The following paper was originally published in the
Proceedings of the USENIX 1996 Annual Technical Conference
San Diego, California, January 1996

Tracking and Viewing Changes on the Web

Fred Douglis and Thomas Ball
AT&T Bell Laboratories

For more information about USENIX Association contact:

1. Phone: 510 528-8649
2. FAX: 510 548-5738
3. Email: office@usenix.org
4. WWW URL: http://www.usenix.org

Tracking

1

Abstract

We describe a set of tools that dete
Wide-Web pages have been modifie
modifications visually to the user
up HTML. The tools consist of th

where on the page, and those changes

the last visit. However, content can

cult to determine visually what chang

tent is added to the end of a page a

sorts documents by modification dat

dynamically. There is a RecentCh

lows multiple users to edit the conte

ronment on the W_3 is the WikiWikiW

environment. One example of a co

interest, it can be of particular use i

While AIDE can help arbitrary use

differences with special HTML mark

HTML pages and creates a “merged”

it has changed. *HtmlDiff* automatica

page and later use a third tool, *Htm*

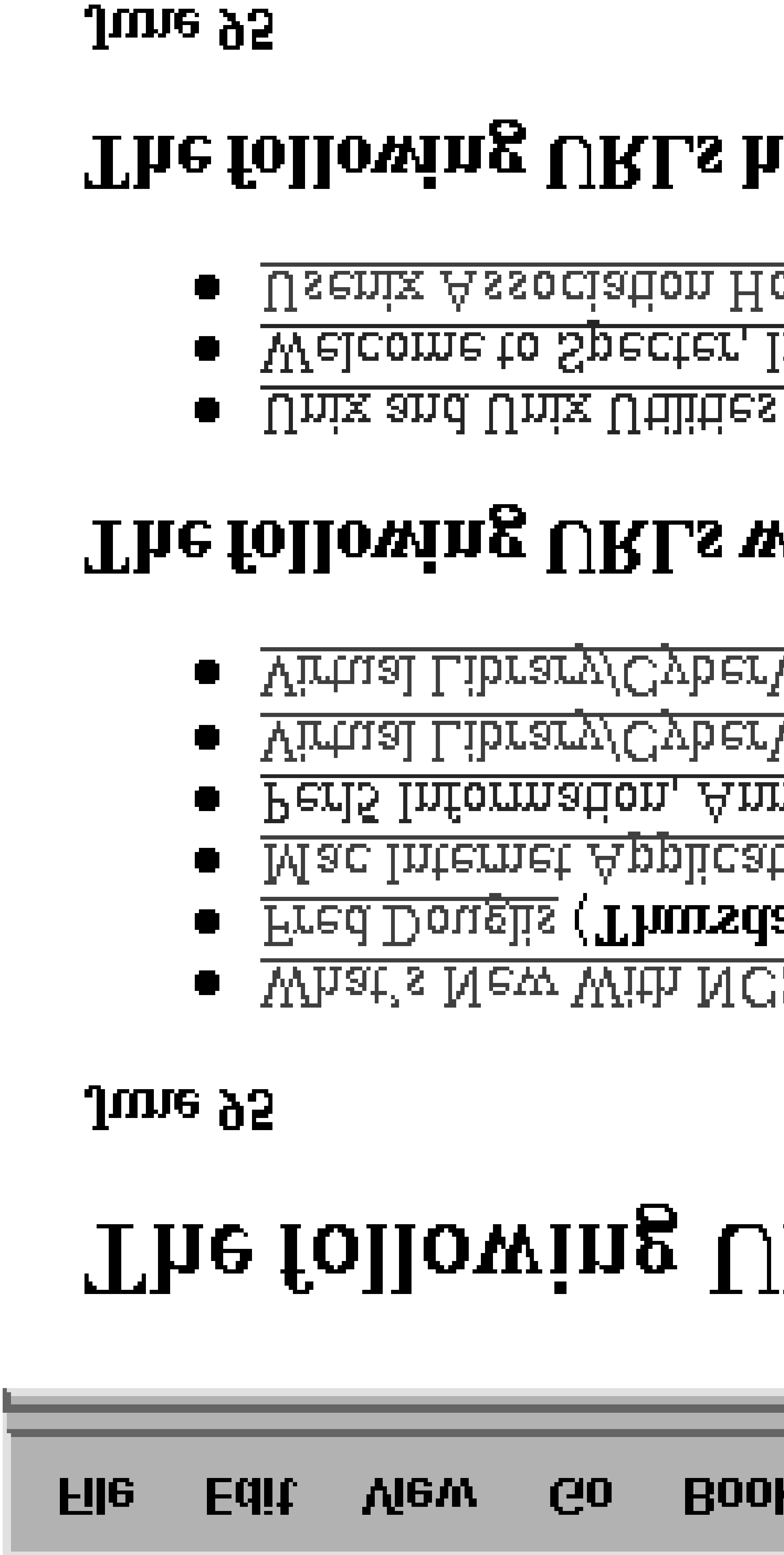
second, *snapshot* allows a user to sa

find when pages on a user's hotlist ha

check-ins of that page by different u
complished outside of RCS by main
control file, allowing quick access t
history.

2.3 HTML Differencing

We know of no existing tools that
pages and present the comparison as
there is much closely related work on
allel document alignment and simila
tween documents [2] that we benefit
comparison utilities such as UNIX
are ill-suited to the comparison of
ments such as HTML. Most modern
programs have a “revision” mode th
and deletions on-line as an author



The following U

June 95

- What's New With NC
- Fred Douglis (**Thursda**
- Mac Internet Applicat
- Perl5 Information, Ann
- Virtual Library/Cyber/
- Virtual Library/Cyber/

The following URLs w

- Unix and Unix Utilities
- Welcome to Specter, I
- Usenix Association Ho

The following URLs h

June 95

Error Conditions

When a periodic task checks the statu
ber of URLs, a number of things can
problems such as network connectivi
a proxy-caching server can cause *all*
fail. Proxy-caching servers are some
to the point of timing out large num
and a background task that retrieve
a short time can aggravate their con
should therefore be able to detect cas
abort and try again later (preferably i
to see an updated report).

At the same time, a number of erro
individual URLs. A URL can mov
out leaving a forwarding pointer.
URL can be deactivated or rename

its previously checked-in versio
for pages that change in many res
storage overhead is minimal be
save a copy of the page in the fi

Drawbacks to the "external servi
that the service must remember the st
that anyone who uses the service has i
est in and must know which user has
sion of each page. The first issue is
resource allocation, and is not expect
icant issue unless the service is used
clients on a number of large pages.
is addressed in our initial prototype
support for datestamps and requestin
isted at a particular time. In the ne
system, a set of version numbers is

cility, however, will still have informa
user accesses which pages, unless the
can be done anonymously.

5 Comparison of HTM

In our experience, only a small fractio
W^3 contain information that allows
how the pages have changed—examp
that highlight recent additions, a link
or a special "What's New" page. A
in the introduction, these approaches
ciencies. They are intended to be vie
but users will visit the pages at diffe
have different ideas of "what's new"
maintainer must explicitly generate

(where words matching exactly agai
signed weight 1, and markups matc
markups, as before). Let W be the
and content-defining markups in the
sentences and let L be the sum of the
sentences. If the percentage $(2 * W)$
large, then the sentences match with
wise, they do not match.

5.2 Presentation of the diffe

The comparison algorithm outline
mapping from the tokens of the old do
kens of the new document. Tokens tha
are termed “common”; tokens that ar
document but have no counterpart in
“old” (“new”). We refer to the “old”

HtmlDiff: Here is the f

USENIX

The Advanced C

SAGE

The System Adm

- ▶ ***Safe on USE***
- Letter to SAGE

saved if it is unchanged from th
was stored away.

Diff Have the *snapshot* facility inv
display the changes in a page
saved away by the user.

History Have *snapshot* display a fu
of this page, with the ability to
any pair of versions or to view a
directly.

Thus, each page that is reported as
diately be passed to *HtmlDiff*, and ar
can be "remembered" for future use.
interface with *snapshot* directly, via a
pages, view differences, or view the

One disadvantage of the current

Automatic archival of new version
context, but it has the disadvantage o
space dramatically when the entire co
changes (such as the "What's New i
When the entire contents are replace
for *HtmlDiff*. Automatic archival wo
in cases when one wants a way to go
old versions, but in many cases (incl
ple), the content provider has its own

8.3 Server-side URL Tracki

Currently, *w3newer* runs on the us
multiple instantiations of the script
same work. Although it runs a relate
same machine as an AT&T-wide prox

While such automation is clearly be
eral context, we expect that AIDE wil
of more focused uses of the W^3, espe
volving collaborative and distributed

Several issues still need to be ad
ticular, many of the complications o
avoided by better integration with W
servers. The increasing availability o
erarchical HTTP repositories such as
also be both an opportunity and a chal
notification mechanisms and version

For more information on AI
http://www.research.att.com/orgs/s
lis/aide.

Acknowledgments

Thomas Ball is a member of techni
Bell Laboratories. His research inte
gramming languages, software tools,
ficiently monitoring system and prog
software and system performance vis
ceived a B.A. from Cornell Universi
M.S. (1989) and Ph.D. (1993) degree
sity of Wisconsin-Madison, all in C
Email: *tball@research.att.com*.