Explaining and Predicting Web Page Status

Gautam Pant
School of Accounting and Information Systems, The University of Utah, Salt Lake City, UT 84112,
gautam.pant@business.utah.edu

Padmini Srinivasan
School of Library Information Science and Department of Management Science, The University of Iowa, Iowa City, IA 52242,
padmini-srinivasan@uiowa.edu

The World Wide Web has become a key intermediary between producers and consumers of information. While the producers strive to increase visibility for their products, services, socio-political messages, etc., the consumers struggle to find reliable and relevant information. Web linkage structure has been exploited by contemporary search engines to decrease the search cost for consumers while usually also rewarding the producers of higher status Web pages. In addition to influencing visibility and accessibility, in-links, as marks of recognition, accord status to a Web page. In this paper we show how Web page status may be determined at least in part by page location and topic specificity. Moreover, we observe that the “philanthropic” contributions of a Web page, specifically contributions of information brokerage function, influence other pages into generating in-links. The observations are made in the presence of quality, domain, and topical effects. Interestingly, all of these features influencing status are “local” to a given Web page and within the control of the owner/author of the page. Since the linkage structure of the Web has affects on browsing, crawling, and search engine ranking, our results have implications for content managers, search engine marketers, and search engines.

Key words: web search, search engine marketing, web mining, hyperlinks

History:

1. Introduction

Visibility on the Web is critical to a wide variety of organizations, be it for-profit, non-profit, governmental, political or others. Through their Web sites these organizations disseminate information for potential consumers. The information disseminated is intended to achieve one or more goals such as increased sales, improved brand awareness, consumer education, or to spread socio-political messages. Regardless of the goal, organizations want to deliver relevant information to potential consumers. Search engines provide a simple and popular interface between the producers and the
consumers of information sifting through billions of documents or Web pages. Hence, the quest for greater online visibility has lead to a multi-billion dollar search engine advertising and marketing industry. Estimated average CPC (cost-per-click) for keywords such as “plasma hdtv,” “top business schools,” and “prolife bumper stickers” can be in excess of $1-per-click if the corresponding advertisers want to consistently appear among the top 3 sponsored results on Google.\(^1\) These example keywords also indicate the variety of entities or organizations interested in Web visibility. Interestingly, organizations are ready to pay for sponsored results despite studies (VanBoskirk 2006, iProspect 2004) indicating that consumers pay more attention to organic (non-sponsored) portions of search results and are less likely to trust and click on sponsored results. Hence, from the point of view of user trust, greater visibility, and reduced advertising costs, information producers have a strong incentive to improve their presence among the organic results. However, the presence and rank of a Web page within the organic results greatly depends on the links to it from other Web pages (aka in-links). Barring link spamming techniques (Henzinger et al. 2002), which are being actively identified and penalized,\(^2\) Web linkage structure (i.e., how Web pages link to each other) develops over time thus creating a steep entry barrier for new pages (Chakrabarti et al. 2005).

Considerable attention has been paid in previous research to Web site design from the perspective of usability (Palmer 2002, Agarwal and Venkatesh 2002, Gupta et al. 2007). However, Web site usability becomes relevant only after a Web site is found by the user. Whether or not a consumer would find a relevant Web site depends on its visibility. A research question that has received limited systematic attention is: Why do some Web pages have higher “status” in terms of attracting in-links that determine its visibility? Research explorations in this direction have the possibility of creating validated best practices and guidelines for content managers, search engine marketers, and search engine designers. For content managers, such guidelines would allow consideration of Web site visibility in their usability focused designs. For search engine marketers involved in search engine optimization such studies could provide an opportunity to step back and take a long term

\(^1\) data retrieved from Adsense Keyword Tool on 03/24/2007

\(^2\) http://www.google.com/webmasters/guidelines.html
approach towards optimizing Web sites instead of a largely short-term focus on reverse engineering search engine algorithms. Search engine designers too can benefit from such research by utilizing predictive models that estimate the potential in-links that a new page may receive over time. Such measurement will allow them to decide on ranks of new pages that may have as yet gathered little or no in-links. A predictive model of potential in-links can be specially beneficial for vertical search engines that cater to niche interests such as “oncology.” Such search engines gather and index pages on a specific topic. This allows them to decrease the search space and hence possibly increase the relevancy and/or freshness of results. However, unlike general purpose search engines, vertical search engines restrict their collection to only those pages that appear related to the topic of interest and hence do not have the more global picture of in-links received by the pages in their index. A predictive model that can help vertical search engines estimate this global picture can benefit their ranking mechanisms. In this paper, we present a first attempt at answering the above mentioned question regarding status of a Web page. We begin by drawing upon more abstract notions in sociology.

Two key notions that have been of interest to sociologists for a long time are power and status. Within the network-exchange perspective, power is the potential to gain resources based on one’s position in the social network (Markovsky et al. 1998). Status, the esteem and honor in which a person is held, results when the individual is perceived as capable or motivated to contribute to the common good. A third notion, particularly relevant in our context, is influence. Influence uses persuasion, information, and advice (Mokken and Stokman 1976) and high status individuals are often perceived as influential because of their competence and their regard for achieving group goals.

So how does all this relate to the Web? The connection is evident when we view the Web as a sociological phenomenon with participants making choices on what links to pursue, what pages to read, and especially what links to build from their own Web pages. Page links and the pathways they define are critical in supporting Web applications. A page with no in-links is socially isolated and will not be reached by those unaware of its existence. As the number of links to a page increases,
so do the individual probabilities of it being reached, fetched, indexed, retrieved, read and its own out-links followed. In fact we may say that as the number of in-links to a page increases so does its status. Thus Web search engines often use in-links or a function thereof to estimate page importance (Brin and Page 1998). Analogously, number of citations is often part of a measure of impact for scholarly journals. Observe also Web pages with large numbers of in-links being called “authorities” (Kleinberg 1999). From a sociological perspective our research question becomes: how is it that some pages garner higher status than others?

Unfortunately, spamming (Henzinger et al. 2002) occurs wherein in-links to a page are manipulated to improve its search engine ranking. Our focus is on more ‘legitimate’ links. In general, since linking is a fairly democratic activity, power as described earlier cannot explain differences in page status. But can we gain some insights by exploring the notion of influence with regards to a Web page? Observe that a link is an acknowledgment of value by the author of the source page to the target page. In sociological terms we may say that the target page has influenced the source page author to a level above a threshold for generating such an acknowledgment.\(^3\) Thus in seeking to answer our earlier question about differences in status one approach could be to ask: what factors make one page more influential than another? If we understand the ways in which a page may influence authors of future pages then we are closer to understanding and possibly shaping page status in legitimate ways.

We explore the research question by defining select page characteristics that may influence page status. The selected characteristics are represented by quantifiable and “local” features of the page that are within the control of the author(s) of the page. With the help of a linear model, we then try to understand the effect and significance of the characteristics on Web page status. We perform the study with two data sets of 122,012 and 82,068 Web pages respectively across 14 broad topics. The linear model provides us with a broad understanding of direction, magnitude, and significance of the individual page characteristics’ effect on page status. However, in an application where the

\(^3\) Note that we do not qualify whether the influence is positive or negative. An example of the latter is where the source page refutes ideas in the target page. These are further refinements on the notion of page influence.
accurate prediction of page status is of primary importance, non-linear models may be better suited despite their potentially opaque nature. This motivates additional analysis with non-linear models, specifically with neural networks and decision trees. We find that the non-linear models do provide significantly better predictions of page status.

2. Related Research

2.1. Web-site Usability

Several studies have been performed on Web site usability. Nielsen (Nielsen 1993) describes several approaches to evaluating software usability. Palmer (2002) uses a questionnaire, completed by a panel of 35 users, to quantify the notion of Web site usability and to identify some Web site attributes that affect usability. Agarwal and Venkatesh (Agarwal and Venkatesh 2002) apply a set of Microsoft Usability Guidelines (Keeker 1997) to measure Web site usability. These studies are based on less than a hundred or a few hundred Web sites, typically of large companies, and their focus is on Web site design parameters that affect usability. More recently, Gupta et al. (2007) propose using user access log data to re-link Web pages on a given Web site to improve its navigability (a dimension of usability). As noted earlier Web site usability becomes relevant only after the Web site is identified or discovered. The focus of the current study is on characteristics influencing Web page visibility, characteristics that increase the likelihood of discovering and then using a Web page. In addition, our study is based on more than a hundred thousand Web pages from a variety of topics.

2.2. Hyperlinks as Status Indicators

A paradigm shift in Web search was introduced in the late 1990s with the use of hyperlink structure to measure the centrality of Web pages. Two influential algorithms were proposed for centrality measurement—Klienberg’s HITS (Kleinberg 1999) and Brin and Page’s PageRank (Brin and Page 1998). Many extensions have been proposed as for example to calculate topic specific PageRank (Haveliwala 2003) and to use links as trust indicators and calculate topical TrustRanks (Wu et al. 2006). The common assumption in these algorithms and their variants is that a page that makes a link to another page confers some notion of “authority” to the latter. Hence, status of a Web page
can be determined by the links that lead to it. In this paper, we do not suggest any new measure or application of Web page status but rather explore the general page characteristics that lead to higher status.

2.3. Web Page Quality

An obvious potential influential characteristic of a Web page is its quality. However, page quality in terms of its design and content is hard to define and then to measure. A large number of guidelines offer opinions based on personal observations and common practices in Web site design. Features such as font size, number of images, presence of section headers, use of color, and presence of sound files are discussed in this context. Researchers have also emphasized the need for more rigorous methods for studying page quality (e.g., (Ivory et al. 2001, Spyridakis et al. 2005)). There have been recent studies that respond both from an empirical viewpoint and using a theoretical framework. For example, quality from the user’s perspective was found to be influenced by specific content, content quality, appearance and technical adequacy (Aladwani and Palvia 2002). Singh et al. (2005) employ an informational model to understand user reactions to home pages while Geissler et al. (Geissler et al. 2001) apply the theory of stimulus complexity to explore user interest. In this study on predicting Web page status we control for page quality by basing our experiments on a sample of pages from a Web directory called the Open Directory Project (ODP). Since this directory is human edited, we expect our pool of pages to be of acceptable quality. In addition we identify a few indirect measures of page quality that are objective in nature and easily scalable to more than hundred thousand pages.

2.4. Web Linkage Dynamics

Any study of page status via page influence is complicated by the fact that the two are also mutually reinforcing. Higher status pages (those with more in-links) over time have a greater chance of being read and therefore a greater chance of influencing future pages into generating in-links. This “success begets success” model (Barabási and Albert 1999) is mirrored in the finding that search
engines, due to their reliance on in-links, direct their users to pages that are already of high status (Hindman et al. 2003). This creates a steep entry barrier for new pages (Chakrabarti et al. 2005). For example, in a study of Web data collected over 7 months, Cho and Roy (Cho and Roy 2004) observed that the top 20% of the pages ranked by the number of incoming links obtained 70% of the new links after 7 months, while the bottom 60% of the pages obtained almost no new incoming links during that time. Theoretically they predicted that it takes 60 times longer for a new page to become popular when search engines are heavily used as compared to a random-surfer model.

There is also a sizeable literature on the nature of Web topology and how it evolves over time through preferential attachment (e.g., (Barabási and Albert 1999, Broder et al. 2000)). Our focus is not on the interaction effects between status and influence that cascade over time. Instead our goal is to identify status influencing characteristics of pages that can be exploited by authors. In other words, we assume that the cascading “success begets success” model applies to all pages and focus on identifying characteristics of a page that are likely to make it influential over time.

2.5. Search Engine Marketing

Currently, a growing search engine optimization (SEO) industry focuses on providing advice and content management to Web site owners with the purpose of improving their site’s rankings on popular search engines (Moran and Hunt 2006). The techniques used by the SEO industry range from spamming (link or content) to more legitimate content changes based on constant (likely non-rigorous) attempts at reverse engineering search engine rankings. For example, a recent article\(^5\) talks about tapping the long tail of searches by considering more than just the top search terms when describing a page. Techniques designed have also to adapt to changes in the ranking mechanisms used by search engines. In contrast to these efforts which rely on understanding how a particular search engine ranks, our paper is on studying the relationships between certain general properties of Web pages and its status.

Finally, as indicated in the introduction section the notions of power, status and influence have been of interest to sociologists for a long time. Factors that influence status of an individual within

Table 1  URLs extracted from various categories of the ODP.

<table>
<thead>
<tr>
<th>URL</th>
<th>Category</th>
<th>Top-level Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.xmlpitstop.com/">http://www.xmlpitstop.com/</a></td>
<td>Computers/Data_Formats/Markup_Languages/XML/Resources</td>
<td>Computers</td>
</tr>
<tr>
<td><a href="http://www.nhlbi.nih.gov/health/prof/sleep/dray_drv.htm">http://www.nhlbi.nih.gov/health/prof/sleep/dray_drv.htm</a></td>
<td>Health/Conditions_and_Diseases/Sleep_Disorders</td>
<td>Health</td>
</tr>
<tr>
<td><a href="http://www.leadershipservices.com/">http://www.leadershipservices.com/</a></td>
<td>Business/Management/Leadership</td>
<td>Business</td>
</tr>
<tr>
<td><a href="http://www.hep.net/sites/directories.html">http://www.hep.net/sites/directories.html</a></td>
<td>Science/Physics/Particle/Research_Groups</td>
<td>Science</td>
</tr>
</tbody>
</table>

the context of a social network are a major focus of study (Markovsky et al. 1998, Mokken and Stokman 1976). Our goal is analogous, i.e., to determine ways in which a Web page may influence future readers into generating in-links towards itself.

3. Page Characteristics

We are interested in page characteristics that are local in nature (i.e., that do not need other Web pages for their operationalization) and have the potential to affect page status. As noted earlier, page quality is an obvious choice for potential page characteristic. However, due to difficulties in defining and measuring page quality across various domains we control for it by using pages from ODP, a manually edited Web directory. ODP arranges millions of select Web pages in a topical hierarchy. At the time of writing ODP has hierarchically arranged more than 4.8 million Web sites and has over 75 thousand editors. Google’s directory service is based on ODP data. Table 1 shows a few examples of ODP URLs and the categories under which they were placed.

The amount of information provided on the page which we call Information Volume (IV) may affect page status. It may be that pages with more information are more influential or perhaps pages fail to influence future authors by being too wordy. Neilsen and Loranger (2006) suggest that text on Web pages should be “short, scannable, and approachable.” They also argue that “Web is a user-directed medium, where people adopt information seeking strategies to save time” (Neilsen and Loranger 2006). Recommendations favoring shorter Web pages for content that is meant to be browsed and read online can also be found in Web style guides (Lynch and Horton
2001). We estimate $IV$ by the number of words (including numbers) in the Web page. Common words or stop words such as “and,” “or,” “of” are not included in the estimation of $IV$. Another characteristic of interest is Information Location (IL). The URL of a page located deep within a Web server will be long, less rememberable, and therefore possibly less influential. The idiom ‘hiding one’s light under a bushel’ aptly describes this feature. There is some evidence (Upstill et al. 2003) that home pages tend to have higher PageRank (Brin and Page 1998) than other pages within a site. Search engine marketing practitioners suggest that URLs of pages should be “link-friendly” by being short and rememberable (Moran and Hunt 2006). Web usability experts note that complex (long) URLs “hurt both usability and search engine optimization” (Neilsen and Loranger 2006). Clearly, the value of smaller and hence easily recognizable URLs has not been lost on Web designers. Search engine marketing practitioners have also noted that a page located deep (e.g., http://www.some_domain.com/press/public/month/top.html) within a site might rank lower (among search engine results) as opposed to the same page being located closer to the home page (e.g., http://www.some_domain.com/press/top.html) (Moran and Hunt 2006). Thus we choose to formally study the effect of this characteristic on page status. We estimate $IL$ by the number of sub-directories from the host name to the actual file name in the URL of the Web page.

A third feature of interest is Information Specificity (IS). More general topics may appeal to a larger audience and therefore have greater influence. Alternately, readers may be attracted by depth in the discussion which is more likely in pages dealing in specific topics. Microsoft Usability Guidelines (Keeker 1997) also consider the depth and breadth (i.e., specificity) of topics covered by a Web site from a usability perspective. In particular the guidelines argue that the content must be of interest to more than a niche community. Here, we test specificity effects on the status of Web pages. We estimate $IS$ by the depth of the Web page in the ODP hierarchy (see Figure 1). Both location and specificity potentially determine, in a sense, the accessibility and outreach of a page respectively. Greater accessibility and outreach can be expected to increase influence and hence ultimately the status of a Web page.
A fourth Web page feature that could influence status is one that we call Information Brokerage (IB). This feature derives from recent sociology research by Lovaglia et al. (2003) exploring a number of propositions regarding power, status and influence. The one that attracts our attention is that philanthropy increases influence and status. By voluntarily contributing to a common good, individuals with power can reduce perceptions of greed and thereby raise their status. This suggests an almost parallel proposition concerning Web page status. Namely, voluntary contributions and services made for the common good of the Web community increases the level of influence and therefore the status of a Web page.

Web pages offer, often for free, many kinds of services such as access to a specialized database or the current weather. A more subtle service, however, is that of being a broker or gateway for information. This occurs when a Web page provides select out-links to other relevant information. The philanthropic aspect is apparent because it takes time and effort to find, evaluate, select and build appropriate out-links. Moreover, these out-links are generally offered at no charge (barring the pay for access sites) as sites seldom charge exit fees for following their out-links. Using the sociologist’s viewpoint, which sees influence as using persuasion, information and advice, we see parallels. In addition to the Web page content, the author may influence the reader via the information and advice provided about other relevant pages available on the Web. Implicitly the author may persuade the reader who seeks more information to follow these lucrative paths. Consistent with these sociological theories we conjecture that the greater the number of out-links from a Web page, an estimate of a property that we call Information Brokering, the higher its ability to influence readers into linking to it from their own pages. As a related note, it is widely recognized that review articles in scholarly journals tend to attract higher than normal number of citations. These review articles aggregate previous research within a given topic and hence make references to a large body of work (analogous to large number of out-links). Hence, the value of information brokering has been observed in contexts other than the Web. We estimate $IB$ of a page by the number of other pages it links to or the out-link count. We do not count a link from a page to itself for $IB$. 
As noted earlier defining a direct measure of page quality across various topics on the Web is difficult. However, we identify a few variables that may be considered surrogates for several dimensions of page quality. The choice of these quality-based variables is guided by their objectivity and scalability. The first of these variables is one that we call Content Inertia (CI). A Web page that remains mostly unchanged over time can be considered to have high content inertia. Usability experts Neilsen and Loranger (2006) suggest against keeping outdated content. The Web is a dynamic medium of communication and a page that updates and changes over time can be expected to hold users' interest better than a page that does not change. Greater user interest in turn can be expected to increase the influence and hence status of a page. We estimate $CI$
by taking snapshots of a given page at two different times and computing the textual similarity between the two snapshots. The textual similarity metric we use is the term-frequency-based cosine similarity (Salton and McGill 1983) that varies between 0 and 1 where 1 is indicative of highest content inertia. Web pages that are either no longer updated or are archival in nature can be expected to have a \( CI \) value of 1. We explore a second variable that is likely related to page quality called \textbf{Links Ratio (LR)}. This is measured as a ratio of number of links on a page to the number of words on the page (hence, \( LR = IB/IV \)). Metrics similar to link ratio have been used to differentiate between content (lower \( LR \)) and navigation (higher \( LR \)) pages (Cooley et al. 1999). We expect \( IB \) identified earlier to positively effect status. However, the presence of too many links without adequate accompanying textual content may convey a negative signal about the quality of the page and hence its status. Similarly, the presence of mostly quantitative information on a page may limit user interest and hence page status. This observation leads us to the third variable likely related with quality which is \textbf{Quantitative Ratio (QR)}. \( QR \) is measured as the ratio of the number of numbers appearing on a page and the number of words (i.e., \( IV \)) appearing on the page. Hence, if 25% of a words on a page are numbers, then \( QR \) for the page is 0.25.

4. Controlling for Other Page-level Variables

The seven main variables of interest (\( IV, IL, IS, IB, CI, LR, \) and \( QR \)) are easy to compute, local in nature, and importantly may be manipulated by Web content designers. There is also enough evidence, at least anecdotally, to make them reasonable candidates in our study on explaining and predicting Web page status. In order to robustly estimate the explanatory power of these seven main variables we also include the following page-level variables that could possibly also influence page status. However, the extra variables are different in that they cannot be directly manipulated by page designers. Our seven main variables of interest previously identified must be able to explain page status above and beyond these ‘control’ variables.

The first control variable is \textbf{Domain Traffic (DT)} which measures the traffic (visits by users) to a given domain. Would a page located in a high traffic domain such as \texttt{yahoo.com} automatically
gain greater visibility and hence greater status? Or, do high traffic domains such as yahoo.com act mainly as conduits to a large number of pages satisfying numerous niche interests? In other words does the domain traffic and the consequent visibility on such high traffic domains spread thinly across millions of hosted pages? We estimate DT through reach data from Alexa Web Information Service (AWIS), one of the Web services provided by Amazon.\(^6\) AWIS measures reach for a given domain as the proportion of a panel of users who visit that domain. The proportion is provided as “per million users” and is computed on a daily basis. Hence, a reach of 198 for the domain utah.edu means that for every one million users in the AWIS panel 198 visited some page on utah.edu site. We use a 3 month average of reach provided by AWIS for different domains as an estimate of their corresponding DT.

The other control variable that we identify is Topical Share (TS). As noted earlier, ODP is a large directory with millions of listed URLs and tens of thousands of editors. Hence, we expect that a URL listed in an ODP category (see Table 1 for examples) that has many other URLs listed has a lower share of the topical subspace and hence is facing greater competition for status than a URL that is listed in an ODP category with only few other URLs. The idea of topical share is a simplistic analogue to the concept of market concentration in organization theory (Tirole 1988). We estimate TS as the inverse of the size of an ODP category, where size of an ODP category is the number of URLs appearing in it. For example the URL http://www.xmlpitstop.com/ appears in the ODP category ‘Computers/Data_Formats/Markup_Languages/XML/Resources’ which lists 45 URLs in total. Hence, the TS for http://www.xmlpitstop.com/ is 1/45 or 0.022. We expect TS, which increases with decreasing competition for status, to effect page status positively.

Finally we estimate status (S) of a Web page by its number of in-links. This number is obtained by querying Google through its Web Search API.\(^7\) Previous literature (Amento et al. 2000, Upstill et al. 2003) has shown very high correlation (> 0.75) between different measures of Web page status such as number of in-links, PageRank (Brin and Page 1998), and Authority (Kleinberg 1999) scores.

\(^6\)aws.amazon.com

\(^7\)http://code.google.com/apis/soapsearch/reference.html
Since reasonable computation of PageRank or Authority scores for tens of thousands of pages will require downloading at least several million pages, we choose the number of in-links measure for Web page status.

Figure 1 illustrates the computation of the described variables. Variables $S$, $IB$, $IV$, and $DT$ are known to have skewed distributions with a small number of Web pages having large values (Broder et al. 2000). Hence, we apply a log transformation ($\ln(1 + x)$) on each of our 10 variables ($x$) and use the transformed variables for the rest of our study.

The data used in our analysis was collected at two time points (2006 and 2007). We used two Web services (Google and Amazon), multi-threaded Web crawlers, and HTML/XML parsers to obtain and process the data.

5. Data Sets

Our goal is to explore the effect on page status of seven page variables introduced in section 3. Our study includes two control variables described in section 4 and is done using a sample of Web pages drawn from ODP. We obtained a version of ODP from February 2006 and another from January 2002 (ODP provides downloadable archives of its directory listings). We removed the ODP topics with single character sub-categories (e.g. “A,” “B”) and ones that include the words, “Adult,” “Regional,” “International,” or “World.” These topics generally contain either alphabetical listings, spam pages, redundant topics, or non-English text. We then identified the URLs (along with the ODP topics they are listed under) that are in the 2006 as well as the 2002 archive. While not guaranteed, the pages of these URLs can be expected to be fairly stable, having existed for more than 4 years (the half-life\(^8\) of Web pages is estimated to be around 2 years (Koehler 2002)). Since status can only be built over time we do not want to mix in our analysis for example, pages that are two days old with others available for over two years. We picked a random sample of 150,000 URLs and attempted to download up to first 1000 KB (HTML) of each page in June 2006. We filtered out pages with frames due to ambiguity in defining their $IB$ variable, and the need for

\(^8\) Time required by half of the pages in a sample to disappear from the Web.
additional downloading to obtain all their content. We tried to obtain the number of in-links or 
in-links count for the downloaded pages within the 10 days after downloading the pages. With
these steps we obtain an effective sample of 122,012 ODP pages and we call it Data Set 2006. We
can compute six of the seven main variables and one of the two control variables for each of the
pages in the Data Set 2006. We cannot compute CI since it requires snapshot of the page at two
time-points before obtaining the in-links data. We also do not have the traffic data from 2006 to
compute DT. Hence, any analysis with Data Set 2006 will have to omit these two variables.

More than a year later, in July 2007, we once again attempted to download the same pages that
we had downloaded in June 2006. As before, we filtered out pages with frames. We obtained in-links
count for each URL in September 2007 which was about 2 months after downloading the pages.
Considering successful downloads, this gave us a sample of 113,676 ODP pages with corresponding
in-links data. For each of the pages, we tried to obtain the reach data (3-month average) for its
second-level domain name from AWIS Web service as explained in Section 4. For example, for
the URL http://www.microsoft.com/office/visio/ we obtain a reach value of 64675 for its
second-level domain, i.e. microsoft.com. The 3-month average reach data for several domains was
not available and this shrunk our effective sample to 82,068 ODP pages and we call this Data
Set 2007. We can obtain all of the variables described in Section 3 and Section 4 for each of the
URLs (and corresponding pages) in Data Set 2007. Hence, this data set will be used for most of
the following analysis.

6. Modeling Status

Our objective is to model status (variable S) in terms of some or all of the identified variables. We
first divide our Data Set 2007 into two subsets—modeling set (66%) and validation set (34%). We
use the modeling set to build a model of page status and test the resulting model on the validation
set (holdout data). We first model S using all of the seven main variables as follows:

\[ S = \beta_0 + \beta_1 \cdot IS + \beta_2 \cdot IL + \beta_3 \cdot IB + \beta_4 \cdot IV + \beta_5 \cdot CI + \beta_6 \cdot LR + \beta_7 \cdot QR + \epsilon \]  

(1)
Table 2: Regression Results for Main Variables: p-value < 0.001 for all $\beta$s

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\beta}$±std. err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant ($\beta_0$)</td>
<td>3.884±0.065</td>
</tr>
<tr>
<td>IS ($\beta_1$)</td>
<td>-0.567±0.028</td>
</tr>
<tr>
<td>IL ($\beta_2$)</td>
<td>-1.080±0.015</td>
</tr>
<tr>
<td>IB ($\beta_3$)</td>
<td>0.529±0.009</td>
</tr>
<tr>
<td>IV ($\beta_4$)</td>
<td>-0.125±0.008</td>
</tr>
<tr>
<td>CI ($\beta_5$)</td>
<td>-1.612±0.048</td>
</tr>
<tr>
<td>LR ($\beta_6$)</td>
<td>-0.698±0.057</td>
</tr>
<tr>
<td>QR ($\beta_7$)</td>
<td>-0.977±0.106</td>
</tr>
</tbody>
</table>

where $\beta_1, \beta_2, \ldots, \beta_7$, are the coefficients, $\beta_0$ is a constant, and $\epsilon$ is the error term. Using ordinary least square regression with the modeling set data we derive the values of model coefficients ($\beta$s) as shown in Table 2. All coefficients are significant with $p < 0.001$ and the model has an adjusted R-square of 0.224.

The first significant effect is that the more specific pages have lower status and vice versa. This suggests that general topics may attract larger audiences and hence garner higher status. The coefficient for $IL$, as expected, is negative and significant indicating that location is important. If submerged deep within the directory structure of a Web server even a high quality page’s URL is unlikely to be noticed and/or remembered resulting in lower status over time. The positive and significant coefficient for $IB$ means that as the number of out-links from a page increases so does its status. This supports our conjecture that the philanthropic information brokerage service offered to the larger Web community influences status positively. The negative coefficients of $IV$ and $CI$ indicate the merits of brevity and updating, respectively, on Web pages and are in concurrence with the suggestions of usability experts. However, the coefficient of $LR$ suggests the need for a healthy balance of links and text on a Web page. In other words, an increase in $IB$ without an adequate amount of accompanying text ($IV$) affects page status negatively. Similarly, a high proportion of quantitative information ($QR$) is associated with a negative signal with respect to page status.

We use the regression model to predict the value of S for pages in the validation set. We then compute the correlation between the predicted and observed values of S. Interestingly, this correlation is found to be 0.464, which is significant at the 0.01 level.
We now add the control variables, \( DT \) and \( TS \), to the linear model described in Equation 1.

The main purpose of adding these control variables is to check if our main variables retain their sign and significance and hence explain the page status beyond the two control variables. Note that unlike our main variables the control variables cannot be manipulated by the page designer. Hence, we now model \( S \) as:

\[
S = \beta_0 + \beta_1 \cdot IS + \beta_2 \cdot IL + \beta_3 \cdot IB + \beta_4 \cdot IV + \beta_5 \cdot CI + \beta_6 \cdot LR + \beta_7 \cdot QR + \beta_8 \cdot DT + \beta_9 \cdot TS + \epsilon \quad (2)
\]

Again, using ordinary least square regression with the modeling set data we derive the values of model coefficients (\( \beta \)s) as shown in Table 3. The model has an adjusted R-square of 0.229 (slightly higher than the adjusted R-square with only the main variables). We observe that all of the main variables retain their sign and significance. Therefore, our analysis for these variables is robust.

We note that the control variable \( DT \) negatively affects page status. Hence, popular domains such as \textit{yahoo.com} do spread their traffic thinly across millions of hosted pages and hence the status of most pages is not expected to be high. In other words, while the entire domain benefits from the aggregated traffic over millions of pages, many pages do not gain high status. Also, as expected, the higher the \( TS \) of a page within its topic (represented by the corresponding ODP category) the higher its status. This suggests that pages benefit when the competition for status is lower in their topical subspace.

We use the regression model with the main and control variables to predict the value of \( S \) for pages in the validation set. We then compute the correlation between the predicted and observed values of \( S \). This correlation is found to be 0.470, which is again significant at the 0.01 level.

We also apply the linear model in Equation 2 (without the \( CI \) and \( DT \) variables) to Data Set 2006 and find the coefficients to be consistent in sign and significance with those shown in Tables 2 and 3. Similar to Data Set 2007, the Data set 2006 is also split into modeling and validation sets and the regression is performed using the modeling set data. Table 4 shows the model coefficients. The model has an adjusted R-square of 0.208. We use the regression model to predict the value
of $S$ for pages in the validation set. We then compute the correlation between the predicted and observed values of $S$. This correlation is found to be 0.457, significant at the 0.01 level.

Since 2004 we have conducted several preliminary explorations that concentrated on just three of the current seven main variables, namely, $IS$, $IB$, and $IL$. These explorations used data sets of different sizes derived from ODP archives from a different time period. We note that the coefficients of $IS$, $IB$, and $IL$ derived in these preliminary exploration were consistent in the sign and significance with those presented in Tables 2, 3, and 4. Hence, the current results with Data Set 2007 and Data Set 2006 as well as the preliminary explorations confirm the robustness of our findings. All of the further analysis in this paper is based on Data Set 2007.

Based on our findings a good general strategy for hosting a Web site would be to provide some pages that link to several relevant resources and thereby act as information brokers. For example, a small company may provide pages with links to tutorials, white-papers, FAQs, or expert advice. Pages should be as general as possible within the limits of information relevance and quality. Also,
it is prudent to locate them conveniently with few sub-directories in their path. These strategies must be accompanied with a general aim toward quality such as by not flooding a page with links without an adequate amount of accompanying text. The status gained by appropriately shaping these local properties of Web pages may positively affect the outlook towards the entire site or even the organization that hosts them. Moreover, the additional in-links will likely improve the placement of the specific Web pages among the search engine results.

Another implication is for search engine designers. For example, Google relies, at least partially, on information derived from the in-links of a page to judge its (query-independent) status (Brin and Page 1998). However, some (Kleinberg 1999) have proposed the use of both out-links and in-links to obtain two different measures of page importance. The first is a “hub” score (a type of information brokerage measure) and the other is an “authority” score (a type of status measure). Our results for IB suggest that it may be unnecessary to have both scores since high status pages are likely to be potential hubs. Indeed, it may well be that Google is providing us with good hubs and authorities in its top results.

Much importance has been given to Web page authority measurements for ranking retrieved results (Brin and Page 1998, Kleinberg 1999) and building good collections. Page authority measurements rely on “global” information (such as in-links) derived through linkage structure on the Web. This is hard to gather (requiring exhaustive Web crawling) or as in the case of new Web pages, simply unavailable. Our results indicate that due to the sociological underpinnings of the Web the “global” information relating to a page may be estimated using properties of the “local” content (e.g., specificity, out-links, location) of the page. As noted earlier such an estimation would be particularly useful for vertical search engines.

7. Topical Effects

We have shown that the model described above is effective in explaining status in a generic sense, i.e., without concern for the particular topic of the page. We now explore the question as to whether there are differences in how the identified variables effect or predict page status across broad topics. The ODP data that makes up our collection has 14 top-level topics (see examples in Table 1).
The 14 top-level topics represent broad subject areas with significant semantic differences. Hence, by capturing topical cues, we may be able to improve upon our generic model for explaining and predicting page status. Specifically, we may be able to build topic-specific models that are allowed to capitalize on the nuances within each top-level topic. We explore this aspect by first splitting our modeling and validation data based on the 14 top-level topics of the ODP. We then obtained topic-specific models based on Equation 2 using the modeling data for each top-level topic. We found the resulting models for all of the top-level topics to be consistent with each other (and the generic model) in terms of the sign (i.e., direction) of the coefficients ($\beta$s) for the main variables. The only exception is ‘Games’ for which the coefficient of $QR$, unlike all other top-level topics, was positive. This means that the presence of a higher proportion of quantitative information on a ‘Games’ page has a positive effect on its status. We note that several ‘Games’ pages include statistics that may be of interest to the players. The significance of the coefficients is, however, not consistent across the top-level topics. In fact, the only variables whose coefficients are significant (at 0.05 level) across all top-level topics are $IS$, $IB$, $IL$, and $CI$. Hence, these four properties significantly determine page status not only at a generic level but also at a more specific topical level. Our observations for these variables based on the generic model also apply to the topic-specific models. Table 5 shows, for each top-level topic, the correlation between the predicted and observed values of status, $S$, using the generic or the topic-specific model. The correlations are based on validation set data for each top-level topic. For all top-level topics (except ‘News’) the topic-specific model yields higher correlation than the generic model. For the combined validation set, the topic-specific models produce a correlation of 0.555 which is 18% higher than that produced by the generic model (0.470). We note that the combined result for the topic-specific models can be viewed as that for an ensemble of models where an appropriate model (based on top-level topic) is used for each validation set URL. The key conclusion here is that, having knowledge of the broad topic of the page enhances our ability to estimate Web page status.

As noted earlier the coefficients of $IS$, $IL$, $IB$, and $CI$ are consistent in sign and significance across the 14 broad topics. However, there are significant differences between the magnitudes of the
Table 5 Correlation using the generic and the topic-specific models.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Generic Model</th>
<th>Topic-Specific Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>0.5316</td>
<td>0.5573</td>
</tr>
<tr>
<td>Business</td>
<td>0.4050</td>
<td>0.4254</td>
</tr>
<tr>
<td>Computers</td>
<td>0.4561</td>
<td>0.4637</td>
</tr>
<tr>
<td>Games</td>
<td>0.5169</td>
<td>0.5684</td>
</tr>
<tr>
<td>Health</td>
<td>0.4655</td>
<td>0.4805</td>
</tr>
<tr>
<td>Home</td>
<td>0.4538</td>
<td>0.5113</td>
</tr>
<tr>
<td>Kids and Teens</td>
<td>0.4631</td>
<td>0.4675</td>
</tr>
<tr>
<td>News</td>
<td>0.4394</td>
<td>0.3933</td>
</tr>
<tr>
<td>Recreation</td>
<td>0.5595</td>
<td>0.5955</td>
</tr>
<tr>
<td>Reference</td>
<td>0.5039</td>
<td>0.5258</td>
</tr>
<tr>
<td>Science</td>
<td>0.4765</td>
<td>0.5326</td>
</tr>
<tr>
<td>Shopping</td>
<td>0.4235</td>
<td>0.4284</td>
</tr>
<tr>
<td>Society</td>
<td>0.5449</td>
<td>0.5718</td>
</tr>
<tr>
<td>Sports</td>
<td>0.5525</td>
<td>0.5556</td>
</tr>
</tbody>
</table>

coefficients for different top-level topics. This suggests that the degree to which the these independent variables effect Web page status can vary significantly with top-level topic. For example, the top-level topic with the smallest magnitude for the $IS$ coefficient is ‘Computers’ ($-0.447 \pm 0.123$) while ‘News’ has the largest ($-3.845 \pm 0.933$). ‘Computers’ includes URLs leading to information that is technical in nature. Whereas in addition to URLs of various news sources, ‘News’ has a large number of URLs leading to specific news items. The more specific news items are less likely to attract (and retain) attention in terms of in-links when compared to the more specific technical articles. Thus specificity in ‘News’ strongly and negatively influences status. Whereas, with ‘Computers’ although status is significantly and negatively related to specificity, this is to a much lesser degree when compared with ‘News.’ Similarly, ‘Reference’ has the highest value for $IB$ ($0.734 \pm 0.051$). The ‘Reference’ topic is typified by pages such as those providing access to bibliographies, encyclopedias, and almanacs. Out-links in reference articles attract in-links — this in a sense underlines the “referential” nature of such writings. Thus although the topic-specific models are generally consistent with our generic model of status, topic-specific analysis of coefficients offers special insights on how better to exploit topical features during Web page design.
8. Non-linear Models

The linear regression based analysis helped us understand the direction and significance of the seven main page characteristics and 2 control variables both across and specific to 14 broad topics. However, if the emphasis is placed on predictive power, some non-linear models such as neural networks and decision trees could out-perform linear regression. The emphasis on predictive power would be typical in an application such as a general purpose or vertical search engine where an automated system is expected to judge the status of millions of Web pages. In that scenario, the question about “why” some Web pages more easily gain status is secondary to “how” to predict page status well. This emphasis on predictive power for some applications motivates the following analysis.

Neural networks and decision trees are amongst the most popular data mining techniques. A three-layer feedforward neural network can approximate any bounded continuous function to arbitrary precision (Hornik et al. 1989). We use the implementation of neural network with backpropagation learning that is included in the Weka machine learning software (Witten and Frank 2005). The neural network is used to predict the value of variable $S$, while the nine variables identified earlier are provided as inputs to the network. Hence, the neural network has nine input nodes and one output node where the output node provides a continuous value corresponding to the (predicted) $S$.

We first tune the neural network by varying two of its important parameters—the number of hidden layer nodes ($H$) and the learning rate ($L$). While the former controls the complexity of the model, the latter affects the incremental changes made to the model during the learning process. For the purpose of this parameter tuning we divide the modeling data into a training (66%) and a tuning set (34%). Hence, we do not use the validation set while tuning the neural network. We train the neural network using the training set, and measure its performance using the tuning set. Our performance measurement is the correlation between the actual and predicted values of $S$ in the tuning set. To avoid large permutations of experiments, we first vary $H$ to find its best value while keeping the value of $L$ fixed. We then fix the $H$ to the best value found and vary $L$. For each
Figure 2  Correlation between predicted and actual $S$ (status): (a) for various values of $H$ (number of hidden layer nodes) (b) for various values of $L$ (learning rate).

value of $L$ and $H$, we train the neural network on the training set and use it to predict the value of $S$ for the tuning set. We experiment with the following values for $H$: 2,4,8,16,32,64,128,256,512 while keeping the learning rate fixed at 0.1. Figure 2 (a) shows the correlation between the predicted and the actual values of the variable $S$ (using the tuning set data) as $H$ varies. We find the best value for $H$ to be 128. We now fix the value of $H$ to 128 and vary the learning rate to be one of the following: 0.2,0.1,0.01,0.001,0.0001, 0.00001. Figure 2 (b) shows the variation in correlation between the predicted and actual values of $S$ (using the tuning set data) for various values of $L$. Based on the results shown in the Figure 2, we set the $H = 128$ and $L = 0.001$ for our further experiments.

Widely used decision tree algorithms such as C4.5 (Quinlan 1993a) are applied to classification problems (where the variable to be predicted is a set of classes). Since $S$ is a continuous variable, we chose the M5’ algorithm (Wang and Witten 1997) that can handle regression, and its implementation is available in Weka (Witten and Frank 2005). M5’ closely follows ideas of model tree (M5) that were suggested by Quinlan (Quinlan 1993b, Wang and Witten 1997). A model tree is similar to a decision tree where each node is a decision point based on the input variables. However, unlike a typical decision tree, where the leaves correspond to discrete classes, the leaves of the model tree are linear models that estimate the output variable. In other words a model tree provides a
piecewise linear function of input variables (nine variables identified earlier) to estimate the output variable \( S \) (Quinlan 1993b).

Figure 3 shows the performance of the three different techniques—linear regression, neural network, and M5'—for modeling the status of Web pages. The vertical axis shows the average correlation between the predicted and the actual S over 10 runs of 5-fold cross-validation. The 5-fold cross-validation is done by splitting the validation set into 5 equal parts and using 4 parts to train and the remaining part to test the performance (measure correlation between actual and predicted S). The process is repeated 10 times to gather 50 different correlation values for each technique. The average of the correlation values for different models are shown in the Figure 3. The neural network and M5' provide correlations between the predicted and actual values of \( S \) that are more than 32% and 36% higher than the linear regression respectively and the improvements are statistically significant (at 0.01 level of significance). Hence, non-linear models do provide better predictive power than the linear model. In fact the correlations provided by the non-linear models are high (>0.6) and hence valuable for predictive applications. On the other hand, the lucid nature of linear regression allows for a broad understanding of the implications of page characteristics.

![Figure 3](image_url)

**Figure 3** Average correlation between predicted and actual S (status) based on 10 runs of 5-fold cross-validation.

Using a process paralleling the one described in Section 7 we build (train) topic-specific models
using the neural network and the M5’ algorithms. Figure 4 shows performance of the topic-specific models using the three techniques on the fourteen top-level topics in ODP. Each technique uses the part of the modeling set corresponding to a given top-level topic to train or regress, and the part of the validation set corresponding to the same top-level topic to measure performance. Again we find that the non-linear models consistently outperform the linear models across all top-level topics. Hence, if either a generic or a topic-specific application requires high predictive power for estimating the status of Web pages, non-linear models such as neural networks are the appropriate choice.

![Figure 4](image)

**Figure 4** Correlation between predicted and actual S (status) on various broad topics using the three different modeling techniques.

In Section 7 we had observed that topic-specific linear models can provide a 18% higher correlation (between the predicted and actual values of S) than the generic linear model. In Figure 5 we compare and contrast the performance of generic and topic-specific models using each of the three techniques (linear, neural networks, and M5’). The models are built using the modeling set data and their performances are measured on the validation set. To measure performance we apply the appropriate topic-specific model for each page in the validation set. We note that again the topic-specific models outperform the generic model across all techniques. Hence, regardless
of whether the model is linear or non-linear, it can exploit topical knowledge to provide better predictions of Web page status. Both neural network and M5’ achieve high correlation of 0.659 and 0.665 respectively. However, unlike the 18% improvement observed for the linear regression, topic-specific models based on neural network and M5’ provide much smaller improvements of 5% and 1.7% respectively over their generic counterparts. This suggests that the generic non-linear models capture some topic level cues from the page attributes without requiring the explicit topic level partitions. Hence the extra knowledge about the topic is of lesser value for these models as compared to the simple linear model.

![Figure 5](image)

**Figure 5** Correlation between predicted and actual S (status) using general and topic-specific models.

9. **Discussion**

We now discuss the results of our study on page status along several dimensions. Some of our discussion involves further data collection and analysis. Our discussion is primarily focused on three variables ($IB$, $IL$, and $IS$) that significantly and consistently explain Web page status both in general and topic-specific contexts using both linear and non-linear models.

9.1. **Information Brokerage: Local Versus Remote Links**

Thus far we have measured $IB$ using the number of out-links from a given page. However, not all out-links are the same. Some of the out-links may lead to resources within the local site (in which
We differentiate between local and remote out-links by identifying the second-level domain of the outlink URL. As an example, the second-level domain of \texttt{http://www.dell.com/} is \texttt{dell.com}. If the domain of the outlink URL is the same as the domain of the current page, then the outlink is considered a local outlink, otherwise it is considered a remote outlink. We count the number of local and remote out-links on a page and use the values for two component variables of \( IB \), \( IB_l \) and \( IB_r \) respectively. The variables are transformed \((\ln(1+x))\) as before. The general \( IB \) variable is replaced in Equation 2 by the component variables giving us the following model of page status \( S \):

\[
S = \beta_0 + \beta_1 \cdot IS + \beta_2 \cdot IL + \beta_3 \cdot IB_l + \beta_4 \cdot IV + \beta_5 \cdot CI + \beta_6 \cdot LR + \beta_7 \cdot QR + \beta_8 \cdot DT + \beta_9 \cdot TS + \epsilon
\]

(3)

Table 6 shows the coefficient values obtained through regression on the modeling set. All coefficients are significant and the adjusted R-square is 0.235. We find that, consistent with \( IB \), both local and remote components of \( IB \) have positive and significant effect on page status. Hence, by providing links to local or remote resources a Web page can improve its status. In a manner similar to Figure 3, we measure the performance (through cross-validation) of the linear and non-linear models that use the two component \( IB \) variables instead of a single generic \( IB \) variable. The linear
model achieves an average correlation of $0.476 \pm 0.001$ while neural network and M5’ achieve average correlations of $0.633 \pm 0.002$ and $0.641 \pm 0.002$ respectively. We find that the linear and neural network models significantly outperform (with $p < 0.001$) their counterparts in Section 8 that used one generic $IB$ variable instead of the two component variables used here. For M5’ we do observe a better performance with component $IB$ variables, however, the improvement is not statistically significant as compared to using a single $IB$ variable. Hence, in general we find that splitting $IB$ into components (based on local vs. remote nature of out-links) improves the prediction of page status.

9.2. Information Location: Beyond Home Page Effect

As noted earlier there is some evidence (Upstill et al. 2003) that home pages tend to have higher PageRank (Brin and Page 1998) than other pages within a site. Typically, home pages correspond to $IL = 0$ (i.e., no sub-directories in the URL). Although, home pages of some sites may be located deeper ($IL > 0$) and many pages that are not home pages may be located at $IL = 0$. In fact, the definition of home page itself is at best subjective and vague (Craswell et al. 2001). However, we want to investigate if much of effect on status is purely due to $IL = 0$ pages (we call this the “home page effect”) versus $IL > 0$, or do higher levels of $IL$ continue to show significant impact on page status. Hence, we split our modeling data by $IL$ values, and compute the mean $S$ over all pages with the same $IL$. In addition we calculate $\pm 1$ standard error round the mean $S$. For statistical robustness, we compute mean $S$ and standard error for only those $IL$ values for which at least 20 pages are available. Hence, we do not include $IL > 2.08$ (i.e., pages with more than 7 sub-directories in their URL) in our analysis since there are only 4 pages with such values.

Figure 6 shows how mean $S$ varies with different $IL$ values. The first point from the left corresponds to pages with 0 sub-directories in their URL, the second point corresponds to pages with 1 sub-directory in their URL and so on. We find that the mean $S$ of pages with 3 sub-directories in their URL (4th point from the left on Figure 6) is significantly lower (at 0.01 level) than mean $S$ of pages with 2 sub-directories, which is turn is significantly lower than mean $S$ of pages with 1
sub-directory. Finally, the mean $S$ of pages with 1 sub-directory is significantly lower than the mean $S$ of pages with 0 sub-directories. As seen in Figure 6 the general trend for pages with 4 or more sub-directories is that the mean $S$ continues to decrease with increasing number of sub-directories (although the decrease is not monotonic). Hence, our analysis shows that the effect of $IL$ goes beyond the “home page effect” and that the increasing number of sub-directories in the URL are associated with significantly decreasing page status.

9.3. Information Specificity: An Automated Approach

Out of the 7 main variables considered in this study the only variable that was not computed from the page content or its URL is $IS$. In fact, our measurement of information specificity is overly dependent on the ODP hierarchy. Thus our research capitalizes on the intellectual effort of ODP’s editors resulting in a relatively reliable measurement of information specificity. However, in order to extend the measurement of information specificity to arbitrary pages on the Web, we would like to now explore a substitute for the current $IS$ metric that is not dependent on ODP hierarchy. With that in mind, we present preliminary investigation on estimating information specificity as a function of simple term weights. In order to estimate information specificity we first try to identify the characteristic terms in the page. We do this by representing the page in vector space using

![Figure 6](image-url)
the TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme (Salton and McGill 1983). We then choose the terms corresponding to the top-N components of the page vector as the characteristic terms. The average IDF of these characteristic terms is used as an estimate of information specificity ($IS'$) for the given page. We note that inverse document frequency (IDF) is considered a measure of term specificity (Robertson 2004). Hence, we use the average of the IDF's of the characteristic terms within a page as its information specificity measure. We use the following standard TF-IDF formulation:

$$w_{ij} = \left( \frac{tf_{ij}}{\max_{i' \in T_j} tf_{i'j}} \right) \cdot \left( \frac{\ln |E|}{df_i} \right)$$  \hspace{1cm} (4)

where $w_{ij}$ is the weight of term $i$ in page $j$, $tf_{ij}$ is the number of times term $i$ appears in page $j$, $T_j$ is the set of all terms that appear in page $j$, and $|E|$ is the size of the collection (in number of pages) used for computing the document frequencies. The document frequency of term $i$, $df_i$, is the number of pages (among $|E|$ pages) in which the term $i$ appears. Hence, a given page $j$ can be represented as a vector $\vec{w}_j = (w_{1j}, w_{2j}, \ldots, w_{ij}, \ldots, w_{nj})$ where $n$ is the number of unique terms in $j$. As mentioned earlier, we find the top-N components of this vector and use the corresponding terms as the characteristic terms for the page.

Based on Equation 4, IDF of term $i$ is the negative log of the probability of finding the term $i$ in a page within the collection of $|E|$ pages ($\ln \frac{|E|}{df_i} = - \ln \frac{df_i}{|E|}$) (Robertson 2004). Hence, a highly specific term can be expected to appear in very few Web pages (low probability) and therefore have a large value of IDF. We assume that a page that is characterized by highly specific terms has high information specificity. This forms the basis for our choice of using average IDF of characteristic terms for measuring the $IS'$ variable.

Some of the important terms describing a Web page are often found in its title or a heading at the top of the page. Hence, it may be worthwhile to concentrate on just the first few words of a page (instead of the entire page) for obtaining the characteristics terms. With that in mind, we also try variants of the above mentioned scheme for computing the $IS'$. The variants differ in
the extent of the Web page ($T$ words from the top) used for obtaining the characteristic terms. In order to quantify how well $IS'$ approximates the original $IS$ variable we measure correlation between the two. Unlike all of the previous variables, we do not transform $(\ln(1 + x))$ $IS'$ since it is already an average of log values (IDFs).

For the purpose of this preliminary exploration we set $N$ (number of characteristic terms) to 3. Also, $T$ is varied such that we use the first 5, 25, 125, and all words in a given page for obtaining the $N$ characteristic terms. The IDFs are computed using the document frequencies made available by the digital library project at UC Berkeley. These document frequencies are based on a collection of 49,602,191 Web pages. The large size of this collection should provide a good approximation for the true document frequencies of various words on the Web. We remove words with document frequencies less than 400 to avoid common misspellings and other noise terms. We also remove numbers, words containing numbers, and stopwords from pages.

Figure 7 shows the correlation between the $IS'$ and $IS$ for various values of $T$ using the modeling set. While all correlations are positive and statistically significant (with $p < 0.01$), they only show a weak relationship between $IS'$ and $IS$. Hence, $IS'$ is only a noisy approximation for $IS$. This is not completely unexpected given the simplistic nature of measuring information specificity from term weights. We choose the value of $T$ that shows the highest correlation with $IS$ (see Figure 7) for further analysis. In particular, we replace the $IS$ variable with $IS'$ ($T = 25$) in Equation 2 and perform the regression using the modeling set.

As shown in Table 7 all of the coefficients are statistically significant and the adjusted R-square is 0.230. We find that the alternate measure of information specificity ($IS'$) based on term weights has a negative and significant effect on page status. This is consistent with a similar effect observed for the original $IS$ measure in Section 6.

We use the generic linear and non-linear (neural network and M5') models learned from the modeling set to predict the values of $S$ for the validation set. The correlation between the predicted

---

9 http://elib.cs.berkeley.edu/docfreq/index.html
and actual values of $S$ is 0.473 for the linear model, while neural network and M5' achieve correlations of 0.625 and 0.647 respectively. All correlations are significant at 0.01 significance level. Hence, the new measurement of information specificity ($IS'$) is a good alternative to the original variable ($IS$).

### 9.4. Causation and Content Control

Our results suggest two broad implications. The first is that the variables identified in this work are good predictors of Web page status. In fact, with the use of non-linear models the variables provide predictions of page status that have remarkably high correlation ($>0.6$) with actual status. The second implication is that our main variables explain the status achieved by Web pages and hence
suggest causation. The first implication (predictability) is obvious based on the results, however, the second implication (causation) is harder to establish with observational data such as ours. Hence, similar to other research that are based on market or transactional data, we can only suggest causation. An alternate research design (i.e., controlled experiment) that may be considered for follow up research could be to somehow simulate the Web over time in a laboratory to test for causation. However, such a “simulated Web” could at best be a surrogate for an intranet and it would be hard to scale its findings to the public World Wide Web.

Our main variables indicate the potential degrees of freedom for content authors. In practice, a content author or manager would be under constraints while operating with those degrees. For example, departments or units within a firm may have access to only a sub-directory on the firm’s Web server. Hence, they may only reorganize content (to say lower $IL$) within the assigned sub-directory and would not have the luxury to place their content within the parent or grand-parent directories. Many other constraints based on pragmatic issues such as organization hierarchy, IT policy, and costs can be similarly identified. Hence, our results provide a set of guidelines whose application would depend on specific organizational context.

10. Conclusion

Gaining page status through greater in-links is important given a Web that is characterized by the “success begets success” syndrome. It is well known that new Web pages face a steep entry barrier to gaining status. We have presented a systematic study of page status and factors that influence it. The page properties that we have chosen for study are those that may be shaped by the page designer in order to positively influence status. We find that a page’s specificity, location, information brokerage, and its content’s inertia determine its status in both general and within specific topics. The key results from our study are as follows.

1. Using a regression model we find that a page is less likely to gain status if it is located deeper in a directory structure or if it is more specific in its content. Also, as its information brokerage function increases, i.e., as the number of out-links from the page increases so does its status. In
addition, as expected, indicators of higher page quality, positively effect status. These relationships are statistically significant and are observed in two data sets (2006 and 2007) and even in the presence of variables that control for domain traffic and topical competition. Moreover, the results for $IS$, $IB$, and $IL$ obtained from the two data sets are consistent with our preliminary explorations that used data from an earlier time period.

2. Topic-specific models are more capable of explaining and predicting status in different broad topics as compared to the generic regression model. Correlations between predicted and actual status scores are higher using the topic-specific models. The relationship between location, specificity, information brokerage, and content inertia on the one hand and status on the other are consistent with those observed with the generic model and with each of the 14 broad topics considered. Consistency is seen both in terms of sign and significance of coefficients.

3. Non-linear models are significantly better at predicting status compared with the linear regression models. In fact, non-linear models show high correlation ($> 0.6$) between predicted and actual values of page status ($S$) for the held-out validation set. Hence, non-linear models are effective and clearly preferable to linear models in applications where predictive power is of primary importance. However, the linear regression model due to its simplicity and transparency gives us a better understanding of Web page status as a phenomenon.

4. Separating the information brokerage function by local (same site) versus remote (another site) out-links gives us more powerful regression models that have better predictive power. The analysis also informs us that both local and remote out-links contribute positively to page status.

5. The effect of page location on status is seen beyond the “home page effect” (i.e., $IL = 0$ versus $IL > 0$). In other words, an increasing number of sub-directories in the URL is associated with significantly decreasing page status.

6. Finally, we proposed an alternate measure of specificity which is the average TF-IDF of the characteristic terms from the page. Unlike the original measure that relies on manual decisions by the editors of ODP, the alternate measure of specificity is fully automatic as it is based on page content. We find that the alternative measure shows a similar effect on page status as the original.
The main implications of the findings for content managers is the importance of providing some pages that appeal to a broader community and that are conveniently located on the Web server. Also, these pages are favored if they provide links to other relevant information and act as brokers of information. These implications are accompanied with an obvious suggestion for attention to quality such as regular page updates and to avoid flooding the page with links without adequate accompanying page text. The resulting increase in visibility of some pages could possibly spill over to increase the visibility of the entire site. The increased status of pages will likely also have direct effect on search engine marketing objectives for the organization that owns the site. The ranking of the pages with increased in-links can be expected to improve among the organic part of search engine results. Our findings also inform search engine designers that, in the absence of in-links data (as in the case of new pages), they may use non-linear predictive models that use local features of Web pages identified here to estimate their relative status. Such a predictive approach can be especially beneficial for vertical search engines that may not have the luxury to exhaustively crawl the Web on a continual basis to gather accurate in-links data.

While we control for the quality of pages by restricting our sample data to manually edited ODP and by including a few surrogate measures of quality in our analysis, we do not include a direct and integrated measure of quality in our analysis. This limitation of the current study could be addressed through a user panel that answers a quality-related questionnaire about each page. However, coming up with a reasonable quantitative measure of quality (from answers to a questionnaire) that may effect Web page status is an interesting research question in itself. In addition, such a study would be rather limited in the number of sample pages since obtaining answers from users for a hundred thousand pages would be infeasible in terms of time and cost involved.

In future work we will explore additional approaches to estimating page specificity as also additional perspectives on information brokerage. We will also like to further explore the role of text itself in determining status. For example, it would be interesting to determine if textually similar web pages acquire similar status scores.
To close, some key characteristics that influence status are within the control of the page designer. This has important implications in terms of building effective connections between information producers and consumers. Besides page designers, our results are of potential value to content managers, search engine marketers and search engine designers. A key general finding is that the important Web page properties are surrogates of properties that bring status and influence in human society—accessibility through the right location, outreach, and philanthropy.

Acknowledgments
The authors thank the anonymous reviewers and the editors for their helpful comments and suggestions.

References


iProspect. 2004. iprospect search engine user attitudes. iProspect.com.


