Presentation 1

The Deep Web: Surfacing Hidden Value

Michael K. Bergman

Searching for Hidden-Web Databases

Luciano Barbosa & Juliana Freire

John Berlin

September 15, 2016

Old Dominion University
Introduction to Information Retrieval
CS734/834

Table of contents

Introduction

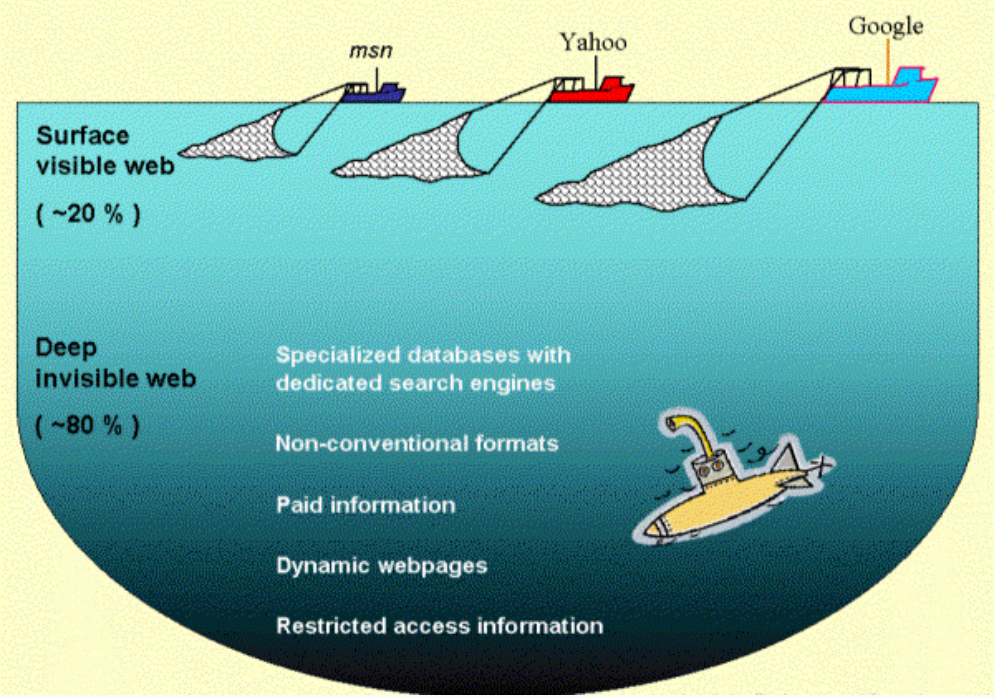
What is the Deep Web?

Content not indexed (crawled) by search engines

This content is characterized as *dynamic* and is generally generated as the result of a specific query

Where does the dynamic content come from?

- Databases
- Forms



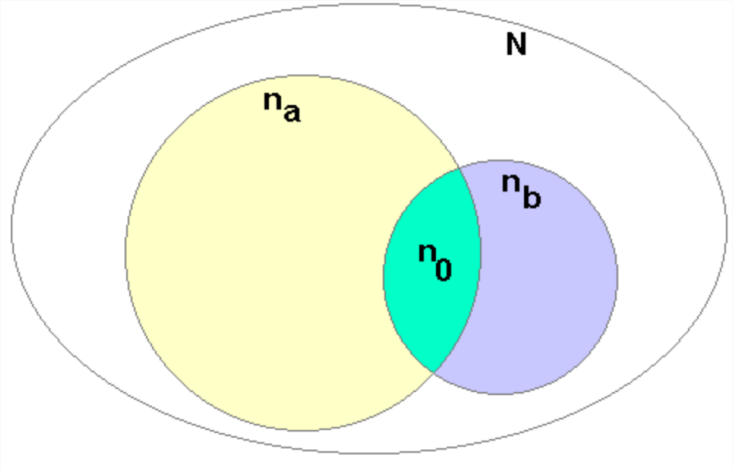
The Deep Web: Surfacing Hidden Value

Bergman's Contribution

- A quantification for the size of the deep Web
- A characterization of the deep Web's contentent
- Initial enumeration of the difficulties for retrieving deep web content

Quantification of the deep web

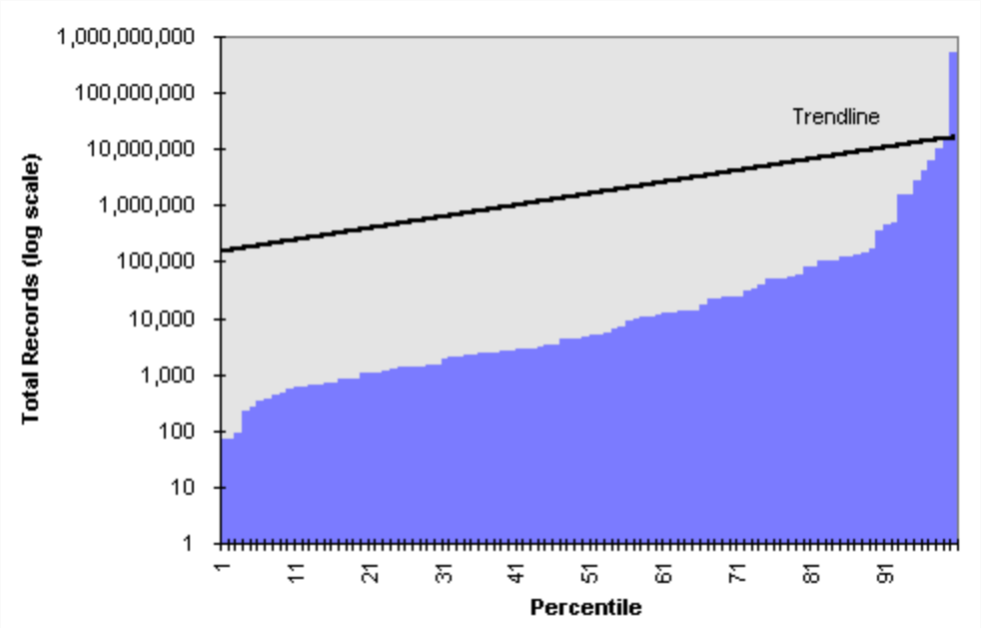
To quantify the deep web a pool of 53,220 urls was used
43,348 retrieved and 700 were randomly selected
13.6% $\approx$ 100 were found not to be search sites i.e. Google like
but provided a lower bounds size estimation by content overlap



Remember Lecture 1 WebSci

Quantification of the deep web

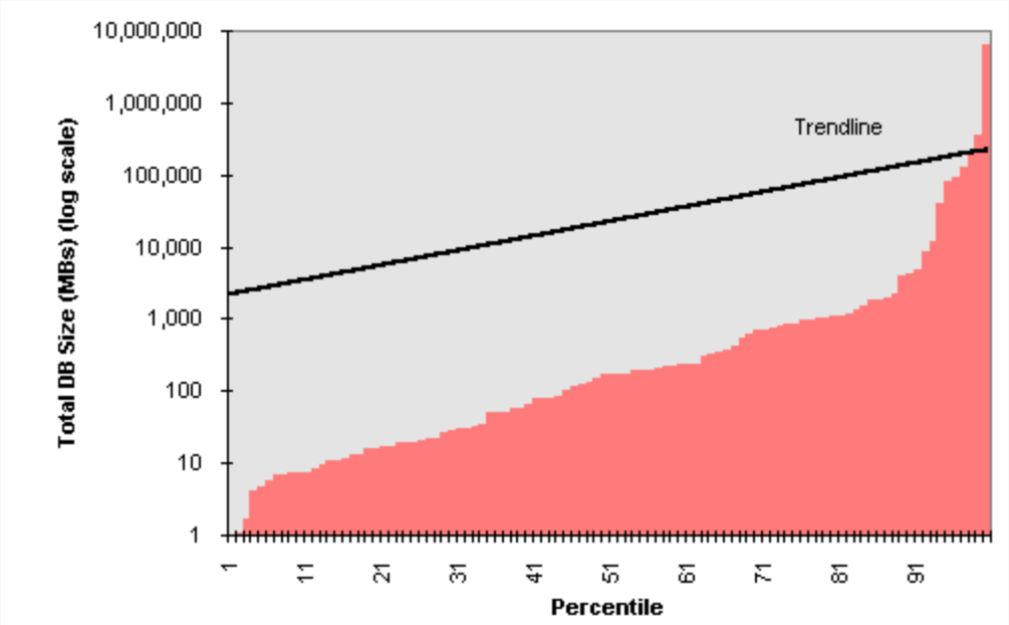
Another 100 random sites were chosen for content analysis
The html documents (records) per site was retrieved
Mean size of 13.7KB, median 19.7KB
Mean #documents of 5.43 million, median 4.95 thousand
From this they estimated > 200,000 total deep web sites
For a total of 543 billion documents



Inferred Distribution of Record Size

Quantification of the deep web

Along with the documents the databases were retrieved
Mean size 74.4 MB with median of 169 KB
Estimated total database size of 7.44 petabytes
Compared to 18.7 terabytes of the surface web at the time
60 deep web sites had already known database size totaling 750 terabytes



Inferred Distribution of Database Size

Characterization of the deep web

Revisiting the initial 43,348 urls

17,000 sites were selected

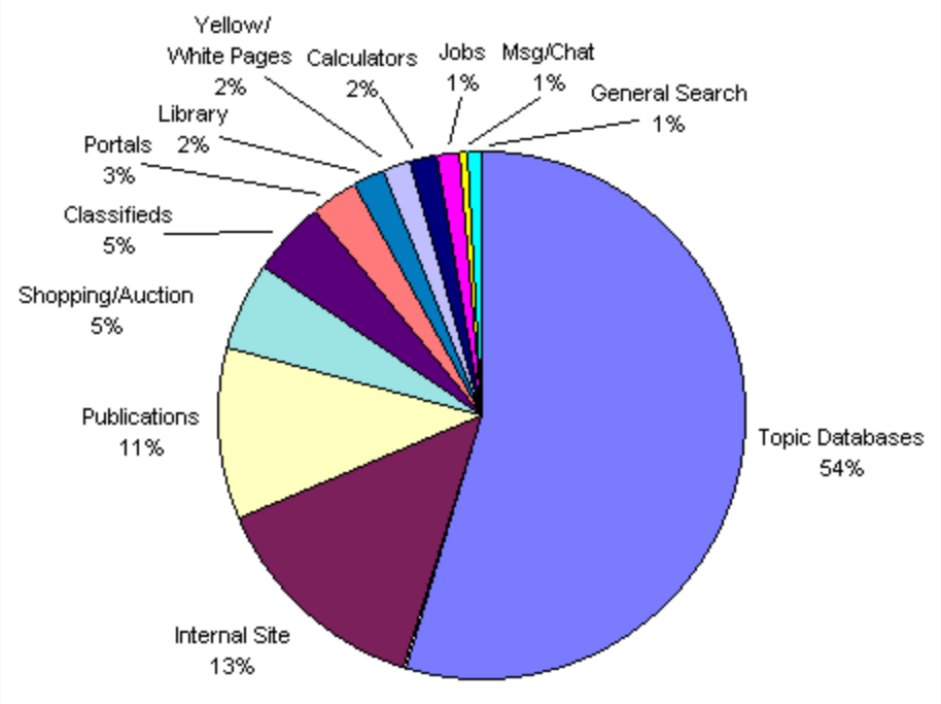
For subject and content analysis

It was found that they contained an uniform subject distribution

Deep Web Coverage	
Agriculture	2.7%
Arts	6.6%
Business	5.9%
Computing/Web	6.9%
Education	4.3%
Employment	4.1%
Engineering	3.1%
Government	3.9%
Health	5.5%
Humanities	13.5%
Law/Politics	3.9%
Lifestyles	4.0%
News, Media	12.2%
People, Companies	4.9%
Recreation, Sports	3.5%
References	4.5%
Science, Math	4.0%
Travel	3.4%
Shopping	3.2%

Characterization of the deep web

Topical databases, internal site documents and archived publications make up 80% of all deep web sites
E-commerce along with auction and classified sites 10%
Remaining sites 10%



Difficulties in retrieving deep web content

Database Content Retrieval Used In Study

Directed queries are necessary using 21m terms, 430k unique

For each new database 430k queries are needed

To get all of their contents

An infeasible task at scale

Difficulties in retrieving deep web content

Search Engines Use Breath Crawls

The query *URL:dmoz.org* was made to four major search engines

Dmoz or Open Directory had at the time subject structure of 248k categories

The search engines returned only a small percentage of expected results

Engine	OPD Pages	Yield
Open Directory (OPD)	248,706	---
AltaVista	17,833	7.2%
Fast	12,199	4.9%
Northern Light	11,120	4.5%
Go (Infoseek)	1,970	0.8%

Difficulties in retrieving deep web content

Leaving the question of how to effectively access the contents of the deep web databases and crawl sites in order to find links to the deeper content open

Searching for Hidden-Web Databases

Barbosa & Freire's Contribution

New Crawling Strategy to automatically discover hidden-web databases

Form-Focused Crawler