Evaluation of a System for Personalized Summarization of Web Contents*

Alberto Díaz[1], Pablo Gervás[2], and Antonio García[3]

[1] CES Felipe II – Universidad Complutense de Madrid
adiaz@cesfelipesegundo.com
[2] SIP – Universidad Complutense de Madrid
pgervas@sip.ucm.es
[3] Departamento de Comunicación - Universidad Rey Juan Carlos
an.garcia@cct.urjc.es

Abstract. Existing Web personalized information systems typically send to the users the title and the first lines of the chosen items, and links to the full text. This is, in most cases, insufficient for a user to detect if the item is relevant or not. An interesting approach is to replace the first sentences by a personalized summary extracted according to a user profile that represents the information needs of the user. 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 system-oriented evaluation developed in this paper indicates that personalized summaries perform better than generic summaries in terms of identifying documents that satisfy user preferences. We also considered a user-centred qualitative evaluation indicating a high level of user satisfaction with the summarization method described, in consonance with the quantitative results.

1 Introduction

Web content personalization is a technique for reducing information overload through the adaptation of contents to each type of user. A Web personalization system is based on 3 main functionalities: content selection, user model adaptation, and content generation. For these functionalities to be carried out, they must be based on information related to the user that must be reflected in his user model or profile [8].

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 is necessary because user needs change over time as a result of his interaction with information [1]. For this reason the user model must be dynamic to adapt to those interest changes.

Content generation involves generating a new result web document that contains, for each selected document, some extract considered indicative of its content. Existing

* This research has been partially funded by the Ministerio de Ciencia y Tecnología (TIC2002-01961).

L. Ardissono, P. Brna, and A. Mitrovic (Eds.): UM 2005, LNAI 3538, pp. 453 – 462, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Web personalized information systems typically send to the users the title and the first lines of the chosen items, and links to the full text. This is in most cases insufficient for a user to detect if the item is relevant or not, forcing him to inspect the full text of the document. An interesting approach is to replace the first sentences sent as a sample of a document by a proper summary or extract.

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 document that would be useless if summarized in a generic manner may be useful if the right sentences are selected that match the user interest.

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.

In this paper we focus on a system-oriented and user-centred evaluation of the content generation (summarization) process. Section 2 describes previous work. The multi-tier selection process employed for evaluation is described in section 3. Section 4 describes the personalised summarization method. The experimental set up and results are given in section 5. Section 6 outlines the main conclusions.

2 Relevant Previous Work

Automatic summarization is the process through which the relevant information from one or several sources is identified in order to produce a briefer version intended for a particular user - or group of users - or a particular task [6]. This paper considers indicative summaries of single documents, intended to help the user to decide on the relevance of the original document. Summaries can be *generic*, if they gather the main topics of the document and they are addressed to a wide group of readers, or *user adapted*, if the summary is constructed according to the interests of the particular reader that the system is addressing.

Techniques for selection of phrases extract segments of text that contain the most significant information, selected based on linear combination of the weights resulting from the application of a set of heuristics applied to each of the units of extraction. These heuristics may be *position dependent*, if they take into account the position that each segment holds in the document; *linguistic*, if they look for certain patterns of significant expressions; or *statistical*, if they include frequencies of apparition of certain words. The summary results from concatenating the resulting segments of text in the order in which they appear in the original document [4].

There are similar works that use personalized summaries in information retrieval. In this case, the personalization is based on the user query [7, 11]. In particular, in [11] the initial segment of the documents is compared with query oriented summaries using a IR system. The results are shown to the users as title and initial segment or title and automatic summary. The evaluation was performed with 50 TREC queries with 50 documents per query. Measures were taken on precision, recall, speed in the

decision process, number of access to the full document and subjective opinion of the user about the received information (initial segment or summary). The results show that the query oriented summaries are significantly more effective that the initial segment for the information retrieval task.

Work on evaluation of item summarization has already shown that indirect evaluation methods of summarization - where summaries are evaluated in terms of their ability to recreate the ranking obtained by the full items when submitted to a given information selection process - provide reasonable means of measuring the amount of information loss involved in summarization. In particular, the selection process used in [7] was keyword-based single-tier over a corpus of 5000 news items and 50 queries from the TREC collection. Generic and personalized summarization heuristics are considered. The results show that the query oriented summaries are better than the first sentences and the generic summaries.

On the other side, existing literature provides different techniques for defining user interests: keywords, stereotypes, semantic networks, neural networks, etc. A particular set of proposals [1, 8] model users by combining long term and short term interests: the short term model represents the most recent user preferences and the long term model represents those expressed over a longer period of time. Various classification algorithms are available for carrying out content selection depending on the particular representation chosen for user models and documents. The feedback techniques needed to achieve a dynamic modeling of the user are based on feedback given by the user with respect to the information elements selected according to his profile. The information obtained in this way can be used to update accordingly the user models in representation had been chosen.

3 Multi-tier Content Selection

The multi-tier content selection process [2] to be employed in this paper 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 first tier of selection corresponds to a domain specific given classification (for digital newspapers, the assignment of news items to sections). 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 general 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 [5]. 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 [2]. The term weight vector for each user (t_u) represents the short-term interests of that user, information needs that loose interest to the user over time, so their weight must be progressively decreased.

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 [9].

Each document is assigned the weight associated with the corresponding specific category associated to it in the particular user model, which represents the similarity between a document d, belonging to a specific category c, and a user model u (s^c_{du}). The similarities between a document d and a general 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 [9]:

$$s_{dg} = sim(d_d, g) \qquad s^k_{du} = sim(d_d, k_u) \qquad s^t_{du} = sim(d_d, t_u) \tag{1}$$

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

$$s^g_{du} = \sum_{i=1}^{14} G_{iu} s_{dg_i} \Big/ \sum_{i=1}^{14} G_{iu} \tag{2}$$

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:

$$s_{du} = \frac{\delta s^c_{du} + \varepsilon s^g_{du} + \phi s^k_{du} + \gamma s^t_{du}}{\delta + \varepsilon + \phi + \gamma} \tag{3}$$

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

4 Applying Long and Short Term User Models to Personalize Summaries

Our system uses three phrase-selection heuristics to build summaries: two to construct generic summaries, and one for personalized summaries. To generate summaries a value is assigned to each phrase of the text being summarized, obtained as a weighted combination of the results of the three heuristics. This value is used to select the most relevant phrases, which will be used to form an extract of the news item later used as summary.

The *position heuristic* assigns the highest value to the first five phrases (1, 0.99, 0.98, 0.95, 0.9) of the text [3]. These provide the weights A_{pd} for each phrase p of a news item d using the position heuristic. These values are independent of the user u being considered.

Each text has a number of thematic words, which are representative of its content[1]. To obtain the M most significant words of each document, documents are indexed to

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

provide the weight of each word in each document using the *tf · idf* method [9]. The *thematic words heuristic* extracts the M non-stoplist most significant words of each text. To obtain the value for each phrase p within the document d using the thematic words heuristics (B_{pd}), the number of thematic words appearing in the phrase is divided by the total number of words in the phrase. This is intended to give more weight to sentences with a higher density of thematic words [10]. The values obtained in this way are also independent of the particular user u being considered. We have chosen M=8.

The *personalization heuristic* boosts those sentences that are more relevant to a particular user model. The user model provides a vector of weighted terms (k_u) corresponding to the chosen keywords of the long-term model and a vector of weighted terms (t_u) corresponding to the feedback keywords of the short-term model. This information is used to calculate the similarity (C_{pdu}) between the user model u and each phrase p of news item d, assigning the final weight to the sentence as:

$$C_{pdu} = \frac{\chi sim(p_{pd}, k_u) + \beta sim(p_{pd}, t_u)}{\chi + \beta} \tag{4}$$

where p_{pd} is the term weight vector representing the phrase p of news item d, and sim is the cosine formula of the Vector Space Model [9].

The values resulting from each of the three heuristics are combined into a single value (Z_{pdu}) for each phrase p of each news item d for each user u:

$$Z_{pdu} = \frac{\mu A_{pd} + \nu B_{pd} + \sigma C_{pdu}}{\mu + \nu + \sigma} \tag{5}$$

The parameters μ, ν and σ allow relative fine-tuning of the different heuristics, depending on whether position (μ), thematic key words (ν) or similarity to the user model (σ) is considered more desirable. Values of σ determine the degree of personalization of the summaries: if σ is 0, the resulting summaries are generic, and for σ greater than 0 personalization increases proportionally to σ. Again, to ensure significance, the relevance obtained for each framework must be normalized.

The summary is constructed by selecting the top 20% of the ranking of sentences by the value Z_{pdu} and concatenating them according to their original order of appearance in the document.

5 Evaluation

We have performed two kinds of evaluations. System-oriented evaluation is based on the precision and recall metrics obtained through different configurations of the system, and intends to identify which is the best way of carrying the content generation process through the effect in the selection process. User-centred evaluation collects the opinions of the users about the use of summaries instead of the complete news items.

5.1 System-Oriented Evaluation

Experiments are evaluated over data collected for 106 users 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 [2]. The set of users includes 18 lecturers, 4 teachers, 77 students and 7 professionals from no education areas. The students come from the fields of computer science, journalism and advertising. The average of news item per day is 78.5.

To carry out the system-oriented evaluation, judgments from the user are required as to which news items are relevant or not for each of the days of the experiment. To obtain these judgments users were requested to check the complete set of news items for each day, stating for each one whether it was considered interesting (positive feedback) or not interesting (negative feedback).

As the evaluation process involved an effort for the users, only 37.4 users per day actually provided judgments. Additionally, some users only perform feedback for less than 10 news items per day. These users have been eliminated for the evaluation in order to obtain more significant results. The final collection contains, on average, 28.6 user per day.

For evaluating summarization, the effect of selection (formula (3) with $\delta=\varepsilon=\phi=\gamma=1$) 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.

Normalized recall and precision are used as evaluation metrics, given the users binary relevance judgments are compared against the ranking provided by the system [9]. These metrics measure the difference between an ideal ranking, with the relevant documents at the top, and the actual ranking provided by the system. On the other hand, the recall and precision metrics are computed with respect a selected fixed number of documents and they don't use the information about the ranking.

Data are considered statistically significant if they pass the *sign-test*, with paired samples, at a level of significance of 5% ($p \leq 0.05$) [9].

5.1.1 Experiment 1. Personalized Summaries

The generation of personalized summaries (formula (5) with $\mu=\nu=0$ y $\sigma=1$) combines the long-term model (keywords provided by the user) and short-term model (feedback terms obtained from the interaction with the user).

Several evaluation collections have been generated for each user. Each one of them is obtained by summarizing the complete set of original news items according to a particular method for generating personalized summaries of those indicated above (formula (4)). There is a collection for each user of personalized summaries generated using the short term model (Ps(S): $\chi=0,\beta=1$), a different collection for each user generated using the long term model (Ps(L): $\chi=1,\beta=0$) and a third different collection for each user generated using a combination of long term and short term models (Ps(LS): $\chi=1,\beta=1$). In each case, values of normalized recall and precision have been computed. These experiments have been repeated for all users during the 14 days of evaluation. The results for the three types of personalized summaries have been compared only from the second day on, to allow for the fact that on the first day there is no short-term model based on user feedback.

If different summarization methods lead to different degrees of loss of relevant information, the resulting rankings will differ amongst them in a proportional way. The results shown in Table 1 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 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.

The use of both heuristics 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.

Table 1. Normalized precision (P) and recall (R) for different combinations of long and short-term model for generating personalized summaries

	P	R
Ps(LS)	0.592	0.684
Ps(S)	0.583	0.678
Ps(L)	0.576	0.674

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.

5.1.2 Experiment 2. Heuristic Combination for Summary Generation

Experiment 2 tests whether summaries obtained by using only the personalization heuristic are better in terms of precision (formula (3) with $\delta=\varepsilon=\phi=\gamma=1$) with respect to information selected by the user than other summaries (including the first lines of the document) but worse than the complete news item.

The following types of summaries are involved (formula (5) with (4) with $\chi=\beta=1$): Fs (baseline reference), 20% first phrases of the corresponding news item; Gs, using generic heuristics ($\mu = 1$, $\nu = 1$, $\sigma = 0$); Ps, using personalization heuristics ($\mu = 0$, $\nu = 0$, $\sigma = 1$); GPs, using both types of heuristics ($\mu = 1$, $\nu = 1$, $\sigma = 1$).

Several different evaluation collections – consisting each one of summaries obtained from the news items in the original collection by applying a different summarization method – are built for each user. 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, the values of normalized recall and precision have been computed in experiments that have been repeated over the 14 days for all users.

Table 2. Normalized precision (P) and recall (R) for news item (N), personalized (Ps), generic-personalized (GPs), generic (Gs) and first phrases (Fs) summaries

	N	Ps	GPs	Fs	Gs
P	0.603	0.593	0.584	0.581	0.577
R	0.694	0.686	0.680	0.678	0.675

Personalized summaries (Ps) offer better results (table 2) with respect to normalized precision and recall than generic-personalized summaries (GPs), though the difference is not significant. With respect to baseline summaries (Fs) and generic summaries (Gs) the difference is significant. Generic-personalized summaries (GPs) are better than baseline summaries (Fs), and baseline summaries (Fs) are better then generic summaries (Gs), but the differences involved are not statistically significant. Personalized summaries are worse than full news items (N) under the same criteria.

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

This technique has been used in similar works with similar results. In [7] the query oriented summaries (title, location, thematic and query heuristics) obtained significant better average precision than generic summaries and first sentences, and the full document improve the adapted summaries but no significantly. In [11] the query oriented summaries show better effectiveness that the initial segment.

5.2 User-Centred Evaluation

The qualitative user-centred 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 content generation in the face of possible problems of clarity, coherence and redundancy.

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%. 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.

6 Conclusions

We can conclude that personalized summaries that use a combination of long and short term models are better than other types of summaries 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 heuristic is overpowering the thematic word heuristic, 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. This explains why the generic-personalized summaries perform so poorly in spite of being a combination of good techniques: given a fixed limit on summary length, the inclusion of sentences selected by the generic heuristics in most cases pushes out of the final summary information that would have been useful from the point of view of personalization.

The user centred evaluation further sanctions the concept that offering users summaries of the news items helps to decrease information overload on the users. As shown in these results, the possible problems of sentence extraction as a summary construction method do not affect performance in the present context of application. 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 content generation method described in this paper.

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

1. Billsus, D. & Pazzani. M.J.: User Modeling for Adaptive News Access. User Modeling and User-Adapted Interaction Journal 10(2-3) (2000) 147-180
2. Díaz, A. & Gervás, P.: Adaptive User Modeling for Personalization of Web Contents. Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2004). LNCS 3137. Springer-Verlag (2004) 65-75
3. Edmundson, H.: New methods in automatic abstracting. Journal of the ACM 2(16) (1969) 264–285

4. Kupiec, J., Pedersen, O., Chen, F.: A trainable document summarizer. Research and Development in Information Retrieval (1995) 68–73
5. Labrou, Y. & Finin, T.: Yahoo! As an Ontology: Using Yahoo! Categories to Describe Documents. Proceedings of the 8th International Conference on Information Knowledgement (CIKM-99). ACM Press (2000) 180-187
6. Mani, I. & Maybury, M.: Advances in Automatic Text Summarization. The MIT Press (1999)
7. Maña, M., Buenaga, M., Gómez, J.M.: Using and evaluating user directed summaries to improve information access. Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL1999). LNCS 1696. Springer-Verlag (1999) 198–214
8. Mizarro, S. & Tasso, C.: Ephemeral and Persistent Personalization in Adaptive Information Access to Scholarly Publications on the Web. Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2002). LNCS 2347. Springer-Verlag (2002) 306-316
9. Salton, G.: Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley Publishing (1989)
10. Teufel, S. & Moens, M.: Sentence extraction as a classification task. Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization. Madrid, Spain (1997) 58–65
11. Tombros, A. & Sanderson, M.: Advantages of query-biased Summaries in IR. Proceedings of the 21st ACM SIGIR Conference (1998) 2-10