Adversarial Content Manipulation Effects

Fiana Raiber
Faculty of Industrial Engineering and Management
Technion – Israel Institute of Technology
Haifa 32000, Israel
fiana@tx.technion.ac.il

ABSTRACT

We address a question that has been somewhat overlooked throughout the transition from classical ad hoc retrieval to Web search: *how is the performance of classical retrieval approaches affected by the presence of content manipulation?* Our initial experiments have shown that the relative performance patterns of some classical retrieval strategies might change in the transition from non-manipulated to manipulated corpora. A natural future venue to explore is how to mix these strategies and make (some of) them more robust under presumed content manipulation conditions.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms: Algorithms, Experimentation

Keywords: ad hoc retrieval, keyword stuffing, content manipulation, cluster-based retrieval, passage-based retrieval

1. INTRODUCTION

Classical ad hoc (query-based) retrieval approaches focus on the document content for determining relevance to queries. On the other hand, the abundance of information over the Web has driven the utilization of additional information sources such as anchor text [9] and hyperlink structure [2]. Such information sources can potentially help search engines to cope, for example, with the *adversarial* nature of Web search. That is, the fact that some Web-page creators manipulate the content of pages so as to have them ranked high in response to popular queries, even if the pages do not pertain to the expressed information needs. However, as is typical in adversarial settings, manipulation with respect to these additional information sources, e.g., artificial creation of hyperlinks, has emerged [4]. Indeed, various forms of manipulation have been acknowledged as among the main challenges that search engines have to address [5]. Naturally, then, there is a large body of work on automatically detecting and handling different forms of manipulation [1, 3].

Throughout this ever evolving process of an adversarial search setting, a highly interesting question has been somewhat overlooked: how do classical retrieval methods that depend solely on the content of documents and some corpus-based term statistics perform in the presence of content manipulation? Intuitively, such approaches, which heavily rely on within-document term-frequency information, can be easily "mislead" by the simplest form of manipulation — *keyword stuffing* of popular query terms [4]. Indeed, early experiments in the Web setting have shown that the performance of classical retrieval methods is inferior to that of Web search engines that utilize, among others, hyperlink information [10].

Our initial experiments have shown that the relative performance patterns of some classical retrieval strategies can change in the transition from clean (non-manipulated) corpora to corpora containing content-manipulated documents. More specifically, we contrasted standard document-based retrieval with (i) passage-based document retrieval [7], (ii) cluster-based document retrieval [8], and (iii) pseudo-feedback-based retrieval [6].

There are numerous research challenges that rise from our findings. For example, improving the performance of (some of) these classical retrieval methods under presumed content manipulation conditions, and combining the different strategies and making them more robust to content manipulation. Developing new content-based retrieval methods that are resistant to manipulation and are better suited for the adversarial nature of the Web is another interesting research direction.

Acknowledgments This research is based upon joint work with Oren Kurland and Moshe Tennenholtz, and is supported in part by the Technion-Microsoft Electronic-Commerce Research Center and the Miriam and Aaron Gutwirth Memorial Fellowship.

2. REFERENCES

[1] *Proceedings of the AIRWeb workshop:adversial information retrieval on the Web*, 2005-2009.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW*, pages 107–117, 1998.

[3] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, 2006.

[4] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of AIRWeb*, pages 39–47, 2005.

[5] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[6] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of SIGIR*, pages 120–127, 2001.

[7] X. Liu and W. B. Croft. Passage retrieval based on language models. In *Proceedings of CIKM*, pages 375–382, 2002.

[8] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proceedings of SIGIR*, pages 186–193, 2004.

[9] O. A. McBryan. GENVL and WWWW: Tools for taming the Web. In *Proceedings of WWW*, 1994.

[10] A. Singhal and M. Kaszkiel. A case study in web search using TREC algorithms. In *Proceedings of WWW*, pages 708–716, 2001.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.