Inter-Search Engine Lexical Signature Performance

Martin Klein
Department of Computer Science
Old Dominion University
Norfolk, VA, 23529
mklein@cs.odu.edu

Michael L. Nelson
Department of Computer Science
Old Dominion University
Norfolk, VA, 23529
mln@cs.odu.edu

ABSTRACT

We generate lexical signatures (LSs) from web pages and acquire the mandatory document frequency values from three different search engine (SE) indexes. We cross-query the LSs against the two SEs they were not generated from and compare the retrieval performance by parsing the result set and analyzing the rank of the source URL.

Categories and Subject Descriptors

H.3.0 [**Information Storage and Retrieval**]:

General Terms

Measurement, Performance, Design

1. INTRODUCTION

A lexical signature (LS) is a small set of terms derived from the content of a document capturing its "aboutness". LSs of web pages have been shown to be suitable as search engine (SE) queries for discovering the page [3]. This scenario is useful when the content of a web pages has moved and the original URL returns the HTTP response code 404 or "Page Not Found". A LS consists of the top n terms with the highest term frequency (TF) - inverse document frequency (IDF) scores. IDF (unlike TF) depends on global knowledge. The number of documents in the corpus and the number of documents the term occurs in are needed. We compute LSs for web pages hence the corpus is the live web and values for the IDF computation can only be estimated.

2. THE EXPERIMENT

A common approach researchers have taken to estimate the term's DF is to query it against SEs and use the returned estimated number of results [1]. In this poster we use Google, Yahoo and MSN for this approach and generate LSs of 309 URLs randomly sampled from *DMOZ*. We cross-query all LSs against the two SEs they were not generated from. We compare the result sets in respect to the source URL. We query 5- and 7-term LSs since this setup has been shown to perform best for discovering missing pages [2]. Figure 1 shows the 5-term LS performance in all SEs. The labels on the axes indicate what SE the LSs were derived from and what SE they were queried against. G, M and Y stand for

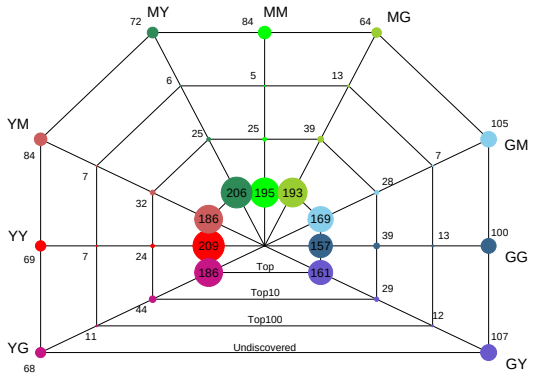

Figure 1: LS Performance in All Three SEs

Google, MSN and Yahoo respectively. The label GM, for example, represents LSs based on Google and queried against MSN. The size of the dots is proportional to the number of URLs returned within one of our categories (top, top10, top100, undiscovered). The absolute values are also plotted in the graph. MSN based LSs perform better when queried against Yahoo or Google than against MSN itself. They return more top ranked URLs (MY), more URLs in the top 10 and top 100 (MG) and leave fewer URLs undiscovered in both scenarios. Yahoo and Google based LSs perform best when queried against the SE they were generated from. Even though YG returns almost twice as many URLs in the top 10 than YY its performance in the top ranks is much worse.

3. REFERENCES

[1] F. Keller and M. Lapata. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, 29(3):459–484, 2003.

[2] M. Klein and M. L. Nelson. Revisiting Lexical Signatures to (Re-)Discover Web Pages. In *Proceedings of ECDL '08*, pages 371–382, 2008.

[3] S.-T. Park, D. M. Pennock, C. L. Giles, and R. Krovetz. Analysis of Lexical Signatures for Improving Information Persistence on the World Wide Web. *ACM Trans. Inf. Syst.*, 22(4):540–572, 2004.

JCDL'09, June 15–19, 2009, Austin, Texas, USA.
ACM 978-1-60558-322-8/09/06.