Evaluating a Visual Navigation System for a Digital Library

Anton Leouski and James Allan

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

leouski@cs.umass.edu, allan@cs.umass.edu

Abstract. In this paper we investigate a general purpose interactive information organization system. The system organizes documents by placing them into 1-, 2-, or 3-dimensional space based on their similarity and a spring-embedding algorithm. We begin by designing a method for estimating the quality of the organization when it is applied to a set of documents returned in response to a query. We show how the relevant documents tend to clump with each other in space. We proceed by presenting a method for measuring the amount of structure in the organization and we explain how this knowledge can be used to refine the system. We also show that increasing the dimensionality of the organization generally improves its quality. We introduce two methods for modifying the organization based on the information obtained from the user and show how such feedback improves the organization. All the analysis is done off-line without direct user intervention.

1 Introduction

An important part of a digital library is the ability to access the stored information effectively. Methods for interactive information presentation and organization have been getting much attention in recent years. In this paper we present a non-interactive approach for evaluating a graphical document organization system. This is a general purpose tool that could find important applications in the context of digital libraries when it is necessary to locate what one is interested in rapidly or to assess the content of a document group quickly. Our study of this system in an information retrieval setting is motivated by automatic search and retrieval being an important way of accessing the content of a digital library and also by the large amount of readily available experimental data.

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 ten retrieved documents.

In particular, we are interested in helping a searcher find all of the relevant material in the ranked list without forcing him or her to wade through all of the non-relevant material. We believe that in this case an information organization technique that arranges the retrieved data and reveals how individual documents relate to each other will help the user to isolate relevant material quickly.

In this study we investigate an interactive visualization technique where retrieved documents are placed in 3-dimensional space and positioned according to the similarity among them [5]. To begin with, the documents are represented as vectors of terms, the vector size is equal to the vocabulary size of the retrieved set. Each retrieved document's vector defines a point in a high dimensional space. The distances between these points and their relative position are strong indicators of the similarity among the corresponding documents. Unfortunately, it is difficult to visualize objects in more than three dimensions. To display the points properly and show the relationships to the user we need to reduce the number of dimensions to 1, 2, or 3. There are many different algorithms that do dimensionality reduction. We use spring-embedding in our system.

The following questions are investigated in this study:

- The Cluster Hypothesis of Information Retrieval states that "closely associated documents tend to be relevant to the same requests." [20, p.45]. In our experiments we found continued support for the hypothesis' truth in retrieved documents – relevant documents tend to appear in close proximity to each other, often forming tight "clumps" that stand apart from the rest of the material. Is the observed separation between relevant and non-relevant a natural attribute of the visualization?
- 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 be enhanced – among both the marked documents and the unmarked part of the retrieved set?
- Is a high dimensional visualization more useful than a low dimensional one for the purpose of isolating the relevant documents? That is, is a 2-D picture more helpful than its 1-D counterpart; is 3-D better than 2-D? The documents exist 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 relationships among documents more accurately and the relevant documents will be better isolated from non-relevant ones.

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 the 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 [10]. More recently Hearst and Pedersen showed the same effect by using Scatter/Gather to cluster the top-ranked documents presented to searchers [14].

2.1 Textual Presentations

The Scatter/Gather interface [14] presents the document clusters as text. It groups the documents into five (or another preselected number) 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 [16]. Their method is similar to the one used by Scatter/Gather, but the number of clusters is based on a similarity threshold. 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 clusters 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-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 [11] is a 2-D display that shows how documents relate 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 inside the circle and along its edge, 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 [15] 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 [8] 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 [14] 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 Moran [7]): 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 high-dimensional vectors, they have to be approximated by vectors in a smaller number of dimensions. We used a spring-embedding approach [12]. Our choice was rather arbitrary and any other technique (e.g., Linear Programming) is entirely possible.

The idea of spring-embedder is simple but elegant: Consider a set of points in a high-dimensional space and a function that defines the distance between two points. We will call this high-dimensional space t-space, where t is the actual dimension of the space. Consider also a low-dimensional space where this point set is going to be visualized e.g., 1, 2, or 3 dimensions. We call this space v-space (as in visualization). The algorithm creates a point configuration in v-space space that "mimics" the configuration in t-space – it attempts to preserve the relative distances and positions of the points in t-space. Generally, it is impossible to reproduce the same configuration *exactly* in low dimensions.

Each object in t-space is modeled with a steel ring in v-space. The rings repel each other with a constant force: the rings are pushing away from each other and the system is striving to break apart. The "break-away" does not happen because the rings are inter-connected with springs. The force constant of a spring is proportional to the original distance between points in t-space. This way a "mechanical" model is created. Left to itself the model oscillates and assumes an "optimal" final state. If two points were very close to each other in t-space, the corresponding rings are connected with a very strong spring, and

they are very likely to end up close to each other. On the other hand, the rings that correspond to pairs points that are far apart have a weak link and the general repulsive force among the rings will push them apart.

Although ring placements may vary widely across oscillations, the final configuration does not usually depend on the original ring locations and these locations are randomly selected. For N objects there are $(N^2 - N)/2$ springs. If all springs are presented in the model, all rings are connected very strongly and the final configuration tend to resemble a tight "soccer-ball." To prevent this from happening and to reduce computational expense, we impose a limit on the inter-point distances in t-space. If a distance between two points in t-space exceeds a predefined threshold, such points are consider to be infinitely far apart and the corresponding rings are not connected with a spring. Indeed, this allows us to model a situation when two documents are said to be different, when at the same time they have some terms in common.

Unfortunately, selecting the right threshold is a difficult task. Changing the threshold value adds or removes springs in the model and can have a dramatic effect on visualization. This leads us to another question that we investigate in this study:

- For N objects there are $(N^2-N)/2$ springs, so there are $(N^2-N)/2$ different threshold values as each threshold allows an additional spring into the model. In the absence of any information, a threshold has to be chosen randomly across all $(N^2 - N)/2$ values. How much effect has a threshold value on the final configuration? How can we select the "best" threshold value, or at least limit our choice only to "good" thresholds?

4 Evaluation Method

Our proposed visualization analysis is task-oriented. For this study we define the task as *fast identification of relevant material.* Given that the user already knows some of the relevant documents, the question is: how fast could he or she identify the rest of the relevant documents from the visual image created by the system? There are two main components to this problem. First, there is the spatial configuration of the points representing the retrieved data – the amount of structure in the image, the separation between relevant and non-relevant objects. Second, there is a question of how the user goes about finding the relevant information – the user's strategy. Both of these aspects are the variables in the evaluation.

4.1 Spatial Properties

The system used in this study represents documents with objects that are floating in space. These objects form visual patterns. 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 the 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 any structure, what is the extent of the structure? 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.

4.2 K Function

The theory of point fields (i.e., point processes) [9, 19] 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 [6] defines the K function as

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

where $E(\cdot)$ is the expectation operator on the point field. In other words, the K function is the average number of points in the point field within distance h of any point in this field, normalized by the number of points in a unit volume of space. Practically, it measures a local concentration of points, or what part of the point field on average is within distance h of any point in the point field. Ripley [17, 18] 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 [19, p. 224]. Indeed, $K(h)$ is proportional to the number of points

[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.

at most h away from an arbitrary point. If this number is unusually high, we find many points in close proximity of every given point – i.e., we have clumps or clusters of points in the point field. If the number is low, we have few points in close proximity – i.e., 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.

The K function is just a metric for comparing one point field to another. To decide if the the point field has clusters, we compare this field to some configuration that is known not to have clusters. Generally, a completely random arrangement of points with neither clumps nor gaps is selected. This configuration of points is called a "random point field."

It is customary [9] to model this random point field with a Poisson point field, a configuration where a point is equally likely to occupy any location in the space of the field. The only condition is that no two points can occupy the same location in space. 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 $K(h)$:

$$L(h) = \sqrt[d]{K(h) \frac{\Gamma(1 + \frac{d}{2})}{\pi^{d/2}}}, \text{ note that } L_{Poisson}(h) \equiv h \tag{3}$$

When $L(h)$ is greater than $L_{Poisson}(h) \equiv h$, there are clumps in the point field; $L(h) < h$ implies gaps in the configuration.

The test variable

$$\tau = \max_{h \leq h_0} |L(h) - h| \ , \tag{4}$$

is used to test the amount of structure in the point fields, here h_0 is the upper bound on the interpoint distance. The outcome of the test is based on comparing τ with its table values [19, p. 225].

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

$$\hat{K}(h) = \hat{\lambda}^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} I(\|s_i - s_j\| \leq h)/N, \quad i \neq j, h \geq 0 \ . \tag{5}$$

Here $\hat{\lambda} = N/v$ is the estimator of the intensity, v is the volume that contains the point field, s_i is the location of the ith point, and $I(x)$ is the indicator function: 1 if x is true, 0 otherwise.

4.3 Average Spatial Precision

Evaluation using recall and precision has a long history in Information Retrieval. Using the same ideas that are behind the K function we define a spatial statistic that closely resembles precision. We call this measure *spatial precision*.

Suppose we have a point field Ω_R – a set of points representing the retrieved documents. As defined by our task, some of the relevant documents are known to us, some of them are not. We are interested in how fast we could find the unknown relevant documents. We select three subsets of the point field:

- $\Omega_{KR} \subset \Omega$ – points that represent known relevant documents.
- $\Omega_{UR} \subset \Omega$ – points that represent unknown relevant documents.
- $\Omega_{UN} \subset \Omega$ – points that represent unknown non-relevant documents.

Let us define $N(r; \Omega_{KR}, \Omega_{UR}, \Omega_{UN})$ as the proportion of the documents of set Ω_{UR} among documents of both Ω_{UR} and Ω_{UN} that are at least as close to an arbitrary document of set Ω_{KR} as are the closest r documents of set Ω_{UR}. Then the spatial precision is defined as:

$$P(r; \Omega_{KR}, \Omega_{UR}, \Omega_{UN}) = E(N(r; \Omega_{KR}, \Omega_{UR}, \Omega_{UN})) \tag{6}$$

For example, pick a random known relevant document and from its location start to grow a d-dimensional sphere. Let it grow until it includes two unknown relevant documents. $P(2; \Omega_{KR}, \Omega_{UR}, \Omega_{UN})$ is the expected fraction of the unknown 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. It is can be thought of as traversing a "multidimensional" ranked list. One difference is that we have several starting points that we are equally likely to choose from (i.e., members of Ω_{KR}). In this case we average the performance over all these starting points.

The *average spatial precision* is then obtained by averaging $P(\cdot)$ in (6) over the set of possible values for r:

$$\bar{P}(\Omega_{KR}, \Omega_{UR}, \Omega_{UN}) = E(P(r; \Omega_{KR}, \Omega_{UR}, \Omega_{UN})) \tag{7}$$

To compute $\bar{P}$ we replace the expectation operator in (7) with an empirical average:

$$\bar{P}(\Omega_{KR}, \Omega_{UR}, \Omega_{UN}) = \frac{1}{|\Omega_{UR}|} \sum_{i=1}^{|\Omega_{UR}|} \frac{1}{|\Omega_{KR}|} \sum_{\forall k \in \Omega_{KR}} \frac{i}{i + \sum_{\forall n \in \Omega_{UN}} I(||k-n|| \leq \rho_{i,k})}, \tag{8}$$

where $\rho_{i,k}$ is such that

$$\sum_{\forall r \in \Omega_{UR}} I(||k-r|| \leq \rho_{i,k}) = i \tag{9}$$

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_{UR} and Ω_{UN}). It also assumes a starting point for the computation: the top of the ranked list (Ω_{KR}). In the following text unless otherwise noted we use term "precision" to mean the average spatial precision.

4.4 User's Strategy

As we mentioned at the beginning of this section, the evaluation analysis requires some assumptions about the user's strategy or how the user is looking for the relevant material. It is impossible to define the degree of separation between the relevant documents and the non-relevant documents without assuming some searching strategy first. Generally, the strategy is rather intuitive and goes unspecified. For example, consider a linear separation test – two sets of points in 2-dimensions are considered well-separated if it is possible to draw a straight line between them. Here the assumed strategy is "draw the line; consider all the points that are on one side of the line."

We have assumed a particularly user's strategy when we defined the average precision. We assumed that the user begins from an arbitrary known relevant document and looks at the closest unknown document. He or she then proceeds to the next closest document and so on. In other words, the unknown documents are reordered based on their proximity to the starting point. Average precision serves as the measure of the visualization effectiveness. It characterizes not only the spatial configuration presented by the visualization but also the user's strategy. Some experiments to validate the suggested strategy are clearly needed. This was not done for this paper.

Note that this strategy depends to some degree upon the visualization we have described. Other presentation approaches might require a different model of interaction.

5 Experiments

For this study we used TREC ad-hoc queries with the corresponding collections and relevance judgments [13]. 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. We are interested in situations when there was not *enough* relevant material in the top ten documents. We ignored each run that contained too many relevant

documents – it is a success already and the visualization is unnecessary. We also discarded complete failures, or runs that had just a few relevant documents. We are interested in how the visualization changes when the user's feedback about both relevant and non-relevant documents is provided. A small amount of either relevant or non-relevant data renders such analysis uninteresting. Therefore, 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 in Section 3. Distance between vectors was measured by the sine of the angle between the vectors. The embedded structure depended 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.

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 Threshold Selection

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.

We also determine what is the probability of randomly selecting a "good" threshold value. For all queries in question and for all possible spatial embeddings (for all threshold values) we count the number of of times each average precision value occurs and normalize them over the total number of embeddings. This

gives us probability distribution for precision values. If we take this distribution, fix some value of the precision ($prec_0$), and add all the values in the distribution for each point that exceeds $prec_0$, we compute the probability for an arbitrary selected spatial configuration among all possible embeddings (remember, that the embeddings vary because of the threshold) to exceed $prec_0$, or $P(prec|prec > prec_0)$

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 "pure" clusters of relevant documents. The clusters of non-relevant documents are also possible, but "mixed" clusters are less likely. 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.

5.3 Warping

One hypothesis of this work was that if the system had information about the relevance or non-relevance of some documents, it could adjust the visualization to emphasize the separation between the two classes. To this end, we implemented a form of relevance feedback to create a new set of vectors.

A subset of the 50 documents being used was presumed known and marked as relevant or not using the TREC relevance judgments. We experimented with the subsets of different sizes. The known relevant documents were averaged to create a representative relevant vector, V_R. Similarly, the remaining known non-relevant documents were averaged to create a representative non-relevant document, V_N. With $\Delta V = V_R - 0.25 \cdot V_N$, each known relevant vector was modified by adding ΔV to it and the known 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 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 would move towards the known relevant, and unjudged non-relevant would shift towards the known non-relevant.

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 easily recognizable – particularly since 10 of the documents in the presentation have already been judged.

5.4 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 spring-embedding, 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 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 using the spatial information.

6.1 Threshold Selection

In the absence of any other information, a threshold value would have 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 information supplied by the user. 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 Comparison to Ranked List

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

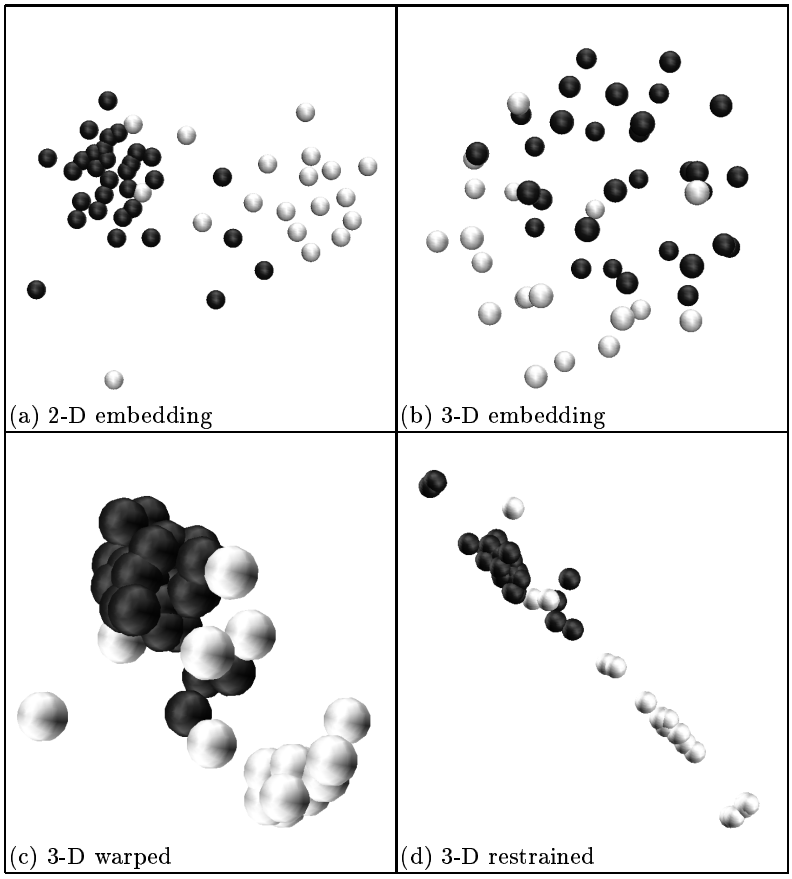

Fig. 1. Visualization of retrieved documents for one of the queries. Both 2- and 3-space embeddings are shown, plus two variations on the 3-space. Relevant documents are shown as black spheres; non-relevant as grey.

to its rank value. In this table, the numbers for the ranked list are always better than the numbers for the embedded structures. However, there are several important points to consider. First, the precision numbers for spatial embeddings are averaged across many different configurations. If only the best possible precision value is considered for each query, the precision numbers are about 75-80% (20% higher than for the ranked list). Second, the threshold selection procedure significantly improves the result. Third, the actual numbers are not that different. For example, compare the first and the last columns with numbers. For each query we are considering 50 documents. There are 18 relevant documents on average. A difference of 6% means that the user will consider one extra non-relevant document before finding all the relevant ones. Last, in this experiment we assumed that only the top ranked relevant document is known. Table 3 shows that when

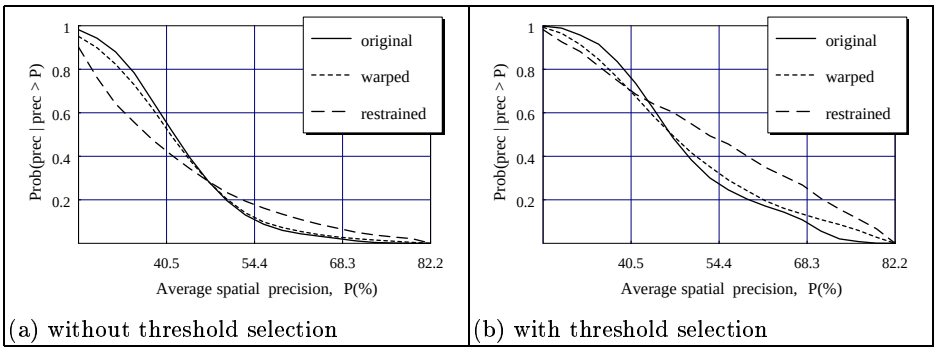

Fig. 2. Probability of selecting an embedding at random with a given precision value or higher for the full queries on TREC5 collection in 2 dimensions. It illustrates the effect of different user's feedback techniques. (a) No restrictions are imposed on the set of possible embeddings. (b) The set of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

all the relevant documents from the top 10 are known both ranked list and visualization perform equally – in 12 out of 24 points visualization outperforms ranked list.

6.3 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.

6.4 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 space and apply the restraining spheres. Then we add the information about the next relevant/non-relevant

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 τ measure. The relevance judgments for two documents are known to the system – the top ranked both relevant and non-relevant documents.

Queries		Rank List	t-D space	Embedding					
				w/o threshold selection			w/ threshold 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

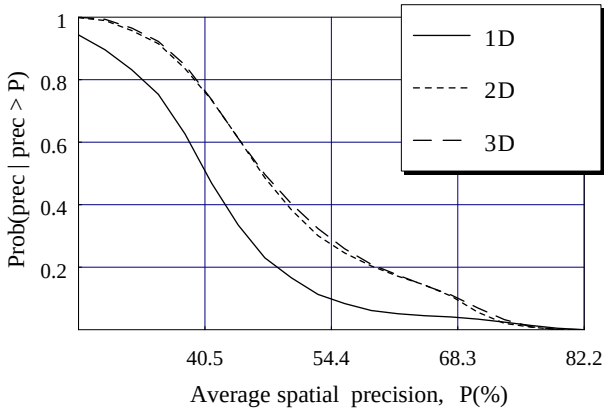

Fig. 3. Probability of selecting an embedding with a given precision value or higher for the full queries on TREC5 collection. The effect of different dimensions on the original embedding is illustrated. The set of embeddings is limited by the threshold selection procedure. The values on X-axis are averaged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

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.

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.

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

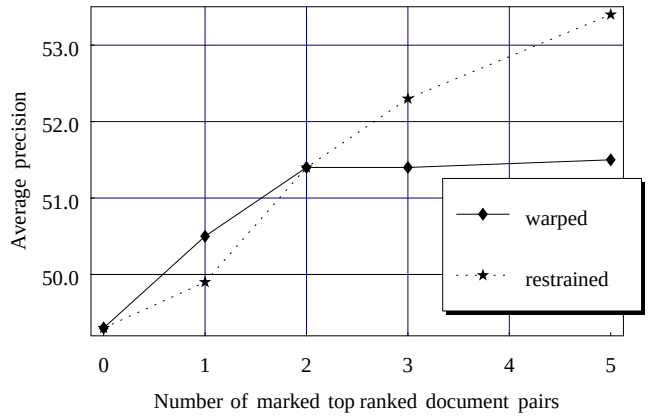

Fig. 4. Average precision computed starting from the first 5 relevant documents. The retrieved documents from TREC5/full queries are embedded in 2 dimensions.

6.5 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.

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

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%)

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

6.6 Best Case Analysis

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 and the system beats ranked list hands down. There are good embeddings out there – it is just difficult to find them.

7 Conclusion

In this paper we presented a non-interactive analysis of graphical interface for an information retrieval system that might be part of a digital library. Such analysis could help the researcher to isolate what part of the performance in a system-user pair is attributed to the visualization as compared to the user's skill. It could help the researcher to form an objective opinion of the system's abilities, generate clear expectations of the system performance in real life situations. We envision this approach as a companion to a user study – clearly stated hypotheses about the system actions should lead to a more accurate and potentially more productive user study.

- 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.

  However, we made an important assumption that a clear separation between relevant and non-relevant material will help the user to find the relevant material. This assumption is rather intuitive, but nevertheless a user study is necessary for its validation.

  We have also made some assumptions about the user's search strategy. An alternative strategies could result from observations of real users. Our analysis allows us to compare different strategies and select the best one for the visualization.
- 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.
- The "best case" analysis shows that the suggested visualization has a very high potential. It seems to be worthwhile to attempt a deeper investigation in how to make the threshold selection process more robust.

– The suggested visualization method in its current state (no robust threshold selection procedure) 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. That study also found no difference in precision between ranked list and 3-D visualization. We provide additional support that suggests visualization is neither better nor worse than a ranked list. In this study we showed that although the visualization does not help, it at least does not hinder the actual effectiveness of the system and it has much potential to be better.

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 situations when the user places the relevant documents into multiple classes. That task is modeled after the interactive TREC task of "aspect retrieval."

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

In this study we assumed that the user has already found some of the relevant documents (e.g., by means of the ranked list). We plan to look into the problem of helping the user to establish these first relevant documents. One way is to check the "clumpiest" areas of the visualization.

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 on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. 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 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. J. Allan, A. Leouski, and R. Swan. Interactive cluster visualization for information retrieval. Technical Report IR-116, CIIR, Department of Computer Science, University of Massachusetts, Amherst, 1996.
6. M. S. Barlett. The spectral analysis of two-dimensional point processes. *Biometrika*, 51:299–311, 1964.
7. 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.
8. M. Chalmers and P. Chitson. Bead: Explorations in information visualization. In *Proceedings of ACM SIGIR*, pages 330–337, June 1992.
9. N. A. C. Cressie. *Statistics for Spatial Data*. John Willey & Sons, 1993.
10. W. B. Croft. *Organising and Searching Large Files of Documents*. PhD thesis, University of Cambridge, October 1978.
11. D. Dubin. Document analysis for visualization. In *Proceedings of ACM SIGIR*, pages 199–204, July 1995.
12. T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software–Practice and Experience*, 21(11):1129–1164, 1991.
13. D. Harman and E. Voorhees, editors. *The Fifth Text REtrieval Conference (TREC-5)*. NIST, 1997.
14. 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.
15. 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.
16. 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.
17. B. D. Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13:255–266, 1976.
18. B. D. Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society*, 39:172–192, 1977.
19. D. Stoyan and H. Stoyan. *Fractals, Random Shapes and Point Fields*. John Willey & Sons, 1994.
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.