# Modelling users' interests and needs for an Adaptive On-line Information System*

Enrique Alfonseca and Pilar Rodríguez

Computer Science Department, Universidad Autonoma de Madrid,
Carretera de Colmenar Viejo, km. 14,5,

## 2 User modelling based on interests and available time

The aim of the system is the transformation of a static web site into an adaptive site that shows different kinds of information to different users. To do that, the contents of the web site will be modified, but the link structure will be kept the same as much as possible. There are two kinds of transformations that are performed to the textual contents: a filtering of the information that is deemed irrelevant, and a summarisation of the remaining information, according to the instructions provided by the user. To achieve that aim, two different facts will be modelled in the users' profile: (a) their personal interests, and (b) the amount of information that they are willing to read.

### 2.1 User interests

A single hypermedia site may contain information about different topics. For example, one of the web sites that we have processed with our system dealt with Charles Darwin's *The Voyages of the Beagle*. It is a multi-disciplinary document in which he describes both animals and places, and the adventures he lived. It seems natural to pre-define three possible user interests for this text: biology, about animals and plants; geography, which includes all the descriptions of places; and history, which includes the fragments in which Darwin explains his adventures and other events. Other web sites have different topics predefined.

The system allows the definition of predefined interests or *stereotypes*. A user might be interested in one or several of these, or in some particular topic that had not been thought of beforehand by the designer of the web site. Therefore, it is necessary that the way in which the user's interests are encoded allows the definition of new possible topics. This section describes the design decisions that were taken in order to tailor the page contents to the user interests. The process performed consists in eliminating all the information that is considered irrelevant for each particular user, a process that can be called *topic filtering*.

In order to train the topic classifier we used the vector model, which has been used very often in Information Retrieval (IR) applications. One hundred paragraphs from each web site were classified by hand with one of the pre-defined label. Next, all the open-class words (nouns, verbs, adjectives and adverbs) appearing in those paragraphs were collected, together with their frequencies of appearance. Finally, for each stereotype, the frequencies were changed into weights indicating the support that a word gives to the decision of classifying a paragraph in each stereotype. The list of words and weights is called a *topic signature*. In our case, we weighed the words with the $\chi^2$ metric [1], which seemed to give better results than other standard metrics such as tf·idf.

### 2.2 Topic filtering

As said above, when a new user registers into the system there are two ways in which his or her interests can be specified. It is possible to select one or several stereotypes, or to ask for a user-defined profile. In the first case, the pre-defined

topic signatures are copied into the new profile, and they are marked either as relevant or irrelevant. In the second case, the system will start by showing the users all the available information, and they can press small icons in the page to indicate whether a paragraph was interesting or not, so their models are updated as they browse.

While browsing a web site, for each paragraph in a hypertext page, its relevancy is calculated in the following way:

1. The paragraph's signature is formed by collecting all the open-class words.
2. The dot product is performed between that signature and each of the topic signatures in the user's profile. The metrics are normalised so they sum 1.
3. The paragraph is chosen if, for any of the topics that the user has selected as relevant, the similarity metric is above $\frac{1}{N}$, where $N$ is the number of topics.

Note that, with this algorithm, the same paragraph may be considered relevant for different stereotypes, if it contains keywords relevant to more than one topic. The classification algorithm was evaluated with a test set that contained 58 additional paragraphs, labelled by hand. 91.38% of them were chosen as relevant for the right stereotypes.

When the user is navigating the web site, for every paragraph in the text there will be two small icons, that can be clicked to inform the system that the information in that paragraph is either very relevant or irrelevant. This information will be used to update the frequency counts in the topic signatures in the user's profile, in order to provide better information in the future.

Finally, when a user registers, all the pages in the web site are analysed and those which do not contain any relevant information are marked in the user's profile. When browsing the web site, the links to those pages will be removed.

### 2.3 Available time and reading speed

The second aim of our adaptive site is to provide the quantity of information that the users are willing to read. A user can request different compression rates to the system by indicating the compression rate to be applied; the total number of words that the target site should contain; or the amount of time available.

The first situation is the simplest one, when the compression rate is directly provided by the user; in the second case, the rate is calculated by dividing the total number of words that the user is willing to read by the number of words in the relevant portions of the web site. The last case is somewhat more complicated. The users provide their *availability of time* as a number of hours that they intend to browse the site. Next, the users' *reading efficiency* is collected with a short test, in which the site server asks the user to read a small passage of text ($\sim$3 minutes) and fill in a questionnaire. The *reading efficiency* is calculated as the product of the reading speed and comprehension [5], measured in words per minute. Finally, if the user interests change while browsing, the compression rate is automatically re-calculated.

The summarisation is performed with a common sentence extraction procedure [3], taking into consideration the positions of the sentences in the document, their length, and their relevancy according to the user's profile.

**Table 1.** Answers of the users. Each question was evaluated from 1 to 5 (very low, low, medium, high and very high).

| Question | Mean | Std dev. |
|---|---|---|
| Easiness to register in the system | 4 | 2.83 |
| Did any of the stereotypes equalled your personal interests? | 3.8 | 2.76 |
| Did you find useful the possibility of creating new stereotypes? | 4.9 | 0.95 |
| The length of the summaries is appropriate | 4.22 | 1.89 |
| The summaries are coherent and easy to read | 3.5 | 2.55 |
| What is your overall opinion about the system? | 4.2 | 1.26 |

## 3 Evaluation and results

The evaluation has been performed with a controlled experiment, in which twelve people have used the system and provided comments. Their backgrounds are somewhat different: there were one linguist, three electrical engineers, one industrial engineer, two mathematicians and five computer scientists. The different options of the system were taught to them, and the profile of each person was collected. In general, they all had much experience using a web browser, but only some of them knew what adaptive hypermedia is; fluency reading English was very variable, as some had a high proficiency, and others could only read around 60 words per minute. Finally, few of them knew about the texts used in the experiments.

The evaluation centred on the usability of the new tool. All the users spent some time using the system, and when they felt that they had explored the different possibilities, they filled a questionnaire. The questions that concern the user model are listed in Table 1. The users considered most features of the system as highly or very highly usable (the mean score was between 4 and 5). The ones that received the lowest weights were the interest of the pre-defined stereotypes, and the coherence of the generated summaries, a fact that could be expected, as extracts are sometimes incoherent.

## 4 Related work

There are several automatic algorithms to partition full-length expository text into a sequence of subtopical discussions [4], but it is also possible to partition the text in paragraphs, and to classify the paragraphs separately, considering that it is not frequent to find a topic shift in the middle of a paragraph. The problem of finding relevant topics in a text is not a new one; [6] start with a pre-defined set of topics and lists of words that are relevant for each of them, and use the lists to classify the text fragments. [7] described a supervised algorithm for which there is, initially, a set of text fragments, each one labelled with its topic, and the algorithm learns the lexical items that are useful for classifying new texts. Finally, the topics themselves can be automatically induced from the set of texts, e.g. by clustering them, or using a conceptual ontology.

Concerning adaptive hypermedia systems, there are several approaches to adapting a web site to the needs of a user using natural language processing techniques, such as those described by [8] and [9].

## 5 Conclusions and future work

A new approach has been implemented for automatically adapting a complete web site to the interests of a user. The users can specify their needs by choosing from a list of pre-defined interests, or by selecting the paragraphs that are of interest to them. Secondly, by indicating a compression rate to be performed to the pages, their contents are summarised using a sentence extraction procedure, which also takes into account the user's interests. This technique performs well for web sites that contain large amounts of text, such as those that have been generated from linear text. In this case, the system selects just the paragraphs that meet the user's interests; if a page is left empty, it is removed from the user's hyperspace, until a change in the user's profile makes it relevant again.

When used by some potential users, the system had a good acceptance. Future work includes improving the user-friendliness of the system with the suggestions gotten from the users that participated in the evaluation; and making it easier to generate new user-defined stereotypes.

## References


1. E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching very large ontologies using the www. In *Ontology Learning Workshop, ECAI*, Berlin, Germany, 2000.
2. E. Alfonseca and P. Rodríguez. Automatically generating hypermedia documents depending on user goals. In *Workshop on Doc. Compression and Synthesis, AH-2002*, 2002.
3. H. P. Edmundson. New methods in automatic abstracting. *Journal of the Association for Computational Machinery*, 16(2):264–286, 1969.
4. M. A. Hearst. Texttiling: A quantitative approach to discourse segmentation, 1993.
5. M. D. Jackson and J. L. McClelland. Processing determinants of reading speed. *Journal of experimental psychology*, 108:151–181, 1979.
6. P. Jacobs and L. Rau. Scisor: Extracting information from on-line news. *Communications of the ACM*, 33(11):88–97, 1990.
7. B. Masand, G. Linoff, and D. Waltz. Classifying news stories using memory based reasoning. In *Proceedings of SIGIR 92*, pages 59–65, 1992.
8. M. Milosavljevic, R. Dale, S. J. Green, C. Paris, and S. Williams. Virtual museums on the information superhighway: Prospects and potholes. In *CIDOC'98*, 1998.
9. J. Oberlander, M. O'Donell, C. Mellish, and A. Knott. Conversation in the museum: experiments in dynamic hypermedia with the intelligent labeling explorer. *The new review of multimedia and hypermedia*, 4:11–32, 1998.
10. H. Wu and P. de Bra. Link-independent navigation support in web-based adaptive hypermedia. In *Proceedings of the 11th International WWW Conference*, 2002.