Journal of Web Engineering, Vol. 3, No.3&4 (2004) 297-313


# TOWARD ENABLING USERS TO VISUALLY EVALUATE THE EFFECTIVENESS OF DIFFERENT SEARCH METHODS

ANSELM SPOERRI

*Rutgers University, New Brunswick*
*aspoerri@scils.rutgers.edu*

Received July 18, 2004
Revised December 8, 2004

This paper explores how information visualization can provide insights into the effectiveness of different query formulations or the same query submitted to multiple search engines. Different queries or search methods are deemed more effective if the fusion of their results leads to a new result set that contains an increased number of relevant documents. The MetaCrystal toolset can be used to visualize the degree of overlap and similarity between the results returned by different queries or engines. The work presented is guided by two working hypotheses: 1) documents found by multiple methods are more likely to be relevant; 2) high degrees of overlap and/or systematic relationships between the ranked lists being compared will not lead to fusion results that contain more relevant documents. MetaCrystal visually identifies documents found by multiple queries or engines. The number and distribution patterns of documents found by multiple methods can be used as a visual measure of the fusion effectiveness. Examples, using Internet and TREC data, are presented that support both in a qualitative and quantitative way the working hypotheses.

*Key words*: Web search, evaluation, user satisfaction
*Communicated by*: A Spink & C Watters

## 1. Introduction

When searching the Internet, users are often confronted with an overwhelming number of potentially relevant documents to sift through. It is difficult for users to determine the overall effectiveness of their search. Researchers tend to use measures, such as precision and recall, to capture the effectiveness of an information retrieval system [13]. Searching the Web is a highly interactive process and additional measures are needed to capture the richness of user interactions and their experiences of ease of use. Spink and Wilson [16] have proposed that some of these new measures need to reflect how users progress through the different stages of their information seeking process and how their interactions change their understanding of their information problem. Spink [15] employed such a user-based approach to evaluate the meta search interface Inquirus. This study showed that users experienced some change in their information problem and information seeking stage. The results also showed that search precision did not necessarily correlate with the user-based evaluation measures or change in the search process. Another study [23] evaluated the performance of four search engines by using these user-centered criteria: relevance, efficiency, utility, user satisfaction, and connectivity. This paper explores how MetaCrystal, a set of visual tools, can provide insights into the effectiveness of different query formulations or the same query submitted to multiple search engines. It aims to show that visual abstractions, such as the degree of overlap or similarity between different ranked lists, can be used as indicators about the effectiveness of fusing different search methods.

Users tend to create short queries when searching the Internet and they rarely formulate advanced queries [17]. Eastman and Jansen [4] have shown that the use of most query operators in short Internet queries had no significant impact on the effectiveness of the search results. MetaCrystal can visualize this high degree of similarity between different formulations of simple Internet searches. It enables users to see that these related queries are not very effective in finding more relevant documents and that the relevance of the found documents can not be corroborated by these related queries.

Users employ meta search engines because individual search engines only index 20% of the Internet [11] and therefore return different documents for the same query. Meta search engines address this limitation by combining the results by different engines. MetaCrystal visualizes the precise overlap between the top documents retrieved by different search engines. It makes it easy to identify how many and which documents have been found by more than one search engine.

This paper is organized as follows: section 2 briefly reviews related visualization work. Section 3 describes the MetaCrystal toolset. Section 4 shows how the degree of similarity between different ranked lists can be visualized. Section 5 describes how MetaCrystal can provide visual feedback about the effectiveness of the search and fusion effort. Section 6 discusses issues related to MetaCrystal's implementation and addresses future research.

## 2. Related Visualization Work

Large sets of documents can be visualized by showing how the documents are related to specific "Points of Interest" (POIs), such as the query terms or search engines used to retrieve them [12]. The POIs act as magnets and the ratio of the "forces of attraction" between a document and the POIs determines its location in the display. Specifically, a document's position is equal to the sum of the position vectors of the POIs, where each vector is scaled by the strength of a document's relatedness to the POI. Thus, a document is placed closer to the POIs that it is more related to. It can be shown that a standard POI visualization can produce undesirable clustering results, because the distance from its center is not necessarily a reliable visual cue of a document's potential relevance. MetaCrystal's Cluster Bulls-Eye tool overcomes this problem by using a polar mapping, where only the angular value is determined by the POIs.

Sparkler [8] enables users to compare the results returned by multiple search methods. It combines a bull's eye layout with star plots, where a document is plotted on each star spoke based on its rankings by the different methods. Sparkler spreads the multiple results that have the same position on a spoke, and thus would overlap, to show their distribution pattern. It does not explicitly represent which documents have been retrieved by multiple search methods and users need to examine individual documents to determine how many and which retrieval methods found them.

Beadplots [1] aims to visualize the shared subpatterns in ranked lists of retrieved documents, because the similarity scoring used in most retrieval systems tends to find documents in groups. For a specific topic, the rows in a beadplot correspond to the different systems, and the "beads", gray and colored diamonds, along each row represent 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. Beads with the same color 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. The colors assigned to the documents use spectral (ROYGBIV) coding. The ordering ranges from most relevant (dark red) to least relevant (light violet), where color reference 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.

Several meta search engines have been developed that visualize the combined retrieved documents. Vivísimo [24] organizes the retrieved documents using a hierarchical folders metaphor. At the end of each document summary, the search engines are listed that retrieved the document, together with the ranking by each of these engines. Kartoo [10] creates a 2-D map of the highest ranked documents and displays the key terms that can be added or subtracted from the current query to modify it. Grokker [7] uses nested circles or rectangles to visualize a hierarchical grouping of the search results. MetaSpider [3] uses a self-organizing 2-D map approach to classify and display the retrieved documents. None of these visual meta search tools provide users with a compact visualization of the precise overlap between the search engines. Instead, they require substantial user interaction to infer the degree of overlap and users can not control how the results by the different engines are combined.

## 3. MetaCrystal Toolset

MetaCrystal [20, 21] consists of several linked tools that enable users to compare and combine the search results returned by different query formulations or different search methods processing the same query. Documents found by multiple retrieval methods are more likely to be relevant [5, 13]. MetaCrystal's tools map the documents found by multiple methods toward the center of their display and visualize all the retrieved documents in a compact and structured way. The *Category View* displays the precise overlap between the top result sets returned by different queries or search methods and shows the number of documents retrieved by different retrieval combinations. (see Figure 1). The *Cluster Bulls-Eye* enables users to see how *all* the found documents are related to the different search methods being compared. It clusters documents retrieved by multiple engines toward its center and at the same time helps users scan the top documents found by a single query or method (see Figure 1). This tool can also be used to visualize the degree of similarity between different ranked lists (see Figure 3). The *RankSpiral* places *all* the documents sequentially along an expanding spiral based on their total ranking scores to enable users to rapidly scan large numbers of documents and their titles.

Informal usability tests have been conducted and guided the development of the current version of MetaCrystal, [22]. Implemented in Flash using ActionScript, MetaCrystal supports flexible exploration and advanced filtering operations and guides users toward relevant information [21]. Its direct manipulation interface enables users to iteratively create crystals of increasing complexity that show the precise overlap between up to five search methods or queries [22]. Users can apply different weights to the queries or methods to create their own ranking functions. Users can control the degree of overlap between the different engines or queries by modifying the URL directory depth used for matching web pages or by changing the number of documents being compared. MetaCrystal can use the document scoring data provided by search methods, but these scores need to be normalized and, more importantly, are not always available. The presented examples use only the position or *ranking* of a document in the result list, since this data is always available. The Category View and Cluster Bulls-Eye tools will be described in more detail, because they enable users to identify documents found by multiple search methods and to see how their rankings by the different methods are related. This type of visual feedback can be used to determine the effectiveness of the search and fusion process.

### *3.1. Category View*

In Figure 1, the Category View displays the precise overlap between the top documents retrieved by the search engines Google, Teoma and AltaVista, when searching for 'information visualization'.

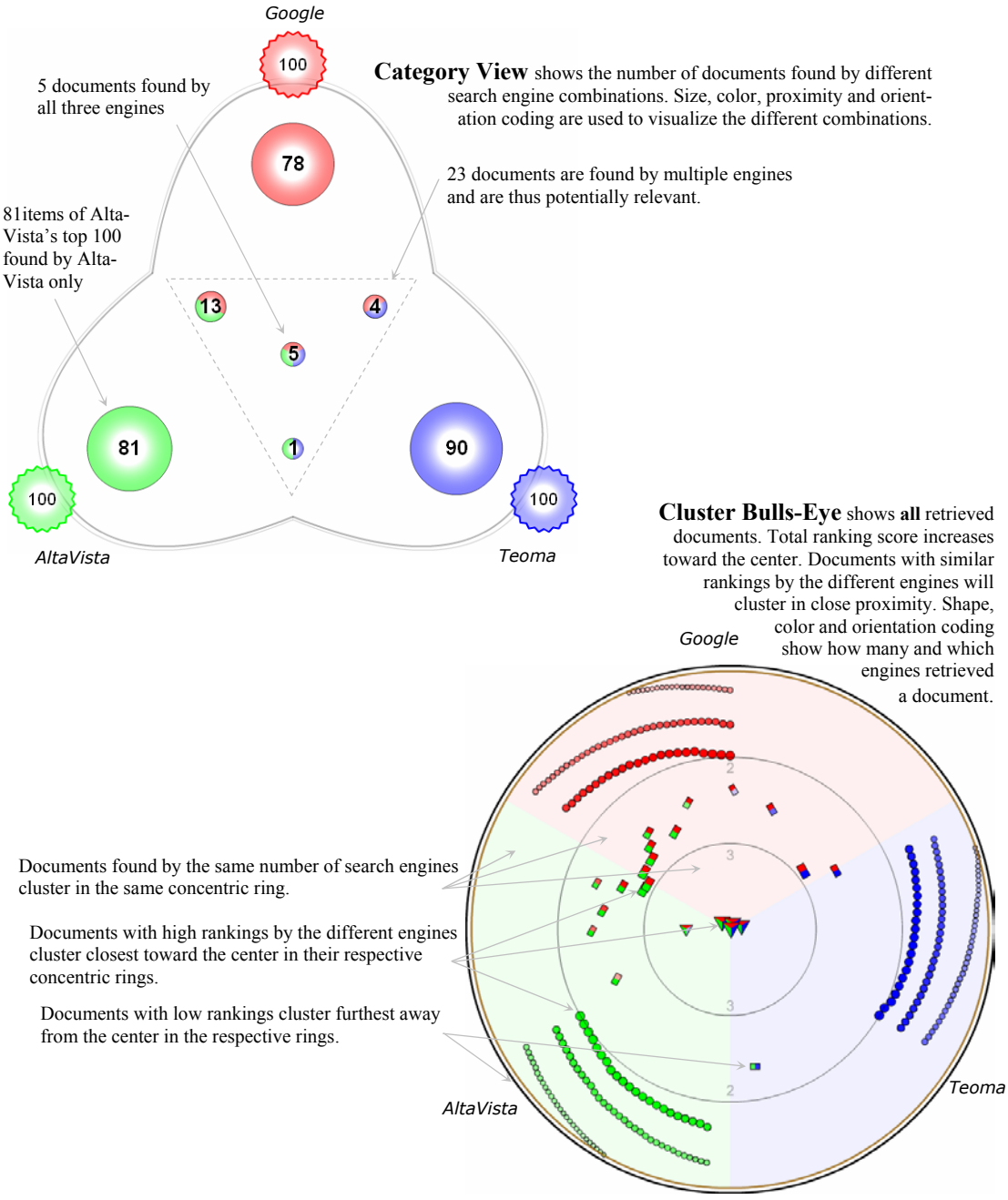



**Figure 1**: The ***Category*** and ***Cluster Bulls-Eye*** tools provide users with complementary ways to explore the precise overlap between the top 100 documents found by Google, Teoma and AltaVista, when searching for 'information visualization'. These overviews make it easy for users to identify how many and which documents have been retrieved by more than one search engine.

Modeled on the InfoCrystal layout [19], the interior consists of *category icons*, whose shapes, colors, positions and orientations encode different search engine combinations. At the periphery, colored and star-shaped *input icons* represent the different engines, whose top 100 results are compared to compute the contents of the category icons. The icon in the center of the Category View displays the number of documents retrieved by all engines. The number of engines represented by a category icon decreases toward the periphery. Shape coding is used for a category icon if we want to emphasize the number of search engines it represents (see Figure 4); size coding is employed to emphasize the number of documents retrieved by a search engine combination (see Figure 1). The Category View supports these tasks: a) identification of the number of documents found by multiple engines and by which combinations; b) selection of specific search method combinations to specify Boolean constraints, since the category icons represent all possible Boolean queries in disjunctive normal form [18, 19].

### *3.2. Cluster Bulls-Eye*

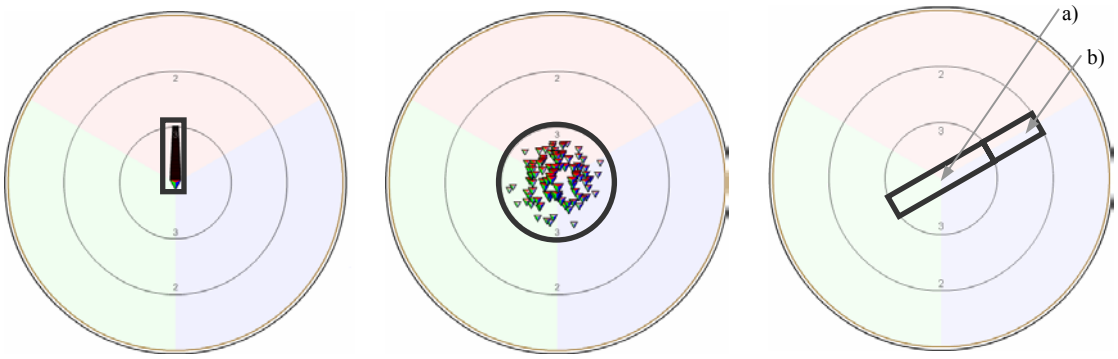
In Figure 1, the Cluster Bulls-Eye tool shows how *all* the retrieved documents are related to the different engines, because a document's position reflects the relative difference between its rankings by the different search engines. Documents with similar rankings by the different engines will be placed in close proximity. Shape, color and orientation coding indicate which search engines retrieved a document. The Cluster Bulls-Eye uses polar coordinates to display the documents: the *radius* value is related to a document's total ranking score so that the score increases toward the center; the *angle* reflects the relative ratio of a document's rankings by the different engines. The *total ranking score* of a document is equal to the sum of the number of engines that retrieved it and the normalized average of its rankings by the different engines that found it. This causes documents retrieved by the same number of engines to cluster and to be contained in the same concentric ring (see Figure 1). Specifically, documents with high rankings by the different engines cluster in their respective concentric rings so that they are closest to the center of the display and the size of their icons is set to the largest value. Documents with low rankings cluster furthest away from the center in their respective rings and the size of their icons is set to the smallest value.

The Cluster Bulls-Eye tool combines a POI visualization with a "bull's eye" mapping to ensure that users will always find documents with high total ranking scores toward its center. Although not always explicitly shown in Cluster Bulls-Eye, the input icons act as "points of reference" that pull a document toward them based on the document's rankings by the different engines. The documents that are only retrieved by a single engine represent a special case. Their angle value would be equal to the angle of the position vector of its related input icon, causing the documents to overlap on a straight line. Instead, document icons are "fanned out" so that they don't overlap, making it easy for users to scan the top documents only retrieved by a specific engine (see Figure 1). The Cluster Bulls-Eye tool supports these tasks: a) identification of documents retrieved by multiple engines and by which combinations; b) exploration of top documents found by a single engine; c) identification of the degree of similarity between different ranked lists.

## 4. Visualizing Degree of Similarity Between Multiple Ranked Lists

The Cluster Bulls-Eye tool can be used to visualize the degree of similarity between different ranked lists. If a document is contained in all the search results being compared, then it will be placed inside the inner most circle of the Cluster Bulls-Eye. The greater a document's rankings in the lists being compared, the closer toward the center its position will be. If a document has the same ranking in all of

**Visualizing Degree of Similarity between Three Ranked Lists**




[ i ] Identical [ ii ] Random [ iii ] High Similarity Areas

**Figure 2**: shows the "visual signatures" and how documents will cluster in the *Cluster Bulls-Eye* if the three ranked lists being compared: [i] are identical; [ii] contain the same documents, but their ranking orders are randomized. [iii] a) shows the area where the documents found by all three queries will cluster if the rankings for the first two queries are identical; [iii] b) highlights the area where the documents found by only two queries will cluster if the rankings for these two queries are identical.

**Comparing Results by Different Query Formulations**

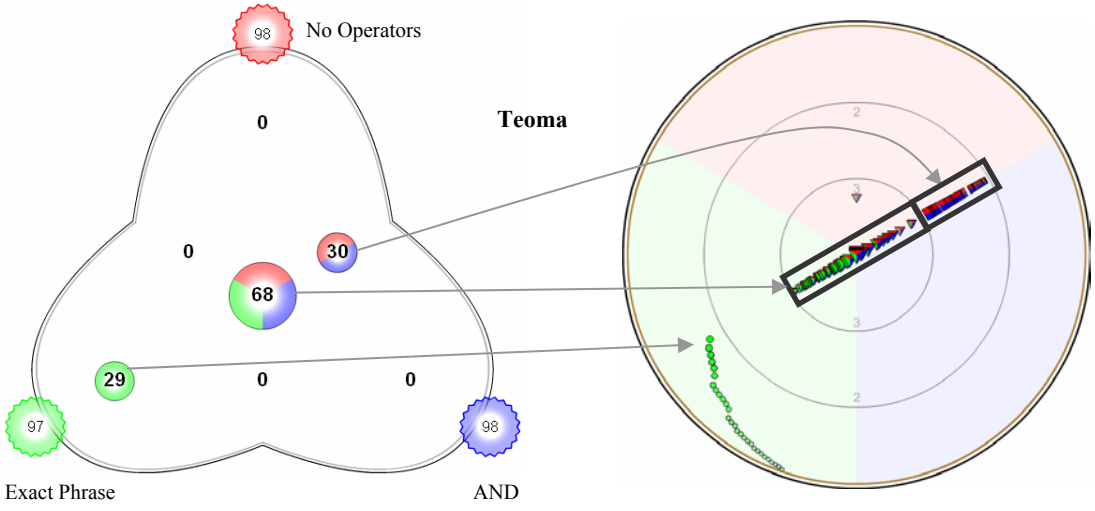



**Figure 3**: shows the degree of overlap between the top documents retrieved if we search for 'information visualization' and compare queries that employ "no operators", a "Boolean AND" or an "exact phrase" constraint, using the Teoma search engine. The *Category View* on the left uses size coding for the category icons and shows that there is a great deal of overlap between the queries. In particular, the same documents are retrieved by the queries that use "no operator" or the "Boolean AND", because the category icons related to only one of these queries are empty. The *Cluster Bulls-Eye* on the right shows that the ranking order for the documents found by more than one query is identical or very similar. Specifically, the documents found by only two queries have identical rankings. The documents found by all three queries have identical rankings for the "no operator" and "Boolean AND" queries. Thus, these different query formulations do not provide independent "sources of evidence" to infer the potential relevance of the documents.

the results being compared, then by default its angle will be set equal to 90 degrees (see Figure 2 [i]). If the rankings of the documents are not correlated in the different result sets, then they will cluster as shown Figure 2 [ii]. If the rankings are identical for two of three results lists being compared, then documents will cluster along the midline between the two inputs associated with each input query, as shown in Figure 2 [iii] a). If documents are only contained in two of the three result lists being compared and their rankings are identical, then they will cluster as shown in Figure 2 [iii] b). Figure 2 demonstrates that certain types of similarity relationships between ranked lists give rise to unique "visual signature" patterns in the Cluster Bulls-Eye tool. Thus, it can be used to determine the degree of similarity between the ranked lists returned by different queries or search engines.

## 5. Visualizing the Effectiveness of Different Query Formulations or Search Methods

In this section, it will be shown how MetaCrystal can provide insights into the effectiveness of different query formulations or search methods as well as if the fusion of their result sets is effective. Different query formulations or search methods will be deemed *more effective* if the fusion of their search results leads to a new result set that contains an increased number of relevant documents. Research has shown that documents found by multiple methods are more likely to be relevant [5, 13]. However, it will be shown in this section that the relationships between the rankings by the different methods, which found the same document, needs to be considered as well in order to assess the effectiveness of the different search results that are being compared. This section is organized as follows: 1) different query formulations of the same search topic are created and their results compared; 2) the same query is submitted to multiple Internet search engines and their results are compared; 3) for two different search topics, the results sets of the top five search methods, which participated in the TREC 8 automatic Ad Hoc task, are compared [25]. For the latter, it will be possible to provide quantitative effectiveness measures.

### *5.1. Visual Comparison of Different Query Formulations Results*

The Category View and Cluster Bulls-Eye tools can be used to visualize and support the finding that the use of most query operators in short Internet queries leads to very similar search results [4], both in terms of the retrieved documents and their respective rankings. In Figure 3, searching for 'information visualization', we visually compare the search results returned if a query that uses "no operators", a "Boolean AND" or an "exact phrase" constraint, respectively, is submitted to the Teoma search engine. The Category View shows that there is a great deal of overlap between the different formulations, which would suggest that the documents found by multiple queries are more likely to be relevant [5, 13]. However, the Cluster Bulls-Eye tool shows that the queries which use "no operators" or the "Boolean AND", retrieve the same documents and their rankings are identical, because they cluster in Figure 3 in the same areas that are highlighted in Figure 2 [iii] a) and b), respectively. The high number of documents retrieved by more than one query formulation can not be interpreted in isolation. The relationship between the rankings by the different queries needs to be considered. This helps users determine if the different queries actually represent sufficiently independent "sources of evidence" to infer potential relevance. Because the different query formulations are submitted to the same Teoma database, the high number of documents found by multiple queries in the Category View and their highly structured distribution pattern in Cluster Bulls-Eye display indicate that the different query formulations are not very effective in finding more relevant documents. In the Cluster Bulls-Eye tool,

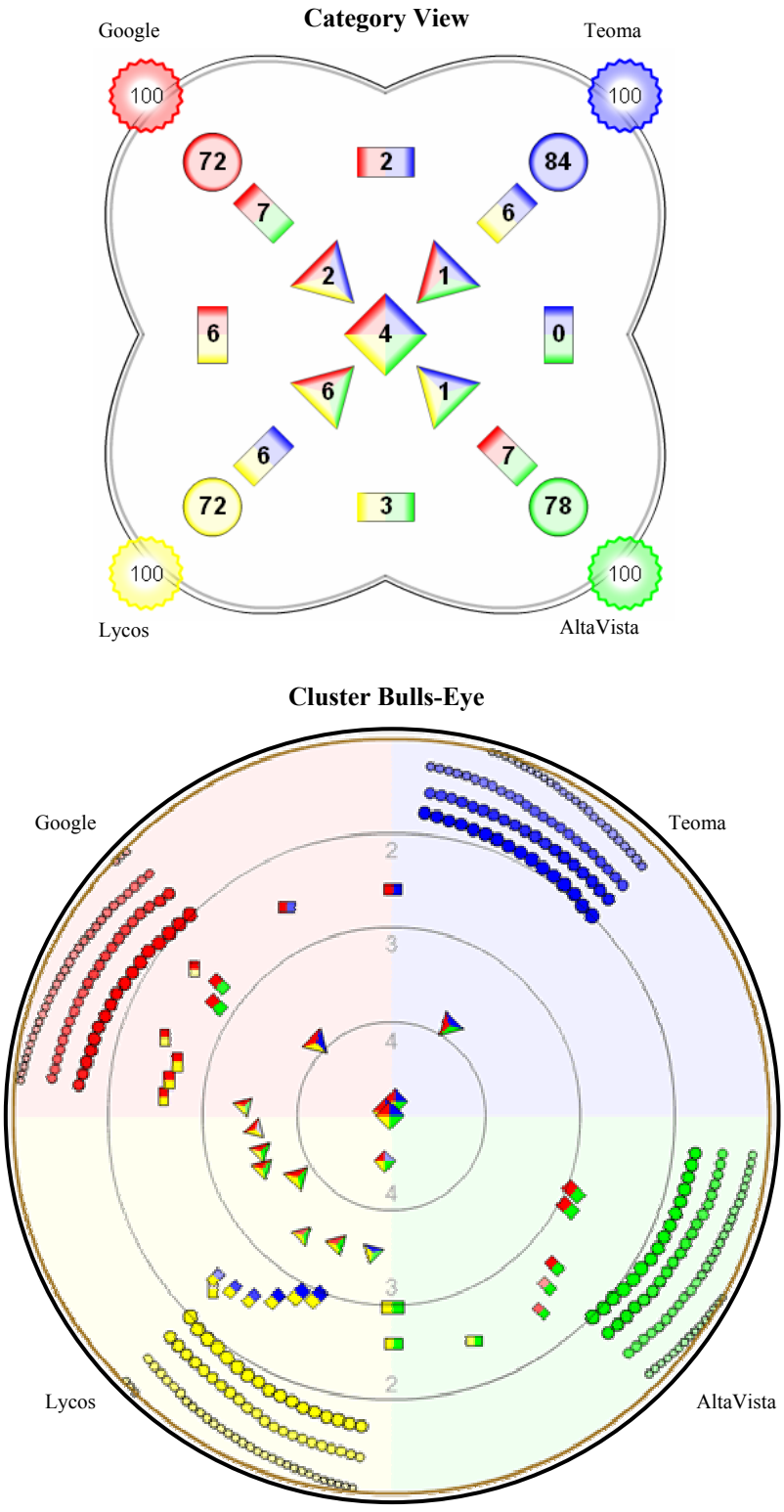



**Figure 4**: shows the visual comparison of the meta search results for the top 100 documents found by Google, Teoma, AltaVista and Lycos when searching for 'information visualization'.

the distribution patterns of the documents found by a single method can be used to infer, for example, whether or not highly ranked documents were also retrieved by the other methods. Figure 3 shows that the highly ranked documents found by the "exact phrase" query are also retrieved by the other query formulations, because there are no document icons in the vicinity of the border to the area containing documents found by two different query formulations.

### *5.2. Visual Comparison of Meta Search Results*

Meta search engines combine the results by different engines, because individual engines tend to index 20% of the Internet [11] and thus return different documents for the same query. In meta search context, the fact that a document has been retrieved by multiple engines is significant, because each search engine uses a unique retrieval method and indexes different parts of the Internet. Hence, each engine can be understood as an independent "source of evidence" that can be used to corroborate the potential relevance of a document. MetaCrystal makes it easy for user to identify how many and which documents have been found by more than one search engine. The Category View groups all the documents found by the same combination of engines and displays the number of documents retrieved by different search engine combinations. Figures 1 and 4 show that there is some overlap between the result sets, but at least 70% of documents retrieved by an engine are not found by the others. The Cluster Bulls-Eye tool visualizes the relationship between a document's rankings by the different engines that retrieved it. Figures 1 and 4 show that there is no systematic pattern of relatedness between the rankings for the documents found by multiple search engines. The Category View and Cluster Bull's-Eye displays suggest that the meta search has been effective in retrieving documents likely to be relevant. An exploration of the top results generated by the MetaCrystal fusion suggests that most of documents retrieved by multiple engines are relevant. In particular, the top document "OLIVE: Online Library of Information Visualization Environments", which has been retrieved by all search engines and had the highest total ranking score, is an excellent top choice.

### 5.3. Visual Comparison of TREC results

The TREC workshops provide information retrieval researchers with large documents collections, a set of search topics and ways to compare the search results [25]. Participating retrieval systems search the collections for each of the 50 provided topics, and then submit a ranked list of 1000 documents for evaluation. For each topic, NIST pools the top 100 retrieved documents from each run. The person who proposed a topic then determines the relevance of each document. The list of relevant documents for each topic is publicly available and the systems are evaluated based upon different measures of recall and/or precision. *Recall* assesses the fraction of relevant documents that were found by a system, while *precision* assesses the fraction of a system's retrieved documents that are actually relevant.

In this section, the top five retrieval systems from the *automatic short Ad Hoc track* in *TREC 8* are compared for two of the 50 topics used to evaluate the systems. These top systems use different approaches to identify 1000 potentially relevant documents and specific information about the systems can be found in the TREC 8 proceedings [25]. Figure 5a shows the precise overlap between result sets by the top five systems when searching for *topic 401*. The Category View shows that 99 documents are found by all systems; more than 50% of documents returned by the top performing retrieval system are not found by the other systems. The Cluster Bulls-Eye shows that there is no systematic relationship between the rankings for documents found by multiple systems and the document icons are mostly scattered in a uniform fashion (see Figure 5b which does not display the documents found by a single system). As stated previously, different search methods will be deemed *more effective* if the fusion of

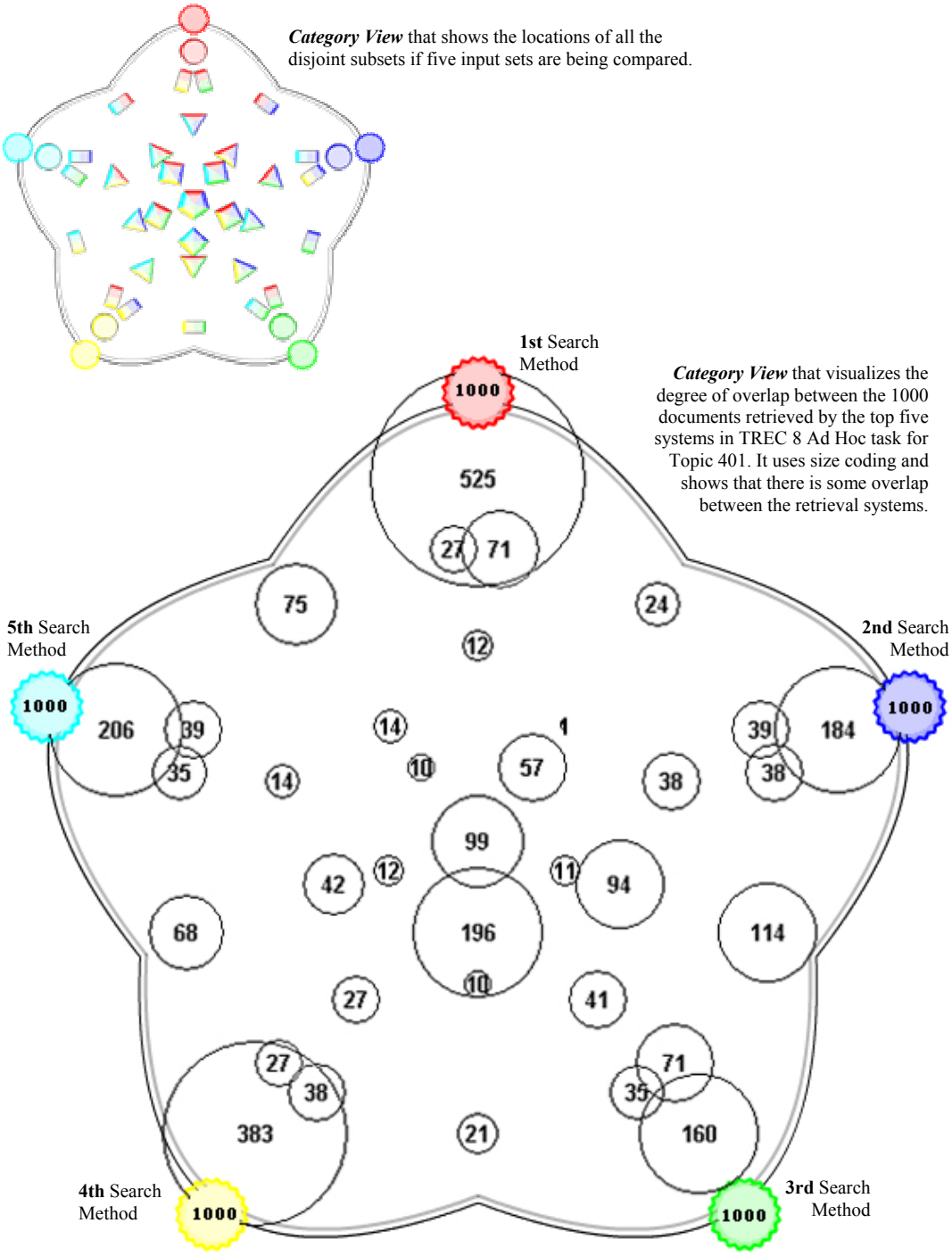



**Figure 5a**: ***Category View***, using size coding, of the top five search methods in TREC 8 Ad Hoc task for Topic 401.

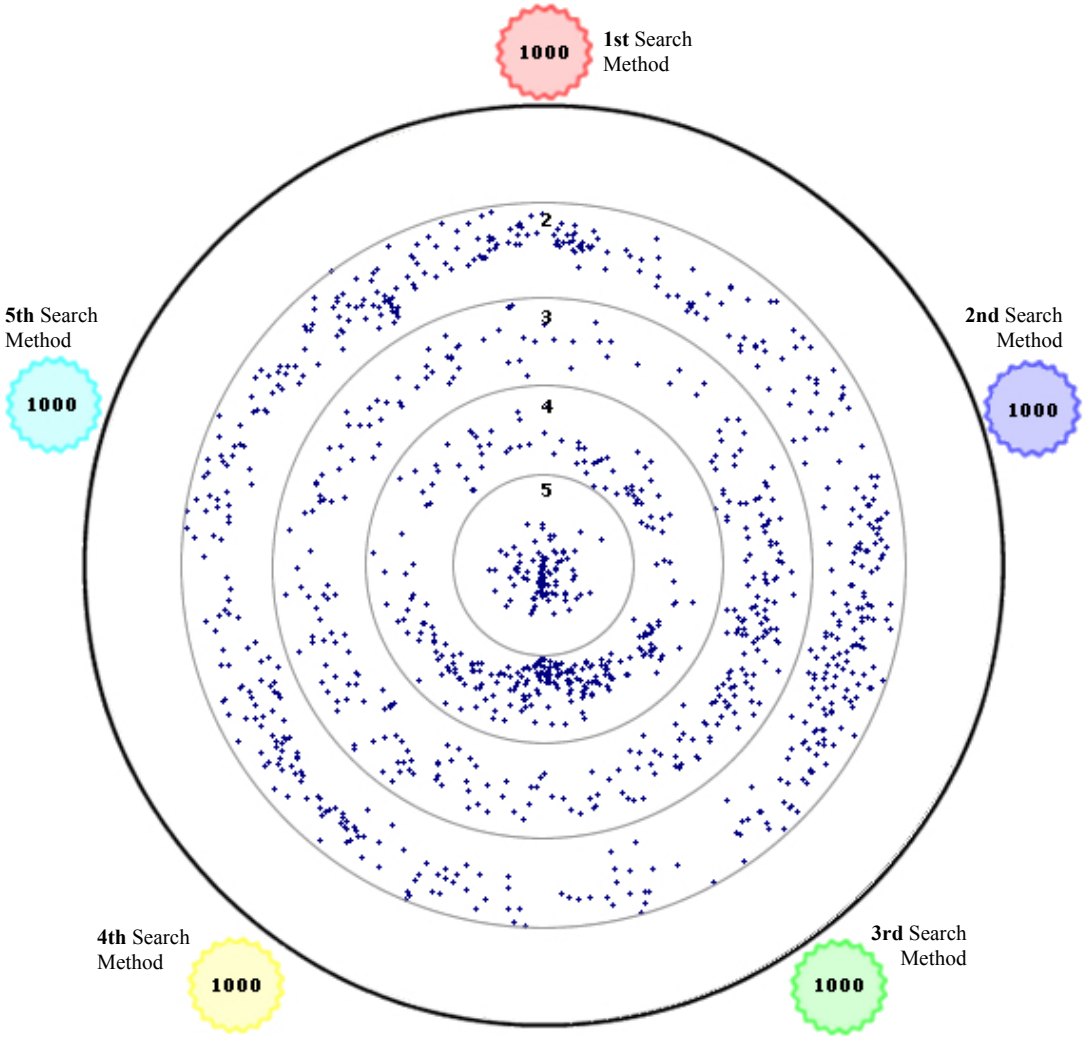



**Figure 5b**: ***Cluster Bull's Eye*** for the top five search methods in TREC 8 Ad Hoc task for Topic 401.

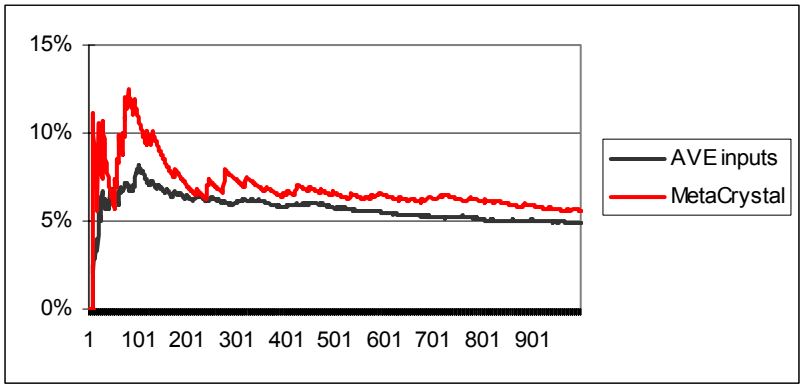



**Figure 5c**: Average of the percentage of the relevant documents in the 1000 documents retrieved by the top five methods in TREC 8 Ad Hoc task for Topic 401 and the MetaCrystal fusion method.

their search results leads to a new result set that contains an increased number of relevant documents. Figure 5c displays the average number of relevant documents that have found as a function of the documents examined that have been returned by the top five search systems. Figure 5c also displays the resulting precision curve if the top 1000 documents, based on the total ranking score, are selected in the Cluster Bulls-Eye display; it shows that 5.6% of the 1000 documents returned by the MetaCrystal fusion approach are relevant, which is 14.75% better than the average of the precisions values of the top five systems.

In Figure 6a, the Category View shows that 832 documents are found by all five systems for *topic 404*. As mentioned earlier, documents found by multiple methods are more likely to be relevant, but the relationships between the rankings need to be considered. The Cluster Bulls-Eye does not reveal a systematic relationship between the rankings for documents found by multiple systems (see Figure 6b). However, if so many documents are retrieved by all systems, then it decreases the probability that the relevant documents found only by a single system will be included in the MetaCrystal fusion set and thus it will not lead to a fused result set that contains more relevant documents than are included in any of the input sets. Figure 6c displays the average of the precision values of the top five search systems and the MetaCrystal precision curve; it shows that 12.6% of the 1000 documents returned by the MetaCrystal fusion approach are relevant, which is only 0.32% better than the average of the precisions values of the top five systems.

## 6. Discussion and Future Work

The MetaCrystal has been implemented in Flash using ActionScript 1.0. This has the advantage that it can be deployed using a Web browser and the file size of the application is small. The current implementation can compare different ranked lists that each contains at most 200 documents. This limitation is due to the fact that Flash "times out" if a computation takes longer than a certain time period. MetaCrystal is currently being reimplemented using ActionScript 2.0, which has better performance characteristics. The Category View and Cluster Bulls-Eye displays have been implemented using Microsoft Excel to be able to compare the TREC 8 Ad Hoc result sets, which contain 1000 documents. Figures 5 and 6 were created using Excel's charting tools.

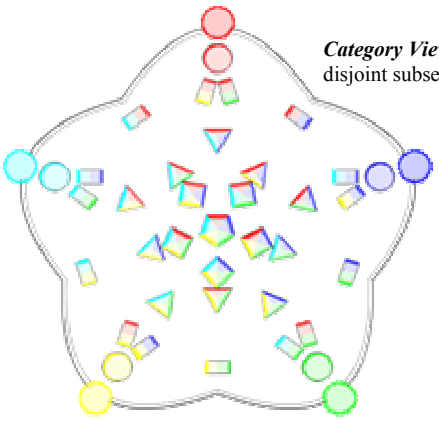
As the MetaCrystal toolset was being developed, an informal usability test was conducted, where 15 graduate students interacted with MetaCrystal's different tools [22]. They rated the *Category View* as most effective, especially when shape coding was used (see Figure 4 top). The students had some difficulty understanding the conceptual difference between the *Category View*, which displays groupings of documents and the *Cluster Bulls-Eye,* which displays individual documents. The latter tool initially represented the individual documents as circles with colored slices, which may have contributed to the confusion the students experienced. A document is now visualized using also shape and size coding (see Figure 4 bottom). Shape encodes the number of engines that retrieved a document and size reflects the rankings by the engines. The only difference between a category icon and a document icon is that the former displays the number of documents retrieved by the search engine combination it represents.

Furthermore, a formal user study was conducted to test if users can interpret the category icons in terms of their associated Boolean meaning based on the shape, color, orientation and proximity coding employed [18]. The user study compared a standard, text-based Boolean query language with the InfoCrystal interface, which is the precursor to the Category View. InfoCrystal represents all the

possible queries among its inputs in *disjunctive normal form* and each category icons represents a distinct Boolean relationship. Subjects had to perform a *recognition* and *generation* task. The former asked them to recognize either the correct Boolean or InfoCrystal query from three possible queries. In the latter task subjects had to generate a Boolean or InfoCrystal query that captured a given information need. For both tasks each subject was presented with each query in both modes. Hence, the paired-differences in performance between the two query languages could be computed to reduce the noise in the collected data. In the recognition task there is no statistically significant difference between the two query languages both in terms of the scores or the time measurements. In the generation task there is significant difference for the scores in favor of the Boolean mode, but the experiment favored this mode. In terms of the time measurements there was a statistically significant difference in favor of the InfoCrystal. The collected user feedback regarding the InfoCrystal (IC) was positive, ranging from "The IC was absolutely clear." to "The IC was actually not that bad, even usable." This formal user study demonstrates that novice users, who received a short training tutorial, can interpret and distinguish between the category icons successfully. This study also suggests that users should be able to infer how many documents have been found when all search engines and how many documents have been found by single engines when using the Category View.

The presented examples lend support to the working hypotheses and guiding principles of this paper: 1) documents found by multiple methods are more likely to be relevant; 2) a high degree of overlap and/or systematic relationships between the ranked lists being compared will not lead to fusion results that contain more relevant documents and thus do not improve the effectiveness of the search effort. The former hypothesis has been supported in a qualitative way by the meta search example, where the documents found by multiple Internet search engines are mostly relevant, and in a quantitative way by the two TREC 8 examples, where the precision at 1000 documents for the MetaCrystal fusion approach is as good as the average of the precisions of the top five retrieval systems being fused. The latter hypothesis has been supported in a qualitative way by comparing different formulations of a short Internet query and in a quantitative way for the two TREC examples, where it could be shown that a high degree of overlap and/or systematic relationships between the document rankings limit the effectiveness of the fusion process.

More extensive research is needed to prove the validity of the hypotheses and principles guiding the research presented in this paper. The primary goal of future research is to quantify and show how visual features, such as the number of documents found by all or just single search methods as well as systematic relationships between the rankings, are correlated to the effectiveness of different fusion methods. Specifically, future research will seek to quantify how an increase in the number of methods that retrieved a method affects the probability that the document is relevant. Once it has been established that there are statistically significant and robust visual features in the MetaCrystal displays, as the presented examples suggest, then a formal user study will be conducted to test if users are able to perceive and use these visual cues to determine the effectiveness of the search and fusion process. It is currently being investigated if and how the MetaCrystal fusion method leads to improved precision performance if all the 50 topics and random subsets of all systems participating in the TREC 8 Ad Hoc track are considered. Preliminary results indicate that the MetaCrystal fusion method leads to improved precision. Future work will also compare different fusion methods with the MetaCrystal fusion approach.



***Category View*** that shows the locations of all the disjoint subsets if five input sets are being compared.

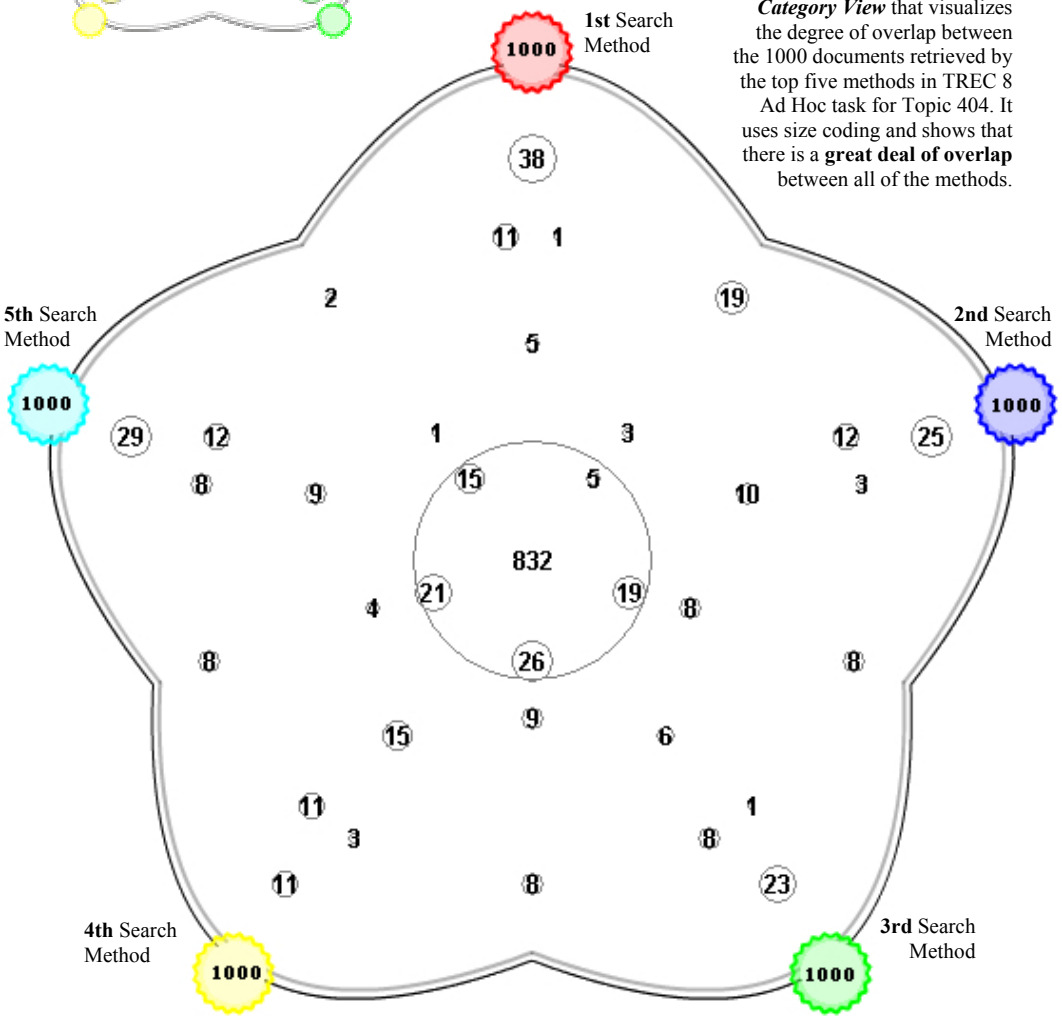



**Figure 6a**: ***Category View***, using size coding, of top five search methods in TREC 8 Ad Hoc task for Topic 404.

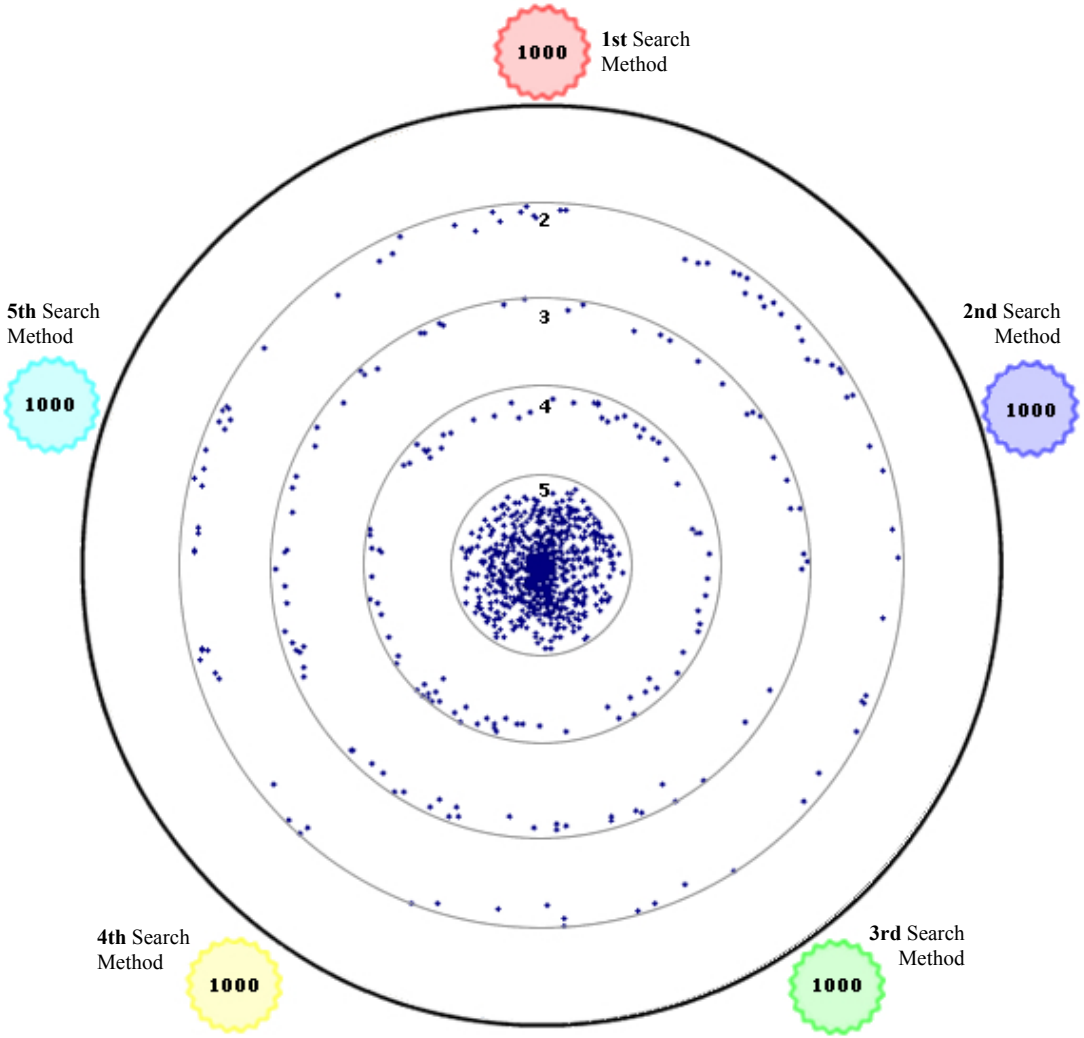



**Figure 6b**: ***Cluster Bull's Eye*** for the top five search methods in TREC 8 Ad Hoc task for Topic 404.

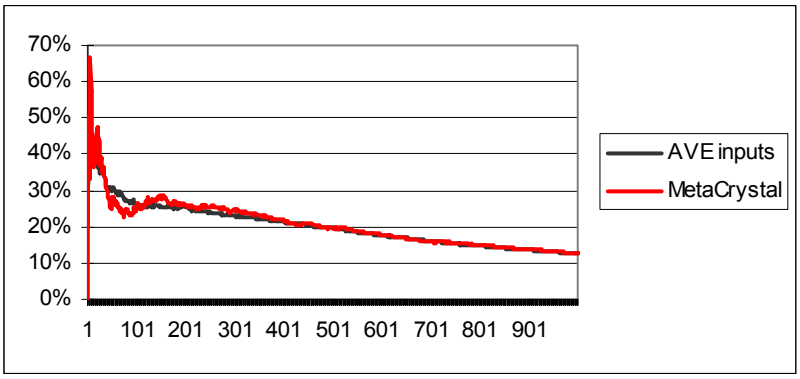



**Figure 6c**: Average of the percentage of the relevant documents in the 1000 documents retrieved by the top five methods in TREC 8 Ad Hoc task for Topic 404 and the MetaCrystal fusion method.

## 7. Conclusions

This paper addressed how information visualization can provide insights into the effectiveness of the search and fusion process. Different query formulations or search methods were deemed more effective if the fusion of their search results leads to a new result set that contains an increased number of relevant documents. MetaCrystal and its Category View and Cluster Bulls-Eye tools were used to visualize the degree of overlap and similarity between the results returned by different query formulations or search methods. It was shown that certain types of similarity relationships between ranked lists give rise to unique visual patterns in the Cluster Bulls-Eye display.

The presented research has been guided by these two hypotheses: 1) documents found by multiple methods are more likely to be relevant; 2) a high degree of overlap and/or systematic relationships between the ranked lists being compared will *not* lead to fusion results that contain more relevant documents and thus do not improve the effectiveness of the search and fusion effort. The Category View enables users to identify how many and which documents have been retrieved by more than one method. It was shown that the relationship between a document's rankings by the different retrieval methods needs to be also considered to infer a document's potential relevance. The Cluster Bulls-Eye tool enables users to visually examine and spot specific distribution pattern for the rankings of the retrieved documents. If there is a systematic pattern of relatedness between the rankings for the documents found by more than one method, then it is likely that the fusion of different retrieval methods will not retrieve more relevant documents, especially if they search the same database.

Multiple examples were presented: 1) different query formulations of the same search topic were created and their results compared; 2) the same query was submitted to multiple Internet search engines and their results compared; 3) for two different search topics, the results sets of the top five search methods, which participated in the TREC 8 automatic short Ad Hoc task, were compared. MetaCrystal was used to visualize and support the finding that the use of query operators in short Internet queries leads to very similar search results [4] and thus the different query formulations are ineffective in retrieving more relevant documents. The Category View showed that there is a great deal of overlap between the documents retrieved by different formulations of the same query, but little overlap when the same query is submitted to different Internet search engines. The Cluster Bulls-Eye tool was used to show that the rankings for the documents retrieved by different query formulations are highly related, whereas the rankings in the meta search context are not related in a structured way. Precision plots were presented for the two TREC 8 topics to address in a quantitative way the appropriateness of the hypotheses and principles guiding the presented research. Future research aims to quantify and show how visual features, such as the number of documents found by all search methods or systematic relationships between document rankings, are correlated to the effectiveness of different fusion methods. Once it has been established that there are statistically significant and robust visual features in the MetaCrystal displays, as the presented examples suggest, then a user study will be conducted to test if users are able to perceive and use these visual cues to determine the effectiveness of the search and fusion process.

## Acknowledgements

This work extends a paper presented at the *Workshop on Measuring Web Search Effectiveness: the User Perspective* at WWW 2004. The TREC 8 data used has been provided by NIST and can be downloaded at www.nist.org. This research has been supported by a Rutgers Research Council Grant.

## References


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

[2] Belkin, N. & Croft, B. (1992). Information Filtering and Information Retrieval: Two Sides of the Same Coin. Comm. of the ACM, Dec., 1992.

[3] Chen H., Fan H., Chau M. and Zeng D. MetaSpider: (2001). Meta-Searching and Categorization on the Web. JASIS, Volume 52 (13), 1134 - 1147.

[4] Eastman, C. and Jansen, B. J. (2003) Coverage, relevance, and ranking: the impact of query operators on Web search engine results. ACM Transactions on Information Systems. 21(4), 383 - 411.

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

[6] Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. Information Processing and Management, 35 (2), 141–180.

[7] Grokker – www.groxis.com

[8] Havre, S., Hetzler, E., Perrine K., Jurrus E., and Miller N. (2001). Interactive Visualization of Multiple Query Results. Proc. IEEE Information Visualization Symp. 2001.

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

[10] Kartoo – www.kartoo.com

[11] Lawrence, S., & Giles, C.L. (1999). Accessibility of information on the Web. Nature, 400, 107–109.

[12] 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.

[13] Saracevic, T. and Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches and overlap. JASIS. 39, 3, 197-216.

[14] Saracevic, T. (1995). Evaluation of evaluation in information retrieval. Proceedings of ACM SIGIR ' 95.

[15] Spink, A. (2002). A user centered approach to the evaluation of Web search engines: An exploratory study. Information Processing and Management, 38(3), 401-426.

[16] Spink, A., Wilson, T. D. (1999). Toward a theoretical framework for information retrieval (IR) evaluation in an information seeking context. Proceedings of MIRA 99:

[17] Spink, A., Wolfram, D., Jansen, B. J., and Saracevic, T. (2001). Searching of the Web: the public and their queries. JASIS, 52 (3) (2001), 226 - 234 .

[18] Spoerri, A. (1995). InfoCrystal: A Visual Tool for Information Retrieval. Interdepartmental Ph.D. Thesis. Massachusetts Institute of Technology. February, 1995.

[19] Spoerri, A. (1999). InfoCrystal: A Visual Tool for Information Retrieval. In Card S., Mackinlay J. and B. Shneiderman (Eds.), Readings in Information Visualization: Using Vision to Think (pp. 140 – 147). San Francisco: Morgan Kaufmann.

[20] Spoerri, A. (2004). MetaCrystal: A Visual Interface for Meta Searching. Proceedings of ACM CHI 2004.

[21] Spoerri, A. (2004). Coordinated Views and Tight Coupling to Support Meta Searching. Proceedings of IEEE CMV 2004.

[22] Spoerri, A. (2004). Visual Search Editor for Composing Meta Searches. Proceedings of ASIS&T 2004.

[23] Su. L. T., Chen, H. L., & Dong, X. Y. (1998). Evaluation of Web-based search engines from an end-user's perspective: A pilot study. Proceedings of ASIS&T 1998.

[24] Vivisimo – www.vivisimo.com.

[25] Voorhees E. & Harman D (2000). Overview of the eighth Text REtrieval Conference (TREC-8). In E.M.Voorhees and D.K. Harman, editors, Proceedings of the Eighth Text REtrieval Conference (TREC-8), 2000. NIST Special Publication 500-246.