A Visualization System using Data Mining Techniques for Identifying Information Sources on the Web

Richard H. Fowler, Tarkan Karadayi, Zhixiang Chen, Xiaodong Meng, Wendy A. L. Fowler
Department of Computer Science
University of Texas – Pan American

Abstract

The World Wide Web provides great opportunities for serving as a repository of knowledge, yet presents great challenges in accessing this knowledge. Limitations of current search technologies are well known and include shortcomings in information filtering and authentication of sources. Recently, the application of data mining and information visualization techniques in Web search has provided new tools to complement conventional search methodologies. The Visual Analysis System (VAS) was developed to couple emerging successes in data mining with information visualization techniques in order to create a richly interactive environment for information retrieval from the Web. VAS's retrieval strategy operates by first using a conventional search engine to form a core set of retrieved documents. This core set is expanded by crawling from links in these documents to form a much larger set. The link topology of the larger set is then examined using data mining techniques to identify pages likely to be both relevant to the query and reliable as sources of information. Information visualization techniques are used to display the filtered set in a form amenable to the use of perceptual processes to filter and guide further user directed search.

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 a "quality assessment" scheme, to mention a few elements.

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 [Ingwerson & Wormell, 1986]. Retrieving information that meets users' information needs is an iterative process, and techniques which explicitly incorporate users' judgments, such as relevance feedback [Maron & Kuhn, 1960], 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 [Bates, 1986].

1.1 Data Mining

Data mining refers to the analysis of typically very large data sets to reveal underlying structures or relations [Cios, Pedrycz, & Swiniarski, 1998]. 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 amount of work directed at the integration of textual content and link information for the purpose of organizing and searching in 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 [Chakrabardi et al., 1999] 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. Research group developed an algorithm to compute the hubs and algorithms called HITS (Hyperlink Induced Topic Search) algorithm. Beginning with a search topic, specified by one or more query terms, the HITS (Hyperlink Induced Topic Search) algorithm applies two main steps; a sampling component, which constructs a focused collection of several thousand 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 interactive procedure. Kleinberg and colleagues have continued to

refine this basic approach [Gibson, Kleinberg, & Raghavan, 1998a; Gibson, Kleinberg, & Raghavan, 1998b; Kleinberg, 1998].

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 [Chakrabarti, Dom, & Indyk, 1998] embodies one approach to this problem, making use of robust statistical models such as Markov 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 [Brin & Page, 1998]. 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 Information 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 [McCahill & Erickson, 1995; Hendley et al., 1995; Munzner & Burchard, 1996].

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 participate in 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 no longer be automated [Wong, 1999]. VANISH [Kazman & Carriere, 1996] 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 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 Visual Analysis System: Mining and Visualizing WWW Structure and Information Sources

The Visual Analysis System (VAS) was developed to couple emerging successes in data mining with information visualization techniques in order to create a richly interactive environment for information retrieval from the Web. VAS's retrieval strategy operates by first using a conventional search engine to form a core set of retrieved documents. This core set is expanded by crawling from links in these documents to form a much larger set. The link topology of the larger set is then examined using data mining techniques to identify pages likely to be both relevant to the query and reliable as sources of information. Information visualization techniques are used to display the filtered set in a form amenable to the use of perceptual processes to filter and guide further user directed search. More detail on VAS is available in [Karadayi & Fowler, 2000].

The Visual Analysis System is made up of several components. Initially, VAS receives a user's query as a keyword or keyword set. The system then submits the query to other search engines, such as Alta Vista, retrieving the best matches (usually 200) to serve as a starting point for VAS directed search. Visual Analysis System then starts its own search by following links in these pages to retrieve a second level set (typically 2000-3000). Links in these pages can also be followed to retrieve a third set of pages. Typically, pages retrieved in the second level set contains a large number of links to pages within the retrieved sets, as would be expected starting from a single query. As page sets are retrieved, a graph showing connections among pages is formed and displayed, as shown in Figure 1. This dynamically formed and displayed graph is the primary means of interaction with the system.

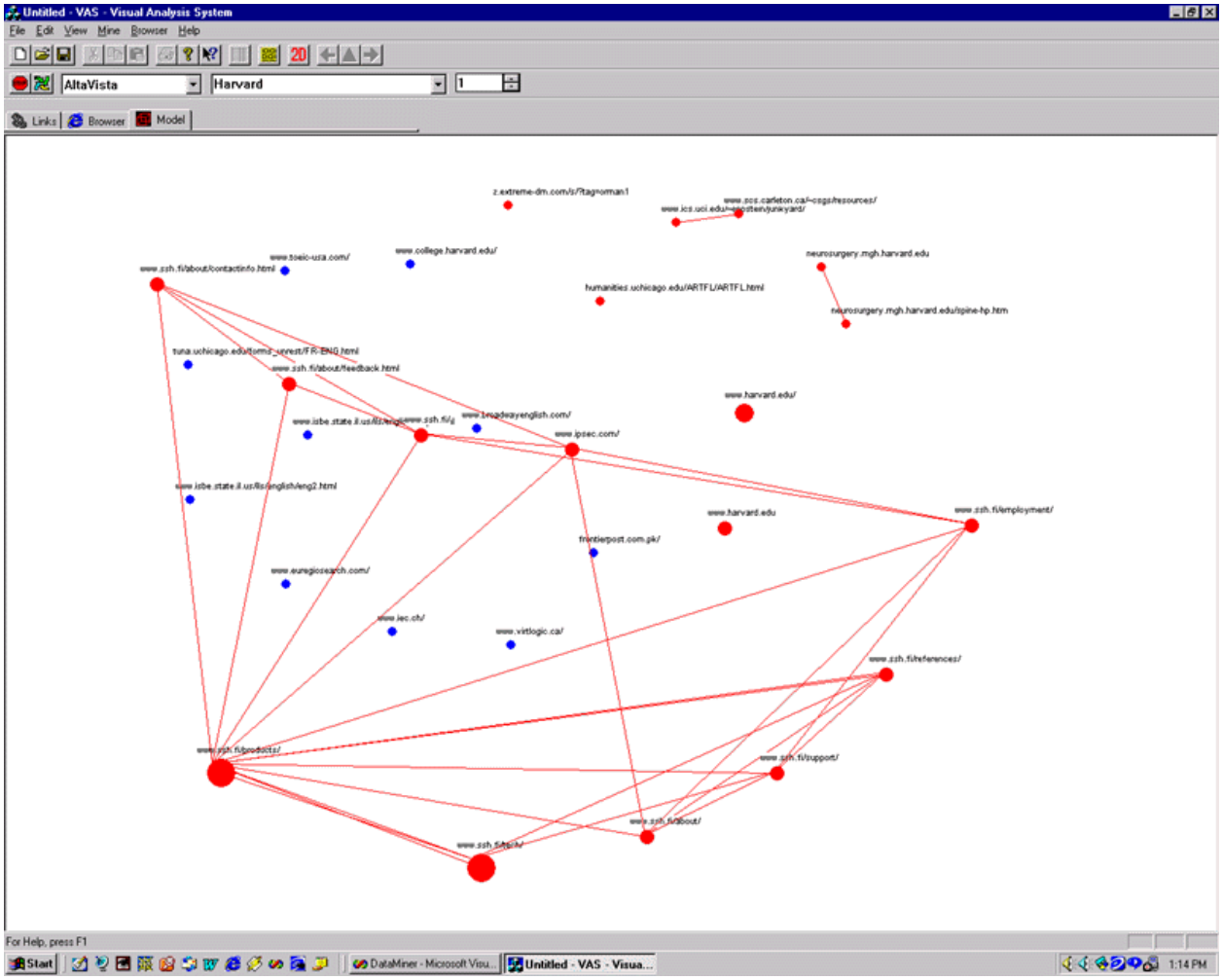

Figure 1. A typical Visual Analysis System Display. The 30 pages of highest authority found in a two level search are displayed. Size and color are used to indicate degree of authority and connectivity. Users can select a displayed document and go directly to the document while search and analysis continues. Users can also select to display hub documents.

Other researchers [cf., Kleinberg, 1998] have characterized pages as *authorities* (many links to the page), which provide the best source of information on a given topic, and *hubs* (many links from the page), which provide useful lists of possibly relevant pages. VAS first distills a large World Wide Web search topic to a size that makes sense to the human user: a means of identifying the topic's most definitive or authoritative Web pages. That is, not only a set of relevant pages is located, but also those relevant pages of the highest quality are identified. VAS exploits the fact that the Web consists not only of pages, but also hyperlinks that connect one page to another.

Just computing authorities and hubs in a query is insufficient, however, if the iterative nature of the information retrieval process is to be accommodated. VAS provides a dynamic visual summary of the systems data mining and search results for interaction. This enables users to get the information they need, make sense of it, and reach decisions in a relatively short time. It takes advantage of the human's natural pattern recognition ability by creating visual representation of the data mining results the system computes. By providing a 2-dimensional view, users can navigate through visual space to select and manipulate objects easily. Interaction with the resulting visualization allows users to focus their attention directly on the results of the query in an immediate and compelling way.

Determining the documents displayed to users as authorities and hubs is based on a graph theoretic analysis of the hyperlink structure of the page sets retrieved in the initial query to the search engine and subsequent crawling of links in the sets. Loosely stated, authorities are pages with many links to them (high in degree) and hubs are pages which supply many links to other documents in the same semantic domain (high out degree). Given any subset S of nodes, the nodes induce a subgraph containing all edges that connect two nodes in S. The algorithm starts by constructing the subgraph in which AVS will search for hubs and authorities. To construct the subgraph, the system first uses the query terms to collect a root set of pages, which is about 200, from an index based search engine such as Alta Vista. This set of pages does may not necessarily contain authoritative pages. However, since many of these pages are presumably relevant to the search topic, it is expected that some will have links to most of the prominent authorities. Now AVS can expand the root set into base set by including all the pages that the root set pages link to, and all pages that link to a page in the root set, up to designated set size. This approach follows the intuition that the prominence of authoritative pages derives typically from the endorsement of many relevant pages that are not in themselves prominent. Since links between two pages with the same Web domain frequently serve a purely navigational function, and thus do not confer authority, we need to delete all links between pages with the same domain from the subgraph induced by the base set, and then apply the remainder of the algorithm to this modified subgraph.

Giving a concrete numerical interpretation to authorities and hubs, we can extract good hubs and authorities from the base set. After these calculations, we can update the authority and hub weights as follows: If many good hubs point to a page, it increases its authority weight. Intuitively, the pages with large weights represent a very dense pattern linkage, from pages of large hub weight to pages of large authority weight. System outputs a short list consisting of the pages with the largest hub weights and the pages with the largest authority weights for the given search topic. A text based display, which is not shown, is available to the user to present this state of the analysis.

The visual representation, shown in Figure 1, is constructed by using a user-selectable criterion for the level of authority or hub to display to filter the complete sets, e.g., 30 highest ranked authorities in this example. Underlying link structure is used to position nodes, and size and color reflect authority, in this example, or hub rank. Connections are only shown when there is a direct link among pages displayed.

3. Acknowledgements

Work on this project has been supported by NASA grants NAG9-842 and NAG9-871.

4. Conclusions

For broad topics on the Web the amount of relevant information is growing extremely rapidly, making it difficult for individual users, and even individual search engines, to filter the available resources. To deal with this problem a way to distill a topic for which there may be millions of relevant pages down to a representation of very small size is needed. It is for this purpose that VAS uses the notion of authoritative and hub sources based on the link structure of the Web. The goal is to produce results that are of as a high a quality as possible in the context of what is available on the Web globally. For VAS the underlying domain is not restricted to a focused set of pages, or those residing on a single Web site.

VAS infers global structure without directly maintaining an index of the Web or its link structure. It requires only an interface to any of a number of standard Web search engines and uses techniques for producing enriched samples of Web pages to determine notions of structure and quality that make sense globally. This helps in dealing with problems of scale for handling topics that have large representations on the WWW. VAS discovers authoritative pages, and in fact identifies a more complex pattern of social organization on the WWW, in which hub pages link densely to set of thematically related authorities.

The results are useful from a number of prespectives. Analysis of the link structure of the WWW suggests that the on-going process of page creation and linkage, though difficult to understand at a local level, results in structure that is considerably more orderly than is typically assumed. Thus, it gives us a global understanding of the ways in which independent users build connections to one another in hypermedia that arises in a distributed setting. It also suggests some of the types of structured, higher-level information that designers of information discovery tools may be able to provide both for, and about, user populations on the Web.

Finally, VAS is user-centric, taking user's needs into account by allowing them to interact with the information contained in large numbers of documents. The visualization process is an integral part of the overall process. VAS focuses on visualization and visual interfaces in support of information retrieval and data mining.

5. References

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

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

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

[Chakrabarti, et al., 1999] Chakrabarti, S., Dom, B., Kumar, S. Raghavan, P., & Rajagopalan, S. A. Tomkins, D. Gibson, J. Kleinberg, *Mining the Web's Link Structure,* IEEE Computer, August 1999.

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

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

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

[Hendley et al., 1995] Hendley, R. J., Drew, N. S., Wood, A. M. & Beale, R. (1995). Narcissus: Visualizing information. *Proceedings of Information Visualization,* 90-97. IEEE.

[Ingwerson & Wormell, 1986] Ingwerson, P., & Wormell, I. (1986). Improved subject access, browsing and scanning mechanisms in modern online ir. *Proceedings of ACM SIGIR,* 68-76. ACM.

[Karadayi & Fowler, 2000] Karadayi, T. & Fowler, R., Data Mining and Visualization Engine Based on Link Structure of WWW, Technical Report CS-00-20, University of Texas - Pan American, Edinburg, TX USA, http://www.cs.panam.edu/ /TR/cs-tr.html

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

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

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

[McCahill & Erickson, 1995] McCahill, M. P. & Erickson, T. (1995). Design for a 3D spatial user interface for Internet Gopher. *Proceedings of ED-MEDIA 95,* 39-44. AACE.

[Munzner & Burchard, 1996] Munzner, T. & Burchard, P. (1996). 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.

[Wong, 1999] Wong, P. *Visual Data Mining,* IEEE Computer Graphics and Applications, September/October 1999.