Web08-PR Dataset

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao
Language Technologies Institute
Carnegie Mellon University
callan@cs.cmu.edu

September 26, 2008

Introduction

Carnegie Mellon University plans to create a web dataset of approximately 1 billion web pages that can be distributed to the research community under a TREC-style data license. This dataset will consist of high PageRank pages in 10 languages. The dataset is intended as a companion to the 1 billion page dataset created by Fetterly, Craswell, and Vinay (SIGIR 2008). We call these two datasets Web08-PR (PageRank) and Web08-BR (breadth-first) in this document, to reflect the crawl order of each dataset.

The Web08-PR dataset is primarily a collection of web pages. It will not include queries, click data, or other information about search activity. We expect that such data will be generated by others in subsequent efforts. Some of the decisions described below are intended to support those future activities. For example, the dataset will cover pages important to some existing query log data, and display-related data will be collected to support relevance assessments.

This document describes our current thinking and plans, which are based on preliminary conversations with a dozen or so colleagues at universities, web search companies, and NIST. The dataset is intended to be of use to a broad research community, so before we begin crawling, we seek one last round of feedback and advice prior to actually creating it. There is still time to adjust our plans based on the feedback we receive from the research community.

Dataset Goals

The Web08-PR dataset is intended to approximate Tier 1 of a web search engine index. Although the details are proprietary and vary from company to company, it is generally believed that Tier 1 consists of web pages that have high PageRank and/or significant search and clickthrough activity. We do not have comprehensive and reliable information about search and clickthrough activity, which in any case will vary over time, so we restrict our attention to high PageRank.

A secondary goal is a dataset that provides good coverage of several languages besides English. Selecting pages by PageRank alone may not achieve this goal.

A third goal is a dataset that will be comparable in quality and utility to datasets used by TREC, useful to researchers for 5-10 years, and support a variety of research agendas.

Web Crawling

The seed URLs for the web crawl will be created using two sources of information.

The Web08-BR breadth-first crawl is one source of seed URLs. A web graph was created from this dataset by Dennis Fetterly. The 100 million URLs with the highest PageRank [1]will be one set of seeds for the Web08-PR PageRank crawl. This set of URLs starts the crawl in high PageRank portions of the web, and seeds the crawler with a relatively good web graph.

Several commercial web search engines will also provide seed URLs. The most frequent queries will be extracted from one or more generally-available web search logs (e.g., the AOL query log). These queries will be submitted to several commercial search engines (e.g., G, Y, M). The top 100 URLs for each query will be added to the list of seed URLs. This set of URLs guarantees that the dataset has good coverage of documents that match query logs used by the research community.

All of the seed URLs will be downloaded, subject to language restrictions (see below). This set is the first iteration of the crawl.

Each subsequent iteration begins by creating a web graph from the pages downloaded so far, extracting the *n* uncrawled URLs that have the highest OPIC (On-line Page Importance Computation) value, and downloading all of those pages (subject to language restrictions). The only significance to the iteration size is that OPIC is recomputed only at iteration boundaries. Each iteration is about 10% the size of the number of URLs crawled so far; for example, the second iteration is 10 million URLs. A comment by Baeza-Yates, et al. (WWW 2005) suggests that this property is important to OPIC accuracy. About 25 iterations will be required to reach 1 billion pages.

Note that after the first 100 million URLs, the crawl is actually ordered by OPIC, not PageRank. This decision is due to a variety of factors: The default behavior of Nutch, the greater efficiency of OPIC, and observations by Baeza-Yates, et al. (WWW 2005) and others that OPIC is at least as effective as PageRank once the crawl has reached a certain size.

The crawl will be accomplished within a four-week timespan, if possible, to provide improved temporal cohesion. The crawl is currently scheduled to begin in late September or early October. An explicit goal is to complete data collection by the first week of November because computer hardware may be less available in November. Data processing may extend for several months after data collection is complete.

The crawler will have the following limits:

- Ignore dynamic web pages beyond depth 5;
- Ignore static web pages beyond depth 15;
- Limit the number of pages from any site to 25,000; and
- Truncate any file of more than 100 megabytes in size.

[1] A standard, plain-vanilla version of the PageRank algorithm was used.

The crawl will be conducted by a CMU-modified version of the open-source Nutch crawler. The crawl will be conducted from the Google/IBM cluster.

Languages

We want to provide very good coverage of English, but also to provide good coverage of a small number of other languages. We intend to over-sample English somewhat, so that we have very good coverage of at least one language. Nine other languages will be included in the dataset, providing coverage of the ten major languages used on the Internet (today).

The dataset percentages are determined as follows. The percentage of Internet users by language was obtained from Internet World Stats. We dropped languages not in the top 10, and renormalized. We set the percentage of English content at 50%. The remaining 50% was distributed across the other nine languages proportional to the number of Internet users who use that language. See the table below.

Rank	Language	Internet Users By Language[2]	Language as a % of the Top 10	Dataset Proportion
1	English	29.4%	34.7%	50.0%
2	Chinese	18.9%	22.2%	17.0%
3	Spanish	8.5%	10.0%	7.7%
4	Japanese	6.4%	7.6%	5.8%
5	French	4.7%	5.5%	4.2%
6	German	4.2%	4.9%	3.8%
7	Arabic	4.1%	4.8%	3.7%
8	Portuguese	4.0%	4.7%	3.6%
9	Korean	2.4%	2.8%	2.1%
10	Italian	2.4%	2.8%	2.1%
Rest	Others	15.1%	0.0%	0.0%

If we were to use the Wikipedia statistics for Internet Users By Language, we would have i) a larger skew towards English, ii) Arabic would be dropped, and iii) Russian would be included.

Language identification is performed by the TextCat language identification software[3]. In preliminary testing, it was 99.7% accurate, when given 400 or more bytes of text; accuracy is lower on shorter texts, for example 60-70% on 20 characters (very rough estimates). Thus, the actual distribution will vary slightly from our goals.

Language preferences are enforced in two ways. First, a preliminary estimate of page language is determined based on the language of the referring page and the URL domain. If the preliminary estimate is that the page is from a language that is not desired or over quota, its PageRank is discounted, so that crawling focuses on other languages, but it remains in the list of uncrawled URLs. If

[2] Internet World Stats (http://www.internetworldstats.com/stats7.htm).
[3] http://odur.let.rug.nl/~vannoord/TextCat/.

the page is subsequently crawled, i.e., its PageRank is high enough that it cannot be ignored even though the language is doubtful, it is crawled and its language reassessed based on its full contents. If the page is in a not desired or over quota language, it is dropped from the crawl, otherwise it is retained.

Display-Related Data

Relevance assessments are potentially more reliable when assessors can see complete page images, as opposed to the text-only portions of the page. We will attempt to capture the images, css style files, and other objects that affect how a page is displayed. We believe that such data is 2.2 to 3.0 times the size of the text data, i.e., 25 terabytes of text requires 60-75 terabytes of display-related data.

Our intent is to enable pages to be rendered accurately later, potentially much later (e.g., years after the data is collected). We considered rendering pages on-the-fly during crawling, but doing so would require significant computational effort, render many pages that nobody would ever look at, and perhaps most importantly, double the amount of image-related data (because even text-only pages would generate large image files).

Much of the display-related data consists of small graphics files, but some of it is large files. We may save only degraded copies of large images, to reduce storage costs.

Duplication of display-related data within a single site will be recognized by duplicated URLs. This decision assumes that display objects have stable URLs, and one display object does not have multiple distinct URLs within a single site. Although these assumptions will be violated occasionally, we do not expect that they will be violated sufficiently often to warrant more complex solutions. Monitoring software will check after each iteration to determine whether this assumption is causing problems.

Duplication of display-related data across sites (i.e., one image appearing on multiple sites) will be ignored. De-duplication, for example based on MD5 hashes, would seem to make sense for display-related data, due to its size. However, in preliminary testing, de-duplication involves significant costs (e.g., keeping track of MD5 hashes for billions of objects), while delivering surprisingly small storage savings (e.g., about 4%). Given the short timeline, it isn't clear that this capability is worth the effort.

A Few Things We Won't Do

We list a few things that we won't be doing to the dataset, just to be explicit.

- We will not make any attempt to identify or remove spam pages.
- We will not make any attempt to identify or remove pages with offensive content.

Dataset Organization

The crawl will be distributed as a large number of files in TREC web format, e.g., as are wt10g and gov2.

We estimate that the text portion of the crawl will be about 25 terabytes of uncompressed text, which can be distributed in compressed form on five 1-terabyte hard drives (approximate cost, US$ 1,000).

We recognize that some research groups will prefer to work with subsets of the data. The dataset will be organized into 5 segments of approximately 200 million documents each (about 5 terabytes, uncompressed, 1 terabyte compressed). Our analogy is the TREC CDs.

The segments can be organized either by language, or by crawl order. Our initial plan is to organize them by language, for example, English on segments 1-3, and other languages on segments 3-5 (note that segment 3 is 50% English), however this is an arbitrary choice; we have no real preference.

Most research groups will not need display-related data, so display-related data will be stored separately from the pages it is associated with. We estimate that this data will require 60-75 terabytes of storage, and will be of interest to only a very small group of researchers. It isn't clear that it will be practical to distribute this data to other organizations in the next few years, due to storage costs. It may be more practical for CMU or another organization to set up a web service that provides page images upon request.