Surfing the Web by Site

David Gibson
davgib@us.1bm.com
IBM Almaden Research Center
650 Harry Rd
San Jose, CA 95120

ABSTRACT

We provide a system for surfing the web at a high level of abstraction, which is an analogy of the web browser, but which displays entire sites at a time. It allows a principled investigation of what is present, based on an overview of all available information. We show a site's relation to other sites, the broad nature of the information contained and how it is structured, and how it has changed over time. Our current system maintains a continuously updated archive of 40 million sites representing 1.9 billion web pages, and enables real-time navigation through the sea of web sites.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Information Search and Retrieval—Information filtering; H.5.4 [Information Systems]: Hypertext/Hypermedia—Navigation,User issues

General Terms: Algorithms, Measurement, Experimentation, Human Factors.

Keywords: Novel Browsing Paradigms, Web Navigation Strategies, Large Scale Systems

1. INTRODUCTION

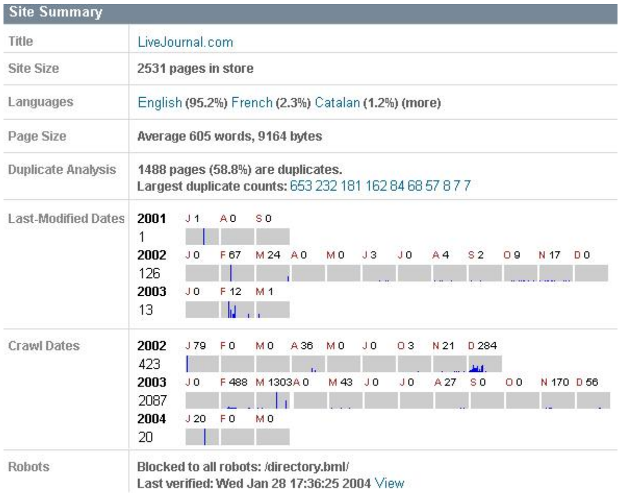
In much the same way as search engines, this work improves the usefulness of the web not by improving the interface, but by enabling access to new information. While a search engine is able to find a good site about a favorite hobby, it cannot tell how extensive and well-maintained it is, and how it relates and compares to other hobbyist sites. These are issues which involve aggregating information from across all the pages on the site, and they are usually left to manual exploration. Automating this process completely is very dependent on particular users' needs. Site Browser attempts to provide as much assistance as possible, for a variety of possible user objectives [7]:

Navigation While many sites provide site maps and navigation aids, they are most often limited or altogether missing. Site Browser builds alternative navigation structures with a consistent interface.

Discovery Many information retrieval researchers seek to recapture the serendipity effect of physical libraries [3]. Viewing a particular site at a time very naturally exploits the selectivity of the site author, and the topical coherence of a site means that related documents are likely to be "nearby".

Assessment By gathering information about the site that contains a particular page, we can better form judgements about freshness, authorship, audience, reliability, and scope.

The Site Browser is, in this initial form, a very general purpose tool. It will be useful to professionals engaged in such fields as business competitive intelligence, current events analysis, or other research tasks, as well as to the general public browsing for news, goods, or entertainment.



Site Summary	
Title	LiveJournal.com
Site Size	2531 pages in store
Languages	English (95.2%) French (2.3%) Catalan (1.2%) (more)
Page Size	Average 605 words, 9164 bytes
Duplicate Analysis	1488 pages (58.8%) are duplicates. Largest duplicate counts: 653 232 181 162 84 68 57 8 7 7
Last-Modified Dates	**2001** J 1 A 0 S 0 1 **2002** J 0 F 67 M 24 A 0 M 0 J 3 J 0 A 4 S 2 O 9 N 17 D 0 126 **2003** J 0 F 12 M 1 13
Crawl Dates	**2002** J 79 F 0 M 0 A 36 M 0 J 0 O 3 N 21 D 284 423 **2003** J 0 F 488 M 1303 A 0 M 43 J 0 J 0 A 27 S 0 O 0 N 170 D 56 2087 **2004** J 20 F 0 M 0 20
Robots	Blocked to all robots: /directory.bml/ Last verified: Wed Jan 28 17:36:25 2004 View

Figure 1: LiveJournal Summary Page

2. FEATURES TO MEASURE

Site Browser is built on the WebFountain [2] platform for large scale text analytics, which gives us large volumes of current web page data. While comprehensiveness is essential, this large data set poses challenges in making the system responsive. In a tradeoff between wide-coverage low-content approaches [1] and richer interfaces [4, 5, 6], we choose to precompute site summaries for the entire web, since many useful statistics are very rapid to compute:

Measures of size We aggregate the total number of pages, words, and bytes on the site.

Page attributes We maintain simple counters for many features of pages, such as the language detected to be on the page, the HTTP return code, and the presence of media files such as .mpg and .wav.

Date ranges The web crawler reports the date of the last page fetch, and the last-modified date of the page, as returned by the web server. We can thus plot crawl frequency and update frequency.

Directory structure Each URL consists of a path into the logical directory tree of pages on the site. By assembling all URLs on the site we can reconstruct the directory structure. In many cases this gives a very useful overview of the site contents.

Link structure We record the hostname in each link, to show which other sites are referenced, and with what frequency.

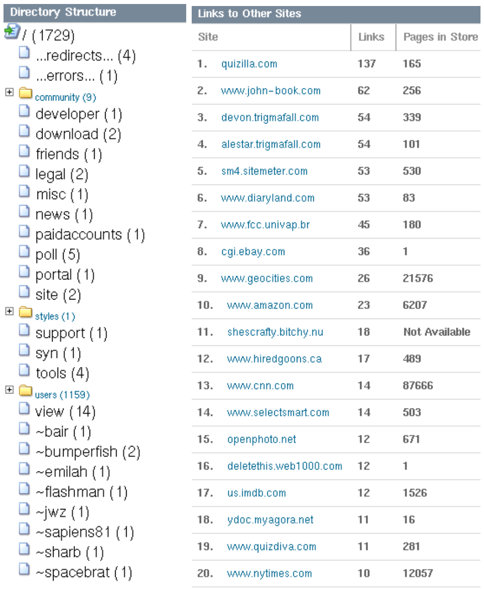
Keyword frequency We track all words appearing on the site, and show the most frequent.

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

3. EXPERIENCES

The interface design for Site Browser follows principles of minimality, interactivity, and performance. There are four basic views: a Site Summary view, Directory Structure, Links to Other Sites, and Top Keywords. The median response time is 0.13 seconds.

Throughout the interface, results are "clickable" to show for example the pages in a directory, or matching a keyword, or containing links to the given site. This enables users to verify that the overview statistics are indeed correct, and gives a starting point for continued browsing.



Directory Structure

/ (1729)
- ...redirects... (4)
- ...errors... (1)
- ⊞ community (9)
- developer (1)
- download (2)
- friends (1)
- legal (2)
- misc (1)
- news (1)
- paidaccounts (1)
- poll (5)
- portal (1)
- site (2)
- ⊞ styles (1)
- support (1)
- syn (1)
- tools (4)
- ⊞ users (1159)
- view (14)
- ~bair (1)
- ~bumperfish (2)
- ~emilah (1)
- ~flashman (1)
- ~jwz (1)
- ~sapiens81 (1)
- ~sharb (1)
- ~spacebrat (1)

Links to Other Sites

	Site	Links	Pages in Store
1.	quizilla.com	137	165
2.	www.john-book.com	62	256
3.	devon.trigmafall.com	54	339
4.	alestar.trigmafall.com	54	101
5.	sm4.sitemeter.com	53	530
6.	www.diaryland.com	53	83
7.	www.fcc.univap.br	45	180
8.	cgi.ebay.com	36	1
9.	www.geocities.com	26	21576
10.	www.amazon.com	23	6207
11.	shescrafty.bitchy.nu	18	Not Available
12.	www.hiredgoons.ca	17	489
13.	www.cnn.com	14	87666
14.	www.selectsmart.com	14	503
15.	openphoto.net	12	671
16.	deletethis.web1000.com	12	1
17.	us.imdb.com	12	1526
18.	ydoc.myagora.net	11	16
19.	www.quizdiva.com	11	281
20.	www.nytimes.com	10	12057

Figure 2: LiveJournal Directories and Links

Figure 1 shows the summary statistics for a large community site. From this page we can see how frequently it has been crawled, and view miscellaneous statistics. This view contributes to the assessment task, and also allows navigation into particular subcommunities, such as the French language subcommunity.

Figure 2 shows the directory structure for this community site. Most pages are in the `/users/` directory, but distinguished users and administration pages are clear. This is the best point of entry for navigation within the site: indeed, the directory structure can be as useful as the site's own site map, and often more complete.

Figure 2 also shows that LiveJournal users link most to quizilla and to news sites, and, of course, to other blogging sites. This is an ideal starting point for discovery of related sites.

4. ARCHITECTURE

The system is designed to minimize the number of times a given piece of data is read or written to disk. The first processing stage, *Gather*, assigns each site to one of 8 dual-2.4GHz CPU, 4 GB RAM build machines. A stream of new pages is extracted from the main cluster of 256 machines, where each machine stores a distinct set of pages partitioned by hash of the URL. Thus, on each of the build machines we begin with 256 separate input "packets". For each page we store a record consisting of counts of all the features described in section 2. Each packet is sorted by site, by using hostnames. In the *Update* phase we accumulate this newpages data into the main pages store, which is also sorted by site. Older versions of pages are discarded. Finally, the *Merge* phase performs a 256-way merge to extract a contiguous sequence of pages on each site, accumulates the statistics, and writes to the site store, which is currently 300GB in size, using compression.

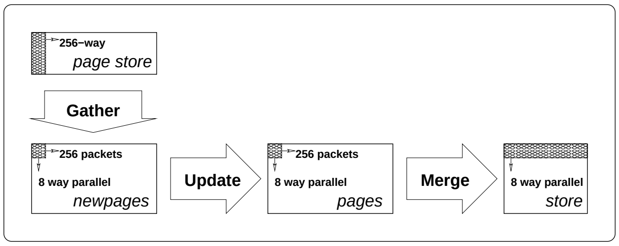


Figure 3: Data flow

5. REACTIONS AND CONCLUSIONS

The Site Browser system has been presented to a initial group of users, consisting primarily of professional information analysts. Their initial reactions are that it has clear potential to aid their tasks. The directory structure display has been the most useful: our users have found some sites with "particularly interesting directory structure ... which tells quite a bit about the site".

Search engines revealed things about the web which were not obvious before, although the information was in principle accessible, such as finding out what a particular person is involved in. And this has lead to changes in the way pages are published and promoted. In the same way, Site Browser reveals details about sites and their interrelations, which again may change the behavior of site creators.

Directory Structure, for example, reveals the conceptual layout of the site in a consistent way, as it appears to the creator, and not as it appears on the site's own navigation system. It will show directories which are intended to be less commonly accessed, as readily. Top Keywords can show the balance of topics more accurately than what is portrayed directly, and bring out interesting topics that may be hidden on the site.

The notion of browsing by overviews introduces a new kind of information to users of the web. Page-level information merely states that some datum exists: overview information provides its prevalence and context. Global indexes became useful because most creators do not provide comprehensive site indexes. We expect global overview services to become popular in the same way.

6. REFERENCES

[1] Alexa. `http://www.alexa.com`.

[2] S. Dill, et al. Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proc 12th International WWW Conf*, Budapest, Hungary, May 2003.

[3] M. Hearst. User interfaces and visualization. In *R. Baeza-Yates and B. Ribeiro-Neto (Eds.) Modern information retrieval*. NY: ACM Press., 1999.

[4] Y. Maarek and I. Shaul. Webcutter: A system for dynamic and tailorable site mapping. In *Proceedings of the 6th International World Wide Web Conference*, 1997.

[5] G. Marchionini and B. Brunk. Toward a general relation browser: A GUI for information architects. In *Journal of Digital Information*, volume 4, 2003.

[6] D. Nation, C. Plaisant, G. Marchionini, and A. Komlodi. Visualizing websites using a hierarchical table of contents browser: WebTOC. In *Designing for the Web: Practices and Reflections*, 1997.

[7] A. J. Sellen, R. Murphy, and K. L. Shaw. How knowledge workers use the web. In *Proc. SIGCHI*, 2002.