Lexical and Semantic Clustering by Web Links

Filippo Menczer
Department of Computer Science, School of Informatics, Indiana University, Bloomington, IN 47408. E-mail: fil@indiana.edu

Recent Web-searching and -mining tools are combining text and link analysis to improve ranking and crawling algorithms. The central assumption behind such approaches is that there is a correlation between the graph structure of the Web and the text and meaning of pages. Here I formalize and empirically evaluate two general conjectures drawing connections from link information to lexical and semantic Web content. The *link-content conjecture* states that a page is similar to the pages that link to it, and the *link-cluster conjecture* that pages about the same topic are clustered together. These conjectures are often simply assumed to hold, and Web search tools are built on such assumptions. The present quantitative confirmation sheds light on the connection between the success of the latest Web-mining techniques and the small world topology of the Web, with encouraging implications for the design of better crawling algorithms.

Introduction

Search engines use a combination of information retrieval techniques and Web crawling algorithms to index Web pages. These allow users to search for indexed information by querying the resulting databases through Web interfaces. Although each search engine differentiates itself from the rest by offering some special feature, they all basically perform the same two functions: crawling (which includes indexing) and ranking (in response to queries). The most successful engines, apart from marketing issues, are those that achieve a high coverage of the Web, keep their index fresh, and rank search results in a way that correlates with the user's notion of relevance.

Ranking and crawling algorithms to date have used mainly two sources of information: words and links. Thinking of the Web as a physical space, one can associate word cues with a *lexical topology,* in which two pages are close to each other if they are similar in terms of their content. Similarity metrics of this sort are derived from the vector space model (Salton & McGill, 1983), that represents each document or query by a vector with one dimension for each term and a weight along that dimension that estimates the contribution of the corresponding term to the meaning of the document. Lexical topology therefore attempts to infer the semantics of pages from their lexical representation. The *cluster hypothesis* behind this model is that a document close in vector space to a relevant document is also relevant with high probability (van Rijsbergen, 1979). Lexical metrics have been traditionally used by search engines to rank hits according to their similarity to the query (Pinkerton, 1994).

Although lexical topology is based on the textual content of pages, *link topology* is based on the hypertextual components of Web pages—links. Link cues have traditionally been used by search engine crawlers in exhaustive, centralized algorithms. However the latest generation of Web search tools is beginning to integrate lexical and link metrics to improve ranking and crawling performance through better models of relevance. The best known example is the *PageRank* metric used by Google: Pages containing the query's lexical features are ranked using query-independent link analysis (Brin & Page, 1998). In this scheme, a page confers importance to other pages by linking to them. Links are also used in conjunction with text to identify hub and authority pages for a certain subject (Kleinberg, 1999), determine the reputation of a given site (Mendelzon & Rafiei, 2000), guide search agents crawling on behalf of users or topical search engines (Ben-Shaul et al., 1999; Chakrabarti, Punera, & Subramanyam, 2002; Chakrabarti, van den Berg, & Dom, 1999; Menczer & Belew, 2000; Menczer, Pant, Ruiz, & Srinivasan, 2001; Menczer, Pant, & Srinivasan, 2004), and identify Web communities (Flake, Lawrence, & Giles, 2000; Flake, Lawrence, Giles, & Coetzee, 2002; Gibson, Kleinberg, & Raghavan, 1998; Kumar, Raghavan, Rajagopalan, & Tomkins, 1999).

The assumption behind all of these retrieval, ranking, and crawling algorithms that use link analysis to make semantic inferences is a correlation between the Web's link topology and the meaning of pages. Thinking of the Web as a directed graph, one can define a distance metric based on the shortest path between two pages. A link-based analog of the cluster

Accepted January 23, 2004

© 2004 Wiley Periodicals, Inc. ● Published online 13 August 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20081

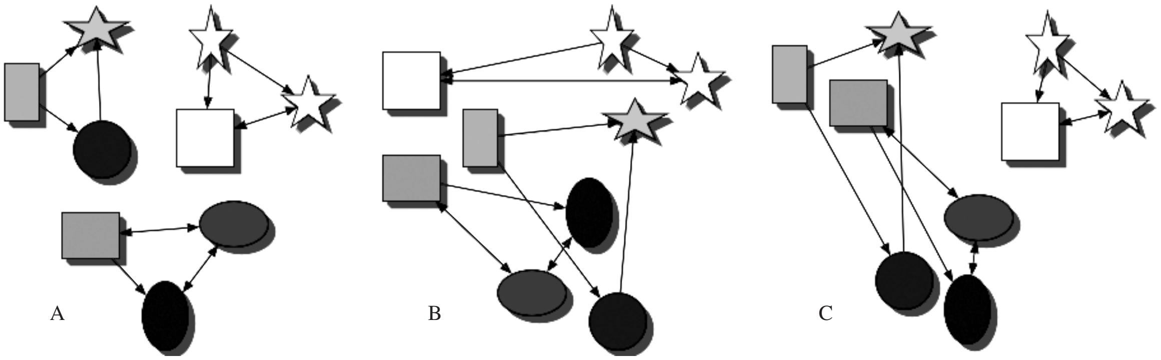

FIG. 1. Correlation among (A) link, (B) lexical, and (C) semantic topology.

hypothesis can be quantitatively stated as follow: Decreasing the number of links separating a page *p* from a relevant source increases the probability that *p* is also relevant. This *link-cluster conjecture* draws a connection from link topology to semantics—we can infer the meaning of a page by looking at the pages that link to it.

Figure 1 qualitatively illustrates the relationship between lexical, link, and *semantic* topology that is implied by the cluster hypothesis and link-cluster conjecture. In link space, pages with links to each other (represented as arrows) are close together, whereas in lexical space, pages with similar textual content (represented as shapes) are close to each other. Imagine a semantic space in which pages with similar meanings (represented as shades of gray) are clustered together. In such a space, a distance metric should be positively correlated with lexical distance (by the cluster hypothesis) and with link distance (by the link-cluster conjecture). The correlation between the distance metrics means that the semantic relationship is approximated by, and can be inferred from, both lexical and link cues.

In this article I formalize, quantitatively validate, and generalize the cluster hypothesis and link-cluster conjecture. These are empirical questions that may lead to a better understanding of the cues available to Web search agents and help build smarter search tools. Such tools will rely on local cues and thus will have the potential to scale better with the dynamic nature of the Web.

Background

This is by no means the first effort to draw a formal connection between Web topologies driven by lexical and link cues, or between either of these and semantic characterizations of pages. Recently, for example, theoretical models have been proposed to unify content and link generation based on latent semantic and link eigenvalue analysis (Achlioptas, Fiat, Karlin, & McSherry, 2001; Cohn & Hofmann, 2001). The more local flavor of the present formulation makes it easier to validate empirically.

Various forms of the cluster and link-cluster hypotheses have been implied, stated, or simply assumed in various studies analyzing the Web's link structure (Bharat & Henzinger, 1998; Chakrabarti et al., 1998; Dean & Henzinger, 1999; Gibson, Kleinberg, & Raghavan, 1998; Henzinger, 2000) as well as in the context of hypertext document classification (Chakrabarti et al., 1998; Chakrabarti, Dom, & Indyk, 1998; Kumar et al., 1999; Getoor, Segal, Taskar, & Koller, 2001). However, none of these studies consider empirical measures to quantitatively validate such hypotheses.

The textual similarity between linked pages has been analyzed by Davison (2000), who only considers page pairs separated by a single link. The present paper generalizes Davison's work to further link distances and characterizes how content relatedness decays as one crawls away from a start page.

The correlation between page meaning across links has been studied by Chakrabarti, Joshi, Punera, and Pennock (2002). In that study various page sampling strategies are considered. Some of them cannot be compared directly with the results of this paper or implemented in a Web crawler, because they rely on search engines to provide inlinks, and because they introduce random jumps to avoid the bias created by popular pages with many inlinks, such as www.adobe.com/products/acrobat/readstep2.html. Chakrabarti et al. (2002) do analyze one breadth-first crawl, but stop at depth 2. Here we extend their work by reaching depth 3. Another important difference is that Chakrabarti et al. (2002) use an automatic classifier to estimate the topics of crawled pages (in a predefined taxonomy) and then measure semantic distance based on the different classifications. Here I use a simpler conditional probability calculation to directly estimate the semantic similarity between pages in any topic and characterize how this relatedness decays as one crawls away from a start page.

Navigation models for efficient Web crawling have provided another context for our study of functional relationships between link probability and forms of lexical (Kleinberg, 2000; Menczer, 2002) or semantic similarity (Kleinberg, 2002; Menczer, 2002; Watts, Dodds, & Newman, 2002). I have also analyzed the dependence of link probability on lexical similarity to interpret the Web's emergent structure through a content-based growth model (Menczer, 2002; Menczer, 2004b).

The Link-Content Conjecture

The first step toward making a connection between lexical and link topologies is to note that given any pair of Web pages (p_1, p_2), we have well-defined distance functions δ_l and δ_t in link and lexical (text) space, respectively. To compute, $\delta_l(p_1, p_2)$, we use the Web hypertext structure to find the length, in links, of the shortest path from p_1 to p_2. There are a few caveats. First, this is not a metric distance because it is not symmetric in a directed graph; a metric version would be $\min(\delta_l(p_1, p_2), \delta_l(p_2, p_1))$, but for convenience δ_l will be referred to as "distance" in the remainder of the paper. Second, I intentionally consider only outlinks in the directed representation of the Web because this is how the Web is navigated—I do not assume that a crawler has knowledge of inlinks because that would imply free access to a search engine during the crawl. Third, this definition requires that we build a minimum spanning tree and therefore crawl pages in exhaustive breadth-first order. The large fan-out of Web pages therefore imposes a serious limit to the maximum δ_l that we can measure, in a practical sense.

To compute $\delta_t(p_1, p_2)$ we can use the vector representations of the two pages, where the vector components (weights) of page p, w_p^k are computed for terms k in the textual content of p, given some weighting scheme. One possibility would be to use Euclidean distance in this word vector space, or any other L_z norm:

$$\delta_t^z(p_1, p_2) = \left(\sum_{k \in p_1 \cup p_2} |w_{p_1}^k - w_{p_2}^k|^z \right)^{\frac{1}{z}}. \qquad (1)$$

However well-defined, L_z metrics have a dependency on the dimensionality of the pages; i.e., larger documents tend to appear more distant from each other than shorter ones. This is because of the fact that documents with fewer words have more zero weights (for words that are not included), which do not contribute to the distance. For this reason L_z distance metrics are not used in information retrieval. Instead, similarity measures are used, focusing on the words in the documents rather than absent ones. Therefore we define a distance measure based on the similarity between pages:

$$\delta_t(p_1, p_2) = \frac{1}{\sigma(p_1, p_2)} - 1 \qquad (2)$$

where $\sigma(p_1, p_2) \in [0, 1]$ is the similarity between the content of p_1 and p_2. Let us use the *cosine similarity* function (Salton & McGill, 1983), because it is a standard measure used in the information retrieval community:[1]

$$\sigma(p_1, p_2) = \frac{\sum_{k \in p_1 \cap p_2} w_{p_1}^k w_{p_2}^k}{\sqrt{\sum_{k \in p_1} (w_{p_1}^k)^2 \sum_{k \in p_2} (w_{p_2}^k)^2}}. \qquad (3)$$

We can now formally restate the cluster hypothesis.

Conjecture 1 *σ is anticorrelated with δ_l (link-content conjecture).*

The idea is to measure the correlation between the two distance measures across pairs of Web pages. The collection used for this purpose was obtained by starting from 100 topic pages in the Yahoo directory and performing a breadth-first crawl from each. Yahoo was selected as a starting hub owing to its wide popularity as a portal.

Figure 2 illustrates the data collection process. It is important to note that Yahoo was used to obtain seed pages for the crawls and approximate relevant sets, but Yahoo pages themselves were not part of the crawl data used in our analysis.

To obtain meaningful and comparable statistics at $\delta_l = 1$, only Yahoo pages with at least five external links were used to seed the crawls, and only the first 10 links for Yahoo pages with over 10 links. (These restrictions do not apply to any of the pages in the crawl.) Topics were selected in breadth-first order and therefore covered the full spectrum of Yahoo top-level categories. Each crawl reached a depth of $\delta_l = 3$ links from the start page and was stopped if 10,000 pages had been retrieved at the maximum depth. A timeout

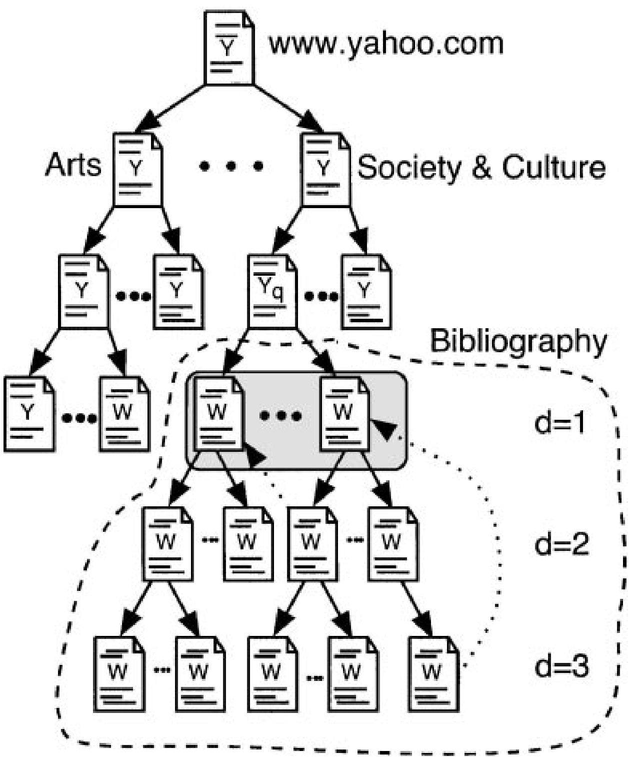

FIG. 2. Representation of the data collection. 100 topic pages were chosen in the Yahoo directory. Yahoo category pages are marked "Y," external pages are marked "W." The topic pages were chosen among "leaf" categories, i.e., without subcategories. This way the external pages linked by a topic page ("Y_q") represent the relevant set compiled for that topic by the Yahoo editors (shaded). In this example, the topic is SOCIETY CULTURE BIBLIOGRAPHY. Arrows represent hyperlinks and dotted arrows are examples of links pointing back to the relevant set. The crawl set for topic q is represented inside the dashed line.

[1] The remainder of the paper will focus on σ rather than δ_t due the intuitive familiarity of similarity measures.

of 60 seconds was applied for each page. The resulting collection comprised 376,483 pages. The text of each fetched page was parsed to extract links and terms; terms were conflated using a standard stemming algorithm (Porter, 1980).

A common TFIDF (term frequency—inverse document frequency) weighting scheme (Sparck Jones, 1972) was employed to represent each page in word vector space. This model assumes a global measure of term frequency across pages. To make the measures scalable with the maximum crawl depth (a parameter), inverse document frequency was computed as a function of distance from the start page, among the set of documents within that distance from the source. Formally, for each topic q, page p, term k, and depth d:

$$idf(k, d, q) = 1 + \ln\left(\frac{N_d^q}{N_d^q(k)}\right) \tag{4}$$

$$w_{p,d,q}^k = f(k, p) \cdot idf(k, d, q) \tag{5}$$

where N_d^q is the size of the cumulative page set $P_d^q = \{p : \delta_l(q, p) \leq d\}$, $N_d^q(k)$ is the size of the subset of pages in P_d^q containing term k, and $f(k, p)$ is the frequency of k in page p.

Correlation Between Lexical and Link Distance

The weights in Equation 5 were used in Equation 3 to compute the similarity $\sigma(q, p)$ between each topic q and each page in the set P_d^q. The link distances and the corresponding similarity measures were averaged over these cumulative page sets for each depth:

$$\delta(q, d) \equiv \langle\delta_l(q, p)\rangle_{P_d^q} = \frac{1}{N_d^q}\sum_{i=1}^{d} i \cdot (N_i^q - N_{i-1}^q) \tag{6}$$

$$\sigma(q, d) \equiv \langle\sigma(q, p)\rangle_{P_d^q} = \frac{1}{N_d^q}\sum_{p \in P_d^q} \sigma(q, p). \tag{7}$$

The 300 measures of $\delta(q, d)$ and $\sigma(q, d)$ from Equations 6 and 7, corresponding to 100 queries by 3 depths, are shown in the scatter plot of Figure 3. Note that the points are clustered around $\delta_l = 1, 2, 3$ because the number of pages at distance $\delta_l = d$ typically dominates P_d^q ($N_d^q \gg N_{d-1}^q$). The two metrics are indeed well anticorrelated (correlation coefficient $\rho = -0.76$). The two metrics are also predictive of each other with high statistical significance ($p < 0.0001$). This result quantitatively confirms the link-content conjecture.

Decay of Content Similarity

To analyze the decrease in the reliability of lexical content inferences with distance from the topic page in link space, one can perform a nonlinear least-squares fit of these data to a family of exponential decay models:

$$\sigma(\delta) \sim \sigma_\infty + (1 - \sigma_\infty)e^{-\alpha_1\delta^{\alpha_2}} \tag{8}$$

using the 300 points as independent samples. Here σ_∞ is the noise level in similarity, computed by comparing each topic

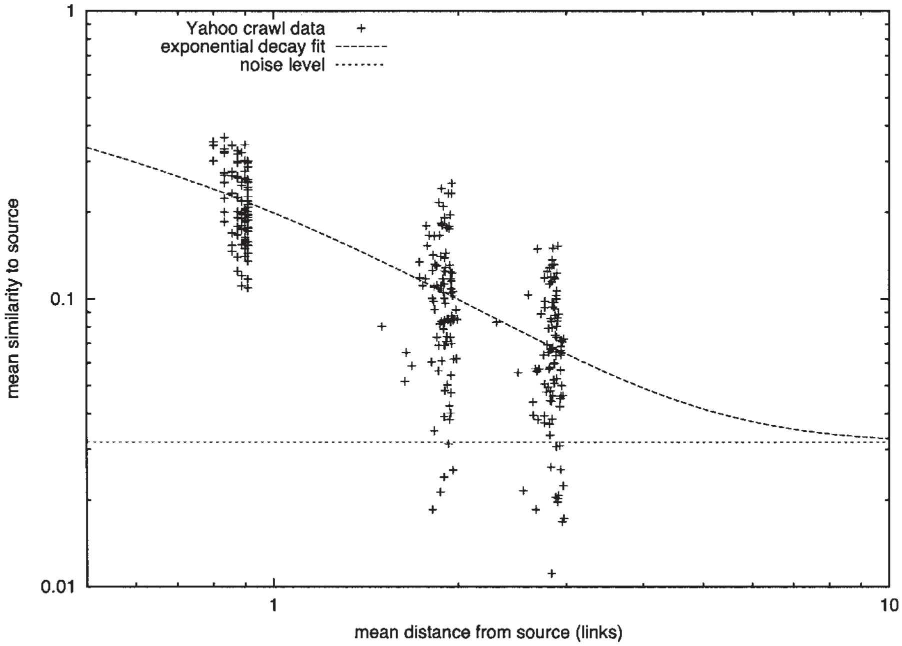

FIG. 3. Scatter plot of $\sigma(q, d)$ versus $\delta(q, d)$ for topics $q = 0, \ldots, 99$ and depths $d = 1, 2, 3$. An exponential decay fit of the data and the similarity noise level are also shown.

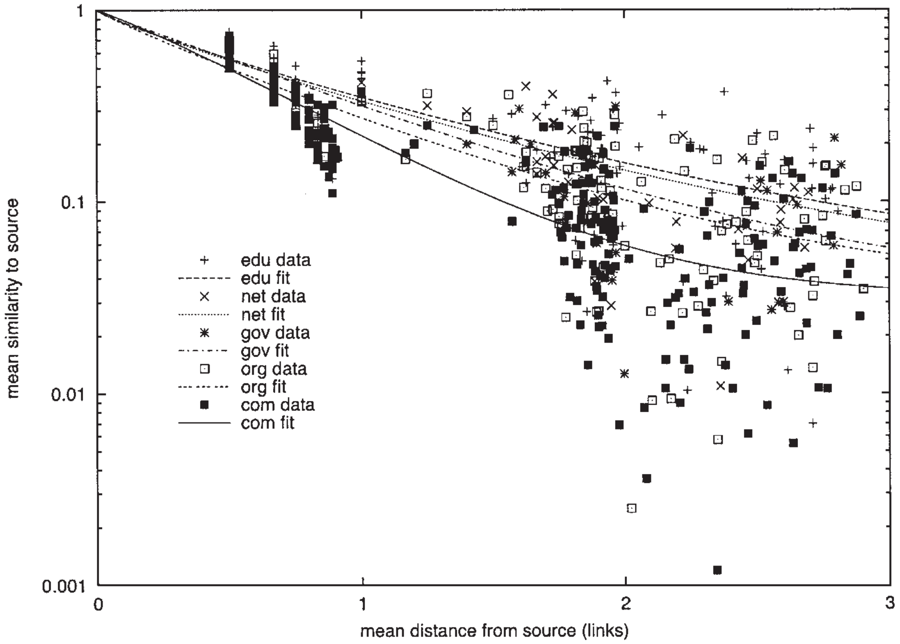

FIG. 4. Scatter plot of $\sigma(q, d)$ versus $\delta(q, d)$ for topics $q = 0, \ldots, 99$ and depths $d = 1, 2, 3$, for each of the major us top-level domains. An exponential decay fit is also shown for each domain.

page to external pages linked from different Yahoo categories:

$$\sigma_\infty \equiv \left\langle \frac{1}{N_1^{q'}} \sum_{p \in P_1^{q'}} \sigma(q, p) \right\rangle_{\{q,q':q \neq q'\}} \approx 0.0318 \pm 0.0006. \tag{9}$$

Note that while starting from Yahoo pages may bias σ ($\delta < 1$) upward, the decay fit is most affected by the constraint $\sigma(\delta = 0) = 1$ (by definition of similarity) and by the longer-range measures $\sigma(\delta > 1)$. The regression yields parametric estimates $\alpha_1 \approx 1.8$ and $\alpha_2 \approx 0.6$. The resulting fit is also shown in Figure 3, along with the noise level σ_∞. The similarity decay fit curve provides us with a rough estimate of how far in link space one can make inferences about lexical content.

Heterogeneity of Content Decay

The crawled pages were divided up into connected sets within top-level Internet (DNS) domains (.com, .gov, .edu, .uk, and so on). The resulting sets are equivalent to those obtained by breadth-first crawlers that only follow links to servers within each domain. The scatter plot of the $\delta(q, d)$ and $\sigma(q, d)$ measures for these domain-based crawls is shown in Figure 4. The plot illustrates the heterogeneity in the reliability of lexical inferences based on link cues across domains. The parameters obtained from fitting each domain data to the exponential decay model of Equation 8 estimate how reliably links point to lexically related pages in each domain. The parametric estimates are shown in Figure 5 together with a summary of the statistically significant differences among them. This result suggests that, for example, academic Web pages are better connected to each other than commercial pages in that they do a better job at pointing to other similar pages. Such a finding is not surprising considering the different goals of the two communities. This result can be useful in the design of general crawlers (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001) as well as topical crawling algorithms that prioritize links based on the textual context in which they appear; one could weight a link's context based on its site domain.

Domain	α_1	α_2
edu	1.11 ± 0.03	0.87 ± 0.05
net	1.16 ± 0.04	0.88 ± 0.05
gov	1.22 ± 0.07	1.00 ± 0.09
org	1.38 ± 0.03	0.93 ± 0.05
com	1.63 ± 0.04	1.13 ± 0.05

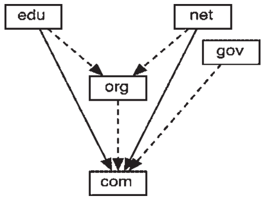

FIG. 5. (Left) Exponential decay model parameters obtained by nonlinear least-squares fit of each domain data, corresponding to the curves in FIG. 4, with asymptotic standard errors. (Right) Summary of statistically significant differences (at the 95% confidence level) between the parametric estimates; dashed arrows represent significant differences in α_1 only, and solid arrows significant differences in both α_1 and α_2.

The Link-Cluster Conjecture

The link-cluster conjecture has been implied or stated in various forms (Brin & Page, 1998; Chakrabarti et al., 1998; Davison, 2000; Dean & Henzinger, 1999; Gibson et al., 1998; Kleinberg, 1999). One can most simply and generally state it in terms of the conditional probability that a page p is relevant with respect to some query q, given that page r is relevant and that p is within d links from r:

$$R_q(d) \equiv \Pr[rel_q(p)|rel_q(r) \wedge \delta_l(r, p) \leq d] \qquad (10)$$

where $rel_q()$ is a binary relevance assessment with respect to q. In other words, $R_q(d)$ is the posterior relevance probability given the evidence of a relevant page nearby. $R_q(d)$ allows one to ask, Does a page have a higher than random chance of being related to a certain topic if it is within a few links of other pages on that topic? The simplest form of the link-cluster conjecture is stated by comparing $R_q(1)$ to the prior relevance probability G_q:

$$G_q \equiv \Pr[rel_q(p)] \qquad (11)$$

also known as the *generality* of the query. Finally, define a likelihood factor:

$$\lambda(q, d = 1) \equiv \frac{R_q(1)}{G_q}. \qquad (12)$$

If link neighborhoods allow for semantic inferences, then the following condition must hold:

Conjecture 2 λ (q, d = 1) > 1 *(weak link-cluster conjecture).*

To illustrate the importance of the link-cluster conjecture, consider a random crawler (or user) searching for pages about a topic q. Call $\eta_q(t)$ the probability that the crawler hits a relevant page at time t. One can define $\eta_q(t)$ recursively:

$$\eta_q(t + 1) = \eta_q(t) \cdot R_q(1) + (1 - \eta_q(t)) \cdot G_q. \qquad (13)$$

The stationary hit rate is obtained for $\eta_q(t - 1) = \eta_q(t)$. Solving Equation 13:

$$\eta_q^* = \frac{G_q}{1 + G_q - R_q(1)}. \qquad (14)$$

The weak link-cluster conjecture is a necessary and sufficient condition for such a random crawler to have a better than chance hit rate, thus bounding the effectiveness of the crawling (and browsing!) activity:

$$\eta_q^* > G_q \Leftrightarrow \lambda(q, 1) > 1. \qquad (15)$$

Definition 12 can be generalized to likelihood factors over larger neighborhoods:

$$\lambda(q, d) \equiv \frac{R_q(d)}{G_q} \xrightarrow{d \to \infty} 1 \qquad (16)$$

and a stronger version of the conjecture can be formulated as follows:

Conjecture 3 $\exists \delta^* > 1$ *s. t.* $\lambda(q, d) \gg 1$ *for* $\delta(q, d) < \delta^*$ *(generalized link-cluster conjecture)*

where δ^* is a critical link distance beyond which semantic inferences are unreliable.

Preservation of Relevance in Link Space

In 1997 I attempted to measure the likelihood factor $\lambda(q, 1)$ for a few queries and found that $\langle \lambda(q, 1) \rangle_q \gg 1$, but those estimates were based on very noisy relevance assessments (Menczer 1997). To obtain a reliable quantitative validation of the stronger link-cluster conjecture, such measurements were repeated on the larger and more recent data set from the crawl described in the previous section.

To estimate $R_q(d)$, one can use the relevant sets compiled by the Yahoo editors for each of the 100 topics:

$$R_q(d) \simeq \frac{|P_d^q \cap Q_q|}{N_d^q} \qquad (17)$$

where Q_q is the relevant set for q. In other words, we count the fraction of links out of a set that point back to pages in the relevant set. For G_q one can use:

$$G_q \simeq \frac{|Q'_q|}{|\cup_{q' \in Y} Q'_{q'}|}. \qquad (18)$$

Note that all of the relevant links for each topic q are included in Q'_q, even for topics where only the first 10 links were used in the crawl ($Q'_q \supseteq Q_q$), and the set Y in the denominator includes all Yahoo leaf categories. Finally the measures from Equations 17 and 18 were plugged into Definition 16 to obtain the $\lambda(q, d)$ estimates for $1 \leq d \leq 3$.

The 300 measures of $\lambda(q, d)$ thus obtained are plotted versus $\delta(q, d)$ from Equation 6 in the scatter plot of Figure 6. Closeness to a relevant page in link space is highly predictive of relevance, increasing the relevance probability by a likelihood factor $\lambda(q, d) \gg 1$ over the range of observed distances and queries.

Decay of Expected Relevance

We also performed a nonlinear least-squares fit of this data to a family of exponential decay functions using the 300 points as independent samples:

$$\lambda(\delta) \sim 1 + \alpha_3 e^{-\alpha_4 \delta^{\alpha_5}}. \qquad (19)$$

Note that this three-parameter model is more complex than the one in Equation 8 because $\lambda(\delta = 0)$ must also be estimated from the data ($\lambda(q, 0) = 1/G_q$). Further, the correlation between link distance and the semantic likelihood factor ($\rho = -0.1$, $p = 0.09$) is smaller than between link distance and lexical similarity. The regression yields parametric estimates $\alpha_3 \approx 1000$, $\alpha_4 \approx 0.002$ and $\alpha_5 \approx 5.5$. The resulting fit is also shown in Figure 6. Remarkably, fitting the data to the exponential decay model provides us

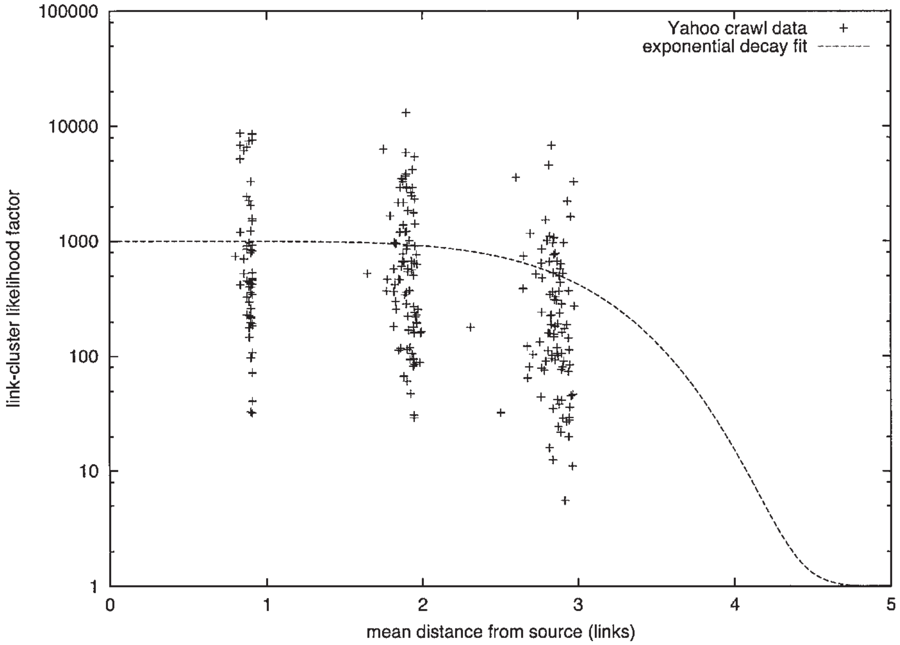

FIG. 6. Scatter plot of $\lambda(q, d)$ versus $\delta(q, d)$ for topics $q = 0, \ldots, 99$ and depths $d = 1, 2, 3$. An exponential decay fit of the data is also shown.

with quite a narrow estimate of how far in link space we can make inferences about the semantics (relevance) of pages, i.e., up to a critical distance δ^* between four and five links.

Implications for Topical Web Crawlers

To consider localized crawlers let us focus on the pages within a depth of $d = 1$ link. From Equation 14 we can quantify the relative increase in the hit rate of a random crawler over the chance rate:

$$\frac{\eta_q^*}{G_q} - 1 = \frac{R_q(1) - G_q}{1 + G_q - R_q(1)} \tag{20}$$

Using the 100 points from $d = 1$ sets as independent samples, we find that for the topics in our data set, simply starting from good seed pages gives a random crawler an advantage corresponding to a hit rate increase between 50% and 1000%. This increase is roughly linear in $\lambda(q, 1)$ indicating that the degree to which the link-cluster conjecture is valid for a particular topic has a significant impact on the performance of a random crawler searching for pages about that topic. Such an effect is likely to be amplified by smarter topical crawlers (Chakrabarti et al., 1999; Chakrabarti et al., 2002; Menczer & Belew, 2000; Menczer et al., 2001; Menczer et al., 2004).

Discussion

The main contributions and results of this paper are summarized as follows:

- The link-content and link-cluster conjectures have been formalized in a general way that characterizes the relationships between lexical and link topology and between semantic and link topology.
- The link-content conjecture has been empirically validated by quantifying the correlation between lexical and link distance.
- Lexical similarity displays a smooth exponential decay over a range of several links.
- Considerable heterogeneity was found in the reliability of lexical inferences based on link cues across Web communities broadly associated with server domains.
- The link-cluster conjecture has been empirically validated by showing that two pages are significantly more likely to be related if they are within a few links from each other.
- Relevance probability is preserved within a radius of about three links, then it decays rapidly.
- Being in the vicinity of relevant pages significantly affects the performance of a topical crawler.

There are a number of limitations that must be acknowledged in this study. First, it would be desirable to extend the present analysis to depths $d > 3$. Unfortunately, as already mentioned, the accurate measurement of link distances requires the knowledge of shortest paths and therefore the use of exhaustive breadth-first crawls. If we sampled the links in our crawls, we could reach greater distances but the link distance measurements would overestimate true distances because shortest paths would not be guaranteed. Therefore, given the exponential growth of the crawl set with d, the maximum depth is limited by our current computational and bandwidth resources.

A second limitation is our use of a popular directory such as Yahoo to identify the starting pages. This choice may boost the popularity of our seed pages (those linked from the Yahoo topic pages), perhaps leading to an overestimation of the posterior relevance probability R_q. While it is very difficult to reliably identify relevant sets for large numbers of topics on the Web without resorting to manually maintained directories such as Yahoo or the Open Directory, such an effect deserves further study.

Third, our analysis only considers pages found in forward crawls and thus we do not account for incoming links from pages that are not visited. One could imagine extending the analysis by considering the inlinks obtained from a search engine. Unfortunately this approach is made difficult by the fact that search engines typically limit access, even when access is facilitated by tools such as the Google API.[2] In an alternative approach (Menczer, 2004a) this limitation has been sidestepped by approximating the link distance via a neighborhood function that integrates cocitation (Small, 1973) and bibliographic coupling (Kessler, 1963). Such an approximation allows to map the relationship among independent content, link, and semantic similarity distributions across larger numbers of page pairs. Here we are limited by exhaustive breadth-first crawling—but for this price we obtain more reliable link distance measurements. Furthermore, the approach based on forward crawls makes our results most directly relevant for crawling applications.

With the above caveats in mind, the results of the measurements presented in this paper confirm the existence of a strong connection between the Web's link topology and its lexical and semantic content. In spite of the Web's decentralized structure, diversity of information, and freedom of content and style, hyperlinks created by Web authors create a signal that is detectable over the background noise within a distance of at least three links. There is remarkable agreement between this observation and the dramatic drop in performance displayed by adaptive crawling agents when the target pages are more than three links away from the start page (Menczer & Belew, 2000).

The results presented here provide us with a new way to interpret the success of algorithms such as PageRank (Brin & Page, 1998) and HITS (Kleinberg, 1999). In each of these, lexical topology is used as a filter to gather a set of potential pages, then link topology is used to identify the top resources (e.g., most relevant or authoritative pages). These techniques typically look within a small distance in link space (i.e., one or two links away) or rapidly converge if they recursively compute the eigenvector of a link adjacency matrix. This is consistent with the short range of the link neighborhoods in which significant lexical and semantic signals can be detected. If Web pages were not clustered in link space in a way that correlates with their meaning, link analysis would not help in identifying relevant resources.

The correlation between Web links and content takes on additional significance in light of link analysis studies that tell us the Web is a "small world" network with a power law distribution of link degree (Albert, Jeong, & Barabási, 1999; Barabási & Albert, 1999; Broder et al., 2000; Huberman & Adamic, 1999; Kumar et al., 1999). Small world networks have a mixture of clustered local structure and random links that create short paths between pages. The present results suggest that the Web's local structure may be associated with semantic clusters resulting from authors linking their pages to related resources.

The link-cluster conjecture may also have important normative implications for future Web search technology. The short paths predicted by the small world model can be very hard to navigate for localized crawling algorithms in the absence of geographic or hierarchical clues relating links to target pages (Kleinberg, 2000; Kleinberg, 2002; Watts et al., 2002). The results presented here suggest that the clues provided by links and words may be sufficient. While such theories are further explored elsewhere (Menczer, 2002), smart crawling algorithms exploiting textual and categorical associations between links and targets are being actively developed (Chakrabarti et al., 2002; Menczer et al., 2004; Pant & Menczer, 2002).

At a more general level, the present findings should foster the design of better search tools by integrating traditional search engines with topic- and query-driven crawlers (Menczer et al., 2001; Menczer et al., 2004) guided by *local* lexical and link clues. Because of the size and dynamic nature of the Web the traditional approach of keeping query processing separate from crawling, indexing and link analysis is efficient only in the short term and leads to poor coverage and recency (Brewington & Cybenko, 2000; Lawrence & Giles, 1999). Finite network resources imply a trade-off between coverage and recency. When crawling is not informed by the users, the trade-off can be very ineffective, for example, updating pages that few users care about while not covering new pages with a lot or potential interest. Closing the loop from user queries back to crawling will lead to more dynamic and scalable search engines that may better match the information needs of users (Pant et al., 2003).

Acknowledgments

The author is grateful to Dave Eichmann, Padmini Srinivasan, Nick Street, Alberto Segre, Rik Belew, and Alvaro Monge for helpful comments and discussions, and to Mason Lee and Martin Porter for contributions to the crawling and parsing code. This work is funded in part by NSF CAREER Grant No. IIS-0133124/0348940.

References

Achlioptas, D., Fiat, A., Karlin, A., & McSherry, F. (2001). Web search via hub synthesis. Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (pp. 500–509). Silver Spring, MD: IEEE Computer Society Press.

Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the World Wide Web. Nature, 401(6749), 130–131.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. ACM Transactions on Internet Technology, 1(1), 2–43.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286, 509–512.

Ben-Shaul, I., Herscovici, M., Jacovi, M., Maarek, Y., Pelleg, D., Shtalhaim, M., et al. (1999). Adding support for dynamic and focused search with Fetuccino. Computer Networks, 31(11–16), 1653–1665.

[2] http://www.google.com/apis/

Bharat, K., & Henzinger, M. (1998). Improved algorithms for topic distillation in hyperlinked environments. Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 104–111). New York: ACM Press.

Brewington, B., & Cybenko, G. (2000). Keeping up with the changing Web. IEEE Computer, 33(5), 52–58.

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

Broder, A., Kumar, S., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the Web. Computer Networks, 33(1–6), 309–320.

Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press.

Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., & Kleinberg, J. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. Computer Networks, 30(1–7), 65–74.

Chakrabarti, S., Joshi, M., Punera, K., & Pennock, D. (2002). The structure of broad topics on the Web. In D. Lassner, D. De Roure, & A. Iyengar (Eds.), Proceedings of the 11th International World Wide Web Conference (pp. 251–262). New York: ACM Press.

Chakrabarti, S., Punera, K., & Subramanyam, M. (2002). Accelerated focused crawling through online relevance feedback. In D. Lassner, D. De Roure, & A. Iyengar (Eds.), Proceedings of the 11th International World Wide Web Conference (pp. 148–159). New York: ACM Press.

Chakrabarti, S., van den Berg, M., & Dom, B. (1999). Focused crawling: A new approach to topic-specific Web resource discovery. Computer Networks, 31(11–16), 1623–1640.

Cohn, D., & Hofmann, T. (2001). The missing link—A probabilistic model of document content and hypertext connectivity. In T.K. Leen, T.G. Dietterich, & V. Tresp (Eds.), Advances in Neural Information Processing Systems, 13 (pp. 430–436). Cambridge, MA: MIT Press.

Davison, B. (2000). Topical locality in the Web. Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 272–279). New York: ACM Press.

Dean, J., & Henzinger, M. (1999). Finding related pages in the World Wide Web. Computer Networks, 31(11–16), 1467–1479.

Flake, G., Lawrence, S., & Giles, C. (2000). Efficient identification of Web communities. Proceedings of the Sixth ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 150–160). New York: ACM Press.

Flake, G., Lawrence, S., Giles, C., & Coetzee, F. (2002). Self-organization of the Web and identification of communities. IEEE Computer, 35(3), 66–71.

Getoor, L., Segal, E., Taskar, B., & Koller, D. (2001). Probabilistic models of text and link structure for hypertext classification. Proceedings of the IJCAI Workshop on Text Learning: Beyond Supervision.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring Web communities from link topology. Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia (pp. 225–234). New York: ACM Press.

Henzinger, M. (2000). Link analysis in Web information retrieval. IEEE Data Engineering Bulletin, 23(3), 3–8.

Huberman, B., & Adamic, L. (1999). Growth dynamics of the World-Wide Web. Nature, 401, 131.

Kessler, M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14, 10–25.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604–632.

Kleinberg, J. (2000). Navigation in a small world. Nature, 406, 845.

Kleinberg, J. (2002). Small-world phenomena and the dynamics of information. In T.G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, Cambridge, MA: MIT Press.

Kumar, S., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. Computer Networks, 31(11–16), 1481–1493.

Lawrence, S., & Giles, C. (1999). Accessibility of information on the Web. Nature, 400, 107–109.

Menczer, F. (1997). ARACHNID: Adaptive retrieval agents choosing heuristic neighborhoods for information discovery. Proceedings of the 14th International Conference on Machine Learning (pp. 227–235). San Francisco: Morgan Kaufmann.

Menczer, F. (2002). Growing and navigating the small world Web by local content. Proceedings of the National Academy of Science USA'99 (22), 14014–14019.

Menczer, F. (2004a). Correlated topologies in citation networks and the Web. European Physical Journal B., 38, 211–221.

Menczer, F. (2004b). The evolution of document networks. Proceedings of the National Academy of Science USA, 101, 5261–5265.

Menczer, F., & Belew, R. (2000). Adaptive retrieval agents: Internalizing local context and scaling up to the Web. Machine Learning, 39(2–3), 203–242.

Menczer, F., Pant, G., Ruiz, M., & Srinivasan, P. (2001). Evaluating topic-driven Web crawlers. In D.H. Kraft, W.B. Croft, D.J. Harper, & J. Zobel (Eds.), Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 241–249). New York: ACM Press.

Menczer, F., Pant, G., & Srinivasan, P. (2004). Topical Web crawlers: Evaluating adaptive algorithms. ACM Transactions on Internet Technology. In press.

Mendelzon, A., & Rafiei, D. (2000). What do the neighbours think? Computing Web page reputations. IEEE Data Engineering Bulletin, 23(3), 9–16.

Pant, G., Bradshaw, S., & Menczer, F. (2003). Search engine–crawler symbiosis. Proceedings of the European Conference on Digital Libraries (ECDL). Berlin, Germany: Springer Verlag.

Pant, G., & Menczer, F. (2002). MySpiders: Evolve your own intelligent Web crawlers. Autonomous Agents and Multi-Agent Systems, 5(2), 221–229.

Pinkerton, B. (1994). Finding what people want: Experiences with the WebCrawler. Proceedings of the Second International World Wide Web Conference.

Porter, M. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

Salton, G., & McGill, M. (1983). An introduction to modern information retrieval. New York: McGraw-Hill.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between documents. Journal of the American Society for Information Science, 42, 676–684.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28, 111–121.

van Rijsbergen, C. (1979). Information Retrieval (2nd ed., Chapter 3, pp. 30–31). London: Butterworths.

Watts, D., Dodds, V., & Newman, V. (2002). Identity and search in social networks. Science, 296, 1302–1305.