deusto, Espanha. Universidade do Vale do Rio dos Sinos, Brasil.
Email: avanti@unisinos.br

1. INTRODUCTION

The circumstances that characterize the current business environment can be summarized in the internationalisation, the globalisation of the markets and the Information Society. Four powerful worldwide changes have altered the business environment: the emergence and strengthening of the global economy; the transformation of industrial economies and societies into knowledge and information-based service economies; the transformation of the business enterprise; and the digital firm. (LAUDON; LAUDON, 2006). In this context, the work of managers in enterprises is very information-intensive and the global environment in which it is done is very information rich, a lot of information is presented in text documents written in different languages that can be obtained from various sources. The development of the Information Society, have forced the companies to improve their competitiveness, and the managers have to be able to exploit the wealth of information which surrounds them and to use it for improving business performance. Definitively, Information Technology (IT) is transforming the way

organizations and people do business, and the way managers use information in the decision-making, planning and control processes.

The ever increasing amount of text documents written in different languages and the ever increasing dependence of people and organisations on such information require effective multilingual document retrieval and classification mechanisms. Also the popularity of Internet, and the rapid growth of the World Wide Web, has caused the increasing amount of available textual documents written in different languages that can be useful in decision-making processes. The Web expansion means electronically accessible information is now available in an ever-increasing number of languages.

In this context, the organizations need for Information Systems (IS) and computer tools that can help them to manage, consult and extract information in large sets of documents, and discover global knowledge in a multilingual environment. It is widely recognized that IS knowledge is essential for managers and IS can help companies to change the way they conduct business. However, it also necessary that these systems integrate tools with text mining and processing capabilities. Examples of such tools are the news summarizers (CHEN; KUO, SU, 2003), (EVANS; MCKEOWN; KLAVANS, 2005) or topic detection and tracking systems (SPITTERS; KRAAIJ, 2002) that are automatic systems for locating topically related material in streams of data such as newswire and broadcast news. The history of the industry is full of examples of companies who succumbed to the sudden appearance of a new technology. News summarizers and topic detection and tracking systems can help companies implement efficient technology watching systems and keep an eye on the latest technologies that are being developed, and the latest products coming into the market; to take advantage of new opportunities.

Text mining refers to the process of deriving high-quality information from text, including categorization, text clustering, concept/entity extraction and document summarization tasks. Text categorization is a classification problem of deciding whether a document belongs to a set of pre-specified classes or categories of documents; in text clustering, however, the categories are not pre-defined and the objective is to find sets or groups of related documents with a high similarity between them and high dissimilarity with the documents in the other groups. Some examples of practical applications of text mining techniques include identification of patterns and market trends through the analysis of text-based information, spam filtering, automatic suggestion and recommendations systems, monitoring public opinions (for example in blogs or review sites), automatic labelling of documents in libraries, measuring customer preferences, fraud detection or fighting cybercrime.

Text clustering is a powerful technique for topic discovery from text that has been gaining popularity with the increasing availability of digital documents in various languages from all around the world. It involves two phases: first, feature extraction maps each document to a point in high-dimensional space, and then clustering algorithms automatically group the points. Most text clustering tools mainly focus on processing monolingual documents; however in this work we focus on techniques to handle multilingual sets of documents. We present a multilingual document representation method for economic and business documents using two available linguistic tools, and then we try to apply Swarm Intelligence techniques to the problem of grouping related documents written in different languages.

Swarm Intelligence (SI), also referred to as collective intelligence, is an innovative distributed artificial intelligence paradigm for solving optimization problems that originally took its inspiration from the collective behaviour of social insects such as ants, termites, bees, and wasps, as well as from other animal societies such as flocks of birds or schools of fish. These algorithms use multiple interacting agents in order to exploit the benefits of cooperation in situations where you do not have global knowledge of an environment. In these situations individuals within the group (agents) interact to solve the global objective exchanging locally available information, which in the end propagates through the entire group such the problem is solved more efficiently than can be done by a single individual (ENGELBRECHT, 2005). Different mathematical models inspired by such behaviours have been successfully applied for solving a wide range of real problems. Ant

Colony Optimization (ACO) was one of the first techniques for approximate optimization inspired by SI. It was introduced as a technique for combinatorial optimization in the early 1990s (DORIGO, 1992). The inspiring source of ant colony optimization is the foraging behaviour of real ant colonies. In this work we propose an ACO algorithm to solve text clustering problems and compare it with a more classical SI-based clustering algorithm: ant clustering.

2. DOCUMENT PROCESSING, REPRESENTATION AND SIMILARITY MEASURE

There are several ways to model a text document; traditional text clustering methods are based on "bag of words" representation based on frequency statistics in a set of documents. This model is widely used in information retrieval and monolingual text mining (BAEZA; RIBEIRO, 1999). Each word corresponds to a dimension in the resulting data space and each document then becomes a vector consisting of non-negative values on each dimension. This process involves a indexing and pre-process to extract the terms. In the indexing phase the text is tokenized and stopwords are removed to keep only potentially interesting words; stopwords are words like articles, prepositions, etc., that don't carry any information from a linguistic point of view. After stopword elimination the remaining terms are lemmatized or stemmed, and weighted. In our work we use the Tree-Tagger tool, developed by the University of Stuttgart, to extract terms from the document with part-of-speech and lemma information.

However, with multilingual collections alternative approaches are needed in order to implement language independent retrieval, classification or clustering systems. The first alternative is the use of standard machine translation techniques to create a monolingual text corpus in the target language of the user (EVANS; KLAVANS, 2003), (RAUBER; DITTENBACH; MERKL, 2001). In spite of the inaccuracies introduced by the machine translation, the translated documents can be classified or organized correctly using appropriated text mining techniques and standard similarity measures can be used. To avoid automatic translation of every document in a pivot language, bilingual dictionaries can be used to design a cross-lingual similarity measure (MATHIEU; BESANCON; FLUHR, 2004). Dictionary-based translation of terms are not optimal, however is less costly than the complete translation using machine translation systems.

Others alternatives are the extraction of language-independent features of the document or the representation over conceptual spaces using techniques like LSI (Latent Semantic Indexing) of FCA (Formal Concept Analysis). Examples of language-independent features are names of places, people and organisations, dates, numerical expressions, cognates (words, in one or more languages, that have a common origin). Proper names have been widely studied in the field of Information Extraction, and can play an important role in Information Retrieval systems. In newspaper articles, by instance, proper names represent about 10% of the words and their informational quality can be used in similarity computation between documents written in different languages (FRIBURGER; MARUEL, 2002).

Finally, the semantic similarity of documents written in the same or different languages can also be achieved by representing the documents contents using the descriptor terms of a multilingual thesaurus (STEINBERGER; POULIQUEN; IGNAT, 2004). Thesaurus databases, created by international standards, are generally arranged hierarchically by themes and topics. An example of commonly used thesaurus is Eurovoc. Eurovoc is a multilingual thesaurus covering the fields in which the European Communities are active; it provides a means of indexing the documents in the documentation systems of the European Institutions and of their users. Eurovoc exists in 22 official languages of the European Union and is used in different information retrieval, text clustering and classification projects (STEINBERGER; POULIQUEN; IGNAT, 2005). The thesaurus is a structured list of more than 6,600 descriptors and 127 microthesauri in 21 thematic fields, by example, politics, international relations, economics, trade, finance, business and competition, employment and working conditions, production, technology and research, energy, and industry.

In our work, in order to obtain a language-independent representation in a corpus of

economic documents, we use a strategy successfully applied in a previous work (COBO; ROCHA, 2011). We extract four kinds of elements or features from any document using different linguistic resources:

- Words in the native language. Only the verbs, adjectives and nouns identified by TreeTagger are selected as features of the document.
- Proper names identified by TreeTagger. This tool classifies as proper names the words that start with capital letters; obviously, some erroneous names are obtained, however the experimental results show a good performance in the similarity computation. For place name recognition, linguistic resources like gazetteers can also be used.
- Terms in an economic multilingual glossary. We use a glossary with 18.724 Spanish-English entries. This glossary contains over 11,500 records of terms that include words, phrases, and institutional titles commonly encountered in documents of the International Monetary Fund (IMF) in areas such as money and banking, public finance, balance of payments, and economic growth. It provides versions of terms in a number of languages, without definitions. The Language Services of the IMF has granted us a license to use the glossary for research purpose.
- Microthesauri of Eurovoc thesaurus. The Office for Official Publications of the European Communities has also granted us a non-exclusive licences to use it in our research. We look for Eurovoc descriptors inside the document text and associate to the document the most appropriate microthesaurus.

Using these features, a document is represented by four vectors using a modified Vector Space Model (VSM) with the classical weighting schema TF-IDF. Traditionally, every document is represented by a vector of weighted features (BAEZA; RIBEIRO, 1999), (SALTON, 1971). Using word based features is the most popular and, despite of its simplicity, very effective feature construction method. Vector Space Model (VSM) has become a standard tool in IR systems since its introduction by Salton (1971). Given a set of index terms, not all of them are equally useful for describing the contents of a particular document, this yield to the assignment of numerical weights $w_{ij} \geq 0$ to each index term or keyword k_i of a document d_j. According VSM a document $\boldsymbol{d}_j$ is represented by a vector $\boldsymbol{d}_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$ where t is the number of index terms or keywords considered. In our approach we extract four kinds of elements or features from a document: associated terms in the economic glossary, descriptors of the thesaurus, proper names and words in the native language, so 4 vectors are used.

We use a *tf-idf schema* that computes the weight of a term in a document as the product of two factors; the first one measure the raw frequency of the term inside the document, and the second one are motivated by the fact that a term which appears in many documents is not very useful for distinguishing documents. The Term Frequency Inverse Document Frequency weighting (TF-IDF) is defined by

$$w_{\boldsymbol{ij}} = f_{\boldsymbol{ij}} \times idf_i = \frac{freq_{i,j}}{\max_p freq_{p,j}} \log \frac{N}{n_i} \qquad (1)$$

where $freq_{i,j}$ represents the number of times that keyword k_i appears in the text of document d_j, N is the total number of documents in the collection and n_i is the number of documents in which the keyword k_i appears. There are many variations of the TF-IDF formula, but all of them are based on the same idea: term weighting must reflect the relative importance of a term in a document with respect to other terms in the document as well as how important the term is in other documents.

To compute similarities between documents, or between documents and queries, several measures can be used (EGGHE; MICHEL, 2002), (STREHL; GHOSH; MOONEY, 2000). When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors and can be quantified as the cosine of the angle between vectors, known as cosine similarity or angular separation. This cosine similarity is one of the most popular similarity measure applied to text documents because it is relatively simple to compute and tends to lead to better results than other distances. Given two vectors v(**p**) and

v(**q**), representing two documents the angular separation is defined by the following expression:

$$Sim(v(\mathbf{p})\ v(\mathbf{q})\ = \cos(\sigma) = \frac{\mathbf{p} \circ \mathbf{q}}{\|\mathbf{p}\| \times \|\mathbf{q}\|} = \frac{\sum_{i=1}^{t} w_{ip}\ w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{p}^{2}}\ \sqrt{\sum_{i=1}^{t} w_{q}^{2}}} \quad (2)$$

Since the weights are non-negative, *Sim(v(**p**),v(**q**))* varies from 0 to 1. When the similarity is 0, it means that the two vectors are totally dissimilar. When the similarity is 1, it means that the two vectors are totally equal. If the vectors are normalized, this score is computed as the inner product of the vectors.

In the proposed approach, each document is represented by four vectors and we estimate the similarity score between a pair of documents (**p**,**q**) using a convex linear combination of four similarity scores, namely, similarity of entries in the glossary, similarity of Eurovoc microthesauri, proper names similarity and terms similarity, that is:

$$SimML(\mathbf{p},\mathbf{q}) = \lambda_1\ Sim(V_{glossary}(\mathbf{p})\ V_{glossary}(\mathbf{q})\ + \lambda_2\ Sim(V_{Eurovoc}(\mathbf{p})\ V_{Eurovoc}(\mathbf{q})\ + \lambda_3\ Sim(V_{pnames}(\mathbf{p})\ V_{pnames}(\mathbf{q})\ + \lambda_4\ Sim(V_{words}(\mathbf{p})\ V_{words}(\mathbf{q}) \quad (3)$$

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1 \qquad \text{with} \qquad \lambda_i \geq 0$$

This approach is also used in other research works that use multi-similarity measures and a set of vectors in the representation of each document (FRIBURGER; MARUEL, 2002), POULIQUEN ET AL., 2004); (STEINBERGER; POULIQUEN; HAGMAN, 2005).

3. SWARM INTELLIGENCE CLUSTERING ALGORITHMS: TWO ANT-INSPIRED APPROACHES

The roots of SI are deeply embedded in the biological study of self-organized behaviours in social insects (BONABEAU; DORIGO; THERAULAZ, 1999). Colonies of social insects have fascinated researchers for many years, and the mechanisms that govern their behaviour remained unknown for a long time. Biological swarm systems that have inspired computational models include ants, termites, bees, and spiders. Even though the single members of these colonies are non-sophisticated individuals, they are able to achieve complex tasks in cooperation. Examples of collaborative work are nest building, task allocation, recruitment of colony members for prey retrieval, foraging behaviours, larvae organization, clustering task of corpses… Many aspects of these collective activities are self-organized and work without a central control (BLUM; LI, 2008).

In this work we apply two ant based approaches for multilingual clustering in document collections. The first one is an Ant Colony Optimization (ACO) approach, and the second one is based on the Ant Clustering (AC) algorithms introduced by (DENEUBOURG AT AL, 1990), (LUMER; FAIETA, 1994). An ant colony has many characteristics that are considered useful; it is composed of many simple agents that can perform rather complex tasks as a group, but without central coordination. Real life ants do perform clustering and sorting of objects among their many activities.

The ACO metaheuristic has been applied successfully to solve various combinatorial optimization problems (DORIGO; STÜTZLE, 2004). In ACO principles of communicative behaviour occurring in real ant colonies are used, and several generations (populations) of artificial ants search for good solutions. Every ant of a generation builds up a solution step by step using information provided by the previous ants (pheromone trails) and heuristic information that represents a priori information about the problem. Once a solution is completed, pheromone trails are updated according to the quality of the solution built, so that the following ants are attracted by the pheromone and will likely search in the solution space near good solutions.

We propose such a population based metaheuristic to solve document clustering problems. Given a predefined number of clusters *k*, we can see the clustering problem as a combinatorial optimization problem where

we have to decide the document distribution between the clusters. An artificial ant can build a clustering step by step going through several decisions. Each decision implies a randomly selection of a non-classified document and the decision of the cluster where the document will be placed. The heuristic information in the problem is composed by the similarity matrix, and the pheromone trails are based on the quality of clustering solutions found.

The goal of the process is the maximization of the similarity function between the documents and the centroids $\mathbf{c}_j$ of their groups. This function is defined using (3) as follows:

$$W(C) = \frac{1}{N}\sum_{j=1}^{k}\sum_{\mathbf{d}_i \in C_j} SimML(\mathbf{d}_i, \mathbf{c}_j) \quad (4)$$

The probability that the ant places document *i* in cluster *j* is defined by

$$\Pr_{ij} = \frac{\tau_{ij}^{\alpha}\ ms_{ij}^{\ \beta}}{\sum_{r=1}^{k}\tau_{ir}^{\alpha}\ ms_{ir}^{\ \beta}} \quad (5)$$

where τ_{ij} is the pheromone value in the pheromone matrix, ms_{ij} is the mean similarity of document *i* with the documents in the cluster *j*, and α and β are two parameters that determine the relative influence of the pheromone and the heuristic information in the decision.

In the initialization phase, the same randomly selected value is assigned to all components in the pheromone matrix, however once an ant has constructed its solution the matrix is updated according to the following expression:

$$\tau_{ij} = (1-\rho)\tau_{ij} + \Delta_{ij} \qquad i = 1,...,N;\ j = 1,...,k \quad (6)$$

where ρ is a evaporation rate that avoids old pheromone from having a too strong influence on future decisions, and the increment applied is defined by:

$$\Delta_{ij} = \begin{cases} W(C) & if\ \ d_i \in C_j \\ 0 & if\ \ d_i \notin C_j \end{cases} \quad (7)$$

Additionally, we use an elitist strategy, at the end of each generation; the pheromone trails laid on the clustering solution with best quality found by the ants are reinforced to facilitate the search around this solution.

The other ant-inspired technique used is the *ant clustering algorithms* (HANDL; KNOWLES; DORIGO, 2006), (MONMARCHÉ, 1999). Ant-based clustering has been applied variously, in commerce, circuit design, text mining and different studies show evidences that ant-based clustering is a robust and viable alternative, compared with more classical techniques (HANDL; KNOWLES; DORIGO, 2006). A study on the performance of ant-based clustering can be found in (HANDL; DORIGO, 2003).

In this case, the clustering operation happens on a toroidal bidimensional grid, where the objects (documents) are placed randomly and a set of artificial ants explore the grid picking and dropping the objects. The probabilities of picking and dropping a document are based on the disparity between that document and other documents in its neighbourhood. When a document is similar with its neighbours in the grid, the probability of picking up it is low; however, if the similarity is low the artificial ants will pick it with high probability and will look for a good position in the grid to reallocate it. These probabilities are defined using the following expressions:

$$P_{pick}(\mathbf{d}_i) = \left(\frac{k^+}{k^+ + f(\mathbf{d}_i)}\right)^2 \qquad P_{drop}(\mathbf{d}_i) = \left(\frac{f(\mathbf{d}_i)}{k^- + f(\mathbf{d}_i)}\right)^2 \quad (8)$$

where k^+ is a pick up threshold parameter, k^- is a drop threshold parameter and $f(\mathbf{d}_i)$ is a similarity function in the neighbourhood:

$$f(\mathbf{d}_i) = \frac{1}{\sigma^2}\sum_{\mathbf{d}_j \in \Omega}\frac{SimML(\mathbf{d}_i, \mathbf{d}_j)}{\alpha} \quad (9)$$

After picking or dropping a document, the ant will move to a random adjacent position on the grid and the process continues. By following these rules, related documents will be likely to be dropped in neighbouring positions in the grid and a graphical visualization of the clusters will be obtained.

In order to speed up the clustering process a short-term memory is implemented. Each ant remembers the last few carried documents and their respective dropping positions and searches his local memory for a "best matching" document and their position to drop a new document. In (COBO; ROCHA, 2011) this clustering technique was used for the identification of related documents.

4. EXAMPLE OF APPLICATION: EXPERIMENTAL RESULTS

We present the results of a computational experiment conducted to test the performance of the proposed algorithms. We have compared the two ant-based algorithms with the classical *k-means* clustering algorithm and comparative results are presented for a document test corpus. Of course, other clustering algorithms can be used for this comparison; however, the *k-means* algorithm is one of the simplest unsupervised learning algorithms to solve clustering problems and that has been adapted to many problem domains. *K-means* constructs a partition into the final required *k* clusters. First, *k* cluster centroids are randomly selected; next, it examines each document and assigns it to one of the clusters depending on the minimum distance to the centroid. The centroids are recalculated after all documents are assigned and this continues until all the documents are grouped into the final required number of clusters. *K-means* has a time complexity which is linear in the number of points, which complies with the speed requirement of clustering a large amount of documents.

4.1. Document corpus

The test corpus contains 250 research papers in different areas of economics and management extracted from scientific journals of the involved areas. There are 125 documents in Spanish and 125 in English. We selected 50 documents about information systems, 50 of human resource management, 50 of marketing, 50 of economic theory and 50 about accounting and finance. The five thematic group are equally distributed (i.e., each group contains 25 English documents and 25 Spanish documents).

The documents in the corpus were processed and represented as described above. The extraction of glossary terms and thesaurus terms done with each document was automatically, applying an exact matching technique.

In spite of this corpus is small, we can derive some interesting conclusions and we can observe the difficulty of the topics selected for the evaluation, with documents that can belong to two different thematic groups. A study with a bigger corpus will be a future work.

4.2. QUALITY MEASURES

The clustering results of the different algorithms on the bilingual test corpus are compared using three external quality measures: purity, F-measure and entropy. The purity measures how much a cluster is "specialized" in a class or category; and is defined as the ratio of the number of documents in the dominant category to the total number of documents in the cluster. To evaluate an entire clustering we compute the average of the cluster purities weighted by cluster size. The F-measure uses the ideas of precision and recall from information retrieval. The precision and recall of a cluster *j* with respect to a category *i* are defined as:

$$P(i,j) = \frac{n_{ij}}{n_j} \qquad R(i,j) = \frac{n_j}{n_i} \tag{10}$$

where n_{ij} is the number of documents of category *i* in cluster *j*, n_j is the number of documents of cluster *j*, and n_i is the number of documents of category *i*. The overall F-measure for the clustering is computed as

$$F = \sum_i \frac{n_i}{N} \max_j \{F(i,j)\} \qquad with \qquad F(i,j) = 2\frac{P(i,j)R(i,j)}{P(i,j)+R(i,j)} \tag{11}$$

The higher the overall F-measure, the better the clustering. The optimal F-measure value is 1.

Finally, the entropy tells us how homogeneous a cluster is, its optimal value is zero. The entropy of a cluster and the overall entropy are defined by the following expressions.

$$E_j = -\sum_i \frac{n_j}{n_j} \log(\frac{n_j}{n_j}) \qquad E = \frac{1}{N}\sum_j n_j E_j \quad (12)$$

4.3. Parameter settings

The proposed algorithms require a number of different parameters to be set. One of the main difficulties of applying heuristic methods to a given problem is to decide on an appropriate set of parameter values. After several previous experiments we decide to use the values shown in Table 1, because a better performance was observed.

ACO algorithm		AC algorithm	
Similarity coefficients (λ_i)	0.45, 0.45, 0.05, 0.05	Similarity coefficients (λ_i)	0.45, 0.45, 0.05, 0.05
Number of clusters	5	Grid size	60 x 60
Population size (number of ants)	10	Population size (number of ants)	25
Pheromone parameter (α)	2.5	Picking-up parameter (k+)	0.0015
Heuristic parameter (β)	5.0	Dropping parameter (k-)	0.05
Evaporation rate (ρ)	0.15	Neighbourhood size	5 x 5
		Scale similarity parameter (α)	1.0
		Maximum length step	25
		Short-term memory size	20

Table 1: Parameter settings

In the case of the computation of overall document similarity score, the parameter setting gives each of the first two components (similarity score based on multilingual glossary terms and similarity score based on Eurovoc) a weighting factor nine times larger than each of the last two (similarity score based on proper names and similarity score based on words). The reasons behind this choice are the fact of the glossary and the thesaurus provides the connection between the terms from two documents in different languages (multilingual characteristic), and the necessity of a good similarity calculation when we compare two documents in different languages.

4.4. Evaluation

The algorithms were run 20 times with the parameter settings shown in Table 1. The obtained results are summarized in Table 2, which shows the averages over the 20 runs obtained for each of the quality measures using the *k-means* algorithm and the two proposed ant-based algorithms. The following observations can be made from this table:

ACO and AC algorithms perform very well under all quality measures and outperform *k-means* algorithm. Concerning the computational time the ACO algorithm's runtime is considerably higher than those *k-means* and AC algorithms. It is worth observing that AC algorithm scales linearly and becomes the fastest algorithm on big document collections. Another interesting feature of the AC algorithm is its robustness to the effects of outliers within the collection. By example, documents that cannot be clearly classified in a particular thematic category, the algorithm can identify them and doesn't include them in any cluster.

	k-means	ACO	AC
F-measure	0.6380	0.6432	0.6669
Entropy	0.9305	0.9114	0.9203
Purity (%)	65.10	65.56	65.42
Execution time (milliseconds)	4 945	18 577	5 581

Table 2: Comparative quality values.

The computational times of ACO algorithm are considerably high; however it has an interesting feature that can be used in small document collections: the pheromone trails obtained as additional information. Figure 1 shows the evolution of the mean distance to centroids in 30 iterations of ACO algorithm and a density plot of the pheromone trails after the run. In the density plot the cluster structure of the collection can be observed, and the different pheromone values can be seen as a membership degrees of the documents to the clusters as in a fuzzy clustering approach. This feature can be especially interesting when the structure within the data is not very pronounced, as happen in the case of a corpus of documents of related areas. In our case there are documents in the corpus that can be assigned to different thematic categories, and the pheromone information is very useful to discover them.

Both ACO and *k-means* algorithms need a priori knowledge of the correct number of cluster. However the results demonstrate that the ant clustering algorithm is quite reliable at identifying this number. Figure 2 represents the spatial distribution of the documents on the 60x60 toroidal grid after the execution of 750000 ant basic operations in a run of the AC algorithm. In this image the cluster structure of the corpus can be clearly observed. Finally, Table 3 presents a summary of the clustering solutions obtained in an execution of each proposed algorithm.

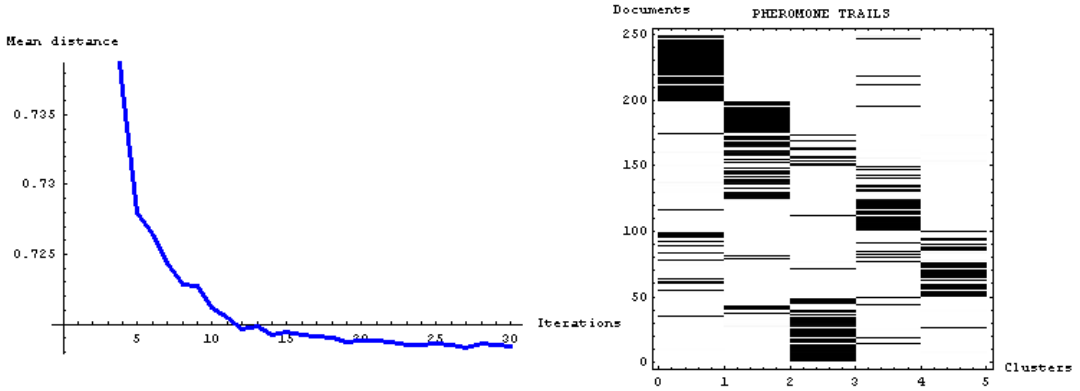

Figure 1: ACO clustering process. Evolution of mean distance to centroids and density plot

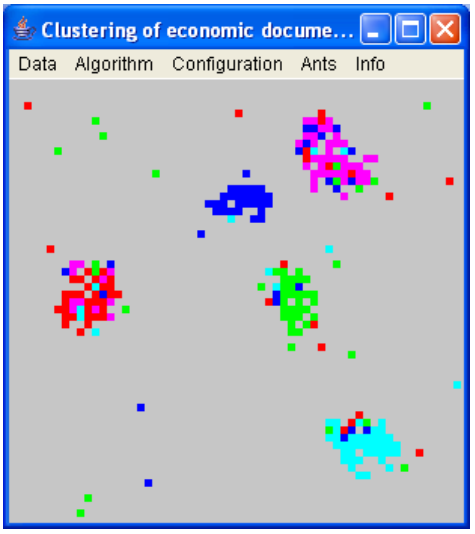

Figure 2: Ant clustering process in the bidimensional grid

	Num. Docs. ACC	Num. Docs. MKT	Num. Docs. HRM	Num. Docs. ECO	Num. Docs. INF	Dominant category	Cluster purity (%)
K-means Clustering							
Cluster 1	3	0	30	1	1	HRM	85.71
Cluster 2	7	5	10	24	0	ECO	52.17
Cluster 3	0	2	0	24	0	ECO	92.31
Cluster 4	38	30	1	1	1	ACC	53.52
Cluster 5	2	13	9	0	48	INF	66.67
ACO Clustering							
Cluster 1	2	12	1	1	47	INF	74.60
Cluster 2	5	2	17	38	0	ECO	61.29
Cluster 3	38	2	1	10	0	ACC	74.51
Cluster 4	4	5	31	1	3	HRM	70.45
Cluster 5	1	29	0	0	0	MKT	96.67
AC Clustering							
Cluster 1	2	10	7	4	39	INF	62.90
Cluster 2	1	29	0	0	0	MKT	96.67
Cluster 3	4	3	32	3	11	HRM	60.38
Cluster 4	3	3	4	33	0	ECO	76.74
Cluster 5	39	3	4	3	0	ACC	79.59
Non-classified documents	1	2	3	7	0		

Table 3: Summary of clustering solutions obtained by k-means, ACO and AC algorithms

5. CONCLUSIONS

In this work we have presented two ant-based algorithms and a vector multilingual representation model that allow the user to cluster Spanish-English collections of economic documents. We describe the text processing performed and the identification of cross-lingual features using linguistic tools, as a specialized economic multilingual glossary and Eurovoc thesaurus. We compute similarities between documents written in different languages using four types of features.

The clustering algorithms implemented are based on observed behaviours in ant colonies and the experimental results show a good performance in comparison with the classical k-means algorithm. The ACO approach is a population based algorithm that allows obtaining a set of good cluster solutions and additional information that can be seen as membership degrees. The main features of the second approach (AC) are the automatically determination of the number of clusters, the identification of outliers in the corpus, good runtimes and efficient clustering results.

Our evaluation of the algorithms was based on measurements for a corpus of 250 documents extracted from scientific journals in different business and economic areas and in two languages. Clearly a thorough evaluation for other corpora and a higher number of categories and languages are important questions to be addressed in the future.

In global context organizations need methodologies and computational tools that allow them generate new knowledge from unstructured information in textual form and in different languages. The classic techniques of text mining have convincingly shown their worth and SI techniques have left their mark in recent years and are demonstrating their efficiency for solving a large number of complex problems. In this work we have shown as the integration of these new methodologies into technological tools for the management of documents allows optimising knowledge generation processes.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the Office for Official Publications of the European Communities and the Language Services of the International Monetary Fund, because the use of original editions of Eurovoc Thesaurus and IMF multilingual Glossary (http://eurovoc.europa.eu), respectively. Thanks to Prof. Julia Handl for allowing us the use of her java source code for the implementation of AC.

Artigo recebido em 30/01/2013 e aceito para publicação em 10/03/2013

GESTÃO DA INFORMAÇÃO EM AMBIENTES GLOBAIS: computação bio-inspirado em repositórios de documentos multilingues econômicos

RESUMO: *A informação é um recurso estratégico primordial para qualquer organização, por isso é essencial dispor de metodologias sistemáticas que permitam gerir adequadamente e gerar conhecimento a partir dela. Estratégias de geração de conhecimento também são necessárias para gerar informações textuais não-estruturadas advindas de fontes e idiomas diferentes. Este artigo apresenta duas abordagens bio-inspiradas para agrupamento de documentos multilíngues em economia e negócios. O problema abordado é de particular interesse e necessário para organizar grandes volumes de informação administradas pelas organizações em um contexto global caracterizado, este caracterizado pelo uso intensivo da tecnologia da informação e comunicações. Os algoritmos de agrupamento propostos são inspirados pelo comportamento observado em colônias de formigas e pode ser usado para identificar grupos de documentos escritos em diferentes linguagens dentro da esfera da economia e administração de empresas. Com o objetivo de se obter uma representação dos documentos, independentemente da língua, se utiliza de diferentes ferramentas e recursos linguísticos. A eficácia dos algoritmos é analisada através de um corpus de 250 documentos escritos em espanhol e inglês e associados a diferentes áreas funcionais da empresa. Vários resultados experimentais são apresentados, mostrando a sua utilidade e eficiência como técnicas de agrupamento documental.*

Palavras-chave: *clustering, algoritmos baseados em formigas, documentos multilíngues, mineração de texto.*

REFERENCES

BAEZA, R.; RIBEIRO, B. **Modern Information Retrieval**. NY: Addison Wesley, 1999.

BLUM, C.; LI, X. Swarm Intelligence in Optimization, in BLUM, C.; MERKLE, D. (ed.), **Swarm Intelligence: Introduction and Applications**, Berlin: Springer, p.43-85, 2008.

BONABEAU, E.; DORIGO, M.; THERAULAZ, G. **Swarm Intelligence: From Natural to Artificial Systems.** NY: Oxford University Press, 1999.

CHEN, H.; KUO, J.; SU, T. Clustering and visualization in a multilingual multi-document summarization system. In **Advances in Information Retrieval: Proceedings of the 25th European Conference on IR Research ECIR 2003,** Lecture Notes in Computer Science vol. 2633, p. 266-280, Pisa, Italy,14-16 April, 2003.

COBO, A.; ROCHA, R.. Identification of related multilingual documents using ant clustering algorithms. **Ingeniare. Revista chilena de ingeniería,** vol. 19, n. 3, p. 351-358, 2011.

DENEUBOURG, J.; GOSS, S.; FRANKS, N.; SENDOVA-FRANKS, A.; DETRAIN, C.; CHRETIEN, L. The dynamic of collective sorting robot-like ants and ants-like robots. In **Proceedings of the First Conference on Simulation of Adaptive Behavior,** MIT Press Cambridge, p. 356-363, 1990.

DORIGO, M. **Optimization, learning and natural algorithms**. Ph.D. Thesis, Politecnico di Milano, Italy, 1992.

DORIGO, M. ; STÜTZLE, T. **Ant Colony Optimization**. Bradford: MIT Press, 2004.

EGGHE, L.; MICHEL, C. Strong similarity measures for ordered sets of documents in information retrieval. **Information Processing and Management** n.38, p. 823–848, 2002.

ENGELBRECHT, A.P. **Fundamental of Computational Swarm Intelligence.** Chichester, UK: John Wiley & Sons, 2005.

EVANS, D. ; KLAVANS, J. **A platform for multilingual news summarization. Technical report, Computer Science Technical Reports**. University of Columbia, 2003.

EVANS, D.; MCKEOWN, K.; KLAVANS, J. Similarity-based multilingual multi-document summarization. **Technical report, Computer Science Technical Reports** (num. CUCS-014-05). University of Columbia, 2005.

FRIBURGER, N.; MARUEL, D. Textual similarity based on proper names. **In Proceedings of Workshop on Mathematical/Formal Methods in Information Retrieval at the 25th ACM SIGIR Conference**, Tampere, Finland, p. 155-167. ACM New York, August 11-15, 2002.

HANDL, J.; DORIGO, M. On the performance of ant-based clustering. In **Proceedings of the 3rd International Conference on Hybrid Intelligent Systems.** Melbourne, Australia, 14-17, IOS Press, Dec. 2003.

HANDL, J.; KNOWLES, J.; DORIGO, M. Ant-based clustering and topographic mapping. **Artificial Life**, vol. 12, n.1, p. 35-61, 2006.

LAUDON, K., LAUDON, J. **Management Information Systems: Managing the Digital Firm.** NJ: 9th ed. Prentice Hall, 2006.

LUMER, E.; FAIETA, B. Diversity and adaptation in population of clustering ants. In **Proceedings of 3rd International Conference on Simulation of Adaptive Behaviour: From Animals to Animats**, Brighton, UK, MIT Press Cambridge, p. 501-508, 8-12 August, 1994.

MATHIEU, B.; BESANCON, R.; FLUHR, C. Multilingual document cluster discovery. In Proceedings of Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - **RIAO 2004, 7th International Conference, University of Avignon, France,** April 26-28, p.1-10, CID, 2004.

MONMARCHÉ, N. On data clustering with artificial ants. In Freitas, A., editor, **AAAI-99 & GECCO-99** Workshop on Data Mining with Evolutionary Algorithms: Research Directions, Orlando, Florida, p. 23-26, 1999.

POULIQUEN, B.; STEINBERGER, R.; IGNAT, C.; KÄSPER, E.; TEMNIKOVA, I. Multilingual and cross-lingual news topic tracking. In **Proceedings of the 20th International Conference on Computational Linguistics.** Association for Computational Linguistics Stroudsburg, PA, USA, p. 23-27, 2004.

RAUBER, A.; DITTENBACH, M.; MERKL, D. Towards automatic content-based organization of multilingual digital libraries: An english, french and german view of the russian information agency novosti news. **In Third All-Russian Conference Digital Libraries: Advanced Methods and Technologies,** Digital Collections Petrozavodsk, Russia, RCDI, 2001.

SALTON, G. **The SMART Retrieval System - Experiments in Automatic Document Processing**. NJ: Prentice Hall, 1971.

SPITTERS, M.; KRAAIJ, W. Unsupervised clustering in multilingual news streams. **In Proceedings of the LREC. Workshop: Event Modelling for Multilingual Document Linking**, ELRA, Paris, p. 42-46, 2002.

STEINBERGER, R.; POULIQUEN, B.; HAGMAN, J. Cross-lingual document similarity calculation using the multiligual thesaurus Eurovoc. In **Proceedings of Computational Linguistic and Intelligence Text Processing. Third International Conference, CICLing'2002.** Springer Lecture Notes in Computer Science, LNCS 2276, p. 415-424. Mexico-City, Mexico, Springer-Verlag.17-23 February 2002.

STEINBERGER, R., POULIQUEN, B., IGNAT, C. Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. **In: Information Society 2004 (IS'2004) - Proceedings B of the 7th International Multiconference - Language Technologies**, p. 2-12. Ljubljana, Slovenia, 13-14 October 2004.

STEINBERGER, R.; POULIQUEN, B.; IGNAT, C. Navigating multilingual news collections using automatically extracted information. **Journal of Computing and Information Technology,** CIT n. 13:p. 257-264, 2005.

STREHL, A.; GHOSH, J.; MOONEY, R. Impact of similarity measures on web-page clustering. In **AAAI-2000: Workshop on Artificial Intelligence for Web Search**. Austin, Texas, p. 58-64, july 2000.