Interactive Cluster Visualization for Information Retrieval

James Allan, Anton V. Leouski, and Russell C. Swan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003 USA

{allan,leouski,swan}@cs.umass.edu
+1 413 545 0463
+1 413 545 1789 (fax)

July 31, 1997

Abstract

This study investigates the ability of cluster visualization to help a user rapidly identify relevant documents. It provides added support for the truth of the Cluster Hypothesis on retrieved documents and shows that clustering of relevant documents is readily visible. The study then shows the visual effect of a technique similar to relevance feedback and shows how to enhance that effect to further help the user locate relevant material.

A ranked list returned by a text search engine purports to present the documents in the order they are most likely to be relevant: the first document is the best match for the user's query, the second is the next most likely to be helpful, and so on. We are interested in situations where this simple model breaks down—where the user is unable to find enough relevant material in the first or second screens of the list. In particular, we are interested in helping a searcher find all of the relevant material in the top ranked list without forcing him or her to wade through all of the non-relevant material.

Our approach is based on a combination of document clustering and visualization. We have observed that when documents are clustered and their relationships are visually displayed, the relevant documents generally clump together in the visualization. In this study, we investigate several hypotheses related to this observation:

1. Clustering is useful for separating relevant and non-relevant documents. This hypothesis is critical to our work, but not at all novel or surprising

because it has been examined several times in the past and shown to be true. We confirm prior results as part of this study, but beyond reviewing the problem do not view this as a significant contribution of this work.

2. Graphical representations of clustering highlight the clumping of relevant documents. It is difficult for a searcher to assess the relationships between documents rapidly in a simple ranked list presentation of documents. We show evidence that 2- and 3-D presentations allow the clusters to be identified quickly.

3. Feedback techniques enhance the separation between relevant and non-relevant documents, and visualizations can capitalize on that improvement. If a searcher expends the effort to mark some documents as relevant and others as non-relevant, the separation between the two sets can be enhanced—among both the marked documents and also the unmarked part of the retrieved set.

4. 3-D visualizations of clustering are more useful for this purpose than 2- or 1-D presentations. (A 1-D presentation would be, for example, a ranked list of documents.) Document clustering is usually done in an extremely high-dimensional space (e.g., thousands of dimensions). When the relationships are presented graphically in 2 or 3 dimensions, some documents are necessarily shown "nearby" when they are actually unrelated. We examine whether the 3rd dimension helps with this problem by providing more "elbow room" for the embedding and thereby reducing accidental incorrect visual associations.

In the following sections we discuss our investigations of the hypotheses above and find support for the first three; evidence for the fourth is weak. The proposal and investigation of hypothesis 3 is the main contribution of the work.

Document clustering and visualization

The Cluster Hypothesis of Information Retrieval states that "closely associated documents tend to be relevant to the same requests". [15, p.45] The implication of the hypothesis' truth is that if one document is relevant to a query, then it is reasonable to include documents that are highly similar to that one: they, too, are likely to be relevant.

The Cluster Hypothesis was originally conceived as applying to an entire collection where it holds for only some collections.[16] There is strong evidence, however that the hypothesis is valid within a set of documents retrieved in response to a query. Two decades ago, Croft showed that the top-ranked documents usually contained a "best" cluster—one that had most of the relevant documents.[6] Hearst and Pedersen showed the same effect by using Scatter/Gather to cluster the top-ranked documents presented to searchers.[10]

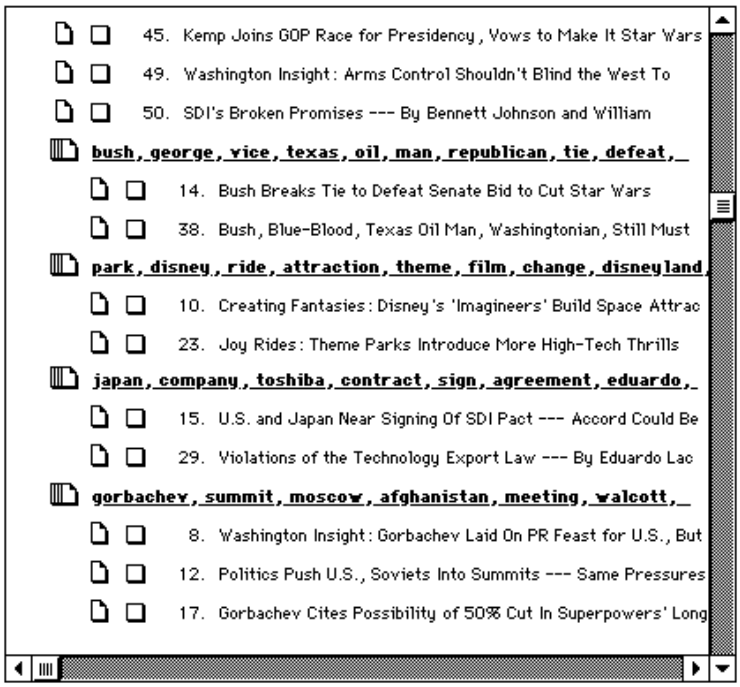

Figure 1: Clusters of documents resulting from the query "Star Wars" with the clusters presented textually.

Textual presentations

The Scatter/Gather interface[10] presents the document clusters as text. It groups the documents into five clusters and displays them simultaneously as lists. On a large enough screen, the top several documents from each cluster are clearly visible. The Superbook project includes a hierarchically clustered set of returned documents.[4] The use of document clustering in Information Retrieval has been extensive,[7, 16, 13] though surprisingly little of it has survived into actively used systems.

Another text-based visualization is presented by Leouski and Croft.[12] Their method is similar to the one used by Scatter/Gather, but does not fix the number of clusters to five. Instead, they cluster documents that are strongly related and allow documents to remain singletons if they are not well associated with other texts. Their display looks more like a standard ranked list because they can have an arbitrarily large number of clusters (limited only by the size of the retrieved set). Figure 1 shows sample output of the system for the query "Star Wars".

Graphical presentations

It is very common for clustering to be presented graphically. The documents are usually presented as points or objects in space with their relative positions indicating how closely they are related. Links are often drawn between highly-related documents to make their relationships clearer.

2-D visualization

Allan[1, 2] developed a visualization for showing the relationship between documents and parts of documents. It arrayed the documents around an oval and connected them when their similarity was strong enough. Allan's immediate goal was not to find the groups of relevant documents, but to find unusual patterns of relationships between documents.

The Vibe system[8] is a 2-D display that shows how documents related to each other in terms of user-selected dimensions. The documents being browsed are placed in the center of a circle. The user can locate any number of terms along the edge of the circle, where they form "gravity wells" that attract documents depending on the significance of that terms in that document. The user can shift the location of terms and adjust their weights to better understand the relationships between the documents.

3-D visualization

High-powered graphics workstations and the visual appeal of 3-dimensional graphics have encouraged efforts to present document relationships in 3-space. The LyberWorld system[11] includes an implementation of the Vibe system described above, but presented in 3-space. The user still must select terms, but now they are placed on the surface of a sphere rather than the edge of a circle. The additional dimension should allow the user to see separation more readily.

Our system is similar in approach to the Bead system[5] in that both use forms of spring embedding for placing high-dimensional objects in 3-space. The Bead research did not investigate the question of enhancing the separation of relevant and non-relevant documents. Figure 2 shows sample visuals of our system (they are explained in more detail in later sections).

Combining text and graphics

The system used for this study is described in detail elsewhere.[14] It is a tightly coupled set of displays including both a 3-D representation of clustered documents and a list presented in rank order of relevance. The displays are designed so that manipulating documents in one display has appropriate impact on the other displays—e.g., marking a document as "relevant" (in preparation for relevance feedback) causes a green check-box to be marked in the ranked list, but also makes the corresponding icon in the 3-D visualization turn green.

Clustering for evaluation

To study how well relevant and non-relevant documents are separated by clustering and visualizations, we used the TREC-5 ad-hoc queries and the corresponding database and relevance judgments.[9] Specifically, TREC topics 251-300 were converted into queries and run against the documents in TREC volumes 2 and 4. Three forms of query were used: (1) the title of the topic, (2) the description field of the topic, and (3) a query constructed by extensive analysis and expansion.[3] The intent was to examine the effect that different types of queries had on the results.

The top 50 documents for each query were selected. Because each version of the query behaved differently, there were three different lists per query. Any query that had fewer than 6 relevant documents in the top 50 or fewer than 3 in the top 10 was discarded. This resulted in 20 queries for the title-only version, 24 for the description level, and 26 for the full version.

Vector generation

Each document was transformed into a vector V such that v_i was a $tf{\cdot}idf$ weight of term t_i in the document. The result of this process is a set of vectors in t-space, where t is the size of the vocabulary of the 50 documents (around 3,000 in most cases). Distance between vectors was measured by $\sin\theta$, where θ is the angle between the vectors (identical vectors have a sine of 0.0, orthogonal vectors have a sine of 1.0, so all "distances" fall into that range).

Embedding vectors in 1- 2- and 3-space

In order to display t-dimensional vectors, they have to be approximated by vectors in a smaller number of dimensions. We used a spring embedding approached that is described elsewhere.[14] We generated spring-embeddings in 1-, 2-, and 3-space. 1-D "vectors" lie on a line; 2-D vectors in 2-space, and so on. For each dimension, we created vectors at every possible threshold–that is, every threshold value that resulted in a different set of documents being below the threshold (roughly 1,300 values). In the results below, only the best threshold value for each set of vectors is used. ("Best" was selected using the precision measure as the criterion, ties being broken by the recall measure.)

Space warping

One of our hypotheses is that if the system has information about the relevance or non-relevance of some documents, it can adjust the visualization to emphasize the separation between the two classes. To that end, we implemented a form of relevance feedback to create a new set of vectors.

The 10 top-ranked documents of the set of 50 being used were marked as relevant or not using the TREC relevance judgments. The relevant documents

in the top 10 (by design, there must be at least 3) were averaged to create a representative relevant vector, R. Similarly, the remaining of the 10 documents were averaged to create a representative non-relevant document, N. With $\Delta V = R - \frac{N}{4}$, the relevant vectors were modified as $V_i = V_i + \Delta V$ and the non-relevant vectors were modified by subtracting ΔV. Any resulting negative values were replaced by zero.

This approach is very similar to relevance feedback methods traditionally applied in Information Retrieval, but rather than modifying the query, the relevant documents themselves are modified to be brought "closer" to each other.

The enhanced vectors were in t-space and were then embedded in 1-, 2-, and 3-space as described previously. A second "embedding" in 1-space was created by placing the documents on a line, according to their retrieval belief value.

Restraining spheres

We are interested in visualizations that help the user find the relevant documents as rapidly as possible. We found that simple space warping was valuable, but that it tended to group the documents too tightly. We developed a modified version of the space warping that used "restraining spheres" to encourage separation. During the spring embedding phase, judged-relevant documents were forced to remain on the interior of a small sphere. Similarly, non-relevant documents were forced to remain *outside* a larger, enclosing sphere. Unjudged documents could appear anywhere, though they tend to settle within the non-relevant sphere and outside the relevant sphere.

Evaluation of generated clusters

We start with a set $\mathcal{D}$ of 50 vectors for one of the queries, one of the choices of dimension, and either the original or warped case. We find the center of relevance by averaging *all* relevant documents in the top 50 (not just those in the top 10). The center is the point R_{avg}. We then determine the standard deviation σ of the distances of relevant documents from R_{avg}.

Consider three spheres, each centered at R_{avg}, but with different radii. Sphere S_0's radius is the same as the average distance of relevant documents from the center point; S_1 has a radius that is larger by σ; S_2's radius is 2σ more than S_0's.

Let C_i represent all documents contained within sphere S_i. Note that $C_0 \subseteq C_1 \subseteq C_2 \subseteq \mathcal{D}$. Each of C_i is treated as a cluster and evaluated using several measures (presented as an average of the measures over the entire query set):

- **recall** measures the proportion of the relevant documents in the top 50 that made it into the cluster.
- **precision** is a measure of the "purity" of the cluster; it is the fraction of the documents in a cluster that are actually relevant.

		Rank	Regular embedding				Warped by feedback			
		List	t-D	3-D	2-D	1-D	t-D	3-D	2-D	1-D
Title	(%R)	85.7	81.3	88.8	87.2	85.7	81.2	87.9	89.5	85.9
queries	(%P)	50.0	95.2	81.5	79.6	71.7	95.8	88.7	87.8	81.9
	(N)	27.5	14.0	17.3	17.8	18.7	14.2	16.0	16.2	16.7
Desc.	(%R)	83.6	82.6	89.2	87.1	85.4	81.0	86.2	85.9	85.5
	(%P)	49.7	95.2	79.4	77.9	69.7	96.0	87.0	86.5	79.0
	(N)	28.0	14.7	18.6	19.0	20.0	14.3	16.5	16.8	17.7
Full	(%R)	83.2	82.8	87.0	86.2	85.7	81.5	86.8	87.0	85.8
	(%P)	55.6	87.0	74.0	72.3	69.0	89.1	79.4	77.3	72.3
	(N)	30.7	18.9	22.8	23.1	23.7	18.2	20.8	21.7	22.5

Table 1: Evaluation of documents within S_1. Each entry includes percent recall, percent precision, and cluster size. The first column of numbers is for the system's ranked list; the second group is a normal embedding; the last group is the result of warping space after feedback.

- **cluster size** counts the number of documents that are part of the cluster. Since we are always considering 50 retrieved documents, it is important that the cluster be of a reasonable size.

Results

It has been known for at least two decades that the Cluster Hypothesis is true within the top-ranked retrieved documents. We confirm those results even though our notion of "cluster" is different than that typically used: we are considering only the group of relevant documents and not actually partitioning the retrieved set into distinct clusters. Table 1 clearly shows that the relevant documents are tightly grouped in one place across all variants of the embedding. The data in the table were extracted by using the relevance sphere S_1 as described above. Very similar results were obtained for other spheres, with expected variations: the tighter S_0 sphere had higher precision and lower recall, while the broader S_2 sphere increased recall at the expense of precision.

The high precision values show that the grouped documents contain very few non-relevant documents. Equally important is the size of the recall value. The S_1 spheres contain almost all of the relevant material with very little useless information. That these high-quality groups occur with such a small cluster size (a third to a half of the retrieved set of 50 documents) means that only a small proportion of the 50 documents needs to be examined if the user can find the relevant group.

The "ranked list" column of Table 1 illustrates that the document ranking function used did not group similar documents together. In that case, the relevant documents are much more widely spread out, resulting in a higher cluster size and correspondingly low precision.

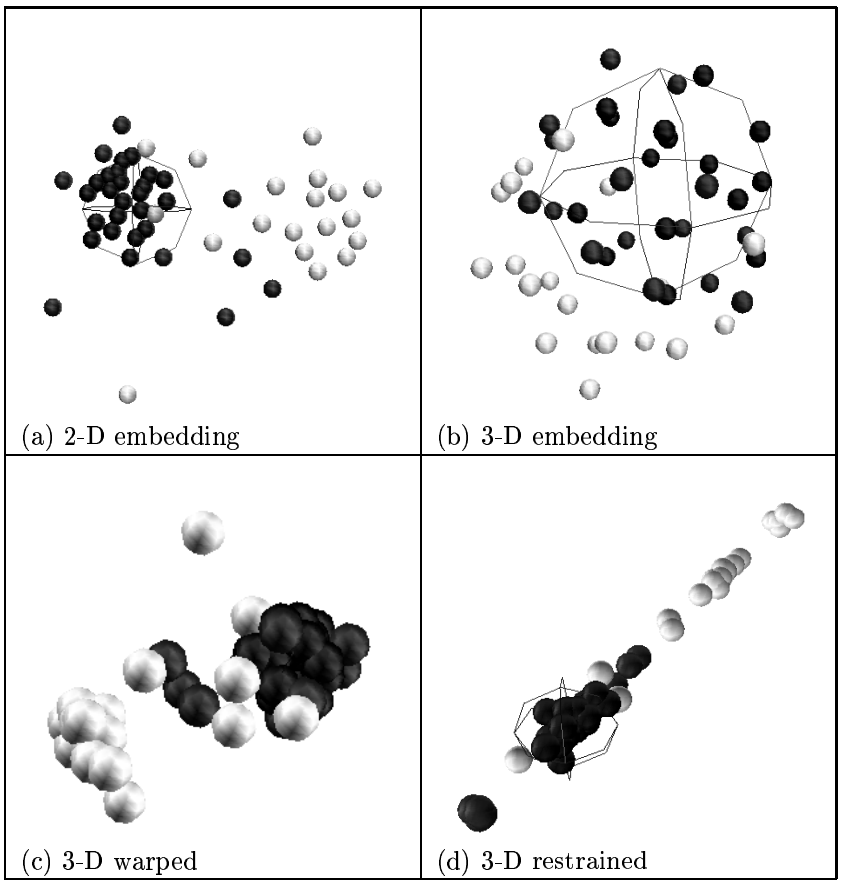

Figure 2: Visualization of retrieved documents for one of the queries. Both 2- and 3-space embeddings are shown, plus two variations on the 3-space. Relevant documents are shown as black spheres; non-relevant as grey. The wire-frame sphere is S_0, centered at at the average relevant document (R_{avg}).

Visualization of separation

Our second hypothesis was that graphical visualizations can show the clumping of relevant documents. Figure 2 shows several presentations of the 50 documents retrieved in response to a representative query. Figures 2a and 2b show that the relevant documents (dark spheres) are very well separated from the non-relevant documents (light spheres) in both 2- and 3-D embeddings of the visualization.

Space warping

The relevant documents are strongly grouped, but it will usually be difficult to find that group without some clues. The warping process described earlier strengthened the relationship between relevant documents and weakened that involving non-relevant documents. The hope was that by showing the user the location of the judged documents in the embedding, and by adjusting their location, the user would be able to find the separation point between relevant and non-relevant.

Table 1 shows the different values for before and after the warping was applied. For the full queries, the clusters on average became 5% smaller, increasing precision 6–7%, and having almost no impact on recall. The description queries showed a more pronounced version of the same effect: the cluster size dropped 10%, precision increased 10%, and recall dropped very little or not at all. The title queries performed similarly.

The change in effectiveness measures was most pronounced for the 1-D, 2-D, and 3-D spring embedding results. There was almost no change in the t-D measurements.

Figure 2c shows how the warping process can improve the separation between relevant and non-relevant documents. It shows the same documents as those in Figure 2b, but with space warping added. The relevant and non-relevant documents are still grouped apart from each other, but the location of the groups is much more readily seen—particularly since 10 of the documents in the presentation have already been judged.

The effectiveness measures confirm what the visual suggests. Our hypothesis is confirmed, that warping by a feedback process can improve the separation between relevant and non-relevant documents.

Advantages of 3-D

We have hypothesized that 3-D visualizations are more useful for presenting clustering than 1- or 2-D visualizations are because the extra dimension(s) minimize accidental but forced proximity of unrelated material. Unfortunately, our results do not support this hypothesis in any substantial way.

Compare the 2-D and 3-D numbers, both with and without warping, as presented in Table 1. For full queries, recall barely changes and precision rises 2.3%; with warping the numbers are -1.4% and 2.8%. For description queries the

results are slightly better, but still in the 2% range (closer to zero for warping). Title queries are similar.

The sample visualizations in Figures 2a and 2b also fail to support our hypothesis. The relevant documents are clearly separated in *both* the 2- and 3-D displays.

Restraining spheres

To enhance the separation of relevant documents, we modified the spring embedding procedure to restrain the relevant and non-relevant documents away from each other. Figure 2d shows the effect of the restraining spheres by contrasting it with normal space warping. In this particular case, the normal space warping would probably be useful, but the location of the unjudged relevant documents is even more obvious since the documents have been stretched apart.

One of the problems with our spring embedding algorithms is that they are based upon a threshold: document matches above the threshold are maintained; those below had their attractive force significantly dropped. Choosing the threshold is a difficult task and can induce wild changes in the resulting embedding. A side-effect of the restraining spheres is a substantial drop in the variability of the embedding in response to small changes in threshold. To show that, we measured the precision of sphere S_1 at every threshold value for three different embeddings over the 26 full queries. We then considered the change in precision for each change in threshold and measured the averages:

	3D	warped	restrained
mean	3.75%	3.92%	2.23%
stdev	4.67%	4.80%	4.19%

The lower mean and standard deviation of the restrained set show that the effect of changing the threshold is less pronounced, meaning the user will see fewer wild fluctuations as a threshold is adjusted. We believe that stability is an important aspect of interactive systems, so feel this advantage of the restrained graph is important.

Conclusions and Future work

Our hypotheses have been clearly supported, with the exception of the value of 3-D visualization over lower dimensionalities. 3-D appears to have a minimal advantage, but the evidence is too weak for any definitive conclusions. Visualizations in both 2- and 3-D successfully capture the clustering effect of relevant documents, and make it simpler for a user to identify new relevant documents given a few examples. Space warping and restraining spheres are important tools to help the user identify the unjudged relevant documents rapidly.

We are continuing to investigate visualization of clustering for the purpose of identifying relevant documents. The restraining spheres represent a form of

user-directed clustering that we hope to expand upon in the near future. We believe this early work on this idea is the core contribution of this study.

We are also interested in visualizations that show how new documents relate to previously known material, and have begun investigating appropriate visualizations for that question. Finally, we expect to apply these approaches to concept clustering; we have preliminary results that concepts (terms and phrases) from the top-ranked documents cluster as well as the documents themselves, and we hope that they may provide a means for more rapid identification of relevant material.

Acknowledgements

The authors thank Victor Lavrenko for his help with this study.

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, and by the National Science Foundation under grant number IRI-9619117. This material is also based on work supported in part by Defense Advanced Research Projects Agency/ITO under ARPA order number D468, issued by ESC/AXS contract number F19628-95-C-0235. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

References

[1] J. Allan. *Automatic Hypertext Construction.* PhD thesis, Cornell University, January 1995. Also technical report TR95-1484.

[2] J. Allan. Building hypertext using information retrieval. *Information Processing and Management*, 33(2):145–159, 1997.

[3] J. Allan, J. Callan, B. Croft, L. Ballesteros, J. Broglio, J. Xu, and H. Shu. Inquery at trec-5. In *Fifth Text REtrieval Conference (TREC-5)*, 1997. Forthcoming.

[4] R. Allen, P. Obry, and M. Littman. An interface for navigating clustered document sets returned by queries. In *Conference on Organizational Computing Systems*, pages 166–171, Milpitas, CA, Nov. 1993.

[5] M. Chalmers and P. Chitson. Bead: Explorations in information visualization. In *Proceedings of ACM SIGIR*, pages 330–337, June 1992.

[6] W. B. Croft. *Organising and Searching Large Files of Documents.* PhD thesis, University of Cambridge, October 1978.

[7] D. B. Crouch, C. J. Crouch, and G. Andreas. The use of cluster hierarchies in hypertext information retrieval. In *Hypertext '89 Proceedings*, pages 225–237, Pittsburgh, Pennsylvania, 1989. ACM Press.

[8] D. Dubin. Document analysis for visualization. In *Proceedings of ACM SIGIR*, pages 199–204, July 1995.

[9] D. Harman. The fifth Text REtrieval Conference (TREC-5). In *Fifth Text REtrieval Conference (TREC-5)*, 1997. Forthcoming.

[10] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of ACM SIGIR*, pages 76–84, Aug. 1996.

[11] M. Hemmje, C. Kunkel, and A. Willet. LyberWorld - a visualization user interface supporting fulltext retrieval. In *Proceedings of ACM SIGIR*, pages 254–259, July 1994.

[12] A. V. Leouski and W. B. Croft. An evaluation of techniques for clustering search results. Technical Report IR-76, Department of Computer Science, University of Massachusetts, Amherst, 1996.

[13] G. Salton, J. Allan, C. Buckley, and A. Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264:1421–1426, June 1994.

[14] R. C. Swan and J. Allan. Improving interactive information retrieval effectiveness with 3-d graphics. Technical Report IR-100, Department of Computer Science, University of Massachusetts, Amherst, 1996.

[15] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 1979. Second edition.

[16] E. M. Voorhees. The cluster hypothesis revisited. In *Proceedings of ACM SIGIR*, pages 188–196, June 1985.