Issues with Web Search Engines
SI650: Information Retrieval

Winter 2010

School of Information
University of Michigan
A Typical IR System Architecture

docs

INDEXING

Doc Rep

Query Rep

SEARCHING

Ranking

Feedback

INTERFACE

query

results

User

judgments

QUERY MODIFICATION
What about a Web Search Engine?
Web Search is Hard
The Data-User-Service (DUS) Triangle

**Lawyers**
**Scientists**
**Umich employees**
**Online shoppers**
...

**Users**

**Data**

- Web pages
- News articles
- Blog articles
- Literature
- Email
...

**Services**

- Search
- Browsing
- Mining
- Task support, ...

2010 © University of Michigan
Millions of Ways to Connect the DUS Triangle!

- Web pages
- Literature
- Organization docs
- Blog articles
- Product reviews
- Customer emails

Web Search

Enterprise Search

Literature Assistant

Opinion Advisor

Customer Rel. Man.

Everyone

Umich Employees

Scientists

Online Shoppers

Customer Service People

Search

Browsing

Alert

Mining

Task/Decision support
Indexer

• Standard IR techniques are the basis
  – Basic indexing decisions
  – Indexing efficiency (space and time)
  – Updating
• Additional challenges (in addition to scaling up)
  – Recognize spams/junks
  – Index selection – decide what to include in the index
  – Multi-level index
  – How to support “fast summary generation”?
• Google’s contributions:
  – Google file system: distributed file system
  – Big Table: column-based database
  – MapReduce: Software framework for parallel computation
  – Hadoop: Open source implementation of MapReduce (used in Yahoo!)
Redundancy Removal

• Detect duplications, and remove them.
• Strict duplication is easy to handle (why?), but what about near-duplication?
• 30% of web pages are near-duplicates (Broder et al 1997)
  – mirrors, local copies, page updates, …
  – spam, spider-traps, …
  – Crawling errors, …
• How to detect near-duplicates with limited storage and fast running-time?
Duplication Detection with Shingling

- Segment text into shingles
- TO BE OR NOT TO BE:
  - TO BE OR
  - BE OR NOT
  - OR NOT TO
  - NOT TO BE
- Large size of intersection/size of union → near-duplication
- Use Min-Hash to estimate the size of intersection
Overload of Data

Index size: Google/Yahoo/Bing ~ 20 Billion pages
Cuil ~ 100 Billion pages
MapReduce

• Provide easy but general model for programmers to use cluster resources
• Hide network communication (i.e. Remote Procedure Calls)
• Hide storage details, file chunks are automatically distributed and replicated
• Provide transparent fault tolerance (Failed tasks are automatically rescheduled on live nodes)
• High throughput and automatic load balancing (E.g. scheduling tasks on nodes that already have data)

This slide and the following slides about MapReduce are from Behm & Shah’s presentation
MapReduce Flow

Input = \{ Key, Value, Key, Value, \ldots \}

Map

Key, Value

Key, Value

\ldots

Split Input into Key-Value pairs.

For each K-V pair call Map.

Each Map produces new set of K-V pairs.

Sort

Reduce(K, V[ ])

Key, Value

Key, Value

\ldots

For each distinct key, call reduce. Produces one K-V pair for each distinct key.

Output as a set of Key Value Pairs.

Output = \{ Key, Value, Key, Value, \ldots \}
MapReduce WordCount Example

**Input:**
File containing words

Hello World Bye World
Hello Hadoop Bye Hadoop
Bye Hadoop Hello Hadoop

**Output:**
Number of occurrences of each word

Bye 3
Hadoop 4
Hello 3
World 2

**How can we do this within the MapReduce framework?**

**Basic idea:** parallelize on lines in input file!
MapReduce WordCount Example

Input
1. “Hello World Bye World”
2. “Hello Hadoop Bye Hadoop”
3. “Bye Hadoop Hello Hadoop”

Map Output
- <Hello,1>
- <World,1>
- <Bye,1>
- <World,1>
- <Hello,1>
- <Hadoop,1>
- <Bye,1>
- <Hadoop,1>
- <Bye,1>
- <Hadoop,1>

```java
Map(K, V) {
  For each word w in V
  Collect(w, 1);
}
```
MapReduce WordCount Example

Map Output

<Hello,1>
<World,1>
<Bye,1>
<Hadoop,1>
<Bye,1>
<Hadoop,1>
<Bye,1>
<Hadoop,1>
<Hello,1>
<Hadoop,1>

Internal Grouping

<Bye → 1, 1, 1>
<Hadoop → 1, 1, 1, 1>
<Hello → 1, 1, 1>
<World → 1, 1>

Reduce(K, V[ ]) {
    Int count = 0;
    For each v in V
        count += v;
    Collect(K, count);
}

Reduce Output

<Bye, 3>
<Hadoop, 4>
<Hello, 3>
<World, 2>
User Behaviors on Web

• The better you understand me, the better you can serve me

• Sample problems:
  – Identifying sessions in query logs
  – Predicting accesses to a given page (e.g., for caching)
  – Recognizing human vs. automated queries
  – Recommending alternative queries, landing pages, …
Query Log analysis

- Main idea: log the user behaviors/actions in web search
- Analyze the log to better understand the users

<table>
<thead>
<tr>
<th>AaxonID</th>
<th>Query</th>
<th>QueryTime</th>
<th>ItemRank</th>
<th>ClickURL</th>
</tr>
</thead>
<tbody>
<tr>
<td>100218</td>
<td>tennessee department of transportation</td>
<td>2006-03-01 11:08:30</td>
<td>1</td>
<td><a href="http://www.tdot.state.tn.us">http://www.tdot.state.tn.us</a></td>
</tr>
<tr>
<td>100218</td>
<td>state of tennessee emergency communications</td>
<td>2006-03-01 12:56:28</td>
<td>2</td>
<td><a href="http://www.tennessee.gov">http://www.tennessee.gov</a></td>
</tr>
<tr>
<td>100218</td>
<td>dixie youth softball</td>
<td>2006-03-02 10:36:48</td>
<td>2</td>
<td><a href="http://www.dixie.org">http://www.dixie.org</a></td>
</tr>
<tr>
<td>100218</td>
<td>cdwg</td>
<td>2006-03-03 14:29:07</td>
<td>1</td>
<td><a href="http://www.cdwg.com">http://www.cdwg.com</a></td>
</tr>
<tr>
<td>100218</td>
<td>cdwg scam cdwge</td>
<td>2006-03-03 14:30:11</td>
<td>1</td>
<td><a href="http://www.cdwg.com">http://www.cdwg.com</a></td>
</tr>
<tr>
<td>100218</td>
<td>escambia county sheriff’s department</td>
<td>2006-03-07 09:26:51</td>
<td>1</td>
<td><a href="http://www.escambiaiso.com">http://www.escambiaiso.com</a></td>
</tr>
<tr>
<td>100218</td>
<td>escambia county sheriff’s department</td>
<td>2006-03-07 09:26:51</td>
<td>2</td>
<td><a href="http://www.escambiaiso.com">http://www.escambiaiso.com</a></td>
</tr>
<tr>
<td>100218</td>
<td>escambia county sheriff’s department</td>
<td>2006-03-07 09:26:51</td>
<td>1</td>
<td><a href="http://www.escambiaiso.com">http://www.escambiaiso.com</a></td>
</tr>
<tr>
<td>100218</td>
<td>pensacola police department</td>
<td>2006-03-07 09:34:28</td>
<td>1</td>
<td><a href="http://www.penscolapolice.com">http://www.penscolapolice.com</a></td>
</tr>
<tr>
<td>100218</td>
<td>memphis pd</td>
<td>2006-03-07 09:42:33</td>
<td>1</td>
<td><a href="http://www.memphispolice.org">http://www.memphispolice.org</a></td>
</tr>
<tr>
<td>100218</td>
<td>nashville metro pd</td>
<td>2006-03-07 09:44:43</td>
<td>1</td>
<td><a href="http://www.police.nashville.org">http://www.police.nashville.org</a></td>
</tr>
<tr>
<td>100218</td>
<td>florida highway patrol</td>
<td>2006-03-07 09:48:35</td>
<td>1</td>
<td><a href="http://www.fhp.state.fl.us">http://www.fhp.state.fl.us</a></td>
</tr>
<tr>
<td>100218</td>
<td>tennessee highway patrol</td>
<td>2006-03-07 09:49:52</td>
<td>1</td>
<td><a href="http://www.state.tn.us">http://www.state.tn.us</a></td>
</tr>
<tr>
<td>100218</td>
<td>florida bureau of investigations</td>
<td>2006-03-07 09:51:08</td>
<td>2</td>
<td><a href="http://www.fbi.com">http://www.fbi.com</a></td>
</tr>
<tr>
<td>100218</td>
<td>florida bureau of investigations</td>
<td>2006-03-07 09:51:08</td>
<td>1</td>
<td><a href="http://www.fhp.state.fl.us">http://www.fhp.state.fl.us</a></td>
</tr>
<tr>
<td>100218</td>
<td>government finance officers association</td>
<td>2006-03-07 21:16:11</td>
<td>1</td>
<td><a href="http://www.comptroller.state.tn.us">http://www.comptroller.state.tn.us</a></td>
</tr>
<tr>
<td>100218</td>
<td>state of tennessee controllers manual</td>
<td>2006-03-07 21:17:12</td>
<td>1</td>
<td><a href="http://www.comptroller.state.tn.us">http://www.comptroller.state.tn.us</a></td>
</tr>
<tr>
<td>100218</td>
<td>state of tennessee controllers manual</td>
<td>2006-03-07 21:17:40</td>
<td>3</td>
<td><a href="http://www.fbr.state.tn.us">http://www.fbr.state.tn.us</a></td>
</tr>
<tr>
<td>100218</td>
<td>state of tennessee controllers manual</td>
<td>2006-03-07 21:17:40</td>
<td>4</td>
<td><a href="http://www.tbr.state.tn.us">http://www.tbr.state.tn.us</a></td>
</tr>
<tr>
<td>100218</td>
<td>state of tennessee controllers manual</td>
<td>2006-03-07 21:17:40</td>
<td>9</td>
<td><a href="http://audit.tennessee.edu">http://audit.tennessee.edu</a></td>
</tr>
<tr>
<td>100218</td>
<td>internal controls for municipalities under 10 000</td>
<td>2006-03-07 21:38:04</td>
<td>1</td>
<td><a href="http://www.nysscpa.org">http://www.nysscpa.org</a></td>
</tr>
<tr>
<td>100218</td>
<td>internal controls for municipalities under 10 000</td>
<td>2006-03-07 21:38:04</td>
<td>4</td>
<td><a href="http://www.massdor.com">http://www.massdor.com</a></td>
</tr>
<tr>
<td>100218</td>
<td>municipal fraud audit detection internal controls</td>
<td>2006-03-07 21:41:40</td>
<td>1</td>
<td><a href="http://www.whitehouse.gov">http://www.whitehouse.gov</a></td>
</tr>
<tr>
<td>100218</td>
<td>evaluating internal controls a local government managers guide</td>
<td>2006-03-07 21:51:18</td>
<td>5</td>
<td><a href="http://www.allbusiness.com">http://www.allbusiness.com</a></td>
</tr>
</tbody>
</table>
Query Log Analysis

The 250 most frequent queried terms in the “famous” AOL query log!
Thanks to http://www.wordle.net for the tagcloud generator
Query Log Analysis in Literature

<table>
<thead>
<tr>
<th>Query log name</th>
<th>Public</th>
<th>Period</th>
<th># Queries</th>
<th># Sessions</th>
<th># Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excite ‘97</td>
<td>Y</td>
<td>Sep ‘97</td>
<td>1,025,908</td>
<td>211,063</td>
<td>~ 410,360</td>
</tr>
<tr>
<td>Excite ‘97 (small)</td>
<td>Y</td>
<td>Sep ‘97</td>
<td>51,473</td>
<td>N.D.</td>
<td>~ 18,113</td>
</tr>
<tr>
<td>Altavista</td>
<td>N</td>
<td>Aug 2nd - Sep 13th ‘98</td>
<td>993,208,159</td>
<td>285,474,117</td>
<td>N.D.</td>
</tr>
<tr>
<td>Excite ‘99</td>
<td>Y</td>
<td>Dec ‘99</td>
<td>1,025,910</td>
<td>325,711</td>
<td>~ 540,000</td>
</tr>
<tr>
<td>Excite ‘01</td>
<td>Y</td>
<td>May ‘01</td>
<td>1,025,910</td>
<td>262,025</td>
<td>~ 446,000</td>
</tr>
<tr>
<td>Altavista (public)</td>
<td>Y</td>
<td>Sep ‘01</td>
<td>7,175,648</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>Tiscali</td>
<td>N</td>
<td>Apr ‘02</td>
<td>3,278,211</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>TodoBR</td>
<td>Y</td>
<td>Jan - Oct ‘03</td>
<td>22,589,568</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>TodoCL</td>
<td>N</td>
<td>May – Nov ‘03</td>
<td>N.D.</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>AOL (big)</td>
<td>N</td>
<td>Dec 26th ‘03 – Jan 1st ‘04</td>
<td>~ 100,000,000</td>
<td>N.D.</td>
<td>~ 50,000,000</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>N</td>
<td>Nov ‘05 – Nov ‘06</td>
<td>N.D.</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>AOL (small)</td>
<td>Y</td>
<td>Mar 1st – May 31st ‘06</td>
<td>36,389,567</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
</tbody>
</table>

Mei and Church 08: MSN Search – 18 months, 637 million unique queries, 585 million unique urls, 193 million unique IP addresses

- Enhance ranking – retrieval, advertisement
- Query suggestion; refinement; expansion; substitution, …
- Spelling check
- Other tasks …
Main Results

- Average number of terms in a query is ranging from a low of 2.2 to a high of 2.6
- The most common number of terms in a query is 2
- The majority of users don’t refine their query
  - The number of users who viewed only a single page increase 29% (1997) to 51% (2001) (Excite)
  - 85% of users viewed only first page of search results (AltaVista)
- 45% (2001) of queries is about Commerce, Travel, Economy, People (was 20% 1997)
  - The queries about adult or entertainment decreased from 20% (1997) to around 7% (2001)

- This slide is from Pierre Baldi
Main Results

• All four studies produced a generally consistent set of findings about user behavior in a search engine context
  – most users view relatively few pages per query
  – most users don’t use advanced search features

- Query Length Distributions (bar)
- Poisson Model (dots & lines)

\[
P(r) = \frac{\lambda^{r-1}}{(r-1)!} e^{-\lambda}
\]
Power-law Characteristics

- Frequency $f(r)$ of Queries with Rank $r$
  - 110000 queries from Vivisimo
  - 1.9 Million queries from Excite
- There are strong regularities in terms of patterns of behavior in how we search the Web

This slide is from Pierre Baldi
A sketch of a searcher… “moving through many actions towards a general goal of satisfactory completion of research related to an information need.” (after Bates 90)

(Slide from “Marti Hearst, UCB SIMS, Fall 98)
Query Session Detection

• Roughly defined as queries that are submitted by the same user in a short period of time

• Hypothesis:
  – Queries in the same session are related
  – Queries in the same session reflect the same mission/task, etc.
  – Queries in the same session reflect the “modification” relationship

• How to segment query sequence into sessions?
• Heuristic methods; Machine learning methods (hidden Markov models, conditional random fields, etc)
Example – A Poet’s Corner

AOL User 23187425 typed the following queries within a 10 minutes time-span:

• you come forward 2006-05-07 03:05:19
• start to stay off 2006-05-07 03:06:04
• i have had trouble 2006-05-07 03:06:41
• time to move on 2006-05-07 03:07:16
• all over with 2006-05-07 03:07:59
• joe stop that 2006-05-07 03:08:36
• i can move on 2006-05-07 03:09:32
• give you my time in person 2006-05-07 03:10:07
• never find a gain 2006-05-07 03:10:47
• i want change 2006-05-07 03:11:15
• know who iam 2006-05-07 03:11:55
• curse have been broken 2006-05-07 03:12:30
• told shawn lawn mow burn up 2006-05-07 03:13:50
• burn up 2006-05-07 03:14:14
• was his i deal 2006-05-07 03:15:13
• i would have told him 2006-05-07 03:15:46
• to kill him too 2006-05-07 03:16:18
Query Reformulation – Spelling Correction

[Cucerzan and Brill, 2004]
Query Suggestions

Yahoo!

Web Images Video Local Shopping More v

digital camera

Search

Explore related concepts:
digital camera reviews
megapixels

Nikon

digital camera ratings

canon digital camera

explore related concepts:
digital camera reviews
efficiency

az

Samsung

optical zoom
easyshare

Sony

Digital Cameras

Digital Camera Reviews and News: Digital Photography Review...

A digital camera is a camera that takes video or still photographs, or both, digitally by recording images via an electronic image sensor. En.wikipedia.org/wiki/Digital_camera - 117k - Cached

Digital Camera Reviews and News: Digital Photography Review...

Camera and accessory reviews, digital photography and imaging news, discussion forums, sample images, buyer's guides with side-by-side comparisons, and a database. www.dpreview.com - 60k - Cached

Digital Cameras - High-End, Advanced Digital Cameras...

Shoot with ease and style with the Canon PowerShot cameras including Digital ELPH series. PowerShot digital cameras incorporate the creative performance of a... usa.canon.com/consumer/controller?act=ProductCatIndexAct&... - 100k - Cached

Best Digital Cameras in the Kodak EASYSHARE line including...

Capture Top Quality Pictures With Digital Cameras Ranging from 8MP, 9MP, 10MP, 12MP & 14MP. Whether You are Looking for Performance, Sleek & Stylish, or Point ... store.kodak.com/store/ekconsus/en_US/ist/... - 86k - Cached

Digital Cameras

The Easiest Way To Shop Online. Compare Prices & Discounts. www.google.com/Products

Kodak EasyShare Cameras


SONY Digital Cameras

Free Shipping & Exclusive Deals at SONY Cyber-shot Official Site. www.SonyStyle.com/Cyber-shot

Canon PowerShot G11 Digital...

Every day low prices on your favorite Canon camera. Order online. www.SaveHereDigital.com

Top 10 Digital Cameras

Compare Top Brand Digital Cameras! We List the Lowest Price Stores. Digital-Cameras.compare247.us

PENTAX Digital Cameras

From easy-to-use to advanced. Learn more at the PENTAX Official Site. www.pentaximaging.com

Nikon Coolpix S1000PJ $419

Buy Now & Take Advantage of Our
Generating Query Suggestion using Click Graph

- Construct a (kNN) subgraph from the query log data (of a predefined number of queries/urls)
- Compute transition probabilities $p(i \rightarrow j)$
- Compute hitting time $h_i^A$
- Rank candidate queries using $h_i^A$
Beyond Query Logs?

- Browsing logs
- Eye-tracking logs
- Social bookmarks?

Nested search sequences – Mei et al. 09
Eye Tracking (Golden Triangle)

- Google Eye Tracking Heat Map, Eyetools Eyetracking Research
Retriever

- Standard IR models apply but aren’t sufficient
  - Different information need (home page finding vs. topic-driven)
  - Documents have additional information (hyperlinks, markups, URL)
  - Information is often redundant and the quality varies a lot
  - Server-side feedback is often not feasible

- Major extensions
  - Exploiting links (anchor text, link-based scoring)
  - Exploiting layout/markups (font, title field, etc.)
  - Spelling correction
  - Spam filtering
  - Redundancy elimination

- In general, rely on machine learning to combine all kinds of features
Learning to Rank

• In a web search engine, relevance models like BM25 is just one of the many features. How to balance the features? Tune parameters? - Machine learning!

• Pointwise (e.g., logistic regression)
  – Input: single documents
  – Output: scores or class labels

• Pairwise (e.g., RankSVM)
  – Input: document pairs
  – Output: partial order preference

• Listwise (e.g., LambdaRank)
  – Input: document collection
  – Output: ranked list of documents
Personalized Search

• Ambiguous query: MSR
  – Microsoft Research
  – Mountain Safety Research

• Disambiguate based on user’s prior clicks

• If you know who I am, you should give me what I want

• Research issues:
  – What if we don’t have enough history?
  – History v.s. new information needs
  – Privacy, privacy, privacy!
Computational Advertising

• Main Challenge:
  – Find the "best match" between a given user in a given context and a suitable advertisement.

• Examples
  – Context = WEB search query → Sponsored search
  – Context = publisher page → Content match, banners
  – Other contexts: mobile, video, newspapers, etc.
Computational Advertising

• A new scientific sub-discipline that provides the foundation for building online ad retrieval platforms
  – To wit: given a certain user in a certain context, find the most suitable ad

• At the intersection of
  – Large scale text analysis
  – Information retrieval
  – Statistical modeling and machine learning
  – Optimization
  – Microeconomics
Placing Ads - Sponsored Links

query

slots of ads, the higher the better
Research Issues in Computational Advertising

• How to price? How to charge? How to bid? How to encourage bidding?
  – CPM, CPC, CPT/CPA
• How to evaluate?
• How to order ads? What to optimize?
  – Revenue?
  – Relevance?
  – What else?
• Exact match v.s. advanced match
• Understanding the intent of queries
• When to show ads, when not to show ads?
Adversarial Information Retrieval

• We looked at spamming in the context of Naïve Bayes
• Let’s now consider spamming of hyperlinked IR
• The main idea: artificially increase your in-degree/pagerank, etc.
• Question: How to increase your pagerank?
• Link farms: groups of pages that point to each other.
• Google penalizes sites that belong to link farms
User Interfaces

- Offer informative feedback
- Reduce working memory load
- Provide alternative interfaces for novice and expert users
- Basic elements:
  - Document set selection
  - Query specification
  - Result examination
  - Interaction support (feedback)
Example - rich document surrogate information

The small molecule Mek1/2 inhibitor U0126 disrupts the chordamesoderm to notochord transition in zebrafish

Distinct Functions for Different sd Isoforms in Zebrafish Primitive and Definitive Hematopoiesis
Example - Cuil
Next Generation Search Engine?

- Better support for query formulation
  - Allow querying from any task context
  - Query by examples
  - Automatic query generation (recommendation)
- Better search accuracy
  - More accurate information need understanding (more personalization and context modeling)
  - More accurate document content understanding (more powerful content analysis)
- More complex retrieval criteria
  - Consider multiple utility aspects of information items (e.g., readability, quality, communication cost)
  - Consider collective value of information items (context-sensitive ranking)
- Better result presentation
  - Better organization of search results to facilitate navigation
  - Better summarization
- More effective and robust retrieval models
  - Automatic parameter tuning
- More scalable retrieval architecture
  - P2P
Next Generation Search Engines?

• More specialized/customized
  – Special group of users (community engines, e.g., Citeseer)
  – Personalized (better understanding of users)
  – Special genre/domain (better understanding of documents)
• Learning over time (evolving)
• Integration of search, navigation, and recommendation/filtering (full-fledged information management)
• Beyond search to support tasks (e.g., shopping)