



PERGAMON

Information Processing and Management 37 (2001) 459–484

**INFORMATION
PROCESSING
&
MANAGEMENT**

www.elsevier.com/locate/infoproman

Using clustering and classification approaches in interactive retrieval

Mingfang Wu^{a,b,*}, Michael Fuller^b, Ross Wilkinson^a

^a *CSIRO, Mathematics & Information Sciences, 723 Swanston Street, Carlton, VIC 3053, Melbourne, Australia*

^b *Department of Computer Science, Royal Melbourne Institute of Technology, Melbourne, Australia*

Accepted 19 September 2000

Abstract

Satisfying non-trivial information needs involves collecting information from multiple resources, and synthesizing an answer that organizes that information. Traditional recall/precision-oriented information retrieval focuses on just one phase of that process: how to efficiently and effectively identify documents likely to be relevant to a specific, focused query. The TREC Interactive Track has as its goal the location of documents that pertain to different instances of a query topic, with no reward for duplicated coverage of topic instances. This task is similar to the task of organizing answer components into a complete answer. Clustering and classification are two mechanisms for organizing documents into groups. In this paper, we present an ongoing series of experiments that test the feasibility and effectiveness of using clustering and classification as an aid to instance retrieval and, ultimately, answer construction. Our results show that users prefer such structured presentations of candidate result set to a list-based approach. Assessment of the structured organizations based on the subjective judgement of the experiment subjects suggests that the structured organization can be more effective; however, assessment based on objective judgements shows mixed results. These results indicate that a full determination of the success of the approach depends on assessing the quality of the final answers generated by users, rather than on performance during the intermediate stages of answer construction. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Clustering; Classification; Interactive retrieval

* Corresponding author. Tel.: +61-3-8341-8200; fax: +61-3-8341-8222.

E-mail addresses: mingfang.wu@cmis.csiro.au (M. Wu), msf@mds.rmit.edu.au (M. Fuller), ross.wilkinson@cmis.csiro.au (R. Wilkinson).

1. Structured topics and structuring answers

Individuals have *information needs* that they need resolved. Satisfying a non-trivial information need involves more than simply locating a specific datum; it typically involves collecting information from one or more resources, and synthesizing an *answer* that organizes that information (Wu & Fuller, 1997).

Information retrieval represents one stage in this task: the identification and retrieval of partial or whole documents from a collection. Traditional information retrieval research has focused on this issue: the question of how to efficiently and effectively identify those (partial) documents most likely to be relevant to a specific, focused query. The main TREC ‘ad hoc’ track is representative of this work (Voorhees & Harman, 2000). However, this task is only part of the overall information seeking process, which has as its goal the construction of an answer to the overall information need.

A variation from the standard information retrieval focus is the TREC Interactive Track’s ‘instance retrieval’ task. This task uses an interactive framework to explore query topics whose resolution requires locating multiple independent data items. In direct contrast to the main ad hoc track, where the goal is to locate as many documents as possible that are relevant to the topic, the goal in the Interactive Track is to locate documents that pertain to different instances of the query topic, with no reward for duplicated coverage of topic instances. Interestingly, this task is significantly closer to our own overall goal – the production of complete answers to non-trivial information needs – than the ad hoc track. Although the parameters of the Interactive Track do not state it explicitly, the task of the Interactive Track can be considered to be the production of a single answer for each topic, where an answer consists of multiple sub-components, one per topic instance.

The assessment of a TREC Interactive Track session measures the ability of the interactive subject to identify documents that contain topic instances. It is measured by applying standard recall/precision to an interactive session, where recall equates to the proportion of the known topic instances contained in the documents identified by a subject, and precision to the proportion of the documents identified by a subject that were deemed to contain topic instances. The assessment process, therefore, provides indirect or, more accurately, circumstantial evidence of the effectiveness of the interactive system’s ability to help the subject develop an answer to the information need represented by the interactive topic. That is, it does not evaluate the answer itself (and, in fact, no such answer is explicitly instantiated during the experimental procedure), but it does attempt to determine the potential of the selected information sources – the identified documents – to be used to generate such an answer.

A key point may be drawn from the foregoing. The TREC Interactive Track topics are structured. This, when combined with our goal of structuring and organizing information to form ‘answers’, suggests an interesting working hypothesis: *that organizing information with regard to task structure is helpful to users.*

Intuitively, this makes sense. As previously discussed, the goal of an interactive subject is to locate documents that pertain to as many different instances of the topic as possible. Given that there is no benefit¹ in locating documents that cover previously discovered topic instances, it

¹ In fact, because Interactive Track experiments are conducted within a fixed time limit, it is counter-productive to locate or view documents that only address previously identified instances of a topic.

would seem desirable to organize the candidate documents in such a way that documents addressing different instances of the query topic were separated into different groups. Ideally, the interactive user could then simply select a single representative document from each instance group. Further, these instance groups could help the user to organize the discovered information as components of their final answer.

How, therefore, should the candidate information be organized? The approaches we have chosen to explore are clustering and classification. The remainder of this paper explores this hypothesis through a sequence of thematically linked experiments. The experiments address the use of document clustering and, later, document classification within the context of the TREC Interactive Track; each experiment examines components of the above hypothesis, and leads into the subsequent experiments. The questions that we attempt to answer include:

- Can automatic clustering reflect topic structure?
- Can users recognize good clusters?
- Do users prefer a clustering approach?
- Are users more effective with a clustering approach?
- Are variations in users' mental maps significant for instance retrieval tasks?
- Can users recognize appropriate classification axes?
- Are users more effective with a classification approach?
- Is the use of information delivery strategies that reflect topic structure beneficial?

In Section 2, we address two of the fundamental issues that underlie this approach. One, can we cluster documents based on their relevance to a topic, versus can we cluster documents by their relevance to separate instances of a topic? Two, given that documents have been clustered, are users capable of identifying the cluster or clusters most appropriate to their information needs?

In Section 3, we examine whether clustering helps users carry out an instance retrieval task. The assessed outcome from this experiment raises the question of variations in mental maps from experiment subject to subject, and between subjects and objective assessors; an experiment exploring this issue is presented in Section 4.

After that, we explore the use of simple document classification in place of unguided clustering. Experiments and analysis pertaining to this are presented in Section 5.

Finally, the limitations of the presented work are discussed in Section 6. We conclude with a general discussion in Section 7 on the basic hypothesis and an analysis of the outcomes of all the experiments as a group.

2. Can users recognize good clusters?

Cluster analysis is a method for revealing structure and relationships in data (Kaufman & Rousseeuw, 1990). Clustering is normally used in information retrieval to organize documents in a collection into topic-coherent groups (Rijsbergen, 1979; Salton, 1989). Recently, clustering has been used as an alternate organization of retrieved documents, aiming to help users better understand the retrieved documents and therefore be better able to focus their search (Cutt, Karger, Pedersen, & Tukey, 1992; Hearst & Pedersen, 1996; Rose et al., 1993).

The purpose for which we seek to use clustering is to investigate whether the implicit structure discovered (or inferred) by a clustering method can help users with structured tasks such as

instance retrieval and answer generation. The standard clustering hypothesis is that relevant and irrelevant documents tend to fall into different clusters (Croft, 1978). The TREC instance retrieval task is slightly different from the standard retrieval task in that it introduces a second level of relevance: a document is relevant or irrelevant to an *instance* of a topic. For a given topic, there is a general level of relevance, or *topic relevance*, and relevance to each instance of a topic, or *instance relevance*. For example, two documents that are both relevant to the topic as a whole (topic relevant) may not be relevant to the same instances of the topic (instance relevant). Previous work, including Croft (1978) and Hearst and Pedersen (1996), has indicated the ability of clustering to group documents with respect to topic relevance; such findings are the basis for the clustering hypothesis. To group a set of documents into clusters of documents relevant to different instances of a topic requires clustering with respect to instance relevance. Given a clustering algorithm able to do so, our hypothesis was that the candidate documents retrieved by an instance topic query could be re-organized into a structure that reflected the desired answer, and that such a structure would help users to resolve their information needs more efficiently and more effectively.

There are many variants of clustering algorithms, which fall into one of two groups: hierarchical and non-hierarchical (Frakes & Baeza-Yates, 1992). To our knowledge, there is no previous work that has evaluated clustering algorithms with respect to instance retrieval. Our purpose of using a clustering method is to group and order a set of documents with regard to certain instances. The concept of documents being “about” a certain instance is already inexact, so hierarchical clustering is probably not appropriate in this case (Rose et al., 1993). We thus chose a non-hierarchical, single-pass algorithm. To reduce document order dependence, after the set of cluster centroids had been selected, documents were reassigned to the nearest centroids. The number of clusters was controlled to be between seven and ten; the size of each cluster was not controlled. Within a cluster, documents were ranked according to their similarity to the query. Clusters were ranked according to the similarity to the query of the highest ranked document they contained.

Each cluster was represented by its cluster description. A cluster description was formed from the ten highest-weighted terms from the cluster vector, the five most frequent word pairs from all documents in the cluster, and the titles of the three documents in the cluster that were most similar to the query.

2.1. Experiment I: user-selection of relevant clusters

The first experiment was to investigate how well the implemented clustering algorithm grouped the retrieved documents, and whether users could successfully distinguish clusters likely to contain relevant documents from those not likely to contain relevant documents. (In the TREC Interactive Track, a document is considered relevant if it is relevant to at least one instance of the topic.)

Using the TREC *Financial Times of London 1991–1994* document collection,² we selected eight topics from TREC-7 Interactive Track. For each topic, the MG search engine (Witten, Moffat, & Bell, 1994) was used to retrieve the 300 highest ranked documents from the collection. These documents were then clustered.

² All experiments reported in this paper used this collection.

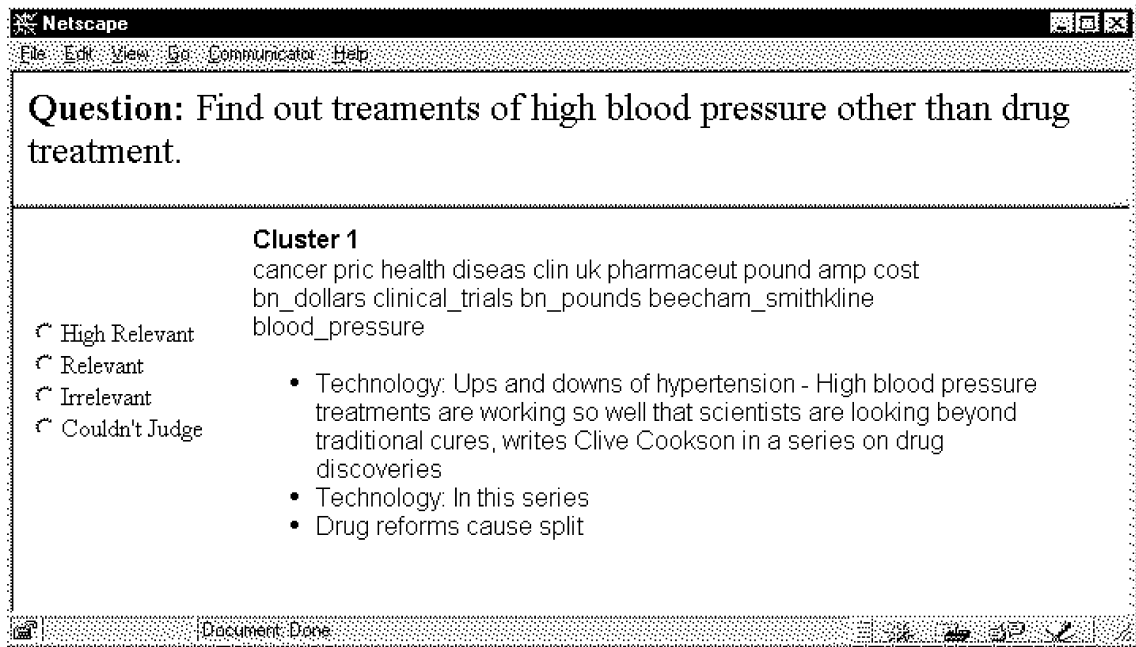


Fig. 1. Part of the interface for the experiment I.

Fig. 1 shows a part of the experiment interface. As in all of our experiments, the clusters were presented as a list of textual cluster descriptions. Although graphical presentation of a cluster structure in 2-D and 3-D has been shown to help users to understand the clustering effect (Swan & Allan, 1998; Leuski & Allan, 1998; Sebrecchts, Cugini, Vasilakis, Miller, & Laskowski, 1999), we wished to focus solely on the two alternate result structures. For this reason, we chose a deliberately simple textual presentation for our experimental interfaces.

Four postgraduate students volunteered to take part in the experiment. Their task was to judge the relevance of a cluster to the topic based solely on the given topic and the only description of the cluster. The possible judgements were: “High Relevant” (the cluster contained instances to the topic); “Relevant” (the cluster was relevant to the topic, but may not contain instances to the topic); “Irrelevant” (the cluster was not relevant to the topic); or “Could not Judge” (the provided information was not enough to make one of above three choices).

2.2. Results and findings

For the eight topics, there were 66 clusters. According to the TREC/NIST assessors’ judgements, there were 16 clusters containing relevant documents; these clusters are here referred as the relevant clusters. Among the eight topics, two of them had only one relevant cluster, four of them had two relevant clusters and the remaining two topics had three relevant clusters. These results confirm the cluster hypothesis with respect to topic relevance. However, the relevant documents were not separated into relevant clusters according to different instance relevance: for all topics, one or two clusters contained all retrieved instances.

Table 1
Summary of relevance judgement of clusters

	Agree, right	Agree, wrong	Disagree
Relevant clusters	9	2	5
Irrelevant clusters	37	0	13

Table 1 summarizes the four subjects' judgements for all clusters. Nine of 16 relevant clusters were judged either "High Relevant" or "Relevant" by all four subjects; two of the relevant clusters (in topic 357 and 392 respectively) were judged "Irrelevant" by all four subjects;³ the subjects disagreed on five of the relevant clusters – they were judged "Irrelevant" by at most two subjects, either "Relevant" or "Could not Judge" by the other subjects.

For 50 irrelevant clusters, 37 were judged "Irrelevant" by all four subjects, while the subjects disagreed on only 13 clusters. Of the latter 13 clusters, only one cluster (in topic 387, which three of the four subjects judged relevant) can be regarded as wrongly judged. For 11 of the other 12 clusters where the subjects' judgement was not unanimous, three of the judgements were "Irrelevant" with the other "Could not Judge" for 11 clusters; for the remaining cluster, the breakdown was three "Irrelevant" judgements and one "Relevant".

This initial experiment showed two things. First, the cluster algorithm could group topic relevant documents, but could not separate documents with different instance relevance. Second, subjects were able to correctly determine from the cluster description which clusters were likely to contain relevant (topic or instance) information, and which were not.

3. Can clustering be used effectively?

Although the clustering algorithm did not group the retrieved documents into instances, subjects were able to successfully judge the potential of a cluster from its description. This suggested that the cluster structure was likely to be an effective way of organizing query results into mostly relevant and mostly non-relevant segments; Hearst and Pedersen (1996) had shown this was the case for ad hoc query topics. The next question is therefore whether the cluster structure is equally effective for an instance retrieval task. In our second experiment, using the same document sets and the same clustering algorithm, the experiment subjects' task became "save relevant documents, which, taken together, covered as many different instances of a topic as possible within a 15 min time limit".

3.1. Experiment II: using clustering for interactive instance retrieval

Two interfaces were implemented: one based on clusters, the other on ranked lists. Given the goal of comparing two alternative organizations of the same data, it was important that the two

³ Interestingly, the two clusters were the only two relevant clusters that were the lowest ranked and therefore last displayed clusters for the respective topics.

interfaces be as consistent as possible, differing only in their presentation of the alternate organizations. The design of the interfaces also assumed that relatively large monitors would be available for the interactive experiments, sufficient to permit side-by-side viewing of documents and result organizations. To minimize variation between searches, no mechanism for providing relevance feedback or for supplying a new query was provided; subjects were restricted to exploring the pool of pre-selected candidate documents.

Although the TREC task description required subjects to save only those documents that covered at least one unsaved instance, we asked our subjects to describe all instances they found in documents. Therefore, in our experiment, a document was saved only when a subject explicitly identified an instance within it.

Fig. 2 shows the interface for the ranked list. The interface was divided into two panels. The left-hand panel displayed a ranked, scrollable list of the titles of the top 300 documents for a topic; each title could be selected by a single clicking. The upper part of the right-hand panel, initially

The screenshot shows a Netscape browser window titled "MDS TREC7 Interactive Experiment - Netscape". The address bar contains "File Edit View Go Communicator Help". The main content area is divided into two panels. The left panel, titled "Topic (example): Find out treatments of high blood pressure other than drug treatment.", contains a "Next Topic" button and a list of retrieved documents. The right panel displays the selected document (FT931-2736) and its content.

Topic (example): Find out treatments of high blood pressure other than drug treatment. Next Topic

A list of retrieved documents:

- FT 18 MAR 93 / Technology: Ups and downs of hypertension - High blood pressure treatments are working so well that scientists are looking beyond traditional cures, writes Clive Cookson in a series on drug discoveries
- FT 29 NOV 94 / Technology: In this series
- FT 07 FEB 92 / Technology (Worth Watching): Relief in sight for asthma sufferers
- FT 14 NOV 92 / Drug reforms cause split
- FT 05 APR 93 / Leading Article: Drugs on trial
- FT 02 JAN 93 / UK Company News: Three groups' shares rise on drugs approval
- FT 29 JUL 94 / Technology: Brains on their minds - Drug researchers are seeking a stroke treatment that could transform current therapy
- FT 02 SEP 94 / Technology: Towards a cure for blindness
- FT 08 FEB 94 / UK Company News: Glaxo asthma drug wins US approval
- FT 01 NOV 94 / UK Company News: British Biotech new cancer drug - Third promising treatment makes company one of best in sector
- FT 07 SEP 93 / Technology: A renaissance in treatment - New drugs to treat schizophrenia are finally becoming available
- FT 20 FEB 93 / New cancer drugs show promise
- FT 20 JUL 92 / Wellcome expects good news on Aids drug tests
- FT 31 MAR 94 / Technology: Deadly challenge proves costly - Daniel Green examines the continuing search for an effective sepsis treatment, in a series on drugs
- FT 03 NOV 93 / International Company News: Merck drug withdrawn
- FT 28 MAR 94 / Glaxo drug rival wins licence for UK
- FT 19 APR 91 / World News in Brief: Drug money fund
- FT 07 DEC 93 / Drug row flares as calls for tighter guidelines grow
- FT 18 MAR 94 / New cancer drug shows promise in first tests
- FT 10 FEB 94 / Clinton in new drugs push
- FT 14 FEB 92 / Technology (Worth Watching): Watching high blood pressure
- FT 28 MAR 92 / Trials and tribulations: Large-scale clinical drugs testing
- FT 08 APR 94 / Effective Aids drugs 'a long way off': Doctors divided over HIV treatment as study casts doubts on leading

FT931-2736
_AN-DCRCHAGAFT
930318

FT 18 MAR 93 / Technology: Ups and downs of hypertension - High blood pressure treatments are working so well that scientists are looking beyond traditional cures, writes Clive Cookson in a series on drug discoveries

By CLIVE COOKSON

Drugs to bring down high blood pressure are one of the great successes of pharmaceutical research. Over the past decade the industry has given doctors dozens of new drugs to treat hypertension - the medical name for the condition - by several different mechanisms. Their sales are worth more than Dollars 10bn (Pounds 7bn) a year, three times as much as the total market for cancer drugs. 'The treatment of hypertension is very good now and the side effects are minor,' says Desmond Julian, medical director of the British Heart Foundation, 'and because there is a range of drugs, you can normally find one to suit any particular patient.'

In industrialised countries, 15 to 20 per cent of the adult population has high blood pressure. Julian says patients with mild or moderate hypertension should not be put on drugs straightaway; their doctors should urge them to make changes in diet and lifestyle. But for the 5 per cent of people with severe hypertension, drugs are usually required to bring blood pressure down to a safe level. Clinical trials have shown that the greatest benefit of hypertension treatment is a 40 per cent reduction in the risk of suffering a stroke, which is caused by the rupture of blood vessels in the brain. The effects on other forms of cardiovascular disease are less clear cut; indeed there is

Add New Aspect

Relevant aspects already identified for this topic:

1. calcium
2. regular exercise
3. biofeedback

Fig. 2. Experiment II: the list-based interface.

blank, displayed a scrollable view of any document selected in the left-hand panel. The lower part of right-hand panel, the instance selection panel, displayed a list of saved instances; subjects could use this panel to record all identified instances for a topic, and to supply a description for an instance (via a pop-up dialogue box).

Fig. 3 shows the interface for the cluster structure. The interface was also divided into two panels. The left-hand panel shows an ordered, scrollable list of cluster descriptions; each cluster could be selected by a single clicking. The right-hand panel was sub-divided into three parts. Compared with the right-hand panel of the list-based interface, it had an additional section in which a scrollable, ranked list of the titles of documents in any cluster selected in the left-hand panel could be displayed; each title in the top part could be selected by a single clicking. The function of the other sections matched that of the corresponding components of the list-based interface.

The screenshot shows a web browser window titled "MDS TREC7 Interactive Experiment - Netscape". The interface is divided into two main panels. The left panel, titled "Topic (example): Find out treatments of high blood pressure other than drug treatment.", contains a list of seven clusters (Group 1 to Group 7). Each cluster is represented by a list of keywords and a list of document titles with their publication dates. For example, Group 1 includes keywords like "cancer pric health diseas clin uk pharmaceut pound amp cost" and document titles such as "FT 18 MAR 93 / Technology: Ups and downs of hypertension - High blood pressure treatments are working so well that scientists are looking beyond traditional cures, writes Clive Cookson in a series on drug discoveries". Each cluster has a link to view more documents. The right panel is titled "A list of documents in Group 1:" and displays a list of document titles for the selected cluster. Below this list, there is a section titled "By CLIVE COOKSON" which contains a paragraph of text discussing hypertension treatments. At the bottom of the right panel, there is a section titled "Relevant aspects already identified for this topic:" with a list of three aspects: "1. calcium 2. snakeroot plant 3. regular exercise".

Fig. 3. Experiment II: the cluster-based interface.

This experiment was the basis for our participation of TREC-7 Interactive Track (Fuller et al., 1998). The experiment design thus followed the Latin Square arrangement as stipulated by the TREC-7 Interactive Track (Lagergren & Over, 1998). In this design, eight search topics were broken evenly into two blocks, the order of four topics was fixed within each block. Each subject was assigned to use one system on a block of four topics and then use the other system on another a block of four topics. This design is intended to minimize the effect of inter-subject and inter-topic variations, making it possible to focus solely on inter-system variations. The arrangement required a minimal number of four subjects because of the sequence and the combination of two blocks and two systems. We augmented this design by adding another three groups of four subjects.

Sixteen paid subjects were recruited via an internal university newsgroup. All of them were undergraduate computer science students, aged from 17 to 23, and had an average 3.3 years of online search experience.

When subjects arrived at the experiment site, they completed a pre-search questionnaire and psychometric test. The subjects were then given a quick demonstration of the main functions of each interface. During the experiment, prior to using a system for the first time, the subjects attempted an example topic to familiarize themselves with its interface; they were free to ask any question at this point. Each subject was required to fill in a post-topic questionnaire after completing each topic, a post-system questionnaire after completing their three allocated topics on each system, and an exit questionnaire at the conclusion of their session. Subjects were permitted up to 15 min to complete each topic; at the 15th minute mark they were informed that the time allocated had expired and were directed to complete their current action, complete the appropriate questionnaire, and move on to the next topic. All actions time-stamped outside the allocated 15 min were discarded. During each search session, every significant event was automatically logged and time-stamped. Participants were not informed which interface was the control system and which was the experimental system.

3.2. Results and findings

3.2.1. Is clustering approach more effective than the ranked list?

Evaluation of a system within the TREC Interactive Track is based on instance recall and instance precision of the saved documents. Instance recall is the fraction of total known instances (as determined by the assessor) for the topic that are covered by the saved documents, instance precision is the fraction of the saved documents which contain at least one instance. Here we focus on instance recall. Those documents where a subject saved at least one instance were identified from search logs. The instance relevance judgements of the TREC/NIST assessors were then used to determine the instance coverage of each such document. From this, the average instance recall across all subjects, per topic, per interface could be calculated. Table 2 shows the average instance recall across all subjects per topic. As it can be seen, the average instance recalls of two interfaces are very close, with no statistically significant variation present ($P < 0.05$).⁴

⁴ A two-tailed *t*-test has been used to determine the statistical significant difference throughout, unless otherwise stated.

Table 2

Instance recall per topic of the saved documents, as judged relevant by assessors

	352	353	357	362	365	366	387	392	Mean
Cluster	0.089	0.102	0.231	0.146	0.687	0.214	0.111	0.184	0.221
List	0.053	0.068	0.279	0.135	0.677	0.071	0.250	0.285	0.227

Table 3

Subject–system interaction

		352	353	357	362	365	366	387	392	Mean
Instances saved by subjects	Cluster	3.750	3.250	4.375	4.125	4.000	2.625	3.125	6.750	4.000
	List	2.875	2.625	5.375	6.125	2.750	1.625	3.375	8.750	4.188
Documents saved	Cluster	3.750	3.375	4.000	4.125	1.875	2.750	3.250	5.875	3.625
	List	2.750	2.625	5.250	4.750	1.625	1.750	3.250	7.250	3.656
Documents read	Cluster	15.00	19.75	14.75	20.63	8.375	15.50	19.88	17.38	16.41
	List	15.13	19.88	14.38	17.00	5.000	19.63	19.25	15.25	15.69

So based on the instance recall, the hypothesis that “the clustering approach is more effective than the ranked list” is rejected.

Although there is almost no difference in terms of average instance recall between two interfaces, inspection of the topic-by-topic results suggests that there is in fact a variation in performance for a subset of the topics. Table 2 shows that, for the five topics (352, 353, 362, 365, and 366) for which subjects using the list organization saved fewer instances, the subjects using the cluster organization saved more instances, especially for topic 366 and 353. Conversely, for the three topics (357, 387, and 392) for which subjects using the list organization saved more instances, fewer instances were saved by the subjects using the cluster organization.

Interestingly, the topic in which subjects of the cluster interface most out-performed subjects of the list interface – topic 366 – is also the topic with which all subjects claimed to be least familiar (see Section 3.2.2, for the definition of familiarity), while the topics with which subjects claimed to be most familiar – topics 387 and 392 – are also the topics in which subjects of the list interface most out-performed subjects of the cluster interface. However, no correlation between familiarity and system performance could be detected at a statistically significant level.

Table 3 shows the interaction between the subject and the system. On the average, subjects read more documents from cluster interface than the list interface (List $M = 15.69$ and Cluster $M = 16.41$), saved similar number of documents (List $M = 3.656$ and Cluster $M = 3.625$) and similar number of instances (List $M = 4.188$ and Cluster $M = 4.0$). There is also no statistical significance found between two interfaces for any above measure.

No significant correlation was found between instance recall and the number of instances saved by subjects, and between instance recall and the number of documents saved. No significant difference was found between instance recall of individual subjects, or the number of instances saved by individual subjects.

3.2.2. Feedback from subjects

During the experiment, subjects filled in a post-search questionnaire after searching each topic. We extracted three questions from the post-search questionnaire that related to performance and subject–system interaction. The three questions are: “Are you familiar with this topic?”, “Are you satisfied with your search results?”, and “Are you confident that you identified all of the different instances for this topic?”. Each of these questions was measured on a 5-point Likert scale, where 1 = not at all; 3 = somewhat; and 5 = extremely. Table 4 shows the average responses per system for the three selected questions. There was little difference between two interfaces in the subjects’ response to each question.

Table 5 shows the correlation between the selected questions, the performance (instance recall), and the number of documents saved. There is no significant correlation between instance recall and any selected question. The number of documents saved is significantly correlated to the familiarity and satisfaction, but not to the confidence. Familiarity is also significantly correlated to the satisfaction and confidence.

The pre-experiment psychometric test attempted to gauge subjects’ verbal skills in terms of their ability to identify synonyms of eight stimulus words (Ekstrom, French, Harman, & Dermen, 1976). The mean score for the 16 subjects was 23.2 correct of 34.9 total terms, with a standard deviation of 8.1 terms. There appeared to be a linear correspondence between subjects’ score and their average instance recall; no significant correlation was found with performance with either interface.

From the exit questionnaire, 12 of the 16 subjects preferred the cluster-based interface than the list-based interface, and 13 subjects rated the cluster-based interface as easy to use.

A fairly clear preference for the cluster structure was also shown in subjects’ comments, such as:

- It showed me all the list of the topics in a screen.
- It is easy to search and has a topic or summary in each group.
- The group narrows down the scope of the searching task.
- Specific group of articles is associated with one another.

Table 4
Average scores for three selected questions from post-search questionnaire

	Cluster	List
Familiarity	2.344	2.469
Satisfaction	2.890	2.890
Confidence	2.766	2.860

Table 5
Correlation matrix

	Familiarity	Satisfaction	Confidence
Instance recall (assessor’s judgement)	0.17	0.30	0.41
Documents saved	0.52*	0.72*	0.28
Familiarity		0.77*	0.56*

* Correlation is significant at the 0.05 level.

These comments aligned with our motivation of exploring the clustering organization of information. As discussed in Kaufman and Rousseeuw (1990), classification and grouping like things together are some of the most primitive and common activities of human beings. (Miller, 1956) also explores the benefit of grouping as an aid to human information processing.

In contrast, comments on the simple list organization included:

- The list is too long to explore all items.
- Everything was just in a list and it was difficult to concentrate on the actual topic.
- Long list, sometimes frustrated in could not find suitable topic.
- Hard to search, depends on the topic.

However, subjects did note some inadequacies of the cluster structure as implemented, such as:

- The keywords in each group are not clear.
- They will make users confused for the first time.

This is probably reflective of the fact that the terms in cluster descriptions were stemmed, rather than complete, words. We also observed that apparently not all subjects understood the cluster structure. For example, some subjects commented:

- Unclear how groups are determined.
- I did not really understand the way the categories (clusters) were grouped. Perhaps if I did, it would have been better.

This may indicate that subjects needed more training and experience to make best use of the clustering structure.

This experiment showed us that although most subjects liked the cluster structure, the overall performance of using two structures was very similar. Subjects tended to browse all clusters, but generally saved documents from clusters that contained instances (as determined by the NIST assessors); in contrast, subjects generally did not save documents from clusters that contained few topic instances. While we hope that the subjects may have gained a greater understanding of what the collection, as a whole, contained about a topic, this understanding did not help them to complete the instance task more successfully than the list-based approach did.

4. Are variations in mental maps significant?

In the experiment presented in Section 3, the subjects filled the role of search intermediary. They were given an information need, then searched for documents that were relevant to that information need. Analysis of the results revealed a substantial disagreement as to what was and was not a relevant instance between the experiment subjects and the TREC/NIST assessors. Subject-determined instances sometimes covered multiple assessor-determined instances; the reverse was also true. Subjects also perceived some relevant instances in documents that the assessors felt did not contain any relevant instances.

This kind of conflict in judgement has also been seen in other previous TREC investigations (Voorhees, 1998); it is just more conspicuous for the instance finding task, as the determination of what is an instance appears a more open, subjective decision. However, in contrast to the main TREC ad hoc Track – where the impact of this phenomenon has been shown to be minimal (Zobel, 1998) – in the context of the TREC Interactive Track with its relatively shallow pool of relevant documents and instance judgements, it may make the calculation of instance recall and

precision unstable, with a concomitant deleterious effect on confidence in quantitative evaluation using those judgements. This is an open question.

A possible explanation is that we were observing a variation in mental maps; each user had a particular view of how the topic should be split into instances, and where the dividing lines should be drawn between those instances. Because each individual's map differed, there could be no consistent correlation between the instances identified by users and those identified by assessors.

Our next experiment sought to prove or disprove the presence and impact of variations in mental maps. Rather than trying to match each subject's instances with those of the assessors, we chose to have the subjects use the instance framework provided by the assessors. By presenting the instances identified by the assessors as the definitive partition of a topic into its sub-topics, we could simplify the users' task to that of identifying occurrences of each sub-topic. In this experiment, we tried to investigate:

- Is there any significant judgement difference between assessors and subjects if the subjects are provided the assessor instance sets?
- Can the evaluation of two experiment systems be made more reliable based on assessor's instance sets?

We therefore made the following changes to the experimental interfaces:

Rather than asking subjects to identify instances, we provided them with the instances identified by the TREC/NIST assessors. These instances were permanently listed in the right-hand panel of both experimental interfaces; an example is shown in Fig. 4. When the text of a document was being displayed, the list of instances was active: during that time, subjects could select from the pre-determined list those instances that they considered were present in the visible document. In keeping with the Interactive Track goal of avoiding repeated instances, once an instance had been marked as present, it was removed and added to a list of discovered instances (see Fig. 4).

Because the right-hand panel permanently displayed the assessor-identified instances (to reinforce the foreign mental map that the experimental subjects were being asked to explore), the left-hand panel was used to present all delivered information. For the list-based interface, the left-hand panel showed the retrieved documents in the ranked order, with each document title a link to its full content. For the cluster-based interface, the left-hand panel showed the cluster descriptions, with each cluster description containing a link to all documents in the cluster; the document titles that formed part of a cluster description were linked directly to the document content. Selecting a document or cluster link caused the appropriate information to replace the content of the left-hand panel. As noted, while the text of a document was being viewed, the list of instances in the right-hand panel was active. Once a subject had finished viewing a document, they could return to the list of documents or clusters.

To know how much information a subject can capture during a search, the presentation of the left-hand frame was paged. For the list structure, titles of 20 documents appeared on each page, similar to the style of Web interfaces to search engines such as Alta Vista and Excite. For the cluster structure, the cluster descriptions were also paged, four or five clusters per page. The lists of documents in each cluster were also paged in the same way as the document lists of the list-based interface.

When subjects read a document but did not save any instance, the color of the link to the document changed to red; if the subject read and saved at least one instance from a document, the link to the document changed to red and was ticked.

Topic (357): Identify documents discussing international boundary disputes relevant to the 200-mile special economic zones or 12-mile territorial waters subsequent to the passing of the "International Convention on the Law of the Sea".

Group 1 ▶ Show Group 1 (210 documents)
 issu intern argentin rus minister foreign compan offic se south
 territorial_waters foreign_minister law_sea falkland_islands china_sea
 • [Law of the Sea promises many disputes](#) ✓
 • [Compromise over islands](#)
 • [Australia extends offshore zone](#)

Group 2 ▶ Show Group 2 (15 documents)
 drift fishes vessel tun catch fisherm spain quot span net
 drift_nets fishing_rights bay_biscay barents_sea fishing_illegal
 • [Icelandic trawler detained in disputed Rockall waters](#)
 • [Fleets fight in over-fished waters: Fishing disputes haverisen up the diplomatic agenda. FT reporters examine the conflicts worldwide](#)
 • [Outdoors: Europe's fishing fleets prepare for war - Fishsupplies are running out, but the authorities seem powerless to act](#)

Group 3 ▶ Show Group 3 (31 documents)
 izetbegov nat karadz owen fight belgrad baz krajin republ europ
 bosnian_serbs bosnian_serb forces_serb peace_plan nations_united
 • [Plan for Bosnia stirs disputes: The Moslems object tosuggested boundaries but at least all sides are prepared to discuss theEC-UN proposals](#)
 • [Moslems say sea access agreed: Deal with Croats could clearobstacle to Bosnian peace](#)
 • [UN reports new wave of 'ethnic cleansing'](#)

Group 4 ▶ Show Group 4 (9 documents)
 carg traff southern mountain equival port maca mw construc pearl
 hong_kong economic_zone government kong economic_special china_hong
 • [Survey of Building for Asia's Future \(11\): Another Asiandragon rises](#)
 • [World Trade News: Alternative port in storm over Hong Kong -Simon Davies on the implications of development plans for a mainlandentrepot](#)
 • [Tourism deal on disputed islands](#)

Possible new topic aspects:

- oil exploration / "political" or "adm" boundary may require intl. jurisdiction: Egypt - Sudan
- Kurle islands: Russ - Jap
- Falkland islands: Falkland - Chile
- St.Pierre & Miquelon: Canada - France
- St.Kilda: U.K. - Ireland
- Faroe islands: U.K. - Denmark
- Bahrain - Qatar
- United Arab Emirates - Iran
- So. Georgia & So Sandwich Islands: U.K., Argentina
- Rockall island: U.K., Iceland, Denmark
- Chistmas island: Indonesia - Australia

Previously located topic aspects:

- Spratly, So. China Sea: V.N., China, Phil, Indonesia, Japan, Taiwan, Malaysia, Brunei
- extended territorial waters - delineation of Aegean continental shelf: Grc - TURk

More Groups: Page 1 Page 2

Fig. 4. Interface for experiment III.

4.1. Experiment III: clustering and retrieval of fixed instances

In the TREC-7 experiment reported in the Section 3, eight topics were used. This was likely to be too small a sample to allow statistically significant results to be observed, so in this experiment another four topics from TREC-6 were added.

Sixteen new subjects were recruited to undertake the experiment. The experiment design and procedure were the same as that in the previous experiment, but the subjects attempted six topics using each interface.

In the previous experiment, the average instance recall was 0.646 for the pooled documents, and only 0.312 for the first 20 highly ranked documents; this seemed low. We therefore used Rocchio-based relevance feedback (Salton, 1971) to improve the quality of the set of candidate documents. The average instance recall for 12 topics was 0.879. For the eight queries (352–392) used in TREC-7, the instance recall (at first 300 documents) increased 22.5%.

4.2. Results and findings

Table 6 shows the average instance recall per topic per interface for saved documents from which a subject saved at least one instance. Interestingly, while fewer instances were covered in documents saved by users of the clustering interface than of the list interface according to the TREC/NIST assessors’ relevance judgements, subjects selected more instances using the clustering interface than the list interface. However, the difference in instance recall between two interfaces was not statistically significant in either assessor’s judgement or subject’s selection.

The correlation between assessors’ judgement and subjects’ selection for the clustering interface is highly significant ($P < 0.02$); the correlation for the list interface is not significant. Overall, the correlation between assessors’ judgement and subjects’ selection is highly significant ($P < 0.01$).

Table 6 quantifies the differences in judgement between assessors and subjects in terms of instance recall. Instance recall, as used in the TREC Interactive Track does not directly correlate to the standard TREC ad hoc recall measure. In the TREC Interactive Track, instance recall is defined as “the fraction of total instances (as determined by the assessor) for the topic that are covered by the submitted documents” (Hersh & Over, 1999). That is:

$$\text{TREC instance recall} = \frac{\text{number of relevant instances found in saved documents}}{\text{total number of relevant instances}}.$$

True instance recall, akin to the standard TREC ad hoc recall metric, would be:

$$\text{true instance recall} = \frac{\text{number of relevant instances saved}}{\text{total number of relevant instances}},$$

that is, the number of instances explicitly identified by a subject, divided by the total number of instances. Similarly, whereas

Table 6
Instance recall for saved documents

Topic	Assessors’ judgement		Subjects’ selection	
	Cluster	List	Cluster	List
303	0.518	0.544	0.375	0.589
307	0.217	0.174	0.250	0.201
326	0.583	0.569	0.555	0.528
339	0.775	0.700	0.663	0.425
352	0.183	0.094	0.161	0.125
353	0.239	0.329	0.466	0.511
357	0.404	0.394	0.519	0.442
362	0.208	0.229	0.240	0.219
365	0.797	0.766	0.469	0.214
366	0.250	0.429	0.571	0.462
387	0.042	0.333	0.278	0.389
392	0.163	0.306	0.208	0.257
Mean	0.365	0.406	0.396	0.364

$$\text{TREC instance precision} = \frac{\text{number of saved documents with relevant instances}}{\text{total number of saved documents}}$$

we can say that:

$$\text{true instance precision} = \frac{\text{number of relevant instances saved}}{\text{total number of instances saved}}.$$

Under normal Interactive Track conditions, true instance recall and true instance precision cannot be determined, because only information about which documents were saved, not specific instances, is recorded; further if specific instances were saved, it would still be necessary to reconcile or consolidate the overlapping or conflicting topic instances that could be nominated by each interactive subject. In this experiment, however, neither problem applies: all subjects were working from a single, agreed upon set of possible topic instances, and the exact instances associated with each document were recorded.

Table 7 shows the distribution of average true instance precision and average true instance recall, per topic for each interface. The overall average true instance precision was 0.63 for the cluster interface, and 0.70 for the list interface (no significant difference). The overall average true instance recall was 0.65 for the cluster interface, and 0.66 for the list interface (no significant difference).

In contrast, there was a significant correlation between the TREC instance precision and true instance precision, and between TREC instance recall and true instance recall, suggesting that the standard TREC Interactive Track measures are acceptable substitutes.

The two “true” metrics can also be considered as measuring the level of agreement between the experimental subjects and the TREC/NIST assessors. From this perspective, the data do not show a significant variation in the level of agreement based on interface; further, the overall level of agreement is consistent with that reported elsewhere for similar multi-assessor experiments (Voorhees, 1998; Cormack, Palmer, & Clarke, 1998).

5. Can classification be used effectively?

One of the goals of this work has been to organize retrieved data in ways that reflect the structure of the topic and its answer. Static clustering, as explored in the preceding sections, failed to achieve that as a measurable result. This section describes a different way, classification, of achieving that goal.

Clustering attempts to group documents that are internally similar to each other. An alternative is to group documents based on their similarity to some external set of criteria. One source of such criteria is the query topic itself. By analyzing the query topic, it may be possible to identify an appropriate set of classification constraints that correspond to the desired set of topic instances. This would allow the set of candidate documents (retrieved in response to the query) to be classified into categories that correspond to different instances of the answer. Such classification constraints can easily be automatically identified in the query topic with the help of the interactive users themselves. While users may not know the exact details of the answer to their query, they usually know *characteristics* of the answer that they are looking for. For example, consider the topic: “Which countries import sugar from Cuba?”. While users may not know beforehand that

Table 7
True instance precision and true instance recall

Topic	True instance precision		True instance recall	
	Cluster	List	Cluster	List
303	0.54	0.58	0.50	0.63
307	0.91	0.90	1.00	1.00
326	0.98	1.00	0.92	0.92
339	0.77	0.84	0.67	0.50
352	0.46	0.05	0.30	0.19
353	0.54	0.51	0.80	0.73
357	0.55	0.70	0.80	0.78
362	0.71	0.63	0.71	0.67
365	0.97	0.94	0.56	0.26
366	0.47	0.83	0.70	0.94
387	0.13	0.74	0.50	0.72
392	0.48	0.70	0.41	0.59
Mean	0.63	0.70	0.65	0.66

Russia, Latvia, or Iran are actual instances of the answer, they do know that all instances should be country names. In this example, by using a set of country names (the potential instances) to classify the retrieved documents, subjects may be able to more easily find instances of the topic.

5.1. Experiment IV: classification and instance retrieval

In the classification approach, several possible axes for categorization are identified from the query topic. From the alternative axes, the user interactively selects categorization that is most appropriate. The retrieved documents are then dynamically classified into categories. This approach is shown in Fig. 5, and contains several significant stages: a category generation stage, a category selection stage, and a document classification and ordering stage.

Category generation. The category generator extracts keywords from each query, and uses WordNet (Fellbaum, 1998) to identify a set of hyponyms for each keyword. These hyponym sets form the basis for candidate category sets. We do not attempt to distinguish between alternate senses when identifying sets of hyponym.

Category selection. For a given query, there may be multiple ways of classifying the set of retrieved documents. Automatically determining the appropriate classification axis is in itself a difficult procedure. Instead, we let the user select the categorization appropriate for organizing the retrieved documents. In our implemented system, a window (shown in Fig. 6) shows users the extracted keywords and the sample of their associated categories; users can consider the alternative classifications before selecting the most appropriate.

Classification and ordering. The retrieved documents are then matched against the selected set of categories. Classification is based on ranking the set of retrieved documents by their similarity to the terms describing each category; in our initial implementation, each category is restricted to the 10 highest-ranked documents for that category. Documents may belong to more than one

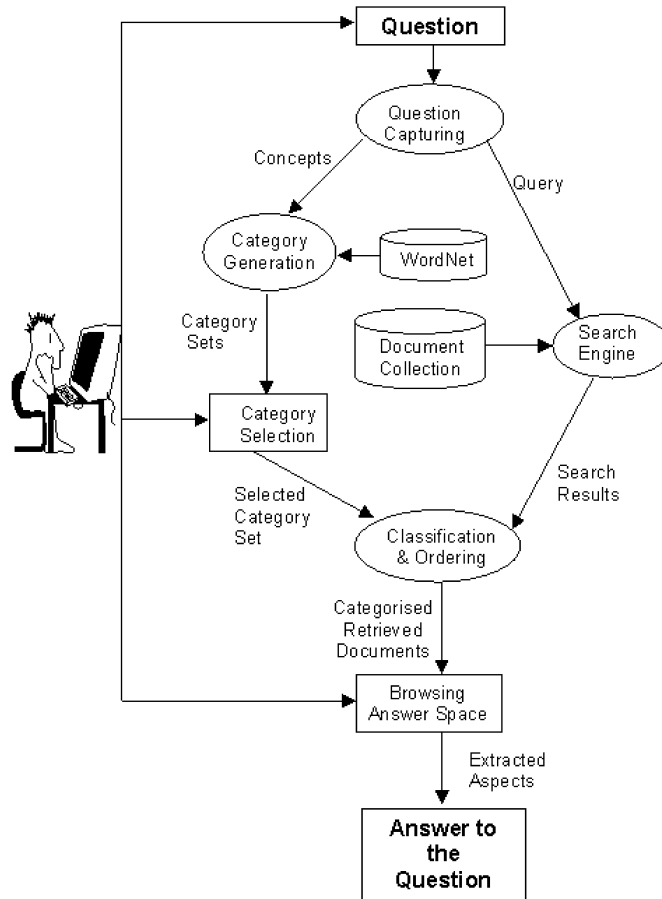


Fig. 5. System architecture for classification approach.

category; those that do not match any specific category are allocated to a category of miscellaneous documents. Within each category, the documents are ordered according to their similarity to the original query. Overall, categories are ranked by the similarity of their first-ranked document to the query.

After the search results have been categorized and ranked, they are presented to a user as shown in Fig. 7. The interface is divided into two panels. In the left-hand panel, the upper frame shows the document categories. Each category is expandable and collapsible; in Fig. 7, the first category is shown collapsed, and the second expanded. The middle frame shows the already discovered instances, along with the saved documents relevant to each instance. A button in the bottom frame enables users to add new categories into which documents may be classified. When any document is selected from the upper-left or middle-left frame, its content is shown in the right-hand panel of the window. Any terms that match the currently expanded category are highlighted in red; terms that match the descriptions of other categories are highlighted in blue. This highlighting is intended to help users more easily locate potential answers from within what may be

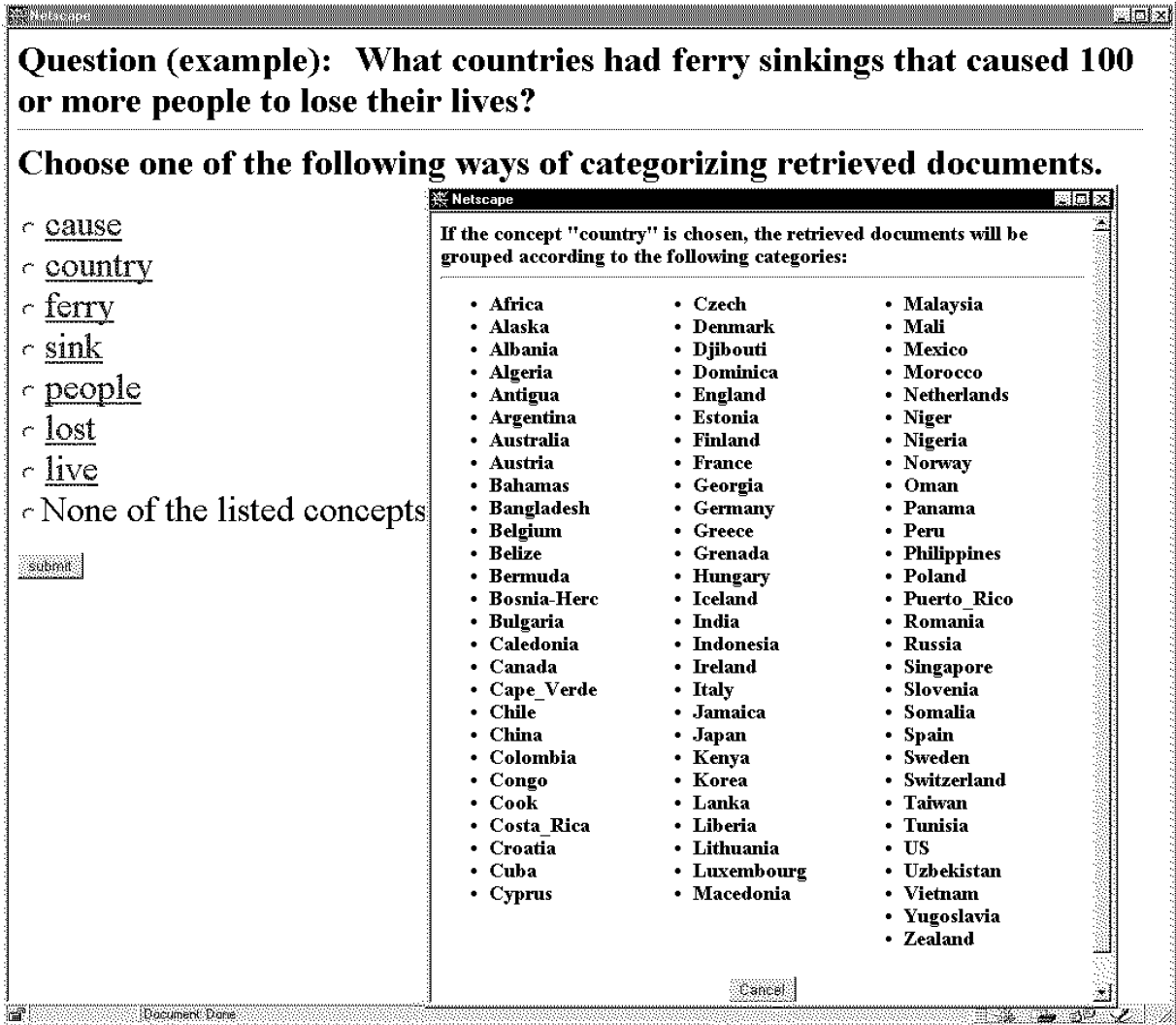


Fig. 6. Interface for category selection.

lengthy documents. When the user finds information relevant to an instance of the topic in a document, they can click on “Save Instances” button. This causes a pop-up window to appear in which the user can note the instances to which the document is relevant. The discovered instances and their associated document are then added to the middle-left frame. Whereas the upper-left frame helps the user to search for information that can contribute to their answer, the information in the middle-left frame helps the user synthesize their answer.

The ranked list interface differed from the classification interface in three ways. One, where the classification-based interface contained a list of expandable categories of retrieved documents (the upper-left frame), the ranked list interface contained a simple ranked list of retrieved documents. Two, the classification-based interface allowed users to interactively add additional

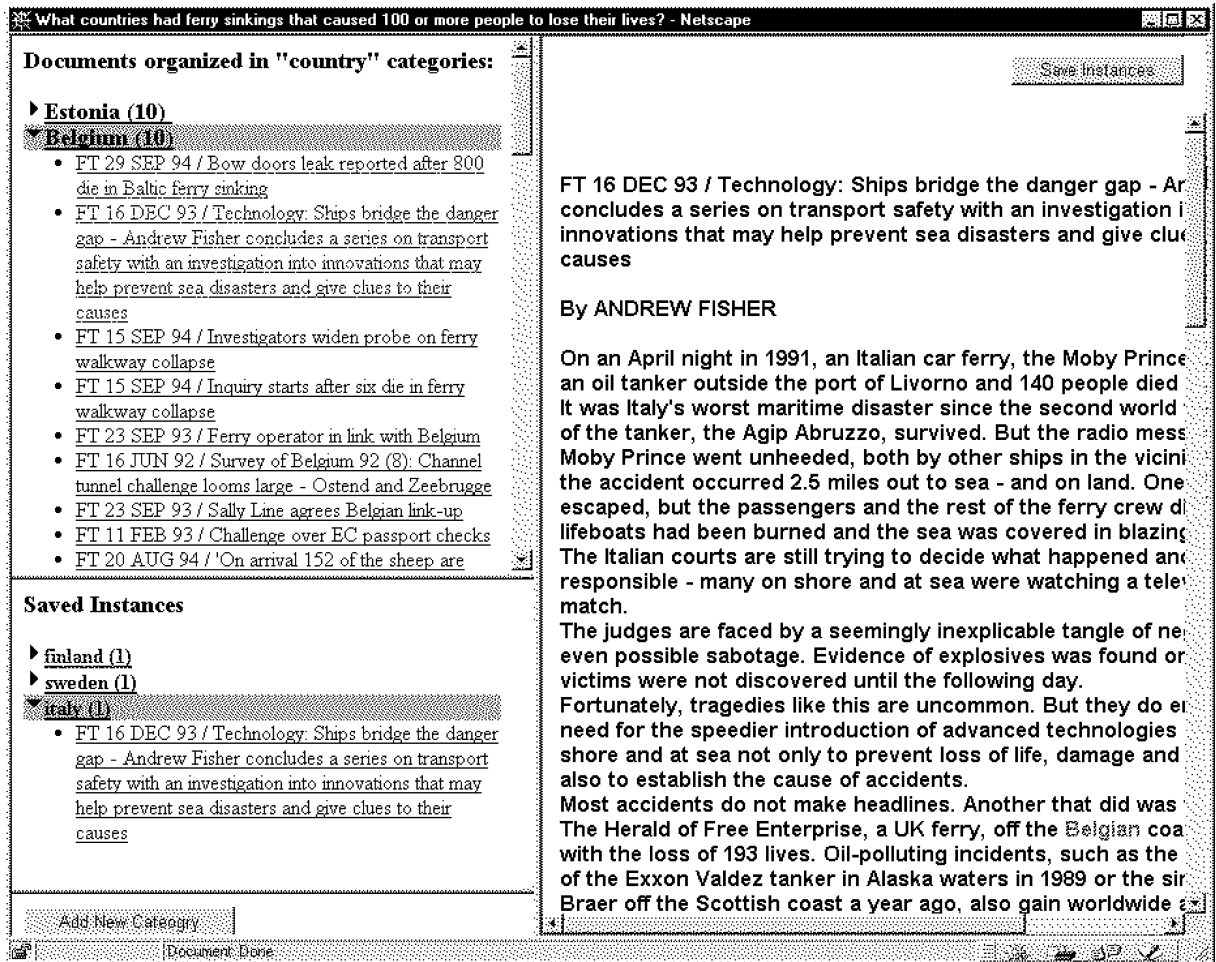


Fig. 7. Interface for classification.

categories. Three, no term highlighting was used when displaying documents using the ranked list interface.

This experiment was the basis for our participation of TREC-8 interactive track (Fuller et al., 1999), which used the same document collection as TREC-7, but a different set of six topics. There is also slight difference in experimental design between TREC-7 and TREC-8. In TREC-7, the order of topics within each block was fixed; in TREC-8, each topic was searched at a different position within each block. A complete round of the TREC-8 experiment required 12 subjects. We augmented the experiment design by adding an additional group of twelve subjects. The twenty-four subjects were computer science undergraduate and postgraduate students, with an average age of 23, three years online search experience, an average FA-1 (Controlled Associations) score of 28.6, and an average VZ-1 (paper folding) score of 15 (Ekstrom et al., 1976). None of the subjects had previously participated in a TREC interactive experiment.

5.2. Results and findings

Form Table 8, we can see that subjects saved more instances on average using the classification-based interface. The mean number of saved instances is 9.3 ($SD = 5.2$) for the classification-based interface, and 8.8 ($SD = 4.1$) for the ranked list interface; this difference is not statistically significant (two tail, paired t -test).

That fewer instances were saved using the classification-based interface in topics 438 and 446 was unexpected. Topics 438 and 446 were both instances of “country” topics: topics where the instances consisted of a list of different countries. Before the experiment, we anticipated that such topics would be ideally supported by the classification approach, as the categories instantiated from country match well with a reasonable division of the topics into instances. On closer examination, we found that the categories for topic 438 turned out not to have been ranked as expected due to a coding error (the other five topics were not affected by this mistake); as a result, some categories containing documents that were very similar to the topic were not highly ranked. Discarding the results of this topic, on the average, the subjects saved 7.98 instances from the list interface and 8.84 instances from the category interface. This difference was not statistically significant.

Table 9 shows the average instance recall for each topic, based on the TREC/NIST assessors’ judgement (topic 438 is excluded). According to the assessors’ judgement, the saved documents from category interface covered more instances than the saved documents from list interface. However, this difference is also not statistically significant. The comparison of Tables 8 and 9 shows the difference between the objective evaluation and the users’ subjective performance.

Closer inspection of the experiment logs revealed a somewhat surprising occurrence. For topic 446, we expected that subjects would choose the concept country as the classification axis. However, five of the twelve subjects did not do so. This (in our opinion) mis-selection of classification axis also occurred with two other country topics (414 and 428): six out of twelve subjects for topic 414, and three out of twelve subjects for topic 428 did not choose country as the basis for classification. What we had expected as a given – that subjects would trivially select an apparently obvious classification axis for a given topic – turned out not to be the case. This makes comparing the effectiveness of the list-based and classification-based interfaces difficult. Taking specific searches in isolation, for topics 414, 428 and 446, if we consider only those searches where subjects selected the “correct” categories, significantly more instances were saved using the classification-based interface than using the list based interface ($P < 0.03$, one tail, unpaired t -test), although again this does not carry on to an improved objective performance. However, such isolated analyses are fraught with danger: the use of a subset of searches compromises the integrity of the Latin Square design that the Interactive Track experiments are built around.

Table 8
The average number of saved instances for each topic in subjects’ view

	408	414	428	431	438	446	Mean
List	8.5	6.6	8.3	8.9	13	7.6	8.8
Category	10.1	7.3	8.8	11.3	11.6	6.7	9.3

Table 9

The average instance recall of the saved documents per topic in assessors' view (excludes topic 438)

	408	414	428	431	446	Mean
List	0.292	0.535	0.263	0.317	0.219	0.325
Category	0.323	0.514	0.179	0.273	0.172	0.292

5.3. Measuring user satisfaction

In this experiment we also gauge users' impressions of the two interfaces. After completing a set of searches with an interface, members of one group of 12 subjects were asked to complete a questionnaire evaluating their experience. The questionnaire was adapted from (Doll & Torkzadeh, 1988), and focused on subjects' satisfaction with the presentation format, the delivered data, an interface's ease-of-use, and the time available for the topics.

The results from the questionnaires are shown in Fig. 8. Q1-Q3 shows subjects' satisfaction with the displayed contents; Q4 the time available; Q5-Q6 the ease of use; Q7-Q9 the way the data was organized; and Q10 overall satisfaction with the interface. Note that for Q1, a lower score indicates greater satisfaction. Subjects responded using a five-point scale. From Fig. 8, we can see that for all questions the satisfaction scores for the classification-based interface are higher than the ranked list interface. This difference is statistically significant ($P < 0.001$, paired, one tail t -test).

Fig. 8 also suggests that the organization of retrieved data may influence the subjects' perception of it. Although both interfaces offered the same number of documents, albeit differently organized, subjects nonetheless felt that the ranked list interface showed too much information, and felt able to find neither enough instances nor sufficiently precise instances to answer the topics. Given that subjects saved approximately the same number of instances with each interface, this shows a strong discrepancy between subjects' preferences and their performance.

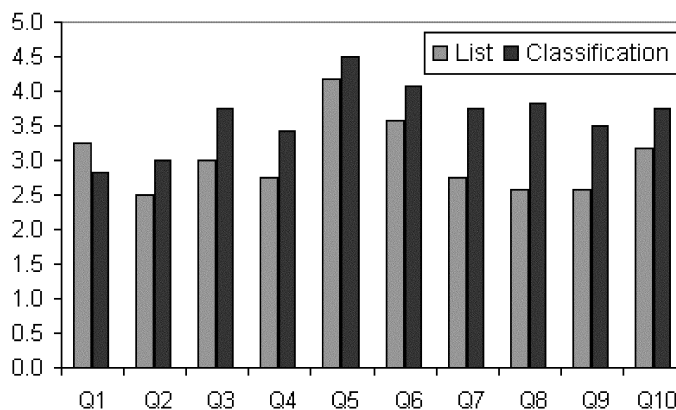


Fig. 8. The results from the user satisfaction questionnaire (Q1: Too much information, Q2: Precise instances, Q3: Sufficient instances, Q4: Enough search time, Q5: Easy to use, Q6: User friendly, Q7: Clear organization, Q8: Useful format, Q9: Organization what is needed, Q10: Satisfaction with the interface.)

6. Limitations

Only one or two experiments are not convincing enough to accept or reject an experimental hypothesis, especially when the experiment sample is small. The experiments presented in this paper were our first attempt to explore how to conduct interactive experiments, and how to evaluate and interpret the interactive experiment results. We have identified some limitations of the presented experiments which will be addressed in our future work.

The interfaces we have explored have been static: each cluster or classification was based on the pre-specified query. Users were able neither to re-formulate the query terms used to generate the clusters or classifications, nor to re-cluster or re-classify. Given that the experimental framework is inherently an interactive one with a human user available at all times to guide or refine the selection and organization of data, an improved approach would enable users to dynamically control the information presented over the course of a session. We believe that the structured delivery does help users understand better about the set of retrieved documents than a ranked list. We will be in a better position to prove this hypothesis if we let users have more interaction with the system in our future experiments.

The subjects for each experiment were all first time users of the two delivery systems. Through the post experiment interview, some subjects admitted that they did not understand why the documents were clustered or classified. We believe if subjects understand a little bit more about the concept of clustering or classification structure, they could adopt more suitable search strategies. We will try to verify this through an experiment that involves both “experienced” subjects and new subjects.

The experiments introduced in this paper were basically conducted under TREC Interactive Track framework. Compared with the TREC ad hoc data, the Interactive Track data has an extremely low proportion of alternative, equivalently relevant documents for each topic instance. Taken to an extreme, this means that it is simply not possible to outperform a list interface for such a collection. Consider a collection that contained exactly one relevant document per topic instance. For such a collection, the ideal delivery organization would be a list of the relevant documents, one after the other. Any re-structuring of that list can only add a level of indirection between the user and the data, without adding any possible benefit. In contrast, a collection that contained many relevant alternatives or non-relevant near-alternatives would be ill-served by a simple list: a user must sift through a large amount of redundant or irrelevant information in order to locate documents containing all relevant instances. In such a case, a suitable form of re-organization may be of significant benefit.

We observed that the experiment systems (clustering/classification) worked well for some topics. However, examination of the subject–system interactions and of subjects’ perception of topics and systems provided no explanation for this phenomenon. It may be possible to discover some hidden relationships if the number of topics and the number of subjects are considerably larger.

In the presented experiments, subjects played the role as intermediary searchers. We observed that the subjects lacked a genuine motivation to pursue their solutions, and lacked a good understanding of the search topics and search tasks. It would be interesting to investigate how subjects performed if asked to carry out searches based on their own, real information needs. We expect this will bring up more challenges to the experiment design and evaluation.

7. Discussion

Clustering and classification are two mechanisms for organizing documents into groups. In this paper, we have presented an ongoing series of experiments that tested the feasibility and effectiveness of using clustering and classification as an aid to instance retrieval and, ideally, answer construction.

Our results have shown that static clustering can organize intermediate result sets into subsets that are (mostly) relevant or non-relevant to the topic, but not into instance groups. Users were able to find the clusters that contained the relevant documents.

Users preferred a cluster-based interface to a list interface for the interactive instance retrieval tasks set for them. They believed they were performing better; objective assessments showed that they did not improve their performance in terms of instance recall. However, it is possible that the structuring provided by the cluster-based interface delivered additional benefits (not measured by these experiments) in the larger task of synthesizing and constructing an answer to the topic.

Several hypotheses exist for the variation between subjective and objective assessment of user performance. One is simply that if users like something, particularly something new, they believe they perform better. An alternative is that the discrepancy is caused by the inevitable differences in world view between the objective assessors and the experimental subjects. This second hypothesis was tested and invalidated by an experiment that removed the potential variance in world views by pre-determining the potential instances to be located.

Apparently, although the static clustering explored here could separate documents into groups containing more or fewer relevant and non-relevant documents, and although users could successfully identify those groups likely to contain more relevant documents, it simply was not an appropriate mechanism for instance retrieval. There are several possible remedies for this. One is to use a hierarchical mechanism to cluster documents according to topic relevance at higher levels, while clustering according to instance relevance at deeper levels. An alternative is dynamic approaches, using interactive clustering and re-clustering, perhaps along the lines of Hearst and Pedersen (1996).

An alternative to clustering is classification. Simple classification using WordNet was attempted. Under experimental conditions, subjects did not always select the apparently most appropriate classification axis from the other alternatives. Nonetheless, even when an inappropriate classification was selected, there was no great deterioration in performance. As was the case with the clustering approach, subjects preferred the classification interface to a simple list interface, and believed their performance to be better. Again however, objective assessments showed that their performance with the classification interface was not improved.

These approaches flow naturally into the task of constructing answers themselves, as opposed to simply identifying valuable information resources. The classification approach, in particular, can be seen as an initial stage in helping users with the task of organizing and constructing answers. Arguably, the content and quality of those constructed answers are what should be assessed in order to determine the effectiveness of the interactive process itself. The ultimate challenge, of course, is the fully automated construction of synthetic answer documents with minimal user input.

Acknowledgements

The authors wish to acknowledge the helpful suggestions made by anonymous referees.

References

- Cormack, G. V., Palmer, C. R., & Clarke, C. L. A. (1998). Efficient construction of large test collections. In *Proceedings of the 21st ACM-SIGIR international conference on research and development in information retrieval*, Melbourne, Australia (pp. 282–289).
- Croft, W. B. (1978). *Organising and searching large files of documents*. Ph.D. thesis, University of Cambridge.
- Cutt, D. R., Karger, D. R., Pedersen, J., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM/SIGIR conference*, Copenhagen, Denmark (pp. 318–329).
- Doll, W. J., & Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. *MIS Quarterly*, June, 259–274.
- Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual of kit of factor-referenced cognitive tests*. Educational Testing Service, Princeton, NJ (tests used by permission of ETS).
- Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
- Frakes, W. B., & Baeza-Yates, R. (1992). *Information retrieval: Data structures and algorithm*. Englewood Cliffs, NJ: Prentice-Hall.
- Fuller, M., Kaszkiel, M., Kim, D., Ng, C., Robertson, J., Wilkinson, R., Wu, M., & Zobel, J. (1998). TREC7 ad Hoc, speech, and interactive tracks at MDS/CSIRO. In *Proceedings of the seventh text retrieval conference (TREC-7)*, Gaithersberg, MD, USA (pp. 465–474).
- Fuller, M., Kaszkiel, M., Kimverley, S., Ng, C., Wilkinson, R., Wu, M., & Zobel, J. (1999). The RMIT/CSIRO ad Hoc, Q&A, web, interactive, and speech experiments at TREC 8. In *Proceedings of the eighth text retrieval conference (TREC-8)*, Gaithersberg, MD, USA.
- Hearst, M. A., & 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, 18–22 August, ACM (pp. 76–84).
- Hersh, W., & Over, P. (1999). TREC-8 Interactive Track. *SIGIR Forum*, 33(2), 8–11.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley-Interscience.
- Lagergren, E., & Over, P. (1998). Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiments. In *Proceedings of the 21st ACM-SIGIR international conference on research and development in information retrieval*, Melbourne, Australia (pp. 164–172).
- Leuski A., & Allan, J. (1998). Evaluating a visual navigation system for a digital library. In *Proceedings of the second European conference on research and advanced technology for digital libraries*, Heraklion, Crete, Greece.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.
- Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.
- Rose, D. E., Mander, R., Oren, T., Ponceon, D. B., Salomon, G., & Wong, Y.Y. (1993). Content awareness in a file system interface: Implementing the 'Pile' metaphor for organizing information. In *Proceedings of the 16th annual international ACM/SIGIR conference* (pp. 260–269).
- Salton, G. (Ed.). (1971). *The smart retrieval system: Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall.
- Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

- Sebrechts, M. M., Cugini, J. V., Vasilakis, J., Miller, M. S., & Laskowski, S. J. (1999). Visualization of search results: A comparative evaluation of text, 2D, and 3D interfaces. In *Proceedings of the 22nd ACM-SIGIR international conference on research and development in information retrieval*, Berkley, CA, USA (pp. 3–10).
- Swan, R. C., & Allan, J. (1998). Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. In *Proceedings of the 21st ACM-SIGIR international conference on research and development in information retrieval*, Melbourne, Australia (pp. 173–181).
- Voorhees, E. (1998). Variations in relevance judgements and the measurement of retrieval effectiveness. In *Proceedings of the 21st ACM-SIGIR international conference on research and development in information retrieval*, Melbourne, Australia (pp. 315–323).
- Voorhees, E., & Harman, D. (2000). Overview of the sixth text retrieval conference (TREC-6). *Information Processing and Management*, 36, 3–35.
- Witten, I., Moffat, A., & Bell, T. (1994). *Managing gigabytes: Compressing and indexing documents and images*. New York: Van Nostrand Reinhold.
- Wu, M., & Fuller, M. (1997). Supporting the answering process. In *Proceedings of the second Australian document computing symposium*, Melbourne, Australia (pp. 65–73).
- Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st ACM-SIGIR international conference on research and development in information retrieval*, Melbourne, Australia (pp. 307–314).