Computer Networks and ISDN Systems 28 (1996) 963–980

COMPUTER NETWORKS and ISDN SYSTEMS

An investigation of documents from the World Wide Web

Allison Woodruff *, Paul M. Aoki [1], Eric Brewer [2], Paul Gauthier [3], Lawrence A. Rowe [4]

Computer Science Division [5], *University of California* [6], *Berkeley* [7], *CA 94720-1776, USA*

Abstract

We report on our examination of pages from the World Wide Web. We have analyzed data collected by the **Inktomi** [8] Web crawler (this data currently comprises over 2.6 million HTML documents). We have examined many characteristics of these documents, including: document size; number and types of tags, attributes, file extensions, protocols, and ports; the number of in-links; and the ratio of document size to the number of tags and attributes. For a more limited set of documents, we have examined the following: the number and types of syntax errors and readability scores. These data have been aggregated to create a number of ranked lists, e.g., the ten most-used tags, the ten most common HTML errors.

Keywords: HTML; Statistics; Tools; World Wide Web

1. Introduction

We report the results of an extensive analysis of HTML documents from the World Wide Web. Our data set, collected by the **Inktomi** [9] Web crawler, currently comprises **over 2.6 million** [10] HTML documents. We present a broad range of statistics pertaining to these pages.

Such an analysis of the content of HTML documents is of interest for several reasons:

- **Evolution of HTML.** Unused features and extensions that do not achieve a reasonable level of acceptance should be deprecated and, eventually, eliminated. This prevents the accretion of useless language features.
- **Improving Web content.** Widespread awareness of poor natural and markup language usage will promote the spread of helpful tools and practices.
- **Control of HTML.** The marketplace perceives the relative ability of vendors to force acceptance of new, non-standard language extensions as market "strength." Understanding the true acceptance level of such extensions can help fight vendor disinformation.

* Corresponding author. Email: woodruff@cs.berkeley.edu, http://HTTP.CS.Berkeley.EDU/ woodruff/

[1] Email: aoki@cs.berkeley.edu, http://HTTP.CS.Berkeley.EDU/ aoki/

[2] Email: brewer@cs.berkeley.edu, http://HTTP.CS.Berkeley.EDU/ brewer/

[3] Email: gauthier@cs.berkeley.edu, http://HTTP.CS.Berkeley.EDU/ gauthier/

[4] Email: rowe@cs.berkeley.edu, http://HTTP.CS.Berkeley.EDU/ larry/

[5] http://www.CS.Berkeley.EDU/

[6] http://www.Berkeley.EDU/

[7] http://www.ci.berkeley.ca.us/

[8] http://inktomi.berkeley.edu/

[9] http://inktomi.berkeley.edu/

[10] http://inktomi.berkeley.edu/counting.html

PII S0169-7552(96)00064-5

- **Sociological insights**. Many interesting sociological observations may be derived from the content of Web pages.

Despite these motivations, however, previous studies relating to the Web have either focused on other topics or have been limited in scope. The most closely related work includes:

- **User studies**. User surveys [5,16–19,21] and browser usage studies [2,15] have become very common. Such studies focus on high-level user issues (e.g., choice of software, available connectivity) and low-level user-browser interaction (e.g., use of the `back` button). The information extracted, though valuable, is wholly user-centric.
- **Content analyses of small data sets**. There have been some attempts to perform simple analyses of the content of the Web. For example, the original Lycos project at Carnegie Mellon University's Center for Machine Translation [12] tracked a number of interesting statistics while their data set was relatively small. These included:
  - content of title and headings
  - 100 top keywords and first 20 lines
  - word frequency count
  - file size (bytes, words)
  - URL types
  - most-linked-to URLs
- **Structural analysis**. The CMU Lycos project generated at least one **complete graph** [11] of their data set. The project's commercial successor, Lycos, Inc., now tracks the 250 most-linked-to sites as a side-effect of their indexing [11]. Other projects have focused on (graph-oriented) structural analysis as well. These include several Web visualization systems (e.g., Webspace [4] and the Navigational View Builder [13]). For the most part, such visualization has been very small-scale and limited in scope. More sophisticated analyses are possible, combining both structural analysis and semantic modelling. A project at Xerox PARC [14] is conducting such analyses over small data sets.

To complement the above work, we have conducted a large-scale investigation of the content of HTML documents from the Web. The remainder of this paper is structured as follows. First, we describe the tools we used to perform our study. We next discuss the scope of our study and our results. Finally, we present some lessons learned and possible future directions.

2. Tools

The tools used to perform the data collection and data analysis for this study represent the integration of software from a variety of sources. Specifically, we have developed or adapted software to perform the following tasks:

- Web data collection (see Section 2.1)
- Data extraction and manipulation (see Section 2.2)
- Natural (English) language analysis (see Section 2.3)
- Markup (HTML) language analysis (see Section 2.4)

We discuss each set of tools in turn.

2.1. Web data collection

The **Inktomi** [12] research project at Berkeley, consisting of Prof. Eric Brewer and graduate student Paul Gauthier, conducts research in the construction of **scalable Web servers** [13] using parallel processing technology. To date, the project has produced two major software components: a parallel **Web crawler** [14] and a parallel Web index search engine. In this paper, where we mention Inktomi, it may be assumed that we refer to the crawler.

The data presented in this study comes entirely from Inktomi. The high speed of the crawler enables us, for the first time, to consider taking "snapshots" of the Web and analyzing them. As of this writing, the Inktomi team has crawled twice. The first set of runs, from July to October 1995, collected 1.3 million unique HTML documents. The second set of

[11] ftp://nl.cs.cmu.edu/usr/mlm/ftp/url-graph.Z

[12] http://inktomi.berkeley.edu/

[13] http://inktomi.berkeley.edu/scalable.html

[14] http://info.webcrawler.com/mak/projects/robots/robots.html

runs, in November 1995, collected 2.6 million unique HTML documents.

2.2. HTML data extraction and manipulation: libink

Although toolkits such as the W3C Reference Library [8] already exist for manipulating HTML and HTTP objects, we have developed our own special-purpose library, libink. This was necessitated by the fact that our performance and functionality needs were very different from those of the other toolkit developers.

libink consists of four major subcomponents:

- **HTML parser.** libink contains a simple flex-based HTML scanner. We found existing parsers too slow (especially true in the case of parsers written in scripting languages) or difficult to modify. The libink scanner is small, enabling us to make it both fast and relatively robust, as well as highly configurable. Like the W3C SGML/HTML lexical analyzer [6], our scanner uses a callback interface to handle various events (e.g., recognition of a tag and its attributes). The W3C lexical analyzer, however, is not configurable.
- **URL parser.** The URL parser, unlike many freely-available implementations, conforms to RFC 1808 [7].
- **Domain name service (DNS) translation and caching.** We use Internet addresses to reduce hostname aliasing in our data. To speed up the lookup process, we provide a wrapper around the standard name service routines that caches *all* URL hostnames.
- **General hash table services.** The various lookup tables on which libink relies sometimes exceed the capacity of a single machine's physical memory. Therefore, in addition to in-memory hash tables, libink provides interfaces to striped on-disk hash tables (using GNU DBM) as well as hash-partitioned distributed hash tables (using ONC RPC). The distributed hash tables support 1ms turnaround on hash table lookups, which is far better than the 20–30 ms required to fetch a hash table page from secondary storage.

2.3. Natural language analysis: style

We scored English language documents using the standard UNIX style program [3]. style reports a variety of statistical properties of each document, such as the average sentence length and the number of complex sentences. It also scores the document using four readability metrics. These metrics indicate the nominal educational (grade) level a reader would need to understand the document.

Since most HTML documents do not conform to an internationalization standard, we applied heuristics to screen out non-English documents. We filtered out documents that contained any character with the high bit set (indicating a non-ASCII character set) or containing character sequences indicating known encodings (such as the Shift-JIS encoding of the Japanese character set).

2.4. Markup language analysis: weblint

We scored documents using weblint [1], an analogue to the standard UNIX lint utility, written in Perl. We modified weblint to report the classes of errors in a document rather than a line-by-line analysis.

3. Results

We examined over 2.6 million HTML documents collected by the Inktomi crawler in November of 1995. Although Inktomi occasionally downloads non-HTML documents, the results presented reflect only HTML documents. (For example, we filtered out all binary files, such as images.) Furthermore, because Inktomi implements the **Robot Exclusion Standard** [15], the contents of automated databases which follow the standard (e.g., genome data sets) have also been excluded. The distribution of the documents in the data set by domain appears in Table 1.

Here, "other" includes all domains other than the

[15] http://info.webcrawler.com/mak/projects/robots/norobots.html

Table 1
Documents studied by domain

Domain	# of HTML documents	% of total
other	1064318	41%
com	516709	20%
edu	698616	27%
gov	117125	4%
net	113595	4%
mil	14734	1%
org	89939	3%
Total	2615036	100%

given top-level domains. For example, "other" contains all non-US top-level domains (such as Germany's `.de`).

We analyzed a variety of properties of these documents.In this paper, we present results on the following:

- Document size (see Section 3.1)
- Tag/size ratio (see Section 3.2)
- Tag usage (see Section 3.3)
- Attribute usage (see Section 3.4)
- Browser-specific extension usage (see Section 3.5)
- Port usage (see Section 3.6)
- Protocols used in child URLs (see Section 3.7)
- File types used in child URLs (see Section 3.8)
- Number of in-links (see Section 3.9)
- Readability (see Section 3.10)
- Syntax errors (see Section 3.11)

3.1. Document size

After all markup had been extracted, the size of each HTML document was measured. For the entire data set, the mean size was 4.4KB, the median size was 2.0KB, and the maximum size was 1.6MB.

Fig. 1 presents different views of the size distribution. On first inspection, this distribution appears to be exponential (the magenta line represents the location of the mean). However, further zooming indicates a curve before the distribution begins to taper off. The final graph in Fig. 1 contains a semilog plot of the same data (in which the sizes are plotted logarithmically and the number of documents is plotted arithmetically).

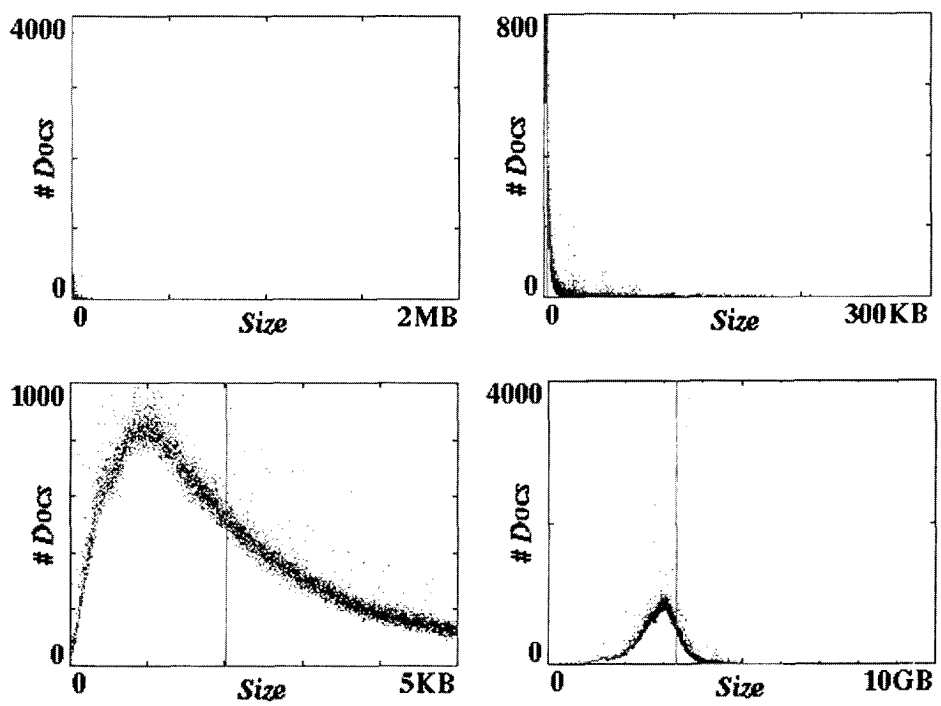


Fig. 1. Size distribution.

These simple size distribution plots proved to be very useful in detecting several problems with the data set. Many of the outliers were caused by one of two major classes of errors:

- **Problematic URLs**: when faced with incorrect URLs that contain valid prefixes, some HTTP servers return the file matching the valid prefix. For example, the data set contains hundreds of documents with URLs of the form http://bazaar.com/underground2.html/..., all of which are identical to http://bazaar.com/underground2.html. There does not appear to be a general way for a client program (such as a crawler) to differentiate this situation from a site containing a large number of identical files.
- **CGI Error Responses**: some of the most popular CGI programs, such as NCSA `imagemap` and CERN `HTImage`, report errors with messages containing HTTP status "200" (success). Because the image map programs all happen to return fixed error messages, we were able to detect and eliminate those particular messages, but there (again) does not appear to be any general way for a client to distinguish "200" error messages from valid documents.

3.2. Tag / size ratio

For each document we examined the ratio of the total number of tags to its size. Fig. 2 contains the results. An interesting pattern emerges—rays radiating out from the origin, indicating a number of documents with constant tag/size ratios. One such ray is indicated by the green ellipse. We examined a number of these rays and determined that they represented different versions of the same document (occurring in archives or mirrored sites). This suggests that the tag/size ratio might be used as a component of a signature for an HTML document, e.g., for purposes of copy detection.

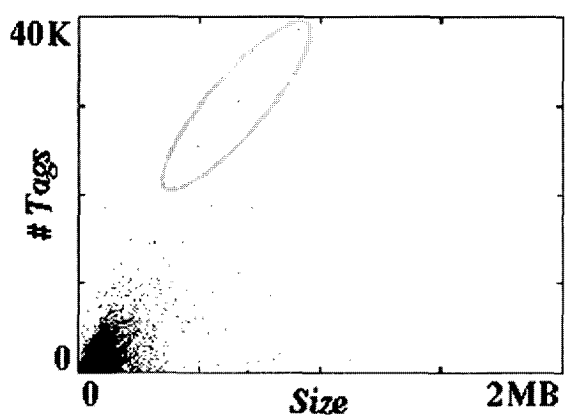


Fig. 2. Tag/size ratio.

3.3. Tag usage

We examined the distribution of tags (see Table 2). We obtained a list of valid tags from the Sandia HTML Reference Manual [9]. The average number of total tags per document was 71. The average number of unique tags per document was 11.

We examined the most popular tags. The top graph of Fig. 3 shows the top ten tags (ranked according to the number of documents in which the tag appeared at least once). The bottom graph indicates the average number of occurrences of the tag per document.

We also examined the least popular tags. Several tags, **BDO** [16], **COLGROUP** [17], and **NOEMBED** [18] were used zero times in our data set of over 2.6 million HTML documents. A number of other tags appeared a very limited number of times.

3.4. Attribute usage

We examined the distribution of attributes (see Table 3). The average number of total attributes per document was 29. The average number of unique attributes per document was 4.

We examined the most popular attributes. Fig. 4 shows the top ten attributes (ranked according to the number of documents in which the attribute appeared at least once). `HREF` appeared an average of 14 times per document.

We also examined the least popular attributes.

[16] http://www.sandia.gov/sci_compute/elements.html#BDO

[17] http://www.sandia.gov/sci_compute/elements.html#COLGROUP

[18] http://www.sandia.gov/sci_compute/elements.html#NOEMBED

Table 2
Tag usage [a]

Tag	% of docs	Avg. per doc	# occurrences	# docs
TITLE	92%	0.9726424	2543495	2417887
A	88%	14.951125	39097731	2314111
P	77%	9.4494152	24710561	2002651
HR	72%	2.3940473	6260520	1889400
BODY	69%	0.8122745	2124127	1812134
IMG	64%	3.9899294	10433809	1663111
HEAD	61%	0.615572	1609743	1595265
HTML	59%	0.6049091	1581859	1548459
H1	59%	0.7017005	1834972	1537981
BR	54%	7.222327	18886645	1422696
H2	42%	1.0677226	2792133	1098753
B	40%	3.991433	10437741	1053086
LI	39%	7.5248077	19677643	1032132
UL	35%	1.2868622	3365191	908006
I	29%	1.9286232	5043419	768457
H3	27%	0.9172428	2398623	709168
CENTER	27%	0.6873944	1797561	706649
ADDRESS	24%	0.3019805	789690	632869
PRE	20%	0.5063383	1324093	527146
DL	15%	0.4328973	1132042	384868
DD	15%	2.1735162	5683823	382317
FONT	13%	0.7305456	1910403	347282
DT	13%	1.8391598	4809469	345886
H4	11%	0.3590604	938956	296218
STRONG	11%	0.7684024	2009400	290080
EM	10%	0.5073047	1326620	258627
TABLE	6%	0.133267	348498	159006
LINK	6%	0.0707355	184976	154507
TD	6%	1.5339754	4011401	152557
TR	6%	0.5344542	1397617	150226
FORM	5%	0.069269	181141	138973
INPUT	5%	0.4806029	1256794	138210
H5	5%	0.105423	275685	133887
other	5%	0.1310414	342678	125167
BLOCK-QUOTE	4%	0.1384639	362088	106026
MENU	4%	0.0732177	191467	105025
OL	4%	0.0844646	220878	104557
META	4%	0.1074532	280994	92801
H6	3%	0.0530042	138608	90048
TT	2%	0.2167913	566917	63736
BASE	2%	0.0217485	56873	56556
CITE	2%	0.1036387	271019	56489
CODE	2%	0.2712456	709317	52470
BLINK	2%	0.0315028	82381	50694
TEXTAREA	2%	0.0226073	59119	44867
U	2%	0.085922	224689	43413
TH	2%	0.1778886	465185	39611
SELECT	1%	0.0315556	82519	36680
OPTION	1%	0.3199746	836745	36623
BASEFONT	0%	0.0059135	15464	13052
ISINDEX	0%	0.0043556	11390	11364
NOBR	0%	0.0132128	34552	9990
KBD	0%	0.0255503	66815	8602

Table 2 (continued)

Tag	% of docs	Avg. per doc	# occurrences	# docs
CAPTION	0%	0.0055173	14428	7981
SAMP	0%	0.0288099	75339	7842
DIR	0%	0.0136426	35676	7291
VAR	0%	0.0313931	82094	6530
SUP	0%	0.0135914	35542	5256
LISTING	0%	0.0055384	14483	4694
XMP	0%	0.0093268	24390	3950
DFN	0%	0.0054026	14128	2388
LH	0%	0.0028382	7422	2388
NEXTID	0%	0.0009235	2415	2236
SMALL	0%	0.0020378	5329	2168
SUB	0%	0.0069999	18305	1856
FRAME	0%	0.0023155	6055	1545
APP	0%	0.001039	2717	1411
FRAMESET	0%	0.0008658	2264	1364
DIV	0%	0.0009575	2504	1307
AREA	0%	0.0039303	10278	1061
PLAINTEXT	0%	0.0004922	1287	1060
MAP	0%	0.0004558	1192	1034
NOFRAMES	0%	0.0003644	953	944
WBR	0%	0.0012898	3373	936
BIG	0%	0.0009235	2415	763
BANNER	0%	0.0002979	779	760
TAB	0%	0.0023185	6063	589
TBODY	0%	0.0003396	888	553
BGSOUND	0%	0.0002034	532	509
NOTE	0%	0.0002608	682	465
S	0%	0.0008952	2341	455
MARQUEE	0%	0.0002191	573	449
APPLET	0%	0.0001954	511	400
AU	0%	0.0004482	1172	390
PERSON	0%	0.0007832	2048	360
PARAM	0%	0.0005266	1377	318
STRIKE	0%	0.0008103	2119	290
Q	0%	0.0005247	1372	245
FIG	0%	0.0002096	548	237
ACRONYM	0%	0.0002929	766	140
CREDIT	0%	4.551E-05	119	96
THEAD	0%	6.769E-05	177	91
COL	0%	0.0001231	322	73
BQ	0%	3.939E-05	103	63
HP	0%	3.174E-05	83	61
FN	0%	3.939E-05	103	60
DEL	0%	1.95E-05	51	44
EMBED	0%	2.141E-05	56	43
ABBREV	0%	3.633E-05	95	36
INS	0%	1.3E-05	34	15
LANG	0%	3.059E-06	8	7
TFOOT	0%	4.589E-06	12	4
OVERLAY	0%	7.648E-07	2	2
SPAN	0%	3.824E-07	1	1
BDO	0%	0	0	0
COLGROUP	0%	0	0	0
NOEMBED	0%	0	0	0

[a] http://www.sandia.gov/sci_compute/elements.html

Several attributes, **ACCEPT-CHARSET** [19], **AXIS** [20], **CHAROFF** [21], and **CONTROLS** [22], were used zero times in our data set of 2.6 million HTML documents. A number of other attributes appeared a very limited number of times.

3.5. Browser-specific extension usage

We also studied the use of browser-specific extensions. These consist of HTML features (i.e., tags or attributes) added by vendors rather than by the standards process. Here, we contrast the use of such extensions in the first Inktomi data set (1.3 million documents, collected in mid-1995) and the second Inktomi data set (2.6 million documents, collected in November 1995).

Fig. 5 shows the percentage of documents in which the four most popular extensions are used. The usage of most of these features has risen dramatically, indicating wide user acceptance. Other features, such as **BLINK** [23], have not experienced such growth.

Fig. 6 indicates the popularity of various proposals for dynamic addition of functionality to browsers. **APP** [24] and **APPLET** [25] support SunSoft's Java "applet" language, **DYNSRC** [26] supports VRML markup, and **EMBED** [27] supports Netscape's third-party "plug-in" modules. All have enjoyed significant growth, though the oldest and most popular method

[19] http://www.sandia.gov/sci_compute/elements.html#FORM
[20] http://www.sandia.gov/sci_compute/elements.html#TH
[21] http://www.sandia.gov/sci_compute/elements.html#TH
[22] http://www.sandia.gov/sci_compute/elements.html#IMG
[23] http://www.sandia.gov/sci_compute/elements.html#BLINK
[24] http://www.sandia.gov/sci_compute/elements.html#APP
[25] http://www.sandia.gov/sci_compute/elements.html#APPLET
[26] http://www.sandia.gov/sci_compute/elements.html#IMG
[27] http://www.sandia.gov/sci_compute/elements.html#EMBED

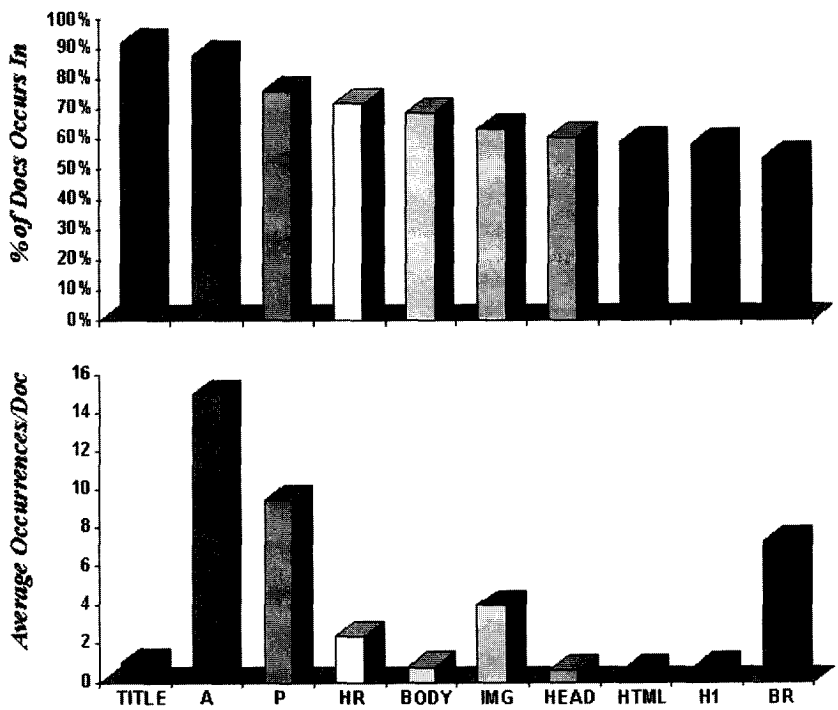


Fig. 3. Ten most-used tags.

Table 3
Attribute usage [a]

Attribute	% of docs	Avg. per doc	# occurrences	# docs
HREF	88%	13.89630315	36339333	2312905
SRC	64%	3.995921662	10449479	1663955
ALIGN	32%	1.726451567	4514733	839392
ALT	29%	1.742787862	4557453	770333
NAME	22%	2.367810998	6191911	575374
SIZE	20%	1.148610574	3003658	527465
BORDER	16%	0.661218431	1729110	405878
WIDTH	14%	0.742229552	1940957	358322
BACKGROUND	12%	0.122979569	321596	318096
BGCOLOR	9%	0.161771004	423037	236135
HEIGHT	9%	0.452991087	1184588	230890
TEXT	7%	0.07621463	199304	189552
LINK	7%	0.07161431	187274	184225
VLINK	7%	0.070186032	183539	181668
VALUE	6%	0.410621881	1073791	152915
TYPE	6%	0.399540962	1044814	149019
REV	6%	0.061423629	160625	147019
ISMAP	6%	0.066939805	175050	146603
ACTION	5%	0.067888167	177530	137476
HSPACE	5%	0.196408004	513614	132911
CLEAR	4%	0.095963497	250948	104855
METHOD	4%	0.048018077	125569	99147
ALINK	4%	0.037972709	99300	98728
other	4%	0.147613647	386015	92501
CONTENT	3%	0.07844978	205149	71979
CELLPADDING	2%	0.045978335	120235	63245
VSPACE	2%	0.061642746	161198	53753
CELLSPACING	2%	0.035331062	92392	49046
COLS	2%	0.023403884	61202	45997
ROWS	2%	0.023466981	61367	45892
VALIGN	2%	0.107953772	282303	39786
NOSHADE	1%	0.045230352	118279	36800
COLSPAN	1%	0.111472653	291505	33263
SELECTED	1%	0.017915241	46849	22799
REL	1%	0.013695031	35813	19736
COMPACT	1%	0.026405755	69052	17140
MAXLENGTH	1%	0.032788459	85743	16552
CHECKED	1%	0.014750466	38573	16211
ROWSPAN	1%	0.021920157	57322	14378
HTTP-EQUIV	0%	0.007987653	20888	11555
VERSION	0%	0.003984649	10420	10391
LOWSRC	0%	0.003255022	8512	4842
NOWRAP	0%	0.019980222	52249	4540
TITLE	0%	0.006284426	16434	3803
TARGET	0%	0.005155187	13481	3142
COLSPEC	0%	0.002059627	5386	2948
COLOR	0%	0.003419838	8943	2321
N	0%	0.001250843	3271	2230
CLASS	0%	0.002427118	6347	2188
MULTIPLE	0%	0.001422925	3721	1745
TO	0%	0.001285642	3362	1645
USEMAP	0%	0.000730774	1911	1528
SHAPE	0%	0.003825569	10004	1104
COORDS	0%	0.003836276	10032	1058
FACE	0%	0.000758307	1983	1050

Table 3 (continued)

Attribute	% of docs	Avg. per doc	# occurrences	# docs
SCROLLING	0%	0.000630202	1648	979
UNITS	0%	0.000814903	2131	941
ID	0%	0.001266139	3311	934
PROMPT	0%	0.000321219	840	837
METHODS	0%	0.000364813	954	830
URN	0%	0.000293304	767	694
CODE	0%	0.000762131	1993	678
LANG	0%	0.000404966	1059	514
BGPROPERTIES	0%	0.000188908	494	493
NORESIZE	0%	0.000349135	913	458
WRAP	0%	0.000224854	588	425
MAX	0%	0.000951421	2488	420
PLAIN	0%	0.000624466	1633	406
URL	0%	0.000652381	1706	395
START	0%	0.000291392	762	389
LOOP	0%	0.000162139	424	353
MARGINWIDTH	0%	0.000282979	740	351
MARGINHEIGHT	0%	0.00024971	653	327
ENCTYPE	0%	0.000260799	682	293
CODEBASE	0%	0.000101337	265	198
BEHAVIOR	0%	7.954E-05	208	174
MIN	0%	0.000490242	1282	157
SCROLLDELAY	0%	5.54486E-05	145	125
DYNSRC	0%	5.04773E-05	132	117
IMAGEMAP	0%	4.12996E-05	108	106
DIRECTION	0%	5.00949E-05	131	100
NOHREF	0%	6.65383E-05	174	96
NOFLOW	0%	4.01524E-05	105	93
Y	0%	7.07447E-05	185	73
SCROLL-AMOUNT	0%	3.21219E-05	84	70
X	0%	7.91576E-05	207	70
FRAME	0%	4.32116E-05	113	61
INDENT	0%	0.0001304	341	53
DIR	0%	1.83554E-05	48	40
DINGBAT	0%	3.17395E-05	83	34
CHARSET	0%	2.3709E-05	62	33
RULES	0%	1.33841E-05	35	30
SPAN	0%	1.68258E-05	44	30
MD	0%	1.56786E-05	41	28
CONTINUE	0%	1.49138E-05	39	27
SKIP	0%	1.72082E-05	45	26
DISABLED	0%	5.27717E-05	138	22
CHAR	0%	2.06498E-05	54	21
SCRIPT	0%	6.11846E-06	16	15
SEQNUM	0%	1.10897E-05	29	11
DP	0%	8.79529E-06	23	10
ERROR	0%	4.58885E-06	12	10
ACCEPT	0%	2.29442E-06	6	4
LOOPDELAY	0%	1.52962E-06	4	3
AXES	0%	1.14721E-06	3	2
ACCEPT-CHARSET	0%	0	0	0
AXIS	0%	0	0	0
CHAROFF	0%	0	0	0
CONTROLS	0%	0	0	0

[a] http://www.sandia.gov/sci_compute/elements.html

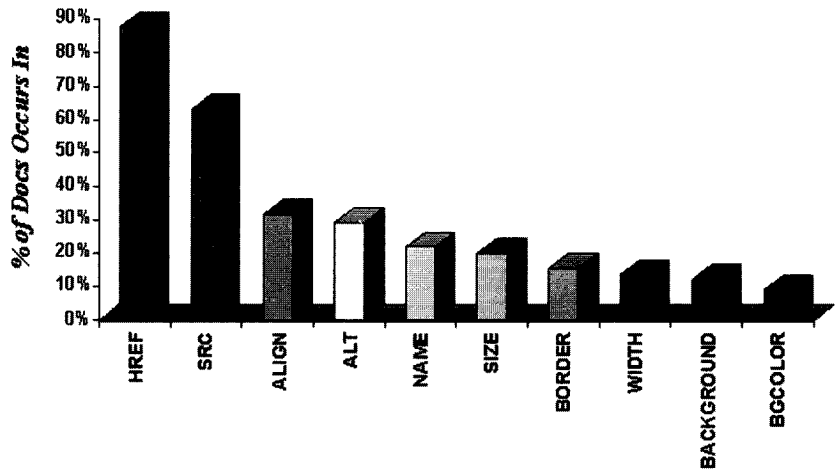


Fig. 4. Ten most-used attributes.

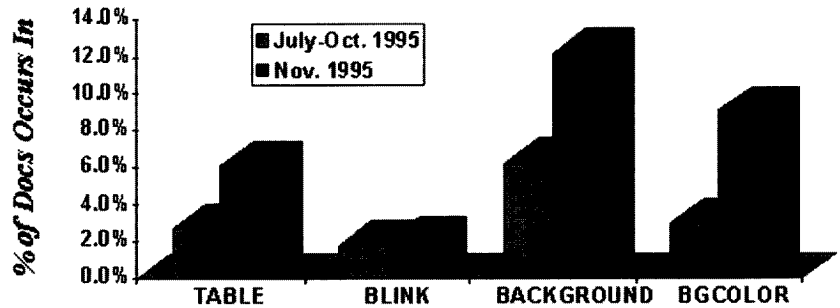


Fig. 5. Browser-specific extensions usage.

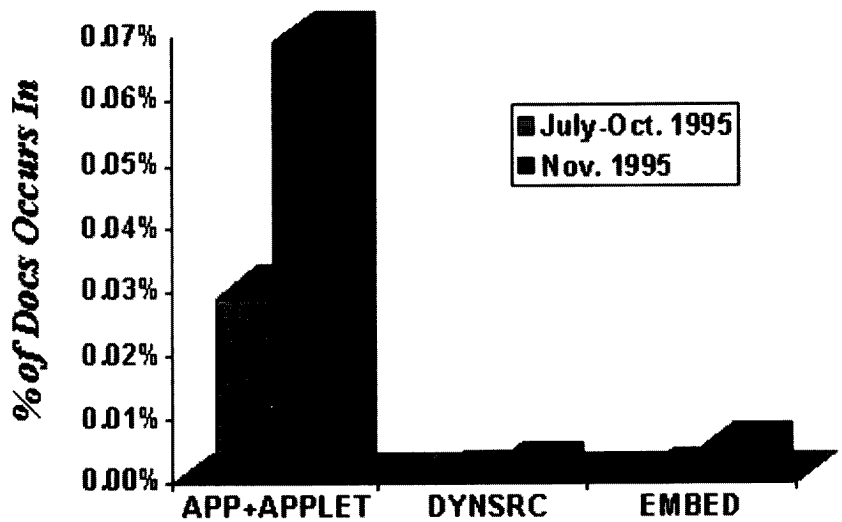


Fig. 6. Popularity of various proposals for dynamic addition of functionality to browsers.

Table 4
Port usage

Category	Port	% of docs
Standard	80	93.6%
< 1024	70	0.3%
≥ 1024	8000	0.5%
	8001	0.5%
	8080	0.7%
	8888	2.8%

(Java, first released in May 1995 [10]) still has very low usage.

3.6. Port usage

For each of the HTML documents in our data set, we extracted the port number used to access the document. We analyzed the distribution of port numbers. While 418 unique ports were observed, six ports accounted for over 98% of the documents. Table 4 presents the most popular ports.

Port 80, the standard HTTP port, was used for approximately 94% of the documents. Port 70 (the standard Gopher port) was used for approximately 0.3% of the documents (this number is slightly lower than the 1% usage of port 70 observed in our earlier data set). We checked many of the documents being served from port 70; all the ones we examined were in fact HTML documents. Ports 8000, 8001, and 8080, and 8888 accounted for the majority of the remaining documents. The strong preference for "8" and "80" in the non-standard ports is presumably related to the standard port number "80"

Table 5
Child URL protocols

Protocol name	% of docs	Avg per doc	# occurrences	# docs
HTTP	91%	16.9071994	44212935	2374512
MAILTO	28%	0.5465669	1429292	722263
FTP	5%	0.3049927	797567	120919
GOPHER	4%	0.1553810	406327	100764
NEWS	1%	0.1211023	316687	36914
TELNET	1%	0.0236899	61950	21879
WAIS	0%	0.0067051	17534	4170
other	0%	0.0167279	43744	4045
HTTPS	0%	0.0010482	2741	1737
TN3270	0%	0.0004742	1240	840
RLOGIN	0%	0.0003430	897	338
FILE	0%	0.0003442	900	105
NNTP	0%	0.0000080	21	6
AFS	0%	0.0000004	1	1
PROSPERO	0%	0.0000000	0	0

3.7. Protocols used in child URLs

As discussed above, we extracted child URLs from all HTML documents in our data set. Fig. 7 presents the distribution of protocols in this set of child URLs. By far, the most dominant protocol observed was HTTP (there were an average of 17 HTTP URLs per document). (See also Table 5.)

3.8. File types used in child URLs

We also studied the distribution of file types described in the set of extracted child URLs. We inferred the file type from the file name extensions (e.g., ".gif") found in the URL path. In Table 6, the

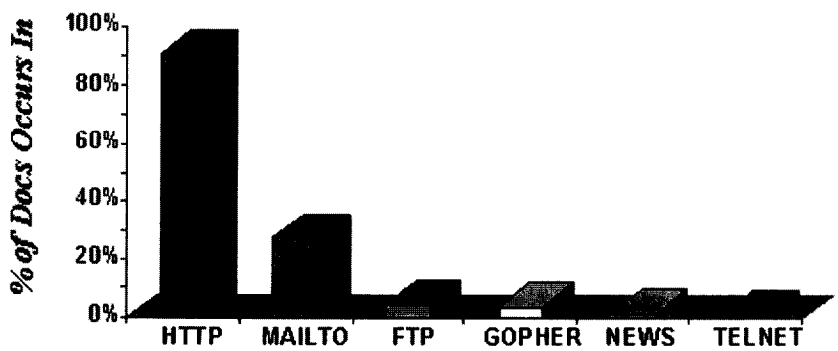


Fig. 7. Protocol usage.

Table 6
File type and file name extensions

Type (extension)	% of docs	# of occurrences	# of docs
Compression / Archive (see also Fig. 8)			
GNU zip (gz/gzip/taz/tgz)	0.7%	126839	18694
Zip (zip)	0.7%	157918	17277
compress (Z)	0.6%	121519	16857
BinHex (hqx)	0.3%	138259	7188
StuffIt (sea)	0.1%	5290	2615
LHArc (lha/lharc)	0.0%	20985	597
ARC archive (arc)	0.0%	432	129
Document (see also Fig. 9)			
HTML (htm/html)	76.3%	21982792	1995731
text (txt)	2.2%	325165	57476
PostScript (eps/ps)	1.8%	239949	46977
MS Word (doc)	0.2%	20153	5959
Adobe Acrobat (pdf)	0.2%	30640	5360
TeX DVI (dvi)	0.2%	14680	4163
Tex (tex)	0.1%	11998	2993
TROFF (man/me/ms)	0.1%	6488	2191
Rich Text (rtf)	0.0%	3921	1184
Maker Interchange (mif)	0.0%	262	113
Audio (see also Fig. 10)			
Sun audio (au)	0.7%	60405	18865
MS WAVE (wav)	0.3%	24361	7325
Audio IFF (aif/aifc/aiff)	0.1%	7761	2611
MIME audio (snd)	0.0%	1839	600
Amiga MOD (mod/nst)	0.0%	4202	254
IRCOM (sf)	0.0%	353	161
IFF (iff)	0.0%	322	47
SoundBlaster (voc)	0.0%	122	27
U-law (ul)	0.0%	21	19
FSSD (fssd/hcom)	0.0%	3	3
Image (see also Fig. 11)			
GIF (gif)	61.7%	9990239	1614244
JPEG (jpe/jpeg/jpg)	7.8%	811353	205088
X bitmap (xbm)	2.9%	968410	75825
TIFF (tif/tiff)	0.2%	22546	5416
X pixmap (xpm)	0.0%	3448	814
RGB (rgb)	0.0%	985	259
portable pixmap (ppm)	0.0%	646	124
portable graymap (pgm)	0.0%	219	78
portable bitmap (pbm)	0.0%	114	70
X window dump (xwd)	0.0%	277	66
raster (ras)	0.0%	221	54
portable anymap (pnm)	0.0%	51	7
Movie (see also Fig. 12)			
MPEG (mpe/mpeg/mpg)	0.3%	21496	7460
QuickTime (mov/qt)	0.2%	15026	5199
MS video (avi)	0.1%	5589	1742
SGI (movie)	0.0%	538	313

"% of docs" column indicates the percentage of documents which contained a file of a given type. The "# of occurrences" column shows the total number of extensions of the given file type that were observed. The "# of docs" column indicates the number of documents which contained one or more extensions of the indicated type. Note that files can be counted multiple times, e.g., `file.ps.Z` would be counted as a file having both ".ps" and ".Z" extensions. (See also Figs. 8–12.)

3.9. Number of in-links

We sorted the child URLs which we extracted according to the number of times they occurred in our data set. This showed us the most "popular" sites, as measured by the number of in-links observed. These appear in Table 7.

The in-link entries marked with `(*)` indicate sites that arehighly self-referential. That is, these sites (by inspection) appearto contain a great number of links to their own top-level pages. It would probably be instructive to count only links from outside a given site.

3.10. Readability

The UNIX utility `style` was used to assess the readability level of a subset of the HTML documents in our data set (approximately 150,000). We remove HTML markup before invoking `style` on each document. We do this for two reasons. First, `style` does not understand HTML, so the extra punctuation would confuse its analyzer. Second, breaking English text into sentences and sentence fragments can be tricky and we need to provide the `style` analyzer with some assistance. For example, it is not always clear when a bulleted list should be ignored, treated as a single long sentence, or treated as a list of individual sentences. When invoked on `troff` documents, `style` uses a set of heuristics to insert punctuation into text, using the markup to assist it [3]. This information is then used by later passes of

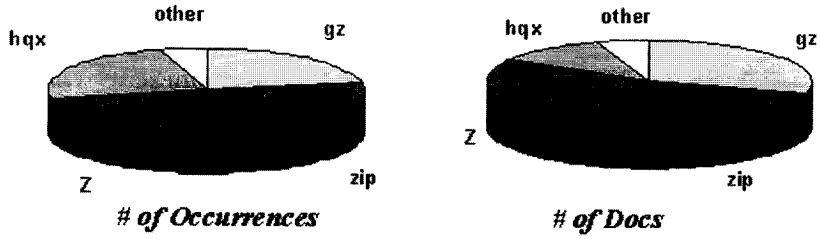


Fig. 8. Distribution of compression/archive files.

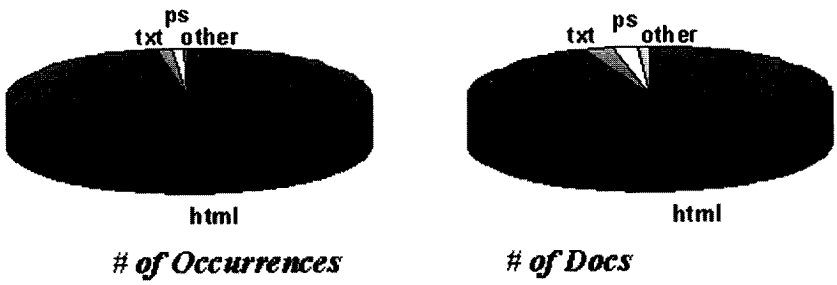


Fig. 9. Distribution of document files.

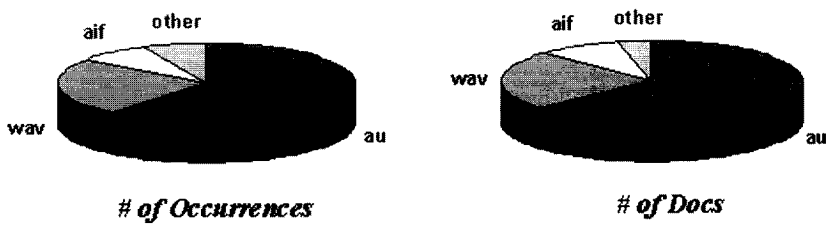


Fig. 10. Distribution of audio files.

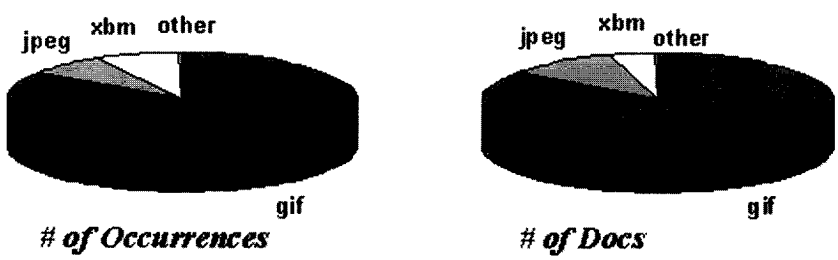


Fig. 11. Distribution of image files.

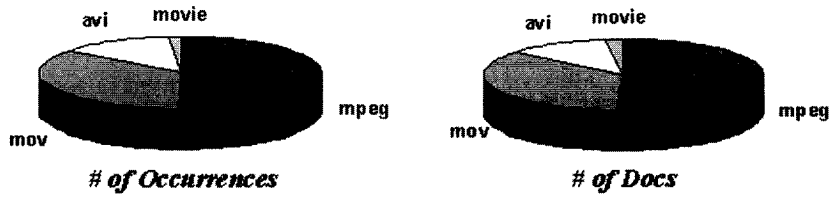


Fig. 12. Distribution of movie files.

Table 7
Most-linked-to URLs

Site	Description	In-links
www.xerox.com [a]	Xerox PARC	28188 *
www.yahoo.com [b]	Yahoo	19424
cool.infi.net [c]	Cool Site of the Day	19028
hamsterix.funet.fi [d]	Bible (in Finnish)	17243 *
sundarssrv2.cern.ch [e]	CERN preprint service	16049 *
wings.buffalo.edu [f]	Best of the Web '94	14685
wings.buffalo.edu [g]	U.S. Gazetteer	14369
www.ist.unige.it [h]	Cell database	12750 *
home.netscape.com [i]	Netscape Communications	12081
www.american.recordings.com [j]	Ultimate Band List	11014
jasper.ora.com [k]	Comprehensive TeX Archive Network	10650
www.ibm.com [l]	IBM Corp.	10617
www.informatik.uni-trier.de [m]	Bibliography Server on Database Systems & Logic Programming	10212 *
siva.cshl.org [n]	wusage 3.2 (WWW usage statistics)	9038
curly.cc.utexas.edu [o]	Jane Austen's Pride & Prejudice	8928 *
www.starwave.com [p]	StarWave	8721
allison.clark.net [q]	Rob & Jen's Genealogy Page	8476 *
helios.jicst.go.jp [r]	Japan Information Center of Science and Technology	8331
neoteny.eccosys.com [s]	NetSurf mailing list	8036 *

* Highly self-referential sites.
[a] http://www.xerox.com:80/
[b] http://www.yahoo.com:80/
[c] http://cool.infi.net:80/
[d] http://hamsterix.funet.fi:80/pub/doc/religion/christian/Bible/html/finnish/1992/
[e] http://sundarssrv2.cern.ch:80/cgi-bin/ppbyauthor.sh
[f] http://wings.buffalo.edu:80/contest/
[g] http://wings.buffalo.edu:80/geogw
[h] http://www.ist.unige.it:80/cldb/spe16.html
[i] http://home.netscape.com:80/
[j] http://www.american.recordings.com:80/WWWoM/cgi-bin/ubl
[k] http://jasper.ora.com:80/cgi-bin/ftp_ctan.cgi
[l] http://www.ibm.com:80/
[m] http://www.informatik.uni-trier.de/~ley/db/index.html
[n] http://www.boutell.com/wusage/
[o] http://curly.cc.utexas.edu:80/~churchh/ppdrmtis.html
[p] http://www.starwave.com:80/
[q] http://allison.clark.net/pub/rj/
[r] wais://helios.jicst.go.jp:210/jitr
[s] http://neoteny.eccosys.com/cgi-bin/hotmess

Table 8
Average readability broken down by domain

Domain	Readability score
com	10.3
edu	11.0
gov	10.0
net	12.3
mil	12.1
org	11.2

the analyzer to determine sentence and sentence fragment breaks. We use a similar set of heuristics to insert periods and commas into HTML documents as we strip out markup.

The numbers presented in Table 8 represent the scores of the different domains on the Kincaid readability test. Higher numbers represent more grammatical and lexical complexity. Lower numbers represent more simple structure and word choice. Documents with lower numbers are considered to be more "readable". The "other" domain is excluded because it represents extraordinarily diverse sources.

3.11. Syntax errors

`weblint` was used to assess the syntactic correctness of a subset of the HTML documents in our data set (approximately 92,000). Fig. 13 presents the top ten syntax errors ranked according to the percentage of documents in which they appear. (See also Table 9.) (Note that "netscape-attribute" is not necessarily an error, but rather indicates the percentage of documents using Netscape-specific extensions.) Observe that over 40% of the documents in our study contain at least one error. Descriptions of the errors appear in Tables 10 and 11.

4. Conclusions

We have reported the results of our examination of pages from the World Wide Web. Additional data not presented in the hardcopy version of this paper may be found at **http://www.cs.berkeley.edu/~woodruff/inktomi/**.[28]

4.1. Truisms

There are two maxims which are particularly apropos of our experience. First, dealing with large data sets is difficult and time-consuming. None of the existing tools which we used scaled adequately to dealing with a data set on the order of millions of documents.

[28] http://www.cs.berkeley.edu/~woodruff/inktomi/

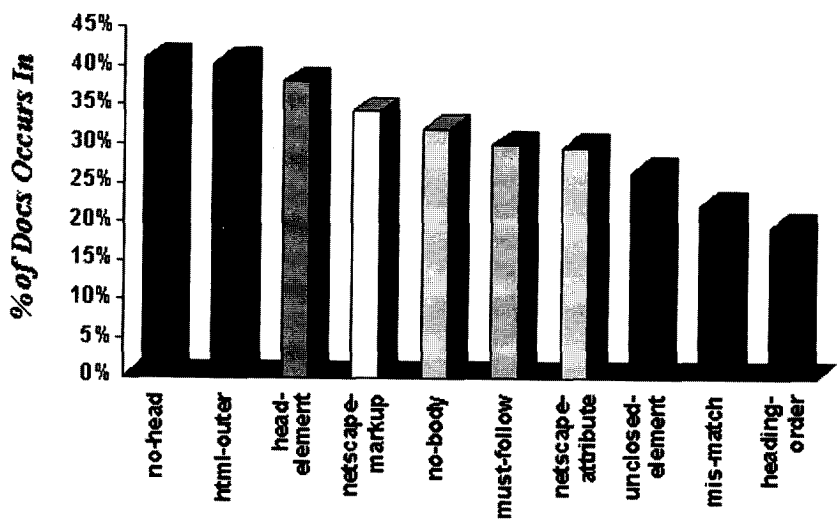


Fig. 13. Ten most common syntax errors.

Table 9
Syntax errors

Error	% of docs	# docs
no-head	41%	37498
html-outer	40%	37124
head-element	38%	34882
netscape-markup	34%	31433
no-body	32%	29308
must-follow	30%	27556
netscape-attribute	29%	27145
unclosed-element	27%	24439
mis-match	22%	20404
heading-order	19%	17886
unknown-element	16%	15040
body-no-head	12%	10942
unknown-attribute	8%	7542
odd-quotes	7%	6147
element-overlap	6%	5925
required-context	6%	5586
once-only	5%	5018
nested-element	5%	4980
no-title	5%	4514
non-head-element	5%	4394
literal-metacharacter	2%	1912
unexpected-open	2%	1669
required-attribute	1%	1269
illegal-closing	1%	1257
repeated-attribute	1%	1175
require-head	1%	1104
closing-attribute	1%	996
markup-in-comment	1%	759
unclosed-comment	1%	563
expected-attribute	1%	492
obsolete	0%	307
leading-whitespace	0%	211
attribute-delimiter	0%	38
mixed-case	0%	0
directory-index	0%	0

Second, we observed empirically that the Web changes exceptionally quickly. Many properties of the documents in our first data set have altered in the months since the data was collected. The largest document in our data set was 1.6Mbytes; we checked the current size of that same document. It has grown to 9 Mbytes. As another example, many of the most popular URLs in the first data set no longer exist.

4.2. Future directions

A longitudinal study examining trends would be extremely interesting. Our limited observation reveals that while certain charactertistics change fairly quickly (e.g., new features are introduced) others appear to change more slowly (e.g., average document size and reading level did not appear to change between the time periods we observed). One could also consider how the introduction of new tools impact these characteristics. For example, as authoring tools become more common, one could study their impact on the number and type of syntax errors.

Structural graph analysis has many applications in this area. In particular, analysis of the kind practiced by sociologists in *structural network analysis* [20] promises insight. However, existing social network algorithms are several orders of magnitude more complex than is viable for a data set of this size. Significant work would have to be done to make such analysis feasible.

It would also be interesting to allow user-defined queries against the data set. The simplest functional-

Table 10
List of `weblint` errors

Error name	Explanation
html-outer	outer tags should be <HTML>..</HTML>
no-head	missing <HEAD>
head-element	heading-only tag (TITLE, NEXTID, LINK, BASE, META) found outside of heading
no-body	missing <BODY>
must-follow	required tag does not immediately follow another
unclosed-element	unclosed elements (e.g., <H1>...)
netscape-markup	Netscape-specific tag
empty-container	empty container element
mis-match	mis-matched tag (e.g., <H1>...</H2>)
heading-order	order of headings (e.g., <H3> following <H1>)

ity would be to allow a user to ascertain how a form-specified URL compared with the data set. A more interesting and complex interface would allow the user to define arbitrary queries on the data set.

References

[1] N. Bowers, Weblint Home Page (version 1.013), Khoral Research, Inc., Albuquerque, NM, Jan. 1996. http://www.khoral.com/staff/neilb/weblint.html

[2] L.D. Catledge and J. E. Pitkow, Characterizing Browsing Strategies in the World-Wide Web, *Proc. 3rd Int. World Wide Web Conf.*, Darmstadt, Germany, Apr. 1995. http://www.igd.fhg.de/www/www95/proceedings/papers/80/userpatterns/UserPatterns.Paper4.formatted.html

[3] L.L. Cherry, Writing Tools - The STYLE and DICTION Programs, Computer Science Technical Report No. 91 (TM 79-1271-13), Bell Laboratories, Murray Hill, NJ, Feb. 1981. Revised version reprinted as L.L. Cherry and W. Vesterman, Writing Tools - The STYLE and DICTION Programs, 4.4 BSD User's Supplementary Documents, Computer Science Research Group, Berkeley, CA, 1994.

[4] E.H. Chi, Webspace Visualization, The Geometry Center, Univ. of Minnesota, Minneapolis, MN. http://www.geom.umn.edu/docs/weboogl/webspace/webspace.html

[5] CommerceNet Consortium, The CommerceNet/Nielsen Internet Demographics Survey, Menlo Park, CA, 1995.

Table 11
`weblint` errors

Error name	Explanation
html-outer	outer tags should be <HTML>..</HTML>
no-head	missing <HEAD>
head-element	heading-only tag (TITLE, NEXTID, LINK, BASE, META) found outside of heading
no-body	missing <BODY>
must-follow	required tag does not immediately follow another
unclosed-element	unclosed elements (e.g., <H1>...)
netscape-markup	Netscape-specific tag
empty-container	empty container element
mis-match	mis-matched tag (e.g., <H1>...<H2>)
heading-order	order of headings (e.g., <H3> following <H1>)
netscape-attribute	Netscape-specific attribute
unknown-element	unknown tag
body-no-head	<BODY> but no <HEAD>
unknown-attribute	unknown attribute
no-title	missing <TITLE>
element-overlap	overlapped elements
required-context	failed context check (where a tag must appear within a certain element)
nested-element	illegally nested element
odd-quotes	odd number of quotes in tag (unclosed quotes)
non-head-element	tag other than a valid heading tag (ISINDEX, TITLE, NEXTID, LINK, BASE, META, RANGE, STYLE) found in heading
once-only	catches elements which should only appear once
unexpected-open	unexpected < (potentially unclosed element)
closing-attribute	closing tag should not have any attributes specified
require-head	expects to see a TITLE in the HEAD element
illegal-closing	element is not a container
repeated-attribute	repeated attribute
expected-attribute	missing expected attribute
unclosed-comment	unclosed comment
obsolete	obsolete element
markup-in-comment	markup embedded in comments (can confuse some browsers)
required-attribute	missing required attribute
leading-whitespace	should not have whitespace between < and tag
attribute-delimiter	use of for attribute value delimiter is not supported by all browsers
directory-index	directory does not have an index file
literal-metacharacter	use of a literal metacharacters instead of an &-symbol

http://www.commerce.net/information/surveys/execsum/exec_sum.html

[6] D. Connolly, A Lexical Analyzer for HTML and Basic SGML, W3C Working Draft, World Wide Web Consortium, Cambridge, MA, Dec. 1995. http://www.w3.org/pub/WWW/TR/

[7] R. Fielding, Relative Uniform Resource Locators, RFC 1808, June 1995. http://www.cis.ohio-state.edu/htbin/rfc/rfc1808

[8] H. Frystyk and H.W. Lie, Towards a Uniform Library of Common Code: A Presentation of the World Wide Web Library, *Proc. 2nd Int. World Wide Web Conference*, Chicago, IL, Oct. 1994. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/DDay/frystyk/LibraryPaper.html

[9] M.J. Hannah, HTML Reference Manual, Sandia National Laboratories, Albuquerque, NM, Dec. 1995. http://www.sandia.gov/sci_compute/html_ref.html

[10] R. Karpinski, Hot Java Arrives: Sun Aims to Revolutionize the Web, *InteractiveAge*, May 22, 1995. http://techweb.cmp.com/ia/15issue/15hotjava.html

[11] Lycos, Inc., The Lycos 250 and Hot Lists, Pittsburgh, PA, Sep. 1995. http://www.lycos.com/lists/index.html

[12] M.L. Mauldin and J.R.R. Leavitt, Web Agent Related Research at the Center for Machine Translation, 1994 Meeting of the ACM Special Interest Group on Networked Information Discovery and Retrieval, McLean, VA, Aug. 1994. http://fuzine.mt.cs.cmu.edu/mlm/signidr94.html, Carnegie Mellon Univ., Jul. 1994.

[13] S. Mukherjea and J.D. Foley, Visualizing the World-Wide Web with the Navigational View Builder, *Proc. 3rd Int. World Wide Web Conf.*, Darmstadt, Germany, Apr. 1995. http://www.igd.fhg.de/www/www95/proceedings/papers/44/mukh/mukh.html

[14] P. Pirolli, J. Pitkow and R. Rao, Silk from a Sow's Ear: Extracting Usable Structures from the Web, Xerox PARC, Palo Alto, CA, Nov. 1995. Submitted for publication.

[15] J.E. Pitkow and K. Bharat, WEBVIZ: A Tool for World Wide Web Access Log Visualization, *Proc. 1st Int. World Wide Web Conf.*, Geneva, Switzerland, May 1994. http://www1.cern.ch/WWW94/PrelimProcs.html

[16] J.E. Pitkow and M.M. Recker, Results From The First World-Wide Web User Survey, Georgia Institute of Technology, Atlanta, GA, Jan. 1994. http://www.gatech.edu/pitkow/survey/survey-1-1994/survey-paper.html

[17] J.E. Pitkow and M.M. Recker, Using the Web as a Survey Tool: Results from the Second WWW User Survey, *Proc. 3rd Int. World Wide Web Conf.*, Darmstadt, Germany, Apr. 1995. http://www.igd.fhg.de/www/www95/proceedings/papers/79/survey/survey_2_paper.html

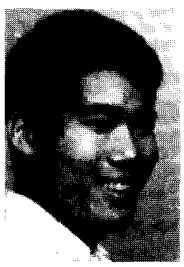
[18] J.E. Pitkow and C. Kehoe, The GVU Center's 3rd WWW User Survey, Georgia Institute of Technology, Atlanta, GA, Apr. 1995. http://www.cc.gatech.edu/gvu/user_surveys/survey-04-1995/

[19] M. Rissa and C. Oy, WWW User Survey Results, Helsinki, Finland, Feb. 1995. http://www.mroy.fi/dec94.htm

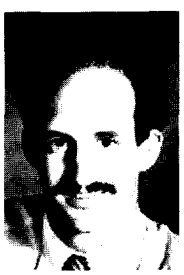
[20] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, UK, 1994.

[21] Yahoo, Inc., Survey Says... Mountain View, CA, Aug. 1995. http://www.yahoo.com/docs/survey/first.html and http://www.yahoo.com/docs/survey/index.html

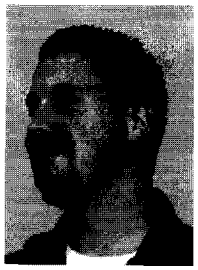
Allison Woodruff is a PhD student in the Electrical Engineering and Computer Science Department at the University of California, Berkeley. Her research interests include spatial information systems, multimedia databases, visual programming languages, and user interfaces. She has worked as a geographic information systems specialist for the California Department of Water Resources. Woodruff holds a BA in English from California State University, Chico and an MA in Linguistics and an MS in Computer Science from the University of California, Davis.



Paul M. Aoki is a PhD student in the Department of Electrical Engineering and Computer Sciences at the University of California at Berkeley. He holds a B.S. in Electrical Engineering and a M.S. in Computer Science from the University of California at Berkeley. His research interests include query optimization for parallel and distributed databases and index support for non-traditional data types.



Eric Brewer is an Assistant Professor of Computer Science at the University of California at Berkeley, and received his PhD in CS from MIT in 1994. Interests include mobile and wireless computing (the InfoPad and Daedalus projects); scalable servers (the NOW and Inktomi projects); and application- and system-level security (the ISAAC project and Netscape security holes). Previous work includes multiprocessor-network software and topologies (Strata, metabutterflies), high-performance multiprocessor simulation (Proteus).



Paul Gauthier has served as Directo. and Vice President of Research and Development of Inktomi Corporation since February 1996. Mr. Gauthier is also in the doctorate program in the Department of Electrical Engineering and Computer Sciences at the University of California at Berkely, where he is working towards a doctoral degree in computer science. Mr. Gauthier holds a Bachelor of Science degree, with honors, in Computer Science from Dalhousie University (located in Nova Scotia, Canada).

Lawrence A. Rowe received a BA in mathematics and a PhD in information and computer science from the University of California at Irvine in 1970 and 1976, respectively. Since 1976 he has been on the faculty at the University of California at Berkeley where he is now a Professor of Electrical Engineering and Computer Science and the founding director of the Berkeley Multimedia Research Center. His current research interests are multimedia applications, systems, and databases on which he has published over fifty papers. He is an editorial board member for the ACM Multimedia Systems Journal. Professor Rowe heads the research group that developed the Berkeley Distributed Video-on-Demand System, algorithms to compute special effects on compressed images, the Berkeley Continuous Media Toolkit, and the Berkeley MPEG1 video tools.