CCIT REPORT #743 August 2009

Efficient Search Engine Measurements*

Ziv Bar-Yossef † Maxim Gurevich ‡

August 30, 2009

Abstract

We address the problem of externally measuring aggregate functions over documents indexed by search engines, like corpus size, index freshness, and density of duplicates in the corpus. The recently proposed estimators for such quantities [5, 8] are biased due to inaccurate approximation of the so called "document degrees". In addition, the estimators in [5] are quite costly, due to their reliance on rejection sampling.

We present new estimators that are able to overcome the bias introduced by approximate degrees. Our estimators are based on a careful implementation of an approximate importance sampling procedure. Comprehensive theoretical and empirical analysis of the estimators demonstrates that they have essentially no bias even in situations where document degrees are poorly approximated.

By avoiding the costly rejection sampling approach, our new importance sampling estimators are significantly more efficient than the estimators proposed in [5]. Furthermore, building on an idea from [8], we discuss *Rao-Blackwellization* as a generic method for reducing variance in search engine estimators. We show that Rao-Blackwellizing our estimators results in performance improvements, while not compromising accuracy.

1 Introduction

Background. In this paper we focus on external methods for measuring aggregate functions over search engine corpora. These methods interact only with the public interfaces of search engines and do not rely on privileged access to internal search engine data or on specific knowledge of how the search engines work.

One application of our techniques is external evaluation of global quality metrics of search engines. Such evaluation provides the means for objective benchmarking of search engines. Such benchmarks can be used by search engine users and clients to gauge the quality of the service they get

*A preliminary version of this paper appeared in the proceedings of the 16th International World-Wide Web Conference (WWW2007) [3]. Supported by the Israel Science Foundation.

†Google Haifa Engineering Center, Israel, and Dept. of Electrical Engineering, Technion, Haifa 32000, Israel. Email: zivby@ee.technion.ac.il.

‡Dept. of Electrical Engineering, Technion, Haifa 32000, Israel. Email: gmax@tx.technion.ac.il. Supported by the Eshkol Fellowship of the Israeli Ministry of Science.