Working Paper Version
Published Version available at http://pubsonline.informs.org/doi/abs/10.1287/isre.1120.0457.
G. Pant, P. Srinivasan. Status Locality on the Web: Implications for Building Focused Collections. Information Systems Research, Vol. 24, No. 3, 2013.

Status Locality on the Web: Implications for Building Focused Collections

Gautam Pant
Management Sciences Department, The University of Iowa, Iowa City, IA 52242, gautam-pant@uiowa.edu
Padmini Srinivasan
Department of Computer Science, The University of Iowa, Iowa City, IA 52242, padmini-srinivasan@uiowa.edu

Topical locality on the Web is the notion that pages tend to link to other topically similar pages and that such similarity decays rapidly with link distance. This supports meaningful web browsing and searching by information consumers. It also allows topical web crawlers, programs that fetch pages by following hyperlinks, to harvest topical subsets of the Web for applications such as those in vertical search and business intelligence. We show that the Web exhibits another property that we call "status locality." It is based on the notion that pages tend to link to other pages of similar status (importance) and that this status similarity also decays rapidly with link distance. Analogous to topical locality, status locality may also be exploited by web crawlers. Collections built by such crawlers include pages that are both topically relevant and also important. This capability is crucial because of the large numbers of web pages addressing even niche topics. The challenge in exploiting status locality while crawling is that page importance (or *status*) is typically recognized through global measures computed by processing link data from billion of pages. In contrast, topical web crawlers depend on local information based on previously downloaded pages. We solve this problem by using methods developed previously that utilize local characteristics of pages to estimate their global status. This leads to the design of new crawlers, specifically of utility-biased crawlers guided by a Cobb-Douglas utility function. Our crawler experiments show that status and topicality of web collections present a tradeoff. An adaptive version of our utility-biased crawler dynamically modifies output elasticities of topicality and status to create web collections that maintain high average topicality. This can be done while simultaneously achieving significantly higher average status as compared to several benchmarks including a state-of-the-art topical crawler.

Key words: status locality, predictive models, topical crawlers, homophily
History:

1. Introduction

From the very inception of the hyperlink, its purpose was to connect related pieces of information contained in different documents thus creating a cohesive and traversable information "trail" (Bush 1945). This design supports the well-acknowledged property called *topical locality* on the Web which refers to the notion that pages tend to link to other topically similar pages and that such similarity decays rapidly with link distance (i.e., the number of links to be followed to navigate

from one page to another). There are at least three approaches by which this property has been demonstrated. One is through direct study: Davison (2000) has empirically shown that web pages linked to each other are much more likely to have similar content than random pages and Menczer (2004) has shown that such similarity exponentially decreases with link distance creating a strong topic localization effect. Second, the property is self-evident in the success of manual browsing (imagine browsing a web where pages are randomly connected, independent of content). Third, the property is reflected by the effectiveness of *topical web crawlers*, programs guided by topical criteria that traverse the Web and are used to build focused collections (Chakrabarti et al. 1999, Diligenti et al. 2000, Menczer et al. 2001, Pant and Srinivasan 2005). Specifically, this property allows crawlers to exploit cues from within downloaded pages to fetch more on-topic pages by selectively following hyperlinks. Thus topical locality represents a type of regularity that is part of the foundation required for search engines and higher level applications.

We propose the presence of a second fundamental regularity on the Web that we call *status locality* and that is analogous to topical locality. Status locality refers to the notion that pages tend to link to other pages of similar status (importance) and this similarity also decays rapidly with link distance. Our interest in this phenomenon was triggered by notions of *homophily* in Sociology where there is substantial support for two types of homophily – value homophily and status homophily (McPherson et al. 2001). Value homophily refers to the tendency of people or groups to connect when they share similar thoughts, ideologies, beliefs and attitudes. Status homophily refers to the tendency of people or groups to associate when they have similar social status. Value homophily which relies on the "content" so to say of individual preferences, strongly parallels topical locality. Analogously status homophily may also draw a parallel on the Web and therefore we should find that status locality also holds. We expect this since the Web is, generally speaking, an expression of society at large (e.g., web designers may have a propensity to link to pages of similar status). Based on this motivation we conjecture the presence of status locality on the Web and our first goal in this paper is to test the presence of this phenomenon.

Another significant contribution here is that we take crawler design research a major step forward. Thus far crawler algorithms (e.g., (Chakrabarti et al. 1999, Diligenti et al. 2000, Menczer et al. 2001, Pant and Srinivasan 2005)) have relied greatly on topical locality. After directly showing that status locality holds, we take the next logical step and design crawlers that exploit both topical and status locality properties.

In addition, recent work showing how to estimate status using page *local* properties (Pant and Srinivasan 2010) offers an efficient way to make run-time estimates of status (generally operationalized as number of in-links or PageRank) during the crawl for the growing collection of unvisited

URLs. This ability to estimate status using local properties combined with the Web's status locality property allows us to explore new and significantly different crawlers exploiting both topic and status regularities on the Web. Specifically, we propose new *utility-biased* crawlers and their adaptive versions. In the latter, for example, a user may specify a minimum topicality level desired and if the crawler reaches this level it could start to increase the emphasis on status and so on. It was not previously possible to build focused collections with this type of control over crawls. Note that in contrast to the prior work, which was solely on status estimation strategies (Pant and Srinivasan 2010), our status locality conjecture and the new utility-biased crawlers are unique to this paper. In summary our specific contributions here are as follows.

- Motivated by theories of homophily in Sociology we conjecture the presence of status locality as a property of the Web that is analogous to the widely-acknowledged topical locality property.
- We provide direct, empirical evidence supporting our conjecture and thereby make a theoretical contribution to the overall study of the Web.
- We propose an entirely new family of web crawlers, specifically utility-biased web crawlers, that takes advantage of both topical and status localities. The challenge here is that if the crawler pursues status alone we put topicality at risk and vice versa. Thus there is the important requirement of balance between the two.
- Addressing the balance between topicality and status we propose an advanced adaptive version of our utility-biased crawler where the user specifies the overall topical goals for the focused collection and status is maximized within that constraint. This offers a natural approach for balancing status and topicality.
- An obstacle to building crawlers that use status in the crawl logic is that one needs an efficient and effective approach for run-time estimation of page status during the crawl. We show that estimation methods developed in prior research using page local cues can overcome this obstacle.

Why might crawlers that consider status locality be of interest? The user, be it an individual or an application, is best served with the most topical *and* highest status web pages available. Given the Web's vastness and its coverage of almost every topic imaginable, the current challenge is not so much about retrieving sufficient topical documents. Instead, it is to retrieve topical sets that also excel in an orthogonal dimension such as status. To illustrate, there are millions (if not more) of Web pages relating to proteomics.[1] Crawlers launched to create a topical collection must focus on capturing the important subsets of this topic. We present a novel and pragmatic web crawling framework capable of doing so by balancing the dual concerns of topicality and page status. We show this can be done in real-time by harnessing local regularities including the phenomenon of status locality that we propose.

[1] a search for the word "proteomics" alone on Google returns more than 4 million pages.

In the next section we cover related work. Then we present our status locality conjecture and direct empirical evidence in support. In Section 4 we present our utility-biased crawlers and the methods used to estimate both topicality and status. In Section 5 we present our test bed, performance metrics and the results for our utility-biased crawlers. A new adaptive version of our crawler is presented and tested in Section 6. Section 7 considers additional features of our crawlers (robustness, diversity and reliability) and also presents comparisons against a more competitive benchmark. We make our conclusions in the last section.

2. Related Work

2.1. Topical Locality

Topical locality on the Web has been well-documented (Davison 2000, Menczer 2004). Davison (2000) has shown that pages are significantly more similar in topic to pages they link to than random or other nearby pages. Menczer (2004) quantifies the rapid decline in topical similarity between pages with increasing link distance using an exponential function. It is hard to imagine browsing a web where such topical locality does not hold (e.g., if pages were randomly linked, independent of content). This property has been a key component of the theoretical underpinnings of web crawler design since the earliest topical crawlers. The success of such crawlers in building topically focused web collections is in itself key evidence in support of the existence of the topical locality property on the Web. Topical locality has been found to be short-range, i.e., it decays rapidly with link distance (Menczer 2004). Hence it is imperative to apply topical cues of the closest pages (ideally, those that are 1-link away) to score URLs of pages that are yet to be downloaded.

A key point is that the topical locality property does not suggest that pages on a topic will directly link to all other pages on the topic (due to obvious space and design constraints, a page typically links to only a few other pages). In fact, under some scenarios, most notably, competition, a page on a topic may avoid linking to some of the pages on the same topic. However, this is not a contradiction since the property does not imply close connections between all pages on the same topic; instead it affirms the topical similarity of pages that are connected.

2.2. Crawl Prioritization

A critical component of a web crawler is the logic used to select the next URL to visit from a list of unvisited URLs referred to as the *frontier*. The simplest is *breadth-first crawling* where the crawler picks the first URL from a few starting *seed URLs* in the frontier, downloads the corresponding page, extracts and adds its out-link URLs to the frontier, picks the next URL in a FIFO (first-in-first-out) manner and so on. Breadth-first crawling is appropriate when the purpose is to exhaustively crawl the Web as in the case of a general-purpose search engine. However, when the goal is to build a topic focused collection, breadth-first crawling has been shown to be a rather

poor strategy (Chakrabarti et al. 1999, Pant and Menczer 2003). Therefore, topical crawlers utilize a *best-first* crawling strategy. In particular, topical crawlers assign a score to each URL in the frontier based on the estimated potential of the URL leading to a topical page. At each step of the crawl the URL with the highest score in the frontier is picked next for fetching. Within this broad framework, many crawlers have been designed that essentially differ in the algorithms and the range of criteria used to estimate topical relevance for scoring URLs in the frontier.

Cho et al. (1998) have suggested prioritizing crawls so that the more important pages are downloaded. They suggested metrics such as in-link count and PageRank of URLs in the frontier to prioritize the crawl. Unfortunately the link-based metrics utilized by them were computed using the limited link structure information available from the downloaded pages. In other words, they can be seen as using *partial* in-link count or *partial* PageRank (Baeza-Yates et al. 2005). One may expect these partial metrics to be noisy at best since they are computed with just a small number of downloaded pages (Boldi et al. 2004). Also, Menczer et al. (2001) show that, in addition to being computationally expensive, a crawling strategy that continuously updates the PageRanks of frontier URLs using downloaded pages serves is a weak crawl strategy. Thus when we seek to build focused collections it is clear that crawl criteria based solely on global notions of status (e.g., in-links or PageRank) is both insufficient and inefficient. This lacuna in part motivates our research on utility-biased crawlers designed to consider both topicality and status in the design and evaluation of web crawlers. Moreover in the adaptive version we propose methods to dynamically adjust the balance between status and topicality during the crawl within bounds set by overall topicality expectations for the focused collection.

2.3. Estimates of Page Topicality and Status

Chakrabarti et al. (1999) suggested using a bayesian classifier for scoring topicality of the URLs; such a classifier is trained (before the crawl) using example topical pages. They found this method of crawl prioritization to be effective for building topical collection as compared to breadth-first crawlers. Pant and Srinivasan (2005) compared a variety of classification techniques (under various parameter settings) for guiding topical crawlers and highlighted the importance of distribution of scores provided by different classification techniques. They found support vector machine (Vapnik 1995) to provide better performance as compared to naive bayesian classifiers. Therefore we use SVM classifiers for the topical portion of the proposed utility-biased crawlers.

To estimate *status* or importance of a web page, in-links to the page have been used as an indicator in the same manner as citations to scholarly articles have been used to measure article impact. Pages with large numbers of in-links have been described as "authorities." The PageRank (Brin and Page 1998) measure is based on the same notion. A high correlation between in-link

count and PageRank has been observed in previous studies (Amento et al. 2000, Fortunato et al. 2008). In addition, for several information retrieval tasks, in-link count and PageRank have been found to perform similarly (Upstill et al. 2003b,a). Both in-link count and PageRank are global measures. That is, they require a very large snapshot of the Web (i.e., billions of pages) in order to identify most if not all of the links leading to the page being assessed. In contrast, the information available to web crawlers, particularly to topical web crawlers, is local in nature. A crawler accesses only those web pages that it has downloaded up to a given point in time. The crawlers that we suggest circumvent the mammoth task of computing the global page status by relying on empirical work by Pant and Srinivasan (2010) that proposes predictive models of global page status using page-specific features.

2.4. Building Focused Web Collections

Kelley coined the term "Marketing Intelligence" during the 1960s (Kelley 1965) to encompass a wide variety of external as well as internal information gathering and analysis by businesses. Kelley noted a burgeoning trend in which managers were no longer plagued by the lack of data but instead by their overabundance. This problem has only increased since the 60s. The example with 'proteomics' used in Section 1 is only one of possibly hundreds of thousands if not millions of topics that have an abundance of relevant information amongst the billions of publicly available web pages (Lyman and Varian 2003)[2]. While many of these pages (and their temporal trends) are of potential value for business intelligence and vertical search applications (Chen et al. 2002), finding the most appropriate pages remains a challenge. Adding to that challenge is the distributed and dynamic nature of the Web which theoretically makes any snapshot of a subspace of the Web incomplete and obsolete as soon as it is downloaded. To keep up with the dynamism of the Web the collection gathered for an application must be dynamic requiring repeated Web crawls in order to be current. Clearly manual searching and browsing will not scale well for building or maintaining a collection of even a few thousand pages. Therefore topical crawlers (e.g., (Chakrabarti et al. 1999, Diligenti et al. 2000, Pant and Srinivasan 2005, 2006)) have been suggested for automatically building focused collections rich in information about a topic or theme. Such an automated approach is more scalable than a manual technique.

However, the present situation is that given the large number of pages on even niche topics collecting topical pages is not enough; topical crawlers now face the daunting task of capturing the *most important* of the topical pages. Specialized applications such as in vertical search and business intelligence need access to web collections that are not just topical but also focused on the more important subspaces of the topic's presence on the Web. In other words, focused web collections

[2] The official Google blog claims that the search engine is aware of a trillion unique URLs on the Web.

must capture important topical pages. Our status locality conjecture and our utility-biased crawlers point the way towards satisfying this requirement.

3. Status Locality

In this section we first present the arguments supporting our conjecture of status locality, i.e., that pages linking to each other tend to be of similar status as compared to random pages and that such similarity in status decays rapidly with link distance. We then present direct empirical evidence supporting this conjecture.

The web *graph*, with pages as nodes and hyperlinks as edges, has been shown to exhibit various properties of social networks. Social networks are representations of people and their relationships. For example, similar to social networks, the Web has been shown to have a "small world" structure (Albert et al. 1999), exhibit the presence of communities (Girvan and Newman 2002) and reflect power-law distributions among links (Albert et al. 1999, Adamic and Huberman 2002). Our status locality hypothesis identifies another property when the Web is viewed as a social network. The hypothesis is inspired by the field of Sociology where there is substantial evidence to support the classical thought that "similarity breeds connection" (McPherson et al. 2001), a concept formally referred to as *homophily*. Homophily has been observed in social networks that vary widely in scale and properties of its constituents. In particular two different types of homophily have been identified—value homophily and status homophily (McPherson et al. 2001). Value homophily refers to the tendency of people or groups to connect to other people or groups of similar thought, ideologies, beliefs and attitudes. As mentioned before we find a strong parallel between value homophily and topical locality on the Web.

The second type of homophily, i.e, status homophily is central here. It refers to the tendency of people or groups to associate with other people or groups of similar social status. Status is the honor or esteem in which an individual is held. Status homophily manifests itself in 'social hierarchies' that are pervasive in social groups and noted to be a basic property of social relations (Magee and Galinsky 2008). Literature in sociology, psychology, and organizational behavior have suggested that hierarchies can emerge spontaneously and informally in social groups (Blau and Scott 1962, Berger et al. 1980, Eagty and Karau 1991) and these can sustain themselves despite attempts to undermine them (Magee and Galinsky 2008). Works by Gould (2002) and Anderson et al. (2001) argue that there is asymmetry of attention between low status and high status individuals. Specifically, high status individuals pay less attention (are weakly connected) to low status individuals than the reverse (Gould 2002, Magee and Galinsky 2008). We see strong parallels between the appearance of informal hierarchies in social groups and the emergence of status of web pages as measured by in-links. Similar to social groups where a few central individuals

attract most of the status (Magee and Galinsky 2008), a few "authoritative" pages on the Web are known to attract most of the links (and hence status) as indicated by power-law distributions of in-link counts (Albert et al. 1999, Adamic and Huberman 2002). Hence, similar to the argument by Gould (2002) about relationship asymmetry in social groups, we conjecture that high status web pages are more likely to link to other high status pages than to low status pages. Similar to topical locality, this status similarity effect is expected to be localized and decay rapidly with link distance creating a parallel notion of status locality. These arguments form the motivations of our conjecture about the presence of status locality on the Web.

As with the topical locality, status locality does not imply that high status pages directly link to all other high status pages. In fact a typical page links to only a few other pages and sometimes a high status page may avoid linking to another high status page for reasons such as competition. Again, this is not contradictory to the status locality property since the property does not enforce a connection but only suggests similarity in status of connected pages.

We test our conjecture using approaches that parallel tests of topic locality. First we put it to direct empirical test by examining changes in status similarity with link-distance changes. This parallels the direct test of topical locality done in previous studies (Davison 2000, Menczer 2004). Second we test the conjecture by testing the effectiveness of crawlers built using this property. In particular we build a utility-biased web crawler that considers the hypothesized status locality property in addition to topical locality. Effectiveness of the resulting crawler design will be evidence in support of the hypothesized property. This approach parallels the well-recognized fact that traditional topical crawler designs assume the topical locality property and in turn the success of these crawlers is evidence of the existence of that property.

3.1. Direct Empirical Test of Status Locality

To put the status locality conjecture to direct test we study the change in status similarity of page pairs as their link-distance changes. Link distance is measured as the number of links to be followed to navigate from one page to another. The conjecture suggests that as link distance increases status similarity will decrease. Also, the lowest similarity will be exhibited by random pairings of pages. In order to conduct this test we need a collection of web pages that are at various link-distances. We also need the global status information for each web page in the collection. With these two pieces of information we can explore the relationship between link-distances and status similarity. In particular, we would like to compare the status correlation of pages that are 1-link away to pages that are at different link-distances including random link-distances. This would parallel the observations made previously with regards to empirically establishing the topical locality on the Web (Davison 2000, Menczer 2004).

To build our collection we first identify all of the unique URLs listed under the "Business" category of the open directory project or ODP.[3] ODP is a web directory with hierarchically arranged topics that list corresponding relevant URLs. At the time of writing, ODP is maintained by more than 80 thousand human editors. It is referred to as the "largest, most comprehensive human-edited directory of the Web" (ODP 2010). It was utilized by Google for its own directory service. Previous work (Pant 2005, Pant and Srinivasan 2006) has suggested that competitive communities pose greater challenge for web crawling due to reluctance in linking to other relevant yet competitor pages. Hence we expect that our choice of "Business' topics from ODP will provide a conservative estimate of status locality as we would expect greater linking among similar status pages across web sites that belong to more collaborative communities (such as academia).

From the unique URLs identified we sample 5000 URLs randomly. We call this set of URLs the *root* URLs since we begin our exploration of status locality from these URLs. We then attempt to download each of the 5000 URLs and extract the unique links (URLs) appearing in them. We identify a random sample of 5000 URLs among the extracted links and call these the *distance 1* URLs since they are one link away from the root URLs. Each distance 1 URL is associated with a root URL from which it was obtained. Next we attempt to download the pages corresponding to distance 1 URLs and extract all of the unique URLs appearing in them. We again identify a random sample of 5000 URLs among the extracted links and call the sample *distance 2* URLs to indicate that these URLs are at most 2 links away from the root URLs.[4] Each distance 2 URL is associated with a root URL which led to it. We repeat the process to obtain *distance 3* URLs and associate them with appropriate root URLs. Now, from amongst the root, distance 1, distance 2, and distance 3 URLs, we create a random sample of 5000 URLs that we call *random distance* URLs and associate them randomly with the root URLs. In other words unlike the URLs at various distances, the random distance URLs are at an unknown distance from the root URLs they are associated with since they are not obtained based on known links. The schematic in Figure 1 shows the links at various distances from root URLs. The circles indicate the pages while solid lines with arrows are the hyperlinks. The dashed lines with arrows shows virtual associations between the root URLs and the random distance URLs.

Next we obtain the in-link counts of URLs at various distances including the random distance URLs using the Google API[5] that provides limited non-commercial access to Google's database. The in-link count as previously noted is a global measure of page status that Google computes

[3] `dmoz.org`

[4] note that we are not exhaustively searching for every path from root to distance 2 URLs and hence we can only guarantee that the upper bound on the distance between the two sets of URLs is 2.

[5] `http://code.google.com/apis/soapsearch/`

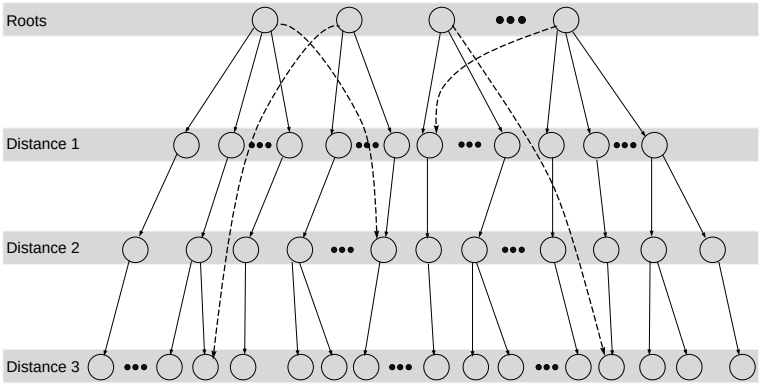

Figure 1 Pages at various distances from the root URLs. The dashed lines represent random distance.

through processing billions of web pages. Figure 2 shows the spearman's rank correlation (ρ) between the status of root URLs with associated URLs at various link distances. We find that the correlation between the status of the root URLs and the distance 1 URLs is 0.36 and statistically significant. The correlation drops by more than 45% as we move to the distance 2 URLs and the distance 3 URLs but still remains statistically significant. At random distance the correlation becomes close to 0 ($\rho = -0.01$) and is statistically insignificant. Clearly, two pages that are directly linked are much more similar in status than those at higher or random link-distances. These observations support our status locality conjecture. These observations are also similar to that made for topical locality on the Web (Davison 2000, Menczer 2004). To further check that this phenomenon holds even for pages from different (web) domains, we filter the distance 1, distance 2, and distance 3 URLs to only include those that are from a different second-level domain than the corresponding root URL. Using this filtered set of URLs we recompute the correlations as before. We find that the spearman correlation between the status of the root URLs and the filtered distance 1 URLs is 0.17 and statistically significant and drops by more than 61% as we move to the filtered distance 2 and distance 3 URLs. At random distance the correlation becomes close to 0 and statistically insignificant. Hence we again observe status locality wherein pages that are directly linked are much more similar in status than those at higher or random link-distances. The lower status correlation between different domain URLs (as compared to all URLs in the initial sample) is consistent with similar observation for topical similarity between different domain pages while uncovering the topical locality phenomenon (Davison 2000).

So we know now that given a page of interest both status similarity and topic similarity are localized to directly connected pages. To the best of our knowledge this is the first direct evidence of status locality on the Web. As with topicality, we find that the status locality signal falls quickly (see Figure 2) as we move away from the root page. In other words, the *radius* of status locality is small and hence it provides a short-range guidance for status of linked pages. The main implication

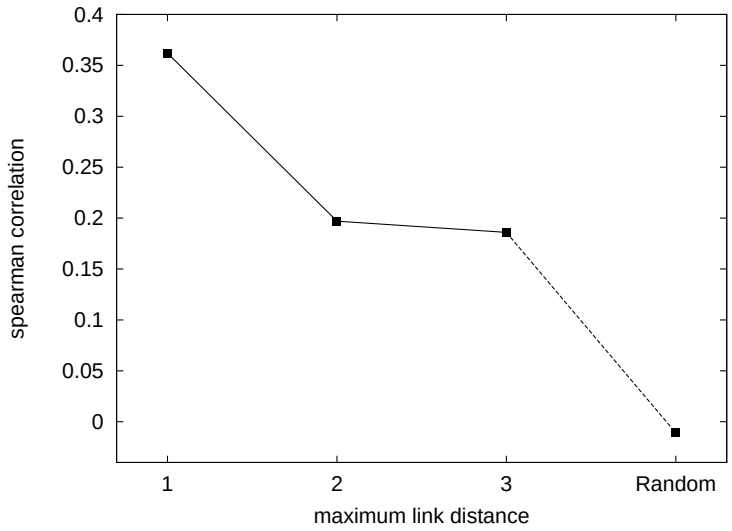

Figure 2 **Spearman correlation between status of root URLs and URLs as various distances (including random distance) from their roots**

of the observed status locality for crawler design is that one should restrict the use of status cues to score URLs that appear directly on the page whose status estimate has been computed. In other words, we will use estimated status of pages to rank URLs of unexplored pages that are 1-link away. Our ability to make judgements on status of pages at larger link-distances is much weaker.

4. Utility-biased Web Crawler

Here we present the design of a new family of crawlers called utility-biased crawlers that take advantage of both the status locality and the topical locality properties of the Web. Note that effectiveness of these crawlers will provide additional evidence of the presence of the status locality property. Figure 3 describes the basic design of our new crawler. The crawler begins by learning a topical model and a status model using example web pages. Both models use machine learning techniques to learn a mapping between local properties (independent variables) of a web page and its topicality or status (dependent variable). The crawler then loads the frontier (i.e., a dynamic list of URLs that are yet to be visited) with seed (initial) URLs. Next a multi-threaded loop is executed that stops when an adequate number of pages have been downloaded. The frontier is shared across many threads executing this loop. Each thread of execution is a crawler instance that first picks the URL with highest utility, downloads the corresponding page, and uses the two previously learned models to estimate the topicality (τ) and the status (σ) of the downloaded page. Using these two estimates, the crawler computes the utility ($U_c(p)$) of the page (p) based on the following Cobb-Douglas function:

$$U_c(p) = \tau^{\alpha} \cdot \sigma^{\beta} \tag{1}$$

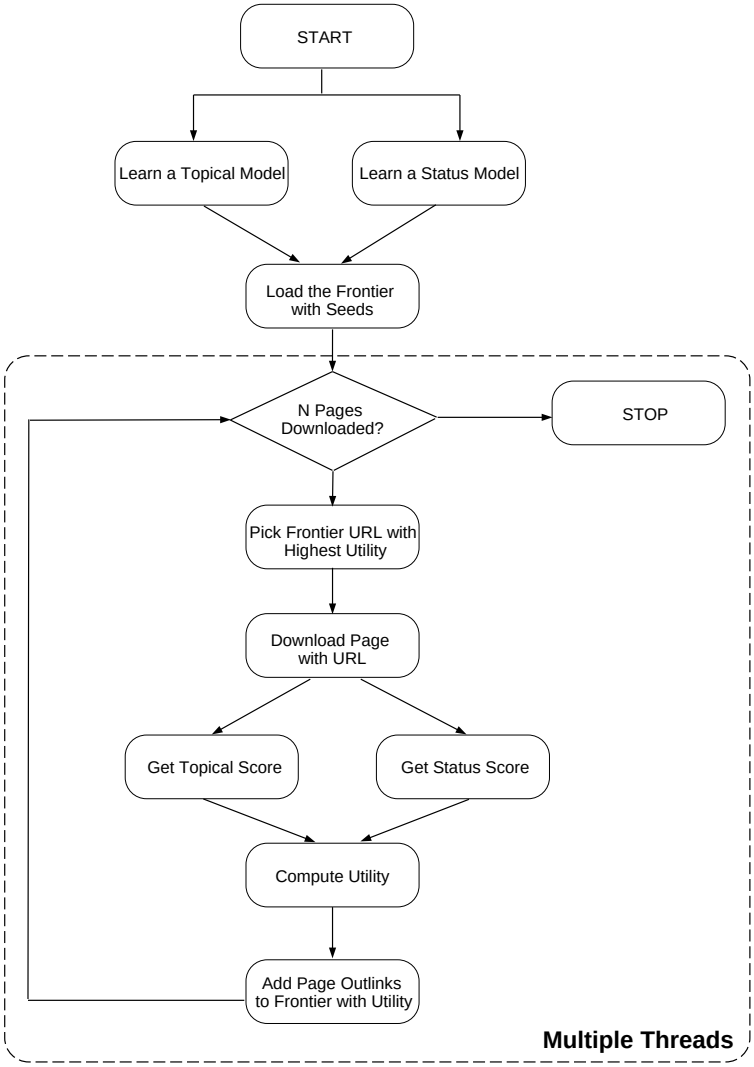

Figure 3 Algorithmic Design of the Utility-Biased Web Crawler

where α and β are the output elasticities of topicality and status. We assume $\alpha+\beta=1$ and hence we will only specify α for the rest of paper. The above function provides an intuitive interpretation; if a page is twice as high in topicality and twice as high in status than another, then it has twice as high utility as well. Also, the function's concave nature ensures that increases in topicality (or status) cause smaller increases in utility (i.e., diminishing returns) as the topicality (or status) increases. We also experiment with the following alternate definition of utility that is linear in topicality and status:

$$U_l(p) = \alpha \cdot \tau + (1-\alpha) \cdot \sigma \tag{2}$$

The crawler computes the utility of each page that is downloaded and then associates that utility value with each of the unexplored URLs (out-links) appearing on the page. The URLs are then added to the frontier along with the utility values. When the crawler needs to fetch the next page,

it again selects the URL with the highest utility value. We follow the basic multi-threaded crawler setup as described in Pant and Srinivasan (2005) that (for online politeness) implements the robot exclusion protocol and avoids accessing the same web domains too frequently.

We obtain different versions of the utility-biased crawler by varying α. In the current study we experiment with α values of 0, 0.25, 0.5, 0.75, and 1 for both utility functions. We note that for $\alpha = 1$, the crawler is relying purely on topicality (τ) and hence it is equivalent to the best-performing state-of-the-art topical crawler (Pant and Srinivasan 2005). For $\alpha = 0$, it is relying purely on status (σ). For all other α values it is guided by a mixture of the two components. As noted earlier, the success of topical crawlers relies on the topical locality on the Web. Since we associate the status of a downloaded page with its unexplored out-links through the utility function, the utility-biased crawler's computation of the utility of the out-links is additionally based on the validity of status locality on the Web. Therefore, the success of the utility-biased crawler in obtaining higher status collections (especially for low values of α) will be additional concrete evidence in support of the status locality phenomenon.

4.1. Estimating Topicality

The crawler estimates topicality with an SVM-based text classifier that has been used effectively to guide topical crawlers (Pant and Srinivasan 2005). We use the Weka implementation of SVM (Witten and Frank 2005). Before crawling begins, an SVM-based classifier is trained using the positive and the negative examples (web pages) for a given topic. Given a vocabulary $\mathcal{V}$ of n terms, a training example j is represented as a vector $\mathbf{p_j} = (p_{j1}, p_{j2} \ldots p_{jn})$ in $\mathcal{R}^n$ space. Each term weight (p_{jk}) for the training example is based on standard TF-IDF representation (Manning et al. 2008) of the web page. The SVM algorithm finds a hyperplane in n-dimensional space that optimally separates (if possible) the positive and negative example vectors. Hence the training process constitutes solving an optimization problem (quadratic programming) to get the parameters that define the discriminating hyperplane (Burges 1998). The trained classifier is used at crawl time to estimate the topicality of downloaded pages. Each downloaded page is also represented as a vector of TF-IDF term weights. The classifier is then applied to the vector and its output (proportional to the distance of the page vector from the hyperplane) transformed through a sigmoid function ($1/[1 + e^{-x}]$). The transformed score ranges between 0 and 1, where 1 indicates highest relevance or topicality level.

4.2. Estimating Status

We discussed earlier the difficulties of efficiently calculating global status measures such as in-link count and PageRank using global data. Moreover, we pointed to recent work by Pant and Srinivasan (2010) presenting effective strategies for predicting global page status using properties

local to the page. We use the terms local and global to contrast the levels of information needed. Global page characteristics (such as in-link count) by definition necessitate processing of all or a large fraction of pages on the Web. Local characteristics of a page on the other hand, such as those related to its content, require much more limited information. This is an ideal setup for a web crawler that is building a topical collection with no efficient access to global status information.

Our utility-biased crawler estimates the global page status from its local characteristics with the M5' decision tree model (Pant and Srinivasan 2010). We use the implementation of M5' provided by the Weka machine learning software (Witten and Frank 2005). M5' is closely based on the ideas of model tree which is a type of decision tree where each node represents a decision point based on input variables (page characteristics). However, unlike a typical decision tree, the leaf nodes do not correspond to discrete classes but rather linear models that map input variables to the output (estimated status). Hence M5' can be used for regression problems (instead of classification problems) and provides a piece-wise function that maps inputs to a continuous output.

We use the following page characteristics as specified by Pant and Srinivasan (2010) for the predictive model (summarized here for reader convenience). They motivate the choice of these characteristics by citing sociological theories along with observations of search engine marketing and usability experts.

1. *Information Volume* is measured as the number of words (ignoring stopwords such as "and", "the") appearing in the page. It measures the amount of textual information on a given page.

2. *Information Location* is measured as the number of sub-directories from the host name to the actual file name in the URL of the page. For example, the page with URL `http://www.somedomain.com/products/news/daily.html` has an Information Location of 2. This variable quantifies how deeply a page is submerged in the directory structure of the web server.

3. *Information Specificity* is measured as the average inverse document frequency (IDF) of the characteristic terms on the page. We follow the procedures specified in (Pant and Srinivasan 2010) to identify the characteristic terms and compute the IDFs. IDF is high if the term has low probability of appearing in pages of a given collection. In other words higher IDF indicates higher specificity.

4. *Information Brokerage* is measured as the number of other pages a page links to or the out-link count. Links from a page to itself are not counted. This variable quantifies the subtle service provided by web pages in terms of suggesting to users links to follow.

5. *Link Ratio* is a ratio of the number of links on a page to the number of words on the page. Similar metrics have been used to differentiate between primarily content (lower Link Ratio) and navigation (higher Link Ratio) pages (Cooley et al. 1999).

6. *Quantitative Ratio* is measured as the ratio of the number of numbers and the number of words appearing on the page. For example, if 10% of words on a page are numbers, then Quantitative Ratio for the page is 0.1. This variable tries to differentiate between pages that have mainly quantitative information from those with mainly qualitative information.

7. *Domain Traffic* is measured through *reach* data for the domain (e.g, `merck.com`) in which the page is hosted. The reach data is obtained from Alexa Web Information Service (AWIS).[6] AWIS measures reach for a given domain as the proportion of a panel of users who visit that domain. We note that this is a domain characteristic that we associate with a given page.

The M5' decision tree is trained using a topic's example pages before crawling begins for the topic. Each example page is represented by these seven page characteristics. The model's output is the estimated page status. While training we use the in-link count of the example pages as the target output variable. We have discussed earlier the use of in-link count as an indicator of page status, its correlation with PageRank and the similar performances obtained on several tasks using these measures (Amento et al. 2000, Upstill et al. 2003a,b). We obtain the in-link counts using the same Google API as in Section 3.1. In-link counts and some of the page characteristics are known to have skewed distributions (Albert et al. 1999, Adamic and Huberman 2002). Hence, log-transformed ($\ln[1+x]$) variables are used with the M5'-based regression model. During the crawl, the output of the model for a page is transformed through a sigmoid function ($1/[1+e^{-(x-\theta)}]$, $\theta = 2.82$)[7] to get an estimate of the page's status which is then used in Equations 1 and 2.

Note that the training of the topical and status models happens only once for a topic with its initial set of example pages. This happens before starting the crawling phase (see Figure 3). Hence the only global page status information that the crawler uses (i.e., in-link counts from Google) is for learning the status model. This is a one time cost for the utility-biased web crawler. For the entire duration of the crawl the utility-biased web crawler depends only on the topicality and status estimates provided by the learned models.

5. Performance of Utility-Biased Crawlers

5.1. Test Bed and Performance Metrics

The crawlers are evaluated on topics derived from the open directory project (ODP). We choose the top-level category of ODP called "Business." We then identified the 37 sub-categories directly under it such as "Business→Information Technology" that have 500 or more URLs listed under them (down the entire topic hierarchy below it). We used all 37 sub-categories as topics. These range from "Transportation and Logistics" to "Investing." For each topic, we randomly pick 250

[6] We use the URLInfo interface of the AWIS to obtain the reach data.

[7] θ is the median of status values predicted by the M5' model in an exploratory crawl.

URLs as positive training examples and another random but disjoint set of 250 URLs as held-out positive testing examples. Using ODP topics other than the given topic, another random set of 250 URLs of negative training examples is created. Similarly, yet another set of disjoint 250 URLs is used as held-out negative testing examples. The positive and negative training examples are used to train the SVM-based text classifier for estimating page topicality. The positive training examples are also used to train the M5'-based non-linear regression model to estimate page status. The SVM-based topical classifier and the M5'-based predictive model are used by the utility-biased web crawler. In addition, the positive training examples are used as seeds for the crawl.

We test 5 different utility-biased crawlers that only differ in the value of α (0, 0.25, 0.5, 0.75, 1.0) as explained in Section 4. Again, for $\alpha = 1$, the crawler is equivalent to a state-of-the-art topical crawler, while for $\alpha = 0$ the crawler relies purely on status estimates and the status locality property. For each topic we crawl and download up to 10000 pages using each of the 5 utility-biased crawlers. In other words we crawl more than a million pages across topics and crawlers. We need performance metrics that can scale to millions of pages and we use the following two metrics:

1. *Average Topicality*: This metric is built using harvest rate (Chakrabarti et al. 1999), a popular metric to measure overall topicality after a crawler harvests a collection. Harvest rate is the fraction of crawled pages that is relevant. An SVM-based evaluation classifier is used to classify each page as relevant or not relevant. We train it using the held-out positive and negative testing examples which were never provided to the crawler. Average topicality is the average harvest rate for the crawler across the 37 topics. We note that the average topicality metric utilizes the manual judgements of ODP editors and can be applied to millions of crawled pages.

2. *Average Status*: Although our crawler uses an *estimate* of page status based on local features of the page with this metric we measure performance using the *actual* status of the pages. Ideally, we would like to obtain the in-link counts (from Google) for each crawled page and use the mean of (log-transformed) in-link counts as a measure of collection level status. However, this is not possible due to Google API limits on the number of queries per day. Instead, we find in-link counts for 30 random pages from every 500 pages crawled (i.e., random 30 pages from the first 500 pages crawled, 30 from the next 500 and so on). We use the mean of the (log-transformed) in-link counts of these random pages as the status of the pages crawled. We compute status at various points of the crawl (first 500, 1000, ..., 10000 pages) for each of the crawlers and for each topic. We then average the status for a crawler across the 37 topics.

We measure the accuracy of the evaluation classifiers (used for computing the average topicality) on each of the topics using the positive and negative examples used for training the crawler (but not for training the evaluation classifier). We find that the average accuracy of the evaluation classifiers for the 37 topics is 78.14 ± 0.70 (note that the prior is 50%). In other words, the evaluation

classifiers largely capture the topical signals. Further, they are not expected to be biased towards any particular crawler and hence can serve as a neutral judge.

We do not consider a utility-like formulation combining status and topicality of crawled pages into a single score (similar to Equations 1 and 2) for a performance metric. Such a metric would require us to a priori decide on an "appropriate" α value which would bias the resulting performance metric (since we also use different α values as the basis for different types of crawlers). The current set of performance metrics focus on the main concerns relating to the performance of the crawlers without having any dependence on the setup of those crawlers. This helps us clearly understand the trade-offs between the dual concerns of topicality and status. We provide an analysis of the joint distribution of topicality and status in the online companion (Appendix).

5.2. Performance

We now present results after the 5 different utility-biased crawlers have crawled 10000 pages for each of the 37 test topics. Figure 4(a) shows the time-series plot of average topicality achieved by the crawlers with different values of α using the Cobb-Douglas utility (results with linear utility are described later). To avoid clutter we only include crawlers using 3 of the 5 α values. Figure 4(b) shows the average topicality after crawling 10000 pages for all of the 5 different values of α. Not surprisingly, $\alpha = 1$ (i.e., the state-of-the-art topical crawler) achieves the highest average topicality. But interestingly, the crawler with $\alpha = 0.5$ (the same is true for $\alpha = 0.75$) maintains consistent average topicality of close to or above 70% for the entire crawl of 10000 pages. Hence, as long as a reasonable weight is given to topicality, the crawler is able to build collections that are largely on the topic. The $\alpha = 0$ crawler, which emphasizes only status, quickly deviates from the topic (see Figure 4(a)).

Figure 5(a) shows the time-series of average status achieved by utility-biased crawlers with different values of α. Again, to avoid clutter, we only include crawlers using 3 of the 5 α values in the figure. Figure 5(b) shows average status for the 5 different values of α after crawling 10000 pages. We see that $\alpha = 0$ (i.e., crawler driven purely by status) achieves the highest average status. Also $\alpha = 0.25$ and $\alpha = 0.5$ have higher average status than $\alpha = 0.75$ and $\alpha = 1$. We note that the lower α values indicate greater reliance on the status estimates and the status locality. In particular, at $\alpha = 0$ the utility-biased web crawler is simply using the estimated status of downloaded pages as scores of the corresponding unvisited pages (URLs) that are 1-link away from them. In other words the crawler depends only on status locality (and ignores topical locality) to guide itself. This strategy achieves the highest average status among all variations of the utility-biased web crawler thus validating the presence of status locality.

We note that the highest average topicality for $\alpha = 1$ (see Figure 4) is achieved at the cost of lower status. This $\alpha = 1$ crawler is a state-of-the-art topical crawler. Note also that the highest

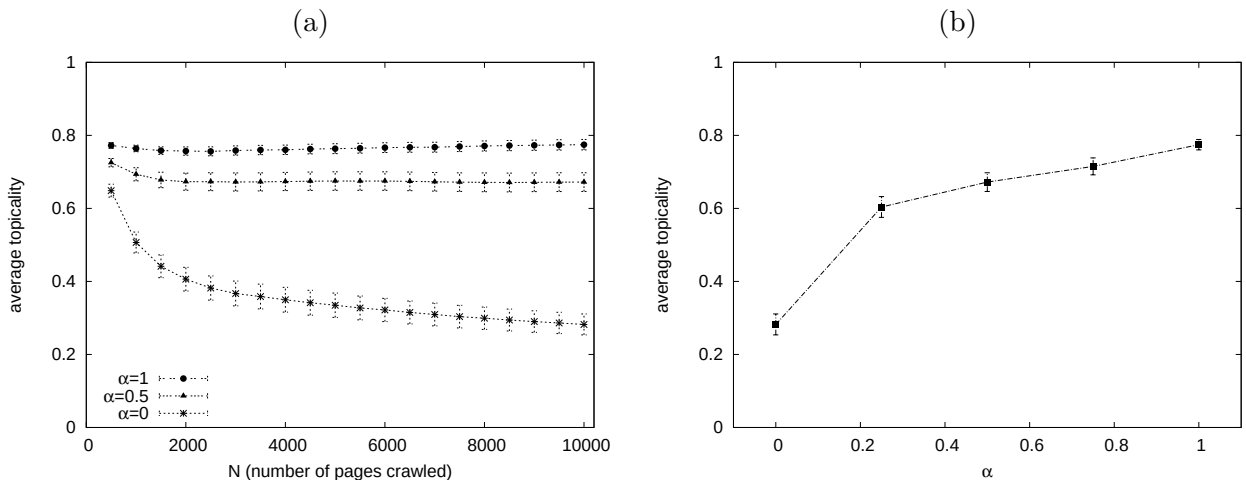

Figure 4 **Average Topicality: (a) at different points in the crawl (b) for different values of α after crawling 10000 pages. The error bars indicate ± 1 std. error.**

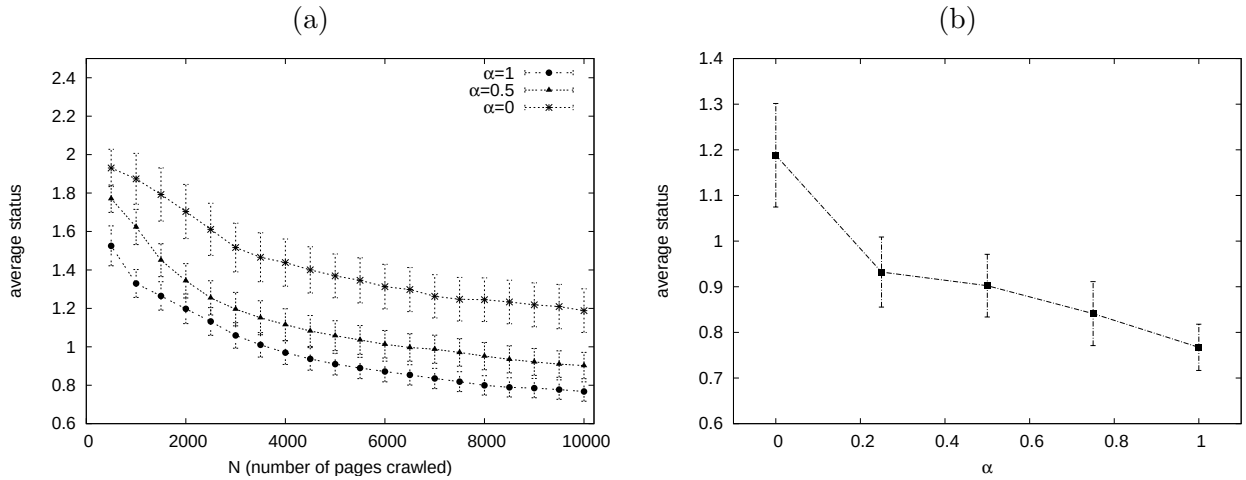

Figure 5 **Average Status: (a) at different points in the crawl (b) for different values of α after crawling 10000 pages. The error bars indicate ± 1 std. error. The performance measurement is based on the actual status of a sample of crawled pages.**

average status by $\alpha = 0$ (see Figure 5) is achieved at the cost of lower collection topicality. Hence, on average there is a trade-off between creating the most topical and the highest status collection. This trade-off is not bound to the utility function as in the extreme cases ($\alpha = 1$ and $\alpha = 0$) the utility function is immaterial. In contrast to these two extremes, Figure 5 shows that $\alpha = 0.5$ maintains consistently higher average status than $\alpha = 1$ for the entire crawl. In fact, average collection status after crawling 10000 pages is significantly higher for $\alpha = 0.5$ than for $\alpha = 1$ ($p = 0.02$, two-tailed paired t-test). At the same time, as seen in Figure 4, $\alpha = 0.5$ also creates largely topical collection with average topicality consistently close to 70%. In other words, *with appropriate values of output elasticities the utility-biased web crawler produces largely topical collections that*

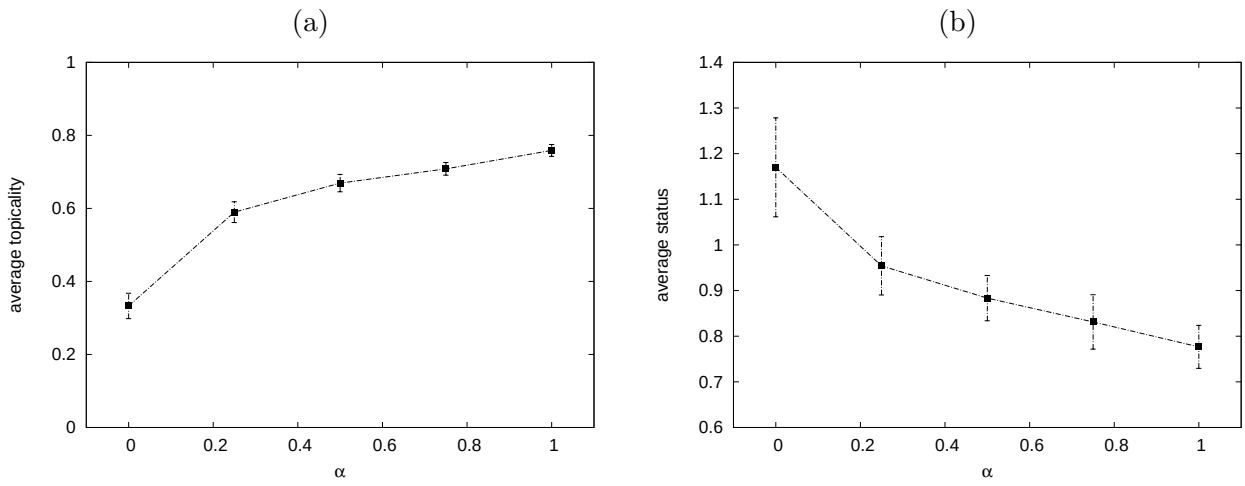

Figure 6 **Linear Utility: (a) average topicality (b) average status for different values of α after crawling 10000 pages. The error bars indicate ± 1 std. error.**

have significantly higher average status than those produced by a state-of-the-art topical crawler. This is a key conclusion that indicates the merits of our new family of utility-biased crawlers.

As compared to the time series plots of average topicality (see Figure 4(a)), the time series plots of average status (see Figure 5(a)) show a bigger drop as the crawl progresses ($\alpha = 0$ is a notable exception). We believe this is due to the different nature of topicality versus status. In particular the power-law distribution of status on the Web (Albert et al. 1999, Adamic and Huberman 2002) is expected to create a different type for trajectory for average status. Hence it is important to individually analyze the relative performance of the crawlers given a performance metric without mixing the metrics.

Performance results with the linear utility function guiding the utility-biased crawlers were similar to the ones with the Cobb-Douglas utility function. Figure 6 summarizes the performance of these crawlers using various values of α under the same conditions for crawl size and test topics. Again we observe the trade-off between status and topicality and note that for $\alpha = 0.5$ the utility-biased web crawler produces largely topical collections that have significantly higher average status than those produced by a state-of-the-art topical crawler (i.e., $\alpha = 1$). In general for values of α that do not lie on the extremes (i.e., neither 0 or 1) the crawler with Cobb-Douglas utility performs slightly better (on both of the evaluation metrics) than the crawler with linear utility. The only exception is $\alpha = 0.25$ where the crawler with linear utility was slightly better on average status than the crawler with Cobb-Douglas utility. Hence we will use utility-biased web crawler with Cobb-Douglas utility for the rest of our experiments. We believe that the concave nature of the Cobb-Douglas utility formulation provides a slight edge over linear utility as it can handle extreme values of status and topicality more robustly.

The above results (especially those for low values of α) provide empirical evidence supporting our conjecture of the status locality property of the Web. This is in addition to the direct evidence provided in Section 3. Figure 4 and Figure 5 along with the statistical tests provide a strong quantitative understanding of the design of the utility-biased web crawler. We see clearly the trade-off between status and topicality. For a more qualitative understanding of the design's sensitivity to α parameter we plot the joint distribution of topicality and status among the crawled pages from the performance data. These plots are provided in the online companion (Appendix).

6. Adaptive Utility-biased Web Crawler

We observed, both quantitatively and qualitatively, the trade-off between topicality and status with our utility-biased web crawler. However, it seems unrealistic to expect the nature of the optimal trade-off to remain constant throughout the duration of the crawl and that too across different topics. By fixing α (as we have done thus far) this is what is implicitly assumed. The question we ask now is as follows: Is there an "ideal" constant α value or should α be sensitive to the local subspace of the Web in which the crawler finds itself positioned? Building upon our basic design of the utility-biased web crawler and its evaluation we now suggest a more adaptive version of the crawler. The difference is we allow α (as seen in Equation 1) to change during the crawl and adjust the trade-off between status and topicality given the local conditions. The intuition for this is as follows. If, for example, the crawler is in a region of the Web that offers a lot of topically relevant pages then the crawler can afford to lower its α value so as to concentrate more on harvesting high status pages. However, when it is hard to find topical pages the crawler would be able to increase the α value so as to concentrate on topicality. We believe that such an adaptive design would be able to better manage the trade-off between topicality and status by being sensitive to local subspaces of the Web. This is in contrast to our "one-size-fit-all" design described in Section 4 where the α remains constant for different topic domains and for all portions of a given crawl. In the spirit of the assess-refine loop of design science (Hevner et al. 2004) we now suggest a refined adaptive design for the utility-biased web crawler.

6.1. Adaptive Design

The adaptive crawler that we now propose is based on the concept of *target topicality level*, δ. This represents the desired minimum topicality of pages among the last M pages crawled. It captures the notion of desired topical makeup of pages among the local (recent) pages that have been crawled.[8] We conjecture that if a crawler is currently in a topical subspace of the Web (i.e., it recently downloaded enough topical pages) then it can afford to be relaxed about seeking topical pages and

[8] Of course the analogous notion of target status level also arises. However, since we are interested in largely topical collections we choose target topicality level as the initial guide for our crawler.

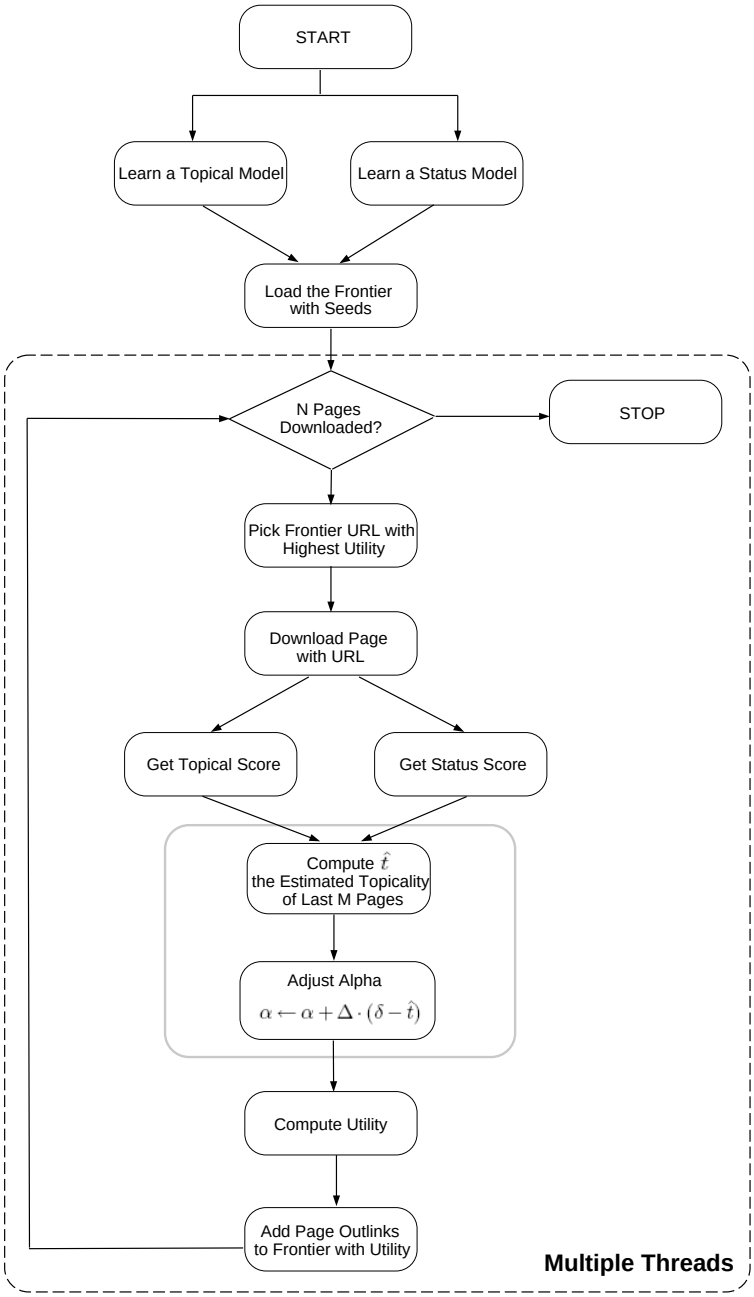

Figure 7 Algorithmic Design of the Adaptive Utility-biased Web Crawler

pay more attention to downloading higher status pages. Due to topical locality the neighborhood of the recently crawled pages is expected to contain many more topical pages. This would make the job of harvesting topical pages, easier. Thus it can afford to lower α as long as it remains in such a topical subspace. In contrast when the crawler moves to a subspace that is not as dense in topical pages it would need to adjust its priorities by boosting the importance of topical heuristics (i.e., increase α). These arguments are formalized in the design of our adaptive utility-biased web

crawler. Specifically, α is allowed to change based on the gap between desired topicality for the collection (δ) and the estimated topicality of the last M pages crawled. Note that the crawler makes topicality estimates about the pages it has downloaded using the SVM-based text classifier that it uses to guide the crawl. Given a desired value of δ (set before the crawl), α controling the relative importance of topicality and status is modified after downloading each page as follows:

$$\alpha \leftarrow \alpha + \Delta \cdot (\delta - \hat{t}), 0 \leq \alpha \leq 1 \tag{3}$$

where $\hat{t}$ is the average estimated topicality of the last M pages and Δ is the step size (α is bounded by 0 and 1). We set $M = 25$ and $\Delta = 0.01$. We now evaluate the performance of our adaptive crawler with these settings using the test bed and metrics described earlier. Figure 7 summarizes the algorithm followed by the adaptive crawler. The main difference between the utility-biased web crawler (see Figure 3) and the adaptive utility-biased web crawler are the steps circled by the gray box in Figure 7.

6.2. Performance: Adaptive Utility-biased Web Crawler

Using the test bed and evaluation metrics described in Section 5.1 we now test 4 versions of the adaptive utility-biased crawler that differ in the target topicality levels desired (i.e., the δ values). In particular, we use $\delta = 0.6, 0.7, 0.8, 0.9$ where the increasing values of δ indicate a greater amount of topicality desired in the collection. We compare performance of these adaptive crawlers with that of the utility-biased web crawler with constant α value ($\alpha = 0.5$). This setting was found to present a good trade-off creating largely topical web collections that were significantly higher in average status than that created by a purely topical crawler (Section 5.2).

Figure 8 (a) shows the average topicality of pages downloaded by the different versions of the adaptive utility-biased crawler as time-series data over the length of the 10000 page crawls. Increasing the target topicality level, δ, leads to more topical collections throughout the length of the crawls. Figure 8 (b) shows the average topicality for the adaptive utility-biased web crawlers (different δs) and the utility-biased web crawler with constant $\alpha = 0.5$ after crawling 10000 pages. The crawlers with δ =0.8 and 0.9 achieve higher average topicality as compared to the crawler with constant $\alpha = 0.5$ (the difference is statistically significant for δ=0.9). While the crawler with δ =0.7 achieves a lower average topicality than the crawler with constant $\alpha = 0.5$, the difference between them is not statistically significant. We note again that δ and α serve different purposes. While the former sets the target topicality level, the latter controls the relative weight of page topicality versus status in guiding the crawler. In our adaptive crawler design, the user sets the value of δ which in turn (automatically) effects the dynamic values of α as indicated by Equation 3. Our results show that with appropriate values of δ it is possible for our adaptive utility-biased web

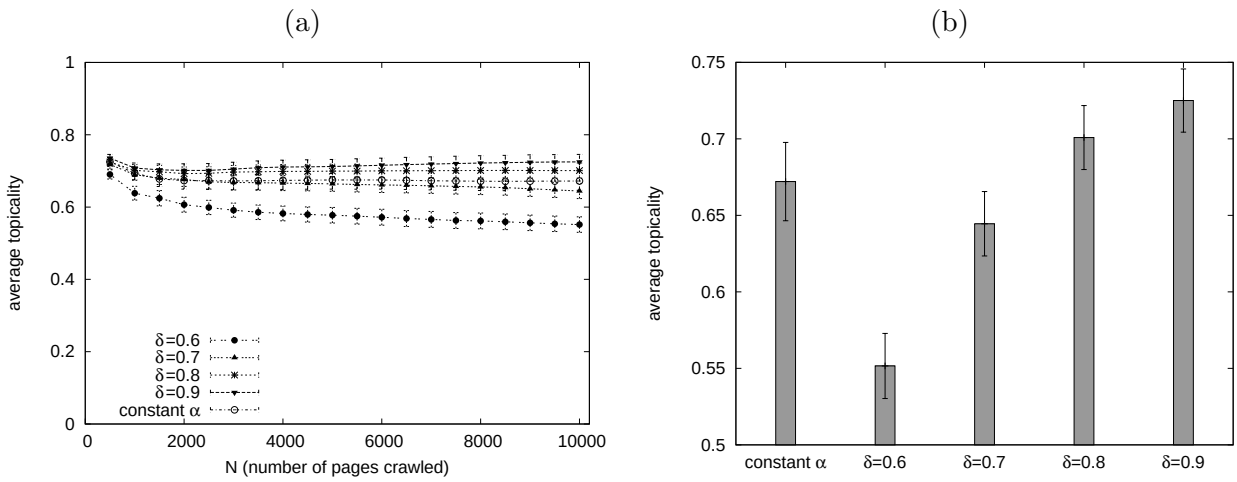

Figure 8 **Average Topicality: (a) at different points in the crawl (b) for constant $\alpha = 0.5$ and different values of δ after crawling 10000 pages. The error bars indicate ± 1 std. error.**

crawler to perform as well or better on average topicality than our utility-biased web crawler with constant $\alpha = 0.5$.

Figure 9 (a) shows the average status of pages downloaded by the different versions of the adaptive utility-biased as time-series over the length of the 10000 page crawls. We observe that lower values of target topicality level (δ) produces higher status collection. In essence, results in Figure 8 (a) and Figure 9 (a) reinforce the topicality-status trade-off noted before. Figure 9 (b) shows the average status for the adaptive utility-biased web crawlers (different δs) and the utility-biased web crawler with constant $\alpha = 0.5$ after crawling 10000 pages. The crawler with constant $\alpha = 0.5$ is the worst performer on this metric. In fact, the average status for constant $\alpha = 0.5$ is statistically significantly ($p < 0.01$, two-tailed paired-t test) lower than that for δ=0.6, 0.7, and 0.8.

Based on the results in Figure 8 (b) and Figure 9 (b) we conclude that allowing α to change dynamically according to a preset target topicality level (δ) allows the adaptive utility-biased crawler to better manage the topicality-status trade-off. Adaptive crawlers with δ =0.7 and 0.8 create web collections that on an average are of significantly higher status than collections created using utility-biased web crawlers with constant $\alpha = 0.5$. Moreover this improvement in status is achieved while maintaining comparable or higher average topicality. Thus we have demonstrated that *the adaptive design provides better management of the status-topicality tradeoff within the general framework of utility-biased web crawlers.*

6.3. Analysis

To further understand the dynamics of adaptive utility-biased web crawler we plot the average values of α as it changes at each step of the crawl (see Figure 10). The α values are averaged over the 37 topics in the test bed. All of the various versions of adaptive utility-biased web crawler

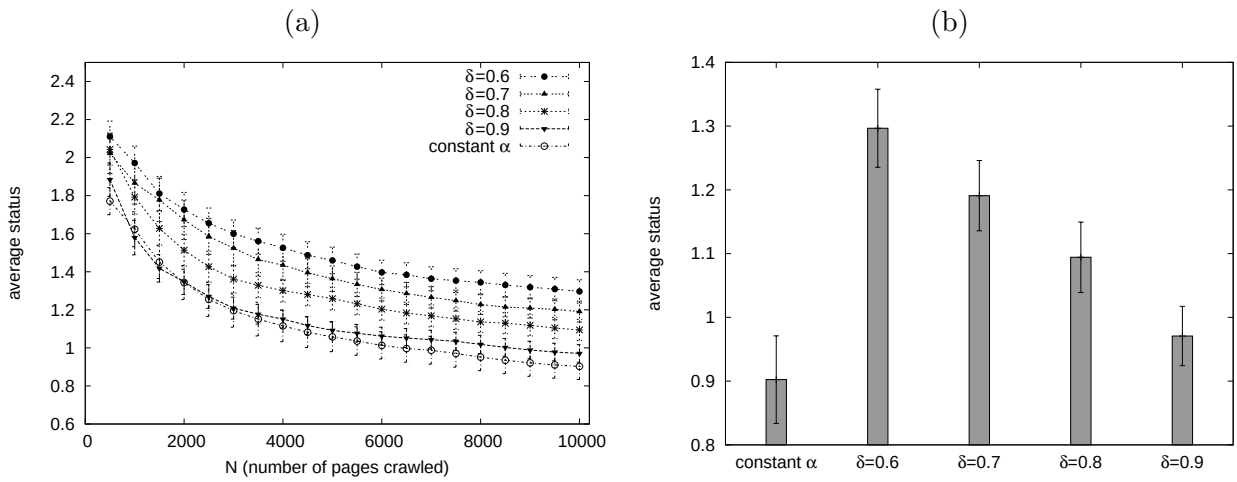

Figure 9 **Average Status: (a) at different points in the crawl (b) for constant $\alpha = 0.5$ and different values of δ after crawling 10000 pages. The error bars indicate ± 1 std. error.**

start with the same value of $\alpha = 0.5$ but each has a different target topicality level, δ. In the initial stages of the crawls the trajectories of average α for δ =0.8 and 0.9 diverge away quickly from the trajectories of average α for δ =0.6 and 0.7 (see Figure 10). In particular, the trajectories of δ =0.8 and 0.9 spike upwards indicating that the target topicality levels are hard to meet and hence there is a strong pressure on α to move upwards thus giving increasing importance to topicality in utility computations. Similarly average α initially drops downwards in case of δ =0.6 and 0.7 indicating that the these target topicality levels are relatively easy to meet (at least initially) and hence the crawler can afford to give greater importance to page status. After the initial startup stage that lasts a few hundred pages (see Figure 10), the crawlers with δ =0.6, 0.7, 0.8 correct themselves in the reverse direction before showing relative stability in their levels of α values. In contrast, the crawler with δ =0.9 does not show substantial correction in the values of average α after the startup period and the average α values for this crawler hover close to 1 for the rest of the crawl. This indicates that $\delta = 0.9$ is a target topicality level that is extremely hard to meet and hence the adaptive utility-biased web crawler with $\delta = 0.9$ on an average tends to remain close to $\alpha = 1$.

In the adaptive utility-biased crawler, the parameter that needs to be set by the user is δ (i.e., the target topicality). Once set, δ controls the dynamics of α (i.e., status versus topicality focus) as the crawler adjusts to the local subspaces on the Web. This adaptive design is simpler for the user since it focuses their attention to a single dimension (topicality) and the crawler then takes care of the adjustments between topicality and status. The above analysis informs the user that if the target topicality is set to be arbitrarily high (say $\delta = 0.9$), the crawler will be largely unable to meet such a target and hence be forced to put inordinate amount of importance on the topicality of the pages it is downloading. Given the status-topicality tradeoff this will lead to considerably

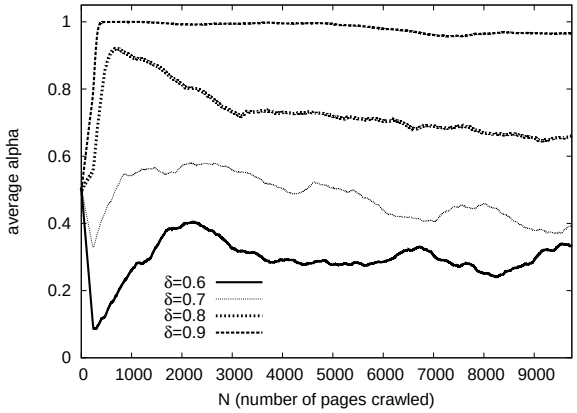

Figure 10 **Average α: for various versions of adaptive utility-biased web crawler at various points in the crawl.**

lower status of web collections. However, for moderate values of δ such as 0.7 or 0.8, the crawler can adjust α values despite an initial upward or downward spike (see Figure 10). In other words, these topicality target are not unrealistic and can be met, thus providing wiggle room for the crawler to focus its attention on the status of pages.

7. Additional Features and Comparative Results

7.1. Robustness: Crawl Size and Status Data

We now test the robustness of our adaptive utility-biased crawler by asking the following two questions: (1) Does the crawler continue to maintain its status advantage beyond 10000 pages, and (2) What would happen if we used a different search engine than Google both for obtaining the status data for our crawling and evaluation framework?

To answer these questions we run the adaptive utility-biased crawler ($\delta = 0.7$) and the state-of-the-art topical crawler (as described in Section 4) for a 20000 page crawl—double the size of our previous crawls—on each of the 37 topics in our test bed. In addition, we obtain the in-links count data using the Yahoo! Search Boss API. We use the in-links count data for computing status of pages as before. We used the page status obtained from Yahoo! both for training the status estimate model and for computing the average status performance measure for evaluating the crawled pages. We note that different search engines have been shown to have a low overlap in terms of their coverage of the Web (Lawrence and Giles 1998, Spink et al. 2006). Hence the use of Yahoo! data tests the robustness of not just the crawler design but also the evaluation framework to varying sources of data on page status.

Figure 11 shows the average performance of the two crawlers over the 20000 page crawls. As seen before in Figure 4 and Figure 5, the topical crawler outperforms on the average topicality, however, the adaptive utility-biased web crawler continues to build web collections that are largely topical as well. More importantly, the adaptive utility-biased web crawler continues to maintain

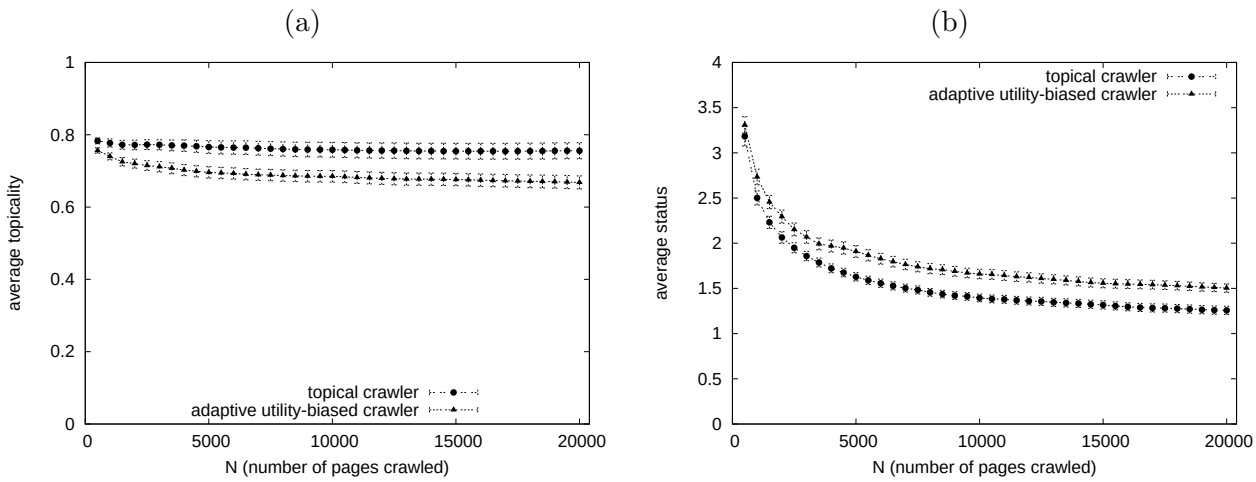

Figure 11 **Long Crawls: (a) average topicality (b) average status after crawling** 20000 **pages. The error bars indicate** ± 1 **std. error.**

its statistically significant advantage on average status of web collections as compared the topical crawler. In other words, we find that the pattern of results seen for the 10000 page crawls and with status data from Google qualitatively holds for longer crawls of 20000 pages and with data from Yahoo!. These results indicate consistency in our observations and robustness of our crawlers .

7.2. A Competitive Benchmark using In-links

We now propose and evaluate a competitive benchmark called the *in-links crawler* that differs in one key aspect from the adaptive utility-biased web crawler. It uses partial in-link count to estimate status instead of our predictive model. This benchmark crawler tracks the in-link counts of each downloaded page via a *structure index* (Arasu et al. 2001) which is updated after each new page is fetched. This benchmark crawler is exactly the same as the adaptive utility-biased crawler except it uses the in-link count from the structure index instead of the page status estimate from M5'-based predictive model. Everything else is kept the same between the two crawlers. Comparing the two will help us tease out the relative merit of using our predictive model. The benchmark crawler also corresponds to crawlers suggested earlier using 'local snap shots' of in-link counts (Cho et al. 1998, Baeza-Yates et al. 2005). These are partial in-links count compared to more global estimates from a search engine. We set the $\delta = 0.7$ in both the crawlers as this value was found to be effective in Section 6.

Figure 12 shows the performance of the two crawlers. We find that both achieve similar average topicality rates (no statistically significant difference at 10000 pages). However, the utility-biased web crawler significantly outperforms the in-link crawler on the average status ($p < 0.02$, two-tailed, at 10000 pages). Both of the crawlers utilize the same topical model (using SVM) and are able to guide themselves to equally topical content. The only and key difference is in the measurement of

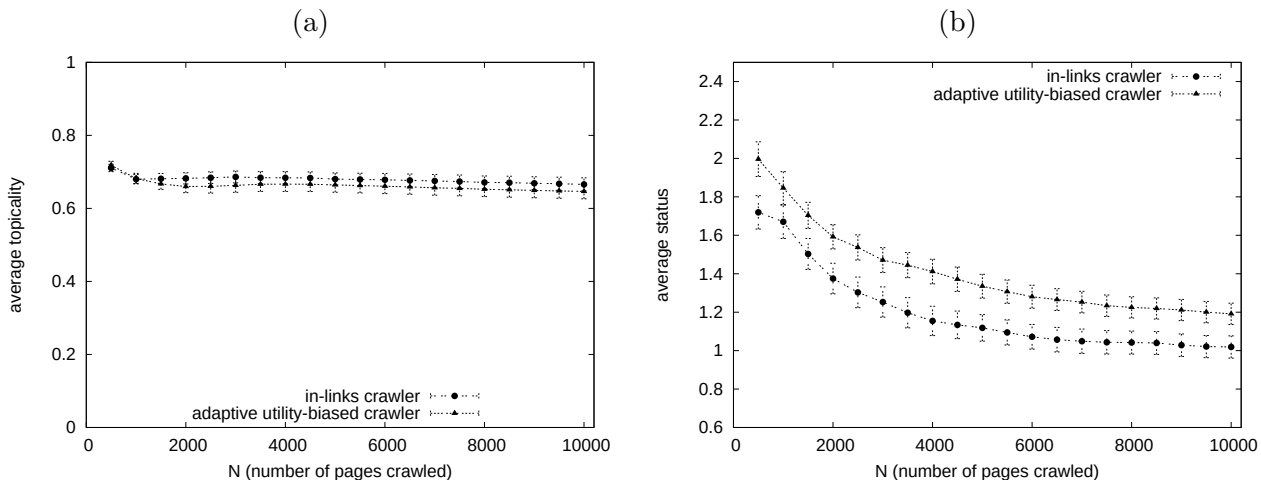

Figure 12 **Performance of the in-links crawler and the the adaptive utility-biased crawler: (a) average topicality (b) average status at different points in the crawl. The error bars indicate ± 1 std. error.**

the page status. While both measure page status based on local information about pages that have been downloaded, Figure 12 (b) reveals that information used by the in-links crawler is clearly less helpful than the estimate of page status provided by the M5' predictive model. In other words, we find that page status measurements designed to estimate global status (i.e., our predictive model) more effectively guide a web crawler towards important topical content than the partial information contained in the local snap shots of the in-link counts.

7.3. Diversity and Reliability of Crawled Sites

Average topicality and average status assess the quality of web collections created by crawlers based on the text-based topicality and link-based importance of the crawled pages. Similar measures have been popular in the past (Cho et al. 1998, Chakrabarti et al. 1999) for measuring the quality of collections obtained through web crawling. We now consider additional dimensions of quality that can be measured. The first is diversity of sources from which the web collection is obtained. This is particularly important for business intelligence tasks where different sources of similar information can add to the validity of the intelligence extracted from a web collection. To measure the diversity of a collection we count the number of unique web sites (i.e., second-level domains) from which the crawled pages are obtained. Figure 13 (a) shows the diversity of web sites for the adaptive utility-biased crawler and the competitive benchmark (i.e., in-links crawler). We find that both crawlers, on average, derive the 10000 page collection from more than 1100 different web sites thus attesting to the diversity of sources from which the collection is built. This essentially indicates that on an average less than 10 pages are obtained from each site. Given that we can expect many web sites to have a much larger number of pages (than 10), we note that the crawlers maintain a healthy diversity of sources. At the same time, we would like to also note that a crawler could achieve

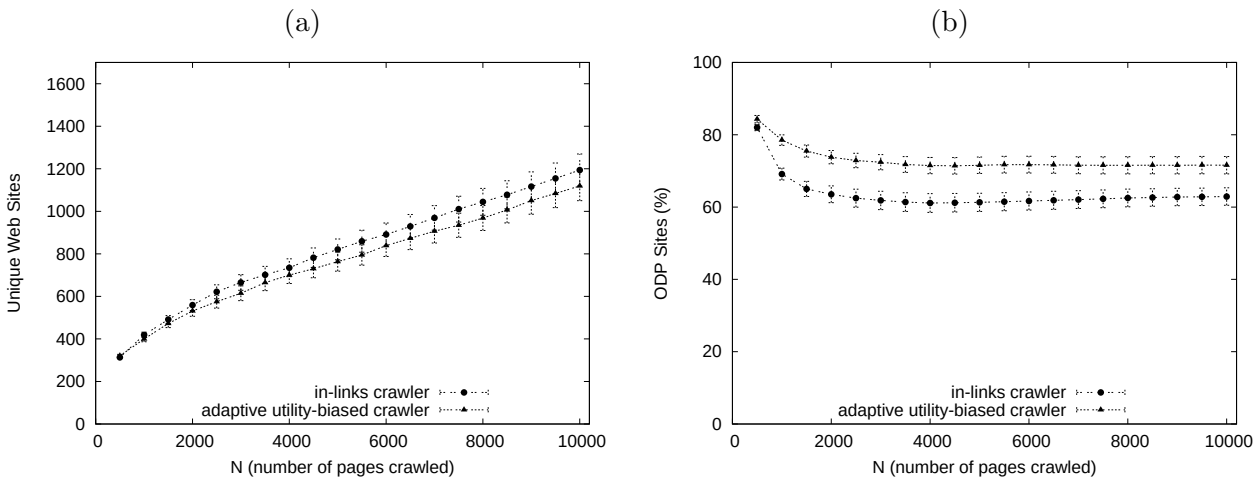

Figure 13 **Other metrics of quality: (a) Diversity (b) Reliability. The error bars indicate ± 1 std. error.**

high diversity by randomly crawling a large number of unrelated sites. However, the crawlers we evaluate in Section 7.2 obtain largely topical collections. Hence they are harvesting relevant pages from diverse sources.

The second dimension of interest is reliability or reputation of web sites crawled. Again from the perspective of business intelligence applications it is important to build collections that are seen as largely reliable. It is extremely hard to obtain an independent measure of reliability for a large and arbitrary set of web sites. However, we take advantage of the intellectual effort of ODP editors to design a metric of reliability. We identify all of the web sites (second-level domains) appearing in the URLs listed in the ODP. We call these over 2.7 million web sites *ODP sites*. Since URLs from ODP sites go through a manual editorial filter we can consider the corresponding sites to be on average more reliable than arbitrary web sites. We measure the reliability of sources behind a web collection by computing the percentage of crawled web sites that are ODP sites. Figure 13 (b) shows this measurement for the adaptive utility-biased crawler and the in-links crawler. We find that the collections built by the adaptive utility-biased crawler, on an average, are derived from more reliable web sites. On an average, throughout the length of the crawl, 70-80% of the sites crawled by the adaptive utility-biased crawler are from the reliable ODP sites. Given that human editors curate these sites, the results indicate that a large majority of sites covered by the crawler are of higher reliability than arbitrary web sites. This result is an important additional validation for the higher quality of web collections produced by the adaptive utility-biased crawler.

8. Conclusion

Given the presence of status homophily in social networks (McPherson et al. 2001), we hypothesize that a similar phenomenon must occur on the Web (which after all is possibly the largest social artifact) as well. In other words, we expect pages that are directly linked to have a much greater

similarity in status than pages that are further away in terms of link distances. We call this phenomenon status locality on the Web. We note that such a phenomenon would signal a type of localized regularity on the Web that can be exploited in the design of applications that depend on local information on the Web such as a topical web crawler. Our empirical exploration provides the first direct evidence for the presence of status locality on the Web. The success of our utility-biased web-crawler design that exploits status locality further validates the same phenomenon while demonstrating its usefulness.

We proposed and tested a utility-biased web crawler that exploits localization of both status and topicality of web pages. Guided by a Cobb-Douglas utility function that combines estimates of status and topicality of pages encountered, the crawler at each step decides on the next URL to fetch. Both status and topicality estimates are made real-time during the crawl using local properties of the pages. This makes the proposed design well suited for vertical search and business intelligence applications that utilize focused collections of web pages. We experiment with various versions of the utility-biased web crawler design by altering the a priori output elasticities of topicality and status (i.e., α). More specifically we test 5 different versions of the utility-biased web crawlers that range from a state-of-the-art topical crawler ($\alpha = 1$) to a crawler that is purely driven by estimated status of downloaded pages ($\alpha = 0$). Based on the evaluation of our basic design and our improved understanding of the problem we refine the idea of utility-biased web crawler by allowing α to change dynamically and adapt during the crawl based on environmental conditions. This requires the crawler to sense and react to changes in the subspaces of the Web as it crawls through them.

This work makes an important theoretical contribution by suggesting and empirically validating the status locality property on the Web. However, it goes beyond that and utilizes the theoretical phenomenon in an effective and practical design of a new family of web crawlers. We now summarize our main findings:

1. We find direct empirical support for status locality on the Web. Status locality has a short-range and is best applied to pages that are 1-link away.

2. We propose a web crawler that exploits both status locality and the well-known topical locality on the Web. Our crawler overcomes the challenge of estimating global status by using strategies presented in previous research for using local cues to predict global status.

3. We find that on an average there is a trade-off between building the most topical collection and building the collection with highest status. However, with appropriate values of output elasticities (within a Cobb-Douglas utility formulation) one can obtain largely topical collections that are also of significantly higher status than those obtained using a state-of-the-art topical crawler. This facility is important for applications inundated with topical information on the Web.

4. An important observation is with regards to crawlers with low values of α (i.e., more weight on page status). Their relative success in terms of achieving higher average status of downloaded pages further validates the existence of status locality as a phenomenon on the Web; it also provides a systematic demonstration of the phenomenon's effect within an application.

5. The adaptive utility-biased web crawler more effectively manages the trade-off between topicality and status as it dynamically reacts to changing environmental conditions. Adaptive crawlers with δ =0.7 and 0.8 create largely topical web collections of significantly higher status than collections created using our best utility-biased web crawler with constant $\alpha = 0.5$. The adaptive design is also simpler for users since it requires their input on a single dimension (topicality) and the crawler then takes care of the adjustments between topicality and status.

6. Our adaptive utility-biased web crawlers are robust even when crawl length is doubled and the status data is derived from a different search engine (i.e., Yahoo!). Importantly our crawler continues to maintain its advantages over the state-of-art topical crawler under these conditions.

7. We compare the adaptive utility-biased web crawler with a crawler that differs only in that it uses in-link counts to estimate page status instead of our status predictive models (that uses page local cues). Our predictive model provides better guidance to the crawler in terms of harvesting a higher status collection. This result again emphasizes that in-link counts computed from downloaded pages are only partial and hence noisy estimates of status.

8. We suggest diversity and reliability of web sites from which the collection is derived as additional measures of collection quality. Both the adaptive utility-biased web crawler and the competitive benchmark create collections that derive from a diverse set of web sites. However, site reliability in the collection created by the adaptive utility-biased web crawler is significantly better.

We have provided a framework for building topical web collections through web crawlers that focus on important subspaces of the topic on the Web. Such a framework is greatly needed to support the information needs of vertical search or business intelligence without becoming inundated by the scale of the Web. Also our framework builds on local information about web pages and hence does not need exhaustive large scale (billions of pages) crawls of the Web or constant reliance on general-purpose search engines for global status measures of web pages. Thus it is relevant not just to large technology-oriented firms but also to smaller organizations with limited technology resources.

Our work makes a larger theoretical contribution by conjecturing and testing the status locality property for the Web. Status locality has broad implications for information search since it suggests that one is likely to find unknown important information by greedily acquiring known important information and following the hyperlinks (more generally, citations) therein. In the future we plan to design search ranking mechanisms that explicitly utilize the status locality phenomenon. We also

aim to explore the current adaptive utility-biased web crawler in the context of a search engine-crawler design where the underlying collection that supports a vertical search engine evolves over time to adjust to the changing information needs of a user community.

References

Adamic, L. A., B. A. Huberman. 2002. Zipfs law and the internet. *Glottometrics* **3** 143–150.

Albert, Réka, Hawoong Jeong, A.-L. Barabási. 1999. Diameter of the world-wide web. *Nature* **401** 130.

Amento, B., L. Terveen, W. Hill. 2000. Does authority mean quality? predicting expert quality ratings of web documents. *Proc. 23rd ACM SIGIR Conference*. 296–303.

Anderson, C., O. P. John, D. Keltner, A. M. Kring. 2001. Who attains social status? effects of personality and physical attractiveness in social groups. *Journal of Personality and Social Psychology* **81**(1) 116–132.

Arasu, Arvind, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, Sriram Raghavan. 2001. Searching the web. *ACM Transactions on Internet Technology* **1**(1) 2–43.

Baeza-Yates, R., C. Castillo, M. Marin, A. Rodriguez. 2005. Crawling a country: better strategies than breadth-first for web page ordering. *Proceedings of the 14th international conference on World Wide Web*. 864–872.

Berger, J., S. J. Rosenholtz, Jr. M. Zelditch. 1980. Status organizing processes. *Annual Review of Sociology* **6** 479–508.

Blau, P. M., W. R. Scott. 1962. *Formal Organizations: A Comparative Approach*. Stanford University Press.

Boldi, P., M. Santini, S. Vigna. 2004. Do your worst to make the best: Paradoxical effects in pagerank incremental computations. *Proceedings of the third Workshop on Web Graphs (WAW), Lecture Notes in Computer Science*. 168–180.

Brin, S., L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* **30**(1–7) 107–117.

Burges, C. J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**(2) 121–167.

Bush, V. 1945. As we may think. *The Atlantic Monthly* .

Chakrabarti, S., M. van den Berg, B. Dom. 1999. Focused crawling: A new approach to topic-specific Web resource discovery. *Proc. 8th WWW Conference*.

Chen, H., M. Chau, D. Zeng. 2002. CI spider: A tool for competitive intelligence on the Web. *Decision Support Systems* **34** 1–17.

Cho, J., H. Garcia-Molina, L. Page. 1998. Efficient crawling through URL ordering. *Computer Networks* **30**(1–7) 161–172.

Cooley, R., B. Mobasher, J. Srivastava. 1999. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems* **1**(1) 5–32.

Davison, B. D. 2000. Topical locality in the web. *Proc. 23rd ACM SIGIR Conference*.

Diligenti, M., F. Coetzee, S. Lawrence, C. L. Giles, M. Gori. 2000. Focused crawling using context graphs. *Proc. 26th Intl. Conference on Very Large Data Bases (VLDB 2000)*. Cairo, Egypt, 527–534.

Eagty, A. H., S. J. Karau. 1991. Gender and the emergence of leaders: A meta-analysis. *Journal of Personality and Social Psychology* **60**(5) 685–710.

Fortunato, Santo, Marián Boguñá, Alessandro Flammini, Filippo Menczer. 2008. Approximating pagerank from in-degree. William Aiello, Andrei Broder, Jeannette Janssen, Evangelos Milios, eds., *Algorithms and Models for the Web-Graph*. Springer-Verlag, Berlin, Heidelberg, 59–71.

Girvan, M., M. E. J. Newman. 2002. Community structure in social and biological networks. *Proc. National Academy of Sciences (PNAS)* **99**(12) 78217826.

Gould, R. V. 2002. The origins of status hierarchies: A formal theory and empirical test. *American Journal of Sociology* **107**(5) 1143–78.

Hevner, A. R., S. T. March, J. Park, S. Ram. 2004. Design science in information systems research. *MIS Quarterly* **28**(1) 75–106.

Kelley, W. T. 1965. Marketing intelligence for top management. *Journal of Marketing* **29** 19–24.

Lawrence, Steve, C. Lee Giles. 1998. Searching the world wide web. *Science* **280**(5360) 98–100.

Lyman, P., H. R. Varian. 2003. How much information. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on 07/26/2005.

Magee, J. C., A. D. Galinsky. 2008. Social hierarchy: The self-reinforcing nature of power and status. *The Academy of Management Annals* **2** 351 – 398.

Manning, C. D., P. Raghavan, Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge.

McPherson, M., L. Smith-Lovin, J. M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* **27** 415–444.

Menczer, F. 2004. Lexical and semantic clustering by web links. *Journal of the American Society for Information Science and Technology* **55**(14).

Menczer, F., G. Pant, M. Ruiz, P. Srinivasan. 2001. Evaluating topic-driven Web crawlers. *Proc. 24th Annual Intl. ACM SIGIR Conference on Research and Development in Information Retrieval*.

ODP. 2010. *About the Open Directory Project*. URL `http://www.dmoz.org/about.html`.

Pant, G. 2005. Building web collections for vertical markets. *Proc. 15th Workshop on Information Technologies and Systems (WITS)*.

Pant, G., F. Menczer. 2003. Topical crawling for business intelligence. *Proc. 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*.

Pant, G., P. Srinivasan. 2005. Learning to crawl: Comparing classification schemes. *ACM Transactions on Information Systems* **23**(4) 430–462.

Pant, G., P. Srinivasan. 2006. Link contexts in classifier-guided topical crawlers. *IEEE Transactions on Knowledge and Data Engineering* **18**(1) 107–122.

Pant, G., P. Srinivasan. 2010. Predicting web page status. *Information Systems Research* **21**(2).

Spink, Amanda, Bernard J. Jansen, Vinish Kathuria, Sherry Koshman. 2006. Overlap among major web search engines. *Internet Research* **16**(4) 419–426.

Upstill, T., N. Craswell, D. Hawking. 2003a. Predicting fame and fortune: Pagerank or indegree? *Proc. 8th Australasian Document Computing Symposium*. Canberra.

Upstill, T., N. Craswell, D. Hawking. 2003b. Query-independent evidence in home page finding. *ACM Transactions on Information Systems* **21**(3) 286–313.

Vapnik, V. N. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc.

Witten, I. H., E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. 2nd ed. Morgan Kaufmann.

Appendix. Map of Topicality and Status

For a more qualitative understanding we plot as heat maps the joint distribution of topicality and status among the crawled pages from the performance data. We use the relevance score (between 0 and 1) provided by the evaluation classifier (see Section 5.1) as the topicality of a crawled page. The status of a crawled page, as explained in Section 5.1, is the (log transformed) in-link count. Since the status for only a sample of crawled pages (30 random pages out of every 500 crawled) is known, we only use these sampled pages to plot the joint distribution. For each α value (i.e., crawler) we gather all sampled crawled pages across the 37 topics. We plot the distribution in a region of topicality and status that covers more than 90% of the sampled crawled pages for each of the crawlers. Figure 14 (a), (b), and (c) present color coded distribution plots for α values of 1 (purely topical crawler), 0 (purely status-based crawler), and 0.5 (an even mixture). The color corresponds to the (log transformed) number of pages in a given region of topicality and status. The black and blue areas have either no or few pages respectively. Red and orange shaded areas have larger numbers of pages. Figure 14 (a), (b), and (c) show the contrast in the pages collected by the corresponding crawlers. Notice the black-blue region that is concentrated in the lower-right portion of Figure 14 (a) ($\alpha = 1$) moves to upper-right portion of Figure 14 (b) ($\alpha = 0$). This qualitatively indicates that the crawler with $\alpha = 1$ tends to focus on topicality and misses out on high status pages. On the other hand, the crawler with $\alpha = 0$ tends to focus on status and misses out on highly topical pages. Figure 14 (c) ($\alpha = 0.5$) shows a mixture of these two extreme policies. As we have seen in Figure 4 and Figure 5, such a mixture is capable of producing a largely topical collection that is also of significantly higher status than a purely topical crawler. Our observations further indicate that α is playing its necessary role in shifting the focus of the crawler between status and topicality. Since we associate the status and topicality estimates of a downloaded page with its unexplored out-links, the role of α would not be effective in the absence of status and topical locality. In Figure 14 (d) we show the distribution for adaptive utility-biased web crawler ($\delta = 0.7$) where the α values are allowed to

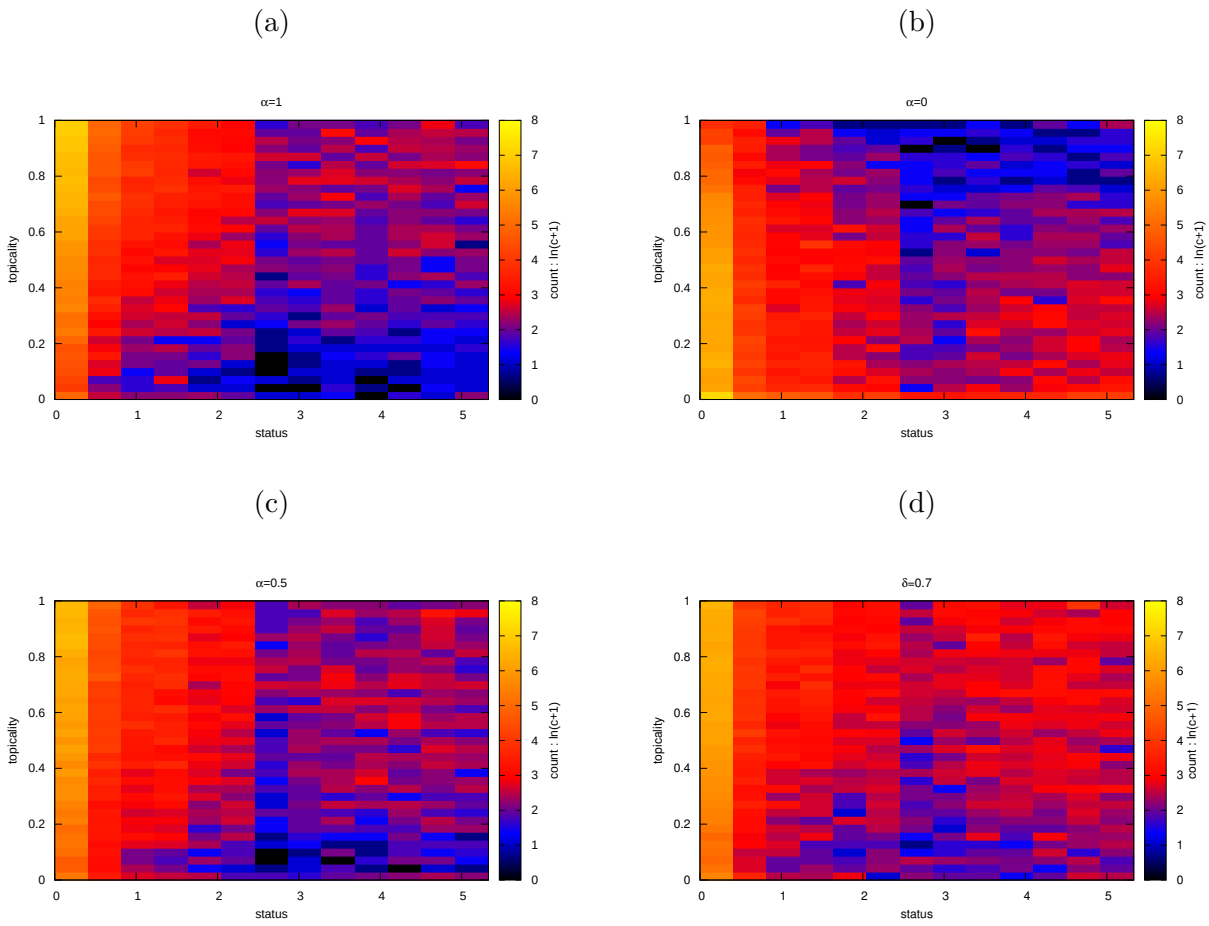

Figure 14 **Distribution of Topicality and Status (color-coded): (a)** $\alpha = 1$ **(b)** $\alpha = 0$ **(c)** $\alpha = 0.5$ **(d)** $\delta = 0.7$**.**

dynamically change over the the crawl period. As compared to Figure 14 (a), (b), and (c), we note a strong shift towards collecting high status and high topicality pages which highlights the better performace of the adaptive utility-biased crawler.