Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection

Peter Pirolli, Patricia Schank, Marti Hearst, Christine Diehl

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304, USA
pirolli@parc.xerox.com, schank@unix.sri.com,
hearst@parc.xerox.com, cdiehl@violet.berkeley.edu

ABSTRACT

Scatter/Gather is a cluster-based browsing technique for large text collections. Users are presented with automatically computed summaries of the contents of clusters of similar documents and provided with a method for navigating through these summaries at different levels of granularity. The aim of the technique is to communicate information about the topic structure of very large collections. We tested the effectiveness of Scatter/Gather as a simple pure document retrieval tool, and studied its effects on the incidental learning of topic structure. When compared to interactions involving simple keyword-based search, the results suggest that Scatter/Gather induces a more coherent conceptual image of a text collection, a richer vocabulary for constructing search queries, and communicates the distribution of relevant documents over clusters of documents in the collection.

Keywords

Information retrieval; Scatter/Gather; Clustering; Browsing.

INTRODUCTION

When faced with ill-defined problems requiring information access, we often want to explore the resources available to us before exploiting them. This exploration may be partly aimed at refining our understanding of the potential space of content that is available, and partly aimed at formulating a concrete course of action for getting specific documents. Interfaces that support the *browsing* of a collection, as opposed to *searching* a collection, are aimed at satisfying this need to learn more about a collection before taking action.

Scatter/Gather is a document browsing technique aimed at supporting such exploratory learning [3, 4]. The emphasis in this browsing technique is to present users with an automatically computed overview of the contents of a document collection, and to provide a method for navigating through this summary at different levels of granularity. The central primitive operation in Scatter/Gather involves document clustering based on pairwise document similarity. The technique aims to place similar documents into the same cluster. Recursively clustering a collection produces a *cluster hierarchy.*

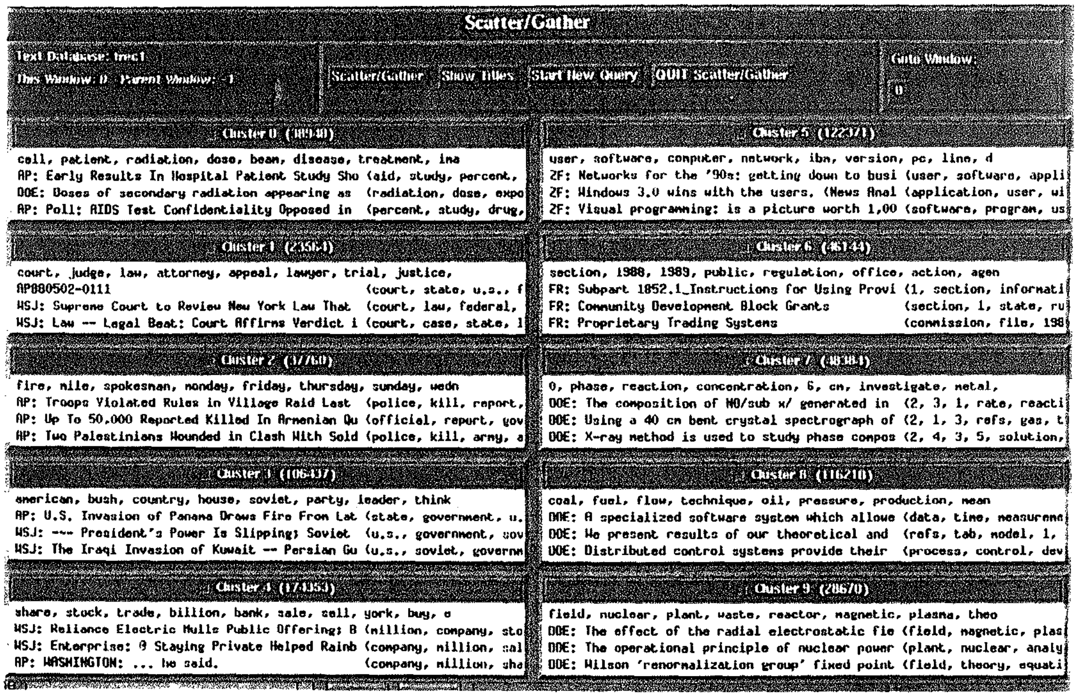
For each cluster, at each level of this hierarchy, the user is presented with summary information about the cluster that presumably communicates something about the kinds of documents it contains. The user may then select (*gather*) those clusters that seem interesting or relevant. This subset of clusters can then be reclustered (*scattered*) to reveal more fine-grained clusters of documents. With each successive iteration of scattering and gathering clusters, the clusters become smaller and more detailed, eventually bottoming out at the level of individual documents.

In this paper we present experimental evidence to support the view that Scatter/Gather interaction induces the development of a better understanding of the contents of large text collections, refines users' formulations of search queries, and is a feasible information retrieval technique in its own right when compared to standard search techniques.

METHODOLOGICAL ISSUES

Much of the discussion on the evaluation of textual information access systems has focused on how to improve the traditional measures of *precision* and *recall* (e.g., [6, 11]). Precision is the proportion of relevant documents retrieved among a set of retrieved documents, whereas recall is the proportion of relevant documents in the original collection that were retrieved. Some researchers argue the need to move away from strict measures of overall precision and recall toward other measures of search success or to what is termed user-centric evaluation [5]. However, most of the analyses that offer alternatives to precision and recall measures focus on how well the user's information need is satisfied (and in some cases on the progression of familiarity with the system under study), rather than what is learned incidentally about the collection itself. Pirolli and Card [8] suggest evaluating information access search strategies using a theoretical model based on optimal foraging theory from biology. Foraging often involves a

CHI 96 Vancouver, BC Canada
© 1996 ACM 0-89791-777-4/96/04..$3.50



Text Database: trec1
This Window: 0 Parent Window: -1

```
Cluster 0 (38940)
cell, patient, radiation, dose, bean, disease, treatment, ina
AP: Early Results In Hospital Patient Study Sho (aid, study, percent,
DOE: Doses of secondary radiation appearing as  (radiation, dose, expo
AP: Poll: AIDS Test Confidentiality Opposed in  (percent, study, drug,

Cluster 1 (23564)
court, judge, law, attorney, appeal, lawyer, trial, justice,
AP880502-0111                                   (court, state, u.s., f
WSJ: Supreme Court to Review New York Law That  (court, law, federal,
WSJ: Law -- Legal Beat: Court Affirms Verdict i (court, case, state, 1

Cluster 2 (37760)
fire, mile, spokesman, monday, friday, thursday, sunday, wedn
AP: Troops Violated Rules in Village Raid Last  (police, kill, report,
AP: Up To 50,000 Reported Killed In Armenian Qu (official, report, gov
AP: Two Palestinians Wounded in Clash With Sold (police, kill, army, a

Cluster 3 (106437)
american, bush, country, house, soviet, party, leader, think
AP: U.S. Invasion of Panama Draws Fire From Lat (state, government, u.
WSJ: --- President's Power Is Slipping; Soviet  (u.s., government, sov
WSJ: The Iraqi Invasion of Kuwait -- Persian Gu (u.s., soviet, governm

Cluster 4 (174353)
share, stock, trade, billion, bank, sale, sell, york, buy, o
WSJ: Reliance Electric Mulls Public Offering; B (million, company, sto
WSJ: Enterprise: @ Staying Private Helped Rainb (company, million, sal
AP: WASHINGTON: ... he said.                    (company, million, sha

Cluster 5 (122371)
user, software, computer, network, ibm, version, pc, line, d
ZF: Networks for the '90s: getting down to busi (user, software, appli
ZF: Windows 3.0 wins with the users. (News Anal (application, user, wi
ZF: Visual programming: is a picture worth 1,00 (software, program, us

Cluster 6 (46144)
section, 1988, 1989, public, regulation, office, action, agen
FR: Subpart 1852.1_Instructions for Using Provi (1, section, informati
FR: Community Development Block Grants          (section, 1, state, ru
FR: Proprietary Trading Systems                 (commission, file, 198

Cluster 7 (48384)
0, phase, reaction, concentration, 6, cm, investigate, metal,
DOE: The composition of NO/sub x/ generated in  (2, 3, 1, rate, reacti
DOE: Using a 40 cm bent crystal spectrograph of (2, 1, 3, refs, gas, t
DOE: X-ray method is used to study phase compos (2, 4, 3, 5, solution,

Cluster 8 (116210)
coal, fuel, flow, technique, oil, pressure, production, mean
DOE: A specialized software system which allows (data, time, measureme
DOE: We present results of our theoretical and  (refs, tab, model, 1,
DOE: Distributed control systems provide their  (process, control, dev

Cluster 9 (28670)
field, nuclear, plant, waste, reactor, magnetic, plasma, theo
DOE: The effect of the radial electrostatic fie (field, magnetic, plas
DOE: The operational principle of nuclear power (plant, nuclear, analy
DOE: Wilson 'renormalization group' fixed point (field, theory, equati
```

Figure 1. The Scatter/Gather interface presents a user with clusters of like documents.

phase of exploration to gain knowledge about the distribution and cost structure of access to resources and this knowledge determines later patterns of foraging [1].

Although there are some subtle issues involved in the assessment of precision and recall measures (such as deciding on how to evaluate documents as being "relevant"), even more difficulties arise when trying to assess whether a browsing system is communicating knowledge to a user about the structure of a document collection. As will become clear, this was a major challenge to our research. We have had to develop multiple methodologies and analyses in order to converge on an assessment of the mental models that people get from interacting with Scatter/Gather.

DOCUMENT CLUSTERING IN SCATTER/GATHER

Clustering techniques have been used in the past as an enhancement to keyword-search retrieval algorithms [3, 4]. The novelty of the Scatter/Gather approach is its use of clustering as the basis of an interactive method that supports the browsing of a summary of the structure and content of a text collection--a summary intended to be like a table of contents. This raises two essential requirements [4]. First, the clustering method has to be fast enough to support real user interaction. Second, there must be a method of cluster summarization that enables a user to make sense of the text collection content and to navigate its structure.

Clustering depends on some measure of inter-document similarity. A common approach [10], with a number of variants, is to represent documents as vectors of equal length, where each component of a vector is associated with one of the unique content words in the document collection. In some schemes, the component may contain a value indicating the presence of a word in the document (i.e., a binary coding). In other schemes, a vector component may indicate the frequency or some normalized frequency of a word in the document. The similarity of two documents can then be computed by a *cosine measure*, which is the cosine of the angle between two vectors, which is sometimes also known as a *normalized correlation.*

Single-linkage *hierarchical clustering* is a commonly used method outside of information access. For instance, we use it below in our data analysis. It is, however, too slow for even moderately large document collections. Beginning with individual documents, single-linkage hierarchical clustering iteratively agglomerates the most similar pair of clusters built so far into a new cluster. The global consideration of all pairwise similarities at each stage of clustering leads to running times of $\Omega(n^2)$ in the number, n, of documents.

Speedier algorithms [4] were developed for Scatter/Gather based on a nonagglomerative *partitional clustering* scheme. The essence of this approach is to flatly partition a collection into k subsets, and recursively partition the subsets as needed to induce a hierarchical structure. At each partitioning, k document seeds are selected and remaining documents are clustered with the most similar seed. At each stage the procedure may be run further to iteratively improve the selection of seeds and improve the clustering. The developed schemes have running times of O(*kn*).

Linear O(*kn*) run-times are still too slow for very large collections. Constant-time interaction costs for Scatter/Gather navigation through a cluster hierarchy are achieved by precomputing a cluster hierarchy to be used in Scatter/Gather interaction [3]. The off-line precomputation uses the linear-time algorithms and summarizes document clusters by *meta-documents* containing profiles of topical words and the most typical titles.

These topical words and typical titles are also used to present users a summary of the documents in a cluster. Topical words are those that occur most frequently in a cluster, and typical titles are those with the highest similarity to a centroid of the cluster. Together, the topical words and typical titles form a *cluster digest*. The cluster digest scheme combined with the constant-time interaction cost satisfy the two requirements desired for a cluster-based browsing method.

A SCATTER/GATHER INTERFACE

Figure 1 presents a typical view of the Scatter/Gather interface. There are ten clusters, with each cluster represented by a separate area in Figure 1. For each cluster, the user is presented with the topical words that occur in the text contents of documents in a cluster, as well as the titles of the three most typical documents. The user may *gather* some subset of those clusters, by pointing and selecting buttons above each cluster, and then ask the system to *scatter* that subcollection into another 10 subclusters, by selecting a Scatter/Gather button at the top of the display in Figure 1.

Users could also select one or more clusters and then select the Show Titles button at the top of Figure 1, which would then display all of the document titles in the selected clusters, in a separate Show Titles Window. The user can scan the displayed list of titles, seeking those that appear relevant to the task at hand. Relevant documents can be selected by pointing, and then saved to file by cut and paste techniques.

OVERVIEW OF THE EXPERIMENT

Although Scatter/Gather has not been developed as a stand-alone information access technique [4], we felt that a stern trial of its efficacy would be to test it as a stand-alone search tool and determine how it fared in support of (a) finding relevant documents and (b) incidental learning about the text collection. We compared Scatter/Gather to a word-based search interface we called SimSearch. This interface allowed users to enter an arbitrary collection of words based on a topic query and it retrieved a fixed small set of documents most similar to the query, as determined by a similarity in a vector space model [2].

METHOD

Participants

Sixteen adults participated in this study. Three of the participants were Xerox PARC employees, one was a PARC employee family member, and the remaining 12 participants were graduate students from Stanford University. PARC employees volunteered their time for the study; the remaining subjects were paid $10/hour for their participation.

Materials and Procedure

Participants were asked to read the instructions for the experiment and then use one of the interfaces to find articles relevant to given topics in a large and complex collection of text documents. There were four phases to the experiment: (1) finding articles relevant to a topic using Scatter/Gather or Similarity Search, (2) writing search queries and drawing a diagram representing the topic structure of the collection, (3) a replication of Phase 1, and (4) a replication of Phase 2. Phases 1 and 3 were aimed at providing participants with exposure to one of the information access techniques, and Phases 2 and 4 were aimed at assessing the impact of that exposure on query formulation and the users' conception of the text collection.

The experiment used the 2.2 gigabyte TIPSTER text collection created for the TREC conference [7]. Twelve topics were drawn from the first 100 topics used in the TREC conference. These twelve were chosen based on a level of difficulty measured by the mean number of relevant documents in the Tipster collection as identified by information retrieval experts associated with TREC [7]. The four topics with the fewest (expert-identified) relevant documents ($M = 46$) were placed in the "hard" group, the four topics with the most relevant documents ($M = 865$) were placed in the "easy" group, and the four topics about the median number of relevant documents were placed in the "medium" group ($M = 303$).

Participants spent between two and five hours total on study activities, with two phases a day over the course of two days. Four blocks of topics were constructed for presentation over the four phases of tasks. Each topic-block contained one easy topic, one medium topic, and one hard topic, in that order. The order of blocks across phases was counterbalanced over sets of four participants, within groups, according to a randomized Latin square.

Participants were randomly assigned to one of three study conditions: Scatter/Gather Speeded (SGS, $N = 4$), Scatter/Gather with Relevance Ratings (SGR, $N = 4$), and SimSearch (SS, $N = 8$). Phases 1 and 3 varied by condition, and Phases 2 and 4 were identical across conditions. In the SGS condition, subjects were given one hour to find articles in each of Phases 1 and 3. In the SGR condition, subjects were not given a time limit, and were asked to complete additional classification and relevance activities: Given worksheets, they were asked to indicate how they would classify each presented cluster (i.e., using words or short phrases), and to estimate what percentage of texts in a cluster seemed relevant to the topic. In the SS condition, subjects used SimSearch rather than Scatter/Gather to find their relevant articles in phases one and three. Note that SGR and SGS conditions were combined for some analyses into one group (SG; see below).

RESULTS

Our first set of analyses will concentrate on comparisons of Scatter/Gather versus SimSearch, with respect to their effectiveness as information retrieval tools. Subsequent analyses will focus on comparisons with respect to enhancement of users' understandings of the text collection.

Finding Relevant Articles (Phases 1 and 3)

Table 1 presents summary data for the mean time spent per query, the mean number of documents selected (saved) for retrieval, mean number of those that were relevant (according to TREC identification), and the mean *precision* of the retrieved documents (No. relevant + No. saved). For each type of data in Table 1, (i.e., each row) we conducted an analysis of variance (ANOVA) for the two Groups (SG, SS) × Query Difficulty (Easy, Medium, Hard) × Phase (1, 3) factorial design.

Interface Effects

Examining Table 1, it is apparent that participants who used the Scatter/Gather interface to answer queries took substantially longer than those using SimSearch, and the SG vs SS group difference in Table 1 was significant, $F(1, 14) = 34.07$, $MSE = 297.79$, $p < .001$. The SGR participants took about twice as long as the SGS participants to answer queries. This is attributable to the ancillary rating task performed by the SGR participants. Still, SGS participants took about twice as long as SS participants, so that, even when instructed to use the Scatter/Gather interface as quickly as possible to answer queries, Scatter/Gather users were substantially outhustled by SimSearch users.

Although there was not a significant difference in the total number of documents saved by SG ($M = 12.26$) versus SS ($M = 16.44$) participants [$F(1, 14) = .24$, $MSE = 2355.77$], there was a difference in number of relevant documents saved, $F(1, 14) = 5.44$, $MSE = 108.55$, $p = .04$, with SS participants ($M = 10.40$) outgaining SG participants ($M = 5.44$). This is reflected in the differences in precision scores in Table 1.

The SG group showed substantially more variation in number of documents saved ($SD = 65.21$; $SD = 17.94$ with largest outlier removed) than the SS group ($SD = 11.27$). Furthermore, 27% of SG queries (13 of the 48), resulted in no documents saved, whereas all the SS participants saved at least one document on each query. These results suggest that the chance of finding relevant documents is a bit more "hit and miss" for participants using a pure SG system.

Query Effects

Both the total number of documents saved and the number of relevant documents saved decreased significantly with increases in Query Difficulty: (a) Easy No. saved $M = 34.91$ and No. relevant $M = 15.06$, (b) Medium No. saved $M = 15.72$ and No. relevant $M = 5.81$, and (c) Hard No. saved $M = 6.00$ and No. relevant $M = 2.88$ [for No. saved, $F(2, 28) = 3.99$, $MSE = 1737.52$, $p = .03$; for No. relevant $F(2, 28) = 10.04$, $MSE = 128.98$, $p = .001$]. Query Difficulty had a significant, although complex main effect on task time, $F(2, 28) = 3.73$, $MSE = 75.71$, $p < .05$, with participants spending more time on Easy ($M = 23.42$ mins) and Hard ($M = 20.15$ mins) queries than on Medium queries ($M = 17.50$ mins). This main effect did not interact with the type of interface used, $F(1, 14) = 1.56$. Perhaps participants monitored the rates at which they encountered relevant documents and this rate did not exhibit substantial depression before the 30 min cutoff for Easy and Hard queries, but did so for the Medium queries. That is, the rate of encounter with relevant documents should be highest for Easy queries, less for Medium, and less still for Hard queries, but the point at which there is a noticeable drop-off in this rate (indicating a near-depletion of relevant documents) may occur earliest for the Medium queries. Clearly the users' judgments about when to give up is complex and requires further investigation.

Table 1. Per-query averages for SimSearch (SS) and Scatter/Gather (SG) groups (SGS = Speeded, SGR = Rating conditions).

	SS	SG		
		SGS	SGR	Both
Time (min)	10.10	22.10	39.18	30.64
No. Saved	16.44	7.71	17.09	12.26
No. Relevant	10.40	2.45	8.42	5.44
Precision	.63	.32	.49	.44

Practice Effects

Although there were trends indicating that participants increased the number of documents saved as they gained more experience (Phase 1 $M = 11.92$; Phase 3 $M = 25.83$) and increased the number of relevant documents saved with increased practice (Phase 1 $M = 6.81$; Phase 3 $M = 9.02$), neither increase was significant, nor did the trends show a significant interaction with the type of interface used. There was also no main effect of practice on time per query (Phase 1 $M = 19.78$; Phase 3 $M = 20.94$). It could be that our participants found both SimSearch and Scatter/Gather to be rather easy to learn during the warm-up exercises and therefore showed relatively little additional learning during the experimental phases.

Summary

As a stand-alone information retrieval tool, Scatter/Gather is not as effective as a common word-based search technique. As we noted, however, this is a stem test of the technique, since it has really been aimed at communicating the topic structure of a collection. In the following sections we examine the incidental learning about topic structure that Scatter/Gather induced in the SG participants.

Communicating the Distribution of Relevant Texts

We expect the Scatter/Gather browser to provide users who are working on a query with a sense of the distribution of relevant documents across clusters. The SGR participants were asked to rate the precision of each cluster encountered (No. relevant documents + No. documents). Ideally, one would like to match these ratings against the actual precisions to assess how well Scatter/Gather communicates the location of relevant documents, but it is currently infeasible to carry out this computation.

As an approximation, however, we were able to formulate the expected distribution of relevant documents across clusters and to compare this against the observed distribution of ratings. The distribution of relevant documents across clusters was examined for each of 29 queries used in TREC. Sets of the 200 most similar documents were retrieved using the SimSearch technique and then these retrieved sets were clustered using the Scatter/Gather clustering algorithm into five partitions. The clustering, however, proceeded without any information about the nature of the query used to select the initial 200 documents. Clusters larger than a criterion size were clustered again, and this process recursed. For each query, the clusters were ranked $i = 1, 2, ..., 5$ by the number of relevant documents u_i assigned to cluster i from the U relevant documents contained in the original set that was clustered. Among regressions of linear, exponential, and power functions, the best fit was obtained by an exponential distribution of relevant documents across clusters,

$$u_i = .47Ue^{-.57(i-1)}, \qquad (1)$$

with $R^2 = .92$.

Using Equation 1, we estimated the distribution of relevant documents (those identified by TREC experts) across the 10 topmost clusters and compared this against the mean precision estimates given by SGR participants, and these data are presented in Figure 2. (For the sake of clarity, Figure 2 omits estimates of zero precision by our participants.) This makes the assumption the rank ordering of people's ratings mirror the rank ordering of No. relevant documents in clusters.

The linear relationships apparent in the log-log coordinates of Figure 2 suggest that the Scatter/Gather users perceive a distribution of relevant documents across clusters that has a power law relation to the expected distribution of relevant documents. Such relations are common in many domains where people must assess the strength of evidence to judge the probability of events [9]. Users apparently show the same biases in estimating how many relevant documents are in a collection as they do in estimations of events in other domains, such as sports or health risks.

It seems, then, that three conclusions can be drawn about how well Scatter/Gather communicates the location of relevant documents among the clusters presented to users: (1) even though the clustering algorithm obtains no information about users' queries, most of the documents relevant to a typical (TREC) query will tend to be grouped into few clusters, (2) users judge most of the relevant documents to be grouped into few clusters, and (3) the users' judgments appear to have a well-defined power-law relationship to the actual distribution of relevant documents.

Communicating Effective Query Terminology

We might also expect that exploring a text collection with the Scatter/Gather interface improves the ability of users to formulate better key-word search queries. We examined this expectation through an analysis of the search queries formulated by participants in our study. Both SG and SS subjects were given the task to generate search words for given query topics after having used their systems. Table 2 shows the mean number of keywords participants used in their keyword queries, and the mean number of new keywords in their query (i.e., keywords that didn't appear in the topic description). SG participants used significantly more new terms not given in the topic description than SS participants, $t(15) = 2.01$, $p < .05$, and this was more apparent early in Phase 2, $t(15) = 1.84$, $p < .05$. In Phase 2, SG subjects also generated significantly more terms than SS subjects, $t(15) = 1.79$, $p < .05$, although SS subjects marginally increased the number of terms they used from Phase 2 to Phase 4. These results support the hypothesis that Scatter/Gather users are learning about the effective topic language from which to generate their search queries.

Communicating a Conceptual Model of Topic Structure

In Phases 2 and 4 we had asked all participants to draw tree diagrams representing the topic structure of the text collection. Descriptive statistics for subjects' drawn diagrams are given in Table 3. SGR participants' diagrams had the most nodes and links, significantly more than SGS [$t(7) = 2.78$, $p < .05$], but not significantly more than SS [$t(11) = 1.45$, $p > .10$]. For all three groups, the number of nodes did not significantly differ from number of links. SGR diagrams were significantly more deep than SGS diagrams [$t(7) = 2.82$, $p < .05$), but not significantly deeper than SS diagrams. SGR diagrams were significantly broader than both SGS and SS diagrams [$t(7) = 2.22$ and $t(11) = 4.77$, respectively, both p's $< .05$]. Although SGS diagrams were not significantly broader than SS diagrams

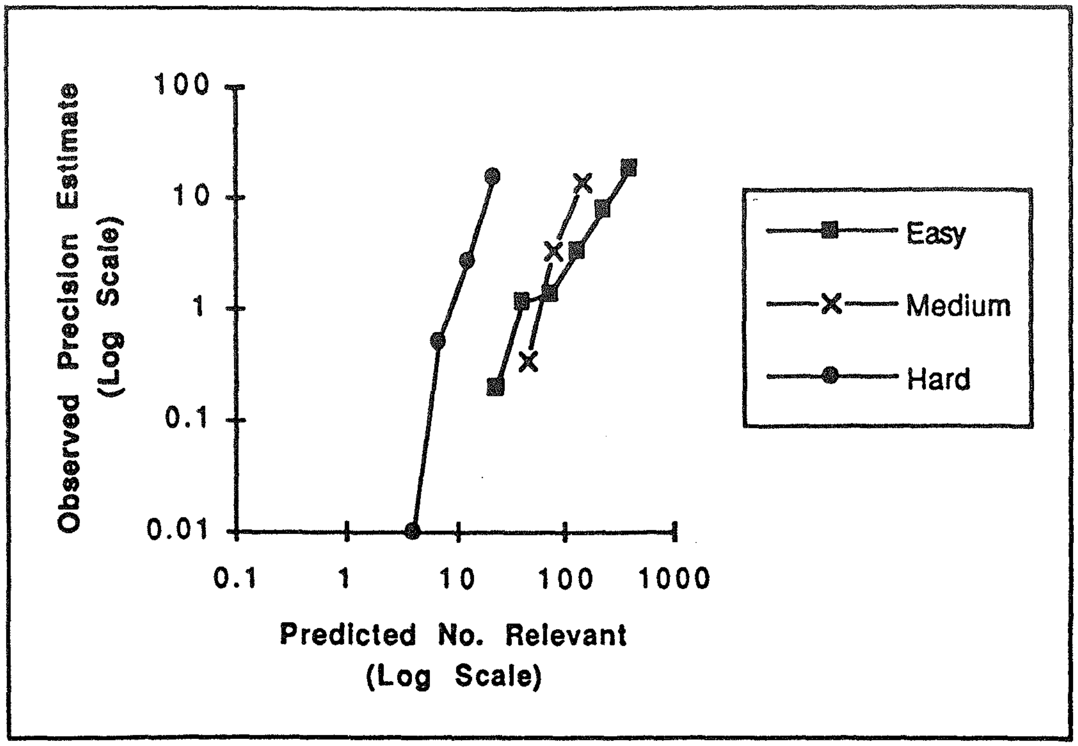


Figure 2. The match of the perceived precision (No. relevant + No. total documents) of clusters to the predicted distribution of relevant documents for different degrees of Query Difficulty.

overall, both SGS and SGR diagrams were significantly broader than SS diagrams at the highest level [$t(11) = 2.81$, $p < .01$, and $t(11)=4.71$, $p<.001$, respectively]. That is, the SG subjects used a greater number of primary (or organizing) topics in their diagrams (see Table 3).

Table 2. Search terms used, Phases 2 and 4.

Group	Number of Terms Used	Number of New Terms
SS	5.77	0.73
Phase 2	4.75	0.54
Phase 4	6.79	0.91
SG	6.44	1.84[a]
Phase 2	6.67[a]	2.04[a]
Phase 4	6.21	1.63

[a]significantly greater than other group, same condition, $p < .05$

If it is the browsing interaction with Scatter/Gather that induces the topic structure in users, then we should expect the richness of the diagrams drawn by SG participants to increase with increased browsing. Figure 3 shows that, indeed, the breadth of SG diagrams was highly correlated with the number of Scatter/Gather cluster windows (windows such as Figure 1) that participants saw during a block of queries. The number of cluster windows was not correlated at all, however, with the depth of the diagrams or the number of nodes or links.

Regarding content, there seemed to be more diversity in topics listed in SGR diagrams, e.g., in addition to terms used in the given query topics. The SS group listed items more related to their specific query topics, as well as more general nodes like "news," "AP," "WSJ," etc. Finally, the SGS diagrams seem to have a few new topics not given in the queries (e.g., "medicine," "science"), but not as many as the SGR group. Two judges rated the topics appearing in the diagrams on a three-point "specific-general" scale to determine the diversity of topics contained in the diagrams (a rating of one was assigned to a topic judged to be very specific, e.g., "corrupt government officials", and a rating of three was assigned to a very general topic, e.g., "law"). There was an 82% inter-rater agreement overall and 85% agreement on the most general topics (i.e, those having a rating of three). Forty diagram topics (20% of the distinct topics used) were rated as most general by both judges, and the groups were compared on the proportion of "general" topics their diagrams contained. Overall, the SG subjects used a significantly greater percentage of general topics than the SS subjects (52% and 31% respectively; $t(15)=2.56$, $p < .05$). The SGS and SGR subjects did not differ significantly in the proportion of general topics they included in their diagrams (54% and 49% respectively). There were no significant between- or within-group differences in the number of general topics for diagrams drawn in Phase 2 versus Phase 4.

Table 3. Diagram analyses, Phases 2 and 4.

Group	Mean #Nodes	Mean #Links	Mean Depth	Mean Breadth	Breadth to Depth Ratio	Mean #Primary Topics
SS	12.5	11.8	2.8	3.9	1.4	3.1
SGS	7.5	7.3	1.9	4.4	2.3	4.9[b]
SGR	15.1[a]	14.4[a]	3.0[a]	6.0[ab]	2.0	6.6[ab]

[a] significantly greater than SGS group, $p < .05$
[b] significantly greater than SS group, $p < .05$

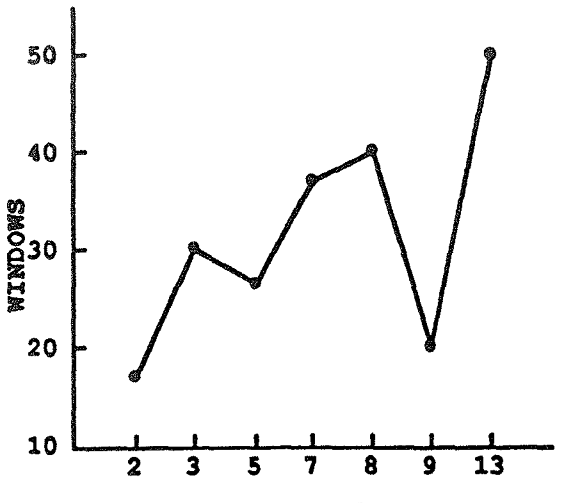


Figure 3. Relation of No. Scatter/Gather windows seen to the breadth of a subject's topic structure diagram.

A multi-dimensional scaling analysis was used to assess the similarity of participants' diagrams based on the number of shared topics. Such an analysis attempts to arrange entities in a space such that the distances between the entities in that space correspond (as much as possible) to some measured differences (or similarities) between all pairs of entities. We wanted to look at the output of such an analysis to see which users (the entities in our analysis) seemed to cluster together in terms of their conception of the topics contained in the text collection. That is, multi-dimensional analysis was used to lay out a kind of semantic space for users conceptions of topics in the collection.

The similarity matrix we used for this analysis indicated the proportion of shared topics between any two diagrams (number of topics in common divided by the number of topic nodes in the larger diagram). One SS participant did not use topics and was not included in the analysis. On average, SGS and SGR participants' diagrams were more similar to each other than they were to SS participants' diagrams (sharing 10% versus 5% of topics, respectively). SS participants' diagrams were no more similar to each other than they were to SG participants' diagrams (sharing only 5% of topics). The SGS diagrams were the most similar of the three groups, sharing an average of 14% of their topics. Figure 4 shows a three-dimensional multi-dimensional scaling solution for these diagram topic similarity data. The circles represent SG participants and the boxes represent SS participants. In addition, a minimal spanning tree has been laid over the points in Figure 4 to highlight clusterings. Overall, it appears that SG participants are more central and closer to one another in Figure 4 than SS participants. As a measure of incoherence, we computed the root mean squared (*RMS*) distances among points in the Figure 4. This incoherence measure was lower for the SGS (*RMS* = 1.08) and SGR (*RMS* = 1.22) participants than for the SS participants (*RMS* = 1.68). The people using Scatter/Gather seemed to be closer to one another in terms of their conception of the topics contained in the collection

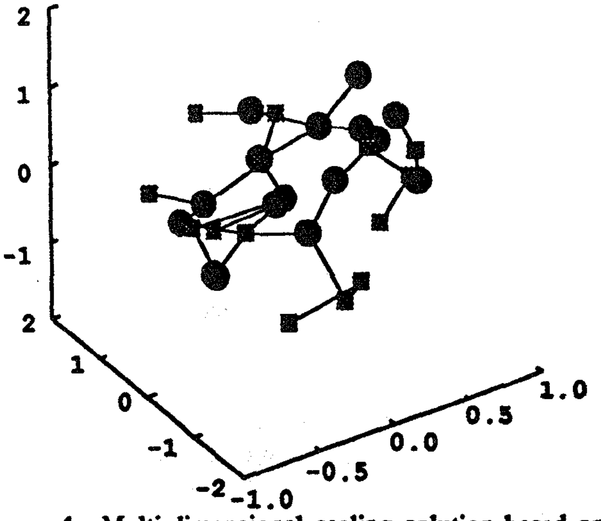


Figure 4. Multi-dimensional scaling solution based on diagram content similarity for the Scatter/Gather (circles) and SimSearch (boxes) participants.

Figure 5 shows a hierarchical clustering based on the similarity data for the diagrams from Phase 4. This is another way of seeing how individual people cluster together in terms of their conception of the topics in the collection. This tree representation clearly shows the strong similarity grouping of the SGS and SGR diagrams. Overall then, it appears that the Scatter/Gather interface is inducing a more coherent view of the text collection than SimSearch.

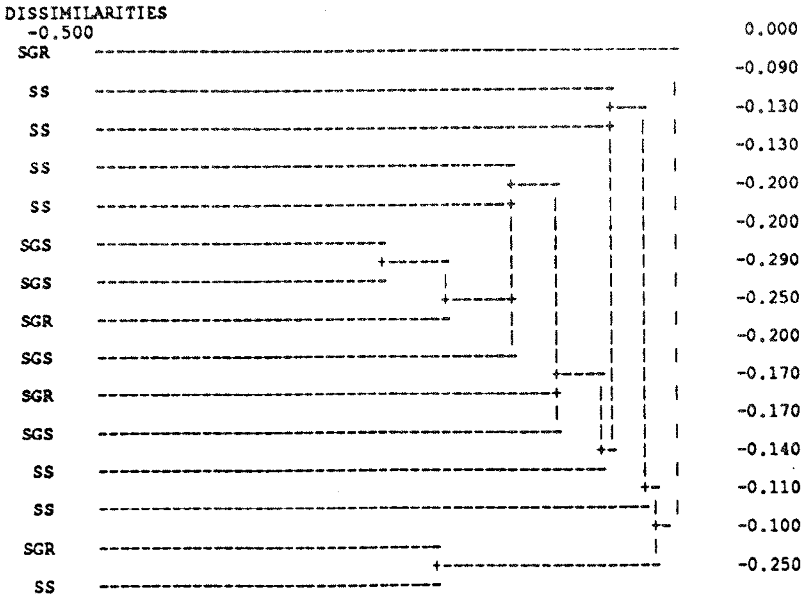


Figure 5. Hierarchical clustering of diagram topic similarity data from Phase 4 (single linkage method, nearest neighbor). SG diagrams (SGS and SGR) appear to be the most similar grouping.

GENERAL DISCUSSION

The results indicate that Scatter/Gather communicates several kinds of information about the topic structure of a large text collection. It is, by itself, not a superior information retrieval tool when the goal is to locate specific documents. Scatter/Gather may be useful in support of the kind of exploratory sensemaking activities that occur when users encounter large unknown text collections, and it should be coupled with other kinds of retrieval techniques, such as SimSearch, that can be enhanced by the knowledge that users gain through preliminary Scatter/Gather browsing. We are currently in the process of testing such a multi-functional browsing/search engine.

In conducting this research, we also faced numerous methodological problems. It is easy to collect measures based on counts of documents retrieved, but not so easy to assess the knowledge about a collection that is communicated by a browsing technique. We developed several convergent assessment measures in this regard, though clearly a great deal of refinement remains to be done.

ACKNOWLEDGMENTS

We thank Jan Pedersen for discussions and advice throughout this study, and Stuart Card for similar interactions during the initial experimental design.

REFERENCES

1. Bell, W.J., Searching behavior: The behavioral ecology of finding resources. Chapman and Hall, New York, 1991.

2. Buckley, C., J. Allan, and G. Salton. Automatic routing and ad-hoc retrieval using SMART. in Second Text Retrieval Conference TREC-2. (1994), National Institute of Standards and Technology.

3. Cutting, D.R., D.R. Karger, and J.O. Pedersen. Constant interaction-time Scatter/Gather browsing of very large document collections. in SIGIR '93. (1993).

4. Cutting, D.R., D.R. Karger, J.O. Pedersen, and J.W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. in SIGIR '92. (1992), pp. 318-329.

5. Froehlich, T.J., Relevance reconsidered -- Towards an agenda for the 21st century. Journal of the American Society for Information Science, 45 (1994). 124-134.

6. Harman, D., Evaluation issues in information retrieval. Jorunal of the American Society for Information Science, 28 (1992). 439-440.

7. Harman, D. Overview of the first text retrieval conference. in 16th Annuam International ACM/SIGIR Conference. (1993), Pittsburgh, PA. ACM. pp. 36-38.

8. Pirolli, P. and S. Card. Information foraging in information access environments. in Conference on Human Factors in Computing Systems, CHI-95. (1995, Association for Computing Machinery.

9. Tversky, A. and C.R. Fox, Weighing risk and uncertainty. Psychological Review, 102 (1995). 269-283.

10. vanRijsbergen, C.J., Information retrieval. Butterworth & Co., Boston, MA, 1979.

11. Wildemuth, B.M. and R.D. Bliek, Measures of searcher performance: A psychometric evaluation. Information Processing and Management, 29 (1993). 533-550.