



# Searching the Web

Frank McCown  
Introduction to Web Science  
Harding University  
Spring 2011

# How do you locate information on the Web?

- When seeking information online, one must choose the best way to fulfill one's *information need*
- Most popular:
  - Web directories
  - **Search engines** – primary focus of this lecture
  - Social media

# Web Directories

- Pages ordered in a hierarchy
- Usually powered by humans
- Yahoo started as a web directory in 1994 and still maintains one: <http://dir.yahoo.com/>
- Open Directory Project (ODP) is largest and is maintained by volunteers  
<http://www.dmoz.org/>

Search the entire directory

**Top: Sports: Football: American: NFL** (1,260)

[Description](#)

- [Arizona Cardinals](#) (26)
  - [Atlanta Falcons](#) (27)
  - [Baltimore Ravens](#) (33)
  - [Buffalo Bills](#) (45)
  - [Carolina Panthers](#) (20)
  - [Chicago Bears](#) (31)
  - [Cincinnati Bengals](#) (21)
  - [Cleveland Browns](#) (53)
  - [Dallas Cowboys](#) (54)
  - [Denver Broncos](#) (30)
  - [Detroit Lions](#) (36)
  - [Green Bay Packers](#) (47)
  - [Houston Texans](#) (17)
  - [Indianapolis Colts](#) (25)
  - [Jacksonville Jaguars](#) (21)
  - [Kansas City Chiefs](#) (33)
  - [Miami Dolphins](#) (42)
  - [Minnesota Vikings](#) (33)
  - [New England Patriots](#) (36)
  - [New Orleans Saints](#) (27)
  - [New York Giants](#) (41)
  - [New York Jets](#) (27)
  - [Oakland Raiders](#) (51)
  - [Philadelphia Eagles](#) (30)
  - [Pittsburgh Steelers](#) (52)
  - [San Diego Chargers](#) (32)
  - [San Francisco 49ers](#) (29)
  - [Seattle Seahawks](#) (29)
  - [St. Louis Rams](#) (30)
  - [Tampa Bay Buccaneers](#) (35)
  - [Tennessee Titans](#) (24)
  - [Washington Redskins](#) (56)
- 
- [NFL Europa@](#) (19)

# Search Engines

- Most often used to fill an information need
- Pages are collected automatically by web crawlers
- Users enter search terms into text box get back a SERP (search engine result page)
- Queries are generally modified and resubmitted to the SE if the desired results are not found on the first few pages of results
- Types of search engines:
  - Web search engines (Google, Yahoo, Bing)
  - Metasearch engines – includes Deep Web ([Dogpile](#), [WebCrawler](#))
  - Specialized (or focused) search engines ([Google Scholar](#), [MapQuest](#))

# Components of a Search Engine

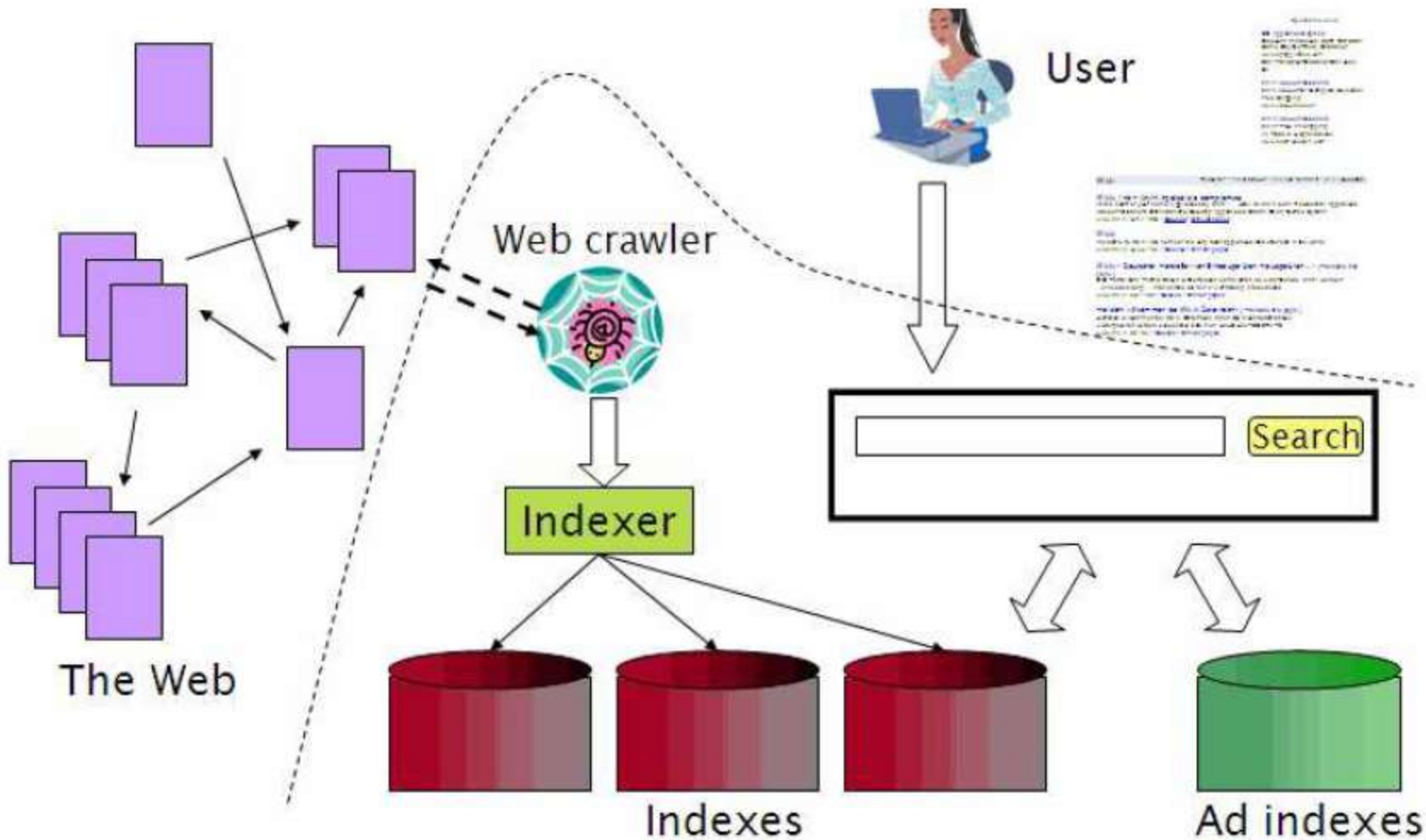
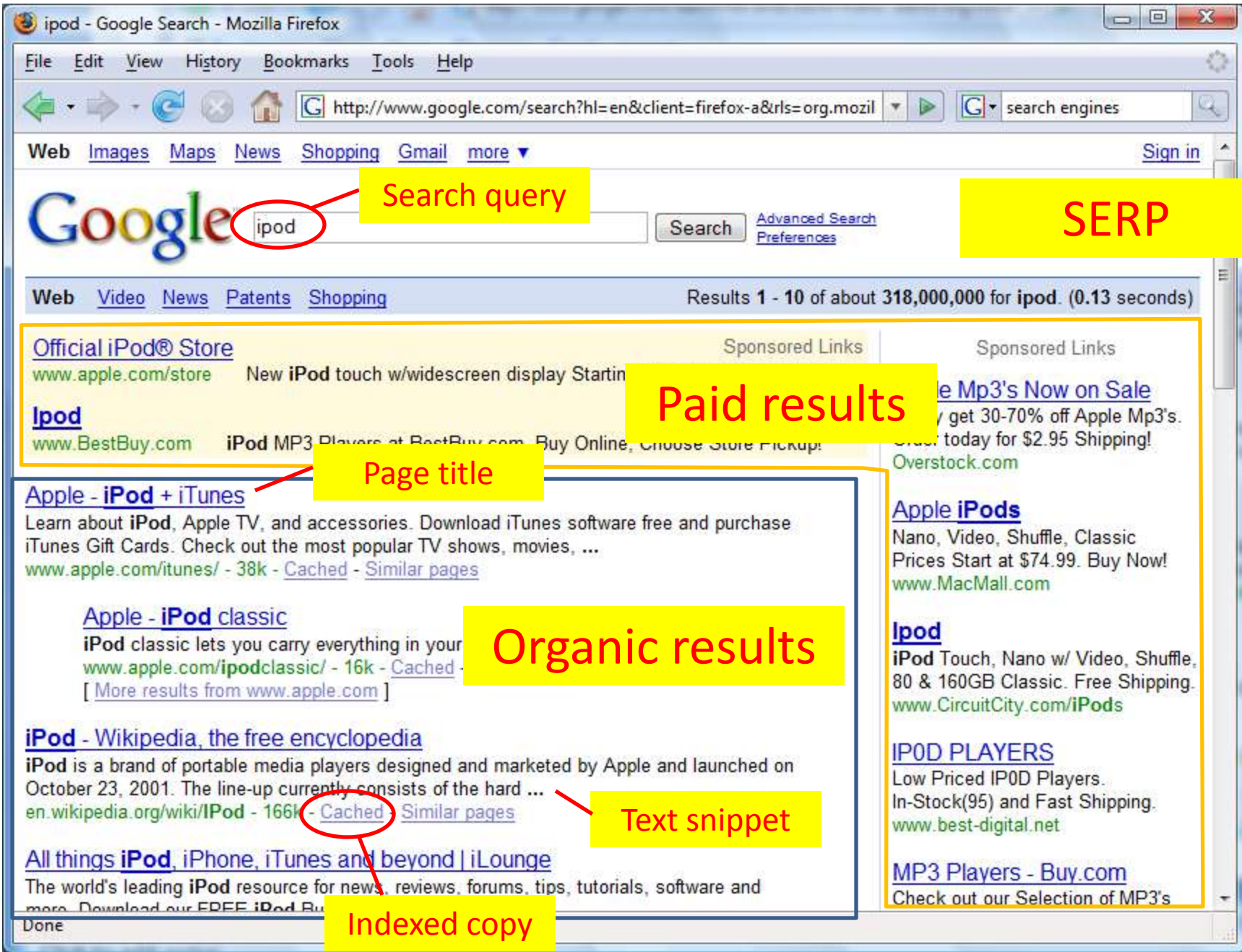


Figure from [Introduction to Information Retrieval](#) by Manning et al., Ch 19.



Search query

SERP

Paid results

Page title

Organic results

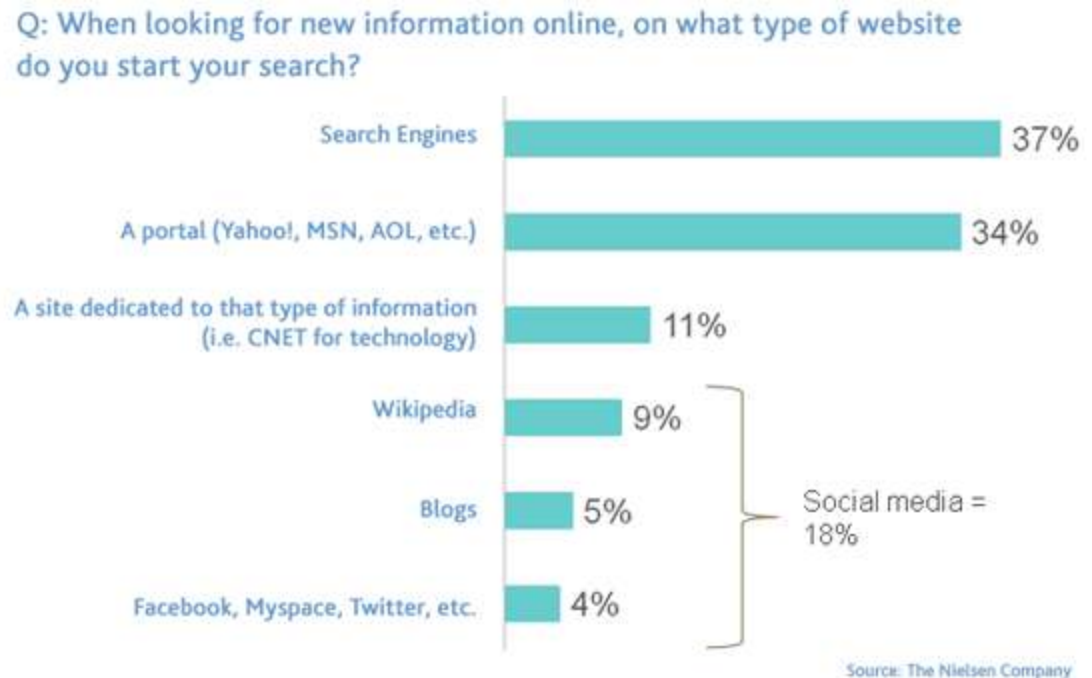
Text snippet

Indexed copy

# Social Media

- Increasingly being used to find info
- Limits influence of results to trusted group

Nielsen study (August 2009)



# Search Queries

- Search engines store every query, but companies usually don't share with the public because of privacy issues
  - [2006 AOL search log incident](#)
  - [2006 govt subpoenas Google incident](#)
- Often short: 2.4 words on average<sup>1</sup> but getting longer<sup>2</sup>
- Most users do not use advanced features<sup>1</sup>
- Distribution of terms is long-tailed<sup>3</sup>

<sup>1</sup> Spink et al., [Searching the web: The public and their queries](#), 2001

<sup>2</sup> <http://searchengineland.com/search-queries-getting-longer-16676>

<sup>3</sup> Lempel & Moran, WWW 2003

# Search Queries

- 10-15% contain misspellings<sup>1</sup>
- Often repeated: Yahoo study<sup>2</sup> showed 1/3 of all queries are repeat queries, and 87% of users click on same result

<sup>1</sup> Cucerzan & Brill, 2004

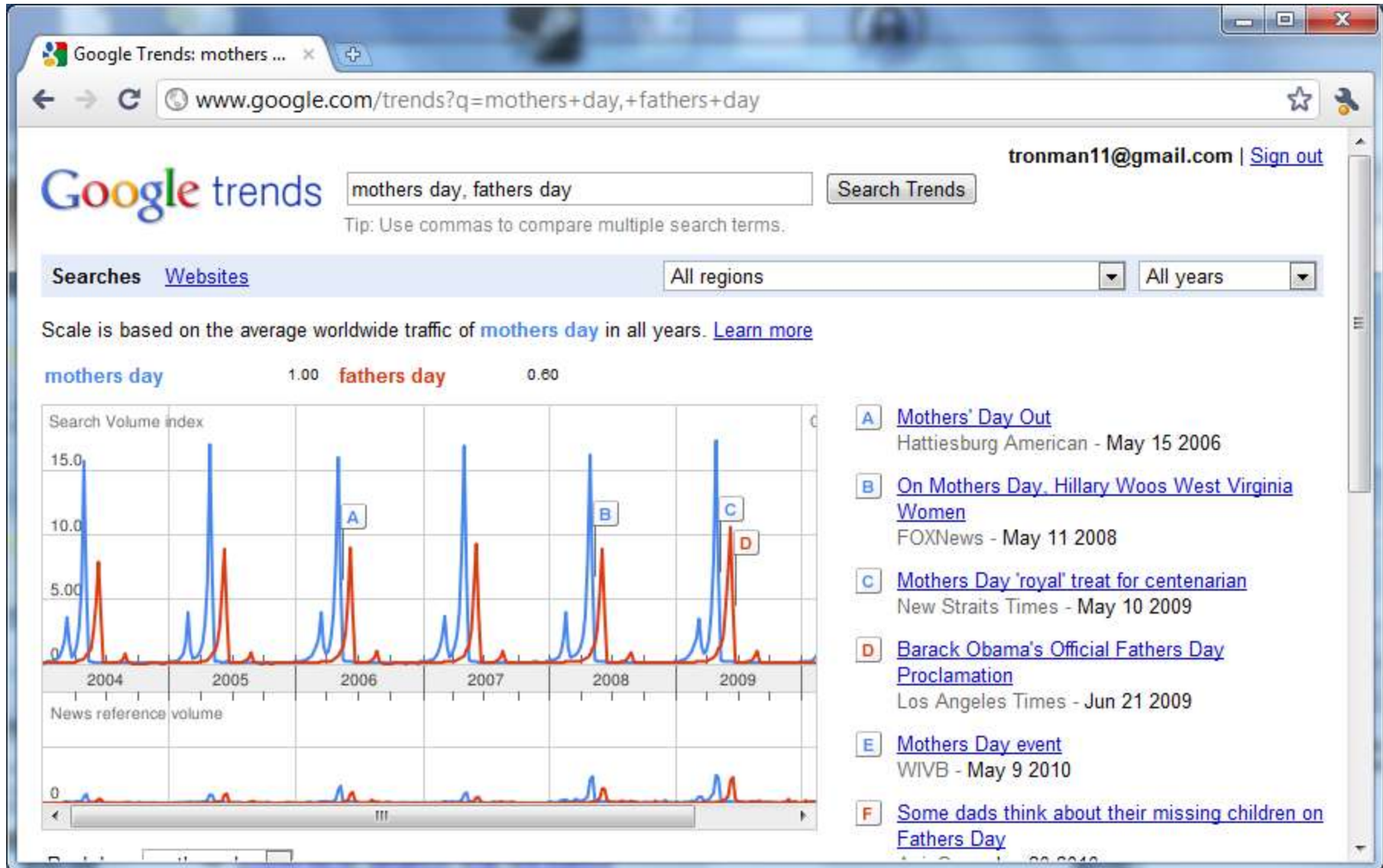
<sup>2</sup> Teevan et al., History Repeats Itself: Repeat Queries in Yahoo's Logs, *Proc SIGIR 2006*

# Query Types

- Informational
  - Intent is to acquire info about a topic
  - Examples: safe vehicles, albert einstein
- Navigational
  - Intent is to find a particular site
  - Examples: facebook, google
- Transactional
  - Intent is to perform an activity mediated by a website
  - Examples: children books, cheap flights

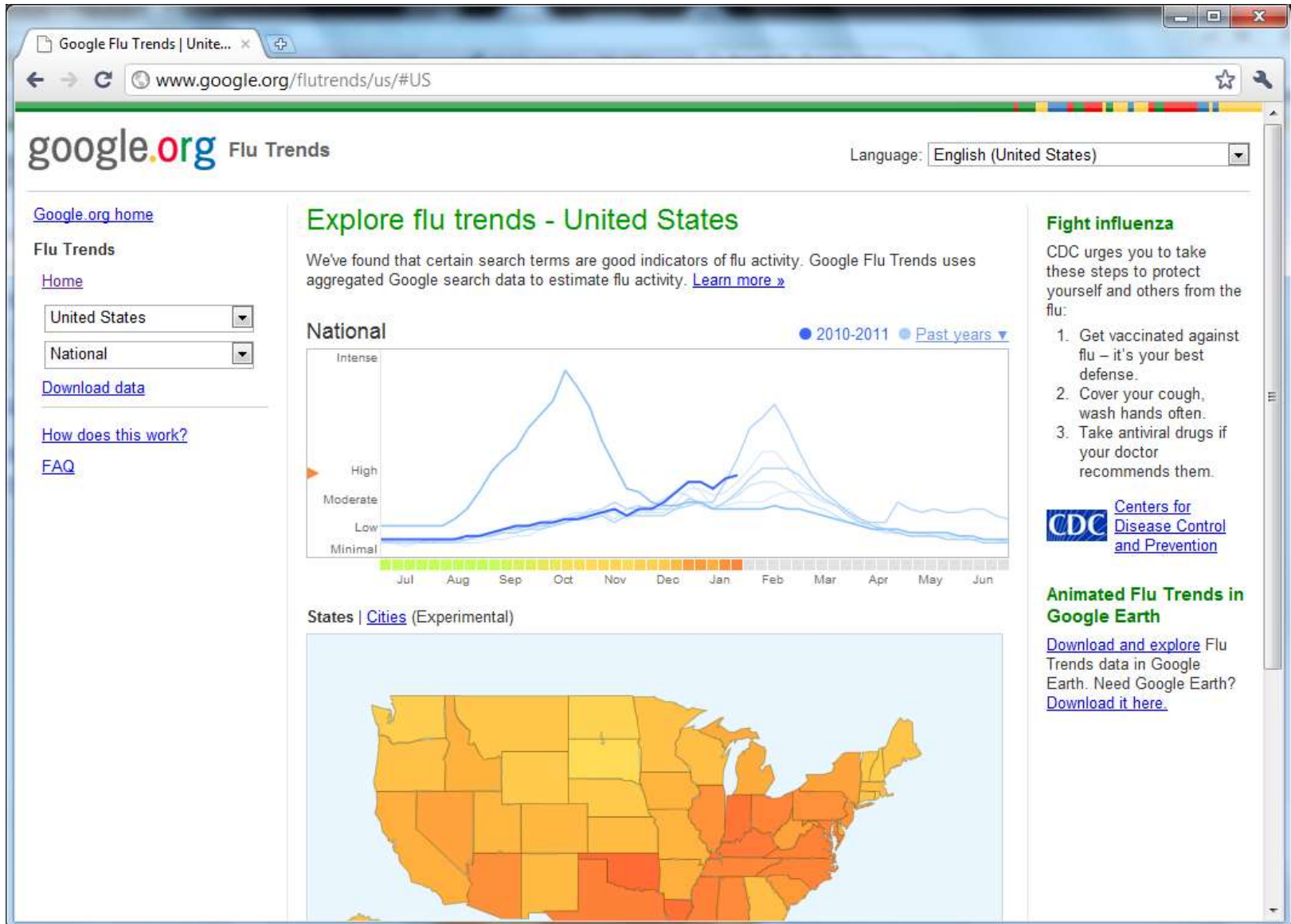
# Google Trends

<http://www.google.com/trends>



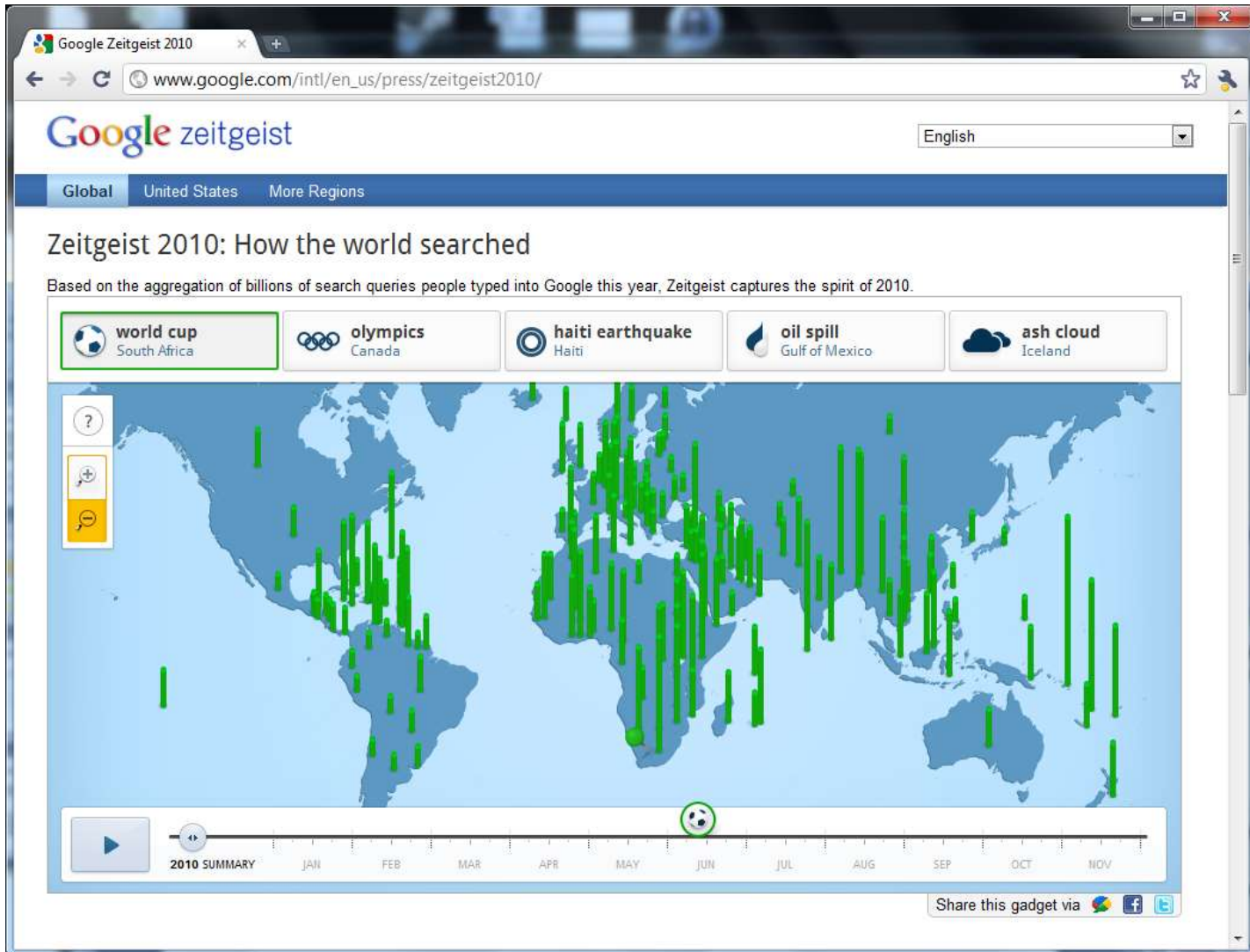
# Google Flu Trends

<http://www.google.org/flutrends/>



# Google Zeitgeist

[http://www.google.com/intl/en\\_us/press/zeitgeist2010/](http://www.google.com/intl/en_us/press/zeitgeist2010/)



# My 2010 Top Queries, Sites, & Clicks

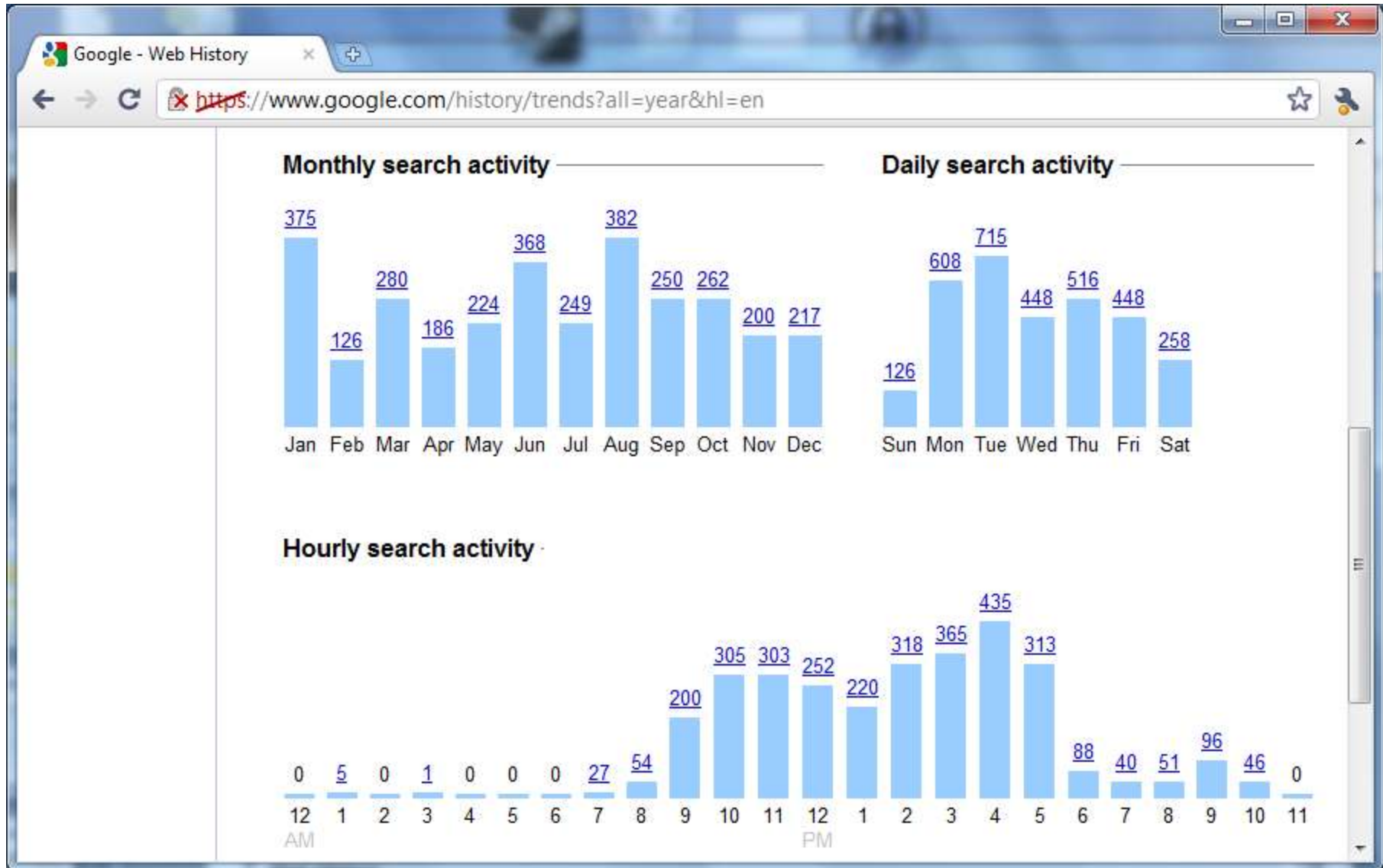
The screenshot shows a Google Web History page in a browser window. The address bar displays the URL <https://www.google.com/history/trends?all=year&hl=en>. The page features the Google logo and a search bar with "Search History" and "Search the Web" buttons. A notification states, "Your web history is limited to searches. [Expand your web history.](#)"

**Web History**

Show trends for: [Last 7 days](#) | [Last 30 days](#) | [Last year](#) | [All time](#)

Web History	Top queries	Top sites	Top clicks
<a href="#">Web</a>	1. <a href="#">weather searcy. ar</a>	1. <a href="#">en.wikipedia.org</a>	1. <a href="#">Common Tasks and How to</a>
<a href="#">Images</a>	2. <a href="#">android linkify</a>	2. <a href="#">developer.android.com</a>	2. <a href="#">Searcy, Arkansas (72143) C</a>
<a href="#">News</a>	3. <a href="#">searcy. ar weather</a>	3. <a href="#">stackoverflow.com</a>	3. <a href="#">VB.NET and C# Comparisor</a>
<a href="#">Products</a>	4. <a href="#">memento browser</a>	4. <a href="#">msdn.microsoft.com</a>	4. <a href="#">Android SDK   Android Deve</a>
<a href="#">Sponsored Links</a>	5. <a href="#">web science</a>	5. <a href="#">www.amazon.com</a>	5. <a href="#">memento-browser - Project I</a>
<a href="#">Video</a>	6. <a href="#">searcy cinema</a>	6. <a href="#">java.sun.com</a>	6. <a href="#">WebView   Android Develop</a>
<a href="#">Maps</a>	7. <a href="#">scrabble two letter words</a>	7. <a href="#">www.harding.edu</a>	7. <a href="#">Welcome to the Searcy Cine</a>
<a href="#">Blogs</a>	8. <a href="#">frank mccown</a>	8. <a href="#">code.google.com</a>	8. <a href="#">http://developer.android.com</a>
<a href="#">Books</a>	9. <a href="#">android</a>	9. <a href="#">video.google.com</a>	9. <a href="#">International Baseball Feder</a>
Pause	10. <a href="#">vb.net c#</a>	10. <a href="#">groups.google.com</a>	10. <a href="#">Yahoo! Sports Fantasy Colle</a>
Remove items			
<b>Trends</b>			
<a href="#">Bookmarks</a>			

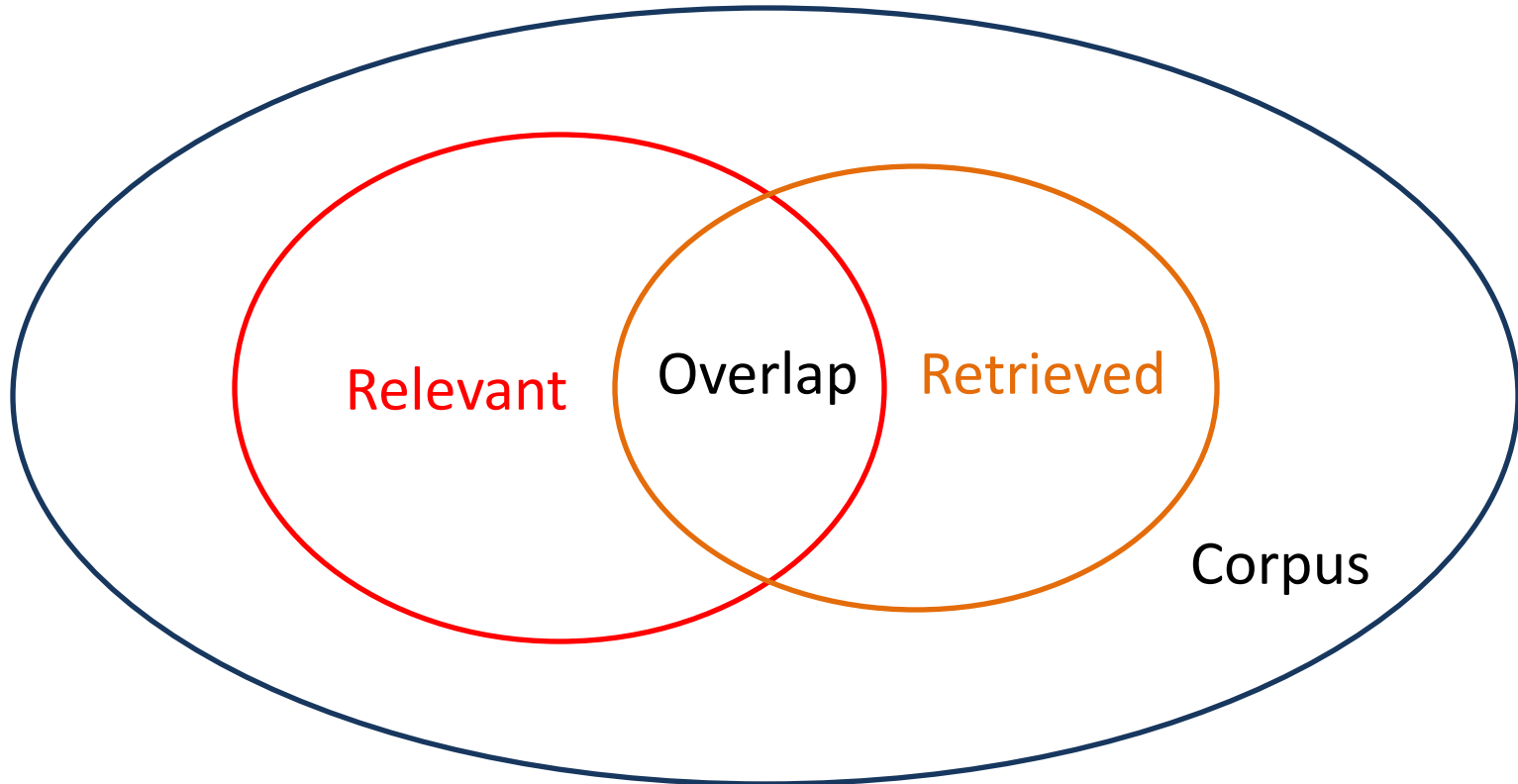
# My 2010 Monthly, Daily, & Hourly Search Activity



# Relevance

- Search engines are useful if they return *relevant* results
- Relevance is hard to pin down because it depends on user's intent & context which is often not known
- Relevance can be increased by personalizing search results
  - What is the user's location?
  - What queries has this user made before?
  - How does the user's searching behavior compare to others?
- Two popular metrics are used to evaluate whether the results returned by a search engine are relevant: **precision** and **recall**

# Precision and Recall



Precision =  $\text{Overlap} / \text{Retrieved}$

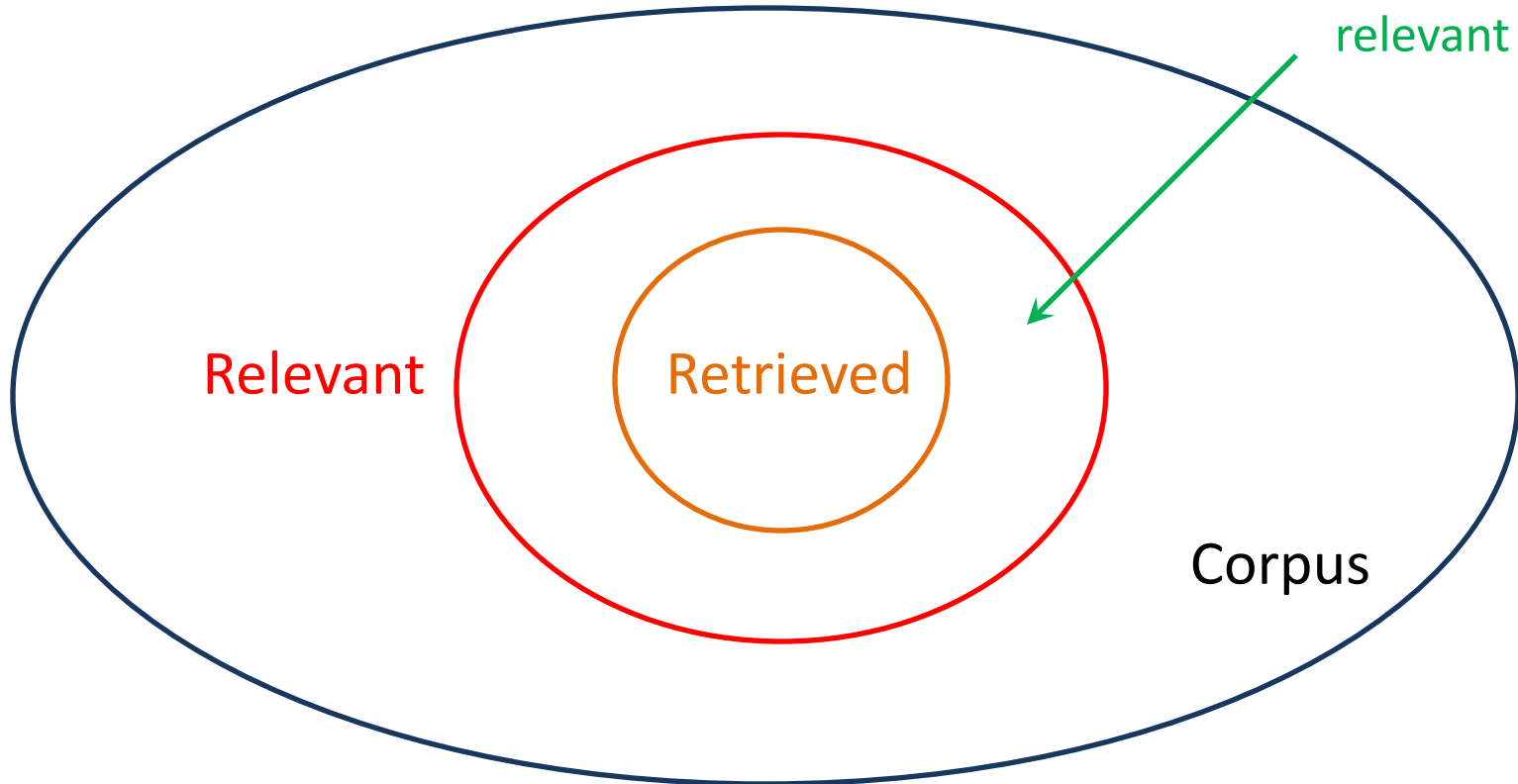
Recall =  $\text{Overlap} / \text{Relevant}$

# Example

- Given a corpus of 100 documents
- 20 are about football
- Search for *football* results in 50 returned documents, 10 are about football
- Precision = Overlap / Retrieved =  $10/50 = .2$
- Recall = Overlap / Relevant =  $10/20 = .5$
- Note: Usually precision and recall are at odds

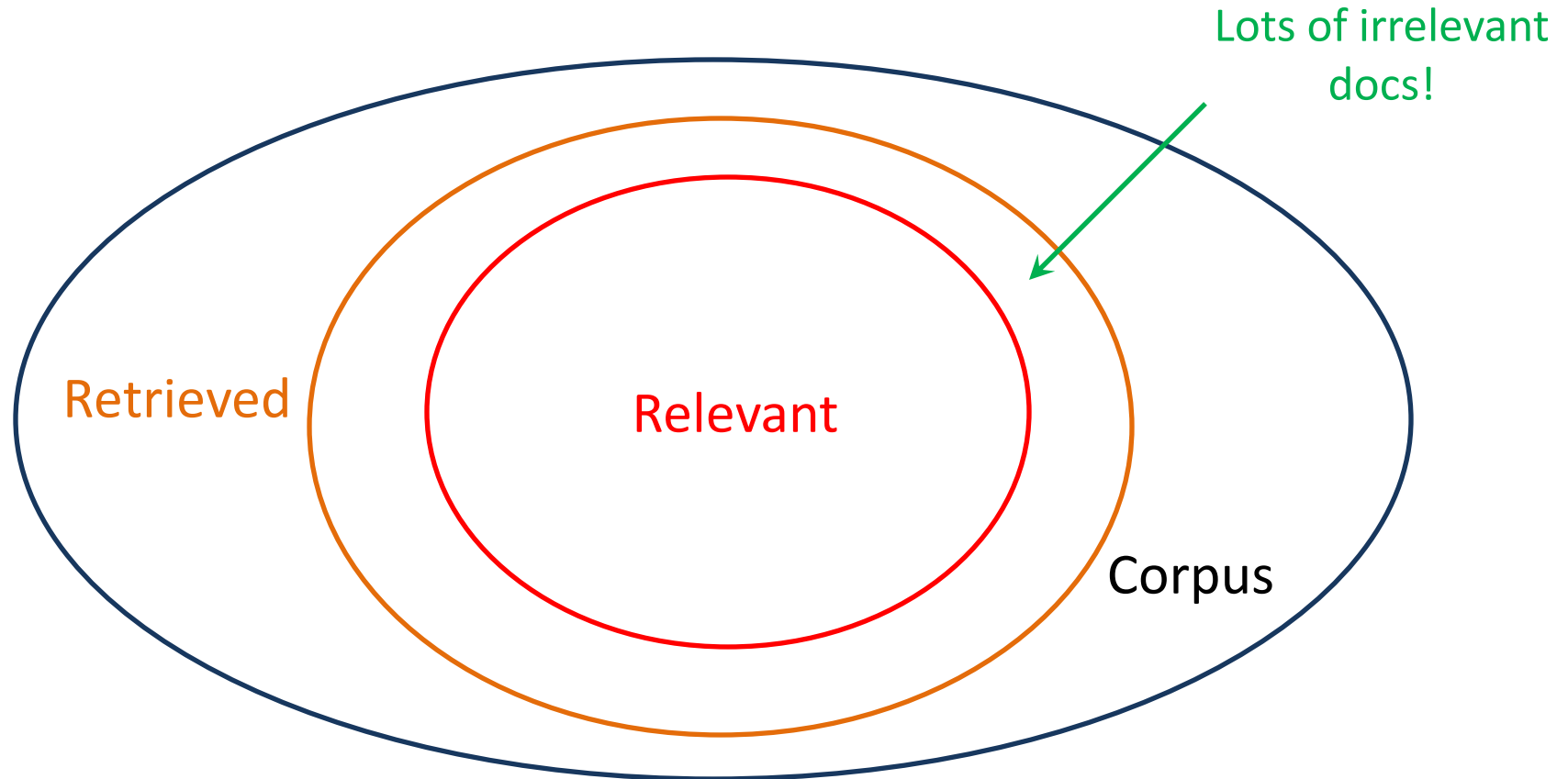
# High Precision, Low Recall

Missing a lot of relevant docs!



$$\text{Precision} = \text{Overlap} / \text{Retrieved}$$
$$\text{Recall} = \text{Overlap} / \text{Relevant}$$

# Low Precision, High Recall



$$\text{Precision} = \text{Overlap} / \text{Retrieved}$$
$$\text{Recall} = \text{Overlap} / \text{Relevant}$$

# Evaluating Search Engines

- We don't usually know how many documents on the entire Web are about a particular topic, so computing recall for a web search engine is not possible
- Most people view only the first page or two of search results, so the top  $N$  results are most important where  $N$  is 10 or 20
- $P@N$  is the precision of the top  $N$  results

# Comparing Search Engine with Digital Library

- McCown et al.<sup>1</sup> compared the P@10 of Google and the National Science Digital Library (NSDL)
- School teachers evaluated relevance of search results in regards to Virginia's Standards of Learning
- Overall, Google's precision was found to be 38.2% compared to NSDL's 17.1%

<sup>1</sup>McCown et al., Evaluation of the NSDL and Google search engines for obtaining pedagogical resources, *Proc ECDL 2005*

# F-score

- F-score combines precision and recall into single metric
- F-score is harmonic mean of precision and recall

$$F = 2 \cdot \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

- Highest = 1, Lowest = 0

# Quality of Ranking

- Issue search query to SE and have humans rank the first  $N$  results in order of relevance
- Compare human ranking with SE ranking (e.g., Spearman rank-order correlation coefficient)
- Other ranking methods can be used
  - Discounted cumulative gain (DCG)<sup>2</sup> which gives higher ranked results more weight than lower ranked results
  - M measure<sup>3</sup> similar function as DCG which gives sliding scale of importance based on ranking

<sup>1</sup>Vaughan, New measurements for search engine evaluation proposed and tested, *Info Proc & Mang* (2004)

<sup>2</sup>Järvelin & Kekäläinen, Cumulated gain-based evaluation of IR techniques, *TOIS* (2004)

<sup>3</sup>Bar-Ilan et al., Methods for comparing rankings of search engine results, *Computer Networks* (2006)