Free-link Topology Navigation on Statistical Table Objects: Metadata Schema and User Interface

Xiangming Mu and Gary Marchionini
Interaction Design Lab, School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360
Phone: (919)962-8366; Fax: (919)962-8071
Email: {mux, march}@ils.unc.edu

- Type: Full paper
- Demo as well?: yes
- Author(s): see above
- Address: see above

Abstract

A table object, defined and described in XML, is treated as the basic unit in a free-link topology that enables multiple inheritances for each node. In this paper, a novel table navigation system called TableHunter is introduced. Each table object is presented in the TableHunter as a node of a map structure, and supports a context+focus view.

1. Introduction

Statistical data is an important type of information that is essential to the day-to-day lives of citizens. The table is the common format to represent such statistical data. An increasing amount of statistical tables is now moving to the WWW due to the fast development of the Internet. For example, in the FedStats (www.fedstats.gov) website, statistical information including tables, surveys, reports, and analysis from more than 100 federal agencies is available for public access. One of the agencies, the Energy Information Administration (EIA, http://www.eia.doe.gov), maintains more than ten thousand web pages and tables covering various fuels (petroleum, natural gas, coal, electricity, nuclear, and all other energy types) in terms of prices, usages, forecasts, analysis, reports, policies, and historical data. In order to provide high quality services for the average information seeker, who is not a statistical data expert, government agencies are challenged to develop interfaces to help novice users find the information and tables they seek.

There are currently two major approaches in the table finding and retrieval process: free table search and hierarchical table navigation. Free table search treats tables as documents and employs current text information retrieval technologies to find tables. The table level metadata such as title, description, date, authors, and column/raw titles, are utilized to aid the search engine in the ranking process. Hierarchy table navigation enables users to browse a "tree" structure that represents different levels of concepts. Tracing the tree will narrow down the concept from general descriptions to concrete definitions until the target tables are found. For example, to get the most recent gasoline import data, the user has to go to the EIA website, the root of the tree hierarchy, and follow fuel->petroleum->gasoline->recent weekly data (imports) to get the target table.

Free search is effective and efficient only if a very specific query can be easily formed, otherwise the target tables are just "buried" in the large number of "noisy" returned web pages. This is partly the reason why most statistical agencies have adopted hierarchal navigation structures to help users narrow down their searching step by step. However, one disadvantage for hierarchy navigation is that it builds on the promise that users fully understand the hierarchy; otherwise users might get confused and fail to reach the right tables and many rounds of "trial-and-error" might be expected before the needed tables are found or the users give up. Another disadvantage for hierarchy tables is that the rigidity of hierarchical structure creates some level of inflexibility: it is virtually impossible for a user to alter the current hierarchy or introduce an alternate structure.

In this paper, a novel free-link topology structure (FTS) is proposed to facilitate table searching tasks. Instead of classifying various "nodes" as roots, "hubs", or "leaves" and forming a hierarchy "tree" structure, the FTS treats each node as an equal "table object" and connects each related table through links. "It is a web not a tree". In this structure, each "table object" is self-contained and self-explainable: it maintains not only the raw table information, including the table itself, related webpages, documents, or graphics, but also "neighbor table objects" in terms of links. The advantage of adopting table objects is that the data is de-coupled from its presentation format and thus flexibility is increased.

An application level system called TableHunter was developed based on the FTS. Part of the EIA hierarchy table index system was analyzed and implemented in the TableHunter prototype. Each table object is represented as a node and links are represented as lines between nodes. TableHunter provides a means for fast "trial-and-error" operations and thus speeds table hunting tasks.

This paper is organized as follows: after defining the free-link topology structure (FTS), we present a schema to accommodate the table objects of the FTS in XML. Then, the user-interface and functionalities of the TableHunter are presented to demonstrate the FTS of EIA tables. Finally, a brief discussion and plans for future research are given.

2. Related work
Considerable research has been conducted to help users understand online tables [1-6]. BEST [1-2], formerly called the table browser, provides frozen column and row headings in viewing large online tables. Metadata is integrated and presented with tooltips, submenus, and active links to aid table comprehension. FOCUS [3] uses a focus + context technique to permit viewing a large table without scrollbars. Table Lens [4] suggests a number of user operations to view and sort tables. Yang et al. [5] proposed a framework for web table mining using HTML tags and attribute-value pairs to extract online tables.

Another related research domain is web page navigation and visualization using a "map" structure. Each web page in this map is treated as a node and entire large web sites are viewed as a map. Durand et al. [6] listed problems that currently exist for navigating a number of documents on the Internet: lack of physical context, lack of organizational paradigms, need for orientation, and need for link summaries. They suggested that maps are the most basic visual orientation tool and present ten requirements for a map structure interface to guide the MAPA design. MAPA is a map navigation system that extracts a hierarchical structure from a web site. The front-to-back arrangement layout and colored grouping are utilized to represent the links and clustering across pages.

InXight (www.inxight.com) is a commercial tool that visualizes the web hierarchy in the form of a "tree" structure. Each node can be moved around and zooming functions are supported. Components near the center become magnified while nodes positioned away from the center decrease in size. Similar technology was presented in other research [7, 8] and has been referred to as Focus + Context technology [9].

However, the visualized map structure for web nodes is completely built on a hierarchy tree structure, which limits its applications. Limited visualization research has been conducted on the more natural free-link topology. In the above approaches, the basic unit for the map structure is a web page. Tables can only "belong to" a web page and cannot be identified independently. In this paper, we separate tables from the web pages and treat tables as well as webpages as the basic unit of our free-link topology. Each of these basic units is defined as an object. A metadata schema is suggested to "wrap" the table object and to add "link" attributes.

3. Table Object and XML Schema
A table is defined as an object having the following attributes:

- Table description attributes, such as title, author, date, size, format (PDF, HTML or text), number of rows, fields, and data types.
- Table relational attributes: number of links, link URLs, and link Titles.
- Table contextual attributes: web page URLs

Table description attributes provide descriptive information about a table, its columns/rows, and table cells. Table relational attributes describe the associated sibling tables. Table contextual attributes describe the webpage(s) with which the table is associated. All these attributes are wrapped into a table object and represented in XML.

A structured metadata schema is employed in this paper in terms of the XML schema definition. "*Structured metadata accompanying statistical tables from different NSI's can be compared and their level of harmonization assessed automatically*" [10]. The following is a sample schema.

```
<Program xmlns = http://www.ils.unc.edu/Table/ExampleSchema>
<DescriptionInfo>
        <Title>…</Title>
        <Author>…</Author>
        <Date format=TableDate>…</Date>
        <Size >…</Size>
        <Format value="PDF, HTML, TEXT">…</Format>
        <RowNumber format=TableNumber>…</RowNumber>
        <DataType value="TableString, TableNumber">…</DataType>
        <Field format="TableString" >…</Field>
        <Field format="TableString" >…</Field>


        …
</DiscriptionInfo>
<RelationInfo>
        <LinkNumber format="TableNumber">…</LinkNumber>
        <Link>
                <LinkURL format="LinkString"> … </LinkURL>
                <LinkTitle format="TableString">…</LinkTitle>
```

```
        </Link>
        <Link>
        ...
        </Link>
        ...
</RelationInfo>
<ContextInfo>
        <PageNumber format="TableNumber">...</PageNumber>
        <Page>
                <PageURL format="LinkString"> ... </PageURL>
                <PageType value="Http,Ftp,Email">..</PageType>
        </Page>
        </Page>
        ...
        </Page>
        ...
</ContextInfo>

...
</program>
```

## 4. TableHunter: Interface and Functions

The TableHunter is a map-based user interface that implements table objects as nodes. The goal is to provide a visual tool to help users in their table hunting tasks. TableHunter employs a more natural free-link topology that supports multiple inheritances and is distinctly different from traditional hierarchy or tree structures. The map is utilized to show both focus and context.

To illustrate the user interface, data from Energy Information Administration (EIA) was employed as an example. Parts of the web pages and tables extracted from EIA were transformed into "table objects" in accordance with the XML schema defined in the previous section. These table objects were parsed by the TableHunter Java applet and then were presented as a map based on attributes like the table title and links. The starting user-interface is the EIA homepage, which is displayed as the center node in red (fig. 1a). A single click on any individual node in the map will drag that node to the center, change its color to red, and refresh the map to build new view (fig.1b). The corresponding webpage, if the current central node has the URL, will be displayed on user's web browser automatically.

On the top of the map is a URL information field that dynamically displays the URL of any node when users hover the mouse over that node. Beneath the map is a series of "map viewing" manipulation controls (fig. 2) that help users interact with the map. The "Turn" button enables users to adjust their viewpoint of the map by a specific degree in a counterclockwise direction. The default value for the turning degree is 45^0; however, users may select other degrees in the "angle" drop-list to get various granularities. "Zoom In" and "Zoom Out" buttons make the map larger or smaller and address the context+focus.
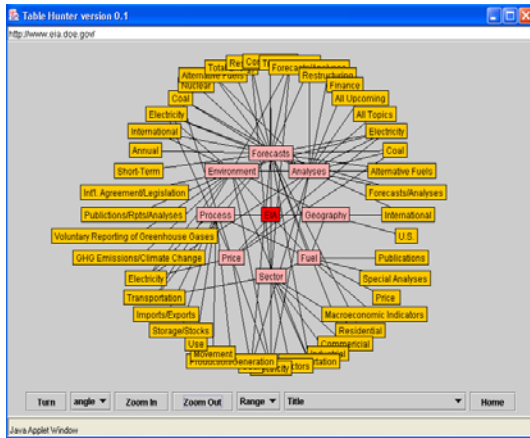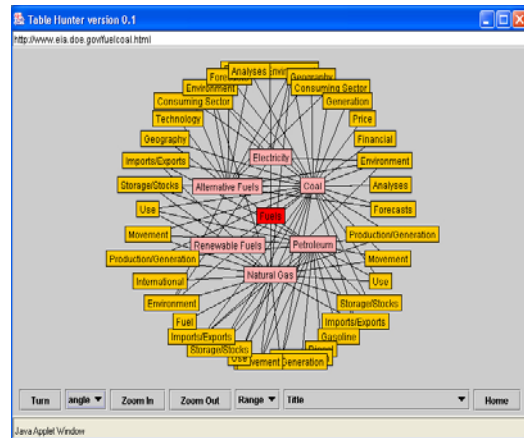
(a)   (b)

Figure 1. User-interface of the TableHunter: Tables and Links in the EIA. The red node in the center represents the current table object. A single click on a node will bring that node into the center with red color. All the links related to the node are also updated and re-positioned in the map.

Figure 2: manipulation controls for TableHunter

The default TableHunter presents three levels of nodes: the current node is located in the center and represented in red. The direct links revealed from this current "table object" are treated as neighbors and represented in pink. The links from the neighbors are treated as a third level and represented in yellow. Users are allowed to present more than three levels by making selections from the drop-list entitled "Range" (fig.2) An alphabetic list of all available nodes is also available in the "Title" drop-list, which enables users to bring that node into the center of the map with a single click. The "Home" button will reset the map to the default map view. In the sample case, the default page is the home page of the EIA.

5. Discussion and Future Research

The prototype of the TableHunter was demonstrated for faculty members and graduate students in the School of Information and Library Science. Revisions were made and the revised version of the TableHunter was then informally tested by some students. Students conducted table searching tasks and then participated in a discussion about their experience with TableHunter. From our observations and the talks, we found that users were excited by the interface and its interactive functions. All the students claimed that it was easy to learn and use. Some students mentioned they preferred this way of displaying many nodes in a very small space and one student commented, "[TableHunter] made me quickly understand the relationships among these concepts and helped me find the right track for the table". Compared to the homepage of EIA which can only present eight links, the default interface (three levels) of the TableHunter for EIA presents 53 nodes.

TableHunter is implemented in Java and presented as a Java applet, which means it is can be embedded into the web browser directly for access. TableHunter may also be integrated with other online table browsing tools such as the BEST [1-2]. By adopting the current web structures

and extract links, TableHunter does not change the web structure. The TableHunter will be demonstrated at the conference.

As a prototype for visualizing table objects and their links, TableHunter needs formal evaluations in terms of both usability and system performance. Current research under consideration involves developing provisional means that would automatically authorize table objects from the database schema.

Acknowledgements

References:
1. Mu,X. & Marchionini,G.: "*An Architecture and Prototype Interface for an Online Statistical Table Browser*", Proceedings of the 64TH annual meeting of the American Society for Information Science and Technology.Vol. 38, 2001. pp156-170.
2. Marchionini,G. & Mu,X., (in press) "User *Studies Informing E-Table Interfaces*", Information Processing & management.
3. Spenke,M., Beilken,C., and Berlage,T., "*Focus: the interactive table for product comparison and selection*", proceedings of UIST'96, pp. 41-50, 1996
4. Rao,R. and Card, S., "*The table lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information*", proceedings of CHI'94, pp.318-322.
5. Yang,Y. and Luk, W.S., "*A framework for web table mining*", WIDM'02, November 8,2002.
6. Durand, D., and Kahn,P. , 1998, "*MAPA: a system for inducing and visualizing hierarchy in websites*". Proceeding of the HyperText 98.
7. Robert, L. and Lecolinet, E. , 1998, "*Browsing hyperdocuments with multiple focus+context views*", Proceedings of the HyperText 98,
8. Lamping J. and Rao R. The hyperbolic browser: "*A focus+context technique for visualizing large hierarchies*". Journal of visual languages and computing, 7:33-55,1996.
9. Mukherjear, S., and Hara, Y., 1997, "*Focus+Context views of world-wide web nodes*", Proceedings of the HyperText 97
10. Petrakos, M., Farmakis, G., Koumanakos, G., Bisiela, P. "*A structured approach to a statistical metadata base*", NTTS 2001 International Conference in Official Statistics, Creta 2001.