Identification of related multilingual documents using ant clustering algorithms

Identificación de documentos multilingües relacionados mediante algoritmos de clustering de hormigas

Ángel Cobo[1] Rocío Rocha[2]

Recibido 15 de marzo de 2011, aceptado 3 de noviembre de 2011
Received: March 15, 2011 Accepted: November 3, 2011

RESUMEN

Este artículo presenta una estrategia de representación documental y un algoritmo bioinspirado para realizar procesos de agrupamiento en colecciones multilingües de documentos en las áreas de la economía y la empresa. El enfoque propuesto permite al usuario identificar grupos de documentos económicos relacionados escritos en español o inglés usando técnicas inspiradas en comportamientos de organización y agrupamiento de objetos observados en algunos tipos de hormigas. Para conseguir una representación vectorial de cada documento independiente del idioma, se han utilizado dos recursos lingüísticos: un glosario económico y un tesauro. Cada documento es representado usando cuatro vectores de rasgos: palabras, nombres propios, términos económicos del glosario y descriptores del tesauro. La identificación de los nombres propios y la extracción y lematización de palabras se realizan usando herramientas específicas. El esquema *tf-idf* es utilizado para medir la importancia de cada rasgo en el documento, y se utiliza una combinación lineal convexa de separaciones angulares de los vectores de rasgos como medida de similitud de documentos. El trabajo muestra resultados experimentales de aplicación del algoritmo propuesto sobre un corpus español-inglés de documentos científicos de áreas económica y de gestión empresarial. Los resultados demuestran la utilidad y efectividad de las técnicas de *ant clustering* y del esquema de representación propuesto.

Palabras clave: Clustering, algoritmos basados en hormigas, documentos multilingües, minería de texto, gestión documental.

ABSTRACT

This paper presents a document representation strategy and a bio-inspired algorithm to cluster multilingual collections of documents in the field of economics and business. The proposed approach allows the user to identify groups of related economics documents written in Spanish and English using techniques inspired on clustering and sorting behaviours observed in some types of ants. In order to obtain a language independent vector representation of each document two multilingual resources are used: an economic glossary and a thesaurus. Each document is represented using four feature vectors: words, proper names, economic terms in the glossary and thesaurus descriptors. The proper name identification, word extraction and lemmatization are performed using specific tools. The tf-idf scheme is used to measure the importance of each feature in the document, and a convex linear combination of angular separations between feature vectors is used as similarity measure of documents. The paper shows experimental results of the application of the proposed algorithm in a Spanish-English corpus of research papers in economics and management areas. The results demonstrate the usefulness and effectiveness of the ant clustering algorithm and the proposed representation scheme.

Keywords: Clustering, ant-based algorithms, multilingual documents, text mining, document management.

[1] Departamento de Matemática Aplicada y Ciencias de la Computación. Universidad de Cantabria. Avda Los Castros s/n, 39005 Santander, España. E-mail: acobo@unican.es

[2] Departamento de Administración de Empresas. Universidad de Cantabria. Avda. Los Castros s/n, 39005 Santander, España. E-mail: rochar@unican.es

INTRODUCTION

Information Technology (IT) is transforming the way organizations and people do business. The information systems capture and store data from the organization and its environment, and managers use them in the decision-making, planning and control processes. The information has become a strategic resource of first order for organizations, since proper information management can allow them to understand the reality of the environment in which they operate and to obtain competitive advantages. Technologies of Information and Communications have intensified the use of information as support in economic activities, allowing information to be processed, stored, recovered and communicated without taking distance, time or volume into account. Also, the rapid growth of the World Wide Web has profoundly affected the culture of information technology in business and the way that information is delivered via a computer. Today, anyone with a personal computer and access to the Web can access huge volumes of information distributed over computer networks. This Web expansion means electronically accessible information is now available in an ever-increasing number of languages.

Present day business activity is characterized by the internationalization and globalization of markets. In this context, organizations obtain great volumes of information from several sources in an automatic way (subscription to news, contents syndication, information retrieval from databases, Internet queries, etc.) and often this information is reflected in documents written in different languages. The result is a need for information systems and computer tools that can help the organizations to manage, consult and extract information in large sets of documents, and gather global knowledge in a multilingual environment. Examples of such tools are the news summarizers [2, 5] or topic detection and tracking systems [16, 10] which are automatic systems for locating topically related material in streams of data such as newswire and broadcast news.

Text mining generally refers to the process of deriving high quality information from texts through the divining of patterns and trends using statistical pattern learning. It also usually involves the process of structuring the input text. Text mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. Typical text mining tasks include text categorization, text clustering, concept/entity extraction and document summarization. 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 a high dissimilarity with the documents in the other groups. Clustering techniques in text mining have been gaining popularity with the increasing availability of digital documents in various languages from all around the world. However, most text clustering tools currently focus primarily on processing monolingual documents only; little attention has been paid to applying the techniques to handle multilingual sets of documents. In this work we try to apply bio-inspired techniques to the problem of grouping related documents written in different languages.

DOCUMENT CORPUS AND LINGUISTIC RESOURCES

In this work we are using a bilingual corpus of 250 categorized documents extracted from databases of scientific articles in management and economics. The articles are published in international journals of the involved areas. We have selected articles in Spanish and English of different functional areas within business: marketing, accounting and finance, information systems, economic theory and human resource management. We have 25 articles in each language and functional area.

The documents are stored in a MySQL database, and a web-based application allows the user to manage them and to perform the following processing operations on the documents:

- *Text extraction from the document*. The original documents are in PDF format and the web application uses the *pdftotext* tool to extract the text from the file; this text is inserted in a database field.

- *Stopword elimination*. A stopword list is used to eliminate very common terms like articles, prepositions, etc. The application has a predefined list of 463 stopwords (359 in Spanish and 104 in English), and allows the user to add new ones.

- *Lemmatization, word classification and proper noun identification*. The *TreeTagger* tool is used to identify the lemmas of the words and their grammatical category (verbs, adjectives, nouns,…). This tool classifies as proper nouns the words that start with capital letters.

- *Search of terms in an economic bilingual glossary*. We use a glossary with 18,724 entries in Spanish and English. 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[3]) 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 purposes.

- *Application of Eurovoc thesaurus*. Eurovoc[4] 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 4.2 exists in 21 official languages of the European Union and is used in different information retrieval, text clustering and classification projects [17]. The thesaurus is a structured list of more than 6,600 descriptors and 127 microthesauri in 21 thematic fields, for example, politics, international relations, economics, trade, finance, business and competition, employment and working conditions, production, technology and research, energy, and industry. The Office for Official Publications of the European Communities has granted us a non-exclusive licence to use it in our research. Our web application searches Eurovoc descriptors inside the document text and associates the most appropriate microthesauri to the document.

LANGUAGE-INDEPENDENT DOCUMENT REPRESENTATION AND SIMILARITIES

In order to apply text mining and cluster analysis techniques to the collection of multilingual documents, we can divide the process into three steps: feature extraction, similarity computation and grouping. The feature extraction allows us to obtain a language-independent representation of the documents in the corpus. We employ a modified vector-space model to represent a document. Traditionally, every document is represented by a vector of weighted terms (features) [14, 1]. Using word based features is the most popular and, despite its simplicity, a very effective feature construction method. The Vector Space Model (VSM) has become a standard tool in IR systems since its introduction by Salton [14]. Given a set of index terms, not all of which are equally useful for describing the contents of a particular document, numerical weights $w_{ij} \geq 0$ are assigned to each index term or keyword k_i of a document d_j.

According VSM a document $\mathbf{d}_j$ is represented by a vector:

$$\mathbf{d}_j = (w_{1j}, w_{2j}, ..., w_{tj}) \tag{1}$$

Let N be the number of documents in the collection and t the number of index terms or keywords considered; a t×N matrix $\mathbf{W} = (w_{ij})$ can be constructed, where each column represents a document, and each entry represents the weight of a keyword in a document. Matrix **W** is known as *term-document matrix*. The vectors can be normalized in order to facilitate the compute of similarities.

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 nouns (persons, places and organisations) and words in the native language. Using these features, a document is represented by four vectors. Whenever a new document is inserted in the database, it is converted into such vector representation. Because we are using bilingual linguistic resources such as the IMF glossary and the Eurovoc thesaurus, the obtained vector representations are language-independent. Each coordinate in the vector has associated to the primary key of the corresponding record in the table, instead of a word or sentence in a specific language.

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, known as the *tf factor*, measures the raw frequency of the term inside the document, and the second one, usually referred to as the *inverse document frequency* or the *idf factor*, is motivated by the fact that a term which appears in many documents is not very useful for distinguishing

[3] http://www.imf.org/external/np/term/index.asp

[4] http://europa.eu/eurovoc/

documents. The *Term Frequency Inverse Document Frequency weighting* (TF-IDF) is defined by

$$w_{ij} = f_{ij} \times idf_i = \frac{freq_{i,j}}{\max_p \ freq_{p,j}} \log \frac{N}{n_i} \tag{2}$$

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 is 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.

The web application computes the weights for the four feature vectors of the document and allows us to obtain a complete report about the features of the document, as shown in Figure 1.

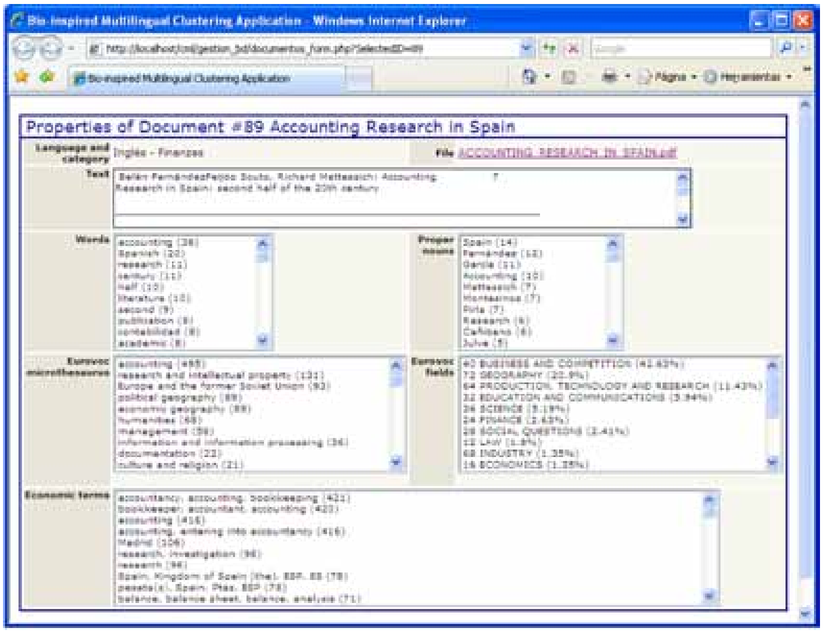
With the four vectors associated to any document, we can estimate the similarity score between a pair of documents (**p**,**q**). The similarity function basically consists of four similarity scores, namely, similarity of entries in the glossary, similarity of Eurovoc microthesauri, proper nouns similarity and terms similarity, which are combined using a convex linear combination:

$$\begin{aligned} SimML(\mathbf{p},\mathbf{q}) = {} & \lambda_1 \ Sim(V_{glossary}(\mathbf{p}), V_{glossary}(\mathbf{q})) + \\ & \lambda_2 \ Sim(V_{Eurovoc}(\mathbf{p}), V_{Eurovoc}(\mathbf{q})) + \\ & \lambda_3 \ Sim(V_{pnouns}(\mathbf{p}), V_{pnouns}(\mathbf{q})) + \\ & \lambda_4 \ Sim(V_{words}(\mathbf{p}), V_{words}(\mathbf{q})) \\ \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1 \quad & \text{with} \quad \lambda_i \geq 0 \end{aligned} \tag{3}$$

The classical cosine similarity score, called angular separation, is used. Given two vectors v(**p**) and v(**q**), representing two documents, the angular separation is defined by the following expression:

$$Sim(v(\mathbf{p}), v(\mathbf{q})) = \cos(\sigma) = \\ = \frac{\mathbf{p} \circ \mathbf{q}}{\|\mathbf{p}\| \times \|\mathbf{q}\|} = \frac{\sum_{i=1}^{t} w_{ip} w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ip}^2} \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \tag{4}$$

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. The metric (4) is relatively simple to compute and experimental results have indicated that it tends to lead to better results than Euclidean distance.


Bio-inspired Multilingual Clustering Application - Windows Internet Explorer
http://localhost/cml/gestion_bd/documentos_form.php?selectedID=89
Bio-inspired Multilingual Clustering Application
Página
Herramientas

Properties of Document #89 Accounting Research in Spain

Language and category: Inglés - Finanzas

File: ACCOUNTING_RESEARCH_IN_SPAIN.pdf

Text: Belén FernándezFeijóo Souto, Richard Mattessich: Accounting Research in Spain: second half of the 20th century

Words:
accounting (28)
Spanish (20)
research (11)
century (11)
half (10)
literature (10)
second (9)
publication (8)
contabilidad (8)
academic (8)

Proper nouns:
Spain (14)
Fernández (12)
García (11)
Accounting (10)
Mattessich (7)
Montesinos (7)
Pirla (7)
Research (6)
Cañibano (6)
Julve (5)

Eurovoc microthesaurus:
accounting (455)
research and intellectual property (131)
Europe and the former Soviet Union (92)
political geography (89)
economic geography (89)
humanities (68)
management (59)
information and information processing (36)
documentation (22)
culture and religion (21)

Eurovoc fields:
40 BUSINESS AND COMPETITION (42.62%)
72 GEOGRAPHY (20.9%)
64 PRODUCTION, TECHNOLOGY AND RESEARCH (11.43%)
32 EDUCATION AND COMMUNICATIONS (5.94%)
36 SCIENCE (5.19%)
24 FINANCE (2.63%)
28 SOCIAL QUESTIONS (2.41%)
12 LAW (1.8%)
68 INDUSTRY (1.35%)
16 ECONOMICS (1.35%)

Economic terms:
accountancy, accounting, bookkeeping (421)
bookkeeper, accountant, accounting (422)
accounting (416)
accounting, entering into accountancy (416)
Madrid (106)
research, investigation (96)
research (96)
Spain, Kingdom of Spain (the), ESP, ES (78)
peseta(s), Spain, Ptas, ESP (78)
balance, balance sheet, balance, analysis (71)

Figure 1. Document features information.

Numerous other distances have been proposed, but the cosine of the angle is the most commonly used.

ANT-BASED CLUSTERING ALGORITHMS

Ant-based sorting and clustering algorithms, introduced in [3] and [11], were among the first metaheuristics to be inspired by the behaviour of ants. 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 web application implements an ant-based clustering algorithm. In the *ant clustering algorithms* [12, 8], 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 to its neighbours in the grid, the probability of picking it up is low; however, if the similarity is low the artificial ants are highly likely to pick it up and will look for a good position in the grid to reallocate it. These probabilities are defined using the following expressions:

$$P_{pick}(\mathbf{d}_i) = \left(\frac{k^+}{k^+ + f(\mathbf{d}_i)} \right)^2$$
$$P_{drop}(\mathbf{d}_i) = \left(\frac{f(\mathbf{d}_i)}{k^- + f(\mathbf{d}_i)} \right)^2 \tag{5}$$

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:

$$f(\mathbf{d}_i) = \frac{1}{\sigma^2} \sum_{\mathbf{d}_j \in \Omega} \frac{SimML(\mathbf{d}_i, \mathbf{d}_j)}{\alpha} \tag{6}$$

After picking-up 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 are likely to be dropped in neighbouring positions in the grid and a graphical visualization of the clusters is obtained, as can be observed in Figure 2. In this figure colours represent thematic categories in the corpus.

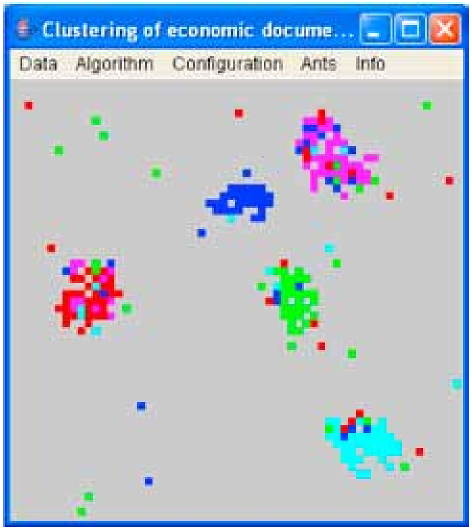


Figure 2. Ant clustering process in the bidimensional grid. Groups of related documents can be observed.

Ant-based clustering has been applied in various areas: commerce, circuit design, text mining, and different studies offer proof that ant-based clustering is a robust and viable alternative, compared with more classical techniques [8]. A study on the performance of ant-based clustering can be found in [7].

QUALITY MEASURES

The clustering results of the proposed algorithm on the bilingual test corpus have been compared with the results of the classical k-means algorithm. Different measures can be used for evaluating the quality of system output in text mining tasks; in [18] an overview of prevalent measures can be found, and their strengths and weaknesses are discussed. [6] presents the state of the art concerning quality measures for data mining and summarizes recent developments and original research on this topic.

In our comparative study we used 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 [13] 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:

$$P(i,j)=\frac{n_{ij}}{n_j} \qquad R(i,j)=\frac{n_{ij}}{n_i} \tag{7}$$

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:

$$F=\sum_i \frac{n_i}{N}\max_j\{F(i,j)\} \quad with$$
$$F(i,j)=2\frac{P(i,j)R(i,j)}{P(i,j)+R(i,j)} \tag{8}$$

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

Finally, the entropy [15] tells us how homogeneous a cluster is, its optimal value being zero. The entropy of a cluster and the overall entropy are defined by the following expressions.

$$E_j=-\sum_i \frac{n_{ij}}{n_j}\log\left(\frac{n_{ij}}{n_j}\right); \qquad E=\frac{1}{N}\sum_j n_j E_j \tag{9}$$

PARAMETER SETTINGS

The proposed algorithm requires a number of different parameters to be set. Table 1 shows the parameter values used in the experiments.

Table 1. Parameter settings.

AC algorithm	
Similarity coefficients (λ_i)	0.45, 0.45, 0.05, 0.05
Grid size	60 x 60
Population size (number of ants)	25
Picking-up parameter (k+)	0.0015
Dropping parameter (k-)	0.05
Neighbourhood size	5 x 5
Scale similarity parameter (α)	1.0
Maximum length step	25
Short-term memory size	20

RESULTS

The algorithm was 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 proposed ant-based algorithm. There are several other clustering algorithms presented in the literature, but we decided to use k-means to make this comparison because it is a simple and fast algorithm. The following observations can be made from Table 1.

The AC algorithm performs very well under all quality measures and outperforms the k-means algorithm.

As regards the computational time, it is worth observing that AC algorithm scales linearly and becomes faster than the k-means algorithm on big document collections.

Another interesting feature of the AC algorithm is its robustness to the effects of outliers within the collection. For example, in the case of documents that cannot be clearly classified in a particular thematic category, the algorithm can identify and does not include them in any cluster.

Table 2. Comparative quality values.

	K-means	**Ant Clustering**
F-measure	0.6380	0.6669
Entropy	0.9305	0.9203
Purity (%)	65.10	65.42
Execution time	4945	5581

The k-means algorithm needs a priori knowledge of the correct number of clusters; 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 750,000 ant basic operations in a run of the AC algorithm. In this image the cluster structure of the corpus can be clearly observed. In the figure each point on the grid represents a document, and its colour represents the thematic category of the associated document in the corpus.

Table 3 summarizes the clusters obtained in the execution. As can be observed, 5 groups of related documents are obtained, each one corresponding to a thematic category in the corpus. The degree of purity of the clusters is quite high, although several documents are not correctly classified. It is worth observing the difficulty of a clustering process in a corpus of documents in very close areas, as happens in our corpus, where a marketing document about sales forces, for example, can also be included in the category of human resource management.

Table 3. Summary of clustering solutions obtained by AC algorithm.

	Number of documents						
	ACC	MKT	HRM	ECO	INF	Categ.	Purity (%)
C1	2	10	7	4	39	INF	62.90
C2	1	29	0	0	0	MKT	96.67
C3	4	3	32	3	11	HRM	60.38
C4	3	3	4	33	0	ECO	76.74
C5	39	3	4	3	0	ACC	79.59
Rest	1	2	3	7	0		

Figures 3 and 4 show other execution results of the algorithm; the cluster structure is also observed.

CONCLUSIONS

In this work we have presented an application that allows the user to manage Spanish-English collections of economics documents. The application extracts features of the documents using linguistic tools, such as a specialized glossary and thesaurus, and computes similarities between documents written in different languages using four types of features. The user can change the parameters that define the relative importance of the different types of features, and analyse the effect of these changes over the similarity between documents. The clustering algorithm implemented in the application is based on observed behaviours in ant colonies. Ant clustering algorithms have been applied in document retrieval contexts [9], but no application in multilingual clustering has been found. The experimental results, using a corpus of 250 Spanish-English documents, show a good performance of this methodology.

ACKNOWLEDGEMENTS

This work has been partially supported by SistIng-Alfa project, ref: ALFA II-0321-FA of the European Union and Project Ref. TIN2006-13615 of the Spanish Ministry of Education and Science.

We gratefully acknowledge the support of the Office for Official Publications of the European Communities and the Language Services of the International Monetary Fund, for the use of original editions of Eurovoc Thesaurus and IMF multilingual Glossary, respectively.

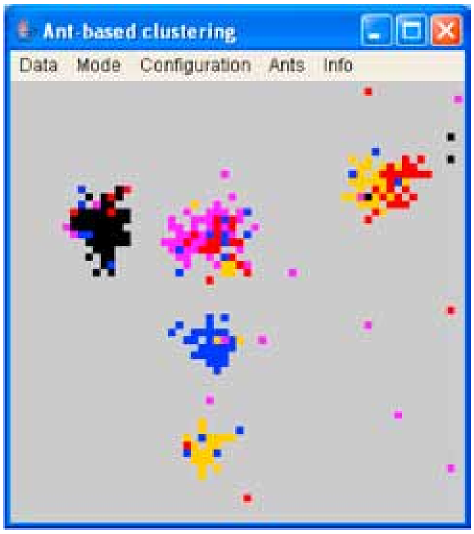


Figure 3. Clusters after 750,000 ant basic operations.

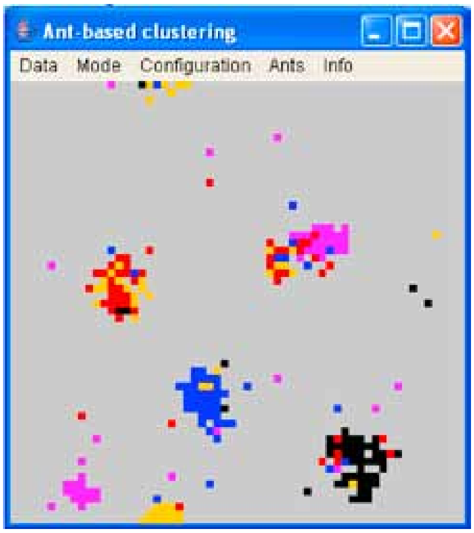


Figure 4. Clusters after 750,000 ant basic operations (another run).

REFERENCES

[1] R. Baeza and B. Ribeiro. "Modern Information Retrieval". Addison Wesley. 1999.

[2] H. Chen, J. Kuo and T. Su. "Clustering and visualization in a multilingual multi-document summarization system". In Advances in Information Retrieval. Proceedings of the 25th European Conference on IR Research, pages 266-280. 2003.

[3] J. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain and L. Chretien. "The dynamic of collective sorting robot-like ants and ants-like robots". In Proceedings of the First Conference on Simulation of Adaptive Behavior, pp. 356-363. 1990.

[4] M. Dorigo and T. Stützle. "Ant Colony Optimization". Bradford - MIT Press. 2004.

[5] D. Evans, K. McKeown and J. Klavans. "Similarity-based multilingual multi-document summarization". Technical report, Computer Science Technical Reports (num. CUCS-014-05). University of Columbia. 2005.

[6] F. Guillet and H.J. Hamilton. "Quality Measures in Data Mining". Studies in Computational Intelligence. Vol. 43. Springer. 2007.

[7] J. Handl and M. Dorigo. "On the performance of ant-based clustering". In Proceedings of the 3rd International Conference on Hybrid Intelligent Systems. 2003.

[8] J. Handl, J. Knowles and M. Dorigo. "Ant-based clustering and topographic mapping". Artificial Life. Vol. 12, pp. 35-61. 2006.

[9] J. Handl and B. Meyer. "Improved ant-based clustering sorting in a document retrieval interface". In Proceedings of 7th International Conference on Parallel Problem Solving from Nature. Lecture Notes in Computer Science 2439, pp. 913-923. Springer-Verlag. 2002.

[10] W. Lam, M. Meng, K. Wong and J. Yen. "Using contextual analysis for news event detection". International Journal of Intelligent Systems. Vol. 16, pp. 525-546. 2001.

[11] E. Lumer and B. Faieta. "Diversity and adaptation in population of clustering ants". In Proceedings of 3rd International Conference on Simulation of Adaptive Behaviour: From Animals to Animats, pp. 501-508. 1994.

[12] N. Monmarché. "On data clustering with artificial ants". In Freitas, A., editor, AAAI-99 & GECCO-99 Workshop on Data Mining with Evolutionary Algorithms: Research Directions, pp. 23-26. 1999.

[13] C.J. van Rijsbergen. "Information Retrieval". Butterworths. 1979.

[14] G. Salton. "The SMART Retrieval System-Experiments in Automatic Document Processing". Prentice Hall. 1971.

[15] C. Shannon. "A Mathematical Theory of Communication". Bell System Technical Journal. Vol. 27, pp. 379-423. 1948.

[16] M. Spitters and W. Kraaij. "Unsupervised clustering in multilingual news streams". In Proceedings of the LREC 2002 Workshop: Event Modelling for Multilingual Document Linking, pp. 42-46. 2002.

[17] R. Steinberger, B. Pouliquen and C. Ignat. "Navigating multilingual news collections using automatically extracted information". Journal of Computing and Information Technology, CIT. Vol. 13, pp. 257-264. 2005.

[18] H. Suominen. "Performance Evaluation Measures for Text Mining". Handbook of Research on Text and Web Mining Technologies. IGI Global, pp. 724-747. 2009.