

Visual Search Editor for Composing Meta Searches

Anselm Spoerri

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

MetaCrystal is visual tool for creating and editing meta search queries. Users can visually combine the top results retrieved by different search engines to create crystals of increasing complexity. MetaCrystal consists of several overview tools to help users gain insight into how to filter the results from the different engines. It shows users the degree of overlap between the top results returned by different search engines. Users can apply different weights to the search engines to create their own ranking functions and can perform advanced filtering operations visually. Users can modify the URL directory depth or change the number of top documents used to increase or decrease the degree of overlap between the different engines. MetaCrystal addresses the problem of the effective fusion of different search engine results by enabling user to visually compose and refine meta searches.

Introduction

Users can choose from a multitude of Internet search engines, which tend to return different documents for the same query, in part because they cover different portions of the Internet (Lawrence & Giles 1999). Meta search engines address this limitation by combining the results returned by different engines. The automatic and effective fusion of different search results can be difficult (Callan 2000). While meta search engines, such as Kartoo, Grokker or MetaSpider (Chen et al. 2001), visually organize the retrieved documents, no meta search interface exists that provides users with an overview of the precise overlap between the search engines. Meta searching can benefit from such a visualization, because: a) documents found by multiple engines are likely to be more relevant (Saracevic & Kantor 1988, Foltz & Dumais 1992); b) it is difficult to predict the quality of coverage for single search engines, which tend to index less than 20% of the Internet (Lawrence & Giles 1999); c) some engines are more effective than others depending on the search domain (Gordon & Pathak 1999); d) users prefer or trust some engines more than others. This suggests that active user

involvement can make a difference when it comes to deciding how to combine the search results. While documents found by multiple engines are likely to be more relevant, users may also want to examine the documents only found by their most trusted engines. This range of requirements can be addressed by MetaCrystal.

User Need

One of the key challenges users face when conducting a meta search is how best to take advantage of the different strengths of the major search engines. Each engine can be understood as providing a different “point of view”, because it uses a unique retrieval method and indexes different parts of the Internet. Research in information retrieval and information filtering has shown that documents found by multiple retrieval methods are more likely to be relevant (Saracevic & Kantor 1988, Foltz & Dumais 1992). This suggests that an effective meta search interface needs to help users identify documents found by multiple search engines. However, the number of documents retrieved by more than one search engine tends to be small (Lawrence & Giles 1999). Users need ways to increase and control the number of documents found by multiple engines. Further, some of the top documents found by a single search engine will be relevant, but users don’t want to have to sift through multiple screens to find them (Hearst 1999, Mann 2002). A meta search interface needs to help users scan the top documents only retrieved by a specific engine, especially since users may prefer some engines more than others.

Similar to the “poly-representation” approach proposed for documents (Ingwersen 2002), MetaCrystal uses complementary visual representations to support users in the meta search process. The hallmark of an effective visualization is that it guides users toward relevant information. Ranked lists have the advantage that users know where to start their search for potentially relevant documents. However, users have to move sequentially through the list and only a small subset of the documents is visible in a single screen. MetaCrystal’s consists of several overview tools that all use a “bull’s eye” layout so that users can expect to find relevant documents toward the center of the displays. These tools also show all the docu-

ments in a compact and structured way. MetaCrystal's tools share a key advantage of a ranked list and overcome its limitations; they guide users toward to potentially relevant documents, while displaying a large number of documents.

This paper is organized as follows: First, previous work is briefly reviewed. Second, MetaCrystal's functionality is described and how its tools support users in the meta search process. Third, it is illustrated how MetaCrystal can be used as a visual search editor to compose meta queries that compare the top results of up to five search engines. Fourth, it is discussed how users can: a) perform complex filtering operations visually; b) apply different weights to the search engines to create their own ranking functions; and c) control the number of documents found by multiple engines. Finally, the lessons learned from an informal evaluation and future research are discussed.

Previous Work

Several meta search engines have been developed that visualize the combined retrieved documents. Vivísimo (www.vivisimo.com) 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 score by each of the engines. Kartoo (www.kartoo.com) 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 broaden or narrow it. Grokker (www.groxis.com) uses nested circular or rectangular shapes 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 display the retrieved documents. Sparkler (Harvé 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 different engines. None of these visual meta search interfaces 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; users have to examine individual documents to be able to determine which search engine combinations retrieved them.

MetaCrystal

The MetaCrystal interface enables a user to formulate a query, submit it to several search engines and use their top 100 results to compose a crystal that visualizes the precise overlap between the different search engines (see Figure 1). The retrieved documents can be progressively filtered to create a result set to be explored in more detail. MetaCrystal consists of several linked tools to help users gain insight into how to filter and combine the top results returned by the different search engines. The *Category View* displays the number of documents found by different search engine combinations. The *Cluster Bulls-Eye* tool enables users to see how *all* the retrieved documents are

related to the different engines. The *RankSpiral* tool places *all* the documents sequentially along a spiral based on their total ranking scores.

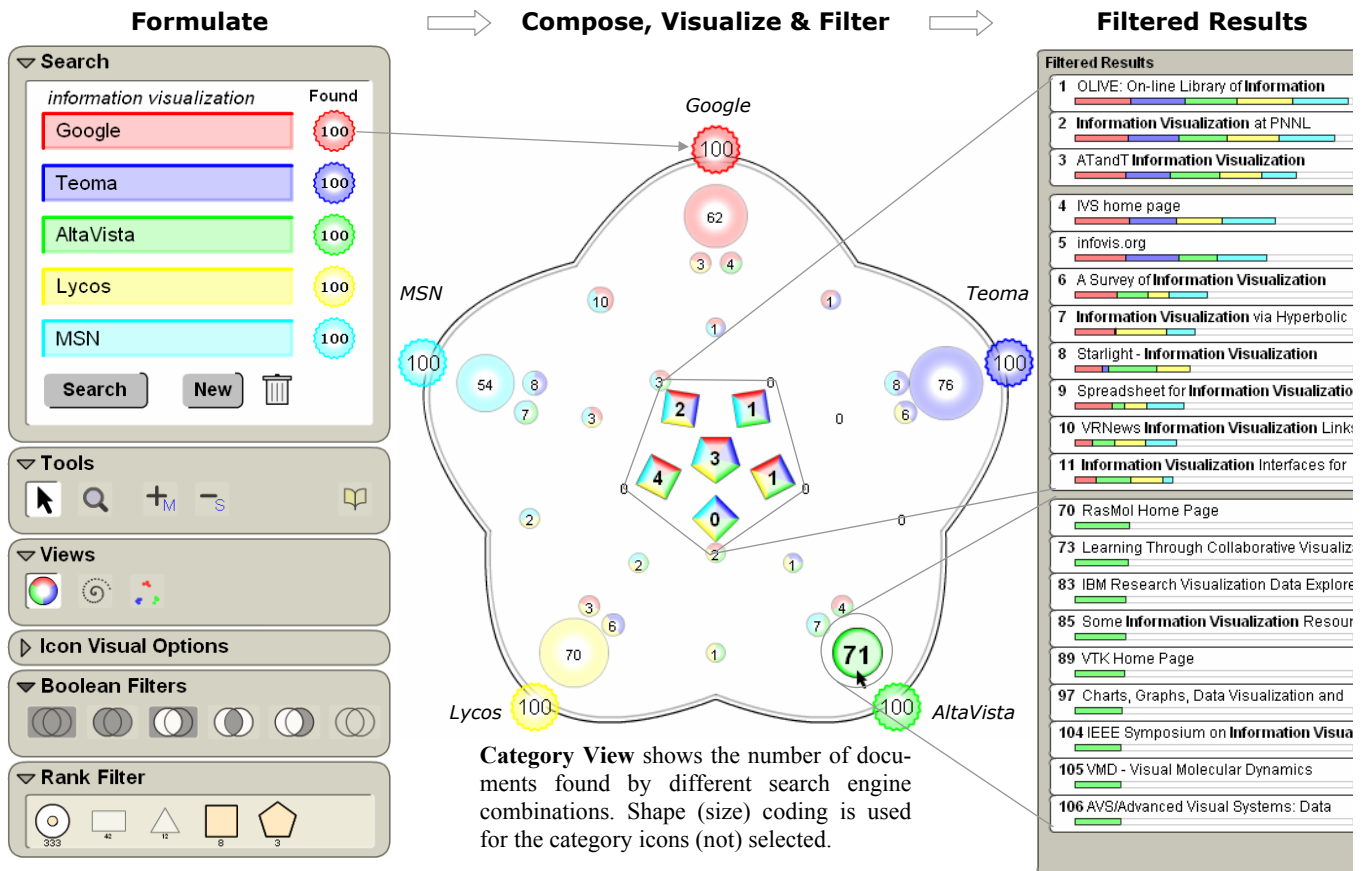
MetaCrystal's overview tools are designed to support flexible exploration, to enable advanced filtering operations and to guide users toward relevant information: 1) the *Category View* helps users decide where to begin their exploration and makes easy for them to identify relevant documents, because it groups all the documents found by the same combination of search engines. 2) The *Cluster Bulls-Eye* causes documents with high ranking scores to cluster toward the center. It also makes it easy for users to scan the top documents that are only retrieved by a specific engine. 3) The *RankSpiral* helps users identify the top documents found by a specific number of engines.

Category View

Modeled on the InfoCrystal layout (Spoerri 1999), 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 search engines, whose top results are "flowing" into the crystal and 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. Figure 2 shows the overlap between the top 100 documents found by Google, Teoma, AltaVista, Lycos and MSN, when searching for 'information visualization': three documents are found by all engines; four documents by Google, AltaVista, Lycos and MSN but not Teoma; and most of the documents are retrieved by a single engine. The Category View uses shape (size), color, proximity and orientation coding to visually organize the category icons and show how they are related to the input icons. Each search engine is assigned a unique color code, because color is a good choice for encoding categorical data (Ware 2000). Shape coding is used for a category icon if we want to emphasize the number of search engines it represents. Size coding is employed to emphasize the number of documents retrieved by a search engine combination.

Cluster Bulls-Eye

The *Cluster Bulls-Eye* enables users see how *all* the retrieved documents are related to the different search engines, because a document's position reflects the relative difference between its rankings by the different engines. Documents with similar rankings by the different engines are placed in close proximity (see Figure 2 and 4). Shape, color and orientation coding indicate which engines retrieved a document. Documents selected by user filtering operations are visually emphasized. In Figure 4, the documents found by at least three engines are emphasized.



Rank Filter is used to select only category icons, which represent documents retrieved by *at least four engines*. In addition, the **category icon** representing the documents found only by AltaVista is selected.

Documents found by many engines can be explored at the same time as the top documents found by only one engine.

Figure 1: shows how MetaCrystal can be used to specify the query ‘information visualization’, submit it to the search engines Google, Teoma, AltaVista, Lycos, MSN and use their top 100 results to compose a crystal that visualizes their precise overlap. The retrieved documents can be progressively filtered to create a result set to be explored in more detail. MetaCrystal enables users to gain insight into how to filter and combine the results returned by different search engines.

The Cluster Bulls-Eye tool 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 differences between a document’s rankings by the different engines. By default, the *total ranking score* of a document is calculated by adding the number of engines that retrieved it and the average of its different ranking scores. This results in documents that are retrieved by the same number of engines to cluster in the same concentric ring. 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 possible 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 use of size coding makes it easy for users to identify the top documents found by a specific number of search engines. In addition, a document’s

position is influenced by the input icons. Although not shown explicitly in this tool, the input icons act as “points of interest” that pull a document toward them based on the document’s rankings by the different engines. The Cluster Bulls-Eye combines a “Points-of-Interest” (POI) visualization with a “bull’s eye” mapping to ensure that users will always find documents with high total ranking scores toward its center. It overcomes a limitation of standard POI visualizations, such as VIBE (Olsen et al. 1993), where documents with different total ranking scores can be mapped to the same location and/or have the same distance from the center (Spoerri 2004).

In the Cluster Bulls-Eye, the distribution patterns of the documents, which are found by only one search engine, make it possible for users to infer if primarily documents with high rankings are also retrieved by the other search engines. Figure 4 (left) shows that many documents highly

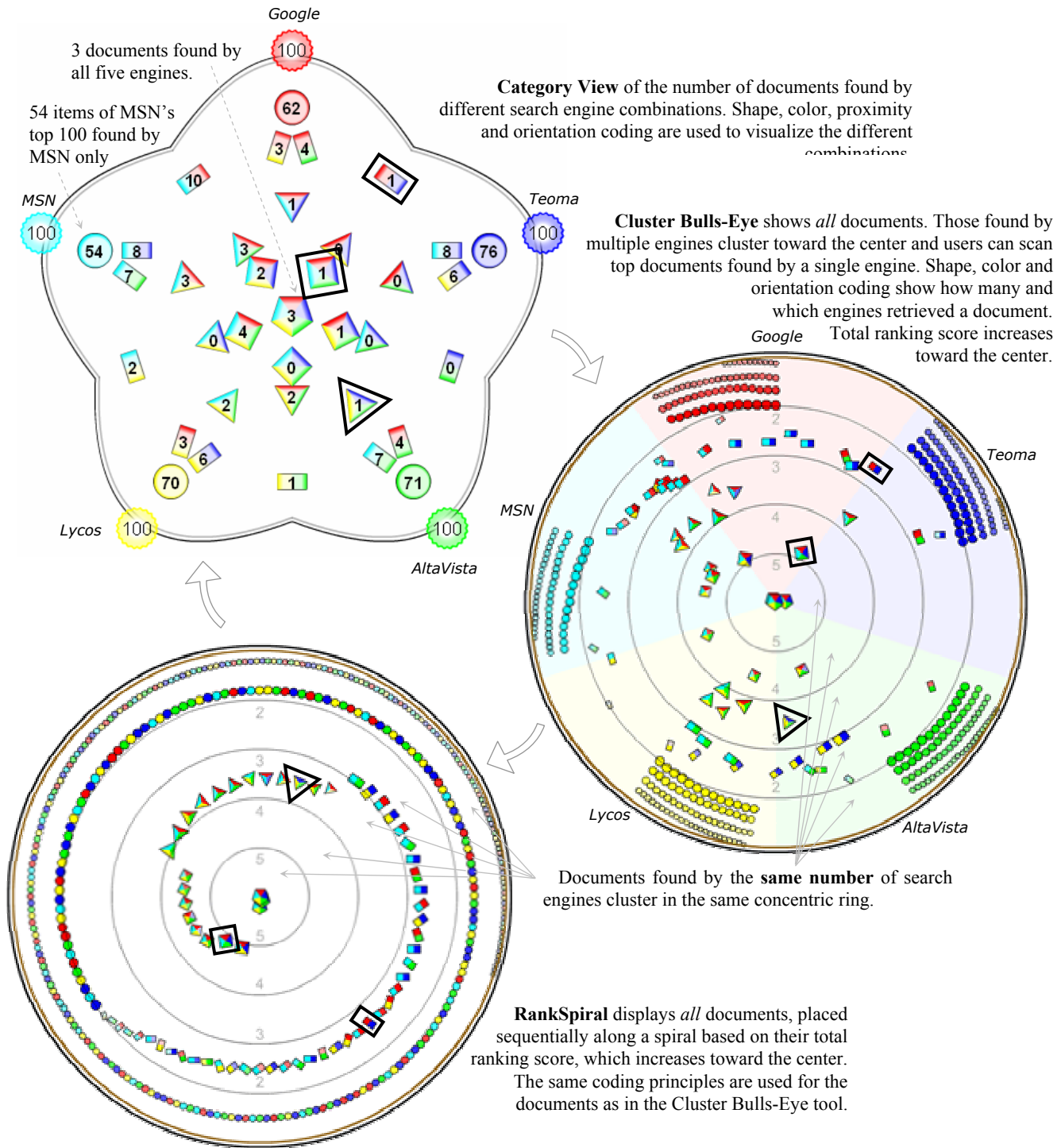


Figure 2: shows MetaCrystal’s linked overview tools: the *Category View*, *Cluster Bulls-Eye*, and *RankSpiral*, which provide users with complementary ways to explore the precise overlap between the top 100 documents found by Google, Teoma, AltaVista, Lycos and MSN, when searching for ‘information visualization’. Several documents are highlighted to show their respective location in the three linked overviews.

ranked by Google, Teoma or MSN are also retrieved by other engines, because the curves formed by the documents only found by Google, Teoma or MSN have a greater curvature than the same curves for AltaVista or Lycos.

RankSpiral

Search engines tend to display their results as ranked lists, which can only show a limited number of documents in a single screen. The RankSpiral overcomes this limitation by placing *all* documents sequentially along an expanding spiral (see Figure 2). A document's distance from the center is inversely related to its total ranking score. The score increases toward the center, ensuring that users can expect to find relevant documents in the center's vicinity. Consecutive documents are placed adjacent to each other so that they do not overlap, even if they have the same total ranking score. Shape, color and orientation coding are used to visualize which engines retrieved a document. Documents retrieved by the same number of engines are placed consecutively along the spiral and in the same concentric ring as in the Cluster Bulls-Eye tool. The RankSpiral makes it easy for users to identify the top documents found by a specific number of engines. The structure of the spiral can be used to solve the labeling problem (Spoerri 2004). For each document, the radial distance to the icon that has the same angle as the document in question can be computed. This distance can be used to display title (fragments) so that they not occlude any document icons. Thus, the RankSpiral makes it possible for users to rapidly scan large numbers of documents and their titles in a way that minimizes occlusions and maximizes information density.

Document Content Views

MetaCrystal supports two views to give users an immediate sense of a document's content. 1) The *Filtered Results* view displays a ranked list of the documents selected by user filtering operations (see Figure 1). It displays a document's overall rank, its title and a stacked bar chart, which shows how the different engines contributed to its total ranking score. This view enables users to simultaneously explore documents found by many engines and the top documents found by only one engine. 2) *Details-on-Demand* provides users with more detailed content information. Besides a document's title and overall rank, it displays a web page snippet and bar charts of the rankings by the different search engines, because quantitative data is best visualized using length coding (Ware 2000).

Visual Composition of Meta Queries

The MetaCrystal interface can be used as a visual search editor to compose meta queries. When users initiate a new search, MetaCrystal returns the top 100 results for up to five different search engines. Each search engine is assigned a unique color code and its result set is represented by a colored, star-shaped icon. Users can select a *result set icon* and drag it onto the "stage" to create an

instance of it. Figure 3 shows how users can visually merge the top documents retrieved by the five search engines Google, Teoma, AltaVista, Lycos and MSN to create crystals of increasing complexity. First, the result icon for Teoma is selected, dragged and placed over an instance of the Google search results to create a new crystal, which has both Google and Teoma as its inputs and visualizes how their result sets overlap. The category icon in the center of the MetaCrystal indicates that nine documents are retrieved by both Google and Teoma. When a user drags a crystal and places it on top of another crystal, the message "+merge" appears to inform the user that a new crystal will be created if the currently selected object is released. The "merge" operation is animated to help users track and assimilate the occurring change; the existing crystals fade away as the new combined crystal fades into view. Next in Figure 3, the AltaVista's result icon is merged with the crystal that has Google and Teoma as its inputs to create a MetaCrystal with three inputs. Now, only five documents are found by all three search engines, and we can see that Google and AltaVista have eighteen documents in common (by adding the numbers, five and thirteen, associated with the category icons related to both search engines). The number of category icons increases exponentially as more result sets are combined. If the search results of Google, Teoma, AltaVista and Lycos are compared, then four documents are retrieved by all four search engines. The most complex MetaCrystal shown in Figure 3 visualizes the overlap between five search engines. Only three documents are found by all five search engines. As we shift our focus away from the center, we can see that two documents are retrieved by Google, Teoma, Lycos and MSN, but not by AltaVista. Three is the highest number of documents found by a combination of three engines. We can also see that almost half of the documents found by MSN are also retrieved by some combination of the other search engines.

Users can reduce the number of MetaCrystal inputs by selecting an input icon, dragging it away from the crystal and releasing it. This "split" operation is animated, where the existing crystal fades away as these two new crystals fade into view: 1) a "simpler" crystal, which has one less input than the original crystal; and 2) a crystal, which represents the search results of the input being split off.

Filtering and Editing Meta Searches

MetaCrystal's tools enable users see the "big picture", while they apply filtering operations and focus on the selected search engine combinations and documents. Users can perform complex filtering operations visually and apply different weights to the search engines to create their own ranking functions. Users can control the degree of overlap between the different engines by modifying the URL directory depth used for matching documents or by changing the number of top documents used to compute the overlap between the different engines.

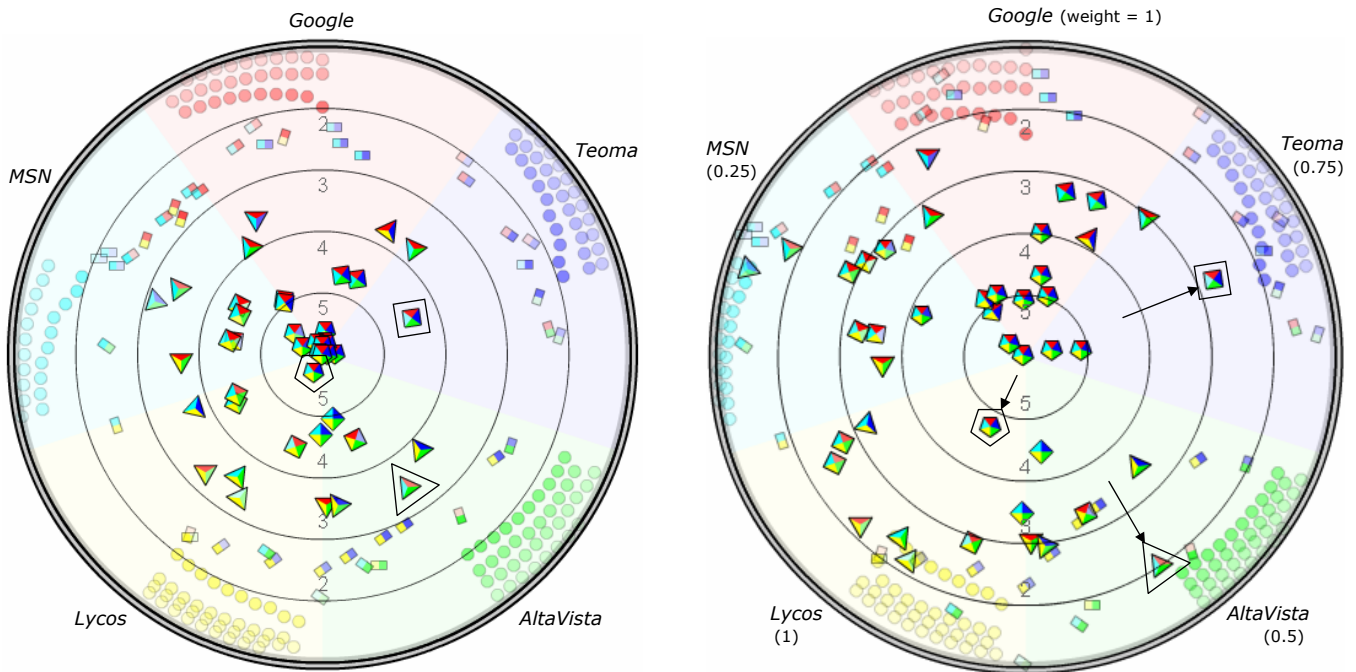


Figure 4: shows the *Cluster Bulls-Eye* for the case when domain names are used to compute the overlap between the different engines. (Left) This tool uses polar coordinates: the *radius* is equal to the number of engines that retrieved a document and the average of its different ranking scores; the *angle* reflects the relative differences between a document's rankings by the different engines. (Right) Shows how the documents cluster if the engines are assigned different weights (1, 0.75, 0.5, 1, 0.25) and the radius is now equal to the *weighted average* of a document's rankings. In both figures, the documents found by at least three engines are selected and *no* size coding is used. Three documents are highlighted to show how their locations change when the weights are applied.

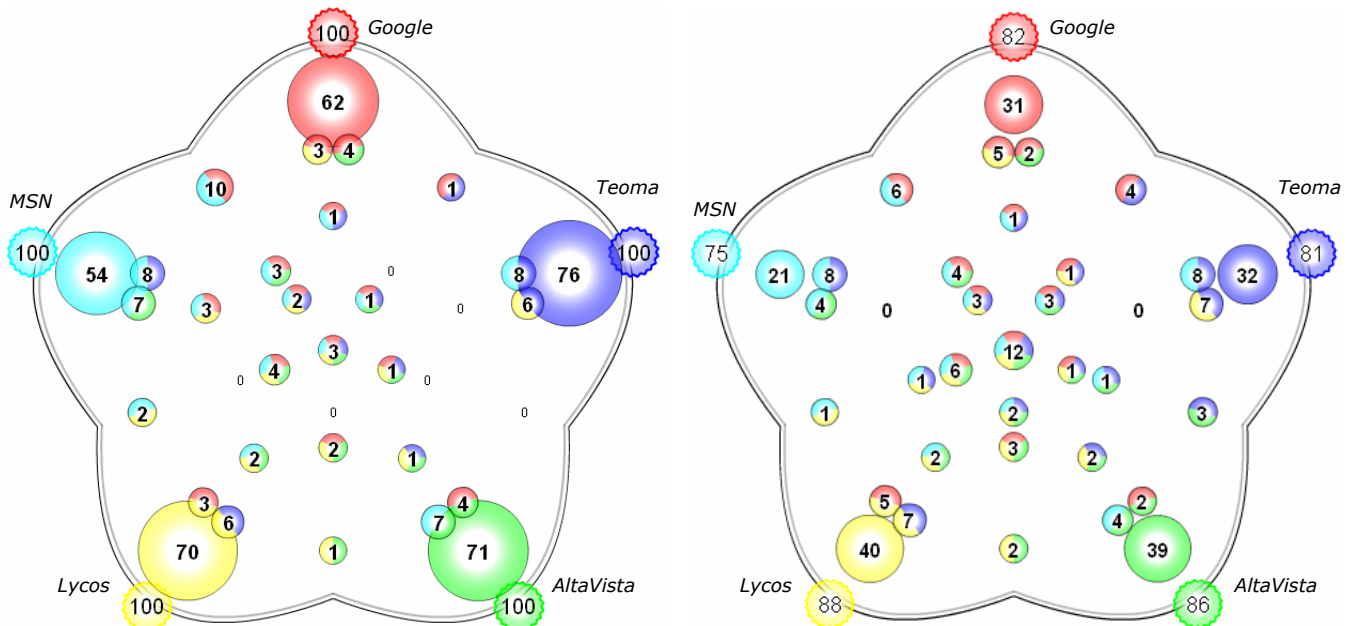


Figure 5: (Left) Shows the *Category View* if the full URL is used for matching the top 100 documents found by five different search engines. Size coding is used for the category icons. (Right) Shows how the overlap increases when an URL equal only to the domain name is used for matching purposes. The shortened URL directory depth reduces the number of unique documents in the result sets, as indicated by the input icons.

Filtering

Users can create a short list of potentially relevant documents by: a) requiring category icons and documents to be retrieved by a specific number of engines (see Figure 1 and the Rank Filter); b) specifying Boolean requirements, since the category icons represent all possible Boolean queries in disjunctive normal form (Spoerri 1999); c) applying a threshold based on the total ranking score; d) selecting individual category icons or documents by clicking on them.

Apply Search Engine Weights

Users can apply different weights to the engines to create their own ranking functions. The total ranking score of a document is now equal to the *weighted average* of its rankings by the different engines. This change in how the total ranking score is calculated impacts the linked tools as follows: 1) in the Cluster Bulls-Eye, documents retrieved by different numbers of engines can now be placed in close spatial proximity. Figure 4 shows the resulting Cluster Bulls-Eye when Google, Teoma, AltaVista, Lycos and MSN are assigned the weights of 1, 0.75, 0.5, 1, and 0.25, respectively. 2) In the RankSpiral, documents found by different numbers of engines will now be interspersed along the spiral. However, the top documents found by a specific number of engines can visually “pop out”, because the human visual system can use size and shape differences to pre-attentively segment a display (Ware 2000).

Control Degree of Overlap

Users need ways to control the number of documents found by multiple engines, because such documents are more likely to be relevant (Saracevic & Kantor 1988, Foltz & Dumais 1992). MetaCrystal enables users to shorten the URL directory depth used for matching documents to increase the number of documents or partial URLs found by more than one engine. However, this reduces the number of unique elements in the result sets being compared. Figure 5 (right) shows how the overlap increases when an URL equal only to the domain name is used for matching purposes. Twelve domains are now retrieved by all engines, and the percentage of unique domains found by a search single engine is reduced.

MetaCrystal also makes it possible for user to specify the number of top documents to be used to compute the overlap between the different engines. Increasing the number of top documents to be compared increases the probability that more documents are found by multiple engines. Decreasing the number of top documents increases the probability that the documents retrieved by several engines are relevant. This suggests that users start by comparing the top 10 documents and explore the ones found by more than one search engine. Next, they can increase the number of top documents to be compared and observe how the overlap changes. Figure 6 shows how the overlap changes if the top 10, 25, 50 and 100 documents, respectively, are compared.

Discussion & Future Research

MetaCrystal has been implemented in Flash using its ActionScript programming language. This has the advantage that it can be deployed using a Web browser and its file size is small. The goal of future MetaCrystal versions is to provide users with a versatile search interface that enables them to maintain a search history so that previous searches can be repurposed and combined in an interactive fashion. MetaCrystal currently receives its input data in XML. The current input / output interface will be formalized so that new databases or search engines can be easily added and supported. Users can control the number of top documents that are retrieved by a search engine. The current maximum number is 100, which is a reasonable choice since search engines make every effort to list the “most relevant” documents as high as possible in their ranked lists and users very rarely explore more than 100 documents (Silverstein et al. 1998).

A preliminary usability test has been conducted, where 15 graduate students interacted with MetaCrystal’s different tools. They rated the *Category View* as most effective, especially when shape coding was used. When size coding was used for the category icons, the students observed that the color contrast needed to be increased to make the colored slices easier to perceive. The students had some difficulty understanding the conceptual difference between the *Category View*, which displays groupings of documents, and the *Cluster Bulls-Eye* and *RankSpiral*, which display individual documents. The latter two tools 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. 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. The students also wanted to get an immediate sense of a document’s content to judge its relevance, regardless of the tool selected. Their feedback inspired the creation of improved *Filtered Results* and *Details-on-Demand* displays.

The next step is to conduct a series of formal evaluations of the MetaCrystal interface. As mentioned, its design is guided by the fact that documents found by multiple search engines are more likely to be relevant. User studies will be conducted to test if guiding users’ attention toward documents found by multiple engines will lead to improved search performance and user satisfaction. More generally, it will be investigated if MetaCrystal and the flexibility it provides lead to greater user satisfaction and help users find relevant information more easily. In particular, it will be tested if the different tools support their intended tasks. It will also be investigated how best to animate the display

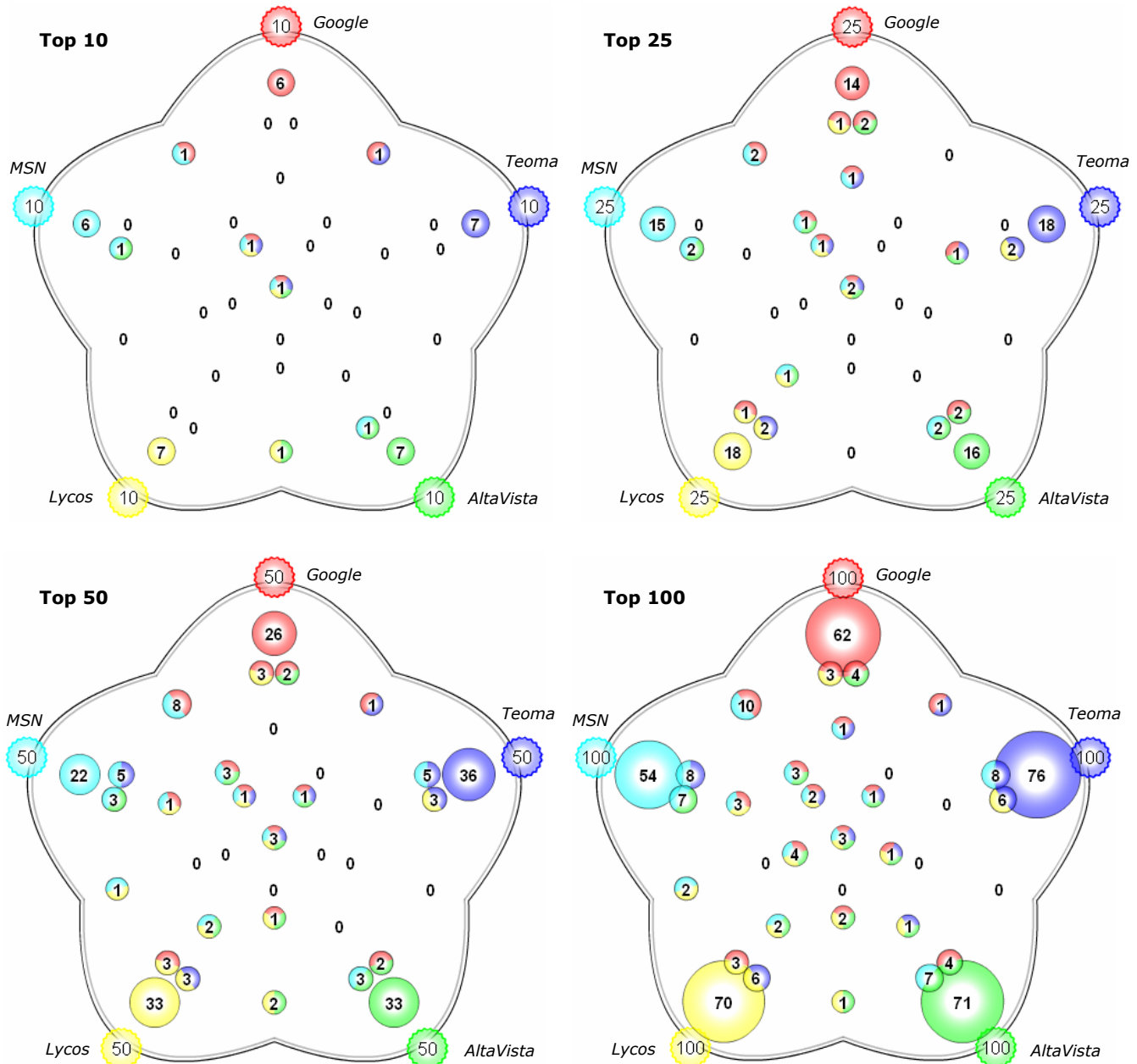


Figure 6 shows how the overlap changes in the Category View when the top 10, 25, 50 or 100 documents are compared. Size coding is used for the category icons to visually emphasize the number of documents found by the different search engine combinations.

changes when users modify search engine weights, the number of top documents to compare or the URL depth used for matching purposes so that users can gain insights from the motion cues. Finally, it will be studied how relevance feedback can be implemented in the context of the MetaCrystal interface to improve the user experience.

Summary

MetaCrystal addresses the difficult problem of the effective fusion of different search engine results by enabling users to interact with a visual search editor to compose and refine meta searches. Research has shown that web search engines index different parts and less than 20%

of the Internet (Lawrence & Giles 1999). Further, documents found by multiple retrieval methods are more likely to be relevant (Saracevic & Kantor 1988, Foltz & Dumais 1992). MetaCrystal helps users take advantage of the different strengths of major Internet search engines, because it makes easy for users to identify documents found by multiple engines and at the same time scan the top documents retrieved by a single engine. It offers users complementary overview tools to support them in the meta search process. Its tools are designed to support flexible exploration, to enable advanced filtering operations and to guide users toward relevant information. The tools all use a “bull’s eye” layout so that users can expect to find relevant

documents toward the center of the displays. The *Category View* groups and displays the number of documents retrieved by different search engine combinations. The *Cluster Bulls-Eye* shows how the documents are related to the search engines, causing documents with high ranking scores to cluster toward the center. The *RankSpiral* places all documents along a spiral based on their total ranking scores, making it easy for users to identify the top documents found by a specific number of search engines. *Details-on-Demand* give users an immediate sense of a document's content and how the different search engines contributed to its total ranking score. In terms of workflow, users formulate a query, submit it to several search engines and then "drag & merge" the top 100 results to compose crystals that visualize the precise overlap between the different search engines. The retrieved documents can be progressively filtered to create a result set to be explored in more detail. MetaCrystal can be used to apply different weights to the search engines so that users can create their own ranking functions. Users can control the degree of overlap between the different engines by modifying the URL directory depth used for matching documents or by changing the number of top documents that are being compared. A preliminary usability test was conducted, where 15 graduate students provided useful feedback regarding MetaCrystal's different tools. The next step is to conduct a series of formal evaluations of the MetaCrystal interface and its extensive functionality.

COLOR FIGURES

Color plays a crucial role in the MetaCrystal visualizations and the color figures are available online at: <http://www.scils.rutgers.edu/~aspoerri/Publications.htm>.

REFERENCES

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

Chen H., Fan H., Chau M. and Zeng D. MetaSpider: Meta-Searching and Categorization on the Web. *Journal of the American Society for Information Science*, Volume 52, Number 13 (2001), 1134 - 1147.

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

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.

Groxis – www.groxis.com

Havre, S., Hetzler, E., Perrine K., Jurrus E., and Miller N. (2001) *Interactive Visualization of Multiple Query Results*. Proceedings of the IEEE Information Visualization Symposium 2001 (InfoVis 2001), San Diego, CA. October 22-23, 2001

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

Ingwersen, P. (2002). Cognitive perspectives of document representation. Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4), Seattle, July, 2002. Colorado: Libraries Unlimited, 2002, 285-300.

Kartoo – www.kartoo.com

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. *Journal of the American Society for Information Science*. 39, 3, 197-216.

Silverstein, C., Henzinger, M., Marais, J. & Moricz, M. (1998). Analysis of a very large Alta Vista query log. Technical Report 1998-014, COMPAQ Systems Research Center, Palo Alto, Ca, USA.

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.

Spoerri, A. (2004) Cluster Bulls-Eye and RankSpiral: Enhancing Points-of-Interest and Search Results Visualizations. In preparation.

Vivisimo – www.vivisimo.com

Ware C. (2000) *Information Visualization: Perception for Design*. San Francisco: Morgan Kaufmann.