

Evaluating a User-Model Based Personalisation Architecture for Digital News Services

Alberto Díaz Esteban¹, Pablo Gervás Gómez-Navarro¹, Antonio García Jiménez²

¹ Departamento de Inteligencia Artificial, Escuela Superior de Informática, Universidad Europea-CEES
Villaviciosa de Odón, Spain
{alberto, pg2}@dinar.esi.uem.es

² Departamento de Periodismo Especializado, Facultad de Ciencias de la Información, Universidad Europea-CEES
Villaviciosa de Odón, Spain
antonio.garcia@fcp.cin.uem.es

Abstract. An architecture that provides personalised filtering and dissemination of news items is presented. It is based on user profiles and it provides mechanisms that allow the user to control and tailor to his own needs the interaction between three different sources of relevance judgements: the existing newspaper categorisation by sections, basic information retrieval on user selected keywords, and an additional operation of automatic categorisation against an alternative hierarchy of categories. These three tiers cover some of the most promising access methods for digital libraries. The proposed architecture has been implemented and evaluation results are presented, covering user response, system efficiency, and user preferences regarding the set of methods made available to them.

1 Introduction

Personalised information services are becoming popular in the Web. Many major digital newspapers around the world now offer to send users by electronic mail a customised selection of news items. In all the instances surveyed by our research team, the selection of information is carried out by simple methods: users choose which sections of a newspaper they are interested in or certain key words that they would like to appear in their selected items. We have performed an evaluation on 15 news services, mainly Spanish, to extract conclusions about the current state of the art. Based on this study we have constructed a system that incorporates these basic methods, together with existing techniques in the fields of natural language processing and user modelling, into the process of selecting the particular information items that are relevant to a given user. The system allows readers of the newspaper to receive a periodic e-mail message containing the news items that the system finds particularly relevant to the interests of the user, previously defined during registration. The resulting architecture provides an additional control layer that allows users to specify how important each of the different methods of selection is to his particular interests. In this

way, the system constitutes an integrated and customisable option of combining document relevance information from sources of three different kinds:

- A prior categorisation by the system owner (documents, i.e. news are sorted into sections by the editor)
- Keyword search information over the content of the documents
- Automatic categorisation results against an alternative (less domain specific) set of categories

As such, it presents a flexible, multidimensional and browsable user model, and a set of well founded techniques for user models and information items matching. The user model captures the user information interests by means of complementary aspects like newspaper sections, categories extracted from a general purpose search engine, and keywords extracted from read news articles. The integration of text retrieval [7] and categorisation [4], [5], [11] based on the Vector Space Model provide a solid theoretic framework for the implementation of the service.

2 Analysis of other systems

We looked for systems that provide news service on the Internet to be evaluated [9]. The following systems were selected: Crayon, Diario de Navarra, Economyweb, El Diario Vasco, El Heraldo de Aragón, El Mundo Digital, El País Digital, Expansión, Hispanidad, La Razón, La Vanguardia Digital, Politicsweb, Telépolis, The Economist, The New York Times.

Each system was monitored over a period of time by a team of four research students, performing a quantitative and qualitative analysis. Following [6], a quantitative analysis was carried out in which a form of 103 questions was filled out by each evaluator for each system. The questions were grouped around four specific topics: interface evaluation, categorisation evaluation, summary evaluation and actual relevance of received documents. For each of these parameters a numerical value was worked out from the evaluators responses. In each case a total percentage value was obtained for each system. Qualitative analysis was based on the description and valuation of each system, and personal interviews with the evaluators describing their overall impression of the different systems.

The following conclusions regarding the systems themselves arise from the evaluation: a) evaluators consider it important for systems to provide categories and keywords as means of selection (a second level of categories is considered a plus); b) systems that allow some personalisation rate much higher than others; c) personalisation should allow a choice of when to receive information and in what format; d) the effectiveness of some systems is low; e) access to the final documents must be provided; f) including a summary is a key positive feature; g) the presence of unrequested or irrelevant information (noise) can severely spoil otherwise good ratings; h) users dislike systems that provide too large a number of items; i) subscribing and unsubscribing must be easy tasks; j) ease of access and use is important.

All these results were taken into account when designing the personalisation architecture described below. The questionnaires and general evaluation method employed were later applied to the resulting prototype.

3 Modelling the users

After the study of some commercial systems we propose a browsable user model designed to represent the user information interests in a wide variety of ways. This describes the user information needs, that is, what a user is really looking for.

Of the ingredients of a user model proposed in [1], for this particular application information type, document structure, means of delivery and source of the information need not be modelled, since these aspects are invariable in a typical digital news service.

The user model stores three main kinds of information:

- The personal information of the user, which includes name, login, password and email address.
- The format information for the messages, which includes the weekdays the user wants to get a message, an upper bound on the number of news items per message, and an “on holidays” binary value (which allows to put the system on hold for specific periods of time).
- The information about the user interests, which covers newspaper sections, general categories, and keywords.

The format information is very useful for the users. Both imposing an upper bound on the number of items per message and the possibility of putting the system on hold avoid message overloading. Establishing a lower bound has been considered counter-productive, since it may lead to the inclusion of noise in messages whenever the lower bound cannot be met with relevant information.

As described in section 2, news services usually support personalisation by means of sections and/or keywords selection. Our model supports both features:

- The users can select their preferred sections. The sections of a newspaper are a primitive set of content-based categories. This set of 9 different categories comes from the traditional organisation of a newspaper. Users can assign a weight to each section, in order to prioritise the news items coming from it. Examples of sections in the newspaper are “International” or “Culture”.
- Users can also provide a set of chosen keywords by typing them in. Each keyword provided has to be given a weight that represents its relative importance for the users interests.

The system allows users to edit their selections: to modify the weights for sections or keywords, or to add or remove keywords.

As an additional selection feature, an alternative classification of the news items, obtained by means of automatic categorisation of the documents against a different set of categories, is provided. Internet users are already familiar with the categories systems employed in search engines directories. The category system of Yahoo! Spain was chosen as an alternative way of representing interests. Since it is designed for a

wider purpose and a wider domain than newspaper section headings, it constitutes a good second opinion. The first level of 14 categories from Yahoo! Spain is presented as a choice. Users can assign a weight to each category, to represent their interest in it.

Too many methods of selection available simultaneously can lead to confusion. Unless additional control features are provided users get at most a blurred picture of the operation of the system. For this reason, our personalisation architecture allows an extra level of user specification. A general control mechanism has been included to make the results more predictable for the user. Each of the three features (keywords, sections and additional categories) has a weight that represents its importance for the user interests. For example, if the weight of sections is low and the weight of additional categories is high, relevance values concerning additional categories will be considered more important for selecting news items.

In this way, each of the three dimensions considered in the user profiles can be defined and controlled by the user, providing a fine-tuning mechanism to obtain a flexible characterisation of his interests.

4 Information Filtering

The user model is applied daily by the system to the news items of the day, using Text Classification techniques (Text Categorisation and Information Retrieval [2], [4], [3]). A ranking of the news items is obtained according to their relevance for the given user. 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.

4.1 Representing information

The news item are downloaded daily from the newspaper website in the form of html documents. The title, section, URL and text for each document are extracted and stored. The representation of the news is obtained applying Vector Space Model (VSM) [7] to the text.

The VSM was originally developed for Information Retrieval (IR), but it provides support for many text classification tasks. The VSM for IR is applied by representing natural language expressions as term weight vectors. Each weight measures the importance of a term in a natural language expression, which can be a document or a query. Semantic closeness between documents and queries is computed by the cosine of the angle between document and query vectors. The terms selected for the representation are those that do not appear in a stop list for Spanish. We use the formulae based on term frequencies to compute their weights.

A representation for each category can be obtained by applying text categorisation techniques [4], [5] and using a set of training documents¹. In our case, the set of training documents used were the web pages indexed by Yahoo! within these categories. Thus, each category can be represented by a term weight vector that is obtained from

¹ A set of documents labeled manually with the suitable categories.

the name of the category, the name of its subcategories, if there are any, and the names and short descriptions of the web pages associated to the category.

The keywords also are represented with VSM, using the weight assigned for each word in the model.

4.2 Integrating Text Classification Tasks

We applied Text Categorisation using category-pivoted categorisation [4], [5], [8] 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 [3] with all the keywords against the news to obtain a list of relevant documents for the user.

Also all the news 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.

4.3 Information Dissemination

A message is generated for each user with the selected documents respecting the preferences stated in their profile. The message is sent by email early in the morning if the current day has been selected in the profile. The exact time is chosen to guarantee that the news of the day have been just placed at the disposal of the public in the corresponding web page. The user receives a message that matches the newspaper's design.

The automatic summary is generated by extracting from the HTML document the sub-heading, which tends to be a short summary in itself, provided by the editor. If there is no sub-heading, the first paragraph of the document is extracted. This heuristic gives good results due to the typical structure of news features, where the key ideas are presented at the beginning of the document.

A message is composed of: a) Title of the message with the current date and the name of the user; b) A link to the user model to permit change it if the user wants ; c) Various links to the newspaper (homepage, sections..); d) Brief description of the interests of the user (as featured in his profile); e) The selected documents, presented ordered by relevance and respecting the upper bound selected by the user (for each one: title, hyperlink to the original document, name of the section that it belongs to, final relevance value obtained, short automatic summary of the document, extra link explicitly stating it allows access to the full document).

5 Evaluation

We describe and discuss the three kinds of evaluation that were carried out: an evaluation carried out by a set of different users, a system evaluation that considers the performance of the system in measurable parameters, and an evaluation of the user model provided and how the evaluators have fared in dealing with it.

A controlled evaluation environment was established to allow analysis of the results with respect to the different kinds of user involved. Evaluation was carried out by 44 users in four categories: A) Collaborators; B) Researchers; C) University lecturers (both on Computer Science and Journalism); D) External users with no professional relationship with the fields involved.

The system was evaluated following the working pattern applied to other existing systems (see section 2). 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.

5.1 User-centred evaluation

During the first stage, the evaluation was centred on user response and the vision that users develop of the system. The aim was to harvest explicit evaluations provided by the users about system response-time, ease of use, system efficiency, and conceptual and physical presentation. This information was compiled on the basis of a closed questionnaire with specific questions on the relevant main topics. For each of these parameters a numerical value was worked out from the users responses.

In general, users found the system suitable although had some differences between different groups of users. This were the results for the interface evaluation: System Access: (high); General Interface, User Adaptation, and Integration into User Environment: (medium-high); Management of Contents, Query and Retrieval Schemes and User Help (medium). With respect to newspaper sections the following results were obtained: Expressive Faithfulness, Objectivity and Relevance (high). With respect to categories the following results were obtained: Expressive Faithfulness and Objectivity (medium-high); Relevance (medium). With respect to summaries the following results were obtained: Summary Content (high); Summary Structure (medium-high).

Recall and Precision rates have been estimated based on user impressions (see table 1), under the assumption that aim is not to obtain conclusive results but to draw roughly significant conclusions.

Group	Precision	Recall
A & B	0.9	0.8
C	0.9	0.6
D	0.9	0.6
Average	0.9	0.7

Table 1. User estimated Recall and Precision (by groups)

Additionally, the qualitative analysis showed that users were satisfied with the system characteristics, personalisation quality, formal quality, and categories system. On the other hand, the users' familiarity with similar systems influenced their understanding of the basic mechanisms. Some users found it could be more visual, but most of them understood it after receiving the first message.

5.2 Evaluation of observed user profiles

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	14,0	1,9	2,6	3,4	2,3
Max	20,0	3,0	9,0	14,0	15,0
Min	5,0	1,0	0,0	0,0	0,0
Selected values	44,0	44,0	30,0	26,0	18,0
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 method 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.

Some user select a method (section information for instance) but do not select any particular criteria for it (mark no specific sections). This results in an empty user model. This issue appeared for 14 users, all of which chose only sections. This has been identified as a problem that needs further work.

Regarding differences in profile development between user groups, it has been observed that groups A and B (which had taken part in the development of the project) tended to restrict their selections more than groups C and D: the number of selected sections and categories on average rose steadily from A to D.

5.3 System evaluation

We computed the values of recall and precision and others features for all the users on the last of the three specific days that the user had to review exhaustively. This allows a comparison between user evaluation and system evaluation to check the exhaustiveness of the user judgements and check the true performance of the system.

We have obtained the following results for each user: recall, precision, number of news selected by the information filtering system, that is, news with relevance greater than zero, number of truly relevant news. The day of this evaluation had 108 news.

Table 3 shows the average results and the maximum and minimum values for each feature.

	Upper bound	Relevant news	Recall	Not relevant news received	Precision	News selected
average	14	69,2	0,2	0	1,0	96
max	20	88	1	1	1,0	107
min	5	11	0,1	0	0,9	11

Table 3. Recall and Precision figures from system logs

5.4 Results discussion

Studying the results we can see that our system refines the information of the sections with the categories and keywords. The average precision is close to one because the fact that a document belongs to a section is enough for the document to have a high relevance value. Moreover, the relevance for belonging to a section is always greater than the relevance for belonging to a category or containing a keyword. Since most users have selected at least two sections, a section holds an average of eleven documents, and the average upper bound of documents per message is 14, most users get messages where all selected documents are relevant.

If a user marked sections as selection method in his profile (most do, in fact 50% of the users rely on sections altogether to select), the selected documents that appear first belong to these sections. They are shown ordered according to relevance computed with respect to categories and keywords. They are followed by documents that do not belong to these sections but are relevant in terms of categories and keywords. These show much lower relevance values nonetheless.

However, a user that does not use sections obtains documents sorted by the information relative to categories and keywords, and so obtains relevant documents from different sections. Only one user operated in this way, and he obtained similar value of precision and a low value of recall. This is because he had selected 7 categories and

14 keywords and the number of relevant documents under such wide criteria is above average.

If the relevance value computed using the categorisation method were greater than it currently is, documents relevant according to this source might find their way to the top of the ranking. This is at present unlikely because the categorisation system yields always very low relevance values, but we hope to improve it by developing a richer representation for categories.

If we compare the results of the user evaluation and the system evaluation we can see 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.

6 Conclusions

We have constructed a system that implements a personalisation architecture based on a complete user model that captures different interests of a user. The system operates in the domain of digital news services. It improves on existing personalisation methods by relying on the user model and text classification techniques to filter the particular information items that are relevant to a given user.

Based on an evaluation of 15 digital news services the following features have been found desirable: use of more information than sections or key-words to describe the user's interests, effectiveness, personalisation, ease of use of the system and absence of noise. We performed an evaluation of our proposal in three ways: evaluation by users, evaluation of observed user profiles and system evaluation.

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. They have also shown that more careful guidance of the user is needed (help documentation, contact persons). It has been noted that users have difficulties in filling out the model in a correct way, generating bad profiles in some cases. 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; and evaluation processes that allow some means of measuring the effect of integration should be developed.

The technology employed is not domain specific and relies solely on general techniques. Therefore it can very easily be ported to other domains where there is an important volume of new textual information to be processed periodically, like digital libraries where we can use the same method on articles of an author, filtering the documents that belong to a category and with some keyword.

As lines of future work, we are considering: the possibility of creating personal categories to allow the users to define extra categories suited to their own information

needs, the use of relevance feedback [11] in our system, a second level of categories to get a more specific model, the use of a stemmer for Spanish and Spanish Wordnet to take into account different lexical relations between words, and extending our system to a multilingual framework by taking advantage of Eurowordnet [10] and other multilingual resources.

References

1. Amato, G., Straccia, U. (1999). User Profile Modelling and Applications to Digital Libraries. Third European Conference, ECDL'99, Paris, France, September 1999.
2. Díaz, A., Buenaga, M., Ureña, L. A., García, M. (1998) Integrating Linguistic Resources in an Uniform Way for Text Classification Tasks. First International Conference on Language Resources & Evaluation, Granada (Spain), 1998.
3. Frakes, W. & Baeza, R. (1992), Information retrieval: data structures and algorithms, Prentice Hall, London. 1992.
4. Gómez Hidalgo, J. M. and Buenaga, M. (1997). Integrating a Lexical Database and a Training Collection for Text Categorisation. ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP, Madrid (Spain), 1997.
5. Lewis, D.D., Schapire, R.E., Callan, J.P. & Papka, R. (1996). Training algorithms for linear text classifiers. In Proceedings of the ACM SIGIR, 1996.
6. Pastor, J. A., Asensi, V. (1999). Un modelo para la Evaluación de Interfaces en Sistemas de Recuperación de Información, IV Congreso Iско-España Eoconsid'99, Granada (Spain), 1999.
7. Salton, G. (1989). Automatic Text Processing: the transformation+, analysis and retrieval of information by computer. Addison Wesley. 1989
8. Sebastiani, F. (1999). A Tutorial on Automated Text Categorisation. Proceedings of the First Argentinean Symposium on Artificial Intelligence (ASAI-99)
9. Theng, Y.L., Duncker, E., Mohd-Nasir, N., Buchanan, G., Thimbleby, H. (1999). Design Guidelines and User-Centred Digital Libraries. Third European Conference, ECDL'99, Paris, France, September 1999.
10. Vossen, P. (1996). Eurowordnet: building multilingual wordnet database with semantic relations between words. technical and financial annex. Technical report, EC-funded project LE # 4003.
11. Yang, Y. (1999), "An Evaluation of Statistical Approaches to Text Categorisation", Information Retrieval Journal