Applications (1 of 2): Information Retrieval

Kenneth Church
Kenneth.Church@jhu.edu
Pattern Recognition Problems in Computational Linguistics

• Information Retrieval:
  – Is this doc more like relevant docs or irrelevant docs?

• Author Identification:
  – Is this doc more like author A’s docs or author B’s docs?

• Word Sense Disambiguation
  – Is the context of this use of bank
    • more like sense 1’s contexts
    • or like sense 2’s contexts?

• Machine Translation
  – Is the context of this use of drug more like those that were translated as drogue
  – or those that were translated as medicament?
Applications of Naïve Bayes

Word Sense Disambiguation (WSD)

\[
score(context) = \prod_{\text{word in context}} \frac{Pr(\text{word}|\text{sense}_1)}{Pr(\text{word}|\text{sense}_2)}
\]

Author Identification

\[
score(doc) = \prod_{\text{word in doc}} \frac{Pr(\text{word}|\text{author}_1)}{Pr(\text{word}|\text{author}_2)}
\]

Information Retrieval (IR)

\[
score(doc) = \prod_{\text{word in doc}} \frac{Pr(\text{word}|\text{relevant})}{Pr(\text{word}|\text{irrelevant})}
\]

Sentiment Analysis

\[
score(doc) = \prod_{\text{word in doc}} \frac{Pr(\text{word}|\text{positive review})}{Pr(\text{word}|\text{negative review})}
\]
Classical Information Retrieval (IR)

• Boolean Combinations of Keywords
  – Dominated the Market (before the web)
  – Popular with Intermediaries (Librarians)

• Rank Retrieval (Google)
  – Sort a collection of documents
    • (e.g., scientific papers, abstracts, paragraphs)
    • by how much they “match” a query
  – The query can be a (short) sequence of keywords
    • or arbitrary text (e.g., one of the documents)
Motivation for Information Retrieval
(circa 1990, about 5 years before web)

• Text is available like never before
• Currently, $N \approx 100$ million words
  – and projections run as high as $10^{15}$ bytes by 2000!
• What can we do with it all?
  – It is better to do something simple,
  – than nothing at all.
• IR vs. Natural Language Understanding
  – Revival of 1950-style empiricism
How Large is Very Large?

From a Keynote to EMNLP Conference, formally Workshop on Very Large Corpora

<table>
<thead>
<tr>
<th>Year</th>
<th>Source</th>
<th>Size (words)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1788</td>
<td>Federalist Papers</td>
<td>1/5 million</td>
<td></td>
</tr>
<tr>
<td>1982</td>
<td>Brown Corpus</td>
<td>1 million</td>
<td></td>
</tr>
<tr>
<td>1987</td>
<td>Birmingham Corpus</td>
<td>20 million</td>
<td></td>
</tr>
<tr>
<td>1988-</td>
<td>Associate Press (AP)</td>
<td>50 million</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(per year)</td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>MUC, TREC, Tipster</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rising Tide of Data Lifts All Boats

If you have a lot of data, then you don’t need a lot of methodology

• 1985: “There is no data like more data”
  – Fighting words uttered by radical fringe elements (Mercer at Arden House)

• 1993 Workshop on Very Large Corpora
  – Perfect timing: Just before the web
  – Couldn’t help but succeed
  – Fate

• 1995: The Web changes everything

• All you need is data (magic sauce)
  – No linguistics
  – No artificial intelligence (representation)
  – No machine learning
  – No statistics
  – No error analysis
“It never pays to think until you’ve run out of data” – Eric Brill

Moore’s Law Constant:
Data Collection Rates → Improvement Rates

No consistently best learner

More data is better data!

Fire everybody and spend the money on data

Dec 2, 2009
Benefit of Data

LIMSI: Lamel (2002) – Broadcast News

Supervised: transcripts
Lightly supervised: closed captions
The rising tide of data will lift all boats!

TREC Question Answering & Google:

*What is the highest point on Earth?*

The following words are very common and were not included in the search results:

**Altitude of the Highest Point on Earth**


**The Sun and its Highest Point**

... If the Earth had a perfectly circular orbit, the Analemma ... perfectly symmetrical Fig 8 with the cross-over point ... One way to determine when the Sun is highest ... imagine.gsfc.nasa.gov/docs/ask_astro/answers/970714.html - 22k - Cached - Similar
The rising tide of data will lift all boats!

Acquiring Lexical Resources from Data:
Dictionaries, Ontologies, WordNets, Language Models, etc.

http://labs1.google.com/sets

England
France
Rising Tide of Data Lifts All Boats
If you have a lot of data, then you don’t need a lot of methodology

• More data \(\Rightarrow\) better results
  – TREC Question Answering
    • Remarkable performance: Google and not much else
      – Norvig (ACL-02)
      – AskMSR (SIGIR-02)
  – Lexical Acquisition
    • Google Sets
      – We tried similar things
        » but with tiny corpora
        » which we called large

http://labs1.google.com/sets

<table>
<thead>
<tr>
<th>Cat</th>
<th>Dog</th>
<th>more</th>
<th>England</th>
<th>France</th>
<th>Japan</th>
<th>China</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horse</td>
<td>is</td>
<td>Germany</td>
<td></td>
<td></td>
<td>India</td>
<td></td>
</tr>
<tr>
<td>Fish</td>
<td>rm</td>
<td>Italy</td>
<td></td>
<td></td>
<td>Indonesia</td>
<td></td>
</tr>
<tr>
<td>Bird</td>
<td>mv</td>
<td>Ireland</td>
<td>Spain</td>
<td>Korea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rabbit</td>
<td>cd</td>
<td>Scotland</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cattle</td>
<td>cp</td>
<td>Belgium</td>
<td>Thailand</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rat</td>
<td>mmdir</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Livestock</td>
<td>man</td>
<td>Canada</td>
<td>Singapore</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td>tail</td>
<td>Austria</td>
<td>Australia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>pwd</td>
<td></td>
<td>Bangladesh</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Applications

- What good is word sense disambiguation (WSD)?
  - Information Retrieval (IR)
    - Salton: Tried hard to find ways to use NLP to help IR
      - but failed to find much (if anything)
    - Croft: WSD doesn’t help because IR is already using those methods
    - Sanderson (next two slides)
  - Machine Translation (MT)
    - Original motivation for much of the work on WSD
    - But IR arguments may apply just as well to MT

What good is POS tagging? Parsing? NLP? Speech?

- Commercial Applications of Natural Language Processing, CACM 1995
  - $100M opportunity (worthy of government/industry’s attention)
    1. Search (Lexis-Nexis)
    2. Word Processing (Microsoft)
  - Warning: premature commercialization is risky

Don’t worry; Be happy
Sanderson (SIGIR-94)

http://dis.shef.ac.uk/mark/cv/publications/papers/my_papers/SIGIR94.pdf

• Could WSD help IR?
• Answer: no
  – Introducing ambiguity by pseudo-words doesn’t hurt (much)

Short queries matter most, but hardest for WSD

Not much?

Dec 2,
Soft WSD?

- Resolving ambiguity badly is worse than not resolving at all
  - 75% accurate WSD degrades performance
  - 90% accurate WSD: breakeven point

Query Length (Words)
IR Models

- **Keywords (and Boolean combinations thereof)**
- **Vector-Space “Model”** (Salton, chap 10.1)
  - Represent the query and the documents as V-dimensional vectors
  - Sort vectors by $\cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}$
- **Probabilistic Retrieval Model**
  - (Salton, chap 10.3)
  - Sort documents by $\text{score}(d) = \prod_{w \in d} \frac{\Pr(w \mid \text{rel})}{\Pr(w \mid \text{rel})}$

Dec 2, 2009
Information Retrieval and Web Search

Alternative IR models

Instructor: Rada Mihalcea

Some of the slides were adopted from a course taught at Cornell University by William Y. Arms
Latent Semantic Indexing

Objective

Replace indexes that use *sets of index terms* by indexes that use *concepts*.

Approach

Map the term vector space into a lower dimensional space, using singular value decomposition.

Each dimension in the new space corresponds to a latent concept in the original data.
Deficiencies with Conventional Automatic Indexing

**Synonymy:** Various words and phrases refer to the same concept (lowers recall).

**Polysemy:** Individual words have more than one meaning (lowers precision)

**Independence:** No significance is given to two terms that frequently appear together

Latent semantic indexing addresses the first of these (synonymy), and the third (dependence)
Bellcore’s Example

http://en.wikipedia.org/wiki/Latent_semantic_analysis

c1 Human machine *interface* for Lab ABC *computer* applications

c2 A *survey* of *user* opinion of computer *system* *response time*

c3 The EPS *user interface* management *system*

c4 *System* and *human system* engineering testing of EPS

c5 Relation of *user*-perceived *response time* to error measurement

m1 The generation of random, binary, unordered *trees*

m2 The intersection *graph* of paths in *trees*

m3 *Graph minors* IV: Widths of *trees* and well-quasi-ordering

m4 *Graph minors*: A *survey*
Term by Document Matrix

<table>
<thead>
<tr>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| human | interface | computer | user | system | response | time | EPS | survey | trees | graph | minors |

Dec 2, 2009 21
Query Expansion

Query:
Find documents relevant to *human computer interaction*

**Simple Term Matching:**

<table>
<thead>
<tr>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Matches c1, c2, and c4
Misses c3 and c5

Dec 2, 2009
• It is generally assumed that terms are independent. That is, $\rho_{i,j} = 0$ when $i \neq j$.

• In practice, this assumption is often problematic.

• Positive correlations arise when two words share similar distributions:
  • synonymous terms: computer, machine
  • morphological variants: computer, computers
  • spelling variants: IBM, I.B.M.
  • upper and lower case: computer, Computer
  • strong collocations: computer scientist

• Negative correlations arise when two words have complementary distributions.
Correlations: Too Large to Ignore

<table>
<thead>
<tr>
<th>human</th>
<th>interface</th>
<th>computer</th>
<th>user</th>
<th>system</th>
<th>response</th>
<th>time</th>
<th>EPS</th>
<th>survey</th>
<th>trees</th>
<th>graph</th>
<th>minors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.4</td>
<td>0.4</td>
<td>-0.4</td>
<td>0.4</td>
<td>-0.3</td>
<td>-0.3</td>
<td>0.4</td>
<td>-0.3</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>0.4</td>
<td>1.0</td>
<td>0.4</td>
<td>0.2</td>
<td>0.04</td>
<td>-0.3</td>
<td>-0.3</td>
<td>0.4</td>
<td>-0.3</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
<td>1.0</td>
<td>0.2</td>
<td>0.04</td>
<td>0.4</td>
<td>0.4</td>
<td>-0.3</td>
<td>0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>-0.4</td>
<td>0.2</td>
<td>0.2</td>
<td>1.0</td>
<td>0.2</td>
<td>0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>0.2</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.4</td>
</tr>
<tr>
<td>0.4</td>
<td>0.04</td>
<td>0.04</td>
<td>0.2</td>
<td>1.0</td>
<td>0.04</td>
<td>0.04</td>
<td>0.8</td>
<td>0.04</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.3</td>
</tr>
<tr>
<td>-0.3</td>
<td>-0.3</td>
<td>0.4</td>
<td>0.8</td>
<td>0.04</td>
<td>1.0</td>
<td>1.0</td>
<td>-0.3</td>
<td>0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>-0.3</td>
<td>-0.3</td>
<td>0.4</td>
<td>0.8</td>
<td>0.04</td>
<td>1.0</td>
<td>1.0</td>
<td>-0.3</td>
<td>0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
<td>-0.3</td>
<td>0.2</td>
<td>0.8</td>
<td>-0.3</td>
<td>-0.3</td>
<td>1.0</td>
<td>-0.3</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>-0.3</td>
<td>-0.3</td>
<td>0.4</td>
<td>0.2</td>
<td>0.04</td>
<td>0.4</td>
<td>0.4</td>
<td>-0.3</td>
<td>1.0</td>
<td>-0.4</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>1.0</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>0.2</td>
<td>0.5</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>-0.3</td>
<td>-0.3</td>
<td>-0.3</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.3</td>
<td>-0.3</td>
<td>-0.3</td>
<td>0.4</td>
<td>0.2</td>
<td>0.8</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Dec 2, 2009
• One can compute the correlation for each pair of terms, and adjust the cos calculation appropriately.

• Unfortunately, this is generally not practical since there are $V^2$ correlations to consider.

• For any particular pair of documents, one can look at the terms that contribute the most and adjust for their correlations. (I don’t think this has been tried.)

• It is also quite common to merge terms that have "similar" distributions, either for linguistic or statistical reasons.

• For example, it is common to treat morphologically related words (e.g., *computer* and *computers*) as a single term.

• Treating two words ($i$ and $j$) as the same term is equivalent to assuming $\rho_{i,j} \approx 1$.  

Correcting for Large Correlations
Thesaurus

• Merge terms that cluster together
  • human/ interface/ computer
  • user/ response/ time
  • system/ EPS
  • graph/ minors
### Term by Doc Matrix: Before & After Thesaurus

#### Before

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Terms:
- human
- interface
- computer
- user
- system
- response
- time
- EPS
- survey
- trees
- graph
- minors

#### After

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Terms:
- com/hum/inter
- user/res/time
- system/EPS
- survey
- trees
- graph/minors

Dec 2, 2009
Singular Value Decomposition (SVD)

\[ X = U D V^T \]

- \( m \) is the rank of \( X < \min(t, d) \)
- \( D \) is diagonal
  - \( D^2 \) are eigenvalues (sorted in descending order)
- \( U U^T = I \) and \( V V^T = I \)
  - Columns of \( U \) are eigenvectors of \( X X^T \)
  - Columns of \( V \) are eigenvectors of \( X^T X \)
Dimensionality Reduction

\[ \hat{X} = U^T \]

\( k \) is the number of latent concepts

(typically 300 ~ 500)
SVD

\[ B B^T = U D^2 U^T \]

\[ B^T B = V D^2 V^T \]
The term vector space

The space has as many dimensions as there are terms in the word list.
Latent concept vector space

- term
- document
- query
- cosine $> 0.9$
Recombination after Dimensionality Reduction

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>0.16</td>
<td>0.40</td>
<td>0.38</td>
<td>0.47</td>
<td>0.18</td>
<td>-0.05</td>
<td>-0.12</td>
<td>-0.16</td>
<td>-0.09</td>
</tr>
<tr>
<td>interface</td>
<td>0.14</td>
<td>0.37</td>
<td>0.33</td>
<td>0.40</td>
<td>0.16</td>
<td>-0.03</td>
<td>-0.07</td>
<td>-0.10</td>
<td>-0.04</td>
</tr>
<tr>
<td>computer</td>
<td>0.15</td>
<td>0.51</td>
<td>0.36</td>
<td>0.41</td>
<td>0.24</td>
<td>0.02</td>
<td>0.06</td>
<td>0.09</td>
<td>0.12</td>
</tr>
<tr>
<td>user</td>
<td>0.26</td>
<td>0.84</td>
<td>0.61</td>
<td>0.70</td>
<td>0.39</td>
<td>0.03</td>
<td>0.08</td>
<td>0.12</td>
<td>0.19</td>
</tr>
<tr>
<td>system</td>
<td>0.45</td>
<td>1.23</td>
<td>1.05</td>
<td>1.27</td>
<td>0.56</td>
<td>-0.07</td>
<td>-0.15</td>
<td>-0.21</td>
<td>-0.05</td>
</tr>
<tr>
<td>response</td>
<td>0.16</td>
<td>0.58</td>
<td>0.38</td>
<td>0.42</td>
<td>0.28</td>
<td>0.06</td>
<td>0.13</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td>time</td>
<td>0.16</td>
<td>0.58</td>
<td>0.38</td>
<td>0.42</td>
<td>0.28</td>
<td>0.06</td>
<td>0.13</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td>EPS</td>
<td>0.22</td>
<td>0.55</td>
<td>0.51</td>
<td>0.63</td>
<td>0.24</td>
<td>-0.07</td>
<td>-0.14</td>
<td>-0.20</td>
<td>-0.11</td>
</tr>
<tr>
<td>survey</td>
<td>0.10</td>
<td>0.53</td>
<td>0.23</td>
<td>0.21</td>
<td>0.27</td>
<td>0.14</td>
<td>0.31</td>
<td>0.44</td>
<td>0.42</td>
</tr>
<tr>
<td>trees</td>
<td>-0.06</td>
<td>0.23</td>
<td>-0.14</td>
<td>-0.27</td>
<td>0.14</td>
<td>0.24</td>
<td>0.55</td>
<td>0.77</td>
<td>0.66</td>
</tr>
<tr>
<td>graph</td>
<td>-0.06</td>
<td>0.34</td>
<td>-0.15</td>
<td>-0.30</td>
<td>0.20</td>
<td>0.31</td>
<td>0.69</td>
<td>0.98</td>
<td>0.85</td>
</tr>
<tr>
<td>minors</td>
<td>-0.04</td>
<td>0.25</td>
<td>-0.10</td>
<td>-0.21</td>
<td>0.15</td>
<td>0.22</td>
<td>0.50</td>
<td>0.71</td>
<td>0.62</td>
</tr>
</tbody>
</table>
Document Cosines

(before dimensionality reduction)

$$sim(x, y) = \cos(x, y) = \frac{\sum_{i=1}^{V} x_i y_i}{|x| \cdot |y|}$$

$$\cos(c1, c2) = \frac{1}{\sqrt{3} \times \sqrt{6}} = 0.2$$

<table>
<thead>
<tr>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>1.0</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c2</td>
<td>0.2</td>
<td>1.0</td>
<td>0.4</td>
<td>0.3</td>
<td>0.7</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c3</td>
<td>0.3</td>
<td>0.4</td>
<td>1.0</td>
<td>0.6</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c4</td>
<td>0.2</td>
<td>0.3</td>
<td>0.6</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c5</td>
<td>0.7</td>
<td>0.3</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.7</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>m2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.7</td>
<td>1.0</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>m3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>m4</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td>0.7</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

Dec 2, 2009
## Term Cosines

*(before dimensionality reduction)*

<table>
<thead>
<tr>
<th>term</th>
<th>human</th>
<th>interface</th>
<th>computer</th>
<th>user</th>
<th>system</th>
<th>response</th>
<th>time</th>
<th>EPS</th>
<th>survey</th>
<th>trees</th>
<th>graph</th>
<th>minors</th>
</tr>
</thead>
<tbody>
<tr>
<td>cosine</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cosine</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cosine</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.4</td>
<td>0.3</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cosine</td>
<td>0.4</td>
<td>0.4</td>
<td>1.0</td>
<td>0.5</td>
<td>0.8</td>
<td>0.8</td>
<td>0.4</td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cosine</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
<td>0.5</td>
<td>1.0</td>
<td>0.3</td>
<td>0.3</td>
<td>0.9</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cosine</td>
<td>0.5</td>
<td>0.8</td>
<td>0.3</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cosine</td>
<td>0.5</td>
<td>0.8</td>
<td>0.3</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cosine</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>0.9</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.4</td>
</tr>
<tr>
<td>cosine</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.4</td>
<td>0.5</td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>cosine</td>
<td>1.0</td>
<td>0.4</td>
<td>0.8</td>
<td>1.0</td>
<td>0.8</td>
<td>0.4</td>
<td>0.7</td>
<td>1.0</td>
<td>0.8</td>
<td></td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>cosine</td>
<td>0.5</td>
<td>0.4</td>
<td>0.8</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.7</td>
<td>0.4</td>
</tr>
</tbody>
</table>

**Dec 2, 2009**
## Document Cosines

*(after dimensionality reduction)*

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>1.0</td>
<td>0.9</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
<td>−0.2</td>
<td>−0.2</td>
<td>−0.2</td>
<td>−0.02</td>
</tr>
<tr>
<td>c2</td>
<td>0.9</td>
<td>1.0</td>
<td>0.9</td>
<td>0.9</td>
<td>1.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>c3</td>
<td>1.0</td>
<td>0.9</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
<td>−0.2</td>
<td>−0.2</td>
<td>−0.2</td>
<td>−0.01</td>
</tr>
<tr>
<td>c4</td>
<td>1.0</td>
<td>0.9</td>
<td>1.0</td>
<td>1.0</td>
<td>0.8</td>
<td>−0.3</td>
<td>−0.3</td>
<td>−0.3</td>
<td>−0.1</td>
</tr>
<tr>
<td>c5</td>
<td>0.9</td>
<td>1.0</td>
<td>0.9</td>
<td>0.8</td>
<td>1.0</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>m1</td>
<td>−0.2</td>
<td>0.2</td>
<td>−0.2</td>
<td>−0.3</td>
<td>0.3</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>m2</td>
<td>−0.2</td>
<td>0.2</td>
<td>−0.2</td>
<td>−0.3</td>
<td>0.3</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>m3</td>
<td>−0.2</td>
<td>0.3</td>
<td>−0.2</td>
<td>−0.3</td>
<td>0.3</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>m4</td>
<td>−0.02</td>
<td>0.4</td>
<td>−0.01</td>
<td>−0.1</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Dec 2, 2009
• Useful display for eye-balling a similarity matrix, e.g., cos of terms, cos of docs

• Can also play a role in IR

  • Term clusters can be used to reduce dimensionality ($V \rightarrow$ number of clusters)

  • Doc clusters can be used to reduce search space (score clusters, rather than documents)

• Input: a similarity matrix (square, symmetric)

• Output: a tree of clusters

• Methods: single linkage, complete linkage, k-means, and many more
Clustering
(before dimensionality reduction)
Clustering
(after dimensionality reduction)
Stop Lists & Term Weighting

- Emphasize content words and de-emphasize function words

\[
sim(x, y) = \frac{\sum_{t=1}^{V} (w_t x_t) (w_t y_t)}{|w_x| \cdot |w_y|}
\]

- IDF (inverse document freq)

\[
w_t = - \log_2 \left( \frac{\text{number of documents with term } t}{\text{number of documents } (= N)} \right)
\]
Evaluation

precision = \frac{\text{number of relevant & retrieved documents}}{\text{number of retrieved documents}}

recall = \frac{\text{number of relevant & retrieved documents}}{\text{number of relevant documents}}

Standarized Datasets (Ed Fox’s CDROM):
• MED, CACM, ADI, CISI, CRAN, TIME
• Queries, Documents, Relevance Judgments

Private Datasets:
• Bellcore Memos (Who Knows)
• Associated Press Newswire
Experimental Results: 100 Factors

MED: Precision-Recall Curves
Means across Queries

- LSI-100
- SMART
- TERM
- VO

Dec 2, 2009
Experimental Results: Number of Factors
• IR Problem: sort docs by $sim(d, q)$

• Vector-space “Model” (cosine similarity)
• Probabilistic Retrieval Model

• Clustering

• Correlations “Fixes”
  • Merge morphologically related words
  • Merge synonymous words using a thesaurus
  • Singular Value Decomposition (SVD)

• Evaluation

• Term Weighting: IDF & Entropy
Entropy of Search Logs

- How Big is the Web?
- How Hard is Search?
- With Personalization? With Backoff?

Qiaozhu Mei†, Kenneth Church‡
† University of Illinois at Urbana-Champaign
‡ Microsoft Research
How Big is the Web?

• What if a small cache of millions of pages
  – Could capture much of the value of billions?

• Could a Big bet on a cluster in the clouds
  – Turn into a big liability?

• Examples of Big Bets
  – Computer Centers & Clusters
    • Capital (Hardware)
    • Expense (Power)
    • Dev (Mapreduce, GFS, Big Table, etc.)
  – Sales & Marketing >> Production & Distribution

Dec 2, 2009
Millions (Not Billions)
Population Bound

- With all the talk about the Long Tail
  - You’d think that the Web was astronomical
  - Carl Sagan: Billions and Billions...
- Lower Distribution $$ \rightarrow $$ Sell Less of More
- But there are limits to this process
  - NetFlix: 55k movies (not even millions)
  - Amazon: 8M products
  - Vanity Searches: Infinite???
    - Personal Home Pages << Phone Book < Population
    - Business Home Pages << Yellow Pages < Population
- Millions, not Billions (until market saturates)
It Will Take Decades to Reach Population Bound

• Most people (and products)
  – don’t have a web page (yet)

• Currently, I can find famous people
  • (and academics)
  • but not my neighbors
  – There aren’t that many famous people
    • (and academics)...
  – Millions, not billions
    • (for the foreseeable future)
Equilibrium: Supply = Demand

• If there is a page on the web,
  – And no one sees it,
  – Did it make a sound?
• How big is the web?
  – Should we count “silent” pages
  – That don’t make a sound?
• How many products are there?
  – Do we count “silent” flops
  – That no one buys?
Demand Side Accounting

• Consumers have limited time
  – Telephone Usage: 1 hour per line per day
  – TV: 4 hours per day
  – Web: ??? hours per day

• Suppliers will post as many pages as consumers can consume (and no more)

• Size of Web: $O(\text{Consumers})$
How Big is the Web?

• Related questions come up in language
• How big is English?
  – Dictionary Marketing
  – Education (Testing of Vocabulary Size)
  – Psychology
  – Statistics
  – Linguistics
• Two Very Different Answers
  – Chomsky: language is infinite
  – Shannon: 1.25 bits per character
Chomskian Argument: Web is Infinite

• One could write a malicious spider trap

• Not just academic exercise

• Web is full of benign examples like
  – http://calendar.duke.edu/
  – Infinitely many months
  – Each month has a link to the next
How **Big** is the Web?

- More (Chomsky)
  - http://successor?x=0
- Less (Shannon)

Entropy (H)

MSN Search Log
1 month

Query
21.1

URL
22.1

IP
22.1

More Practical Answer

Comp Ctr ($$$$) ➔ Walk in the Park ($)

Cluster in Cloud ➔ Desktop ➔ Flash

Millions (not Billions)

All But IP
22.1 ➔ 22.6

All But URL
26.0

All But Query
27.1

All Three
27.2

Millions
(not Billions)

Desktop ➔ Flash

Dec 2, 2009
Entropy (H)

- \( H(X) = - \sum_{x \in X} p(x) \log p(x) \)
  - Size of search space; difficulty of a task
- \( H = 20 \rightarrow 1 \text{ million items distributed uniformly} \)
- Powerful tool for sizing challenges and opportunities
  - How hard is search?
  - How much does personalization help?
How Hard Is Search?
Millions, not Billions

- Traditional Search
  - $H(\text{URL} \mid \text{Query})$
  - $2.8 \ (= 23.9 - 21.1)$

- Personalized Search
  - $H(\text{URL} \mid \text{Query}, \text{IP})$
  - $1.2 \ (= 27.2 - 26.0)$

<table>
<thead>
<tr>
<th>Entropy (H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
</tr>
<tr>
<td>URL</td>
</tr>
<tr>
<td>IP</td>
</tr>
</tbody>
</table>

Personalization cuts $H$ in Half!

Dec 2, 2009
Difficulty of Queries

• Easy queries (low $H(\text{URL}|Q)$):
  – google, yahoo, myspace, ebay, ...

• Hard queries (high $H(\text{URL}|Q)$):
  – dictionary, yellow pages, movies,
  – “what is may day?”
How Hard are Query Suggestions?
The Wild Thing?  C* Rice → Condoleezza Rice

• Traditional Suggestions
  – \( H(\text{Query}) \)
  – 21 bits

• Personalized
  – \( H(\text{Query} | \text{IP}) \)
  – 5 bits (= 26 – 21)

Entropy (H)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td>21.1</td>
</tr>
<tr>
<td>URL</td>
<td>22.1</td>
</tr>
<tr>
<td>IP</td>
<td>22.1</td>
</tr>
<tr>
<td>All But IP</td>
<td>23.9</td>
</tr>
<tr>
<td>All But URL</td>
<td>26.0</td>
</tr>
<tr>
<td>All But Query</td>
<td>27.1</td>
</tr>
<tr>
<td>All Three</td>
<td>27.2</td>
</tr>
</tbody>
</table>

Personalization cuts \( H \) in Half!

Twice
Personalization with Backoff

• Ambiguous query: MSG
  – Madison Square Garden
  – Monosodium Glutamate
• Disambiguate based on user’s prior clicks
• When we don’t have data
  – Backoff to classes of users
• Proof of Concept:
  – Classes defined by IP addresses
• Better:
  – Market Segmentation (Demographics)
  – Collaborative Filtering (Other users who click like me)
Backoff

- Proof of concept: bytes of IP define classes of users
- If we only know some of the IP address, does it help?

| Bytes of IP addresses | H(URL| IP, Query) |
|-----------------------|------------------|
| 156.111.188.243       | 1.17             |
| 156.111.188.*         | 1.20             |
| 156.111.*.*           | 1.39             |
| 156.*.*.*             | 1.95             |
| *.*.*.*               | 2.74             |

Some of the IP is better than none

Cuts H in half even if using the first two bytes of IP
- Personalization with Backoff
- \( \lambda \)s estimated with EM and CV
- A little bit of personalization
  - Better than too much
  - Or too little

\[
P(\text{Url} \mid \text{IP}, Q) = \sum_{i=0}^{4} \lambda_i P(\text{Url} \mid \text{IP}_i, Q)
\]

\( \lambda_4 \): weights for first 4 bytes of IP
\( \lambda_3 \): weights for first 3 bytes of IP
\( \lambda_2 \): weights for first 2 bytes of IP

......
Personalization with Backoff

Market Segmentation

• Traditional Goal of Marketing:
  – Segment Customers (e.g., Business v. Consumer)
  – By Need & Value Proposition
    • Need: Segments ask different questions at different times
    • Value: Different advertising opportunities

• Segmentation Variables
  – Queries, URL Clicks, IP Addresses
  – Geography & Demographics (Age, Gender, Income)
  – Time of day & Day of Week
Business Queries on Business Days

Consumer Queries (Weekends & Every Day)
Business Days v. Weekends:
More Clicks and Easier Queries

More Clicks

Easier

Jan 2006 (1st is a Sunday)

Total Clicks

H(Url | IP, Q)
Day v. Night:
More queries (and easier queries) during business hours

More clicks and diversified queries

Less clicks, more unified queries

Day time: More & Diversified Queries

(Sundays: 01/08 & 01/15)
Harder Queries during Prime Time TV

Weekends are harder

21:00pm

05:00am

Search is hard at TV

Harder queries

Jan 2006 (Sundays: 01/08 & 01/15)
Conclusions: Millions (not Billions)

• How Big is the Web?
  – Upper bound: \(O(\text{Population})\)
    • Not Billions
    • Not Infinite
• Shannon >> Chomsky
  – How hard is search?
  – Query Suggestions?
  – Personalization?
• Cluster in Cloud ($$$$) \(\rightarrow\) Walk-in-the-Park ($)
Conclusions:
Personalization with Backoff

• Personalization with Backoff
  – Cuts search space (entropy) in half
  – Backoff ➔ Market Segmentation
    • Example: Business v. Consumer
      – Need: Segments ask different questions at different times
      – Value: Different advertising opportunities

• Demographics:
  – Partition by *ip, day, hour, business/consumer query*...

• Future Work:
  – Model combinations of surrogate variables
  – Group users with similarity ➔ collaborative search
Noisy Channel Model for Web Search
Michael Bendersky

• Input → Noisy Channel → Output
  – Input’ ≈ ARGMAX_{Input} Pr( Input ) * Pr( Output | Input )

• Speech
  – Words → Acoustics
  – Pr( Words ) * Pr( Acoustics | Words )

• Machine Translation
  – English → French
  – Pr( English ) * Pr( French | English )

• Web Search
  – Web Pages → Queries
  – Pr( Web Page ) * Pr( Query | Web Page )
Document Priors

• **Page Rank** *(Brin & Page, 1998)*
  – Incoming link votes

• **Browse Rank** *(Liu et al., 2008)*
  – Clicks, toolbar hits

• **Textual Features** *(Kraaij et al., 2002)*
  – Document length, URL length, anchor text
Query Priors: Degree of Difficulty

• Some queries are easier than others
  – Human Ratings (HRS): Perfect judgments ➔ easier
  – Static Rank (Page Rank): higher ➔ easier
  – Textual Overlap: match ➔ easier
    – “cnn” ➔ www.cnn.com (match)
  – Popular: lots of clicks ➔ easier (toolbar, slogs, glogs)
  – Diversity/Entropy: fewer plausible URLs ➔ easier
  – Broder’s Taxonomy:
    • Navigational/Transactional/Informational
    • Navigational tend to be easier:
      – “cnn” ➔ www.cnn.com (navigational)
      – “BBC News” (navigational) easier than “news” (informational)
Informational vs. Navigational Queries

- Fewer plausible URL’s → easier query
  - Click Entropy
    • Less is easier
  - Broder’s Taxonomy:
    • Navigational / Informational
    • Navigational is easier:
      - “BBC News” (navigational) easier than “news”
  - Less opportunity for personalization
    • (Teevan et al., 2008)
Informational/Navigational by Residuals

ClickEntropy ~ Log(#Clicks)

Informational

Navigational

Dec 2, 2009
Informational Vs. Navigational Queries

Residuals – Highest Quartile

"bay"   "car insurance"
"carinsurance"   "credit cards"
"date"   "day spa"
"dell computers"   "dell laptops"
"edmonds"   "encarta"
"hotel"   "hotels"
"house insurance"   "ib"
"insurance"   "kmart"
"loans"   "msn encarta"
"musica"   "norton"
"payday loans"   "pet insurance"
"proactive"   "sauna"

Residuals – Lowest Quartile

"accuweather"   "ako"
"bbc news"   "bebo"
"cnn"   "craigs list"
"craigslist"   "drudge"
"drudge report"   "espn"
"facebook"   "fox news"
"foxnews"   "friendster"
"imdb"   "mappy"
"mapquest"   "mixi"
"msnbc"   "my"
"my space"   "myspace"
"nexopia"   "pages jaunes"
"runescape"   "wells fargo"
Alternative Taxonomy: Click Types

• Classify queries by type
  – Problem: query logs have no “informational/navigational” labels

• Instead, we can use logs to categorize queries
  – **Commercial Intent** → more ad clicks
  – **Malleability** → more query suggestion clicks
  – **Popularity** → more future clicks (anywhere)
    • Predict future clicks (anywhere)
      – Past Clicks: February – May, 2008
      – Future Clicks: June, 2008
Aggregates over \((Q,U)\) pairs

\[
P(U) = \sum_Q P(Q,U)
\]

- Improve estimation by adding **features**
- Improve estimation by adding **aggregates**

<table>
<thead>
<tr>
<th>MODEL</th>
<th>Q/U Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Static Rank</td>
</tr>
<tr>
<td>Aggregates</td>
<td>max</td>
</tr>
</tbody>
</table>

Dec 2, 2009
Page Rank (named after Larry Page) aka Static Rank & Random Surfer Model
Page Rank = 1st Eigenvector

http://en.wikipedia.org/wiki/PageRank

So, the equation is as follows:

$$PR(p_i) = \frac{1 - d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

where $p_1, p_2, \ldots, p_N$ are the pages under consideration, $M(p_i)$ is the set of pages that link to $p_i$, $L(p_j)$ is the number of outbound links on page $p_j$, and $N$ is the total number of pages.

The PageRank values are the entries of the dominant eigenvector of the modified adjacency matrix. This makes PageRank a particularly elegant metric: the eigenvector is

$$R = \begin{bmatrix}
PR(p_1) \\
PR(p_2) \\
\vdots \\
PR(p_N)
\end{bmatrix}$$

where $R$ is the solution of the equation

$$R = \frac{\begin{bmatrix}
(1 - d)/N \\
(1 - d)/N \\
\vdots \\
(1 - d)/N
\end{bmatrix}}{N} + d \begin{bmatrix}
\ell(p_1, p_1) & \ell(p_2, p_1) & \cdots & \ell(p_N, p_1) \\
\ell(p_1, p_2) & \cdots & \vdots & \vdots \\
\ell(p_1, p_N) & \cdots & \ell(p_i, p_j) & \ell(p_N, p_N)
\end{bmatrix} \cdot R$$

where the adjacency function $\ell(p_i, p_j)$ is 0 if page $p_i$ does not link to $p_j$, and normalised such that, for each $i$,

$$\sum_{i=1}^{N} \ell(p_i, p_j) = 1,$$

i.e. the elements of each column sum to 1. This is a variant of the eigenvector centrality measure used commonly in network analysis.
Document Priors are like Query Priors

• Human Ratings (HRS): Perfect judgments $\rightarrow$ more likely
• Static Rank (Page Rank): higher $\rightarrow$ more likely
• Textual Overlap: match $\rightarrow$ more likely
  – “cnn” $\rightarrow$ www.cnn.com (match)
• Popular:
  – lots of clicks $\rightarrow$ more likely (toolbar, slogs, glogs)
• Diversity/Entropy:
  – fewer plausible queries $\rightarrow$ more likely
• Broder’s Taxonomy
  – Applies to documents as well
  – “cnn” $\rightarrow$ www.cnn.com (navigational)
Task Definition

• What will determine future clicks on the URL?
  – Past Clicks?
  – High Static Rank?
  – High Toolbar visitation counts?
  – Precise Textual Match?
  – All of the Above?

• ~3k queries from the extracts
  – 350k URL’s
  – Past Clicks: February – May, 2008
  – Future Clicks: June, 2008
## Estimating URL Popularity

<table>
<thead>
<tr>
<th>URL Popularity</th>
<th>Normalized RMSE Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Extract</td>
</tr>
<tr>
<td><strong>Linear Regression</strong></td>
<td></td>
</tr>
<tr>
<td>A: Regression</td>
<td>.619</td>
</tr>
<tr>
<td>B: Classification + Regression</td>
<td>-</td>
</tr>
<tr>
<td><strong>Neural Network (3 Nodes in the Hidden Layer)</strong></td>
<td></td>
</tr>
<tr>
<td>C: Regression</td>
<td>.619</td>
</tr>
</tbody>
</table>

- **B is better than A**
- **Extract + Clicks: Better Together**
Destinations by Residuals
Real and Fake Destinations

Residuals – Highest Quartile

- actualkeywords.com/base_top50000.txt
- blog.nbc.com/heroes/2007/04/wine_and_guests.php
- everyscreen.com/views/sex.htm
- freesex.zip.net
- fuck-everyone.com
- home.att.net/~btuttleman/barrysite.html
- jibbering.com/blog/p=57
- migune.nipox.com/index-15.html
- mp3-search.hu/mp3shudownl.htm
- www.123rentahome.com
- www.automotivetalk.net/showmessages.phpid=3791
- www.canammachinerysales.com
- www.cardpostage.com/zorn.htm
- www.driverguide.com/drilist.htm
- www.driverguide.com/drivers2.htm
- www.esmimusica.com

Residuals – Lowest Quartile

- espn.go.com
- fr.yahoo.com
- games.lg.web.tr
- gmail.google.com
- it.yahoo.com
- mail.yahoo.com
- www.89.com
- www.aol.com
- www.cnn.com
- www.ebay.com
- www.facebook.com
- www.free.fr
- www.free.org
- www.google.ca
- www.google.co.jp
- www.google.co.uk

Dec 2, 2009
Fake Destination Example

actualkeywords.com/base_top50000.txt

Clicked ~110,000 times
In response to ~16,000 unique queries

Dictionary Attack
Learning to Rank with Document Priors

• **Baseline: Feature Set A**
  – Textual Features (5 features)

• **Baseline: Feature Set B**
  – Textual Features + Static Rank (7 features)

• **Baseline: Feature Set C**
  – All features, with click-based features filtered (382 features)

• **Treatment: Baseline + 5 Click Aggregate Features**
  – Max, Median, Entropy, Sum, Count
Summary: Information Retrieval (IR)

• Boolean Combinations of Keywords
  – Popular with Intermediaries (Librarians)
• Rank Retrieval
  – Sort a collection of documents
    • (e.g., scientific papers, abstracts, paragraphs)
    • by how much they “match” a query
  – The query can be a (short) sequence of keywords
    • or arbitrary text (e.g., one of the documents)
• Logs of User Behavior (Clicks, Toolbar)
  – Solitaire \(\rightarrow\) Multi-Player Game:
    • Authors, Users, Advertisers, Spammers
  – More Users than Authors \(\rightarrow\) More Information in Logs than Docs
  – Learning to Rank:
    • Use Machine Learning to combine doc features & log features