

SURFACING THE DEEP WEB CONTENT

Presented By: Udochukwu Nweke
CS734: Introduction to Information Retrieval
Dr. Michael Nelson

September 14, 2017

Google's Deep-Web Crawl (Second Paper) (2008)

Jayant Madhavan
Google Inc.
jayant@google.com

David Ko
Google Inc.
dko@google.com

Łucja Kot *
Cornell University
lucja@cs.cornell.edu

Vignesh Ganapathy
Google Inc.
vignesh@google.com

Alex Rasmussen *
University of California, San
arasmuss@cs.ucsd.edu

Alon Halevy
Diego Google Inc.
halevy@google.com

White Paper: The Deep Web: Surfacing Hidden Value (Paper 2)

Bergman, Michael K. (2001)

Outline:

1. Introduction
 - a. What is the Deep Web
 - b. What is Surface Web
 - c. What Information Are In the Deep Web
 - d. The Deep Web Is 400 to 500 Times Larger Than The Surface Web
2. Deep Web Crawling Approaches
 - a. Naive Strategy Of Enumerating Cartesian Product of Possible Inputs
 - b. Google's Deep Web Crawling Approach
3. Limitations Of Google's Deep-Web Crawling Approach
4. Conclusion
5. Question and Answer

WHAT IS DEEP WEB?



Definition:

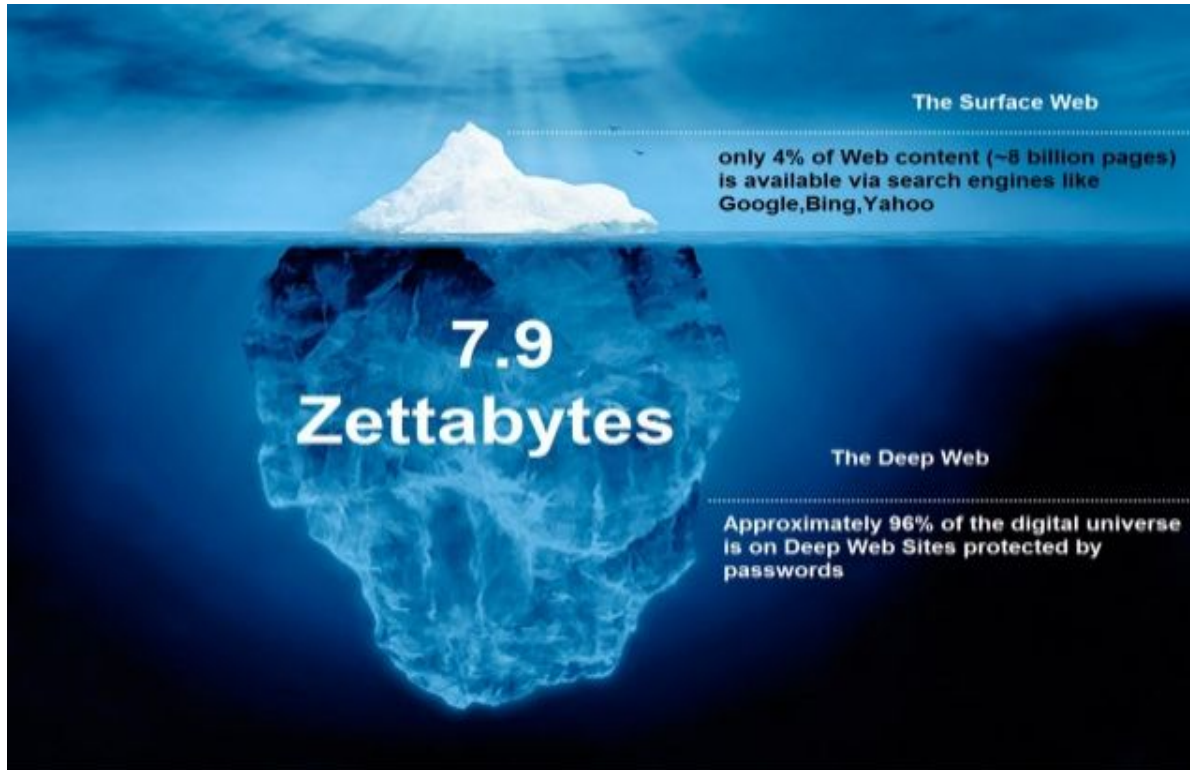
Web contents hidden behind HTML forms.

Deep Web Example:

https://www.odu.edu/directory?F_NAME=&L_NAME=n&SEARCH_IND=A

<https://www.deepweb-sites.com/>

What Is Surface Web?



Definition:

The surface web are web contents that are accessed by search engines or web crawlers.

Surface web example:

<http://www.cs.odu.edu/~mln/>

The Deep Web is 400 To 500 Times Larger Than The Surface Web

Surface Web

Content in the Deep Web is
massive— approximately

**500 Times
Greater**

than the content indexed by
conventional search engines

Journal of
Electronic
Publishing

A more recent study by Cyveillance has shown that the surface web is estimated to be 2.5 billion document in size and this is just a fraction of the deep web content which is about 550 billion individual documents.

What Information Are In the Deep Web?

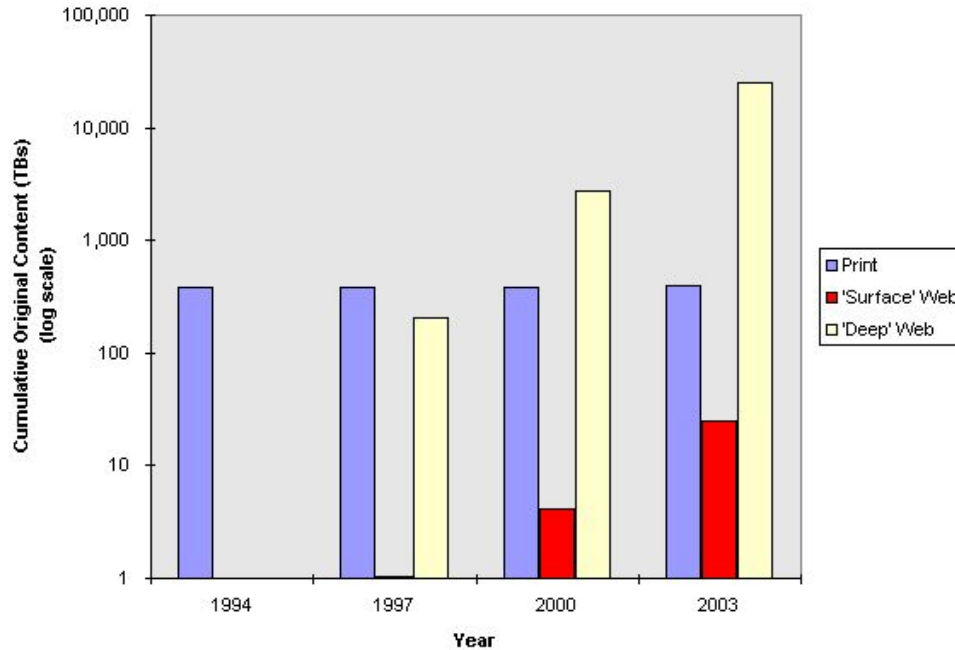


Why do we care about the deep web content: we miss out on valuable contents!

The deep web houses valuable contents such as legal documents, government resources, etc..

<https://www.deepwebtech.com/deepweb-not-darkweb/>

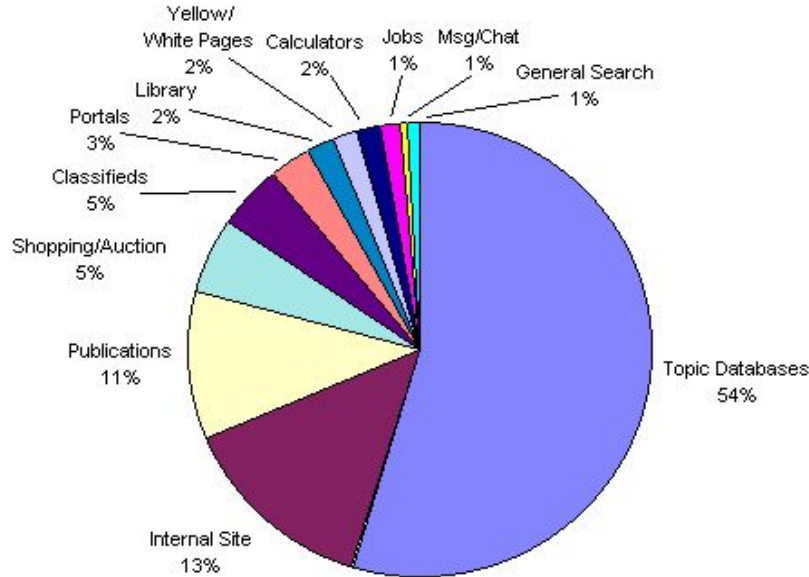
The Deep Web Is Growing Exponentially



If the surface web is growing at the rate of 7.5 million documents per day and the deep web is growing exponentially, what will be the size of the deep web?

Comparing Paper 1 and Paper 2 In Terms of Deep Web Contents

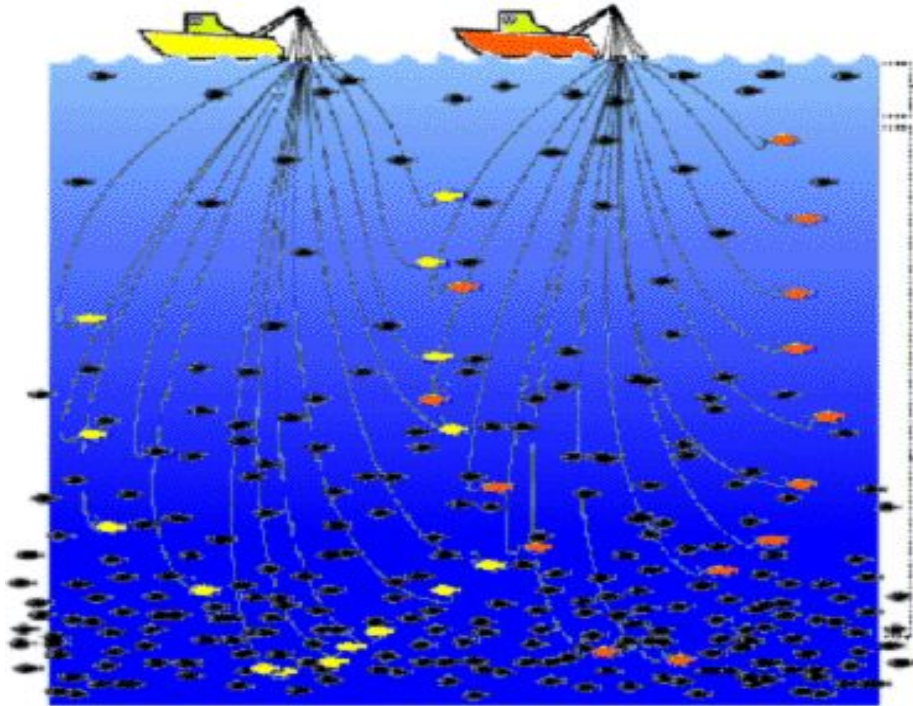
Paper 1



Paper 2

Structured Data:
This topical
databases

Deep Web Crawling Approaches



Surface Web

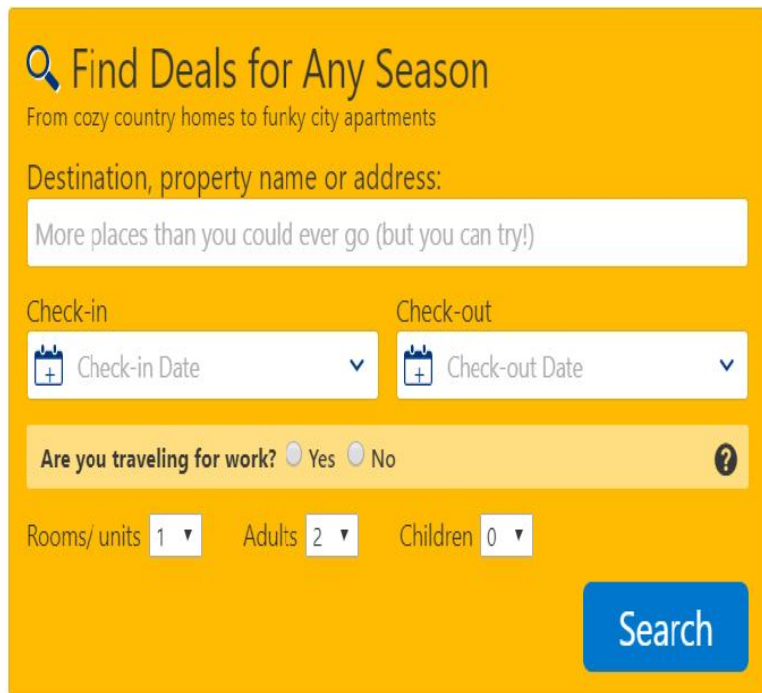
Can the deep web be crawled?

Deep Web

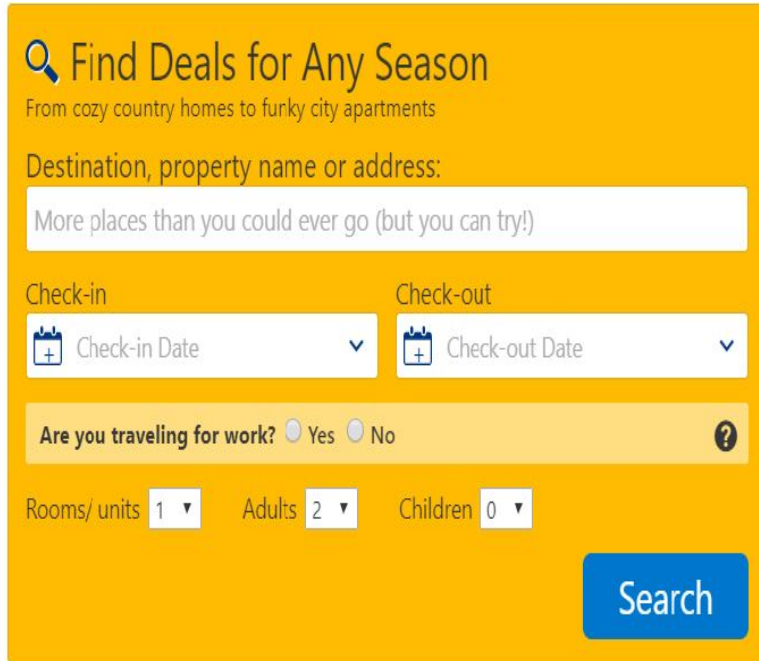
Comparing Paper 1 and Paper 2 In Terms Of Deep Web Crawling

Paper 1 Identifies The Impossibility of Indexing Deep Web Content Because:

- Specific queries have to be issued in order to retrieve database content.
- It is impractical to issue exponential number queries to surface the deep web content.
- Paper 1 proposed the naive way for surfacing the deep web content.

A screenshot of the Booking.com search interface. The header is orange with the text 'Find Deals for Any Season' and a subtext 'From cozy country homes to funky city apartments'. Below this is a search bar with the placeholder text 'Destination, property name or address:' and a hint 'More places than you could ever go (but you can try!)'. There are two date pickers for 'Check-in' and 'Check-out' with calendar icons. Below the date pickers is a section for 'Are you traveling for work?' with radio buttons for 'Yes' and 'No' and a help icon. At the bottom, there are dropdown menus for 'Rooms/ units' (set to 1), 'Adults' (set to 2), and 'Children' (set to 0). A blue 'Search' button is located at the bottom right of the form.

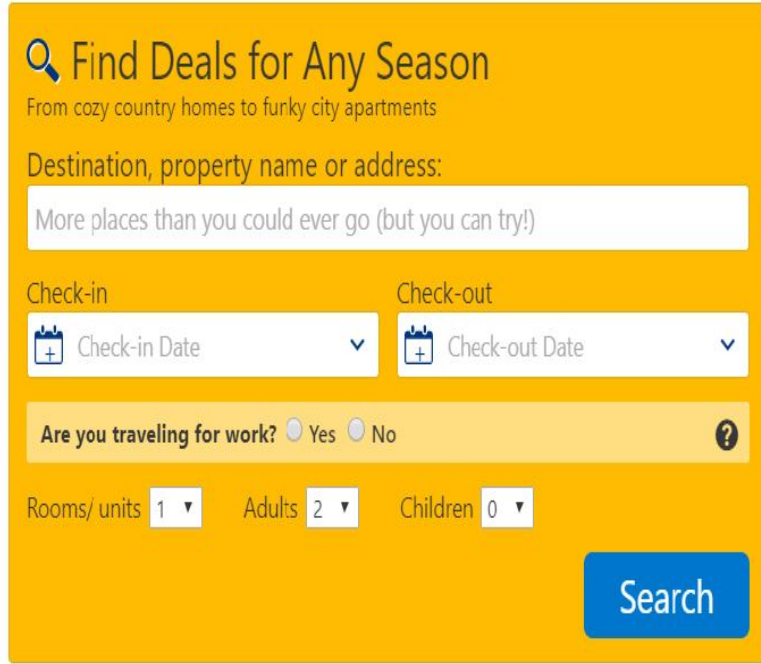
Comparing Paper 1 and Paper 2 In Terms Of Deep Web Crawling



The image shows a yellow search interface for Booking.com. At the top, it says 'Find Deals for Any Season' with a magnifying glass icon and a subtitle 'From cozy country homes to funky city apartments'. Below this is a text input field for 'Destination, property name or address:' with the placeholder text 'More places than you could ever go (but you can try!)'. Underneath are two date pickers for 'Check-in' and 'Check-out', each with a calendar icon and a dropdown arrow. Below the date pickers is a section for 'Are you traveling for work?' with radio buttons for 'Yes' and 'No', and a help icon. At the bottom are three dropdown menus for 'Rooms/ units' (set to 1), 'Adults' (set to 2), and 'Children' (set to 0). A large blue 'Search' button is located at the bottom right of the form.

- Paper 2 proposed an algorithm that efficiently navigates the search space of the input combination.
- This approach improves on the naive strategy proposed in paper 1

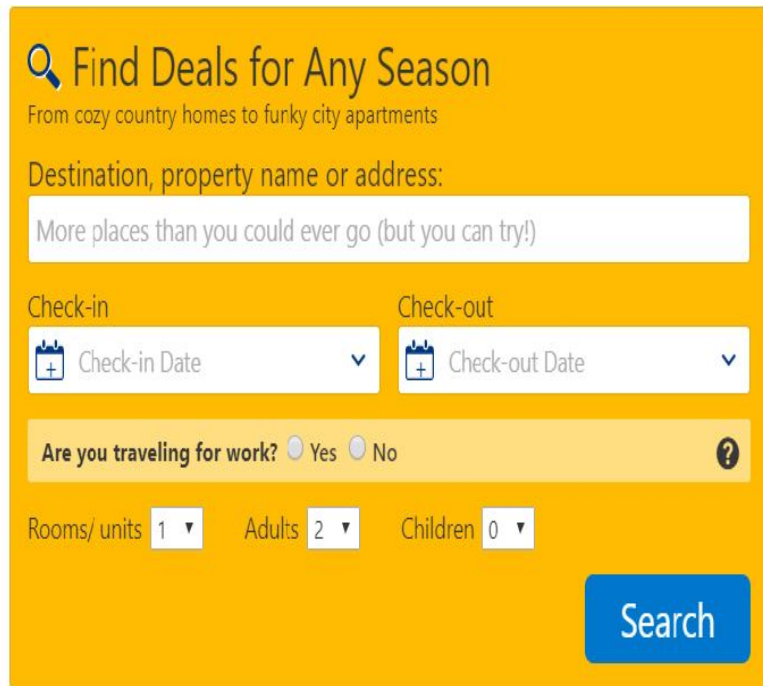
The Naive Strategy And challenges Of Deep Web Crawling



The image shows a screenshot of the Booking.com search interface. It features a yellow background with a magnifying glass icon and the text 'Find Deals for Any Season' and 'From cozy country homes to funky city apartments'. Below this is a text input field for 'Destination, property name or address:' with the placeholder text 'More places than you could ever go (but you can try!)'. There are two date selection fields for 'Check-in' and 'Check-out', each with a calendar icon and a dropdown arrow. Below these is a section for 'Are you traveling for work?' with radio buttons for 'Yes' and 'No', and a help icon. At the bottom, there are three dropdown menus for 'Rooms/ units' (set to 1), 'Adults' (set to 2), and 'Children' (set to 0). A blue 'Search' button is located at the bottom right.

- HTML pages have multiple input forms, enumerating the entire cartesian product of this form will generate a very large number of URL.
- The naive strategy cannot exhaust $2^n - 1$ possibilities
- To be explained on the board

Deep Web Surfacing Face Two Major Challenges



The image shows a yellow search form for Booking.com. At the top, it says 'Find Deals for Any Season' with a magnifying glass icon. Below that is a subtitle 'From cozy country homes to funky city apartments'. The main input field is labeled 'Destination, property name or address:' and contains the placeholder text 'More places than you could ever go (but you can try!)'. Below the input field are two date pickers: 'Check-in' and 'Check-out', each with a calendar icon and a dropdown arrow. Below the date pickers is a section for 'Are you traveling for work?' with radio buttons for 'Yes' and 'No', and a help icon. At the bottom are three dropdown menus for 'Rooms/ units' (set to 1), 'Adults' (set to 2), and 'Children' (set to 0). A blue 'Search' button is located at the bottom right of the form.

Find Deals for Any Season

From cozy country homes to funky city apartments

Destination, property name or address:

More places than you could ever go (but you can try!)

Check-in

Check-out

Check-in Date

Check-out Date

Are you traveling for work? ☐ Yes ☐ No

Rooms/ units 1

Adults 2

Children 0

Search

- Deciding the right input form to fill when submitting queries to the form
- Determining the right values to fill in these inputs

Google's Deep Web Crawling Approach

Google has presented an algorithm for:

- a. Selecting input values for text search input that accepts keywords
- b. Identifying inputs which accept only values of a specific type
- c. Efficiently navigating the search space of possible input combinations to identify only those that generate URLs suitable for inclusion into their web index.



The Goal Of Google's Deep Web Approach

The goal of this algorithm is to use a limited number of web inputs and get a reasonable coverage of the deep web while keeping content unique

Incremental Search For Informative Template Algorithm

```
GetInformativeQueryTemplates (W: WebForm)
  I: Set of Input = GetCandidateInputs(W)
  candidates: Set of Template =
    { T: Template | T.binding = {I}, I ∈ I }
  informative: Set of Template =  $\phi$ 
  while (candidates  $\neq \phi$ )
    newcands: Set of Template =  $\phi$ 
    foreach (T: Template in candidates)
      if ( CheckInformative(T, W) )
        informative = informative  $\cup$  { T }
        newcands = newcands  $\cup$  Augment(T, I)
    candidates = newcands
  return informative

Augment (T: Template, I: Set of Input)
  return { T' | T'.binding = T.binding  $\cup$  {I},
    I ∈ P, I  $\notin$  T.binding }
```

To be explained on
the board

Figure 2: Algorithm: Incremental Search for Informative Query Templates (ISIT)

Google's Deep-Web Crawling Limitations

Google's surfacing approach is encouraging but this approach didn't consider HTML forms powered by Javascript. Since there are millions of Javascript deep website, the deep web remains a challenge for web crawler and search engines

[Javascript Website](#)

References

1. Bergman, Michael K. "White paper: the deep web: surfacing hidden value." *Journal of electronic publishing* 7.1 (2001)
2. Madhavan, Jayant, et al. "Google's deep web crawl." *Proceedings of the VLDB Endowment* 1.2 (2008): 1241-1252.