WebCiao: A Website Visualization and Tracking System

Yih-Farn Chen
Eleftherios Koutsofios

AT&T Labs - Research
600 Mountain Ave.
Murray Hill, New Jersey

Abstract

WebCiao is a system for visualizing and tracking the structures of websites by creating, differencing, and analyzing archived website databases. The architecture of WebCiao allows users to create customized website analysis tools by combining a set of query and analysis operators on a *virtual database pipeline*. Each virtual database sent on the pipe can be converted to directed graphs, database views, or HTML reports. Within a graph view, operators can be fired from any graph node to study a selected neighborhood. WebCiao helps creators of large websites to monitor the dynamics of structural changes closely. It also helps web surfers to quickly identify new products and services from a website. An on-line demo, *Website News*, based on the WebCiao technology, has helped sharpen our focus with its daily analysis of new web contents from the internet and telecommunications industries.

1. Introduction

The complexity and ever-changing nature of major websites are presenting problems to both website creators and frequent visitors to those sites. For website creators, detecting structural changes and maintaining website integrity are critical before publishing new web contents outside the firewall. On the other hand, links to new contents frequently go unnoticed by visitors because they cannot locate new stuff easily in the web hierarchy.

WebCiao is a system that analyzes web pages of selected websites, stores their structure information in a database, and then allows users to query and visualize that database with graphs or HTML pages. WebCiao can also be used to analyze the structural differences of archived web databases by highlighting added, deleted, and changed pages or links.

WebCiao was created to help both surfers and maintainers of complex websites. Web surfers can use customized queries to quickly identify new products or services related to a particular topic without manually going through individual hyperlinks and pages. Website maintainers can use it to (a) visually track structural changes made by various web page authors and (b) perform global analysis to detect missing links or orphan pages before moving pages from staging machines to external servers.

The visualization component of CIAO-html, the predecessor or WebCiao, was used in WebGUIDE [5] as a visual aide to the textual differencing capability of AIDE[4]. This paper focuses on how to combine WebCiao's query, analysis, and visualization operators in a *virtual database pipeline* to perform various website visualization and tracking tasks. It also presents a new interface, *Website News*, as an alternative to deliver updates on website changes without complex user interactions.

We shall describe the architecture of WebCiao in Section 2, the basic query and visualization capabilities in Section 3, a major application of WebCiao, *Website News*, in Section 4, and discuss related work in Section 5, followed by summary and future work in Section 6.

2. Architecture of WebCiao

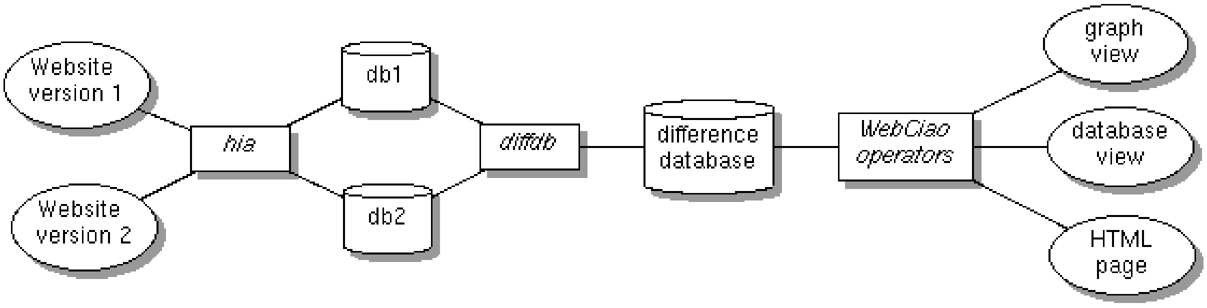


Figure 1: Architecture of WebCiao

Figure 1 shows the architecture of WebCiao, which consists of three major components:

- *HTML Information Abstractor*: *hia* extracts web pages from a website and converts them into a CQL[7] database according to an Entity-Relationship model, which includes URL entities such as HTML and image files, relationships such as image and text hyperlinks, and their associated attributes such as URL addresses and anchor text. *hia* allows users to specify, with regular expressions or lists, what pages to include or exclude during recursive retrievals of a website’s pages. It also allows users to specify the depth of recursive search, similar to that provided by WebCopy[15], but WebCopy simply gets pages, while *hia* also converts the pages into a database.
- *Database Differencing Tool*: *diffdb* takes two versions of a database, compares page checksums and links, and creates a difference database that consists of all pages and links with tags that specify each as added, deleted, changed, or unchanged.
- *WebCiao operators*: The WebCiao system consists of a set of query and analysis operators that read and write *virtual databases*. Each virtual database consists of a subset of entities(pages) and relationships(links) retrieved from the complete database. A set of *view operators* takes any virtual database and converts it to a directed graph, a database view, or an HTML page report. Since query and analysis operators are interchangeable, a *virtual database pipeline* can be constructed to perform complex operations before the results are turned into graphs or other forms of reports. These operators can be used on command lines, in shell scripts, or invoked by WebCiao’s graphical interface shown in Section 3, or a web interface discussed in Section 4.

WebCiao inherits our years of software reverse engineering [14] experience in querying, analyzing, and visualizing large and complex software structures. WebCiao is an instance of Ciao[6], a multi-language graphical navigator for software and document repositories. Ciao has been instantiated for C, C++, Java, Ksh, HTML, and some other languages and business databases. The architecture style shown in Figure 1 applies to all languages. Except for HTML-specific tools like *hia* and operators that communicate with web browsers, the complete set of GUI, query, and analysis tools is generated automatically from a CIAO specification file less than 200 lines long.

3. Query and Visualization Operators

WebCiao consists of several operators that can be combined on its virtual database pipeline:

- *Selection* operators: **ciao_eset** and **ciao_rset** retrieve a set of entity or relationship records according to the selection criteria.
- *Closure* operator: **ciao_closure** performs reachability analysis according to the specified level of recursion.
- *Focus* operator: **ciao_focus** performs fan-in/fan-out analysis in the neighborhood of selected pages.
- *Database View* operators: **ciao_eview** and **ciao_rview** generate database views.
- *Graph View* operators: **ciao_egraph** and **ciao_rgraph** generate graph views.
- *Visit* operator: **ciao_visit** sends requests to a web browser to retrieve corresponding pages.

Except for *View* and *Visit* operators, all operators read and write virtual databases. Additional analysis and view operators can be written to interface with the virtual database, which is simply an archive of plain text database files that can be unpacked easily.

The following examples demonstrate how WebCiao operators can be combined for different analysis and visualization tasks by using a pipeline. In a three-layer content analysis of AT&T’s website (four-layer link analysis), we captured 1,019 web pages and 13,649 links, and created a 1.24 MB CQL[7] database. Here are some sample queries we can make on a command line:

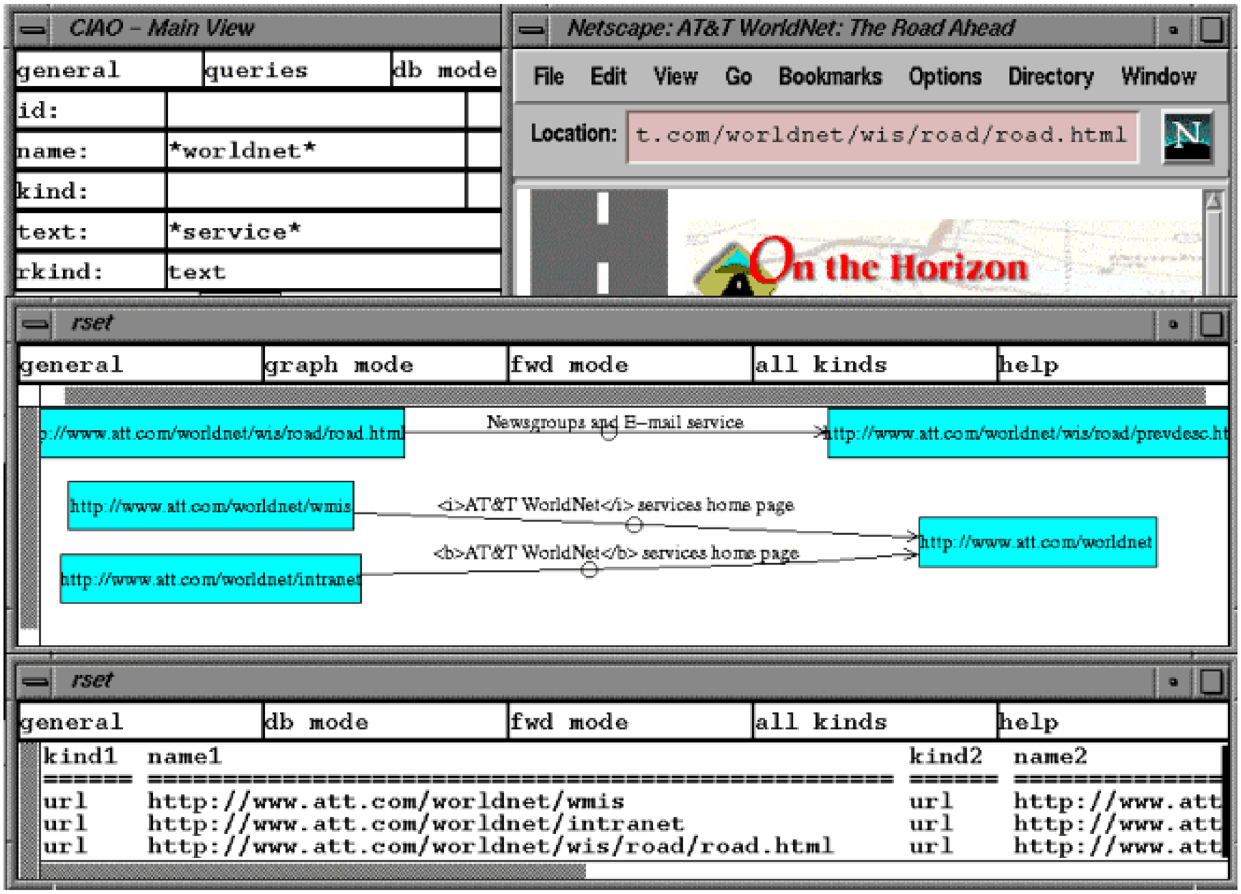
- Show URL addresses that match the pattern ’mailto:*lucent*’; we were surprised to find that there were still a few on AT&T’s site:

```
$ ciao_eset url ’mailto:*lucent*’ | ciao_eview url -
name                                                                      kind
========================================================================= ======
mailto:pbk@lucent.com                                                     url
mailto:rzardetto@lucent.com                                               url
mailto:pbk@lucent.com,michaelmills@att.com,josepharias@lucent.com         url
mailto:josepharia@lucent.com                                              url
```

- Find all links from worldnet pages whose anchor text match the pattern ’*service*’:

```
$ ciao_rset url ’*worldnet*’ url - rkind=text text=’*service*’
```

Figure 2 shows the multiple views generated from the last query, which was formulated in WebCiao’s main query table in the top left corner. By selecting the *graph mode* before executing the query, the resulting database was converted to the graph shown in the middle. Selecting the *database mode* with the same query specification created the database view in the bottom. The user also visited one of the graph nodes, *road.html*, through the *visit* operator on its node menu. Additional query and analysis operators can be fired from any graph node through the node menu.


CIAO – Main View
general
queries
db mode
Netscape: AT&T WorldNet: The Road Ahead
File
Edit
View
Go
Bookmarks
Options
Directory
Window
Location:
t.com/worldnet/wis/road/road.html
On the Horizon
rset
general
graph mode
fwd mode
all kinds
help
p://www.att.com/worldnet/wis/road/road.html
Newsgroups and E–mail service
http://www.att.com/worldnet/wis/road/prevdesc.ht
http://www.att.com/worldnet/wmis
<i>AT&T WorldNet</i> services home page
http://www.att.com/worldnet
http://www.att.com/worldnet/intranet
<b>AT&T WorldNet</b> services home page
rset
general
db mode
fwd mode
all kinds
help

id:	
name:	*worldnet*
kind:	
text:	*service*
rkind:	text

```
kind1  name1                                                    kind2  name2
====== ======================================================== ====== =============
url    http://www.att.com/worldnet/wmis                         url    http://www.att
url    http://www.att.com/worldnet/intranet                     url    http://www.att
url    http://www.att.com/worldnet/wis/road/road.html           url    http://www.att
```

Figure 2: A WebCiao Query and the Resulting Database and Graph Views

Note that Figure 2 shows a WebCiao snapshot for a *single-version database*, while the following two examples were run on a *difference database* created for AT&T's website based on the changes from 11/27/96 to 12/02/96.

- Show new web pages that match the pattern '*press*':

```
$ ciao_eset url '*press*' etag=added | ciao_eview url -
name                                                   kind   etag
====================================================== ====== ==========
http://www.att.com/press/1196/961127.csa.html         url    a
http://www.att.com/press/1196/961125.bsb.html         url    a
http://www.att.com/press/1196/961127.cia.html         url    a
```

- Use a graph to show changes in the neighborhood of AT&T's Easy Commerce page:

```
$ ciao_focus -l2 url http://www.att.com/easycommerce | ciao_rgraph url - url -
```

Figure 3 shows the result of the last query. Changed web pages are marked yellow, deleted web pages are marked white, while green pages are those that stay the same. The picture allows us to easily identify incoming links. If the Easy Commerce page has to be modified, deleted, or moved, we know what other pages need to be checked or updated. It also allows us to identify new or deleted topics occurred on that page.

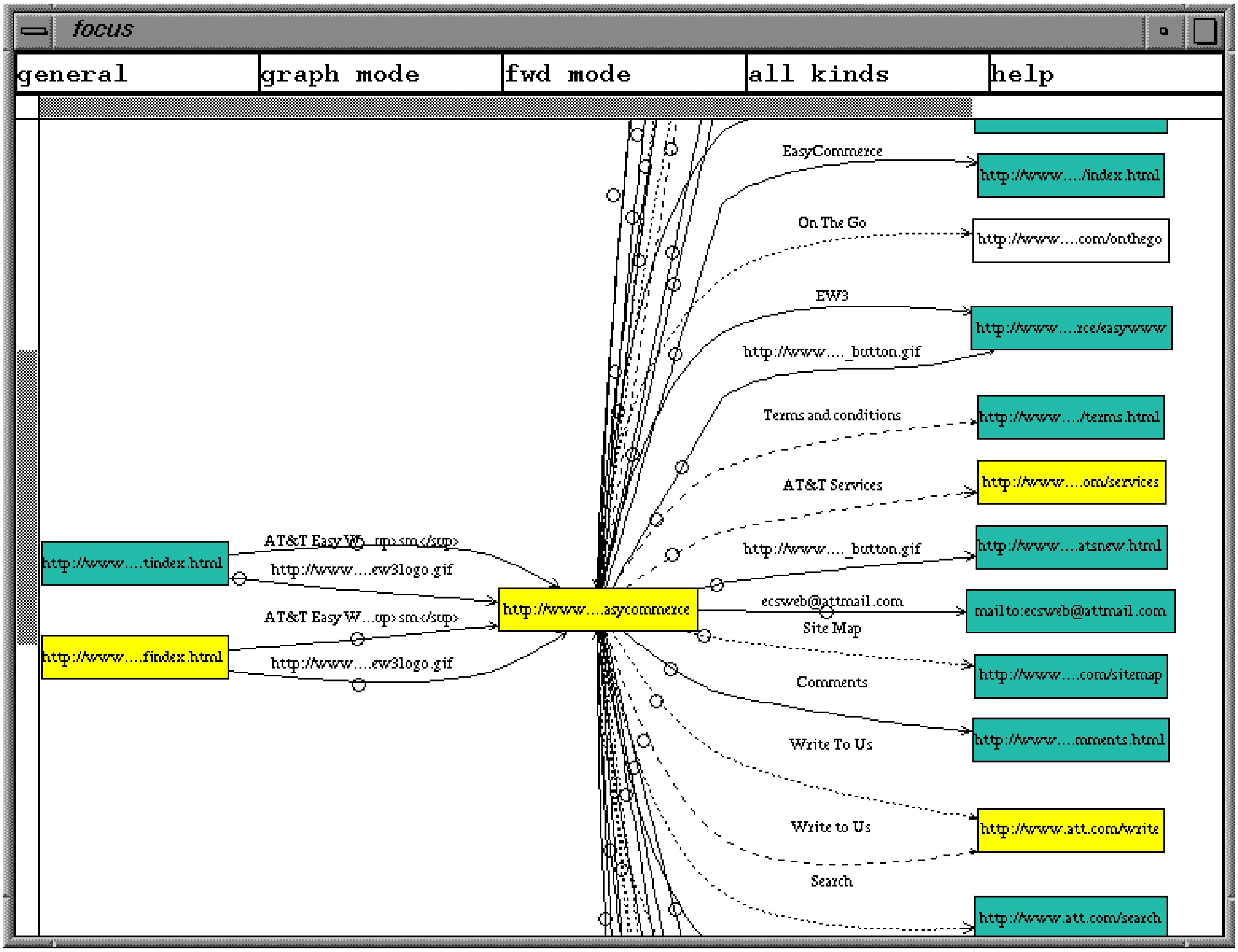

focus
general
graph mode
fwd mode
all kinds
help
EasyCommerce
http://www..../index.html
On The Go
http://www....com/onthego
EW3
http://www....rce/easywww
http://www...._button.gif
Terms and conditions
http://www..../terms.html
AT&T Services
http://www....om/services
http://www...._button.gif
http://www....atsnew.html
ecsweb@attmail.com
mailto:ecsweb@attmail.com
Site Map
http://www....com/sitemap
Comments
http://www....mments.html
Write To Us
http://www.att.com/write
Write to Us
Search
http://www.att.com/search
http://www....tindex.html
AT&T Easy W5...up>sm</sup>
http://www....ew3logo.gif
http://www....findex.html
AT&T Easy W...up>sm</sup>
http://www....ew3logo.gif
http://www....asycommerce

Figure 3: Changes in the Neighborhood of AT&T's Easy Commerce Web Page (11/27/96 to 12/02/96)

As an example of more complex operations, suppose we are interested in finding information under a particular node on a website, similar to the functionality provided by GlimpseHTTP[12] (and recently, WebGlimpse[13]). In WebCiao, we can simply run a *closure* operator performing reachability analysis on the selected node followed by a *selection* operator based on the URL addresses, anchor text of each link, or page contents (if archived). For example, the following virtual database pipeline reports the set of URL's in the first three layers of pages reachable from *http://www.att.com/news* whose addresses match the pattern "**worldnet**" on December 10, 1996:

```
$ ciao_closure -l3 url 'http://www.att.com/news' | ciao_eset url '*worldnet*'
| ciao_eview url -

name                                                                   kind
====================================================================== ======
http://www.att.com/w3403/attworldnetservice/crystal.html               url
```

```
http://www.att.com/worldnet/wis/sky/signup.html                     url
http://www.att.com/worldnet                                         url
http://download.worldnetall.com/mainL54Q.htm                        url
http://www.att.com/w3403/attworldnetservice/legal1.html             url
http://www.worldnet.att.net                                         url
http://www.att.com/worldnet/wis/game/gamstrt.html                   url
http://www.att.com/worldnet/wis/                                    url
...
```

WebCiao is also very effective in computing metrics on website structures or their changes. For example, to compute the change rate of pages on a website (such as similar work done in [18], which used a dynamic trace), we can simply run separate difference database queries that limit the selections to added, deleted, or changed links, and then pipe the results to a simple counting or statistics tool. For example, the following query reports that 9 links were added in the first three layers of AT&T's website from 11/27/96 to 12/02/96:

```
$ ciao_closure -l3 url 'http://www.att.com' | ciao_rflat url - url - rtag=added
| wc -l
       9
```

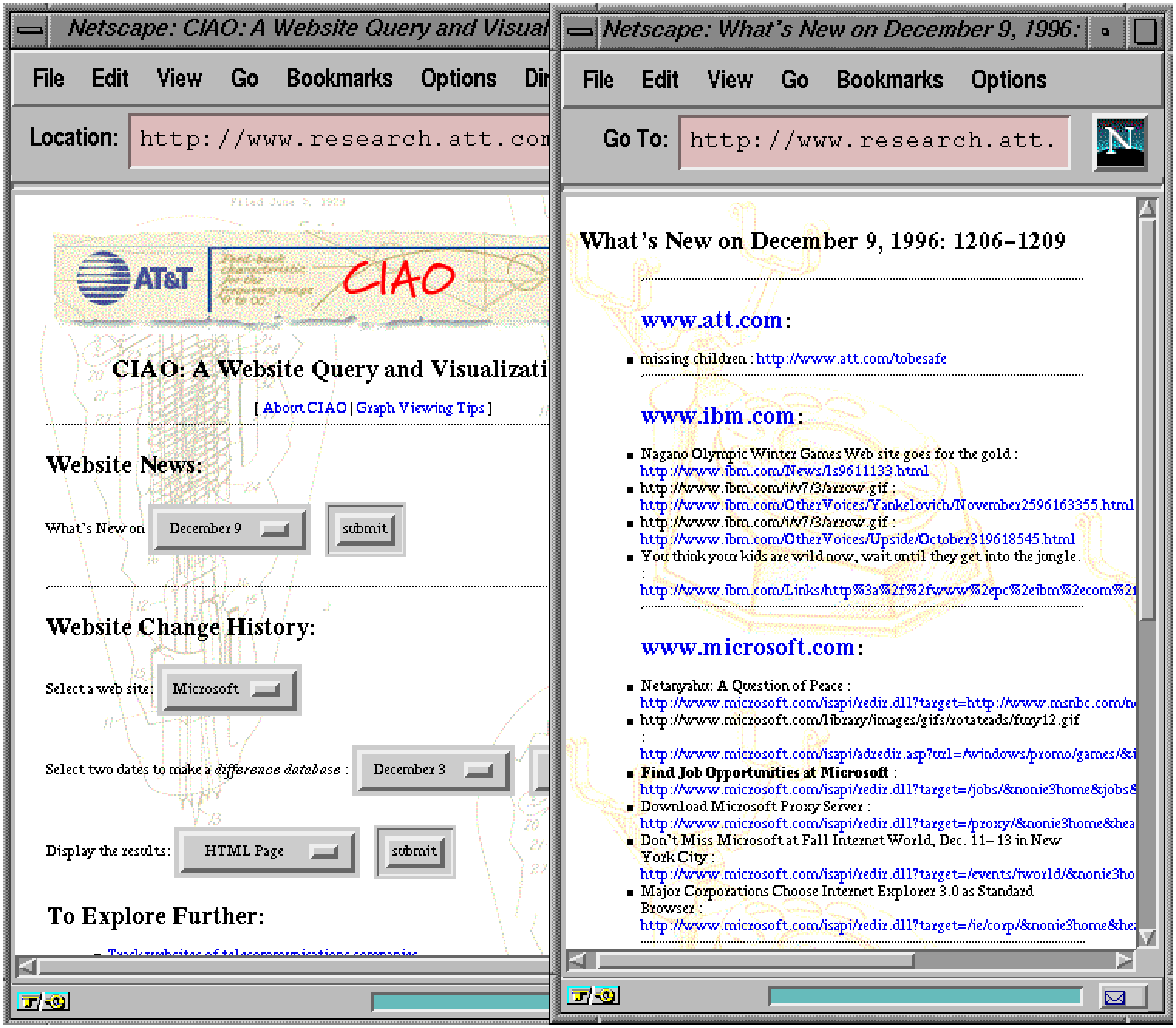
ciao_rflat converts a virtual database to a flat file database that can be easily processed by line-oriented tools like *wc*, a word-counting tool.

4. Application: Website News

A demo called Website News[9] has been set up to demonstrate applications of the WebCiao difference databases. We archive the home pages of a selected set of frequently-visited websites (such as AT&T, Microsoft, Netscape, and IBM) on a daily basis. Users can find new links added everyday on these websites by using the *Website News* web interface page shown on the lefthand side of Figure 4. The righthand side shows the news report on Monday, December 9, 1996 for the selected websites. Users can click on any of these new links directly to get to the new pages without going through the home page and figuring out the new links added. The report shows, for example, that AT&T has added a link to a page about *missing children*, Microsoft added a web page about *Job Opportunities*, etc. since last business day (Friday, December 6).

A user can also review the change history of a website by picking any two dates with archived databases and make a *difference database* to see all the links added and deleted. For example, the lefthand side of Figure 5 displays the HTML form for changes in the URL links on Netscape's home page from December 4 to December 9, 1996. This information is useful for users who last visited the Netscape site on December 4 and are interested in finding out changes occurred on Netscape's site since then.

Since a website usually has layers of web pages, we have started archiving several layers of web pages for selected websites so that deep analysis on these websites can be performed. The righthand side of Figure 5 shows that while AT&T's home page itself did not have any new or deleted links on 11/18/96, one of its links, the AT&T News page, has some substantial changes in terms of news items deleted and added. Reports like these have allowed us to monitor the dynamics of each website closely and recursively.



Netscape: CIAO: A Website Query and Visual
File Edit View Go Bookmarks Options Dir
Location:
AT&T
CIAO

http://www.research.att.com

CIAO: A Website Query and Visualizati

[About CIAO | Graph Viewing Tips]

Website News:

What's New on December 9 submit

Website Change History:

Select a web site: Microsoft

Select two dates to make a *difference database* : December 3

Display the results: HTML Page submit

To Explore Further:

Netscape: What's New on December 9, 1996:
File Edit View Go Bookmarks Options
Go To:

http://www.research.att.

What's New on December 9, 1996: 1206-1209

www.att.com:

- missing children : http://www.att.com/tobesafe

www.ibm.com:

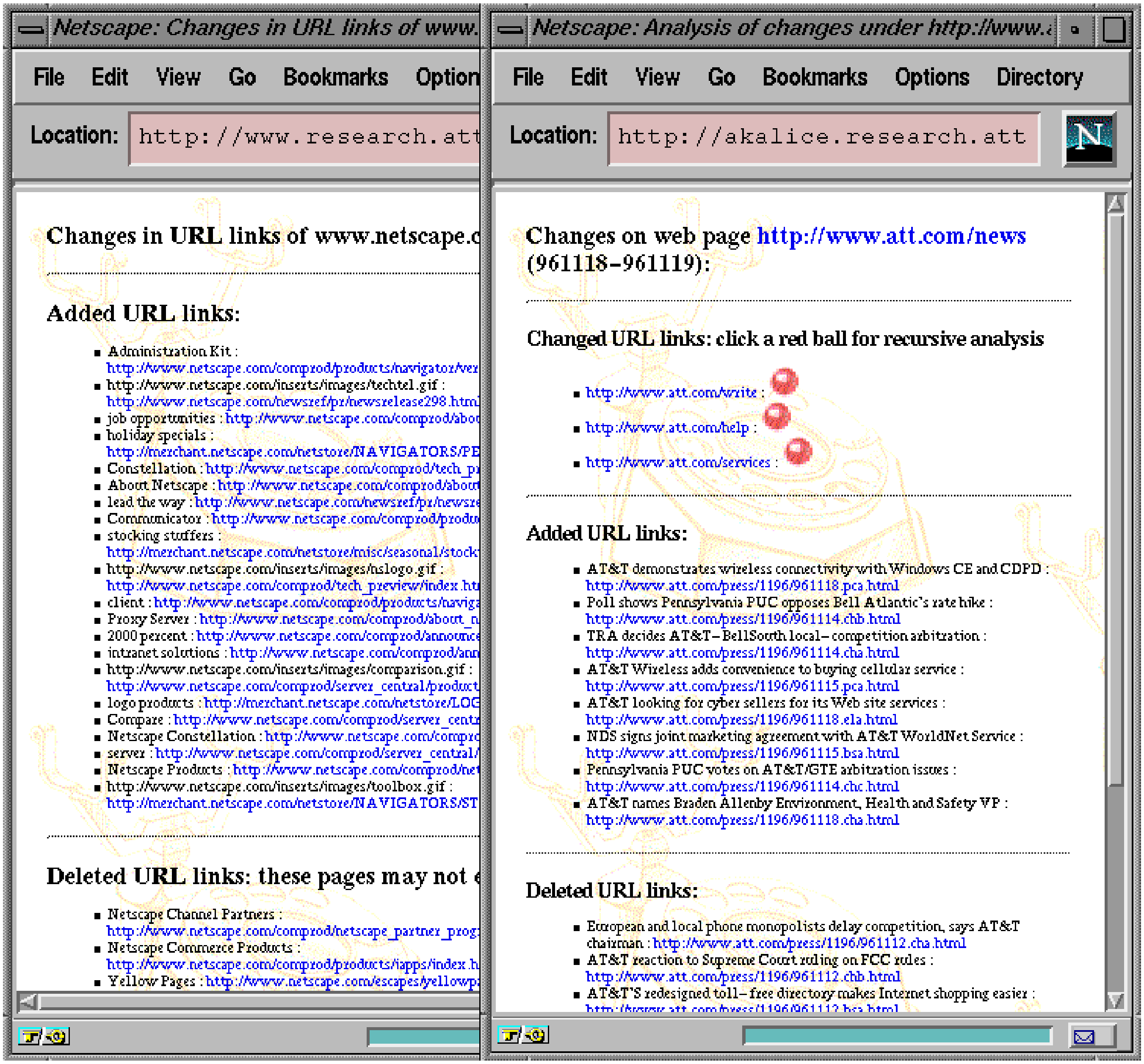
- Nagano Olympic Winter Games Web site goes for the gold : http://www.ibm.com/News/ls9611133.html
- http://www.ibm.com/i/v7/3/arrow.gif : http://www.ibm.com/OtherVoices/Yankelovich/November2596163355.html
- http://www.ibm.com/i/v7/3/arrow.gif : http://www.ibm.com/OtherVoices/Upside/October319618545.html
- You think your kids are wild now, wait until they get into the jungle. : http://www.ibm.com/Links/http%3a%2f%2fwww%2epc%2eibm%2ecom%2f

www.microsoft.com:

- Netanyahu: A Question of Peace : http://www.microsoft.com/isapi/redir.dll?target=http://www.msnbc.com/n
- http://www.microsoft.com/library/images/gifs/rotateads/fury12.gif : http://www.microsoft.com/isapi/adredir.asp?url=/windows/promo/games/&i
- **Find Job Opportunities at Microsoft** : http://www.microsoft.com/isapi/redir.dll?target=/jobs/&nonie3home&jobs&
- Download Microsoft Proxy Server : http://www.microsoft.com/isapi/redir.dll?target=/proxy/&nonie3home&hea
- Don't Miss Microsoft at Fall Internet World, Dec. 11-13 in New York City : http://www.microsoft.com/isapi/redir.dll?target=/events/iworld/&nonie3ho
- Major Corporations Choose Internet Explorer 3.0 as Standard Browser : http://www.microsoft.com/isapi/redir.dll?target=/ie/corp/&nonie3home&he

Figure 4: Website News: Interface Page (left) and News Report (right)

In response to a request from an AT&T marketing group, we have created a separate Website News demo for the telecommunications industry group [10], which includes the websites of AT&T, MCI, Sprint, and the seven Regional Bell Operating Companies (RBOCS). It also demonstrates that the same set of CGI scripts we developed can be reused on a collection of websites that might be of interest to a community of users. Website News could also be useful in reducing the network usage of major corporations or internet service providers by providing the change information of popular websites upfront and thus eliminating many unnecessary downloads by users.


Netscape: Changes in URL links of www.
File
Edit
View
Go
Bookmarks
Option
Location:
http://www.research.att
Netscape: Analysis of changes under http://www.
File
Edit
View
Go
Bookmarks
Options
Directory
Location:
http://akalice.research.att
N

Changes in URL links of www.netscape.c

Added URL links:

- Administration Kit : http://www.netscape.com/comprod/products/navigator/ver
- http://www.netscape.com/inserts/images/techtel.gif : http://www.netscape.com/newsref/pr/newsrelease298.html
- job opportunities : http://www.netscape.com/comprod/abo
- holiday specials : http://merchant.netscape.com/netstore/NAVIGATORS/PE
- Constellation : http://www.netscape.com/comprod/tech_p
- About Netscape : http://www.netscape.com/comprod/about
- lead the way : http://www.netscape.com/newsref/pr/newsre
- Communicator : http://www.netscape.com/comprod/produ
- stocking stuffers : http://merchant.netscape.com/netstore/misc/seasonal/stock
- http://www.netscape.com/inserts/images/nslogo.gif : http://www.netscape.com/comprod/tech_preview/index.ht
- client : http://www.netscape.com/comprod/products/naviga
- Proxy Server : http://www.netscape.com/comprod/about_n
- 2000 percent : http://www.netscape.com/comprod/announce
- intranet solutions : http://www.netscape.com/comprod/ann
- http://www.netscape.com/inserts/images/comparison.gif : http://www.netscape.com/comprod/server_central/product
- logo products : http://merchant.netscape.com/netstore/LOG
- Compare : http://www.netscape.com/comprod/server_centr
- Netscape Constellation : http://www.netscape.com/comprod
- server : http://www.netscape.com/comprod/server_central/
- Netscape Products : http://www.netscape.com/comprod/net
- http://www.netscape.com/inserts/images/toolbox.gif : http://merchant.netscape.com/netstore/NAVIGATORS/ST

Deleted URL links: these pages may not e

- Netscape Channel Partners : http://www.netscape.com/comprod/netscape_partner_prog
- Netscape Commerce Products : http://www.netscape.com/comprod/products/iapps/index.h
- Yellow Pages : http://www.netscape.com/escapes/yellowp

Changes on web page http://www.att.com/news (961118-961119):

Changed URL links: click a red ball for recursive analysis

- http://www.att.com/write :
- http://www.att.com/help :
- http://www.att.com/services :

Added URL links:

- AT&T demonstrates wireless connectivity with Windows CE and CDPD : http://www.att.com/press/1196/961118.pca.html
- Poll shows Pennsylvania PUC opposes Bell Atlantic's rate hike : http://www.att.com/press/1196/961114.chb.html
- TRA decides AT&T–BellSouth local–competition arbitration : http://www.att.com/press/1196/961114.cha.html
- AT&T Wireless adds convenience to buying cellular service : http://www.att.com/press/1196/961115.pca.html
- AT&T looking for cyber sellers for its Web site services : http://www.att.com/press/1196/961118.ela.html
- NDS signs joint marketing agreement with AT&T WorldNet Service : http://www.att.com/press/1196/961115.bsa.html
- Pennsylvania PUC votes on AT&T/GTE arbitration issues : http://www.att.com/press/1196/961114.chc.html
- AT&T names Braden Allenby Environment, Health and Safety VP : http://www.att.com/press/1196/961118.cha.html

Deleted URL links:

- European and local phone monopolists delay competition, says AT&T chairman : http://www.att.com/press/1196/961112.cha.html
- AT&T reaction to Supreme Court ruling on FCC rules : http://www.att.com/press/1196/961112.chb.html
- AT&T'S redesigned toll–free directory makes Internet shopping easier : http://www.att.com/press/1196/961112.bsa.html

Figure 5: Website News for Netscape: 12/4/96 to 12/9/96 (left); Deep and Recursive Analysis on *http://www.att.com/news*: 11/18/96 to 11/19/96 (right)

While Website News is an attractive service, we need to consider the copyright issue: is it OK to report changes, even just the structural aspect, on a website since it might be considered a form of derivative work? We decided that, for the set of websites we selected, it benefits both their visitors (finding new information faster) and the website creators (getting new information out more effectively). However, we will withdraw the news report of any website upon the request of its creator. This is similar to the policy adopted by many free information services (such as AltaVista[11]).

5. Related Work

Recently, there have been growing interests in visualizing the complex structures of major websites -- mainly to help web users locate information faster without getting lost. Examples include WebMap[1], which captures a user's dynamic interactions with the web pages and visualizes the navigation history, and NetCarta's Web Mapper[2], which performs a static analysis of the structure of any selected website. Other examples include Web Analyzer[16], which presents a *wavefront* view, and Hy+[17], which is based on the visual query language GraphLog and, like WebMap, uses dynamic trace information obtained during a Mosaic session. WebCiao is similar to NetCarta as it also maps a website, but it allows users to make customized database queries and visualize changes in website structures.

Tracking website changes is critical for both website maintainers and clients. A website maintainer needs to make sure that there are no missing links or orphan pages after changes are made to a website. A frequent visitor to a website may prefer to be notified when changes occur on that website. Most website change-tracking systems or notifiers such as Smart Bookmarks[3] or AIDE[4] focus on textual changes, while WebCiao focuses on structure changes on a website. WebGuide[5] combines AIDE and the visualization component of CIAO-html, the predecessor of WebCiao, to allow the examination of both textual and structure changes in web repositories. However, the current framework of WebGuide does not allow global database queries and analysis operators to be performed on a set of web pages.

Website News was inspired by both WebGuide and Brewster Kahle's Internet Archive[8], which has the vision of building a complete running snapshot of the public world-wide-web so that the history of anyone's favorite sites can be preserved. If the complete Internet Archive becomes a reality, our vision is that one day a user can use Website News and WebCiao not only to analyze the history of any changed websites, but to construct a search engine like AltaVista[11] on WWW *deltas*.

6. Summary and Future Work

We have found WebCiao to be quite flexible in querying and visualizing the structures of complex websites. The difference database created by WebCiao allows us to monitor the dynamics of many major websites closely and effectively. The change information is useful in tracking evolving products and services on the web, browsing the web with limited bandwidth, and maintaining large websites. The on-line demo, *Website News*, has been serving many customers world-wide on a daily basis to deliver updates on the structure changes of several major websites. We believe that WebCiao could become extremely useful in identifying new web contents if it is applied to generate web deltas for an internet archive of public websites.

References

[1] Peter Dömel. **Webmap - a graphical hypertext navigation tool**. In *Proceedings of the Second International WWW Conference*, 1994.

[2] NetCarta. http://www.netcarta.com/.

[3] First Floor Software. http://www.firstfloor.com/.

[4] Thomas Ball and Fred Douglis. **An internet difference engine and its applications**. In *Proceedings of 1996 COMPCON*, February 1996, pp. 71-76.

[5] Fred Douglis, Tom Ball, Yih-Farn Chen, and Eleftherios Koutsofios. **WebGUIDE: Querying and Navigating Changes in Web Repositories.** In *Proceedings of the Fifth International WWW Conference*, Paris, France, April 1996. Also appears in Computer Networks and ISDN Systems, 28(1996), pages 1335-1344.

[6] Yih-Farn Chen, Glenn S. Fowler, Eleftherios Koutsofios, and Ryan S. Wallach. **Ciao: A Graphical Navigator for Software and Document Repositories.** In *International Conference on Software Maintenance*, pages 66-75, 1995. See also the home page at http://www.research.att.com/~chen/ciao

[7] Glenn S. Fowler, **cql -- A Flat File Database Query Language.** In *USENIX Winter 1994 Conference,* San Francisco, January 1994.

[8] Internet Archive. http://www.archive.org/webarchive96.html.

[9] Website News. http://www.research.att.com/~chen/web-demo

[10] Website News: Telecommunications Group.
http://www.research.att.com/~chen/web-demo/telecom.html

[11] AltaVista. http://altavista.digital.com.

[12] GlimpseHTTP http://glimpse.cs.arizona.edu/ghttp.

[13] Udi Manber, Michael Smith, Burra Gopal, **WebGlimpse** -- Combining Browsing and Searching, Proceeding of the Usenix Conference, January, 1997. Also visit the web page at http://glimpse.cs.arizona.edu/webglimpse.

[14] Yih-Farn Chen, **Reverse Engineering.** Chapter 6, in " *Practical Reusable UNIX Software*," edited by Balachander Krishnamurthy, pages 177-208, John Wiley & Sons, New York, 1995. .

[15] WebCopy. http://www.inf.utfsm.cl/~vparada/webcopy.html.

[16] Web Analyzer. http://www.incontext.com/products/analyze.html.

[17] Masum Hasan, Dimitra Vista, Alberto Mendelzon, **Visual Web Surfing with Hy+**, Proceedings of CASCON'95, Toronto, Canada, 1995. Also visit the web page at http://www.db.toronto.edu:8020/webvis.html.

[18] Fred Douglis, Anja Feldmann, Balachander Krishnamurthy, and Jeffrey Mogul, **Rate of Change and other Metrics: a Live Study of the World Wide Web**, submitted for publication, December 1996.