Sections, Categories and Keywords as Interest Specification Tools for Personalised News Services[1]

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

[1] Departmento de Inteligencia Artificial, Universidad Europea de Madrid – CEES, 28670 Villaviciosa de Odón, Madrid, Spain
alberto@dinar.esi.uem.es
[2] Departamento de Sistemas Informáticos y Programación, Universidad Complutense de Madrid, Ciudad Universitaria, 28040 Madrid, Spain
pgervas@sip.ucm.es
[3] Departamento de Periodismo Especializado, Universidad Europea de Madrid – CEES, 28670 Villaviciosa de Odón, Madrid, Spain
{antonio.garcia,inmaculada.chacon}@fcp.cin.uem.es

Abstract. Through a evaluation of system performance and user satisfaction for the Mercurio system - a system that sends personalised news selections via email -, the general applicability and usefulness of different methods of specifying user interest (sections, categories and keywords) are considered for the general case of digital news services. The specific characteristics distinguishing such systems from more general information systems are outlined and their effect is discussed. An evaluation blueprint for them is proposed starting from information retrieval procedures, existing work on search engine evaluation, and a close study of the working principles and the required evaluation according to the particular properties and conditions of the services under consideration. Actual evaluation results for system tests based both on real users and custom tailored test cases are presented and discussed. Conclusions cover the nature of the information handling tasks that digital news services are faced with, the relative merits of sections, categories, and key words with respect to this particular set of tasks, and the risks of careless application of recall and precision measures in systems such as these.

Introduction

The recent boom in the popularity of the Internet has resulted in a rapid expansion of the range of paradigms of information services available to the common user. One such paradigm is that of systems offering to send users a selection of the daily news by electronic mail. These systems are currently classed as information filtering applications, yet differ from more general information filtering or information retrieval systems in that their contents are completely renewed at regular intervals - daily in the case of most digital newspapers - and remain stable during intervening periods. This simple difference has considerable implications on the general characteristics of the databases that house these contents: they are perishable, throughout their lifetime they remain static, they are usually small, and they hold no claim to universal coverage of any information domain. Although existing systems of this sort are applying well-known techniques from the fields of information retrieval and information filtering, these particular restrictions governing their operation may affect the general applicability and usefulness of different methods of information access. Furthermore, it is at present unclear to what extent generally accepted evaluation measures for information retrieval, such as recall and precision, can be meaningfully applied in these circumstances.

In order to explore these issues we have carried out thorough evaluations of a system that incorporates these basic methods (newspaper sections, categories and keywords) into the process of selecting the particular information items that are relevant to a given

[1] The research work presented in this paper was partially funded by grant REF TS203/1999 of the ATYCA program, Ministerio de Industria y Energía, Spain

user. The Mercurio system (Díaz et al., 2000) allows readers of the newspaper to receive a periodic e-mail message containing the news items that the system finds particularly relevant to the interests of the user, previously defined during registration.

The system constitutes an integrated and customisable option of combining document relevance information from sources of three different kinds:
- A prior categorisation by the system owner (documents, i.e. news are sorted into sections by the editor)
- Keyword search information over the content of the documents
- Automatic categorisation results against an alternative (less domain specific) set of categories

An additional control layer is provided that allows users to specify how important each of the different methods of selection is to his particular interests. As such, it presents a flexible, multidimensional and browsable user model, and a set of well founded techniques for user models and information items matching. These features make it an ideal test bed for the questions described above.

Evaluation is carried out in two distinct stages. During the initial stage, evaluation of the system is carried out according to standard practice in the field of information retrieval. The results fall specially short with respect to determining relative performance of the different methods of information access involved. The second stage of the evaluation includes additional sets of experiments specially designed to shed light on this issue.

Evaluation of Systems for Information Access

At present, Internet is characterised by a proliferation of information systems. These systems present a multitude of innovations, both regarding the nature of the information they deal with and the actual form of service they provide. From many of these, new ways of understanding information services and information systems are arising. In this paper we are concerned with the type of services that send periodic news selections to subscribers of a digital newspaper by means of electronic mail. These systems present certain variations, however there is a set of characteristics that are common to them all:

a) Personalisation is a key factor, specially if carried out with respect to a user profile or user model employed to specify the information that is desired and to avoid unnecessary processing of unwanted documents or services. These profiles or models are usually based on the possibility of selecting newspaper sections. Although user comments show that additional features allowing use of keywords or categories are an important success factor, most systems still do not have any means of categorisation other than newspaper sections. On the other hand, personalisation also plays a role with respect to the format of the message sent (for instance, having the message headed with the reader's name, as a personalised newspaper), the choice of the days of the week in which the message should be sent, the presentation of the results, or the possibility of specifying either a number of news items or threshold relevance values to shape the personalised message.
b) Most systems pay special attention to ease of access and manipulation by the users. This includes speedy transmission, ease of subscription and un-subscription. The working context always remains close to the digital newspaper from which the system springs, allowing the user to consult, for instance, the database of back issues. The presentation schema - headline, abstract, and text, together with a

relevance value with respect to the user profile - rates the highest in terms of user satisfaction, and yet it is not the most frequent.

c) Information agents tend to present a friendly interface (fonts, colours, etc), specially where it concerns the design of each message. In most cases, they also include some means of contacting the service provider, a facility that improves efficiency and competitiveness

The lack of uniformity in these services shows up more easily in those that specialise on graphical information (photographs). These systems place more emphasis on design and commercial aspects, and give less importance to sending documents by electronic mail.

Evaluation of these new instruments of information retrieval requires: a validation of traditional evaluation measures within the new field of Internet, consideration of the knowledge acquired during evaluation of search engines, and a close study of the working principles and the required evaluation according to the particular properties and conditions of the services under consideration. With respect to search engine evaluation, the criteria that have been employed for qualitative evaluation to the present time are (Maldonado y Fernández, 1998):

- Coverage: evaluators are concerned with the number of web pages that the search engine has access to. Other relevant parameters are: geographical and content scope of the database, harvesting models, specific processing of the web documents and the possibility of accessing different types of information and resources.
- Search forms: providing the possibilities of searching at different levels. At present, their versatility and different possibilities are being considered.
- Search fields: the existence of different fields that can be used to guide user searches must be considered. Possible examples are: title, description, URL, keywords, language, information type, owner type, etc.
- Search tools: there are several possible instruments for search and retrieval of information, like categories, key words, stemming, boolean operators, locating compound terms, searching by phrases, proximity, searching by fields, restricting to certain dates, to certain domains, to certain languages, to certain file types, the ability to recognise meta-information, etc.
- Thematic classification and vocabulary control: existence of some way of structuring information, such as categories or other forms of control over vocabulary, as well as their applicability with respect to different types of information.
- Detection of novelties: systems should be able to identify and locate registers newly incorporated to the database, by means of special labels, delimitation and organisation of documents by date of inclusion, etc.
- Shaping results: users should be allowed to chose or define the format of presentation and the criteria used to determine the order of presentation of the results.

Reviewing quantitative research carried out over Internet search engines, it becomes apparent that there is no unified method of procedure. Further work is required on: the influence on results of the particular methods used for harvesting and compiling data, the varying nature of search engines - and the dynamic character of the database they are operating on -; the problems presented by classical instruments of measurement; the need for new evaluation measures, etc. Following (Olvera Lobo, 2000), and having

reviewed existing literature on this topic (Dong and Su, 1997; Gordon and Pathak, 1999; Schwartz, 1998; Leighton and Srivastava, 1999; Clarke and Willet, 1997) the different mechanisms, methods and trials carried out so far all agree on the significance of the following phases in an evaluation:

a) **Determination and subsequent formulation of the information needs of the users.** After an initial stage in which questions about evaluation were provided by the researchers themselves - with an obvious risk of partiality - a new trend imposes the collection of questions posed by 'real users of information'. This initiative faces the limitations imposed by laboratory research and an additional problem of loss of information in the process of translating the informative interests of the users. In as much as it is a starting point, the questions must obey the following ideas: previous knowledge of the existence of specific resources on the Web; combining different levels of difficulty and aims; combining different degrees of technical coverage and question specificity; and delimitation of the number of questions selected according to the aim of the endeavour.
b) **Studying the syntax of the query**. This factor (use of logical operators, parenthesis, etc.) acquires a certain importance. Even so, the achievement of correct results is dependent on an adequate selection of the terms employed in the query. One must determine how many and which particular terms are used, and specify whether boolean logic or natural language is employed to structure the query. Other important factors are: use of upper case letters, stemming, etc. One should take into account throughout that the best results will probably be obtained by posing queries with a relatively simple syntax.
c) **Monitoring the timing of the searches**. Given the dynamic character of the Web and search engines in general, it is convenient to carry out all searches simultaneously, running the same query on each search engine. This is because any delay may result in changes in the set of documents available, and thereby, in changes in the results.
d) **Specifying the set of relevance judgements to be assigned**. To study the effectiveness of the system, the relevance of each retrieved document must be stated. This represents a problematic aspect that may be resolved by resorting to external sources, such as TREC, etc. As a first stage, four levels can be distinguished: a) duplicated, inactive or irrelevant links; b) links that are relevant from a technical point of view; c) potentially useful links; d) links that are probably most useful.
e) **Selecting the means for the analysis of the results**. Recall and precision measurements remain useful depending on the number of items considered (different studies impose different thresholds). Nonetheless, recall has presented problems due to the difficulty of ascertaining the total number of relevant documents with respect to a specific query. This problem has been tackled either by applying the relevance values to a restricted controlled subset of a collection of documents, or by carrying out several different searches to obtain a set of relevant documents that is taken as the total set.

An extrapolation of the evaluation methodology applied to search engines is proposed as a starting point in the effort to sketch an evaluation method for services providing news selections via electronic mail. This extrapolation results in the following considerations:

a) Qualitative criteria:

- Coverage: this issue is taken into account in most existing evaluation frameworks. In the case under consideration, since a specific newspaper is taken as the subject of research, the discussion will be centred on the algorithms used for retrieval, categorisation, and learning and their effect on the accessibility of available documents.
- Search forms and fields: the analysis is oriented to the observation of how users understand the working and the aim of the forms.
- Search tools: the evaluation concentrates in the creation of user profiles and their usefulness. Thematic categories, key words, and sections are employed. For this reason, parameters deemed relevant in this respect are: the adequacy of the number of categories presented, whether the profile reflects the way in which the user specifies the information that he desires, whether categories are well organised, whether there is any overlap between them, whether there are enough of them, whether available means of representing relevance are adequate, etc.
- Detection of novelties: given the working context - set of news available in a given digital newspaper on a specific day – this parameter plays no role in the general evaluation.
- Shaping results: we are referring to parameters related with giving the user the possibility of specifying the format of presentation and the ordering criteria. For instance, having results ordered according to their relevance values with respect to the user profile in a given system.

b) Quantitative criteria:

- Determining user needs: an adequate number of users is selected, taking special care to ensure that they cover a wide range of possible dispositions – in terms of computer literacy and familiarity with Internet applications. At the same time, the evaluation exploits the information available in terms of different user profiles registered during testing, which yields important insights.
- Query syntax: does not apply in this case, since no actual query is provided by the users. Nonetheless, it is relevant to study different user behaviours with respect to profile creation and their relationship with final results.
- Search timing: unlike general purpose search engines, the services under discussion behave statically from one day's edition to the edition of the following day. Therefore no special attention need be paid to the timing of the application of the user model to the newspaper contents on each particular day. It is interesting, however, to carry out tests on different days
- Relevance judgements: in some cases, relevance judgements for a given document depend on feedback provided by the users themselves, and in other cases they are based on diagnose by the researchers.
- Analysis of results: precision and recall measurements are retained. In this case there are no problems with the recall measurement because it is possible to determine the total number of relevant news items available on the newspaper site on a given day.

Mercurio: A Personalization System for Digital News Services

Mercurio (Díaz et al., 2000) is a personalised news service system that applies existing techniques from the field of text classification, text categorization (Sebastiani, 1999) and information retrieval (Baeza-Yates and Ribeiro-Neto, 1999), besides user modelling (Gervás et al., 1999; Amato and Straccia, 1999), to the selection of items relevant for the user. Each user can create a profile with his preferences and receive daily the news items that interest him.

A user accesses the information server and registers for the service. During registration essential data are noted (email address, login, password). A profile for the user (or user model) is built, containing information such as the days of the week in which he wants to receive news, the number of items he wants to receive each time, and the user's interests. These interests are represented with respect to three systems of reference: the sections of the newspaper, a set of categories presented as an alternative system of classification (the first level of categories from Yahoo Spain), and terms chosen by the user as interesting. There are 8 sections in the chosen newspaper[II]: opinion, national, international, economy, society, culture, sports and people. The first level of categories from Yahoo Spain consists of the following 14 categories: Arts & Humanities, Science, Social Science, Recreation & Sports, Business & Economy, Education, Entertainment, Computers & Internet, Reference, News & Media, Government, Health, Society & Culture, and Regional. The system imposes no restrictions on the set of keywords that the user may choose.

Too many methods of selection available simultaneously can lead to confusion. Unless additional control features are provided, users get at most a blurred picture of the operation of the system. For this reason, our personalisation architecture allows an extra level of user specification. A general control mechanism has been included to make the results more predictable for the user. Each of the three features (keywords, sections and additional categories) has a weight that represents its importance for the user interests. For example, if the weight of sections is low and the weight of additional categories is high, relevance values concerning additional categories will be considered more important for selecting news items. In this way, each of the three dimensions considered in the user profiles can be defined and controlled by the user, providing a fine-tuning mechanism to obtain a flexible characterisation of his interests.

The message received by the user contains: the name of the user, the date, and a list of news ranked according to user information interests and according to the upper bound defined. Each news item is presented with a title, a short summary, the relevance, and a link to the news item in the digital newspaper. At the end of the message appear the interests of the user as features in his profile in order to allow the user to check the true relevance of the received news.

The representation of the news items is obtained applying the Vector Space Model to their texts (Salton, 1989). A representation for each category can be obtained by applying text categorisation techniques (Gómez and Buenaga, 1997; Lewis et al., 1996) and using a set of training documents (set of documents labelled manually with the suitable categories). In our case, the set of training documents used were the web pages indexed by the Spanish version of Yahoo! within these categories. Thus, each category can be represented by a term weight vector that is obtained from the name of the

[II] ABC: http://www.abc.es/

category, the name of its subcategories, and the names and short descriptions of the web pages associated to the category. The keywords also are represented with VSM, using the weight assigned for each word in the model.

To perform the selection we applied category-pivoted categorization (Sebastiani, 1999; Yang, 1999) with the categories and Information Retrieval (Baeza-Yates and Ribeiro-Nieto, 1999) with all the keywords. Also all the news are processed to check if they belong to one of the sections selected in the user model. When all the documents have been sorted according to the different sources of relevance, the resulting orderings are integrated by using the level of interest that the user assigned to each of the different reference systems. In order to make the relevance values provided to the user easy to interpret, they are normalised over the number of selection methods involved in obtaining them. This allows the system to quote a final relevance value in the range 0-100% to every user regardless of the number of selection methods that each particular user chose.

Initial Evaluation

We describe and discuss the three kinds of evaluation that were carried out: an evaluation carried out by a set of different users, a system evaluation that considers the performance of the system in measurable parameters, and an evaluation of the user model provided and how the evaluators have fared in dealing with it.

A controlled evaluation environment was established to allow analysis of the results with respect to the different kinds of user involved. Evaluation was carried out by 44 users in four categories: A) Collaborators; B) Researchers; C) University lecturers (both on Computer Science and Journalism); D) External users with no professional relationship with the fields involved.

The system was evaluated by the users following a working pattern previously developed for the analysis of existing systems (García et al., 2000; Díaz et al., 2000; Pastor and Asensi, 1999). For the relevance of the received documents the users had to check the performance of the system against the actual set of documents available on the newspaper website on three particular days. Additionally, on those particular days, logs of system operation (available documents, user profiles at the time, and system selections for each user) were kept to allow objective results to be obtained. With this data we worked out two sets of recall and precision figures: one based on user criteria as put down in the forms, and one based on subsequent close analysis of system logs.

User-centred Evaluation

During the first stage, the evaluation was centred on user response and the vision that users develop of the system. The aim was to harvest explicit evaluations provided by the users about system response-time, ease of use, system efficiency, and conceptual and physical presentation. This information was compiled on the basis of a closed questionnaire with specific questions on the relevant main topics. For each of these parameters a numerical value was worked out from the users responses.

In general, users found the system suitable although the results showed some differences between different groups of users. These were the results for the interface evaluation: System Access: (high); General Interface, User Adaptation, and Integration into User Environment: (medium-high); Management of Contents, Query and Retrieval Schemes and User Help (medium). With respect to newspaper sections the following

results were obtained: Expressive Faithfulness, Objectivity and Relevance (high). With respect to categories the following results were obtained: Expressive Faithfulness and Objectivity (medium-high); Relevance (medium). With respect to summaries the following results were obtained: Summary Content (high); Summary Structure (medium-high).

Recall and Precision rates have been estimated based on user impressions (see Table I), under the assumption that the aim is not to obtain conclusive results but to draw roughly significant conclusions.

Group	Precision	Recall
A &B	0.9	0.8
C	0.9	0.6
D	0.9	0.6
Average	0.9	0.7

Table I. User estimated Recall and Precision (by groups)

Additionally, the qualitative analysis showed that users were satisfied with the system characteristics, personalisation quality, formal quality, and categories system. On the other hand, the users' familiarity with similar systems influenced their understanding of the basic mechanisms. Some users found it could be more visual, but most of them understood it after receiving the first message.

Evaluation of Observed User Profiles

The analysis of the 44 user models logged with the system yields the following data (see Table II).

	Upper bound	Selection methods	Sections	Categories	Keywords
Average	14	1,9	2,6	3,4	2,3
Max	20	3	9	14	15
Min	5	1	0	0	0
Selected values	44	44	30	26	18
Selected average	14	1,9	3,9	5,8	5,7

Table II. Analysis of user profile development

The average selection of a user has approximately 14 as upper bound of documents per message, 2 methods of selection (in most cases, sections and categories), 3 sections, 3 categories and 2 keywords.

All the users selected the sections method, with or without other methods of selection, except one that chose to use only categories and keywords.

All the users select some method and some upper bound, but not all select all methods. Thirty chose sections, 26 chose categories and only 18 chose keywords. It seems that less intuitive methods are less favoured. The users that chose the sections method choose an average of 4 sections. Those that chose categories marked 6, and those that chose keywords marked 6. When the user opts for a method, he tends to select more than one possibility.

Some users select a method (section information for instance) but do not select any particular criteria for it (mark no specific sections), which results in an empty user model. This happened in the case of 14 users, all of which chose only sections. It has been identified as a problem that needs further work.

Regarding differences in profile development between user groups, it has been observed that groups A and B (which had taken part in the development of the project) tended to restrict their selections more than groups C and D: the number of selected sections and categories on average rose steadily from A to D.

System Evaluation

We computed the values of recall and precision and others features for all the users on the last of the three specific days that users were asked to review exhaustively both the daily edition of the newspaper and the message they had received. This allows a comparison between user evaluation and system evaluation to check the exhaustiveness of the user judgements and check the true performance of the system.

We have obtained the following results for each user: recall, precision, number of news selected by the information filtering system, that is, news with relevance greater than zero, number of truly relevant news. The day of this evaluation had 109 news.

Table III shows the average results and the maximum and minimum values for each feature.

	Upper bound	Relevant news	Recall	Not relevant news received	Precision
average	14	69,2	0,2	0	1
max	20	88	1	1	1
min	5	11	0,1	0	0,9

Table III. Recall and Precision figures from system logs

Results Discussion

Studying the results we can see that our system refines the information of the sections with the categories and keywords. The average precision is close to one because the fact that a document belongs to a section is enough for the document to have a high relevance value. Moreover, the relevance for belonging to a section is always greater than the relevance for belonging to a category or containing a keyword. Since most users have selected at least two sections, a section holds an average of eleven documents, and the average upper bound of documents per message is 14, most users get messages where all selected documents are relevant.

If a user marked sections as selection method in his profile (most do, in fact 50% of the users rely on sections altogether to select), the selected documents that appear first belong to these sections. They are shown ordered according to relevance computed with respect to categories and keywords. They are followed by documents that do not belong to these sections but are relevant in terms of categories and keywords. These show much lower relevance values nonetheless.

However, a user that does not use sections obtains documents sorted by the information relative to categories and keywords, and so obtains relevant documents from different sections. Only one user operated in this way, and he obtained a similar value of precision and a low value of recall. This is because he had selected 7 categories and 14

keywords and the number of relevant documents under such wide criteria is above average.

If the relevance value computed using the categorisation method were greater than it currently is, documents relevant according to this source might find their way to the top of the ranking. This is at present unlikely because the categorisation system yields always very low relevance values, but we hope to improve it by developing a richer representation for categories.

If we compare the results of the user evaluation and the system evaluation we can see that the precision obtained is very similar but the recall is lower in the system evaluation. The reason is that a user considers a document as relevant if it refers to something that is interesting for him, whether or not it belongs to a category or contains a word. However, the low recall value is a consequence of the upper bound imposed by the user: with a user model with a few sections and few categories the number of relevant documents is too high to be captured in a maximum recall fixed for the user by the upper bound.

Exhaustive Evaluation: Pros and Cons of the Different Methods

Given the lack of clear results at the end of the initial evaluation, a second stage of evaluation was designed to study the relative advantages and disadvantages of each of the information specification methods applied to the task in hand. As a first step, a more thorough system evaluation was carried out for the results of the first experiment. The 30 non-empty user profiles from the first experiment were selected. The values for recall and precision at 4 points of recall - those that correspond to the most popular choices as upper bound on the number of news items per message - are worked out. These results are shown in Table IV.

	Recall	Precision
max=5	0,11	1,00
max=10	0,22	1,00
max=15	0,31	0,99
max=20	0,39	0,98
Average	**0,26**	**0,99**

Table IV. Recall and precision from first experiment

The results show coincidences with those obtained in the first evaluation, but also show the variation in recall and precision with the different upper bounds. The precision is always very close to 1, whereas the recall has an increment of 10% for each increment of 5 news items in the upper bound.

Thirty six new profiles were constructed to try to represent all the different possibilities that can appear in our system. The total working set consisted of these 36 profiles together with the 30 non-empty user models of the first experiment. The results for recall and precision at the same 4 points of recall were computed. Table V shows the average for recall and precision.

	Recall	Precision
max=5	0,19	0,81
max=10	0,35	0,79
max=15	0,46	0,75
max=20	0,53	0,70
Average	**0,38**	**0,76**

Table V. Recall and precision from second experiment

With this set of profiles the precision decreases by 25% but the recall has an absolute increment of 12%. A closer analysis of our results was performed by considering separately the different parts of a user profile: sections, categories and keywords.

Profiles were constructed with only one section per user and the values for recall-precision at the 4 points of recall were computed. The results are shown in Table VI.

	Recall	Precision
max=5	0,40	1,00
max=10	0,76	0,96
max=15	0,94	0,83
max=20	1,00	0,68
Average	**0,77**	**0,87**

Table VI. Recall and precision from sections

The precision is slightly better (+ 11%), but the recall has a very big increment (+ 39%). These results are based on the numbers of news items per section that are shown in Table VII.

SECTION	ITEMS
culture	17
sports	13
economy	12
people	7
international	13
national	18
opinion	10
society	19
Average	**13,63**

Table VII. News items per section

In table VI there are dramatic variations both in recall and precision are apparent at cut off values around the average number of items per section (table VII). Above this value, the precision is less than one because there are not enough news items in just one section to provide the required number of relevant items. In contrast, the recall is close to one because with 20 news items per message all the relevant items are captured.

Profiles with only one category per user were constructed and the values for recall-precision at the 4 points of recall were computed. The results are shown in Table VIII.

	Recall	Precision
max=5	0,16	0,33
max=10	0,24	0,29
max=15	0,31	0,26
max=20	0,34	0,23
Average	**0,27**	**0,27**

Table VIII. Recall and precision per categories

In this case, the results are low both for recall and for precision. To justify these results it is important to consider the actual distribution of the news items over the categories, in the same way as was discussed for the sections. Table IX shows this information.

CATEGORIES	ITEMS	SECTIONS	WORDS
Arte y cultura (Arts & Humanities)	17	3	107
Ciencia y tecnología (Science)	5	2	101
Ciencias sociales (Social Science)	1	1	41
Deportes y ocio (Recreation & Sports)	13	1	42
Economía y negocios (Business & Economy)	14	3	49
Educación y formación (Education)	1	1	211
Espectáculos y diversión (Entertainment)	10	2	45
Internet y ordenadores (Computers & Internet)	3	2	140
Materiales de consulta (Reference)	3	3	85
Medios de comunicación (News & Media)	2	2	55
Política y gobierno (Government)	62	6	31
Salud (Health)	6	2	52
Sociedad (Society & Culture)	21	4	40
Zonas geográficas (Regional)	25	7	5
Average	**13,07**	**2,79**	**71,71**

Table IX. News items and words per category

One can see that the distribution of the news over categories is very irregular. This implies that some categories, those with few news items, are very difficult to get a good precision for, and they result in a low global value of precision. On the other hand, there is a category with 62 relevant news items, belonging to six different sections. With an upper bound of 20 news items - 20 as maximum point of recall - it is impossible to get a good value for recall - at most a third of all the relevant documents can be retrieved.

An additional problem with the categorisation process that may affect these low values is a poor representation for the categories. Each category is trained only with the name of the subcategories of Yahoo's second level and the descriptions of pages in the web page of the category. This resulted in a representation where each category was represented only with 71,71 words on average, but with big differences between categories. This information is also shown in table IX. For instance, Education (Educación y formación) has 211 words, whereas Regional (Zonas geográficas) has only 5 words. The significance of these numbers becomes apparent when compared with the actual number of terms used to represent a news item. The number of words per new item is shown in Table X. The values shown that news items belonging to the opinion section (leading articles) have more words, whereas those from the sports section have the minimum number of words. The average is approximately of 441 words per news item. Such a number provides a meaningful representation in the VSM of the content of the news item, whereas the number of words used to represent the

categories falls well below in most cases. This could explain poor evaluation results for category based searches.

culture	381,35
sports	260,23
economy	422,33
people	355,56
international	548,00
national	487,28
opinion	645,50
society	431,26
Average	**441,44**

Table X. Words per news item (per section)

Additional problems may spring from the fact that categories were trained with documents that did not belong to the same domain over which the categorisation was later to be performed. Documents indexed in Yahoo! are not news items, and categories trained over them will produce worse results than they might have done if they had been trained with documents belonging to the same domain.

Using only the specific interesting terms provided by the user to drive key word searches always gives good precision. The received news items contain the terms, but the values for recall depend on the specificity of the terms. If the terms chosen are very specific, recall will be low, and if the terms are very general, recall will be high.

Conclusions and Further Work

The experience of evaluating system performance and user satisfaction for the Mercurio personalised news server has provided important insights concerning three different aspects: the nature of the information handling tasks that digital news services are faced with, the relative merits of the three most popular methods of specifying information interests (sections, categories, and key words) with respect to this particular set of tasks, and the risks of careless application of recall and precision measures in systems such as these where different methods of specifying interests are combined.

Digital news services face a double challenge in providing users with efficient, useful and easy to understand systems. On one hand, they have to ensure that the tools they provide for the user to specify his interest in information items of a particular type are sound according to traditional information retrieval measurements. On the other hand, they face a competitive market where different methods of specifying user interest are continuously competing for the user's eye. In most cases a compromise is reached either by providing a single well-tried means of specifying information (usually newspaper sections or key words) or by providing different methods for separate services (section selection for email services, keywords for Internet searches). The Mercurio system shows that integration is a real possibility with great chances of capturing user sympathies. However, it also showed that a certain degree of clouding of the efficiency parameters of the search procedures may be involved.

The study of the different behaviours of the three methods used to specify user interests (sections, categories and keywords) has shown how delicate the interaction is between the complete set of parameters involved: news items per section, news items per category, maximum number of news items per message required by the user, general relevance of the contents of a given day for a given user... Because most of these

parameters change daily in the case of digital news services, and their values on any particular day are independent of previous or future values, it becomes impossible to assert that any given method is best overall. A wise combination of various different ways of specifying a particular interest may have a better chance of consistently coming up with a reasonable personal selection over a number of days.

The discussion of observed recall and precision values for the specially prepared experiments where the influence of each particular method was carefully delimited explains many of the possible pitfalls when interpreting raw overall calculations of such measurements. The main obstacle to their successful application is the fact that users are allowed to specify a maximum number of news items per message. Since no minimum relevance threshold is specified (though non-zero is taken as a default), this upper bound usually acts as a minimum number as well. However, additional non trivial problems arise from the very nature of the task involved in these systems: just as newspaper must come full of news everyday, regardless of whether anything interesting has happened or not, newspaper sections will carry a certain amount of news independently of their relevance to the section heading, and a personalised news message will feature some news item distantly related to the user profile.

Considering these various issues, it is clear that more qualitative and non-reductionist research on this topic is needed. Such non-reductionist research should endeavour to describe and generalise about the informational values of different subject access points in databases (in this case newspapers databases). The present paper has considered the relative merits of sections, categories and key words. Further work ought to consider what the different values for retrieval are for a whole list of different access points: headings, sections, introductions, full text words, value-added descriptors and categories. It should also take into account the possibility of these access points having different values for the different kinds of tasks forming the information needs.

References

Amato, G. and Straccia, U. (1999). "User Profile Modelling and Applications to Digital Libraries". Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Lectures Notes in Computer Science, Springer Verlag, Paris, pp. 184-197

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. ACM Press Books, New York.

Clarke, S and Willet, P. (1997). “Estimating the recall performance of web search engines”, Aslib Proceedings, vol. 4, nº 7, pp. 184-189.

Díaz, A., Gervás, P. and García A. (2000). "Evaluating a User-Model Based Personalisation Architecture for Digital News Services". Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries, Lectures Notes in Computer Science, Springer Verlag, Lisbon, pp. 259-268

Dong, X. and Su, L. T. (1997). “Search Engines on the World Wide Web and Information Retrieval from the Internet: a review and evaluation”. Online & CD-ROM Review, vol. 21, nº2, 1997, pp. 67-81.

García, A., Chacón, I., Díaz, A. and Gervás, P. (2000). “Nuevos sistemas de información: tendencias y evaluación”, Cuadernos de Documentación Multimedia, nº9, http://www.ucm.es/info/multidoc/multidoc/revista/num9/prensa/jime-chacon.htm

Gervás, P., San Miguel, B., Díaz, A. and García, A. (1999) "Mercurio: un servidor personalizado de noticias basado en modelos de usuario obtenidos a través de la

WWW"*, III Congreso de Investigadores Audiovisuales (Los medios del tercer milenio)*, 10-12 noviembre 1999, Facultad de Ciencias de la Información, Universidad Complutense de Madrid, Madrid

Gómez, J. M. and Buenaga, M. (1997). "Integrating a Lexical Database and a Training Collection for Text Categorisation". ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP, Madrid (Spain), 1997.

Gordon, M. and Pathak, P. (1999). "Finding information on the World Wide Web: the retrieval effectiveness of search engines". Information Processing and Management, vol. 35, pp. 141-180.

Leighton, V.H. and Srivastava, J. (1999). "First 20 Precision among World Wide Web Search Services (Search Engines)", Journal of the American Society for Information Science, vol. 50, nº 10, pp. 870-881.

Lewis, D.D., Schapire, R.E., Callan, J.P. and Papka, R. (1996). "Training algorithms for linear text classifiers". In Proceedings of the ACM SIGIR, 1996.

Maldonado, A. and Fernández, E. (1998) "Evaluación de los principales buscadores desde un punto de vista documental: recogida, análisis y recuperación de recursos de información". Actas de las VI Jornadas Españolas de Documentación, vol. 2, pp. 529-552.

Olvera Lobo, M. D. (2000), "Rendimiento de los sistemas de recuperación de información en la world wide web: revisión metodológica", Revista Española de Documentación científica, vol. 23, nº 1, pp. 63-77

Pastor, J. A. and Asensi, V. (1999). "Un modelo para la Evaluación de Interfaces en Sistemas de Recuperación de Información", IV Congreso Isko-España Eoconsid'99, Granada (Spain), 1999.

Salton, G. (1989). Automatic Text Processing: the transformation, analysis and retrieval of information by computer. Addison Wesley. 1989

Sebastiani, F. (1999). "A Tutorial on Automated Text Categorisation". Proceedings of the First Argentinean Symposium on Artificial Intelligence (ASAI-99)

Schwartz, C. (1998). "Web Search Engines", JASIS, vol. 49, n. 11, pp. 973-982

Yang, Y. (1999), "An Evaluation of Statistical Approaches to Text Categorisation", Information Retrieval Journal

Alberto Díaz Esteban (alberto@dinar.esi.uem.es)
Bachelor in Physics. Lecturer in the Artificial Intelligence Department at the School of Computer Science in the Universidad Europea-CEES in Madrid.
Research: Information filtering, user modelling and multilingual text classification tasks. Recently I have collaborated as researcher in three funded projects: "Perseo: an automatic personalised marketing system", "Digital Object Brokering over the Internet", y "Mercurio: Development of a User Oriented Content Provider and Intelligent Search Engine" As result of this project, I have presented recently the article "Evaluating a User-Model Based Personalization Architecture for Digital News Services", in the *Fourth European Conference on Research and Advanced Technology for Digital Libraries* (ECDL 2000, Lisbon).

Pablo Gervás Gómez-Navarro (pgervas@sip.ucm.es)
PhD in Computer Science and lecturer at the Computer Science Faculty of the Universidad Complutense de Madrid. As a former member of the Artificial Intelligence

Department of the School of Computer Science in the Universidad Europea-CEES in Madrid, he has taken part in many of their research projects oriented to intelligent information access and dissemination over the Internet.
Currently involved in work on multilingual information filtering, expert systems applied to language generation and timetabling, and case based retrieval for the reuse of software frameworks.
Recent publications include an expert system for the composition of poetry in Spanish presented at ES 2000 international conference on Expert Systems (Peterhouse College, Cambridge).

Antonio García Jiménez (antonio.garcia@fcp.cin.uem.es)
PhD in Information Science and lecturer on Documentation at the Universidad Europea-CEES, Madrid. Remarkable among his latest works are: a working document titled *Procedimientos de Evaluación: el caso del Análisis Documental y de los lenguajes documentales en la actividad periodística*, a paper presented at the Fourth Conference on Digital Libraries (ECDL 2000), *Evaluating a User-Model Based Personalisation Architecture for Digital News Services* and the paper *Nuevos sistemas de información: tendencias y evaluación*, published in the Cuadernos de Documentación Multimedia journal. He has taken part in several funded projects related with information handling over the Internet, concentrating specially on evaluation. At present he is taking part in a project oriented to the development of a Thesaurus for a media company. His research is related to Knowledge Organisation and Management and Hypermedia Dcoumentation.

Inmaculada Chacón Gutiérrez (inmaculada.chacon@fcp.cin.uem.es)

PhD in Information Science, Universidad Complutense de Madrid. She lectures on Documentation at the Universidad Europea de Madrid to undergraduate and postgraduate students of several faculties. She has taken part in three externally funded research projects: "Perseo: an automatic personalised marketing system", "Digital Object Brokering over the Internet", y "Mercurio: Development of a User Oriented Content Provider and Intelligent Search Engine". At present she is taking part in a project oriented to the development of a Thesaurus for a media company. Her research is related to hypermedia Documentation and its manipulation possibilities. Latest publications: Proyecto Mercurio: un servicio personalizado de noticias basado en técnicas de clasificación de texto y modelado de usuario, Proceedings of the XVI Meeting of the SEPLN (Sociedad española para el procesamiento del Lenguaje Natural), Vigo, España, September 2000). NUEVOS SISTEMAS DE INFORMACIÓN: TENDENCIAS Y EVALUACIÓN, Cuadernos de Documentación Multimedia, nº 9, 2000, http://www.ucm.es/info/multidoc/multidoc/revista/num9/prensa/jime-chacon.htm

Keywords
evaluation, user modelling, personalised news services, information retrieval