Measuring Hyperlink Distances: Wikipedia Case Study*

[Extended Abstract]

Rodrigo R. Paim
ECI/Poli
Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro, Brazil
paim.rodrigo@land.ufrj.br

Daniel R. Figueiredo
PESC/COPPE
Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro, Brazil
daniel@land.ufrj.br

ABSTRACT

Hyperlinks are a fundamental aspect of the Web, as they play a major role in accomplishing important functions such as document clustering and document ranking. Despite various facets of hyperlink analysis, in this work we consider a novel aspect of hyperlinks, namely their *distance*. How far in terms of contextual similarity will a hyperlink take you? We consider classical distance functions that capture the similarity between documents as well as propose a new distance function, an IDF-based generalization of Jaccard distance. We characterize the distance distribution of hyperlinks considering Wikipedia as a case study. Our results indicate that hyperlink distances are strongly skewed, with the majority of hyperlinks exhibiting very long distances.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia,Theory

General Terms

Information Systems

Keywords

hyperlink analysis, hyperlink distance, document similarity, wikipedia

1. INTRODUCTION

Hyperlinks are a fundamental aspect of the Web, as they play a major role in accomplishing important functions such as document clustering, document ranking and document search [6]. Despite various facets of hyperlink analysis, in this work we consider a novel aspect of hyperlinks, namely their *distance*. How far in terms of contextual similarity will a hyperlink take you? Consider a hyperlink (a, b) from webpage a to webpage b. How far apart are a and b in terms of their content?

The content-based distance between two webpages strongly depends on the metric used to capture the notion of content similarity. We will consider two classical distance functions: Jaccard distance and cosine distance. Moreover, we introduce a new distance function to better reflect webpage similarity based on an IDF generalization of Jaccard distance [1]. Our goal is to characterize the distance distribution of hyperlinks using different distance functions. We use the Wikipedia web graph as a case study. Surprisingly, our results indicate that hyperlink distances are strongly skewed, covering a wide range of distances, but with the majority of hyperlinks exhibiting very long distances.

A closely related work has investigated the correlation between contextual proximity of webpages and hyperlink proximity in terms of the webgraph [5]. We note that altough related, such analysis does not reflect hyperlink distances in the sense intended in this work.

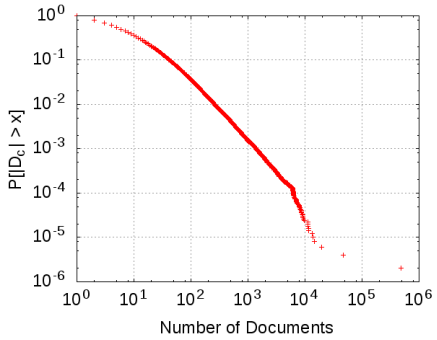
Finally, the distribution of the distance of links in networks has important implications for navigation and search [3, 4, 2]. This aspect has also been studied in the context of the web graph, with hyperlinks serving as directed network edges [5]. We believe our work can shed light on the navigability of the web by posing the simple notion of hyperlink distance.

1.1 Wikipedia Dataset

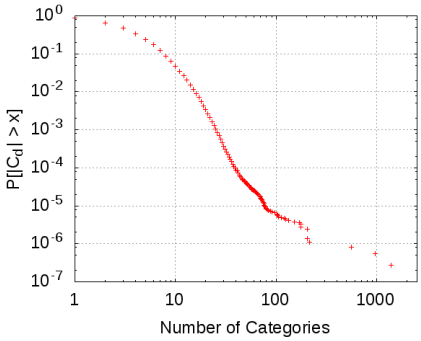
We consider the English Wikipedia as a case study (dataset of January 2011, 3.6 million documents). In Wikipedia, each document has a set of categories and each category appears in one or more document. Let D denote the set of documents in Wikipedia and C the set of categories. Let $C_d \subset C$ denote the set of categories of document $d \in D$. Let $D_c \subset D$ denote the set of documents that have category $c \in C$ associated with them.

Categories play an important role in organizing Wikipedia's content behaving as keywords for documents. Thus, contextual information about a document is directly reflected by its categories. Therefore, we will use the set C_d of document d to indicate what is document d about. In particular, the distance functions that measure similarity between two

*This research receives financial support from CNPq and FAPERJ (Brazil), through grants CNPq 557.128/2009-9 and FAPERJ E-26/170028/2008 (INC&T Program – *Instituto Brasileiro de Pesquisa em Ciência da Web*).



(a) Distribution (CCDF) of number of documents in a category (avg = 23.9, std = 702.1).



(b) Distribution (CCDF) of the number of categories in a document (avg = 3.24, std = 3.38).

Figure 1: Empirical distributions of documents and categories.

documents will operate over the set of categories of the documents. Before continuing, we first characterize the documents and categories of the dataset.

Figures 1a and 1b show the empirical distribution of the size of set C_d and D_c, respectively, for all documents and categories in the dataset. Not surprisingly, the number of documents that have a given category exhibits a heavy tail distribution spanning five orders of magnitude, with some categories appearing in more than 10,000 documents[1]. The number of categories in a document has a much shorter tail, but also exhibits outliers, which are essentially documents that lists categories[2]. Note that both averages are rather small, with only 23.9 documents per category and 3.24 categories per document.

2. MEASURING DISTANCE

We consider two well-known distance functions for measuring content similarity which are Jaccard distance and cosine distance [8, 7]. As we show, these distances functions are not capable of capturing distances over a wide range of scales. We then introduce a third distance function based on IDF and Jaccard distance that can yields distance values at a wide range of scales.

2.1 Jaccard Distance

Jaccard distance is used to measure dissimilarity between two sets and can be interpreted as the complement of the Jaccard index. Consider the set C_a and C_b of categories of documents a and b, respectively. Their Jaccard distance is defined as follows:

$$d_J(a,b) = 1 - \frac{|C_a \cap C_b|}{|C_a \cup C_b|} \tag{1}$$

Clearly, $d_J(a,b)$ assumes values over a short range, in the interval $[0,1]$. Moreover, the zero value means that the intersection of categories is equal to their union, and, therefore, a and b are declared equal, since their sets of categories are identical. The one value occurs when a and b have no categories in common, independent of the categories each document has. Thus, the maximum distance occurs whenever two documents have no categories in common.

2.2 Cosine Distance

The cosine distance can be interpreted as the complement of the cosine similarity [7] between two documents. In this metric, each document d is represented as a $|C|$-dimensional vector, V_d, where C is the set of categories considered. For each dimension $c \in C$, vector V_d assumes value 1 if document d has category c (i.e., $c \in C_d$), or 0 otherwise. Consider V_a and V_b as the two category vectors of documents a and b, respectively. Intuitively, the cosine distance is the complement of the cosine of the angle θ formed by vectors V_a and V_b in the C-dimensional space. More precisely, we have:

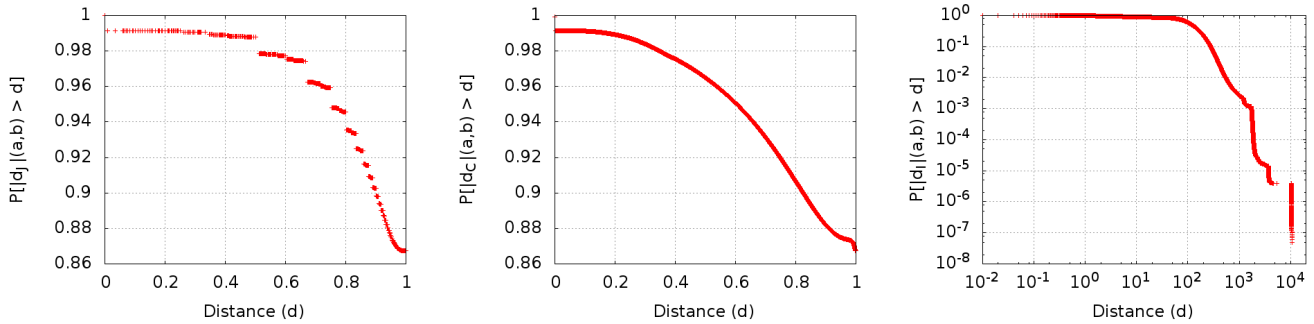
$$d_C(a,b) = 1 - \cos(\theta_{ab}) = 1 - \frac{V_a \cdot V_b}{\|V_a\|\|V_b\|} \tag{2}$$

As with the Jaccard distance, this metric ranges from 0 to 1. Also similarly to the Jaccard distance, this metric is 0 when the set of categories are identical and 1 when they have no category in common (if the intersection is empty then the dot product is zero).

We characterize the empirical distance distribution of hyperlinks by considering all hyperlinks in the dataset[3]. For each hyperlink (a,b) from document a to document b, we have a distance value of $d(a,b)$. Thus, we compute the empirical hyperlink (complementary) distance distribution, which denotes the fraction of hyperlinks that have distance greater than a given value d.

Figures 2a and 2b show the empirical distance distribution of hyperlinks for both the Jaccard and cosine distances, respectively. Surprisingly, note that 87% of the hyperlinks have the maximum possible distance of 1, which means that 87% of the hyperlinks are between documents that have no categories in common. Moreover, only 2% and 4% of the

[1]The most popular category is "Living people", which appears in 482,405 documents.
[2]The document with the most categories is "List of mathematics categories", with 1391 categories.
[3]Each hyperlink in the dataset is considered, even if there are multiple hyperlinks between the same two documents.



(a) Distribution (CCDF) of hyperlinks distance using Jaccard distance (avg = 0.97, std = 0.12).

(b) Distribution (CCDF) of hyperlinks distance using cosine distance (avg = 0.95, std = 0.16).

(c) Distribution (CCDF) of hyperlinks distance using the proposed distance function (avg = 143.6, std = 131.8).

Figure 2: Empirical distance distributions of hyperlinks.

hyperlinks have a distance that is less than half when considering the Jaccard and cosine distances, respectively. This indicates that the vast majority of hyperlinks span very large distances when using these metrics over categories.

2.3 An IDF-Based Metric

The Jaccard and cosine distance functions have two undesirable properties. First, they consider all categories as being identical. Second, the largest possible distance is given when two documents have no categories in common. Clearly, categories are not identical when used to compare documents. Intuitively, a category that appears in many documents is less important than a category that appears in few documents, when it comes to measuring the similarity of documents. Moreover, two documents that have many categories but none in common are probably less similary than two documents that have few categories but none in common. In the following, we propose a metric that captures these two aspects, avoiding the pitfalls of the Jaccard distance and cosine distance.

We associate with each category a positive weight that represents its importance in measuring similarity among documents. Intuitively, a category that is rare in the dataset has more importance. We use the standard IDF (inverse document frequency) metric [1] in determining the weights of the categories. Let $w(c)$ denote the weight of category $c \in C$, which according to IDF is defined as follows:

$$w(c) = \log_2 \left(\frac{|D|}{|D_c|} \right) \tag{3}$$

where D and D_c denote the set of documents in the dataset and the set of documents that have category c.

Using the weights, we generalize the idea behind the Jaccard index and define a distance that is its inverse (as opposed to its complement). In particular, the similarity between two documents is defined as the ratio of the sum of the weights in the intersection and the union, generalizaing the Jaccard index. However, we consider the distance to be the inverse of this similarity metric. Indeed, since the sum of the weights in the intersection can be zero (and will be in 87% of the cases), we introduce a small constant weight $w_u = 1$ and assume that all documents have this universal category. Thus, the distance metric is defined as:

$$d_I(a,b) = \frac{\sum_{c \in C_a \cup C_b} w(c) + w_u}{\sum_{c \in C_a \cap C_b} w(c) + w_u} - 1 \tag{4}$$

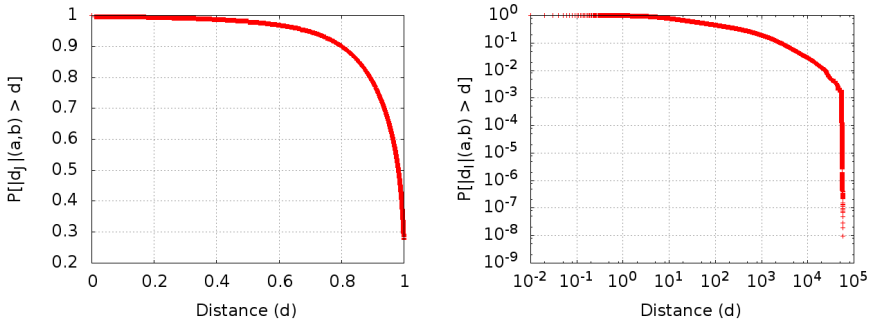
We subtract one in the above expression to have a zero-based metric, in the sense that $d_I(a,b) = 0$ when the two sets of categories C_a and C_b are equal.

Figure 2c shows the hyperlink distance distribution using the above metric, exhibiting intersting properties. Note that the distribution spans six orders of magnitude, a property that would certainly be desired. However, the distribution decays very little for the first four orders magnitude, and then as a heavy tail distribution for about one decade (from 10^2 to 10^3), dropping suddenly as an exponential tail afterwards. In part, this occurs because hyperlinks between documents have hundreds of categories but very few in common, leading to very large distances. Moreover, 87% of the hyperlink distances also depend on the value of w_u since they have no categories in common. This accounts for the entire tail of the distribution, starting at about 10^2.

3. OBTAINING MORE CONTEXT

We have seen that the average number of categories per document is rather low and number of common categories between documents that have a hyperlink is extremely low. In fact, 87% of hyperlinks are between documents that have no common categories. An alternative explanation for this surprising fact is that the categories of a document fail to capture its context. In particular, the categories of a document may be too coarse, not actually reflecting what the document is about. Thus, we would like to obtain a broader context for the documents, better capturing what the document is about.

We accomplish this by introducing the notion of document context. Let Γ_d denote the context of document $d \in D$. Intuitively, the context can be defined by the document itself. However, we define the context using the same set of categories in the dataset, without increasing its size. Thus, to



(a) Distribution (CCDF) of hyperlinks distance using Jaccard distance (avg = 0.93, std = 0.13).

(b) Distribution (CCDF) of hyperlinks distance using proposed metric (avg = 1247.0, std = 4400.4).

Figure 3: Empirical distance distributions of hyperlinks using document context.

define the context of document d, we use the categories of documents that are very close to d. In particular, we consider a document a to be very close to d if there exists at least one hyperlink from a to d and one from d to a. Intuitively, the requirement for a symmetric relationship increases the chances of the two documents being rather related. More formally, we have:

$$\Gamma_d = C_d \cup C_a | a \in D,\ (d,a) \text{ and } (a,d) \text{ exists} \qquad (5)$$

where (d,a) and (d,a) are directed hyperlinks between the documents.

Using this notion of context, we compute the distances of hyperlinks assuming the context Γ_d for each document d. Thus, the distance function $d(a,b)$ between documents a and b now operates over the sets Γ_a and Γ_b. Figure 3a shows the hyperlink distance distribution for the Jaccard distance. Surprisingly, 28% of hyperlinks are among documents that have no common categories in their context (maximum distance of 1). Moreover, only 3% of the hyperlinks are among documents that have a distance less than half. Thus, even when considering a broader notion of context for the documents, hyperlink distances are still very far.

Figure 3b shows the hyperlink distance distribution using the proposed IDF-based metric using document context. We again observe an interesting distribution, but with a much smoother trend (compare with Figure 2c). The distance distribution decays as a heavy tail distribution for several orders of magnitude, from 10^1 to $10^{4.5}$. Note that the probability of finding a hyperlink with any distance in this range is non-neglibile, yielding an average distance value of 1247, much smaller than the maximum obtained. For larger distances, we observe a very sharp drop in the distribution, corresponding to a small fraction of the hyperlinks (0.001), all of which have approximately the same and largest distance. As before, these hyperlinks are among documents that have a very large context but very few categories in common.

4. CONCLUSION

The webgraph (documents and hyperlinks of the web) play a fundamental role as it serves as input to several algorithms that provide various functions, such as document clustering and document ranking. In this work, we investigated the contextual distance of hyperlinks using various distance functions and different methods for determining the context of a document, using Wikipedia as a case study. Surprisingly, when considering classical distance functions such as Jaccard and cosine distance, we observe that hyperlink distances are heavy skewed towards very large values, even when the context of a document is augmented. This indicates that hyperlinks in Wikipedia are more likely to point to documents that are not related, at least from the perspective of the document categories and distance functions considered. Under the IDF-based proposed distance function, hyperlink distances are less skew and attain a much wider range of values with non-negligible probability. This finding is interesting because it allows exploiting this characteristic to guide algorithm design, for example, in document search and navigability.

5. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] M. Boguna, D. Krioukov, and K. C. Claffy. Navigability of complex networks. *Nature Physics*, 5, 2009.

[3] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *ACM STOC*, 2000.

[4] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proc. of the National Academy of Sciences (PNAS)*, 102, 2005.

[5] F. Menczer. Growing and navigating the small world web by local content. *Proc. Natl. Acad. Sci. (PNAS)*, 99, 2002.

[6] H. W. Park and M. Thelwall. Hyperlink analyses of the world wide web: A review. *Journal of Computer-Mediated Communication (JCMC)*, 8, 2003.

[7] Wikipedia. Cosine similarity. http://en.wikipedia.org/wiki/Cosine_similarity, 2011.

[8] Wikipedia. Jaccard index. http://en.wikipedia.org/wiki/Jaccard_index, 2011.