CCIT Report #598
September 2006

Random Sampling from a Search Engine's Corpus*

Ziv Bar-Yossef† Maxim Gurevich‡

August 23, 2006

Abstract

We revisit a problem introduced by Bharat and Broder almost a decade ago: how to sample random pages from the corpus of documents indexed by a search engine, using only the search engine's public interface? Such a primitive is particularly useful in creating objective benchmarks for search engines.

The technique of Bharat and Broder suffers from a well-recorded bias: it favors long documents. In this paper we introduce two novel sampling algorithms: a lexicon-based algorithm and a random walk algorithm. Our algorithms produce *biased* samples, but each sample is accompanied by a *weight*, which represents its bias. The samples, in conjunction with the weights, are then used to *simulate* near-uniform samples. To this end, we resort to four well-known Monte Carlo simulation methods: *rejection sampling*, *importance sampling*, the *Metropolis-Hastings* algorithm, and the *Maximum Degree* method.

The limited access to search engines force our algorithms to use bias weights that are only "approximate". We characterize analytically the effect of approximate bias weights on Monte Carlo methods and conclude that our algorithms are *guaranteed* to produce near-uniform samples from the search engine's corpus. Our study of approximate Monte Carlo methods could be of independent interest.

Experiments on a corpus of 2.4 million documents substantiate our analytical findings and show that our algorithms do not have significant bias towards long documents. We use our algorithms to collect fresh comparative statistics about the corpora of the Google, MSN Search, and Yahoo! search engines.

1 Introduction

The latest round in the search engine size wars (cf. [36]) erupted in August 2005 after Yahoo! claimed [33] to index more than 20 billion documents. At the same time Google reported only 8 billion pages in its index, but simultaneously announced [5] that its index is three times larger than

*A preliminary version of this paper appeared in the proceedings of the 15th International World-Wide Web Conference (WWW2006) [4].

†Department of Electrical Engineering, Technion, Haifa 32000, Israel. Email: zivby@ee.technion.ac.il. Supported by the European Commission Marie Curie International Re-integration Grant.

‡Department of Electrical Engineering, Technion, Haifa 32000, Israel and IBM Research Lab in Haifa, Haifa 31905, Israel. Email: gmax@tx.technion.ac.il.