Analysis of the Diversity of Search Engine Results

Knowledge Lab

Birkbeck UNIVERSITY OF LONDON

Research Student
Suneel Kumar Kingrani

Supervisors
Mark Levene
Dell Zhang

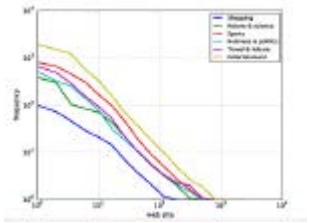

Figure 1. log-log plot - website frequency of Top-10 results

Birkbeck UNIVERSITY OF LONDON

Department of Computer Science and Information Systems

Introduction

Are web search results usually dominated by major websites e.g. www.youtube.com or www.wikipedia.en, and therefore lacking diversity? We aim to answer this question by quantitatively modelling the diversity of search results for popular queries. We use two diversity measures well-studied in ecology i.e. Simpson's diversity index and Shannon's diversity index. Our theoretical analysis shows how the diversity of search results is determined by the Zipfian distribution of websites. Our empirical analysis reveals that comparing Google and Bing, the former is more diverse in the top-50 search results, while the latter is more diverse in the top-10 search results.

Diversity

- Richness: Number of different species present.
- Evenness: Relative abundance of each species.
- We use diversity measures that take both Richness and Evenness into account.
  - Simpson's diversity index

$$D = \left(\sum_{i=1}^{N} p_i^2\right)^{-1}$$

  - Shannon's diversity index

$$H = -\sum_{i=1}^{N} p_i \ln p_i$$

Data

- Popular queries for 114 months in six representative categories of Google Top Charts.
- Shopping, Nature & Science, Sports, Business & Politics, Travel & Leisure, Entertainment.
- Top-k search results for those queries from Google and Bing.
- Extract host name from each search result's corresponding URL as its website address.

Theoretical Analysis

Our theoretical analysis shows how the diversity of search results is determined by the Zipfian distribution of websites as shown in figure 1.

- High skewness of the websites' distribution, a few popular websites (such as Wikipedia, Youtube, and Amazon) occur very frequently.

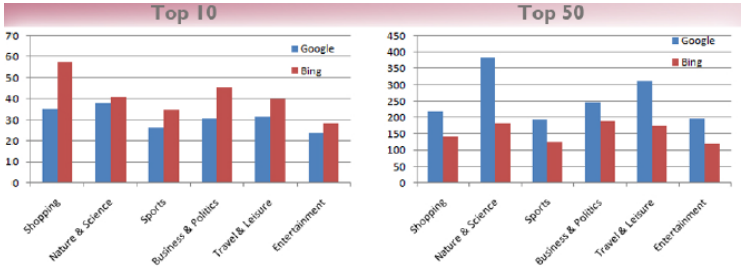

Figure 2. Website Diversity of Google and Bing

- Occurrence of each distinct website in the top-k search results is governed by the Zipf's law
- Able to compute the two diversity measures analytically e.g. Simpson's diversity index $D = G^2_{N,s}/G_{N,2s}$, where $G_{n,m} = \sum_{i=1}^{n} 1/i^m$ is the generalised harmonic number
- For larger s, the Zipfian distribution of websites is more skewed, thus there is less evenness and consequently less diversity.

For a given Zipfian distribution of websites with a specific ***s***, it turns out that the website evenness D_E or H_E actually decreases as the website richness ***N*** increases.

Empirical Analysis

How do Google and Bing compare against each other in the diversity of their search results?

- Richness:
  - Bing has higher richness than Google in top-10 search results
  - Google has higher richness than Bing in top-50 search results
- No clear winner for website evenness. But it is different from category to category.
- Since Google and Bing are roughly equivalent on the website evenness, their overall website diversity levels would just depend on the website richness. Diversity of Google and Bing is shown in figure 2.
  - Bing has higher diversity than Google for top-10 search results
  - Google has higher diversity than Bing for top-50 search results
- Using a randomisation test, we can confirm that the diversity difference between Google and Bing is statistically significant: the p-value is far less than 0.001 for almost all of the categories.

Publications

S.K. Kingrani, M. Levene, and D. Zhang. Diversity analysis of web search results. In Proceedings of the Annual International ACM Web Science conference, Oxford, UK, 2015.