

# News Retrieval through a MultiAgent System

Andrea Addis, Giuliano Armano, Francesco Mascia, and Eloisa Vargiu

University of Cagliari

Piazza d'Armi

I-09123, Cagliari, Italy

Email: {addis,armano,f.mascia,vargiu}@diee.unica.it

**Abstract**—The continuous growth of information sources on the web, together with the corresponding volume of daily-updated contents, makes the problem of finding news and articles a challenging task. This paper presents a multiagent system aimed at creating press reviews from online newspapers by progressively filtering information that flows from sources to the end user, so that only relevant articles are retained. Once extracted, newspaper articles are classified according to a hierarchical text categorization approach. Moreover, an optional feedback provided by the user is exploited to improve the overall performances. The system is built upon a generic multiagent architecture that supports the implementation of personalized, adaptive and cooperative multiagent systems devised to retrieve, filter and reorganize information in a web-based environment.

## I. INTRODUCTION

The World Wide Web offers a growing amount of information and data coming from different and heterogeneous sources. As a consequence, it becomes more and more difficult for Internet users to select contents according to their interests, especially if contents are frequently updated (e.g., news, newspaper articles, Reuters, RSS (Really Simple Syndication) feeds, and blogs). Supporting users in handling the enormous and widespread amount of web information is becoming a primary issue. To this end, several online services have been proposed (for instance Google News <sup>1</sup> and PRESSToday <sup>2</sup>). Unfortunately, they allow users to choose their interests among macro-areas (e.g. *economics*, *politics*, and *sport*), which is often inadequate to express what the user is really interested in. Moreover, existing systems typically do not provide a feedback mechanism able to allow the user to specify non-relevant items—with the goal of progressively adapting the system to her / his actual interests.

In this paper, we propose a multiagent system devised to handle the task of generating press reviews. To this end, the system (i) extracts articles from online newspapers, (ii) classifies them using hierarchical text categorization, and (iii) provides suitable feedback mechanisms to the end user. The motivation for adopting a multiagent system lies in the fact that a centralized classification system might be quickly overwhelmed by a large and dynamic document stream, such as daily-updated online news [18]. Furthermore, Internet is intrinsically a distributed system and offers the opportunity to take advantage of distributed computing paradigms and distributed knowledge resources.

<sup>1</sup><http://news.google.com/>

<sup>2</sup><http://www.presstoday.com/>

The remainder of the paper is organized as follows: Section II recalls some relevant related work. Section III describes the proposed multiagent system. In Section IV the underlying motivation in adopting a multiagent system are briefly pointed out. Section V illustrates the experiments that has been performed during the training phase. Section VI shows the main functionalities of the system. Section VII draws conclusions and points to future work.

## II. RELATED WORK

This section is two-tiered, being aimed at recalling and summarizing relevant topics on multiagent systems used in information retrieval and on hierarchical text categorization.

### A. MultiAgent Systems for Information Retrieval

In the literature, several centralized agents architectures aimed at performing information retrieval tasks have been proposed. Among others, let us recall NewT [38], Letizia [29], WebWatcher [4], and SoftBots [16].

NewT [38] has been designed as a society of information-filtering interface agents, which learn user preferences and act on their behalf. To filter information agents use a keyword-based filtering algorithm, whereas the adopted adaptive techniques are relevance feedback and genetic algorithms. Letizia [29] is an intelligent user interface agent able to assist a user while browsing the Web. The search for information results as a cooperative venture between the user and the software agent: both browse the same search space of linked web documents, looking for interesting ones. WebWatcher [4] is an information search agent that follows web hyperlinks according to user interests, returning a list of links deemed interesting to the user. In contrast to systems for assisted browsing or information retrieval, SoftBots [16] accept high-level user goals and dynamically synthesize the appropriate sequence of Internet commands according to a suitable ad-hoc language.

Despite the fact that a centralized approach could have some advantages, in information retrieval tasks it may encompass several problems, in particular how to scale up the architectures to large numbers of users, how to provide high availability in case of constant demand of the involved services, as well as how to provide high trustability in case of sensitive information, such as personal data. To this end, in the literature, suitable multiagent systems devoted to perform information retrieval tasks have been proposed. For the sake of

brevity, let us recall here CEMAS [8], IR agents [24], and the cooperative multiagent system for web information retrieval proposed in [37].

In CEMAS (Concept Exchanging Multi-Agent System) the basic idea is to have specialized agents for each main task, the main tasks being: (i) exchanging concepts and links, (ii) representing the user, (iii) searching for new relevant documents matching existing concepts, and (iv) agent coordination. IR agents implement an XML-based multiagents model for information retrieval. The corresponding framework is composed of three kinds of agents: (i) managing agents, aimed at extracting the semantics of information and at performing the actual tasks imposed by coordinator agents, (ii) interface agents, devised to interact with the users, and (iii) search agents, aimed at discovering the information on the web. Finally, in [37] the underlying idea is to adopt intelligent agents that mimic everyday-life activities of information seekers. To this end, agents are also able to profile the user in order to anticipate and achieve her/his preferred goals.

### B. Hierarchical Text Categorization

Research interest in text categorization has been growing in machine learning, information retrieval, computational linguistics, and other fields. This reflects the importance of text categorization as an application area of machine learning, also facilitated by the availability of several document collections [28], [43] to which domain experts have assigned categories from a predefined (flat) set. These collections are in fact a benchmark that allow researchers to test their approaches while comparing the corresponding results.

Hierarchical text categorization deals with problems where categories are organized in form of a hierarchy. Many information sources are organized as hierarchies, e.g. web repositories, digital libraries, patent libraries, email folders, product catalogs. In particular, several web repositories encompass an underlying taxonomy, such as DMOZ<sup>3</sup> and the Google directory<sup>4</sup>. Taxonomies are also very useful in the field of news categorization, such as the one provided by the International Press Telecommunications Council<sup>5</sup> and the RCV-taxonomy (proposed by Lewis [27] to perform hierarchical text categorization on the Reuters standard document collection).

Until the mid-1990s researchers mostly ignored the hierarchical structure of categories that occur in several domains. In 1997, Koller and Sahami [22] carried out the first proper study of a hierarchical text categorization problem on the Reuters-22173 collection. First, a small hierarchical subset of Reuters-22173 has been generated by identifying labels that subsume other labels. Then, experiments have been performed by comparing a Naive Bayes classifier with two limited-dependency Bayes net classifiers –both on flat and hierarchical models. Documents were classified into the hierarchy by first filtering them through the single best-matching first-level class and then sending them to the appropriate second level. This

approach showed that hierarchical models perform well when a small number of features per class is used. No advantages were found using the hierarchical model for large numbers of features. After this work several approaches to hierarchical text categorization have been proposed (see for instance [10], [30], [42], [14], [40], [34], [9]).

### III. THE PROPOSED MULTIAGENT SYSTEM FOR NEWS RETRIEVAL

Generally speaking, a system devoted to perform information retrieval tasks might encompass three main steps: (i) extract the required information from web sources, (ii) categorize items according to a given taxonomy, and (iii) provide suitable feedback mechanisms. The proposed multiagent system is organized in three layers, each aimed at performing one of the above information-retrieval steps, as sketched in Figure 1.

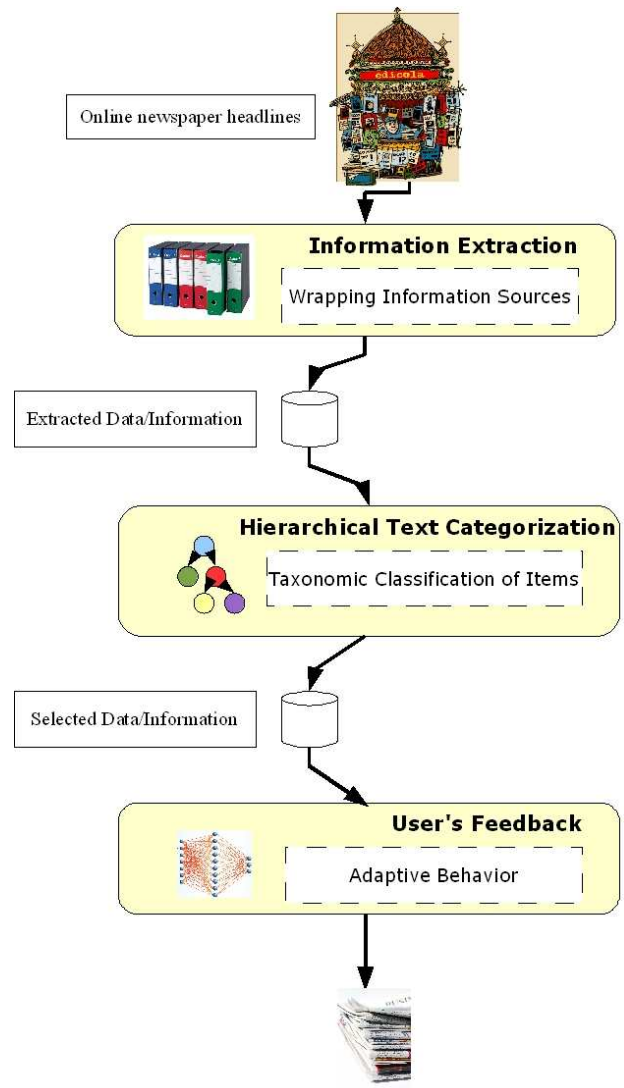


Fig. 1. A functional view of the proposed multiagent system

<sup>3</sup><http://www.dmoz.org/>

<sup>4</sup><http://www.google.com/dirhp?hl=eng>

<sup>5</sup><http://www.iptc.org/>

### A. Information Extraction

The information extraction activity is essential to retrieve documents provided by heterogeneous and distributed sources, such as web sites, digital archives, and web services.

In the literature, several tools have been proposed to better address the issue of generating wrappers for web data extraction [25]. Such tools are based on distinct techniques, such as declarative languages [19], [12], HTML structure analysis [13], [35], natural language processing [17], [39], machine learning [20], [23], data modeling [1], [7], and ontologies [15].

To perform information extraction, we use several wrapper agents, each associated with a specific information source. In particular, three wrappers able to deal with the RSS format have been implemented so far. These wrappers are devoted to process the news feed provided by the Reuters portal <sup>6</sup>, The Times <sup>7</sup>, and The New York Times <sup>8</sup>. Other wrappers, able to embed the Reuters document collection used to train the system and to embed the adopted taxonomy have also been implemented.

Once extracted, all the information is suitably encoded to facilitate the text categorization task. To this end, all non-informative words such as prepositions, conjunctions, pronouns and very common verbs are removed using a stop-word list. After that, a standard stemming algorithm [33] removes the most common morphological and inflectional suffixes. Then, for each category of the taxonomy, feature selection, based on the information-gain heuristics [26], has been adopted to reduce the dimensionality of the feature space.

### B. Hierarchical Text Categorization

A main issue in news categorization is how to deal, for each category, with an unbalance between relevant and non-relevant items. In particular, one may expect that most documents are non relevant to the user, the ratio between negative (e.g., non-relevant) and positive (e.g., relevant) examples being high (typical orders of magnitude are  $10^2 - 10^3$ ). Unfortunately, this aspect has a very negative impact on the precision of a text-categorization system. With the aim of coping with this phenomenon, we adopted a solution that exploits the ability of a pipeline of classifiers to progressively filter out non relevant information.

To better illustrate the underlying mechanism, let us consider the adopted taxonomy, i.e., the RCV1-taxonomy (Figure 2 reports part of the branch corresponding to the *economics* topic). Each node of the taxonomy represents a classifier entrusted with recognizing all corresponding relevant inputs. Any given input traverses the taxonomy as a “token”, starting from the root (in the example in Figure, ECAT). If the current classifier recognizes the token as relevant, it passes it on to all its children (if any). The typical result consists of activating one or more pipelines of classifiers within the taxonomy. As an example, let us consider Figure 3 that illustrates the

pipeline activated by an input document, which encompasses the categories economics (ECAT), government finance (E21), and expenditure/revenue (E211). This means that *all* involved classifiers recognize the input as relevant.

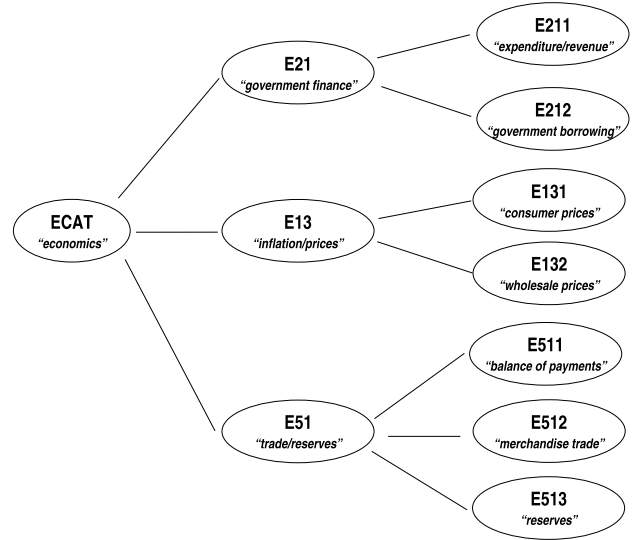


Fig. 2. A portion of the RCV1-taxonomy.

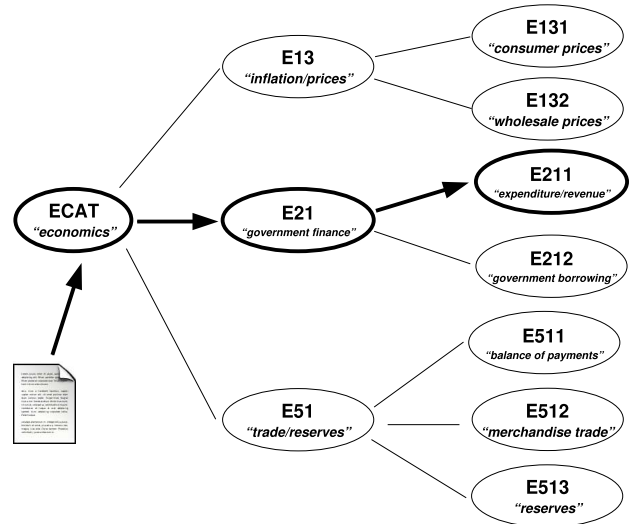


Fig. 3. An example of pipeline (highlighted in bold).

Each item in the taxonomy is implemented by an agent that embeds the corresponding classifier, in the current implementation the underlying classification technique being *k*-NN [44] –in its “weighted” variant [11]. The motivation for adopting this particular technique stems from the fact that it does not require specific training and is very robust with respect to noisy data.

### C. User’s Feedback

So far, a simple solution based on the *k*-NN technology has been implemented and experimented to deal with the problem

<sup>6</sup><http://www.reuters.com>

<sup>7</sup><http://www.the-times.co.uk/>

<sup>8</sup><http://www.nytimes.com/>

of supporting the user’s feedback. When a non-relevant article is evidenced by the user, it is immediately embedded in the training set of a  $k$ -NN classifier that implements the user feedback. A suitable check performed on this training set after inserting the negative example allows to trigger a procedure entrusted with keeping the number of negative and positive examples balanced. In particular, when the ratio between negative and positive examples exceeds a given threshold (by default set to 1.1), some examples are randomly extracted from the set of “true” positive examples and embedded in the above training set.

#### IV. UNDERLYING MOTIVATION IN ADOPTING A MAS

An information retrieval system must take into account several issues, the most relevant being: (i) how to deal with different information sources and to integrate new information sources without re-writing significant parts of it, (ii) how to suitably encode data in order to put into evidence the informative content useful to discriminate among categories, (iii) how to control the unbalance between relevant and non relevant articles (the latter being usually much more numerous than the former), (iv) how to allow the user to specify her / his preferences, and (v) how to exploit the user’s feedback to improve the overall performance of the system.

The above problems are typically strongly interdependent in state-of-the-art systems. To better concentrate on these aspects separately, we adopted a layered multiagent architecture, able to promote the decoupling among all aspects deemed relevant. In particular, the proposed system has been built upon PACMAS (Personalized Adaptive and Cooperative MultiAgent System), a generic multiagent architecture aimed at retrieving, filtering, and reorganizing information according to the users’ interests [2]. The adoption of PACMAS is motivated by the willing of better concentrating on the above aspects separately, as it is in fact a layered architecture capable of promoting the decoupling among all relevant aspects of a complex task aimed at performing information retrieval. The PACMAS generic architecture has been implemented on top of the well known JADE [6] agent-based infrastructure. PACMAS encompasses four main levels:

- *information level*, aimed at wrapping information sources. The ability of the system to deal with new information sources affects only this level (i.e., a corresponding adapter or wrapper agent must be devised and implemented for each new kind of information source to be processed);
- *filter level*, devoted to suitably encode the text content according to an information-gain heuristics. Agents belonging to this architectural level encode (and embed) the text content of an article into a vector of words, which in turn is used to discriminate among existing categories;
- *task level*, devoted to identify relevant articles depending on the user interests. Agents belonging to this architectural level are aimed at performing two-tiered action: first the input is classified in accordance with the existing taxonomy, then the intended category (defined by composing

existing categories with *and*, *or*, and *not* operators) is used to decide whether it interests the user or not. The former action embeds suitable policies aimed at controlling the negative impact of the unbalance between relevant and non-relevant articles, whereas the latter allows the user to explicitly specify her / his preferences about the set of relevant vs. non-relevant articles;

- *user interface level*, agents belonging to this level are aimed at performing the last check in order to decide whether the given input is of interest for the user and –optionally– at providing a feedback by the user, which can be exploited to improve the overall ability of discriminating relevant from non relevant inputs.

Finally, let us put into evidence that the adoption of a multiagent system allows to distribute the computation among several nodes. More the number of involved classifiers grows, more the distribution becomes an important issue to be taken into account. To this end, let us note that the adopted RCV1 taxonomy is composed of 103 classes and each node of the taxonomy represents a classifier entrusted with recognizing all corresponding relevant inputs.

#### V. TRAINING THE SYSTEM

The system has been trained using RCV1-v2, the standard document collection proposed in [27], which is organized in four hierarchical groups: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets).

Before studying the progressive filtering technique, several experiments devoted to set the system parameters have been performed. In particular, experimentally we found that the optimal number of features is 200, and that the number of nearest neighbors to be taken into account by the  $wk$ -NN (e.g., the value of  $k$ ) is in the range 7..11.

To assess the capabilities of the proposed progressive filtering technique, suitable pipelines composed of three classifiers<sup>9</sup> have been considered. First, each node of the pipeline is trained with a balanced data set by using 200 features (TFIDF) selected according to the information gain method. Then, for each node of the taxonomy, a learning set of 500 articles, with a balanced set of positive and negative examples, has been selected to train a classifier based on the  $wk$ -NN technology.

A complete discussion on the progressive filtering technique being out of the scope of this paper (see [3] for a detailed discussion), let us briefly summarize our experimental results.

As for testing, several randomly selected sets of 1000 documents have been generated –characterized by a different ratio between relevant and non-relevant inputs. In particular, the ratio between positive and negative examples has been set to 1:2, 1:10, 1:20, and 1:100 (50%, 10%, 5%, and 1%), say  $TS_{50}$ ,  $TS_{10}$ ,  $TS_5$ , and  $TS_1$ , respectively. To study the impact of progressively filtering information with pipelines of  $wk$ -NN classifiers (denoted as PIPE), we tested with the

<sup>9</sup>We considered pipelines of only three classifiers due to the limited depth of the adopted RCV1-taxonomy.

TABLE I  
MICRO- AND MACRO-AVERAGING.

pos	$f_1^\mu$ WKNN	$f_1^M$ WKNN	$f_1^\mu$ SVM <sup>1</sup>	$f_1^M$ SVM <sup>1</sup>	$f_1^\mu$ SVM <sup>2</sup>	$f_1^M$ SVM <sup>2</sup>	$f_1^\mu$ Pipe	$f_1^M$ Pipe
50	0,883	0,883	0.831	0.832	0,898	0,897	0.905	0.905
10	0,646	0,647	0.507	0.521	0,719	0,722	0.721	0.720
5	0,513	0,514	0.412	0.428	0,535	0,543	0.683	0.682
1	0,165	0,169	0.169	0.190	0,344	0,349	0.412	0.431

above test sets some relevant pipelines, each concerning three nodes of the taxonomy ( $k = 3$ ). Results have been compared with those obtained by running the same tests on stand-alone classifiers based on the following technologies: *wk*-NN (denoted as WKNN),<sup>10</sup> linear SVM (denoted as SVM<sup>1</sup>), and RBF-SVM (denoted as SVM<sup>2</sup>).

Table I summarizes the experimental results illustrating the micro- and macro-averaging of  $F_1$  obtained by moving the acceptance threshold of the classifier(s) under investigation over the range  $[0, 1]$ . A concise recall of the corresponding definitions follows (the interested reader may consult the corresponding literature, e.g. [36]).

As for micro- and macro-averaging, they are aimed at obtaining estimates of precision ( $P$ ) and recall ( $R$ ) relative to the whole category set. In particular, micro-averaging evaluates the overall  $P$  and  $R$  by globally summing over all individual decisions. In symbols:

$$P^\mu = \frac{TP}{TP + FP} \quad (1)$$

$$R^\mu = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^m TP_i}{\sum_{i=1}^m (TP_i + FN_i)} \quad (2)$$

where the “ $\mu$ ” superscript stands for microaveraging. On the other hand, macro-averaging first evaluates  $P$  and  $R$  “locally” for each category, and then “globally” by averaging over the results of the different categories. In symbols:

$$P^M = \frac{\sum_{i=1}^m mP_i}{m} \quad (3)$$

$$R^M = \frac{\sum_{i=1}^m mP_i}{m} \quad (4)$$

where the “ $M$ ” superscript stands for macroaveraging.

As for  $F_1$  [41], it is obtained from a more general definition by imposing that  $P$  and  $R$  have the same degree of importance. In symbols:

$$F_1 = \frac{2PR}{P + R} \quad (5)$$

Table I highlights that, in all selected samples, the solution based on multiple classifiers has reported better results than those obtained with flat models. Summarizing, experimental results show that in presence of unbalanced inputs, a pipeline of three classifiers is able to counteract an unbalance of up to 100 non relevant articles vs. one relevant article.

<sup>10</sup>The technique based on *wk*-NN has been used with both the hierarchical classification (PIPE) and the flat model (WKNN).

## VI. THE CURRENT PROTOTYPE OF THE SYSTEM

Figure 4 illustrates the current user interface of the system. Through it, the user can set (i) the source from which news will be extracted, and (ii) the topics s/he is interested in. As for the newspaper headlines, the user can choose among the Reuters portal, The Times, and The New York Times. As for the topics of interest, the user can select one or more categories in accordance with the given RCV1 taxonomy.

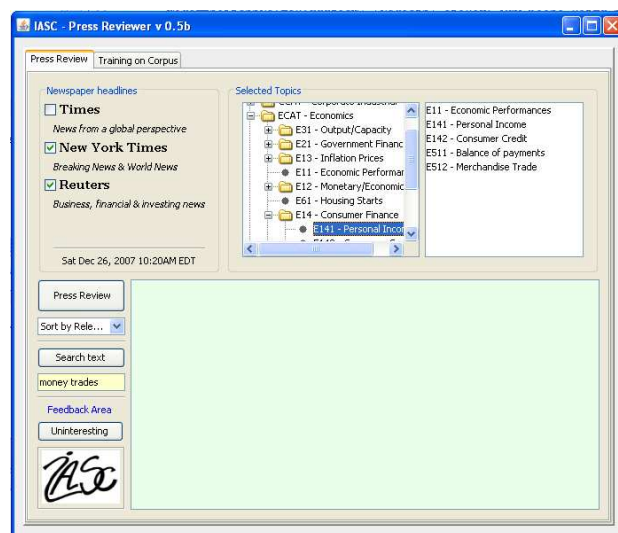


Fig. 4. The current user interface of the system.

The search for relevant news is activated by clicking on the *press review* button. First, information agents able to handle the selected newspaper headlines extract the news. Then, all agents that embody a classifiers trained on the selected topics are involved to perform text categorization. Finally, the system supplies the user with the selected news through suitable interface agents (see Figure 5).

Let us recall that the user can provide a feedback to the system by selecting all non-relevant news (i.e false positives). This feedback is important to let the system adapting itself to the actual interests of the corresponding user.

## VII. CONCLUSIONS AND FUTURE WORK

In this paper, a multiagent system devised to generate press reviews has been presented. The system encompasses three main tasks: (i) extracting articles from online newspapers, (ii) classifying them using hierarchical text categorization, and (iii) providing suitable feedback mechanisms.

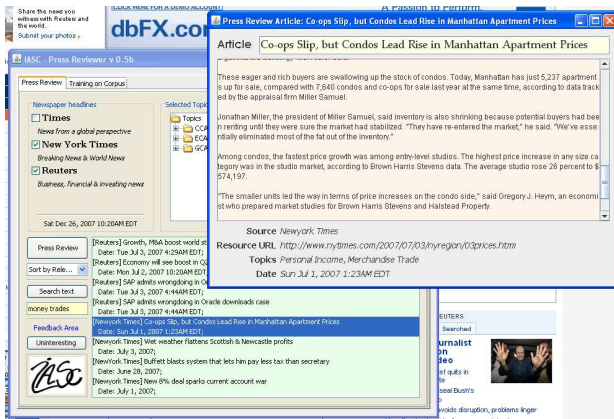


Fig. 5. An example of results provided by the system.

As for the future work, more sophisticated strategies to provide personalization are currently under investigation. Moreover, we are implementing a new release of the system with improved text categorization functionalities by adopting Support Vector Machines.

#### ACKNOWLEDGMENTS

This work has been supported by the Italian Ministry of Education, under the project “DART - Distributed Architecture for Semantic Search and Personalized Content Retrieval”.

#### REFERENCES

- [1] B. Adelberg, “NoDoSEa tool for semi-automatically extracting structured and semistructured data from text documents”, in *Proceedings of the 1998 ACM SIGMOD international Conference on Management of Data* (Seattle, Washington, United States, June 01 - 04, 1998). A. Tiwary and M. Franklin, Eds. SIGMOD '98. ACM Press, New York, NY, 1998, pp. 283-294.
- [2] G. Armano, G. Cherchi, A. Manconi, and E. Vargiu, “PACMAS: A Personalized, Adaptive, and Cooperative MultiAgent System Architecture”, in *Workshop dagli Oggetti agli Agenti, Simulazione e Analisi Formale di Sistemi Complessi (WOA 2005)*, 2005, pp. 54-60.
- [3] G. Armano, F. Mascia, and E. Vargiu, “Using Taxonomic Domain Knowledge in Text Categorization Tasks”, *International Journal of Intelligent Control and Systems, special issue on Distributed Intelligent Systems*, 2007, in press.
- [4] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell, “Webwatcher: A learning apprentice for the world wide web”, in *AAAI Spring Symposium on Information Gathering*, 1995, pp. 6-12.
- [5] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene Ontology: Tool for the Unification of Biology”, *Nature Genetics*, 25(1), 2000, pp. 25-29.
- [6] F.L. Bellifemine, G. Caire, and D. Greenwood, *Developing Multi-Agent Systems with JADE (Wiley Series in Agent Technology)*, John Wiley and Sons, 2007.
- [7] B. Ribeiro-Neto, A.H. Laender, and A.S. da Silva, “Extracting semi-structured data through examples”, in *Proceedings of the Eighth international Conference on information and Knowledge Management* (Kansas City, Missouri, United States, November 02 - 06, 1999). S. Gauch, Ed. CIKM '99. ACM Press, New York, NY, 1999, pp. 94-101.
- [8] M. Bleyer, “Multi-Agent Systems for Information Retrieval on the World Wide Web”, Diploma Thesis, University of Ulm, Germany, 1998.
- [9] L. Cai, and T. Hofmann, “Hierarchical Document Categorization with Support Vector Machines”, in *Proceedings of the ACM Conference on Information and Knowledge Management*, 2004, pp. 78-87.

- [10] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan, “Using Taxonomy, Discriminants, and Signatures for Navigating in Text Databases”, in *Proceedings of the 23rd international Conference on Very Large Data Bases* (August 25 - 29, 1997), M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld, Eds. Very Large Data Bases, Morgan Kaufmann Publishers, San Francisco, CA, 1997, pp. 446-455.
- [11] W. Cost, S. Salzberg, “A weighted Nearest Neighbor Algorithm for Learning with Symbolic Features”, *Machine Learning*, Vol. 10, 1993, pp. 57-78.
- [12] V. Crescenzi and G. Mecca, “Grammars Have Exceptions”, *Information Systems*, Vol. 23 (8), 1998, pp. 539-565.
- [13] V. Crescenzi, G. Mecca, and P. Merialdo, “Roadrunner, Towards Automatic Data Extraction from Large Web Sites”, in *Proceedings of the 27th International Conference on Very Large Data Bases*, 2001, pp. 109-118.
- [14] S. Dumais, and H. Chen, “Hierarchical Classification of Web Content”, in *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 2000, pp. 256-263.
- [15] D. W. Embley, D. M. Campbell, Y. S. Jiang, S. W. Liddle, Y. K. Ng, D. Quass, and R. D. Smith, “Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages”, in *Data Knowledge Engineering*, Vol. 31(3), 1999, pp. 227-251.
- [16] O. Etzioni and D. Weld, “Intelligent agents on the internet: fact, fiction and forecast”, *IEEE Expert*, Vol. 10 (4), 1995, pp. 44-49.
- [17] D. Freitag, “Machine Learning for Information Extraction in Informal Domains”, Ph.D. dissertation, Carnegie Mellon University, 1998.
- [18] Y. Fu, W. Ke, and J. Mostafa, “Automated text classification using a multi-agent framework”, *Proceedings of JCDL*, 2005, pp. 157-158.
- [19] J. Hammer, H. Garcia-Molina, S. Nestorov, R. Yerneni, M. Breunig, and V. Vassalos, “Template-Based Wrappers in the TSIMMIS system”, in *Proceedings of the 1997 ACM SIGMOD international Conference on Management of Data* (Tucson, Arizona, United States, May 11 - 15, 1997), J. M. Peckman, S. Ram, and M. Franklin, Eds. SIGMOD '97. ACM Press, New York, NY, 1997, pp. 532-535.
- [20] C. N. Hsu and M. T. Dung, “Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web”, *Information Systems*, Vol. 23(8), 1998, pp. 521-538.
- [21] B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, and G.O. Barnett, “The Unified Medical Language System: an informatics research collaboration”, *Journal of the American Medical Informatics Association*, 5(1) (Jan-Feb 1998), 1998, pp. 1-11.
- [22] D. Koller, and M. Sahami, “Hierarchically Classifying Documents Using Very Few Words”, in *Proceedings of the International Conference on Machine Learning (ICML)*, 1997, pp. 170-178.
- [23] N. Kushmerick, “Wrapper Induction: Efficiency and Expressiveness”, *Artificial Intelligence*, Vol. 118 (1-2), 2000, pp. 15-68.
- [24] W. Jirapanthong and T. Sunetnanta, “An XML-Based Multi-Agents Model for Information Retrieval on WWW”, in *Proceedings of the 4th National Computer Science and Engineering Conference (NCSEC2000)*, Bangkok, Thailand, 2000.
- [25] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and Juliana S. Teixeira, “A Brief Survey of Web Data Extraction Tools”, *SIGMOD Rec.*, Vol. 31 (2), 2002, pp. 84-93.
- [26] D. D. Lewis. “An evaluation of phrasal and clustered representations on a text categorization task”, in *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval*, (Kobenhavn, DK, 1992), 1992, pp. 37-50.
- [27] D.D. Lewis, Y. Yand, T. Rose, F. Li, “Rcv1: A New Benchmark Collection for Text Categorization Research”, in *Journal of Machine Learning Research*, Vol. 5(Dec.2004), 2004, pp. 361-397.
- [28] D.D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka, “Training Algorithms for Linear Text Classifiers”, in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Zurich, Switzerland, August 18 - 22, 1996). SIGIR '96. ACM Press, New York, NY, 1996, pp. 298-306.
- [29] H. Lieberman. “Letizia: An agent that assists web browsing”, in C.S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Quebec, Canada, 1995. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1995, pp. 924-929.
- [30] A. McCallum, R. Rosenfeld, T. Mitchell, and A. Ng, “Improving Text Classification by Shrinkage in a Hierarchy of Classes”, in *Proceedings of the International Conference on Machine Learning (ICML)*, 1998, pp. 359-367.

- [31] S. Nelson, M. Schopen, A. Savage, J. Schulman, N. Arluk, "The MESH translation maintenance system: Structure, interface design, and implementation", in Fieschi, M.e.a., ed., *Proceedings of the 11th World Congress on Medical Informatics*, IOS Press, 2004, pp. 67–69.
- [32] H.T. Ng, W.B. Goh and K.L. Low, "Feature Selection, Perceptron Learning, and a Usability Case Study for text Categorization", in *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, July 27-31, Philadelphia, 1997, pp. 67–73.
- [33] M. Porter, "An Algorithm for Suffix Stripping", *Program*, Vol. 14(3), 1980, pp. 130-137.
- [34] M. Ruiz, and P. Srinivasan, "Hierarchical Text Categorization Using Neural Networks", *Information Retrieval*, 5, 2002, 87–118.
- [35] A. Sahuguet and F. Azavant, "Building Intelligent Web Applications Using Lightweight Wrappers", *Data Knowledge Engineering*, Vol. 36(3), 2001, pp. 283–316.
- [36] F. Sebastiani, "Machine learning in automated text categorization", *ACM Computing Surveys*, 34(1) (Mar. 2002), 2002, pp. 1–47.
- [37] K. Shaban, O. Basir, M. Kamel, "Team Consensus in Web Multi-agents Information Retrieval System", *World Automation Congress*, Vol. 17, 2004, pp. 68–73.
- [38] B. Sheth and P. Maes, "Evolving agents for personalized information filtering", In I. Press, editor, *9th Conference on Artificial Intelligence for Applications (CAIA-93)*, 2003, pp. 345–352.
- [39] S. Soderland, "Learning Information Extraction Rules for Semi-Structured and Free Text", *Machine Learning*, Vol. 34(1-3), 1999, pp. 233–272
- [40] A. Sun, and E.P. Lim, "Hierarchical Text Classification and Evaluation", in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2001, pp. 521–528.
- [41] C. van Rijsbergen, *Information Retrieval*, Butterworths, London, 1979.
- [42] K. Wang, S. Zhou, and S.C. Liew, "Building hierarchical classifiers using class proximity", in M.P. Atkinson, M.E. Orłowska, P. Valduriez, S.B. Zdonik, and M.L. Brodie, eds, *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB'99)*, Morgan Kaufmann, 1999, pp. 363–374.
- [43] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization", in *Information Retrieval*, Vol. 1(1-2), 1999, pp. 69–90.
- [44] Y. Yang and X. Liu, "A re-examination of text categorization methods", in *Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, (Berkeley, California, United States, August 15 - 19, 1999). SIGIR '99. ACM Press, New York, NY, 1999, pp. 42–49.