A Web Based Visualization for Documents

Manu Konchady, Ray D'Amore, Gary Valley,
Mitre Corporation,
1820 Dolley Madison Blvd., M/S W431,
McLean, VA 22102.

Abstract

Locating relevant information on the Internet is a challenging task. The use of commercial search engines such as Altavista and Hotbot simplify the problem [2], however, an user may scan many pages before finding a relevant document. This process is tedious and users may miss important documents if insufficient information is retrieved from a search engine. We have developed a set of tools to collect, organize, and display information from the Internet. Our tools range from experimental visualizations to a working prototype with multiple users. The tools reduce the time to obtain relevant information by automating the collection and organization of data.
Keywords: Information search and retrieval, Digital libraries, Group and Organization Interfaces

Introduction

In our current prototype, we built a system for users to submit queries and search a collection of existing queries. Each user created a folder to maintain a set of queries. A query consisted of a string of keywords or phrases with optional boolean operators. The query was submitted without modification to the search engines. Up to 6 search engines could be used concurrently. The results from all engines were collected and built into a composite list. Duplicates were eliminated and the results were categorized based on the URL address. The last two characters, if available, were used to determine the country. Other characters in the URL were used to locate the university, company, or government organization. Based on the information extracted from the URL, the query results were organized by category.

In addition to collecting and organizing results, capabilities to cache, monitor, and mail a list of URLs were provided. The cache for a query consisted of a set of URLs deemed to be relevant by an user. An URL could be monitored for changes and the owner of the folder notified by e-mail when the content of the page associated with the URL changes. An URL or a list of URLs can be mailed with an annotation to other users.

Document Visualization

While a table driven approach for the prototype described above was easy to understand, we used visualization to display an overview of document relationships and quickly locate relevant documents.

Our visualization was based on a cube (Fig.1). The axes of the cube represented either keywords or documents. The user assigns a document or keyword to one or more axes. The ten most prominent documents of the collection are shown followed by the ten most frequent keywords in the documents. The keyword frequencies are directly proportional to the number of occurrences of the keyword in all documents and inversely proportional to the number of documents in which it occurs.

An axis is assigned a keyword or document by drawing a line from the axis label to the keyword or document. If one axis is used, then the documents will be plotted on a line. If two axes are used, then documents will be plotted on a plane. If three axes are used, then documents will be plotted in the cube.

Documents are plotted as dots in the cube (Fig. 2). The position in the cube is based on the similarity [4] to the axis document or keyword. In the example below, the word Mitre has been assigned to the z-axis. All documents have an (x,y,z) location within the cube. The z value is determined by the number of times the word Mitre occurs in the document. A higher z value implies that the word Mitre occurs more often. The x and y values are computed similarly. If a document represents an axis, then all the positions of other documents are calculated based on the similarity to the axis document. The similarity between documents is computed using the cosine similarity measure. A threshold can be used to limit the number of dots shown in the cube. A higher threshold can be used to show documents away from the origin. Non-zero position (x,y, and z) values must exceed the the threshold to be displayed in the cube. Buttons to zoom, rotate, and position the cube in 3D can be used to study relationships. When the mouse is placed over a document in the cube, a small window in the upper region of the cube displays the title. In the example above, the mouse is placed near the top of the z-axis, and the selected document is colored white. The title for this particular document is Intelligent Human-Computer Interaction

at Mitre. If the user clicks on the red box before the title, a new window is shown with the text of the document. From the window above (Fig.3), the user can load the web page in a browser using the 'link to page' button or make the documen one of the 10 axes documents. Additional keywords can be chosen for the axes by pressing the 'New Keywords' button in the previous image. When a new keyword is assigned to an axis, the entire document collection is scanned for the keyword and a list of documents in descending order of frequency is built. The new keyword is added to the list of 10 keywords and the oldest keyword is replaced.

When a new document is added to the list of 10 axes documents, the similarity matrix is used to compute locations for th new document in the cube. A document which has a higher similarity to an axis document is positioned proportionally away from the origin. All position values in the cube are normalized.

People Visualization

Within Mitre, there are a number of groups studying various information technologies. This tool is an initial effor towards understanding the common interests among groups of users. The visualization we developed also called the 'people map', shows links between people having a common interest. The topics of interest are shown in a window.

For our experiment, we compiled a list of queries which were submitted to search engines such as Altavista and Hotbot. For each query, we assigned an owner and built a vector to represent the query. The vector consisted of keywords from the query minus stopwords. Phrases were maintained in the vector. The vectors were improved by using keywords from the titles of documents which were manually determined to be relevant (cache documents).

Next, we ran a clustering algorithm against the query vectors. The resulting set of clusters were used to build the 'people map'. For every cluster, the member query vectors represented a collection of queries with some common theme. This common theme was extracted from the centroid for the cluster. For every query vector pair in the cluster with a different owner, a relationship between the two owners was established. The basis for the relationship was the common theme of the cluster. Given the information from the clusters, we built a people map (Fig.4).

A node of the graph represents a person. The photograph of the person is shown at the node. A line is drawn between any two persons having a common interest. When the mouse is placed over the blue circle in the middle of the link, the common themes are shown in a message box. The color of the ball changes to red to correspond with the message box. In this case, the common topics of interest are data mining, graph drawing, and xml.

We wrote a simulated annealing algorithm [1] to compute positions for the nodes on the graph in such a manner that number of edges which intersect are minimized and the nodes were approximately equi-distant. For the small graph shown above, it is possible to manually construct a graph which looks attractive. When the number of nodes is large (> 15), it becomes hard to draw the graph and automated techniques must be used. The simulated annealing technique is well known and has been used for drawing graphs.

All the code for the visualizations was written in Java and ran as applets on a web browser. We have tested the software on UNIX and Windows platforms.

Link Analysis

We developed link analysis tools (Fig.5) to study the relationships between a set of documents retrieved from search engines. Link analysis [3] can provide insight into the structure of documents. From a collection of documents, we compiled a list of links and stored them in a table. We analyzed the links and determined which documents contained references to common documents within the collection and outside the collection. A 'red' link between two documents indicated a reference from a document in the collection to another document in the collection. A 'blue' link between two documents indicated a reference in two documents to a common document outside the collection. When document 'A' contained a link to document 'B' and document 'B' contained a link to document 'C', where documents 'A', 'B', and 'C' were all located in the collection, a 'green' link was drawn between documents 'A' and 'C'

The new documents label indicates the number of documents available using the links within the current collection of 100 documents. The similarity selector can be used to limit the number of links shown by requiring a minimum similarity between any two documents before plotting a link between the two documents. Domains can be selected to limit the documents shown on the circumference of the circle. Keywords can be entered to select a group of documents which will be plotted on the circle.

Conclusions

Current developments include building a temporal view of the people map. This would mean we track the history of query submissions to identify new topics of interest and aging common interests. Periodically, a snapshot of the query vectors will be stored. A new set of clusters will be built at each instance. The new and old links between people will be noted. Th age of links between people can be represented by color. If frames are built for the people map at each instant, then an animation showing the changes can be made.

References

[1] Davidson R. and Harel D., Drawing Graphs Nicely using Simulated Annealing, ACM Transactions on Graphics, Vol.15, No.4, 1996, pp 301-331.

[2] DeJesus E., The Searchable Kingdom, Byte Magazine, 1997.

[3] Freeman L.C., Visualizing Social Networks, AAAI Fall Symposium on Artificial Intelligence and Link Analysis, 1998.

[4] Salton G. and McGill M.J., Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

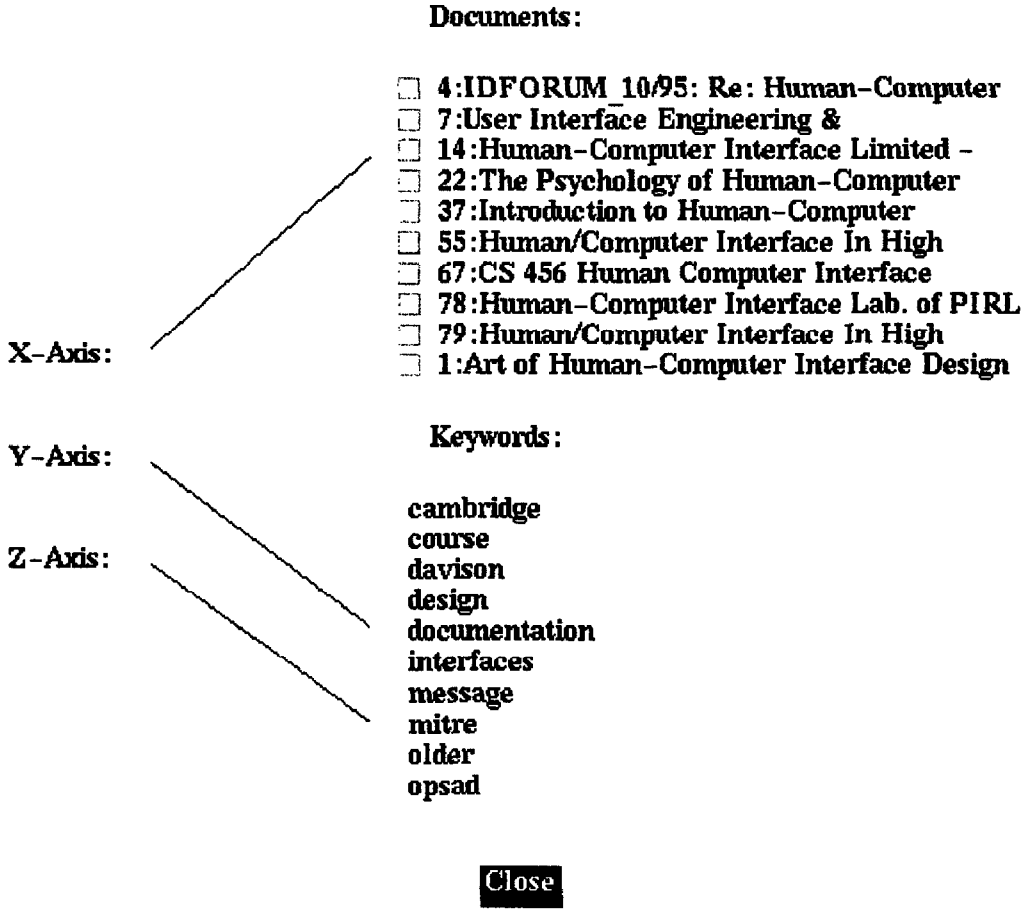

Fig. 1 Axes dimension definition

Fig. 2 A plot of documents in a cube

URL: Link to Page

Title: doc 76: Intelligent Human-Computer Interaction at MITRE

Text:

Intelligent Human-Computer Interaction
at the MITRE Corporation

Overview
Projects
People

Overview

[INLINE] When we say "HCI" at MITRE, we mean a number of different things. For instance, the group you're about to meet is not a human factors group (although MITRE has a very good one). Rather, this group is interested in inserting intuitive and conversational interface capabilities into multimodal environments, using "intelligent"

Make Axis: No

Fig. 3 Text from selected document

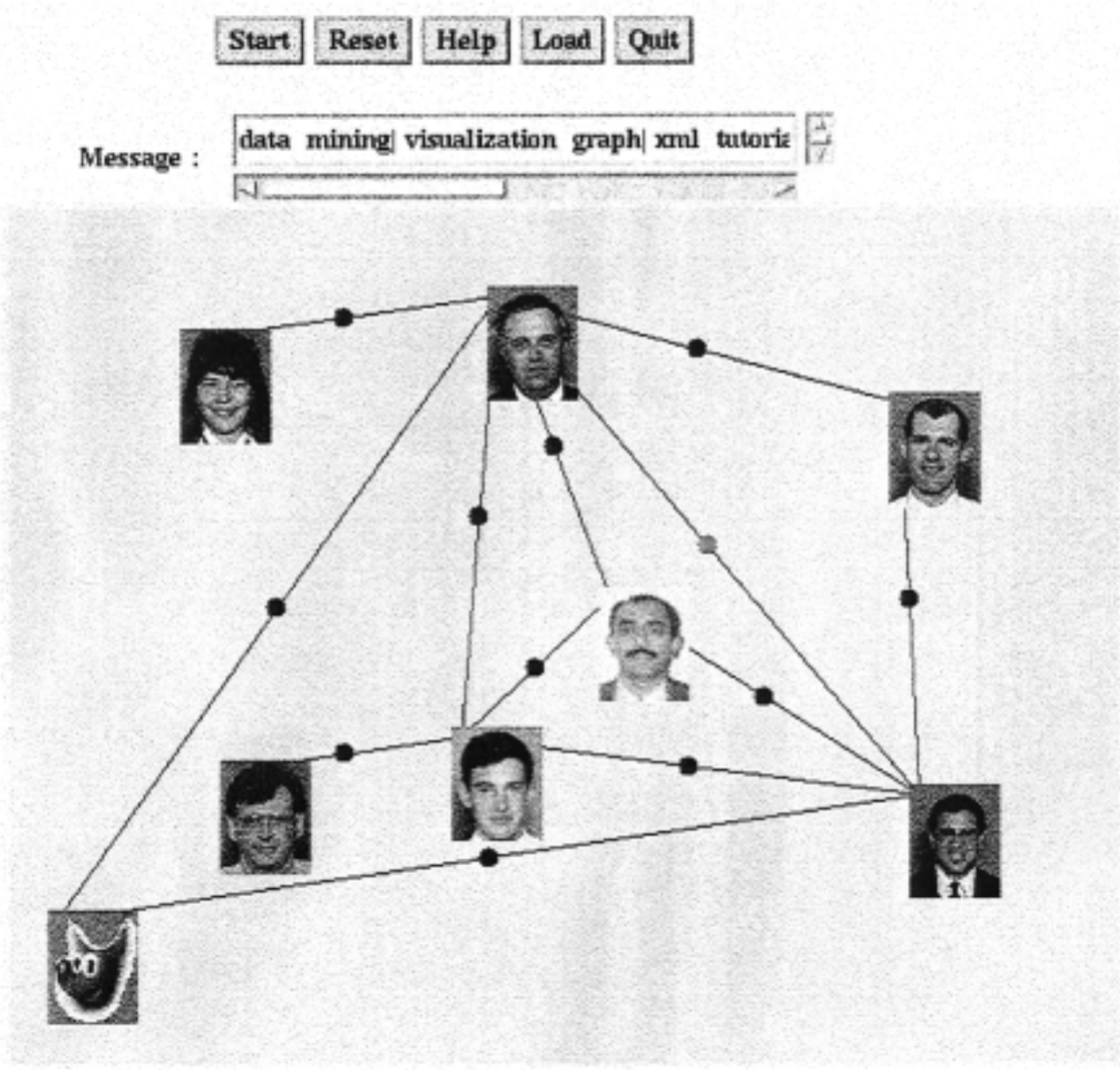

Fig. 4 Relationships between people based on queries

Fig. 5 Link analysis for a set of 100 documents

NPIV 98 Bethesda MD USA
Copyright ACM 2000 1-58113-179-8/00/1...$5.00