WebTheme™: Understanding Web Information through Visual Analytics

Mark A. Whiting and Nick Cramer

Pacific Northwest National Laboratory, PO Box 999, Richland, WA 99352 USA
{Mark.A.Whiting, Nick.Cramer}@pnl.gov

Abstract. WebTheme combines the power of software agent-based information retrieval with visual analytics to provide users with a new tool for understanding web information. WebTheme allows users to both quickly comprehend large collections of information from the Web and drill down into interesting portions of a collection. Software agents work for users to perform controlled harvesting of web material of interest. Visualization and analysis tools allow exploration of the resulting document space. Information spaces are organized and presented according to their topical context. Tools that display how documents were collected by the agents, where they were gathered, and how they are linked further enhance users' understanding of information and its context. WebTheme is a significant tool in the pursuit of the Semantic Web. In particular, it supports enhanced user insight into semantics of large, prestructured or ad-hoc, web information collections.

1 Introduction

Information workers from many domains are discovering the Web as a convenient repository for both casually and formally published information. The Web is now often the primary reference for many information collection activities. Web information grows rapidly, changes frequently, and can be well-advertised and apparent to users or inconspicuous and hidden within the depths of a Web site. For an information user, it is often difficult to develop an overall understanding of a site or discover the most interesting nuggets of information without extensive and time-consuming manual processing.

WebTheme provides an alternative to manual browsing or searching via general search engines to help users understand large collections of Web pages. WebTheme uses both abstract display formats and visual interaction tools to facilitate user understanding. Expressive visualizations engage peoples' perceptual abilities to grasp structure and discern patterns and relationships within information collections. Analytic tools allow both quick, high level investigation to understand document sets as a whole, but also more detailed, specific investigations.

WebTheme is a harvester and a visual analytic tool. Given a URL, a list of URLs, or a query string, WebTheme launches parallel software agents to collect web pages. Pages are processed by text analysis software, clustering and visualization projection

I. Horrocks and J. Hendler (Eds.): ISWC 2002, LNCS 2342, pp. 460–468, 2002.

software, and then made available to to the user for analysis. Specifics about the tool and examples on its use are presented in the following sections.

2 Background

WebTheme is one component of Pacific Northwest National Laboratory's (PNNL) information analytics product offering. The foundation of this product line is the Spatial Paradigm for Information Retrieval and Exploration (SPIRE) system [1], that provides innovative visual tools and approaches to analyzing large sets of textual information. Other components of that product line include tranSPIRE, a tool for translingual visualization, and Topic Islands, a tool that identifies natural transition points at various levels of detail within large individual documents. WebTheme extends the capabilities of the SPIRE software to use the web in two ways – first, to harvest and analyze web information, and second, to deliver this capability and information over the web via a browser. WebTheme was created as an internal research and development project by the Pacific Northwest National Laboratory (PNNL). Subsequent sponsorship from the NASA Goddard Space Flight Center advanced WebTheme from proof-of-concept to the current prototype version. WebTheme is currently deployed at several government installations and the system is now being rolled out at our laboratory as a common desktop tool for our information workers.

3 WebTheme Overview

WebTheme is a tool that allows web information workers to see and interact with information in an uncommon – visual – manner. Web information semantics are better understood for both large document collections and for individual documents within a collection. WebTheme visualizations allow a user to grasp what a document collection describes and represents much more quickly than does text browsing or viewing search engine query results. The visualizations also allow documents to be seen in context within a collection, that is, identifying how they relate to the other documents.

WebTheme consist of two primary components:

- a *web harvester*, which collects web information from both shallow and deep web sources, and
- *visualizations* and a set of *visual analytics tools*, that allow users to see and analyze the harvest results.

3.1 Web Harvester

Users may specify Web-based retrieval in two ways[1]. First, an anchor URL may be specified, from which WebTheme agents will crawl down into the site to retrieve pages. Second, the user may specify a search engine query, to be sent to a general search engine or a site-specific search page. Documents resulting from that search will then be retrieved and processed. Users may set several other parameters to control retrieval agent behavior, such as:

1. Block harvesting of certain Web sites that are known to be of no interest.
2. Limit the search to a particular Internet domain.
3. Specify how many layers or levels of URL links to be followed from the initial target set.
4. Specify minimum and maximum numbers of pages to harvest.
5. Specify that the search should proceed for a specified period of time, rather than setting a target number of documents to be retrieved.
6. Set filters to eliminate certain kinds of unwanted items, including those in foreign languages.

In the case of the URL-based or search engine query, the harvester behaves as a specialized Web client, making contact with Web servers and then requesting and receiving Web documents. Harvesting agents are tasked in parallel from the WebTheme server. This significantly increases harvest speed, because each retrieval may involve delays from the remote servers. The harvesting process retrieves the documents in the initial list, searches those documents for HTML links, and continues by following links on retrieved documents to whatever depth the user requested. Harvesting continues until a user-specified number of documents are retrieved or a user-specified time period has elapsed.

3.2 Document Processing

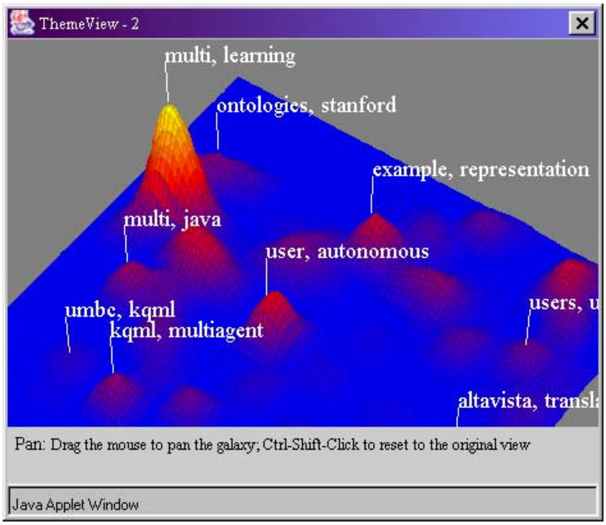
The WebTheme text engine automatically produces a suitable knowledge base of themes (key words) that can be used to distinguish groups within the document collection under analysis. The system creates n-dimensional signature vectors characterizing each document with respect to those themes or topics. The document vectors are clustered and projected from n-space into 2-space, and the lower order projection is used to create the visual representations. The specifics of the signature vectors and visualization projection algorithms are beyond the scope of this paper. However, PNNL is doing extensive research into intelligent software agents that use document and concept signatures to performed enhanced information discovery. We anticipate using this research in conjunction with the use of DAML and other context and semantic enhancement approaches.

[1] A user may also specify a Z39.50 query for retrieval from a digital library, but this functions more like a database query and response than the other approaches.

3.2 Visualizations and Analytics Interface

WebTheme provides two visualizations of the analyzed information space. First, in a ThemeView™ visualization, themes of the document space are shown as a relief map of natural terrain, where taller peaks show dominant themes. This display is particularly good for helping users orient their understanding of the collection. It conveys the main themes in a collection and an overall sense of how they are related. Second, a Galaxy visualization shows the document space with individual web pages presented as stars in on a black space-like background.

We will introduce the visualizations and analytic tools using a practical web information analysis. In this example, we were interested in exploring associations between ontologies and software agents. We began in a typical fashion for most information workers – we entered a query into a general search engine. Unfortunately, merely entering the terms resulted in over 13,000 hits and scanning the resulting links and text quickly became tedious. When we processed this query through WebTheme, we first generated the ThemeView depiction in Figure 1.



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

Java Applet Window

Fig. 1. ThemeView

From this ThemeView, we quickly see strong peaks related to multi (agents), learning, ontologies, and KQML, and we note these as potentially interesting topical areas to explore. Note that words appearing together as a peak label are not processed as a phrase; they are simply terms that are both strongly evident at that point in the collection, not necessarily in the same documents.

The Galaxy visualization for the same data set is shown in Figure 2. Each white dot in the visualization represents an individual Web page that was harvested. The distance between points indicates their thematic similarity. Thus, if points in the visualization are close together, then it is likely that the corresponding documents will contain thematically similar information. If they are far apart, the documents will probably be very different. The spatial layout of the points in the X-Y plane is not meaningful to users—only proximity.

The blue cloud-like areas are a 2-D presentation of the ThemeView peaks. These blue ThemeClouds, resembling nebulae, carry along the ThemeView peak labels to the Galaxy display. Labels are automatically shown for peaks, or locations of greatest density, to provide some orientation to the set when switching between a ThemeView and a Galaxy display.

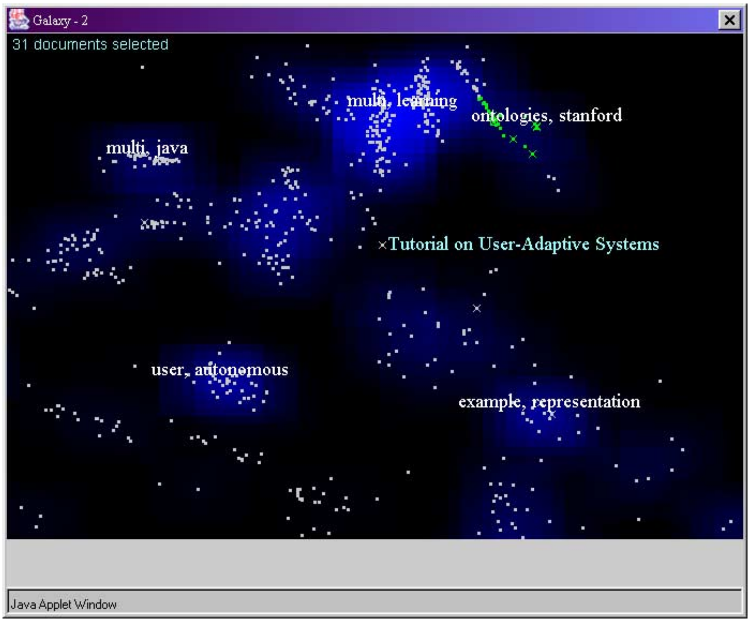
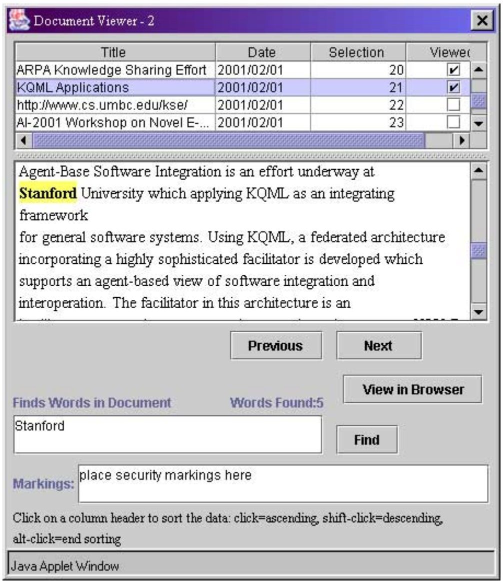


Fig. 2. Galaxy

We may further explore the set using the visual analytic tools. At the simplest level, document titles may be revealed individually or in groups. In Figure 2, one document label is turned on in light blue text color, "Tutorial on User-Adaptive Systems" toward the center of the display. This is one form of browsing this space. Groups of documents may also be selected and reviewed. Curious about the information contained around the ThemeCloud of "ontologies, Stanford", we can select a batch of these documents (selections are highlighted in green on the Galaxy display) and peruse them using the Document Viewer tool (Figure 3). The Document Viewer shows the



Title	Date	Selection	Viewe
ARPA Knowledge Sharing Effort	2001/02/01	20	☑
KQML Applications	2001/02/01	21	☑
http://www.cs.umbc.edu/kse/	2001/02/01	22	□
AI-2001 Workshop on Novel E-...	2001/02/01	23	□

Agent-Base Software Integration is an effort underway at **Stanford** University which applying KQML as an integrating framework
for general software systems. Using KQML, a federated architecture incorporating a highly sophisticated facilitator is developed which supports an agent-based view of software integration and interoperation. The facilitator in this architecture is an

Finds Words in Document Words Found:5

Stanford

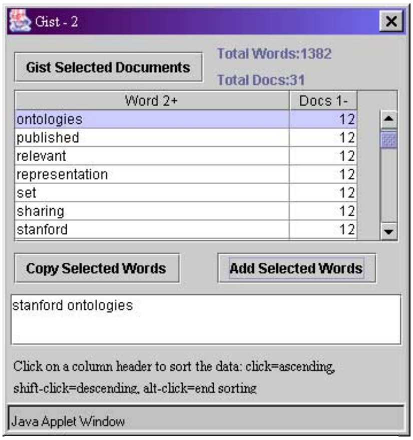
Markings: place security markings here

Click on a column header to sort the data: click=ascending, shift-click=descending, alt-click=end sorting

Fig 3. Document Viewer

list of all of the selected documents in the top panel, displays the text of the document in the middle, highlights words of interest that we would like to search for within documents, and allows us to view the original document in a web browser by clicking on the "View in Browser" button.

There are a couple of tools we can use to better understand the context of both individual and groups of documents. The Gisting Tool (Figure 4) lists frequently occurring terms in our selected documents, reporting the number of documents in that set which contain each word. The list is ordered from highest frequency to lowest. From our selected around the ontologies and Stanford region, we find associated frequent terms such as "representation" and "sharing", possibly new context information for our examination.



Total Words:1382

Total Docs:31

Word 2+	Docs 1-
ontologies	12
published	12
relevant	12
representation	12
set	12
sharing	12
stanford	12

stanford ontologies

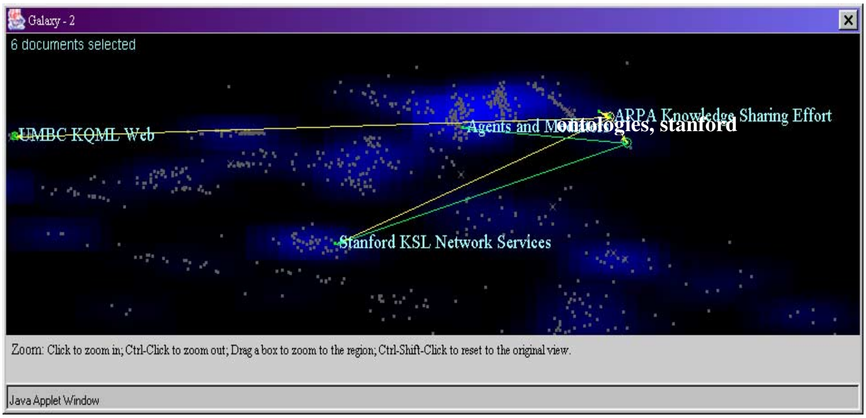
Click on a column header to sort the data: click=ascending, shift-click=descending, alt-click=end sorting

Java Applet Window

Fig. 4. Gist Tool

A Probe Tool allows users to explore the thematic space of the information set, such that when a user clicks anywhere on the Galaxy with the Probe tool, a panel shows the list of themes associated with that position, whether or not a document is located in that exact spot.

At this point we may be interested in using information in ways not anticipated by the authors. To look at how the authors originally linked information, we use the Link tool (Figure 5) to see how documents in the harvested set are hyperlinked to one another. Clicking on a document with the link tool displays these links. Several documents may be displayed with their links at the same time, or the tool may display multiple levels of links. In Figure 5 these links are shown as yellow and green arrows. In our figure, we see links between a document discussing the Knowledge Sharing Effort and the web page for UMBC's KQML web. Using the link tool allows



6 documents selected

Zoom: Click to zoom in; Ctrl-Click to zoom out; Drag a box to zoom to the region; Ctrl-Shift-Click to reset to the original view.

Java Applet Window

Fig. 5. Galaxy documents showing links

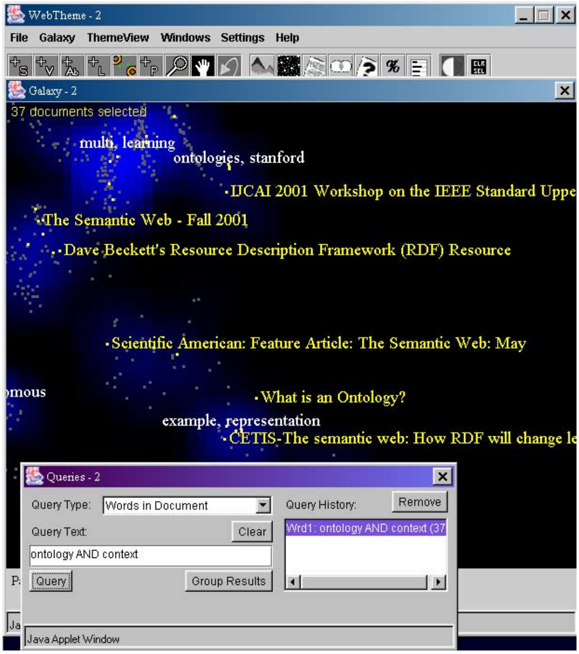
us to understand associations the web page authors found important to explicitly incorporate in their documents, overlaid on the automatically generated topical associations found by WebTheme. It provides us suggestions as to how we may want to use the information and guide our continuing investigations, in this case, we may wish to explore other information from both of these topical areas.

Once we have perused the overall information space, we may be interested in specific information within the harvested collection. WebTheme includes the capability to search the harvested collection using two types of Query Tool searches: Words in Document and Query-by-Example. The Query Tool window is shown in Figure 6.

Words in Document allows Boolean queries. If the Document Viewer is opened, titles matching the selected documents appear in the top of the viewer. Users also have the option to open documents in their Web browser.

Query-by-Example triggers a vector space search and selects the document that is the best match to the query — the one that is closest in the n-dimensional vector space to the query vector. A slider on the Query Tool window allows the user to vary the number of documents selected. By manipulating the slider and thus changing which dots are highlighted, the user can distinguish the location of documents that are closest to the query in the vector space from those that are further away.

A Query History pane is provided on the right side of the Query Tool window. The Query History is a line-by-line record of queries made during a search session. Each line of the query history represents a single query. The first letter on the line signifies the type of query, followed by a colon and the first few words of the query string. The number of documents retrieved is shown in parentheses at the end of the line. The


WebTheme - 2
File Galaxy ThemeView Windows Settings Help
Galaxy - 2
multi, learning
ontologies, stanford
•IJCAI 2001 Workshop on the IEEE Standard Uppe
•The Semantic Web - Fall 2001
•Dave Beckett's Resource Description Framework (RDF) Resource
•Scientific American: Feature Article: The Semantic Web: May
omous
•What is an Ontology?
example, representation
•CETIS-The semantic web: How RDF will change le
Queries - 2
Remove
Clear
Query
Group Results
Java Applet Window

37 documents selected

Query Type: Words in Document

Query History: Wrd1: ontology AND context (37

Query Text: ontology AND context

Fig. 6. Query Tool with selected document titles

most recently executed query will be highlighted in the Query History. In Figure 6, we see the results of a query on "ontology" and "context". The results are highlighted yellow dots on the Galaxy display. We have displayed titles of several of the identified documents (some discussing the Semantic Web).

There are several other tools in the WebTheme toolbox, including a Grouping tool that provides subsetting capabilities, a display that shows the cluster centroids and titles of the visualization, and a Twilight tool which provides an animation of how the harvesting agents collected the documents from the web. Twilight is particularly interesting in that it provides high level visual insight into how web site developers group related web pages as well as how they are discovered by the harvesting agents.

4 Related Work

Much attention is being paid to large scale web harvesting agents. HiWE [2] is a crawler developed by Garcia-Molina to extract content from the hidden web. This capability is invaluable for agents to make use of the Semantic Web. WebTheme seeks to associate this enhanced harvesting capability with the power of visualization. Several systems exploit the advantages of information visualization and information retrieval to various extents. Many visualization interfaces for information retrieval systems present ranked query-document similarity and clustering. VIBE [3] allows users to input query terms, which are associated with a portion of the visualization window, with document icons positioned to illustrate the relevance of documents to the selected terms. TileBars [4] developed at Xerox PARC allows the user to enter search terms as topics. After the system retrieves documents, a graphical, tiled bar is displayed next to the title of each document showing the relationship between the document and query terms. These tools are not presenting the same degree of analytical capability present in WebTheme. Other efforts have focused on creating navigational maps of Web site content. Mappucino allows visual mapping and exploration of web sites [5]. WebTheme can be used in a similar way, but is more focused on both the harvesting and analytical capabilities.

5 Current Status and Future Work

WebTheme is a functional prototype in use at PNNL and NASA. We are strongly encouraging its use as a common desktop system at PNNL to support information workers dealing with various aspects of information overload in their science and technology endeavors. WebTheme is of the class of tools that can help make the goals of the Semantic Web become reality. Combining the behind the scenes work of software agents with the power of information visualization and analytics allows users to truly engage with their information spaces and discover and mold those elements meaningful to their work.

WebTheme works with existing web page descriptions, primarily HTML and text. Enhancements to semantic understanding of web information that will be provided by

DAML and other Semantic Web efforts will enable WebTheme to be even more exciting to users. We are currently working to enhance WebTheme in two particular areas of interest to the Semantic Web community. First, we want to employ much more sophisticated agents to work for the WebTheme user. We are already able to interact with our agents' activities to a certain extent; we envision these agents becoming very able research assistants. We have active research projects in agents working with large document spaces and ontology development. As it stands, WebTheme is valuable to the Semantic Web community, particularly in the area of unintended use of information. WebTheme's ability to unveil implicit context via the text analysis and visualizations provides a complementary capability to the developers of semantic description technologies for web documents. WebTheme also provides a capability to help understand ad-hoc collections where no semantic description exists. This ability can enable automatic generation of semantic descriptions following the text analysis activities.

References

1. Battelle Memorial Institute. 2001. "SPIRE - Spatial Paradigm for Information Retrieval and Exploration" http://www.pnl.gov/infoviz/spire/spire.html .
2. Garcia-Molina, H., Raghavan, Sriram. "Crawling the Hidden Web." *Stanford Database Group Publication Server*. http://dbpubs.stanford.edu:8090/pub/2001-19. May 2001.
3. Olsen, K. A., Korfhage, R. R., Sochats, K.M., Spring, M. B, & Williams, J. G. Visualization of a document collection: The VIBE system. Information Processing and Management, 29, 1 (1993) 69-81
4. Hearst, M. A. TileBars: Visualization of Term Distribution Information in Full Text Information Access, in Proceedings of CHI '95 (Denver, Colorado, May 7-11, 1995) pp. 59-66.
5. IBM Alphaworks. 1999. "Mappucino". http://www.alphaworks.ibm.com/tech/mapuccino.