Visualization Viewpoints

Editor:
Theresa-Marie Rhyne

Adding a Visualization Feature to Web Search Engines: It's Time

Pak Chung Wong
Pacific Northwest National Laboratory

In the dynamic and interactive world of the Internet, we as a technology community have learned that we can make good Web-based applications better by adding rich visualization and analysis capabilities to control and navigate the applications' user interface. Noteworthy examples include Google Earth (http://earth.google.com) which, through visualization, turns hundreds of terabytes of raw satellite images and aero-photographs into actionable information shared with and enjoyed by millions every day. But for every problem the community addresses, plenty more go unrecognized or unexplored. In this Visualization Viewpoints article, I examine a longstanding Web-application problem. My hope is to stimulate readers to consider the issue and offer an innovative solution.

Search Engine Results Pages

Since the first World Wide Web search engine (http://en.wikipedia.org/wiki/Web_search_engine) quietly entered our lives in 1993, the "information need" behind Web searching has rapidly grown into a multibillion-dollar business that dominates the Internet landscape, drives e-commerce traffic, propels the global economy, and affects the lives of the human race. Today's search engines are faster, smarter, and more powerful than those released just a few years ago. With the vast investment pouring into research and development by leading Web technology providers and the intense emotion behind corporate slogans such as "Win the Web" or "Take Back the Web," I can't help but ask—why are we still using the very same "text-only" interface that we've used for almost two decades to browse our search engine results pages (SERPs)? Why has the SERP interface technology lagged so far behind in the Web evolution when the corresponding search technology has advanced so rapidly? Here I explore some current SERP interface issues, suggest a simple but practical visual-based interface design approach, and argue why such an approach can be a strong candidate for tomorrow's SERP interface.

In the Realm of Information Visualization

Like many who have witnessed the birth of the Internet and the WWW technical inventions that followed, I've spent countless hours surfing the Web and trying every search engine that I have encountered. Professionally, I'm an information technology researcher working at a national laboratory on various issues of data visualization and knowledge discovery. Among the work that I've been involved in was the development of a series of information visualization technologies that explore a large amount of free-text files. It was the result of interaction with users that led me to understand and appreciate the importance of simple, yet practical data navigation interfaces. Many of the same design concepts and principles can be applied to the next generation of SERP interface.

Longstanding SERP Interface Issues

It's fascinating that all Web search engines today are almost identical in presentation layout and behavior: each Web search can return up to 1,000 URLs, each SERP can display from 10 to 100 URLs, and each URL comes with the page title and portions of the page that contain the keywords. (Figure 1, next page, shows nearly identical screenshots of three of the most popular search engines today, namely Google, Yahoo, and Live Search.) Companies adopted this interface design when Lynx (http://lynx.browser.org) was the unanimous choice of nongraphical Web browsers and arrow keys were the primary means of cursor movement on a display. Amazingly, the same presentation approach has survived the entire Web evolution and became the de facto standard for all major search engines today.

A main weakness of this interface is that no matter how thematically rich the SERPs' contents are, users have little knowledge of what the Web pages contain until they bring up and read individual pages. And when users find a desired Web page, there's no guarantee that the next few pages in the queue will share the same thematic contents. For example, when I search for "jobs," I might be looking for an employment

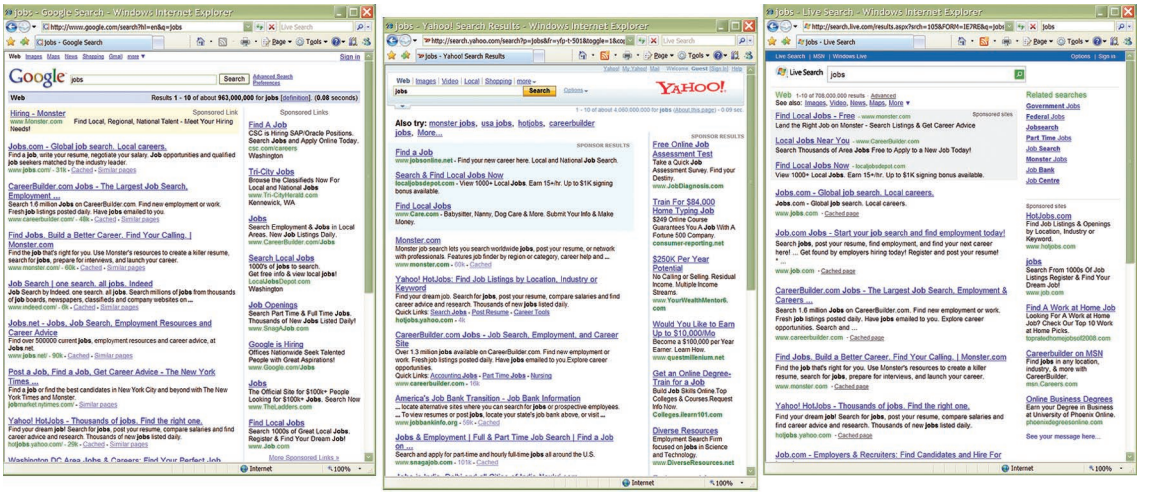

Figure 1. Screenshots of (from left to right) Google, Yahoo, and Live Search. Note that their layouts are nearly identical.

opportunity, Steve Jobs, or jobs that are associated with Steve Jobs. Today's search engines will jumble all of these very different results in one long list, as Figure 1 depicts. The lack of options to preview the similarities among the SERPs, which would let users skip the irrelevant ones, has become one of the most desirable features for tomorrow's search engine.

Major Web search technology providers such as Google and Yahoo have recently released experimental technologies that support a certain degree of thematic preference for Web searches. Notable examples are Mindset by Yahoo (http://mindset.research.yahoo.com) and Google's Experimental Search (www.google.com/experimental). These are strong indications that the Web search industry is technically ready to identify SERPs' thematic contents in real time for users. The next logical hurdle then is to effectively organize these slightly complicated multidimensional results.

Many would agree that the current text-based SERP interface was primarily designed to support 1D top-to-bottom listings, which can be accomplished if the SERPs are organized on the basis of a single ranking scale, such as Google's PageRank. It is, however, rather difficult for the same interface to effectively display SERPs' multidimensional thematic contents. Perhaps a bigger challenge is to show the gray areas in which SERPs are tied to multiple themes. A potential solution to this problem could come from the information visualization community, where researchers for years have studied how to interactively visualize multidimensional information.

A Technologically Feasible Solution

While there might be other visualization techniques that researchers could apply to today's SERP problem, I have chosen the In-Spire visualization system (http://in-spire.pnl.gov) as an example because of my familiarity with it (http://in-spire.pnl.gov; also see the sidebar on the next page). This visualization system, which was developed at the Pacific Northwest National Laboratory (www.pnl.gov), can interactively harvest Web pages and generate a so-called "galaxy" view of the SERPs. Figure 2 shows the visualization of a Google search on "jobs." SERPs (shown as white dots) that are thematically related are clustered together, and the pairwise Euclidean distances between the blue clusters indicate their theme contents' dissimilarities. The white labels are the most content-bearing terms extracted from the SERP text. Figure 2's visualization accurately separates Web pages related to Steve Jobs (upper left); job seeking

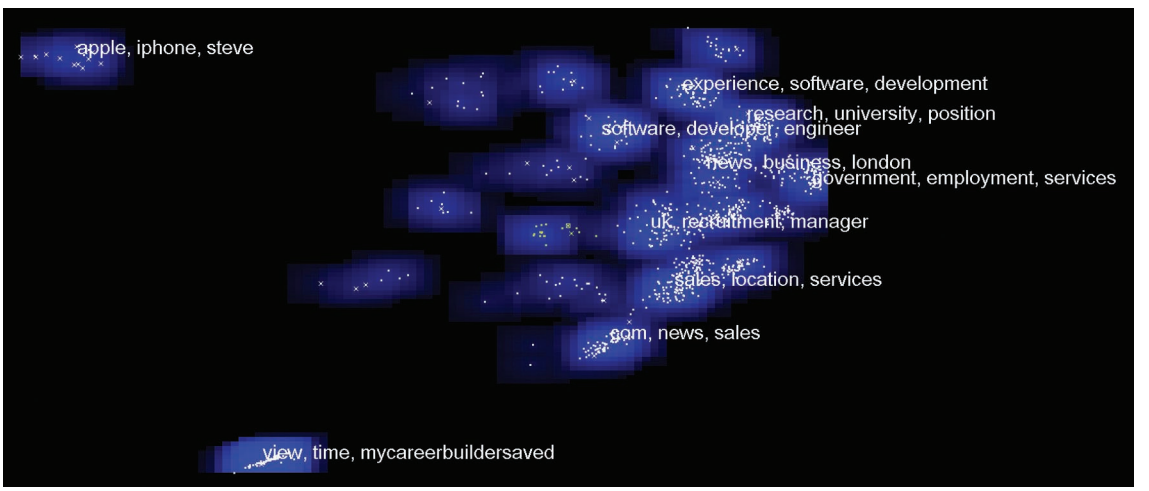

Figure 2. A scatterplot-based visualization of a Google search on "jobs." White dots are search engine results pages (SERPs); blue clusters represent thematically similar SERPs.

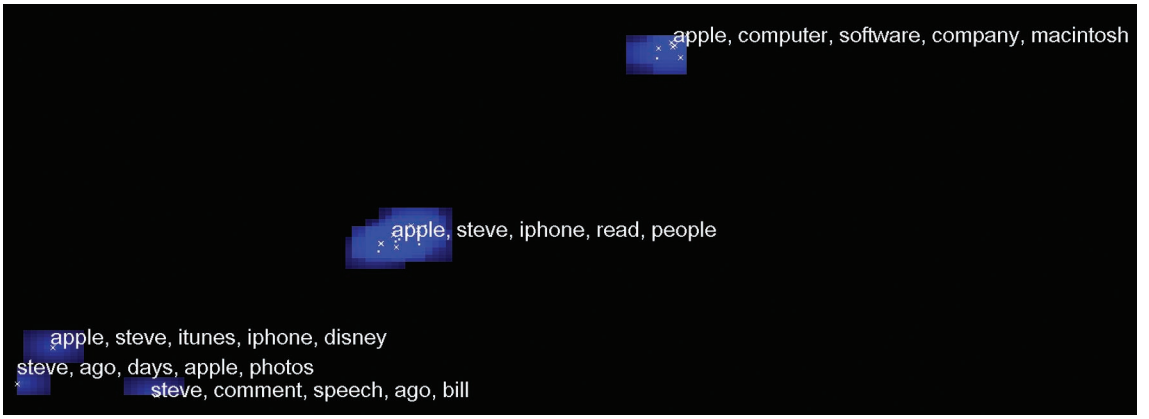

Figure 3. The Steve Jobs cluster in the upper-left corner of Figure 2 is reprojected here to bring out the finer details of the cluster.

and offering (lower left); and academic jobs, employment news, government services, and jobs in foreign countries (the cloud of clusters at middle right).

On the basis of the themes the system suggests, users can quickly narrow the set of Web pages for browsing, potentially saving both user time and network bandwidth for sequentially or randomly page browsing. For example, if users want to learn more about the "Steve Jobs" cluster in Figure 2, they can either zoom in on the scatterplot and browse the details or regenerate a new scatterplot based solely on the SERPs found in the "Steve Jobs" cluster, as Figure 3 shows.

Without the thematic influences of the non-Steve Jobs SERPs, the corresponding Steve Jobs cluster in Figure 2 breaks off into three smaller groups with similar but subtly different theme contents in Figure 3. The lower-left cluster in Figure 3 now contains SERPs about Steve Job's life and career, which include his biography and his commencement speech at Stanford. The middle cluster focuses mainly on news and stories surrounding gadgets such as the iPhone and iPod. The upper-right one is mostly about Apple Computer's business and operations.

While this visualization example is demonstrated on the basis of the SERPs thematic content, other potential options to categorize SERPs include file type, data types, or any metadata tied to the SERPs. For example, Google's VisualRank (www.nytimes.com/2008/04/28/technology/28google.html) can be a key enabling technology to support a full-blown image search engine and reveal the thematic contents of individual images.

After using metasearch sites such as PolyCola (www.polycola.com) for an extended period of time to compare SERPs harvested from different engines, my general evaluation of today's search engines is that although their top-ranked SERPs might be different for the same search, their overall performance in terms of SERP relevance and quality are very comparable. I would argue that the major differences reside mainly on design preference rather than technical justification (or merit).

More on In-Spire

A QuickTime video on the In-Spire technology can be found at www.cs.umd.edu/hcil/InfovisRepository/contest-2004/3/pakwonk-PNNL_Video_2004_InfoVis_A.mov.

On the basis of this observation, I make two linked predictions to conclude my discussion:

- Because major Web technology providers generally have a good understanding of each others' search engine algorithms, the next technological breakthrough could come from the front-end SERP interface design instead of the back-end search algorithm.
- There's little room at the top of the current 1D SERP list for manipulating the rankings returned by the search engine. One future solution is a multidimensional SERP list that probably would require a visual-based interface similar to the one in Figure 2 to display such multidimensional information.

These predictions provide a view of the future as the Web continues to evolve and revolutionize our lives. Information visualization will synergistically improve the performance and capabilities of search engines, which in turn will lead to creation of better Web search technology, and ultimately provide Internet users a vastly improved Internet surfing experience.

Acknowledgments

I thank Lee Ann Dudley, Sharon Eaton, Theresa-Marie Rhyne, and the anonymous reviewers for the comments they provided on this article. The Pacific Northwest National Laboratory is managed for the US Department of Energy by Battelle Memorial Institute under contract DE-AC06-76R1-1830.

Readers may contact Pak Chung Wong by email at pak.wong@pnl.gov.