

PACKAGE ID - 001120SGIIP00 WEBTHEME

KWIC TITLE - Thematic World Wide Web Visualization System

AUTHORS - Pottier, M.
Pacific Northwest Lab., Richland, WA (United States)

Adams, K.
Pacific Northwest Lab., Richland, WA (United States)

Crow, V
Pacific Northwest Lab., Richland, WA (United States)

Pottier, N.
Carnegie Mellon University, Pittsburgh, PA (United States)

LIMITATION CODE -COPY **AUDIENCE CODE** - LIM

COMPLETION DATE - 07/31/1996 **PUBLICATION DATE** - 07/01/1996

DESCRIPTION - WebTheme is a system designed to facilitate world wide web information access and retrieval through visualization. It consists of two principal pieces, a WebTheme Server which allows users to enter in a query and automatically harvest and process information of interest, and a WebTheme browser, which allows users to work with both Galaxies and Themescape visualizations of their data within a JAVA capable world wide web browser. WebTheme is an Internet solution, meaning that access to the server and the resulting visualizations can all be performed through the use of a WWW browser. This allows users to access and interact with SPIRE (Spatial Paradigm for Information Retrieval and Exploration) based visualizations through a web browser regardless of what computer platforms they are running on. WebTheme is specifically designed to create databases by harvesting and processing WWW home pages available on the Internet.

PACKAGE CONTENTS - Media Directory; Software Abstract; Media Includes Source Code;

SOURCE CODE INCLUDED? - Yes

MEDIA QUANTITY - 1 CD Rom

METHOD OF SOLUTION - WebTheme is a complete harvesting, processing and visualization solution. it includes three primary processing components: harvesting Engines, a Text Engine (SID), and Visualizations. Harvesting engines allow WebTheme to access existing world wide web information services such as Lycos and Yahoo to automatically perform queries and collect the resulting documents. In the case of Lycos, WebTheme uses a Perl script called Query underscore Lycos which opens a connection to the Lycos server, submits the query and gathers the web page abstracts. In

PACKAGE ID - 001120SGIIP00 WEBTHEME

METHOD OF SOLUTION - (CONT) the case of Yahoo, WebTheme utilizes a script called LinkGoat, which submits a query to Yahoo, gathers a list of http addresses from the results of the query, and visits each site referenced in the list and collects its text. The gathered information, whether this be web page abstracts or the full text version of the pages are stored locally on the WebTheme server and can be accessed by the user through the WebTheme browser. Once information is harvested, WebTheme formats the information and processes it through SID (System for Information Discovery), a text engine developed at PNNL. The text engine automatically extracts the themes found in the web pages, and generates numeric vectors describing the contents of each document. The document vectors which were produced by SID are then further processed to create both a Galaxies and Themescape visualization. These visualizations are based on the SPIRE system, and allow users to rapidly understand the relationship between web pages. The visualizations are accessed through a WebTheme browser created in JAVA.

COMPUTER - SILICON GRAPHIC

OPERATING SYSTEMS - WebTheme Server IRIX5.3; WebTheme Browser Windows NT, Windows 96, Solaris, SunOS, IRIX

PROGRAMMING LANGUAGES - WebTheme Server C, C++, Perl; WebTheme Browser JAVA

SOFTWARE LIMITATIONS - WebTheme has computational and disk space limitations. The processing for creating new databases and running them through SID all take place on the machine running the WebTheme server. In our tests and work with an SGI Indy (64MB RAM, 10Gb Hard Drive), three concurrently processing datasets can easily cause a machine to have a significantly degraded processing performance. JAVA is still currently under development in both the Microsoft and Netscape browsers. Due to the partial implementation of this code, the WebTheme Browser Applets written in JAVA may not run reliably on some platforms (MacOS, Windows 3.1, IRIX)

SOURCE CODE AVAILABLE (Y/N) - Y

UNIQUE FEATURES - WebTheme is a complete solution for querying, harvesting and visualizing the contents of web based searches. It can be used both for visualizing the contents of world wide web pages as well as accessing the contents of intranet databases. From WebTheme, users can set up a query, create custom made filters for eliminating unwanted documents from the database, control the SID processing parameters, and determine how visualizations are created. The resulting databases and visualizations can then be accessed remotely from WebTheme through an Internet browser and interacted with to retrieve and analyze web pages contained in the database. The visualizations WebTheme creates are content based

PACKAGE ID - 001120SGIIP00 WEBTHEME

UNIQUE FEATURES - (CONT) visualizations. The information and organization of both the Galaxies and Themscapes views are based on the actual text of the harvested pages. This is unique because most web visualization systems rely on the actual link structure (hyperlinks connecting pages) to organize and visualize their information. Content based visualizations are superior to hyperlink visualizations because they actually reveal the themes and relationships found between the pages. WebTheme is truly a cross platform solution. Because of its JAVA based visualizations (Galaxies/Themscapes), it can be used on a wide range of different platforms. The harvesting engine which is part of WebTheme (LinkGoat), was specifically tailored to harvest information for visualizations. It employs simple but effective algorithms for weeding out pages with too many links, as well as some novel approaches to harvesting entire web sites by remaining within a specified domain name.

RELATED SOFTWARE - WebTheme utilizes SID, a text engine developed to characterize the contents of textual documents for the purpose of visualization. SID was developed on internal LDRD funds at PNNL. As WebTheme is an Internet based application, it requires the use of a JAVA enabled WWW browser to access the server and interact with the visualizations. Both the Netscape and Internet Explorer browser have been tested.

HARDWARE REQS - WebTheme Server: Silicon Graphics Workstation, 64MB, 10GB hard drive. WebTheme Browser: Sun Workstation, PC, Mac, SGI.

TIME REQUIREMENTS - There are two distinguishable bottlenecks for harvesting and creating visualizations: Harvesting Engines, The Yahoo and http approaches to harvesting information require WebTheme to sequentially visit and harvest the text from many different web sites. Based on Internet performance, bandwidth, and the size of the site, the process of harvesting can easily take several hours; SID Processing/MDS, The processes used to compute the document vectors in SID, and then create the MDS Galaxies plots are also computationally expensive. For a 2000+ document corpus, computations can often times exceed a half hour.

ABSTRACT STATUS - Submitted October 16, 1996. Released AS-IS 12/3/96

SUBJECT CLASS CODE - P

KEYWORDS -

COMPUTER PROGRAM DOCUMENTATION
W CODES
INFORMATION RETRIEVAL
COMPUTER GRAPHICS
DATA PROCESSING
DATA COMPILATION
INTERNET

EDB SUBJECT CATEGORIES -

990200

E S T S C
ENERGY SCIENCE & TECHNOLOGY SOFTWARE CENTER
SOFTWARE ABSTRACT

PAGE 4

DATE 03/13/2002

PACKAGE ID - 001120SGIIP00 WEBTHEME

SPONSOR - DOE/ER

PACKAGE TYPE - AS - IS