The Connectivity Sonar: Detecting Site Functionality by Structural Patterns

Einat Amitay David Carmel Adam Darlow Ronny Lempel Aya Soffer
IBM Research Labs
Haifa 31905, Israel
[einat/carmel/darlow/rlempel/ayas]@il.ibm.com

ABSTRACT

Web sites today serve many different functions, such as corporate sites, search engines, e-stores, and so forth. As sites are created for different purposes, their structure and connectivity characteristics vary. However, this research argues that sites of similar role exhibit similar structural patterns, as the functionality of a site naturally induces a typical hyperlinked structure and typical connectivity patterns to and from the rest of the Web. Thus, the functionality of Web sites is reflected in a set of structural and connectivity-based features that form a typical signature. In this paper, we automatically categorize sites into eight distinct functional classes, and highlight several search-engine related applications that could make immediate use of such technology. We purposely limit our categorization algorithms by tapping connectivity and structural data alone, making no use of any content analysis whatsoever. When applying two classification algorithms to a set of 202 sites of the eight defined functional categories, the algorithms correctly classified between 54.5% and 59% of the sites. On some categories, the precision of the classification exceeded 85%. An additional result of this work indicates that the structural signature can be used to detect spam rings and mirror sites, by clustering sites with almost identical signatures.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods, Clustered index; H.3.3 [**Information Search and Retrieval**]: clustering, information filtering, selection process; H.5.1 [**Multimedia Information Systems**]: Hypertext navigation and maps

General Terms

Algorithms, Measurement, Experimentation

Keywords

Web graphs, link analysis, Web IR

HT'03, August 26–30, 2003, Nottingham, United Kingdom.

1. INTRODUCTION

The number of Web sites is ever-growing, as is the utility and functionality of the Web. Web sites today serve many different functions. Some sites deliver content, others serve as shopping malls; some contain information concerning a specific corporation, while others provide interfaces for searching the Web. Since sites are created for different purposes and by different people, it should come as no surprise that they sport different designs: the sizes of the sites, the organization of the pages in directories and subdirectories, the internal linkage patterns within the sites' pages and the manner in which the sites link to the rest of the Web, are all suited to and stem from the sites' functionality. The connectivity patterns of "the Web" to sites with different roles are also far from similar.

Although Web sites are diverse in terms of their look and feel, there are many sites that serve similar roles. Consider, for example, sites of two unrelated universities, two different Web directories, or two competing corporations. This paper explores whether such sites exhibit similar structural patterns, despite being designed by different Web masters. In other words, does the functionality of a site somehow induce (1) a typical hyperlinked structure, and (2) typical connectivity patterns to and from the rest of the Web. We argue that indeed, the functionality of large and well-connected sites is reflected in a set of structural and connectivity-based features. Furthermore, the values attained by a Web site across this set of features often results in a typical signature, from which the functionality of the site can be deduced[1].

This research aims to categorize sites by functionality. We want to detect what a site *is*, and not what it *is about*. Our goal, of discerning the *type* of a site, should not be confused with the challenging task of categorizing Web pages and sites by *topical content* (surveyed in Section 7.2). Specifically, we categorize sites into the following eight functionality categories: corporate sites, content & media sites (such as sites of major newspapers and TV networks), search engines, Web hierarchies & directories, portals (both general Web portals and community-specific portals), E-stores, virtual hosting services and universities. These categories will be denoted CORP, CONTENT, SEARCH, DIR, PORTAL, ESTORE, VHOST and UNIV, respectively. Note that these eight categories, while covering the functionality of many of the larger Web sites, do not encompass the entire range of site functionalities found on the Web. The Web's ecosystem

[1]This is similar to the way dolphins use sonar to identify objects based on their unique pattern of reflecting sound waves, hence the name Connectivity Sonar.

is home to other site types, some of which are discussed in Section 8.

We purposely limit our categorization efforts to features that are extracted from (1) aggregate structural properties of the sites (the organization of pages into directories, and the internal links connecting those pages), as well as (2) the connectivity patterns between the site and the rest of the Web. We make no use of any content analysis, neither of the Web pages nor of their URLs. Consequently, our method can be applied to sites in any language and regardless of any local conventions for URLs. Our results show that with precision exceeding 55%, we can classify sites correctly into the above eight categories. Some of the categories were very difficult to discern compared to others, most likely indicating that they are not well defined in terms of structure. However, the precision attained on some categories, such as virtual hosting sites and Web hierarchies, exceeded 85%. An additional result of this work indicates that the structural signature can be used to detect spam rings and mirror sites. By clustering sites with almost identical signatures together, we were able to detect numerous spam rings in our data.

Several search-engine related applications could make immediate use of site tagging using our proposed method. For example, search engines could automatically tag each returned search result with the type of site associated with the result. Type tagging will allow search engines to provide users with a better search experience in cases where the user's information need is unclear. In another example, the crawling policy of a search engine could vary based on the type of the site. Sites which are tagged as *content* sites might merit frequent crawls in order to enable the engines to keep track of current events, trends and other issues of volatile temporal nature. On the other hand, university sites, which have been reported to change less frequently than pages on the ".com" top level domain [19], can be crawled less often. Finally, we envision a more realistic random surfer model based on the site type. While current random surfing models assume a uniform stochastic behavior across all Web pages, we argue that browsing patterns in sites of different roles exhibit different stochastic properties. For example, a person browsing the search page of a search engine is much more likely to enter a query and browse one of the dynamically-generated results, than to follow the static link to the search company's "about" page.

The rest of this paper is organized as follows. Section 2 expands on the intuition and motivation driving our efforts. Section 3 further elaborates on several possible applications where knowledge of sites' functionalities could come into play. Section 4 defines the terminology used in this paper, and details the connectivity and structural data which we had at our disposal, courtesy of the search engine AltaVista [2]. Section 5 presents the specific features which were used to distinguish between sites of different functionalities. Section 6 reports the results of our experiments. Related work is covered in Section 7. Our conclusions and suggestions for future research are brought in Section 8.

2. OUR APPROACH - INTUITION

Before delving into the details of our proposed categorization, we explore the intuition behind the conjecture that sites of similar functionality exhibit similar structural signatures. The following considers several different types of Web sites. For each type, we intuitively list structural properties that it should ideally exhibit. These examples are somewhat simplistic and ignore the harsh realities of the Web. However, they convey the intuition behind this research and its methodology.

Search engines

Our first example, in Figure 1(a), concerns a *pure* search engine, one which does not include a Web directory (e.g. Teoma[3]). Such an engine, if popular among Web users, will seem as a *black hole* in cyberspace: (1) its *static* site will be very small (condensed), consisting of the search page and a few other pages such as "corporate info" and "advanced search"[4]; (2) it will attract a vast amount of inlinks from other Web sites, as if having a strong gravitational pull, and (3) the site's pages will contain very few outlinks to other sites (hardly anything escapes from the site).

Web Hierarchies and Directories

Large scale Web hierarchies and directories categorize sites and pages into rich taxonomies, and link to the categorized sites. Typically, every node of the taxonomy will be represented by a Web page with outlinks to each of the pages that are deemed as pertaining to that node. Thus, directories will have many thousands of outlinks. Furthermore, this outflow of links will be structured: since taxonomies resemble trees, the number of outlinks from directories will increase with the depth of the pages. Additionally, leaf nodes of the taxonomy tree will tend to have many outlinks (See Figure 1(b)).

The pattern of linking to a directory is also unique. While directories certainly attract many inlinks to their main (top-level) page, many inlinks also point to deeper pages of the site. This is not surprising, since internal nodes of directories serve as hubs[27] for their corresponding part of the taxonomy, and are thus attractive to the relevant cyber-community.

Corporate sites

Sites of large corporations (Figure 1(c)) will typically consist of thousands of pages, and will be organized in a complex directory structure. Corporate sites usually have a robust infrastructure of internal linkage, since the pages are created using a template which usually has both "home" and "up" buttons. Many templates also include navigational panels, thus adding *cross links* between different branches of the corporation (and its site tree) [3].

The commercial and competitive nature of corporate sites implies that they will tend to link significantly less to other (external) sites, as compared with the outpour of external links from Web hierarchies.

Virtual hosting services

Intuitively, sites which provide virtual hosting services should have very loose internal linkage. Usually, there is no linkage from the corporate part of the site to the homepages of the individual hosted sites (no directory of hosted entities). The hosted sites may be required to link back to the corporate

[2]http://www.altavista.com/

[3]http://www.teoma.com

[4]Note that the result pages that are served by search engines are *dynamic* pages, generated on the fly by the demands of the query streams, are are not an integral part of the site.

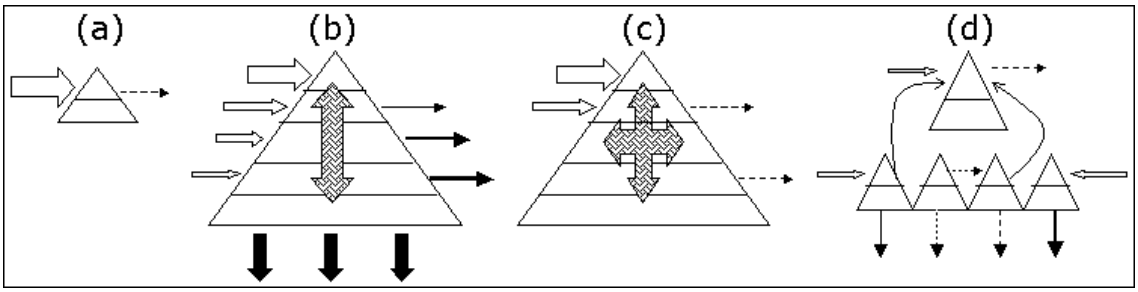

Figure 1: Schematic drawings of (a) search engines, (b) directories, (c) corporate sites and (d) virtual hosting services

site of their host, but that is as far as the requirements go. Spam sites aside, cross-links between separate entities that happen to be hosted on the same site appear at random, and are usually greatly outnumbered by links to external Web sites. Incoming links will tend to point to the individual hosted sites, and not to the corporate site of the hosting entity. See Figure 1(d).

Universities and research institutes

University Web sites are hybrids of corporate sites and virtual hosting services. The administrative parts of a university's site will often resemble a corporate site, whereas the individual home pages of the faculty and students are as chaotic and loosely connected as virtually hosted sites.

3. POTENTIAL APPLICATIONS

The methodology of our research entails examining prominent sites of various roles, and understanding the connectivity and structural features that characterize each role and differentiate it from other site types. Identifying the said characteristics of various site types can enhance the area of Web-site design: understanding the design patterns of successful e-stores, popular directories and well-organized corporate sites, for example, can pave the road for the design of better sites of similar nature. This reasoning has been voiced throughout the history of hypertext, as will be surveyed in Section 7.1. Furthermore, below we identify three applications where the ability to detect the role of sites may potentially have high impact.

3.1 Type-Tagging of Search Results

We envision a paradigm where search engines would automatically tag each returned search result with the type of site that contributed the result. Such tags can take form of a small icon to the side of each result. Type tagging will allow search engines to provide users with a better search experience in cases where the user's information need is unclear [10]. Consider, for example, a query such as "operating systems", where the engine would tag results from (1) corporations that sell operating systems, (2) content sites and developer portals that contain relevant discussion boards and forums, (3) universities that offer courses or papers concerning operating systems, and so forth. Users would then, at a glance, be able to select the type of source that best fits their need. Alternatively, in cases where users are willing to disambiguate their information need, type-tagging will allow search engines to improve their precision: site-type information can be used as a filter for advanced search features, allowing users to ask for specific types of results. These options are severely limited today. For example, users interested in search results of academic nature can today limit their requested results to the *.edu* domain. However, this leaves out research institutions that are not universities, not to mention the entire body of non-US universities. Another example would have users specifically request results from e-stores, when looking to buy a product.

3.2 Setting Crawling Policies

One of the major challenges with which search engines are faced is keeping their indices fresh. With the size of major indices reaching billions of pages and multimedia files, crawl cycles today may require several weeks to complete. Some sites, however, merit more frequent crawls. For example, the major search engines recrawl news sites daily, or even several times a day, in order to keep on top of developing stories [24]. In general, sites which serve as *content* sites might merit frequent crawls in order to enable the engines to keep track of current events, trends and other issues of volatile temporal nature.

Academic sites can provide another example for a functionality-dependent crawling policy. On one hand, university pages have been reported to change less frequently than pages on the ".com" top level domain [19]; on the other hand, there are specific times where university sites should definitely be recrawled, e.g., at the beginning of each term.

Another issue in Web crawling is ensuring the comprehensiveness of the crawl. To this effect, crawlers are often seeded with the top level of a Web directory [13]. Search engines capable of automatically recognizing directories will be able to seed their crawlers with a broader, more comprehensive list of seeds.

3.3 Refining Random Surfer Models

One of the best known link-based algorithms for ranking Web pages is PageRank [9], used by the search engine Google[5]. PageRank assigns each Web page p an importance score, which equals the probability that p is visited by a *random surfer* of the Web. The random surfer moves from Web page to Web page in an infinite sequence of steps, where each step is of one of the following two types:

Browsing Step: leave the current Web page by following one of the page's outlinks, where the followed outlink is chosen uniformly at random.

[5]http://www.google.com/

Random Jump: Choose a Web page q at random according to some distribution (e.g., uniformly at random), and jump to q.

A specific invocation of PageRank is defined by a parameter $d, 0 < d < 1$; each random step is of the first type with probability d, and of the second type with probability $1 - d$.

This basic random surfing model is rather simplistic. More elaborate random surfing models, which better represent reality, may yield "better" flavors of PageRank. In what follows, we suggest three refinements of PageRank that depend on the ability to detect the functionality of Web sites.

It is common practice among many link analysis algorithms to assign less weight to internal links (links connecting pages within a site) than to external links (links connecting pages of different sites). A simple heuristic to classify a link as internal or external involves examining the URLs of the pages on both sides of the link, along with, perhaps, the IP addresses of the hosts on which the pages reside. However, virtual hosting services such as Geocities [6] would be mistreated by such a heuristic, since URLs which seem to be part of the same site are actually logically independent, belonging to unrelated entities that are hosted, by chance, by the same service. The ability to automatically identify virtual hosting sites, along perhaps with the ability to distinguish between the different logical sites hosted therein, could help refine the random surfing model within such sites.

Another possible refinement of PageRank stems from the role that search engines play on the Web. We contend that links to any search engine should be treated in a realistic random surfing model as random jumps with probability (almost) 1, since search engines are simply gateways to (query dependent) random Web pages. As argued in the Introduction, a surfer visiting the home page of Google will almost surely submit a query and continue the browsing session from one (or more) of the search results. Accordingly, the surfer is statistically very unlikely to follow the link from Google's search page to Google's "about" page. In "vanilla" PageRank, Google's "about" page will receive a significant fraction of the PageRank of Google's search page, while in practice, surfers seldom visit that page.

Next, consider surfers at the top level of some Web hierarchy, e.g. Yahoo! [7]. Such surfers will either enter a query in a search box (resulting in a random jump, as argued above) or will patiently browse the directory structure to one of the category pages. As Yahoo! contains thousands of categories, its directory contains many branches, and so browsing paths to some categories are quite long. In PageRank's model of surfing, the probability of sustaining a long browsing sequence decreases exponentially, because of the random jumps that are performed probabilistically at each step. Thus, the contribution of the PageRank of Yahoo!'s home page to a category decreases exponentially in the category's depth. Perhaps a more appropriate model of browsing a hierarchy is to consider it as being "flat", with all categories linked directly from the home page. Then, surfers will either randomly jump away (by submitting search queries), or will reach their chosen category in a single modeled step.

[6] http://www.geocities.com/

[7] http://www.yahoo.com/

4. THE DATA SET

Our experiments are based on Web graph data obtained from AltaVista during the fall of 2001, for a crawl of over 500 million pages. We used AltaVista's connectivity server [6] to extract a 73-field record of aggregated structural statistics for all the pages on each host. These statistics were computed during a single linear pass on the data in the connectivity server, and thus the process is scalable. In order to explain the nature of these statistics, we define the following terms, exemplified by Figure 2.

Host: the host of a page is the host part of the page's URL.

Logical host: a string extracted from the host name using several heuristics and parsing rules. Often, this process produces a string which represents the organization or entity behind the host. For example, both hosts `www.yahoo.com` and `shopping.yahoo.com` are mapped to the logical entity `yahoo.com` . In general, there is a many-to-one mapping between hosts and logical hosts.

Level of page: number of slashes following the host name in the (expanded) url of the page. For example, the level of `http://www.foo.bar/` is 1, and that of `http://www.foo.bar/dir1/dir2/page.html` is 3. This notion is similar to the *depth* notions of [34, 8].

Internal link: a link connecting two pages of the same logical host.

External link: a link between two pages of different logical hosts, e.g., link 1 in Figure 2.

Uplink: an internal link $p \to q$, where the path of q is a proper prefix p's path. See curved links 2, 3 in Figure 2.

Downlink: an internal link $p \to q$, where the path of p is a proper prefix of the page of q. The unnumbered links in Figure 2 are downlinks.

Sidelink: an internal link $p \to q$, where both pages reside on the same directory. See dashed link 4 in Figure 2.

Crosslink: an internal link which is neither an uplink, a downlink nor a sidelink. See dotted links 5, 6 in Figure 2.

Leaf page: a page with no outgoing downlinks.

As noted above, we collected 73 statistics per host. The first three statistics are the total number of pages on the host, the total number of external inlinks to pages on the host, and the total number of external outlinks from the pages of the host. The next 64 fields are comprised of 8 blocks of 8 counters, as follows: let $P(h, \ell)$ denote the set of pages belonging to level ℓ of host h. For each level $\ell = 1, \ldots, 8$, we count the number of pages in level ℓ, $|P(h, \ell)|$, the number of external links to and from the pages of $P(h, \ell)$, the number of internal inlinks to the pages of $P(h, \ell)$, and the number of uplinks, downlinks, sidelinks and crosslinks from the pages of $P(h, \ell)$. The last 6 fields in each record are the average and standard deviation of the number of internal uplinks per page, the average and standard deviation of the number of internal crosslinks per page, and the average and standard deviation of the number of external outlinks per leaf page.

We then sorted the sites according to the sum of external inlinks and outlinks, and printed records for the top several hundred sites according to this criterion. Next, we manually examined 1108 sites. 783 of them were either pornographic, related to online gambling, or judged to be spam. Of the remaining 325 "virtuous" sites, 296 were classified into one of our 8 designated categories.

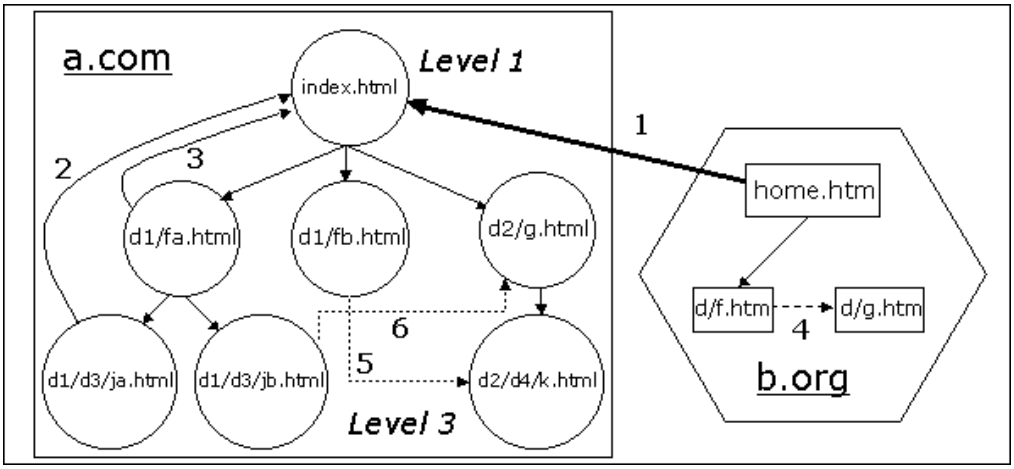

Figure 2: Two Web sites with internal and external links

4.1 Reliability of the Data

The data is extracted from AltaVista's connectivity server, which is derived from a crawl of the Web. Obviously, no crawl covers the entire Web and so our statistics are essentially derived from a large sample of the Web. Furthermore, the portion of the Web which is covered is highly dependent on the configuration of the crawler, e.g., its seed list and the manners by which it (1) prioritizes which sites to crawl, and (2) sets the depth to which it will crawl each site. Additionally, there are other factors which introduce noise into the collected statistics:

Restricted crawling by hosts: Some hosts restrict the crawling of their sites using the Robots Exclusion Protocol, which is respected by AltaVista. Other hosts (e.g., adult sites) limit access to major parts of their sites to paying account members. In either case, AltaVista's crawl will contain a small and non-representative sample of the site.

Dynamic-looking links: The connectivity server does not contain dynamic links. In practice, such links are filtered using syntactic rules, for example the exclusion of links containing '?' and '=' characters. The links of some sites contain such characters by default, even when linking to static pages. The outlink counts for such sites (both internal and external outlinks) will thus be wrongly reported in our data.

Our envisioned categorization is based on structural and connectivity properties of large and popular sites. The above factors may severely limit the amount of structural data that is available to us. We thus discarded from our analysis sites for which any of the following conditions held: (1) less than 40 pages were crawled, (2) pages belonging to only 1 or 2 levels were crawled, (3) less than 10 external inlinks were discovered, and (4) less than 10 external outlinks were found. This filtering lowered the number of classification-eligible sites from 296 to 202, as shown in Table 1.

5. FEATURE SELECTION

We derived 16 features from our 73 raw statistics per host by carefully examining the values for a sample of these sites.

CONTENT	CORP	DIR	ESTORE
49	35	22	17
PORTAL	SEARCH	UNIV	VHOST
23	8	23	25

Table 1: Number of classification-eligible sites, by category

Section 5.1 lists the features, along with the site types that achieve extreme values for each one. The features are discussed in Section 5.2.

5.1 List of Features

Features F1-F3 are based on the distribution of pages into the site's levels. Features F4-F13 are based on the external links to and from the site, and feature F14-F16 are based on the internal linkage patterns within the site.

F1-Average level of page: this feature is a measure of the shape of the site's tree. Deep sites, such as directories and many content sites, attain high values for this feature. Shallow sites, such as virtual hosts, attain low values.

F2-Percentage of pages in most populated level: most virtually hosted pages tend to be on the same level of the hosting site. In contrast, the university sites we examined had a much smoother distribution of page depths, and no single level usually held more than 45% of the pages.

F3-Top level page expansion ratio: the number of pages in the second level, divided by the number of pages in the top level ($|P(h,2)|/|P(h,1)|$). Search engines and portals exhibited low ratios, while universities and virtual hosting services usually attained high ratios.

F4-Inlinks per page: this feature measures the "mass density" of the site. Search engines, which are small sites with many inlinks, typically attain high values for this feature.

F5-Outlinks per page: measures the *emission* per page. Directories attain high values for this feature.

F6-Outlinks per inlink: the ratio between the push and the pull of each site. This ratio in search engines and corporate sites is usually very low.

F7-Top-level inlink portion: this feature measures the fraction of all inlinks to a site that link to pages on its top level. While some corporations have nearly all of their in-

links pointing to the top level, less than 30% of the inlinks to universities and virtual hosting sites we examined were to the top level.

F8-Outlinks per leaf page: leaf pages of directories, which are mostly category pages, usually contain many external outlinks. On the other hand, leaf pages of e-stores are usually product pages, and thus contain few outlinks.

F9-Average level of inlinks: here we average the level which attracts external inlinks. Corporations and search engines usually attract inlinks to their top levels, while virtual hosting sites and universities attract deeper inlinks to their sites.

F10-Average level of outlinks: portals and directories tend to have many outlinks originating from their deeper levels.

F11-The difference between F10 and F9: virtual hosting sites typically have the majority of both their inlinks and their outlinks associated with the top level of the individual hosted sites. Therefore, the difference between the average levels of the inlinks and outlinks will be small.

F12-Percentage of inlinks to most popular level: this feature examines the height of the tallest entry in the histogram of inlinks by levels.

F13-Percentage of outlinks from most emitting level: likewise, for the distribution of outlinks by levels.

The internal-linkage based features, F14-F16, measure how tightly or loosely knit is the design of the site, and how much browsing flexibility it supports.

F14-Crosslinks per page: many content corporate sites attain large values for this feature, while the examined universities and virtual host sites exhibited low values.

F15-Robustness: in the spirit of [8], this feature is defined as the ratio between the structural, navigational links of a site and the the semantic links of a site. We heuristically consider internal uplinks, downlinks and sidelinks and navigational, and internal crosslinks as semantical. We sum the number of navigational links in levels $2-4$, and divide by the number of semantic links in the same three levels. Low values, as were attained by content sites and portals, are indicative of robustness.

F16-Top level internal inlinks per page on site: this feature counts the number of internal inlinks received by pages on the site's top level, divided by the total number of pages in the site. The resulting ratio is an indication of the strength of connection of each page to the top level of the site. Virtual hosts have a weak connection, while corporate sites tend to have a "home" link on each and every page.

5.2 Discussion of Features

Some of our 8 designated categories, such as virtual hosting sites and directories, attained extreme values for many of the above features. Consequently, we expect the identification of those site types to be relatively easy. On the other hand, content sites and E-stores had relatively few typical characteristics, and so should be difficult to identify. For most features, sites of these two categories ranged over the entire spectrum of observed values.

Note that webmasters, by designing a site, control the features that are based on its external outlinks, internal linkage and page-level distribution. In contrast, the features that are based on the external inlinks to a site are indicative of the manner in which the rest of the Web views the sites, and are not controlled by its webmaster.

Most of the features do not rely on the sheer number of incoming links, and thus are not biased towards popular and well established sites. The two exceptions are the *F4-inlinks per page* and *F6-outlinks/inlinks* features, where well established sites may behave differently than new sites, regardless of their functionality. This was not a major concern in this study, though, since most sites we examined in our experiments were indeed well established.

As mentioned earlier, these features were chosen by a heuristic process. We do not claim that they are optimal in any sense. Section 8 presents several ideas for future improvements of the feature extraction and selection process.

6. RESULTS

This section describes the results of some of the experiments we conducted while constructing the Sonar classifier – an automatic tool that classifies Web sites into one of the eight pre-defined categories, using the 16 features described in Section 5.

Our 202 classification-eligible tagged Web sites were used as a training set for building a decision-rule classifier using the *See5* package [37]. The output of *See5* is a set of rules that can be used to classify arbitrary sites into one of the pre-defined categories. Figure 3 shows one of the induced decision-rules. The number in the square bracket (0.739) is the confidence value that *See5* attributes to the rule.

In addition to producing a classifier, *See5* estimates the expected classification error for an un-seen example, using 20 fold cross-validation. According to this estimate, the average expected error of the classifier is 45.5%, i.e., on average about 55% of new unseen sites will be classified correctly. This is clearly much better than a random classifier whose expected accuracy is 12.5%, and as such is an indication that the roles of Web sites are indeed reflected in the proposed features. Nevertheless, these results are not precise enough for a fully automated system.

In order to better understand the strengths and weaknesses of the classifier, we further analyzed these results. Table 2 presents the confusion matrix of the classifier, i.e., the distribution of categories assigned by the classifier to the sites belonging to each given category. Examining Table 2, we see that most of the VHOST and DIR sites were classified correctly. The classification of UNIV sites was quite precise as well, with 4 of the 7 mistakes classified as VHOST and CORP sites. This is due to university sites being a hybrid of CORP and VHOST sites, as explained in Section 2. Portals were mainly confused as content sites, which is predictable due to the nature of portals, as many of them

Rule 1:

crosslinks_per_page	$\leq$	2.996589	&&
log(inlinks_per_page)	$\leq$	8.434424	&&
log(outlinks_per_page)	$\leq$	2.621998	&&
F11	$\leq$	0.510355	&&
F2	$>$	0.4067201	
	$\Longrightarrow$	class VHOST	[0.739]

Figure 3: One of the decision-rules induced by *See5*.

Category	C1	C2	C3	C4	C5	C6	C7	C8
C1. CONTENT	**23**	5	7	1	5	3	3	2
C2. CORP	7	**21**	3	1	1	3	0	0
C3. DIR	0	2	**19**	0	0	0	1	0
C4. ESTORE	6	4	2	**1**	2	0	0	2
C5. PORTAL	9	2	2	1	**8**	0	0	1
C6. SEARCH	1	4	0	0	2	**1**	0	0
C7. UNIV	1	3	0	1	1	0	**16**	1
C8. VHOST	2	0	0	0	1	0	0	**22**

Table 2: Confusion matrix of the decision-rules classifier.

Recall	0.11	0.22	0.37	0.83	1.00
Precision	0.909	0.822	0.787	0.641	0.590

Table 3: Precision@recall of the Bayesian classifier.

provide CONTENT functionality. The opposite mistake, of tagging CONTENT as PORTAL, was also common. Other often-confused pairs include CORP with both CONTENT and SEARCH, although by and large, CORP sites were classified well. The classifier had difficulties in classifying ESTORE and SEARCH sites. ESTORES were most often confused with CONTENT, CORP and DIR sites, as they indeed share many aspects of those site types. It seems that the structural features used by the classifier are not powerful enough to distinguish between those categories. The failure in classifying search engines is surprising, since on the intuitive level, many aspects of search engines are quite unique. The results may be due to the small number of tagged SEARCH samples (8).

In a second experiment, we built a Bayesian classifier based on the same 16 features. The average expected error of the Bayesian classifier for the same set of sites was 41.0%, a small improvement over the rule-based classifier. Its success on the the CONTENT, CORP, PORTAL, UNIV and VHOST categories was practically identical to that of the rule-based classifier, while the SEARCH and ESTORE categories were identified better (at the cost of a decrease in precision for the DIR category).

The Bayesian classifier associates confidence levels to its decisions. We used these figures to measure the precision of the Bayesian classifier at different recall levels (shown in Table 3). We sorted the classified sites in descending order according to the confidence in their associated class(es), and measured the precision at different recall levels. *Recall at level* n is the defined as $\frac{n}{202}$, and *precision at level* n is defined as the ratio of correct decisions at the top n samples in the sorted order of the sites. The results of Table 3 indicate that when the classifier is confident in its own decision, it is usually correct. In other words, a conservative classification that only tags sites upon high confidence levels, leaving some sites untagged, would err significantly less than a process which must classify each and every sample. In Section 8 we discuss some possible approaches for improving the overall accuracy.

6.1 Identifying Link Spammers

Search engine spamming can be loosely described as the creation of Web content (pages and interconnecting links) with the intention of manipulating the search engine rankings of some pages, rather than for delivering content or facilitating human browsing [18]. With the growing emphasis on link-based ranking factors, and the growing economic ramifications of achieving top-10 rankings in search engines, *link spamming* has proliferated. This is defined in [14] as the creation of links "for reasons other than merit", namely to gain unjustifiably high search engine rankings. Spamming has grown to the extent that it severely threatens the quality of search engine rankings. Accordingly, commercial search engines are in search of measures to detect spammers [25].

Although not an original goal of this research, we noticed that our connectivity features allow the identification of link spammers. We used the set of 16 connectivity features to represent sites as vectors in a vector space. By clustering the entire set of 1100 sites using the cosine between vectors as a similarity measure, we were able to detect 183 sites in 31 clusters, where each cluster appears to be a spam ring. We predict that by applying the same technique (without some of the features based on external inlinks), we would be able to detect (legitimate) mirror sites as well. We conclude that in addition to site classification, the set of connectivity features suggested in this work can be used for site representation and for measuring similarity between sites.

7. RELATED WORK

The methodology of this research involves collecting comprehensive data on the connectivity characteristics of Web sites, deriving features from the observed data, learning the typical values that Web sites of certain function exhibit for each feature, and discerning the functionality of each site on the Web. Our work overlaps with three well-established areas of research: the study of hypertext, theme-based categorization of Web pages, and the study of the Web's graph. In the following, we briefly survey these fields.

7.1 Characteristics of Hypertexts

Hypertext predates the advent of the World Wide Web, and so do studies of the characteristics of authored hypertexts. When hypertext systems began to emerge, they were used to present highly structured information (reference books, manuals, etc.) in a flexible computer format which supported browsing. In order to preserve coherence it has been suggested very early on that some general linking characteristics should be observed. Botafogo et al. [8] provided authors of hypertexts with tools and metrics (based on the link structure of the hypertexts) to analyze the hierarchical structure of their documents during the authoring phase. Bernstein [5] defined a vocabulary of common structural patterns found in hypertexts, and argued that understanding such patterns will aid progress in designing (authoring) and analyzing hypertexts. Another comprehensive discussion of hypertext design patterns can be found in [22]. The observations presented in the current study amplify the notion that there exists already a latent taxonomy with which authors choose to display and interconnect their hypertexts. Such connectivity patterns shape and determine our navigational experience of a site. Those patterns also define the site's functionality as a structured information entity (be it a corporate site, a portal, or a virtual host).

Viewed from 30,000 feet, the goals and methodology of our research bears resemblance to the work of Pirolli et al [34]. There, the authors used a set of features to categorize Web *pages* into several functional categories, such as organizational and personal home pages, index and source pages,

and reference and content pages[8]. The features of each page included connectivity-based quantities such as its number of inlinks and outlinks, as well as non structural features such as the number of times it was requested by users in a given time period and the textual similarity between its contents and that of the pages to which it links. While [34] focused on the roles that individual pages serve in a site or in a community of thematically related pages, our research focuses on the roles that sites serve on the Internet as a whole.

7.2 Thematic Categorization of Web pages

Connectivity and structural analyses, in many forms, have been applied to the problems of finding thematically-related Web pages and sites, clustering pages of similar themes, classifying Web pages according to some ontology, and automatically generating descriptions of Web pages. We briefly survey the main techniques that proved useful in these efforts.

Co-citation based techniques [26, 38] have been shown to be useful for finding communities of thematically-related Web pages and sites. Pitkow and Pirolli applied co-citation algorithms as early as 1996 for clustering pages within a site [35]. In separate works, Weiss et al. [39] and Modha and Spangler [32] combined term-based factors with link-based factors when clustering Web pages. Dean and Henzinger [16] adapted Kleinberg's HITS algorithm [27] for identifying pages that are thematically similar to a sample page. Ruhl et al. applied the same approach to the host graph of the Web, in order to find thematically related sites [36].

Many works have identified anchor-text as a major contributor to classification and categorization efforts of Web pages [21, 23]. The effectiveness of anchor text is attributed to the fact that it often describes the target page much better than the text on the target page itself [31, 2, 15].

Flake et al. used Max-flow techniques to identify Web communities [20]. A community was defined there as a set of pages such that for each set member, the number of links it has to other members is at least as large as the number of links it has to non-members. Chakrabarti et al. [12] applied machine learning techniques for classifying Web pages. Their main observation was that when classifying page p, knowledge of the classes to which p's hyperlinked neighbors belong can greatly aid the classification.

7.3 Structural Properties of the Web

A great deal of research has been devoted to understanding the dynamic, ever changing structure of the Web's graph. These efforts revealed many fascinating topological properties of the Web, and gave rise to models that try to explain the manner in which this vibrant gigantic graph evolves.

Kumar et al. [30], in their work on identifying emerging Web communities, found that the Web contains many small complete bipartite subgraphs, which they called *bipartite cores*. These cores are essentially small interconnected sets of hubs and authorities, that will potentially grow to be the heart of future topical communities. They also confirmed the earlier findings of Barabasi and Albert [4] that the in- and out-degrees of Web pages follow power-law distributions. In two follow-up works [28, 29], the authors argued that Erdös-Renyi, $\mathcal{G}_{n,p}$ random graphs [7] would explain neither the degree distributions nor the large observed number of bipartite cores. They then proposed evolutionary models of random graphs that explain these characteristics of the Web. Recently, Pandurangan et al. [33] expanded these models to also support the observed distribution of PageRank scores on the Web.

Large scale connectivity properties of the Web were mined by Broder et al. in [11], using crawls of over 200 million pages. They reported that the Web, as a directed graph, has the following "bow-tie" structure: roughly a quarter of the Web's pages form a strongly connected component (SCC). An additional quarter of the Web includes pages from which directed paths lead to the SCC ("IN"), and yet another quarter includes pages that are accessible from the SCC ("OUT"). The remaining quarter of pages form "tendrils" [9] and small disconnected components. They reported the directed diameter of the SCC, and estimated many statistics regarding the lengths of shortest (directed and undirected) paths between Web pages. Recently, Dill et al. [17] have mined the structural properties of large thematically coherent slices of the Web's graph. They found the underlying structure to be similar across many of their tested graphs, and termed this behavior as the "fractal" nature of the Web.

In separate works, Adamic and Huberman [1] and Ruhl et al. [36] analyzed connectivity characteristics of sites, along with the distribution of the number of pages on sites. They found that in many aspects, the Web's graph behaves similarly at both site and page granularities, and so seems to be devoid of scale. In light of their findings, both works proposed evolutionary models for the manner in which pages (with links) are added to sites over time.

Chakrabarti et al. [13] used a 482-class topic taxonomy to classify about 16 million pages, and explored intra-topic and extra-topic linkage patterns. They also examined the mixture of topics visited while performing random walks on the Web's graph.

8. CONCLUSIONS AND FUTURE WORK

This paper tapped connectivity and structural data alone for the purpose of determining site functionality. After identifying a set of 16 structural and connectivity based features, we applied two classification algorithms to a set of 202 sites of eight functionality categories. The algorithms correctly classified between 54.5% and 59% of the sites. In particular, we were able to identify university sites, Web directories and virtual hosting services with precision exceeding 67%. Additionally, we reported preliminary results of identifying link spammers by finding clusters of sites with exceedingly similar structural fingerprints.

In practice, site functionality can be determined not only by connectivity patterns, but rather by content analysis as well. E-stores, which our classifiers had trouble identifying, may be identified by references to "shopping cart" on many pages. In particular, some site types can be easily identified by simple properties of their URLs. For example, American university sites commonly belong to the ".edu" top level domain, while many international universities belong to ".ac.*" domains. Our structure-based methodology can certainly be augmented by considering content information, thereby improving the classification, reducing its errors and enabling the classification of smaller sites, where the structural patterns are less pronounced. Note, however, that

[8]Some of these functions are closely related to the concepts of hubs and authorities, introduced in [27].

[9]Pages from which members of "OUT" are accessible, or which are accessible from pages of "IN".

content analysis is necessarily language dependent, whereas the Sonar system is indifferent to language issues.

The following research directions are left for future work:

Improving feature extraction and selection. There are several ideas which might yield more informative features. One example is normalizing statistics by the length of the pages, rather than by the number of pages: "links per kilobyte" may turn out to be a more distinctive measure than "links per page". Another direction is to better separate the navigational links of a site from its semantic links. One possible approach is to detect the design templates that are dominant in each site's pages, and then rule template links as navigational while treating the other links as semantical [3].

Identifying additional site types. Naturally, the eight categories handled in this paper do not cover the entire range of site functionalities found on the Web. During the course of this research we encountered several additional site categories. One example are *download sites*, sites that the rest of the Web views primarily as venues from which software can be downloaded. The sites themselves, however, may also serve as corporate sites for the firms that offer the software. One of the most notable examples of this phenomenon is `www.winzip.com`. The site is a bona-fide corporate site, but that is not the perception of the average Web page creator: the patterns by which the rest of the Web links to the site are heavily biased towards the pages from which the Winzip product can be obtained. In some respects, the inlink patterns of download sites resembles that of search engines: the sites provide primarily a single, specific *service* (access to a certain product). Note that larger corporate sites which offer a wide range of downloadable products (such as `www.microsoft.com`) attain inlink-signatures of regular corporate sites. Another type of sites whose properties were not investigated here are sites of non-profit organizations and institutions, which characterize many of the .gov and .org sites. It would be interesting to see the how the properties of such sites relate to those of commercially motivated corporate sites. Smaller site types which have attracted much interest lately are *blogs* and personal sites. We leave the investigation of these (and other) site types for future work.

Acknowledgments

We thank Farzin Maghoul and Glenn Carrol from AltaVista for helpful discussions on this topic, and for providing us with the connectivity data for our experiments.

9. REFERENCES

[1] L. A. Adamic and B. A. Huberman. The web's hidden order. *Communications of the ACM*, 44(9), September 2001.

[2] E. Amitay. Using common hypertext links to identify the best phrasal description of target web documents. In *Proc of the SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web, Melbourne, Australia*, 1998.

[3] Z. Bar-Yossef and S. Rajagopalan. Template detection via data mining and its applications. In *Proceedings of the 11th International WWW Conference*, pages 580–591, Honolulu, Hawaii, USA, 2002.

[4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.

[5] M. Bernstein. Patterns of hypertext. In *Proc. 9th ACM Conference on Hypertext and Hypermedia*, pages 21–29, 1998.

[6] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The connectivity server: Fast access to linkage information on the web. In *7th International World Wide Web Conference*, 1998.

[7] B. Bollobás. *Random Graphs*. Academic Press, 1985.

[8] R. Botafogo, E. Rivlin, and B. Shneiderman. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180, April 1992.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Proc. 7th International WWW Conference*, 1998.

[10] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2), 2002.

[11] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Proc. 9th International WWW Conference*, 2000.

[12] S. Chakrabarti, B. E. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of SIGMOD-98, ACM International Conference on Management of Data*, pages 307–318, Seattle, US, 1998.

[13] S. Chakrabarti, M. M. Joshi, K. Punera, and D. M. Pennock. The structure of broad topics on the web. In *Proc. 11th International World Wide Web Conference (WWW2002)*, 2002.

[14] B. D. Davison. Recognizing nepotistic links on the web. Technical Report WS-00-01, Artificial Intelligence for Web Search, 2000.

[15] B. D. Davison. Topical locality in the web. In *Proc. 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000)*, 2000.

[16] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. *Proc. 8'th International World Wide Web Conference*, 1999.

[17] S. Dill, R. Kumar, K. S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2(3):205–223, August 2002.

[18] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proc. 10th International World Wide Web Conference*, pages 613–622, May 2001.

[19] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proc. 12th International World Wide Web Conference (WWW2003), Budapest, Hungary*, 2003. To appear.

[20] G. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 2000.

[21] J. Fürnkranz. Using links for classifying web-pages. Technical Report TR-OEFAI-98-29, Austrian Research Institute for Artificial Intelligence, 1998.

[22] D. M. Germán and D. D. Cowan. Towards a unified catalog of hypermedia design patterns. In *Proc. 33rd Hawaii International Conference on System Sciences*, 2000.

[23] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using web structure for classifying and describing web pages. In *Proceedings of the 11th International WWW Conference*, pages 562–569, Honolulu, Hawaii, USA, 2002.

[24] M. Henzinger, B.-W. Chang, B. Milch, and S. Brin. Query-free news search. In *Proc. 12th International World Wide Web Conference (WWW2003)*, May 2003. To appear.

[25] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2), 2002.

[26] M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[27] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:5:604–632, 1999.

[28] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: Measurements, models and methods. *Proceedings of the Fifth International Computing and Combinatorics Conference*, 1999.

[29] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. S. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 41nd Annual Symposium on Foundations of Computer Science (FOCS 2000), Redondo Beach, California*, 2000.

[30] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Proc. 8th International WWW Conference*, 1999.

[31] O. A. McBryan. Genvl and wwww: Tools for taming the web. In *Proc First International World Wide Web Conference, Geneva, Switzerland*, May 1994.

[32] D. S. Modha and W. S. S. W.S. Clustering hypertext with applications to web searching. In *Proc. of the ACM Hypertext 2000 Conference, San Antonio, TX*, 2000.

[33] G. Pandurangan, P. Raghavan, and E. Upfal. Using pagerank to characterize web structure. In *Proc. 8th Annual International Computing and Combinatorics Conference*, 2002.

[34] P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear: Extracting usable structures from the web. *Proc. ACM SIGCHI Conference on Human Factors in Computing*, 1996.

[35] J. Pitkow and P. Pirolli. Life, death, and lawfulness on the electronic frontier. In *Proc. of the Conference on Human Factors in Computing Systems CHI'97*, 1997.

[36] M. Ruhl, K. Bharat, B.-W. Chang, and M. Henzinger. Who links to whom: Mining linkage between web sites. In *IEEE International Conference on Data Mining (ICDM)*, 2001.

[37] RuleQuest Research. Data Mining Tools See5 and C5.0. http://www.rulequest.com/see5-info.html.

[38] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. American Soc. Info. Sci.*, 24:265–269, 1973.

[39] R. Weiss, B. Vélez, M. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, and D. Gifford. Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. *Proc. 7th ACM Conference on Hypertext*, 1996.