Development and Use of an Evaluation Collection for Personalisation of Digital Newspapers

Alberto Díaz, Pablo Gervás, Laura Plaza
Dep. Ingeniería del Software e Inteligencia Artificial
Facultad de Informática - Universidad Complutense de Madrid
c/ Profesor José García Santesmases, s/n, Madrid 28040, Spain
albertodiaz@fdi.ucm.es, pgervas@sip.ucm.es, lplazam@fdi.ucm.es

Antonio García
Departamento de Comunicación y Ciencias Sociales
Facultad de Ciencias de la Comunicación y del Turismo - Universidad Rey Juan Carlos
Camino del Molino, s/n Fuenlabrada, Madrid 28943
antonio.garcia@urjc.es

Abstract

This paper presents the process of development and the characteristics of an evaluation collection for a personalisation system for digital newspapers. This system selects, adapts and presents contents according to a user model that define information needs. The collection presented here contains data that are cross-related over four different axes: a set of news items from an electronic newspaper, collected into subsets corresponding to a particular sequence of days, packaged together and cross-indexed with a set of user profiles that represent the particular evolution of interests of a set of real users over the given days, expressed in each case according to four different representation frameworks: newspaper sections, Yahoo categories, keywords, and relevance feedback over the set of news items for the previous day. This information provides a minimum starting material over which one can evaluate for a given system how it addresses the first two observations - adapting to different users and adapting to particular users over time - providing that the particular system implements the representation of information needs according to the four frameworks employed in the collection. This collection has been successfully used to perform some different experiments to determine the effectiveness of the personalization system presented.

1. Introduction

A personalization system selects, adapts and presents contents according to user models that define information needs (Mizarro & Tasso, 2002; Billsus & Pazzani, 2007). To evaluate such systems one must take into account the following observations: what fascinates one reader may bore another one, what fascinates a certain reader today may bore him tomorrow, and different people prefer different ways of expressing their interests.

Then, the existence of an evaluation collection is essential to allow the systematic evaluation of this kind of systems, beyond showing system operation for a particular set of cases that need not demonstrate real efficiency. Additionally, it provides a framework over which different proposals can be compared.

This type of collection is commonly used for other tasks associated with text classification, such as information retrieval; however there are no such collections for content personalisation tasks.

The choice of how to represent user interests should take into account the particular domain in which the system operates. In our case, this is the domain of digital newspapers. The information collected to document the user models behind user judgments, therefore, is based on a combination of interests expressed in terms of newspaper sections, categories and keywords (Díaz et al 2001). Another fundamental aspect that needs to be collected and represented in a collection is how the interest of the user evolves over time, as featured in the relevance feedback that he provides to the system. The adaptation mechanisms of the system should be able to follow any changes implicit in this feedback.

The collection presented here contains data that are cross-related over four different axes: a set of news items from an electronic newspaper for a particular sequence of days, a set of user profiles that represent the particular evolution of interests of a set of real users over the given days, expressed in each case according to four different representation frameworks: newspaper sections, Yahoo categories, keywords, and feedback keywords obtained from relevance feedback over the set of news items for the previous days.

This work is organized as follows. Section 2 describes the characteristics of evaluation collections for personalization. Section 3 describes the personalization system used as reference framework. Section 4 presents the process of construction of the collection. Section 5 shows the use of the collection to evaluate the personalization system presented. Section 6 outlines the main conclusions.

2. Evaluation collections for personalization

An evaluation collection for text classification is composed of a set of documents with a similar structure – usually restricted to particular domains, such as journalism or finance -, a set of tasks to be carried out over the documents, and a set of results for those tasks cross-indexed with the documents in the set – usually a set of judgments established manually by human experts. For instance, in information retrieval the tasks to be carried

out are queries presented over the documents in the collection, and the results are relevance judgments associated with each query. Typical collections have been used in TREC[1] conferences.

Evaluation collections for personalisation, such as the one described in this paper, present a major difficulty when compared with evaluation collections for other tasks: they require different relevance judgments for each and every one of the users and for every particular day. This is because the task to be carried out is to select the most relevant documents for each user on each day, and each user has different information needs – as featured in his user model – that may vary over time as the user becomes aware of new information. These relevance judgments could either be generated artificially by a human expert by cross checking each user model with the set of documents for a given day – very much in the way the system is expected to do –, or they can be established each day for the given documents by the real user who created the user model. This second option is more realistic, since real users determine the relevance of the given documents with respect to their interests at the moment of receiving them, therefore using their current information needs. In existing evaluation collections for text classification (i.e. the Reuters-21578 Text Collection[2]) this is not done, because judgments are generic for all possible users and they are generated by a human expert that does not know what the particular information needs may be for different users involved in different tasks at different times.

Because personalisation is expected to adapt over time to changes in user interests, in addition to generic judgments about relevance, the specific judgments on document relevance provided by each user (positive/negative/indifferent) must also be considered in terms of relevance feedback on the selection process. This implies that the subsets of documents being studied must be considered in a particular sequence, and the effect of user relevance feedback for previous days is also implicitly recorded in the collection in terms of the relevance judgments of that same user for a given day. Again, this type of information is generally not available in typical evaluation collections for text classification, because their evaluation is generally performed in static contexts that do not contemplate possible changes of relevance over time.

3. A personalization system for digital newspapers

Our personalization system is based on 3 main functionalities: content selection, user model adaptation and presentation of results (Díaz & Gervás, 2005). Content selection refers to the choice of the particular subset of all available documents that will be more relevant for a given user, as represented in his user profile or model. In our case, the documents come from the daily Spanish newspaper ABC. On the other hand, user model adaptation is built upon the interaction of the user with the system, which provides the feedback information used to evolve the profile. At last, results presentation involves generating a new result web document that contains, for each selected news item, a personalized extract considered indicative of its content (Díaz & Gervás, 2007).

In the next sections are described the user model, the multi-tier content selection and the result presentation processes used in the system.

3.1 User Model

The proposed user model consists of the combination of two types of user interests: long term and short. The long term model reflects information needs that remain stable across the time. The short-term model reflects the changes on these needs through the feedback of the user.

In the long term model, the first tier of selection corresponds to the 7 more important sections of the digital newspaper. The user can assign a weight to each section (S_{su}). For the second tier, the user enters a set of keywords, with a weight associated, to characterize his preferences (k_u). For the third tier the user must choose, and assign a weight to them, a subset of the 14 categories in the first level of Yahoo! Spain (C_{cu}). These categories are represented as term weight vectors (c) by training from the very brief descriptions of the first and second level of Yahoo! Spain categories entries. In the fourth tier, short-term interests are represented by means of feedback terms (f_u) obtained from feedback provided by the user over the documents he receives (Díaz & Gervás, 2004).

3.2 Multi-tier content selection and result presentation

Documents are downloaded from the web of the daily Spanish newspaper as HTML documents. For each document, title, section, URL and text are extracted, and a term weight vector representation for the document d (d_d) is obtained after the application of a stop list, a stemmer, and the tf · idf formula for computing actual weights (Salton & McGill, 1983).

Each document is assigned the weight corresponding to the section associated to it in the particular user model, which represents the similarity between a document d, belonging to a section s, and a user model u (s^s_{du}). The similarities between a document d and a category c (s_{dc}), between a document d and the keywords of a user model u (s^k_{du}), and between a document d and the feedback terms of a short-term user model u (s^f_{du}) are computed using the cosine formula for similarity within the vector space model (Salton & McGill, 1983):

$$s_{dc} = sim(d,c) \quad s^k_{du} = sim(d,k_u) \quad s^t_{du} = sim(d,t_u)$$

[1] Text Retrieval Conference (TREC). Home Page: http://trec.nist.gov/

[2] Reuters 21578 Text Collection. Home Page: http://www.daviddlewis.com/resources/testcollections/reuters21578/

The similarity between a document d and the categories of a user model (s^c_{du}) is computed using the next formula:

$$s^c_{du} = \sum_{i=1}^{14} C_{iu} s_{dc_i} \Big/ \sum_{i=1}^{14} C_{iu} \quad (4)$$

Finally, the results are integrated using a particular combination of tiers of selection. The similarity between a document d and a user model u (s_{du}) is computed as:

$$s_{du} = \frac{\delta s^s_{du} + \varepsilon s^c_{du} + \phi s^k_{du} + \gamma s^t_{du}}{\delta + \varepsilon + \phi + \gamma} \quad (1)$$

where Greek letters δ, ε, ϕ, and γ represent the importance assigned to each of the tiers of selection: sections, categories, keywords, and feedback terms, respectively. To ensure significance, the relevance obtained from each tier must be normalized.

The format of the new document generated during the result presentation process is (Figure 1): a title with the date and the name of the user, a brief description of the interests of the user, a link to the user model edition, the selected news items ordered by relevance and, for each document: title, author, section, source, relevance, feedback icons and automatic generated summary adapted to the user (Díaz & Gervás, 2005).

4. An evaluation collection for personalization of digital newspapers

The construction of the collection has been carried out in several stages. To start with, news items were downloaded each day from the webpage of the digital newspaper, and the information considered relevant for the collection was extracted from the HTML format. The interests of several users, reflected in the form of long-term user models, in the sense that they should not change over time, were collected. Finally, the relevance judgments concerning to each news item were collected for each of the days during which the process of constructing the collection took place. User models are defined the day before news items start to be processed, and the process of downloading news items is repeated daily over the duration of the chosen period. The collection, therefore, is built of several subcollections corresponding to different days.

The various steps involved in the construction of the collection are described in detail in the next subsections.

4.1 Obtaining the news items

The first step involved building a program to visit the Web site of the particular electronic newspaper and download the corresponding files. News items were collected for a period of 14 days, excluding weekends and holidays. This produced a subcollection for each of the 14 days. The number of news items downloaded each of those days was 95, 75, 87, 71, 76, 76, 76, 85, 82, 86, 81, 73, 72 and 64.

It was possible to access the news items from the section page, which showed the links to all news belonging to that section. In our case, seven sections were considered: National, International, Sports, Economy, Society, Culture and People.

Afterwards, the downloaded files (in HTML format) were parsed in order to extract the relevant information for our collection. In this case, the title, author, section, body and link to the full text were extracted. It is remarkable that the processing described is not trivial, since a good number of these web pages are automatically generated and the resulting HTML is quite chaotic. Fortunately, the tags title, author and body permitted us to easily track down the relevant information. However, this simplification cannot be applied to other newspapers.

(5)

Finally, the news items for a given day were saved in plain-text format, each item in a different file, and arranged in directories corresponding to the different sections. Additionally, the sections were arranged into a shared directory named as the particular date. A news item file is organized as follows: a first line for the title, a second line for the authors, and the remainder for the body. The whole collection of news is organized in a global directory that includes the directories for each of the 14 days.

Using markup languages (SMGL, HTML, XML…) to save these collections could have been appropriate in order to standardize them, but text files are likewise platform-independent and the simplicity of these news contents made the tagged unnecessary.

The collection also organizes the HTML files in a directory hierarchy similar to the one described above, so as to keep available additional information in the HTML format that could be interesting for others personalization systems (i.e. text in black, italic, etc.).

4.2 Obtaining the user models

The second step consists in obtaining the long-term interests as user models, to get a first content personalization. Initial models are built for each user the day or days before the beginning of the news items recollection. This allows information in the models to be available for personalization since the first day.

These preliminary profiles contain information about the user long-term interests, that is, interests that remain constant over time. These long-term interests are defined using three reference systems: sections, categories and keywords. Sections coincide with the seven top sections in the newspaper: National, International, Economy, Society, Culture, Sports and People. Categories correspond with that used in Yahoo! Spain in its first level: Art & Culture, Science, Technology, Social Sciences, Sports & Leisure, Business & Economy, Education, Entertainment, Internet Leisure & Computers, Consultation, News & Media, Politic & Government, Health, Society and Regional. Keywords match with the words provided by the user while defining his interests. (Díaz et al., 2001).

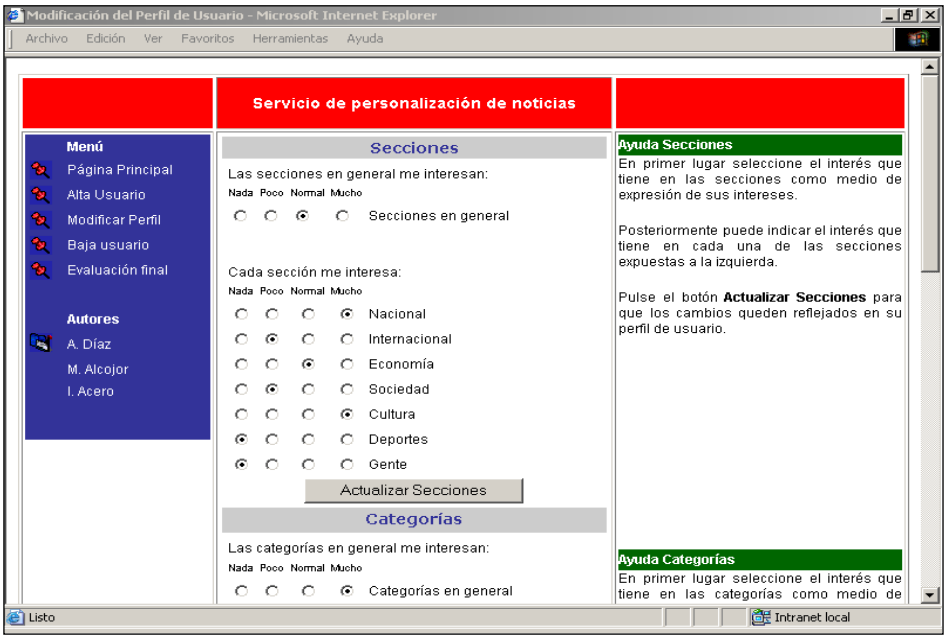

Figure 1: Editing the user model (sections)

Participation in the evaluation was requested sending mails to students and teachers from different faculties. Furthermore, any other additional spreading was permitted. Initially 104 users were registered, but 2 of them registered twice by mistake. Once into operation, 2 new users registered the first day and others 2 the third day, making a total of 106 users. Since the fourth day, users started unregistering, and the number of users suffered significant variations. Since the following day, the number of real participants each day was: 102, 104, 106, 105, 105, 104, 102, 102, 101, 101, 100, 100, 99 and 98.

With respect to the type of users, 77 were students, 22 were university lecturers and 7 were other professionals. Among the lecturers, 17 of them lectured in computing. The largest group of students was studying journalism (36), followed by audiovisual communication (31) and computing (10). In the group of other professional, 2 had relation with computing and the other 5 had no relation either with computing or with journalism. The number of women was exactly the same as that of men, that is, 53.

Interests were introduced manually by the user, assigning a weight from a set of 4 possible values ("Nothing at all", "Not much", "Quite a bit", "Lot"). These values were transformed to the following quantitative weights: 0, 0.33, 0.66 and 1.

These profiles were obtained from a web application where, after registration, the user had to fill in his user model through a series of web pages (Figure 1).

The users chose some sections, some categories and some keywords in their initial profiles, that is, their long-term models (Table 1). These models contained, on average, 5 sections, 10 categories and 3 keywords. The fact that sections and categories were selected marking boxes but the keywords had to be typed may have influenced the low average number of keywords used.

	Sections	Categories	Keywords
average	5.0	9.6	2.8
min	0	0	0
max	7	14	18

Table 1: Number of sections, categories and keywords chosen by the users

It is interesting to observe that 3 users did not select any section, 4 did not choose any category and 21 did not introduce any keyword. Other significant data was that 30 users chose all the sections as relevant and 28 chose all the categories. The maximum number of keywords selected by a user was 18. Finally, there was a user who introduced an empty profile, that is to say, did not introduce any section, category or keyword.

	National	International	Economy	Society	Culture	Sports	People
weight average	0.68	0.58	0.30	0.51	0.68	0.39	0.39
weight <> 0 users	87	86	61	81	94	56	64
weight = 1 users	46	26	9	26	41	24	14

Table 2: Weights assigned to sections by the users

	Art & Culture	Science & Technology	Social Sciences	Sports & Leisure	Business & Economy	Education	Entertainment
weight average	0.59	0.54	0.47	0.54	0.27	0.51	0.72
weight<>0 users	82	79	72	79	56	74	93
weight = 1 users	33	34	24	35	8	31	53
	Internet & Computers	Consultation	News & Media	Politic & Government	Health	Society	Regional
weight average	0.46	0.38	0.69	0.44	0.37	0.41	0.34
weight<>0 users	73	62	85	73	67	66	57
weight = 1 users	24	14	55	18	10	18	9

Table 3: Weights assigned to categories by the users

Table 2 shows the weights assigned to each section. It can be observed that the most important section for the users is Culture, as 94 users chose it, followed by National with 87 users and International with 86 users. The section less often selected was Sports, 56 users, followed by Economy, 61. The section with lower average weight is Economy (0.53), followed by People (0.64).

It can be emphasized that the users varied their weights among the different sections, that is, they did not use the criterion "I'm interested / I'm not interested". This suggests that the system of weight assignments is attractive to the users. The section more often marked with weight 1 was National. It was chosen by 46 users. The less often selected section was Economy, chosen by only 9 users.

Table 3 shows the weights assigned to each category. It can be observed that the most important category for the users is Entertainment, given that 93 users chose it, followed by News & Media with 85. This category has the greatest average (0.72) in the values of the weights assigned by the users. The less selected category was Business & Economy, with 56 users. The lowest average weight category was also Business & Economy (0.51)

As above, it must be emphasized that the users varied their weights among the different categories, which further supports the system of weight assignments as attractive to the users. The category most often selected with weight 1 was News & Media, chosen by 55 users, while the less often selected one was Business & Economy, with only 8.

With respect to the keywords, it is important to say that 80.2% of the users introduced at least one keyword. Users selected weight 1 for all their keywords in 73.6% of the cases. The average weight was 0.98 because, among the 304 keywords typed by the users, only 18 were assigned a weight different to 1 (16 with 0.66 and 2 with 0.33). Of these keywords, 42 were proper nouns and the rest were general terms such as publicity, university, music, sport, television, football, etc.

Proper nouns allow a very specific personalization on topic of interests related with persons and places that rarely stop interesting to the user, while general terms can result in information that could not interest to the user.

User models are finally saved in three different text files, one for each type of interest, that is, sections, categories and keywords. The files for sections and categories present the same format, where each row represents a section or category, a user and the weight assigned to it. All sections and categories are present even if their weight is 0. Keywords file is quite similar. In a fourth file, each row contains the general weights for sections, categories and keywords, assigned by the users.

4.3 Obtaining the relevant judgments

The third step involves obtaining the user relevance judgments with regard to all news items received during the 14 days of the experiment. To that end, every day a mail was sent to the users with the following information associated to each news item: title, author, source, section, personalized summary and link to the complete news item. The users should evaluate each of them as relevant or not.
Relevance judgments are saved in text files containing as many rows as news items, where each row presents a news item with its corresponding relevance judgment. The relevance judgment can either be "relevant" (0) or "not relevant" (1). There exists a relevance judgments file for user and day.

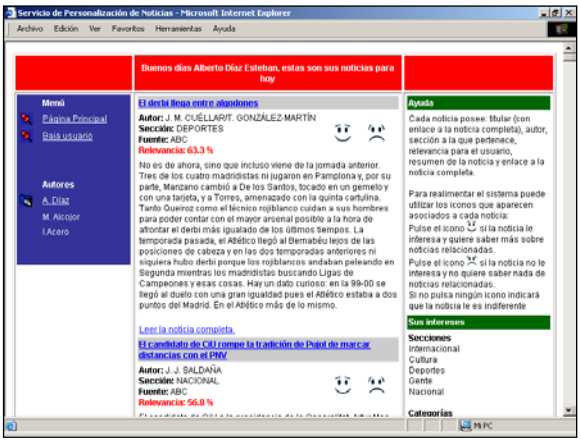

Figure 2: Sample message received by a user.

On the other hand, users can introduce their relevance judgments using two feedback icons (positive feedback/negative feedback) associated to each news item (Figure 2). In fact, the user can also choose not to introduce any relevance judgment (indifferent feedback). Users are advised to base their relevance judgments not only on their long-term profile, but also on their information needs. Actually, the interests not in agreement with the long-term model should be captured by a short-term model. Therefore, this short-term model would keep information about the user's feedback.
The information about the short-term model is not saved into the evaluation collection because it depends on the system used to obtain it from the user relevance judgments. We are interested in obtaining binary relevance judgments, and so we have two possibilities: it can be considered relevant news items only those with a positive feedback, or also those which are indifferent to the user. In this case, we have opted for the first option.
Nonetheless, the analogy established can be questionable. The user can interpret the feedback as a request for more information about a news item and therefore not to feedback the news that are interesting to him if he does not wish to know more about them. This misinterpretation would invalidate the method proposed to get the relevance judgments and would make necessary to have two separated processes: one to get the relevance judgments and another to obtain the feedback. In order to prevent this misapprehension, users are given clear instructions to understand correctly the meaning of the positive and negative feedbacks. Consequently, the procedure proposed for obtaining the relevance judgments and the feedback simultaneously can be used to build the evaluation collection.
The feedback judgments are saved in text files similar to the relevance judgment files. Each file store the feedback associated to each news item for a user for a day. A feedback judgment presents three possible values: 1 for a positive feedback, -1 for a negative feedback or 0 for no feedback at all. These files are stored in the same directory hierarchy together with the relevance judgments.
One must bear in mind that relevance judgments are only generated for those users who introduce them by means of the feedback icons, thus there will not be as many judgments as users, but as many as users had expressed a judgment upon (at least) a news item.
Table 4 shows the number of users that emitted judgments each day. It can be noticed that user involvement were decreasing over time, which is perfectly understandable given the effort needed to judge approximately 100 news items a day. On the other hand, the number of judgments changed considerably between users. There were users who formulated judgments about lots of news items while others formulated just one. It has been considered "valid" user judgments those which affect to 10 news items at least. This means that 9.4 user with judgments, per day, were eliminated. Then, the final collection contains, on average, 28.6 users with judgments per day, which represents a total of 395 different judgments.

day	users	users with judgments	users with less than 10 judgments
1	102	50	12
2	104	53	12
3	106	38	8
4	105	45	10
5	105	44	9
6	104	38	10
7	102	35	9
8	102	37	12
9	101	31	8
10	101	35	9
11	100	33	8
12	100	25	9
13	99	27	10
14	98	28	6
average	102.1	37.1	9.4

Table 4: Number of users that emitted some judgment

Table 5 compiles the statistics about the number of news items marked as relevant by the users with more than 10 judgments per day. The average number of relevant news items per day fluctuates between 33.2 and 47.7, with a global average of 42.1. For the referred reason, the number of judgments was decreasing as days went by.

day	news items	Average	Max.	Min.
1	95	45.7	95	11
2	75	43.3	75	10
3	87	42.9	87	10
4	71	39.2	71	10
5	76	40.3	76	11
6	76	38	75	10
7	76	38	76	10
8	85	44.4	85	12
9	82	47.7	82	13
10	86	45.1	84	11
11	81	41.5	81	12
12	73	45.6	73	10
13	72	45.1	72	11
14	64	33.2	64	10
average	78.5	42.1	78.3	10.8

Table 5: Statistics about the average number of "valid" user judgments per day.

On the other hand, there are 10 users that only give positive feedbacks and 1 user that only gives negative ones. In general, there are more users who formulate more positive than negative feedbacks. Consequently, it can be stated that users give more positive than negative feedbacks on average, even if the number of news items with a negative feedback is slightly higher.

Table 6 presents the average number of news items with feedback, either positive or negative (R+/R-) for users with "valid" judgments and for the 14 days of the experiment. The average of positive feedbacks is situated around 20 news items per day, while the average of negative feedbacks is situated around 23 news items per day. This means that users tended to formulate slightly more negative feedbacks than positive ones.

day	news items	R+ average	R- average
1	95	23.0	23.3
2	75	21.5	22.3
3	87	20.5	23.1
4	71	19.9	19.3
5	76	21.7	19.2
6	76	17.7	21.1
7	76	19.1	19.9
8	85	22.0	23.4
9	82	22.3	26.4
10	86	21.0	25.0
11	81	20.9	21.5
12	73	17.5	28.2
13	72	18.6	29.9
14	64	17.5	17.3
average	78.5	20.2	22.9

Table 6: Statistics about the average number of negative and positive user feedbacks (R+/R-) per day

5. Using the collection to evaluate the personalization system

The results to be obtained are a ranking of documents for each user, obtained from the application of the multi-tier content selection process by means of formula (1), where can be used different combinations of tiers giving different values to the parameters δ, ε, ϕ, and γ.

The comparison between a ranking of documents and binary relevance judgments suggests the use of normalized recall and precision metrics (nR and nP) (Rocchio, 1971). This is justified because rankings of documents rather than groups of documents are compared: one does not simply observe whether the first X documents are relevant or not, but rather their relative order in the ranking.

For the multi-tier selection process (Table 7), the best results are obtained using a combination of a long model based on sections, categories and keywords, together with a short term model (L(SeCaKe)S). The relative order for the rest of combinations of long and short term models is: sections and categories (L(SeKe)S), sections and keywords (L(SeKe)S), categories and keywords (L(CaKe)S), only sections (L(Se)S), only categories (L(Ca)S) and only keywords (L(Ke)S). The worst result appears when only the short term model is used (S) (Díaz & Gervás, 2004).

	L(SeCaKe)S	L(SeCa)S	L(SeKe)S	L(CaKe)S	L(Se)S	L(Ca)S	L(Ke)S	S
nP	0.600	0.583	0.568	0.539	0.535	0.514	0.475	0.421
nR	0.691	0.681	0.669	0.633	0.652	0.614	0.583	0.545

Table 7: Normalized precision (nP) and recall (nR) for different combinations of reference frameworks (Se: sections, Ca: categories, Ke: keywords) in the combination of long (L) and short (S) term model.

	C	Ps	GPs	Fs	Gs
nP	0.600	0.593	0.584	0.581	0.577
nR	0.691	0.686	0.680	0.678	0.675

Table 8: Normalized precision (nP) and recall (nR) for different types of summaries (C: complete news item, Ps: personalized summaries, GPs: generic-personalized summaries, Fs: first sentences summaries, Gs: generic summaries).

Moreover, this collection has been used to evaluate the personalized summarization involved in the result presentation process (Table 8). In this case, the effect of the multi-tier content selection process over the different types of summaries is measured. This involves checking what results are obtained, as compared with user judgments, if instead of selecting news items based on their full text they are selected based on the summaries as input data. The metrics used are again normalized recall and precision. The analysis of the results shown in Table 3 indicates that personalized summaries (Ps) give significantly better results than generic summaries (Gs), generic personalized summaries (GPs) and first sentences summaries (Fs). It can also be seen that personalized summaries are worse than complete news items (N) (Díaz & Gervás, 2007).

6. Conclusions

The evaluation collection contains information about user models and, for several days in a sequence, information about relevance judgments on the sets of documents provided by the users who build the user models for each day. Out of this set of relevance judgments, particular selections can be employed to simulate real relevance feedback judgments as they would have been provided by the users (in a situation of real use of a personalisation system the users are unlikely to consider every day all the documents available for that day, so it would be unrealistic to apply all the relevance judgments available in the collection as relevance feedback for system adaptation; however, they may be used to provide an upper limit on the possible precision of the effect of relevance feedback).

This collection has been successfully used to perform some different experiments to determine the effectiveness of the personalization system described in section 3 (Díaz & Gervás, 2004; Díaz & Gervás, 2007).

On the other hand, there is a recent trend to apply qualitative evaluation based on the opinions of the user, gathered by means of questionnaires. These opinions show the impressions of the users concerning the use of the system and its various aspects. These two approaches to evaluation complement each other, and they visualise the operation of the system from two different points of view: the point of view of the system and the point of view of the user (Díaz et al. 2008).

7. Acknowledgements

This research has been partially funded by the Spanish Ministerio de Ciencia e Innovación (TIN2009-14659-C03-01).

8. References

Billsus, D., Pazzani, M. (2007). Adaptive news access. In P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.), *The Adaptive Web - Methods and Strategies of Web Personalization*, volume 4321 of Lecture Notes in Computer Science. Springer, Berlin / Heidelberg, 2007.

Díaz, A., Gervás, P., García, A., Chacón, I., (2001). Sections, Categories and Keywords as Interest Specification Tools for Personalised News Services. *Online Information Review*, 25(3), pp. 149-159.

Díaz, A. & Gervás, P. (2004). Adaptive User Modeling for Personalization of Web Contents. *Proceedings of the 3th International Conference, AH 2004*, Eindhoven. Lecture Notes in Computer Science, 3137, pp. 65-74.

Díaz, A. & Gervás, P. (2005). Personalisation in news delivery systems: Item summarization and multi-tier item selection using relevance feedback. *Web Intelligence and Agent Systems*, 3(3), pp. 135-154.

Díaz, A. & Gervás, P. (2007). User-model based personalized summarization. *Information Processing & Management*, 43(6), pp. 1715-1734.

Díaz, A., García, A., Gervás, P., (2008). User-centred versus System-centred Evaluation of a Personalization System. *Information Processing and Management* 44(3), pp. 1293-1307.

Mizarro, S. & Tasso, C. (2002). Ephemeral and Persistent Personalization in Adaptive Information Access to Scholarly Publications on the Web. *Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*, Málaga, Spain.

Rocchio, J.J. Jr. (1971). Relevance feedback in information retrieval, The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall.

Salton G. & McGill, M.J. (1983). Introduction to modern information retrieval. McGraw-Hill, New York.