Web Search Basics

Introduction to Information Retrieval
INF 141/ CS 121
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org
Overview

- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
  - Size of the Web
- Web Users
- Spam
Classic IR assumptions

- Corpus: Fixed document collection
- Goal: Retrieve information content relevant to information need
Classic IR Goal

- Classic “Relevance”
  - For each query, Q, and stored document, D, in a corpus there exists a relevance score: \( R(Q,D) \)
  - \( R(Q,D) \) is averaged over users, U, and contexts, C
  - Maximize \( R(Q,D) \) instead of \( R(Q,D,U,C) \)
    - Context is ignored
    - Individuals are ignored
    - Corpus is static
Overview

- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
  - Size of the Web
- Web Users
- Spam
Web IR: Differences from traditional IR

- On the web, search and ads are intricately connected
- The web is huge
- The web is a rapidly changing collection.
- There is spam on the web
  - Adversarial IR
  - Huge difference from traditional IR
- One interface for hugely divergent needs
  - Queries, Maps, Stocks, Weather, Calculations
History

• Early keyword-based engines
  • (1995-1997) Altavista, Excite, Infoseek, Inktomi

• Paid placement ranking
  • Goto.com -> Overture.com -> Yahoo!
    • Results based on auction for keyword placement
1. **Wilmington Real Estate - Buddy Blake**
   Wilmington's information and real estate guide. This is your one and only thing to do with Wilmington.
   www.buddyblake.com (Cost to advertiser: $10.28)

2. **Coldwell Banker Sea Coast Realty**
   Wilmington's number one real estate company.
   www.cbseacoast.com (Cost to advertiser: $10.37)

3. **Wilmington, NC Real Estate Becky Bullard**
   Everything you need to know about buying or selling a home on my website!
   www.iwwc.net (Cost to advertiser: $10.25)
History

- (1998+) Link-based ranking pioneered by Google
- Links added the idea of “authoritativeness” to “relevance”
- Blew away all early engines save Inktomi
- Great user experience looking for a business model
- Meanwhile Goto/Overture’s annual revenues were nearing $1 billion
History

- Result
  - Google:
    - Added paid placement ads on the side
    - Differentiated from search results
  - Yahoo! built a similar architecture
    - Buys Overture for paid placement
    - Buys Inktomi for search
Overview

- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
  - Size of the Web
- Web Users
- Spam
Google has maintained that ads (based on vendors bidding for search queries) do not affect vendors ranking in search results.
Ranking of ads

• Other search engines (Yahoo!, MSN) have made similar statements on occasion

• Any of them can change at any time

• Facebook is currently testing the waters in their “Newsfeeds”

• We will ignore the possibility of paid placement ads being interspersed in search results.
Ranking of ads

- **Goto model:**
  - Rank according to how much advertiser pays

- **Current model:**
  - Balance auction price and relevance
  - Irrelevant ads (few click-throughs)
    - Decrease opportunities for relevant ads
  - Harm the user experience

- Idea: Well-targeted advertising is good for everyone
Paying for advertisements

- CPM
  - “Cost Per Mil”
  - Pay for 1000 eyeballs
  - Important for branding campaigns
- CPC
  - “Cost per Click”
  - Pay for clicking on ads
  - Important for sales campaigns
Overview

- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
  - Size of the Web
- Web Users
- Spam
The Web Corpus

- No design/coordination
- Distributed content creation, linking
- “Democratization of publishing”
- Content includes truth, lies, contradictions, etc.
- Unstructured Data (text, html)
- Semi-Structured (XML, annotated photos)
- Structured (Databases)
- Scale is much larger than previous text corpora
The Web Corpus

- Growth - slowing from “doubling every few months”, but still expanding
Dynamic Content
• Content can be dynamically generated
• There is no static HTML version
• Flight status information, evite responses
• Assembled on request (“?” in URL is a clue)
Web Search Basics

Dynamic Content
- Most (truly) dynamic content is ignored by web spiders
  - Too much to index
  - Static information is more important for search
  - Spider Traps look dynamic
- Actually a lot of “static” content is assembled on the fly
  - ASP, PHP, JSP, ads, etc....