Mapping a Local Web Domain

Weimao Ke
School of Library and Information Science
Indiana University Bloomington
1320 E. 10th St., LI 011
Bloomington, IN 47405-3907, U.S.A.
wke@indiana.edu

Tiago Simas
Cognitive Science
Indiana University Bloomington
819 Eigenmann, 1910 E. 10th St.
Bloomington, IN 47406-7512, U.S.A
tdesimas@indiana.edu

Abstract

In this study, we crawled a local Web domain, created its graph representation, and analyzed the network structure. The results of network analysis revealed local scale-free patterns consistent with previous research on the Web. To discover topical communities of the local domain, link-based co-citation analysis was performed to measure pairwise similarities. Based on the co-citation graph, a visualization was produced to map the networked domain semantically. All of this offers a way to investigate a local Web structure and to make it more intuitively understandable.

1 Introduction

Previous research has found that the World Wide Web is a scale-free network [1] [2] [4]. It resembles a bow-tie structure and is comprised of four components: the Central Core, IN, OUT, and tendrils and tubes [7]. Relying on statistical analysis, these research provide insights to understanding the global Web. Nonetheless, to individual and/or organizational interests, it is also valuable to explore a local Web's structure.

Instead of statistically claiming a Web structure such as a bow-tie, information visualization might be more intuitive in revealing a local topology based on semantic similarities. According to Menczer, the Web has two main classes of cues to approximate semantic similarity: lexical cues (textual content) and link cues (hyperlinks) [10]. Text-based representation and similarity measures have been widely studied in information retrieval research [3]. For example, the TF-IDF (term frequency * inverse document frequency) weighting scheme and the Cosine ranking metric are known to be effective in measuring relevance. However, given the huge number of pages on the Web, it is computationally expensive to measure pairwise textual similarities. Link analysis is more affordable.

In this study, we are primarily interested in applying network analysis and visualization techniques to local Web domains. We wish to know whether the characteristics of the global Web are applicable locally, that is within an organizational domain. Specifically, this study is to provide an exemplary approach to analyzing network properties, identifying topical communities, and visualizing the map of a local domain.

2 Crawling the Local Domain

We restricted ourselves to our own local university domain, namely *.indiana.edu. To retrieve the local graph, we ran a crawler to fetch pages and links within the domain. The crawler crawled 2,710,239 pages/files and found 100,079,061 link pairs before it was terminated. This captured 20% of the indiana.edu domain, which had about 13 million pages according to google. Basic page properties (size, type, url, etc.) and the link data (30GB of pure text) were then loaded into a PostgreSQL database for further analysis.

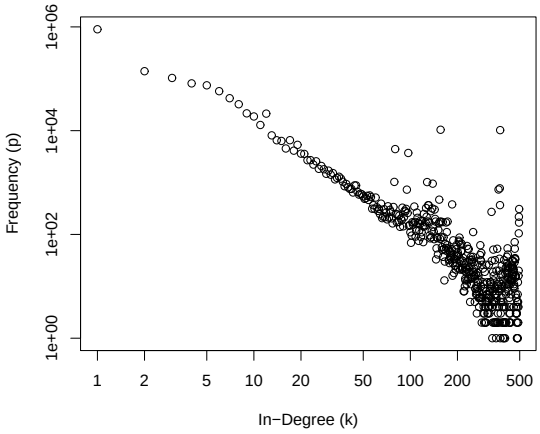
3 Creating the Domain Graph

After the crawl was finished, the data were parsed and loaded into two tables: nodes and links. Web contents were not preserved. We removed all duplicates and URLs longer than 250 characters. For link analysis, we ignored non-text/html files, resulting in a set of $1,707,676$ unique nodes and $43,206,848$ unique links. We then used the remaining hyperlinks to create a directed graph representation of the local Web domain.

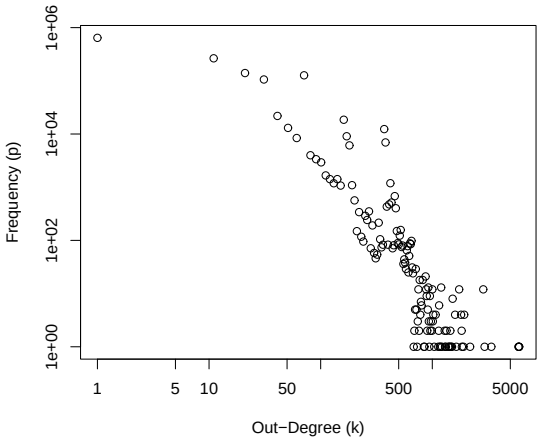
4 Network Analysis

Topological properties of the directed graph were examined. Figure 1 and 2 show the distributions of 1(a) in-

degree, 1(b) out-degree, 2(a) page size (all pages), and 2(b) page size (text/html only) on log-log coordinates, because of the large scale of the plots.



(a) In-Degree Distribution

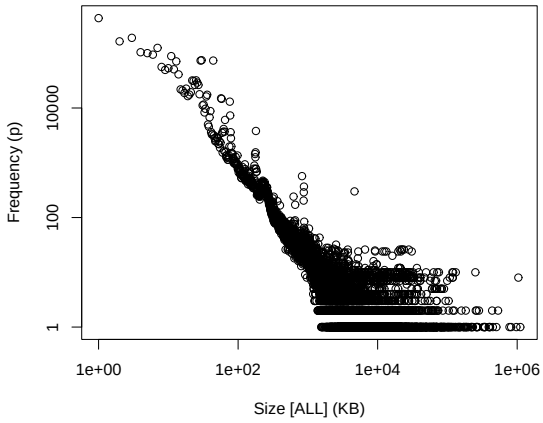


(b) Out-Degree Distribution

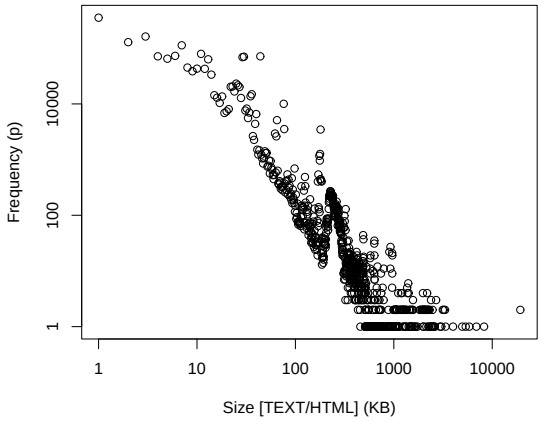
Figure 1. Degree Distributions

As shown in Figure 1(a), the in-degree distribution follows a power law with an exponent $\gamma_{in} = 2.109 \pm 0.0195$. This result is consistent with previous works [1, 7], in which the authors found a $\gamma_{in} = 2.1$. We also notice a few outliers on the power-law distribution. This can be explained by pages that form small groups and are highly connected to each other.

Research have been done on models that characterize scale-free degree distributions [4, 2, 11]. They proposed that a network with this type of distribution grows under "preferential attachment," with the constraint that creators of the Web sites only have partial views of the entire WWW network when connecting to other pages.



(a) Size Distribution (ALL)



(b) Size Distribution (TEXT/HTML)

Figure 2. Size Distributions

In Figure 1(b), the out-degree distribution has a power law cut-off. We obtained a $\gamma_{out} = 2.09 \pm 0.016$. Different values were found in the literature: in [7] $\gamma_{out} = 2.72$, whereas [1] $\gamma_{out} = 2.45$. It is not always clear what the exact mechanisms are behind the out-degree. However, in our case, it makes sense to fit the data as a power-law and explain it with a scale-free model [11].

Interestingly, in Figure 2(a) for all pages and 2(b) for text/html only, file size also follows a power-law. Figure 2(a) has a $\gamma_{size} = 1.308 \pm 0.016$, whereas 2(b) has $\gamma_{size} = 2.001 \pm 0.017$. This tells us that the majority of the Web domain are small-size pages. In 2(a), there is a small jump between 200 - 300 KB, which is likely a particularity of the domain.

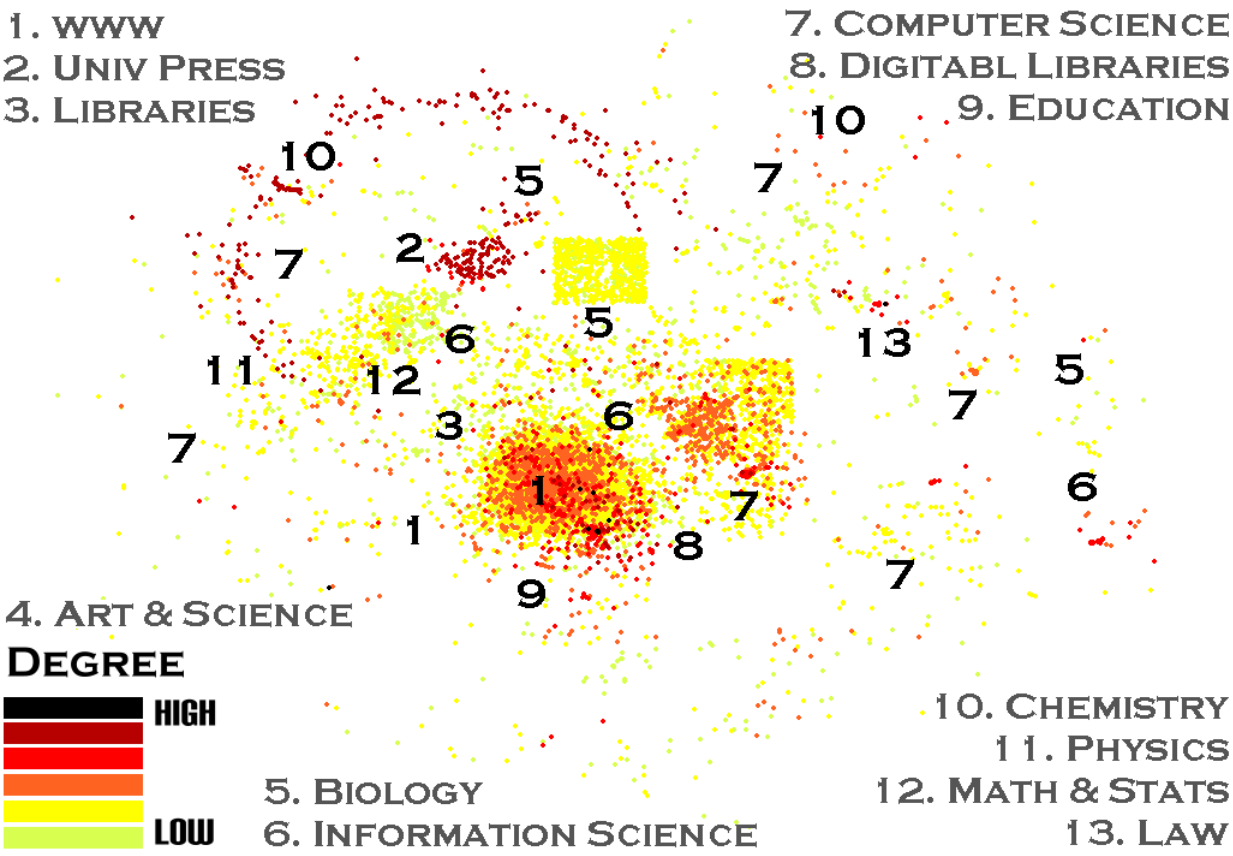


Figure 3. Domain Map Visualization

5 Visualizing the Domain Map

Co-citation analysis has been widely used in bibliographical analysis and, in recent years, has been adopted by researchers to measure link-based similarities on the Web [9]. Dean and Henzinger implemented a co-citation algorithm for finding related pages on the World Wide Web and demonstrated promising results [8]. In our co-citation analysis, hyperlinks were treated as citations. By definition, the number of co-citations between any two nodes n_i and n_j, where $i \neq j$, is the number of nodes that commonly cited (or referred to) both n_i and n_j. Co-citations can thus be derived from directed hyperlinks and used for identifying similar and influential nodes in a given graph. If a web page has many co-citations with many others, it can be treated as a landmark in the topical communities.

For this visualization, we removed all internal links, that is, only keeping links from different hosts. URLs with a length > 250 characters were also removed because these are probably dynamically generated URLs. Then, pages with at least 2 incoming links were retrieved, resulting in 20,074 nodes. Co-citations between these nodes were computed to form an undirected weighted graph. We rendered a layout for the entire graph using VxOrd [6], which tried to place similar objects close together and dissimilar objects far apart in a 2-D space. The coordinates were then loaded into Pajek [5] to produce the visualization shown in Figure 3. Co-citation edges have been removed in order to show the nodes (dots) clearly.

In Figure 3 node color is based on log(in-degree). The darker a node, the larger its in-degree. The layout visualizes co-citational similarities in a 2D space, and is a semantic map of the local domain. Nodes that frequently co-cite each other are usually close together. Therefore, using colors and proximities of the nodes, the map shows not only topical closeness, but also connectivity distribution in the domain.

The majority of the nodes in the visualization are moderately interconnected, shown in green and yellow. It is obvious that there are several visible clusters, among which highly connected nodes (in red and darker colors) are scattered. These high-degree nodes are landmarks of the map. Labeling the nodes enables us to explore it. Close examination of the visualization tells that the main cluster consists of departmental sites, while the others are formed from

of common research areas. For example, information science clusters (6) are close to computer science (7), digital libraries (8), and math & statistics (12). The digital libraries cluster (8) is also adjacent to education (9). This visualization also clearly displays the "small jumps" or outliers on the degree distributions. There are a number of darker dots scattered on the map, which are likely highly-connected nodes of small topical groups.

6 Conclusions

In this study, we analyzed a local Web graph by heuristically examining its network properties. The results were consistent with previous works on the Web, and proposed local scale-free features. Using co-citation analysis and visualization techniques, this paper offered a way to investigate a local Web structure and to visualize its domain map that is more intuitively understandable. Future work will involve scaling up the algorithms and applying the techniques to a broader context.

Acknowledgements

We acknowledge the Laboratory of Applied Informatics Research (LAIR) at Indiana University for computational support and thank Jonathan Warren, Filippo Menczer, Javed Mostafa, and Katy Börner for their valuable feedback. We appreciate constructive comments from the anonymous reviewers of the IV07 conference. This study originated from a class project for the CSCI B659 Web Mining course by Filippo Menczer at Indiana University Bloomington.

References

[1] R. Albert, H. Jeong, and A.-L. Barabasi. The diameter of the world wide web. *Nature*, 401:130, 1999.

[2] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *PNAS*, 97(21):11149–11152, 2000.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman Publishing, 2004.

[4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[5] V. Batagelj and A. Mrvar. Pajek - analysis and visualization of large networks. In P. Mutzel, M. Jünger, and S. Leipert, editors, *Graph Drawing, Vienna, Austria, September 23-26, 2001*, pages pp. 477–478. Springer, 2002.

[6] K. W. Boyack, B. N. Wylie, and G. S. Davidson. Domain visualization using vxinsight for science and technology management. *Journal of American Society for Information Science and Technology*, 53(9):764–774, 2002.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of the 9th international World Wide Web conference on Computer networks*, pages 309–320, 2000.

[8] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1467–1479, 1999.

[9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[10] F. Menczer. Mapping the semantics of web text and links. *IEEE Internet Computing*, 09(3):27–36, 2005.

[11] S. Mossa, M. Barthelemy, H. E. Stanley, and L. A. N. Amaral. Truncation of power law behavior in "scale-free" network models due to information filtering. *PHYS.REV.LETT*, 88:138701, 2002.