Visual Interactions with a Multidimensional Ranked List*

Anton Leouski and James Allan

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003 USA

{leouski,allan}@cs.umass.edu

Abstract

Performance analysis of an interactive visualization system generally requires an extensive user study, a method that is very expensive and that often yields inconclusive results. To do a successful user study, the researcher has to be well aware of the system's possibilities. In this paper we present a different kind of analysis. We show how the system behavior and performance could be investigated off-line, without direct user intervention. We introduce an evaluation measure to assess the quality of a multidimensional visualization. Next, we suggest two methods for dealing with user's feedback. We also discuss the effect the dimensionality of the visualization has on the actual performance.

1 Introduction

An information retrieval system places retrieved documents in a list in the order they are most likely to be relevant: the first document is the best match to 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.

The Cluster Hypothesis of Information Retrieval states that "closely associated documents tend to be relevant to the same requests." [20, p.45]. We have been studying visualization techniques where retrieved documents are placed in 3-dimensional space and positioned according to the similarity among them. In full agreement with the hypothesis' truth we observed that relevant documents tend to appear in close proximity to each other, often forming tight "clumps" that stand apart from the rest of the material. In this study, we investigate several questions related to our observation:

- Although the system in our study does not explicitly create any clusters, the observed separation between relevant and non-relevant is a natural attribute of our visualization. Can we enhance that separation and make it easier for a user to detect?
- 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, can the separation between the two sets can be enhanced—among both the marked documents and also the unmarked part of the retrieved set?
- Is a high dimensional visualization more useful for this purpose than a low dimensional one? That is, is a 2-D picture more helpful than its 1-D counterpart, is 3-D better than 2-D? Document clustering is

*Extended version of work that will be presented as a poster in SIGIR, August 1998

usually performed in an extremely high-dimensional space (e.g., thousands of dimensions). When these configurations are forced down into 2 or 3 dimensions for the purpose of visualization, some documents are shown "nearby" when they are actually unrelated. We expect that visualizing in extra dimensions will show the separation among documents more clearly.

We begin by discussing related studies in clustering and visualization. In Section 3 we briefly summarize the visualization technique at the core of our system and proceed to define evaluation metrics used in this work in Section 4. Section 5 describes experimental setup and we conclude with discussion of the results in Section 6 and plans for future work.

2 Related Work

The Cluster Hypothesis was originally conceived as applying to an entire collection where it holds for only some collections.[21] 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.[9] Hearst and Pedersen showed the same effect by using Scatter/Gather to cluster the top-ranked documents presented to searchers.[13]

2.1 Textual presentations

The Scatter/Gather interface [13] 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. Another text-based visualization is presented by Leouski and Croft. [15] Their method is similar to the one used by Scatter/Gather, but does not fix the number of clusters to five. 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).

2.2 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 the fact that there is a relationship clearer.

2.2.1 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 [10] 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.

2.2.2 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 [14] includes an implementation of the Vibe system described above, but presented in 3-space. The user still must select terms, but now the terms 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 [7] 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 separating

relevant and non-relevant documents. Figure 1 shows sample visuals of our system (they are explained in more detail in later sections).

2.3 Evaluation of presentation

In their user study, Hearst and Pederson [13] showed that users are able to choose the cluster with the largest number of relevant documents using the textual summaries Scatter/Gather creates. This analysis does not apply in our situation, as we neither create clusters nor create any textual representations.

Our approach to evaluation carries some similarity to predictive evaluation (e.g., see Card and Morgan [6]): we define a precise task (the rapid identification of relevant material) and we evaluate the system particularly for this task. We also assume a set of possible strategies for the user. However, instead of predicting the actual time that is required to execute the task using the system, we estimate the *ability* of the system to *support* the task.

3 Spring-Embedder

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 approach, that we will briefly summarize here. (A detailed description of the system used can be found elsewhere. [19]) Other approaches are entirely possible. The idea of a spring-embedder is to model the objects with steel rings and to simulate inter-object relationship with a spring attached to the corresponding rings, creating a "mechanical" system. The system left to itself oscillates and assumes a minimal energy state. The spring-embedding is widely used for graph drawing.[11]

We model the documents with steel rings that repel each other with a constant force. If the distance between two documents is less than a predefined threshold, the corresponding rings are connected with a spring. The force constant of the spring is proportional to the original distance between documents. Starting from a random location this "mechanical" system goes through a number of iterations (usually 50). The resulting spatial structure is presented to the user.

Object placement can vary widely across iterations, but usually settles down. Changing the threshold value adds or removes springs and can have a dramatic effect on visualization. We address that problem with the approach in this study.

4 Evaluation Measure

The system used in this study generates spatial patterns of objects that correspond to the retrieved documents. We are interested in spatial properties of these patterns and require some evaluation technique to quantify these properties. This section discusses what kind of properties we are interested in, establishes requirements for evaluation technique, and suggests some statistics that might be of use. Specifically:

- Do the spatial locations appear to be random, or are they clustered? A spatial point pattern that exhibits some structure provides potentially more information than a set of randomly scattered objects. We require a statistical test to determine if the spatial pattern shows any structure.
- If the spatial pattern shows some structure, how much of the structure is there? We require a way to quantify the amount of "clumpiness" in the point pattern. Such a statistic is crucial for this study: different observers would disagree as to the amount of structure in the point pattern. Further, the process of obtaining such judgments would be enormously expensive.
- Suppose the objects in question are of different type, e.g. relevant and non-relevant documents. Given that we do not define any cluster boundaries, how can we measure the separation between objects of different type? How can we evaluate the "purity" of the spatial structure?

In the following sections we introduce a function K that serves to measure the amount of spatial structure. We also show how the ideas behind the K function could be used to extend and adapt the notion of precision to analyze the quality of these structures.

4.1 K function

The theory of point fields (i.e., point processes) [8, 18] introduces a simple and efficient technique for measuring spatial dependencies between different regions of a point pattern[1]. Consider a set of points in a d-dimensional space and a distance function on this space. Suppose λ is the number of points in a unit volume of space, or the *intensity* of the point field. Let $N(h)$ be the number of extra points within a distance h of a *randomly* chosen point. Then Barlett [5] defines the K function as

$$K(h) = \lambda^{-1} E(N(h)), \quad h \geq 0 \tag{1}$$

Ripley [16, 17] shows that the K function has properties that make it an effective summary of spatial dependence in a point field over wide range of scales.

The main application of the K function is to test if a point field exhibits any structure [18, p. 224]. To do that, we compute the values of K for the point field. We then compare these values against the values of the K function for some baseline point field. This baseline point field defines a completely random arrangement of points with neither clustering nor regularity among them, and consequently is called a "random point field".

It is customary [8] to model the random point field with Poisson point field. The K function for a d-dimensional Poisson field is defined as

$$K_{Poisson}(h) = \frac{\pi^{d/2} h^d}{\Gamma(1 + \frac{d}{2})} \tag{2}$$

It is also customary to use the following statistic $L(h)$ instead of $K(h)$:

$$L(h) = \sqrt[d]{K(h) \frac{\Gamma(1 + \frac{d}{2})}{\pi^{d/2}}} \tag{3}$$

When $L(h)$ is greater than $L_{Poisson}(h) \equiv h$, the system exhibits clusterization; $L(h) < h$ implies regularity in the point field.

To compute the values of the K function the expectation operator in (1) is replaced with an empirical average over the N given points:

$$\hat{K}(h) = \hat{\lambda}^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} I(\|s_i - s_j\| \leq h)/N, \quad i \neq j, h \geq 0. \tag{4}$$

Here $\hat{\lambda} = N/v(A)$ is the estimator of the intensity, $v(A)$ is the volume of observation area A, s_i is the location of the ith point, and $I(\cdot)$ is the indicator function:

$$I(x) = \begin{cases} 1, & \text{if } x \text{ is } true \\ 0, & \text{if } x \text{ is } false \end{cases} \tag{5}$$

To get a better view on how the K function works, recall that $K(h)$ is proportional to the number of points at most h away from an arbitrary point. If this number is high, we find a lot of points in a close proximity – we have clumps or clusters of points in the point field. If the number is low, we have gaps in the field. Because of the expectation operator in (1) these conclusions apply "on average" to the whole point field. Therefore, the K function should not be much affected by outliers. The function does not explicitly depend on point locations, making it independent of the shape of the point field.

[1] Random point fields are mathematical models for irregular 'random' point patterns. We will use this terminology to describe the location pattern of objects corresponding to the retrieved documents.

4.2 Recall and precision

The evaluation measures of recall and precision have a long history in Information Retrieval. We define two spatial statistics that closely resemble those metrics. We begin by extending the definition of the K function. Suppose we have a point field Ω and we have also selected two subsets of the point field ($\Omega_1 \subset \Omega$ and $\Omega_2 \subset \Omega$). Also suppose $N(h; \Omega_1, \Omega_2)$ is the number of points of set Ω_2 within distance h from an arbitrary point of set Ω_1. Then the version of the K function that describes spatial dependencies between members of these two subsets is defined as

$$K(h; \Omega_1, \Omega_2) = \lambda_2^{-1} E(N(h; \Omega_1, \Omega_2)), \quad h \geq 0 \tag{6}$$

Here λ_2 is the intensity of the set Ω_2. If the point field is located in space with a unit volume, λ_2 could be replaced with $|\Omega_2|$. Now we are ready to define *spatial recall* $R(h; \Omega_1)$ as the proportion of relevant documents within distance h from an arbitrary document of set Ω_1:

$$R(h; \Omega_1) = \frac{1}{|\Omega_R|} E(N(h; \Omega_1, \Omega_R)) = K(h; \Omega_1, \Omega_R), \tag{7}$$

where $\Omega_R \subset \Omega$ is the set of the relevant documents.

We will now define *spatial precision* as a function of spatial recall. For this we choose three subsets of the point field ($\Omega_i \subset \Omega,\ i = 1, 2, 3$). As an example think of $\Omega_2 \equiv \Omega_R$ and $\Omega_3 \equiv \Omega_N$ (Ω_N is the set of the non-relevant documents). Let us define $N(r; \Omega_1, \Omega_2, \Omega_3)$ as the proportion of the documents of set Ω_2 among documents of both Ω_2 and Ω_3 that are at least as close to an arbitrary document of set Ω_1 as are the closest r documents of set Ω_2. Then the spatial precision is defined as

$$P(r; \Omega_1, \Omega_2, \Omega_3) = E(N(r; \Omega_1, \Omega_2, \Omega_3)) \tag{8}$$

For example, pick a random relevant document and from its location start to grow a d-dimensional sphere. Let it grow until it includes two other relevant documents. Then $P(2; \Omega_R, \Omega_R, \Omega_N)$ is the expected fraction documents inside the sphere that are relevant. There is a particular similarity with a ranked list: given a starting point we move away from it, marking documents as we encounter them, recreating the ranking. Instead of moving in one direction, as in ranked list, we move out in all directions simultaneously. Think of it as traversing a "multidimensional" ranked list. One difference is that we potentially have several starting points that we are equally likely to choose from. In this case we average the performance over all these starting points.

The *average spatial precision* is then obtained by averaging $P(r; \Omega_1, \Omega_2, \Omega_3)$ over the set of possible value for r:

$$\bar{P}(\Omega_1, \Omega_2, \Omega_3) = E(P(r; \Omega_1, \Omega_2, \Omega_3)) \tag{9}$$

To compute $\bar{P}$ we replace the expectation operator in (9) with an empirical average:

$$\begin{aligned} \bar{P}(\Omega_1, \Omega_2, \Omega_3) &= \frac{1}{|\Omega_2|} \sum_{i=1}^{|\Omega_2|} \frac{1}{|\Omega_1|} \sum_{\forall k \in \Omega_1} \frac{i}{i + \sum_{\forall n \in \Omega_3} I(||k-n|| \leq \rho(i,k,\Omega_2))}, \\ \text{where } \rho(i,k,S) \text{ is such that } &\sum_{\forall r \in S} I(||k-r|| \leq \rho(i,k,S)) = i \end{aligned} \tag{10}$$

The average spatial precision $\bar{P}$ (a function) is a generalization of "conventional" average precision (a number). The conventional definition of the average precision assumes given sets of relevant and non-relevant documents (Ω_2 and Ω_3). It also assumes a starting point for the computation: the top of the ranked list (Ω_1). In the following text unless otherwise noted we use term "precision" to mean the average spatial precision.

5 Experiments

For this study we used TREC ad-hoc queries with the corresponding collections and relevance judgments [12]. Specifically, TREC topics 251-300 were converted into queries and run against the documents in TREC volumes 2 and 4 (2.1GB). Our intent was to study the effect that different types of queries have on the result. For each TREC topic we considered four types of queries: (1) the title of the topic; (2) the description field of the topic; (3) a query constructed by extensive analysis and expansion [3]; and (4) a query constructed from the title by expanding it using Local Context Analysis (LCA) [22].

The top 50 documents for each query were selected. Because each query behaved differently there were four different ranked lists for each topic. The lists with fewer than 6 relevant documents in the top 50 or fewer than 3 or greater than 9 relevant documents in the top 10 were discarded. This resulted in 20 queries for title-only version, 24 for the description queries, 26 for the full versions, and 17 for the expanded title version.

We also collected the same data using a different set of queries on a different collection. We used TREC topics 301-350 to create the queries and ran the queries against TREC volumes 4 and 5 (2.2GB). Again four different types of queries were constructed: (1) the title of the topic; (2) the title and the description field of the topic; (3) the full version constructed by expansion [4]; and (4) the expanded version of title query. The same restrictions were imposed on the retrieved set. This resulted in 25, 27, 25, and 22 queries of each type, respectively.

5.1 Vector generation and embedding

For each document we created a vector V such that v_i was a $tf \cdot idf$ weight of the ith term in the document. For each query this resulted in a set of vectors in t-space, where t is the size of the vocabulary of the top 50 retrieved documents (about 3000 words in most cases).

The t-dimensional vectors were embedded in 1-, 2-, and 3-dimensional space using the spring-embedder described above. Distance between vectors was measured by the sine of the angle between the vectors. The embedded structure depends on the number of springs among objects. This number is determined by a threshold: a maximum distance between documents at which the corresponding objects are connected with a spring. For a set of 50 objects there are 1225 different spring configurations, and therefore, 1225 different embeddings.

Nothing in the spring-embedding approach suggests a way of choosing one threshold value over another (i.e., one embedding over another). In the absence of the information we would have to randomly select one structure to show to the user. We analyze system performance by averaging precision over all possible values of threshold.

Our hypothesis is that embeddings with high spatial structure will have high precision score. Here we rely on the Cluster Hypothesis: if the spatial structure has clusters, it is likely that these clusters are of relevant documents. As an alternative to selecting the threshold value randomly, we choose an embedding with $\tau = \max(L(h) - h)$ in the top 20% of the values ranging over all threshold values.

Figure 1 shows several presentations of the 50 documents retrieved in response to a representative query. Figures 1a and 1b 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.

5.2 Warping

One hypothesis of this work was 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.

A subset of the 50 documents being used were marked as relevant or not using the TREC relevance judgments. The relevant documents in the subset were averaged to create a representative relevant vector, V_R. Similarly, the vectors for remaining selected documents were averaged to create a representative non-relevant document, V_N. With $\Delta V = V_R - 0.25 \cdot V_N$, the relevant vectors were modified as $V = V + \Delta V$ and the non-relevant vectors were modified by subtracting ΔV. Any resulting negative values were replaced by zero.

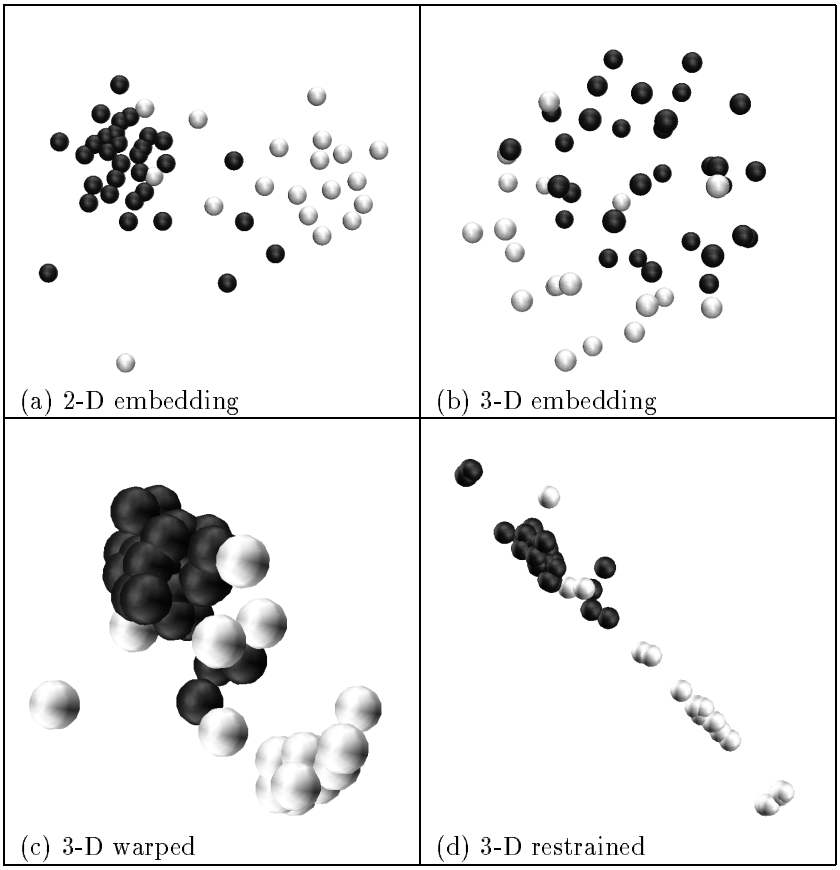

Figure 1: Visualization of retrieved documents for one of the queries. Both 2- and 3-space embeddings are shown, plus two variations on the 3-space. Relevant documents are shown as black spheres; non-relevant as grey.

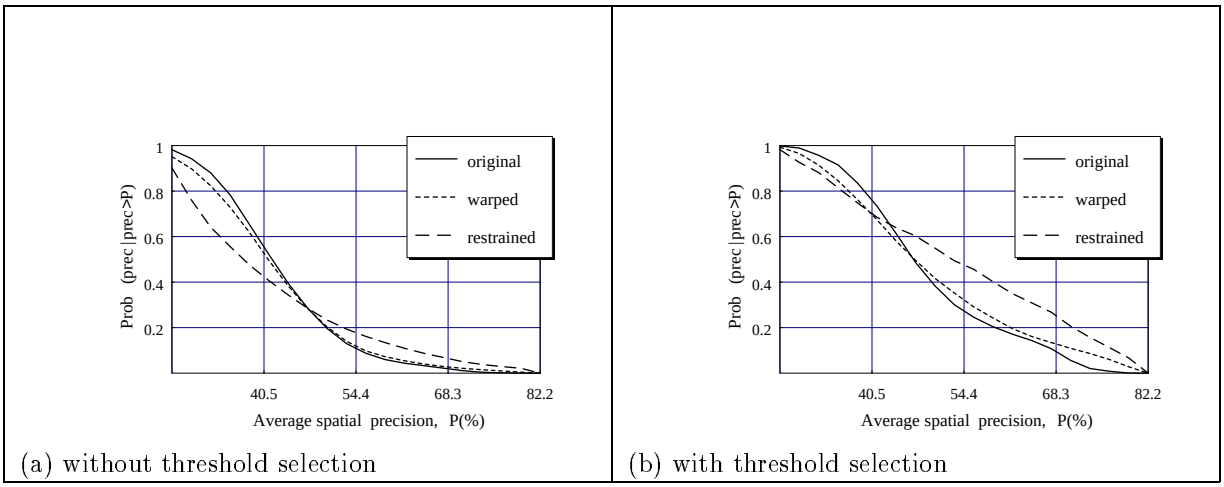

Figure 2: Probability of selecting an embedding at random with a given precision value or higher for the full queries on TREC5 collection in 2 dimensions. It illustrates the effect of different user's feedback techniques. (a) No restrictions are imposed on the set of possible embeddings. (b) The set of embeddings is limited by the threshold selection procedure.

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 vectors were modified in t-dimensional space and the entire set of 50 was then embedded in 1-, 2-, and 3-dimensional space as described previously. The hope was that unjudged relevant documents will move towards the known relevant, and unjudged non-relevant will behave correspondingly.

Figure 1c shows how the warping process can improve the separation between relevant and non-relevant documents. It shows the same documents as those in Figure 1b, 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.

5.3 Restraining spheres

An advantage of a ranked list is the direction it implies: the user always knows where to start looking for relevant information (at the top of the list) and where to go to keep looking (down the list). We observed that the space warping, however effective it is in bringing together relevant documents, tends to "crowd" the objects, making the whole structure more compact and not easily separable. We developed a small modification to the warping approach that enhances separation among documents. At the same time this technique creates a general sense of direction on the object structure.

During the spring-embedding phase, judged relevant documents were forced to lie inside a small sphere. Similarly, judged non-relevant documents were forced into another sphere positioned apart from the first one. The rest of the documents were allowed to assume any location *outside* of these spheres. In other words, we took the spring-embedded structure by the judged documents and "pulled it apart".

Figures 1d shows the effect of restraining spheres. In this particular case, the simple warping would probably be useful, but the location of unjudged relevant documents is even more obvious since the documents have been stretched.

6 Results and analysis

We begin by assuming that the user has identified two documents: one relevant and one non-relevant. (We believe this is a reasonable strategy and almost always could be done by looking at the titles in the ranked

Queries		Rank List	t-D space	Embedding w/o threshold selection			Embedding w/ thresh. selection		
				1-D	2-D	3-D	1-D	2-D	3-D
TREC5	Title	63.0	43.8	38.0	41.8	41.8	42.5	58.2	59.1
	Desc.	54.7	42.1	39.2	42.1	42.1	41.0	51.3	52.2
	Full	58.4	53.1	45.3	46.3	46.7	47.0	49.9	50.9
	Exp. Title	66.6	60.0	46.6	48.5	48.5	49.0	57.3	59.7
TREC6	Title	58.8	52.1	44.7	47.5	47.8	46.9	57.6	59.9
	Desc.	57.7	48.2	39.8	44.0	44.6	41.7	54.8	55.3
	Full	68.6	53.9	42.5	48.8	49.5	43.4	57.9	59.4
	Exp. Title	64.3	52.0	42.3	45.5	45.9	44.0	55.9	59.0

Table 1: Visualization quality evaluation of different query sets in different dimensions. Percent of average precision is shown. The first column is for the system's ranked list. The second column is for the original structure in t-dimensional space. The third column shows the result of spring-embedding. The last column is for embedding with threshold selection done by $L(h)$ measure.

list.) For simplicity, let us assume the user identified the highest ranked relevant and the highest ranked non-relevant document. We evaluate how quickly the user would be able to find the rest of the relevant documents starting from the known relevant one.

Table 1 shows average precision values for different query sets in different dimensions. (Recall that spatial precision is used.) The ranked list is treated as an embedding in 1-dimension where each document is positioned on a line according to its rank value. The numbers for the ranked list are always better than the numbers for the embedded structures.

6.1 Threshold selection

In the absence of any other information, a threshold value would has to be chosen randomly. Limiting our choice to the embeddings with spatial structure (τ) in the top 20% has proved very effective. The average precision across all "eligible" threshold values was significantly increased by 17.2% ($p_{t-test} < 10^{-5}$). The solid line on Figures 2a and 2b show how the threshold selection procedure increases the probability of randomly choosing a high quality spatial structure without any user supplied information. The effect is also consistent across relevance feedback methods. Note that there is almost no change in maximum and minimum values of precision. It means the method does not limit our choices of quality on the spatial structure: it just makes it more probable we will select a "good" one.

6.2 Dimensionality effect

We hypothesized that extra dimensions would prove useful for the task of visualization of separation between documents. Our results support this hypothesis only partially. Indeed, a step from 1 dimension to 2 leads to a statistically significant jump of 23.1% in precision ($p < 10^{-5}$). The difference between 2- and 3-dimensional embeddings is 1.1%, and this result, however consistent, is not significant. (It is significant by the sign test, but not by the t-test. The cut-off value of $p = 0.05$ is used in both tests.)

Figure 3 shows how an increase in dimensionality of embedding leads to a general growth in precision. It is difficult to see, but the maximum precision value for 1-D is higher than for 2-D or 3-D. It means that a better separation between relevant and non-relevant documents could be achieved in 1-dimension than in 2- or 3-dimensions. However, to randomly select a high precision structure in 1-dimension is extremely difficult.

6.3 Interactive embedding

We studied how the quality of the visualization changes as the system is supplied with more and more relevance information. Given the first relevant/non-relevant pair of documents, we use it to warp the embedding

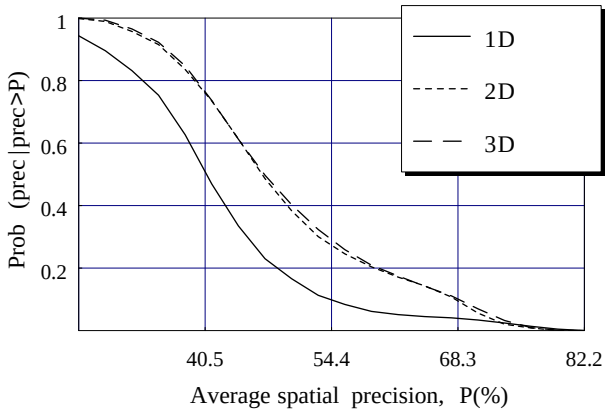

Figure 3: Probability of selecting an embedding with a given precision value or higher for the full queries on TREC5 collection. The effect of different dimensions on the original embedding is illustrated. The set of embeddings is limited by the threshold selection procedure. The values on X-axis are averaged over the query set.

Type of feedback	Number of pairs judged 0	1	2	3	5
warping	49.3	50.5	51.4	51.4	51.5
restraining	49.3	49.9	51.4	52.3	53.4

Table 2: Average precision computed starting from the first 5 relevant documents. The retrieved documents from TREC5/full queries are embedded in 2 dimensions. The first column of numbers is for the case when no feedback has been yet received.

space and apply the restraining spheres. Then we add the information about the next relevant/non-relevant pair. And so on. Table 2 and Figure 4 illustrate how the average precision increases as more data become available to the system. We show the average precision computed starting from five top ranked relevant documents. The warping does not have any effect after the first two steps. The restraining spheres keep pulling the documents apart; however, their influence is also diminishing.

6.4 User's feedback

Our second strategy was to evaluate the effect of a user's feedback on the visualization. Suppose the user extended his or her effort and judged the top 10 documents in the ranked list. We are interested in how fast it is possible to identify the rest of the relevant material starting from the known relevant documents. We compare the effects warping and restraining have on this task.

From Table 3 we conclude that warping did not do as well as we expected. It increased average precision by 1.1% consistently, but not significantly ($p_{sign} < 0.02$ and $p_{t-test} < 0.37$). It actually *hurt* precision in 3D. The effect of warping together with restraining was more profound and nearly always beneficial. The procedure significantly increased precision by 7.4% ($p_{sign} < 0.001$ and $p_{t-test} < 0.037$).

Figures 2a and 2b show that feedback techniques increase the probability of selecting an embedded structure with high precision value. The growth is observed both with and without threshold selection, but with threshold selection the difference between restrained and original cases is more prominent.

We also observed a strong effect that poorly formulated and ambiguous queries have on feedback. The

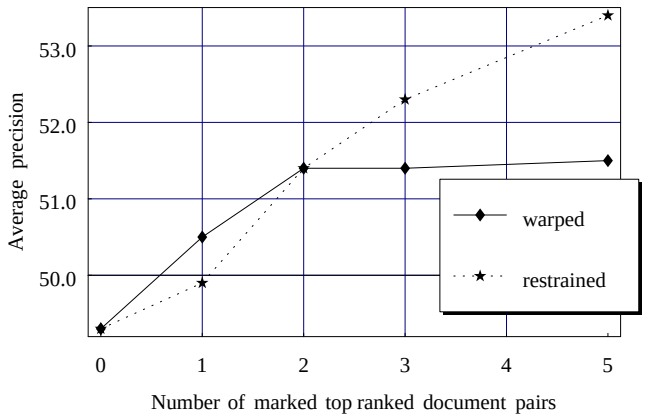

Figure 4: Average precision computed starting from the first 5 relevant documents. The retrieved documents from TREC5/full queries are embedded in 2 dimensions.

restraining spheres largely decreased the precision of the embeddings generated for documents retrieved by the title queries on TREC5 collection. Expanding the "bad" queries (see "TREC5/Exp. Title" row in Table 3) to eliminate the possible ambiguity seems to alleviate the problem. The TREC6 title queries were created to be of higher quality and ranked better.

7 Conclusion

- It has been known for at least two decades that the Cluster Hypothesis is true within the top-ranked retrieved documents. Although the system used in this study does not explicitly generate clusters, we show that the objects representing relevant documents tend to group together. Each query has, on average, about 18 relevant documents in the top 50. If the documents are randomly scattered in space, one would expect an average precision be about 27.8%. The average precision value around 50% speaks of clustering among relevant documents.

- In the context of our visualization, we confirmed the hypothesis that relevance feedback methods can improve separation between relevant and non-relevant documents. Figure 1 shows an example of how these methods can have a significant influence on the embedding structure.

- We have hypothesized that an extra dimension is always helpful for visualization. Our results support this hypothesis only partially. There is a clear advantage in using higher dimensions over 1-D. However, there is almost no improvement in adding an extra dimension to a 2-D visualization.

- We have introduced an evaluation technique to assess the system's performance off-line. That allowed us to collect a large amount of data to make statistically significant claims about the system's quality without requiring an extensive user study. Given the results of this study, a follow-up user study might be useful to confirm the conclusions.

- The suggested visualization method works on average just about as well as a ranked list for finding relevant documents. In another study [4] most of the users loved this visualization: they found it intuitive and fun to use. In this study we showed that although the visualization does not help, it at least does not hinder the actual performance.

Queries		Rank List	Embedded in	Original	Warping		Restraining	
TREC5	Title	46.8	1-D	35.7	36.2	(+1.3%)	31.5	(-11.7%)
			2-D	47.3	48.9	(+3.3%)	40.7	(-14.0%)
			3-D	48.4	50.3	(+3.9%)	40.2	(-17.0%)
	Desc.	40.8	1-D	38.6	39.8	(+3.3%)	37.1	(-3.1%)
			2-D	48.5	48.8	(+0.6%)	49.6	(+2.2%)
			3-D	49.6	48.3	(-2.5%)	47.3	(-4.5%)
	Full	43.1	1-D	41.9	42.4	(+1.2%)	47.3	(+12.8%)
			2-D	45.9	47.1	(+2.8%)	52.0	(+13.4%)
			3-D	46.1	47.0	(+2.0%)	47.5	(+3.0%)
	Exp. Title	42.5	1-D	42.4	42.7	(+0.6%)	51.7	(+22.1%)
			2-D	46.2	46.8	(+1.4%)	54.4	(+17.8%)
			3-D	46.6	46.2	(-0.8%)	52.4	(+12.5%)
TREC6	Title	50.6	1-D	42.9	45.0	(+4.8%)	45.7	(+6.6%)
			2-D	53.6	53.9	(+0.7%)	57.4	(+7.2%)
			3-D	55.9	55.4	(-0.9%)	58.9	(+5.5%)
	Desc.+Title	45.7	1-D	37.6	38.8	(+3.2%)	43.8	(+16.6%)
			2-D	49.8	51.0	(+2.4%)	56.2	(+13.0%)
			3-D	51.3	50.9	(-0.8%)	56.4	(+9.9%)
	Full	53.1	1-D	36.3	37.4	(+2.9%)	44.5	(+22.5%)
			2-D	46.6	47.0	(+0.7%)	55.3	(+18.5%)
			3-D	48.9	47.5	(-2.8%)	54.0	(+10.5%)
	Exp. Title	53.7	1-D	39.1	38.7	(-0.9%)	42.5	(+8.9%)
			2-D	48.4	49.7	(+2.8%)	56.0	(+15.7%)
			3-D	50.6	50.0	(-1.1%)	56.5	(+11.7%)

Table 3: Relevance feedback effect on different queries in different dimensions. Percent of average spatial precision is shown. The threshold selection procedure was applied.

- The Cluster Hypothesis also helped us to select good embedding structures. As a result we show that embeddings with high clumpiness value τ tend to have higher precision.
- We have also done some "best case" analysis, when instead of averaging precision over the set of possible embeddings we considered the structure with highest precision. In this case the values are about 15-20 points higher than in the average case. There are good embeddings out there; it is just difficult to find them.

7.1 Future work

In this study we considered only two classes of documents: relevant and non-relevant. This was caused by the lack of data of any other kind. We are looking into extending our approach into situation when the user places the relevant documents into multiple classes. That task is modeled after the latest interactive TREC task of "aspect retrieval."

We are planning to do more work to investigate different user's strategies before attempting a real user study. The user study is a useful final test of our hypotheses. We are also interested in visualizations that show how new documents relate to previously known material.

Acknowledgments

We would like to thank Russell Swan for the preliminary work on the 3-D spring embedder evaluated in this study.

This material is based in part 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. This material also is based on work supported in part by the National Science Foundation under grant number IRI-9619117, and in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

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)*, pages 119–132, 1997.

[4] J. Allan, J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu. Inquery does battle with trec-6. In *Sixth Text REtrieval Conference (TREC-6)*, 1998. Forthcoming.

[5] M. S. Barlett. The spectral analysis of two-dimensional point processes. *Biometrika*, 51:299–311, 1964.

[6] S. Card and T. Moran. User technology: from pointing to pondering. In Baecker, Grudin, and B. an Greenberg, editors, *Readings in Human-Computer Interaction: towards the year 2000.* Morgan Kaufmann, 1995.

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

[8] N. A. C. Cressie. *Statistics for Spatial Data.* John Willey & Sons, 1993.

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

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

[11] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software–Practice and Experience*, 21(11):1129–1164, 1991.

[12] D. Harman and E. Voorhees, editors. *The Fifth Text REtrieval Conference (TREC-5)*. NIST, 1997.

[13] 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.

[14] 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.

[15] 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.

[16] B. D. Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13:255–266, 1976.

[17] B. D. Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society*, 39:172–192, 1977.

[18] D. Stoyan and H. Stoyan. *Fractals, Random Shapes and Point Fields*. John Willey & Sons, 1994.

[19] 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.

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

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

[22] J. Xu and W. B. Croft. Querying expansion using local and global document analysis. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 4–11, 1996.