Multidocument Summarization: An Added Value to Clustering in Interactive Retrieval

MANUEL J. MAÑA-LÓPEZ
Universidad de Vigo
and
MANUEL DE BUENAGA and JOSÉ M. GÓMEZ-HIDALGO
Universidad Europea de Madrid

A more and more generalized problem in effective information access is the presence in the same corpus of multiple documents that contain similar information. Generally, users may be interested in locating, for a topic addressed by a group of similar documents, one or several particular aspects. This kind of task, called instance or aspectual retrieval, has been explored in several TREC Interactive Tracks. In this article, we propose in addition to the classification capacity of clustering techniques, the possibility of offering a indicative extract about the contents of several sources by means of multidocument summarization techniques. Two kinds of summaries are provided. The first one covers the similarities of each cluster of documents retrieved. The second one shows the particularities of each document with respect to the common topic in the cluster. The document multitopic structure has been used in order to determine similarities and differences of topics in the cluster of documents. The system is independent of document domain and genre. An evaluation of the proposed system with users proves significant improvements in effectiveness. The results of previous experiments that have compared clustering algorithms are also reported.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*abstracting methods*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering, search process*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*evaluation / methodology*

General Terms: Experimentation, Human Factors, Measurement, Performance

Additional Key Words and Phrases: Multidocument summarization, topic segmentation

1. INTRODUCTION

It is more and more usual that users of Information Retrieval (IR) systems must face up to responses of hundreds or thousands of retrieved documents.

Portions of this work was done while M. J. Maña-López was visiting the Universidad Europea de Madrid, partly supported by a fellowship from the Universidad de Vigo.
Authors' addresses: M. J. Maña-López, Escuela Politécnica Superior, Universidad de Huelva, 21071 Palos de la Frontera, Huelva, Spain; email: manuel.mana@diesia.uhu.es; M. Buenaga and J. M. Gómez-Hidalgo, Escuela Superior Politécnica, Universidad Europea de Madrid, 28670 Villaviciosa de Odón, Madrid, Spain; email: {buenaga,jmgomez}@uem.es.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 1515 Broadway, New York, NY 10036 USA, fax: +1 (212) 869-0481, or permissions@acm.org.
© 2004 ACM 1046-8188/04/0400-0215 $5.00

A significant proportion of them are moreover irrelevant or redundant. This is a straight consequence of two well-known facts: the incessant growing of documental databases and the lack of ability of most users to define specific information needs (for this last, see, e.g., Spink et al. [2002]). In such a framework, further effective retrieval techniques are only part of the solution. Interfaces with alternative ways for organizing and browsing the results returned by a search engine would be of great utility assisting users to find relevant documents. Document clustering may play a central role as an effective technique for organizing query results.

Cluster hypothesis states that relevant and irrelevant documents tend to group in different clusters [van Rijsbergen 1979]. Initially, document clustering has been proposed as a method of enhancing the performance of IR systems [Salton 1968]. In this case, the entire document collection was clustered off-line and queries could be compared with a representation of each cluster instead of a representation of each document. This kind of clustering was also investigated as a way of improving effectiveness of IR systems [Jardine and van Rijsbergen 1971]. However, in the latest years, clustering has been applied to documents retrieved for a query (e.g., Hearst and Pedersen [1996], Zamir and Etzioni [1998], Carey et al. [2000], Leuski [2001], Wu et al. [2001]). The purpose of post-retrieval clustering is to organize the results returned by the IR system and to make the browsing through them easier. The clustering of search results is even applied successfully on the Web by search engines, as Teoma, Northern Light or Vivísimo. The impact of clustering techniques on search effectiveness may mean, according to Vivísimo, an increase greater than 40%, measured in the probability of finding relevant information.[1]

Nevertheless, recent works [Fuller et al. 1998; Wu et al. 2001] about clustering evaluation involving users show no significant differences in effectiveness between interfaces using classic ranked list and postretrieval clustering. Among others, authors of these works suggested two enlightening causes. First, in some cases, the information accompanying clusters, the list of keywords, is not clear enough about the content of cluster documents. Second, although the organization in clusters helps users to discover groups with high density in relevant documents, clustering interface does not give further assistance in identifying particular documents with relevant information.

A system that exploits the capacity of organization of clustering techniques but being, also, able to complement groups and documents with useful information, can be of great utility in an interactive retrieval setting. In this article, we propose to integrate multidocument summarization (MDS) techniques with a postretrieval clustering interface. The final result is a system that offers a summary for each cluster reporting document similarities and a summary for each document highlighting the singular aspects that it provides with respect to the common information in the cluster.

The working hypothesis is that the system proposed will help users to feel less overwhelmed by the amount of information and to get a better understanding of the different *aspects* of the information available in the retrieved

[1]`http://vivisimo.com/products/Clustering_Engine/Benefits.html`.

documents. In this sense, the complexity of the information that the user has to cope with is drastically reduced. Just reading some sentences, a user can, on the one hand, select the cluster of documents closest to his or her information needs, discarding the rest of clusters. On the other, the summary generated for this cluster reveals the salient and shared information of its documents. In a second step, an indicative and single-document summary focused on the original issues of each document is showed. Then, the user can select the document that includes the newest information or, in general, satisfies his or her interests.

Usually a document discusses a few main topics or a main topic and a few associated issues. In a collection of documents about the same subject, one or several of these topics describe the central idea of the collection, whereas the rest of subtopics probably cover complementary or marginal matters. Knowledge of the subtopic structure of each document will help discover the shared topic and the original information provided by each one of them. In this article, we use TextTiling system [Hearst 1997] to segment texts into multiparagraph units reflecting its subtopic structure.

A system like the one described in this article can be of great utility in an interactive retrieval setting, but especially in tasks oriented to locate a particular aspect or all aspects of a topic. This kind of task has been explored in the TREC Interactive Track as "instance retrieval" [Over 1998], in conferences from TREC-6 to TREC-8. The goal, in contrast with ad hoc retrieval experiments where users are asked to find all documents that are relevant to a topic, is to locate documents, which, taken together, contain as many as possible different instances of the information wanted. A set of experiments must be designed to contrast working hypothesis, that is, they attempt to probe the usefulness of MDS in a post-clustering interface. Particularly, experiments should investigate factors like the differences of effectiveness (in recall and precision) or the number of groups and documents explored.

Previous works in application of MDS techniques to IR settings (e.g., Ando et al. [2000] and Kan et al. [2001]) propose advanced frameworks of information access. These proposals are focused on new user interfaces (e.g., Ando et al. [2000]) or on new paradigms for browsing and searching (e.g., Kan et al. [2001]). So, in Ando et al. [2000], a system that detects topics underlying a given document collection and a graphical user interface are described. The user interface shows these themes and the relationships among topics and documents. In Kan et al. [2001], informative and indicative summaries are generated to fulfill user's needs in browsing and searching, respectively. The presented system works with highly structured documents.

In summary, we propose an alternative to the classical output of IR systems based on the utilization of MDS techniques. Multitopic structure of documents is considered in order to detect common and original topics in clusters of related documents. A summary indicative of the information overlaps in the cluster is produced. Also, the differences provided for each document are discovered and summarized. The techniques described may be applied to a variety of document domains. A systematic and task-based evaluation has been carried out and results and conclusions are presented.

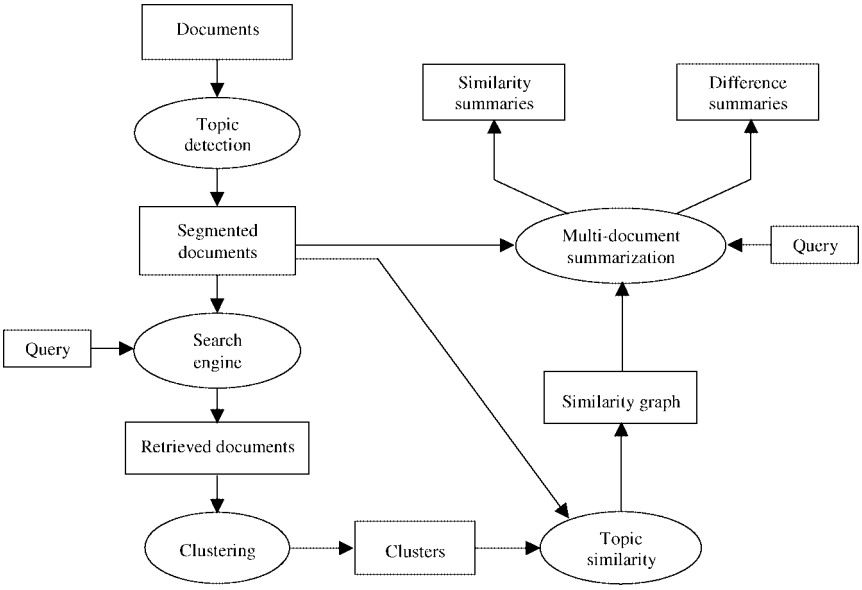

Fig. 1. Diagram of the multidocument summarization system. Documents are segmented into multi-paragraph units, reflecting their subtopic structure. Query results are clustered by semantic closeness and a graph that maps the similarity relationships among topics for each cluster is obtained. Information about document topics in each cluster and their similarity is used to compute a summary for the cluster central topic and also summaries focused on particularities of each document.

The remainder of this article is organized as follows. First, we introduce an overview of the proposed system where clustering and topical structure of documents are presented as the previous steps needed for the application of MDS techniques. Different clustering techniques are shown and the results of a comparison carried out among some of them are provided. Then, we describe how the similarity and difference summaries are generated. Finally, the purpose of the experiments and the methodology followed as well as the obtained results are discussed.

2. SYSTEM DESIGN

Figure 1 outlines the proposed system. Retrieved documents resulting from a query are clustered by semantic closeness. However, in spite of the thematic relationship that keep documents in a cluster, topics addressed typically cover a diversity wide enough to make necessary the use of finer-grained techniques to discover the topic common to all of them. Moreover, a single document may deal with several aspects of a particular issue. As, in general, documents lack of any element of demarcation that denotes its subtopic structure, topics have to be discover automatically. The similarity relationship among topics (i.e., document segments) in a cluster is computed to identify the central topic in this cluster and the original aspects in each one of the documents. This information sets up the basis of summaries focused on similarities and differences, respectively.

The queries submitted by users are usually so wide that they hardly reflect any specific information needs. A recent report on Web searching behaviors [Spink et al. 2002] found that the mean number of terms per query is 2.6, so that more than 50% of queries contain a maximum of two terms and almost 30% have only one term. Such a proportion of short queries cannot be due solely to broad needs but also to the degree of difficulty that supposes for a user to specify an information need a search engine needs. The result of this kind of wide queries is usually a list of thousands of retrieved documents, most of them without semantic connection to the query and, probably, only a few related with user needs. On the other hand, due to the huge volume of information in the Web, a significant degree of overlap among some of the retrieved documents is expected.

The same report reveals that more than 70% of users looked at one or two pages of results per query. Part of this behavior is held to the low tolerance of many users to navigate through a long list of results. The other reason is the deficient precision of search engines.

In such a searching scenario, a multidocument summary standing out similarities among documents would include valuable clues that will assist users to recognize the group or groups of documents related with query real meaning. Probably, users would only have interest about one or a few aspects of a subject. A single-document summary focused on differences would facilitate users to find the particular document or documents that address the aspects related to the query that are of his or her true interest.

In this section, we will deal with clustering techniques and topical structure in a cluster of related documents. Multidocument summarization techniques will be addressed in next section.

2.1 Clustering Techniques

Clustering methods split a set of documents into groups or *clusters*, revealing implicit structure in a corpus. According to the type of cluster structure they produce, clustering methods are usually classified in *partitional* or *nonhierarchical* and *hierarchical* [Rasmussen 1992].

Hierarchical algorithms produce a nested set of partitions that can be displayed graphically as a tree or *dendrogram*. In this hierarchy, intermediate nodes can be viewed as a grouping from the next lower level or a partition from the next higher level. The leaves of the tree are the single documents of the clustered set. This tree can be produced either bottom-up or top-down. Bottom-up or *agglomerative* clustering starts with individual documents and groups the most similar documents or clusters at each step. Top-down or *divisive* clustering starts with all the documents in a single cluster and divides them into groups so as to maximize within-group similarity.

In contrast to previous methods, partitional clustering simply divides the document set into a given number of groups at the same level, where no overlap is allowed. Each cluster may be represented by a *centroid* or *center of gravity*, which is representative of the characteristics of the documents it contains. Among partitional algorithms, *K-means* [Krishnaiah and Kanal 1982] is very

simple and widely used in document clustering. Initially, K documents are chosen at random as centroids. Then, in each iteration, all documents are assigned to the cluster whose center is the closest and next the centroid of each cluster is recomputed. Iteration ends when documents are assigned to the same clusters in consecutive rounds.

A combination of divisive and partitional clustering methods, named *bisecting K-means*, is introduced in Steinbach et al. [2000]. This algorithm starts with a single cluster which contains all the documents and runs as follows. In each iteration, a cluster to split is chosen. Then, K-means algorithm is used to find a partition in two groups of the selected cluster (i.e., $K = 2$). The so-called bisecting step can be repeated a certain number of times with the purpose of taking the split that produces the clustering with the highest overall similarity. The whole process is repeated until the desired number of clusters is reached. The choice criterion used in the first step may be based on the size of the cluster or on the overall similarity. The evaluation carried out in this work shows that bisecting K-means technique is better, in terms of entropy and overall similarity, than standard K-means and the hierarchical agglomerative clustering (HAC) approaches tested.

For our experiments, we are interested in an approach which presents a list of partitions for the set of retrieved documents instead of a nested structure of clusters. Previous works in post-retrieval clustering (e.g., Hearst and Pedersen [1996], Leuski [2001], and Wu et al. [2001]) follow this approach. In Hearst and Pedersen [1996], a hybrid technique between K-means and HAC is presented. The initial centers of the K-means algorithm are found using HAC. The bottom-up process is stopped when K clusters are reached, and then their centroids are computed. In Leuski [2001], an HAC algorithm is applied, setting a threshold on the similarity distance between clusters to obtain a partition of the document set. Finally, the clustering system presented in Wu et al. [2001] used the K-means algorithm.

In order to select an effective clustering algorithm to integrate in the system, a set of experiments involving K-means and bisecting K-means was conducted. These algorithms were chosen because of their availability. The first one is integrated in Weka.[2] [Witten and Frank 1999], a package written in Java, which includes the implementation of many popular machine learning algorithms, and the second one in Cluto[3] [Karypis 2002; Zhao and Karypis 2001], a collection of programs written in C for clustering high-dimensional data sets. After some trials, we could verify, like other works have done previously [Wu et al. 2001], that clustering algorithms could not group documents with different instance relevance in different clusters. Then, the criterion was to choose the algorithm which maximizes the accuracy of relevant class in clusters and minimizes the number of clusters with relevant documents. In this way, a user of the system could access the most relevant documents only inspecting few clusters.

In the experiments, the corpus, topics and relevant judgments which are described in detail in Section 4 were used. In summary, for each one of the

[2]Available at `http://www.cs.waikato.ac.nz/ml/weka/index.html`.

[3]Available at `http://www-users.cs.umn.edu/karypis/cluto/index.html`.

Table I. Results of the Evaluation of Clustering Effectiveness using *K*-Means over the Eight Topics

Feature selection method	Number of clusters with relevant documents	Accuracy of relevant class
$tf \cdot idf$ – 10%	3.38	0.316
$tf \cdot idf$ – 7%	3.25	0.291
$tf \cdot idf$ – 5%	1.50	0.275
$tf \cdot idf$ – 3%	3.13	0.248
tf 40% – df 1	3.00	0.280

Average number of clusters with some relevant document and average accuracy of relevant class are shown. The feature selection methods applied are $tf \cdot idf$ (the percentage of selected terms is specified) and a combination of term and document frequencies [Vaithyanathan and Dom 1999].

8 topics, 300 documents were retrieved and, in all the cases, 10 clusters were obtained. The accuracy of clustering was computed averaging the proportion of relevant documents in each cluster over the total number of clusters with some relevant document for all topics.

In the firsts tests using *K*-means, the clusters obtained were of very irregular size, varying from one to almost two hundreds documents. The existence of clusters so long decreases the effectiveness to the search, because users have to explore, like in the case of ranked list interfaces, a long set of results. A possible cause for this miss-behavior of *K*-means might be the high dimensionality of data sets, typical of text document collections. In our case, the average number of terms of the eight sets of retrieved documents was more than 5600. Previous works have studied different methods of feature selection for document clustering, ranging from heuristics based on term frequencies or weights [Vaithyanathan and Dom 1999; Rüger and Gauch 2000] to more complex models that capture the probabilistic dependencies between features [Sahami 1998].

The application of feature reduction methods to our document sets achieved more homogenous cluster sizes as well as improvements in the accuracy of relevant class. One of these methods was the selection of a certain percentage of terms with the greatest weight. Term weights were computed using the well-known formula of $tf \cdot idf$ [Salton and Buckley 1988]. Table I shows the evaluation results for different percentages of terms selected. Another feature selection method tried was an adaptation of the one presented in [Vaithyanathan and Dom 1999]. In this case, the 40% of terms with the highest frequency and those with a document frequency lower than 2 were removed. Both frequencies are referred to terms in the collection of retrieved documents. The results achieved are also shown in Table I.

The tests using bisecting *K*-means were essentially oriented to select the clustering criterion function, among the total of seven different functions available in Cluto. From Zhao and Karypis [2001], it can be concluded that $\mathcal{I}_2$, $\mathcal{E}_1$, $\mathcal{H}_2$ and $\mathcal{G}'_1$ criterion functions lead to good clustering solutions. In brief, $\mathcal{I}_2$ function maximizes the similarity between each document and the centroid of the cluster to which it is assigned. $\mathcal{E}_1$ tries to minimize the cosine between the centroid vector of each cluster and the centroid vector of the entire collection. $\mathcal{H}_2$ is a hybrid criterion function which is obtained by combining $\mathcal{I}_2$ with $\mathcal{E}_1$. Finally, $\mathcal{G}'_1$ uses a graph obtained by computing the pair-wise similarities between documents.

Table II. Results of the Evaluation of Clustering Effectiveness using Bisecting *K*-Means over the Eight Topics

Clustering criterion function	Refinement	Number of clusters with relevant documents	Accuracy of relevant class
$\mathcal{I}_2$	no	2.25	0.283
	yes	2.00	0.337
$\mathcal{E}_1$	no	2.50	0.265
	yes	2.25	0.248
$\mathcal{H}_2$	no	2.50	0.260
	yes	2.13	0.273
$\mathcal{G}'_1$	no	2.75	0.228
	yes	2.50	0.254

Average number of clusters with some relevant document and average accuracy of relevant class are shown.

Moreover, each function was tested using an additional refinement phase. Basically, it is an incremental refinement strategy which moves a document to a cluster as soon as it is determined that this will lead to an improvement in the value of the criterion function. Table II shows the results of the tests for these functions on our document collection and using the same quality metrics that in the *K*-means case.

The results show that bisecting *K*-means is better than *K*-means when the clustering criterion function used is $\mathcal{I}_2$ and refinement phase is applied. Series of experiments were conducted to test bisecting *K*-means after a dimensional reduction was carried out using previously mentioned methods. The initial results have not been improved, proving the capacity of the algorithm to manage high dimensional data sets.

In short, the clustering algorithm chosen as previous step of the MDS system presented in this paper was bisecting *K*-means, implemented in Cluto, using $\mathcal{I}_2$ as criterion function and later refinement.

2.2 Characterization of Cluster Topical Structure

Finding some kind of topical structure is possible, even in simple texts. That is, authors structure text discourse in function of a set of coherent thematic components which support the main topic discussion. In expository text, this structure usually responds to a sequence of subtopics in which each of them contributes to the final purpose of communicating the entire central idea to the reader. Section headings should break up a text in its structural elements; however, in many cases, sections include more than one subtopic and even a sole subtopic might be broken into two or more sections [Salton et al. 1996]. Additionally, many documents lack any structural sign or they are scarcely marked. Thus, it becomes necessary a segmentation system able to split a text into multiparagraph units reflecting their subtopic structure. These units, called *segments*, are characterized by its high cohesion, that is, they are largely internal tied but essentially disconnected from surrounding units.

In text segmentation, it is of particular interest the TextTiling of Hearst [1997]. This algorithm is based on the hypothesis that a subtopic change in the discourse implies a significant change in the vocabulary used. Text is divided

in pieces of a predefined size (w words), called *token-sequences*. For every gap between a pair of token-sequences, a score, called *lexical score*, is computed. Scores are worked out as the inner product of two vectors of term frequencies corresponding to adjacent fixed long *blocks* of k token-sequences. Topic boundary identification is done adding for each token-sequence gap the differences of lexical scores between this one and its previous and next gaps. These *depth scores* are sorted, according to the obtained value, and segment boundaries are assigned to the largest similarity changes.

These highly cohesive segments are specially valuable in automatic summarization, due to they address a only topic to summarize. On the contrary, automatic summarization of documents covering several topics may produce *unbalanced* summaries, that is, summaries which lack some important topics or which do not reflect structural organization of the original text [Paice 1990]. This is the reason why text segmentation has been used as a previous step before summarization.

In Mittal et al. [1999], the advantages in text summarization of segmenting documents using TextTiling are evaluated. For the experiments, composite documents were created by concatenation of several news. Summaries generated by two different algorithms, neither of the two using positional information, are contrasted with the collection of summaries for the individual texts. A F_1 score averaged across several compression levels is computed with and without topic boundary detection. Results show that text segmentation can significantly improve summary effectiveness using both algorithms, even with a factor near two. It is reasonable to expect that summarizers that exploit sentence position techniques make the best use of segmentation. This is the case of Nakao [2000] which describes a variation of TextTiling algorithm to construct a one-page summary for a long text. The resulting summaries are produced extracting sentences only from the lead part of each segment.

It is expected, then, that a cluster of documents addressing similar contents covers a huge variety of different subtopics in relation to the main subject. For example, a group of news about the Prestige tanker disaster in northwestern Spain coasts may cover, additionally to the main information about the oil spilling, other aspects such as the ones related to the ecologic damage caused, the repercussion on local economy, the clean-up tasks, the government's decisions or the international laws on single-hull tankers. Each document of the cluster may address one or several of these aspects making very difficult a summarization without a previous identification of, at least, topic boundaries. After that, it is possible to find semantic relations among segments of different documents and thus, determine which of them define the common story line in the cluster.

Due to the environment of use decided for our summaries, that is, an IR system with post-retrieval clustering, which uses them with a merely indicative purpose, we propose to exploit only the most significant segment. In this setting, it is not probable a crucial loss of information that hampers the identification of the common idea behind the group of related documents. In contrast, the simplification of MDS problem is evident. The most prominent gain is that reducing to one the number of sources to handle we avoid the problems of

redundancy and inconsistency inherent to summaries extracted from different sources. We will deal with this issue in detail in the next section.

In order to recognize the common topic in the bag of segments corresponding to a document cluster, the similarities among them are characterized using a graph. This is a natural way of representing cohesion in text and has been widely used in previous works. For instance, in Salton et al. [1997] intradocument links, that is, relationships among paragraphs of a document, represented in a graph are used first to segment text in homogenous passages, and then to summarize the text by paragraph extraction. The selection of paragraphs that will compose the summary is determined by different ways of traversing the graph. Any way, *bushy nodes*, defined as nodes linked with many other nodes and therefore with overlapping vocabulary in relation with other paragraphs, are considered good candidates for extraction. This last idea seems to be in concordance with the *Graph Connectivity Assumption* defined in Mani [2001, p. 94] and which states that "nodes which are connected to lots of other nodes are likely to carry salient information". This correlation between connectivity of a node and salience of a text element goes back to the work of Skorochod'ko [1972], who applies it to relationships among sentences to discover the document structure.

In our similarity graph, the nodes are the segments of a cluster. Similarities between segments are computed as the inner product of both vectors of term weights ($tf \cdot idf$ type). An edge is set up between the nodes representing two segments when the similarity between them exceeds certain threshold, which for our experiments was fixed in 0.1. Then, for each cluster, we chose the segment with the highest degree in its similarity graph.

3. MULTIDOCUMENT SUMMARIZATION

A significant challenge for multidocument summarization is the capability to discover the similarities among a group of documents covering related information, but also to highlight the singular aspects provided by each one. In next sections, we will describe how similarities and differences among the documents to summarize are identified and taken into account.

3.1 Summarizing the Similarities

Our purpose in applying MDS to a IR system with post-retrieval clustering is to produce automatically a summary that emphasizes the similarities in each cluster. The goal is to help users to choose the group or groups of documents more related to his/her information needs, discarding the rest of groups.

Journalists usually write articles containing background information extracted from several sources. In Mani [2001, p. 181], this process of manual MDS is described. The journalist selects one of the texts according to different criteria as the credibility of the source, the date or the degree of coverage of the news. This text is used as source for a summary which is complemented with information extracted from the other documents.

This idea of using as basis only one of the texts is behind the MDS system presented in this article. As we have said in the previous section, the more

connected segment in the similarity graph is chosen as representative of the whole cluster. The advantages of this approach are related with the simplification of redundancy and inconsistency problems achieved by the reduction to one of the number of sources to process. When the number of sources to handle is greater than one (that is the habitual scenario in MDS), summarization systems have to avoid including duplicate information. This might reach using techniques like *Cross-Sentence Informational Subsumption* (CSIS) [Radev et al. 2000] or *Maximal Marginal Relevance Multi-Document* (MMR-MD) [Goldstein et al. 2000], which compute penalty measures based on similarity factors to avoid selecting redundant sentences. Moreover, the problems of inconsistency become also worse than in single-document summarization (SDS), because the complexity to connect sentences from different documents. Although the texts to summarize deal with the same subject, each document present a succession of consistent ideas that is hard to break and connect with others in different documents. The result might be a summary of difficult reading because the gaps between sentences.

Sentence extraction techniques allow to construct summaries that are domain independent. The summaries are generated selecting sentences of the original document that contain information highly indicative of its content. The selection is made scoring each of the sentences using a heuristic set. Finally, the sentences with the highest scores are chosen.

Among the most used heuristics for selecting sentences, we have chosen the centroid, title, location and query methods. These techniques have been successfully used in the construction of summaries oriented to IR tasks [Maña-López et al. 1999]. In next sections, we will explain how these techniques are applied to obtain the similarity summaries from the representative segment of each cluster.

3.1.1 *Centroid Method.* The centroid method for MDS is inspired in the keyword method [Luhn 1958] for SDS. This method involves looking for sentences that contain words which appear frequently or *keywords*, believing that these words refer to central topics of the document. This hypothesis is based on the fact that the emphasis in the utilization of certain words by the document's author can be a good indicator of its significance.

To score the sentences, Luhn looks for *clusters of keywords*, arguing that the concentration of keywords in a sentence is a clue about its degree of meaningfulness. These clusters are fragments of sentences delimited by keywords, such that between any two keywords within a cluster there are no more than another four non-keywords. In this way, later works have studied the relationships between words and have proved that 98% of the lexical relations in English occur between words within an interval of five (see Abraços and Lopes [1997]).

In our system, we use clusters with two or more keywords and with a maximum of another five words between them. The keywords are the ten terms with greater $tf \cdot idf$ in the centroid of the most cohesive segments of the cluster of documents, that is, the most connected segment and all the segments to which it is related.

To obtain the score of a sentence, the cluster with more keywords is selected. Using Luhn's method, the square of the keyword number is divided by the total number of words in the cluster. The definitive score in this method is computed multiplying the previous value by the sum of the cluster keyword weights ($tf \cdot idf$).

More formally, let d be a document (or a document segment), K be the set of centroid keywords and their weights $\{(k_1, w_1), (k_2, w_2), \ldots, (k_{10}, w_{10})\}$. Let s_i be a sentence of d and c_{ij} be a keyword cluster of s_i, where $1 \leq i \leq t$ and $0 \leq j \leq p$, being t the number of sentences of d and p the number of clusters of s_i. Further, let N_{ij} and n_{ij} be the total number of words and the number of keywords within the cluster c_{ij}, respectively. If c_{im} is the cluster of s_i that contains more keywords, that is, $n_{im} = \max_{0 \leq j \leq p} \{n_{ij}\}$, then the score of s_i, using centroid method, is defined as:

$$score_C(s_i) = \begin{cases} \frac{n_{im}^2}{N_{im}} \cdot \sum_{l=1}^{10} w_l \cdot b_{im}(k_l) & \text{if} \quad p > 0, \\ 0 & \text{if} \quad p = 0, \end{cases} \tag{1}$$

where,

$$b_{ij}(k_l) = \begin{cases} 1 & \text{if} \quad k_l \in c_{ij}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

3.1.2 *Title Method.* This method supposes that words occurring into document titles, subtitles, and headings could be good indicators of its content [Edmundson 1969]. Moreover, we could think that, when the text author divides the document body in sections, he makes, in some way, a summary of them choosing the suitable titles.

To compute the score for a sentence, the square of the number of words in the title which are also in the sentence is divided between the total number of terms (stop-words excluded) in the title.

3.1.3 *Location Method.* In many cases, the first sentence of a paragraph suggests its central topic, whereas in other cases, the most important sentence is the last one [Baxendale 1958]. This observation is used in Edmundson [1969] to score the sentences occurring in paragraphs at the beginning or end of the document (specially the first and last sentences of each of these paragraphs) or bellow some headings like "Introduction", "Purpose" or "Conclusion".

An important issue when applying this method is to study the characteristics of the corpus used. Particularly, in this article, we deal with a collection of journalistic documents. In this case, the most important information usually appears in the top lines, thus the system scores a positive and decreasing value for the three first sentences of the segment.

3.1.4 *Query Method.* In an IR setting, the user's query is a fundamental element that can show his information needs. The purpose is to select sentences with high semantic contents in relation to the query and, in consequence to user information needs. This approach is followed in Tombros and Sanderson [1998],

scoring the sentences depending on the number of words which are also in the query.

Our system employs also the query as the basis for the user modeling. That is, the summaries include sentences containing *query word clusters*, like the clusters of keywords. To score the sentences, similarly to the keyword method, it finds firstly the cluster with more query words. Then, the sentence score is computed dividing the square of the query-word number of this cluster between the total number of words in the cluster.

3.2 Summarizing the Differences

A summary of each document oriented to highlight the differences with respect to the rest of documents in the cluster is produced. Sentences are scored using heuristics related to the significance of the segment they belong to and to the significance of each sentence within the segment. Previously, CSIS redundancy factor [Radev et al. 2000] is used to discard sentences which repeat information of other ones contained in the similarity summary.

For our interests, the significance of a segment depends on the originality of information that it contains with respect to the information contributed by the similarity summary. Nevertheless, this significance rate must consider the relevance of the segment regarding the rest of the document. Thus, we attain to avoid that segments with a limited relevance in the document stand out in the summary, trying in this way to balance the novelty of the information and its relative significance in the document.

The originality or dissimilarity of the segment information is computed as one minus the cosine between the vectors representing the segment himself and the centroid of the most cohesive segment of the cluster of documents. The relevance of the segment is measured as its relative length among the whole document.

Once the relevance of a segment is measured, the final score of each sentence in the segment is computed adding the values obtained by keyword and location methods. The keyword method works as the centroid method unless, in this case, the ten words with greater $tf \cdot idf$ in the segment are used instead of the words with greater weight in the centroid. Location method, as in the case of similarity summaries, provides a positive and decreasing value for the three first sentences of each segment.

In the case of difference summaries, title and query methods are discarded because the goal in this kind of summaries is to detect the relevant sentences of the documents which contribute with new facts regarding to the common information in the cluster of documents. In this sense, we think that the titles of a cluster of very related documents are focused mostly in the main information and then, they are not useful to identify other minority but remarkable aspects. On the other hand, the query is the element used to retrieve the set of documents to summarize, that is, query words are a common factor in the cluster of documents and therefore they can not be used to detect differences.

The final score in both kinds of summaries is computed as a lineal combination of the partial scores resultant of applying the different selection methods.

The weight of each one of the heuristics was set *ad hoc* using a subcollection of the entire corpus of documents used in the experiments. The three sentences with highest scores are extracted and placed in the summary in the same order in which they appear in the original text.

4. TESTING THE WORKING HYPOTHESIS

The final goal of the experiment is to evaluate how the organization of search results and the use of multidocument summaries affects user efficiency finding relevant documents in an IR setting. The working hypothesis is that using jointly clusters and summaries will help users to feel less overwhelmed by the amount of information and therefore to get a better understanding of the different aspects of the available information in the retrieved documents.

This section deals with the experiment design and the obtained results. An analysis and discussion about these results will be also shown.

4.1 Aim

A kind of search task recently explored in the TREC Interactive Track (see Over [1997, 1998] and Hersh and Over [1999]) has been the *instance* or *aspectual retrieval*. The task of human searcher was to save relevant documents which, taken together, covered as many different *aspects* or *instances* of the query (called *topic* in TREC language) as possible. Searchers were encouraged to avoid saving documents which did not incorporate novel aspects beyond those in documents already kept. The task reflects common situations where the searcher is more interested in finding, within the documents of the collection, distinct examples or instances which answer to an information need than finding all the relevant documents to a query, as it is the case in ad hoc retrieval. It seems that a system like the proposed in this article can contribute positively to this kind of search task and therefore may provide a suitable evaluation setting.

The MDS system proposed is based on a clustering process able to group semantically related documents. An indicative summary about the contents of each cluster and other about the particularities of each document is produced. The evaluation of this system could show variations in the effectiveness that may be due to the organization in clusters of the retrieved documents or to the use of summaries. To isolate the effects of each one of these factors, two control systems (A and B) plus the experimental system (C) were evaluated. The control systems A and B correspond respectively to a traditional interface which shows a list of retrieved documents and an organization in clusters of closeness documents.

In conclusion, the experimentation will provide answer to the following questions:

(1) Does the organization of retrieved documents in clusters improve the retrieval effectiveness?
(2) Do the indicative summaries of clusters and documents improve the retrieval effectiveness?

Table III. Topic Titles of the Interactive Track at TREC-7

352i	British Chunnel impacts
353i	Antarctic exploration
357i	territorial waters dispute
362i	human smuggling
365i	El Nino
366i	commercial cyanide uses
387i	radioactive waste
392i	robotics

4.2 Methodology

The experimental methodology followed is basically the described in TREC-7 [Over 1998]. The collection of documents to be searched was taken from the Financial Times of London 1991–1994 which contains 210,158 documents (newspaper articles) totaling 564 MB. The mean number of terms per document is 412.7. Using this collection, users were asked to carry out eight interactive searches. Each search represents a particular need of information described by a TREC topic and matches with multiple instances within the document collection. Table III collects titles of the eight topics.[4]

The Smart [Buckley 1985] information retrieval system was used as the indexing and searching engine. Documents and topic titles were indexed using, respectively, Smart *ntc* and *atc* weighting methods. For each topic, the first 300 documents in the ranking of retrieved document were selected to be used by the three systems. In the systems that require clustering, the *bisecting K-means* algorithm implemented in Cluto was used (see Section 2.1). The number of clusters was set to ten for all the topics.

For evaluating the effects of search results organization and summaries integration, the implementation of three different systems has been required. System A is based on ranked list traditional interface. System B groups similar documents in clusters. System C adds a summary of the similarities for each group and a summary for each document highlighting its particularities regarding the rest of documents in the cluster it belongs to. The interfaces of the three system were designed in such a way that variations among them were minimal. The differences are motivated by the way retrieved documents are organized (document cluster or list) and depending on whether summaries are available or not.

The interfaces are divided in three main panels. In the upper panel, the title of the current topic is shown; by double clicking, users can look up the whole topic description. The left-hand panel is the results area; it presents the retrieved documents or the clusters of retrieved documents, according to the system. The right-hand panel is the content area; it shows document and cluster contents. In the lower part of the left-hand panel a text box is provided to allow users to keep annotations about the identified instances. Documents contributing new instances can be saved checking the boxes close to their titles.

[4]A detailed description of the topics is available at `http://www.itl.nist.gov/iaui/894.02/projects/t7i/spec.html`.

Fig. 2. Interface of system A. Control system based on a ranked list of retrieved documents.

To control user sessions, the searches was restricted to the set of TREC topics and the collection of pre-retrieved documents. Users can finish a search clicking the “Next topic” button, in the right upper corner.

Figure 2 shows the interface of system A. The results area presents the title of each document and its three first sentences. Due to the fact that documents are articles, it is expected that this leading text is a clue about the subject of the document. A click on the title displays the entire document in the scrollable panel of the left-hand side.

Figure 3 shows the interface of system B. The results area displays a list of groups of documents. The groups are identified by a number and sorted according to its best ranked document. The system provides a description for each group: number of documents in the group, title of the top document in the group and the ten centroid words with the greatest weights. The list of documents in a group is displayed in the upper panel of the content area after a click on the group identification. The document content is shown in the lower panel.

Figure 4 presents the interface of system C. In this case, the changes do not affect the structure of the interface but the descriptions of clusters and documents. Then, the possible variations in the effectiveness among systems B and C can be allotted to the presence of summaries and not to any other cause. Clusters are characterized by the summary of similarities that share

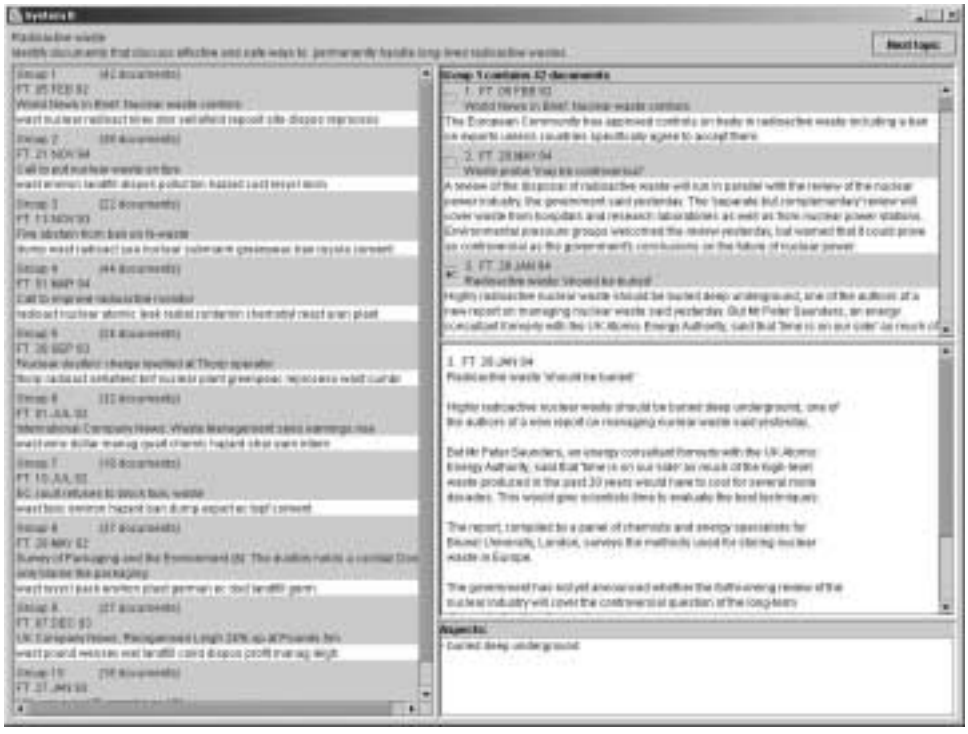

Fig. 3. Interface of system B. Control system based on a post-retrieval clustering.

their documents. As we have said in Section 3.1, this summary is generated starting from the most representative document segment in the cluster. The title of this document is also shown as part of the description. Note that the selection of a document as the most representative of a cluster does not guarantee the top position in the ranked list. For example, in Figure 4, the document most representative for cluster 1, "Waste probe 'may be controversial'", is ranked in second position, as it is showed in the left-hand panel. The summaries of differences are used as additional information to document titles for replacing the selection of three first sentences employed in systems A and B.

Each user carried out eight searches using the TREC topics showed in Table III. Two of the systems were used with three topics and the other system with the two remaining ones. Before starting the first search users were instructed about the task to be done[5] and the three available interfaces. A tutorial using three topics different than those selected to the experiments and some explanations allowed searchers to familiarize with the systems. Subjects had a maximum of fifteen minutes to complete a search but users were allowed to finish a search and pass to the next topic before this time was used up. Users were asked to fill in a presearch questionnaire, a post-topic questionnaire after completing each topic, a postsystem questionnaire after completing the last

[5]The instructions about the task given to the searchers are available at `http://www.itl.nist.gov/iaui/894.02/projects/t7i/spec.html`.

Fig. 4. Interface of system C. Experimental system based on clustering and multidocument summarization.

topic on each system, and an exit questionnaire at the end of their session.[6] During each search session, every significant event was automatically logged and time-stamped. Participants did not know which interfaces constituted the control systems and which one was the experimental system.

The order in which searchers explore topics and systems had to ensure the independence of results of this order. A 8×8 Latin square matrix guarantee that each topic is handled in a different position each time. The repetition of this matrix three times and the allocation of systems using a 3×3 Latin square matrix assure a uniform distribution of the order in which the systems are used. The rows of the final matrix were distributed at random between searchers. In this way, the possible changes observed in the results of the experiment can be attributed to the variations between systems.

The arrangement described required a minimum of 24 participants. The subjects were volunteers recruited at the schools of computer science (18 subjects) and translation (6 subjects). All volunteers were undergraduate students, except a computer science professor. The average age was of 23 years old and the average experience in online searches was of 4.6 years. All subjects were native Spanish speakers and declared, in average, a fluency reading English of 3.4 over 5.

[6]All questionnaires are available at `http://mariner.rutgers.edu/tipster3/trec7.html`.

Table IV. Comparison of Effectiveness Systems Across Subjects

		352i	*353i*	*357i*	362i	365i	*366i*	*387i*	392i	All topics	*Hard topics*	Easy topics
Instance recall	List	0.098	0.216	0.212	0.240	0.771	0.000	0.194	0.358	0.261	0.144	0.456
	Cluster	0.156	0.273	0.317	0.167	0.661	0.232	0.306	0.288	0.300	0.257	0.372
	Summary	0.183	0.227	0.240	0.198	0.630	0.304	0.333	0.250	0.296	0.258	0.359
Instance precision	List	0.506	0.334	0.237	0.885	1.000	0.000	0.350	0.812	0.516	0.285	0.899
	Cluster	0.479	0.423	0.503	0.698	0.813	0.663	0.683	0.549	0.601	0.550	0.686
	Summary	0.646	0.386	0.303	0.575	0.810	0.604	0.705	0.685	0.589	0.529	0.690
Saved instances	List	1.500	1.500	1.000	3.000	2.250	0.000	1.750	4.375	1.922	1.150	3.208
	Cluster	2.750	2.375	3.000	2.000	1.875	1.250	3.875	4.250	2.672	2.650	2.708
	Summary	3.125	2.375	2.250	2.875	2.625	1.875	3.375	3.125	2.703	2.600	2.875
Saved documents	List	3.500	5.375	5.250	3.500	2.250	1.125	4.125	5.750	3.859	3.875	3.833
	Cluster	5.750	5.625	5.750	3.000	2.000	1.875	6.375	7.750	4.766	5.075	4.250
	Summary	4.625	4.750	6.375	6.750	3.750	2.500	5.875	4.875	4.938	4.825	5.125
Read documents	List	9.625	11.250	9.250	9.000	7.625	7.625	7.625	7.500	8.688	9.075	8.042
	Cluster	8.500	10.125	13.125	9.500	2.625	7.000	9.000	8.125	8.500	9.550	6.750
	Summary	4.875	7.125	7.750	9.000	3.875	4.625	7.250	3.875	6.047	6.325	5.583
Browsed clusters	Cluster	3.625	4.250	3.250	5.500	3.375	2.375	2.875	3.000	3.531	3.275	3.958
	Summary	2.250	3.250	3.500	4.250	1.875	3.000	2.250	2.000	2.797	2.850	2.708

Hard topics are shown in italics.

4.3 Results and Analysis

The experiments try to find out the influence of indicative summaries and clustering organization on an interactive retrieval task. These contributions have been measured taken into account the effectiveness improvements and the subjective opinions given by searchers in the questionnaires.

4.3.1 *Effectiveness.* Usually, the evaluation of effectiveness of IR systems is based on recall and precision measures. However, in this kind of experiments, subjects were asked to save documents covering different aspects of a topic instead of all relevant documents. So, in TREC Interactive Track a variation of these measures, called instance recall and instance precision, are used. *Instance recall* is the proportion of established instances covered by saved documents. *Instance precision* is the proportion of saved documents that are relevant to an instance at least. The predefined instances and relevance judgments (which instances contains a document) made by the NIST assessors are used to calculate these effective measures. The experiment focuses in the instance recall because users are encouraged to find out "as many different instances as you can of the needed information for each topic". In any case, the value of instance recall that could be obtained is limited by the number of different instances for each topic that Smart could retrieve among the 300 first ranked documents. The average across topics of the maximum instance recalls for the experiment is 0.77.

Table IV shows a comparison of system effectiveness per topic across all subjects. To test for significant differences among the systems means, we started doing a F-test to establish whether means are equal or not to some confidence level. In cases where there were significant differences, Fisher's least significant difference (LSD) procedure is used to determine which pairs of systems cause these differences. When the normality of the samples could not been supposed or there were statistically significant differences among the standard deviations, we chosen Kruskal–Wallis non-parametric test.

Instance recall of cluster and summary interfaces are slightly greater than in the list system, although these differences are not statistically significant (Kruskal–Wallis test, $H = 1.09$, $p = 0.58$). Then, based on this result, we can conclude that neither the organization of results in clusters nor summaries improve ranked list interface. However, surveying topics individually we can see that list interface improves recall of the other systems only in three topics (362i, 365i, 392i). Like in Wu et al. [2001] we have verified that these topics are the most familiar for searchers.

Familiarity is one of the questions about which subjects express an opinion, over a 5-point scale in post-search questionnaires. So, topics can be grouped in *hard topics*, those with a familiarity lower than 2.5 (352i, 353i, 357i, 366i, 387i), and *easy topics*, those with a familiarity greater than 2.5 (362i, 365i, 392i). Interestingly, this distribution of topics agrees with the recall performance.

Studying separately the recall for hard topics we can see that the differences among cluster and summary interfaces and list interface grow until nearly 80%, which is statistically significant ($F(2, 12) = 3.83$, $p = 0.05$; at 95% confidence level, in pairs list-cluster and list-summary). The evaluation of means of easy topics does not show differences statistically significant ($F(2, 6) = 0.13$, $p = 0.88$).

A valuable conclusion is then that, when topics are unfamiliar for users, there are significant evidences that the cluster or summary systems get better effectiveness than list system. For familiar topics, the interface is irrelevant with respect to differences in instance recall. This is an essential result because our integration model of MDS in an IR interface is based on the organizational capacity of clustering. It seems also to be an evidence of clustering effectiveness in this interactive retrieval environment with subjects, although more work remains to be done in the field. Moreover, we have characterized the kind of topics for which clustering contributed to rise effectiveness of users finding relevant information.

A remarkable example is the topic 366i, *commercial cyanide uses*, which is the topic with lower average familiarity (1.21) and for which users of list interface are not even able of saved a relevant instance, getting a instance recall of 0. The most familiar topic is 392i (2.83), *robotics*, which is also the topic for which list interface achieves the greatest improvement with respect to the best of the other systems. However, only in the case of the list interface has a correlation, at a 90% statistically significant level, between familiarity and recall has been verified. Table V shows correlation between recall and familiarity for all the systems.

Concerning instance precision, Table IV shows, considering all topics, scarce differences between systems that are not statistically significant (Kruskal–Wallis test, $H = 0.38$, $p = 0.83$). However, as in the case of instance recall, the differences became statistically significant when hard topics are considered separately ($F(2, 12) = 4.12$, $p < 0.05$; at 95% confidence level, in pairs list-cluster and list-summary). For easy topics, the differences among means are not statistically significant ($F(2, 6) = 3.31$, $p = 0.11$). Also in this case, the performance of list system is highly correlated with the familiarity of topics.

Table V. Correlation Coefficients among Effectiveness Measures and Familiarity of topics per System

	List	Cluster	Summary
Instance recall	0.64 (90%)	0.30	0.19
Instance precision	0.84 (99%)	0.18	0.36
Saved instances	0.91 (99%)	0.55	0.70 (90%)
Saved documents	0.51	0.38	0.48
Read documents	−0.11	−0.17	0.04
Browsed clusters	—	0.40	−0.24

Percentages among parentheses stand for the statistical confidence level of the regression.

Table VI. Correlation Coefficients among Effectiveness Measures and Satisfaction of Search Results per System

	List	Cluster	Summary
Instance recall	0.42	0.19	−0.06
Instance precision	0.46	−0.18	−0.14
Saved instances	0.70 (90%)	0.76 (95%)	0.46
Saved documents	0.83 (95%)	0.67 (90%)	0.67 (90%)
Read documents	−0.01	0.33	0.18
Browsed clusters	—	0.16	−0.09

Percentages among parentheses stand for the statistical confidence level of the regression.

A similar scenario can be found in the number of different instances saved. Again the differences are only significant for the hard topics ($F(2, 12) = 6.05$, $p < 0.05$; at 95% confidence level, in pairs list-cluster and list-summary).

The number of saved documents per topic can be considered an effectiveness measure complementary to instance recall and precision. Users could save documents containing instances unidentified by assessors or which have not been considered relevant to the topic. In this way, a greater number of documents saved could point out a better effectiveness of the system, at least from the user point of view. The high correlation among number of saved documents and satisfaction in the search results, shown in Table VI, seems to support this idea. Satisfaction is valued by users in the postsearch questionnaire.

Although summary system is ahead of the others systems in the number of saved documents, these differences are not statistically significant in any case (all topics or hard topics).

Two of the events logged during search sessions were the clicks on document and cluster titles. That is, respectively, the full text of documents which have been accessed, and presumably read, and the clusters which contents have been explored by users. Both measures are related directly with the quality of cluster and document summaries. Here, quality is defined as the indicative capacity of summaries to help users in an interactive retrieval system. So, a low number of read documents or browsed clusters would prove, in this sense, a great utility of summaries or, at least, greater than first sentences or centroid keywords, respectively.

Table IV shows that when summary system is used, a reduction in the number of read documents, close to 30% with respect to the two other systems in

average for all topics, is achieved. A F-test confirms that these differences are statistically significant at a 95% confidence level ($F(2, 21) = 3.58$, $p < 0.05$)), being the differences between means significant in the pairs list-summary and cluster-summary at 95% confidence level.

Regarding the number of browsed clusters, summary system shows also a better behavior than cluster system, decreasing by 20% in average for all topics. However these difference are not statistically significant ($t(14) = 1.62$, $p = 0.13$).

In conclusion, experiment results show a significant superiority of the cluster and summary system in instance recall, at least regarding hard topics. Then, the answer to the first question which we outlined in Section 4.1 would be that thanks to the organization in clusters a significant improvement of instance recall, reaching 80% in our experiments, is achieved when users are not very familiar with the topic search. In other cases, our experiments show that the differences between ranked list and cluster interfaces are not significant.

With respect to the summary system, the experiments show that the presence of summaries cause a reduction in the number of browsed clusters and read documents. In the last case, these differences reach 30% and are statistically significant. Moreover, this reduction in the amount of information read by users is attained without losing instance recall. Then, the answer to the second question in Section 4.1 would be that summaries highlighting particular aspects of each document improve the effectiveness since they reduce the number of accesses to full text. As we have hypothesized in the introduction, the results obtained prove that these summaries allow to get a better understanding of the different aspects of the information available in the retrieved documents. At least, this understanding is better than when first sentences of documents, which constitute the usual scenario. Summaries of common information in a cluster also achieves a reduction in the number of clusters explored, in this case above 20%, but the differences with the use of centroid keywords are not statistically significant.

The relevance of these reductions lies in the amount of time that users could save. Furthermore, if summary system is applied to searches in the Internet the saving in time is even greater because of the foreseeable slow access to documents and clusters.

4.4 Questionnaires

As a complement to objective measures, in this section we present subjects' responses to questionnaires done during the experiment session. Table VII shows the averaged results across subjects for the responses to the following questions in the posttopic questionnaire: "Are you familiar with this topic?", "Was it easy to do the search on this topic?", "Are you satisfied with your search results?", "Are you confident that you identified all the different instances for this topic?". Subjects express an opinion, over a 5-point scale, where 1 stands for "not at all", 3 for "somewhat" and 5 for "extremely". As can be seen, summary system outperforms the other two systems in all the issues, although the differences between systems are not very important.

Table VII. Subjects' Responses per System to Posttopic Questionnaires Averaged across Topic and Subject

	List	Cluster	Summary
Familiarity	2.09	2.16	2.33
Easy search	2.73	2.97	3.11
Satisfaction	2.61	3.02	3.02
Confidence	2.50	2.70	2.77

Table VIII. Subjects' Responses per System to Exit Questionnaires Averaged across Subject

	List	Cluster	Summary
Easy learn to use	1.21	2.29	2.33
Easy to use	1.38	2.25	2.13
Like the best	2.25	1.88	1.88

Table VIII shows subjects' opinion about some questions in the exit questionnaire: "Please rank the three systems in order of how *easy they were to learn to use*", "Please rank the three systems in order of how *easy they were to use*", "Please rank the three systems in order of *which system you liked the best*". This time, 1 means the preferred system and 3 the worse system. The results reveal that users meet some difficulties in learning and using cluster and summary systems. This is an expected result because the greatest experience of users of on-line search system comes from Internet search engines (more than 4 years in average), where ranked list is the predominant kind of interface. However, their preferences go to clustering and summary system over ranked list system. These opinions about the difficulty of learning and use would point out that a larger training in summary system can positively influence effectiveness.

5. CONCLUSIONS AND FUTURE WORK

We have argued that text summarization used jointly with post-retrieval clustering could be a viable and effective alternative to the classical interface of IR systems based on ranked lists. The results obtained in an experiment where users carried out an interactive search task are encouraging since they support the initial hypothesis mainly in two ways.

On the one hand, clustering of retrieved documents has been shown to be an effective technique that improves instance recall for a subset of the selected topics. An analysis of instance recall and users outlook on topics has allowed to find out a relationship between the perception that users have about the difficulty of topics and clusters performance. Instance recall is significantly improved over that provided by the list based system in topics which users consider as less familiar.

On the other hand, experiments have shown an enhancement in the indicative power of summaries above first sentences and keywords. This is expressed in the reduction of the number of access to cluster contents and full texts of documents. Only in the former case is the reduction statistically significant.

Therefore, users of a Web search engine could save a considerable time in reading and waiting for documents.

This article has presented an innovative application of MDS to an IR system interface using new summarization techniques. Common and particular information of a group of semantically related documents is presented in two separate summaries. Cluster summaries try to be indicative of the central subject shared by the group of documents. Document summaries are focused on relevant differences that characterize each document with respect to the common information provided by the cluster they belong to. To recognize common and original subjects in a group of documents it is necessary to know which topics are discussed in each document. TextTiling [Hearst 1997] is used to discover topical structure of each document.

Cohesion among topics in a cluster is represented by a similarity graph. The nodes are the segments of the cluster and an edge linking two nodes means a similarity between both segments higher than some threshold. The most connected segment is considered to contain salient information of the cluster and is used as the only source for the similarity summary. This reduction to one in the number of sources implies an evident simplification of the task and avoids the usual problems of inconsistency and redundancy suffered by MDS.

The relevance of the segment, with respect to the rest of topics in the document, next to dissimilarity between a document segment and the central subject of the cluster of documents it belongs to, are the criteria in the generation of document summaries. The goal is to balance the originality provided by the document topic and its relative significance in the document itself.

Due to the fact that our proposal of application of MDS to an IR system interface is based on the availability of groups of similar documents, an analysis of clustering techniques has been conducted. Prior experiments showed that clustering algorithms could not group documents dealing with different aspects in separate clusters. Then a comparison among clustering algorithms was carried out with the goal of selecting one that minimized the number of clusters with some relevant document and, at the same time, maximized the accuracy of the relevant class. Results obtained over the collection of documents and topics selected for the experiment with users are also presented in this article.

Experiments in closed environments, that is, predefined collection of documents and topics or restricted interfaces, like the one presented in this article allow to measure easily the contribution of a particular technique with respect to some baseline. In this way, our work shows that interfaces combining the capacity of organization due to clustering with the indicative information added by summaries can improve the retrieval effectiveness. However, an experimental environment like the one discussed in this article leads to limitations related to the collection of documents and queries used and the profiles of the participating subjects. So, the promising results encourage future experiments that study the behavior of the combination clustering-summarization in more real scenarios. Furthermore, different clustering and summarization techniques may be contrasted in experimental settings as we have presented in this article.

ACKNOWLEDGMENTS

Thanks to Marti Hearst whose observations, specially her suggestions about the kind of evaluation that might be carried out, were very useful. We are also in debt with Alma Gómez that made innumerable comments that helped us a lot to improve this work. Finally, we would like to thank everybody who participated in the experiment.

REFERENCES

ABRAÇOS, J. AND LOPES, G. P. 1997. Statistical methods for retrieving most significant paragraphs in newspaper articles. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th Meeting of the Association for Computational Linguistics, and the 8th Conference of the European Chapter of the Assocation for Computational Linguistics* (Madrid, Spain). I. Mani and M. T. Maybury, Eds.

ANDO, R., BOGURAEV, B., BYRD, R., AND NEFF, M. 2000. Multidocument summarization by visualizing topical content. 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, Wash.).

BAXENDALE, P. B. 1958. Man-made index for technical literature—An experiment. *IBM J. Res. Develop. 2*, 4, 354–361.

BUCKLEY, C. 1985. Implementation of the Smart information retrieval system. Tech. Rep. 85-686. Cornell University.

CAREY, M., KRIWACZEK, F., AND RÜGER, S. 2000. A visualization interface for document searching and browsing. In *Proceedings of CIKM 2000 Workshop on New Paradigms in Information Visualization and Manipulation* (Washington, D.C.).

EDMUNDSON, H. P. 1969. New methods in automatic extracting. *J. ACM 16*, 2 (Apr.), 264–285.

FULLER, M., KASZKIEL, M., NG, C., WU, M., ZOBEL, J., KIM, D., ROBERTSON, J., AND WILKINSON, R. 1998. Ad hoc, speech, and interactive tracks at MDS/CSIRO. In *Proceedings of the 7th Text REtrieval Conference (TREC-7)* (Gaithersburg, Md.). 465–474.

GOLDSTEIN, J., MITTAL, V. O., CARBONELL, J., AND CALLAN, J. P. 2000. Creating and evaluating multidocument sentence extract summaries. In *Proceedings of the 9th ACM International Conference on Information and Knowledge Management (CIKM)* (Washington, D.C.). ACM, New York, 165–172.

HEARST, M. A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computat. Ling. 23*, 1, 33–64.

HEARST, M. A. AND PEDERSEN, J. O. 1996. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (Zurich, Switzerland). ACM, New York, 76–84.

HERSH, W. AND OVER, P. 1999. TREC-8 interactive report. In *Proceedings of the 8th Text REtrieval Conference (TREC-8)* (Gaithersburg, Md.). 57–64.

JARDINE, N. AND VAN RIJSBERGEN, C. J. 1971. The use of hierarchic clustering in information retrieval. *Inf. Stor. Ret. 7*, 217–240.

KAN, M., MCKEOWN, K. R., AND KLAVANS, J. L. 2001. Domain-specific informative and indicative summarization for information retrieval. In *Proceedings of the Workshop on Text Summarization, 24th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, La.). ACM, New York.

KARYPIS, G. 2002. *Cluto: A Software Package for Clustering High Dimensional Datasets. Release 1.5*. Department of Computer Science, University of Minnesota.

KRISHNAIAH, P. R. AND KANAL, L. 1982. *Classification, Pattern Recognition and Reduction in Dimensionality: Handbook of Statistics*. Vol. 2. North-Holland Publishing Company, Amsterdam, The Netherlands.

LEUSKI, A. 2001. Evaluating document clustering for interactive information retrieval. In *Proceedings of 10th International Conference on Information and Knowledge Management (CIKM'01)*. 33–40.

LUHN, H. P. 1958. The automatic creation of literature abstracts. *IBM J. Res. Develop. 2*, 2, 159–165.

MAÑA-LÓPEZ, M. J., DE BUENAGA, M., AND GÓMEZ-HIDALGO, J. M. 1999. Using and evaluating user directed summaries to improve information access. In *Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)* (Paris, France). Springer-Verlag, New York, 198–214.

MANI, I. 2001. *Automatic Summarization*. John Benjamins Publishing Company, Amsterdam/ Philadephia.

MITTAL, V. O., KANTROWITZ, M., GOLDSTEIN, J., AND CARBONELL, J. 1999. Selecting text spans for document summaries: heuristics and metrics. In *Proceedings of the Conference of the American Association of Artificial Intelligence (AAAI'99)*.

NAKAO, Y. 2000. An algorithm for one-page summarization of a long text based on thematic hierarchy detection. In *Proceedings of the 38th Meeting of the Association for Computational Linguistics*. 302–309.

OVER, P. 1997. TREC-6 interactive report. In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)* (Gaithersburg, Md.). 73–82.

OVER, P. 1998. TREC-7 interactive track report. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)* (Gaithersburg, Md.). 65–72.

PAICE, C. D. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Inf. Proc. Manage. 26*, 1, 171–186.

RADEV, D. R., JING, H., AND 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, Wash.).

RASMUSSEN, E. 1992. Clustering algorithms. In *Information Retrieval: Data Structures & Algorithms*, W. Frakes and R. Baeza-Yates, Eds. Prentice-Hall International, London, England, 419–442.

RÜGER, S. AND GAUCH, S. E. 2000. Feature reduction for document clustering and classification. Tech. Rep. DTR 2000/8. Department of Computing, Imperial College, London, England.

SAHAMI, M. 1998. Using machine learning to improve information access. Ph.D. dissertation. Computer Science Department, Stanford Univ., Stanford Calif.

SALTON, G. 1968. *Automatic Information Organization and Retrieval*. McGraw-Hill, New York.

SALTON, G., ALLAN, J., AND SINGHAL, A. 1996. Automatic text decomposition and structuring. *Inf. Proc. Manage. 32*, 2, 127–138.

SALTON, G. AND BUCKLEY, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Proc. Manage. 24*, 5, 513–523.

SALTON, G., SINGHAL, A., MITRA, M., AND BUCKLEY, C. 1997. Automatic text structuring and summarization. *Inf. Proc. Manage. 33*, 2, 193–207.

SKOROCHOD'KO, E. F. 1972. Adaptive method of automatic abstracting and indexing. In *Information Processing 71: Proceedings of the IFIP Congress 71*, C. Freiman, Ed. North-Holland, Amsterdam, The Netherlands, 1179–1182.

SPINK, A., JANSEN, B., WOLFRAM, D., AND SARACEVIC, T. 2002. From e-sex to e-commerce: Web search changes. *Computer 35*, 3, 107–109.

STEINBACH, M., KARYPIS, G., AND KUMAR, V. 2000. A comparison of document clustering techniques. In *Proceedings of the KDD Workshop on Text Mining*.

TOMBROS, A. AND SANDERSON, M. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia). ACM, New York, 2–10.

VAITHYANATHAN, S. AND DOM, B. 1999. Model selection in unsupervised learning with applications to document clustering. In *Proceedings of the 16th International Conference on Machine Learning (ICML-99)* (Bled, Slovenia).

VAN RIJSBERGEN, C. J. 1979. *Information Retrieval*, 2nd ed. Buttersworth, London.

WITTEN, I. H. AND FRANK, E. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan-Kaufmann, San Mateo, Calif.

WU, M., FULLER, M., AND WILKINSON, R. 2001. Using clustering and classification approaches in interactive retrieval. *Inf. Proc. Manage. 37*, 3, 459–484.

ZAMIR, O. AND ETZIONI, O. 1998. Web document clustering: a feasibility demonstration. In *Proceedings of the 21st Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia). 46–54.

ZHAO, Y. AND KARYPIS, G. 2001. Criterion functions for document clustering: Experiments and analysis. Tech. Rep. 01-40, Department of Computer Science, University of Minnesota.

Received March 2003; revised July 2003 and August 2003; accepted August 2003