Personalisation in news delivery systems: Item summarization and multi-tier item selection using relevance feedback

Alberto Díaz and Pablo Gervás[a,b]
[a]*Centro de Estudios Superiores Felipe II, Universidad Complutense, Aranjuez, Madrid, Spain*
[b]*Departamento de Sistemas Informáticos y Programación, Universidad Complutense Madrid, Spain*
Tel.: +34 913947641; Fax: +34 913947529; E-mail: pgervas@sip.ucm.es

Abstract. The designer of an information dissemination system based on user preferences stated as user models is currently faced with three basic design decisions: whether to use categories, keywords – or both – to enable the user to specify his preferences, whether to use a static long-term model or a dynamic short-term model to register those preferences, and what method to use to provide summaries of the available documents without losing information that may be significant to a particular user even if it would not be considered as such in general terms. Current systems tend to provide one specific choice – either taken at design time by the developer or offered as mutually exclusive alternatives to the user. However, most of the options have relative merits. An efficient way of combining the various solutions would allow users to select in each case the combination of alternatives better suited for their needs. In this paper we defend the use of a combined approach that integrates: an enriched user-model that the user can customise to capture his long-term interests either in terms of categories (newspaper sections) or keywords, a personalised summarization facility to maximise the density of relevance of sent selections, and a tailored relevance feedback mechanism that captures short-term interests as featured in a user's acceptance or rejection of the news items received. Controlled experiments were carried out with a group of users and satisfactory results were obtained, providing material for further development. The experimental results suggest that categories and keywords can be fruitfully combined to express user interests, and that personalised summaries perform better than generic summaries at least in terms of identifying documents that satisfy user preferences.

Keywords: Summarization, relevance feedback, personalisation, user modeling, text classification

1. Introduction

Information systems have recently evolved to include applications, usually Web and e-mail based, that provide specific services for searching textual sources of information and/or disseminating information from these sources to interested users. The general challenge is to provide the means for pairing off elements from two sets: a set of interested users and a set of available news items. This is generally attempted through the use of user modelling to represent the interests of the users and the application of automatic classification techniques over such a representation and the available news items.

Most of the newspaper publishers and news agencies supply engines for information search and delivery, as well as different personalization options. Popular newspaper Web sites offer personalization methods specially focused in adaptation based on domain-dependent classification – newspaper sections – or domain-independent content – keyword-based search.

For the development of a personalised information system, several design decisions must be made:

(1) The user model may be controlled exclusively by the user browsing and editing his user profile, or it may be dynamically modified by the system.
(2) If the user is modelled dynamically, either an initially empty user model is progressively ad-

justed to represent user interests or the user is asked to provide a starting profile

(3) The information on user preferences may be used simply to make a personalised selection of news items considered interesting for the user, or used additionally to provide personalised summaries of the interesting news items.

The consequences of the design decisions on system efficiency and user satisfaction are explored in the following sections. Mechanisms are provided to implement the various processes described (Section 3) and formal evaluation is carried out for each of the resulting options (Section 4). The implications of the results obtained in terms of personalised information delivery are discussed in Section 5.

In this paper we defend the use of a combined approach that integrates: an enriched user-model that the user can customise to capture his long-term interests, a personalised summarization facility to maximise the density of relevance of sent selections, and a tailored relevance feedback mechanism that captures short-term interests as featured in a user's acceptance or rejection of the news items received.

2. Personalised and adaptive information systems

Users of any information access system have preferences at different levels. A user may have, for instance, both a general long-term interest in a specific sport, as well as special temporary interest in a particular political issue that affects his home team. 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 utility of having two separate user models – each with a different temporal range – was studied in [1] in the context of student modeling. In spite of differences in the choice of domain, representation of data, and learning algorithms, this work presents a motivation for using a short-term/long-term combination which is similar to the one presented here. The problem that needs solving is that user preferences may vary in time: if recent data are used for modeling they will better reflect current preferences compared with historical data, yet the use of historical data may better reflect a core of firmly based user preferences that last over a long time. Chiu and Webb [1] solve this issue by using in the first instance a model trained on recent data, and resorting to a model trained over a longer time period – thereby taking into account historical data – if the recent model fails.

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 user modeling 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 interests and information. Representative examples of information access systems that integrate this kind of techniques are WebMate [2], News Dude [3], SIFT [4], IFTool and PIFT [5]. WebMate 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. SIFT is an information filtering system that also models user's interest topics using keyword vectors provided by the user. IFTool is based on a matching algorithm specifically designed for taking advantage of the co-occurrence relationship between pairs of terms appearing simultaneously in the documents. This relationship is represented with a weighted semantic network whose nodes correspond to terms and where arcs link together terms, which co-occurred in some document. PIFT is characterized by a probabilistic approach to filtering, based on bayesian networks. The system in [6] combines a text classification algorithm that produces a ranking using keywords defined by the users, with another ranking generated by an algorithm that uses information about the user such as sex, occupation, marital status, hometown, etc. Other examples are [7,8].

The techniques described so far are applied in many systems that provide services related to the ones considered in this paper. The Daily Learner system [9] 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. The functionality is based on combining a nearest neighbour algorithm to model the short-term interests and a naive bayesian classifier to model the long-term interests. The Smart Guide system [10] provides a filtering application that presents each user with a personalised version of a Web-based reference guide.

To achieve a dynamic model that evolves together with user's interest, feedback techniques are applied. Similar techniques have been successfully used to improve the effectiveness of IR systems [11]. The technique works as follows: after retrieving a set of documents, the user provides the system with feedback, designating whether the retrieved documents are relevant or irrelevant. Existing systems solve this problem by resorting to various techniques: applying Boolean matching [4], or weighted matching using contexts of words, based on the cosine measure of the Vector Space Model [2], or updating the weights associated with arcs and nodes of the semantic network that represent the interests of a user [5], or using some strategy of learning [9], or using combined information from the title and the body of the feedback news items [6].

A related method for filtering information with respect to user preferences is collaborative filtering. In this approach, users indicate their opinions by providing relevance values to various information contents, and the collaborative filter correlates these values with those provided by other users to obtain future predictions. If a user A gives a high relevance value to a new document and there are several users whose judgements in the past matched those of user A, the sytem will then recommend the new document to those users. The user profile, therefore, stores relevance values asigned by the user to each of the elements he has considered. Additionally, in collaborative filtering the values in the profile for each user are shared with other users to enable them to apply them for their own predictions. The first systems required the user to explicitly evaluate the document – such as Tapestry [12] –, then systems emerged where user effort was considerably reduced – such as Grouplens [13,14] –, which led to systems where evaluation is done implicitly based on user actions – such as PHOAKS [15].

The use of a collaborative filtering system in isolation can lead to poor efficiency due to several reasons: the problem of the first evaluator, the problem of the scarcity of evaluations, and the problem of the black sheep. The first problem appears when a certain information content appears for the first time, because there are no previous evaluations by anybody on which to base a prediction. The second problem derives from the fact that there are usually many more elements to evaluate than users ready to evaluate them. This may result in not enough evaluations being available on a given item to support adequate predictions. The third problem arises when there are users whose behaviour is not consistent with any other group of users. This kind of user rarely receives predictions according to his profile. Experiments have shown that this type of system can be improved by applying content-based filters [16–18]. The combination of collaborative filtering and content-based filtering may provide the advantages of both methods, including the availability of valid initial predictions for all items and all users, and improved collaborative predictions once the number of users and evaluations increases. For instance, P-Tango [17] is a system that permits the personalisation of the Worcester Telegram and Gazzete Online electronic newspaper. The content-based user model stores the newspaper sections in which the user is interested and a set of explicit keywords provided by the user for each section. Additionally, a set of implicit keywords is stored. These implicit keywords are extracted from the documents that have been highly valued by the user.

Fab [16] is a Web page recommendation system where documents are represented by the 100 terms with highest tf-idf weight. When a new user starts using the system he is presented with a series of Web pages selected randomly from those with highest ratings for other system users. In this way the user does not start with an empty profile. To achieve this, the average rating of all users is stored explicitly in a global profile.

One aspect that is not covered in the systems described so far, and which constitutes an interesting contribution of the present system, is the ability to generate personalised summaries of the items that are being filtered. The normal behaviour of a personalization system is to send to the users the title and the first lines of the items that are detected as interesting, and links to the full text. This information is in most cases insufficient for a user to detect if the item is relevant or not relevant to him. This forces him to inspect the full text of the document, involving an extra time cost. However, even in this case, the size and structure of the document may not be suitable. In such cases, it can be difficult for the user to find the information required to decide about document relevance. An interesting approach is to replace the first sentences sent as a sample of a document by a proper summary or extract [19].

Automatic summarization is the process through which the relevant information from one or several sources is identified in order to produce a briefer version intended for a particular user – or group of users – or a particular task [20]. According to their scope and purpose [21], the summaries considered in this paper will be restricted to a single document – rather than a set of documents – and they will be indicative – their aim is to anticipate for the user the content of the text and to help him to decide on the relevance of the original document. Attending to their focus, we can distinguish between:

Generic abstracts , if they gather the main topics of the document and they are addressed to a wide group of readers, and

User adapted abstracts , if the summary is constructed according to the interests – i.e. previous knowledge, areas of interest, or information needs – of the particular reader or group of readers that the system is addressing.

It is clear that, especially for high levels of compression, a summary that does not take the user's needs into account may be too general to be useful. It has already been shown that in an information retrieval environment summaries adapted to the user query outperform other kinds of summary [19].

Given the variety of types and domains of available documents, techniques for selection and extraction of phrases are very attractive due to their domain- and language-independence. In these techniques the segments of text – usually sentences or paragraphs – that contain the most significant information are selected based on linear combination of the weights resulting from the application of a set of heuristics applied to each of the units of extraction. These heuristics may be *position-dependent*, if they take into account the position that each segment holds in the document; *linguistic*, if they look for certain patterns of significant expressions; or *statistical*, if they include frequencies of occurrence of certain words. The summary results from concatenating the resulting segments of text in the order in which they appear in the original document [22]. In areas like information filtering or personalized information services, in which the system may have available a larger amount of information about the preferences of the user, a more adequate selection of the sentences that make up the summary contents can be expected. Possible inconsistency problems within the resulting summary are the main disadvantage of this solution. One way of reducing them is to take paragraphs rather than sentences as unit of extraction [23] in the hope that within the wider context of paragraphs legibility is improved. Another set of techniques, knowledge rich techniques, uses methods of deep understanding of the text [21]. This approximation may lead to the creation of systems that sidestep the inconsistency problems mentioned above, but they require a great amount of information about the domain. The domains involved tend to be restricted and fulfilling well-known characteristics. As a consequence, the systems are inflexible and difficult to adapt to other situations or other languages.

3. Exploring alternatives in personalised information dissemination

The system described in this paper sends a periodic message to each user containing relevant news with respect to the interests stored in his model. A message is composed of [24]:

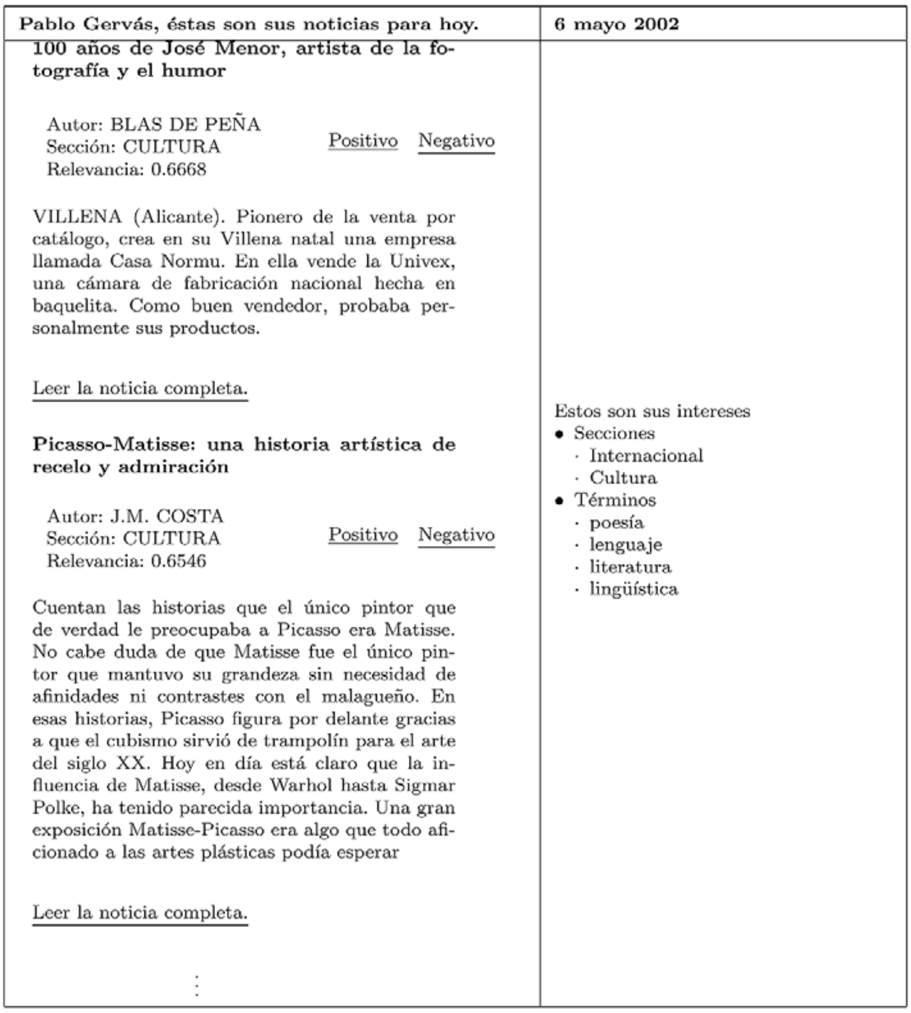
- a title with the current date and the name of the user,
- a link to the user model to allow editing if desired,
- various links to the newspaper (homepage, sections, . . .),
- brief description of the interests of the user (as featured in his profile), and
- the selected documents, presented ordered by relevance and respecting the established upper bound (for each one: title, name of the author, 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).

An example of the typical content of a message received by a user is presented in Fig. 1.

To carry out this task, the user model is applied daily by the system to the news items of the day. 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 to be sent, a system configuration parameter T, set for the work presented here to an arbitrary value of $T = 10$.

3.1. Representing the news items

The news items are downloaded daily from the newspaper Web site in the form of HTML documents. The title, section, URL and text for each document are extracted and stored. The representation of the news item is obtained applying Vector Space Model (VSM) [25] 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. To obtain these vectors, standard stop lists are applied to the texts, and remaining words are reduced to a canonical form by applying a Porter stemmer adapted for Spanish [26]. We use the *tf-idf* method [25] based on term frequencies to compute their weights. Document frequencies are computed over the collection of news items for a given day. Each weight



Pablo Gervás, éstas son sus noticias para hoy. **6 mayo 2002**

100 años de José Menor, artista de la fotografía y el humor

Autor: BLAS DE PEÑA
Sección: CULTURA
Relevancia: 0.6668

Positivo Negativo

VILLENA (Alicante). Pionero de la venta por catálogo, crea en su Villena natal una empresa llamada Casa Normu. En ella vende la Univex, una cámara de fabricación nacional hecha en baquelita. Como buen vendedor, probaba personalmente sus productos.

Leer la noticia completa.

Picasso-Matisse: una historia artística de recelo y admiración

Autor: J.M. COSTA
Sección: CULTURA
Relevancia: 0.6546

Positivo Negativo

Cuentan las historias que el único pintor que de verdad le preocupaba a Picasso era Matisse. No cabe duda de que Matisse fue el único pintor que mantuvo su grandeza sin necesidad de afinidades ni contrastes con el malagueño. En esas historias, Picasso figura por delante gracias a que el cubismo sirvió de trampolín para el arte del siglo XX. Hoy en día está claro que la influencia de Matisse, desde Warhol hasta Sigmar Polke, ha tenido parecida importancia. Una gran exposición Matisse-Picasso era algo que todo aficionado a las artes plásticas podía esperar

Leer la noticia completa.

⋮

Estos son sus intereses
- Secciones
  - Internacional
  - Cultura
- Términos
  - poesía
  - lenguaje
  - literatura
  - lingüística

Fig. 1. Sample message received by a user.

measures the importance of a term in a natural language expression. Semantic closeness between documents and queries is computed by the cosine of the angle between document and query vectors.

3.2. Available data on user preferences: a user model

The user model has to adapt to the different aspects of each domain, in order to allow a better definition of the user interests [27]. We propose a browsable user model designed to represent the user information interests in a wide variety of ways. 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 and an “on holidays” binary value (which allows to put the system on hold for specific periods of time).
– The specific information about the interests of the user defined in terms of a number of reference frameworks used by the system to personalise information provided to the user.

The possibility of putting the system on hold avoids message overloading. Establishing a lower bound has been considered counterproductive, since it may lead to the inclusion of noise in messages whenever the lower bound cannot be met with relevant information.

The system uses various reference frameworks to obtain from the user different views or descriptions of his interests matching these requirements. The first one involves choosing categories (in our case newspaper sections) under which his interests may be classified. The second one involves selecting keywords that may appear in items of interest. The third one is extracted automatically from items for which the user has provided feedback. An important contribution of this paper is to test whether a combination of such reference frameworks provides a good solution, and, if so, which manner of combining the frameworks gives best results. To achieve this purpose, each of the three explicit reference frameworks (categories, keywords and feedback terms) has a weight that controls its influence in the selection of the final results by the system. For example, if the weight assigned to categories is low and the weight for keywords is high, relevance values concerning keywords 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 during the experiments.

3.2.1. *Categories*

A user is asked to select those categories into which the documents that may interest him would be classified. These sets of categories are usually domain-dependent – different sets of categories are considered to be most effective for specific domains – and they tend to be established by the information provider. For instance, newspapers arrange their news items by sections, bookshops arrange their books in terms of literary genres. The categories employed in the work described here correspond to traditional newspaper sections for a typical Spanish daily newspaper, such as: international, national, sports, culture, society . . . . From here on, this set of categories is referred to as *sections*. Users can assign a weight to each section, to represent their interest in it. Numerical values for weights are obtained from an initial assignment by the user, who evaluates their degrees of interest for him on the following scale: *without interest, of some interest, interesting, very interesting*.

3.2.2. *Keywords*

In completing his user profile, each user is asked to type in a number of keywords, whose appearance in a news item may indicate that it will interest him. For each keyword the user also indicates a rough weighting to show the degree of his interest. The keywords typed in by the user are also represented in terms of the VSM, using the weight assigned to each word in the current user model. The system allows users to edit their selections: to modify the weights for sections or keywords, or to add or remove keywords.

3.2.3. *Feedback information*

The messages sent to the user with the daily selection of news items according to his preferences as stated in the user model allow the user to provide positive or negative feedback on each news item received. The user may also decide not to provide feedback on a given news item. Each news items that has been selected appears with two underlined links which the user can click on to send either positive or negative corresponding feedback to the system (in the example message shown in Fig. 1, the links appear as *positivo* and *negativo*).

This information is stored, a set of feedback keywords is extracted from the corresponding news items, and these feedback keywords are used to update the short-term part of the user model. Because this final reference framework is specifically designed to cope with short-term variations in interests without requiring explicit editing of the profile, no browsing facility is provided for it, and mechanisms are provided for a progressive fading of older information of this sort.

3.3. *Personalised selection of relevant news items based on long-term model*

A personalised selection of the news items relevant to a specific user model can be carried out with respect to each of the alternative reference frameworks available in the long-term user model. The system automatically computes relevance for each news item against all the frameworks, and subsequently applies a mathematical formula to calculate an overall relevance value for them. This final relevance value takes into account not only the particular relevance values obtained against each framework, but also the relative weighting of the different frameworks being used for that particular experiment, and it gives rise to a ranking of the news item with respect to the given long-term user model.

3.3.1. *Selection with respect to the section framework*

Given that each news item comes pre-assigned from its source to a certain section, selection against this framework is immediate. Each news item is processed to check which is the value associated in the user model to the section that it belongs to. The relevance value[1] r^s_{ij} between a news item i and a user model j is therefore straightaway the value assigned to its corresponding section s by the user j, referred to as S_{sj}.

$$r^s_{ij} = S_{sj} \tag{1}$$

[1] Super indices show the part of the model that the relevance refers to, that is, super index s in the first relevance indicates that this value is about the sections, k, keywords and f, feedback keywords.

3.3.2. Selection with respect to the keyword framework

The relevance value according to this reference framework between a news item i and a user model j is computed using the following formula [25]:

$$r_{ij}^{k} = sim(d_i, k_j) \tag{2}$$

where k_j is the term weight vector for the keywords for the user j, d_i is the term weight vector for news item i, and sim is the cosine formula of the VSM.

3.3.3. Computing overall relevance

When all the news items have been sorted according to the different sources of relevance, the resulting orderings are integrated by using the particular combination that is assigned to each of the different reference systems. Thus, the overall relevance between a news item i – belonging to a section s – and a user model j is computed using the following formula:

$$r_{ij} = \frac{\alpha r_{ij}^{s} + \gamma r_{ij}^{k}}{\alpha + \gamma} \tag{3}$$

where Greek letters α and γ show the significance assigned for this experiment to each of the different frameworks (α for sections, and γ for keywords). In order for this combination to be significant, the relevance obtained for each framework must be normalised with respect to the best results within that framework for the collection of news items under consideration. For instance, if the highest relevance assigned to docuemtns corresponding to that particular day in terms of keywords is 0.42, all relevances obtained from keywords for that day are normalised over that value before combining them with relevances arising from sections. Intuitively, this corresponds to adjusting the scales on the different frameworks so that the relevance value for the highest item in the ranking for each framework is actually 1. This ensures that different frameworks can rank items competitively on an equal footing, independently of whether the absolute relevance values obtained by means of their respective processes compare well. The sections framework, for instance, always assigns relevance 0 or 1 – an item either is or is not in a given section –, whereas the keywords framework may assign a low relevance to a document with two occurrences of user-defined keywords if other user-defined keywords are not present in it. If no better choice is available in terms of keywords, we want the document with two keywords to be assigned a relevance that can compete with that of documents chosen for belonging to user-selected sections, therefore it is normalised to 1 before combining them together.

The ratio between α and γ represents the relative importance assigned within the long-term model to the domain-dependent (sections) and the domain-independent (keywords) frameworks of classification.

A ranking of the news items is obtained according to their relevance for the given user obtained from the preceding formula. The top T news items of the resulting ranking are selected for the user.

3.4. Personalised selection of relevant news items based on short-term model

The short-term user model is obtained from those news items for which the user has provided either positive or negative feedback. A normal use of the system is assumed, in which each user receives only the T items that the system has considered more relevant to his interests. Therefore, to model accurately real system performance, only feedback information about these T more relevant items is used in each case. The representation of the short-term model is based on the set of feedback keywords obtained from those news items, which are used to dynamically evolve the representation of the user model.

Because this feature is intended to model short-term interests, we use an algorithm to decrease the weight of these terms with the passage of the time. Each day, the new weights are obtained by substracting a certain value D from the weights for the previous day. Whenever the weight for a feedback term becomes less than or equal to 0, the corresponding term is eliminated from the set. For this work a value of $D = 0.1$ has been chosen. Updating the set of feedback terms for a user therefore involves substracting 0.1 from the values of all feedback terms, and eliminating those feedback terms whose value drops below 0. This operation is carried out every day before the new feedback information is taken into account. The values resulting from this operation – for a given user j– are referred to as O_{wj} in the discussion that follows.

To take into account the new feedback information, the following calculations must be performed each day for each user j, since the set of news item available is different for each day, and feedback for a given day is different for each user. Let us assume that feedback information provided by the user j is formalised as two sets of news items: $F_j(+)$, the set of news item for which positive feedback has been provided, and $F_j(-)$, the set of news items for which negative feedback has been provided. A specific item can belong to one or the other set, or not belong to either, but never to both. The set of all news items for which some feedback – whether positive or negative – has been provided by user j is referred to as F_j. The selection of particular keywords to represent a given feedback item for a user j is achieved through the following steps [6].

For the selection/update of the new feedback terms all the documents are preprocessed in the same way as described for the process of selection: all stoplist words are eliminated, and a Spanish stemmer is applied to the remaining terms. The starting point for the process of adaptation is the set of terms representing the documents, with their associated frequency (tf).

The *article access value* (a_{wij}) for the word w in the news item i for user j is defined as:

$$a_{wij} = \begin{cases} E(HT_{wi} + B_{wi}) & \text{if } i \in F_j(+) \\ -R(HT_{wi} + B_{wi}) & \text{if } i \in F_j(-) \end{cases} \tag{4}$$

where T_{wi} is the frequency of keyword w in title area of news item i, B_{wi} the frequency of keyword w in the body of news item i, E the positive feedback weight, R the negative feedback weight, and H the title weight. Following [6], we have used the following values: $E = 0.9$, $R = 0.9$, and $H = 2$.

The *word update rate* (u_{wj}) is computed for each word w and user j, by gathering together the resulting weights for words appearing on several fed back news items and over the complete set of keywords that results from the previous step:

$$u_{wj} = \frac{\sum_{i \in F_j} a_{wij}}{\max(\mid \sum_{i \in F_j} a_{wij} \mid)} \tag{5}$$

The *new interest value* (N_{wj}) for the word w for user j is then computed as:

$$N_{wj} = \begin{cases} O_{wj} + P(1 - O_{wj})u_{wj} & \text{if } u_{wj} \geqslant 0 \\ O_{wj} - PO_{wj} \mid u_{wj} \mid & \text{if } u_{wj} < 0 \end{cases} \tag{6}$$

where *speed* (P) has value from 0 to 1, and represents the modifiable part of interest degree for one day[2], and O_{wj} is the old interest value of word w.

The R words at the top of the ranking based on N_{wj} are chosen as components of the weighted term vector (f_j) representing the feedback keywords for user j, and they provide an additional reference framework for the system during selection of specific news items for the given user. For our work we have chosen a value of $R = 10$.

The relevance value between a news item i and a user model j is computed using the following formula:

$$r^f_{ij} = sim(d_i, f_j) \tag{7}$$

where f_j is the term weight vector for the feedback keywords for the user j, and d_i is the term weight vector for the news item i.

3.5. Personalised selection of relevant news items based on combined short-term and long-term models

When all the documents have been sorted according to the different sources of relevance, the resulting orderings are integrated by using the particular combination that is assigned to each of the different reference systems. Thus, the overall relevance between a news item i – belonging to a section s – and a user model j is computed using the following formula:

$$r_{ij} = \frac{\alpha r^s_{ij} + \gamma r^k_{ij} + \delta r^f_{ij}}{\alpha + \gamma + \delta} \tag{8}$$

where Greek letters α, γ and δ show the significance assigned to each of the different references frameworks (α for sections, γ for keywords, and δ for feedback keywords). In order for this combination to be significant, the relevance obtained for each framework must be normalised with respect to the best results obtained within that framework for the collection of documents under consideration. This allows the keywords framework and the feedback framework – which deal in relevance values covering the whole range between 0 and 1 – to compete in equal terms with the sections framework – which assigns only binary relevance values.

The relative value of δ with respect to α and γ represents the relative importance assigned to the short-term model with respect to the long-term model. The values of α, γ and δ will act as variables to control the experiments. A ranking of the news items is obtained according to their relevance for the given user obtained from the preceding formula. The top T items of the ranking are selected for the user.

3.6. Item summarization

Our system uses three phrase-selection heuristics to build summaries. The first two are used to construct generic summaries, whereas the third one is used for personalized summaries. The three heuristics have a common objective, which is to assign a value to each sentence of the text being summarized. To generate

[2] After [6], we use $P = 0.8$.

Table 1
Values assigned to the first 5 sentences of a document for the position heuristic

Sentence number	1	2	3	4	5
Assigned value	1.00	0.9	0.98	0.95	0.90

summaries we employ the weighted combination of various heuristics. These values are later used to select the most relevant sentences, which will be used to form an extract of the news item later used as summary. We describe each heuristic separately and we comment on how we modify the different parameters in order to obtain generic or personalized summaries.

3.6.1. *Position heuristic*

In journalistic domains, the headline and the first few sentences of a text usually provide a good summary of the content of the following text. This heuristic assigns the highest value to the first five sentences of the text [28]. The specific values chosen in our system for these first 5 sentences are shown in Table 1. Sentences from the 6th on are assigned value 0. These provide the weights A_{oi} for each sentence o of a news item i using the position heuristic. These values are independent of the particular user j being considered.

3.6.2. *Thematic words heuristic*

Each text has a number of thematic words, which are representative of its content.[3] This heuristic extracts the M non-stoplist most significant words of each text and checks how many of these thematic words are found in each sentence. In this way, a higher value will be assigned to sentences that hold a highest number of thematic words [22,29].

To obtain the M most significant words of each document, documents are indexed to provide the weight of each word in each document using the *tf-idf* method [25]. As described when discussing news item representation, document frequencies are computed over the set of news documents for the corresponding day. The eight words with highest weight are selected for each document ($M = 8$).

To obtain the value for each sentence o within the document i using the thematic words heuristics (B_{oi}), the number of document key words (thematic words) appearing in the sentence is divided by the total number of words in the sentence. This is intended to give more weight to sentences with a higher density of thematic words [29]. The values obtained in this way are also independent of the particular user j being considered.

3.6.3. *Personalization heuristic*

The aim of this heuristic is to boost those sentences that are more relevant to a particular user model [26]. The potential of summary personalization is high, because a document that would be useless if summarised in a generic manner may be useful if the right sentences are selected that match the user interest.

The calculation of weights for each sentence o of news item i with respect to user j using the personalisation heuristic (C_{oij}) is done in the following way. The user model provides a vector of weighted terms (k_j) corresponding to the keywords specified by the user for his long-term model. This information is used to calculate the similarity between the user model j and each sentence o of news item i, assigning a final weight to the sentence by means of the following similarity:

$$C_{oij} = sim(s_{oi}, k_j) \tag{9}$$

where s_{oi} is the term weight vector representing sentence o of news item i.

A possible alternative is to take into account the additional short-term information provided by user feedback. In that case, the Eq. (9) for computing sentence value under the personalisation heuristic would be revised as:

$$C_{oij} = \frac{\chi sim(s_{oi}, k_j) + \beta sim(s_{oi}, f_j)}{\chi + \beta} \tag{10}$$

3.6.4. *Combining the three heuristics*

The following equation is applied to combine the values resulting from each of the three heuristics and provide a single value for each sentence:

$$Z_{oij} = \frac{\mu A_{oi} + \nu B_{oi} + \sigma C_{oij}}{\mu + \nu + \sigma} \tag{11}$$

The parameters μ, ν and σ allow relative fine-tuning of the different heuristics, depending on whether position (μ), key words (ν) or relevance to the user model (σ) is considered more desirable. Values of σ determine the degree of personalisation of the summaries: if σ is 0, the resulting summaries are generic, and for σ greater than 0 personalisation increases proportionally to σ. As was the case when combining results from different reference frameworks, in order for this combination to be significant, the relevance obtained for each heuristic must be normalised with respect to

[3] This set of content-based keywords for a document should not be confused with the set of keywords specified by a user to define his interests.

the best results obtained by that heuristic for the set of sentences for the document under consideration.

The summary is constructed by selecting the top L% of the ranking of sentences by the value Z_{oij} and concatenating them according to their original order of appearance in the document. For the work presented here L has been set at 20%.

3.7. Summary of system parameters

As explained above, system operation is determined by a number of parameters.

For the experiments presented in this paper, some system parameters have been fixed at empirically acceptable values in order to concentrate on studying the effect of alterations of the remaining parameters. The parameters assigned fixed values – represented by upper case Latin letters – are shown in Table 2.

The set of parameters for which different configurations are studied in the following experiments – represented by lower case Greek letters – is shown in Table 3. All these parameters take values between 0 and 1.

4. Evaluation of systems for information access

In this paper we explore the different alternatives for the dissemination of personalised information, as presented in the previous sections. The possibilities that are explored relate to the selection of news items and the summarization of the news items sent.

4.1. Experimental set up

The alternatives have been tested over an experimental set up specifically designed for the application of the user modelling, personalization, and selection mechanisms to the field of news item filtering and dissemination.

4.1.1. Experimental data

To evaluate the experiments, we have used the sets of news items corresponding to five consecutive days (Monday to Friday) from the digital edition of the ABC newspaper, a major Spanish daily. These days correspond to the period between the 6th and the 10th May 2002. The number of news items for each day is 128, 104, 87, 98 and 102, respectively. The total number of news items considered is 519, and the average number of news items per day is 104.

Additionally, 11 users with different profiles – in terms of their information interests – have been involved. The set of evaluators included 10 academics and 1 civil servant. Of the academics, 9 lecture on computer science topics, and 1 lectures on business issues. The set overall included 3 women and 8 men, of ages between 25 and 35.

To build the evaluation collection, full evaluations of all news item for the days under consideration where requested from every user. This corresponds to 55 evaluation episodes – 11 users over 5 different days, with a different evaluation episode on each day. In terms of evaluated documents, 11 users considered for evaluation the 519 documents, providing a total of 5,709 samples. These data were used to build the evaluation collection that is subsequently used as basis for all experiments. In these experiments, we resort to the full data obtained from the users only in order to calculate the recall and precision for system performance with respect to this information about real user interests over the whole set of options available.

The actual experiments were carried out under the assumption that the system will only send information to the user regarding the 10 best ranked news items[4] resulting from the filtering task. This implies that only user feedback information concerning news items amongst those 10 is considered as practical user feedback, whereas the rest of the feedback information contained in the collection (concerning items ranked by the system outside the 10 best news items) is only applied to construct the recall and precision metrics.

The initial user models for the profiles were built by the users the day before the start of the experiment. This made them available to be used by the system for selecting news item on the first day of the evaluation. These initial models have information about the long-term interests of the user, in this case represented in terms of sections and keywords.

In order to be able to carry out the evaluation of the system it was necessary to collect judgements from the users regarding the relevant and non-relevant news item for each user on each of the five days of the experiment. To provide these judgements users had to read through the complete sets of news items for each day and decide whether they would have been of interest or not. Users evaluate based on the standard format for presentation of news items in digital newspapers, which includes title, author, section, summary and link

[4] The actual number is determined by system parameter T.

Table 2
Fixed system parameters

System module	Parameter	Description	Range	Value
Short-term modeling	T	relevant items per day		10
	E	positive feedback weight	0 to 1	0.9
	R	negative feedback weight	0 to 1	0.9
	H	title weight		2
	P	speed	0 to 1	0.8
	R	number of words for f_j		10
	D	decay per day	0 to 1	0.1
Item personalization	–	1st sentences position heuristic		see Table 1
	M	thematic words of a document		8
	L	summary length in %		20

Table 3
Adjustable system parameters (values between 0 and 1)

System module	Parameter	Description
Reference framework weight	α	classification by sections
	γ	classification by keywords
	δ	short-term model
Item summarization	ν	weight position heuristic
	μ	weight thematic words heuristic
	σ	weight personalization heuristic
Summary personalization	χ	keyword weight personalization
	β	feedback weight personalization

to the complete news item. This experimental set up provides no way of knowing which or how many of these parameters play a role in user decisions. This is a shortcoming that we intend to address in further work. The interest they were asked to evaluate was not limited to whether the news item was relevant with respect to the initial profile. It also captured relevant news item that were not related to the original profile. It was hoped that information on these interests that had not been made explicit in the original profiles would be progressively taken into account by the system through the mechanisms of relevance feedback.

Table 4 shows the number of sections and keywords initially chosen by each user. In general, several sections and several terms are chosen in each profile. The average size of a choice is approximately 5 sections and 4 keywords. It is interesting to note that one user (user number 2) selected no keywords.

Table 5 shows the number of news items deemed relevant each one of the five days of the experiment. The average number of relevant news items per day varies between 25 and 34, with 28 as average value. The numbers suffer considerable variation, between 5 news items (10th May, user 1) and 80 (7th May, user 8).

There is also a significant difference between a set of more demanding users, who judge few news items as relevant (users 0, 1, 2, 3, 4 and 10 with approximately 17, 11, 11, 15, 15 and 18 relevant news items on average) and users who judge many news items as relevant (users 5, 6, 8 and 9 with 32, 56, 68 and 43 relevant news items on average).

4.1.2. Metrics

The way to evaluate a given selection quantitatively is to compare, according to a chosen metric, the selected documents with a set of relevance judgements provided by the user. There are many metrics in the field of information retrieval that can offer the results of the evaluation in the form of a curve, based on two values or based on a single value. These metrics can be grouped into several categories depending on the type of relevance and the type of retrieval employed [30–32]: binary relevance and binary retrieval, binary relevance and documents retrieved as a ranking, and relevance and retrieval both as ranking of the documents.

In this working framework we are considering binary relevance as stated by the users (whether or not a news item is relevant) and a ranking of the news items provided by the system. This suggested [25,31] the use of normalised precision and recall as metrics.

Normalised precision and recall [11] measure the effectivity of a particular ranking. Intuitively, this is best visualised over a graph where recall or precision are plotted against levels of the ranking. Normalised precision – or recall – represent the size of the area

Table 4
Number of sections and keywords chosen by each user

User	0	1	2	3	4	5	6	7	8	9	10	Average
NumSections	2	4	4	7	4	8	8	7	2	5	6	5.18
NumKeyWords	3	3	0	5	4	3	6	7	4	5	3	3.91

Table 5
Number of news items deemed relevant each day by each user

Day	Items	0	1	2	3	4	5	6	7	8	9	10	Av.
6-5	128	18	23	16	20	30	30	72	39	77	23	22	33.64
7-5	104	16	8	8	15	12	35	72	21	80	47	16	30.00
8-5	87	11	10	10	9	16	33	43	20	72	41	17	25.64
9-5	98	17	9	8	10	6	27	53	24	60	42	17	24.82
10-5	102	21	5	14	22	9	37	41	10	51	61	16	26.09
Av.	103.8	16.6	11.0	11.2	15.2	14.6	32.4	56.2	22.8	68.0	42.8	17.6	28.04

between the curve for the best solution given by relevant documents in the first positions of the ranking and the curve for the solution generated by the system. These metrics can also be considered as approximations to the average recall and precision obtained for all positions of the ranking.

Normalised recall is calculated using Eq. (12):

$$nR = 1 - \frac{\sum_{i=1}^{REL} RANK_i - \sum_{i=1}^{REL} i}{REL(N - REL)} \qquad (12)$$

where REL is the number of relevant documents, $RANK_i$ represents the ranking of the ith most relevant item, and N is the total number of items.

Normalised precision is calculated using Eq. (13):

$$nP = 1 - \frac{\sum_{i=1}^{REL} \log RANK_i - \sum_{i=1}^{REL} \log i}{\log(N!/(N-REL)!REL!)} \qquad (13)$$

The fact that the whole ranking is taken into account requires that these formulas add up the differences in values for the whole ranking and normalise it. The logarithmic terms in the formula for normalised precision arise from the fact that precision steadily decreases as subsequent non-relevant items are found in the ranking – and peaks again whenever a relevant item is found. The corresponding formula for normalised recall is simpler because the values for recall remain steady as subsequent non-relevant items are found in the ranking, and they simply jump up when a relevant item is found.

Additionally, for cases where equal relevance values are obtained for consecutive positions in the ranking, the average position number in the ranking has been chosen as position number for the whole conflicting set [25]. This adjustment avoids the problem of attributing a random relative ordering within the ranking to documents that have obtained equal relevance values.

4.1.3. Statistical significance

For two techniques A and B, if we want to show that A performs better than B ($A > B$) with respect to a parameter V, $V+$ represents the number of times that A has outperformed B, $V-$ represents the number of times that B has outperformed A, and the number of times that similar results have been achieved by both techniques is represented as ties. This is applied to normalised precision ($V = nP$) and normalised recall ($V = nR$).

We consider the results to be statistically significant if they pass the sign-test on paired samples at the 5% level ($p \leqslant 0.05$). This decision is based on the fact that there is no assumption about the underlying distribution, and, given the different normalization procedures being applied at various levels, the relative values rather than the actual magnitudes of relevance should be considered [25].

4.2. Personalised selection of items

We are basically concerned with trying to answer the following questions:

- Which mechanism provides the best results in terms of better selecting the news items that each user finds relevant?
- Within the long-term model, is it better to use only sections, only keywords, or a combination of both?
- Is it better to use a dynamic short-term model, a static long-term model, or a combination of both?

These questions give rise to the following hypotheses:

H1. The precision of news item selection with respect to user preferences, using only a static long-term model, is better if sections and keywords combined are employed than if either one of them is used separately.

H2. The precision of news item selection with respect to user preferences, is better if a static long-term model and a dynamic short-term model are combined than if either one of them is used separately.

These hypotheses are tested over the data obtained in the experiments as shown below.

4.2.1. Experiment 1: Combining categories and keywords for long-term modeling

To test the first hypothesis, we apply to all users the different selection mechanisms of the long-term model. These different mechanisms are configured by giving different values to the parameters of Eq. (8):

Se: only sections ($\alpha = 1, \gamma = 0, \delta = 0$)
Ke: only keywords ($\alpha = 0, \gamma = 1, \delta = 0$)
SeKe: the combination of sections and keywords ($\alpha = 1, \gamma = 1, \delta = 0$)

In each case, normalised recall and precision are observed. These experiments are repeated for the 5 days of the experiment. The average data per day over all the users, and the global average are presented in Table 6.

The results show that the combination of sections and keywords is significantly better than sections (11% on normalised precision) or keywords (33% on normalised precision) when used separately, both in terms of normalised precision and recall. This confirms hypothesis **H1**.

As an additional result, it is shown that the selection is significantly better when using only sections than when using only keywords.[5]

It is clear that the use of sections or keywords serves different purposes. Sections are better suited for situations in which the user has a general broad interest in items related to the section. Keywords are more useful in situations when the user has a specific interest in mind. For purposes of searching, it is convenient to give the user a choice. For specifying interests for a filtering application, an exclusive choice is too restrictive.

None of the features discussed is dependent on the specific domain under consideration. In general terms, this approach is applicable wherever there is an existing set of domain-dependent categories. The current system has so far been tested successfully for a digital newspaper, a virtual bookshop, and the Web site of a news agency [33].

Having established the combination of sections and keywords as the best option for identifying long-term preferences, this mechanism is employed from here on for all subsequent experiments presented in this paper.

4.2.2. Experiment 2: Combining short and long-term models for item selection

To test the second hypothesis, we apply to all users the different combinations of short and long-term models. Again, these different combinations the following values must be assigned to the parameters of Eq. (8):

Lo: only long-term model ($\alpha = 1, \gamma = 1, \delta = 0$)
Sh: only short-term model ($\alpha = 0, \gamma = 0, \delta = 1$)
LoSh: the combination of both models ($\alpha = 1, \gamma = 1, \delta = 1$)

In each case, normalised recall and precision are observed. These experiments are repeated for the last 4 days of the experiment, since there is no relevance feedback available to generate a short-term model for the first day. The average data per day and the global average are presented in Table 7.

The results show that the combination of short- and long-term models is significantly better in terms of normalised precision than using them separately (9% better than the long-term and 33% better than the short-term). This confirms hypothesis **H2**.

As an additional result, it is shown that, at least for periods of relevance feedback as short as this, the long-term model is significantly better than the short-term model when either is used on its own.

Regarding normalised recall, the combination of short-term and long-term models is not significantly better than the long-term model on its own, though one can see an apparent improvement resulting from the addition of the short-term model.

We can therefore conclude that the combination of short- and long-term models is better than the long-term model in terms of precision, but not in terms of recall. The long-term model is better than the short-term model, both regarding normalised recall and normalised precision.

In general, long- and short-term modeling constitute differing alternatives that give services to different

[5] It seems that the traditional organization of newspapers is supported by the experimental data for cases with a specific domain-dependent classification. The current prevalence of keyword search mechanisms may be a mechanism for coping with satisfying user preferences beyond one single domain, where a pre-established set of categories may hinder more than it helps.

Table 6
Average normalised precision (nP) and recall (nR) per day and the global average, for different combinations of sections (Se) and keywords (Ke)

	06-may2		07-may		08-may		09-may		10-may		Av.	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
SeKe	0.616	0.507	0.569	0.436	0.627	0.510	0.630	0.501	0.642	0.508	0.617	0.493
Se	0.595	0.444	0.566	0.410	0.610	0.461	0.625	0.457	0.619	0.444	0.603	0.443
Ke	0.550	0.395	0.505	0.325	0.547	0.376	0.533	0.383	0.553	0.378	0.538	0.371

Table 7
Average normalised precision (nP) and recall (nR) per day and the global average, using short (Sh) and long (Lo) term models

	07-may		08-may		09-may		10-may		Av.	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
LoSh	0.577	0.483	0.619	0.540	0.640	0.555	0.659	0.544	0.624	0.531
Lo	0.569	0.436	0.627	0.510	0.630	0.501	0.642	0.508	0.617	0.489
Sh	0.523	0.422	0.553	0.457	0.498	0.367	0.471	0.350	0.511	0.399

needs. Long-term modeling is possibly better suited to information filtering applications because it does not require the user to provide feedback in order to operate efficiently and it can capture known long-term interests. On the other hand it cannot easily keep track of changes in user interest when they happen frequently. Short-term modeling presents definite advantages in cases where there is no established set of interests but rather a changing need to shift the filtering process along with a frequent change in user interests. However, it requires a reasonable amount of feedback before it can start operating efficiently, and the training process may be long. The system described here would present advantages in those cases where needs of the two types described meet in a single context: there are long-term interests to be tracked as well as recent changes to be taken into account.

The short-term modeling component is not dependent on the domain in any way, however, the long-term modeling component is, as described above, partially dependent on the domain through application of specific categories.

Having established the combination of short- and long-term models as the best option, this mechanism is employed from here on for all subsequent operations of item selection involved in the remaining experiments presented in this paper.

4.3. *Item summarization*

The main issue to be tested regarding the use of summarization in an information filtering and dissemination setting is to what extent the use of summaries instead of the complete document involves a loss of significant information for the user. This is tested in the experiments described below.

4.3.1. *Summary evaluation*

Summaries are evaluated independently of the rest of the experiments using a technique of indirect evaluation proposed elsewhere before [19]. The technique is based on the assumption that if a summarization process is good, the resulting summary should have retained as much as possible of the information that ensures correct retrieval according to the given user profile. For each user, a personalised version of the complete evaluation collection is built by summarising (using the heuristic that is to be tested) each news item. The very same process of selection applied in experiment 2 is repeated (in truth, restricted to the user profile corresponding to the user for whom the collection has been personalised), but using the new personalised (summarised) version of the evaluation collection. The generated summaries have been used as input data in a selection process equivalent to that carried out for the set of complete news items for experiment 2. For each user, on each day the system is requested to rank according to the corresponding user profile a document collection that contains the constructed summaries of the documents in place of the documents themselves. This ranking is then compared against the users relevance judgements as registered for the full documents. The selection mechanism employed has been the one that gave best results for experiment 2, combining short-term and long-term models. The hypothesis is that, if the summarization process employed preserves the information that is relevant to that user profile, the results obtained should mirror exactly those obtained for this user in experiment 2 – where the system is requested to rank according to the corresponding user profile the collection of full documents – , which are taken as reference value. Any deviations from those values indicate loss of information due to "leaks" during summarization, which have forced the resulting ranking for the summarised items to deviate from the one obtained using the complete news item as input. By applying a similar process to a given summarization heuristic, this experiment should provide an explicit – though admittedly indirect – measure of its adequacy for personalised summarization (where *personalised summarization* is understood as a process of summarization that preserves the specific information that is relevant to a given user profile, rather than information that truly summarises the content of the news item).

For this evaluation, summaries have been generated for all the news items for one day for all the users. This means that a summary has been generated for each news item and each user, for all the news items of a given day. This process is repeated for the 5 days of the experiment.

4.3.2. Experiment 3: Personalized summarization

An initial issue involves discovering how different configurations of the system parameters involved can affect the technique described above for the personalization of summaries. Generation of personalised summaries involved a combination of long-term (keywords explicitly provided by the user in his profile) and short-term models (keywords obtained by the system from user feedback and stored as part of his user profile). Although combining these two sources has been shown to be better for item selection this may not necessarily be the case for generation of personalised summaries.

We nevertheless start from the following hypothesis:

H3. Summaries obtained by using only the personalization heuristic are better if a combination of short-term and long-term models is used as source for user interests than if either model is used on its own.

We are considering here summaries obtained using only the personalisation heuristics (assigning the values $\mu = 0, \nu = 0, \sigma = 1$ to the parameters in Eq. (11)). The differences between the summaries arise from the following assignment of values to parameters in Eq. (10):

Ps(**Lo**) personalised summary obtained using only the keywords of the long-term model ($\chi = 1, \beta = 0$)
Ps(**Sh**) personalised summary obtained using only the feedback
Ps(**LoSh**) personalised summary obtained using a combination of keywords of the short- and long-term models ($\chi = 1, \beta = 1$)

The conclusions that can be drawn from the results shown in Table 8 can be summarised as follows. Summaries generated using a combination of short- and long-term models (**Ps**(**LoSh**)) are better than summaries obtained using only the long-term model (**Ps**(**Lo**)), with a statistically significant improvement in precision. Summaries obtained using only the short-term model (**Ps**(**Sh**)) are better than summaries obtained using only the long-term model (**Ps**(**Lo**)), with a statistically significant improvement in precision (3%). Summaries generated using a combination of short- and long-term models (**Ps**(**LoSh**)) are only slightly better in precision (1%) than summaries obtained using only the short-term model (**Ps**(**Sh**)). Although the result is not statistically significant, we are encouraged to retain **H3** as a reasonable working hypothesis.

From here on, when personalised summarization is discussed (**Ps**), the personalisation has been done using a combination of long-term keywords and short-term feedback keywords.

4.3.3. Experiment 4: Combining techniques for summarization

In this case, the questions that need to be answered are:

- How much is lost, in terms of information received, by sending a summary of a news item instead of the complete document?
- Which type of summary is better in that sense?

The hypotheses to be tested are:

H4. Summaries obtained by using only the personalization heuristic are better in terms of precision with respect to information selected by the user than summaries obtained using generic summarization heuristics and summaries obtained using a combination of heuristics.

H5. Summaries obtained by using only the personalization heuristic are worse than complete news items in terms of precision with respect to information selected by the user.

H6. Summaries obtained by using only the personalization heuristic are better in terms of precision with respect to information selected by the user than summaries obtained by extracting the first sentences of the complete news item.

The summaries involved can be of different types, depending on the specific heuristics employed to generate them as determined by the assignment of values to the parameters in Eq. (11):

Table 8
Average normalised precision (nP) and recall (nR) per day, and the global averages, for personalised summaries built with different combinations of models: long- and short-term – Ps(LoSh) –, long-term model only – Ps(Lo) –, and short-term only – Ps(Sh)

	06-may2		07-may		08-may		09-may		10-may		Av.	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
Ps(LoSh)	0.616	0.508	0.581	0.486	0.620	0.538	0.632	0.535	0.657	0.550	0.621	0.523
Ps(Lo)	0.616	0.508	0.574	0.482	0.611	0.532	0.625	0.523	0.631	0.507	0.611	0.510
Ps(Sh)	0.610	0.497	0.588	0.493	0.622	0.550	0.624	0.526	0.642	0.524	0.617	0.518

Table 9
Average normalised precision (nP) and recall (nR) for complete news items (Nw) and various types of summaries (Ps, GPs, Gs, Fs) per day, and the global averages

	06-may2		07-may		08-may		09-may		10-may		Av.	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
Nw	0.616	0.507	0.577	0.483	0.619	0.540	0.640	0.555	0.659	0.544	0.622	0.526
Ps	0.616	0.508	0.581	0.486	0.620	0.538	0.632	0.535	0.657	0.550	0.621	0.523
GPs	0.612	0.497	0.563	0.448	0.603	0.527	0.625	0.521	0.649	0.539	0.611	0.506
Gs	0.607	0.479	0.563	0.452	0.612	0.526	0.631	0.529	0.652	0.538	0.613	0.505
Fs	0.607	0.479	0.557	0.450	0.598	0.513	0.622	0.522	0.649	0.536	0.607	0.500

Fs. (baseline reference) first sentences of the corresponding news item

Gs. using generic heuristics (position and keywords) ($\mu = 1, \nu = 1, \sigma = 0$)

Ps. using personalization heuristics (combining short- and long-term models) ($\mu = 0, \nu = 0, \sigma = 1$)

GPs. using both types of heuristics ($\mu = 1$, $\nu = 1$, $\sigma = 1$)

Several different evaluation collections – consisting each one of summaries obtained from the news items in the original collection by applying a different summarization heuristic – are built for each user. The procedure described in experiment 3 is applied to each one of these collections, using the corresponding user profile as source for user interests. If different summarization heuristics lead to different degrees of loss of relevant information, the resulting rankings will differ amongst them in a proportional way.

Evaluation of personalised summaries is more costly because each user must be evaluated separately. Generic summaries are the same for all users, so all the users can be evaluated simultaneously. The system is run once for all users, using the set of generic summaries as input. Personalised summaries require a different procedure, since the selection for each particular user must be obtained from his very own set of personal summaries.

The analysis of the results shown in Table 9 indicates that personalised summaries (**Ps**) give significantly better results with respect to normalised precision of the selected information than generic summaries (**Gs**) and generic-personalised summaries (**GPs**). In both cases the improvement is statistically significant (4% for **GS** and 3% for **GPs**). This confirms hypothesis **H4**. Generic-personalised summaries (**GPs**) are better than generic summaries (**Gs**), and generic summaries (**Gs**) are better than summaries based on first sentences (**Fs**), but in neither case is the difference statistically significant. It can also be seen that personalised summaries (**Ps**) are worse (1% on normalised precision) than complete news items (**Nw**) under the same view point. This confirms hypothesis **H5**.

Personalised summaries (**Ps**) are better than summaries based on the first sentences of the news item (**Fs**), with a statistically significant improvement (5 % in normalised precision). This confirms hypothesis **H6**.

It seems apparent from these results that generic summaries perform very closely to summaries obtained by taking the first few lines of the news item. This seems to indicate that the position heuristic is overpowering the thematic word heuristic, which may be corrected by refining the choice of weights. In any case, although a first-sentences approach may provide good results for indicative summarization, it does not do so well in terms of personalised summarization – as defined above –, where it is crucial to retain in the summary those specific fragments of the text that relate to the user profile.

This explains why the generic-personalised summaries perform so poorly in spite of being a combination of good techniques: given a fixed limit on summary length, the inclusion of sentences selected by the generic heuristics in most cases pushes out of the final summary information that would have been useful from the point of view of personalisation.

As it has been shown, automatic summarization has the choice of applying generic techniques designed to capture the way in which the more relevant information is distributed over a typical document of the domain under consideration, or to apply personalisation techniques that concentrate on specific contents for which the user is known to have an interest (either because he has explicitly stated it in a long-term model or because it has been dynamically captured in a short-term model). The newspaper domain provides a good example where the structuring of information gives very useful cues as to their relevance (news items are usually built as inverted pyramids in terms of relevance: the most relevant information at the top, with relevance decreasing as they are read). This leads to quite simple summarization techniques providing good results in terms of indicative summarization – allowing the user to get an idea of what a document is about – though such a method would be severely domain-dependent, and might not work as well for different domains. The methods proposed here ensure the selection of the relevant information in terms of information needs, operating efficiently for personalised summarization – providing the user an extract of the specific contents of a document that are related to his interests –, with no domain-dependent assumptions.

5. Conclusions

In this paper, we have presented a user model which represents separately short-term needs and long-term multiple interests. Users can express their long-term preferences both in terms of domain-dependent (sections) and domain-independent (keywords) content-based information. This representation works like a stereotypical definition that avoids starting with an empty user model that is trained by user feedback. When starting from an empty user model, the initial training phase may become frustrating for users if many irrelevant news items are selected. This new approach solves the difficulties presented for some users – beginners mainly – by the method of providing a set of keywords and their associate weights used by some filtering systems. Also, application of feedback allows these initial definitions to be enhanced and evolve together with user's interest.

One possible disadvantage of the method presented here is that the system has no explicit mechanism for recognising as interesting spectacular news items that break suddenly, such as for example, the 11-S terrorist attacks on the World Trade Center in New York or the 11-M Madrid train bombings. The ability of the system to recognise any such news item as interesting is restricted only in as much as it is possible that some users of the system may not receive it, whereas common sense dictates that such news item may be universally interesting. However, the system provides at least three levels of filtering at which specific users may be able to specify their interest on events of a similar nature: users may have explicitly selected the sections under which the news item is classified, they may have provided terms that may occur in them, such as terrorism, or they may have provided feedback on similar items which the system will have used to update their short term model.

An alternative method that would have succesfully identified breaking news items as interesting is collaborative filtering. This method, however, presents a major drawback in the context in which the present system is designed to operate. The mode of delivery considered here is based on that of daily newspapers: the set of relevant news items for a given day is collected at a certain moment in time and bulk processed for the whole list of available user models. As a result, personalised selections are sent simultaneously to all registered system users. At the time of collating a selection for a given user from the news item of that day, no feedback is available from any other user on that particular set of news items. The system is faced with an instance of the first-evaluator problem that makes it impossible to apply a collaborative filtering solution. In the face of this restriction, we consider that the combination of different layers of interest-specification methods provides a better solution than the use of any one of them on its own would have done. A possible enhancement to the system might be to add a number of subsequent mailings throughout the day in which items deemed interesting by means of collaborative filtering techniques are sent to each user, based on user feedback on the news items of the day received up to that moment in time.

Our approach includes a summarization subsystem that generates different kinds of summaries adapted to the user. The idea is to allow the users to decide about the relevance of the received news items without inspecting the full text document.

We have also presented a systematic evaluation of the different alternatives for personalised information dissemination. The use of linear combinations of various frameworks has been shown to improve results in precision both when combining different frameworks for specifying interest (categories and keywords) and when combining long-term and short-term models. With respect to summarization, personalised summaries perform better than other combinations, providing an interesting alternative, better than summaries obtained from the first lines of the text and only slightly worse than the full text. This leaves open the possibility of using a personalised summary in place of the complete news item with an acceptable loss of information where convenient.

The results in recall were less clear, with improvements in the first experiment, but no significant improvements in the others. These results can be due to the behaviour of the normalised recall metric. That is, the normalised recall is sensitive to the classification of the last relevant item, while the normalised precision is sensitive to the classification of the first correct item. Under the circumstances, our algorithms improve the precision because they raise the more relevant documents to the top of the ranking, but they also drop the less relevant documents to the bottom. This can be due to the user's relevance judgements, which include documents very related with his initial profile, but also others that are, at most, very distantly related. These last documents might have been detected by the relevance feedback mechanism and this information used to correct the trend. However, given the peculiarities of the dissemination process, the user can only provide feedback on the documents at the top of ranking – as specified by his upper bound on the number of news item that he is to receive per day. This implies that documents that, if spotted on time, might have risen in the ranking thanks to the feedback process, not only fail to rise in the ranking but fall to the bottom.

However, we are more interested in precision than in recall because we send to the user only the documents at the top of the ranking within the upper bound expressed by the user – or in the particular case of the experiments described in this paper, defined by the system configuration parameter T –, and the number of relevant documents selected for a user is always greater than this limit. On the other hand, the fact that the system allows users to specify explicitly how many news items he wants to receive each day makes it necessary for us to consider how well the system behaves in terms of recall. Additionally, it further justifies the decision to use normalised precision and recall: by computing values over the complete ranking rather than just the fragment of it above the (arbitrary) cut-off point, the obtained results remain relevant whatever cut-off value for parameter T is used in subsequent runs of the system. The impossibility of establishing a fixed cut-off value under this set up also makes it difficult to use other metrics where recall and precision are compounded, such as the F measure, because they rely on establishing such an explicit cut-off value.

It is clear that the task of constructing a personal profile at the start of system use requires a significant effort on the part of the user. There are two possible arguments against considering this a significant drawback.

On one hand, it would be possible to develop a user interface in which the users would read an on-line newspaper, and the items clicked would be deemed of interest to him. This is in fact how the users evaluate the selections provided by the system, and also the way in which the dynamic short term modelling via user feedback is carried out. The implicit extraction of terms is the mechanism underlying the short term modeling described in the paper. The fact that its results differ somewhat from those obtained using terms provided explicitly by the user suggests that replacing one for the other is not trivial. Furthermore, the results presented here indicate that a combination of the various techniques performs better than any one of them on their own. This seems to validate the idea that the effort involved in explicitly constructing a user profile is worthwhile in terms of system performance.

On the other hand, having an explicit user profile allows the user to browse the representation that the systems has of his interests at any moment. Although the possibility has not been considered in the experiments described in this paper, the user of such a system might decide to modify his explicit user profile after some time using the system, either in view of system results or simply to capture recognised shifts in his interests. Were the modeling mechanism totally implicit, these possibilities would not be available.

In general, a key point of the present proposal is that such an – admittedly – information intensive method for information personalisation is indeed validated by improval of system results. Those users willing to trade a slight decrease in precision in exchange for less effort on system start-up can opt for systems based exclusively on implicit modelling.

The system as described in this paper includes a number of parameters that govern its operation. Of all these, the experiments described here attempt to establish adequate operating values for those parameters presented in Table 3. Similar experiments may have to carried out to obtain equivalent experimental support for the values assigned to the parameters presented in Table 2, which for the purposes of the experiments described in this paper have been assigned fixed values. Some of these values have been selected on the basis of similar studies in the bibliography (as outlined in the relevant sections of the paper), others have been chosen because they seemed reasonable in the face of no evidence to the contrary. It has already been discussed above how some of these decisions affect issues relevant to the content of the paper – such as the need for considering recall as well as precision, in view of the existence of parameter T governing the number of news items received each day by each user. Further work will address the experimental determination of optimal values for these parameters wherever possible. In certain cases, the volume of data that must be collected in order to carry out this task implies the effort may not be worth the returns it may provide.

The technology employed is not domain specific and relies solely in general techniques. Therefore it can be very easily ported to other domains as long as they have both a domain-specific classification system and textual descriptions of the items to be selected. Additionally, the message sent to the users can be ported to WAP/PDA technologies by simply changing the markup language (WML/simplified HTML) used to construct the message.

We can therefore conclude that the combined personalization approach of using a combination of long and short-term with user adapted summaries is a useful tool to assist users in a personalization system. Notwithstanding, the information contained in these summaries can not replace the full text document from an information retrieval point of view.

In future work, we will try to explore the possibility of obtaining feedback for the user models from the different kinds of summaries and explore its effectiveness. We are also interested in carrying out experiments with more users and during more days to extract more informative conclusions. Another line of research could be to add more information to the profile to improve the modeling of the users and to explore other techniques to perform the feedback.

References

[1] B. Chiu and G. Webb, Using decision trees for agent modeling: improving prediction performance, *User Modeling and User-Adapted Interaction* **8**(1–2) (1998), 131–152.

[2] L. Chen and K. Sycara, WebMate: A personal agent for browsing and searching, in: *Proceedings of the 2nd International Conference on Autonomous Agents* (*Agents'98*), K.P. Sycara and M. Wooldridge, eds, ACM Press, New York, 1998, pp. 132–139.

[3] D. Billsus and M. Pazzani, *A hybrid user model for news story classification,* in: Proceedings of the Seventh International Conference on User Modeling. Banff, Canada, Springer-Verlag, June 20–24, 1999, 99–108.

[4] T. Yan and H. Garcia-Molina, *SIFT-A tool for wide-area information dissemination,* in: Proc. 1995 USENIX Technical Conference, New Orleans, 1995, 177–186.

[5] F. Asnicar, M.D. Fant and C. Tasso, User model-based information filtering, in: *Fifth Conference of the Italian Association for Artificial Intelligence* (*AI*IA97*), M. Lenzerini, ed., Springer-Verlag, Roma, 1997.

[6] T. Nakashima and R. Nakamura, *Information filtering for the newspaper,* in: IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, BC, Canada, August 20–22, 1997.

[7] V. Luque, C. Fernández and C. Delgado, *Personalizing your electronic newspaper,* in: Proceedings of the 4th Euromedia Conference (WEBTEC), April 26–28, 1999.

[8] P. Chesnais, M. Mucklo and J. Sheena, *The fishwrap personalized news system,* in: IEEE Second International Workshop on Community Networkins Integrating Multimedia Services to the Home, 1995.

[9] D. Billsus and M. Pazzani, User modeling for adaptive news access, *User Modeling and User-Adapted Interaction* **10** (2000), 147–180.

[10] K.F. Gates, P.B. Lawhead and D.E. Wilkins, Towards and adaptive www: a case study in customized hypermedia, *The New Review of Hypermedia and Multimedia* **4**.

[11] J.J. Rocchio, Relevance feedback in information retrieval, in: *The SMART Retrieval System: Experiments in Automatic Document Processing,* G. Salton, ed., Prentice-Hall, 1971.

[12] D. Goldberg, D. Nichols, B.M. Oki and D. Terry, Using collaborative filtering to weave an information tapestry, *Communications of the ACM* **35**(12) (1992), 61–70.

[13] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl, *GroupLens: An open architecture for collaborative filtering of netnews,* in: Proceedings of the Conference on Computer-Supported Cooperative Work, 1994, 175–186.

[14] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon and J. Riedl, GroupLens: Applying collaborative filtering to Usenet news, *Communications of the ACM* **40**(3) (1997), 77–87.

[15] L. Terveen, W. Hill, B. Amento, D. McDonald and J. Creter, PHOAKS: a system for sharing recommendations, *Communications of the ACM* **40**(3) (1997), 59–62.

[16] M. Balabanovic and Y. Shoham, Content-based, collaborative recommendation, *Communications of the ACM* **40**(3).

[17] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes and M. Sartin, *Combining content-based and collaborative filters in an online newspaper,* in: ACMSIGIR Workshop on Recommender Systems, Berkeley, CA, 1999.

[18] N. Good, J.B. Schafer, J.A. Konstan, A. Borchers, B.M. Sarwar, J.L. Herlocker and J. Riedl, *Combining collaborative filtering with personal agents for better recommendations,* in: Proceedings of the Sixteenth National Conference on Artificial Intelligence AAAI/IAAI, 1999, 439–446.

[19] M. Maña, M. de Buenaga and J.M. Gómez, 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* (*ECDL'99*), *Lecture Notes in Computer Science,* (Vol. 1696), S. Abiteboul and A. Vercoustre, eds, Springer-Verlag, 1999, pp. 198–214.

[20] I. Mani and M. Maybury, eds, *Advances in Automatic Text Summarization,* The MIT Press, Cambridge, Massachusetss, 1999.

[21] U. Hahn and U. Reimer, Knowledge-based text summarization: Salience and generalization operators for knowledge-based abstraction, in: *Advances in Automatic Text Summa-*

rization, I. Mani and M. Maybury, eds, The MIT Press, Cambridge, Massachusetss, 1999.

[22] J. Kupiec, J.O. Pedersen and F. Chen, A trainable document summarizer, in: *Research and Development in Information Retrieval,* Proceedings of the 18th ACM SIGIR conference on research and development in information retrieval, 1995, pp. 68–73.

[23] G. Salton, J. Allan, C. Buckley and A. Singhal, Automatic analysis, theme generation and summarization of machine-readable texts, *Science* **264** (1994), 1421–1426.

[24] A. Díaz, P. Gervás and A. García, Evaluating a user-model based personalization architecture for digital news services, in: *Research and Advanced Technologies for Digital Libraries, Lecture Notes in Computer Science 1923,* J. Borbinha and T. Baker, eds, Springer, 2000, pp. 259–268.

[25] G. Salton, *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer,* Addison-Wesley Publishing, Reading, Massachusets, US, 1989.

[26] I. Acero, M. Alcojor, A. Díaz and J. Gómez, Generación automática de resúmenes personalizados, *Procesamiento del Lenguaje Natural* **27** (2001), 281–290.

[27] G. Amato and U. Straccia, User profile modeling and applications to digital libraries, in: *Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries,* ECDL, no. 1696, S. Abiteboul and A.-M. Vercoustre, eds, Springer-Verlag, 1999, pp. 184–197.

[28] H. Edmundson, New methods in automatic abstracting, *Journal of the ACM* **2**(16) (1969), 264–285.

[29] S. Teufel and M. Moens, *Sentence extraction as a classification task,* in: Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, 1997, 58–65.

[30] C. van Rijsbergen, *Information retrieval,* Butterworths, London, UK, 1979.

[31] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval,* McGraw-Hill, New York, US, 1983.

[32] S. Mizzaro, A new measure of retrieval effectiveness (or: What's wrong with precision and recall), in: *International Workshop on Information Retrieval* (*IR'2001*), T. Ojala, ed.), Infotech Oulu, 2001, pp. 43–52.

[33] A. Díaz and P. Gervás, Three information filtering applications on the internet driven by linguistic techniques, *Revue Francaise de Linguistique Appliquee* **2** (2000), 137–149.