# Dynamic Visualization of Hubs and Authorities during Web Search

Richard H. Fowler[1], David Navarro, Wendy A. Lawrence-Fowler, Xusheng Wang
*Department of Computer Science*
*University of Texas – Pan American*
*Edinburg, TX 78539*
*[1]fowler@panam.edu*

## Abstract

*The Information Retrieval Visualizer (IRV) is a web-based retrieval system that provides the user with dynamic visualizations of search results. It combines well-known techniques of information retrieval, including stemming, keyword matching, and cosine similarity, with the new and relatively successful hubs and authority approach for describing Web document relevance based on a document's relationship to other documents. The system develops a new and unique approach to document visualization that encodes these metrics in a single visual representation. This new, easily scannable representation, allows the user to interact with search results as the scope of search is expanded dynamically across the Web.*

## 1. Introduction

The explosive growth of the World Wide Web has created a critical need to automatically filter, organize, and assess the quality of information so that users can efficiently and effectively identify and assess information sources. The availability of commercial search engines provides users one means of accessing information on the Web. Though this may account for much of the Web's popularity, it hides many of the difficulties of effective information access. Searching the Web is difficult due to size, diversity of data, and lack of quality assessment metrics, among other factors.

Though widely available commercial search tools are valuable assets to the Web searcher, it seems likely that these tools alone will not solve the current problems of information access. The challenges of information access on the Internet are issues common to all forms of information retrieval. These longstanding issues include difficulties in using indexing vocabularies, indexing indeterminacy, and the user's inability to completely specify information needs [9]. Retrieving information that meets users' information needs is an iterative process, and techniques that explicitly incorporate users' judgments, such as relevance feedback [12], provide means to automate some aspects of user guided retrieval. It is also clear that mechanisms providing alternative paths of access to information can enhance retrieval effectiveness [1].

### 1.1 Data Mining

Data mining techniques typically operate on very large data sets to reveal underlying structures or relations [5]. The link structure of the World Wide Web represents a form of latent human annotation, and thus offers a promising starting point for structural studies of the Web. There has been a growing body of work directed at the integration of textual content and link information in order to organize and search hypermedia such as the WWW.

Of particular interest for the problem of Web search is the recently emerging focus on the Web's hyperlink structure as a source of semantic information in the knowledge discovery process. Clever [3, 4] is a search engine that analyzes hyperlinks to uncover two types of pages; "authorities", which provide the best source of information on a given topic, and "hubs", which provide collections of links to authorities. Kleinberg and colleagues [6, 7, 11] developed an algorithm computing metrics for hubs and authorities, HITS (Hyperlink Induced Topic Search). Beginning with a search topic specified by query terms, the HITS algorithm performs two main steps: a sampling component, which constructs a focused collection of Web pages likely to be rich in relevant authorities; and a weight propagation component, which determines numerical estimates of hub and authority weights by an iterative procedure.

In addition to finding structural components such as hubs and authorities, hyperlinks can also be used to categorize Web pages. However, exploiting this link information is challenging because it is highly noisy. HyperClass [4] embodies one approach to this problem, making use of robust statistical models using random fields together with a relaxation labeling technique. The methodology of influence weights from citation analysis is similar to a link based search method initially used in the Google search engine [2]. The algorithm first computes a score, PageRank, for every page indexed. Given a query, Google returns pages containing the query terms, ranked in order of these pages' Page Ranks. It focuses on pages identified as "authorities", as in other work. Using a crawler, it searches for hyperlinks to other pages that are deemed relevant to the topic, based on text-matching and other techniques. Google ranks such pages highly and to return them in response to a search query.

### 1.2 Visualization

One promising approach for enhancing information access in large information spaces, such as the Web, is visualization to facilitate users' perception of document relation structure. A number of systems have been developed to provide visually based browsing mechanisms for traversing the link structure of Internet documents [8, 14, 15].

Visualization has also been used routinely in data mining as a presentation tool to generate initial views, navigate data with complicated structures, and convey the result of an analysis. Perhaps a stronger visual data mining strategy lies in tightly coupling the visualizations and analytical processes into one tool. Letting human visualization guide an analytical process and decision-making remains major challenge. Certain mathematical steps within an analytical procedure may be substituted by human decisions based on visualization to allow the same analytical procedure to analyze a broader scope of information. Visualization supports humans in dealing with decisions that can no longer be automated [16]. VANISH [10] is a visualization tool used in this way by supporting the easy integration of new semantic domains for visualization. It was used to implement an algorithm that supplements standard content only searches with structural information collected by a spider. After the query has been post-processed via a three-step process, an information space is constructed from the neighbor set, suitable for visualization. It has the capability to visualize these information spaces structured as both graphs and trees.

## 2. The Information Retrieval Visualizer (IRV) System

The Information Retrieval Visualizer (IRV) combines long-standing techniques of information retrieval with more recent developments in data mining of web contents to dynamically present visual representations of web search results to users. The system proceeds by using the user's query to initiate a search of the web using the hubs and authority approach of root set formation. The system then "crawls" from this set to retrieve further documents for which metrics of authority are determined. An important aspect of the system is that a visual representation of each document is displayed, and with which the user may interact, as the computation of metrics through search continues. This section details IRV document retrieval and data mining functionality, together with display and visualization techniques used by the system. Figure 1 below presents an overview of the system in use.

### 1.1 System Use and Display

Displayed as a search progresses, visual page representations are dynamically added to the results visualization window. In the figure's example the page representation shown on the left on both screen and visualization window reflects a) a relatively long document, as represented by the length of the line, that b) has a moderate number of the search keywords, as represented by its placement midway along the x, or keyword count, axis, and c) has a relatively high authority score, as represented by its placement at the high end of the y, or authority, axis. The page representation shown on the right is a) a shorter document, with a shorter line, that b) has fewer keywords, as it is to the right of the first page, and c) is of lower authority, as it is higher. Additionally, it has one paragraph that is

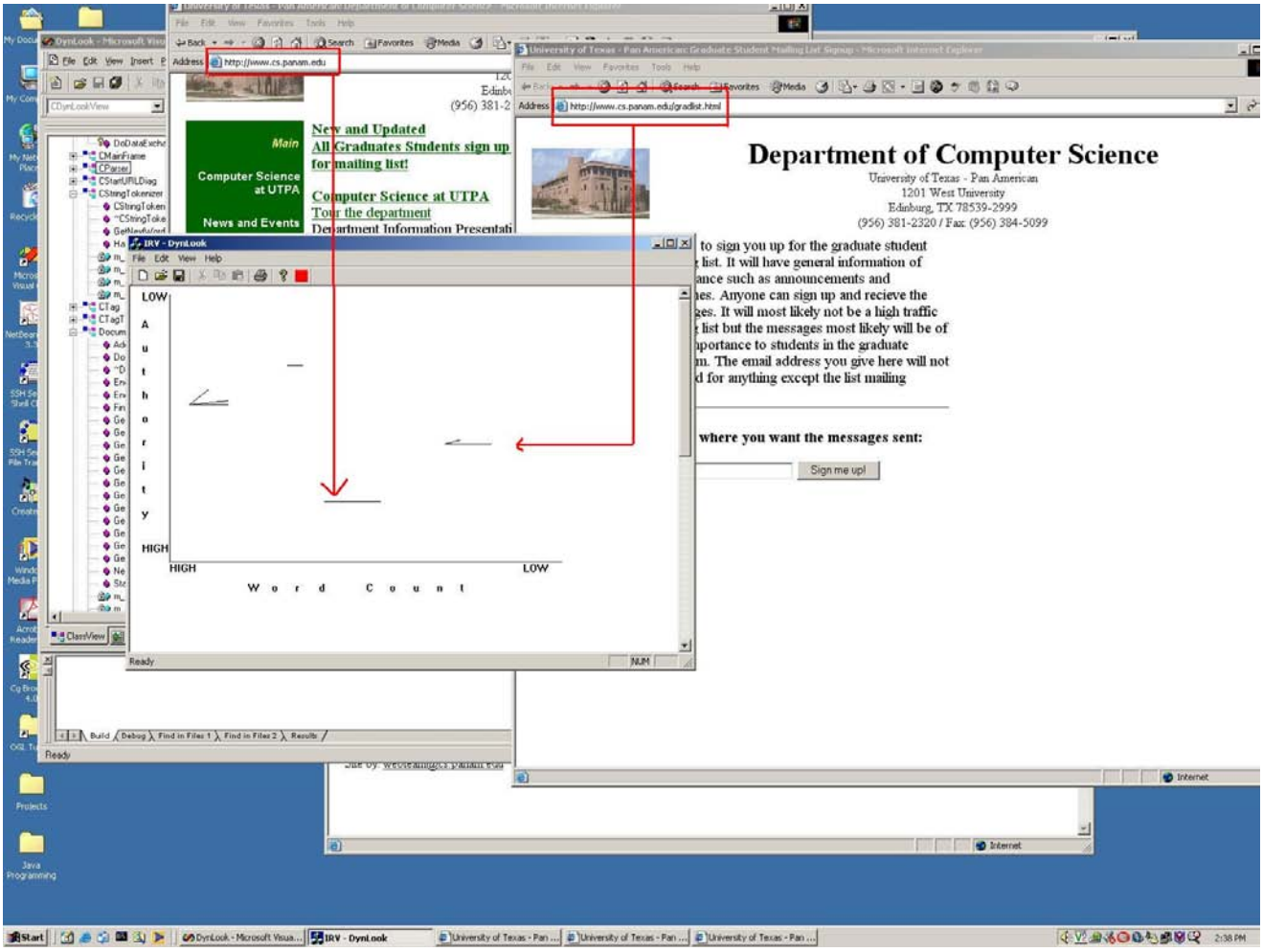


Figure 1. Results of a search using the Information Retrieval Visualization (IRV) system. Displayed as a search progresses, visual representations are dynamically added to the results visualization window. In the example above the page shown on the left on both screen and visualization window has a) relatively many of the search keywords, as represented by the length of the horizontal line, but a low authority score, as indicated by no diagonal line. The page shown on the right has b) fewer keywords than the other page, as the horizontal line is shorter, but a higher authority score, shown by the diagonal line.

closely related to the overall content of the page. As page representations are added to the display, the range of values for authority and number of keywords, of course, changes and it is a pages relative position in the dynamically adjusted range that determines its placement.

As noted, the display element representing a single page is formed from a set of lines. It is the combination of two lines characterizing the system's extraction of information about a page's number of words, i.e., the word count line, and paragraph structure, i.e., paragraph content lines. The display element's form facilitates comparison with other pages' content. For all pages a horizontal line represents the number of words that document has, the word count, or length, line. Given that single web pages may range from a simple list of links to entire reports, this information can be quite useful to the user in determining which document to retrieve and examine.

For pages with multiple paragraphs, paragraph content lines are drawn at different angles to the horizontal line. Each of these lines represents a paragraph in the document. The angle of paragraph content lines represents the relevance individual paragraphs in the document have to the document as a

whole. This provides some information about the structure of the page. It gives an indication of how "focused" the document is, as detailed in the following section. Documents that contain a group of paragraphs closely related to each other will have paragraph lines relatively close to the horizontal and close together. For groups of paragraphs that are not similar, i.e., less "focused", the paragraph lines will spread out more and be farther from the horizontal word frequency line.

### 2.2 Visual Structures

Information Retrieval Viewer shows the user multiple dimensions of information simultaneously. The view is split into two main parts, the X-axis/Y-axis system and P(aragraph)/D(ocument) system. Figure 2 below, shows the visual representation of a single page.

*X-axis/Y-axis System.* Keyword match is a primary element in placement of the visual page representation. The words are stemmed to enhance results and keep higher counts on our values to determine effectiveness. Authority value of a page determines placement on the Y-axis. Placement at X and Y points is relative to the range of keyword matching authority values that is constructed as pages are analyzed and added to the set. Visual representations are positioned at a point on an X/Y plane, where the X-axis represented the matching keywords and the Y-axis represented the authority value on the document.

*P/D System.* While trying to give users as much information as possible without overwhelming them we implemented a way to show how relative a paragraph was to the document it was on. This method is very well known of creating a vector of terms for the entire document and creating a vector of terms on each paragraph and measuring the difference of each. This also gave us the opportunity to show an easy to understand way of representing the length of each document and each paragraph. The result was our ability of instead of plotting generic points we plot points with this structure. This is done with the following function:

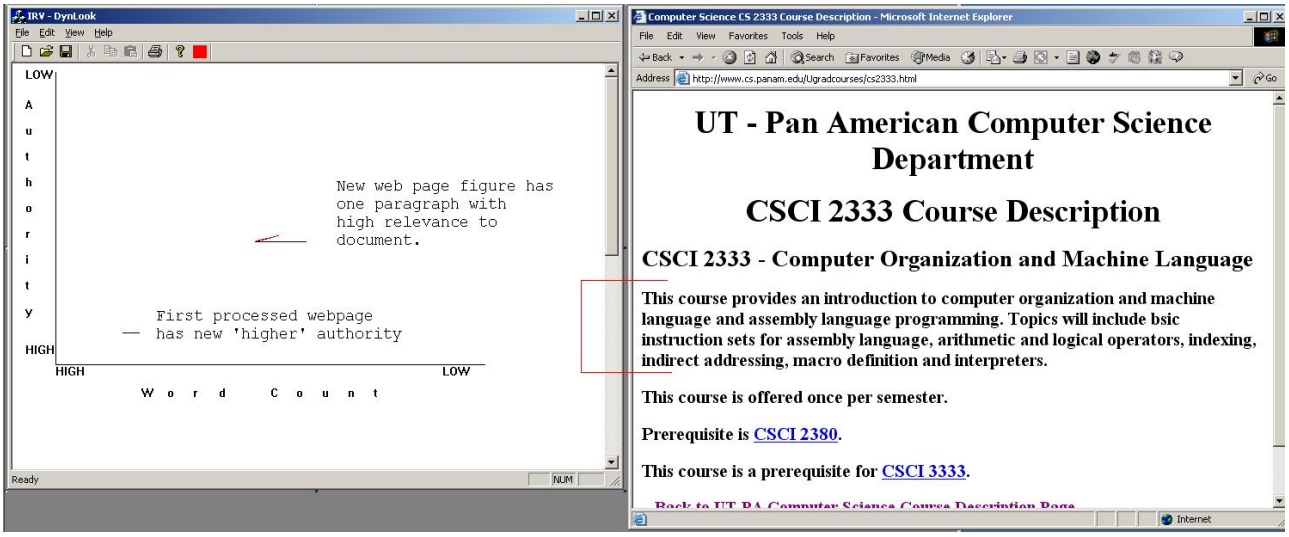


Figure 2. Visual representation of a single page. The visual representation reflects a page's length and paragraph structure. All pages are represented by a horizontal line. The length of the line represents the number of words in the page. Pages with multiple paragraphs have additional, non-horizontal, lines, each of which represents a paragraph in the document. Paragraphs with content close to that of the entire document are represented by lines with small angles from the horizontal line.

$$D(p') = \frac{R(p',d) - \frac{\sum_{p \in d} R(p,d)}{n}}{\sqrt{\sum_{p \in d} \left[ R(p,d) - \frac{\sum_{p \in d} R(p,d)}{n} \right]^2}}$$

Where R(p,d) is the cosine similarity of term vectors of paragraph p and the document d, and *n* is the number of paragraphs in the document. D(p') analyzes the number of steps that the relevance value R(p',d) is away from the mean relevance value. Each step is equal to the standard deviation of the relevancies of paragraphs to the containing document.

## 3. Conclusions

The IRV system uses both well know information retrieval techniques and newly developed data mining techniques coupled with visual representations of retrieved documents to provide users a dynamic, highly interactive system for document retrieval on the web. The combination of techniques is novel and provides an alternative to current static, typically text-based web document retrieval systems.

## 4. Acknowledgements

This work was supported by the University of Texas – Pan American Computing and Information Technology Center and NASA Grant NAG 9-1169 to the first author.

## 5. References

[1] Bates, M. J., "Subject access in online catalogs: A design model", *Journal of the American Society for Information Science, 37(6),* 1986, pp. 357-386.

[2] Brin, S & Page, L., "The Anatomy of Large Scale Hypertextual Web Search Engine", in *Proceedings of the Seventh World Wide Web Conference*, 1998.

[3] Chakrabarti, S, Dom, B. & Indyk, P., "Enhanced Hypertext Classification Using Hyper-Links", in *ACM SIGMOND International Conference Management of Data,* 1998.

[4] Chakrabarti, S., Dom, B., Kumar, S. Raghavan, P., & Rajagopalan, S. A. Tomkins, D. Gibson, & Kleinberg, J., "Mining the Web's Link Structure", *IEEE Computer, 32(3),* 1999, pp. 98-113.

[5] Cios, K., Pedrycz, W. & Swiniarski, R., *Data Mining Methods for Knowledge Discovery*, Kluwer Academic Publishers, 1998.

[6] Gibson, D., Kleinberg, J, & Raghavan, P., "Inferring Web Communities from Link Topology", in *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia,* 1998.

[7] Gibson, D., Kleinberg, J, & Raghavan, P., "Clustering Categorical Data: An Approach Based on Dynamical Systems", in *Proceedings of the 24th International Conference on Very Large Databases*, 1998.

[8] Hendley, R. J., Drew, N. S., Wood, A. M. & Beale, R., "Narcissus: Visualizing information", in *Proceedings of IEEE Information Visualization*, 1995.

[9] Ingwerson, P., & Wormell, I. "Improved subject access, browsing and scanning mechanisms in modern online ir", in *Proceedings of ACM SIGIR*, 1986.

[10] Kazman, R & Carriere, J., "An Adaptable Software Architecture for Rapidly Creating Information Visualizations", in *Proceedings of ACM Graphics Interface*, 1996.

[11] Kleinberg, J., "Authoritative Sources in a Hyperlinked Environment", in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[12] Maron, M. E., & Kuhn, J. L. "On relevance, probabilistic indexing, and information retrieval", *Journal of the Association for Computing Machinery, 7(3),* 1960, pp. 216-244.

[13] Orlean, K., "Visualizing Websites Using a Hierarchical Table of Contents Browser: WebTOC", in *Proceedings of the 3rd Conference on Human Factors and the Web,* 1998.

[14] McCahill, M. P. & Erickson, T., "Design for a 3D spatial user interface for Internet Gopher", in *Proceedings of AACE ED-MEDIA 95*, 1995.

[15] Munzner, T. & Burchard, P., Visualizing the structure of the World Wide Web in 3D hyperbolic space. Available at http://www.geom.umn.edu/docs/weboogl/, The Geometry Center, University of Minnesota.

[16] Wong, P. "Visual Data Mining", *IEEE Computer Graphics and Applications, 19(1),* 1999, pp. 72-81.