Topical Locality in the Web: Experiments and Observations*

Brian D. Davison
Department of Computer Science
Rutgers, The State University of New Jersey
New Brunswick, NJ 08903 USA
davison@cs.rutgers.edu

July 2000

Abstract

Most web pages are linked to others with related content. This idea, combined with another that says that *text in, and possibly around, HTML anchors describe the pages to which they point*, is the foundation for a usable World-Wide Web. In this paper, we examine to what extent these ideas hold by empirically testing whether topical locality mirrors spatial locality of pages on the Web. In particular, we find that the likelihood of linked pages having similar textual content to be high; the similarity of sibling pages increases when the links from the parent are close together; titles, descriptions, and anchor text represent at least part of the target page; and that anchor text may be a useful discriminator among unseen child pages. These results present the foundations necessary for the success of many web systems, including search engines, focused crawlers, linkage analyzers, and intelligent web agents.

1 Introduction

Most web pages are linked to others with related content. This idea, combined with another that says that *text in, and possibly around, HTML anchors describe the pages to which they point*, is the foundation for a usable World-Wide Web. They make browsing possible, since users would not follow links if those links were unlikely to point to relevant and useful content. These ideas have also been noticed by researchers and developers, and are implicit in many of the systems and services found on the Web today.

These ideas are so basic that in many cases they are not mentioned, even though without them the systems would fail to be useful. When one or both are mentioned explicitly (as in [MB00, DH99, GKR98, BS97, Kle98, BP98, CDR+98, Ami98]), their influence is measured implicitly, if at all. This paper is an attempt to rectify the situation — we wish to measure the extent to which these ideas hold.

This paper primarily addresses two topics: it examines the presence of textual overlap in pages near one another in the web, and the related issue of the quality of descriptions of web pages. The former is most relevant to focused web crawlers and to search engines using link analysis, while the latter is primarily of use to web indexers, meta-search tools, and to human browsers of the web

*A shorter version of this paper is available as a conference paper [Dav00].

since users expect to find pages that are indeed described by link text (when browsing the Web) and to find pages that are characterized accurately by the descriptive text presented by search engine results. We show empirical evidence of topical locality in the Web, and of the value of descriptive text as representatives of the targeted page. In particular, we find that the likelihood of linked pages having similar textual content to be high; that the similarity of sibling pages increases when the links from the parent are close together; that titles, descriptions, and anchor text represent at least part of the target page; and that anchor text may be a useful discriminator among unseen child pages.

For the experiments described in this paper, we select a set of pages from the Web and follow a random subset of the links present on those pages. This provides us with a corpus in which we can measure the textual similarity of nearby or remote pages and explore the quality of titles, descriptions, and anchor links with respect to their representation of the document so described. In the next section, we will describe the motivation of this work in further detail, giving examples from many applications, including web indexers, search ranking systems, focused crawlers and web prefetchers. We will then describe our experimental methodology, present the results found, conclude with a summary of our findings and suggest further work.

2 Motivation

The World-Wide Web is not a homogeneous, strictly-organized structure. While small parts of it may be ordered systematically, many pages have links to others that appear almost random at first glance. Fortunately, further inspection generally shows that the typical web page author does not place random links in her pages (with the possible exception of banner advertising), but instead tends to create links to pages on related topics. This practice is widely believed to be typical, and as such underlies a number of systems and services on the web, some of which are mentioned below.

Additionally, there is the question of describing the web pages. While it is common for some applications to just use the contents of the web pages themselves, there are situations in which one may have only the titles and/or descriptions of a page (as in the results page from a query of a typical search engine), or only the text in and around a link to a page. A number of systems could or do assume that these "page proxies" accurately represent the pages they describe, and we include some of those systems below.

2.1 Web indexers

A web indexer takes pages from the web and generates an inverted index of those pages for later searching. Popular search engines including AltaVista[1], Lycos[2], etc. all have indexers of some sort that perform this function. However, many search engines once indexed much less than the full text of each page. The WWW Worm [McB94], for example, indexed titles and anchor text. Lycos, at one time, only indexed the first 20 lines or 20% of the text [KABL96]. More recently Google[3] started out by indexing just the titles [BP98].

Today it is common for the major engines to index not only all the text, but also the title of each page. Smaller services such as research projects or intranet search engines may opt for reduced storage and index less. What is less common is the indexing of HTML META tags containing author-supplied keywords and descriptions. Some search engines will index the text of these fields,

[1] http://www.altavista.com/
[2] http://www.lycos.com/
[3] http://www.google.com/

but others do not [Sul00], citing problems with search engine spamming (that is, some authors will place keywords and text that are not relevant to the current page but instead are designed to draw traffic for popular search terms).

Likewise, while indexers typically include anchor text (text within and/or around a hypertext link) as some of the terms that represent the page on which they are found, most do not use them as terms to describe the page referenced. One significant exception is Google, which does index anchor text. By doing so, Google is able to present target pages to the user that have not been crawled, or have no text, or are redirected to another page. One drawback, however, is that this text might not in fact be related to the target page. A recent publicized example was the query "more evil than satan himself" which, at least for a while, returned Microsoft as the highest ranked answer from Google [Sul99, Spr99].

So, for search engine designers, we want to address the questions of how well anchor text, title text, and META tag description text represent the target page's text. Even when title and descriptions are indexed, they may need to be weighted differently from terms appearing in the text of a page. Our goal is to provide some evidence that may be used in making decisions about whether to include such text (in addition to or instead of the target text content) in the indexing process.

2.2 Search ranking systems

Traditionally, search engines have used text analysis to find pages relevant to a query. Today, however, many search engines incorporate additional factors of user popularity (based on actual user traffic), link popularity (that is, how many other pages link to the page), and various forms of page status calculations. Both link popularity and status calculations depend, at least in part, on the assumption that page authors do not link to random pages. Presumably, link authors want to direct their readers to pages that will be of interest or are relevant to the topic on the current page. The link analysis approaches used by Clever[4] [Kle98] and others [BH98, BP98, DGK+99] depend on having a set of interconnected pages that are both relevant to the topic of interest and richly interconnected in order to calculate page status. Additionally, some [CDR+98, CDG+99] use anchor text to help rank relevance of a query to communities discovered from the analysis.

LASER [BFJ96] demonstrates a different use of linkage information to rank pages. It computes the textual relevance, and then propagates that relevance backwards along links that point to the relevant pages. The goal is to enable the engine to find pages that are good starting points for automated crawling, even if those pages don't rank highly based on text alone.

Our analysis may help to explain the utility of anchor text usage, as well as show how likely neighboring pages are to be on the same topic.

2.3 Meta-search engines

Meta-search engines (e.g. MetaCrawler[5] [SE95, SE97], SavvySearch[6] [DH97, HD97], and DogPile[7]) are search services that do not search an index of their own, but instead collect and compile the results of searching other engines. While these services may do nothing more than present the results they obtained for the client, they may want to attempt to rank the results or perform additional processing. Grouper [ZE98, ZE99], for example, performs result clustering. While

[4] http://www.almaden.ibm.com/cs/k53/clever.html
[5] http://www.metacrawler.com/
[6] http://www.savvysearch.com/
[7] http://www.dogpile.com/

Inquirus [LG98b, LG98a] fetches all documents for analysis on full-text, a simpler version (perhaps with little available bandwidth) might decide to fetch only the most likely pages for further analysis. In this case, the meta-engine has only the information provided by the original search engines (usually just URL, title, and description), and the quality of these page descriptors is thus quite important to a post-hoc textual ranking or clustering of the pages.

2.4 Focused crawlers

Focused crawlers are web crawlers that follow links that are expected to be relevant to the client's interest (e.g. [CvdBD99, BSHJ+99, Men97, MB00, Lie97, RM99] and the query similarity crawler in [CGMP98]). They may use the results of a search engine as a starting point, or they may crawl the web from their own dataset. In either case, they assume that it is possible to find highly relevant pages using local search starting with other relevant pages. Dean and Henzinger [DH99] use a similar approach to find related pages.

Since focused crawlers may use the content of the current page, or anchor text to determine whether to expand the links on a page, our examination of nearby page relevance and anchor text relevance may be useful.

2.5 Intelligent Browsing Agents

There have been a variety of agents proposed to help people browse the web. Many of those that are content-based depend on the contents of a page and/or the text contained in or around anchors to help determine what to suggest to the user (e.g. [AFJM95, JFM97, Mla96, Lie95, Lie97, MB00, BS97, LaM96, LaM97]) or to prefetch links for the user (e.g. [Lie97, PP97, Dav99]).

By comparing the text of neighboring pages, we can estimate the relevance for pages neighboring the current one. We also find out how well anchor text describes the targeted page.

3 Experimental Method

3.1 Data Set

3.1.1 Initial Data Set

Ideally, when characterizing the pages of the WWW, one would choose a random set of pages selected across the Web. Unfortunately, while the Web has been estimated to contain hundreds of millions of pages [LG99], no one entity has a complete enumeration. Even the major search engines, with a few hundred million pages in their databases only know of a fraction of the web, and the pages retained in those datasets are biased samples of the Web. As a result, the unbiased selection of a random subset of the Web is an open question [BB98], although some progress is being made [HHMN00].

Accordingly, the data set used as the starting points in this paper were selected at random from a subset of the web. We randomly selected 100,000 pages out of the approximately 3 million pages that our local research search engine (DiscoWeb [DGK+99]) had crawled and indexed by early December 1999. The pages in the DiscoWeb dataset at that time were generated primarily from the results of inquiries made to the major search engines (such as HotBot[8] and AltaVista) plus pages that were in the neighborhood of those results (i.e. direct ancestors or descendants of pages in those results). Thus, selecting pages from this dataset will bias our sample toward pages

[8] http://www.hotbot.com/

in the neighborhood of high-ranking English-language pages (that is, pages near other pages that have scored highly on some query to a search engine).

3.1.2 Remaining Data Set

From the initial data set, we randomly selected one outgoing link per page and retrieved those pages. We also randomly reselected a different outgoing link per page (where possible) and fetched those pages as well. The latter set was used for testing anchor text relevance to sibling pages and to measure similarity between sibling pages.

3.1.3 Retrieval, Parsing, and Textual Extraction

The pages were retrieved using the Perl LWP::UserAgent library, and were parsed with the Perl HTML::TreeBuilder library. Text extraction from the HTML pages was performed using custom code that down-cased all terms and replaced all punctuation with whitespace so that all terms are made strictly of alphanumerics. Content text of the page does not include title or META tag descriptions, but does include alt text for images. URLs were parsed and extracted using the Perl URI::URL library plus custom code to standardize the URL format (down-casing host, dropping #, etc.) to maximize matching of equivalent URLs. The title (when available), description (when available), and non-HTML body text were recorded, along with anchor text and target URLs. The anchor text included the text within the link itself (i.e. between the <a> and), as well as surrounding text (up to 20 terms but never spanning another link). The basic representation of each textual item was bag-of-words with term frequency.

3.2 Textual Similarity Calculations

To perform the textual analysis, we used three straightforward calculations, which we describe in this section. While each of the measures can be applied to any pair of documents, we will sometimes use the term "query" when we refer to the "document" composed of the words in the source document (e.g. the title words, or description, or anchor text, or in general the first document of a pair).

Note that all measures have the following two properties: they produce scores in the range [0..1]; and identical documents generate a score of 1 while documents having no terms in common generate a score of 0.

3.2.1 TFIDF cosine similarity

The first calculation selected was TFIDF, for its widespread use and long history in information retrieval. Note that the IDF values are calculated from the documents in the combined retrieved sample, not over the entire Web. The specific formulas used were:

$$\mathrm{TFIDF}(w_i, X) = \frac{\mathrm{TF}(w_i, X) * \mathrm{IDF}(w_i)}{\sqrt{\sum_{all\ w}(\mathrm{TF}(w, X) * \mathrm{IDF}(w))^2}}$$

where

$$\mathrm{TF}(w, X) = \lg(\text{number of times } w \text{ appears in } X + 1)$$

and

$$\mathrm{IDF}(w) = \lg(\frac{\text{number of docs} + 1}{\text{number of docs with term } w})$$

So each document has the value 0 or a TFIDF value for each term, which are then normalized (divided by the sum of the values) so that the values of the terms in a document sum to 1. For document similarity, we use the cosine measure.

$$\text{TFIDF-Cos}(X,Y) = \frac{\sum_{all\ w} \text{TFIDF}(w,X) * \text{TFIDF}(w,Y)}{\sqrt{\sum_{all\ w} \text{TFIDF}(w,X)^2 * \sum_{all\ w} \text{TFIDF}(w,Y)^2}}$$

3.2.2 Term probability

The second measure is designed to measure the likelihood of a term in the "query document" being present in the target document. It is simply the sum of the fractions of the query corresponding to terms that are also present in the target document:

$$\text{Fract}(w,X) = \frac{\text{Number of times } w \text{ appears in } X}{\text{Number of terms in } X}$$

$$\text{Prob}(X,Y) = \sum_{all\ w} \begin{cases} \text{Fract}(w,X) & \text{if } w \in Y \\ 0 & \text{otherwise} \end{cases}$$

3.2.3 Document overlap

The third measure used was chosen to measure the amount of overlap of the two documents, after being normalized for differences in length. Thus, to calculate this measure we sum over all terms the smaller of the representative fractions of each document:

$$\text{Overlap}(X,Y) = \sum_{all\ w} \min(\text{Fract}(w,X), \text{Fract}(w,Y))$$

3.3 Experiments Performed

The primary experiments performed include measuring the textual similarity:

- of the title to its page, and of the description to its page
- of a page and one of its children
- of a page and a random page
- of two pages with the same immediate ancestor (i.e. between siblings) and with respect to the distance in the parent document between referring URLs
- of anchor text and the page to which it points
- of anchor text and a random page
- of anchor text and a page different from the one to which it points (but still linked from the parent page)

Additionally, we measured lengths of titles, descriptions (text provided in the description META tag of the page), anchor texts, and page textual contents. We also examined how often links between pages were in the same domain, and if so, the same host, same directory, etc.

We also performed experiments with stop word elimination and Porter term stemming [Por97], with similar results. Thus, for clarity of presentation, graphs for those cases are postponed to the appendix at end of this paper. No other feature selection was used (i.e., all terms were included).

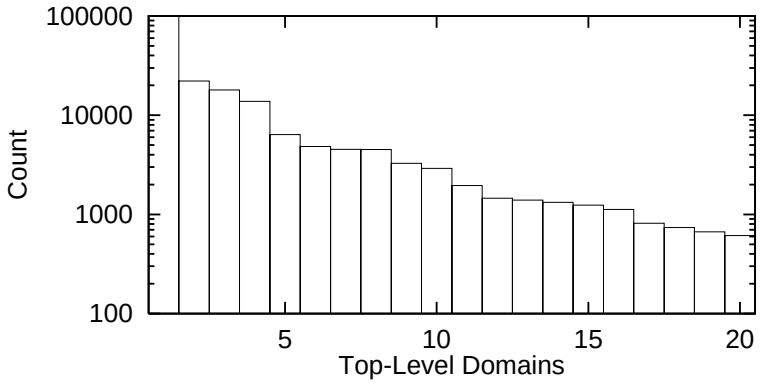


Figure 1: Representation of the twenty most common top-level domain names in our combined dataset, sorted by frequency. The top ten domains are .com, .edu, .org, .net, .uk, .de, .us, .ca, .gov, and .au.

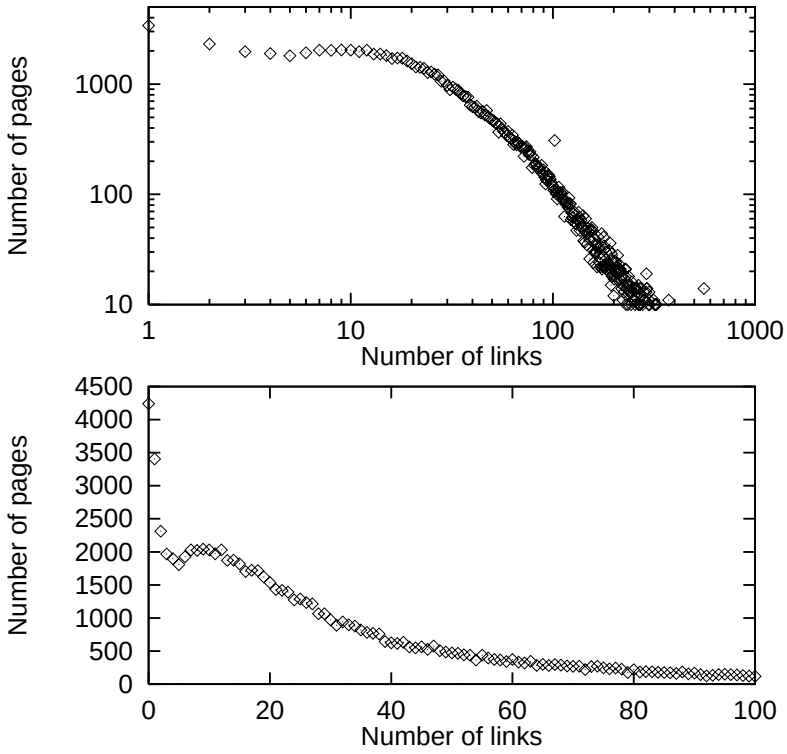


Figure 2: Two views of the distribution of the number of links per web page.

4 Experimental Results

4.1 General characteristics

For a baseline, we first consider characteristics of the overall dataset. Out of the initial 100,000 URLs selected, 89,891 were retrievable. An additional 111,107 unique URLs were retrievable by randomly fetching two distinct child links from each page of the initial set (whenever possible). The

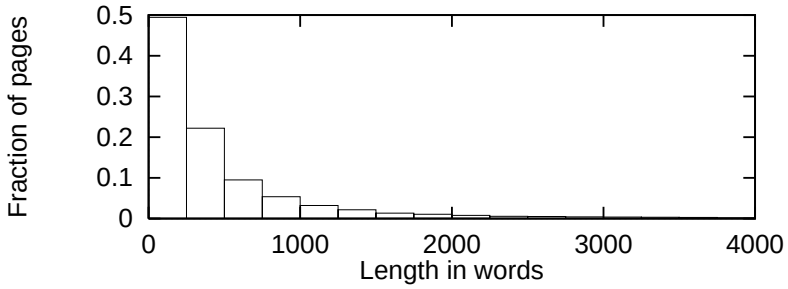


Figure 3: Distribution of content lengths of web pages.

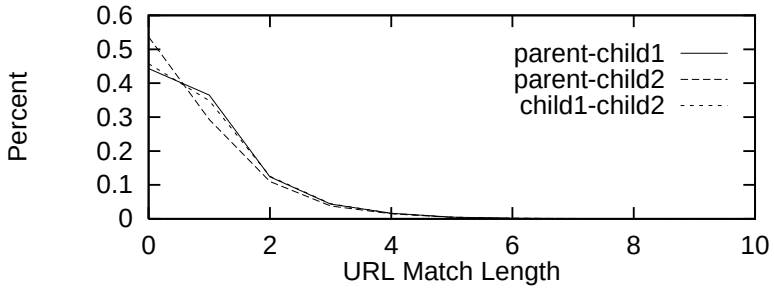


Figure 4: Distributions of URL match lengths are similar for parent-child1, parent-child2, and child1-child2 (siblings).

top five represented hosts were: www.geocities.com (561 URLs), www.webring.com (419 URLs), www.amazon.com (303 URLs), members.aol.com (287 URLs), and www.tripod.com (196 URLs). Combined, they represent less than 1% of the URLs used. Figure 1 shows the most frequent top-level domains in our data; close to half of the URLs are from .com, and another 26.8% of the URLs came from .edu, .org, and .net. Approximately 18% of the URLs represent top-level home pages (i.e. URLs with a path component of just /). The initial dataset contained a mean of approximately 49 links per page, with the distribution shown in figure 2.

With respect to content length, the sample distributions used for source and target pages are similar, so we present one distribution (pages from the initial dataset containing titles), shown in figure 3. Thus it can be seen that almost half of the web pages contain 250 words or less.

For pairings of pages with links between them, the domain name matched 55.67% of the time. For pairings of siblings, the percentage was 46.32%. For random pairings of pages, the domain name matched 0.003% of the time.

We also measured the number of segments that matched between URLs. A score of 1 means that the host name and port (more strict than just domain name matching) matched. For each point above 1, an additional path segment matched (i.e. top-level directory match would get 2; an additional subdirectory would get 3, and so on). The distributions of these segment match lengths for connected pages are shown in figure 4.

Figure 5 shows similarities for the author-supplied same-page descriptors (titles and description META tag contents). Descriptions show poorer performance than titles for term probabilities, suggesting that authors often include terms not present in the page being described. With longer text in descriptions than in titles, we find that descriptions have higher overlap with the content,

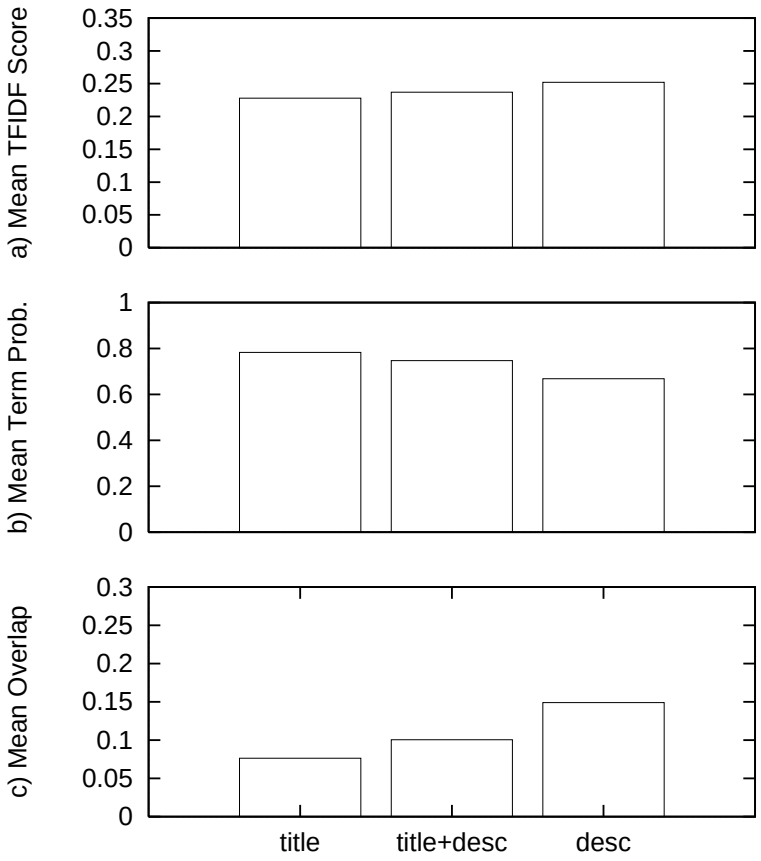


Figure 5: Similarity scores for title, description, and title+description as compared to the text on the same page. Comparisons between scores in a graph are significant ($p < .01$).

but not as much as the increased length of the description would suggest.

4.2 Page to page characteristics

Figure 6 presents the similarity scores of the current page to the linked page, to random pages, between sibling pages, and to subsets of the linked pages. All three metrics demonstrate that random page texts have almost nothing in common, linked page texts have more in common when the links are between pages of the same domain, and that sibling pages are more similar than linked pages of different domains.

In figure 7, we plot sibling page similarity scores as a function of distance between referring URLs in the parent page (where distances is the count of the number of links away). Thus, if two links (A,C) are separated by a third link (B), then C is a distance two away from A. We find that in general, the closer two URLs are, the more likely they are to share the same terms. This is most strikingly found for TFIDF-Cosine similarity, but it is present in all three metrics. This is corroborated by others [DH99, CDI98] who have observed that links to pages on similar topics are often clustered together on the parent page.

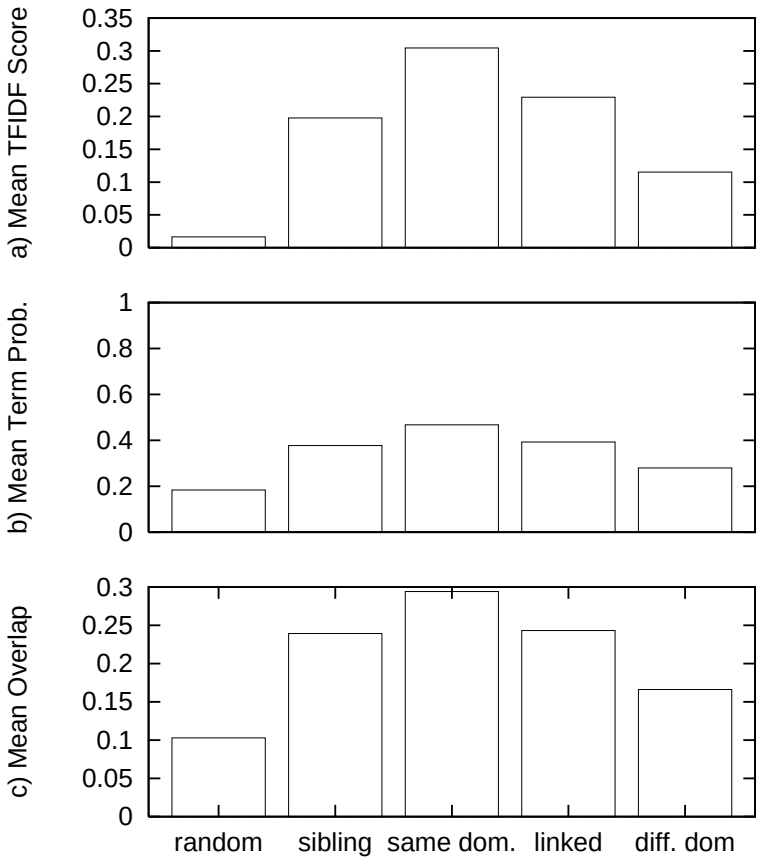


Figure 6: Textual similarity for linked pages, random pages, sibling pages, linked pages in the same domain, and linked pages in different domains. Comparisons between scores in a graph are significant ($p < .01$).

4.3 Anchor to page characteristics

Anchor text, by itself, has a mean length of 2.69 (distribution shown in figure 8) terms, slightly lower than the average reported by Amitay [Ami97]. In comparison, titles have a mean length of 5.27 terms (distribution shown in figure 9). However, we can also consider using text before or after the anchor text, and when we consider using up to 20 terms before and 20 terms after, we get a mean of 11.02 terms.

Figure 10 shows that anchor text scores much higher for non-random pages for each of the metrics. Even the similarity of anchor text to pages that are siblings of the targeted page get scores at least an order of magnitude better than random. There are also some conflicting results: in 10a and 10c, the highest scoring performance goes to anchor text to linked pages of a *different domain* than the source page, but this is not the case for term probabilities in 10b.

The mean TFIDF scores (figure 11a) for anchor text plus varying amounts of surrounding text are almost constant. While there is some improvement as more text is added, it is very small. The term probabilities (figure 11b), on the other hand, show a decline when additional words are used. Apparently the additional text provided has a much lower likelihood of being present in the target page. For example, the additional terms (.76 terms, on average) when allowing one additional word on each side of the anchor, have only a 51% chance of being in the target page (as compared to

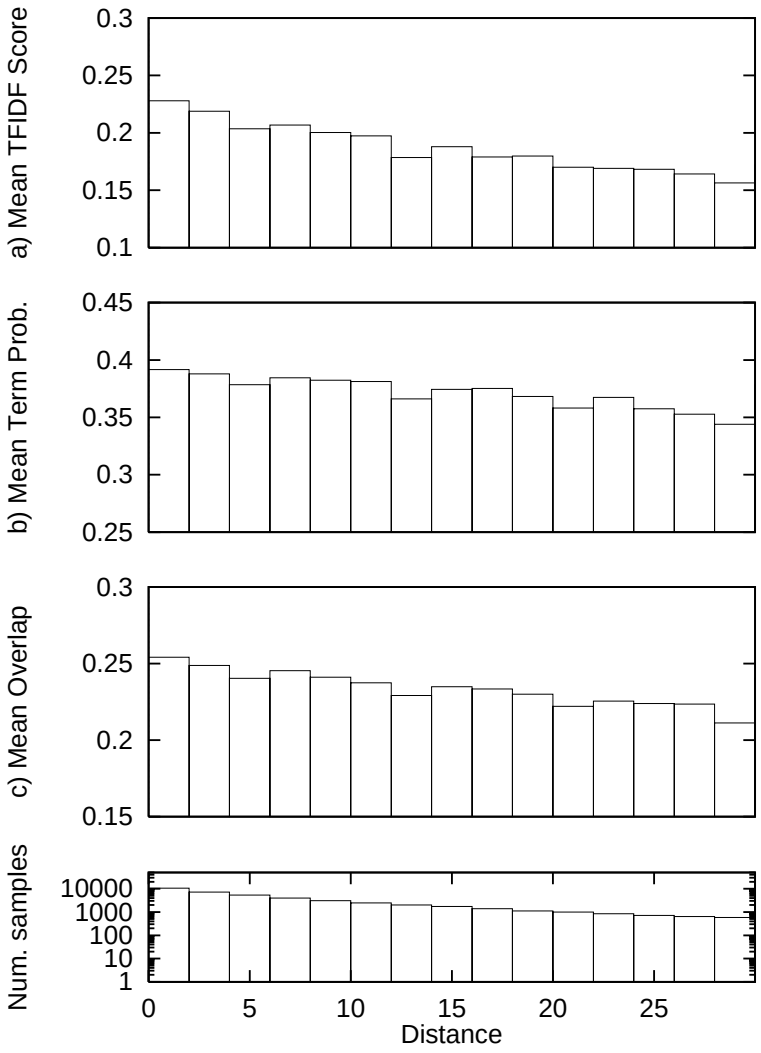


Figure 7: Plots of similarity scores between sibling pages as a function of distance between referring URLs in parent page for TFIDF, Term Probability, and Overlap, respectively. Each has a high negative correlation coefficient ($r \leq -.79$).

the 65% chance for anchor text terms). Unlike the others, overlap scores in figure 11c show some improvement as additional words are used.

While potentially confusing, these results are compatible to those reported by Chakrabarti et al. [CDR+98]. They found that including fifty bytes of text around the anchor would catch most references of the term "Yahoo" for a large dataset of links to the Yahoo home page[9]. Our interpretation is that while additional text does increase the chance of getting the important term(s), it also tends to catch more unimportant terms, lowering the overall term probability scores (as seen in 11b), but almost cancelling each other out in 11a. While these results may not be particularly encouraging, text surrounding the anchor is occasionally quite useful (especially for link text made of low-content terms like "click here").

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

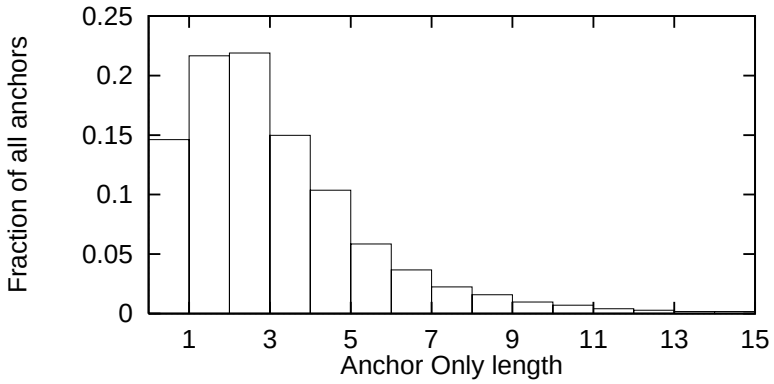


Figure 8: Distribution of the number of terms per anchor.

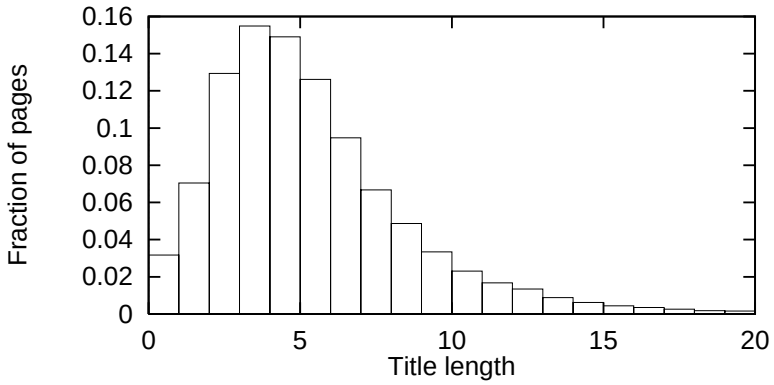


Figure 9: Distribution of the number of terms per title.

5 Conclusions

Text on the Web is not the same as text off the Web. Amitay [Ami97, Ami99] examines the linguistic choices that web authors use in comparison to non-hypertext documents. Without going into the same detailed analysis, we did find some similar characteristics of web pages. The bigrams "home page" and "click here" were the seventh- and thirteenth-most popular (of all raw bigrams), and certainly not typical bigrams of off-Web text. Interestingly, "all rights" and "rights reserved" were the eleventh- and twelfth-most popular, perhaps reflecting the increasing commercialization of the Web. Table 1 contains a list of the most frequent content-bearing bigrams.

This paper provides empirical evidence of topical locality of pages mirroring spatial locality in the Web — that is, WWW pages are typically linked to other pages with similar textual content. We found that pages are significantly more likely to be related topically to pages to which they are linked, as opposed to other pages selected at random, or other nearby pages. Furthermore, we found evidence of topical locality within pages, in that sibling pages are more similar when the links from the parent are closer together.

We also found that anchor text is most similar to the page it references, followed by siblings of that page, and least similar to random pages, and that the differences in scores are statistically significant ($p < .01$) and often large (an order of magnitude or more). This suggests that anchor

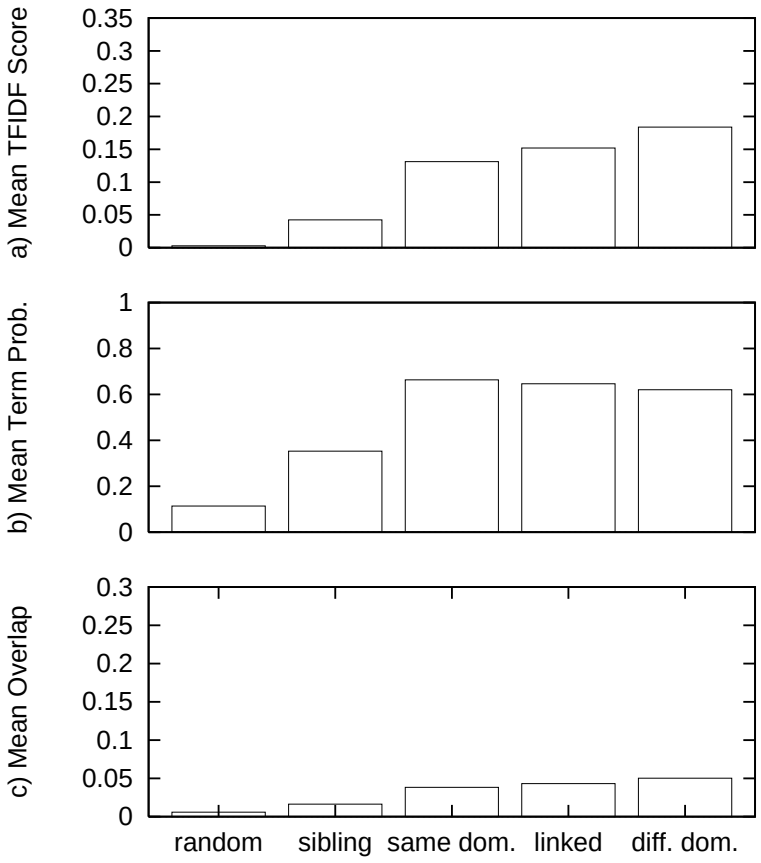


Figure 10: Performance of anchor text only to linked text, linked text in a different domain, linked text in the same domain, text of a sibling of the link, and the text of random pages. Comparisons between scores in a graph are significant ($p < .01$).

1) e mail	6) click here	11) http www	16) new york	21) that you
2) home page	7) web site	12) contact us	17) web sites	22) web page
3) if you	8) you can	13) last updated	18) world wide	23) your own
4) all rights	9) this site	14) more information	19) more than	24) united states
5) rights reserved	10) this page	15) check out	20) copyright 1999	25) wide web

Table 1: The twenty-five most common bigrams found after removing bigrams containing articles, prepositions, and various forms of the verb *to be*.

text may be useful in discriminating among unseen child pages. We note that anchor text terms can be found in the target page close to as often as the title terms on that target page, but that the titles also have better overlap and TFIDF cosine similarity scores. We have pointed out that on average the inclusion of text around the anchor does not particularly improve similarity measures (but neither does it hurt). Finally, we have shown that titles, descriptions, and anchor text all have relatively high mean term probabilities (and high mean TFIDF scores), implying that these page proxies represent at least part of the target page well.

Pitkow and Pirolli [PP97] have observed that "hyperlinks, when employed in a non-random format, provide semantic linkages between objects, much in the same manner that citations link

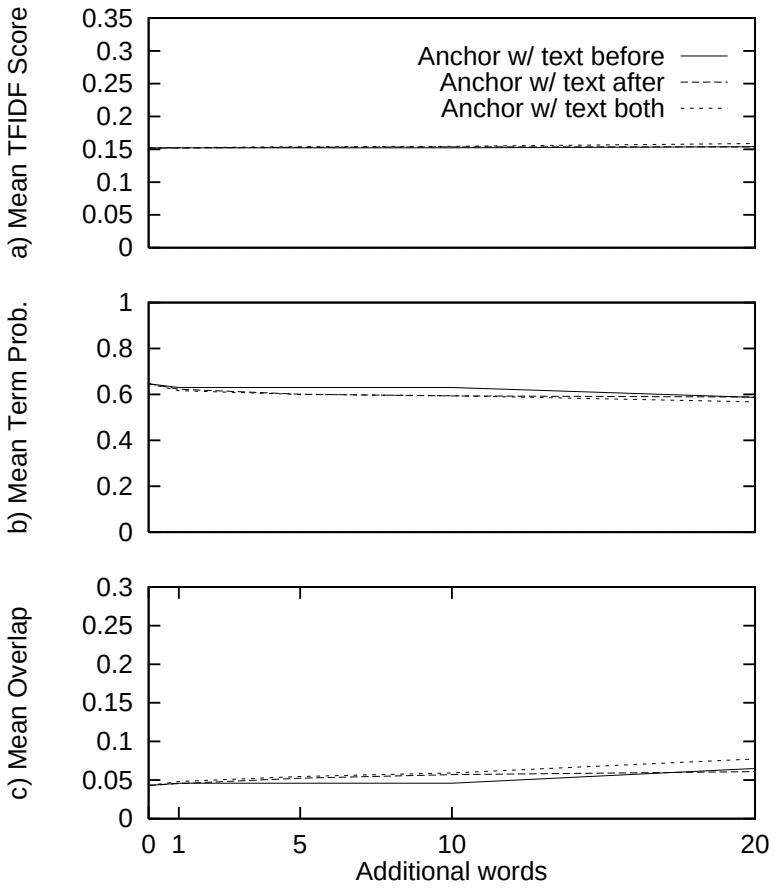


Figure 11: Performance of varying amounts of anchor text to linked text.

documents to other related documents." We have demonstrated that this semantic linkage, as approximated by textual similarity, is measurably present in the Web, thus providing the underpinnings for various web systems, including search engines, focused crawlers, linkage analyzers, and intelligent web agents.

As part of our future work, we plan to extend our analysis to include the textual similarity of a page to its grandchild, great-grandchild, etc. We would also like to differentiate between different types of links (e.g. navigational, advertising, semantic), and to use a better model for determining internal vs. external site links (rather than looking only for matching host or domain names).

Acknowledgments

Thanks are due to Haym Hirsh, Apostolos Gerasoulis, and Paul Kantor for their discussions and helpful comments on earlier drafts of this paper, and to the other members of the DiscoWeb group. This work is supported in part by DARPA under Order Number F425 (via ONR contract N6600197C8534).

References

[AFJM95] Robert Armstrong, Dayne Freitag, Thorsten Joachims, and Tom Mitchell. WebWatcher: A Learning Apprentice for the World Wide Web. In *Proceedings of the AAAI Spring Symposium on Information Gathering from Distributed, Heterogeneous Environments*, Stanford University, March 1995. AAAI Press.

[Ami97] Einat Amitay. Hypertext — The importance of being different. Master's thesis, Edinburgh University, Scotland, 1997. Also Technical Report No. HCRC/RP-94.

[Ami98] Einat Amitay. Using common hypertext links to identify the best phrasal description of target web documents. In *Proceedings of the SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web*, Melbourne, Australia, 1998.

[Ami99] Einat Amitay. Anchors in context: A corpus analysis of web pages authoring conventions. In Lynn Pemberton and Simon Shurville, editors, *Words on the Web - Computer Mediated Communication*. Intellect Books, UK, October 1999.

[BB98] Krishna Bharat and Andrei Broder. A technique for measuring the relative size and overlap of public Web search engines. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998.

[BFJ96] Justin Boyan, Dayne Freitag, and Thorsten Joachims. A Machine Learning Architecture for Optimizing Web Search Engines. In *AAAI Workshop on Internet-Based Information Systems*, Portland, OR, August 1996.

[BH98] Krishna Bharat and Monika R. Henzinger. Improved Algorithms for Topic Distillation in Hyperlinked Environments. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104–111, August 1998.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998.

[BS97] Marko Balabanovic and Yoav Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3), March 1997.

[BSHJ+99] Israel Ben-Shaul, Michael Herscovici, Michal Jacovi, Yoelle S. Maarek, Dan Pelleg, Menachem Shtalhaim, Vladimir Soroka, and Sigalit Ur. Adding support for dynamic and focused search with Fetuccino. In *Proceedings of the Eighth International World Wide Web Conference*, Toronto, Canada, May 1999.

[CDG+99] Soumen Chakrabarti, Byron E. Dom, David Gibson, Jon M. Kleinberg, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Mining the web's link structure. *IEEE Computer*, pages 60–67, August 1999.

[CDI98] Soumen Chakrabarti, Byron E. Dom, and Piotr Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of ACM SIGMOD*, Seattle, WA, 1998.

[CDR+98] Soumen Chakrabarti, Byron E. Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon M. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998.

[CGMP98] Junghoo Cho, Hector Garcia-Molina, and Lawrence Page. Efficient crawling through URL ordering. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998.

[CvdBD99] Soumen Chakrabarti, Martin van den Berg, and Byron E. Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Proceedings of the Eighth International World Wide Web Conference*, Toronto, Canada, May 1999.

[Dav99] Brian D. Davison. Adaptive Web Prefetching. In *Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW*, pages 105–106, Toronto, May 1999. Position paper. Proceedings published as Computing Science Report 99-07, Dept. of Mathematics and Computing Science, Eindhoven University of Technology.

[Dav00] Brian D. Davison. Topical locality in the Web. In *Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000)*, Athens, Greece, July 2000.

[DGK+99] Brian D. Davison, Apostolos Gerasoulis, Konstantinos Kleisouris, Yingfang Lu, Hyunju Seo, Wei Wang, and Baohua Wu. DiscoWeb: Applying Link Analysis to Web Search. In *Poster proceedings of the Eighth International World Wide Web Conference*, pages 148–149, Toronto, Canada, May 1999.

[DH97] Daniel Dreilinger and Adele E. Howe. Experiences with Selected Search Engines Using Metasearch. *ACM Transactions on Information Systems*, 15(3):195–222, July 1997.

[DH99] Jeffrey Dean and Monika R. Henzinger. Finding related pages in the world wide web. In *Proceedings of the Eighth International World Wide Web Conference*, pages 389–401, Toronto, Canada, May 1999.

[GKR98] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (Hypertext'98)*, 1998. Expanded version at http://www.cs.cornell.edu/home/kleinber/.

[HD97] Adele Howe and Daniel Dreilinger. SavvySearch: A MetaSearch Engine that Learns Which Search Engines to Query. *AI Magazine*, 18(2), 1997.

[HHMN00] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. On near-uniform url sampling. In *Proceedings of the Ninth International World Wide Web Conference*, Amsterdam, May 2000.

[JFM97] Thorsten Joachims, Dayne Freitag, and Tom Mitchell. WebWatcher: A Tour Guide for the World Wide Web. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 770–775. Morgan Kaufmann, August 1997.

[KABL96] T. Koch, A. Ardo, A. Brummer, and S. Lundberg. The building and maintenance of robot based internet search services: A review of current indexing and data collection methods. Prepared for Work Package 3 of EU Telematics for Research, project DESIRE; Available from http://www.ub2.lu.se/desire/radar/reports/D3.11/, September 1996.

[Kle98] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA-98)*, pages 668–677, San Francisco, CA, January 1998. Expanded version at http://www.cs.cornell.edu/home/kleinber/.

[LaM96] Brian A. LaMacchia. *Internet Fish*. PhD thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, 1996. Also available as AI Technical Report 1579, MIT Artificial Intelligence Laboratory.

[LaM97] Brian A. LaMacchia. The Internet Fish Construction Kit. In *Proceedings of the Sixth International World Wide Web Conference*, Santa Clara, CA, April 1997.

[LG98a] Steve Lawrence and C. Lee Giles. Context and Page Analysis for Improved Web Search. *IEEE Internet Computing*, 2(4):38–46, 1998.

[LG98b] Steve Lawrence and C. Lee Giles. Inquirus, the NECI meta search engine. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998.

[LG99] Steve Lawrence and C. Lee Giles. Accessibility of Information on the Web. *Nature*, 400:107–109, 1999.

[Lie95] Henry Lieberman. Letizia: An Agent That Assists Web Browsing. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 924–929, Montreal, August 1995.

[Lie97] Henry Lieberman. Autonomous Interface Agents. In *Proceedings of the ACM SIGCHI'97 Conference on Human Factors in Computing Systems*, Atlanta, GA, March 1997.

[MB00] Filippo Menczer and Richard K. Belew. Adaptive Retrieval Agents: Internalizing Local Context and Scaling up to the Web. *Machine Learning*, 39(2/3):203–242, 2000. Longer version available as Technical Report CS98-579, University of California, San Diego.

[McB94] Oliver A. McBryan. GENVL and WWWW: Tools for taming the Web. In *Proceedings of the First International World Wide Web Conference*, Geneva, Switzerland, May 1994.

[Men97] Filippo Menczer. ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery http://dollar.biz.uiowa.edu/~fil/papers.html. In *Proceedings of the 14th International Conference on Machine Learning (ICML97)*, 1997.

[Mla96] Dunja Mladenic. Personal WebWatcher: Implementation and Design. Technical Report IJS-DP-7472, Department of Intelligent Systems, J. Stefan Institute, Univ. of of Ljubljana, Slovenia, October 1996.

[Por97] Martin F. Porter. An algorithm for suffix stripping. In Karen Sparck Jones and Peter Willet, editors, *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco, 1997. Originally published in *Program*, 14(3):130-137 (1980).

[PP97] James E. Pitkow and Peter L. Pirolli. Life, Death, and Lawfulness on the Electronic Frontier. In *ACM Conference on Human Factors in Computing Systems*, Atlanta, GA, March 1997.

[RM99] Jason Rennie and Andrew McCallum. Efficient Web Spidering with Reinforcement Learning. In *In Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, 1999.

[SE95] Erik Selberg and Oren Etzioni. Multi-service search and comparison using the metacrawler. In *Proceedings of the Fourth International World Wide Web Conference*, Boston, MA, December 1995.

[SE97] Erik Selberg and Oren Etzioni. The MetaCrawler Architecture for Resource Aggregation on the Web. *IEEE Expert*, 12(1):8–14, Jan/Feb 1997.

[Spr99] Tom Spring. Search engines gang up on Microsoft. *PC World*, November 1999. Available online at http://www.pcworld.com/pcwtoday/article/0,1510,13760,00.html.

[Sul99] Danny Sullivan. More evil than Dr. Evil? From the Search Engine Report, at http://www.searchenginewatch.com/sereport/99/11-google.html, November 1999.

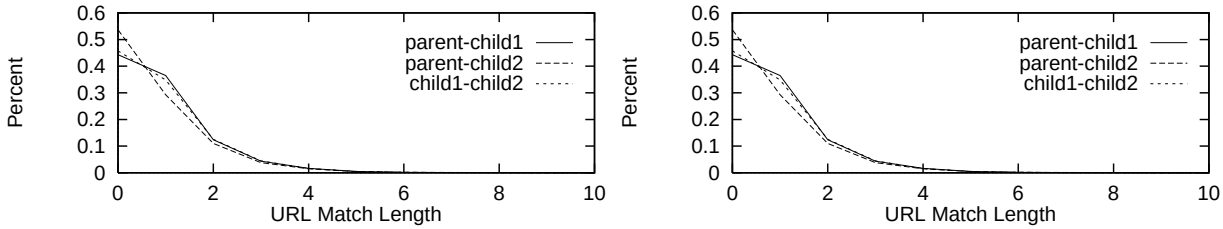
[Sul00] Danny Sullivan. Search engine features for webmasters. From Search Engine Watch, at http://www.searchenginewatch.com/webmasters/features.html, January 2000.

[ZE98] Oren Zamir and Oren Etzioni. Web Document Clustering: A Feasibility Demonstration. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[ZE99] Oren Zamir and Oren Etzioni. Grouper: a dynamic clustering interface to Web search results. In *Proceedings of the Eighth International World Wide Web Conference*, Toronto, Canada, May 1999.

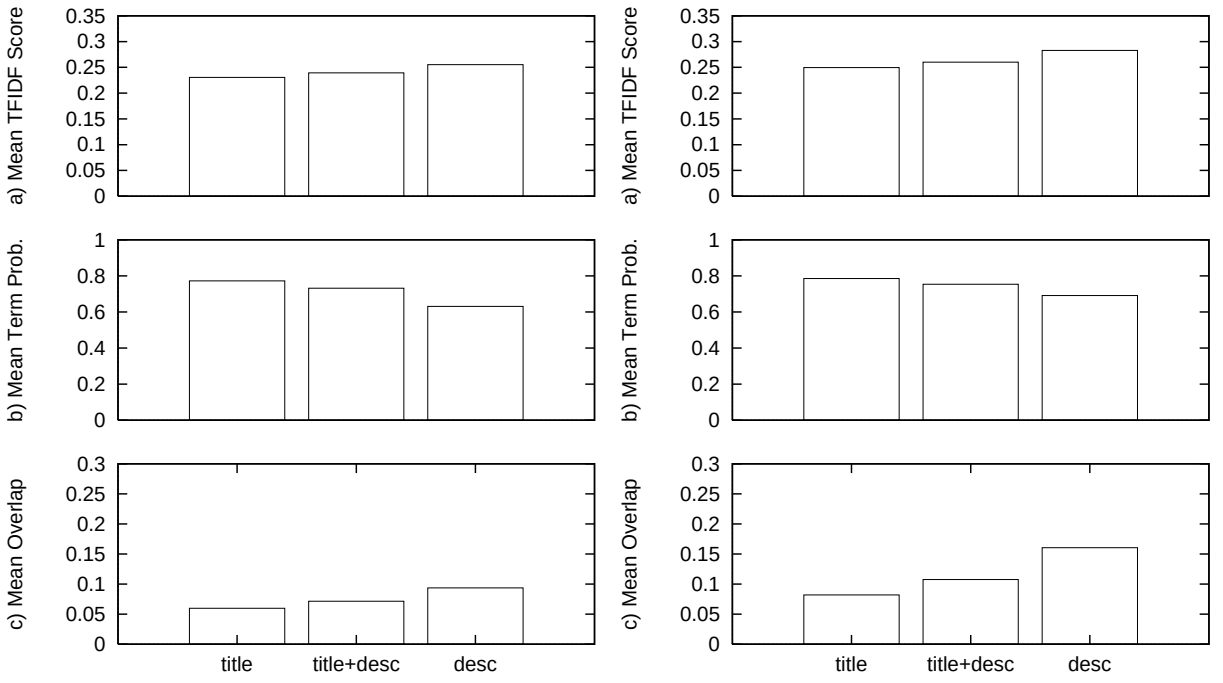
Appendix

In this section we present the results when using stop word elimination and Porter term stemming. Figures 12 through 16 match identically to the corresponding figures in the text, except for figure 15 which has different y-axis scaling.



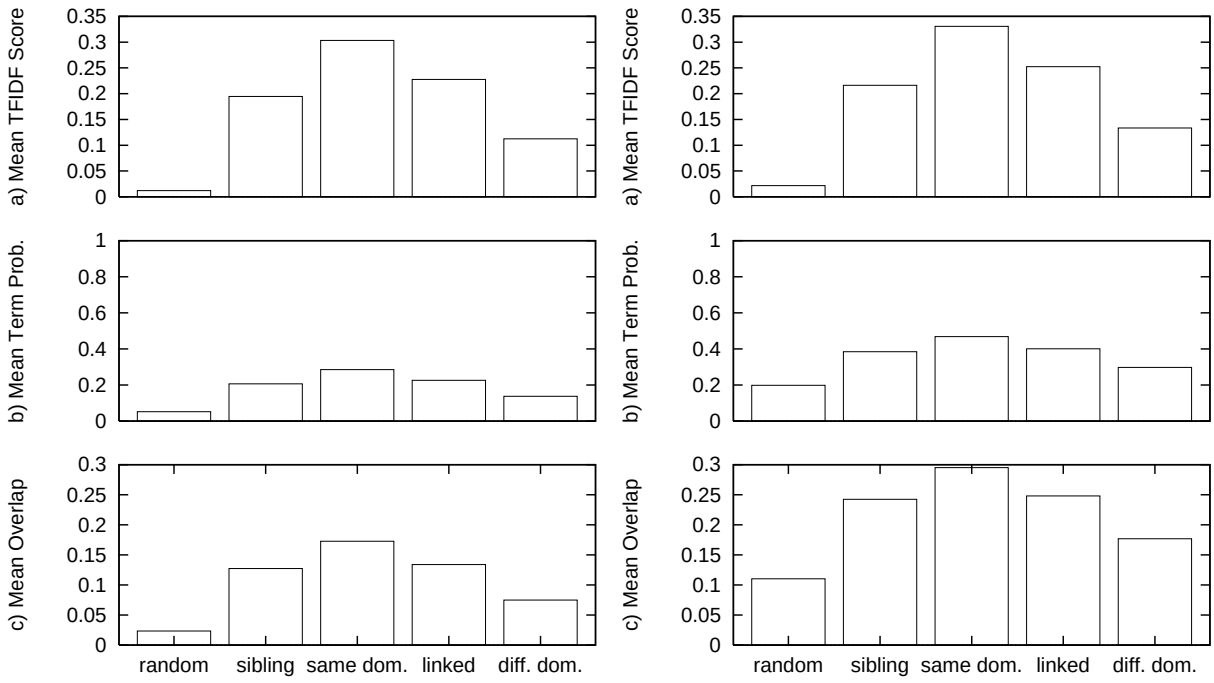
With stop word elimination. With term stemming.

Figure 12: Distributions of URL match lengths are similar for parent-child1, parent-child2, and child1-child2 (siblings).



With stop word elimination. With term stemming.

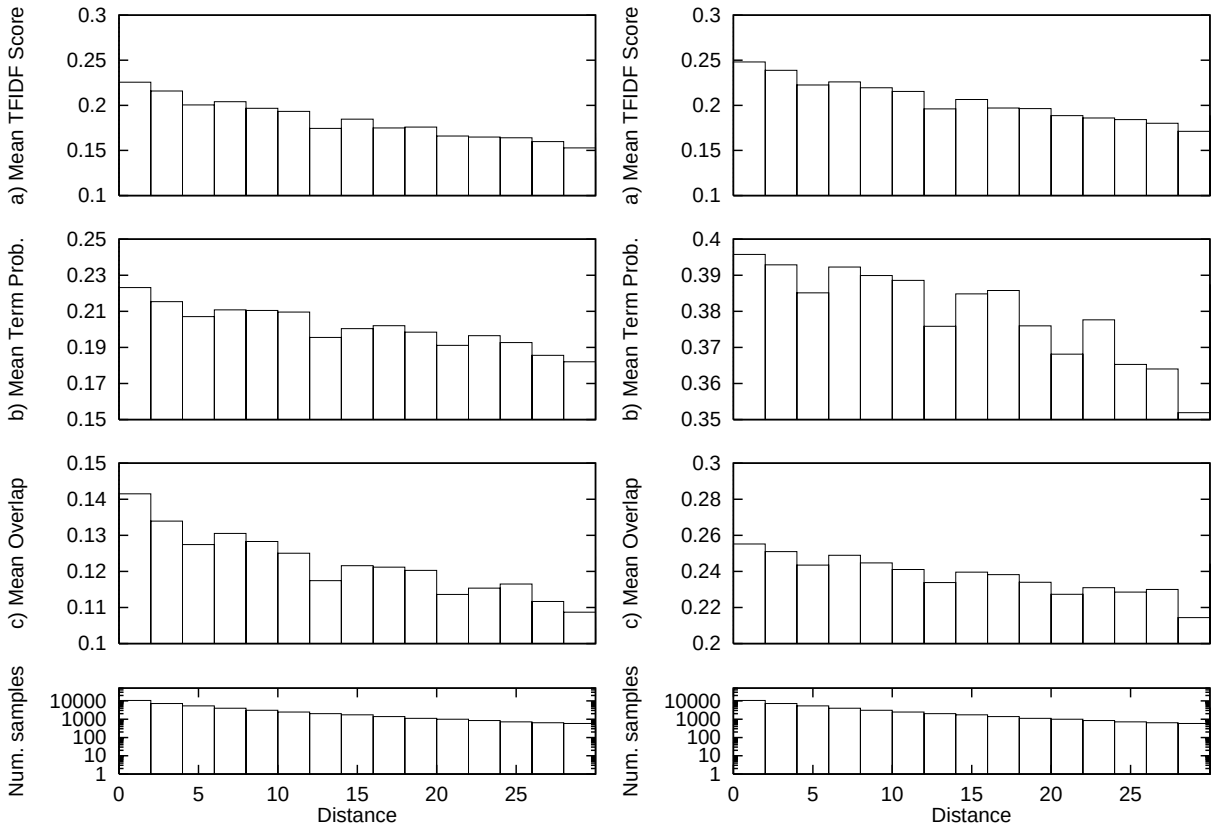
Figure 13: Similarity scores for title, description, and title+description as compared to the text on the same page.



With stop word elimination.

With term stemming.

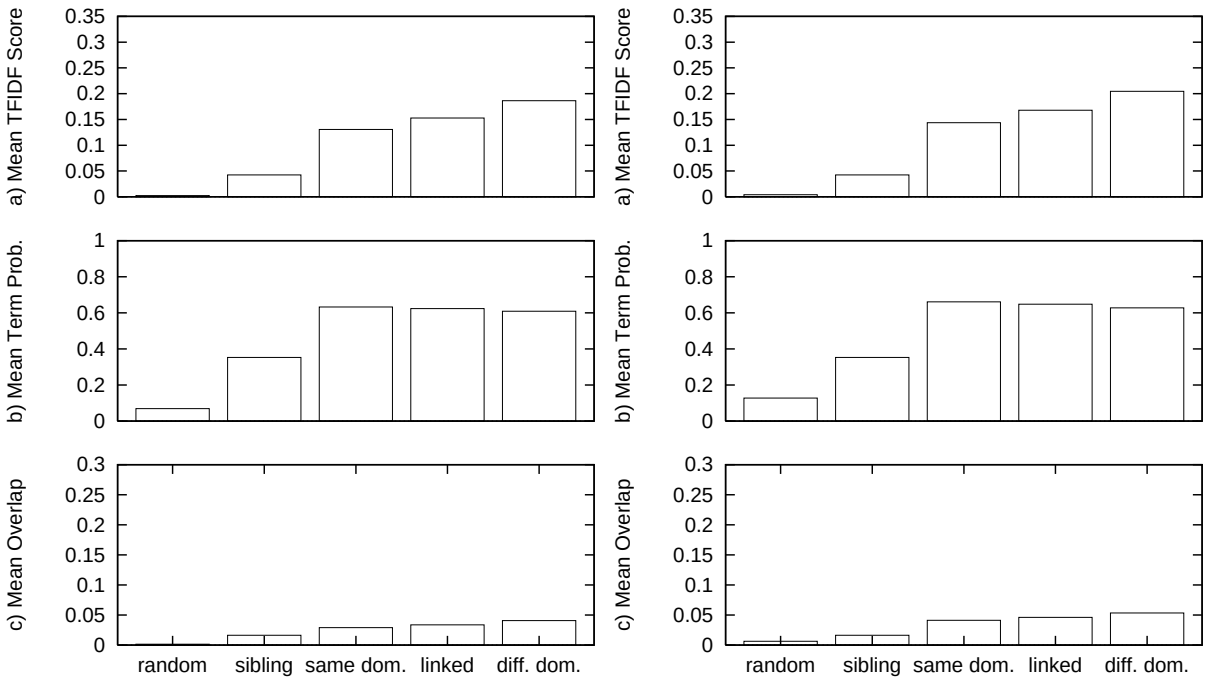
Figure 14: Textual similarity for linked pages, random pages, sibling pages, linked pages in the same domain, and linked pages in different domains.



With stop word elimination.

With term stemming.

Figure 15: Plots of similarity scores between sibling pages as a function of distance between referring URLs in parent page for TFIDF, Term Probability, and Overlap, respectively.



With stop word elimination.

With term stemming.

Figure 16: Performance of anchor text only to linked text, linked text in a different domain, linked text in the same domain, text of a sibling of the link, and the text of random pages.