*Appears in **ICIV'04**, London, England, July 14 - 16, 2004*

How Visual Query Tools Can Support Users Searching the Internet

Anselm Spoerri
Department of Library and Information Science
SCILS, Rutgers University
4 Huntington Street, New Brunswick, NJ 08901, USA
aspoerri@scils.rutgers.edu

Abstract

Visual tools have been developed to help users formulate advanced Internet queries. Users can control the query coordination process at the level of the individual search terms or at the level of the search engines to be consulted. This paper addresses the usability issue of the level at which visual query tools can provide the greatest benefit to users searching the Internet. MetaCrystal and its Category View and Cluster Bulls-Eye tools are used to visualize the degree of overlap between the results returned by different query formulations or search engines. It is shown that certain types of similarity relationships between ranked lists give rise to unique visual patterns in the Cluster Bulls-Eye. The MetaCrystal toolset is used to show why visual query tools may provide greater benefits to users who want to coordinate meta searches rather than individual search terms.

1. Introduction

Users searching the Internet are confronted with several problems such as: which search engine(s) to select (database selection problem), which terms to use (vocabulary problem), which query operators to apply (query coordination problem), how to explore the many retrieved documents (information overload problem) and how to modify the query to find more relevant documents (query modification problem). Many visual tools have been developed to help users overcome the specific problems they encounter in the search process [11]. None of these tools have been adopted by standard Internet search engines. It is only in the domain of meta searching that some commercial systems are employing visualization [9, 13]. This paper addresses the usability issue of whether visual query tools can provide their greatest benefit to users who want to control the coordination process of their Internet searches at the level of the search terms or at the level of the search engines.

1.1. User behavior and needs

Users tend to create short queries when searching the Internet [16]. They rarely formulate advanced queries, using Boolean or proximity operators in 10% of their queries [16]. It is well documented that users find it difficult to formulate Boolean queries [2, 3]. To overcome this problem, visual tools have been developed to help users specify and coordinate Boolean queries [1, 12, 17, 20]. Yet, research has shown that the use of most query operators in short Internet queries had no significant impact on the effectiveness of the search results [6]. This suggests that visual query tools have minimal benefits to offer for the majority of Internet searches conducted.

Users employ meta search engines because individual search engines only index 20% of the Internet [14] and thus return different documents for the same query. Meta search engines address this limitation by combining the results returned by different engines. The automatic and effective fusion of different search engine results can be difficult [4]. While meta search engines exist that visually organize the retrieved documents, [5, 9 10, 13, 18] no meta search interface 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 search engines are more likely to be relevant [7, 15]; b) it is difficult to predict the quality of coverage for single engines, which tend to index less than 20% of the Internet [14]; c) some engines are more effective than others depending on the search domain [8]; and d) users may prefer or trust some search engines more than others. This suggests that active user involvement could make a difference when it comes to coordinating the fusion of the different search results. While documents found by multiple engines are more likely to be relevant, users may also want to examine the documents only found by their preferred engines. This range of requirements can be addressed by MetaCrystal.

This paper is organized as follows: section 2 briefly reviews related work. Section 3 describes the MetaCrystal toolset. In section 4, the Category View and Cluster

Bulls-Eye tools are used to visualize and support the finding that advanced formulations of short Internet queries tend to lead to very similar search results [6]. Section 5 illustrates how MetaCrystal can support users in the meta search process and enables them to control how the different search engines results are coordinated.

2. Related work

Several meta search engines have been developed that use visualization techniques. Vivísimo [18], Grokker [9], MetaSpider [5] and Sparkler [10] address the information overload problem. Vivisimo organizes the retrieved documents using the familiar hierarchical folders metaphor. Grokker uses nested circular or rectangular shapes to visualize a hierarchical grouping of the search results. MetaSpider uses a self-organizing 2-D map approach to classify and display the retrieved documents. Sparkler 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. Kartoo [13] addresses the query reformulation problem. It creates a 2-D map of the highest ranked documents and displays the key terms that can be added or subtracted to modify the current query. 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. Further, none of them enable users to control how the search results by the different engines are combined.

3. MetaCrystal toolset

MetaCrystal consists of several linked tools that enable users to compare and combine the search results returned by different query formulations or different search engines processing the same query. Its design is guided by the fact that documents found by multiple search methods are more likely to be relevant [7, 15]. The *Category View* displays the precise overlap between the top result sets returned by different queries or search engines (see Figures 1 and 3). The *Cluster Bulls-Eye* tool displays how *all* the found documents are related to the different queries or engines being compared (see Figure 1). It clusters documents retrieved by multiple search methods toward its center and at the same time helps users scan the top documents found by a single method. This tool can also be used to visualize the degree of similarity between different ranked lists.

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, AltaVista, Lycos and MSN, when searching for 'information visualization'. Modeled on the InfoCrystal layout [17], 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 100 results are compared to compute the contents of the category icons. Each search engine is assigned a unique color code, because color is a good choice for encoding categorical data [19]. The category 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; size coding is employed to emphasize the number of documents retrieved by a combination of search engines.

3.2. Cluster Bulls-Eye

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 (see Figure 1). Shape, color and orientation coding indicate which search engines retrieved a document. 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 ratio of a document's rankings by the different engines. 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. 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 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 ensures that users will always find documents with high total ranking scores toward the center of the display. It also makes it easy for users to scan the top documents that are only retrieved by a single engine.

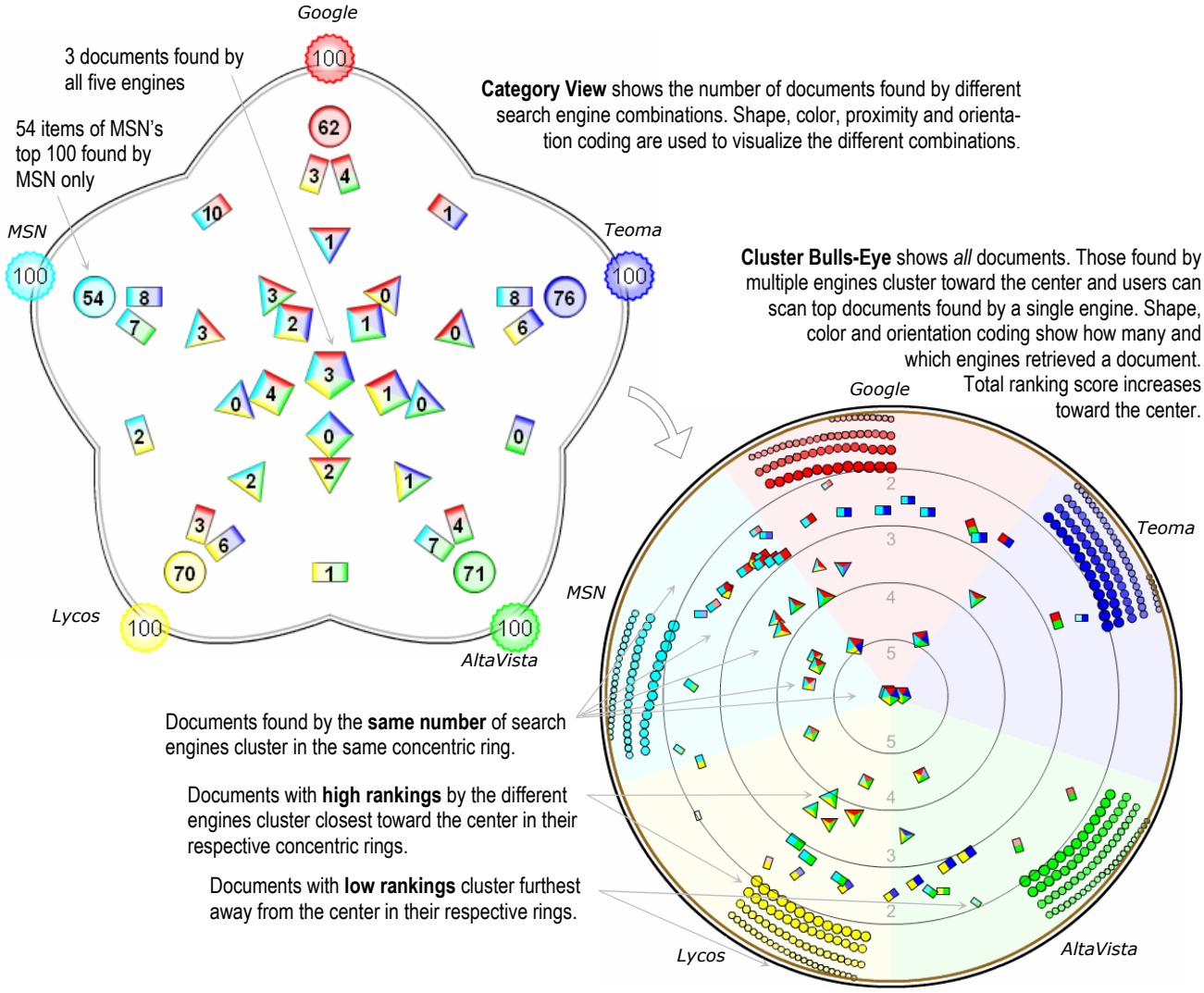

Figure 1: The *Category View* and *Cluster Bulls-Eye* provide users with complementary ways to explore the precise overlap between the top 100 documents found by Google, Teoma, AltaVista, Lycos and MSN, when searching for 'information visualization'.

4. Visualizing similarity between searches

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. If a document has the same ranking in all of the results being compared, then its angle will be 90 degrees (see Figure 2 [i]). If the rankings of 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 line separating the lightly colored areas 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.

The Cluster Bulls-Eye and Category View can be used to visualize and support the finding that the use of most query operators in short Internet queries leads to very similar results [6], 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 queries that use "no operators", a "Boolean AND" or an "exact phrase" constraints are submitted to Google, Teoma and AltaVista.

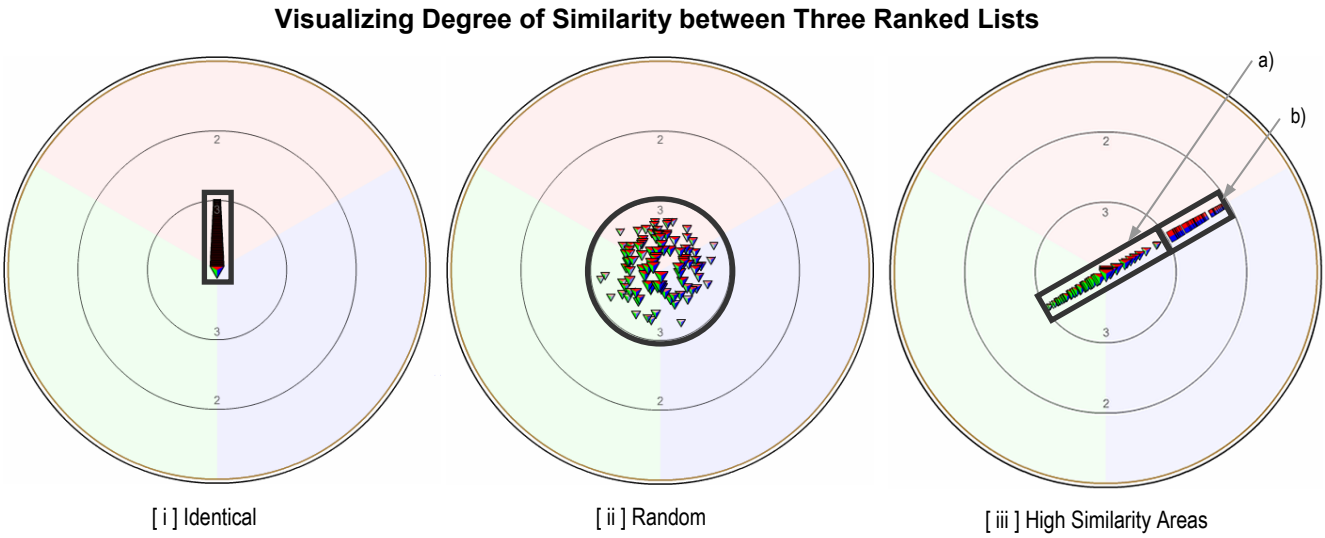

Figure 2: shows the "visual signatures" and how documents will cluster in the *Cluster Bulls-Eye* if the three ranked lists being compared: [i] are identical; [ii] contain the same documents, but their ranking orders are randomized. [iii] a) shows the area where the documents found by all three queries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

For the search engines Google, Teoma and AltaVista, Figure 3 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 degree of overlap between the queries, which use "no operator", "Boolean AND" or "exact phrase" constraint, is greatest for Google and the smallest for AltaVista. Figure 3 illustrates why visual query tools will provide minimal benefits to users who formulate short queries, which represent the majority of Internet searches conducted.

5. Supporting meta search process

Internet search engines use different retrieval methods and index different parts of the Internet [14]. Key challenges users face when conducting a meta search is deciding which engines to select and how to take advantage of and coordinate their respective strengths. Research has shown that documents found by multiple retrieval methods are more likely to be relevant [7, 15]. This suggests that an effective meta search interface needs to help users identify documents found by several search engines. However, the number of documents retrieved by more than one Internet search engine tends to be small [14] (see Figure 1). Users need ways to increase and control the number of documents found by multiple engines. Furthermore, some of the top documents found by a single search engine will be relevant, but users don't want to have to sift through a long list of documents to find them [11]. Thus, a meta search interface needs to help users scan and select the top documents only retrieved by a specific engine, especially since users may prefer some engines more than others.

MetaCrystal helps users gain insight into how to coordinate and filter the top documents retrieved by different search engines. Implemented in Flash using ActionScript, its direct manipulation interface enables users to iteratively create crystals that show the precise overlap between up to five engines. MetaCrystal supports flexible exploration, enables advanced filtering operations and guides users toward relevant information: the *Category View* groups all the documents found by the same combination of engines and displays the number of documents retrieved by different engine combinations. The *Cluster Bulls-Eye* tool helps users identify documents found by multiple search engines and at the same time scan the top documents retrieved by a single engine. Users can perform advanced filtering operations visually to create a short list of potentially relevant documents by: a) requiring documents to be retrieved by a specific number of engines; b) specifying Boolean requirements by selecting specific category icons, which represent all possible Boolean queries in disjunctive normal form [17]; c) assigning different weights to the search engines to create their own ranking functions; and d) controlling the degree of overlap between the engines by modifying the URL directory depth used to match documents or change the number of top documents compared. MetaCrystal also supports "details on demand" to give users an immediate sense of a document's content, such as its title, total ranking score, content snippet and bar charts of the rankings by the different search engines.

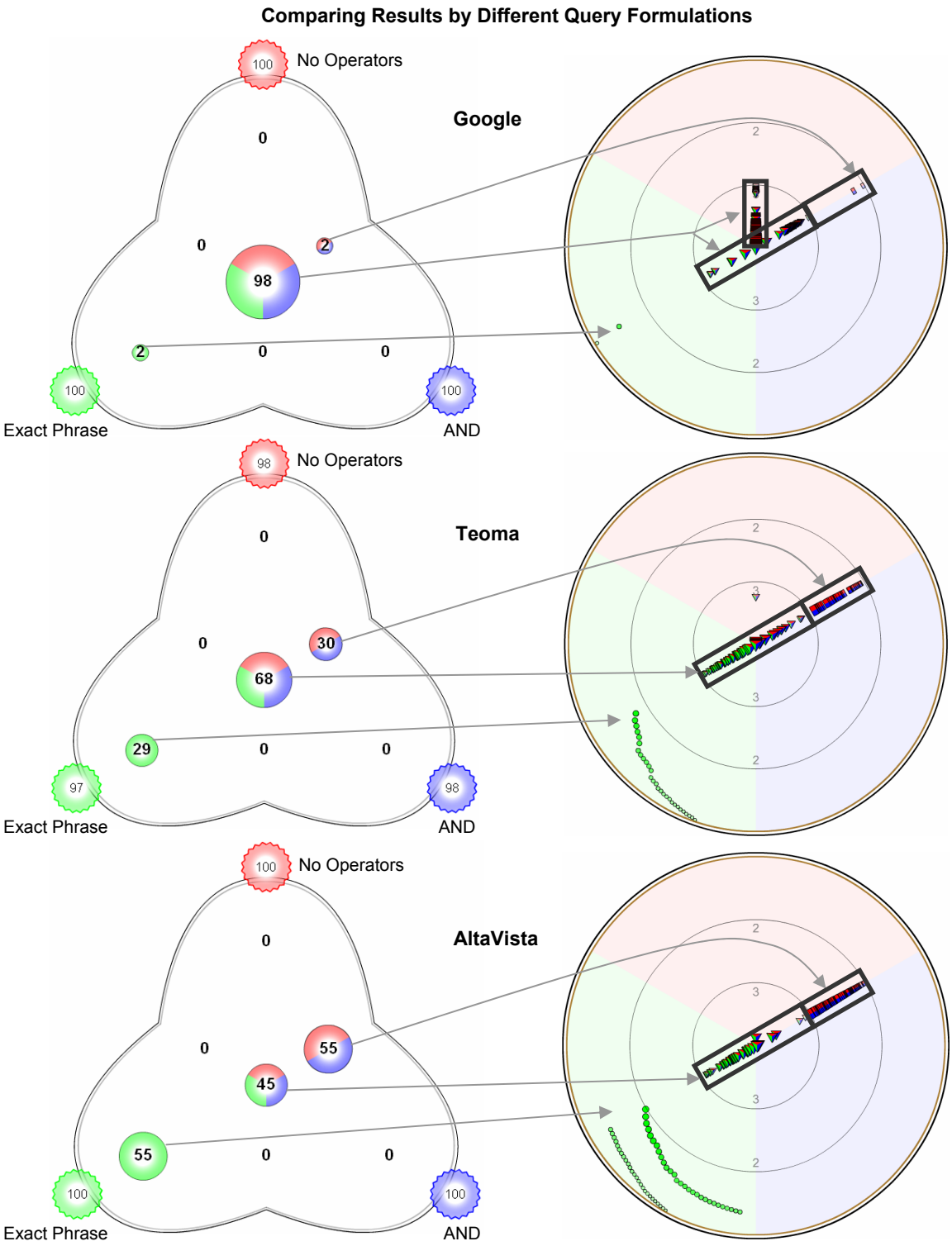

Figure 3: shows the precise overlap between the top documents found if we search for 'information visualization' and compare queries that employ "no operators", a "Boolean AND" or an "exact phrase" constraint, using Google, Teoma and AltaVista, respectively. The left column contains the *Category Views*, which use size coding for their category icons. These views show 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 that are related to only one of these queries are empty. The right column contains the *Cluster Bulls-Eye* displays, which show that the ranking order for the documents found by more than one query is identical or very similar.

6. Conclusions

This paper described and used the MetaCrystal toolset to address the usability issue of whether visual query tools provide their greatest benefit to users who want to control the coordination process of Internet searches at the level of the search terms or the search engines.

It was shown that certain similarity relationships between different ranked lists give rise to unique visual patterns in the Cluster Bulls-Eye tool. The Category View was used to show that there is a great deal of overlap between the documents found by different formulations of the same query, but little overlap when the same query is submitted to different Internet search engines.

MetaCrystal was used to visualize and support the finding that the use of query operators in short Internet queries leads to very similar results [6], both in terms of the documents being retrieved and their respective rankings. Thus, visual query tools have minimal benefits to offer to users who formulate short queries, which represent the majority of Internet searches conducted.

This paper also showed how MetaCrystal can be used as visual meta query tool. It enables users to coordinate how the documents retrieved by different search engines are combined. Research has shown that documents found by multiple retrieval methods are more likely to be relevant [7, 15]. MetaCrystal helps users identify how many and which documents are found by multiple search engines and at the same time scan the top documents retrieved by a single engine. It enables users to gain insight into how to solve the problem of the effective fusion of different search results. Users can visually coordinate and filter the top documents retrieved by different engines to create a short list of potentially relevant documents.

In conclusion, MetaCrystal was used to show *why* visual query tools may provide greater benefits to users who want to coordinate meta searches rather than individual search terms. The next step is to conduct a formal user study to test *if and how* MetaCrystal does support users in the meta search process.

7. References

[1] Anick, P.; Brennan, J.; Flynn, R.; Hanssen, D.; Alvey, B. & Robbins, J. (1990). A Direct Manipulation Interface for Boolean Information Retrieval via Natural Language Query. Proc. ACM SIGIR '90.

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

[3] Borgman, C. (1989). All Users of Information Retrieval Systems Are Not Created Equal: An Exploration of Individual Differences. Information Processing & Management, 25 (3).

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

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

[6] 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.

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

[8] 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.

[9] Grokker – www.groxis.com

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

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

[12] Jones, S. (1998). Graphical Query Specification and Dynamic Result Previews for a Digital Library. Proc. of ACM UIST 1998:

[13] Kartoo – www.kartoo.com

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

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

[16] 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 .

[17] 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.

[18] Vivisimo – www.vivisimo.com

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

[20] Young, D. & Shneiderman, B. (1993). A Graphical Filter / Flow Representation of Boolean Queries: A Prototype Implementation and Evaluation. JASIS, 44 (6).