Presentation 1

The Deep Web: Surfacing Hidden Value

Michael K. Bergman, 2001, Journal of Electronic Publishing

Searching for Hidden-Web Databases

Luciano Barbosa & Juliana Freire, 2005, Proceedings of WebDB

John Berlin

September 15, 2016

Old Dominion University
Introduction to Information Retrieval
CS734/834

Table of contents

1. Introduction
2. The Deep Web: Surfacing Hidden Value
3. Searching for Hidden-Web Databases
4. Conclusion

Introduction

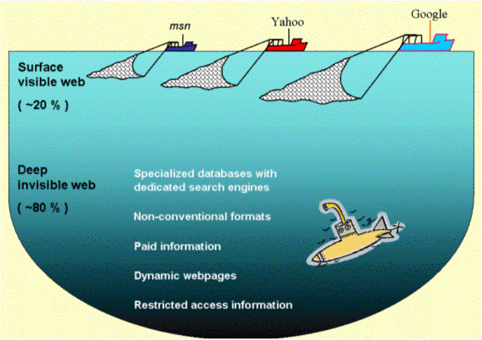
What is the Deep Web?

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

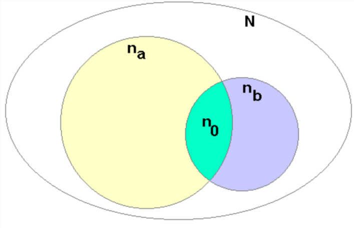
Quantification of the deep web

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
Figure 3 page 4

Quantification of the deep web

Another 100 random sites were chosen for content analysis

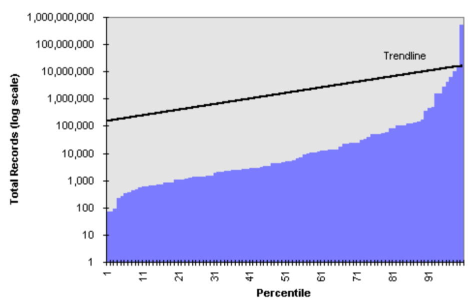
The html documents (records) per site was retrieved

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
Figure 4 page 8

Quantification of the deep web

Along with the documents the databases were retrieved

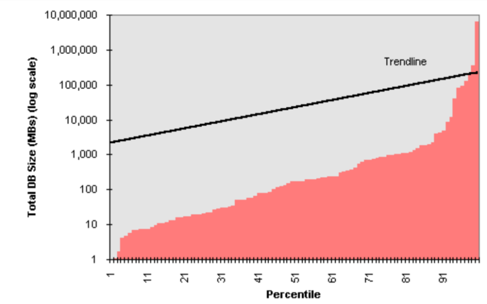
Mean size 74.4 MB with median of 169 KB

Estimated total database size of 7.44 petabytes

Compared to 18.7 terabytes of the surface web at the time

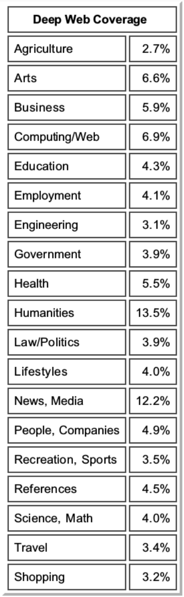
60 deep web sites had already known database size totaling 750 terabytes

Inferred Distribution of Database Size
Figure 5 page 9



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
Table 6, page 9 seen left shows these findings



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%

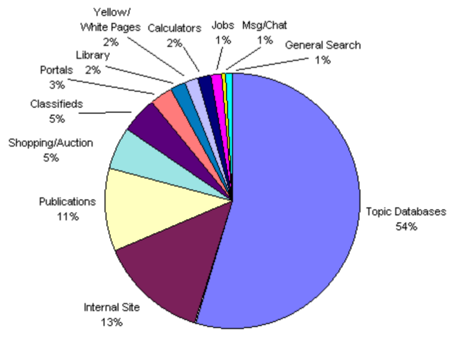


Figure 6 page 10

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

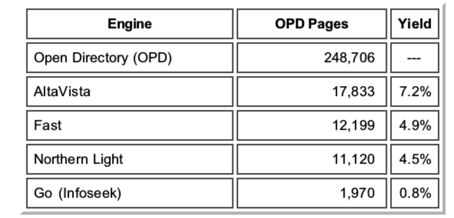
Difficulties in retrieving deep web content

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%

Table 7 page 10

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

Depth Focused Crawling

Avoid links that lead to off-topic regions

Back the crawler with a classifier to determine what is relevant

The classifier is trained on the pages belonging to topics in a taxonomy e.g. *dmoz.org*

Links are then given to another classifier to select the most promising links in the selected page.

Form-Focused Crawler

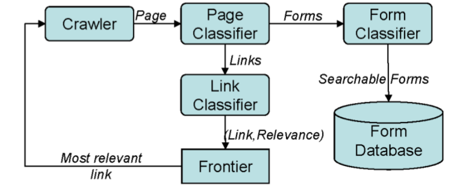
Crawler that understands form interfaces

Deep web use forms as the front end to databases

Must know what forms are searchable or not e.g. logins

And be domain-independent

To do this the crawler uses a third classifier for forms



Form Crawler Architecture

Figure 1 page 3

Classifier Setup

Link Classifier

Forms are sparsely distributed

Selecting links with immediate benefit means you miss forms

This classifier identifies links that bring *delayed benefit*

Or links that will *eventually* lead to forms

In order to know what links will do that depends on training

Classifier Setup

Link Classifier - Feature space by back crawling

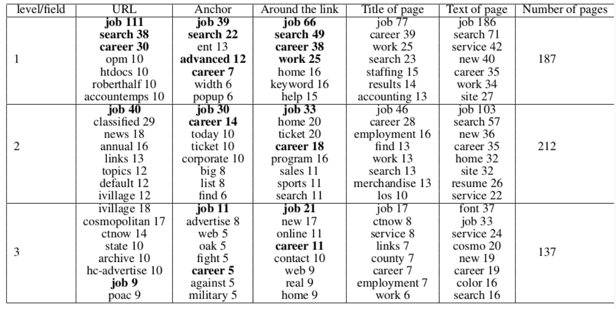
Approximation of the connectivity graph for a site

Using Google's “link:” searches conduct bread-first crawl

Starting with pages that have a searchable form *level 1*

Find links that point to the form level+1

Count features in url string and document text



level/field	URL	Anchor	Around the link	Title of page	Text of page	Number of pages
1	**job 111** **search 38** **career 30** opm 10 htdocs 10 roberthalf 10 accountemps 10	**job 39** **search 22** ent 13 **advanced 12** **career 7** width 6 popup 6	**job 66** **search 49** **career 38** **work 25** home 16 keyword 16 help 15	job 77 career 39 work 25 search 23 staffing 15 results 14 accounting 13	job 186 search 71 service 42 new 40 career 35 work 34 site 27	187
2	**job 40** classified 29 news 18 annual 16 links 13 topics 12 default 12 ivillage 12	**job 30** **career 14** today 10 ticket 10 corporate 10 big 8 list 8 find 6	**job 33** home 20 ticket 20 **career 18** program 16 sales 11 sports 11 search 11	job 46 career 28 employment 16 find 13 work 13 search 13 merchandise 13 los 10	job 103 search 57 new 36 career 35 home 32 site 32 resume 26 service 22	212
3	ivillage 18 cosmopolitan 17 ctnow 14 state 10 archive 10 hc-advertise 10 **job 9** poac 9	**job 11** advertise 8 web 5 oak 5 fight 5 **career 5** against 5 military 5	**job 21** new 17 online 11 **career 11** contact 10 web 9 real 9 home 9	job 17 ctnow 8 service 8 links 7 county 7 career 7 employment 7 work 6	font 37 job 33 service 24 cosmo 20 new 19 career 19 color 16 search 16	137

Feature Space For Job Domain ,Table 1 page 4

Classifier Setup

Page Classifier

Uses the Rainbow classifier, naïve Bayes

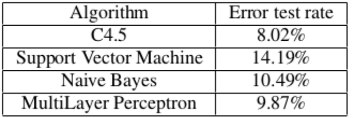
Trained on pages from *dmoz.org*

Gives a score if the page belongs to the focus topic

Form Classifier

Decision Tree classifier (C4.5) to determine searchability

Trained by finding number of tags, input fields, size of text and submission method (post or get)



Algorithm	Error test rate
C4.5	8.02%
Support Vector Machine	14.19%
Naive Bayes	10.49%
MultiLayer Perceptron	9.87%

Test error rates for different learning algorithms
Table 2 page 4

Crawl Strategy

Frontier Generation

N queues determined by the number of levels used by the link classifier

Prioritize links closer to target page

Queues ordered by likelihood of belonging to a level

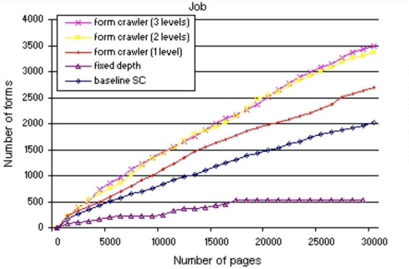
Stopping Criteria

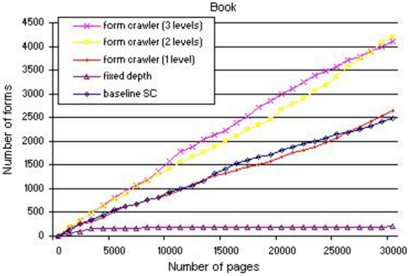
When a predetermined number of forms has been retrieved

Estimated 4.2 query interfaces (form) on any given deep web site

Visited maximum number pages on a site

Crawler Performance





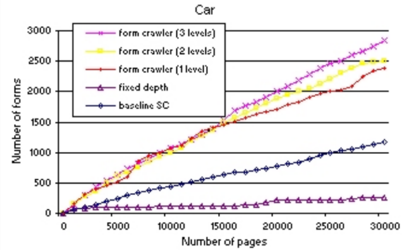


Figure 2 page 5

Using 3 vs 1 level configuration

Gain improvements of

20% to 30%

Conclusion

Conclusion

The Deep Web: Surfacing Hidden Value
Bergman provided a measurement for the deep web
Showed its content is highly relevant to surface web searches
Enumerated on the difficulties of crawling and extracting the content

Searching for Hidden-Web Databases
Barbosa & Juliana built a crawler to address the issue of database discovery that Bergman highlighted