Searching the Deep Web

What is Deep Web?

- ✶ Information accessed *only* through HTML form pages
  - database queries
  - results embedded in HTML pages
- (was) part of invisible Web
  - any information on Web can't search
    - Javascript output
    - unlabeled images, video, music, …
      - extract information?
    - pages sitting on servers with no pathes from crawler seeds

Extent of problem

- Estimates
  - 500 times larger than "surface" Web in terabytes of information
  - diverse uses and topics
    - 51% databases of Web pages behind query forms non-commercial (2004)
      - includes pages also reachable by standard crawling
    - 17% surface Web sites are not commercial sites (2004)
  - in 2004 Google and Yahoo each indexed 32% Web objects behind query forms
    - 84% overlap

Approaches to getting deep Web data

- Application programming interfaces
  - allow search engines get at data
  - a few popular site provide
  - not unified interfaces
- virtual data integration
  - a.k.a. mediating
  - done at time of user query
- Surfacing
  - a.k.a warehousing
  - build up HTML result pages in advance

Virtual Data Integration

- In advance:
  - identify pool of databases with HTML access pages
    - crawl
  - develop model and query mapping for each source
    - domains + semantic models
    - identify content/topics of source
    - develop "wrappers" to "translate" queries
- When receive user query:
  - from pool choose set of database sources to query
    - based on source content and query content
    - real-time content/topic analysis of query
  - develop appropriate query for each data source
  - integrate results for user

Virtual Integration: Issues

- Good for specific domains
- Doesn't scale well

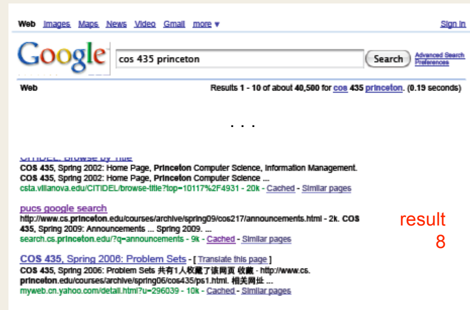
Surfacing

- In advance:
  - crawl for HTML pages containing forms that access databases
  - for each form
    - execute many queries to database using form
      - how choose queries?
    - index each resulting HTML page as part of general index of Web pages
  - pulls database information to surface
- When receive user query:
  - database results are returned like any other

7

Google query: cos 435 princeton

executed April 30 in AM



result 8

8

cached version of pucs google search

9

Surfacing: Google methodology

- Major Problem:

Determine queries to use for each form
  - determine templates
    - SELECT * FROM DB WHERE *predicates*
  - generate values for *predicates*
- Goal:

Good coverage of large number of databases
  - “Good”, not exhaustive
    - limit load on target sites during indexing
    - limit size pressure on search engine index
  - trading off depth within DB site for breadth of sites

10

Google: generating values

- generic text boxes: any words
  - select seed words from form page to start
    - tf-idf analysis
  - extract more keywords from initial form results
  - repeat until …
  - choose subset of keywords found
- typed text boxes: well-defined set values
  - type can be recognized with high precision
    - relatively few types over many domains
      - zip code, date, …
    - often distinctive input names
    - test types using sample of values

11

Google designers’ observations

- semantics not “significant role” in form queries
  - exceptions: correlated inputs
    - min-max ranges - mine collection of forms for patterns
    - keyword+database selection - hard
- user still gets fresh data
  - Search result gives URL with embedded DB query
    - doesn’t work for POST forms
- is now part of Google Search
  - in results of “more than 1000 queries per second”
- impact on “long tail of queries”
  - domain independent approach important

12

Next challenges

- Data behind Javascript code
  - mashups, visualizations
- Combining data from multiple sources
  - general, not custom, solution

13

Closing remarks

14

Where we started

"Google's mission is to organize the world's information and make it universally accessible and useful" Google's mission statement, ~ 1998.

"A memex is a device in which an an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory." Vannervar Bush, As we may think, *Atlantic Monthly*, July 1945.

15

Where we have been: major themes

- Models
  - information contents and queries
    - text
    - Web
    - audio, visual media
- Algorithms and data structures
  - Search: Indexing
  - Large data sets: Distributed computation
  - Clustering
  - Sampling

16

Where we have been: additional core issues

- Humans add information
  - users: characteristics & feedback
  - authors: semi-structured content
- What is the corpus?
  - Web crawling
- Evaluation!

17

What we missed

- Where is the content?
  - information caching
- Permanence of information
  - information preservation projects
- Semantic Web?
  - way beyond XML

18

Semantic Web

From W3C Semantic Web Activity Statement:

http://www.w3.org/2001/sw/Activity

"The goal of the Semantic Web initiative is as broad as that of the Web: to create a universal medium for the exchange of data. It is envisaged to smoothly interconnect personal information management, enterprise application integration, and the global sharing of commercial, scientific and cultural data."

19

Semantic Web Overview

- Initiative of W3C: World Wide Web Consortium - of academic, government and industry
  - begun 1994 by Tim Berners-Lee
- provides common frameworks for data specification allowing sophisticated functionality
  - Allowing automated understanding and use of information
- Open specifications, open source
  - Allow independently written tools interoperate

20

Where are "we" going?

- real semantic-based search
  " ... It is an enlarged **intimate** supplement to his memory." Vannervar Bush
- search everything
  - multi-media
  - data
  - social networks
  - cloud computing
- Deep Web → Semantic Web ?
  - structured data sets
  - interoperability

21

Major concerns

- Data explosion?
- Universal access?
  - Resource limitations
    - developing nations
- Security

22