Desktop Search Engine Visualisation and Evaluation

Schubert Foo and Douglas Hendry

Division of Information Studies, School of Communication and Information
Nanyang Technological University, Singapore 637718
{assfoo,hend0007}@ntu.edu.sg

Abstract. This work investigates the potential of applying a suite of visualisation for query processing in a desktop search environment. While each of these visualizations may not be totally new on its own, we have attempted to add value to each one by endowing it with useful features, and to seamless integrate them to allow easy switching of views, thereby providing the novelty in this work to create a potentially useful means to process search results and carry out query refinements and exploration. These visualisations include a List View, Tree View, Map View, Bubble View, Tile View and Cloud View. A first evaluation was undertaken by 94 M.Sc. participants to gauge the system's potential usefulness and to detect usability issues with its interface and graphical presentations. The evaluation results were encouraging and showed that these views to be both effective and useful on the whole, and support the research premise that a combination of integrated visualisations will result in a more effective search tool.

Keywords: Query result processing, query reformulation, tree view, map view, bubble view, tile view, cloud view, evaluation, search engine, user interface.

1 Introduction

Internet search and information seeking is predominated by large search engines such as Google, Yahoo, MSN and AOL (more commonly referred to as GYMA) where textual search results are largely displayed as text-based URLs and a few lines of document contents to provide the context to the URLs. These established search engines' strength lies mainly in their very sophisticated search and ranking algorithms but lack somewhat in innovation in terms of their interfaces which are traditional text-based. However, we begin to see emergence of online search engines that have started to offer more graphical interfaces to assist users such as Grokker and Ujiko.

In contrast to the developments of online tools to search the Web, desktop search tools for searching increasing large volumes of documents held on local computers hard drives, have been slower to develop. While search tools are now being incorporated into the latest desktop operating systems, such as Spotlight in Apple's OSX and Vista in Microsoft Windows, their search result visualisations have taken the traditional display of textual output. With result lists becoming increasingly longer, the challenge is to improve efficiency and effectiveness of search and result selection. In this respect, the metaphor that "a picture paints a thousand words" neatly encapsulates

the concept that well presented graphical views can convey large amounts of complex information in a simple and easy to understand manner. It is therefore not surprising that graphical visualisations have been employed in search engines to assist users.

This work reports on a number of visualisations that were developed for desktop searching. While each of the individual visualisation might not be new by itself, we believe that the seamless integration of these views and value-added functionality in them are novel to assist in the results review, selection and query refinement. A set of first evaluations was carried out to determine the usefulness of the developed views and to identify areas for further development. The work therefore aims to contribute towards desktop searching research through the provision of a suite of integrated views, eliciting and confirming the characteristics (applicability, usefulness, limitations) of the views, supported by evaluation results.

2   Related Work

In order to overcome the limitations of text-based result lists, a number of researchers have developed a variety of 2D and 3D graphical visualisations in order to allow the user to explore and understand the results of their query. PFNET was an early to attempt to introduce visualisation through a network based tool that used a thesaurus to create associated document networks based on the query results [1]. Users could browse results as a network and view individual documents.

Envision [2] is a tool that was built as an alternative to the query-document similarity ranking. The tool allowed the searcher to graphically display the search results from a bibliographic IRS using an X-Y graph. Different attributes such as author, year, document type, number of citations and relevance can be plotted on the X and Y-axes and the result documents are represented as icons in the main plot area.

GRIDL uses a grid-based approach to present large volumes of digital library search results using categorical and hierarchical axes (hieraxes) to simplify the display [3]. The documents in a search result are clustered together based on metadata. These hierarchical categories are then used as the axes for a grid-based plot that displays the relative number of documents at the intersection of each axis attributes. Users can then drill down these hierarchies to explore the search results. Other examples of visualisation tools include the Visual Information Browsing Environment (VIBE) that presents search results as a 2D map [4], Periscope which is a system for adaptive 3D visualisation of web search results [5], and YAVI that uses a 3D information space to display a vector-space model output of search results [6].

To support query reformulation, visualisations usually present terms that are related to the query terms in use. These related terms often come from a controlled dictionary, thesaurus or other system metadata. The user can review these new terms and use them to modify their query. One such tool is the AquaBrowser Library [7] which shows a visual word cloud that suggests words similar or related to the users query terms.

Whilst some evaluation studies have reported mixed results [8] many have found positive support that the visualisations have aided user performance [9, 10]. Even in cases where performance has not improved users often report better satisfaction with tools incorporating visualizations [11]. Visualisations seem to be particularly effective

where the complexity of the task and volumes of data are at their highest [12, 13]. They also seem to work well when they are kept as simple as possible [14].

3   System Design

DSE is a Java-based lightweight desktop search engine developed to index and search content on a desktop computer. The indexing and searching sub-systems were designed based on traditional IRS principles and incorporates stop word removal, stemming and is based on Boolean logic. DSE uses a plug-in architecture for supporting different file types that include text, HTML, RTF, XML, MS Word, MS Excel and PDF.

The design of the user interface was based on the following research premises:

- Visualisations can assist users to search for documents [10, 13]
- Different visualisations can be used to support different elements of the searching process (results review and query reformulation)
- Different graphical techniques can be used to assist users to visualise different kinds of information
- Visualisations work best when they are kept simple [14].

The search engine GUI has a plug-in view architecture that allows different views to be created independent of the searching mechanism. Six views were constructed for use and evaluation: List View, Tree View, Map View, Bubble View, Tile View and Cloud View. These will be elaborated later in the paper.

4   Evaluation Methodology

After development, a first evaluation was carried out to determine the usefulness of the views, particularly to find out if the availability of such views was useful and which of them would be worth developing further. As such, the work reported here is intended as a proof of concept and to have a first gauge of usability and usefulness of the views.

The evaluation was carried out through a user survey based a questionnaire comprising 51 questions to gather user opinion about the visualisations and their usefulness. The evaluation was split into five tasks – each required the user to perform a search in support of a given information need and the participants worked through these in sequence. For each task they were given five minutes to interact with an interface to find the most relevant documents to satisfy the information need. Following this, they were asked to complete a series of questions pertaining to the interface used.

Prior to the evaluation, a handout describing the search engine and views was provided to the evaluators, followed by a briefing and demonstration. Evaluators were also invited to download the search engine to familiarise themselves with it and the various views. Each evaluator had a minimum of one week of familiarity prior to the evaluation, and each spent 45 to 60 minutes in completing the evaluation.

Since the purpose of this evaluation was to review the visualisation aspects of the search tool, it was important to remove other factors that could influence user responses. Therefore, for each task, the set of search terms to be used was pre-specified.

We used a small set of 30 documents for evaluation since DSE had the functionality to show the top *n* documents in the display, thereby restricting the volume of results displayed even if a larger collection was indexed and searched. We expect future evaluations to be more comprehensive and extended to test for scalability to larger document collections, and to test how the visualisations will lose their effectiveness and usefulness when the displays are increasingly populated with more search results.

The documents used in the evaluation covered topics on information retrieval systems, the World Wide Web, indexing and programming languages. A portion of the ERIC thesaurus was used to create a hierarchical folder structure and these documents were then stored into the folders based upon their categorisation in ERIC.

The post-graduate students of the Nanyang Technological University M.Sc. Information Studies class of 2006/07 participated in the evaluation. There were 94 participants – 57% female, 43% were male. 98% of participants rated themselves having Satisfactory (or better) computer skills, while 76% rated themselves having Satisfactory (or better) searching skills.

5   Search Engine Views and Evaluation Results

This section presents the various views constructed in DSE for presenting search results from the queries. It highlights the characteristics of each display and reports on the salient findings of the evaluation.

5.1   List View

The List View (Fig 1) is the classic search results view. It contains a list of files that, based on Boolean logic, match the users query. Each file name is shown along with the number of "hits" from the query. A hit is defined as one occurrence of one query term in the file contents. The files are listed in descending order of the total number of hits.
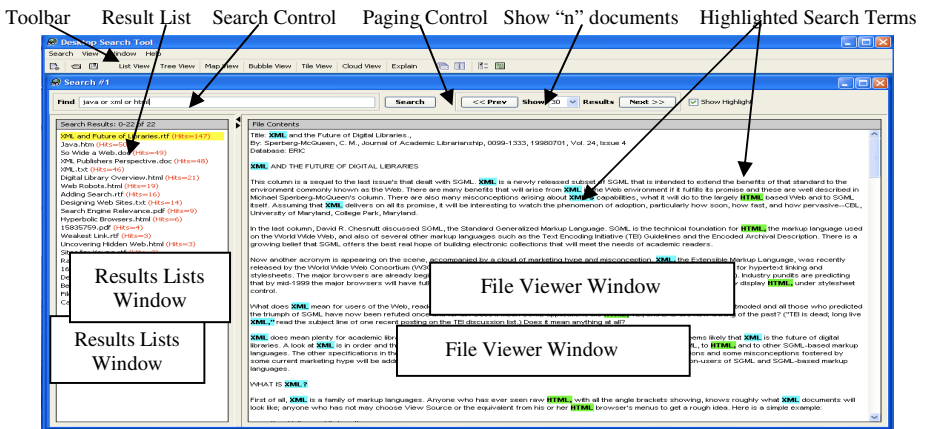

Fig. 1. List View

This view also contains a File Viewer window that will display the textual contents of a selected file in order to allow quick review of the file contents. Matching query terms are highlighted in different colours to aid the user identify where these terms occur in the document. Any non-textual content, such as images, etc. in the actual document are not displayed. Similarly, some of the source document formatting will be lost as only line and paragraph breaks are preserved in the extraction process.

Double clicking on a selected file displays the original document in a separate window. Such a view, although elementary, is simple, intuitive, provides clarity and a quick preview of the documents. Unsurprisingly, most evaluation participants gave strong support to this view as both easy to use (89% who Agree or Strongly Agree) and useful (86%) in reviewing the results. The evaluators liked the highlighting of the search terms in the file viewer and clear indication of the number of hits per result file, and suggested improvements related to more flexible sorting of the results and more document and result information to be made easily available. These results confirm that the basic design and operation of the desktop search engine is effective and useful.

5.2 Tree View

The Tree View is similar to the List View (Fig 2) except the result files are organised based on their underlying folder structure. For each file in the results list, all of its parent folders are added to the folder hierarchy (avoiding duplicates). The Result files are then added into the tree at the appropriate folder for their physical location.
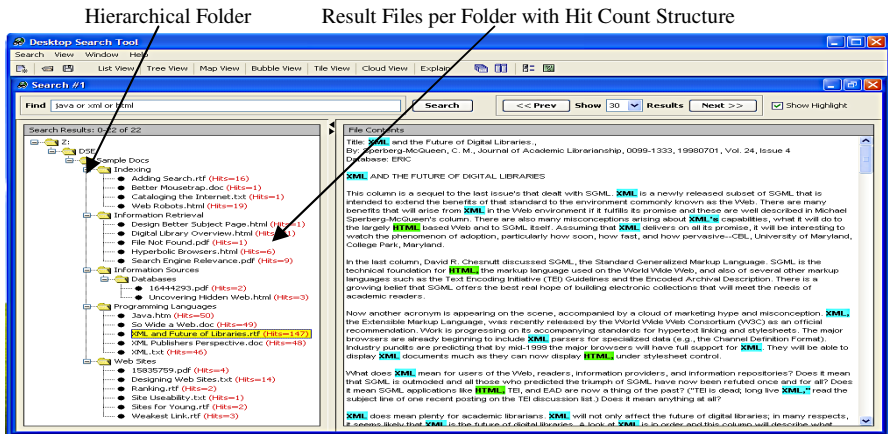

Fig. 2. Tree View

This view is very similar to the Microsoft Windows Explorer view. However, only files that match the query string are displayed and only the parent folders of these files are included in the tree. The purpose of this view is to use the physical file structure as part of the results display. If users have taken the time to organise their documents into meaningful folders and hierarchies then this information may be useful when reviewing results. This view is particularly suited for thesaurus or taxonomy based folder organisations where documents are stored in the respective nodes of this organisation scheme.

As such, related documents would already have been assessed and organised into folder hierarchies that will help users to quickly zoom into documents of interest.

With the familiarity of Windows Explorer, participants strongly indicated that this view was easy to use (93%) and useful (91%) in reviewing the results. They found the view clear and obvious. 87% of them acknowledged that if they had organise their documents logically in folders, then this view would be especially useful for them. This confirms the design premise that the user's folder structure is a useful aid to present search results, as well as a means to logically organise information in thesaurus/taxonomy-like structures that can support browsing as well as searching.

5.3 Map View

The Map View (Fig 3) provides an overview of the relationship between the query terms and the result files. Each query term is depicted as a blue rectangle and each result file as a green ellipse. Lines link related query terms and results files. These are annotated (in red) with the number of occurrences of the query term in the result file.

The view can be zoomed and rotated and individual shapes can be moved around on screen to obtain views that are more legible and avoid cluttering. If the mouse is moved over a query term it will display a popup window that lists all the result files that contain the query term along with their respective number of hits (not shown in Fig 3). Similarly, if the mouse is moved over a result file then a popup window will display all the query terms found in this file with their respective hit counts (Fig. 3).
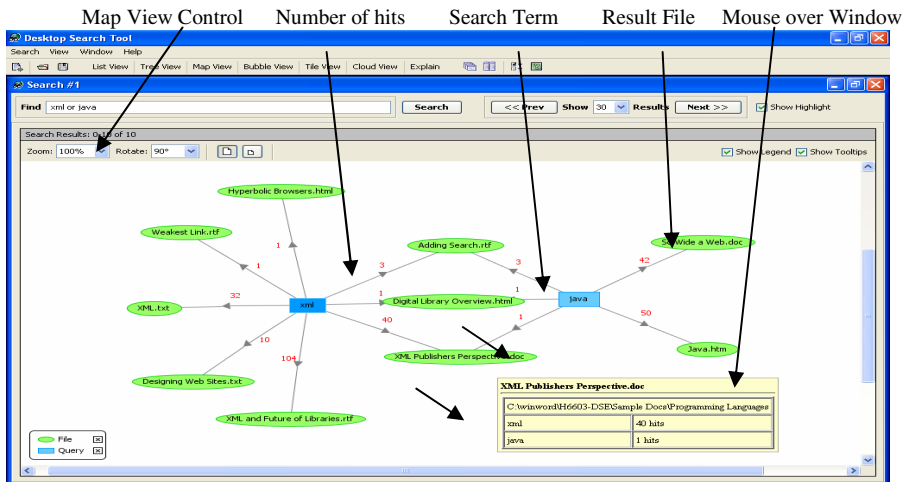

Fig. 3. Map View

This view shows how individual query terms affect the results and which files contain one or more query terms. This bird's eye view can be used to detect problems in the query specification if the required results are not as expected. It will clearly show the relative influence of each query term in producing the result files and therefore help the user in deciding whether the query needs to be reformulated and how to do so.

The evaluation found that slightly over half of the evaluators (51%) agreed or strongly agreed that the Map View was useful in reviewing their query results and reformulating their query. The distribution of responses for ease of use and usefulness are very similar. Qualitative comment analysis indicated that the most useful aspect noted by the evaluators (35%) was the ability to see an overview of the relationship between the query terms and the results files. This was the design premise for the Map View – to provide a clear overview of the query and its effect on matched results. However, the view can become very crowded for complex Boolean queries with a large number of items displayed resulting in overlapping of the graphic objects. A significant number of evaluators (36%) indicated that this caused confusion.

5.4   Bubble View

Boolean logic systems make it difficult to judge the relevance of a result file. The total number of hits alone is not necessarily a good guide to relevance especially when document length is taken into consideration. Therefore, it is desirable to normalise this measure to take into account document size. In this work, a hit density is calculated as the number of hits per 1,000 searchable terms (non stopwords) in the document.

The intention of the Bubble View (Fig 4) is to help the user better assess the relevance of different documents. The axes of the graph are the number of hits and the calculated hit density. These measures are used to distribute the documents along each axis as they provide good document discrimination in order to achieve a better visualisation. The diameter of the bubble is determined by the number of query terms present in the result file and its colour represents its file type.
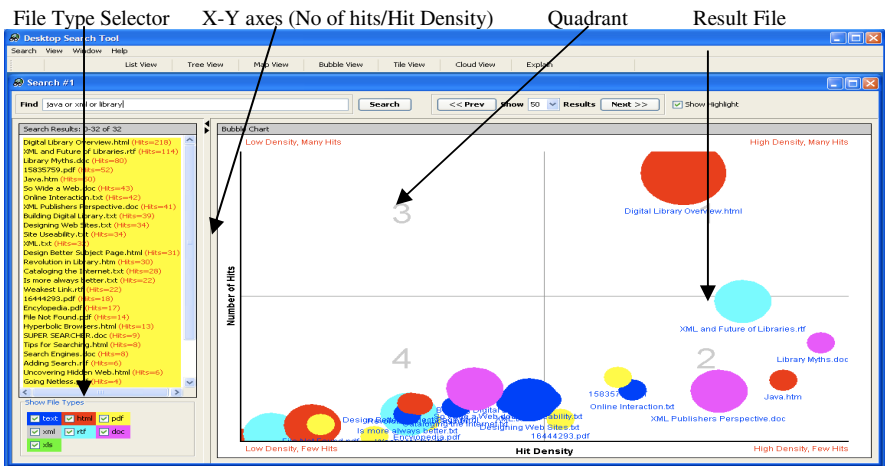

Fig. 4. Bubble View

Quadrant 1 is expected to contain the most relevant documents as both the number and density of hits is greatest. Correspondingly, quadrant 4 will be expected to contain the least relevant documents, as both the hit count and density are smallest. The display

suggests that documents should be explored in priority according to the Quadrant numbers. The view therefore attempts to provide an overview of document relevance for a given query and aids the review of documents most likely to be relevant to the query.

The evaluation results show that 46% of the evaluators found this view useful in reviewing their query results. The majority found the position (65%) and size (59%) of the bubbles gave them useful information, which supports the concept of this view as a means to convey several dimensions about the relevance of the results documents.

The comments analysis showed that useful features were the ability to get a quick and easy overview of the relevancy of the results and the ability to see the hit density. The major confusion factors related to the display of a large number of result documents where the titles overlap and become unreadable and the display was found to be very cluttered and messy. Suggestions for potential improvement relate mainly to improving the layout to increase clarity and for help on how to interpret the view.

5.5   Tile View

The Tile View (Fig 5) presents each result file as a coloured tile using a Treemap. A Treemap is "a space-constrained visualization of hierarchical structures" [15]. The size of each tile is determined by a measure such as Total Number of Hits, File Size, and Hit Density (Hits per 1,000 searchable terms). Using the control panel, the user can change the measure used to determine the size of a tile. As before, the colour of a tile is determined by its file type and the display can be restricted to certain file types.

The Tile View can optionally include the folder hierarchy of the results files (not shown in figure). In this variant, all the result files in a specific folder are grouped together in a "super tile". Each folder is enclosed within a tile representing its parent so that the entire folder structure of the results files can be displayed. This is an alternative display of the tree view but with value-added information in the tiles. The purpose of the Tile View is to allow users to review the results visually and judge their relevance based on different criteria with larger tiles denoting the most relevant documents.
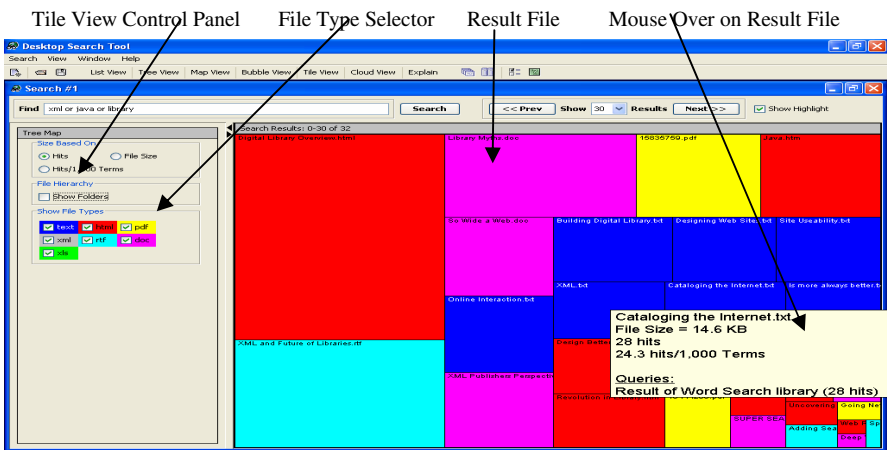

Fig. 5. Tile View

The results of the evaluation of the Tile View show over half the evaluators (59%) agreed or strongly agreed that the Tile View was useful in reviewing their results. Over two thirds found the tiles to be obvious and easy to understand (68%) and the ability to use different criteria to control their sizing was found to be useful (69%). This supports the design objective for this view to easily support the use of different criteria for judging the relevance of the results documents.

The ability to group files by folders also received strong support with 75% of evaluators agreeing or strongly agreeing that this was useful. The comments analysis indicated that useful features were the ability to change tile size based on different criteria, the ability to group files by folder and the use of colour to distinguish file types.

5.6 Cloud View

The Cloud View (Fig 6) is adapted from the Tag Clouds popular on social networking sites such as Flickr. A Tag Cloud is a weighted list which contains the most popular tags used on that site and the relative popularity of each tag is indicated by changing its font size. It is thus easy to see the most popular tags. The Cloud View creates a Word Cloud based on the (indexable) content of the result files. The file contents are examined and stop words and non-indexable terms are removed. The words are then stemmed and a simple term count of the documents contents. The top 300 terms are then displayed in a Word Cloud as they represent the most common indexable terms.

Only files selected in the Results List (in the left hand window) have their contents included in the Word Cloud. If the selection of files is changed, the Word Cloud is dynamically refreshed with information based on the new selection of files.

When the user clicks on a word in the Word Cloud a popup menu appears offering the choice to expand (OR), restrict (AND) or exclude (NOT) the word from the current query or to create a new search (NEW) using the selected word (Fig 6).

Selectable Result List   Word Cloud (Alphabetically arranged) Common word (larger fonts)   Query refinement
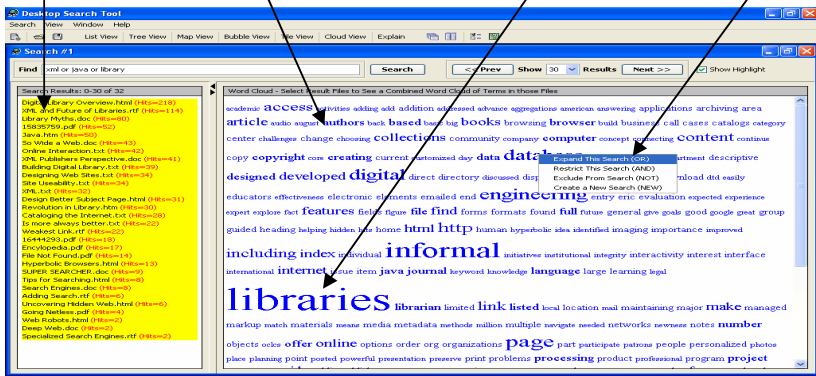

Fig. 6. Cloud View

The purpose of the Cloud View is to provide uses with the most common **n** words (300 in this instance) found in the selected files in the result lists thereby providing an idea of the contents of the result files (i.e. basically a concordancer) and information on potential words that can be used in the query refinement process.

The evaluation results for the Cloud View showed that nearly two thirds of the evaluators found the Cloud View useful in reformulating their query (63%) and easy to use (61%). However, the distribution profile for Question 37 (usefulness of Cloud View) is different with a bi-polar distribution, with a peak for Disagree and Agree. This implies that the evaluators were split into two groups, a conclusion strongly supported by a review of the comments. Those evaluators who scored the usefulness of the Cloud View very low (Strongly disagree or Disagree) reported a lot of confusion as to the contents of the Cloud. In other words, they did not find the view useful because they did not understand what it does. Those who did rated it highly. This implies that some users had not seen this type of visualisation before nor understood its potential.

6 Conclusion

A lightweight desktop search engine, DSE, along with six integrated views was developed to study the effectiveness and usefulness of them to aid the processing of query results and query refinement. The evaluation carried out on 94 participants indicates these visualisations useful and easy to use on the whole. They generally felt that it would help them find their desired results quicker. In particular, the Tree View and Cloud View were rated highly by the evaluators. The Tree View takes advantage of the users own defined hierarchies (their folder structures) to present the search results in a format that significant numbers of the evaluators found useful. The Cloud View, while novel, posed difficulty for some evaluators who did not understand it. Those who did rated the view to be very useful. Our work indicates that the provision of a suite of tightly linked yet different visualisations has the potential to increase the usefulness and ease of use for result processing and query refinement for desktop searching which is very much at its infancy stage of development in contrast to its more established online search engine counterpart. Our next stage of work is to improve the current views in light of the evaluation findings in preparation for the scalability and longitudinal tests for a series of increasingly larger result sets of documents.

References

1. Fowler, R.H., Fowler, W.A.L., Wilson, B.A.: Integrating query thesaurus, and documents through a common visual representation. In: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, New York (1991)
2. Nowell, L.T., et al.: Visualizing search results: some alternatives to query-document similarity. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, New York (1996)

3. Shneiderman, B., et al.: Visualizing digital library search results with categorical and hierarchical axes. In: Proceedings of the fifth ACM conference on Digital libraries, ACM Press, New York (2000)
4. Koshman, S.: Testing user interaction with a prototype visualization-based information retrieval system. Journal of the American Society for Information Science and Technology 56(8), 824 (2005)
5. Wojciech, W., Krzysztof, W., Wojciech, C.: Periscope: a system for adaptive 3D visualization of search results. In: Wojciech, W. (ed.) Proceedings of the ninth international conference on 3D Web technology, ACM Press, New York (2004)
6. Newby, G.B.: Empirical study of a 3D visualization for information retrieval tasks. Journal of Intelligent Information Systems 18(1), 31–53 (2002)
7. Kaizer, J., Hodge, A.: AquaBrowser Library: Search, Discover, Refine. Library Hi Tech News 22(10), 9–12 (2005)
8. Heo, M., Hirtle, S.C.: An empirical comparison of visualization tools to assist information retrieval on the Web. Journal of the American Society for Information Science and Technology 52(8), 666 (2001)
9. Sebrechts, M.M., et al.: Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces. In: Proceedings of SIGIR: International Conference on R&D in Information Retrieval, vol. 22, pp. 3–10 (1999)
10. Veerasamy, A., Heikes, R.: Effectiveness of a graphical display of retrieval results. In: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, New York (1997)
11. Heflin, J., et al.: WebTOC: Evaluation of a Hierarchical Browsing Interface for the World Wide Web (1997), http://www.otal.umd.edu/SHORE/bs11/
12. Morse, E., Lewis, M., Olsen, K.A.: Testing visual information retrieval methodologies case study: comparative analysis of textual, icon, graphical, and "spring" displays. J. Am. Soc. Inf. Sci. Technol. 53(1), 28–40 (2002)
13. Wingyan, C., Chen, H., Nunamaker Jr, J.F.: A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration. Journal of Management Information Systems 21(4), 57–84 (2005)
14. Chen, C., Yu, Y.: Empirical studies of information visualization: A meta-analysis. International. Journal of Human-Computer Studies 53(5), 851–866 (2000)
15. Bederson, B.B., Shneiderman, B., Wattenberg, M.: Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies. ACM Transactions on Graphics (TOG) 21(4), 833–854 (2002)