Measures of “Ignorance” on the Web

Siddhartha Reddy K.   Srinath Srinivasa   Mandar R. Mutalikdesai

International Institute of Information Technology
26/C, Electronics City, Hosur Road
Bangalore, India 560100
{siddhartha.reddy.k, sri, mandar}@iiitb.ac.in

Abstract

The hyperlink is a crucial element that makes up the web. While the hyperlink is seen as an element of recommendation, it also represents something more fundamental – namely, *knowledge.* A page A linking to another page B on the web, indicates *knowledge* about page B that the author of page A, wishes to make public. If either the author doesn't know about page B or does not wish to make this knowledge public, it contributes to “ignorance” on the web. This ignorance is on part of the user or a web crawler or a ranking algorithm considering page A. In this work we investigate whether hyperlinks reflect the latent knowledge in the web, by measuring how much they adhere to the well known “cluster hypothesis.” That is, if a user reading page A is very likely to read page B then page B is linked by a hyperlink from A. Various interpretations are considered for the term “very likely to read” and *ignorance* is measured as the extent to which the cluster hypothesis is violated. The results show a highly entrenched web, with distinct hyperlink-based communities of pages unaware of other related pages elsewhere on the web. The results also highlight the role of collaborative content-management environments like Wikipedia to mitigate this entrenchment effect.

1 Introduction

The hyperlink is the crucial element that forms the fabric of the web. It is an intrinsic part of several activities like browsing, crawling, ranking, classification, information filtering, etc.

While a hyperlink is widely seen as an element of endorsement, it also indicates something more fundamental: *knowledge.* If a page A links to page B, then it denotes that the author of page A knows about page B and *wishes to make this knowledge public.* Users and crawlers visiting page A will now know about page B if they didn't know about it already.

If the author of page A either did not know about the related page B or did not wish to make this knowledge public (maybe because B is a competitor page), we consider this to contribute to the “ignorance” on the web. This is especially so if page B is pertinent to page A, and if any user visiting page A is very likely to be interested in page B as well. It does not matter whether the absence of this hyperlink is intentional or accidental. It will affect the knowledge of users or crawlers browsing the web; or of ranking algorithms ordering their search results.

With a premise that the web comprises of (or *should comprise of*) semantically interrelated documents linked together by hyperlinks, ignorance on the web can be broadly seen as the amount of missed opportunities for connecting related data elements across the web.

Since document ranking in search engines today is predominantly based on hyperlinks, it is believed that ignorance is only growing on the web. This is variously called as the *entrenchment effect*, rich-getting-richer effect, *googlearchy*, etc. [5, 8, 9, 14].

A web page having a high PageRank would be very likely to be linked by new, upcoming web pages on the topic. This *preferential attachment* would contribute to increasing the PageRank of the page further, while the probability for pages down the line to obtain hyperlinks becomes lesser. This property of social networking where famous nodes become more famous is believed to cause the emergence of the power-law degree distribution on the web [2, 4].

However, there are theories that challenge the above notions, by noting the fact that users use *keyword search* to navigate the web, just as much as they use hyperlinks.[1]

Search terms provided on a search engine are also known to follow a power-law distribution with a long tail (cf. [3]). While some terms are searched very fre-

International Conference on Management of Data
COMAD 2006, Delhi, India, December 14–16, 2006

[1] See `http://www.seroundtable.com/archives/001896.html` for a report on a panel discussion on this topic at Search Engine Strategies 2005 Conference.

quently, a vast majority of terms in keyword search are infrequent terms that return results from small, niche areas of the web. This is claimed to offset the dominance of pages with very high PageRank and drive traffic to lesser known pages [13, 16].

Given this, a pertinent question to ask is whether analysing hyperlink structures in the web is still significant. There are several reasons why the answer is in the affirmative. They can be listed as follows.

1. The "egalitarian keyword search" hypothesis rests on the fact that most of the search terms are rarely-used terms, thus driving traffic to lesser known pages that use these terms. However, this would not apply to search terms that are common. Hence, for widely-popular search terms, keyword searching need not be egalitarian. Indeed, this seems to be already verified for political websites [14].

2. Analogously, just because most of the keyword search terms are rare, it does not necessarily mean that they drive traffic to lesser known parts of the web. Pages with high PageRank can very well have rarely occurring terms.

3. Many algorithms for searching similar pages, use the hyperlink neighbourhood of a page in order to look for similar pages [7, 11, 12]. On an entrenched web, providing a "Similar page" link (computed based on hyperlink neighbourhood) would not be very effective.

4. Crucial elements of keyword search engines like crawling and ranking, still rely on hyperlinks and on the assumption that hyperlinks connect related pages.

5. Finally, measures of ignorance help us understand the effectiveness of collaborative platforms like Wikipedia for sharing information.

Measuring ignorance can be seen as a dual of the task of measuring *topic drift* on the web [6, 10, 15]. While measuring topic drift seeks to assess how topically related are pages within a given hyperlink neighbourhood, "ignorance" measures how well connected are pages that are on the same topic.

2 Measuring Ignorance

In order to define "ignorance" on the web, we define the following terms. The web is treated as a graph $G = (V, E)$, where V is a set of pages and E is a set of hyperlinks across pages. The graph is directed, where $(v_1, v_2) \in E \not\Rightarrow (v_2, v_1) \in E$. With support for backlinks for spaces like blogs or Wikipedia and for the web in general, with the Google toolbar, the directed nature of the graph can be relaxed. However, we retain the directed property of the hyperlink graph since it is more general and can apply to situations where backlinks are not available.

The predicate $Related(v_1, v_2)$ is defined for all $v_1, v_2 \in V$ as: *if a user is interested in* v_1 *then the user is very likely to be interested in* v_2 *as well.*

Ignorance can be broadly defined as:

$$\forall v_i, v_j \in V,\ 1 - Pr[(v_i, v_j) \in E | Related(v_i, v_j)] \quad (1)$$

where $Pr[x]$ is the probability of the occurrence of an event, x.

On the web graph, pages need not be directly connected to be reachable. There needs to be only a hyperlink path between pages to reach one from the other. The *hyperlink path distance* between pages v_1 and v_2 is the smallest k such that $(v_1, v_2) \in E^k$. Here E^k is the k^{th} transitive closure of E. The term $hd(v_1, v_2)$ is used to define the hyperlink path distance between pages.

Ignorance is now defined by a more general definition as:

$$\forall v_i, v_j \in V,\ 1 - Pr[hd(v_i, v_j) = k | Related(v_i, v_j)] \quad (2)$$

However, the presence of a hyperlink path from page u to page v does not necessarily mean that a user on page u will definitely reach page v. At each step in the path, the user has to choose one outgoing link from among all the outgoing links on that page. With this in mind, a third definition of ignorance can be given as the probability that the user on page u will reach page v by following hyperlinks beginning at page u.

In order to define ignorance this way, we introduce the following terms. Given $hd(u, v) = k$, the term $path(u, v)$ is defined as the sequence $u, v_1, v_2, \ldots v_{k-1}v$ as the sequence of pages comprising of the shortest path between u and v.

Secondly, suppose that page u is directly linked to v, the term $nav(u, v)$ is defined as the probability that a user visiting page u would click on the hyperlink pointing to page v. In this work, we compute this term in a way similar to PageRank computations, as follows: $nav(u, v) = \frac{1}{|u_*|}$, where $u_* = \{(u, v) | (u, v) \in E\}$ is the set of all outgoing hyperlinks from u.

Given these, ignorance is now defined as:

$$\forall u, v \in V, 1 - \Pi_{i=0}^{k} nav[(v_i, v_{i+1}) | hd(u, v) = k \wedge Related(u, v)] \quad (3)$$

where $v_0 = u$, $v_k = v$ and $\forall v_i, v_i \in path(u, v)$.

In order to define the predicate $Related()$ in a concrete fashion, we take the following interpretations for the term:

Cosine Similarity: Two pages v_i and v_j are said to be related if their cosine similarity is very high.

Co-citation: Two pages v_i and v_j are said to be related if they are co-cited by a large number of other pages.

Tagging: Two pages v_i and v_j are said to be related if they are tagged by a large number of common tags by a large number of users.

Cosine similarity is a measure of relatedness in terms of the contents of the pages, while co-citation is a measure of relatedness in terms of the hyperlink structure among pages. It would be interesting to see how ignorance measured using one relatedness measure compares with ignorance measured using the other relatedness measure.

The third variety of relatedness measure, namely tagging, is an increasing amount of meta-data available from bookmark harvesting engines like del.icio.us. While we have conducted experiments based on the third interpretation as well, we have not been able to obtain enough data in order to compare it with the first two interpretations. Hence in this paper, we shall be primarily focusing on relatedness based on content similarity and based on co-citation.

3 Data Sets

The ignorance experiment was performed on three kinds of datasets:

1. Wikipedia dump
2. Web crawl from Ask.com
3. Analyzing tags from the del.icio.us[2] bookmark sharing engine

Characteristics of these data sets are explained in the following subsections. For measuring hyperlink path distance, a maximum depth of $k = 8$ was used. We define this as k_{max}. This is based on the result by Baeza-Yates and Castillo [1] that users rarely browse beyond a depth of 5. It may be noted that in the charts further on, "a hyperlink path distance greater than k_{max}" does not imply the existence of a hyperlink path. It only means that there does not exist a hyperlink path with length less than or equal to k_{max}.

3.1 Wikipedia

We downloaded a Wikipedia dump of all the article pages in the English language, in December 2005. This dump consists of 2,185,443 pages and 2,283,601 links. This includes main article pages, image pages (a page for every image with additional information and links to the parent article), disambiguation pages (pages with links to pages on different topics represented by the same name), etc.

No. of pairs with $S \geq 0.9$	Total no. of pairs
1	8429
3	10039
0	9877
2	17683
4	11181

Table 1: Content-wise similar pages in Wikipedia

Wikipedia is unlike the rest of Web, in that its content evolves from thousands of people editing and adding information collaboratively. A notable observation is that the more popular pages in Wikipedia are accessed much more than the less popular ones and are consequently more "current." In these pages, if there are missing references, they are quickly added in.

3.2 Ask.com Crawl

We obtained a part of the general Web crawl of the Ask.com search engine from January 2006. This data set consists of about 10,623,000 pages and 85,812,128 links.

3.3 Del.icio.us

Del.icio.us is a service for storing and sharing web bookmarks or tags. Users can "tag" a page using key phrases that they think describe the page best. Del.icio.us also facilitates retrieval of pages that have been tagged with a given key phrase. We used the tags for analyzing popular pairs of related pages on the Web.

A description of the experimental procedures along with the results is presented below.

4 Experiments: Wikipedia

4.1 Cosine Similarity

We began by extracting key phrases from every Wikipedia page and indexing them. Key phrases were extracted from the anchor text of hyperlinks (and Wiki links) and phrases embedded within markup like headings, bold face, italics, bullets, etc.

In order to get an idea of how content-wise similarity is distributed across Wikipedia, we conducted several experiments. For each experiment, we picked anywhere between 8000 to 17000 pairs of pages uniformly at random and calculated the cosine similarity, S, of each pair. Figure 1 shows the result of one such sampling, where cosine similarity values S are plotted against their frequency, F. The cosine similarity values across all these pairs follow a Poisson distribution. This can be observed in the $\ln(S)$ vs $\ln(F)$ plot shown in figure 2. As is evident, a majority of pairs of pages have very low cosine similarity.

For our purposes, we set a cosine threshold of $S \geq 0.9$ for determining content-wise similarity.

[2] http://del.icio.us

Table 1 shows the number of content-wise similar pages that were found in various sampling experiments. On an average, the distribution of content-wise similar pages was found to be 0.0178%.

Given the extremely small percentage of pairs of pages having a cosine of at least 0.9, it would be safe to assume that whenever we find any such pair, they are likely to contain semantically related content as well.

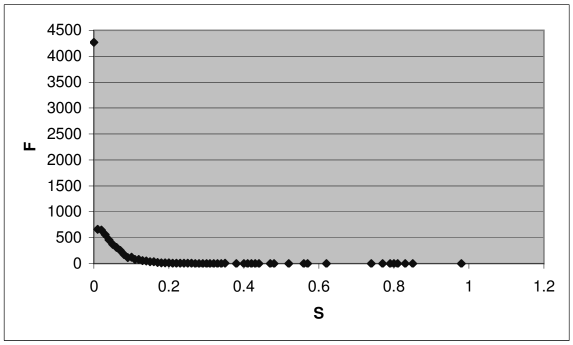


Figure 1: Wikipedia: Cosine Similarity (S) vs Frequency (F)

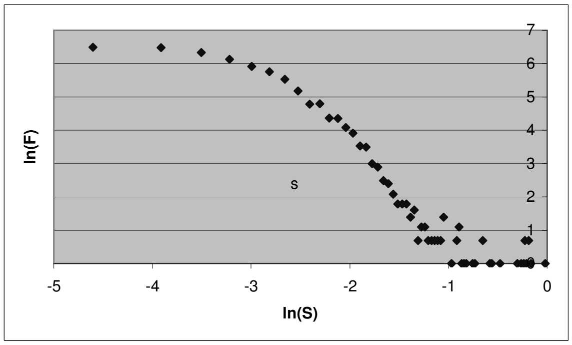


Figure 2: Wikipedia: $\ln(S)$ vs $\ln(F)$

With the index in place, we picked a page u from Wikipedia uniformly at random. With the aid of the index, we found other pages v, with a high Jaccard coefficient: $\frac{|term(u) \cap term(v)|}{|term(u) \cup term(v)|} \rightarrow 1$, where $term(u)$ is the set of all key phrases of page u. For page u and each such page v, we defined $Related(u, v) = true$ if $S(u, v) \geq 0.9$, where $S(u, v)$ is the cosine similarity between pages u and v. For pairs of pages (u, v) where the above predicate holds, we proceeded to find $hd(u, v)$.

Results: The experiment was repeated with 65 pairs of pages having a cosine similarity of 0.9 or more. While 65 seems like a very small number, given the observation that the number of content-wise similar pages are about 0.0178%, 65 pairs of pages would correspond to a sample space of about 365,170 pairs or 730,340 pages. This constitutes a sizeable proportion of the entire Wikipedia dump.

Figure 3 shows a plot of the hyperlink path distance against the frequency of its occurrence among these 65 pairs. Here, the mode is 5, with 37 of the 65 pairs of pages having this hyperlink path distance. The next most frequent hyperlink path distance is 4, with a frequency of 12. There are 5 pairs with a hyperlink path distance beyond k_{max}.

The mode 5, seems to be a very arbitrary number. Most pairs of pages that are content-wise similar are reachable within a distance of k_{max} hops, and most of them are separated by a distance of 5. We discuss this strange result and our interpretation in section 4.3.

For these 65 pairs of pages, a measure of ignorance was also calculated based on equation 3. On an average, the probability of a random surfer reaching one content-wise similar page from another was found to be 0.0021, giving us a high ignorance measure of 0.9979.

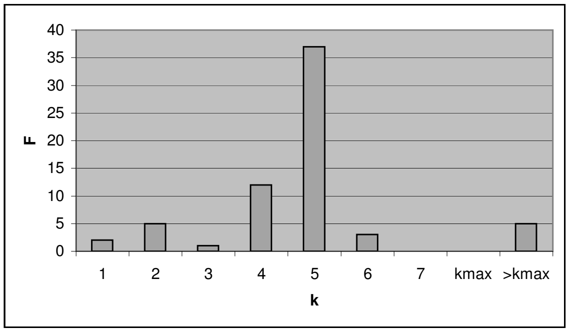


Figure 3: Wikipedia (Cosine Similarity): Hyperlink path distance (k) vs Frequency (F)

4.2 Co-citations

We indexed all the links in the Wikipedia data set. We then picked pairs of pages (u, v) uniformly at random from the data set, and calculated the number of times they have been co-cited, N. We collected more than a million such pairs of pages. The distribution of each N against its frequency F, is a power-law, as shown in figures 4 and 5. It is evident from figure 4 that a majority of pairs of pages have been co-cited very few times.

It is interesting to note here that, while distribution of content across pages follows a Poisson distribution, the co-citation distribution is a power-law. Pois-

No. of pairs with $N(u, v) \geq 1500$	Total no. of pairs
5	1336643
26	1186231
2	1462987
6	1312479
9	1173765

Table 2: Co-citation similarity in Wikipedia

son distributions are indicative of independent random events, while power-law distributions are characteristic of systems having non-linear influences across disparate random events. A page with a given content seems to be created independently, while existence of hyperlinks seems to influence the formation of other hyperlinks.

For this experiment, we defined $Related(u, v) = true$ if $N(u, v) \geq 1500$, where $N(u, v)$ is the number of times pages u and v have been co-cited. Table 2 shows the proportion of related pages found across various sampling experiments. On an average, related pages based on co-citation was found to be 0.00078% of the sample.

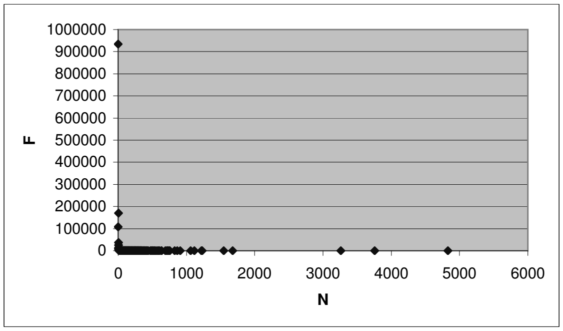


Figure 4: Wikipedia (Co-citations): Number of co-citations (N) vs Frequency(F)

Results: We calculated the hyperlink path distance between 14,004 pairs of pages. Figure 6 shows a plot of the hyperlink path distance against the frequency of its occurrence among these 14,004 pairs. Here, the mode is 2, with 9,747 of the 14,004 pairs of pages having this hyperlink path distance. The next most frequent hyperlink path distance is 1, with a frequency of 4,122. The maximum hyperlink path distance observed was 5, with a frequency of 4.

The hyperlink separation between pairs of highly co-cited pages is very much lower than the hyperlink separation between content-wise similar pages. However, it is interesting to note that mode distance here is 2 and not 1. Most co-cited pages seem to not directly link to one another, but *are linked via an intermediary.*

On an environment like Wikipedia, we imagine that this intermediary could be one of the co-citing pages itself, that is acting as an index. A page that co-cites a pair of pages would be a page on a general topic (like “Delhi”), while the pages that are co-cited would be about specific topics (like say, “History of Delhi” and “Educational Institutions in Delhi”). Pages on the specific topics would have linked back to the general page, giving them a short path to all other pages related by co-citation.

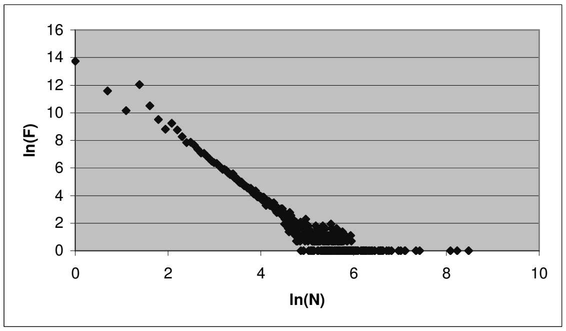


Figure 5: Wikipedia (Co-citations): $\ln(N)$ vs $\ln(F)$

However, even with such a small degree of separation, the average probability that a user can browse from one co-cited page to another was found to be only 0.0007. This gives us an ignorance value of 0.9993.

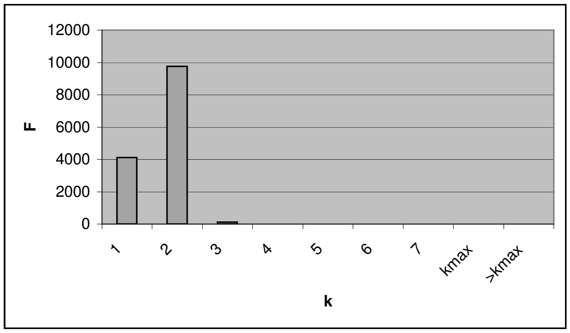


Figure 6: Wikipedia (Co-citations): Hyperlink path distance (k) vs Frequency (F)

4.3 Interpretation of Results

In our experiments using cosine similarity as a relatedness measure, we found that a majority of the pairs of pages are linked with a hyperlink path of length less than k_{max} with a mode of 5. Very few of the related pairs were separated by a distance greater than k_{max}.

This can be seen as an indication that Wikipedia, being a collaborative medium, is much more densely connected compared to the general Web. In section 5.1, we shall see that, on the Web, content-wise similar pages are generally disconnected. Wikipedia in general is much more densely connected and probably has a much smaller diameter than the web. However, a small degree of separation between related pages, does not necessarily mean that users visiting one are likely to browse to the other.

Even if content-wise similar pages are separated within k_{max} degrees of separation, the ignorance factor was still very high: about 0.9979. This high value of ignorance stems from the dense nature of hyperlink connectivity, giving the user several choices to browse away from a page before reaching a content-wise similar page.

However, what is even more interesting is that while the degree of separation between co-cited pages is very low (with a mode of 2), the ignorance value for the co-citation experiments is actually a little higher than that of content-wise similar pages! The ignorance value in the co-citation experiment was found to be 0.9993, while for the cosine experiment, it was 0.9979.

So what does it mean to have a low degree of separation, and a high value of ignorance?

A low degree of separation among co-cited pages means that such pairs of pages are likely to form communities that are discovered by crawlers and ranking algorithms. A similarity search algorithm searching a hyperlink neighbourhood is very likely to chance upon co-cited pages. This is because crawling and automated browsing are likely to follow all hyperlinks from a page. However, the same cannot be said of human browsers who are more likely to click on only a small number of hyperlinks from any given page. Hence, high ignorance and low degree of separation for co-cited pages indicates the presence of communities or clusters of pages, that are too intricate to explore manually, but which can be exploited by search engines.

For content-wise similar pages, not only is the ignorance high, but so is the degree of separation. However, assuming that similarity search algorithms search in hyperlink neighbourhoods till a depth of k_{max}, the high degree of separation (with mode=5) on Wikipedia is still amenable for automated techniques to find similar pages.

5 Experiments: Ask.com

5.1 Cosine Similarity

The data-set from Ask.com comprised 10,623,000 pages from a web crawl. In order to find content-wise similar pages, we conducted several sampling experiments choosing pairs of pages at random. Because of the way the crawl data was organized and the lack of an index, we found it difficult to sample a large number of pages at one go. Hence, a number of experiments were conducted with smaller samples.

No. of pairs with $S \geq 0.9$	Total no. of pairs
65	9957
79	10875
5	993
7	998
2	997
8	995

Table 3: Content-wise similar pages from Ask.com crawl

Table 3 shows results from some of the experiments. The number of pairs of pages having a high cosine ($S \geq 0.9$) was found to be about 0.5981%. Figure 7 shows a plot of each S against its frequency, F. If we ignore the obvious mode at cosine 0, the distribution seems to characterize a Gaussian curve with a mean somewhere around 0.5.

It is also somewhat of a surprise that the proportion of content-wise similar pages on the general web is higher than the corresponding proportion found on Wikipedia. However, this can also be explained by the fact that two or more largely similar pages on Wikipedia are likely to be merged by users over time, whereas on the web, there is a larger tendency to duplicate data and mirror page contents.

Hence it is quite plausible for the general web to contain greater proportion of content-wise similar pages. Whatever be the reason, we shall not be speculating on this further, and would go by what the numbers say.

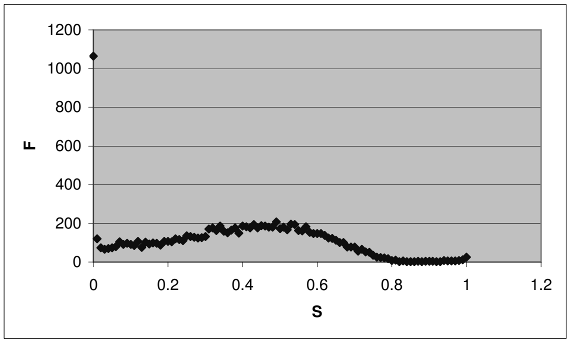


Figure 7: Ask.com: Cosine Similarity (S) vs Frequency (F)

As before, for a pair of pages (u, v), we defined $Related(u, v) = true$ if $S(u, v) \geq 0.9$, where $S(u, v)$ is the cosine similarity between pages u and v. For pairs of pages (u, v) where the above predicate holds, we proceeded to find $hd(u, v)$.

Results: The experiment was repeated with 105 pairs of pages having a cosine similarity of 0.9 or more. This corresponds to a sample space of about 17555 pairs of pages. Ask.com experiments had a smaller sample space simply because of the lack of any supporting index structures and the size of the dataset, making it enormously difficult to find pairs of pages having a high cosine. However, we will see that the results obtained with this sample are such that they are only likely to be strengthened with larger sample sizes.

Figure 8 shows a plot of the hyperlink path distance against the frequency of its occurrence among these 105 pairs. Here, a majority of the pairs of pages (68 of 105 pairs) have a hyperlink path distance greater than k_{max}. This indicates a high degree of separation. The next most frequent hyperlink path distances are 6 and 7, with a frequency of 10 and 11 respectively.

On the web, most of the content-wise similar pages seem to be disconnected from one another. A degree of separation greater than k_{max} amounts to an ignorance factor of 1. For this sample, we obtained an average ignorance factor of 0.9996.

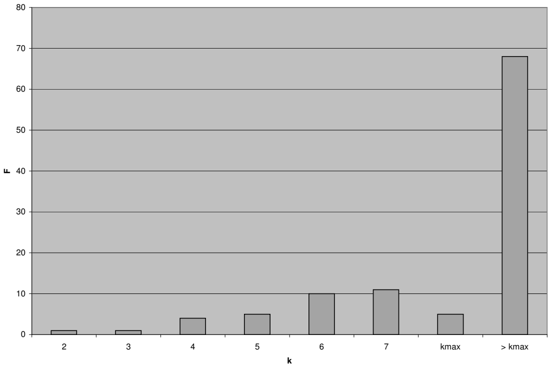


Figure 8: Ask.com (Cosine Similarity): Hyperlink path distance (k) vs Frequency (F)

5.2 Co-citations

We indexed all the links in the Ask.com data set. We then picked pairs of pages (u, v) uniformly at random from the data set, and calculated the number of times they have been co-cited, N. We collected more than a million such pairs of pages. The distribution of each N against its frequency, F, is a power-law, as shown in figures 9 and 10.

We defined $Related(u, v) = true$ if $N(u, v) \geq 1500$, where $N(u, v)$ is the number of times pages u and v have been co-cited.

Table 4 shows the proportion of related pages found from different sampling experiments. On an average, the proportion of similar pages based on co-citation was found to be 0.0121%.

No. of pairs with $N(u, v) \geq 1500$	Total no. of pairs
165	1194752
7	48830
4	32330
6	59585
7	68737

Table 4: Co-citation similarity in crawl data from Ask.com

An interesting point here is that the proportion of pairs of pages having at least 1500 co-citing pages is much higher on the web. However, this could be simply because of the larger size of the web. Because the co-citation distribution is scale-free, it becomes difficult to normalize the amount of co-citations within a bounded interval.

Nevertheless, the results obtained from 1500 as the co-citation threshold would likely to be only strengthened for larger thresholds.

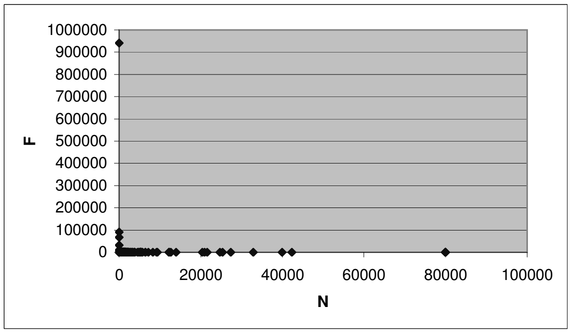


Figure 9: Ask.com (Co-citations): Number of co-citations (N) vs Frequency (F)

Results: We calculated the hyperlink path distance between 443 pairs of pages. Figure 11 shows a plot of the hyperlink path distance against the frequency of its occurrence among these 443 pairs. Here the mode is 1, with 272 of the 443 pairs of pages having this hyperlink path distance. The next most frequent hyperlink path distance is 3, with a frequency of 108. There are 9 pairs of pages with a hyperlink path distance greater than k_{max}.

Despite this, the average ignorance recorded was still quite high: 0.9782. The results here are analogous to the Wikipedia co-citation experiment. In clusters of co-cited pages, the degree of separation is very low, making it amenable for crawlers to discover communities. But the ignorance is still high, making it unlikely for human browsers to reach related pages from one another.

The presence of a direct hyperlink between most pairs of co-cited pages also strengthens the notion that

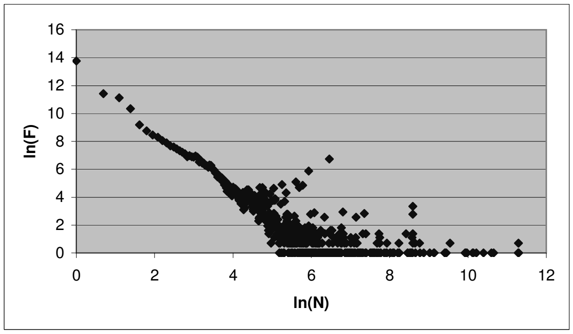


Figure 10: Ask.com (Co-citations): $\ln(N)$ vs $\ln(F)$

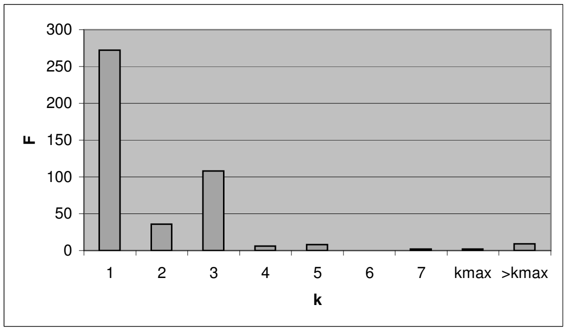


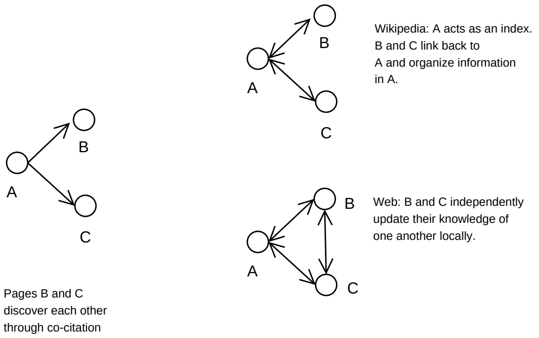
Figure 11: Ask.com (Co-citations): Hyperlink path distance (k) vs Frequency (F)

the Web is a clustered graph. A clustered graph is characteristic of most social networks, where if a node a is connected to nodes b and c, then there is a high probability that b and c are connected themselves.

On the web, usually co-citation appears as a *consequence* of pages being hyperlinked in the first place. That is, the hyperlinks appear first, then followed by the co-citation. Even when this is not the case and two web pages discover each other from a co-citing page, because of the fact that web page owners can only modify their own respective pages, this results in addition of direct links.

Wikipedia in contrast, allows anyone to modify any page. As a consequence, it is a more organized place, where co-citing pages acts like index pages. If two pages b and c discover one another through a co-citing page a, then any knowledge obtained by this discovery is likely to be organized by modifying the index page a itself.

Figure 12 depicts this process schematically. Updation of knowledge in Wikipedia happens by making "global" changes, since anyone can update any page. Eventually the changes converge towards preserving an overall semantic structure of the Wiki web. However on the web in general, each page updates knowledge locally, thus forming clusters among themselves.



A
B
C
Pages B and C
discover each other
through co-citation
A
B
C
Wikipedia: A acts as an index.
B and C link back to
A and organize information
in A.
A
B
C
Web: B and C independently
update their knowledge of
one another locally.

Figure 12: Clustering processes in Wikipedia and the Web

5.3 Interpretation of Results

In our experiments using cosine similarity as a measure of relatedness, we found that a majority of pairs of related pages have a hyperlink path distance greater than k_{max}. This shows the existence of a high degree of separation among content-wise related pages, and as a consequence, high ignorance.

However, co-cited pages seem to know one another *directly*. This indicates the existence of tightly connected web communities that connect inter-related pages, which are unaware of other content-wise related pages elsewhere on the web. The web, hence is highly entrenched. The content-wise related pages are segregated from these web communities, thus causing the entrenchment effect.

6 Experiments: Del.icio.us

The Wikipedia and Ask.com data sets are offline dumps/crawls. In order to measure ignorance on the live Web, we decided to conduct an experiment by performing live crawls on the Web. In order to establish the relatedness of pages, we used the del.icio.us tags, which give us a human interpretation of relatedness.

We chose a random pair of tags from among the tags on the latest bookmarks of del.icio.us. We then obtained the popular pages that have been tagged using both these tags. We picked pairs of pages from among these popular pages that have been tagged with these two tags by more than 100 people. Based on our premise that pairs of pages picked as above are related, we calculated the hyperlink path distance between them.

Results:

We performed this experiment for 20 pairs of pages. We found total ignorance (up to level 5) in all the cases, i.e. none of the pairs were connected to each other within 5 levels. Due to the restrictions on accessing websites by a bot, we were not able to perform the experiments on more pairs of pages.

7 Related Literature

Dean and Henzinger [11] described an algorithm to find pages related to a given page, using its neighbourhood graph. They computed the hub score and authority score for each node in this graph, and returned the nodes with the top-ranked authority score. They also described another algorithm for finding related pages of a given page, by determining its sibling or co-cited pages that have high degrees of co-citation.

Fogaras, et al. [12] showed that the hyperlink structure of vertices within four to five levels from a given page provide more similarity information than single level neighbourhoods.

Davison [10] studied the extent to which the following assumption about the nature of hyperlinks holds true: "If page A and page B are connected by a hyperlink, they are on the same topic." He analyzed whether topical locality mirrors the spatial locality of pages on the Web. He sampled around 200,000 pages from the repository of a research search engine called DiscoWeb. He collected several pairs of pages of various kinds: (1) random pages, (2) random sibling or co-cited pages (pages that are both linked to from some other page), (3) random pages from the same host, and (4) random pages from different hosts. He compared the textual content on these pairs of pages using the cosine between their document vectors, and made the following observations: (1) Random page pairs have almost nothing in common, (2) Linked pages are more similar when the pages are from the same domain, and (3) Sibling pages are more similar than linked pages of different domains.

Menczer [15] conjectured and proved that pages within a few links from a given page are relevant to it with a high probability. He experimentally assessed the extent to which relevance is preserved within hyperlink neighbourhoods and the decay in textual similarity as one browses away from a page.

Chakrabarti, et al. [6] used a topic taxonomy like the Open Directory as a framework for understanding the structure of content-based clusters and communities. Using this framework and a topic classifier, they measured the background distribution of broad topics on the Web. They analyzed the capability of random walk algorithms to draw samples that followed such distributions. They also measured the probability that a page about one broad topic linked to another broad topic. Extending this experiment, they measured how quickly topic context is lost while walking randomly on the Web graph.

Chakrabarti, et al. [5] addressed the issue of entrenchment effect. Their analysis addressed the concern that popular search engines limit the attention of authors of new pages to a small set of "celebrity" URLs, for any query. They showed that, eventually, the celebrity URLs accumulate a constant fraction of all newly created links, and that the other URLs still follow a power-law distribution, but with a steeper power. They concluded that search engines offer new pages a steep and self-sustaining barrier to entry into well-connected web communities.

Cho, et al. [8, 9] also showed that the entrenchment effect exists in the Web. They analyzed how much longer it takes for new pages to attract a large Web traffic when search engines return only celebrity pages at the top of search results. They also concluded that search engines have an adverse impact on the discovery of new pages.

8 Concluding Remarks

The results of this study shows a strong disconnect between the hyperlink structure of the web and the distribution of similar content.

The entrenchment effect of the web is very real, which may have a bearing on several activities like focused crawling and ranking that rely on hyperlinks. The entrenched nature of the web also indicates the significance of collaborative knowledge sharing environments like Wikipedia and del.icio.us to bring similar content closer.

We envisage that browsing using an overlay web like del.icio.us tags or Wiki links (perhaps embedded within the browser) would become just as commonplace if not more, as compared to hyperlink based browsing.

Acknowledgements

We are grateful to Tao Yang (Chief Scientist & Senior Vice President, Ask.com and Professor of Computer Science, UC Santa Barbara) and Ambuj Singh (Professor of Computer Science, UC Santa Barbara) for making the Ask.com data set available to us.

The third author is grateful to Infosys Technologies Limited, Bangalore for his PhD fellowship.

References

[1] R. Baeza-Yates and C. Castillo. Crawling the infinite Web: Five levels are enough. *Proc. of the 3rd Workshop on Web*, 2004.

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

[3] J. Battelle. *The Search: How Google and its rivals rewrote the rules of business and transformed our culture.* Nicholas Brealey Publishing, 2005.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and

J. Wiener. Graph structures in the Web. *Proc. of the 9th International World Wide Web Conference*, 2000.

[5] S. Chakrabarti, A. Frieze and J. Vera. The influence of search engines on preferential attachment. *Proc. of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005.

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

[7] P. -A. Chirita, D. Olmedilla and W. Nejdl. Finding related pages using the link structure of the WWW. *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 632-635, 2004.

[8] J. Cho and S. Roy. Impact of search engines on page popularity. *Proc. of the 13th International World Wide Web Conference*, pages 20-29. ACM Press, 2004.

[9] J. Cho, S. Roy and R. Adams. Page quality: In search of an unbiased Web ranking. *Proc. of the ACM SIGMOD International Conference on Management of Data*, 2005.

[10] B. Davison. Topical locality in the Web. *Proc. of the 23rd ACM SIGIR Conference on Research & Development on Information Retrieval*, 2000.

[11] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Proc. of the 8th International World Wide Web Conference*, pages 1467–1479, 1999.

[12] D. Fogaras and B. Racz. Scaling link-based similarity search. *Proc. of the 5th International World Wide Web Conference*, 2005.

[13] S. Fortunato, A. Flammini, F. Menczer and A. Vespignani. The egalitarian effect of search engines. `http://arxiv.org/abs/cs.CY/0511005`, 2006.

[14] M. Hindman, K. Tsioutsiouliklis and J. A. Johnson. “Googlearchy”: How a few heavily-linked sites dominate politics on the web. *Annual Meeting of the Midwest Political Science Association*, 2003.

[15] F. Menczer. Links tell us about lexical and semantic Web content. *Technical Report Computer Science Abstract CS.IR/0108004, arXiv.org*, 2001.

[16] F. Menczer, S. Fortunato, A. Flammini and A. Vespignani. Googlearchy or Googlocracy? *IEEE Spectrum*, Feb 2006.