Multiform Glyph Based Web Search Result Visualization

Jonathan Roberts, Nadia Boukhelifa, Peter Rodgers
University of Kent at Canterbury
{J.C.Roberts@ukc.ac.uk, N.Boukhelifa@ukc.ac.uk, P.J.Rodgers@ukc.ac.uk }

Abstract

Searching for information on the web is hard; the user may not know what they are looking for, they may refine their search from information gathered by preliminary naive searches, and they may be looking for luminous sites that have many external links so that they can browse further. Information visualization can aid the user in many of these search related tasks. Certainly, the user is familiar with browsing and manipulating the search results through textual style interfaces, but they would gain a better understanding of the information through different presentation methods. Thus, we believe rank ordered lists should be used along-side abstract information visualization presentations. We present a system that displays multiple views of search result information. It provides views for displaying abstract visualization designs using multiform glyphs as well as a ranked text based list. Our engine also retrieves detailed information about the located sites (such as size of page, and quantities of internal and external links); and we describe two glyph designs that display this rich information.

Keywords Visualization, multiple views, query, search results, abstract visualization, glyphs, WWW.

1. Introduction and Motivation

Searching for information on the web can be frustrating. Not only does the web user need to choose the right search engine but also they need to think of and input the right keyword query terms. Moreover, the user is reliant on the search engine displaying the results in a way that is clear and highlights the most relevant information.

From our own experiences the search engine may return tens, hundreds or even hundreds of thousands results. The user may refine their query terms, but, more often the user will browse through many of these result pages looking for pages that of interest, documents that may trigger off ideas useful for searching, pages of links that may enable further browsing. The users may be interested in seeing where their keywords fit in the website or page and where the information is located (or at least who owns the content or domain name).

The current search engines are amazing because they do a good job at quickly finding large amounts of relevant information. According to some recent studies the Web consists of approximately 2.5 billion documents [7, 11] and growing at a rate of more than 7.3 million pages per day. The accuracy of these figures is debatable but it is clear that the World Wide Web is growing at a phenomenal rate. However, the search results are mostly presented in a relevance ordered textual list that the user browses to find interesting information. These traditional representations are presented over multiple pages and the user only views a small proportion of the results in one window, they need to scroll down or move to the next page to see more results.

We believe that graphical visualization of this information, coordinated with traditional text output would allow the user to more quickly drill-down onto the pages that are most interesting and useful. Such a graphical representation could display thousands of results in one view, allowing clustering operations to depict similarities of the search results and allow the user to find relevant information more easily.

We believe that it is useful and possible to present a more data-rich presentation of web search results than solely a textual representation. By data rich visualization we mean more search results as well as more information about those results with minimal occlusion.

1. We require techniques to display a large amount of information on one page; rather than the 5 to 10 links returned by many current search engines. Such a depiction would allow an overview of the information to be presented with subsequent exploration techniques; such as described by Ben Shneiderman's mantra [19] of ‘overview first, zoom and filter, then details-on-demand’.
2. We believe that abstract representations of search result data are possible and would help the user to find what they are looking for. For example, as well as showing some textual information from the retrieved site, quantities could be depicted by lengths, colours, symbols of data such as: how many external or internal URLs or the amount of images present on the site. Such multiform methods are useful because they allow the user to view the same information in different ways and allow the user to better understand the displayed information [18].

3. We believe that there is a benefit in displaying multiple view of search results, including a text view. Current users expect a text representation and find it useful.
4. The views should be coordinated. For example, highlighting some elements in one view would cause highlighting of those same elements in other related views [17].

In this paper we present an experimental visualization tool that displays web search result information utilizing the aforementioned four attributes. This tool has been implemented in Java 1.4. Each search result is represented by a glyph, depicting multiple variables from each part of a search result, the glyphs are placed on the results window by different methods. The system presents the same information in different ways and coordinates the views together, allowing for coupled navigation and exploration of the query results.

In such an application, there is a trade off between more information and more results. This is a central issue in web search result visualization and there were many scenarios in which we had to find a balance between these two key concepts. For instance when choosing the type and number of attributes to assign to each page result.

Visualizations showing more search results are beneficial for many reasons. The current web search engines display a limited set of page results, the user has to scroll down, it is therefore hard to see where a particular result fits in to the whole picture. In addition, rich data visualizations may form clusters, some of which could be meaningful. Furthermore, the user will be provided with a less dense visualization in the form of the classic text display.

On the other hand, a more attribute rich visualization can also be advantageous if not at the expense of the limited display space. One might argue against the rich results display since the ranking techniques for the most relevant pages have proved to be very efficient. Web users hardly go as far as the third page in a web search session. Besides, more results in a 2D environment can cause occlusion.

However, E. R. Tufte argues that the quantity of details is an issue completely separate from the difficulty of reading. Clutter and confusion are failures of design, not attributes of information. “What we seek is a rich texture of data, a comparative context, and understanding of complexity revealed with an economy of means” [23].

Clearly there is a need for a careful mapping between search results attributes and the visual parameters in particular the interaction between these parameters and the effects they have on each other is a major issue that needs to be addressed.

The following section, Section 2, details background and related work. This includes discussing web searching techniques, listing the data that might be visualized, and gives some previous work on visualizing search results. Section 3 discusses the multiform glyphs, how they are placed on the results window, and the coordination between results windows. Section 4 gives our conclusions and further work.

2. Background

In any visualization exercise the developer should evaluate what data is available, what task the user is trying to achieve and how information can be best displayed for that task [23].

Indeed, not only should the design of the system concentrate on the look of the visualization, but also how the user interacts and uses that system by providing operations for the manipulation of the set as well as its visual presentation [6].

Thus, we first look at web searching, data and then visual representation methods.

2.1. Web Searching

Using search engines to find pages on the web is a daily activity for web users. The search engine takes the users keywords ranks the pages that match and displays the first 10, 20 or 100 of the search results in order. The search engine provides a filter mechanism by selecting all the relevant pages from those indexed in the engines database and most relevant to the query terms. This constitutes a search session.

The user can enter more complex search commands such as using the logical operators AND, OR, NOT, some search engines allow Proximity operators such as: ADJACENT, NEAR and FOLLOWED BY, others allow methods to require or exclude certain words (often indicated by the + and – signs).

Each search engine operates on slightly different command syntax, thus sometimes the user does not find relevant information because they have used an erroneous syntax.

In a given session the user often browses some intermediate results then refines the search by exchanging or adapting one or some of the keywords. From studies, such as those by Jansen et al [12] the user typically enters up to four terms with 67% of searches having one term and under 2% containing five or more keywords.

The task of the user is to find relevant information and we postulate that there are various types of searches: a user may be looking for specific pages or looking to find general information about a subject. A user looking for specific pages or domains may have lost the link for an actual page but remembered some aspects of it or believes that there should be a domain registered by some organisation, e.g. “University of Kent at Canterbury", indeed, online search tips such as [22] suggest methods to guess URLs.

Alternatively, a user looking for general information may be searching for a page that holds categorical information and uses this as the jumping off list, rather than browsing through the search results.

2.2. Data

There are many different types and sizes of data to visualize. We group the data into four categories.

The search query, including:

- the number of keywords,
- the keywords themselves,
- operators used (such as boolean, inclusion exclusion operators, etc),
- stemming, other synonyms,
- corrected spelling,
- the session (a series of queries of a user over time) visualized by Sparkler [9],

The search results, including:

- the results themselves
- and the rank of the results.
- domain name and URL,
- title of the page,
- text snippets near the keywords,
- total the number of results,
- results from different search engines.

The content of the information, including:

- media type (html, text, images, sound, pdf, etc)
- content size,
- position of the keywords in the page,
- last modified date.

The structure of the pages, including:

- position of media such as images on the page,
- the quantity of internal links,
- the quantity of external links, often called luminous sites [2].

It is possible to look this data in an alternate manner, by the visualization that each allows. For example, data that contains categories, such as content type, needs to visualized in a different manner from data that can be represented numerically, such as content size.

2.3. Presentation of Results

Most search engines display the data as rank ordered list, such as depicted by some popular search engines (e.g. google, altavista, excite). These tools only represent *search query* and *search result* information – a small subset of the possible information – they omit to depict the *content* and *structure* information.

Indeed, only a small number of query results (often only ten) are displayed at in one presentation; the user is thus missing out on the richness of the information.

However, the information returned from the search result is a textual form, thus there needs to be found an appropriate abstract form to represent this information. There have been techniques to represent similar information in abstract methods, such as simplifying the text to coloured lines, as in SeeSoft [8] and WebTOC [20]. Other methods map various quantities in different perceptual variables.

Some current methods use dots or coloured areas to display the search result information; methods such as Dotfire [21] represent digital library search results by coloured dots; Sparkler [9] represents the search results by coloured dots and relevancy in a bull's-eye formation. Other methods represent multiple variables in a glyph representation, e.g. TileBars [10] map the position of the keywords and the lengths of the documents as used by Mann [14]. Such tilebar information is often laid out adjacent to the text information, moreover, glyphs can be laid out over two or three dimensional axis. For example xFind can plot relevance to y-axis and document size to the x-axis [1]; Cugini et al [5] use different 3D presentation methods from spirals to 3D axis designs; and a target layout method used by DART [4].

Other methods involve visualizing connectivity information, such as SQWID [15], WebQuery [3] and VisIT [13].

3. Multiform Visualization

It is often useful to represent the information in many forms [18]. Indeed, Mann [14] use alternate representations, however, these are presented inline with each result. We choose to display the search results in multiple windows and coordinate the information between views to regain contextual information.

To achieve our data-rich presentation we first retrieve search results from a public domain search engine – such as Yahoo – and then visit each site to gather more detailed information. Our engine gathers information from the search queries including: the number of keywords, the keywords themselves, operators used (such as boolean, inclusion exclusion operators, etc); the search results, including: the results themselves, and the rank of the results, domain name and URL, total the number of results; the content of the information, including: content size, the last modified date; and the structure of the pages, including: the quantity of internal and external links. These were chosen because we perceive them to be useful to the users' exploration.

3.1. Glyph Design

We have been experimenting with different glyph designs for visualizing the data from search results. The first glyph emphasises the domain of the results. Figure 1 shows how the main domains are mapped onto a symbol; the rank is mapped to colour and the size of page to the width of the border of the glyph (the larger the size of the page the thicker the border).

Figure 1: Symbol representations are allocated to different domains.

The external and internal link quantities are mapped to x and y position respectively. Moreover, the country of

origin – if available – is depicted by a small flag of the country to the right of the glyph, as shown in Figure 2.

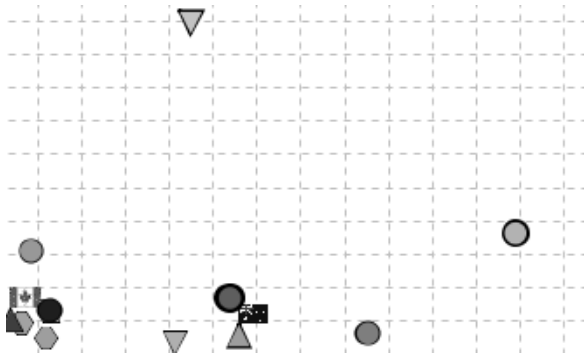

Figure 2: Snapshot of the domain glyph showing some *.com*, .au and *.nz* addresses.

The placement of the glyphs is important as they allow the user to see clusters. Indeed as the amount of links are integer quantities there is a problem with overlap and occlusion in two dimensions. Thus, we jitter the glyph placement by a small random amount in both the x and y coordinates. This allows the user to perceive that there are more dense regions. Otherwise each of the glyphs – with same values – would be directly occluding the other.

The second glyph design is based on quartiles. Each of the quantities is evaluated whether it is in the lower, median or upper quartile, if they are then segments round the side of a cross are filled in (as show in Figure 3). Currently we can visualize two independent data sets using this method, with one represented in the outer quartiles (top of Figure 3) and the other in the inner quartiles (bottom of Figure 3). This concept could be further extended with concentric rings of quartiles to show a number of different variables in one glyph.

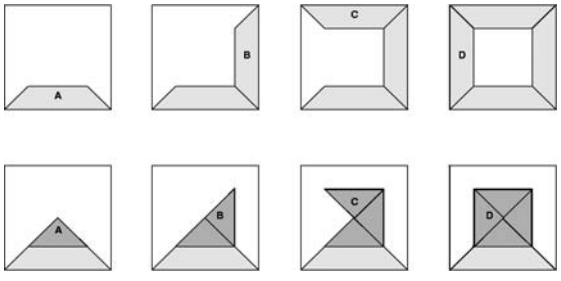

Figure 3: The figure depicts the glyph shape that represents quartile information. The segment parts A, B, C, D represent values in steps of 25%. The inner and outer quartiles can be used to visualize separate variables.

As with the first glyph, colour is used as an indicator of relevance, however it is possible to reassign different variables to each representation. For example, colour could be used to indicate the content type of a page.

The bottom left window of Figure 4 shows a number of such glyphs. The inner quartile indicates the number of external links, and the outer quartile indicates the number of internal links. It would have been possible to use the placement method of the previous glyphs, in a 2D layout based on two variables, however to show the variety of placement techniques the glyphs are in a linear layout, with the most relevant at the top left. The glyphs are wrapped onto the next line. This gives an overview of the search results in a more compact form than the standard textual representation

We believe there are different attitudes to web search: specific needs, general searching and categorical information seeking. We try to cater for these needs by choosing varied types of page result attributes. For more general browsing, users might be satisfied with features such as the title, the number of results and the number of query keywords. For specific requests, more precise information such as the domain name, the URL and the type of information is needed. Finally, for luminous sites seekers, the number of internal and external links could be just what they are looking for.

For the mapping between information attributes and visual attributes, Bertin's six retinal variables of information display have been adopted (shape, orientation, colour, texture, value and size) and some of the preattentive visual features have been investigated (such as the difference in curvature or form, size, closure, hue and flicker) [24].

The most emphasised feature in the domain Glyph is clearly the domain name. In theory, a domain name reflects the country of origin and then the organisation it belongs to. A domain name is like a cyber-state [25]. It is the equivalent of a physical business address, acting as an address where customers can contact the service providers. Meaningful symbols such as logos for known organizations could have been used but this might use up some valuable display space.

3.2. Coordination and Visual Exploration

We have also implemented two further views that display a rendering of the actual search result (as it would appear from that search engine) and also an abbreviated list that solely contains the URLs from the search results. Users are familiar with such lists, and the presence of the WWW page in the multiple view system aids the coordination between windows.

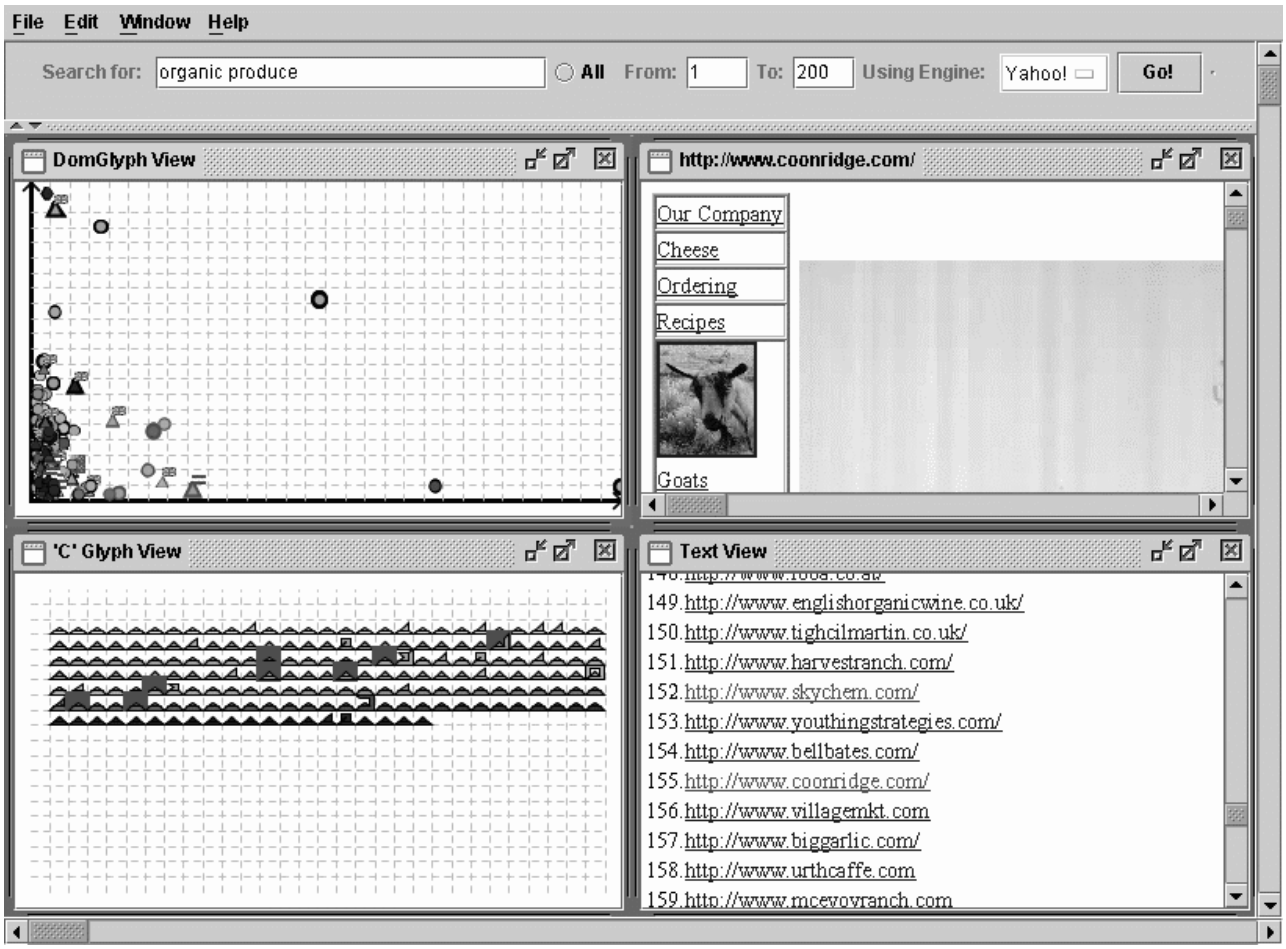

Figure 4: Screen dump of our search result visualization system, showing search results for "organic produce". Some sites are highlighted and the *coonridge* site is shown in the top right.

With any multiple views system coordination is important, as it allows the user to understand how the information in one view relates to another. In our system the user can select items so that a selection made in the domain glyph view (Figure 4) highlights the same items in the other views. Moreover, when the user clicks on a glyph or link the page is loaded in the search result window (Figure 4 top right).

Highlighting of selected pages is via text colour in the text window, which cannot be clearly seen in the monochrome Figure 4. The borders of the glyphs in the top right window are coloured to indicate selection. More clear is the background shading that shows selection in the bottom left window.

4. Conclusions & Further Work

We have developed an experimental software system that retrieves the search results from a traditional search engine, gathers details about the content from the linked sites and displays the information in multiple forms, using glyph techniques. We have introduced two simple glyph designs based on symbol and quartiles.

Recently Google has developed new API that allows developers to issue search requests to the engine to access Google's index. It may be that this would improve the speed of our implementation by making the data we visualize more readily available.

We believe that the quartile glyph has much potential and we are currently developing this idea further. Additionally, we plan to extend the coordination of the system; we have found that certain sites take a long time to load; this affects both the data gathering and the coordination with windows that deal with the actual rendering of that site. Obviously using different threads would help, but also some form of refinement rendering would be useful for such views. The system could easily be implemented directly on top of a search engine. In this case, the relevant data would be stored in the search database, rather than having query the pages directly as present.

It is useful to have the text view in combination with the abstract visualizations. However, it would be beneficial to include some form of hierarchy in the text (such as used in Table Lens [16]) to squash up the non-highlighted information to clearly show the more relevant information.

Acknowledgements

This work has been supported by the EPSRC (grant reference: GR/R59502/01).

References

1. K. Andrews, C. Gütl, J. Moser, V. Sabol, W. Lackner. Search Result Visualization in xFind. In Proceedings UIDIS. IEEE Computer Society. pp. 50-58. 2001.
2. T. Bray. Measuring the Web. Computer Networks and ISDN Systems. Volume 28, Issues 7-11, pp 993.
3. J. Carrière, and R. Kazman. WebQuery: Searching and Visualizing the Web through Connectivity. In Proceedings of WWW6. Santa Clara CA, April 1997 pp. 701-711.
4. E. I. Cho, S. H. Myaeng. Visualization of Retrieval Results using DART, Proceedings of RIAO 2000.
5. J. Cugini and S. Laskowski and M. Sebrechts. Design of 3D Visualization of Search Results: Evolution and Evaluation. In Visual Data Exploration and Analysis VII, Proceedings of SPIE Vol 3960. pp 198-210. 2000.
6. J. C. Cuz. Presenting search results: Design, visualization, and evaluation. Workshop: Information Doors: Where Information Search and Hypertext Link. In conjunction with ACM Hypertext 2000.
7. http://www.cyveillance.com/web/us/ Sizing the Internet. 2000.
8. S. Eick. Graphically displaying text, Journal of Computational and Graphical Statistics, 3(2):127-142, June 1994.
9. K. Perrine, S. Havre, E. Hetzler and E. Battelle. Interactive Visualization of Multiple Query Results. In Proceedings of the IEEE Symposium on Information Visualization 2001 INFOVIS'01, pp 105-112, 2001.
10. M. A. Hearst. TileBars: Visualization of Term Distribution in Full Text Information Access. Proceedings of CHI'95. May 1995.
11. How much information? http://www.sims.berkeley.edu/ research/projects/how-much-info/internet.html. 2002.
12. B. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing and Management, 36(2): 207-227, 2000.
13. D. Kauwell. VisIT. http://www.visit.uiuc.edu/
14. T. M. Mann, Visualization of WWW-Search Results, DEXA Workshop, pp. 264-268, 1999
15. S. McCrickard and C. Kehoe, Visualizing Search Results using SQWID, In Proceedings of the Sixth International World Wide Web Conference, April 1997.
16. R. Rao and S. K. Card. The table lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In Proceedings of ACM CHI'94, ACM Press, pages 318- 482.
17. J. C. Roberts. On Encouraging Coupled Views for Visualization Exploration. In, Visual Data Exploration and Analysis VI, Proceedings of SPIE, volume 3643, pp. 14 - 24. January 1999.
18. J. C. Roberts. Multiple-View and Multiform Visualization. In Visual Data Exploration and Analysis VII, Proceedings of SPIE, volume 3960 pp.176 – 185. IS&T and SPIE, January 2000.
19. B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In Proceedings of the IEEE Symposium on Visual Languages, pages 336-343, Washington, September 1996.
20. B. Shneiderman, D. Feldman, A. Rose. WebTOC: A Tool to Visualize and Quantify Web Sites using a Hierarchical Table of Contents. Technical Report CS-TR-3992, February 1999
21. B. Shneiderman, D. Feldman, A. Rose, X. F. Grau. Visualizing Digital Library Search Results with Categorical and Hierarchical Axes. HCIL Technical Report No. 99-03. June 1993.
22. R. Tyner. Sink or swim: Internet search tools & techniques. http://www.ouc.bc.ca/libr/connect96/search.htm.
23. E.R. Tufte. Envisioning Information. 1990.
24. C. Ware. Information Visualization: Perception for Design. Morgan Kaufmann Publishers, 2000.
25. http://www.aboutdomains.com/News/basics.htm. About domains. 2002.