Does WT10g Look Like the Web?

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, MD
ian.soboroff@nist.gov

ABSTRACT

We measure the WT10g test collection, used in the TREC-9 and TREC 2001 Web Tracks, with common measures used in the web topology community, in order to see if WT10g "looks like" the web. This is not an idle question; characteristics of the web, such as power law relationships, diameter, and connected components have all been observed within the scope of general web crawls, constructed by blindly following links. In contrast, WT10g was carved out from a larger crawl specifically to be a web search test collection within the reach of university researchers. Does such a collection retain the properties of the larger web? In the case of WT10g, yes.

Categories & Subject Descriptors: H.3.m [Information Storage and Retrieval]: Miscellaneous – Test Collections

General Terms: Experimentation, Measurement

Keywords: web test collections, TREC

1. INTRODUCTION

A critical requirement of a retrieval test collection is that it match the task. When the collection in question is a web collection, the issue expands to cover not only the content of the pages, but the broader hypertext structure of the collection as a whole. Since it is impossible to conduct repeatable retrieval experiments as we understand them on the "live web", several static web test collections have been built and used by the retrieval community in the past few years.

Bailey et al. [1] describe the construction of WT10g, the Web Track test collection used for TREC-9 and TREC 2001. This collection is about 10GB in size, and contains 1.69 million web pages. Their goal was to create a testbed for "realistic and reproducible" experiments on web documents with traditional, distributed and hyperlink-based retrieval algorithms. They began with VLC2, a 100GB subset of a 1997 crawl by the Internet Archive. From this they selected documents using a process designed to maximize inter-server connectivity, retain as many pages as possible from each server represented, incorporate documents likely to be relevant to a wide variety of queries, and exhibit a realistic distribution of server sizes. This process is described in detail in [1]. They measured the properties of the resulting collection according to mean in- and out-links per server, fraction of connected servers in the collection, and server "relevance", measured using a large query set.

One question that they did not answer was, does WT10g look like the World Wide Web? To answer that, we first need to understand more about what the web looks like. Singhal and Kaszkiel [4] looked at average in- and out- links, within and across hosts, between the smaller WT2g corpus and their own large crawl. They concluded that linkage in WT2g was inadequate for web experiments. However, the mean is a poor statistic to describe the power-law distributions of links on the web; average linkage is dominated by the many pages with few links and gives little insight into the topology.

Broder et al. [2] analyzed two large web crawls of about 200M pages each done by Altavista in 1999, and compared their structure to two important earlier studies. They looked at the distributions of in-links and out-links in their crawls, illustrating that these distributions obey power laws with exponents close to those observed in other studies. Further, using breadth-first traversals from a large sample of starting points they sketched out the high-level structure of the web in what has become the well-known "bow-tie model". These characteristics seem to hold for the web in general, however, Pennock et al. [3] found that category-specific subsets of the web can deviate strongly from power law scaling.

In order to show that WT10g indeed does resemble the web in many important ways, we measured the collection's link graph using the yardsticks of Broder et al. We show that while WT10g is small, structurally it does resemble larger web crawls that have been studied. This is an important result, because a primary criticism of web test collections is that they are inherently too small to be realistic testbeds of the web. These metrics can also be used to tune the construction methods of future test collections.

2. POWER-LAW DISTRIBUTIONS

Broder et al. found that the distributions of links in their crawls followed a power-law, that is, that the probability that a node has (in- or out-) degree d is proportional to $1/x^d$ for some $d > 1$. The exponents in their crawls was 2.1 for in-degree, and 2.72 for out-degree. Figure 1 shows the degree distributions in WT10g. These graphs are very similar to those found by Broder et al. In particular, notice the linear shape in the log-log plot, the messy tails for those few pages of very high degree, and that out-links diverge from the fitted curve at very low degree. The power-law exponents are 2.03 for in-degree, and 2.49 for out-degree. We are missing some spikes that they found and attributed to a spammer.

Power laws of hyperlink degree have been found in nearly

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

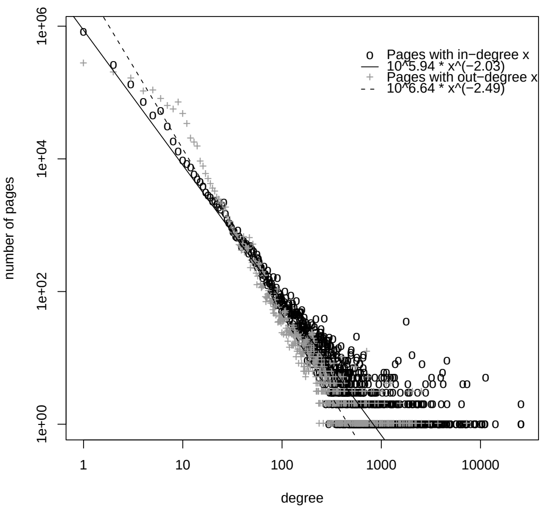

Figure 1: In- and out-degree distributions in WT10g.

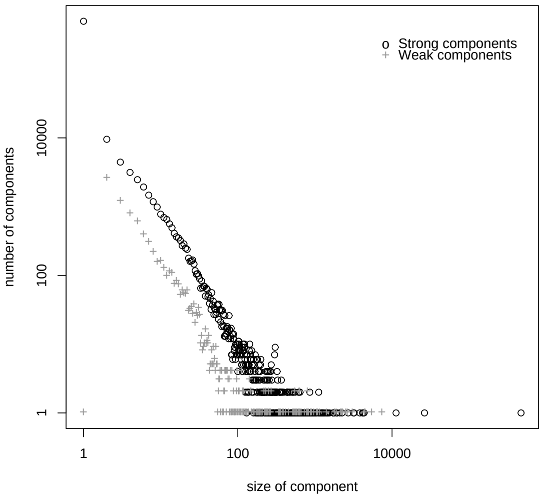

Figure 2: Distributions of strongly and weakly connected components in WT10g.

every study of a web crawl, through a wide variety of crawl sizes. In contrast, WT10g is a subset of a web crawl carefully chosen to incorporate whole servers and highlight inter-site links, but without regard to the overall link distribution.

3. CONNECTED COMPONENTS

Broder et al. also examined strongly- and weakly-connected components of the link graphs of their crawls. A strongly connected component (or, "strong component") of a graph G is a subgraph G' such that every node in G' is reachable from every other node in G' by following forward links through the graph. A weak component is the equivalent structure in an undirected graph; in our web graphs, this means taking the union of in-links and out-links into consideration when finding connected components. Figure 2 shows the distributions of strong and weak components in WT10g.

These graphs also follow a power law (exponents 1.79 for SCCs, 1.37 for WCCs) similar to the distributions found in the Altavista crawls. Our largest weak component contains 91% of the pages in WT10g, the same fraction as in the Altavista crawls. The largest strong component encompasses 29.4% of all the pages in the collection, compared to 28% in the Altavista crawls. The similarity in largest component coverage is striking, but the smaller exponents in the WT10g distributions indicate a more gradual falling-off of component sizes. This probably reflects the tendency of WT10g to favor entire servers while at the same time having many fewer pages overall than the Altavista crawls.

4. EXPLORING WITH BFS

The third and most interesting component of Broder's study was designed to probe the dichotomy in coverage between the largest weak and strong components: if 91% of the collection is connected by undirected links, but only 29.4% by browseable links, what happened to all the other pages? If nothing else, it means that understanding the web to have uniformly small diameter is inaccurate; obviously, some pages are only reachable from certain places in the web, and a large fraction are all reachable from each other within a short distance. To explore this phenomenon, they conducted breadth-first searches backward and forward from random starting nodes, noting the depth of each traversal. We did the same for 500 random starting points.

Our findings again mirror those from the Altavista crawls. The traversal depths are sharply bimodal: either they would stop after reaching a small set of pages (often, fewer than 100), or they would balloon to a huge node set (roughly 740,500 following in-links, 926,500 for out-links). For about 30% of the start nodes, both directions would balloon; 30% would balloon following in-links only, and 10% following out-links. Following Broder's analysis, we find a bow-tie in WT10g with an IN set leading into the large SCC of 270,059 pages, an OUT set of pages reachable from the SCC of 456,059 pages, and 261,828 TENDRILS pages. WT10g's OUT set is larger than IN, compared to the Altavista crawls, where the sets were of roughly equal size. We hypothesize that the strategy in WT10g of selecting by server in order of size is biased somewhat toward SCC+OUT pages.

5. REFERENCES

[1] Peter Bailey, Nick Craswell, and David Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, to appear.

[2] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure on the web. In *Proceedings of the 9th International WWW Conference*, pages 309–320, Amsterdam, The Netherlands, May 2000.

[3] David M. Pennock, Gary W. Flake, Steve Lawrence, Eric J. Glover, and C. Lee Giles. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211, April 2002.

[4] Amit Singhal and Marcin Kaszkiel. A case study in web searching using TREC algorithms. In *Proceedings of the 10th International World Wide Web Conferenece*, pages 708–716, Hong Kong, May 2001.