Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Information Processing and Management 44 (2008) 1293–1307
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

User-centred versus system-centred evaluation of a personalization system

Alberto Díaz [a,*], Antonio García [b,1], Pablo Gervás [a,2]

[a] *Dep. Ingeniería del Software e Inteligencia Artificial, Facultad de Informática – Universidad Complutense de Madrid, c/ Profesor José García Santesmases, s/n, Madrid 28040, Spain*

[b] *Dep. Ciencias de la Comunicación I, Facultad de Ciencias de la Comunicación – Universidad Rey Juan Carlos, Camino del Molino, s/n, Fuenlabrada, Madrid 28943, Spain*

Received 1 February 2007; received in revised form 3 August 2007; accepted 3 August 2007
Available online 19 September 2007

Abstract

Some of the most popular measures to evaluate information filtering systems are usually independent of the users because they are based in relevance judgments obtained from experts. On the other hand, the user-centred evaluation allows showing the different impressions that the users have perceived about the system running. This work is focused on discussing the problem of user-centred versus system-centred evaluation of a Web content personalization system where the personalization is based on a user model that stores long term (section, categories and keywords) and short term interests (adapted from user provided feedback). The user-centred evaluation is based on questionnaires filled in by the users before and after using the system and the system-centred evaluation is based on the comparison between ranking of documents, obtained from the application of a multi-tier selection process, and binary relevance judgments collected previously from real users. The user-centred and system-centred evaluations performed with 106 users during 14 working days have provided valuable data concerning the behaviour of the users with respect to issues such as document relevance or the relative importance attributed to different ways of personalization. The results obtained shows general satisfaction on both the personalization processes (selection, adaptation and presentation) and the system as a whole.

Keywords: User-centred; System-centred; Evaluation; Personalization; Web contents; News services

1. Introduction

Information retrieval or information filtering systems are usually evaluated using a set of criteria that are independent of the users for which they are intended. Some of the most popular measures – recall and precision – are based on a set of relevance judgments that are not usually obtained from the final users of the

[*] Corresponding author. Tel.: +34 913 947 606; fax: +34 913 947 529.
E-mail addresses: albertodiaz@fdi.ucm.es (A. Díaz), antonio.garcia@urjc.es (A. García), pgervas@sip.ucm.es (P. Gervás).
[1] Tel.: +34 914 887 259; fax: +34 914 888 220.
[2] Tel.: +34 913 947 641; fax: +34 913 947 529.

doi:10.1016/j.ipm.2007.08.001

system, but rather from one or several experts in the domain. A typical instance of this method is provided by the TREC[3] conferences on information retrieval. This type of evaluation is more oriented to the system than to the user, in fact, the participation of users is not essential.

These expert judgments are an approximation of the reality, since the relevance of the documents associated to a query performed by a user depends on the specific context in which it is carried out. Other aspects of system use should be taken into account, such as the user experience, his preferences, the aim of the search, the use of the received information, etc. In fact, the concept of relevance has been discussed in a great extent in the bibliography with different interpretations (Barry & Schamber, 1998; Greisdorf, 2003; Mizzaro, 1997; Spink, Howard, & Bateman, 1998).

On the other hand, there is a more recent tendency towards user-oriented evaluation, in which the user' opinions about the use of the information retrieval system are collected, in an effort to obtain the impressions of the users about the system from their own point of view. This kind of evaluation may be aimed at collecting both qualitative and quantitative measures (i.e. recall and precision), but in all cases the participation of the users is always essential.

The work presented in this paper focuses on discussing the problem of user-centred versus system-centred evaluation of a Web content personalization system, with a dual goal: to present a case of how multi-faceted evaluations can be conducted, and to show how the combination of system and user centred evaluations can provide more insightful views in the examination of a system than each of them separately.

Web contents appear in many forms over different domains of application, but in most cases the form of presentation is the same for all users, that is, the contents are static in the sense that they are not adapted to each user. Content personalization is a technique that tries to avoid information overload through the adaptation of web contents to each type of user.

Section 2 describes previous relevant work on user-centred and system-centred evaluation in information seeking environments and outlines an existing personalization system that is used as subject of observation in this paper. Section 3 describes the evaluation methodology, experimental setup and results obtained using user-centred and system-centred evaluations in an experiment performed with the personalization system described in Section 2. In Section 4 the combined use of both types of evaluation is discussed. Finally, Section 5 outlines the main conclusions.

2. Previous work

The evaluation method proposed in this paper relies on previous studies about evaluation in information seeking environments and it is exemplified over an existing personalization system that combines a particularly broad selection of the features of such systems.

2.1. Evaluation in information seeking environments

The idea of applying a double evaluation (user-centred and system-centred) follows the approach of Ingwersen and Järveling (2005) in as much as it aims to evaluate an information system – in this case a personalized information system – without leaving aside a global comprehension of the search process and the real information needs of the users potentially involved.

Below are presented the main ideas covered in existing studies about user-centred evaluation and system-centred evaluation.

2.1.1. User-centred studies

Recent years have seen an increase in the number of information retrieval research efforts that explore new ways of addressing the evaluation of information systems. The appearance of a cognitive and a social tendency, together with an increase in the number of user-centred information seeking works (Beaulieu, 2003;

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

Borrego, 1999; Dalrymple, 2001; Fidel, 1993; Kuhlthau, 2005; Martzoukou, 2005; Rieh, 2004) constitute a good example of this.

The social approach, with a smaller projection, focuses on context and its interrelation with the users of information systems (Harris, 1986; Hjørland & Albrechtsen, 1995; Ørom, 2000). On the other hand, a cognitive research school has also appeared (Ellis, 1996; Ford, 2004; Wilson, 1997). An approach in process of construction that tries to promote a human view opposed to a technical and traditional outlook, including different aspects such as a psychological point of view (Ford, 2000; Heinström, 2003) or new measures (Borlund, 2003). The main factors that it studies are information processing, changes in goals in the strategies of users, effective and contextual elements of information seeking, and the influence of individual characteristics or behaviour patterns.

This research school proposes different theoretical models (Beaulieu, 2003; Järveling & Wilson, 2003; Kuhlthau, 2005; Vargas-Quesada, de Moya, & Olvera, 2002): (a) the Ingwersen global model of poly-representation (Ingwersen, 1996), that makes special emphasis on different aspects about requests, information problem and work task levels; (b) the model of Belkin (1990), that points out interactions between the user and the system during each phase; (c) the stratified model of Saracevic (1996), that tries to improve previous proposals, identifying information search processes in order to incorporate them to system design; (d) the interactive feedback model (Spink, 1997), where the key is located in the effects of different types of feedback in information retrieval. Besides, this perspective relies on different research methodologies (Caro-Castro, Cedeira, & Travieso, 2003; Martzoukou, 2005) such as recorded transactions, verbal questionnaires, interviews as well as discussion groups and empirical observation. With reference to the analytical tasks, both quantitative and qualitative methods are usually used.

Some of these user-centred works pay attention to personalization of information access. It is necessary to underline the work made (Spink, 2002) in the context of human interaction with search engines in Web, based on changes and actions that take place during episodes of information seeking. Another work in this direction is carried out by Salampasis and Diamantaras (2002) or by Kelly and Belkin (2002) and Kelly (2003) that presents a naturalistic user study to understand how an individual's online information behaviour can be used as implicit evidence for the construction and maintenance of a personalized user model.

2.1.2. *System-centred studies*

In this type of studies the comparison between systems or algorithms is based on the similarity between the relevance values assigned to each document by the system and the relevance judgments usually pre-assigned by experts (i.e. TREC conferences).

There are many metrics originating from information retrieval that can present evaluation results in the form of a curve, based on two values or on a single value. These metrics can be grouped into several categories depending on the type of relevance and the type of retrieval that are being considered (Mizzaro, 2001; Salton & McGill, 1983). In particular, normalized recall and normalized precision (nR and nP) (Rocchio, 1971) compare binary relevance with retrieval in the form of a ranking. Plotting recall or precision against the levels of the ranking, these metrics measure the effectiveness of the ranking in terms of the area of the graph delimited by the best possible solution, on one hand, and the solution generated by the system under evaluation, on the other. These metrics are calculated using formulas (1) and (2), where REL is the number of relevant documents, RANK_i represents the position in the ranking of the ith relevant document, and N is the total number of documents in the collection.

$$R = 1 - \frac{\sum_{i=1}^{\mathrm{REL}} \mathrm{RANK}_i - \sum_{i=1}^{\mathrm{REL}} i}{\mathrm{REL}(N - \mathrm{REL})} \tag{1}$$

$$P = 1 - \frac{\sum_{i=1}^{\mathrm{REL}} \log \mathrm{RANK}_i - \sum_{i=1}^{\mathrm{REL}} \log i}{\log N!/(N - \mathrm{REL})!\mathrm{REL}!} \tag{2}$$

2.2. *Personalization system for news services*

A personalization system is based on three main functionalities: content selection, user model adaptation and presentation of results (Díaz & Gervás, 2005; Mizarro & Tasso, 2002). 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. User model adaptation (Belkin, 1997; Billsus & Pazzani, 2000; Saracevic, 1996) is built upon the interaction of the user with the system, which provides the feedback information used to evolve the profile. Results presentation involves generating a new result web document that contains, for each selected document, a personalized extract considered indicative of its content.

The evaluation processes will be applied in the frame of reference of a personalization system for a digital newspaper (Díaz & Gervás, 2004, 2005). This domain has peculiar features that affect the personalization processes and their evaluation. Their main characteristic is that every day a separate collection of new documents is made available that have to be personalized and distributed by means of an e-mail received by all users in the early hours of the morning. The choice of the particular system to be subjected to the evaluation was based on the fact that it integrates a broad number of the available technologies for specifying information needs. This allows the consideration over a single system of evaluation procedures that are applicable to a large number of systems, independently of the technologies that are being used for specifying information needs.

2.2.1. User model

The user model proposed consists of the combination of two types of user interests: long term and short term (Díaz & Gervás, 2005). 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 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 an associated weight, 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).

2.2.2. Multi-tier content selection and results presentation

Documents are downloaded from the web of a daily Spanish newspaper as HTML documents. For each document, title, section, URL and text are extracted, and a term weight vector representation for a document d (d_{d}) is obtained by 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 associated with the corresponding 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 ($s^{\mathrm{s}}_{\mathrm{du}}$). The similarities between a document d and a category c (s_{dc}), between a document d and the keywords of a user model u ($s^{\mathrm{k}}_{\mathrm{du}}$), and between a document d and the feedback terms of a short-term user model u ($s^{\mathrm{f}}_{\mathrm{du}}$) are computed using the cosine formula for similarity within the vector space model (Salton & McGill, 1983):

$$s_{\mathrm{dc}} = \mathrm{sim}(d, c) \quad s^{\mathrm{k}}_{\mathrm{du}} = \mathrm{sim}(d, k_{\mathrm{u}}) \quad s^{\mathrm{t}}_{\mathrm{du}} = \mathrm{sim}(d, t_{\mathrm{u}}) \tag{3}$$

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

$$s^{\mathrm{c}}_{\mathrm{du}} = \sum_{i=1}^{14} C_{i\mathrm{u}} s_{\mathrm{dc}_i} \Bigg/ \sum_{i=1}^{14} C_{i\mathrm{u}} \tag{4}$$

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:

$$s_{\mathrm{du}} = \frac{\delta s^{\mathrm{s}}_{\mathrm{du}} + \varepsilon s^{c}_{\mathrm{du}} + \phi s^{\mathrm{k}}_{\mathrm{du}} + \gamma s^{\mathrm{t}}_{\mathrm{du}}}{\delta + \varepsilon + \phi + \gamma} \tag{5}$$

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 results presentation process is: 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 documents ordered by relevance and for each document: title, author, section, source, relevance, feedback icons and automatically generated summary adapted to the user.

Three phrase-selection methods are used to build summaries. The two first methods are generic: position and thematic words. The third one is based on the personalization of the summary using the information from the user model (Díaz & Gervás, 2007).

3. User-centred and system-centred evaluation of a personalization system

This section presents the evaluation methodology used in the user-centred and system-centred evaluations, the experimental setup, and the results obtained with both types of evaluations.

3.1. Methodology

The methodology presented here has been designed to cover a large number of the approaches discussed in Section 2.1. The particular choice of sample system to evaluate has been chosen to ensure a broad coverage of these various approaches.

3.1.1. User-centred evaluation

This evaluation is based on the use of evaluation questionnaires that try to obtain explicit user opinions about different functionalities of the system. In particular, we propose the use of two evaluation questionnaires, one to be filled in before using the system and the other to be filled in after using the system. The need for the two questionnaires arises from our interest in measuring the degree to which the system has fulfilled the initial expectations of each user.

The evaluation questionnaires contain questions that are grouped under the following headings: interface, user model, summaries, selection and adaptation, measurement of news relevance, global estimation, open questions and comments (see Appendix for more details). The questionnaires are composed mostly of closed questions, where answers are indicated by means of a rank of five levels (very high, high, regular, low, very low) or by means of the duality Yes/No. Additionally, there is also the possibility of answering open questions regarding system performance.

An initial group of questions (questions 1–6) asks about the degree of satisfaction with more important graphical components of the system, the degree in which the system is attractive for users, the facility to use the system and its friendliness, as well as questions on contents management and help facilities. These questions appear both in the initial and in the final questionnaires.

A second group of questions considers the different parts of the user model, that is, sections, categories and keywords. Regarding sections and categories (questions 7–10), users are asked about their suitability in order to reflect user information needs. In this sense, users are also required to answer whether they would introduce new sections or categories to reflect their information needs. On the other hand, in the final evaluation users are asked to what extent documents selected because they belong to sections or categories preferred by the user appear before documents that are not relevant. Finally, questions 13, 14, 16 and 17 enquire about possible changes in the use of sections or categories as personalization methods, as well as identifying the moments in time in which those changes took place. Regarding keywords, the questions try to discover if this option is relevant for specifying information needs (question 11). In the final evaluation (questions 18–24) users are asked to what extent the system is capable of showing news corresponding to selected keywords before news that do not contain selected keywords, and the extent to which documents retrieved according to introduced keywords correspond to information needs. Users are also asked about the clarity that the system offers at time of retrieving documents based on keywords, or about possible changes to their selection of keywords resulting from system use.

There is another group of questions (25–28) in the final evaluation that addresses the information selection and adaptation of the system through time. The following aspects are analysed: the validity of news rankings from the selected profile, the system adaptation with respect to user information needs and judgments, and the changes produced in user information needs.

Another set of questions (29–36) addresses opinions about summaries as a way to present the results to the users. Users are asked to what extent the summaries are well constructed, and about their coherence and clarity. In the same way, questions related to possible redundancies in the summaries or informative elements missing from the summaries are included. These questions appear in the final evaluation when the users have received the messages with the summaries.

The next group of questions serves to determine the criteria by which a user understands if an item is relevant or not. Thus, before using the system (question 12), each user is asked to indicate criteria that will be applied at the time of making this decision. After using the system (question 37), users reply to questions about the criteria they actually used. Similarly, before using the system users are also asked (question 13) about their degree of interest in the information that they are going to receive. On the other hand, in the final evaluation (question 39) users were asked about their real interest in the information received, and about (question 38) which particular elements they have consulted to decide if news were relevant or not: headline, section, relevance, summary or complete item.

The last issue addressed in the final questionnaire concerns a global evaluation of the system in terms of both the level of satisfaction and the confidence that users reach after working with it. This interest translates into questions (40–42) about the extent to which their information needs are solved and the degree of personalization of the system. The users are also asked about the way that they prefer to define their profile (question 43).

Finally, both questionnaires finish with a set of open questions and comments.

3.1.2. System-centred evaluation

The results to be obtained are a ranking of documents for each user, obtained from the application of the selection process by means of formula (5), where different combinations of tiers can be used giving different values to the parameters δ, ε, ϕ, and γ.

These results must be compared with binary relevance judgments collected previously. Evaluation collections for personalization 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 for every particular day. This is because the tasks to be carried out in each case 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) and these information needs 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 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 carrying out different tasks at different times. In our case the relevant judgments between each document and each user model are assigned by the proper users during the normal running of the system for several days.

The comparison between a ranking of documents and binary relevance judgments suggests the use of normalized to recall and precision metrics – formulas (1) and (2). 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 evaluating summarization, the effect of selection (formula (5)) 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. Then, the metrics used are again normalized recall and precision (Díaz & Gervás, 2007).

3.2. Evaluation setup

As an example of web documents for experimentation we have chosen the web pages of the digital edition of a Spanish newspaper. Experiments are evaluated over data collected for 106 users and the news items corresponding to three weeks – 14 working days – of the digital edition of the ABC Spanish newspaper. The average number of news items per day is 78.5, obtained from the main seven sections of the newspaper.

Out of 106 users, 90 filled in the initial evaluation form, whereas only 38 completed the final evaluation. These different numbers are justified because the final questionnaire was provided after the last day and many users left the system before. The main reason why the users abandoned the experiment was the large quantity of news items that they had to judge everyday (78.5 news item per day on average).

In the end, 35 users filled both in the initial and the final evaluations. The user-centred evaluation has been applied to this set of users in order to obtain more significant results. With respect to these users, 60.0% were students, 31.4% were university lecturers (mainly in computing) and 8.6% were other professionals. The largest group of students was studying journalism (28.6% of the total number), followed by audiovisual communication (20.0%) and computing (11.4%). In conclusion, the group is big and heterogeneous enough to allow the extraction of significant conclusions on its behaviour.

On the other hand, the system-centred evaluation has been applied to those users that have provided a significant number of judgments along the 14 working days. The final collection employed for system-centred evaluation has 395 different sets of relevance judgments.

3.3. Results

The evaluation results are presented separately for user-centred and system-centred evaluations in the next sections. As the number of users considered is different in each type of evaluation, the number of respondents associated with each percentage is indicated in each table.

3.3.1. User-centred evaluation results

With respect to the interface evaluation (Table 1), it can be concluded that the initial impressions of the users about the interface were positive enough. In all cases similar trends were repeated in both evaluations, with a small increase in the scores for users' opinions about the graphics components, usability and friendliness. However, a small decrease was also apparent in the scores for content management and system help facilities.

With respect to the user model (Table 2), in the initial evaluation the users considered that keywords reflect better their information needs followed by sections and categories. Moreover, these high results confirm the three tiers of selection of the long term model as adequate for reflecting the users' information needs. On the other hand, in the final evaluation, the users preferred sections, followed by keywords and categories, but with a small difference between all of them. However, a decrease is observed for all the methods.

On the other hand, it is important to note that a great decrease is observed for keywords. This is due mostly to the fact that keywords chosen by the user are in general few and very specific, making it very difficult for the system to find matches for them in the news items. Nonetheless, whenever they do appear in some news item,

Table 1
Percentage of users and average for each reply to questions about the interface evaluation, in the initial and final evaluations ($N = 35$)

Question	Initial evaluation						Final evaluation					
	Very high	High	Medium	Low	Very low	Average	Very high	High	Medium	Low	Very low	Average
(1) Graphic components	5.7	45.7	45.7	2.9	0.0	3.5	2.9	54.3	42.9	0.0	0.0	3.6
(2) Attractiveness	2.9	34.3	60.0	2.9	0.0	3.4	5.7	37.1	48.6	8.6	0.0	3.4
(3) Usability	20.0	71.4	8.6	0.0	0.0	4.1	28.6	62.9	8.6	0.0	0.0	4.2
(4) Friendliness	8.8	44.1	41.2	5.9	0.0	3.6	2.9	62.9	31.4	2.9	0.0	3.7
(5) Content management	11.8	52.9	35.3	0.0	0.0	3.8	2.9	68.6	22.9	5.7	0.0	3.7
(6) Help facilities	12.1	63.6	21.2	3.0	0.0	3.8	8.6	51.4	40.0	0.0	0.0	3.7

Table 2
Percentage of users and average for each reply to questions about usefulness for reflecting information needs of sections, categories and keywords, in the initial and final evaluations ($N = 35$)

Question	Initial evaluation						Final evaluation					
	Very high	High	Medium	Low	Very low	Average	Very high	High	Medium	Low	Very low	Average
(7) Sections	14.3	62.9	20.0	2.9	0.0	3.9	2.9	57.1	34.3	5.7	0.0	3.6
(9) Categories	0.0	71.4	22.9	2.9	2.9	3.6	0.0	67.6	20.6	8.8	2.9	3.5
(11) Keywords	48.6	42.9	8.6	0.0	0.0	4.4	25.0	28.1	28.1	9.4	9.4	3.5

they lead to very high user satisfaction when that item is selected. A possible additional problem associated with keywords is their potential for polysemy. This does not occur frequently due to the high specificity of the words chosen by the users.

The validity of the set of sections and categories is also studied by asking the users whether they would introduce new sections or categories (Table 3). For both sections and categories, users showed less need for introducing new elements after using the system than they had shown before using the system.

Users were also asked in the final evaluation whether the system selected documents corresponding to sections, categories and keywords proposed by the users rather than documents not related to them (Table 4). Most users consider that sections allow a more adequate selection of documents. Slightly lower results are given to categories, but still with mostly positive results. In contrast, users are less convinced about keywords. Again, this can be related either to the high specificity of keywords or to possible problems of polysemy.

In the final evaluation some additional questions concerning keywords were included (Table 5). It is clear that the use of keywords on their own for defining a user profile would have led to much less satisfaction in the use of the system, even though for a great quantity of the users the performance of keywords was satisfactory. Approximately the majority of the users recognize the usefulness of some utility to show words related to their input so as to be able to refine their user profile.

Table 3
Percentage of users and average for each reply to questions about to add new sections or categories, in the initial and final evaluations ($N = 35$)

Question	Initial evaluation					Final evaluation				
	Many	Some	Few	None	Average	Many	Some	Few	None	Average
(8) Sections	0.0	35.3	55.9	8.8	2.3	2.9	11.4	34.3	51.4	1.7
(10) Categories	2.9	34.3	48.6	14.3	2.3	3.2	6.5	32.3	58.1	1.5

Table 4
Percentage of users and average for each reply to questions about selection of documents corresponding to each tier of selection, in the final evaluation ($N = 35$)

Question	Very high	High	Medium	Low	Very low	Average
(12) Sections	5.9	52.9	32.4	8.8	0.0	3.6
(15) Categories	6.5	45.2	35.5	9.7	3.2	3.4
(18) Keywords	12.5	31.3	25.0	25.0	6.3	3.2

Table 5
Percentage of users and average for each reply to another questions about the keywords, in the final evaluation ($N = 35$)

Question	Very high	High	Medium	Low	Very low	Average
(19) Correspondence with information needs	16.1	22.6	41.9	12.9	6.5	3.3
(20) Clarity of retrieved documents	9.7	32.3	45.2	9.7	3.2	3.4
(21) Correspondence with proposed words	3.2	41.9	38.7	9.7	6.5	3.3
(22) Usefulness of related words	14.7	32.4	20.6	32.4	0.0	3.3

Finally, users were asked in the final evaluation for their preferences with respect to a particular method for specifying their information needs (Table 6). It is interesting to note that, in spite of the problems detected in the selection using keywords, users feel that they are still the best way for defining their information needs. This is probably due to the fact that it provides a very specific tier of selection as opposed to sections or categories, where the definition of interests is very broad.

With respect to the users' estimations on the efficiency of the selection and adaptation processes, as given in the final evaluation (Table 7), users consider that the system adapts better to their relevance judgments than to their information needs, even though these adaptations are satisfactory for less than the half of the users. On the other hand, most users do not change their information needs throughout their use of the system. Finally, users judge that they receive news items they are interested in more than news items that they are not interested in. This suggests that the selection and adaptation processes are in most cases judged positively.

With respect to the summaries (Table 8), most users consider that the summaries are of high quality, coherent, and clear, and that they reflect the content and the main ingredients of the corresponding document. Most of them also consider, though to a lesser degree, that the summaries contain no redundancies and that they are well adapted to user profile and user needs. This positive evaluation indicates that the method of sentence selection for the construction of summaries is a valid approach for results presentation in the face of possible problems of clarity, coherence and redundancy.

The criteria employed to decide about the relevance of a document seem to match more or less from the initial to the final evaluation. The criteria selected more often were: relation with information needs and subject of interest, relation with user profile, usefulness and novelty. The criteria selected less often were: style, proximity and degree of motivation from an emotional point of view, proximity and familiarity with the content, and, finally, proximity and familiarity with the language employed. With respect to the part of the news item used to establish the relevance (Table 9), it can be concluded that the summary becomes an important element for defining the relevance of a news item, justifying the proposed schema for presentation of results.

Finally, most users value the system positively. The great majority is satisfied and confident, and thinks that the system solves his information needs directed at a daily online newspaper. On the other hand, they consider the system offers an appropriate way of personalization (Table 10).

Table 6
Percentage of users that prefer each method in the final evaluation ($N = 35$)

Question	Sections	Categories	Keywords	Others
(43) Preference	20.6	29.4	47.1	2.9

Table 7
Percentage of users and average about the questions about the selection and adaptation processes, in the final evaluation ($N = 35$)

Question	Very high	High	Medium	Low	Very low	Average
(25) News reflect user needs	8.8	47.1	38.2	5.9	0.0	3.6
(26) Adaptation to information needs	8.8	29.4	55.9	5.9	0.0	3.4
(27) Adaptation to judgments	5.9	41.2	38.2	14.7	0.0	3.4
(28) Information needs changes	3.1	6.3	25.0	31.3	34.4	2.1

Table 8
Percentages of users and average about the questions about the summaries, in the final evaluation ($N = 35$)

Question	Very high	High	Medium	Low	Very low	Average
(29) Quality	12.1	72.7	9.1	6.1	0.0	3.9
(30) Coherence and clarity	14.7	67.6	11.8	2.9	2.9	3.9
(31) Avoid redundancies	9.1	60.6	27.3	3.0	0.0	3.8
(32) Adaptation to user model	5.9	61.8	23.5	5.9	2.9	3.6
(33) Adaptation to user needs	8.8	61.8	17.6	5.9	5.9	3.6
(34) Document reflection	17.6	64.7	11.8	2.9	2.9	3.9

Table 9
Percentages of users and average about the information related with the news item used to decide the relevance, in the final evaluation ($N = 35$)

Question 38	Often	Some	Few	None	Average
Title	88.2	11.8	0.0	0.0	3.9
Section	29.4	50.0	8.8	11.8	3.0
Relevance	20.6	35.3	23.5	20.6	2.6
Summary	47.1	29.4	23.5	0.0	3.2
Full news item	0.0	29.4	52.9	17.6	2.1

Table 10
Percentage of users and average for each reply to questions about the overall judgments about the system, in the final evaluation ($N = 35$)

Question	Very high	High	Medium	Low	Very low	Average
(40) Satisfaction and confidence	14.7	52.9	29.4	2.9	0.0	3.8
(41) Information needs addressed	0.0	56.7	30.0	10.0	3.3	3.4
(42) Personalization level	9.1	45.5	45.5	0.0	0.0	3.6

3.3.2. System-centred evaluation results

To carry out the evaluation, judgments from the user are required as to which news items are relevant or not for each of the days of the experiment. Because the evaluation is based on these judgments, significant results can only be obtained for those users that have provided judgments over and above a minimum threshold in terms of number of judgments per day. As the evaluation process involved an effort for the users, only 37.4 users per day actually provided judgments. Additionally, some users only perform judgments for less than 10 news item per day. These users have been eliminated for the evaluation in order to obtain more significant results. The final collection employed for evaluation presented, on average, 28.6 users per day, which represents 395 different judgments.

For the multi-tier selection process (Table 11), 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).

For the evaluation of the summaries (Table 12), it can be concluded that personalized summaries (Ps) that use a combination of long and short term models are better than other types of summaries in terms of normalized precision and recall. Complete news item (C) offer only a slight improvement against personalized

Table 11
Normalized precision (nP) and recall (nR) for the different combinations of reference frameworks (Se: sections, Ca: categories, Ke: keywords) in the combination of long (L) and short (S) term model

($N = 395$)	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 12
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)

($N = 395$)	C	Ps	GPs	Fs	Gs
nP	0.603	0.593	0.584	0.581	0.577
nR	0.694	0.686	0.680	0.678	0.675

summaries, which seems to indicate that the loss of information for the user is very small with this type of summary.

More specific details of these evaluation results are given in Díaz and Gervás (2004, 2007).

4. Comparison and discussion

The opinions about each method do suffer significant changes between the initial and final evaluations. Initially, the users prefer keywords followed by sections and categories. At the end, the users prefer sections, categories and keywords, but with similar grades. When the users are asked for the degree in which the system shows documents related with the selected items for each tier of selection before documents that are not related, they value better the sections, followed by categories, and a poor last, keywords. This matches with the results obtained in the system-centred evaluation where the frameworks that offer better results in normalized precision and recall are the sections, followed by categories and keywords. This evaluation already showed the preponderance of category-based methods over keywords. This affects specially the combination of long and short term models, which improves significantly in all cases except where keywords are used (Díaz & Gervás, 2004).

As can be observed, the impressions about the keywords have a particular behaviour that cannot be detected in a system-centred evaluation. The users decrease their view about the suitability of the keywords to reflect their interests. However, they continue preferring this tier of selection. This fact can reflect the preference to determine by themselves their interests in the more specific way.

With respect to the measurement of news relevance, users show a special preference for criteria reflecting interest, that is, that the news item was relevant to the user, followed by novelty criteria. On the other hand, the style, the depth and the proximity have been the criteria less applied by the users. The related criterion of whether the news item added new knowledge showed an increase of 10% after system use, probably due to the fact that the collection included several streams of documents about the same subject. Stated preference for depth and quantity of information as selection criteria decreased by 32%, possibly because users faced with periodic collections of documents shift their preference from more verbose towards more concise alternatives of presentation.

With respect to system generated summaries, user-centred evaluation underlines the idea that offering users a summary of each news item is a good solution that reduces information overload. Additionally, the fact that users claim to have used mainly the summaries to determine if a given news item was relevant for them allow us to conclude that user adapted summaries are a useful tool to assist users in a personalization system. This aspect cannot be detected with the system-centred evaluation.

The evaluation method proposed here is designed to cover a range of means of specifying information needs that captures most general trends in current information systems. This makes it possible to apply for evaluating personalization systems irrespective of whether they employ domain-specific sections, keywords, categories, or relevance feedback (Díaz & Gervás, 2000). Where systems employ only some of these methods, it is enough to omit the parts of the evaluation method that concern those that do not feature. On the other hand, for the type of system used in this paper, based on multi-tier selection, the system oriented evaluation of each method carried out independently may be overridden by significantly contrasting results in the user-based evaluation. For instance, a large proportion of replies stating preference for one method of specifying information needs over another may lead to a decision to select one in favour of the other irrespective of their measured efficiency.

The introduction of objective metrics in the field of information retrieval had a very positive effect in the development of efficient algorithms and the identification of the most successful techniques. However, it also brought about a secondary effect of side-lining issues of usability or user satisfaction with the methods being introduced. System-centred evaluations tend to focus on efficiency issues on abstract terms, as related to a single system run over an evaluation collection with set standard results. The range of variation between the results obtained for the metrics is not necessarily large, and very small improvements are considered significant based only on statistical information. This may be a good alternative for guiding the development of algorithms, but it is a poor approach for the guiding the development of systems designed for human users. The motor car industry has long since accepted that the refinements and tuning possibilities that can make a Formula (1) prototype maximally efficient at the racing track need not be the kind of feature that a user wants in the car he drives to work everyday, irrespective of the objective data presented in the dashboard or provided by the stopwatch. Yet the evaluation methods generally employed for information systems still tend to concentrate mainly in the objective numerical metrics rather than on collecting the impressions of'users in the street'.

5. Conclusions

The results obtained in the evaluation of a Web content personalization system, with 106 users during 14 working days, show general satisfaction on various aspects of the system and they indicate that user perceptions agree with the objective system oriented evaluation. Additionally, user centred evaluation provides valuable data concerning the behaviour of the users with respect to issues such as document relevance or the relative importance attributed to different ways of personalization. In particular, the analysis of the different methods of selection has shown that sections, categories and keywords are useful as means of specifying information needs for different goals. Sections are more appropriate when users want to identify a general interest on documents that belong to a prefixed section in newspapers. Keywords are more useful when a user wants to define a more specific interest. Categories provide users with a more intuitive and general procedure beyond the rigid selection connected to sections. For an application of personalization, to use a combination of tiers of selection allows users to define information needs from different points of view.

With respect to the relative merits of the two evaluation methods that have been applied, it seems clear that the combination of both methods provides more information about the real performance of a system than either one on its own. If only system-centred evaluation is applied, the information obtained concerns only what the best configuration of the evaluated system is in terms of efficiency, but no impressions are gathered about user opinion. On the other hand, if only a user-centred evaluation is used, impressions and efficiency measures for specific runs of the system are collected, but such data may not be extrapolated to other configurations or sets of techniques.

The work undertaken in this paper, in what concerns user-centred evaluation, has taken into account the corresponding theoretical and conceptual models, though they have not been applied formally. Additionally, the evaluation employed combines the different perspectives followed in user-centred studies. It considers the most relevant aspects in terms of users information needs, degree of satisfaction, behaviour, efficiency, utility, friendliness and relevance. In this way, it goes beyond the evaluation of a specific system architecture. For the type of system under evaluation, a personalized system, existing literature shows a shortage of studies that combine user-centred and system-centred evaluation. In subsequent studies, research might focus on proposing and validating new metrics, on identifying and modelling user motivation and expectations, or on relating information tasks with the different types of user information needs.

Acknowledgements

This research has been partially funded by the Spanish Ministerio de Ciencia y Tecnología (TIN2006-14433-C02-01) and by the Universidad Complutense de Madrid (AE5/06-15073).

Appendix. A sample questionnaire of the kind described in Section 3.1.1 is presented below. It corresponds to the final questionnaire, which also includes all the questions given in the initial questionnaire (question numbers in brackets).

Final Questionnaire

(1) Interface evaluation

(1) 1. Degree of satisfaction with the graphical components
(2) 2. In what degree is the system attractive?
(3) 3. To what extent is the system easy to use?
(4) 4. In what degree is the system friendly for the user?

(5) 5. In what degree is the system of links good? (colours, length, title, etc.)
(6) 6. Evaluate the assistance that the system provides to the user.

(2) User model

(7) 7. To what extent has the set of sections provided by the system been useful for reflecting your information needs?
(8) 8. Would you introduce new sections to reflect your information needs? How many?
(9) 9. To what extent has the set of categories provided by the system been useful for reflecting your information needs?
(10) 10. Would you introduce new categories to reflect your information needs? How many?
(11) 11. To what extent has the possibility of introducing keywords provided by the system been useful for reflecting your information needs?
12. To what degree does the system show documents corresponding to selected sections before documents that do not belong to those sections?
13. Did you change your choice of sections during the use of the system?
14. In case of changes, could you indicate the days that they were made?
15. To what degree does the system show documents corresponding to selected categories before documents that do not belong to those categories?
16. Did you change your choice of categories during the use of the system?
17. In case of changes, could you indicate the days that they were made?
18. To what degree does the system show documents corresponding to selected keywords before documents that do not contain them?
19. In what degree do you think that retrieved documents, according to selected keywords, correspond to your information needs?
20. How clear are the reasons for retrieving documents related to the proposed keywords?
21. To what extent do retrieved documents correspond to the specificity level of proposed keywords?
22. Would some instrument to show other related words to improve news selection, such as a dictionary, be useful?
23. Did you change your choice of keywords during the use of system?
24. In case of changes, could you indicate the days that they were made?

(3) Selection and adaptation

25. To what extent does the system show news corresponding to your information needs before other news?
26. To what extent does the system adapt to your information needs during the use of the system?
27. To what extent does the system adapt to your judgement about news?
28. To what extent did your information needs change during the use of the system?

(4) Summaries

29. In what degree summaries have quality?
30. To what extent are summaries well constructed, coherent and clear?
31. In what degree is the system able to avoid redundancies in summaries?
32. In what degree do summaries adapt to your user profile?
33. In what degree do summaries adapt to your information needs?
34. In what degree do you think that summaries reflect the contents of the documents?
35. Are the main components of news items represented in the summary?
36. In case of a negative answer, could you indicate which components were not represented?

(5) Measurement of news relevance

(12) 37. Indicate criteria applied at the time of deciding if a piece of news is relevant or not: perspective and exposition; depth and amount of information, the style; novelty; utility; relationship with the user profile; relationship with information needs and subjects of interest; capacity to add new knowledge in front of to other related documents; proximity and degree of motivation from an emotional point of view; proximity from a geographic point of view; proximity and familiarity in the exposed content; proximity and familiarity with the used language.

38. How many times did you use each one of these elements related to each news item to decide on their relevance: headline; section; relevance; summary; complete item?

(13) 39. How do you describe your level of interest in the received information, after using the system?

(6) Global estimation

40. Which is your general degree of satisfaction and confidence about the system, after using it?

41. In what degree did the system solve your information needs?

42. To what extent does the system personalize information in an appropriate way?

43. What tier of selection do you prefer to define your interests?

(7) Open questions

(14) 44. Which are the more important characteristics of the system?

(15) 45. What elements do you miss in the system?

(16) 46. Could you describe your information needs after using the system?

47. Do you think that the system allows interactivity with the user?

(17) 48. What kind of information has more interest for you, not only from a subject point of view, but from the type of document (article, chronicle, editorial, etc.), after using the system?

49. If you changed your user profile, what were the reasons for doing it?

(8) Comments

References

Barry, C. L., & Schamber, L. (1998). User's criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management, 34*(2/3), 219–236.

Beaulieu, M. (2003). Approaches to user-based studies in information seeking and retrieval: A Sheffield perspective. *Journal of Information Science, 29*(4), 239–248.

Belkin, N. J. (1990). The cognitive viewpoint in information science. *Journal of Information Science, 16*(1), 11–16.

Belkin, N. J. (1997). User modeling in information retrieval. In *Tutorial on sixth international conference on user modeling*, UM97, Chia Laguna, Sardinia, Italy.

Billsus, D., & Pazzani, M. J. (2000). User modeling for adaptive news access. *User Modeling and User-Adapted Interaction Journal, 10*(2–3), 147–180.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3). Available at:http://informationr.net/ir/8-3/paper152.html.

Borrego, Á. (1999). La investigación cualitativa y sus aplicaciones en Biblioteconomía y Documentación. *Revista Española de Documentación Científica, 22*(2), 139–156.

Caro-Castro, C., Cedeira, L., & Travieso, C. (2003). La investigación sobre recuperación de información desde la perspectiva centrada en el usuario: métodos y variables. *Revista Española de Documentación Científica, 26*(1), 40–55.

Dalrymple, P. W. (2001). A quarter century of user-centered study: The impact of Zweizig and Dervin on LIS research. *Library & Information Science Research, 23*(2), 155–156.

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

Díaz, A., & Gervás, P. (2004). Adaptive user modeling for personalization of web contents. Adaptive hypermedia and adaptive web-based systems. *Proceedings of AH 2004* (pp. 65–75), LNCS 3137.

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), 135–154.

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

Ellis, D. (1996). Progress & problems in information retrieval. London: Library Association.

Fidel, R. (1993). Qualitative methods in information retrieval research. *Library & Information Science Research, 15*(3), 219–247.

Ford, N. (2000). Cognitive styles and virtual environments. *Journal of the American Society for Information Science, 51*(6), 543–557.

Ford, N. (2004). Modeling cognitive processes in information seeking: From Poper to Heinström. *Journal of the American Society for Information Science and Technology, 55*(9), 769–782.

Greisdorf, H. (2003). Relevance thresholds: A multi-stage predictive model of how users evaluate information. *Information Processing and Management, 39*(3), 403–423.

Harris, M. (1986). The dialectic of defeat: Antinomies in research in library and information science. *Library Trends, 34*(3), 515–531.

Heinström, J. (2003). Five personality dimensions and the influence on information behaviour. *Information Research, 9*(1). Available at: http://informationr.net/ir/9-1/paper165.html.

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain analysis. *Journal of the American Society for Information Science, 46*(6), 400–425.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3–50.

Ingwersen, P., & Järveling, K. (2005). The turn: Integration of information seeking and retrieval in context. Springer.

Järveling, K., & Wilson, T. D. (2003). On conceptual models for information seeking and retrieval research. *Information Research, 9*(1). Available at: http://informationr.net/ir/9-1/paper163.html.

Kelly, D. & Belkin, N. J. (2002). A user modeling system for personalized interaction and tailored retrieval in interactive IR. In *Proceedings of ASIST'02: Annual conference of the american society for information science and technology* (pp. 316–325).

Kelly, D. (2003). A longitudinal, naturalistic study of information search and use behavior as implicit feedback for user model construction and maintenance. User modeling. In *Proceedings of UM 2003* (pp. 420–422), LNCS 2702.

Kuhlthau, C. C. (2005). Towards collaboration between information seeking and information retrieval. *Information Research, 19*(2). Available at: http://informationr.net/ir/10-2/paper225.html.

Martzoukou, K. (2005). A review of Web information seeking research: Considerations of method and foci of interest. *Information Research, 10*(2). Available at: http://informationr.net/ir/10-2/paper215.html.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48*(9), 810–832.

Mizzaro, S. (2001). A new measure of retrieval effectiveness (or: What's Wrong With Precision And Recall). In *International workshop on information retrieval (IR'2001)* (pp. 43–52), Infotech Oulu.

Mizarro, S., & Tasso, C. (2002). Ephemeral and persistent personalization in adaptive information access to scholarly publications on the Web. Adaptive hypermedia and adaptive web based systems. In *Proceedings of AH 2002* (pp. 302–316), LNCS 2347.

Ørom, A. (2000). Information Science, historical changes and social aspects: A nordic outlook. *Journal of Documentation, 56*(1), 12–26.

Rieh, S. Y. (2004). On the Web at home: Information seeking and Web searching in the home environment. *Journal of the American Society for Information Science and Technology, 55*(8), 743–753.

Rocchio, J. J. Jr. (1971). *Relevance feedback in information retrieval. The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall.

Salampasis, M., & Diamantaras, K. I. (2002). Experimental user-centered evaluation of an open hypermedia system of an open hypermedia system and Web information seeking environments. *Journal of Digital Information, 2*(4). Available at: http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Salampasis.

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

Saracevic, T. (1996). Interactive models in information retrieval (IR): A review and proposal. In *Proceedings of the 59th annual meeting of the American Society for Information Science* (pp. 3–9).

Spink, A. (1997). Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science, 48*(5), 382–394.

Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: An exploratory study. *Information Processing & Management, 38*(3), 401–426.

Spink, A., Howard, G., & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. *Information Processing & Management, 34*(5), 599–621.

Vargas-Quesada, B., de Moya, F., & Olvera, M. D. (2002). Enfoques entorno al modelo cognitivo para la recuperación de Información: análisis crítico. *Ciencia da Informaçao, 31*(2), 107–119.

Wilson, T. D. (1997). Information behaviour: An interdisciplinary perspective. *Information Processing & Management, 33*(4), 552–572.