Visualizing information
Biomedical and Life Sciences Division

Special Libraries Association

94th Annual Conference

New York, New York

June 12, 2003

Visualizing Information in the Biological Sciences:

Using WebTheme to Visualize Internet Search Results

Karen A. Buxton

Mary Frances Lembo

Information Specialists

Hanford Technical Library

Pacific Northwest National Library

Introduction

With the ever growing amount of information available on the Internet, new methods for gathering, analyzing, and interpreting data are needed. Information visualization is an effective method for displaying large data sets in a pictorial or graphical format. Visualizations aid researchers and analysts in understanding data by evaluating the content and grouping documents together around themes and concepts. The human mind can process vast quantities of information in visual format because the mind uses vision to make judgments that allow analysis of an entire information space (Thomas et al., 1999). This makes information visualization a powerful tool for analysts and researchers. WebThemeä is a visualization tool which allows users to harvest thousands of Web pages and automatically organize and visualize their contents.

WebTheme is an interactive Web-based product that provides a new way to investigate and understand large volumes of HTML text-based information. It harvests data from the World Wide Web using either search terms with selected search engines or by following URLs chosen by the user. Researchers use WebTheme and its suite of tools to rapidly identify themes and concepts found among thousands of pages of text and to explore topics and documents of special interest within the data set.

WebTheme is one of several information visualization products created at the Pacific Northwest National Laboratory (PNNL). Developed in 1996 as an internal project, a WebTheme beta version was released in 1998 under sponsorship from NASA. The basis for WebTheme lies in SPIRE (Spatial Paradigm for

Information Retrieval and Exploration) tool (PNNL, 1999).

PNNL is one of the U.S. Department of Energy's nine multiprogram national laboratories, and it conducts research in the fields of environmental science, energy, health sciences and national security. Scientists at PNNL are investigating a variety of significant topics in the biosciences, including biofuels, genomics, proteomics, and bioterrorism. As new research and development topics are discussed in Web-based HTML documents, WebTheme can help analysts determine who the players are in these quickly evolving fields, find project partners, and understand new research areas as they develop by harvesting and visualizing information from the Web.

Through a collaboration with the Information Science and Engineering (IS&E) Division at PNNL, information specialists at the Hanford Technical Library are helping to deploy WebTheme at the lab. They provide demonstrations to interested researchers and provide one-on-one consultations to get researchers started using the product.

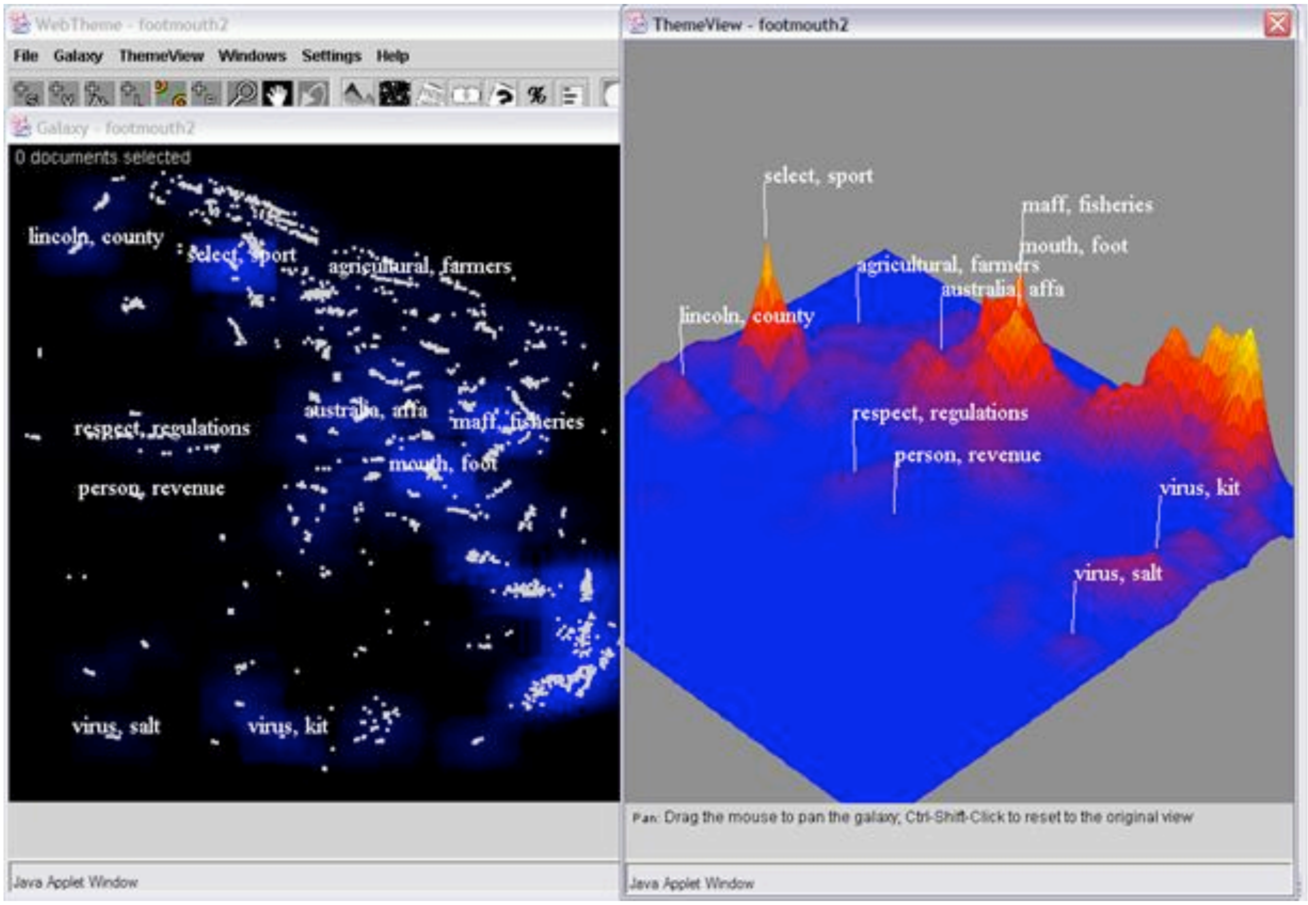
WebTheme Overview

The World Wide Web offers researchers a variety of commercial, technical, scientific, and academic HTML pages as well as casual and authoritative Web sites. Exploring large or complex Web sites can be time consuming, and search engine queries can returns thousands or perhaps millions of results. Researchers usually look through the top several pages of query results or follow the most promising links in a Web site and may miss important details in documents buried deep within query results or several layers into a Web site.

WebTheme uses agent technology that makes contact with Web servers and harvests Web documents (Whiting and Cramer, 2002). A text processor extracts the key descriptors within HTML documents and statistically measures their frequency. A natural language understanding algorithm obtains semantic relationships and creates a visual representation of hundreds or thousands of HTML pages in a graphical display. Researchers and “analysts can quickly gain an understanding of patterns and trends that underlie the documents” (Wise et al., 1999) using map analogies. Content is grouped around themes and concepts that show the documents and their relationships in a star map, known as the Galaxy View, and in a terrain map called the ThemeViewä.

The Galaxy View is a two-dimensional document scatter plot. Documents are displayed as if they were stars in a night sky. The resulting visualization is simple and provides a critical first cut at sifting information and determining how the contents of the Web documents are related. The key feature of this visualization is document similarity. The closer documents are to each other in content and context, the closer they are located in the two-dimensional galaxy (Wise et al., 1999). Blue clouds or nebulae represent areas with large concentrations of similar documents.

The same group of documents can be viewed as a three-dimensional landscape. A relief map displays primary themes in the underlying documents and mountain peaks indicate theme strength (Chen and Paul, 2001). The mountain peaks of ThemeView correspond to the blue nebulae in Galaxy View. Each view offers a different perspective of the same information and aids in quickly grasping overall concepts and topic strength. In addition to creating these two views of the data, WebTheme provides tools for information retrieval, investigation, and further analysis.



Pan: Drag the mouse to pan the galaxy; Ctrl-Shift-Click to reset to the original view

Figure 1. The left image shows the Galaxy View, and the right is the ThemeView of the same data set.

The IS&E/Library Collaboration

WebTheme has powerful data acquisition and analysis capabilities. Understanding Web search techniques as well as the product features are a prerequisite for effectively using it. Two information specialists were provided in-depth training by IS&E staff. In addition, the library staff is experienced in developing, marketing, and delivering product demonstrations. The search skills, WebTheme training, and product demonstration experience equip information specialists with the tools they need to offer WebTheme training and support to PNNL employees. Once a product demonstration was ready, the homegrown tool was rolled out to PNNL employees. In order to provide the level of support needed to fully exploit the product, the library and IS&E agreed to fund a coupon program for demo attendees. Employees who attended a demo were provided with a coupon for 45 minutes of free one-on-one product support.

Creating a Visualization: the Process

Each visualization begins with a question that is appropriate to answer with HTML documents gathered from the Web. There are undoubtedly many questions that can be explored with WebTheme; some questions that have been successfully answered are listed below:

§ What are the key characteristics, structure, and topics of a particular Web site or sites?

§ What is being said about a new technology on the Web?

§ Who are the key players in a field?

§ What organizations are doing complementary work and are good candidates for collaboration? (Lembo and Buxton, 2002)

Prior to creating a visualization, explore the World Wide Web to see what types of results your search will retrieve. WebTheme users have the option of setting up a query using search terms in either Google or AltaVist or exploring a list URLs. In either case, performing an initial investigation outside of WebTheme gives important information about the number of pages likely to be retrieved or the organization of the Web sites you want to investigate. Search terms can be refined and searches broadened or narrowed before using them within WebTheme. For a list of URL harvests, users can follow links on these pages to get an idea of the structure of the pages.

Using WebTheme to Think about Animal Disease

The visualization that we will investigate for this paper is in the field of veterinary science and was developed with a customer who redeemed a coupon for individual support. He was particularly interested in worldwide research on foot and mouth disease and swine fever. He was curious to know who was doing research in the areas of the world that he knew had recently experienced outbreaks of these diseases. We chose to conduct a Google search because our preliminary investigations showed good results with our search strategy: “foot * mouth” “swine fever.” The Google search retrieved over 9,600 hits. We knew that many of these were PDF documents that could not be harvested. The topics on the Web pages we retrieved were on target so we decided to implement the search in WebTheme. This number of hits was adequate to give us a good visualization and provided a focused analysis that revealed important themes in the field.

Settings

Once we decided on the question and a search strategy, WebTheme offered us several harvesting decisions. We have trained users to consider how they want to gather data with these choices in mind as they are conducting preliminary searches and looking at sites. Harvest decisions influence the effectiveness of visualizations. One of the first choices we needed to make was the harvest depth, or in other words, how many layers of Web page links did we want to follow to harvest related documents. We decided to follow Web page links down one level. We could have decided only to look at top-level pages, or we could have opted to delve as deep as ten linking pages. A second choice is whether to harvest links from the local host only or allow the agent to pursue all domains. We chose local host since many sites linked to pages at other domains that we felt would not be useful to visualize. We also selected link mode so that we would be able to see how sites were linked once the visualization was processed.

Next, we needed to decide how much information to harvest. Since the Google relevancy ranking is determined by a citation approach, we felt that since much of what we would like to retrieve may be new and therefore ranked lower on Google, it would be appropriate to select the maximum number of documents possible (10,000). It is possible, however, to limit to as few as 500.

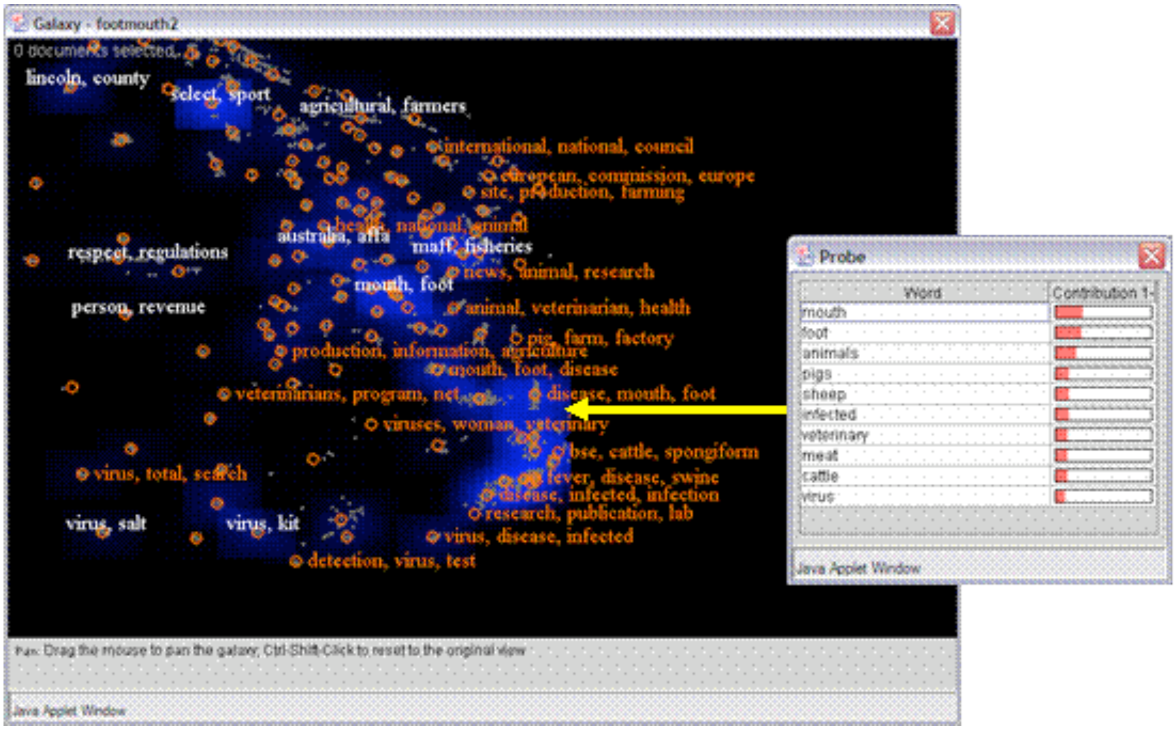
As we reviewed pages in our initial Web investigation, we noted any problematic Web sites that returned results not useful to our investigation. For example, “x” URL kept returning and did not include useful information. Since we can filter unwanted URLs from our visualization, we made a note of the URL so we could filter it later. We can easily filter out foreign language Web sites, newsgroups and FTP sites as well.

Get an overview of subject matter

Once settings and filters had been selected, we processed the visualization. The Galaxy View and the ThemeView taken together offer an effective method for seeing the top-level ideas in a group of HTML pages. Theme labels identify major themes. Nebulae or dark blue clouds represent areas of greatest document density or the “hot topics.” This high-level look at a topic provides analysts with a jumping-off point to pursue subject matter in greater detail.

Narrowing the topic

Several tools are available to conduct further analysis. These tools can be used in any order, but we like to start with the cluster centroids tool. (Keep in mind that software programmers, not librarians, developed these tool names.) The cluster centroids appear in Galaxy View as orange circles and display the top three terms in each document group. The terms specify the unifying concepts within the documents. A second tool that relates to the cluster centroids tool is the probe tool. The probe tool reveals up to 10 top-level terms and their strength among the pages. It allows the analyst to get a broader view of topics within document groups and even identify themes in the gaps between them. The probe tool is one of the few features of the product that may also be used to investigate the ThemeView. Most of the tools available are used only within the Galaxy View.



Pan: Drag the mouse to pan the galaxy; Ctrl-Shift-Click to reset to the original view

Word	Contribution 1-
mouth	
foot	
animals	
pigs	
sheep	
infected	
veterinary	
meat	
cattle	
virus	

Figure 2. The orange circles are cluster centroids. Clicking on the circle shows the top three terms in the document cluster. The probe tool displays the top ten terms anywhere you click. You may click on any region of the galaxy including between document groups.

Once the preliminary exploration has revealed some areas of interest, the zoom tool permits the analyst to focus attention on an area of particular interest. The document title tool allows the user to click individual documents to reveal page titles. Analysts can quickly get a sense of the type of pages in that area.

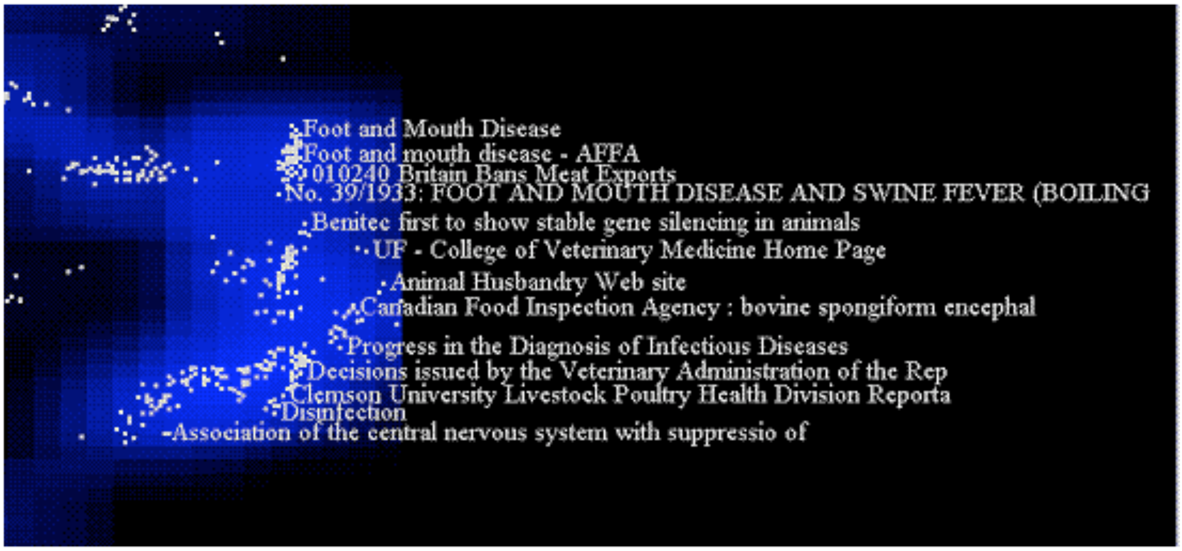

Foot and Mouth Disease
Foot and mouth disease - AFFA
010240 Britain Bans Meat Exports
No. 39/1933: FOOT AND MOUTH DISEASE AND SWINE FEVER (BOILING
Benitec first to show stable gene silencing in animals
UF - College of Veterinary Medicine Home Page
Animal Husbandry Web site
Canadian Food Inspection Agency : bovine spongiform encephal
Progress in the Diagnosis of Infectious Diseases
Decisions issued by the Veterinary Administration of the Rep
Clemson University Livestock Poultry Health Division Reporta
Disinfection
Association of the central nervous system with suppressio of

Figure 3. The zoom tool allows a user to isolate a region and get a closer look at the documents there. The document title tool can then be used to see what types of pages are in the region.

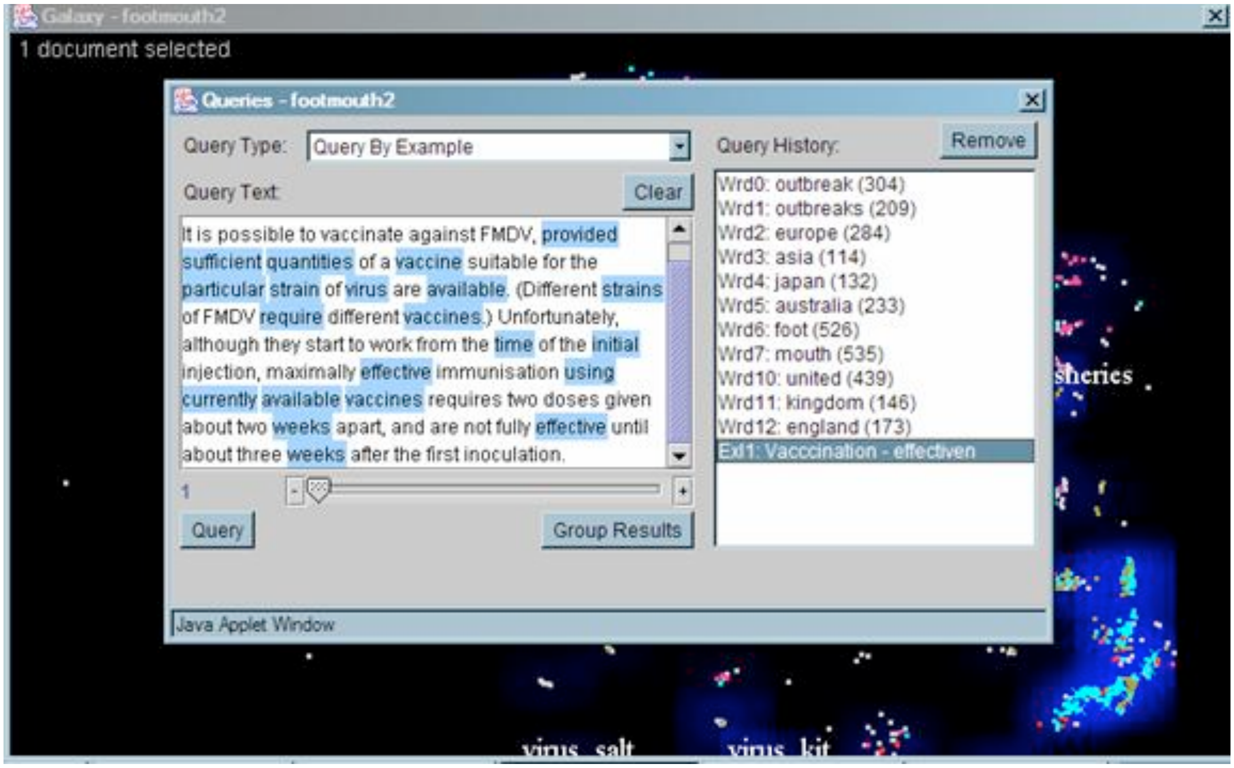
Data Analysis with WebTheme

Isolating the information you need within the data set

As we mentioned earlier, WebTheme offers tools for information retrieval and for further analysis. Since we were looking for information about outbreaks in Europe and Asia we conducted a number of searching using the query tool. We have two query options: a simple word search or a query-by-example search. The query-by-example search allows us to enter a a paragraph or more of text from a particularly interesting document and search for additional documents that are similar to the text entered. We selected the paragraph below and lauched a query-by-example.

It is possible to vaccinate against FMDV, provided sufficient quantities of a vaccine suitable for the particular strain of virus are available. (Different strains of FMDV require different vaccines.) Unfortunately, although they start to work from the time of the initial injection, maximally effective immunisation using currently available vaccines requires two doses given about two weeks apart, and are not fully effective until about three weeks after the first inoculation. (Booth, 2001)

The number under the search window in Figure 4 shows that one document matched our query. We can use the slider just below the query window to move from the minus sign toward the plus sign and choose the number of most closely related documents to investigate. We chose to review 200 documents and moved the slider until the number read 200, then we grouped them by clicking the Group Results button. Once we had run searches and grouped the results, we were ready use the grouping tool to combine our results and focus our investigation.



Query Type: Query By Example

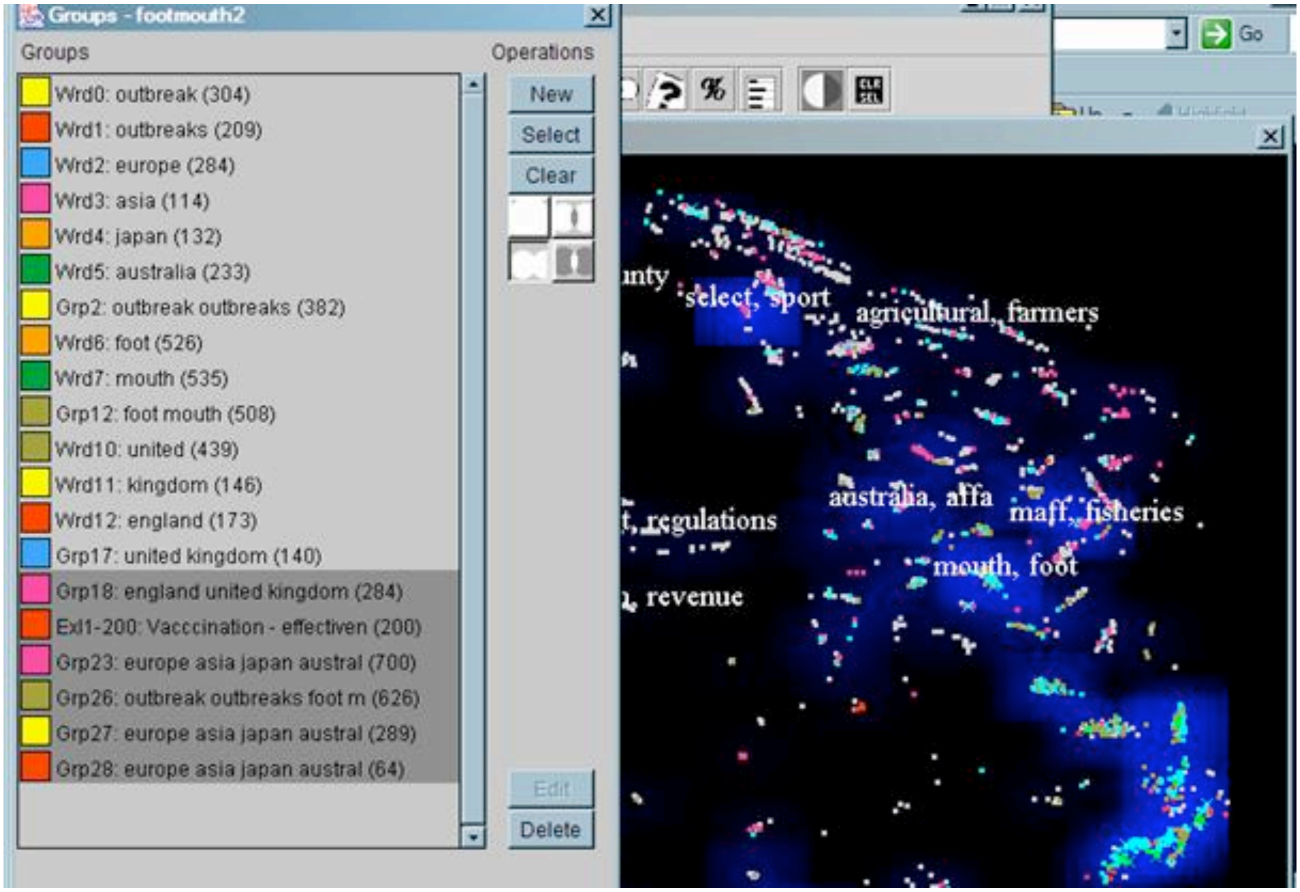
It is possible to vaccinate against FMDV, provided sufficient quantities of a vaccine suitable for the particular strain of virus are available. (Different strains of FMDV require different vaccines.) Unfortunately, although they start to work from the time of the initial injection, maximally effective immunisation using currently available vaccines requires two doses given about two weeks apart, and are not fully effective until about three weeks after the first inoculation.

Wrd0: outbreak (304)
Wrd1: outbreaks (209)
Wrd2: europe (284)
Wrd3: asia (114)
Wrd4: japan (132)
Wrd5: australia (233)
Wrd6: foot (526)
Wrd7: mouth (535)
Wrd10: united (439)
Wrd11: kingdom (146)
Wrd12: england (173)
Ex1: Vacccination - effectiven

Figure 4. A series of twelve word searches followed by a query-by-example search are used to isolate pages with data of interest.

Grouping tool

The grouping tool allows the analyst to combine created sets using set logic. Three grouping tools enable two or more sets to be combined into one set, find the intersection of two or more sets, locate items that are either in one set or the other but not both, or select only one group of records to view.

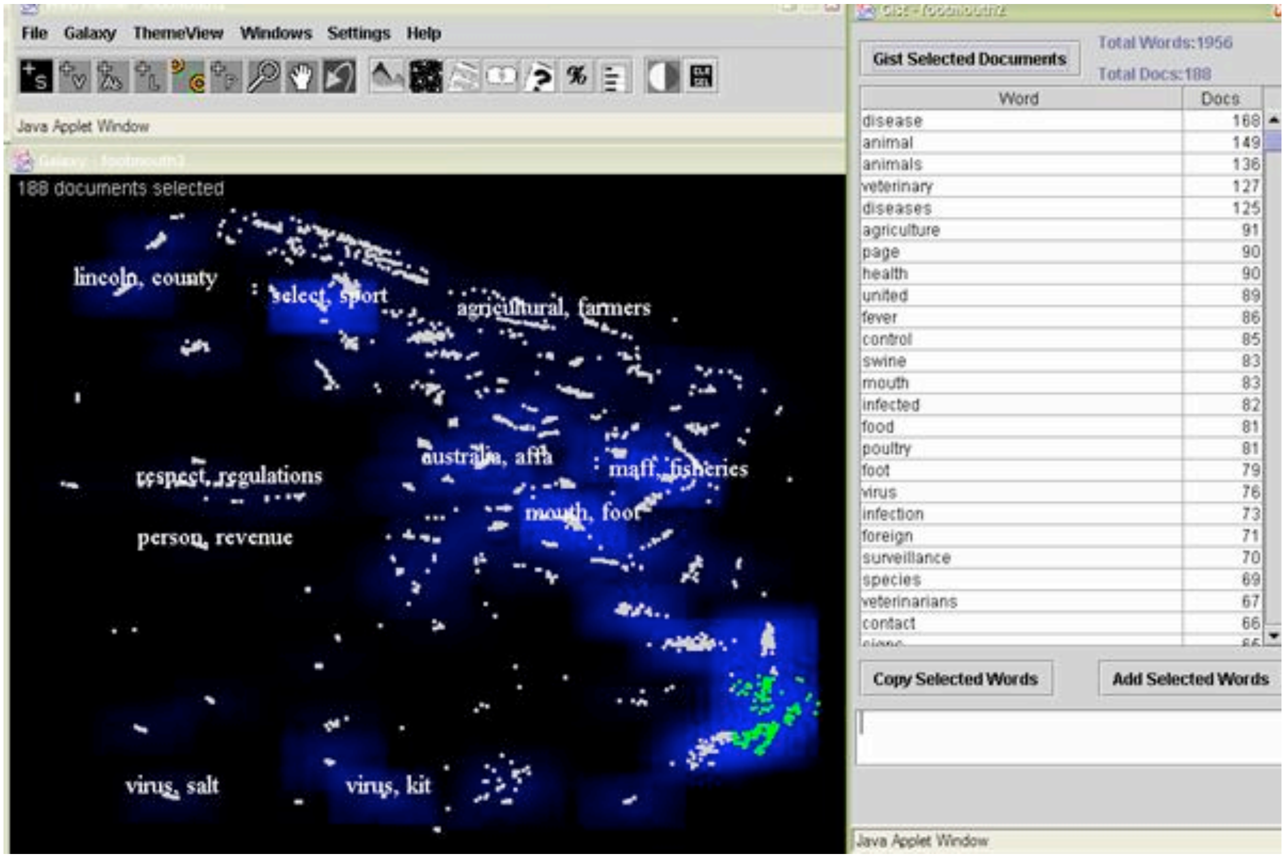


- 🟨 Wrd0: outbreak (304)
- 🟥 Wrd1: outbreaks (209)
- 🟦 Wrd2: europe (284)
- 🟪 Wrd3: asia (114)
- 🟧 Wrd4: japan (132)
- 🟩 Wrd5: australia (233)
- 🟨 Grp2: outbreak outbreaks (382)
- 🟧 Wrd6: foot (526)
- 🟩 Wrd7: mouth (535)
- 🟫 Grp12: foot mouth (508)
- 🟫 Wrd10: united (439)
- 🟨 Wrd11: kingdom (146)
- 🟥 Wrd12: england (173)
- 🟦 Grp17: united kingdom (140)
- 🟪 Grp18: england united kingdom (284)
- 🟥 Exl1-200: Vacccination - effectiven (200)
- 🟪 Grp23: europe asia japan austral (700)
- 🟫 Grp26: outbreak outbreaks foot m (626)
- 🟨 Grp27: europe asia japan austral (289)
- 🟥 Grp28: europe asia japan austral (64)

Figure 5. The sets created in the query tool were combined in various ways within the grouping tool for further analysis.

Gisting tool

The analyst may "get the gist" of a selected group of documents by displaying the top terms in the set and ordering them by the number of times they appear in that particular group with the gisting tool. This tool displays the top 50 terms in a document group and provides an even broader view of the selected pages.



Word	Docs
disease	168
animal	149
animals	136
veterinary	127
diseases	125
agriculture	91
page	90
health	90
united	89
fever	86
control	85
swine	83
mouth	83
infected	82
food	81
poultry	81
foot	79
virus	76
infection	73
foreign	71
surveillance	70
species	69
veterinarians	67
contact	66

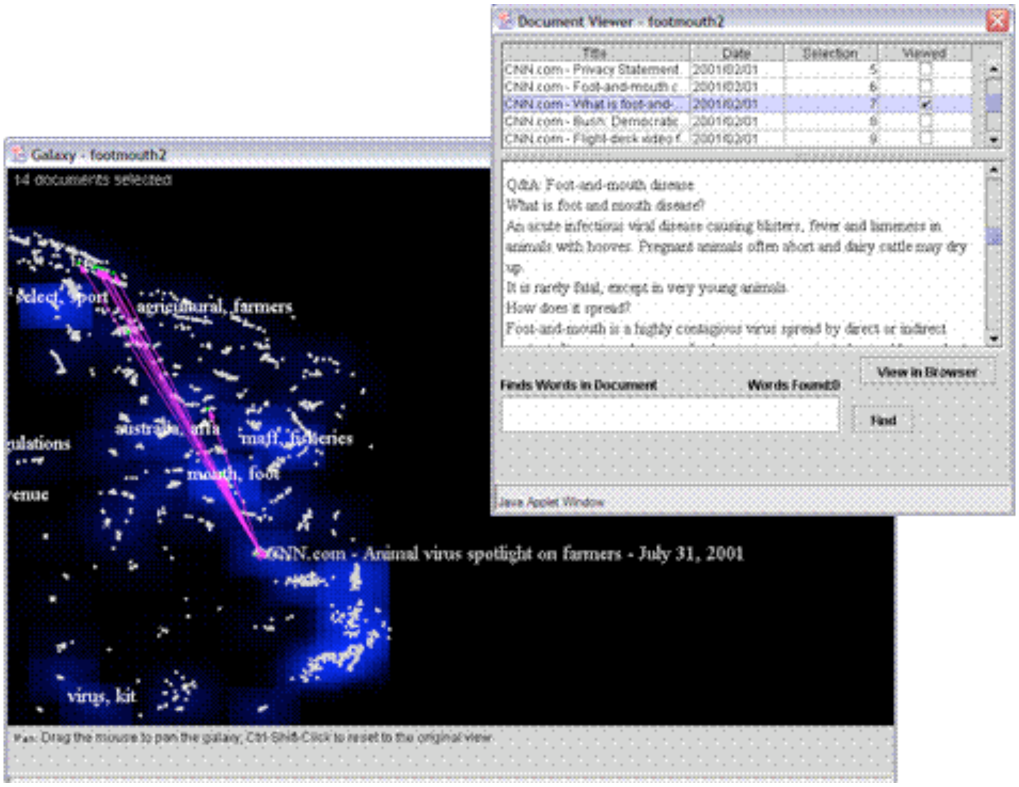
Figure 6. Documents highlighted in green have been selected, and the gisting tool was used to see the top terms or concepts within these documents. This provides the user with a sense of the themes and concepts discussed in these pages.

Document Viewer

Once documents have been identified and selected, users can open them up in the Document Viewer and browse the text, search it for terms of interest, and even see what the Web page looks like in a browser. The document viewer allows you to look at the titles of a group of selected documents and open each one individually. In addition, the tool keeps track of which documents you have looked at by putting a check mark next to those you have opened.

Link mode

This tool lets you see how the Web sites are interconnected and gives you a sense of the structure of your visualization. In addition, the link mode tool selects the documents that are related through hyperlinks so that you can look at them in the document viewer. Link mode gives the user a chance to find documents that may be in other regions of the Galaxy but that are of interest to the subject under investigation.



Title	Date	Selection	Viewed
CNN.com - Privacy Statement	2001/02/01	5	□
CNN.com - Foot-and-mouth c...	2001/02/01	6	□
CNN.com - What is foot-and-...	2001/02/01	7	☑
CNN.com - Bush: Democratic ...	2001/02/01	8	□
CNN.com - Flight-deck video f...	2001/02/01	9	□

Q&A: Foot-and-mouth disease
What is foot and mouth disease?
An acute infectious viral disease causing blisters, fever and lameness in animals with hooves. Pregnant animals often abort and dairy cattle may dry up.
It is rarely fatal, except in very young animals.
How does it spread?
Foot-and-mouth is a highly contagious virus spread by direct or indirect

Figure 7. Link mode will permit researchers to see how pages are linked within a visualization and also selects those documents so that they can be opened in the Document Viewer.

WebTheme is a stand-alone software program that harvests information from the Web rather than a freely accessible search engine. The product is available to users outside PNNL through a license agreement with the laboratory. Government agencies may acquire a license at no cost, though there is a fee for training and installation. Companies, academic institutions, and other non-government agencies may purchase a license, as well as training and installation services.

Conclusions

Using WebTheme allowed us to pinpoint the information our researcher was looking for. The visualization gave us a broad overview of the topic of foot and mouth disease and swine flu. We learned where outbreaks were occurring and what types of foot and mouth disease are being reported. We also were able to find institutions prominent in this field, such as the Department of Agriculture, Fisheries, and Forestry Australia (AFFA) and Ministry of Agriculture, Fisheries and Food (MAFF) in the UK, as well as other universities and research organizations around the world.

The scientist who developed this visualization will use it to determine where to focus his research. In addition, he now is familiar with authors in the field from information gathered from this visualization.

As more and more information becomes available on the World Wide Web, additional tools are needed to aid in making sense of the tangle of data available there. As an information visualization tool, WebTheme effectively displays large sets of HTML pages in a pictorial format. WebTheme visualizations can assist researchers in quickly understanding major themes and concepts of documents retrieved, as well as by providing tools to delve into the data set in order to glean additional insights which generally are not readily evident using standard search engines. New search engines, such as Kartoo (http://www.kartoo.com/) and Vivisimo (http://www.vivisimo.com/) are beginning to incorporate information visualization to assist users in interpreting results. In addition, traditional pay-per-use content vendors such as LexisNexis are introducing visualization tools as part of their product lines (Information Today, 2003)

WebTheme and other visualization applications offer librarians new opportunities to serve customers by providing information professionals with methods to assist in preliminary analysis of rapidly growing amounts of information. As these resources become more widely available in the commercial marketplace, librarians are well positioned to help customers make use of them.

References

Booth, R. K. (2001). *Foot and Mouth Disease Vaccination and Economics.* http://www.ew.ic24.net/fmdv/vacc.htm Announcement accessed from the World Wide Web 3-14-2003.

Chen, C. & Paul, R. J. (2001). Visualizing a Knowledge Domain's Intellectual Structure. *IEEE Computer, 34,* 65-71.

LexisNexis (March 17, 2003). LexisNexis SmartLinx Adds Visualization Technology. *Information Today*, Announcement accessed from the World Wide Web 3/24/2003 http://www.infotoday.com/newsbreaks/wnd030317.shtml

Lembo, M. F. & Buxton, K. A. (2002). "Visualizing Information: Using WebTheme to Visualize Internet Search Results." Slide 8. Presented at the American Society for Information Science, Pacific Northwest Chapter Fall Meeting, Lewis & Clark College, Portland, Oregon. http://www.asis.org/Chapters/asispnc/events/02meeting/presentations/WebThemeASIST.ppt

Pacific Northwest National Laboratory (1999). *SPIRE-Spatial Paradigm for Information Retrieval and Exploration.* Announcement accessed from the World Wide Web. 2-23-2003, http://www.pnl.gov/infoviz/spire/spire.html

Thomas, J. J., Cook, C., Crow, V., Hetzler, B., May, R., McQuerry, D., McVeety, R., Miller, N., Nakamura, G., Nowell, L., Whitney, P., & Wong, P. C. (1999). Human computer interaction with global information spaces - Beyond Data Mining. *Proceedings of British Computer Society Conference*.

Whiting, M. A. & Cramer, N. (2002). "WebTheme^{tm}*:* Understanding Web Information Through Visual Analytics." In J.Hartmanis, G. Goos, & J. van Leeuwen (Eds.), *Lecture Notes in Computer Science* (pp. 460-468). Berlin: Springer-Verlag.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1999). "Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents." In S.K.Card, J. D. MacKinlay, & B. Shneiderman (Eds.), *Readings in Information Visualization: Using Vision to Think.* (pp. 442-449). San Francisco, Calif.: Morgan Kaufmann Publishers, Inc.

**The Hanford Technical Library supports the research of the Pacific Northwest National Laboratory (PNNL) and Hanford contractors. PNNL is operated by Battelle for the Department of Energy.*