Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 43 (2007) 1715–1734

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

User-model based personalized summarization

Alberto Díaz [a,*], Pablo Gervás [b,1]

[a] ITIS CES Felipe II, Universidad Complutense de Madrid, C/Capitán 39, Aranjuez, Madrid 28300, Spain
[b] Dep. Ingeniería del Software e Inteligencia Artificial, Facultad de Informática – Universidad Complutense de Madrid, C/Profesor José García Santesmases, s/n, Madrid 28040, Spain

Received 21 June 2006; received in revised form 11 January 2007; accepted 13 January 2007
Available online 9 March 2007

Abstract

The potential of summary personalization is high, because a summary that would be useless to decide the relevance of a document if summarized in a generic manner, may be useful if the right sentences are selected that match the user interest. In this paper we defend the use of a personalized summarization facility to maximize the density of relevance of selections sent by a personalized information system to a given user. The personalization is applied to the digital newspaper domain and it used a user-model that stores long and short term interests using four reference systems: sections, categories, keywords and feedback terms. On the other side, it is crucial to measure how much information is lost during the summarization process, and how this information loss may affect the ability of the user to judge the relevance of a given document. The results obtained in two personalization systems show that personalized summaries perform better than generic and generic-personalized summaries in terms of identifying documents that satisfy user preferences. We also considered a user-centred direct evaluation that showed a high level of user satisfaction with the summaries.

Keywords: User model; Summarization; Personalization; Evaluation

1. Introduction

There are different types of summary depending on their purpose or function: indicative, informative, aggregative or critical. These summaries can be generated without taking into account the particular user for whom they are intended or, on the contrary, they can be adapted to the peculiarities – previous knowledge, areas of interest, or information needs – of the reader or group of readers that they are generated for.

In this work, personalized summarization is understood as a process of summarization that preserves the specific information that is relevant for a given user profile, rather than information that truly summarizes the content of the news item. The potential of summary personalization is high, because a summary that would be

[*] Corresponding author. Tel.: +34 918 099 200x310; fax: +34 918 099 217.
E-mail addresses: adiaz@cesfelipesegundo.com (A. Díaz), pgervas@sip.ucm.es (P. Gervás).
[1] Tel.: +34 913 947 639; fax: +34 913 947 529.

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.01.009

useless to decide the relevance of a document if summarized in a generic manner, may be useful if the right sentences are selected that match the user interest. In this paper we defend the use of a personalized summarization facility to maximize the density of relevance of selections sent by a personalized information system to a given user.

If automatic summarization is to be used as part of a process of intelligent information access, it is crucial to have some means of measuring how much information is lost during the summarization process, and how that information loss may affect the ability of the user to judge the relevance of a given document with respect to his particular information needs. To study this problem, the effect of summarization on the information content of a document needs to be studied. The aim is to determine if the use of summaries saves time without significant loss of effectivity in the relevance decisions.

It is reasonable to think that, especially for high levels of understanding, a summary that does not take into account the needs of the user may be too general to be useful. In particular, in the context of information retrieval, summaries adapted to the query presented by the user have shown to be superior (Maña, Buenaga, & Gómez, 1999). In related areas in which the system has available a store of information about user preferences – such as information filtering or personalized information systems – user adapted summarization would be a better way of presenting contents than simply using titles or the first few lines of a document.

In this paper we present an automatic personalization summarization process that has been applied in the frame of reference of a personalization system for a digital newspaper. Sections 2 and 3 describe the techniques for the generation of personalized summaries and the evaluation of them, respectively. Section 4 presents the personalized summarization method. Section 5 describes the personalization system used as framework of reference. Section 6 presents the evaluation methodology used. The results of the different experiments are given in Section 7. Section 8 outlines the main conclusions.

2. Techniques for the generation of personalized summaries

Sentence extraction techniques for building generic summaries are described first, followed by the specific techniques applied to the generation of personalized summaries.

2.1. Sentence extraction

Faced with the variety in type and domains of available documents, sentence extraction techniques are attractive for the generation of personalized summaries. The techniques are based on a phase of analysis during which the segments of the text (extraction units, usually sentences or paragraphs) that contain the most relevant information are extracted. The degree of relevance of each one of them can be obtained by means of a linear combination of the weights that result from a set of different features (Edmundson, 1969). These features can be positional (if they take into account the position of the segment within the document), linguistic (if they search for certain patterns of indicative expressions), or statistical (if they include the frequency of occurrence of certain words).

The summary results from concatenating these selected segments in the order in which they appear in the original document. The length of the summary may vary according to the desired percentage of compression. From the experiments presented in (Morris, Kasper, & Adams, 1992) one can assume that a percentage of compression of between 20% and 30% of the original may be informative enough to replace the original document.

The most frequently used methods for sentence extraction are (Hahn & Mani, 2000): position, indicative expressions, thematic words, proper nouns, text typography, and title. In particular, the title method consist on select those sentences that contain the non-empty words (those which are not on a list of stop words) that appear in the title of the document (Edmundson, 1969; Teufel & Moens, 1997).

In most systems the weights associated to each method, within the linear combination that produces the final value associated to each segment, are adjusted manually. This decision is justified by the fact that the contribution of each method depends on the genre of the texts that are being summarized. However, a classifier may be used to determine the optimal values for these parameters by means of machine learning techniques (Kupiec, Pedersen, & Chen, 1995; Teufel & Moens, 1997). The idea is to use a collection of summaries generated by human experts and their associated texts to determine which the best weights for the linear combination are. The problems with this approach are the need for a corpus of summaries and that the resulting weights depend on the genre of the documents in the corpus.

The main disadvantage of sentence extraction techniques is the possible problem of lack of coherence in the resulting summary. One way to reduce this problem is to use a paragraph instead of a sentence as extraction unit (Salton, Allan, Buckley, & Singhal, 1994), in the hope that providing a wider context improves the problems of legibility. If the lack of coherence is due to unsolvable anaphoric references, it may be eliminated by: not including in the summary sentences with such references (Brandow, Mitze, & Rau, 1995), by including additional sentences immediately preceding the anaphoric reference (Nanba & Okumura, 2000), or by including the necessary sentences to solve the anaphora even if they are not immediately preceding the problematic sentence (Paice, 1990). The main disadvantage of adding extra segments is that, given the restrictions on size, they may push out of the summary sentences with relatively higher weight, thereby degrading the quality of the summary. Another way of solving the lack of coherence is to detect expressions that connect separate phrases ("However...", "Thus..." etc.) and eliminate them from the summary if they appear at the beginning of a sentence and the previous sentence is not part of the summary.

On the positive side, the sentence extraction method has some justification: approximately 80% of the sentences included in manually built summaries appear with little or no modification in the original text (Kupiec et al., 1995). On the other hand, there were no significant differences on reading comprehension texts over manual and automatic summaries, when evaluators where asked to select one out of five possible answers to questions about the content of the text (Morris et al., 1992).

2.2. Personalized summaries

Personalized summaries are not a new idea. The early works on summarization (Luhn, 1958) already mention the possibility of personalizing summaries by adapting them to particular areas of interest. To this end, sentences that contain words related to the reader's interests are given more importance. On the other hand, experiments on summary evaluation by users (Paice & Jones, 1993) showed that users tend to select the parts of the text that are more closely related to their interests.

To build a personalized or user-adapted summary a representation of the interests of the corresponding user is required. This representation may vary from a set of keywords (Maña et al., 1999) to a complex user profile where the information needs of the user are represented according to several systems of reference (Díaz, Gervás, & García, 2005).

The method presented above for occurrence of words from the title in the sentences of the body of the text may be generalized to cover the occurrence of words that may be especially important for a given user. For instance, those words that appear in the user query presented to an information retrieval system, or those words present in a user model. In this way, personalized summaries may be generated adapted to the needs of the user expressed as a query or a user model.

Certain approaches (Tombros & Sanderson, 1998) generate user adapted summaries by combining position, title, and words-in-the-user-query methods. In (Carbonell et al., 1997) the method for accounting for words in the user query is used, though the introduction of redundant sentences is avoided. In (Sanderson, 1998) the passage (Callan, 1994) most relevant to the query, previously expanded with the words that occur most frequently in the context in which the words of the query appear in the first documents retrieved, is selected. An expansion of the query using WordNet is carried out in (Maña et al., 1999). Another type of query expansion is applied in (Amini & Gallinari, 2002). In this case WordNet is used together with the most frequent words in the sentences that are more relevant to the query.

3. Techniques for the evaluation of summary generation

From the beginning of research on automatic summary generation to the present day, several proposals have arisen (individual and collective) to try to find a consensus on the various aspects related to evaluation, for instance, the SUMMAC (Mani et al., 2002) or DUC[2] conferences. However, no agreement has been reached yet over which is the best evaluation method, due to the complexity of the task.

Among the reasons for this lack of agreement, the following may be mentioned (Mani, 2001): difficulty for agreeing on when an automatically generated summary is good (comparing with an ideal summary generated by a person is not enough), the need to resort to real humans to judge the summaries (this makes the process difficult to carry out and difficult to repeat, additionally there are problems with the degree of agreement between different human judges), the need for metrics that take into account aspects that go beyond the readability of the summary, such as the relevance of the obtained information, how well the summary meets the user needs, or how well it satisfies its intended use.

The different approaches to summary evaluation are presented below. They can be classified, according to (Sparck-Jones & Galliers, 1996) into direct or intrinsic, and indirect or extrinsic. The former are based on a direct analysis of the summary by means of some metric for establishing its quality. The latter are based on judgments in terms of how useful the summaries are when a specific task is carried out over the summaries instead of over the original documents.

3.1. Direct or intrinsic evaluation

This type of evaluation can take into account (Baldwin et al., 2000) both quality criteria (such as grammatical correctness, or text cohesion), and criteria about information coverage. For some of these aspects the summary is compared with the original document, as in the case of checking for the appearance in the summary of the main ingredients of the original text. For others, such as grammatical correctness or cohesion, the summary is considered in isolation.

Human judges are required for this type of evaluation, to build an initial summary to be used for comparison or to judge the degree of adequacy and correction of the automatically generated summaries. This involves a considerable human effort, especially if we take into account that, for results to be significant, judgments must be applied to large collections of text, such as the ones used for the DUC conferences. Additionally, human judgments may disagree. Large enough disagreements may invalidate the experiments.

With respect to the readability of the summaries, the following linguistic criteria were applied at the DUC'05 conference (Dang, 2005): grammatical errors, redundancy errors, reference errors, focus, structure and coherence. Each criterion was graded on a scale of 5 points – very poor, poor, acceptable, good, very good.

If a good evaluation is obtained according to the criteria above, a certain readability of the summary can be guaranteed. However, this does not take into account what use the summary is intended for. Neither does it guarantee that the summary is good. Other factors must be used to determine the relevance of the information that it includes or its usefulness for a given task.

The most commonly used method for measuring the relevance of the content of an automatically generated summary is the comparison with an 'ideal' summary built manually. However, the lack of agreement between human judges seems to indicate that this 'ideal' summary does not always exist. For this reason, it is possible that a summary contain relevant information but it is not considered good because it does not resemble the 'ideal' summary.

The metrics used for information retrieval, precision, recall and F_1 (Salton & McGill, 1983), have been used frequently to evaluate the relation between automatic summaries and the ideal summary. The relevant parameter is the number of sentences that appear both in the automatic summary and the ideal summary, divided by the number of sentences in the automatic summary (recall) or the number of sentences in the ideal summary (precision).

A first solution, to the direct comparison of sentences, is to leave to the criteria of the judges which sentences of the automatic summary express part of the information of the reference summary. These same judges also determine the percentage of information of the ideal summary that is covered by the automatic summary (Harman & Marcu, 2001). Another possibility is to use the utility index (Radev, Jing, & Budzikowska, 2000)

[2] DUC: http://www.duc.com.

or measures (Donaway, Drummey, & Mather, 2000) of sentence ordering and content similarity based on the cosine measure of the vector space model (Salton & McGill, 1983).

The SEE[3] (Summarization Evaluation Environment) tool, used from the DUC04 conference (Over & Yen, 2004), facilitates the direct evaluation of automatic summaries. This tool compares a model summary with a peer summary, producing judgments of peer linguistic quality (7 questions), coverage on each model unit by the peer (recall) and relevance of peer-only material. The starting point is a manual summary (model summary) divided into model units (MUs). This summary is compared with the automatically generated summaries (peer summaries) divided into peer units (PUs). Manual evaluation consists of determining what is the percentage of similar content between each MU and each set of PUs that is somehow related with the corresponding MU, and providing answers to the 7 questions about linguistic quality.

Another package used for direct evaluation is ROUGE[4] (Recall-Oriented Understudy for Gisting Evaluation) (Lin & Hovy, 2003; Lin, 2004), also used since the DUC04 conference. This package measures the quality of a summary by comparing it with one or several ideal summaries. The metrics are based on the number of overlapping units, using measurements such as N-gram co-occurrence (ROUGE-N), longest common subsequence (ROUGE-L), weighted longest common subsequence (ROUGE-W) or skip-bigram recall metric (ROUGE-S).

3.2. Indirect or extrinsic evaluation

This type of evaluation tries to determine quantitatively how appropriate or useful a given summary is with respect to the full text in the fulfillment of a given task, for example, information retrieval. There are two possibilities to perform this kind of evaluation: to use human judges to determine the relevance of the summaries for the task, or to measure the effectivity of the summaries automatically using them as sources for the task.

Indirect evaluation in information access tasks has been applied in many cases. Most of them try to measure the repercussion of using summaries instead of full texts in decision taking about the relevance of documents. The usual hypothesis is that there is a reduction in the time required with no significant loss of effectivity.

The usefulness of summaries of ad hoc retrieval and categorization task was studied in (Mani & Bloedorn, 1998; Mani et al., 2002). For ad hoc retrieval, the evaluation was carried out over indicative summaries oriented to the query over newspaper articles. Twenty queries were used, and for each one of them a subset of 50 documents was selected out of the first 200 retrieved by a standard retrieval system. The selection was carried out so that the number of relevant documents for each query was between 25% and 75%, and no document appears in more than one subset. A corpus of 1000 documents resulted. Twenty one human judges took part in the evaluation, which consisted in determining the relevance of a document given a query. The documents provided to the evaluators were: original documents, summaries of a fixed length (10% of the number of characters in the source text), summaries of variable length and initial segments of the source text (10% of the sentences). This last item served as control element, since the documents of the experiment all had a lot of relevant information at the beginning of the document. The evaluation showed similar effectivity (no significant differences), in terms of F_1, when the full text and when the variable length summaries are used (average compression between 15% and 30%). However, more than 40% of time is saved when the summaries are used.

For text categorization, the evaluation centred on generic summaries, for there was no query with which to personalize the summary. Two groups of 5 categories belonging to mutually exclusive domains where used. For each category 100 documents were selected. The evaluation was carried out by 24 judges and it consisted of choosing one of five predetermined categories – or "none of them" – for which the content of a given document was deemed relevant. The documents were of the same type as those used for the experiment on ad hoc retrieval. Similar effectivity was obtained with the two types of summary and the full text. Nevertheless, only the fixed length summaries reduced the time employed (by 40%).

In Tombros and Sanderson (1998) the utility of the initial segment of the original documents is compared with that of query oriented summaries. To this end, an information retrieval system is used in which the user is

[3] SEE: http://www.isi.edu/~cyl/SEE.

[4] ROUGE:http://www.isi.edu/~cyl/ROUGE.

shown, next to the title of the retrieved document, either the first sentences or the automatically generated summaries. Fifty TREC queries and 50 documents per query are used. Measured parameters include precision, recall, the time needed for the user to decide, the number of times that the user accesses the full text, and the subjective opinion of the users on the quality of the aid provided (either first lines or generated summary). The results show that user oriented summaries significantly improve user efficiency in the ad hoc retrieval task with respect to the use of the first lines.

In (Dorr, Monz, President, Schwartz, & Zajic, 2005) it is shown that two types of human summaries, Headline Surrogate (non-extractive ''eye-catcher'') and Human Surrogate (extractive headline), can be useful for a relevance assessment task in that they help a user achieve a high agreement in relevance judgments with respect to those obtained with the full texts. On the other hand, a 65% reduction in judgment time is observed between full texts and summaries. In this work a new method for measuring agreement, Relevance-Prediction, is introduced. This method takes a subject's full-text judgment as the standard against which the same subject's summary judgment is measured.

A different approach is followed in (Brandow et al., 1995; Maña et al., 1999; Nomoto & Matsumoto, 2001). In this case, the evaluation method consists of comparing the effectivity of the retrieval when the queries are presented over summarized versions of the documents, instead of over the documents in the original collection. The advantage of this approach is that it does not need judges, avoiding the problems of agreement between judges, and making the evaluation fully automatic.

In Brandow et al. (1995) the effectivity of the automated summaries is compared with that of the initial segment of documents. A corpus of newspaper articles and 12 queries are used. Better effectivity is obtained for the initial segments. Based on clustering techniques, (Nomoto & Matsumoto, 2001), uses a similar approach with a keyword-based automatic summary generator as control element. The corpus contains 5000 news items from a Japanese newspaper and 50 queries. Relevance judgments can be: total relevance, somewhat relevant, and irrelevant. Using the first two types of judgments, the system produces more effectivity than the control element. If only the first level of relevance is used, the system is better only for compression rates of up to 30%. In no case does the effectivity reach that achieved using the full text.

In Maña et al. (1999) similar experiments are carried out using 5000 documents chosen at random out of the Wall Street Journal corpus (news items), and 50 randomly chosen queries with at least one relevant document in the chosen set. The experiments compare in terms of precision-recall curves the effectivity of searches with the full text, with generic summaries, with summaries adapted to the query, with summaries adapted by expanding them with WordNet synonyms, and with the initial segment. Three different compression rates have been used: 10%, 15% and 30%. Results show that query adapted summaries improve the average precision between 30% and 40%, compared with the first sentences of the documents. They also improve upon generic summaries in a similar proportion. Finally, the effectivity of adapted summaries is statistically comparable with that obtained using the original texts. On the other hand, the experiments show that the expansion of the query using WordNet does not improve the effectivity of the summaries.

In conclusion, the indirect evaluation, without human judges, allowed an evaluation less expensive and easier to extend to big text corpora and to different summarization techniques. The evaluation can be performed in a more exhaustive and automatic way.

4. Automatic personalized summarization

The use of summaries allows users to save a significant amount of time by allowing them to correctly identify whether a relevant document is really interesting for them without having to read the full text. If the summary is personalized according to user interests, the user can take even less time not only in deciding whether it is interesting or not, but also in finding the information that is actually interesting to him out of the full text.

The summary of each document is extracted automatically from the full text of the document using techniques for the selection and extraction of sentences, where a sentence is considered the sequence of words between two periods that indicate the end of sentence.

Techniques for selection and extraction of sentences are very attractive because they are independent of the domain and the language. Additionally, these techniques can take into account information stored in a user model, allowing personalization of the summaries, that is, to generate summaries adapted to a user profile.

The different techniques to be used, as well as their possible combinations, are described below. Three different methods for phrase or sentence selection are used to build the summaries. The first two are used two build generic summaries, and the third one is used to generate personalized summaries. Possible combinations of these methods to build summaries with different properties are indicated.

The three methods have a common goal: to assign to each sentence in the text a value that indicates its relevance as a candidate for the summary. The final selection will consist of a percentage, usually 20%, of sentences that have higher relevance. These sentences are concatenated respecting their original order of appearance in the full text, to avoid inconsistencies.

Each method is described, and the various parameters that can be used to customize the final summaries are described in each case.

4.1. Generic summaries

First we describe the two methods that are independent of the user to which the summary is addressed. The first one takes into account the position of the sentence in the document, and the second one takes into account the significant words that appear in the sentence.

4.1.1. Position method

In many domains (i.e journalism), the first few sentences of a text may constitute a good summary of the content of the full text. This method assigns the highest values to the first 5 sentences of the text (Edmundson, 1969). In our work, the specific values chosen for these 5 sentences are shown in Table 1. Sentences from the sixth on are assigned value 0. These values generate the weights A_{sd} for each sentence s within document d. These values are independent of the user u under consideration.

A ranking of the sentences of a document with respect to the position in which they occur in the document is obtained from this method. In the case of this particular version of the method, this ranking corresponds to the order of the sentences in the document.

4.1.2. Thematic word method

Each text has a set of significant or "thematic" words that are representative of its content. This method extracts the 8 non-empty most significant words (non-empty words are those that do not appear in the stop list) for each text and it calculates how many occurrences of those words there are in each sentence of the text. Sentences will be assigned a higher value the more significant words they contain (Kupiec et al., 1995; Teufel & Moens, 1997).

To obtain the 8 most significant words of each document, documents are indexed to provide the weight of each word in each document using the $\mathrm{tf} \cdot \mathrm{idf}$ method (Salton & McGill, 1983).

To obtain the assigned value B_{sd} for each sentence s in document d using the thematic word method, the number of significant words appearing in a sentence is divided by the total number of non-empty words appearing in the sentence. The aim of this calculation is to assign a higher weight to sentences with a higher density of significant words (Teufel & Moens, 1997). These values are also independent of the user.

This method provides a ranking of the sentences in a document according to the number of thematic words that the sentence contains.

Table 1
Values assigned to the sentences in a document according to the position heuristic

Sentence number	Assigned value
1	1.00
2	0.99
3	0.98
4	0.95
5	0.90
Others	0.00

4.1.3. Combination of generic methods

To obtain the value G_{sd} associated with each sentence s in a document d using combinations of the methods described above, formula (1) is applied:

$$G_{\mathrm{sd}} = \frac{\varphi A_{\mathrm{sd}} + \gamma B_{\mathrm{sd}}}{\varphi + \gamma} \tag{1}$$

The parameters φ and γ allow adjustment of the different methods, depending on whether a position method (φ) or a significant word method (γ) is preferred. To ensure significance, the relevance obtained from each method must be normalized.

From formula (1), a ranking of the sentences in each document is obtained with respect to the generic methods, which are independent of the user.

4.2. Personalized summaries

In the next section the personalization method, which uses information dependent on the user for whom the summary is intended, are described.

4.2.1. Personalization method

The aim of this method is to select those sentences of a document that are most relevant to a given user model. The values P_{sdu} associated to each sentence s in a document d for a user u are obtained by calculating the similarity between the user model for user u and each one of the sentences in the document.

This method provides a ranking of the sentences in each document according to the personalization method.

The complexity of a user model that is being used allows various possibilities when customizing this method. Depending on the part of the user model which is selected to participate in calculating the similarity, different types of personalization may be obtained.

This procedure is applicable to any personalization system capable of assigning to sentences in a document a relevance weight that corresponds to the estimated importance of that sentence with respect to the user's needs.

4.3. Combination of generic and personalized methods

To obtain the value Z_{sdu} associated with each sentence s in a document d for a user u, using a combination of all the summarization methods, formula (2) must be applied:

$$Z_{\mathrm{sdu}} = \frac{\chi G_{\mathrm{sd}} + \mu P_{\mathrm{sdu}}}{\chi + \mu} \tag{2}$$

Parameters χ and μ allow the adjustment between the different methods, depending on whether they are generic (χ) or personalized (μ). In this way, the value of the parameters determines the degree of personalization of the summaries: if μ is 0, resulting summaries will be generic, and for values of μ larger than 0, summaries will have a certain degree of personalization. For this combination to be significant, the relevance obtained with respect to each method must be normalized.

From formula (2) a ranking of sentences is obtained for each document with respect to the combination of all proposed methods.

5. The personalization system

The automatic summarization process will be applied in the frame of reference of a personalization system for a digital newspaper. Each day the system sends to the users a selection of the news items more relevant with respect to his user model, defined in the moment they register in. Moreover, the use can interact with the system providing feedback information about the news items he receives (Díaz & Gervás, 2005).

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

5.1. User model

The user model (Billsus & Pazzani, 2000; Díaz & Gervás, 2004a) involves a domain specific characterization, an automatic categorization algorithm and a set of keywords (long-term model), and a relevance feedback tier (short-term model). The long-term model reflects the information needs of the user that remains stable across the time. On the other hand, 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 a specific given classification system related with the application domain, that is, the sections of the digital newspaper. The user can assign a weight to each section: national, sport, etc. This information is stored as a matrix where rows correspond to sections and columns correspond to users (S_{su}). For the second tier, the user enters a set of keywords – with an associated weight – to characterize his preferences. These keywords are stored, for each user u, as a term weight vector (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. This information is stored as a matrix where rows correspond to categories and columns correspond to users (G_{gu}). These categories are represented as term weight vectors (g) by training from the very brief descriptions of the first and second level of Yahoo! Spain categories entries (Labrou & Finin, 2000).

In the fourth tier, short-term interests are represented by means of feedback terms obtained from feedback provided by the user over the documents he receives (Díaz & Gervás, 2004a). The term weight vector for each user (f_{u}) represents the short-term interests of that user, information needs that lose interest to the user over time, so their weight must be progressively decreased.

5.2. Multi-tier content selection

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^{s}_{du}).

The similarities between a document d and a category g (s_{dg}), between a document d and the keywords of a user model u (s^{k}_{du}), and between a document d and a short-term user model u (s^{t}_{du}) are computed using the cosine formula for similarity within the Vector Space Model (Salton & McGill, 1983):

$$s_{\mathrm{dg}} = \mathrm{sim}(d_{\mathrm{d}}, g) \quad s^{\mathrm{k}}_{\mathrm{du}} = \mathrm{sim}(d_{\mathrm{d}}, k_{\mathrm{u}}) \quad s^{\mathrm{t}}_{\mathrm{du}} = \mathrm{sim}(d_{\mathrm{d}}, f_{\mathrm{u}}) \tag{3}$$

The similarity between a document d and the categories of a user model is computed using the next formula:

$$s^{\mathrm{g}}_{\mathrm{du}} = \sum_{i=1}^{14} G_{i\mathrm{u}} s_{\mathrm{dg}_i} \Big/ \sum_{i=1}^{14} G_{i\mathrm{u}} \tag{4}$$

The results are integrated using a particular combination of reference frameworks. The similarity between a document d and a user model u is computed as:

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

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

From formula (5) a ranking of documents is obtained with respect to the complete user model.

5.3. Automatic personalized summarization

In this system the parts of the user model used are the keywords from the long-term model and the feedback terms from the short-term model. From each reference framework a value is obtained by calculating the similarity between the part of the user model being used and each one of the sentences in the document. The values from both frameworks are combined to obtain, from formula (8), the final value P_{sdu} described in the Section 4.2.1. These processes are described in the next sections.

The final value assigned to each sentence in each document is obtained from formula (2) where generic and personalized methods are combined. From this formula the final ranking of sentences for each document is obtained. The sentences selected to build the summary consist of the 20% of sentences that have higher relevance. These sentences are concatenated respecting their original order of appearance in the full text, to avoid inconsistencies.

5.3.1. Personalization using keywords from long-term model

The value P^{k}_{sdu} associated with each sentence s, in document d, for a user u, using keywords k of the long term model is obtained by calculating the similarity between the keyword vector and the vector for each sentence using formula (6):

$$P^{\mathrm{k}}_{\mathrm{sdu}} = \mathrm{sim}\,(s_{\mathrm{sd}}, k_{\mathrm{u}}) \tag{6}$$

where s_{sd} is the term weight vector for sentence s in document d, k_{u} is the term weight vector corresponding to the keywords of the long term model of user u and sim is the similarity measure of the Vector Space Model.

From formula (6) a ranking of sentences is obtained for each document with respect to the personalization method using the keywords of the long term model.

5.3.2. Personalization using the short term model

In a similar way personalized summaries can be obtained using the short term model of user u. In this case, the similarity is calculated between sentence s and the relevance feedback terms f_{u}, of the short term model of user u. In this way, the value $P^{\mathrm{f}}_{\mathrm{sdu}}$ associated with each sentence s in a document d for a user u using the feedback terms of the short term model is calculated as the similarity between the relevance feedback term weight vector and the term weight vector for each of the sentences, as shown in formula (7):

$$P^{\mathrm{f}}_{\mathrm{sdu}} = \mathrm{sim}\,(s_{\mathrm{sd}}, f_{\mathrm{u}}) \tag{7}$$

where s_{sd} is the term weight vector for sentence s in document d, f_{u} is the term weight vector corresponding to the relevance feedback terms of the short-term model of user u and sim is the similarity measure of the Vector Space Model.

From formula (3) a ranking of sentences is obtained for each document with respect to the personalization method using the short term model.

5.3.3. Personalization using combinations of short and long term models

The values obtained for each part of the user model can be combined to obtain a personalized summary that uses the complete user model. In this case, the value P_{sdu} assigned to each sentence s, in a document d, for a user u, is obtained with formula (8):

$$P_{\mathrm{sdu}} = \frac{\eta P^{\mathrm{k}}_{\mathrm{sdu}} + \kappa P^{\mathrm{f}}_{\mathrm{sdu}}}{\eta + \kappa} \tag{8}$$

where Greek letters η and κ show the importance assigned to each of the reference systems (η, for key words in the long term model, κ, for the feedback terms of the short term model). To ensure significance, the relevance obtained from each reference framework must be normalized.

From formula (8), a ranking of the sentences in each document is obtained with respect to the complete user model.

6. Evaluation methodology

To measure the effectivity of personalized summaries a series of methods is needed for measuring to what extent the summary allows the user to obtain the information relevant to his interests without having to read the full text, or at least, allows him to make the correct decision on whether it is worth the trouble to read the full text.

We have used an indirect/extrinsic evaluation based on the differences resulting between parallel processes of selection, using the user models, over a set of parallel collections: one with the original documents and the others built with differently summarized versions of the documents.

We have also carried out a user centred direct/intrinsic evaluation based on a set of questions answered by each one of the users concerning the summaries that they have received.

6.1. Evaluation collection

An evaluation collection 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 with respect all the documents. Typical examples are the collections associated to the conferences TREC or DUC.

Evaluation collections for personalization such as the one describe 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. This 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 during several days. Then, the evaluation collection is composed of the set of news item from a digital newspaper corresponding to several days, and the set of binary (relevant/not relevant) judgments cross-indexed with the news items assigned manually by each user with respect to his proper user model.

On the other hand, there is no collection of ideal summaries with which one could compare automatically generated personalized summaries. This is because it is very difficult to build an ideal personalized summary of each document for each possible user.

6.2. Evaluating summary generation

What needs to be evaluated is the loss of significant information that a user suffers when he is presented with a summary instead of the corresponding full document. Additionally, it will be considered the explicit opinions expressed by the users in their replies to some questions about the quality of the summaries.

6.2.1. Indirect evaluation

The technique is based on the assumption that if a document summarization process is good, then the resulting summary must retain enough of the original information to guarantee a correct judgment concerning the similarity between the summarized item and a given user model. This type of quantitative evaluation is typical of summary extraction systems which operate within more complex systems that carry out other tasks, such as information retrieval or information filtering (Maña et al., 1999).

For each user. To access the evaluation corpus, using the algorithm that is to be evaluated. The selection process (Section 5.2) is applied to this new personalized version of the collection, using as selection criteria the user model corresponding to that user. The hypothesis is that, if the summarization process employed preserves the information in the document that is relevant to the user model, the results of the selection process should be comparable to those obtained for the same user with the full text of the documents, which are taken as upper bound reference values. Any deviation from these reference values indicates information loss due to "leaks" in the generation of summaries, which have produced a variation in the ranking obtained for the collection of summaries with respect to the ranking obtained for the collection of full text documents. By applying a similar process for each algorithm for summary generation, this experiment should show an explicit, though indirect, measure of the adequacy of the personalized summary generation.

6.2.2. User-centred direct evaluation

To perform a direct evaluation about the readability of the summaries, each user might have been asked to evaluate each summary he received. However, the fact that each user receives approximately one hundred summaries per day, during several days, convinced us that we should not consider this possibility. Instead each user was only asked for his opinion at the end of system use.

The questions presented to the user concerned the following aspects: quality of the summary, coherence, clarity and structure, redundancy errors, adaptation to the user model, adaptation to information needs of the user, reflection of the document content and representation of the main components of the document.

Users were also asked about the part of the news item that they had used to decide about the relevance: title, relevance, section, personalized summary or full news item. Of these parts, the first four appeared in the message that was sent to each user. To access the complete news item an additional click over a hyperlink was necessary.

On the other hand, the great quantity of news items processed, together with the associated personalized summaries, made it inadvisable to use a direct evaluation method based on the comparison with an ideal summary.

6.3. Hypothesis

The questions that need to be answered are:

- For the obtention of personalized summaries using only the personalization methods, is it better to use a static long term model or a dynamic short term model or a combination of both?
- How much is lost, in terms of information received by the user, by offering a summary of a document in place of the full document?
- What type of summary is better in this sense?

These questions give rise to the following hypothesis:

H1. Summaries obtained using only the personalization method are better if a combination of the long and short term models is used than if any of them is used on its own.
H2. Summaries obtained using only the personalization method are better in terms of information selected by the users than summaries obtained by extracting the first sentences of the full document.
H3. Summaries obtained using only the personalization method are better in terms of information selected by the users than summaries obtained using the methods for generation generic summaries and the summaries obtained using a combination of methods.
H4. Summaries obtained using only the personalization method are worse than the full document in terms of information selected by the users.

6.4. Design of experiments

To evaluate the generation of summaries the system is tested with the various combinations of parameters associated with the process. This involves the different combinations of methods for generating summaries. In this way results will be obtained associated with the different ways of generating summaries.

6.4.1. Design of experiment 1: generating personalized summaries

To test the first hypothesis (H1) one must evaluate for each user all the different types of summary that can be generated using only the personalization method (P_{sdu}). This corresponds to the different combinations of the use of long and short term models for the generation of personalized summaries using only the personalization method. One must take into account that this will involve using the key words of the long term model and the relevance feedback terms of the short term model. The different possibilities correspond to the following assignment of values to the parameters of formula (8):

- Ps(L): personalized summary using only the keywords of the long term model ($\eta = 1$, $\kappa = 0$).
- Ps(S): personalized summary using only the relevance feedback terms of the short term model ($\eta = 0$, $\kappa = 1$).
- Ps(LS): personalized summary using a combination of the keywords of the long term model and the relevance feedback terms of the short term model ($\eta = 1$, $\kappa = 1$).

Several collections have been generated for each user, each one consisting of a set of summaries of the original documents obtained by means of the application of each one of the indicated methods for generating personalized summaries. In this way, there will be a collection for each user of personalized summaries built using the short term model, another one built with the long term model, and another one built using a combination of both.

As the collection of summaries is different for each user, the selection process must be carried out separately for each particular user, so there will be as many selection processes as the number of users involved.

6.4.2. Design of experiment 2: combining of methods for summary generation

To test the hypotheses (H2, H3 and H4) one must evaluate for each user all the different types of summary that can be generated using the different combinations of method (Z_{sdu}). The summaries to be generated will be of different types, depending on the specific methods used to generate them. This corresponds to the following assignment of values to the parameters of formulas (2) and (8):

- Fs (baseline summary): 20% first sentences of the full text. This summary will act as baseline for the rest of the experiments. In truth, it would correspond to the position method if the values were extended incrementally to all the sentences of a document.
- Gs (generic summary): summary obtained using the methods for generating generic summaries ($\phi = 1$, $\gamma = 1$, $\mu = 0$).
- GPs (generic-personalized summary): summary obtained using both types of methods ($\phi = 1$, $\gamma = 1$, $\mu = 1$, $\eta = 1$, $\kappa = 1$).
- Ps (personalized summary): summary obtained using only the personalization method (combining long and short term models). ($\phi = 0$, $\gamma = 0$, $\mu = 1$, $\eta = 1$, $\kappa = 1$). This corresponds to Ps(LS) of the previous experiment 1.

Several collections have been generated for each user, each one consisting of a set of summaries of the original documents obtained by means of the application of each one of the indicated methods for generating summaries. The collections of baseline summaries and generic summaries are the same for all users, since they do not depend on any user model – they are not personalized. This allows simultaneous evaluation of all users.

However, there will be a different collection per user for personalized and generic-personalized summaries, which do depend on user models. In this way, each user will have to be individually evaluated with respect to his collection of personalized summaries and his collection of generic-personalized summaries.

6.5. Evaluation metrics

The results to be obtained are a ranking of documents (complete texts and summaries) for each user, obtained from the application of the selection process by means of formula (5) to those documents.

These results must be compared with the binary relevance judgments associated with the full texts. This comparison between a ranking of documents and binary relevance judgments (relevant/not relevant) suggests the use of normalized recall and precision metrics. 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.

Normalized precision and recall (Rocchio, 1971) measure the effectivity of a ranking in terms of the area on a recall or precision versus levels of ranking graph encompassed by the best possible solution (relevant documents in the first positions) and the solution generated by the system under evaluation. In those cases where the same relevance value appears at consecutive positions in the ranking, the average position of all the matching values is taken as value for all of them (Salton & McGill, 1983). This adjustment solves the problem of attributing a random relative ordering to the positions that have the same relevance value associated.

For each of the different configurations the values of normalized recall and precision will be obtained, for each user. To obtain the final value, for each configuration, the results will be averaged, obtaining a final average per user, which will serve as effectivity value for the process of generating summaries for each of the established configurations.

6.6. Statistical significance

For two techniques A and B, one must show that technique A achieves better results than technique B ($A > B$) with respect to a parameter V. To this end, the number of times that technique A gives better results than B is represented as $V+$, and the number of times that technique B gives better results than A is represented as $V-$, and the number of times that both techniques give the same results as draws. This will be applied both to normalized recall ($V = R$) and normalized precision ($V = P$).

A result is considered statistically significant if it passes the sign-test, with pairs of values, with a significance level of 5% ($p <= 0.05$). The decision to use this type of statistical significance test is based on the fact that there is no assumption about the underlying distribution, and that, given the different normalization processes that are applied at different levels, it must be used the relative values rather than the absolute magnitudes to establish statistical significance (Salton & McGill, 1983).

7. Results

Two experiments have been carried out over two different personalization systems. The first one had a smaller evaluation collection (11 users, 5 days) and it did not use the categories of the user model. The second one had a fuller evaluation collection (104 users, 14 days) and it used the full user model.

7.1. First personalization system

For summary generation, the keywords of the long term and the relevance feedback terms of the short term model were used. On the other hand, the selection of the different types of summaries was done using formula (5) with variables α, χ and ε set to 1, and β set to 0, and no categories were used (Díaz & Gervás, 2004b; Díaz & Gervás, 2005).

The hypotheses presented for experiments 1 and 2 were evaluated using the indirect evaluation presented in Section 6.2. No direct evaluation was carried out over this system.

7.1.1. Evaluation collection

Experiments were evaluated over data collected for 11 users, mainly computer science lecturers, and the news items corresponding to five days – period 6th–10th May 2002 – of the digital edition of the ABC Spanish newspaper. The number of news item per day was: 128, 104, 87, 98 and 102. It was used the news items from the next sections: national, international, sports, economy, society, culture, people and opinion. The total number of news items was 519.

7.1.2. Results of experiment 1: generating personalized summaries

There was a collection for each user of personalized summaries generated using the short term model (Ps(S)), a different collection for each user generated using the long term model (Ps(L)) and a third different collection for each user generated using a combination of long term and short term models (Ps(LS)). The multi-tier selection process was applied to each one of these collections, using the corresponding user profile as source for user interests. In each case, values of normalized recall and precision were computed.

These experiments were repeated for all users during the 5 days of evaluation. The results for the three types of personalized summaries were compared only from the second day on, for the fact that on the first day there was no short-term model based on user feedback. In total, it was generated 12903 different personalized summaries (11 users, 391 news item, 3 types of summaries per user and per news item).

The results shown in Table 2 indicate that the combination of long and short term models for the generation of personalized summaries provides significantly better results than the use of each model separately, in terms of normalized precision (2.5% against long term only, 1.0% against short term only). As an additional result, it is observed that the short term model on its own is better than the long term model in terms of normalized precision (1.4%), though not significantly so. In terms of normalized recall, results are similar: significant improvement of the long term-short term combination over both short and long on their own, and non-significant improvement of short term only over long term. This confirms hypothesis H1.

The use of both methods adjusts the summaries better to the preferences of the user, as shown by higher values of precision and recall. The slightly better results for the short term could be due to the fact that the terms introduced by the user in his long term model are in general too specific, whereas those obtained through user feedback are terms that appear in the daily news.

From here on, mentions of personalized summaries (Ps) refer to the personalization obtained by means of a combination of the long and short-term models (Ps(LS)).

7.1.3. Results of experiment 2: combining of methods for summary generation

The evaluation collections for the first sentences summaries (Fs) and generic summaries (Gs) are identical for all the users, because they do not depend of the user model, that is, they are not personalized. This allowed evaluating all the users together. However, there are a different collection per user of personalized summaries (Ps) and generic-personalized summaries (GPs).

The multi-tier selection process is applied to each one of these collections, using the corresponding user profile as source for user interests. In each case, values of normalized recall and precision have been computed in experiments that have been repeated over the 5 days for all users. In total, it has been generated 12456 different personalized summaries (24 different summaries per news item, 519 news items).

The analysis of the results shown in Table 3 indicates that personalized summaries give significantly better results with respect to normalized precision of the selected information than generic summaries and generic-personalized summaries. In both cases the improvement is statistically significant (3.7% against generic and 3.3% against generic-personalized). This confirms hypothesis H3. Generic-personalized summaries are better than generic summaries, and generic summaries are better than summaries based on first sentences, but in neither case is the difference statistically significant.

Table 2
Averages normalized recall and precision for different combinations of long and short-term model for generating personalized summaries, in the first experiment

	6-May		7-May		8-May		9-May		10-May		Average	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
Ps(LS)	0.616	0.508	0.581	0.486	0.620	0.538	0.632	0.535	0.657	0.550	0.621	0.523
Ps(S)	0.610	0.497	0.588	0.493	0.622	0.550	0.624	0.526	0.642	0.524	0.617	0.518
Ps(L)	0.616	0.508	0.574	0.482	0.611	0.532	0.625	0.523	0.631	0.507	0.611	0.510

Table 3
Averages normalized recall and precision for different types of summaries, in the first experiment

	6-May		7-May		8-May		9-May		10-May		Averages	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
N	0.616	0.507	0.577	0.483	0.619	0.540	0.640	0.555	0.659	0.544	0.622	0.526
Ps	0.616	0.508	0.581	0.486	0.620	0.538	0.632	0.535	0.657	0.550	0.621	0.523
GPs	0.612	0.497	0.563	0.448	0.603	0.527	0.625	0.521	0.649	0.539	0.611	0.506
Gs	0.607	0.479	0.563	0.452	0.612	0.526	0.631	0.529	0.652	0.538	0.613	0.505
Fs	0.607	0.479	0.557	0.450	0.598	0.513	0.622	0.522	0.649	0.536	0.607	0.500

It can also be seen that personalized summaries are worse (1.7% on normalized precision) than complete news items (N) under the same view point. This confirms hypothesis H4. Personalized summaries are better than summaries based on the first sentences of the news item (Fs), with a statistically significant improvement (4.7% in normalized precision). This confirms hypothesis H2. The improvements on recall are similar but with small percentages.

7.2. Second personalization system

As in the first system, the keywords of the long term and the relevance feedback terms of the short term model were used for summary generation. On the other hand, the selection of the different types of summaries was done using formula (5) with all the reference systems: sections, categories, keywords and feedback terms (Díaz & Gervás, 2004a; Díaz et al., 2005).

The hypothesis presented for experiments 1 and 2 were evaluated using the indirect evaluation presented in Section 6.2. The user-centred direct evaluation showed the opinions of the users about the summaries received.

7.2.1. Evaluation collection

In this case, experiments were evaluated over data collected for 106 users, most of them lecturers and students of different studies, and the news items corresponding to three weeks – the 14 working days of the period 1st–19th Dec 2003 – of the digital edition of the ABC Spanish newspaper. The news items came from the same sections of the first experiment except opinion. The average of news item per day was 78.5. It is also important to indicate that the number of user per day was decreasing in that way that the average number of user per day was only 28.6.

7.2.2. Results of experiment 1: generating personalized summaries

The multi-tier selection process was applied to each one of the collections of personalized summaries and values of normalized recall and precision were computed. The results were compared only from the second day on, for the fact that on the first day there was no short-term model based on user feedback. In total, it was generated 77,598 different personalized summaries (28.6 users on average per day, 1004 news items, 3 types of summaries per user and per news item).

The results shown in Table 4 show that the combination of long and short term models for the generation of personalized summaries provides significantly better results than the use of each model separately, in terms of normalized precision (1.6% against long term only, 2.8% against short term only). As an additional result, it is observed that the short term model on its own is better than the long term model in terms of normalized

Table 4
Averages normalized recall and precision for different combinations of long and short-term model for generating personalized summaries, in the second experiment

	nP	nR
Ps(LS)	0.592	0.684
Ps(S)	0.583	0.678
Ps(L)	0.576	0.674

precision (1.2%), though not significantly so. In terms of normalized recall, results are similar: significant improvement of the long term-short term combination over both short and long on their own, and non-significant improvement of short term only over long term. This confirms hypothesis H1.

7.2.3. Results of experiment 2: combining of methods for summary generation

In each case, normalized recall and precision values were calculated. The experiments have been repeated over the 14 evaluation days. In total, it has been generated approximately 54950 different personalized summaries (58 different summaries per news item, 1099 news items).

Personalized summaries offer better results (Table 5) with respect to normalized precision than generic-personalized summaries, though the difference is not significant (1.5%). With respect to baseline summaries and generic summaries the difference is significant (2.0% and 2.8%, respectively). Generic-personalized summaries are better than baseline summaries (0.5%), and baseline summaries are better then generic summaries (0.7%), but the differences involved are not statistically significant. Personalized summaries are worse than full news items under the same criteria (1.7%). With respect to the normalized recall the results are similar.

This suggests that the personalization method generates the summaries better adapted to the user, followed by a combination of all possible methods. Baseline summaries using the first lines of each news item are better than those generated by a combination of the position and keyword methods. For newspaper articles, the generic method does not improve on simply taking the opening lines.

These results confirm the hypothesis H2 and, although the differences are not significant, the hypothesis H3 and H4.

7.2.4. User-centred direct evaluation

The user-centred direct evaluation was based on a questionnaire that users completed after using the system. In most questions there were 5 options to indicate the degree of satisfaction: very high, high, medium, low and very low. There were 38 users that completed the final evaluation.

Users indicated that the summaries were of high or very high quality in 83.3% of the cases, with 5.6% of very low. Concerning the coherence and clarity of the summaries, the results were as follows: 81.1% valued them as high or very high, and 5.4% as low or very low. With respect to the ability of the system to avoid redundancies, evaluation was high or very high for 69.4% of the users, against 2.8% of low evaluation. At the same time, adaptation of the summary to the user profile was considered high by 59.5% of the users, and low or very low by 8.1%.

The degree of adaptation of the summaries to the information needs was high or very high in 70.3% of the cases, and low or very low in 10.8%. Regarding the extent to which the summaries reflect the content of the original documents, for 81.1% of the users this extent was high or very high, and it was low or very low for 5.4%.

Finally, 89.5% of the users consider that the main ingredients of the news item are represented in the summary. The other 10.5% indicated that at times the summaries were too brief to include them.

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 personalized summarization in the face of possible problems of clarity, coherence and redundancy.

On the other hand, users indicate that they sometimes used the summaries to establish the relevance of a news item. This was said to be often so by 48.6% of the users, sometimes by 29.7% and few by 21.6%.

Table 5
Averages normalized recall and precision for different types of summaries, in the second experiment

	N	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

Table 6
Averages normalized recall and precision for different types of summaries for first and second experiments

	First system		Second system	
	nP	nR	nP	nR
N	0.526	0.622	0.603	0.694
Ps	0.523	0.621	0.593	0.686
GPs	0.506	0.611	0.584	0.680
Gs	0.505	0.613	0.577	0.675
Fs	0.500	0.607	0.581	0.678

Against these data, 89.2% of the users relied on the heading often, and 10.8% only did in some cases. The section heading was used sometime by 45.9%, often by 29.7%, few by 13.5% and none by 10.8%. The stated relevance was used sometimes by 35.1% of the users, few by 24.3%, none by 21.6% and often by 18.9%. Finally, the full news item was used few times by 51.4% of the users, some times by 29.7% and none by 18.9%. In conclusion, the summary becomes an important element for defining the relevance of a news item.

7.3. Comparison of the two personalized systems

It can be observed (Table 6) that the results are better with the second personalization system. This is mainly because the use of the categories as classification system for the long term model and the effect of the short-term model during more days. In whatever case, the results for the second system confirm the results obtained in the first system in the sense that the personalized summaries are the best type of summary.

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

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

8. Conclusions

A summarization process has been presented characterized by the inclusion of a user model to generate personalized summaries adapted to the user. The user model contains different reference systems to represent the information needs of the user: sections, categories, keywords and feedback terms. The aim is to provide a summary oriented to the user that helps him to correctly identify whether a document is really interesting for him without having to read the full text.

The indirect evaluation method used to measure the quality of the summaries has allowed determining that the personalized summaries are the better option in terms of normalized precision and recall. Full news item offer only a slight improvement against personalized summaries, which seems to indicate that the loss of information for the user is very small with this type of summary. Generic summaries perform very closely to summaries obtained by taking the first few lines of the news item. This seems to indicate that the position method is overpowering the thematic word method, which may be corrected by refining the choice of weights. Although a first-sentences approach may provide good results for indicative summarization, it does not do so well in terms of personalized summarization, where it is crucial to retain in the summary those specific fragments of the text that relate to the user profile.

As it has been shown, automatic summarization has the choice of applying generic techniques designed to capture the way in which the more relevant information is distributed over a typical document of the domain under consideration, or to apply personalization techniques that concentrate on specific contents for which the user is known to have an interest (either because he has explicitly stated it in a long-term model or because it has been dynamically captured in a short-term model). The newspaper domain provides a good example where the structuring of information gives very useful cues as to their relevance (news items are usually built as inverted pyramids in terms of relevance: the most relevant information at the top, with relevance decreasing as they are read). This leads to quite simple summarization techniques providing good results in terms of indicative summarization – allowing the user to get an idea of what a document is about – though such a method would be severely domain-dependent, and might not work as well for different domains.

The methods proposed here ensure the selection of the relevant information in terms of information needs, operating efficiently for personalized summarization providing the user an extract of the specific contents of a document that are related to his interests, with no domain-dependent assumptions.

On the other hand, the user centred direct evaluation further sanctions the concept that offering users summaries of the news items helps to decrease information overload on the users. The fact the summaries are said to be employed by users much more often than the full original text or the stated relevance to determine how relevant a news item is to them justifies the use of automatic summaries in a personalization system. This evaluation has also shown that the possible problems of sentence extraction as a summary construction method do not affect performance in the present context of application.

We can conclude that user adapted summaries are a useful tool to assist users in a personalization system. Notwithstanding, the information in these summaries can not replace the full text document from an information retrieval point of view.

References

Amini, M. R., & Gallinari, P. (2002). The use of unlabeled data to improve supervised learning for text summaries. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, Tampere, Finland* (pp. 105–112).

Baldwin, B., Donaway, R., Hovy, E., Liddy, E., Mani, I., Marcu, D., et al. (2000). An evaluation road map for summarization research. technical report, DARPA's TIDES (Translingual Information Detection, Extraction, and Summarization) program.

Billsus, D., & Pazzani, M. J. (2000). User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction Journal, 10*(2–3), 147–180.

Brandow, R., Mitze, K., & Rau, L. F. (1995). Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing and Management, 31*(5), 675–685.

Callan, J. P. (1994). Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval, Amherst, MA*, (pp. 301–310).

Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y., & Lee, D. (1997). Translingual Information Retrieval: A Comparative Evaluation. *IJCAI*(1), 708–715.

Dang, H. T. (2005). Overview of DUC 2005. In Proceedings of DUC 2005 Document Understanding Workshop.

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

Díaz, A., & Gervás, P. (2004b). Item summarization in personalization of news delivery systems. Text, speech and dialogue. LNAI 3206 (pp. 49–56).

Díaz, A., Gervás, P., & García, A. (2005). Evaluation of a system for personalized summarization of web contents. In *Proceedings of the tenth international conference on user modeling, LNAI 3538*, (pp. 453–462).

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

Donaway, R. L., Drummey, K. W., & Mather, L. A. (2000). A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the workshop on automatic summarization at the 6th applied natural language processing conference and the 1st conference of the North American chapter of the Association for Computational Linguistics, Association for Computational Linguistics*, (pp. 69–78).

Dorr, B. J., Monz, C., President, S., Schwartz, R., & Zajic, D. (2005). A methodology for extrinsic evaluation of text summarization: does ROUGE correlate? In *Proceedings of the association for computational linguistics workshop on intrinsic and extrinsic evaluation measures for MT and/or summarization, Ann Arbor, MI*, (pp. 1–8). Association for Computational Linguistics.

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

Hahn, U., & Mani, I. (2000). The Challenges of Automatic Summarization. *Computer, 33*(11), 29–36.

Harman, D.K., & Marcu, D. (Eds.) (2001). In *Proceedings of the 1st document understanding conference, New Orleans, LA*.

Kupiec, J., Pedersen, O., & Chen, F. (1995). A trainable document summarizer. Research and Development in Information Retrieval, pp. 68–73.

Labrou, Y., & Finin, T. (2000). Yahoo! as an ontology: using yahoo! categories to describe documents. In *Proceedings of the 8th international conference on information knowledgment* (pp. 180–187). ACM Press.

Lin, C.-Y. & Hovy, E. H. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of 2003 language technology conference (HLT-NAACL 2003), Edmonton, Canada, May 27–June 1, 2003*.

Lin, C. (2004). Rouge: a package for automatic evaluation of summaries. In M.-F. Moens & S. Szpakowicz (Eds.), *Text summarization branches out: Proceedings of the ACL-04 workshop, Barcelona, Spain*, (pp. 74–81), July. Association for Computational Linguistics.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159–165.

Maña, M., Buenaga, M., & Gómez, J. M. (1999). Using and evaluating user directed summaries to improve information access. In *Proceedings of the third european conference on research and advanced technology for digital libraries* (pp. 198–214). LNCS 1696: Springer-Verlag.

Mani, I. (2001). *Automatic Summarization*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Mani, I. & Bloedorn, E. (1998). Machine learning of generic and user-focused summarization. In *Proceedings of the 15th national conference on artificial intelligence, Menlon Park, California*, (pp. 821–826).

Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., & Sundheim, B. (2002). SUMMAC: a text summarization evaluation. *Natural Language Engineering, 8*(1), 43–68.

Morris, J., Kasper, G., & Adams, D. (1992). The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research, 3*(1), 17–35.

Nanba, H., & Okumura, M. (2000). Producing more readable extracts by revising them. In *Proceedings of the 18th international conference on computational linguistics*, (pp. 1071–1075).

Nomoto, T. & Matsumoto, Y. (2001). A new approach to unsupervised text summarization. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, New Orleans, LA*.

Over, P., & Yen, J. (2004). An introduction to DUC 2004 intrinsic evaluation of generic new text summarization systems. In *Proceedings of DUC 2004 Document Understanding Workshop, Boston*.

Paice, C. D. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management, 26*(1), 171–186.

Paice, C., & Jones, P. A. (1993). The identification of important concepts in highly structured technical papers. In R. Korfhage, E. Rasmussen y & P. Willett (Eds.), *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval*, (pp. 69–78).

Radev, D.R., Jing, H., & Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the workshop on automatic summarization at the 6th applied natural language processing conference and the 1st conference of the North American Chapter of the Association for Computational Linguistics, Seattle, WA*.

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

Salton, G., Allan, J., Buckley, C., & Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts. *Science, 264*(3), 1421–1426.

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

Sanderson, M. (1998). Accurate user directed summarization from existing tools. In *Proceedings of the 7th international conference on information and knowledge management*, (pp. 45–51).

Sparck-Jones, K., & Galliers, J. R. (1996). *Evaluating natural language processing systems: an analysis and review lecture notes in artificial intelligence 1083*. New York: Springer.

Teufel, S., & Moens, M. (1997). Sentence extraction as a classification task. In *Proceedings of ACL/EACL workshop on intelligent scalable text summarization, Madrid, Spain*, (pp. 58–65).

Tombros, A., & Sanderson, M. (1998). Advantages of query-biased summaries in IR. In *Proceedings of the 21st ACM SIGIR conference*, (pp. 2–10).