15th Triennial World Congress, Barcelona, Spain

AN APPROACH TO DOCUMENT ENGINEERING AT FUNDACIÓN HULLERA VASCO-LEONESA

Ángel Alonso, José R. Villar, Carmen Benavides, Isaías García, and Francisco Rodríguez

[dieaaa, diejvf, diecbc, dieigr, diefrs]@unileon.es
Systems Engineering and Control Group
Dept. of Electrical and Electronic Engineering, University of León
Edificio Tecnológico, Campus de Vegazana s/n, 24071 León (SPAIN)

Abstract*:* Fundación Hullera Vasco-Leonesa is a company with a documental department responsible for managing the bibliographic information the company uses. That department manually elaborates and distributes periodic documents (press bulletins, environmental dossiers, etc). This paper describes an intelligent multiagent system as a way to solve the work handled by this department. The objective was to design and implement a digital library with all the tasks needed, like query management, automatic design and generation of electronic documents, selective information distribution, etc. *Copyright © 2002 IFAC*

Keywords: Computer Applications, Agents, Intelligent Knowledge Based Systems, Distributed Artificial Intelligence, Learning Systems, Reasoning.

1. INTRODUCTION

The multiagent system faced in this work is part of a research project founded by Junta de Castilla y León (a local government agency at Spain), with identification number LE038/UA. The aim of this project was to implement the management tasks involved in the documental system at Fundación Hullera Vasco-Leonesa (hereinafter referred to as "the *company*"), which users are not only the rest of the company but also the local small and medium enterprise (SMB). The use of Knowledge Engineering and Management techniques was planned with the premise that the final application must avoid to represent a change in the way the *company* does his work nowadays, but there must be an improvement in the performance obtained.

The documental department (hereinafter referred to as "the *department*") is responsible for the documental system in this *company*. The tasks related to this *department* are the gathering of all kind of bibliographic references (monographs, reviews, articles, etc), their classification and later storing. The *department* also elaborates periodic documents like a press bulletin and an environmental dossier and distributes the information selectively. Finally, it resolves every bibliographic query received. In figure 1 the description of the proposed system is shown.

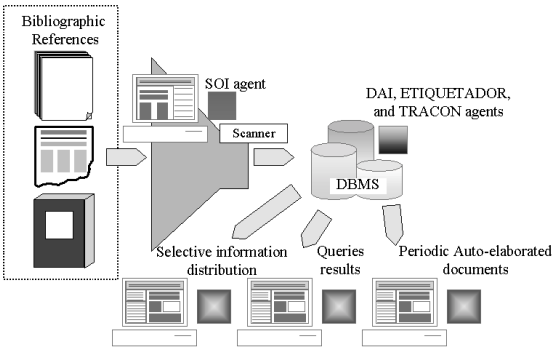

Fig. 1. Graphical description of the project.

The solution proposed to this problem is based on the application of distributed artificial intelligence by means of the multiagent technology (Weiss, 1999; Rao and Georgeff, 1995). The use of a system of agents was decided because of the high level of modularity found in the study of the tasks involved in the documental *department*, making sense to implement those tasks as agents, with or without an intelligent behaviour as necessary.

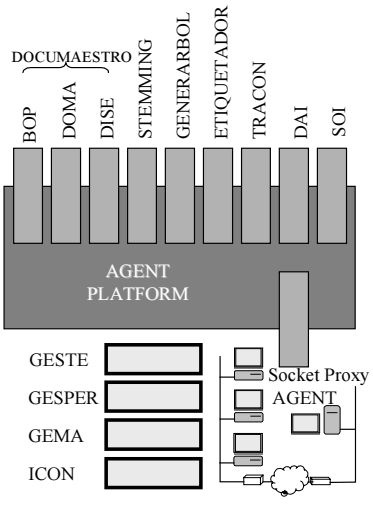

Fig. 2. Planned agent system.

Figure 2 shows a description of the agent system planned. As it can be seen, there are several kinds of agents, each one with different behaviours as follows. There are two types of interfaces: a classical software interface and a world wide web interface. The former is a scanning and optical character recognition interface (SOI) to introduce bibliographic references to the system. The latter is composed of several web interfaces: GESPER is a user management agent, GESTE is a thesaurus management agent, ICON is the query interface agent and GEMA is the document design and management agent.

The rest of the system includes a queries parser agent named TRACON, a morphological parser agent named STEMMING (Porter, 1980), a text document classifying agent named ETIQUETADOR, a database interface agent named DAI, and a general purpose agent (GENERARBOL). Finally, there is an agent named DOCUMAESTRO devoted to elaborate a periodic content-predefined document which will be instantiated for every periodic document defined through the GEMA agent.

There will also be the agents that correspond to the JADE platform (Bellifemine, *et al* 1999), like AMS or DF. There could be more than one instance of the same type of agent running at any time except for the AMS, which is unique in the FIPA platform (Foundation for Intelligent Physical Agents, 1997).

As was explained before, this approach faces the tasks handled by the *department* of the *company*. As it was necessary to keep unchanged the way they work as far as possible, and to maintain their modus operandi as well, the decision was to introduce knowledge and management engineering, trying to develop an ad hoc system (Reese, 2000).

As the department didn't have any database application to log all the work reported it was impossible to apply data mining techniques and so knowledge engineering was applied instead (Palma, *et al*., 2000). The *company* decided to designate two experts to work in the project, so the knowledge sessions were held. Hereinafter, all the decisions about the design of the system were taken using the knowledge extracted from the experts, obviously with the premise of preserving as much as possible their modus operandi. In the same way, it will be referred as *article* any book, article newspaper's report needed to be introduced as a bibliographic reference to the system, and it will be referred as *document* any well-formed, structured, periodic electronic-generated document.

In next sections the description of the different problems faced in this project are explained along with the agent or agents responsible. In Section 2 the classification of articles and document representation implemented is revised. Section 3 explains the design of the automatically generated document and describes its elaboration. Finally, Section 4 introduces the query and selective distribution system. In Section 5 conclusions and future work and research are detailed.

2. THE ARTICLE REPRESENTATION AND CLASSIFICATION. THE FEED SUBSYSTEM

The first task assumed by the group was to decide how to represent and classify the articles introduced to the system. There were some important ideas about the problem to solve. First, the experts did not use a hierarchical set of terms - a thesaurus -. Surprisingly, although they usually used a not fixed, manually selected set of terms, they agreed the thesaurus would be an important tool to make their work.

Moreover, when the department received a query to analyse, they recognised it was necessary to phone the inquirer to fully understand and translate the originally received query to the department's query language and finally solve it. The employees of the company couldn't easily find by themselves the articles needed as the terms used by the documental department and the rest of departments of the company were different.

The solution proposed is based on the following points:

- the articles are represented using inverted file and word frequency, allowing to use the

content of a reference in query analysing and text classification (Sebastiani, 1999; Yang *et al*., 1997),

- the use of a well defined thesaurus on which every term possesses a semantic corpus as proposed in (Riloff and Shepperd, 1997). As a better approach to the problem, every word in the corpus of a term have a credibility,
- the text classifier is a semiautomatic boost technique of two methods: a text classifier as explained in (Riloff 1991, 1996) and (Riloff and Shepperd, 1997) with the modifications due to the use of credibility of every word in a corpus, and a naive CBR (Kolodner, 1993) classification method similar to the one applied as a query analyser method explained in (Ramírez and Coley, 1995). The context of the article to classify (Billsus and Pazzani, 1996) is represented as the user who is classifying plus the content of the article itself.

A list of stop words and a stemming algorithm based on the ideas expressed in (Porter, 1980) with the modifications due to the differences between English and Spanish languages is used, so the reference representation by means of a bag of words excludes the ones in the stop word list and include the words' root. Finally, the text classifier use the thesaurus in a similar fashion to that expressed in (McCallum *et al*., 1998).

The classifier agent is implemented in a modular way so it is possible to use another text classification method as long as the *company* decides to rely this task to an automated machine. Text classification methods include those that treat with word disambiguation (Dagan, *et al*., 1994; Dagan, *et al*., 1999; Krovetz and Croft, 1992; Lin, 2000), machine learning classical methods (Baker and McAllum, 1998; Han, *et al*., 2001; Parekh, *et al*., 2000), probabilistic machine learning methods (Dagan, *et al*., 1997; Nigam, *et al*., 2000), etc.

The feed and classification subsystem is designed as follows. It includes the SOI agent as the scan and OCR interface, the ETIQUETADOR agent, the STEMMING agent, the data access -DAI- agent, and the GENERARBOL agent. When an article is introduced to the system, a user -one of the deparment's classifying experts- fills the article's data, scans as needed, and asks the ETIQUETADOR agent for terms to use as classification tags. The STEMMING is used by the ETIQUETADOR when it analyses the article's data. Every term used to classify an article is selected manually by the user among the ones proposed by the ETIQUETADOR agent or directly from a thesaurus hierarchical view available at the SOI through the GENERARBOL agent. When all article's data is gathered and classified it can then be stored through the DAI agent. The whole article, along with its classification, inverted file and Portable Document Format (Adobe, 1996) version, is stored in the database. In figure 3 the feed and classification subsystem is presented.

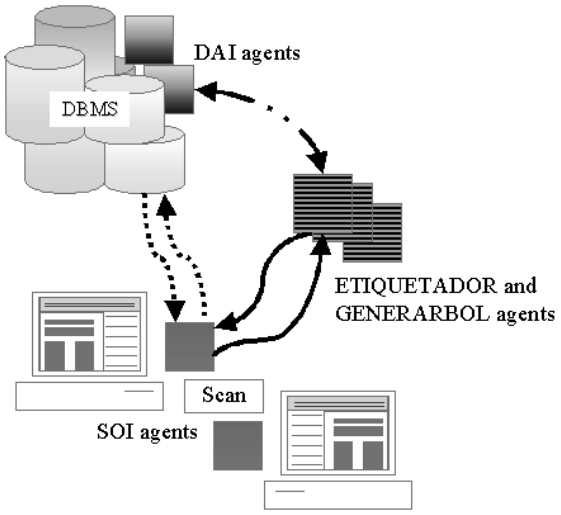

Fig. 3. Feed and classification subsystem.

3. DESIGN OF DOCUMENTS AND THE AUTOMATED GENERATION PROCESS

In this section the process that the *department* uses to elaborate a document is explained. Initially the *department* elaborated two kind of documents: a press bulletin and an environmental dossier.

The first one was a plain text document with the title and abstract of each included report, those reports were grouped by locality, county, administration region, domestic and foreign. The news were manually selected and grouped, typed and printed. In some cases, and only for a very few people, this bulletin was attached with photocopy of some reports. This document was daily generated.

The environmental dossier was a more complex, monthly-edited document. In this case, the document had three sections: the First Page, the Most Relevant and the Document Content. The First Page section contained only the most highlights news, the most important news in this document. The Most Relevant section was a plain listing of the articles considered to be important. Finally, the Content section is a copy of all of the articles about the environment that appeared in the past month and they considered to be important. Each section has a not fixed index, and each index element has its own content.

It was tried to determine the way an article was assigned to an index item by means of knowledge sessions along with the experts. Finally it was found out that this process was content driven. The task was therefore to find out how the content of an index was fixed, and how they introduce a new index element to

the section, and a new section to the document. Another task was to formalize the method used to find the articles of every section of the document as the First Page or Most Relevant sections. The proposed model for the document design has resolved as far as possible all of those items in a similar way the department manually does. All generated document is an instance of this model.

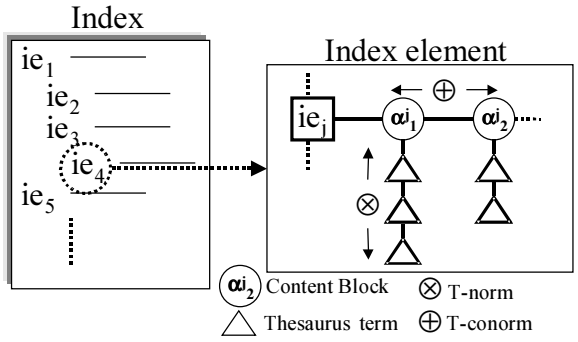

Fig. 4. Content for an index element.

In the approach proposed, each document has an index made from hierarchical index elements. Each element has a defined content as a weighted combination of disjunctive normal form of thesaurus terms. Figure 4 is a representation of the content of an index element.

As it can be seen, there is a hierarchy of index elements, so for example, ie_4 is more specific than ie_3. The triangles represent thesaurus terms, which are operated with a T-norm. Every block, which content is defined by a disjunction of thesaurus terms have a weight (α^j_i) and all of these block are operated with a T-conorm for each index element. Experts have the responsibility for the correct design of index elements and their content.

Once the index was formalized, the management of the different sections was faced. The formalized document's model assumes that it is composed of

- Document's title.
- Rules to format any kind of article to be included in the document.
- The sections of the document.
- A list of articles to include.
- Other characteristics, like the periodicity.

As there are several kinds of articles to include in a document, it is needed to gather different bibliographic data for each one. In the same way, the bibliographic data wanted to appear in the document is different for each kind of article. In this proposal, a document stores the bibliographic data to represent every kind of article, this representation can be defined in a different way for each document. As an example, a document may need for a book its title, authors, year of publication, editor and ISBN, and for an article its title, authors, original source, year, volume, number and page.

As it was previously stated, each document has one or more sections, every section of a document is affected with some attributes like:

- A title.
- An index.
- Rules to assign articles to one or more section index elements.
- Rules to assign the section order.

The rules to assign articles to one or more section index elements evaluate the articles to include in the document, and according to the similarity of the content of the article with the content of each index element the article is assigned to one or more index element. It can be defined for a document to include only once an article or to repeat it in several sections until a maximum. The rules to assign the order to the section implies how the sections are ordered and where an article must appear or not. There are also, some defined thresholds, like the maximum number of times an article is repeated in different sections of the same document, etc. It is important to notice that all rules in the document's design must be introduced by the expert at a specific interface the system has, and those designs can be changed dynamically. In this order, the interface agent GEMA supports this ability. GEMA is a web interface agent, which is responsible for the design, storage, modification, visualization and elimination of any automatically generated document.

Once a document is designed it is possible to launch an instance of the DOCUMAESTRO agent to manage it. DOCUMAESTRO is the one which takes all the articles to be introduced in the document and reorder, classify and assign the articles to each section of the document. A Rete (Forgy, 1982) inference machine is used in those tasks. The Rete machine used was Jess (Friedman-Hill, 2000), a CLIPS (Giarratano, 1993) Java version.

The method used by DOCUMAESTRO to organize the articles is based on the classes and credibility an article belongs to. Through a rule-based system a version of the k-nearest neighbour algorithm is implemented with the index elements and their content blocks as classes and the articles as the subjects to classify. The similarity of an article to each class is computed using the content blocks of the index elements, and setting the similarity between the classes which the article belongs to and any thesaurus term in the content block.

Finally, the arranging of the document is implemented using the index elements, their assigned articles, the rules to format each type of article to be included in the document, and the output medium generated. The output is HTML documents, stored at

the web server. Also, a document can be distributed by e-mail.

4. QUERIES AND SELECTIVE INFORMATION DISTRIBUTION SUBSYSTEM

The last subsystem implemented is the queries management and selective information distribution subsystem which is the one that:

- Resolves the users' queries.
- On-line sends every user or department the references found to be relevant.

To resolve the users' queries this subsystem employs a similar algorithm to the one explained for DOCUMAESTRO, the k-nearest neighbour algorithm is applied to select the better results. In this case, the query is represented as a document with only one section which index is defined as the user profile. Every user can instantiate jobs on which many users work together, and define the profiles for these jobs. In this case, the index is conformed with the job profile. These profiles are updated with the feedback of the user (Billsus and Pazzani, 1996; Liu, *et al.*, 1999). The task is to translate the user query to a system query and generate a document with all the bibliographic references the database has.

To manage the queries there are two agents: the ICON and the TRACON. The former is a query interface to allow the user to look for articles. The latter is the query analyser, the responsible for the translation of the query and updating the profiles.

The selective information distribution is implemented as a document and is managed with an instance of DOCUMAESTRO. The articles introduced to the system are marked to be distributed and the receivers are designated. This agent distributes the information on a daily basis, as soon as it is demanded by the SOI interface or 24 hours after the last action of this DOCUMAESTRO agent.

5. CONCLUSIONS, FUTURE WORK AND RESEARCH

To date, the system has been implemented and is under testing at the company. The idea is to use the system in its real environment almost for six months to obtain the final results. Anyway, some ideas can be extracted. First, a multiagent system is a good platform to implement this type of digital library due to the robustness and modularity reached. Second, intelligent modules could be implemented by code, by rule-based systems or by stored procedures at the data base management system, looking for the better performance. Finally, the knowledge engineering applied results in a good implementation of the knowledge in the company so the implemented system do not represent a big change in the way they work but it has improved the document handling process in the company. As Fundación Hullera Vasco-Leonesa manages information for the local SMBs, the project goodness is the social profits obtainable.

About the future work, it may be cited the inclusion of meta-knowledge at middleware agents on top of the database management system available for the management of different domain documental systems, providing migration and mobility features to the query agents, the integration of this system with other solutions currently existing and the study and implantation of security in mobile agent systems, all of them themes related to the current research area in the group. Another future work and research is to introduce intelligent modules to determine automatically the management to be applied to an article introduced to the documental system, without any human intervention.

REFERENCES

Adobe Systems Inc. (1996) Portable Document Format Reference Manual.

Baker, L. D. and A. K. McCallum (1998). Distributional Clustering of Words for Text Classification In: *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (W. Bruce Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobel Eds.), 96-103, ACM Press, New York, US.

Bellifemine, F., A. Poggi and Y. Rimassa (1999). JADE- A FIPA-compliant agent framework. In: *Proceedings of PAAM'99*, pp. 97-108, London.

Billsus, D. and M. Pazzani (1996) Revising User Profiles: the Search for Interesting Web Pages. In: *Proceedings of the Third International Workshop on Multistrategy Learning (MSL '96).* AAAI Press.

Dagan, I., Y. Karov and D. Roth (1997) Mistake-Driven Learning in text Categorization. In: *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing* (C. Cardie and R. Weischedel eds.), 55-63, Association for Computational Linguistics, Somerset, New Jersey.

Dagan, I., L. Lee and F. Pereira (1999) Similarity-Based Models of Word Cooccurrence Probabilites. *Machine Learning Journal,* **34**, 43-69.

Dagan, I., F. Pereira and L. Lee (1994) Similarity-Based Estimation of Word Coocurrence Probabilities. In: *Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics*, 272-278.

Forgy, C. L. (1982) RETE: A Fast Algorithm for the Many Pattern/Many Object Pattern Matching Problem. Artificial Intelligence **19**, 17-37.

Foundation for Intelligent Physical Agents (1997) Foundation for Intelligent Physical Agents. Specifications. 1997.In: *http://www.fipa.org* , [Checked on May 28th, 2001]

Friedman-Hill, E. J. (2000) Jess, The Java Expert System Shell. Distributed Computing Systems, Sandia National Laboratories, Livermore, CA.

Giarratano, J.C. (1993). CLIPS User's Guide (CLIPS Version 6.0). NASA L.B. Johnson Space Center, Information Systems Directorate, Software Technology Branch.

Han, E. H., G. Karypis and V. Kumar (2001) Text Categorization using weight adjusted k-nearest neighbor classification. In: *Proceedings of the 5th Pacific-Asia Conference of Knowledge Discovery and Data*.

Kolodner, J. (1993). Case-Based Reasoning ; Morgan Kauffmann Publishers, -Inc. San Mateo, California, USA

Krovetz, R. and W. B.; Croft (1992). Lexical Ambiguity and Information Retrieval. *ACM Transactions on Information Systems*,**10**, 115-141

Lin, D. (2000). Word Sense Disambiguation with a Similarity-Smoothed Case Library. *Computer and Humanities*, **34**, 147-152.

Liu, B., W. Hsu, L. Mun, and H. Lee, H. (1999). Finding Interestin Patterns Using User Expectations. *IEEE Transactions on Knowledge and Data Engineering*, **11-6**, 817-832.

McCallum, A., R. Rosenfeld, T. Mitchell, T.and Y. Ng. Andrew (1998). Improving text classification by shrinkage in a hierarchy of classes. In: *International Conference on Machine Learning*, 359-367. Morgan Kaufmann.

Nigam, K., A. McCallum, S. Thrun and T. Mitchell (2000). Text Classification form labeled and unlabeled documents using EM, *Machine Learning*, **39-2**, 103-134.

Palma, J. T., E. Paniagua, F. Martín and R. Martín (2000). Ingeniería del Conocimiento. De la Extracción al Modelado del Conocimiento, *Revista Iberoamericana de Inteligencia Artificial*, **11**, 46-72.

Parekh, R., J. Yang and V. Honovar (2000). Constructive Neural-Network for Learning Algorithms for Pattern Classification, *IEEE Transactions on Neural Networks*, **11-2**, 436-451.

Porter, M. F. (1980). An algorithm for suffix stripping, *Program*, **14-3**, 130-137.

Ramírez, C. and R. Cooley (1995). Case-Based Reasoning Model Applied to Information Retrieval, In: *IEE Colloquium on Case Based Reasoning: Prospects for Applications*, 9/1-9/3.

Rao, A. S. and M. P. Georgeff, M. P. (1995). {BDI}-agents: from theory to practice. In: *Proceedings of the First Intl. Conference on Multiagent Systems, San Francisco 1995*,

IEE Colloquium on Case Based Reasoning: Prospects for Applications, 312-319, MIT Press, San Francisco, CA.

Reese Hedberg, S. (2000) After desktop computing: a progress report on smart environments research, *IEEE Intelligent Systems*, **15-5**, 7-9.

Riloff, E. (1991). Little Words Can Make a Big Difference for Text Classification. In: *Proceedings of the 18th Annual International ACM SIgIR Conference on Research and Development in Information Retrieval* (E. A. Fox, P. Ingwersen and R. Fidel eds.), 130-136, ACM Press, Seattle, US.

Riloff, E. (1996). Using learned extraction patterns for text classification. In: *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing* (S. Wermter, E. Riloff and G. Scheler eds.), 275-289, Springer-Verlag.

Riloff, E. and J. Shepherd (1997) A Corpus-Based Approach for Building Semantic lexicons. In: *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing* (C. Cardie and R. Weischedel eds.), 117-124, Association for Computational Linguistics, Somerset, New Jersey

Sebastiani, F. (1999) Machine Learning in Automated Text Categorization. *Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche*, Pisa, IT.

Weiss, G. (1999). Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. MIT Press, London, England.

Yang, Y., O. Jan and J. O. Pedersen (1997), A Comparative Study on Feature Selection in Text Categorization. In: *Proceedings of the Fourteenth International Conference on Machine Learning*, 412-420, Morgan Kaufmann.