Using linear classifiers in the integration of user modeling and text content analysis in the personalization of a Web-based Spanish News Service*

Alberto Díaz Esteban[1], Manuel J. Maña López[2], Manuel de Buenaga Rodríguez[1], Jose María Gómez Hidalgo[1], Pablo Gervás Gómez-Navarro[3]

[1] Departamento de Inteligencia Artificial, Universidad Europea de Madrid – CEES, 28670 Villaviciosa de Odón, Madrid, Spain. Tel.: (34) 91 6647800 ext. 670. Fax: (34) 91 6168265
{alberto, buenaga, jmgomez}@dinar.esi.uem.es

[2] Departamento de Informática, Universidad de Vigo, Campus As Lagoas s/n, 32004 Orense, Spain. Tel.: (34) 988 387014. Fax: (34) 988 387001
mjlopez@uvigo.es

[3] Departamento de Sistemas Informáticos y Programación, Universidad Complutense de Madrid, Ciudad Universitaria, 28040 Madrid, Spain. Tel.: (34) 91 3944739. Fax: 91 3944662
pgervas@sip.ucm.es

Abstract. Nowadays many newspapers and news agencies offer personalized information access services and, moreover, there is a growing interest in the improvement of these services. In this paper we present a methodology useful to improve the intelligent personalization of news services and the way it has been applied to a Spanish relevant newspaper: ABC. Our methodology integrates textual content analysis tasks and machine learning techniques to achieve an elaborated user model, which represents separately short-term needs and long-term multi-topic interests. The characterization of a user's interests includes his preferences about structure (newspaper sections), content and information delivery. A wide coverage and non-specific-domain classification of topics and a personal set of keywords allow the user to define his preferences about content. Machine learning techniques are used to obtain an initial representation of each category of the topic classification. Finally, we introduce some details about the Mercurio system, which is being used to implement this methodology for ABC. We describe our experience and an evaluation of the system in comparison with other commercial systems.

Keywords. Content analysis, machine learning, information dissemination, short/long-term models, multi-topic user profile, adaptive user model, personalized information service.

1 Introduction

The journalistic field is leading the offer of advanced information access services. Most of the newspaper publishers and news agencies supply engines for information

* This work has been partially supported by Spanish Ministry of Science and Technology (FIT-070000-2000-249)

search and delivery, as well as different personalization options. However, the opportunities for introducing improvements in that context are still important. Thus, popular web sites of outstanding newspapers, e.g. The New York Times, offer personalization methods specially focused in adaptation based on structure, i.e. newspaper sections. Others also provide personalization based on content, but much remains to be said on this issue.

Another point we must consider is that simple user models usually result in the introduction of irrelevant information. The integration of textual content analysis tasks and machine learning techniques can be used to achieve a more elaborate user model, to obtain a suitable representation of document contents, and to evaluate the similarity between user's interests and information. Representative examples of information access systems that integrate this kind of techniques are WebMate [7], News Dude [5] and SIFT [20]. WebMate [7] is a tool that compiles information from a list of URLs that the user wants to monitor (e.g., newspaper home pages) or from the search results using popular engines. The information is selected by its accordance with a user profile, which represents their multiple interests using vectors of terms and their weights. News Dude [5] is a system that compiles a personalized news program. Besides representing user's short-term and long-term interests, it takes into account the news previously heard by the user to avoid presenting the same information twice. Finally, SIFT [20] is an information filtering system that also models user's interest topics using keyword vectors provided by the user which are updated automatically by relevance feedback.

The aim of this paper is twofold. First, a useful methodology to improve the intelligent personalization of news services is presented. Second, the way we have applied it to a Spanish relevant newspaper - ABC - is introduced. Our work goes deeply into the integration of more elements to reach a more complete model and a better personalization. The final purpose is to offer a personalized and especially synthetic version of a newspaper, thus improving similar existing commercial services.

The user model proposed represents separately short-term needs and long-term multiple interests. Users can express their preferences both in terms of structural and content-based information. The sections of the newspaper act as structural information. A wide coverage classification of topics non-specific for the newspaper domain, the first level of categories of Yahoo! Spain, together with a set of keywords, is used to characterize the content based interest of a user. Lastly, application of implicit feedback allows these definitions to be enhanced and to evolve together with user's interest.

Text categorization, the assignment of subject labels to text items, is one of the most prominent text analysis and access tasks nowadays [19]. We have implemented a text categorization module that automatically assigns Yahoo! Spain subject labels to news items based on their text. The text categorization module is based on linear classifiers [6, 12, 14] and a program that mines Yahoo! Spain web pages. We also apply information retrieval with the keywords against all the news items.

2 Modeling User's Interests

In this section we focus on the process of modeling user's interests in our application setting. Users of any information access system have needs of different kinds. A user may have, for instance, a particular curiosity in knowing about the results of the elections in the USA. But, perhaps he does not want his personalized newspaper of subsequent days to hold news stories about the consequences of the elections or other elections in the USA. The very different characteristics of both kinds of needs suggest the inclusion of general and sporadic interests in two separate representations: long and short-term models.

The access method to news depends on the kind of necessity. Short-term needs are handled by the ad hoc retrieval subsystem, whereas long-term needs are handled by the dissemination information subsystem, which selects relevant news according to user's models and sends an e-mail message to the readers containing them.

Both subsystems provide users with a set of elements that places each selected piece of news in its information context. The newspaper section that a news story belongs to is shown together with the news story title. The similarity degree for each news story is also presented. The last element is a summary of the news item adapted to user's interests. This aims to present the most important ideas contained in the document in relation with user's needs.

We integrate a summarization subsystem like the one introduced in [15], that combines heuristics previously used in the generation of generic summaries with others allowing summary adaptation to user's interests. An evaluation of the summarization system using a collection of newspaper articles of different domains proves the utility of user's adapted summaries in an information access setting.

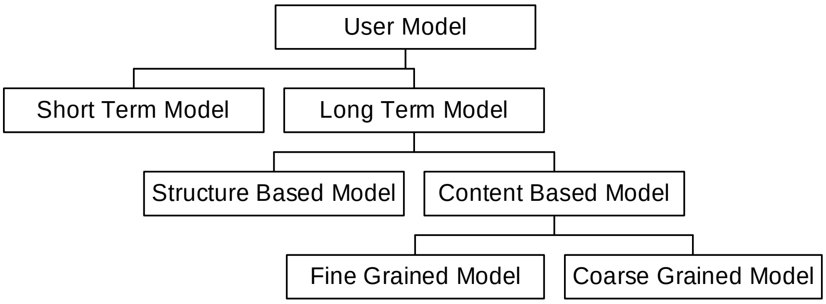


Figure 1. User Model

In the introduction of this paper, we have identified some important requirements of a user model. These are reflected in Figure 1. On the one hand, it is necessary to consider a short and long-term model to handle the different needs of a user. On the other hand, the system must be capable of representing multiple interests on different topics. As for what aspects have to be included in the model, in addition to personal information of the user (e.g. name, login, password, e-mail address) and preferences about delivery [2], the system should store user's preferences about structure (sections in the newspaper scope) and content (categories as coarse grained interests, and keywords as fine grained interests). Finally, the user model must be dynamic to adapt to changes in user's interests

Preferences about News: User's preferences about news may be expressed in terms of contents, i.e. categories and keywords of interest, or in terms of the structural element they belong to, i.e. sections of interest. To allow representation of multiple interests in the user model, users can show their preferences about them giving their degrees of interest on each on: *without interest*, *of some interest*, *interesting*, *very interesting*.

Sections are the way news items are organized in a newspaper. Users can assign a weight to each section, in order to give a value for the news contained in it. Examples of usual sections in a newspaper are "International", "Economy" or "Sports".

We believe sections are not the best candidates to represent user's interests about news contents. An alternative possibility is that the system provide a wide coverage classification about general subjects, non-specific to the newspaper domain, but known and familiar to the users. As the greater part of the readers of newspapers web versions are seasoned Internet users we have settled on choosing one of the category hierarchies supplied by the familiar web directories. Among these we have selected Yahoo! categories because it is one of the larger directories and it offers a specialized version in Spanish. The category system of Yahoo! Spain provides a first level with 14 categories and a second level with more than 200. Users can specify their general interest in some categories of first level assigning weights to them.

As well as the predefined hierarchy of categories, that represents a coarse grained need of a user, the system allows that users to define a set of keywords to represent a fine-grained need.

Too many selection methods available simultaneously (sections, categories and keywords) may lead to confusion. Unless additional control features are provided, system operation can become unpredictable. For this reason, our modeling architecture allows an extra level of specification. All of the features have an associated weight that represents their importance for the user's interests. This is a fine-grain tuning mechanism that allows users to obtain a flexible characterization of their needs.

Preferences about Delivery: In addition to information about user's interests on content and structure, the system also has to know user's preferences about information delivery. The information we refer to is:

- The days of the week the user wants to get a message.
- An upper bound on the number of news items per message (this avoids undesired overloads of information).
- The desired length for the news summaries.
- An "on holidays" binary value, which allows putting the system on hold for specific periods of times.

It is important that a user can access his model and check its features. Thus, the user can know at any time what is the knowledge that the system has on his interests. In addition, he is able to modify this knowledge, to some extent implicitly acquired, when he desires.

3 Information retrieval and machine learning techniques

The information access system sends a periodic e-mail to each user containing relevant news with respect to the interests stored in his long-term model. Each news item is represented as a term weight vector, according to the Vector Space Model (VSM) [17]. In news selection, we have formalized the concept of similarity between a text element and a user (his interests). Thus, the relevance between a news story *i* belonging to a newspaper section *k* and a user model *j* is computed using (1).

$$s_{ij} = a_j S_{kj} + b_j \sum_{h=1}^{n} C_{hj} \cdot sim_c\left(d_i, c_h\right) + g_j sim_k\left(d_i, k_j\right) \tag{1}$$

Where,

a_j is the significance of newspaper sections for user *j*,
S_{kj} is the interest of section *k* for user *j*,
b_j is the significance of categories for user *j*,
n is the number of categories,
C_{hj} is the interest of category *h* for user *j*,
d_i is the vector of weights for the news story *i*,
c_h is the vector of weights for the category *h*,
sim_c is the formula (2),
g_j is the interest of keywords for user *j*, being $a_j + b_j + g_j = 1$,
k_j is the vector of keywords for user *j*, and
sim_k is the cosine formula of the VSM.

A ranking of the news items is obtained according to their relevance for the given user obtained from (1). The top of the ranking is selected for the user in accordance to the upper bound on number of items per message specified in this profile.

We applied Text Categorisation using category-pivoted categorisation [7, 10, 14] with the categories against the news to obtain a ranking of the different news ordered by relevance for each category. We applied also Information Retrieval [17] with all the keywords against the news to obtain a list of relevant documents for the user. Also all the news items are processed to check if they belong to one of the sections selected in the user model.

When all the documents have been sorted according to the different sources of relevance, the resulting orderings are integrated by using the level of interest that the user assigned to each of the different reference systems. This implies that users looking for the same information but having chosen different methods to specify their interest may get different results. For the relevance values provided to the user to be easy to interpret, they are normalised over the number of selection methods involved in obtaining them. In this way, the system can quote a final relevance value in the range 0-100% to every user regardless of the number of selection methods that he chose.

3.1 Automatic Categorization of News Items

An automatic text categorization system has been built to perform this task. The system learns a linear classifier using the information extracted from Yahoo! Spain and then it automatically classifies news items according to their texts.

We have taken the categories in the first level of Yahoo! Spain, that we also represent as a term weight vector. These vectors are built using the information extracted from Yahoo! Spain, and the Rocchio linear classifier [16]. For assigning a category to a news item, the vectors for the category c_j and for the text of the news item p_k are compared using the following similarity formula:

$$sim\left(c_j, p_k\right) = \sum_{i=1}^{N} c_{ij} \cdot p_{ik} \quad (2)$$

Being c_{ij} the weight of the ith term for the vector of category c_j, p_{ik} the weight of the ith term for the vector of the news item p_k, and N the number of different terms. This formula does not take into account the length of news items and categories, because all the news items have got the same number of words (approximately), and also all the categories.

3.1.1 Data acquisition by Mining Yahoo!

For constructing the vectors that represent the categories, we have extracted information from Yahoo! Spain. One of the requirements of the categorization system was that it had to work autonomously, and so, no prelabeled news items were available for learning. Then, information from Yahoo! Spain was taken as training data for the learning process.

We have designed a Java program that mines the Yahoo! Spain web site to get the information necessary for learning the classifier or classification program. The mining program downloads the web pages reachable from the two first levels of the Yahoo! Spain web site, restricted to the Yahoo! domain. For each of the categories, the pages downloaded are taken as training data.

It is important to note that a common assumption made when applying a learning approach is that items to learn from and items to be classified are of the same type. The approach we describe here violates this assumption, but it is forced by the requirement stated above. Also, this approach has been followed by Attardi et al. [3] when classifying web pages, where the documents to be classified are artificially constructed using the text surrounding the hyperlinks to the pages to be categorized.

3.1.2 Machine Learning Linear Classifiers

For training a text categorization system, a number of Machine Learning approaches have been tested, including decision tree and rule-based learners [1, 8], probabilistic classifiers like Naive Bayes [13, 14], neural networks [9, 18], instance-based classifiers like kNN [22], etc. See [19] for other approaches.

An important subclass of learning approaches is that which learn linear classifiers, like Rocchio, Widrow-Hoff, or Winnow algorithms [6, 12, 13, 18]. These approaches examine training instances a finite number of times, and construct a prototype instance (a term weight vector) for each category, which is latter compared to the instances to be classified. Linear classifiers show interesting properties that make them ideal for industrial applications [19]:

- Linear classifiers are very efficient. Both the learning and the classification steps are linear on the number of terms, documents and categories, being far more efficient than most of the other learning approaches.
- Linear classifiers are simple to interpret. When the prototype vector for a category shows a high weight for a term, this term can be considered a good predictor for the category.
- Linear classifiers show good performance. Some of the algorithms are nearly top performing on standard test collections.
- Linear classifiers can take advantage of a standard IR system. The representation of categories is similar to documents in a text collection. The categories vectors can be indexed by an IR system, and documents to classify can be sent as queries to the system. The classification system assigns the most relevant categories to the document.

For our work, we have selected the Rocchio linear classifier. This algorithm is one of the most popular learning approaches tested for text categorization [6, 12, 13, 18]. This algorithm was originally designed for the relevance feedback process in Information Retrieval systems [16]. It was first adapted to text categorization by Hull in [12].

3.2 Adapting the Model to User's Preferences Changes

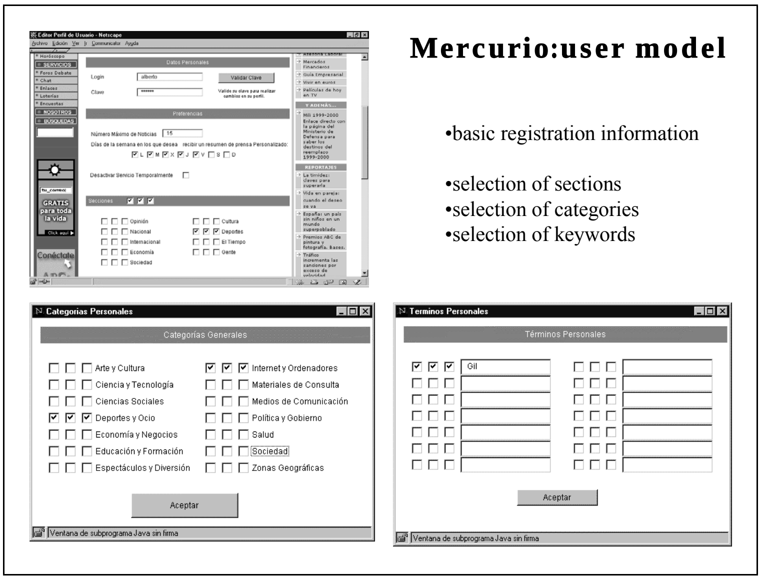
To achieve a long-term dynamic model that evolves together with user's interest, it is necessary to apply feedback techniques that provide information about this evolution. Relevance feedback techniques have been successfully used to improve effectiveness of IR systems [16]. The technique works as follows: after performing a search and retrieving a set of documents, the user provides the IR system with feedback, designating whether the retrieved documents are relevant or not. In VSM, this information may be used by the system for query improvement in two ways: re-weighting the query terms and adding new terms to the query [17].

However in practice, many users are unwilling to provide relevance judgments on retrieved documents. Among other reasons, users may have problems to decide about relevance of some documents. An alternative is to use implicit feedback, i.e. to infer document relevance from user’s behavior, which has been successfully applied in the learning of user models [4]. Then, we use the documents read by the user as feedback. The system provides numerous context elements, including a user-adapted summary that can assist users to decide about document relevance without inspecting the full text. If a user accesses the full text of a piece of news, the system can infer that it is relevant and use its term weight vector to improve the long-term user model.

In order to achieve a dynamic long-term user model, the system has to store for each user the read news. Before a new dissemination process starts, each user model is adapted in its content based part, i.e. the weights assigned to the categories, and the set of keywords and their weights, are updated.

4 Mercurio

Mercurio is the personalization access system for the electronic version of ABC newspaper that has been implemented using the methods described before excepting the automatic adaptation of the long-term model. Figure 2 shows the interface for editing the model.



Mercurio:user model

•basic registration information

•selection of sections
•selection of categories
•selection of keywords

Editor Perfil de Usuario - Netscape

Datos Personales

Login: alberto

Clave: ******

Valide su clave para realizar cambios en su perfil.

Preferencias

Número Máximo de Noticias: 15

Días de la semana en los que desea recibir un resumen de prensa Personalizado:

☑ L ☑ M ☑ X ☑ J ☑ V □ S □ D

Desactivar Servicio Temporalmente □

Secciones ☑ ☑ ☑

□ □ □ Opinión
□ □ □ Nacional
□ □ □ Internacional
□ □ □ Economía
□ □ □ Sociedad
□ □ □ Cultura
☑ ☑ ☑ Deportes
□ □ □ El Tiempo
□ □ □ Gente

Categorias Personales

Categorías Generales

□ □ □ Arte y Cultura
□ □ □ Ciencia y Tecnología
□ □ □ Ciencias Sociales
☑ ☑ ☑ Deportes y Ocio
□ □ □ Economía y Negocios
□ □ □ Educación y Formación
□ □ □ Espectáculos y Diversión
☑ ☑ ☑ Internet y Ordenadores
□ □ □ Materiales de Consulta
□ □ □ Medios de Comunicación
□ □ □ Política y Gobierno
□ □ □ Salud
□ □ □ Sociedad
□ □ □ Zonas Geográficas

Ventana de subprograma Java sin firma

Terminos Personales

Términos Personales

☑ ☑ ☑ Gil

Ventana de subprograma Java sin firma

Archivo Edición Ver Ir Communicator Ayuda
Horóscopo
SERVICIOS
Foros Debate
Chat
Enlaces
Loterías
Encuestas
NOSOTROS
BÚSQUEDAS
tu_correo
GRATIS para toda la vida
Click aquí
Conéctate
Validar Clave
Mercados Financieros
Guía Empresarial
Vivir en euros
Películas de hoy en TV
Y ADEMÁS...
Mili 1999-2000 Enlace directo con la página del Ministerio de Defensa para saber los destinos del reemplazo 1999-2000
REPORTAJES
La timidez: claves para superarla
Vida en pareja: cuando el deseo se va
España: un país sin niños en un mundo superpoblado
Premios ABC de pintura y fotografía. Bases.
Tráfico incrementa las sanciones por exceso de velocidad
Aceptar
Aceptar

Figure 2. User Model for Mercurio

4.1 Evaluation

Mercurio was evaluated following a working pattern that had been applied previously to others 15 existing systems. This working pattern [10], designed to evaluate this kind of services, includes several aspects as interface evaluation, newspaper sections, categories, summaries and user estimated recall and precision. The closed questionnaire used in the evaluation was composed of 103 specific questions centered on these topics.

From the evaluation of these systems the following conclusions were derived. The majority of these systems restrict their service to sending a digital version of the full daily edition. Those that allow some selection provide as criteria only sections or keywords. Only two systems provide together sections and keywords selection, and it is unclear how each one interacts with the other in producing the final selection.

A controlled evaluation environment was established to allow an analysis of the results with respect to the different kinds of user involved. Evaluation was carried out by 44 users grouped in four categories: collaborators, researchers, university lecturers (both on Computer Science and Journalism), and external users with no professional relationship with the fields involved.

We performed an evaluation of our proposal [10] in three ways: evaluation by users, system evaluation and evaluation of observed user profiles. For the relevance of the received documents the users had to check the performance of the system against the actual set of documents available on the newspaper website on three particular days. Additionally, on those particular days, logs of system operation (available documents, user profiles at the time, and system selections for each user) were kept to allow objective results to be obtained. With this data we worked out two sets of recall and precision figures: one based on user criteria as put down in the forms, and one based on subsequent close analysis of system logs.

	Precision	Recall
User criteria	0.9	0.7
System logs	1.0	0.2

Table 1. Recall and Precision figures

The results of the user evaluation and the system evaluation (Table 1) have shown that the precision obtained is very similar but the recall is lower in the system evaluation. The reason is that a user considers a document as relevant if it refers to something that is interesting for him, whether or not it belongs to a category or contains a word. However, the low recall value is a consequence of the upper bound imposed by the user: with a user model with a few sections and few categories the number of relevant documents is too high to be captured in a maximum recall fixed for the user by the upper bound.

The analysis of the 44 user models logged with the system yields the following data (see Table 2).

	Upper bound	Selection methods	Sections	Categories	Keywords
Average	14	1.9	2.6	3.4	2.3
Selected values	44	44	30	26	18
Selected average	14	1.9	3.9	5.8	5.7

Table 2. Analysis of user profile development

The average selection of a user has approximately 14 as upper bound of documents per message, 2 methods of selection (in most cases, sections and categories), 3 sections, 3 categories and 2 keywords. All the users selected the sections method, with or without other methods of selection, except one that chose to use only categories and keywords. All the users select some method and some upper bound, but not all select all methods. Thirty chose sections, 26 chose categories and only 18 chose keywords. It seems that less intuitive methods are less favoured. The users that chose the sections method choose an average of 4 sections. Those that chose categories marked 6, and those that chose keywords marked 6. When the user opts for a method, he tends to select more than one possibility.

The results show the improvements of our system with respect to the other systems analysed: a better personalisation through a more complete user model that integrates text classification techniques. For optimal use, the system should provide specific instructions about which method is better for each kind of search. In particular, more refined ways of integrating the different methods must be explored. Evaluation processes that show some measures of this effect have been developed in [11].

5 Conclusions and Future Work

In this paper, we have presented a set of methods to achieve an advanced user model that allows offering intelligent personalized access services to news. Also, the application of these methods to the development of a personalized newspaper service, the Mercurio system, for the relevant Spanish newspaper ABC has been introduced. An evaluation of Mercurio in comparison with other analogous services shows promising results on user satisfaction and on personalized services effectiveness.

Automatic text content analysis and machine learning techniques have been integrated to achieve an advanced user modeling. The user model takes into account long-term and short-term user's multiple interests and the changing character of these interests. This characterization includes user's interests about content, structure, and information delivery. User's explicit modifications of long-term model are supported.

In future works we plain to extend the machine learning techniques to feedback the user model and to co-training the representation of the categories with the results of the daily text categorization task. We are going to make a continuous effort to incorporate these novel features to commercial versions in ABC and other Spanish newspapers. We are also going to work in user's models in multilingual environments. The existence of resources as EuroWordNet makes it possible.

References

1. Apté, C., Damerau, F.J. and Weiss, S.M. (1994) Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3), pp. 233-251.
2. Amato, G., Straccia, U.: User Profile Modeling and Applications to Digital Libraries. In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Paris, France (1999)
3. Attardi, G., and Gullí, A. and Sebastiani, F. (1999) Automatic Web Page Categorization by Link and Context Analysis. In Proceedings of THAI-99, 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence, pp. 105-119.
4. Balabanovic, M.: An Interface for Learning Multi-topic User Profiles from Implicit Feedback. In: AAAI Workshop on Recommender Systems, Madison, Wisconsin (1998)

5. Billsus, D., Pazzani, M.J.: A Hybrid User Model for News Story Classification. In: Proceedings of the Seventh International Conference on User Modeling, Banff, Canada (1999)
6. de Buenaga, M., Gómez, J.M. and Díaz B. (1997) Using WordNet to Complement Training Information in Text Categorization. In Proceedings of the Second International Conference on Recent Advances in Natural Language Processing (RANLP).
7. Chen, L., Sycara, K.P.: WebMate: A Personal Agent for Browsing and Searching. In: Proceedings of the Second International Conference on Autonomous Agents, Minneapolis, (1998)
8. Cohen, W. and Singer, Y. (1996) Context-sensitive learning methods for text categorization. In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, pp. 307-315, ACM Press, New York, US.
9. Dagan, I., Karov, Y., and Roth, D. (1997) Mistake-driven learning in text categorization. In Proceedings of EMNLP-97, 2nd Conference on Empirical Methods in Natural Language Processing, pp. 55-63, Association for Computational Linguistics, Morristown US.
10. Díaz, A., Gervás P., García, A.: Evaluating a User-Model Based Personalisation Architecture for Digital News Services. In: Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries, Lisbon, Portugal (2000)
11. Díaz, A., Gervás, P., García, A., Chacón, I.: Sections, categories and keywords as interest specification tools for personalised news services. Online Information Review, OIR 2001, no 3. (in press)
12. Hull, D.A. (1994) Improving text retrieval for the routing problem using latent semantic indexing. In Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, pp. 282-289, Springer Verlag, Heidelberg, DE.
13. Larkey, L.S. and Croft, W.B. (1996) Combining classifiers in text categorization. In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, pp. 289-297, ACM Press, New York, US.
14. Lewis, D.D., Schapire, R.E., Callan, J.P. and Papka, R. (1996) Training algorithms for linear text classifiers. In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, pp. 298-306, ACM Press, New York, US.
15. Maña, M.J., Buenaga, M., Gómez, J.M.: Using and Evaluating User Directed Summaries to Improve Information Access. In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Paris, France (1999)
16. Rocchio, J.J. Jr. (1971) Relevance feedback in information retrieval. In G. Salton, editor, The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall.
17. Salton, G. and McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
18. Schütze, H., Hull, D.A. and Pedersen, J.O. (1995) A comparison of classifiers and document representations for the routing problem. In Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, pp. 229-237, ACM Press, New York, US.
19. Sebastiani, F. (1999) A Tutorial on Automated Text Categorisation. In Proceedings of the First Argentinean Symposium on Artificial Intelligence (ASAI-99).
20. Yan, T.W., Garcia-Molina, H.: SIFT – A Tool for Wide-Area Information Dissemination. In: Proceedings of the USENIX Technical Conference, (1995)
21. Yang, Y. (1999) An evaluation of statistical approaches to text categorization. Information Retrieval, Vol. 1, Number 1-2, pp. 69-90.
22. Yang, Y. and Pedersen, J.O. (1997) A comparative study on feature selection in text categorization. In Proceedings of ICML-97, 14th International Conference on Machine Learning, pp. 412-420, Morgan Kaufmann Publishers, San Francisco, US.