Visualizing Meta Search Results: Evaluating the MetaCrystal toolset

Anselm Spoerri

Department of Library and Information Science School of Communication, Information and Library Studies Rutgers University 4 Huntington Street, New Brunswick, NJ 08901 aspoerri@scils.rutgers.edu

MetaCrystal aims to help users find the documents that are most likely to be relevant. It compares the results returned by multiple search engines and displays the documents based on the number of engines that found them and their rank positions in the result lists. First, this paper examines the validity of the MetaCrystal design approach by calculating the probability that a document is relevant based on the number of engines that found it and its rank positions in the result lists. Using TREC data, it is shown that MetaCrystal's visual design is sound because it visually emphasizes documents based on their likelihood of being relevant. Second, this paper investigates if users are able to use the MetaCrystal's visualizations to find the documents that are most likely to be relevant. The results of a user study are reported that show that novice users are able to accomplish this task.

Introduction

Users searching for information are faced with the challenge of how to explore the many documents retrieved by a search engine. Users expect that the results are displayed in a way to make it easy for them to identify the documents that are most likely to be relevant. Commonly, search results are presented as a ranked list, which has the advantage that users know where to start their search for relevant documents. However, users have to move sequentially though the list and only a small subset of the documents is visible in a single screen. Many visual interfaces have been developed to increase the number of documents that users can explore in a single view (Hearst, 1999; Mann, 2002). These visualizations can be useful when users are not just looking for a few relevant documents, but need to find a greater number of relevant documents or want to gain insight into how a large number of documents are related to their search interest.

Meta search methods have been developed to combine the result sets of multiple search engines to increase the number of potentially relevant documents or to move such documents closer to the top of the fused result list (Callan, 2000; Fox & Shaw, 1994). The MetaCrystal toolset has been designed to enable users to visually compare the search results of multiple retrieval engines (Spoerri, 2004). Its tools provide a structured overview of the retrieved documents that reflects the number of engines that found the documents as well as their rank positions in the result lists being compared. This design approach was motivated by research that suggests that documents found by multiple retrieval methods are more likely to be relevant (Foltz & Dumais, 1992). The first goal of this paper is to examine the validity of the MetaCrystal design approach by calculating the probability that a document is relevant as a function of the number of engines that found it and its rank positions in the result lists. The second goal is to investigate if novice users can use MetaCrystal to find the documents that are most likely to be relevant.

This paper is organized as follows: section 2 reviews visual tools that can be used to visualize the results returned by multiple search engines. Section 3 provides a brief overview of MetaCrystal. Section 4 describes the methodology and data sets used to calculate the probability that a document is relevant based on the number of engines that found it and its rank positions. Section 5 describes the user study that has been conducted and provides an analysis of its results. Section 6 discusses lessons learned as well as future research.

Related Work

In this section, visualization approaches that can be used to display the relationships between multiple search results are briefly discussed. In Points of Interest (POIs)

visualizations (Benford et al., 1995; Hemmje et al., 1994; Olsen et al., 1993), the POIs can be used to represent the result sets of different search engines. The POIs act as magnets and the force of attraction is proportional to a document's rank position in the result list of the search engine that is represented by a POI. The ratio of the "forces of attraction" between a document and the POIs determines the document's location in the display. A key limitation of POI visualizations is that the distance from the display's center is not necessarily a reliable visual cue of a document's potential relevance. Further, if a documents is visualized as a simple point then it needs to be selected to determine how many engines retrieved the document.

Sparkler (Harve et al., 2001) combines a bull's eye layout with star plots, where a document is plotted on each star spoke based on its rankings by the search engines. A document is represented on multiple spokes if multiple engines retrieve it. Sparkler spreads the documents that have the same position on a spoke, and thus would overlap, to show their distribution pattern. It visually indicates which documents are more likely to be relevant based on the results of a specific engine by placing them closer toward the center of the display. However, Sparkler does not explicitly represent which documents have been retrieved by multiple engines and by which particular combination. Users need to examine the individual document icons to be able to determine how many and which retrieval methods found them. As will be shown, this information is useful to infer the possible relevance of a document.

Beadplots (Banks & al., 1999) aims to visualize the shared subpatterns in the ranked lists returned by different systems that participate in the Text REtrieval Conferences (TREC) experiments. The rows in a beadplot correspond to the different systems, and the "beads", gray and colored diamonds, along each row represent the documents. The position of a bead along a row indicates its position or rank in the result list of the system associated with the row. Like Sparkler, Beadplots does not explicitly encode the number of systems that retrieve the same document. Beads with the same color in the different rows indicate the same document, enabling users to spot documents retrieved together as a group, which show up as splotches of the same color, at (possibly) different positions along the rows. A spectral color ordering is used to assign colors to the documents. The ordering ranges from most relevant (dark red) to least relevant (light violet), where relevance ordering is based on the top 100 documents found by the University of Waterloo's system or the top 100 composite ranking based on the retrievals from all of the systems that participated in the TREC experiments.

If the results returned by multiple search engines are combined then a new ranked list can be created that can be visualized using a spiral representation. VisDB (Keim & Kriegel, 1994) uses every pixel to represent multi-dimensional data and it places the ordered

items along a space-filling spiral. NIRVE (Cugini et al., 1996) contains a "Document Spiral" tool, which places the highest ranked document in the center. Subsequent document icons are placed and spaced along the spiral proportional to their relative score. The Document Spiral can provide users with an insight into distribution patterns of the relevance scores. However, the placement of documents icons can lead to spurious perceptual groupings and the display space is not optimally used. Similar to the Document Spiral, Torres et al. (Torres et al., 2003) use a spiral layout to display items based on their similarity with the query and the size an item is proportional to its similarity score, which creates a static "focus + context" effect.

Several meta search engines visually organize the retrieved documents. Vivísimo (http://www.vivisimo.com) uses the familiar hierarchical folders metaphor and indicates how many documents are inside each folder. After each document summary, the search engines are listed that retrieved the document, together with the ranking by each engine. Grokker (http://www.groxis.com) uses nested circles or rectangles to visualize a hierarchical grouping of the search results. MetaSpider (Chen et al., 2001) uses a self-organizing 2-D map approach to classify and cluster the retrieved documents. These clustering interfaces, however, do not provide users with visual cues about which specific documents are most likely to be relevant and which folder or cluster may contain them. Instead, the textual descriptions of the clusters are supposed to help users decide which folder or cluster to select for further exploration. Kartoo (http://www.kartoo.com) creates a 2-D map of the highest ranked documents and uses different icons to indicate whether a found web page is a homepage or has related pages from the same site. The icon size provides a weak visual cue about the potential relevance of a webpage. The spatial layout created by Kartoo does not map the documents most likely to be relevant to a specific area, making it difficult for users to decipher the visual map.

MetaCrystal

This section provides a brief overview of the MetaCrystal toolset and focuses on the features that are important for this paper (more detailed information can be found in Spoerri (2004a, 2004b, 2004c)). MetaCrystal consists of several linked tools that enable users to explore the search results returned by multiple search engines. All its tools employ a "bull's eye" layout to guide users toward potentially relevant documents. The Category View aggregates the documents that are found by the same combination of search engines and displays the number of documents retrieved by specific engine combinations (Spoerri 2004a). The Cluster Bulls-Eye enables users to see how all the found documents are related to the different search methods being compared (Figure 1). The RankSpiral places all the documents sequentially along an expanding spiral to enable

users to rapidly scan a large numbers of documents (Figure 1).

The Cluster Bulls-Eye and RankSpiral will be described in more detail because they compared in the user study. Both tools use the following design principles to display the retrieved documents. First, the documents found by the same number of engines are mapped into the same concentric ring; the number of engines increases toward the center of the display. Second, a document is placed within a ring to reflect the average of its positions in the ranked lists that contain it; documents with a high average rank position are placed close to the ring's edge that is closer to center of the overall display. Third, a document is represented by an icon; the shape indicates the number of engines that retrieved the document, color coding is used to represent the specific engines that found the document and the color's intensity reflects the document's rank positions. Fourth, size coding is used to reflect a document's probability of being relevant. This probability is a function of the number of engines that found the document and the average of its rank positions. The next section will address how this probability can be estimated if the retrieval engines search the same database.

Both the Cluster Bulls-Eye and RankSpiral use related radial mappings. The Cluster Bulls-Eye uses polar coordinates to display the document icons and enhances a traditional POI display: the radius value is related to a document's total ranking score; the angle reflects the relative ratio of a document's rankings by the different engines. Specifically, the latter is equal to the angle of the vector that is the sum of the position vectors the POIs (the star-shaped icons at the periphery of the display), where each vector is scaled by the rank position for each related search engine. The total ranking score of a document is calculated by adding the number of engines that retrieved it and the average of its different rankings. As mentioned, this causes documents found by the same number of engines to be placed in the same concentric ring; documents with high rankings by the different engines cluster close to the ring's edge closest to the center of the overall display and their icon's size is set to the largest possible value for the ring in question. Size coding creates a visual hierarchy within each concentric ring, making it easier for users to identify the top documents found by a specific number of search engines. The RankSpiral places documents sequentially along an expanding spiral so that their distance from the display center is equal to the total ranking score. The angle is computed so that consecutive documents are placed adjacent to each other so that they do not overlap (2004b). The use of size coding makes it possible to place more document icons on the spiral.

The user study presented in this paper aims to address how well novices can use the provided visual cues to find the documents that are most likely to be relevant. MetaCrystal has been implemented in Flash using ActionScript to make it accessible via a Web browser. The displays used in the user study can be accessed at

http://www.scils.rutgers.edu/~aspoerri/study/UserStudy.swf .

Testing the Validity of MetaCrystal's Design

MetaCrystal's design was motivated by research that suggests that documents found by multiple retrieval systems are more likely to be relevant (Foltz & Dumais, 1992; Saracevic & Kantor, 1988). This assumption has also guided the design of data fusion methods that combine the result sets of different systems to create an improved ranked list (Callan, 2000; Foltz & Dumais, 1992). A recent study addressed the validity of this assumption by analyzing the overlap between the search results of retrieval systems that participated in the Text REtrieval Conferences (TREC) (Spoerri, 2005). It showed that the potential relevance of a document increases exponentially as the number of systems finding it increases - called the Authority Effect. It also showed that documents higher up in ranked lists and found by more systems are more likely to be relevant - called the Ranking Effect.

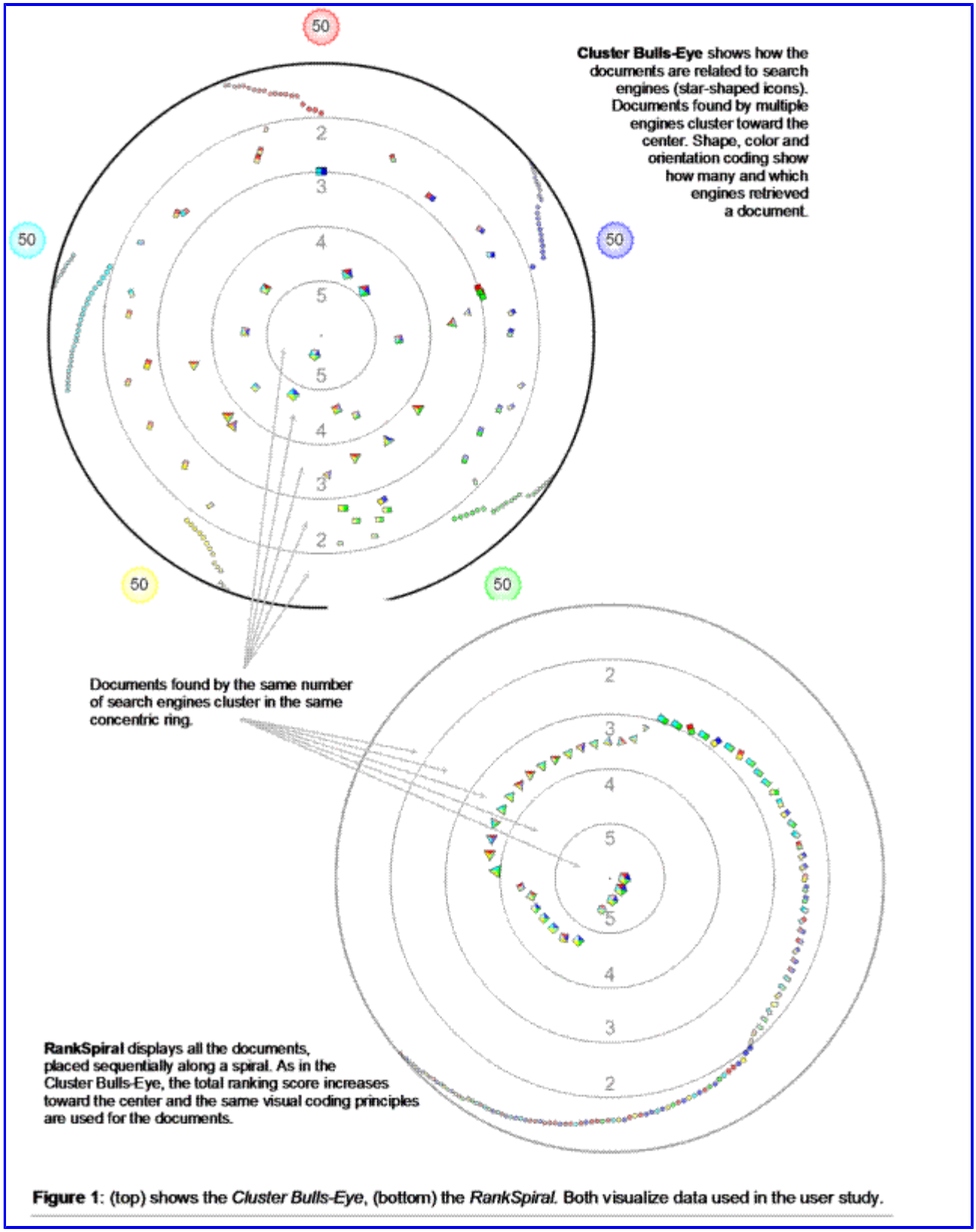

Figure 1: (top) shows the Cluster Bulls-Eye, (bottom) the RankSpiral. Both visualize data used in the user study.

This paper extends this study by estimating the probability of relevance if the results of

only five systems are compared at a time and the systems are randomly selected. When comparing the results returned by five different Internet search engines (Spoerri 2004a), Figure 2 shows how MetaCrystal computes a document's likelihood of being relevant based on the number of engines that found it and its average rank position. The plot resembles a "saw tooth" pattern; it shows that a document, found by three engines and top ranked in all three lists has the same probability of being relevant as a document, which is found by five engines and is top ranked in all five lists. Internet search engines tend to index less than 20% of the Internet, causing the databases they search to only partially overlap (Lawrence & Giles, 1999). This makes it difficult to determine the "correct" probability of relevance when comparing the results of Internet search engines. For this reason, the result sets of systems that participated in TREC are analyzed in this paper, because these systems search the same database and relevance judgments are available.

Methodology

TREC provides information retrieval researchers with large document collections, a set of search topics and ways to compare the search results (Voorhees & Harman, 1999). Retrieval systems participating in the "ad hoc" track search the collections for each of the 50 provided topics, and then submit a ranked list of usually 1,000 documents for evaluation. For each topic, the top 100 retrieved documents are pooled from each system. The evaluator who proposed a topic then determines the relevance of each document. Data fusion researchers commonly use data from the "ad hoc" track, because relevance judgments and result lists by many and diverse retrieval systems are readily available.

In this paper, data from TREC 8 is used to test the validity of MetaCrystal design approach, which is as follows: 1) sort the documents based the number of engines that found them (and display them in concentric rings); 2) sort the documents found by the same number of engines based the average of their rank positions; 3) scale the size of a document icon based on the probability that the document is relevant. In TREC 8, 35 research groups submitted result sets for the 50 topics, using queries that were constructed automatically and that only searched the title and description fields of the document record (Voorhees & Harman, 1999). In this paper, the top 50 results are used in the presented analysis, because users tend to interact with at most 50 to 100 result documents when searching for information.

A goal of this paper is to be able to estimate the probability that a document is relevant based on the number of engines that found it and its rank positions in the result lists. First, a sufficient number of random groupings of five systems need to be compared. Each

system needs to be selected an equal number of times. The randomization process can be constrained so that each of the 35 TREC 8 systems appears exactly five times in 35 random groupings of five systems. Second, the overlap structure between the results of the five randomly selected systems is computed for each of the 50 topics. Specifically, the number of (relevant) documents that are found by 1, 2, 3, 4 and 5 systems is calculated. Third, a document that is found by multiple systems will have multiple rank positions that can be averaged. If the average rank position for the top 50 documents is subdivided into 20 buckets, then a bucket size of 2.5 consecutive average rank positions is used to construct a frequency plot of the number of (relevant) documents found by 1, 2, 3, 4 and 5 systems. Fourth, the percentage of documents that are relevant can be computed for each bucket by dividing the number of relevant documents by the number of documents in a bucket. These percentages are then averaged over all 50 topics. Finally, the resulting percentage is averaged over the 35 groupings for each bucket.

Results

Figure 2 shows the percentage of documents that are relevant as a function of the number of systems that found them and their average rank position, when the percentages for the 35 random groupings of five systems are averaged. The issue arises of how to compute the average if a bucket contains no documents and thus the percentage of documents that are relevant cannot be "defined". The dashed line in Figure 2 shows the resulting average if only "defined" percentages are averaged for the 35 groupings. The green line shows the resulting average if the sum of "defined" percentages is divided by 35, regardless of the actual number of "defined" percentages. Comparing these two lines shows that the number of "undefined" percentages increases significantly for documents that have a low average rank position. There are very few "defined" percentages for such documents and they are not representative. Thus, the "filtered" average is computed as follows: for documents with an average rank position of less than 25, the average is computed by dividing by 35; otherwise, only the "defined" percentages are averaged.

Figure 2 shows that the probability that a document is relevant follows a "saw tooth" pattern, where the maximal height of a saw tooth decreases as the number of systems that found the document decreases. Figure 2 suggests that the document's probability of being relevant can be approximated by a piece-wise linear function of the number of systems that found it and its average rank position. The greater the number of systems that find a document and the greater its average rank position, the greater its probability of being relevant. This supports the Meta-Crystal's design approach, which first visually sorts the documents based on the number of systems that found them, and then sorts the

documents based on their average rank positions.

User Study

The hallmark of an effective text retrieval interface is that it guides users toward potentially relevant documents. This section presents the results of a user study, whose goals are: 1) test how well users, who have received only a short introduction and no training, are able to identify the ten documents that are most likely to be relevant; 2) test if there is a statistically significant performance difference between the Cluster Bulls-Eye and RankSpiral in terms of effectiveness and/or efficiency; 3) test how much the overall distance of a document from the display center will interfere with the size coding used to directly encode its probability of being relevant.

Motivation

The Cluster Bulls-Eye and RankSpiral use icons to represent the documents so that all of the found documents can be shown in a single display. Users need to use their visual reasoning skills to identify icons that represent documents that are most likely to be relevant. In particular, users need to decide over which document icons to place their mouse to receive "details on demand" or which icons to select to view the complete document to determine the document's relevance. A "details-on-demand" display tends to show the document title and a text snippet. Text provides a strong cue for users to decide if a document is relevant. In this user study, this rich source of information is not provided so that subjects do not have to interpret textual information to determine the relevance of a document. Instead, subjects can only use visual information to decide which icons represent documents that are most likely to be relevant. Once subjects have made their selections and submitted them, they receive visual feedback about the location of the documents that are most likely to be relevant ("top 10"). The feedback consists of surrounding the top ten documents with a "green halo." If a subject selects a top 10 document then its icon has a thickened black border surrounded by a green halo. The feedback intends to simulate the inferences subjects would be able to make using the information presented in a "details-on-demand" display and which they can use to learn how to find relevant documents.

The goal of the user study is to test how well users can use the visual cues, such as the icon's shape and position, to decide which icons to explore first to find highly relevant documents. This ability is a prerequisite so that users can make effective use of MetaCrystal's tools. This user study does not show that users can effectively use the full

functionality of MetaCrystal. It addresses whether a key prerequisite is satisfied.

Hypotheses

The task for the subjects is to select the ten documents that are most likely to be relevant. The Cluster Bulls-Eye emphasizes how the documents are related to the engines, causing the documents to be "scattered" in their respective concentric rings based on the "forces of attraction" of the engines that found them. The RankSpiral highlights the ranking of the documents based on the number of engines that found them and their average rank position.

Hypothesis 1: novice users who received no formal training should be able to perform the task. Specifically, the error subjects make should be minimal for the top 5 documents and increase as they have to select the top 6 to top 10 documents. The RankSpiral should produce smaller errors than the Cluster Bulls Eye.

Hypothesis 2: subjects should perform the task in less time using the RankSpiral, because the documents icons are "visually sorted" and placed closer together than in the former. This requires fewer eye movements on the part of the user in the RankSpiral. A user can start in the center of display and follow the spiral. However, the user has to "visually sort" the icons based on their size and needs to skip certain icons, because they are not as large as others further up or down the spiral, instead of just following the spiral. In the Cluster Bulls Eye, subjects have to scan from the center out and need to explore a larger visual area. This requires a more complex visual search strategy than in the RankSpiral. In terms of accuracy, the RankSpiral should outperform the Cluster Bulls-Eye and help users select documents that are more likely to be relevant.

Hypothesis 3: the overall distance of a document icon from the center of the display will interfere with the size coding that is used to directly encode the document's probability of being relevant. An icon's distance from the ring border closest to the center of the display also reflects a document potential relevance. The closer to this border, the higher the potential relevance; but the number of engines that found the document also needs to be taken into account, as Figure 2 illustrates.

Data Used

As in section 4, TREC 8 data is used to create 10 different data sets so that the subjects can interact with the actual search results: for each of the five different TREC systems the top 50 documents are used and compared. The blue line in Figure 2 shows how a

document icon's size is set in the displays presented in the user study. This curve is almost identical to results shown in Figure 2, except for the documents found by one or two systems, where the effect of a declining average rank position is amplified to create a visual hierarchy within the concentric rings for the documents found by one or two systems.

Experimental Design

Nine undergraduate students participated in the user study. The subjects received short instructions, but they had no training prior to being presented the ten different data sets. Each subject viewed the same data set once in both the Cluster Bulls Eye and RankSpiral displays. The order of presentation was randomized in terms of both the data sets and display type, ensuring that the same data set or display type was not presented consecutively. After a subject had selected ten documents, the subject submitted his or her selection and received visual feedback about the correct top 10 documents. However, a subject was not able to change the made selection while or after viewing the feedback. When ready, the subject requested to have the next data set displayed. At the end of the experiment, the subjects were asked to answer a few questions in writing.

The subjects first read these instructions, shown on two consecutive slides (which could not be viewed again once the experiment started): "Five retrieval engines have searched the same database for multiple topics. Each engine returns a ranked list of 50 documents for each topic. The result lists returned by the different engines are compared and the likely relevance of the documents is determined. The likely relevance of a document is affected by two rules: a) the more engines that find the same document, the greater the document's probability of being relevant; b) the higher up a document is placed in the multiple result lists, the greater its probability of being relevant. Your task is to use two different visual displays to select the 10 documents that are most likely to be relevant. A document is represented by a visual icon: a) the shape indicates how many engines retrieved the document; b) the size indicates the document's probability of being relevant: the greater its size, the higher up the document is placed in the multiple result lists. Documents found by the same number of engines are placed in the same circular ring: the position within a circular ring indicates the document's probability of being relevant: the closer towards the center, the greater the probability. Please select the 10 documents as quickly as possible. You can change the documents selected. You will receive feedback when you have selected 10 documents and you request feedback."

For several of the data sets used in the user study, some of the top documents overlap, making it difficult to select all of them, because a selected document can completely

occlude the document icons in close proximity. Subjects received the verbal instruction to first select an icon that lies underneath other icons before selecting an icon that almost completely occludes the former icon. Subjects also received a one-minute demo of the "magnify tool", which lets users "click & drag" the data display to change its scale. Subjects were told that if the data display is magnified so that the control panel is covered, then its size needs to be reduced so that the pointer tool can be selected.

Results

As mentioned, each subject was presented the same data once in both the Cluster Bulls-Eye and RankSpiral. This makes it possible to compute relative difference in performance between the two displays for each subject. This helps to reduce noise and unwanted variability in the collected data. Further, the ten different data sets and two display types were presented in a randomized order to minimize learning effects. Hence, the paired-differences for each subject can be computed and used to make a statistical inference. The performance is measured as follows: 1) a "relevance score" is computed, which is equal to the average of the relevance probabilities for the ten selected documents; and 2) the time it takes to select ten documents is measured. The paired-difference T-test is used to infer if there is statistically significant difference between the Cluster Bulls Eye and RankSpiral displays. The one-sided T-test is used to test if any observed superior performance is due to chance. There were 9 subjects and the degree of freedom of the T-distribution is 9 - 1 = 8.

Hypothesis 1: "Novices can perform the task." The ten documents selected by a subject can be sorted in descending order based the relevance probability associated with the documents. This has the effect of disregarding the order in which the documents were selected. This newly ordered list can be compared with the list of "top 10" documents. The percentage error in relevance between nth document in the subject's and top 10 lists can be computed. Figure 3 shows that the error is minimal for the top 7 documents and increases rapidly after the top 7 documents for both displays. This shows that novice users can perform the task of finding the documents that are most likely to be relevant without prior training. Figure 3 suggests that novice users can use the Cluster Bulls-Eye and RankSpiral displays to select highly relevant documents, especially the top 7 documents.

Hypothesis 2: "RankSpiral outperforms Cluster Bulls-Eye." Eight of the nine subjects performed the task faster using the RankSpiral; the average time difference between the RankSpiral and the Cluster Bulls-Eye was 7.89 seconds. The one-sided T-test value is 0.033, which is significant at the 0.05 level. Seven out of the nine subjects performed the task more effectively using the RankSpiral; the average "relevance score" difference is

0.034. The one-sided T-test value is 0.037, which is significant at the 0.05 level. The "relevance scores" for the RankSpiral and Cluster Bulls-Eye are 5.018 and 4.984, respectively. These scores are 3.5% and 4.1% less than the perfect "relevance score", which suggests that novice users are able to select highly relevant documents using either display.

Hypothesis 3: "Distance from center dominant cue." A hypothesis of the study is that the overall distance from the display center will interfere with the task the subjects have to perform. In order to address this issue, all the document icons can be ranked based on their probability of being relevant (encoded by the icon's size) or based on their overall distance from the display center. A "size rank" and a "distance rank" can be assigned to a document that was selected by a subject. For example, if a subject selected three documents in the following order and with these respective "size" and "distance" ranks [(1, 3), (2, 5), (4, 1)], then the absolute difference between it and the "perfect list" [(1, 1), (2, 2), (3, 3)] is equal to [(0, 2), (0, 3), (1, 2)]. Such "difference vectors" can be computed for each of the data sets and for both display types. Further, these difference vectors can be computed if the order in which documents are selected matters (although the subjects were not instructed to select the documents in a specific order), or the documents are sorted either based on the "size" or "distance" ranks (as was done for the "size rank" when addressing hypothesis 1). Finally, the list of documents selected by the subjects can be compared with the "perfect list" by computing the difference in terms of the rank values or the relevance probability.

If a subject (sequentially) selects all the top ten documents based on the "size rank" then the differences with respect to the ten most relevant documents would be all zero, both for the rank values and the probability of relevance. However, if the "distance rank" differences are equal or close to zero, this suggests that the subjects use the distance from the display center as the primary visual strategy when deciding in which order to select the document icons. Figure 3 indicates that subjects tend to use the distance form the center as the primary visual cue when deciding which icons to select and in which order.

A further experiment was conducted to directly test how much the overall distance of a document from the display center interferes with the size coding used to directly encode its probability of being relevant. The layout algorithms for RankSpiral and the Cluster Bulls-Eye displays were modified so that a document's size and its distance from the center both directly encode the document's probability of being relevant (see Figure 5).

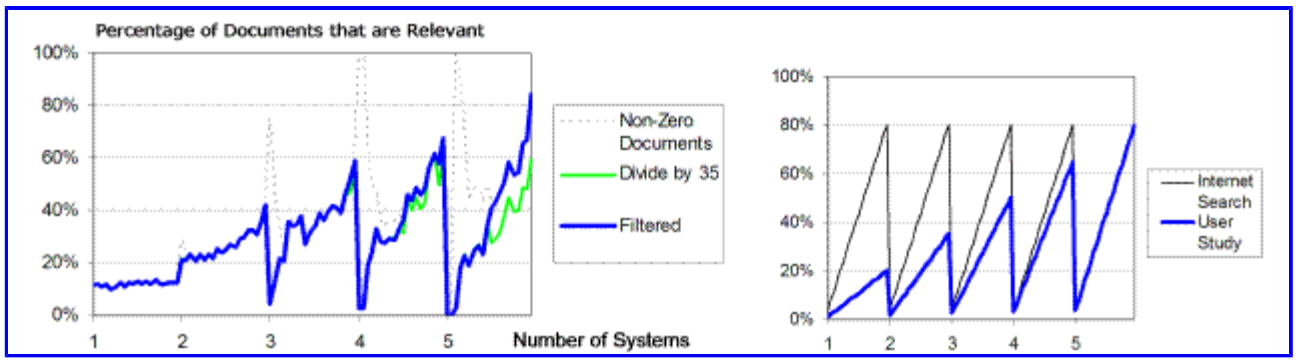

Figure 2: shows the percentage of documents that are relevant as a function of the number of systems that found them and the documents' average rank position, which increases from left to right for each number of systems; TREC 8 data was used to compute these percentages. (Right) shows how the size of a document is specified for Internet searches and for the user study, respectively.

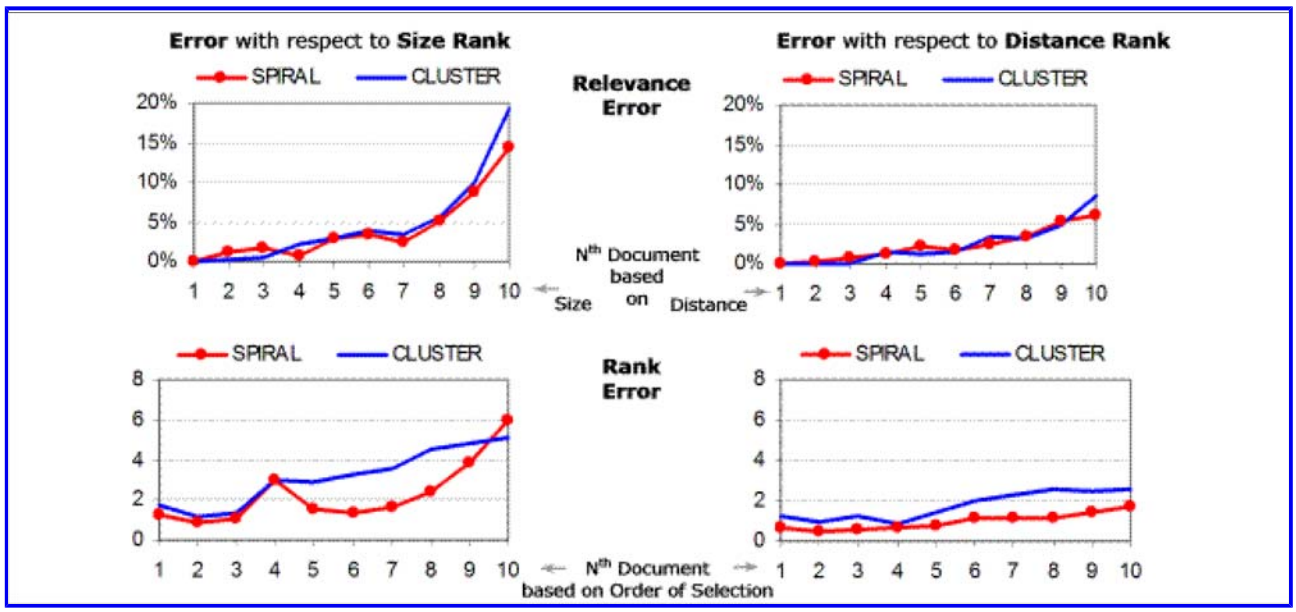

Figure 3: shows the relevance error if the ten documents selected by the subjects are first sorted based on their size or distance from the center and then are compared with the "top 10" documents based on size or distance, respectively; (bottom row) displays the error in rank position if the order of selection matters.

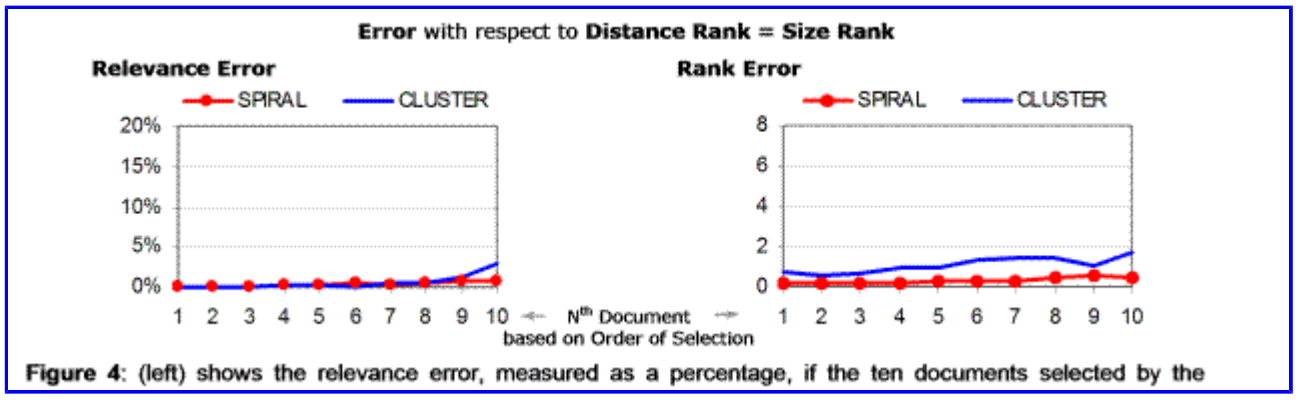

Figure 4: (left) shows the relevance error, measured as a percentage, if the ten documents selected by the

Figure 4: shows the relevance error, measured as a percentage, if the ten documents

selected by the subjects are first sorted based on their distance from the center (and a document's size also decreases away from the center, see Figure 5) and then are compared with the "top 10" documents based on distance; displays the error in rank position if the order of selection matters.

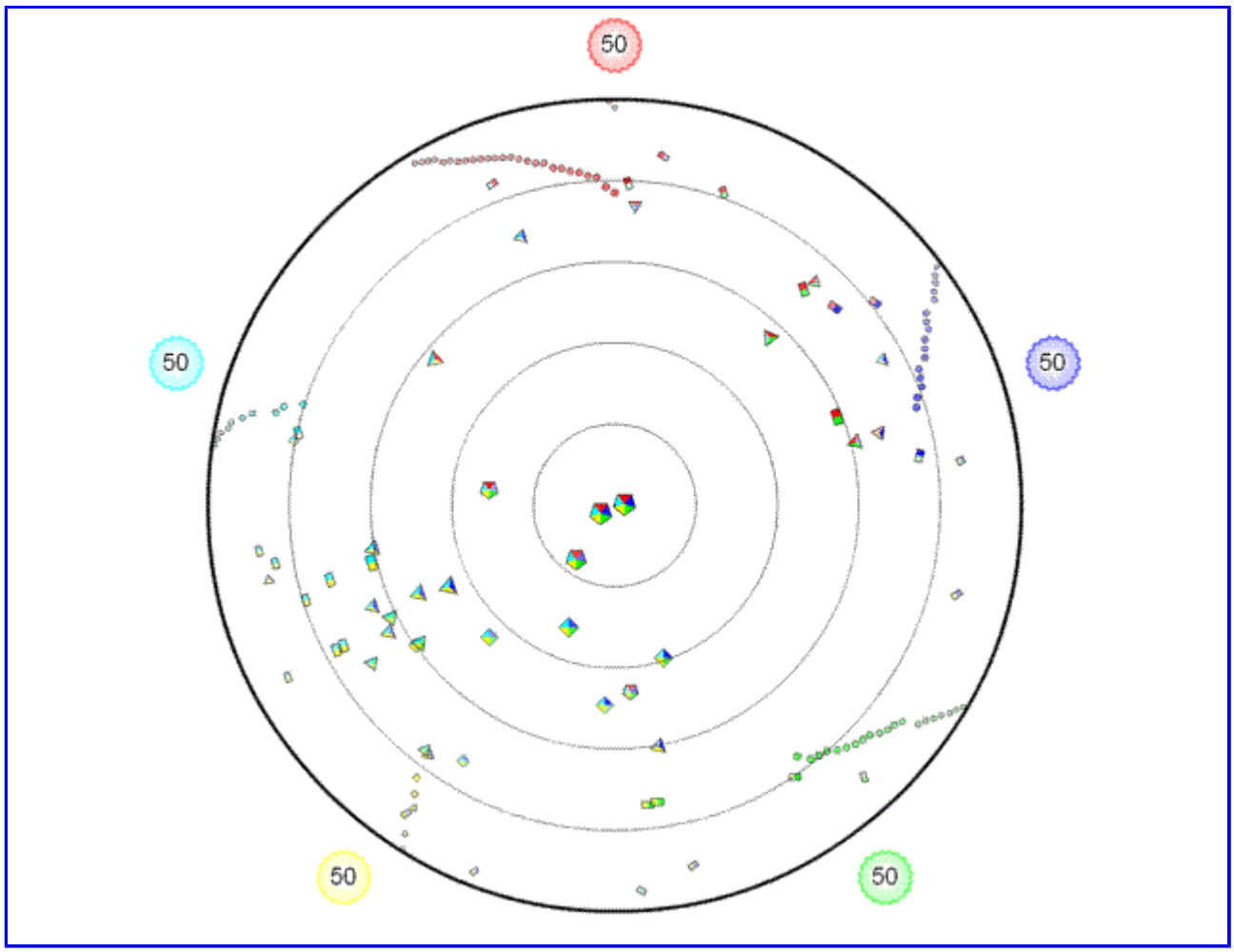

Figure 5: shows the Cluster Bulls-Eye display, when both a document's size and its distance from the center directly encode the document's probability of being relevant.

Figure 4 shows that if such layout algorithm is used, then users make much fewer errors when attempting to identify the top 10 relevant documents. In particular, it shows: a) the "relevance error" is almost zero and there is no significant difference between the Rank Spiral and Bulls-Eye displays; b) the "rank error" is also almost zero and subjects using the RankSpiral made fewer errors than using the Cluster Bulls-Eye display.

User Feedback

After completing the experiment, the subjects were asked to provide written answers to

these questions: Do you think you learned over time how to use the visual tools? What was easy and clear? What was difficult or hard to understand? Some of the key feedback received was:

Learning: “If the icon is large and close to the center, then select it.” “Learned to detect the difference in size and distance from the center. “Color and size difference play an important role and became easy to learn after a while.”

Clear / Easy: “The spiral view was easier; icons are closer together.” “The icons in the center are easiest to spot.” “It is easy to pick the top 5 documents.” “The shape of icons made it easier to figure out how many engines found the document.”

Difficult / Hard: “Could not understand why some choices were wrong: is it size or position or shape that counts most?” “In the cluster view I needed to look all over the display.” “The meaning of the colors was not clear: some are more bold and appear larger than they really were.” “Magnifying the display so that the tool icon became not accessible anymore.” “It is hard to select icons that overlap or are too close together.”

Discussion and Future Work

The analysis of the user studies showed that users tended to use the overall distance from the display center as a primary visual cue to decide which icons to select. As Figure 2 shows this is a reasonable strategy, especially since icons tend be larger towards the center and the size differences between them are not very large. If the subjects had selected the ten documents solely on the basis of the overall distance from the display center, then they would have achieved a relevance score of 5.03 which is slightly better the subject’s results when using the RankSpiral. As Figure 2 shows, there is an interaction between the number of engines that find a document and the average rank position. It could be interesting to create “artificial” displays, where this interaction would be much less and the penalty for using the distance from display center is much greater.

The results from the user study raise the question if it would be advisable to relax some of MetaCrystal’s design principles, such as mapping documents found by the same number of engines into the same concentric ring. Users could be given the option of having the distance from the display’s center and the icon size encode the likelihood that a document is relevant. As Figures 4 and 5 show, users are able to detect the top 10 documents more accurately in both displays and the RankSpiral outperforms the Cluster Bulls Eye, since the former will be equivalent to a ranked list mapped onto a spiral. When comparing Internet search results, the concentric rings are of value, because it is much harder to estimate a document’s probability of being relevant. Future research and user studies will investigate

how MetaCrystal can best support users searching the Internet.

The subjects had to compare the size of icons that could be a circle, rectangle, triangle, square or pentagon. The size coding that is currently employed is not perfectly calibrated so that two icons with different shapes but same probabilities appear the same size. Furthermore, a document icon can have multiple colors. Currently, the intensity of a color reflects a document's position in the ranked list of the search engine associated with the color. If a document has a low rank position for all the engines that found it, then it will appear dimmed and this visual cue is consistent with its small icon size. However, if the document has different rank positions, then the different intensities cues can interfere with the size coding (as mentioned in the user feedback). More work is needed to better calibrate the size coding methods.

The written feedback received in the user study was insightful and helpful. Some key issues were identified, such as how to make icons easier to compare, use color more effectively and how to explain the MetaCrystal's key principles more clearly and succinctly. Finally, further user studies are in preparation that will focus on how users can use MetaCrystal to find documents that they deem relevant; compare MetaCrystal's displays to a standard ranked list; and investigate how to better help users gain new insights when searching for information.

Conclusion

Using TREC 8 data, this paper examined the validity of the MetaCrystal design approach. The overlap structure between the top 50 search results of 35 random groupings of five retrieval systems was computed for 50 topics. The presented analysis showed that a document's probability of being relevant can be approximated by a piece-wise linear function of the number of engines that found it and its rank positions in the result lists. This provides strong empirical support for MetaCrystal's design approach of visually emphasizing documents based on the number of engines that found them and their rank positions. Second, this paper reported the results of a user study, whose goal was to determine if novice users can use two of MetaCrystal's tools to find the documents that are most likely to be relevant. Specifically, it was shown that users can use the provided visual cues, such as the icon's shape and position, to decide which icons to explore first to find highly relevant documents. This ability is a prerequisite for users being able to make effective use of MetaCrystal's full functionality. The user study also showed that the subjects could identify highly relevant documents more rapidly and accurately using the RankSprial than the Cluster Bulls-Eye. The user study helped to bring into better focus how the multiple visual coding schemes employed by MetaCrystal can be better utilized and

optimized.

ACKNOWLEDGMENTS

The author would like to thank the participants of the user study as well as Nick Belkin for his help. This research has been supported by a Rutgers Research Council Grant. The TREC data used in the research reported in this paper has been provided by NIST and can be downloaded at http://www.nist.org .

References

Banks D., Over P. & Zhang N. (1999) Blind Men and Elephants: Six Approaches to TREC data *Information Retrieval* 1, 7-34

Benford, S. D., Snowdon, D N., Greenhalgh, C M., Ingram, R J., Knox, I. & Brown, C C., (1995) VR-VIBE: A Virtual Environment for Co-operative Information Retrieval *Computer Graphics Forum* 14, (3), pp. 349-360

Callan. J. (2000) Distributed information retrieval In Croft W.B. (Ed.) *Advances in Information Retrieval* Kluwer Academic Publishers. pp. 127-150

Chen H., Fan H., Chau M. and Zeng D. MetaSpider: (2001) Meta-Searching and Categorization on the Web *JASIST* 52 (13), 1134 - 1147

Cugini, J., Piatko, C. & Laskowski, S. (1996) Interactive 3D Visualization for Document Retrieval *Proc. of Workshop on New Paradigms in Information Visualization & Manipulation, CIKM '96*

Foltz, P. and Dumais, S. (1992) Personalized information delivery: An analysis of information-filtering methods *Comm. of the ACM* 35 (12):51-60

Fox, E. & Shaw, J. (1994) Combination of Multiple Searches *2nd Annual Text Retrieval Conference (TREC-2), Gaithersburg, MD, 2000* U.S. Government Printing Office, Washington

Grokker http://www.groxis.com

Havre, S., Hetzler, E., Perrine K., Jurrus E., & Miller N. (2001) Interactive Visualization of Multiple Query Results *Proc. IEEE Information Visualization Symposium. 2001*

Hearst M. (1999) User interfaces and visualization *Modern Information Retrieval* R. Baeza-Yates and B. Ribeiro-Neto (eds.). Addison-Wesley, 257-323

Hemmje, M., C. Kunkel, & A. Willet (1994) LyberWorld - a visualization user interface

supporting fulltext retrieval *Proceedings of ACM SIGIR* pp. 254-259

Kartoo http://www.kartoo.com

Keim, D. A.& Kriegel, H.P. (1994) VisDB: Database Exploration Using Multidimensional Visualization *IEEE CG&A, Sept. 1994* pp. 40-49

Lawrence, S., & Giles, C.L. (1999) Accessibility of information on the Web *Nature* 400, 107-109

Mann T. (2002) *Visualization of Search Results from the WWW* Ph.D. Thesis, University of Konstanz

Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., & Williams, J. G. (1993) Visualization of a Document Collection: the VIBE System *Information Processing & Management* 29 (1), 69-81

Saracevic, T. and Kantor, P. (1988) A study of information seeking and retrieving. III. Searchers, searches and overlap *JASIST* Volume 39, 3, 197-216

Spoerri, A. (2004a) Visual Editor for Composing Meta Searches Proc. of the 67th Annual Meeting of the American Society for Information Science and Technology (ASIST 2004)

Spoerri, A. (2004b) RankSpiral: Toward Enhancing Search Result Visualizations *Proc. IEEE Information Visualization Symp. 2004*

Spoerri, A. (2004c) Toward Enabling Users to Visually Evaluate the Effectiveness of Different Queries or Engines *Journal of Web Engineering* 3(3 & 4) 297-313

Spoerri, A. (2005) How the Overlap Between the Search Results of Different Retrieval Systems Correlates with Document Relevance *Proc. of the 68th Annual Meeting of the American Society for Information Science and Technology (ASIST 2005)*

Torres, R., Silva, C., Medeiros, C. &. Rocha, H. (2003) Visual Structures for Image Browsing *CIKM'03, November 3-8, 2003, New Orleans, Louisiana*

Vivisimo http://www.vivisimo.com

Voorhees, E. & Harman, D. (1999) Overview of the Eighth Text REtrieval Conference (TREC-8) *The Eighth Text REtrieval Conference (TREC-8), Gaithersburg, MD* 2000. U.S. Government Printing Office, Washington