Other characteristics of the Web

- Significant duplication
  - Syntactic – 30%-40% (near) duplicates [Brod97, Shiv99b, etc.]
  - Semantic – ???
- High linkage
  - More than 8 links/page in the average
- Complex graph topology
  - Not a small world; bow-tie structure [Brod00]
- Spam
  - Billions of pages
Challenge: “The need behind the query”

- Semantic analysis
  - Query language determination
    - Auto filtering
    - Different ranking (if query in Japanese do not return English)
  - Hard & soft (partial) matches
    - Personalities (triggered on names)
    - Cities (travel info, maps)
    - Medical info (triggered on names and/or results)
    - Stock quotes, news (triggered on stock symbol)
    - Company info
    - Etc.
  - Natural Language reformulation
  - Integration of Search and Text Analysis

Answering “the need behind the query”: Context

- Context determination
  - spatial (user location/target location)
  - query stream (previous queries)
  - personal (user profile)
  - explicit (user choice of a vertical search, )
  - implicit (use Google from France, use google.fr)

- Context use
  - Result restriction
    - Kill inappropriate results
  - Ranking modulation
    - Use a “rough” generic ranking, but personalize later
The spatial context: Geo-search

- Two aspects
  - Geo-coding -- encode geographic coordinates to make search effective
  - Geo-parsing -- the process of identifying geographic context.
- Geo-coding
  - Geometrical hierarchy (squares)
  - Natural hierarchy (country, state, county, city, zip-codes, etc)
- Geo-parsing
  - Pages (infer from phone nos, zip, etc). About 10% can be parsed.
  - Queries (use dictionary of place names)
  - Users
    - Explicit (tell me your location -- used by NL, registration, from ISP)
    - From IP data
  - Mobile phones
    - Many sources of highly accurate location

Answering “the need behind the query”: Context

- Context determination
  - spatial (user location/target location)
  - query stream (previous queries)
  - personal (user profile)
  - explicit (user choice of a vertical search, )
  - implicit (use Google from France, use google.fr)

- Context use
  - Result restriction
    - Kill inappropriate results
  - Ranking modulation
    - Use a “rough” generic ranking, but personalize later
Context transfer

Google search for "brass boot"

Results: 1-10 of about 122,000

Sponsored Links:
- Brass Boots at Zappos
- Shoe Boot Shoes at Shoebuy.com! Free Shipping and No Sales Tax!

No transfer

Google search for "shoe buy"

Sponsored Links:
- Welcome to the Net's largest footwear retailer!
Context transfer

Transfer from search results

1. An analysis of data characteristics that affect \textit{naive Bayes performance}\footnote{\textit{An analysis of data characteristics that affect naive Bayes performance} ...} \footnote{www.research.ibm.com/PAI/cmi01.pdf - 357k - View as HTML - More from this site - Save - Block}

2. Improving the \textit{Performance} of \textit{Naive Bayes} for Text Classification\footnote{Improving the \textit{Performance} of \textit{Naive Bayes} for Text Classification. Yirong Shen and Jing Jiang. CS224N Spring 2003. Abstract. We seek to improve the \textit{performance} of the \textit{naive Bayes} classifier \footnote{www-nlp.stanford.edu/courses/cs224n/2003/fp/yirong99/report.pdf - 135k - View as HTML - More from this site - Save - Block}}\footnote{www-nlp.stanford.edu/courses/cs224n/2003/fp/yirong99/report.pdf - 135k - View as HTML - More from this site - Save - Block}
Answering “the need behind the query”: Context

• Context determination
  – spatial (user location/target location)
  – query stream (previous queries)
  – personal (user profile)
  – explicit (user choice of a vertical search, e.g., Amazon or eBay)
  – implicit (use Google from France, use google.fr)

• Context use
  – Result restriction
    • Kill inappropriate results
  – Ranking modulation
    • Use a “rough” generic ranking, but personalize later
Result Restriction

- Geographic restrictions
  - Holocaust denial in Germany
  - Imagery that may be illegal in some jurisdictions, accepted in others
- Age restrictions
  - COPPA