BROWSING LARGE ONLINE DATA WITH QUERY PREVIEWS

Egemen Tanin[*]
egemen@cs.umd.edu

Catherine Plaisant
plaisant@cs.umd.edu

Ben Shneiderman[*]
ben@cs.umd.edu

Human-Computer Interaction Laboratory and Department of Computer Science[*],
University of Maryland, College Park.

ABSTRACT

Companies, government agencies, and other types of organizations are making their large data available to the world over the Internet. Current front-ends rarely give users information about the contents of the data. This leads many users to waste time and network resources posing queries that have either zero-hit or mega-hit result sets. Query previews form a novel visual approach for browsing large data collections. Query previews supply data distribution information to the users and give continuous feedback about the size of the result set as the query is being formed. On the other hand, initial designs of query previews used only a few pre-selected attributes. The distribution information was displayed only on these attributes. Unfortunately, many data sets are formed of numerous tables and attributes. This paper introduces a generalization of query previews. We allow users to visually browse multiple tables and attributes from a data set using a hierarchical browser. Any of the attributes can be used to display the distribution information, making query previews applicable to many public online data sets.

KEYWORDS

Visual Data Mining, Information Visualization, User Interfaces.

1. INTRODUCTION

Companies, government agencies, and other types of organizations are making their large data sets available over the Internet. IBM, US Census Bureau, NASA, and the World Health Organization are only a few of these organizations. Designers of user interfaces for these data sets rely on command languages or form fillin strategies. These designers make the following assumptions during their design processes:

a) Users are informed about the data that they are working on or they will submit known-item queries rather than probing the data,

b) Users know or have the will to understand a querying environment or fill a lengthy form,

c) Users will have the bandwidth or the time to access large data.

Most of these assumptions are not valid for the users of public online data sets. Many user interfaces do not give users an indication of the distribution of data. However, this is essential for browsing public online data to guide the users in the query formulation process. Unguided novice users may waste time by submitting queries that have zero-hit or mega-hit result sets. Traditional user interfaces require users to fill lengthy forms or form complex queries. However, users of public online data sets do not have the time or the will to learn a query language or they are annoyed when they have to fill a lengthy form. A more efficient, simple, and easy to learn approach for defining queries is needed. Finally, users of a public online data set have to access large amounts of data using a low bandwidth congested network. Hence, strategies that introduce efficient means of communication are needed.

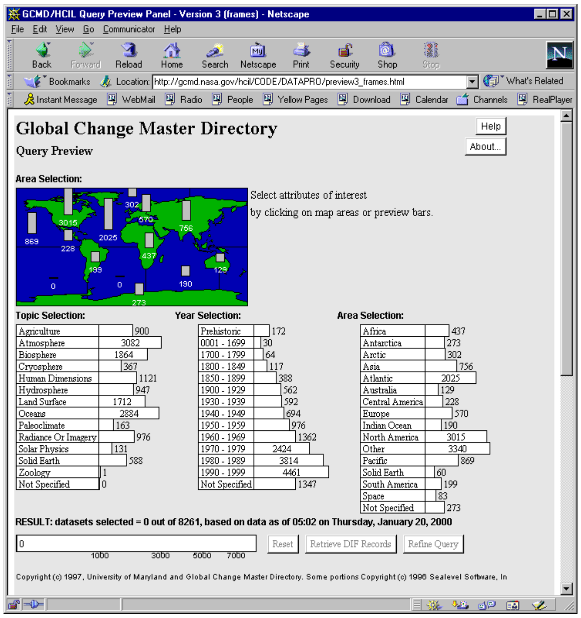
Query previews [3,12] form a novel visual approach for querying large online data collections. Query previews supply data distribution information about the data that is being searched and give continuous feedback about the size of the result set for the query as it is being formed. Queries are incrementally and visually formed by selecting items from a set of charts. Query previews take advantage of the fact that users are generally interested in a subset of the data. Once the scope has been narrowed, a second phase can start with local data. This second phase can be a simple list of results or a sophisticated user interface that will allow users to visualize the result set. This multi-phase approach will increase the network performance of the overall system. Figures 1 and 2 show a sample query preview panel [14] with three commonly used attributes of a NASA data collection, the Global Change Master Directory. The distribution of data over these attributes is shown with bars and the result size is displayed as a separate bar at the bottom.

Recent work by [6] shows that many users prefer query previews and perform better with them. Users report that query previews are easy to use and understand. These results increase our hopes of serving a broad user domain in a more satisfactory manner. Unfortunately, current applications of query previews use just a few pre-selected attributes of the data. The distribution information is displayed only on these attributes. These implementations of query previews work over a single table that has a relatively low number of attributes. The simplicity of the data structures and a user interface with only a few pre-selected attributes are the positive aspects of these implementations. However, data sets are generally formed of numerous tables and attributes. Therefore, pre-selection of a few attributes is an important restriction of the query preview approach. Application of query previews as a visual means to access public online data requires this restriction to be relaxed.

This paper introduces a more general visual approach. We allow users to browse multiple tables and attributes of the data with a hierarchical browser. All the attributes of the data can be used to display the distribution information using bar charts and can be expanded, visualized, and manipulated. Thus, with this generalization, the query previews become a stronger candidate for accessing public online data.

2. RELATED WORK

Many researchers have tried to devise methods for more successful visual querying. The Rabbit system, by Williams [16] and the work of Heppe, Edmondson, and Spence [9] were early demonstrations of the benefits of progressive visual querying. Other systems showed relevance of results: for example Veerasamy and Navathe [15] used histograms, and Hearst [7] used TileBars to visually present relevance of results to the terms used in the query. WebTOC [11] uses a hierarchical outliner and a bar chart presentation to preview the size and type of items (e.g., image, sound, etc.) within each branch. Eick [4] proposes to augment sliders of visualization systems with density plots or bar charts. Antis, Eick, and Pyrce [2] introduce methods for visualizing the schemas of relational data. Dynamic queries [1,5,13,17] use a direct manipulation approach to facilitate query formulation with a visual representation of query components and results. They allow rapid, incremental, and reversible control of the query. Results are presented visually. Continuous feedback guides users in the query formulation process. Marchionini and Greene [10] discuss the importance of visualization issues in public access and use of government statistical information. Hearst [8] lists many approaches to visualization systems for information retrieval purposes.


GCMD/HCIL Query Preview Panel - Version 3 (frames) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://gcmd.nasa.gov/hcil/CODE/DATAPRO/preview3_frames.html
What's Related
Instant Message WebMail Radio People Yellow Pages Download Calendar Channels RealPlayer
Help
About...
3015
302
570
756
869
228
2025
437
199
129
0
0
190
273
Reset
Retrieve DIF Records
Refine Query
1000
3000
5000
7000

Global Change Master Directory

Query Preview

Area Selection:

Select attributes of interest
by clicking on map areas or preview bars.

Topic Selection:

Agriculture	900
Atmosphere	3082
Biosphere	1864
Cryosphere	367
Human Dimensions	1121
Hydrosphere	947
Land Surface	1712
Oceans	2884
Paleoclimate	163
Radiance Or Imagery	976
Solar Physics	131
Solid Earth	588
Zoology	1
Not Specified	0

Year Selection:

Prehistoric	172
0001 - 1699	30
1700 - 1799	64
1800 - 1849	117
1850 - 1899	388
1900 - 1929	562
1930 - 1939	592
1940 - 1949	694
1950 - 1959	976
1960 - 1969	1362
1970 - 1979	2424
1980 - 1989	3814
1990 - 1999	4461
Not Specified	1347

Area Selection:

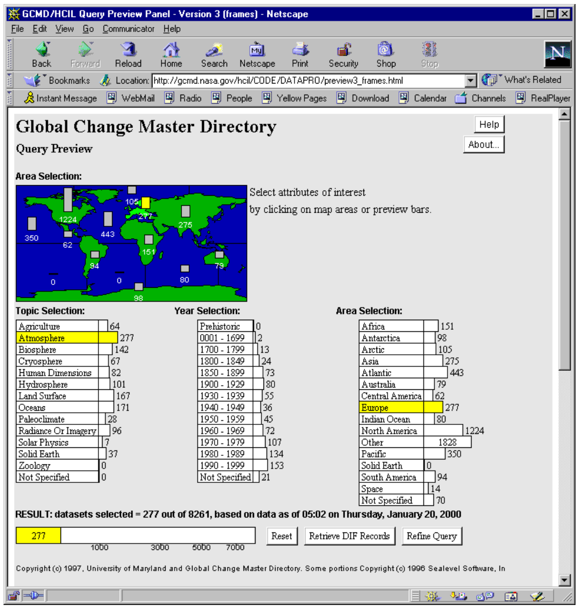
Africa	437
Antarctica	273
Arctic	302
Asia	756
Atlantic	2025
Australia	129
Central America	228
Europe	570
Indian Ocean	190
North America	3015
Other	3340
Pacific	869
Solid Earth	60
South America	199
Space	83
Not Specified	273

RESULT: datasets selected = 0 out of 8261, based on data as of 05:02 on Thursday, January 20, 2000

0

Copyright (c) 1997, University of Maryland and Global Change Master Directory. Some portions Copyright (c) 1996 Sealevel Software, In

Figure 1: A sample query preview developed at the Human-Computer Interaction Laboratory, for NASA's Global Change Master Directory (available at gcmd.nasa.gov). Topic, Year, and Area are three most frequently used attributes of the data. These attributes are selected to show the data distribution information. The distribution is shown with bar charts. The result set size is displayed as a separate bar at the bottom.


GCMD/HCIL Query Preview Panel - Version 3 (frames) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://gcmd.nasa.gov/hcil/CODE/DATAPRO/preview3_frames.html What's Related
Instant Message WebMail Radio People Yellow Pages Download Calendar Channels RealPlayer
Help
About...
1224
350
62
443
105
277
275
151
94
79
0
0
80
98
Reset
Retrieve DIF Records
Refine Query
1000
3000
5000
7000

Global Change Master Directory

Query Preview

Area Selection:

Select attributes of interest
by clicking on map areas or preview bars.

Topic Selection:

Agriculture	64
Atmosphere	277
Biosphere	142
Cryosphere	67
Human Dimensions	82
Hydrosphere	101
Land Surface	167
Oceans	171
Paleoclimate	28
Radiance Or Imagery	96
Solar Physics	7
Solid Earth	37
Zoology	0
Not Specified	0

Year Selection:

Prehistoric	0
0001 - 1699	2
1700 - 1799	13
1800 - 1849	24
1850 - 1899	73
1900 - 1929	80
1930 - 1939	55
1940 - 1949	36
1950 - 1959	45
1960 - 1969	72
1970 - 1979	107
1980 - 1989	134
1990 - 1999	153
Not Specified	21

Area Selection:

Africa	151
Antarctica	98
Arctic	105
Asia	275
Atlantic	443
Australia	79
Central America	62
Europe	277
Indian Ocean	80
North America	1224
Other	1828
Pacific	350
Solid Earth	0
South America	94
Space	14
Not Specified	70

RESULT: datasets selected = 277 out of 8261, based on data as of 05:02 on Thursday, January 20, 2000

277

Copyright (c) 1997, University of Maryland and Global Change Master Directory. Some portions Copyright (c) 1996 Sealevel Software, In

Figure 2: When users select attribute values (e.g. Topic = Atmosphere and Area = Europe in this screenshot), the bars are updated immediately to reflect the distribution of the data that satisfies the query. The result set size is also updated accordingly.

3. QUERY PREVIEWS

The concept of query previews [3,12] was triggered by the need to extend the dynamic querying idea [1,5,13,17] to large networked data collections. Query previews show the contents of the data during the query formulation process. In order to guide users in the query formulation process, query previews provide aggregate information on some pre-selected attributes of the data. Distribution of data over some attribute values is shown graphically using representations such as bar or pie charts. When users select a value on any of the attributes by just clicking on the related representations of a query preview panel, the rest of the user interface is updated immediately. This is called tight coupling. Actions are easily reversible, and error prevention instead of error correction is used. For every action users take, feedback is given continuously. As users see the potential size of their result set before submitting the query, they are less likely to create queries that return zero or mega hits. Users see the trends in the data and they learn where the data has gaps or clusters. (Figures 1 and 2 show a sample query preview panel.) The server load will be reduced if users do not waste their time with zero hit queries or consume network resources in downloading large sets of useless results.

Query previews only need aggregate information about the data. The data distribution information is represented with multidimensional histograms. Each cell of a histogram represents a count of the results mapping to that cell. Hence, whatever the size of the data is, only the counts are needed to form a query preview panel. The size of this information is fixed regardless of the size of the data. Only the counts are incremented with each insertion. This makes query previews a powerful tool for accessing large online data collections.

4. GENERALIZING QUERY PREVIEWS

Current applications of query previews use a few pre-selected attributes of the data. The distribution information is displayed only on these attributes. However, data sets are generally formed of numerous tables and attributes. Therefore, pre-selection of a few attributes is a restriction of the query preview approach.

To relax this restriction, we combine a hierarchical browser and the query preview approach to let users browse all the tables and attributes of a data set. With this generalization all the attributes of the data can be used to display the data distribution information.

Figure 3 presents a sample hierarchical browser. In this prototype, we use the Environmental Protection Agency (EPA) as our sample organization and a fragment of the Toxic Release Inventory from the EPA data collections as our sample data set. This sample data set is formed of approximately 400,000 reports of toxic material releases to the environment from various facilities in the United States. We put together four tables from this data, which are Contact Info, Release Info, Chemical Info, and Facility Info. Each table contains a few sample attributes, e.g., Contact Info contains Contact Phone and Contact Name as its attributes. The root of our browser is tagged with the name of the data set. Each table is represented by a separate branch. Each branch may also have leaves representing different attributes of that branch (table). The result bar is visible on top of the panel showing the total number in the result set (reports for our example with the EPA data) for the current query. At any time, the users can fetch these results by simply pressing the fetch button to the left of the result bar. We attach the distribution information next to the related branch of that attribute.

Some of the attributes do not have the distribution information attached to them. For example, Contact Name of the Contact Info table of Figure 3 does not have anything attached to it. The nature of the Contact Name attribute does not allow a useful representation. There are almost as many names in the data set as the number of reports. In this paper, we focus on other types of attributes, e.g., gender and age. These have useful representations. Figure 4 shows such an attribute. This attribute, the Reporting Year, is an attribute of the Release Info branch. It is represented as a folder. It can have other branches under it. Still, it is an attribute of the Release Info branch. This visual difference is used to show the expandability of this attribute. These types of attributes are expandable into buckets. Buckets show the possible values or ranges of values for that attribute. The distribution of data is shown over them.

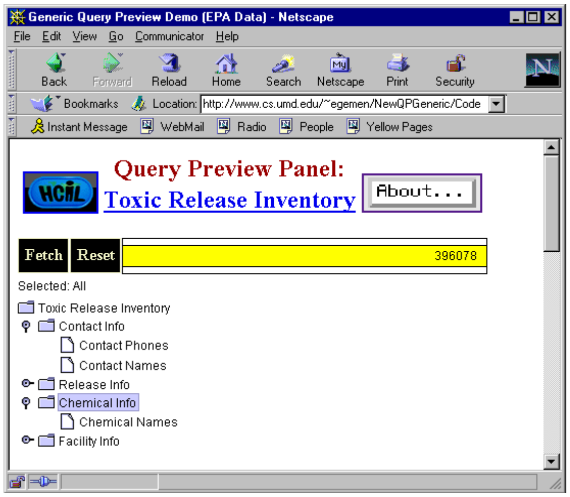

Generic Query Preview Demo (EPA Data) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security
Bookmarks Location: http://www.cs.umd.edu/~egemen/NewQPGeneric/Code
Instant Message WebMail Radio People Yellow Pages
Query Preview Panel:
Toxic Release Inventory
About...
Fetch
Reset
396078
Selected: All
Toxic Release Inventory
Contact Info
Contact Phones
Contact Names
Release Info
Chemical Info
Chemical Names
Facility Info

Figure 3: A hierarchical browser represents the tables of the data set. The root is tagged with the name of the data set. Each table is represented by a separate branch. Each branch may contain a few leaves representing the attributes of that branch. In this example, Contact Info and Chemical Info branches are expanded to demonstrate this feature.

In Figure 4, we see five bars showing the reports submitted in each of the five years of the EPA Data. The users can immediately see the number of reports has declined about 25% over the five years. Other attributes of the data can be expanded similarly. Figure 5 shows such a display that reveals high numbers of reports for the Southeast and Midwest, but relatively few for the Northwest.

Bars also form a mechanism for input to the user interface. Figure 6 shows such an example action. The reporting year 1994 is selected by just clicking on the bar.

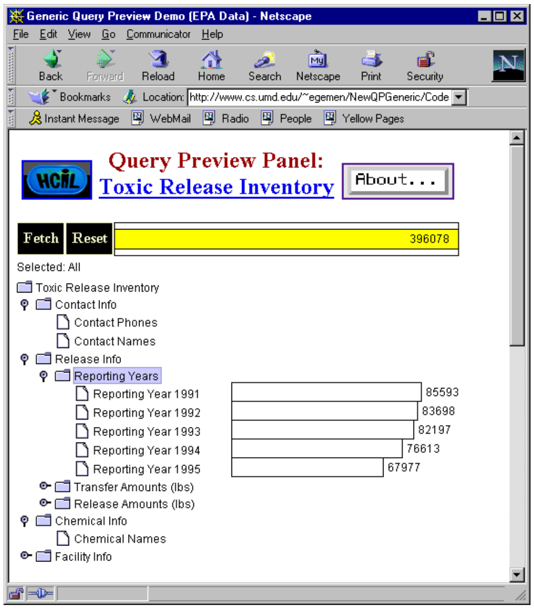


Figure 4: The Reporting Year attribute of the Reporting Info branch is expanded. A set of bars is attached to this presentation to show the distribution of data over this attribute.

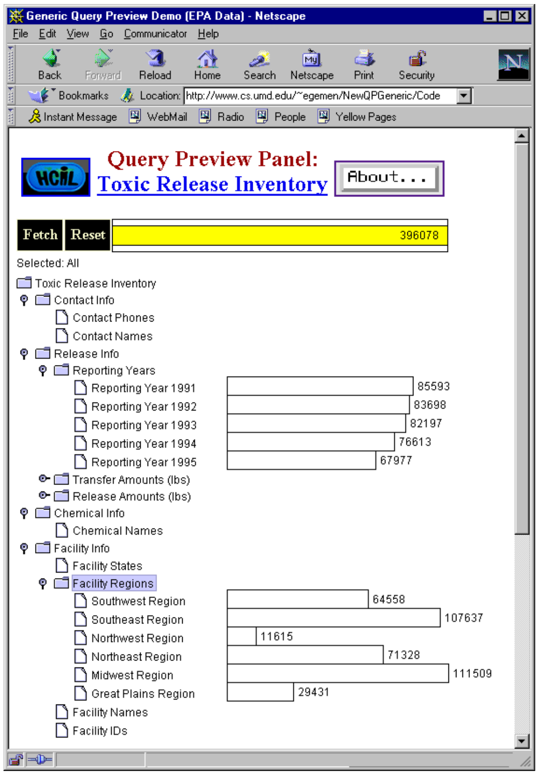


Figure 5: Facility Region attribute is expanded showing the data distribution on another set of bars. Each bar is used to show a different region.

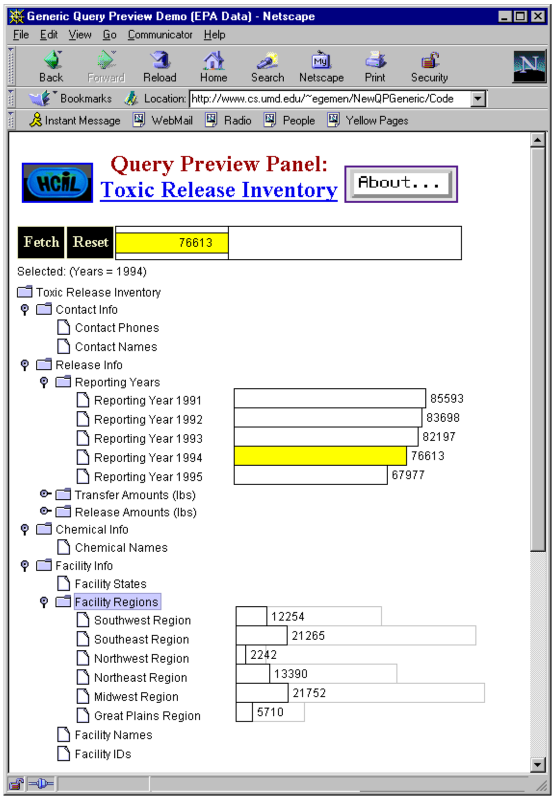


Figure 6: Selections on bars form an input mechanism. This panel shows that for 1994 there were 76,613 reports and shows the distribution by Facility Regions.

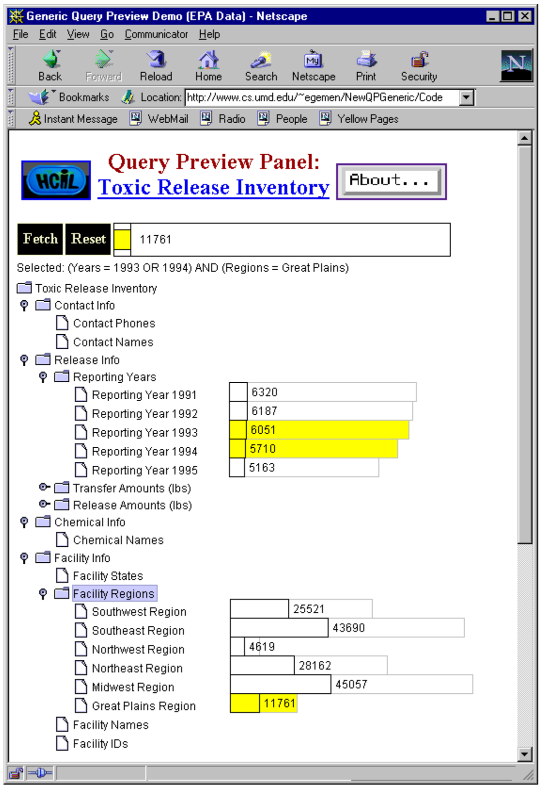
Upon any selection, the distribution information on all the other bars including the result bar is updated. A silhouette of the initial bar settings is kept to show the original sizes of the bars. The text field below the fetch button displays the selected values in text form. This is essential since the bars can also be collapsed to make room for other expansions. In this case, the visual feedback from the collapsed bars may be lost. Hence, the text field is a reminder for the previous selections. Another reminder for selections can be placed next to the attribute names. At any time, users can reset their selections by simply clicking on the reset button next to the fetch button. Figure 7 shows some further selections and updates on the bars. The session can continue as long as the users want to explore the data. When users want to see the results for their query, they can fetch the desired reports from the server matching their selections. They can view this result set as a simple list or they can continue querying on it using local tools.

As bars expand and collapse the desired data distribution information is brought from the server. This creates short delays during the query formulation process. Despite these delays, the amount of data that is downloaded from the network is very small, and does not introduce large interruptions or a significant network load. In general, we do not fetch the results, but only the distribution information about the results of these intermediate collections. In some cases, the distribution information can be cached to improve performance. In other cases, it can be a subset of the previous distribution, so a second network access can be avoided. In our example, the size of the total distribution information is only 8 Kbytes. Thus, the total server overhead is minimal. Also, we can regularly update the distribution information as an offline process.

One limitation on the distribution information is the number of attributes that can be simultaneously displayed. This number is equal to the number of dimensions of data representing the distribution information. As this number grows, the amount of data needed grows exponentially. Thus, manipulating many attributes of the database at the same time may not be feasible. A solution to this problem is downloading the actual attribute values for the matching records after the first few selections. After the initial selections, the size of the result set may be drastically reduced. This avoids downloading large amounts of data from the server. Therefore, this solution can allow users to continue working on a local minimized version of the data set.

5. CONCLUSIONS

Query previews form a novel visual approach for querying large online data collections. On the other hand, current applications of query previews use only a few pre-selected attributes of the data. Unfortunately, many data sets are formed of numerous tables and attributes. Applicability of query previews to many public online data sets requires this restriction to be relaxed. This paper introduces a generalization of query previews. We allow users to browse all the tables and attributes of a data set with a hierarchical browser. All appropriate attributes of the data can be used to display the distribution information using charts and can be expanded, visualized, and manipulated. Hence, with this generalization, we believe that query previews can be used more widely to visually browse large online data.


Generic Query Preview Demo (EPA Data) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security
Bookmarks
Instant Message WebMail Radio People Yellow Pages
Query Preview Panel:
Toxic Release Inventory
About...
Fetch
Reset

Location: http://www.cs.umd.edu/~egemen/NewQPGeneric/Code

11761

Selected: (Years = 1993 OR 1994) AND (Regions = Great Plains)

Toxic Release Inventory
Contact Info
Contact Phones
Contact Names
Release Info
Reporting Years
Reporting Year 1991 6320
Reporting Year 1992 6187
Reporting Year 1993 6051
Reporting Year 1994 5710
Reporting Year 1995 5163
Transfer Amounts (lbs)
Release Amounts (lbs)
Chemical Info
Chemical Names
Facility Info
Facility States
Facility Regions
Southwest Region 25521
Southeast Region 43690
Northwest Region 4619
Northeast Region 28162
Midwest Region 45057
Great Plains Region 11761
Facility Names
Facility IDs

Figure 7: Further selections from the EPA data. Users continue to see updates on the bars as they make their selections.

Acknowledgements: We appreciate support from the National Science Foundation Digital Government Initiative.

REFERENCES

1. Ahlberg, C. and B. Shneiderman, Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays, *Proceedings of the ACM CHI '94 Conf.*, 1994, pp. 313-317.
2. Antis, J., S. Eick, and J. Pyrce, Visualizing the Structure of Relational Databases, *IEEE Software*, January 1996, pp. 72-79.
3. Doan, K., C. Plaisant, and B. Shneiderman, Query Previews in Networked Information Systems, *Proceedings of the Forum on Advances in Digital Libraries*, 1996, pp. 120-129.
4. Eick, S., Data Visualization Sliders, *Proceedings of ACM UIST '94 Conf.*, 1994, pp. 119-120.
5. Goldstein, J. and S. Roth, Using Aggregation and Dynamic Queries for Exploring Large Data Sets, *Proceedings of the ACM CHI '94 Conf.*, 1994, pp. 23-29.
6. Greene, S., E. Tanin, C. Plaisant, B. Shneiderman, L. Olsen, G. Major, and S. Johns, The End of Zero-Hit Queries: Query Previews for NASA's Global Change Master Directory, *International Journal of Digital Libraries*, 2, 2, 1999, pp. 79-90.
7. Hearst, M., TileBars: Visualization of Term Distribution Information in Full Text Information Access, *Proceedings of the ACM CHI '95 Conf.*, 1995, pp. 59-66.
8. Hearst, M., User Interfaces and Visualization, *Modern Information Retrieval*, 1999, ACM Press, Ricardo Baeza-Yates and Berthier Ribeiro-Neto, pp. 257-323.
9. Heppe, D., W. Edmondson, and R. Spence, Helping both the Novice and Advanced User in Menu-driven Information Retrieval Systems, *Proceedings of HCI '85 Conf.*, 1985, pp. 92-101.
10. Marchionini, G. and S. Greene, Public Access and Use of Government Statistical Information, *Presented to the Federal Information Services*, ils.unc.edu/~march/, NSF Workshop, 1997.
11. Nation, D., C. Plaisant, G. Marchionini, and A. Komlodi, Visualizing Websites Using a Hierarchical Table of Contents Browser: WebTOC, *Proceedings of the 3rd Conf. on Human Factors and the Web*, 1997.
12. Plaisant, C., T. Bruns, K. Doan, and B. Shneiderman, Interface and Data Architecture for Query Previews in Networked Information Systems, *ACM Transactions on Information Systems*, 17, 3, 1999, pp. 320-341.
13. Shneiderman, B., Dynamic Queries for Visual Information Seeking, *IEEE Software*, 11, 6, 1994, pp. 70-77.
14. Tanin, E., A. Lotem, I. Haddadin, B. Shneiderman, C. Plaisant, and L. Slaughter, Facilitating Data Exploration with Query Previews: A Study of User Performance and Preference, *Behaviour & Information Technology* (to appear), 2000.
15. Veerasamy, A. and S. Navathe, Querying, Navigating and Visualizing a Digital Library Catalog, *Proceedings of the Second International Conf. on the Theory and Practice of Digital Libraries*, 1995.
16. Williams, M., What Makes RABBIT Run, *International Journal of Man-Machine Studies*, 21, 4, 1984, pp. 333-352.
17. Williamson, C., and B. Shneiderman, The Dynamic Home Finder: Evaluating Dynamic Queries in a Real-Estate Information Exploration System, *Proceedings of ACM SIGIR '92 Conf.*, 1992, pp. 338-346.