Information Retrieval: Queries

Fernando Diaz

Microsoft Research NYC

October 27, 2014
given a **query** and a **corpus**, find **relevant documents**.

- **query**: user’s expression of the information need
- **corpus**: the repository of retrievable items
- **relevance**: satisfaction of the information need
Belkin et al (1982) proposed a model called Anomalous State of Knowledge

ASK hypothesis:
- difficult for people to define exactly what their information need is, because that information is a gap in their knowledge
- Search engine should look for information that fills those gaps
- Interesting ideas, little practical impact (yet)

Based on a slide http://www.search-engines-book.com/slides/
Query Data

- **explicit**: user expression of the information need
- **implicit**: unexpressed portion of the information need
Explicit Query Data

- **description**: a paragraph-length expression
- **question**: natural language expression of query
- **title**: exact description of a known item (e.g. paper title, navigational destination)
- **keyword**: a short set of words
- **structured**: a system-specific description optimized to retrieve relevant documents
Relevant documents will include those on historical exploration and drilling as well as history of regulatory bodies. Relevant are history of the oil industry in various states, even if drilling began in 1950 or later.
What is the history of the U.S. oil industry?
Explicit Query Data

US Oil History Wikipedia
Explicit Query Data

keyword

us oil history
Explicit Query Data

structured

#weight(
0.90 #combine(#syn(us #1(united states)) oil history).title
0.10 #combine(#syn(us #1(united states)) oil history).body
)
Implicit Query Data

- **synonyms**: words missing from the query
- **constraints**: metadata filters missing from the query
- **session**: session-specific information
- **personal**: user-specific information
Implicit Query Data

synonyms

• introduced in Lecture 2 (Vocabulary Mismatch)
• many ways to address synonym expansion
  • acronyms
  • thesauri
  • statistical co-occurrences (e.g. LSI)
• how do we know which technique is best?
Implicit Query Data

constraints

- many unexpressed in queries
  - **temporal**: ‘(most recent) election news’, ‘(oldest) CS publication’
  - **geographic**: ‘(nearby) cafes’
  - **corpus**: ‘japan earthquake (news)’
  - ...
- may be combined
  - ‘(most recent) (local) election (news)’
  - ‘(earliest) search engine invention (in patent database)’
- once detected, constraints are easy to implement
  - very difficult to detect
Implicit Query Data

**search session**: sequence of information-seeking interactions

- queries issued
- documents read, printed, skipped
- documents skipped

- when a query is issued, session-level information can help with,
  - disambiguation (e.g. ['dance lessons'→'local dance lessons'→'salsa'])
  - document-level relevance feedback

- easy to log
  - difficult to exploit
Implicit Query Data
personal

- **user information**
  - demographics
  - searching history
  - browsing history
  - social network
- when a query is issued, user-level information can help with,
  - disambiguation
  - ‘re-finding’ queries
- difficult to log
  - difficult to exploit
Query Representation

• Given all of the explicit and implicit and explicit sources of query data, how do we represent a query for use in a retrieval system?
  • **text**: embed queries in the same text space as documents.
  • **metadata**: embed queries in the same metadata space documents.

• Translating from an under-specified query to a rich query is a fundamental problem in information retrieval.
Outline

Corpus-Driven Query Expansion

Log-Driven Query Expansion
Corpus-Driven Query Expansion
Query Expansion

• **query expansion** refers to a set of techniques for recognizing query terms which *should* be in the query but are not.

• manually-built thesauri tend to be brittle
  • make assumptions about the corpus (i.e. synonyms are not portable across corpora).
  • do not capture all possible synonyms.
  • need to be maintained.

• statistical, data-driven thesauri
  • no assumptions except co-occurrence.
  • high coverage.
  • completely automated.
Query Expansion
Global vs. Local

- **Global methods**: use statistics in the entire corpus.
  - **advantage**: leverages large amounts of data.
  - **disadvantage**: may miss query-specific regularity.

- **Local methods**: use statistics in the top retrieved documents.
  - **advantage**: leverages query-specific statistics.
  - **disadvantage**: poor data for small topics; poor for bad queries.
# Query Expansion

## Global vs. Local

<table>
<thead>
<tr>
<th></th>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global</strong></td>
<td>[bike race] → [bike bicycle race]</td>
<td>[bank account] → [river bank account]</td>
</tr>
<tr>
<td><strong>Local</strong></td>
<td>[bank account] → [savings bank account]</td>
<td>[bike race] → [brooklyn bike race]</td>
</tr>
</tbody>
</table>
term-at-a-time refers to expanding each query term individually.

- advantage: global statistics can be precomputed.
- disadvantage: may miss query-specific regularity (e.g. [rock climbing]).

objective: introduce terms which co-occur with the candidate term.

- e.g. [bike] co-ccurs with [bicycle] in many documents
Query Expansion
Statistical Term Association

Given two terms $a$ and $b$,

- Dice coefficient:
  \[
  \frac{n_{ab}}{n_a + n_b}
  \]

- mutual information
  \[
  \log \frac{p(ab)}{p(a)p(b)}
  \]

where $n_x$ is the frequency of string $x$.

Query Expansion
Statistical Term Association

- expected mutual information (addresses low-frequency terms)

\[ p(ab) \log \frac{p(ab)}{p(a)p(b)} \]

- Pearson’s \( \chi^2 \)

\[ \frac{(n_{ab} - \frac{n_a n_b}{N})^2}{n_a n_b} \]

## Query Expansion

**related terms for ‘tropical’**

<table>
<thead>
<tr>
<th>$MIM$</th>
<th>$EMIM$</th>
<th>$\chi^2$</th>
<th>$Dice$</th>
</tr>
</thead>
<tbody>
<tr>
<td>trmm</td>
<td>forest</td>
<td>trmm</td>
<td>forest</td>
</tr>
<tr>
<td>itto</td>
<td>tree</td>
<td>itto</td>
<td>exotic</td>
</tr>
<tr>
<td>ortuno</td>
<td>rain</td>
<td>ortuno</td>
<td>timber</td>
</tr>
<tr>
<td>kuroshio</td>
<td>island</td>
<td>kuroshio</td>
<td>rain</td>
</tr>
<tr>
<td>ivirgarzama</td>
<td>like</td>
<td>ivirgarzama</td>
<td>banana</td>
</tr>
<tr>
<td>biofunction</td>
<td>fish</td>
<td>biofunction</td>
<td>deforestation</td>
</tr>
<tr>
<td>kapiolani</td>
<td>most</td>
<td>kapiolani</td>
<td>plantation</td>
</tr>
<tr>
<td>bstilla</td>
<td>water</td>
<td>bstilla</td>
<td>coconut</td>
</tr>
<tr>
<td>almagreb</td>
<td>fruit</td>
<td>almagreb</td>
<td>jungle</td>
</tr>
<tr>
<td>jackfruit</td>
<td>area</td>
<td>jackfruit</td>
<td>tree</td>
</tr>
<tr>
<td>adeo</td>
<td>world</td>
<td>adeo</td>
<td>rainforest</td>
</tr>
<tr>
<td>xishuangbanna</td>
<td>america</td>
<td>xishuangbanna</td>
<td>palm</td>
</tr>
<tr>
<td>frangipani</td>
<td>some</td>
<td>frangipani</td>
<td>hardwood</td>
</tr>
<tr>
<td>yuca</td>
<td>live</td>
<td>yuca</td>
<td>greenhouse</td>
</tr>
<tr>
<td>anthurium</td>
<td>plant</td>
<td>anthurium</td>
<td>logging</td>
</tr>
</tbody>
</table>

## Query Expansion

related terms for ‘fish’

<table>
<thead>
<tr>
<th>MIM</th>
<th>EMIM</th>
<th>( \chi^2 )</th>
<th>Dice</th>
</tr>
</thead>
<tbody>
<tr>
<td>zoologico</td>
<td>water</td>
<td>arlsq</td>
<td>species</td>
</tr>
<tr>
<td>zapanta</td>
<td>species</td>
<td>happyman</td>
<td>wildlife</td>
</tr>
<tr>
<td>wrint</td>
<td>wildlife</td>
<td>outerlimit</td>
<td>fishery</td>
</tr>
<tr>
<td>wpfmc</td>
<td>fishery</td>
<td>sportk</td>
<td>water</td>
</tr>
<tr>
<td>weighout</td>
<td>sea</td>
<td>lingcod</td>
<td>fisherman</td>
</tr>
<tr>
<td>waterdog</td>
<td>fisherman</td>
<td>longfin</td>
<td>boat</td>
</tr>
<tr>
<td>longfin</td>
<td>boat</td>
<td>bontadelli</td>
<td>sea</td>
</tr>
<tr>
<td>veracruzana</td>
<td>area</td>
<td>sportfisher</td>
<td>habitat</td>
</tr>
<tr>
<td>ungtt</td>
<td>habitat</td>
<td>billfish</td>
<td>vessel</td>
</tr>
<tr>
<td>ulocentra</td>
<td>vessel</td>
<td>needlefish</td>
<td>marine</td>
</tr>
<tr>
<td>needlefish</td>
<td>marine</td>
<td>damaliscu</td>
<td>endanger</td>
</tr>
<tr>
<td>tunaboot</td>
<td>land</td>
<td>bontebok</td>
<td>conservation</td>
</tr>
<tr>
<td>tsolwana</td>
<td>river</td>
<td>taucher</td>
<td>river</td>
</tr>
<tr>
<td>olivacea</td>
<td>food</td>
<td>orangemouth</td>
<td>catch</td>
</tr>
<tr>
<td>motoroller</td>
<td>endanger</td>
<td>sheepshead</td>
<td>island</td>
</tr>
</tbody>
</table>

## Query Expansion

**related terms for ‘fish’ in a 5 word window**

<table>
<thead>
<tr>
<th>MIM</th>
<th>EMIM</th>
<th>$\chi^2$</th>
<th>Dice</th>
</tr>
</thead>
<tbody>
<tr>
<td>zapanta</td>
<td>wildlife</td>
<td>gefilte</td>
<td>wildlife</td>
</tr>
<tr>
<td>plar</td>
<td>vessel</td>
<td>mbmo</td>
<td>vessel</td>
</tr>
<tr>
<td>mbmo</td>
<td>boat</td>
<td>zapanta</td>
<td>boat</td>
</tr>
<tr>
<td>gefilte</td>
<td>fishery</td>
<td>plar</td>
<td>fishery</td>
</tr>
<tr>
<td>hapc</td>
<td>species</td>
<td>hapc</td>
<td>species</td>
</tr>
<tr>
<td>odfw</td>
<td>tuna</td>
<td>odfw</td>
<td>catch</td>
</tr>
<tr>
<td>southpoint</td>
<td>trout</td>
<td>southpoint</td>
<td>water</td>
</tr>
<tr>
<td>anadromous</td>
<td>fisherman</td>
<td>anadromous</td>
<td>sea</td>
</tr>
<tr>
<td>taiffe</td>
<td>salmon</td>
<td>taiffe</td>
<td>meat</td>
</tr>
<tr>
<td>mollie</td>
<td>catch</td>
<td>mollie</td>
<td>interior</td>
</tr>
<tr>
<td>frampton</td>
<td>nmf</td>
<td>frampton</td>
<td>fisherman</td>
</tr>
<tr>
<td>idfg</td>
<td>trawl</td>
<td>idfg</td>
<td>game</td>
</tr>
<tr>
<td>billingsgate</td>
<td>halibut</td>
<td>billingsgate</td>
<td>salmon</td>
</tr>
<tr>
<td>sealord</td>
<td>meat</td>
<td>sealord</td>
<td>tuna</td>
</tr>
<tr>
<td>longline</td>
<td>shellfish</td>
<td>longline</td>
<td>caught</td>
</tr>
</tbody>
</table>

Based on a slide http://www.search-engines-book.com/slides/
Query Expansion

Vector Space Term Association

- represent individual terms by the documents they occur in,
  - [bike] co-occurs with \( \{d_{231}, d_{134}, \ldots\} \)
- use vector space model representation for each vector of documents.
- term association based on cosine similarity of term vectors.
Query Expansion
Vector Space Term Association

\[ \hat{a} = \mathbf{C}a \quad \hat{b} = \mathbf{C}b \]

\[ \text{sim}(a, b) = \hat{a}^T\hat{b} = a^T\mathbf{C}^T\mathbf{C}b \]

\[ \mathbf{C} \quad n \times m \text{ corpus} \]
\[ a \quad \text{vector of zeros with a 1 for entry } a \]
\[ b \quad \text{vector of zeros with a 1 for entry } b \]
• represent individual terms by the terms they co-occur with,
  ● [bike] co-ccurs with {chain, ride, helmet, derailleur}
• can treat candidate terms as ‘pseudo-documents’
  ● use vector space model representation for each pseudo-document
• term association based on cosine similarity of term vectors.
Query Expansion
Vector Space Term Association

\[ \tilde{a} = C^T C a \quad \tilde{b} = C^T C b \]

\[ \text{sim}(a, b) = \tilde{a}^T \tilde{b} = a^T C^T C C^T C b = \hat{a}^T C C^T \hat{b} \]
Query Expansion
Statistical Query Association

• **query-at-a-time** refers to adding terms related to the entire query.
  • **advantage:** query context preserved.
  • **disadvantage:** inefficient for new queries; poor for bad queries.

• **objective:** introduce terms which co-occur with the entire query.
  • e.g. [river bank] co-occurs with [fresh water] in many documents

• **pseudo-relevance feedback:** extract terms and phrases from an initial retrieval.
Pseudo-Relevance Feedback

\[\text{PseudoRelevanceFeedback}(q)\]

\[\begin{align*}
1 & \quad \mathcal{D} \leftarrow \text{RetrieveDocument}(q) \quad \triangleright \text{Lecture 3} \\
2 & \quad q' \leftarrow \text{ExtractTerms}(\mathcal{D}) \\
3 & \quad \tilde{q} \leftarrow \text{CombineTerms}(q, q') \\
4 & \quad \text{return } \tilde{q}
\end{align*}\]

Usually parameterized by number of documents \(|\mathcal{D}|\), number of terms returned by \text{ExtractTerms}(\mathcal{D})\), and a combination weight for \text{CombineTerms}(q, q')\).
Initial retrieval, $\mathcal{D}$, contains documents with language ‘similar’ to the query.

**Most-frequent**: select the most frequent terms in $\mathcal{D}$.

**Most-distinguishing**: select the most distinguishing terms in $\mathcal{D}$,

$$\text{weight}(w) = \frac{p(w|\mathcal{D})}{p(w|\mathcal{C})}$$
Pseudo-Relevance Feedback

Defining CombineTerms\((q, q')\)

- Feedback terms, \(q'\), can be different-from-but-related-to the original query intent.
  - This is referred to as ‘query drift’.
- CombineTerms\((q, q')\) usually ‘anchors’ the expanded query to the original query intent.
  - Rocchio feedback,

\[
\text{CombineTerms}(q, q') = \lambda q + (1 - \lambda)q'
\]

where a large \(\lambda\) gives more weight to the original query.
Pseudo-Relevance Feedback

[ Lavrenko and Croft 2001 ]

![Pseudo-Relevance Feedback Graph]

- **X-axis**: Recall
- **Y-axis**: Precision

Comparison of models:
- **LM**: Language Model
- **LM + expansion**: Language Model with expansion
- **Relevance Model**: Augmented Relevance Model

Legend:
- **LM**: solid line
- **LM + expansion**: dashed line
- **Relevance Model**: dotted line

Stars indicate statistically significant differences in performance with a 95% confidence according to the Wilcoxon test.
• Similar techniques can be used for inferring unexpressed metadata.
• Historically, corpus-driven query expansion provided the strongest, most robust method for query expansion.
• Access to query-logs changed this...
Log-Driven Query Expansion
Log-Driven Query Expansion

- Instrumentation of search interfaces provides rich query and click data.
- Query logs provide a rich source of information directly related to users’ search intents.
- Two principle ways of exploiting this data,
  - click behavior
  - session behavior
Click Logs

- **Assumption**: users click on documents which are relevant.
  - untrue for accidental clicks.
- **Assumption**: all queries which find a document $d$ relevant are related.
  - untrue for multifaceted documents.
- strong assumptions but effective in practice.
Click Logs

co-clicks

Queries

URLs

q1

u1

20

30

q2

u2

100

100

q3

u3

40

120

q4

u4

10

u5
Click Logs
co-clicks (images)

Figure 1: Click graph. Nodes are queries or images, edges indicate clicks. Images A and B are equidistant from the query 'panda' (distance=3), so retrieval based on a na"ıve shortest-path algorithm could not distinguish them. Our Markov random walk approach sums over paths, so image A benefits from having 7 distinct paths of length 3. In other words, nodes A and 'panda' are connected by a large "volume" of paths.

Application areas for such algorithms include:

- Query-to-document 'search'. Given a query, find relevant documents, as in adhoc search. Relevant documents should be ranked highly regardless of whether they are adjacent to the query. Search is the focus of this paper.
- Query-to-query 'suggestion'. Given a query, find other queries that the user might like to run. This is more difficult to evaluate, but approaches already exist for finding related queries using the click graph [4, 13].
- Document-to-query 'annotation'. Given a document, attach related queries to it. This method for creating document surrogates, based entirely on the click graph, was studied in [15].
- Document-to-document 'relevance feedback'. Given an example document that is relevant to the user, find additional relevant documents.

Two of the above approaches [4, 13] use the graph to identify query-to-query similarity followed by the application of clustering algorithms. Xue et al [15] use the graph to find document-to-document similarities. Annotations (query associations) are then spread amongst similar documents. All three papers make use of co-click/co-visitation information, for example, two queries are similar if they have overlapping sets of clicked documents. Xue et al also try an iterative algorithm, based on query-to-query and document-to-document similarity.

The focus of this paper is the search problem. However, the Markov random walk methods studied here handle all these cases. We discuss other cases briefly in Section 5.

3. RANDOM WALK MODEL

To derive our probabilistic retrieval model, we first propose a basic query formulation model. The model captures a process that starts from an information need and ends with a query.

We assume that query formulation begins with the user imagining a single document, representing their information need. They then think of a query that is associated with the document. The process might stop at that query, at which point they issue the query. Alternatively, the query makes them imagine another document, and that document makes them imagine another query. This thought process of query-document and document-query transition can repeat, or it can stop at a query which is then issued.

Our model, detailed in Section 3.1, makes a number of simplifying assumptions. The user has limited memory, so forgets their previous location after each transition. Although they do not remember their starting point, our model limits the number of transitions to keep them in the vicinity.
Click Logs

co-clicks

\[ \hat{a} = Y a \qquad \hat{b} = Y b \]

\[ \text{sim}(a, b) = \hat{a}^T \hat{b} = a^T Y^T Y b \]

- $Y$  $n \times m$ click log
- $a$ vector of zeros with a 1 for entry $a$
- $b$ vector of zeros with a 1 for entry $b$
Click Logs
[Cao et al. 2008]

<table>
<thead>
<tr>
<th>Test Case</th>
<th>Adjacency</th>
<th>N-Gram</th>
<th>CACB</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.at&amp;t.com">www.at&amp;t.com</a></td>
<td>at&amp;t</td>
<td>at&amp;t</td>
<td>att wireless</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.at.com">www.at.com</a></td>
<td><a href="http://www.att.com">www.att.com</a></td>
<td>at&amp;t online billing</td>
</tr>
<tr>
<td></td>
<td>cingular</td>
<td>cingular</td>
<td>cingular</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.cingular.com">www.cingular.com</a></td>
<td><a href="http://www.cingular.com">www.cingular.com</a></td>
<td>verizon</td>
</tr>
<tr>
<td>msn news</td>
<td>cnn news</td>
<td>cnn news</td>
<td>cnn news</td>
</tr>
<tr>
<td></td>
<td>fox news</td>
<td>fox news</td>
<td>fox news</td>
</tr>
<tr>
<td></td>
<td>cnn</td>
<td>cnn</td>
<td>abc news</td>
</tr>
<tr>
<td></td>
<td>msn</td>
<td>msn</td>
<td>cbs news</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bbc news</td>
</tr>
<tr>
<td><a href="http://www.chevrolet.com">www.chevrolet.com</a></td>
<td>gmc acadia</td>
<td>&lt;null&gt;</td>
<td>ford</td>
</tr>
<tr>
<td>⇒ <a href="http://www.gmc.com">www.gmc.com</a></td>
<td><a href="http://www.chevy.com">www.chevy.com</a></td>
<td></td>
<td>toyota</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.chevrolet.com">www.chevrolet.com</a></td>
<td></td>
<td>nissan</td>
</tr>
<tr>
<td></td>
<td>gmc envoy</td>
<td></td>
<td>pontiac</td>
</tr>
<tr>
<td></td>
<td>gmc cars</td>
<td></td>
<td></td>
</tr>
<tr>
<td>circuit city ⇒ best buy</td>
<td>circuit city</td>
<td>walmart</td>
<td>walmart</td>
</tr>
<tr>
<td></td>
<td>walmart</td>
<td>target</td>
<td>target</td>
</tr>
<tr>
<td></td>
<td>target</td>
<td>sears</td>
<td>sears</td>
</tr>
<tr>
<td></td>
<td>best buy stores</td>
<td>office depot</td>
<td>radio shack</td>
</tr>
<tr>
<td></td>
<td>sears</td>
<td></td>
<td>staples</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Session Logs**

- **Assumption**: multiple queries submitted within a single user’s search session are related.
- **Problem**: detecting a search session.
  - multiple users from a single IP.
  - multiple search tasks for a single user.
  - timeouts are crude.
- strong assumptions but effective in practice.
Table 2. Most frequent queries that precede and follow the queries “walmart” and “target” in user search sessions, and the corresponding precede and follow frequencies.

<table>
<thead>
<tr>
<th>Target query: walmart</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follows</td>
</tr>
<tr>
<td>Freq</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>3932</td>
</tr>
<tr>
<td>2254</td>
</tr>
<tr>
<td>1074</td>
</tr>
<tr>
<td>1046</td>
</tr>
<tr>
<td>868</td>
</tr>
<tr>
<td>523</td>
</tr>
<tr>
<td>371</td>
</tr>
<tr>
<td>357</td>
</tr>
<tr>
<td>331</td>
</tr>
<tr>
<td>326</td>
</tr>
</tbody>
</table>
Session Logs

- Represent each query as a distribution over other queries, \( p(q_t|q_s) \)
- Estimate each query distribution using session log data,

\[
p(q_t|q_s) \propto \frac{\#(q_t, q_s)}{\#(q_s)}
\]

- Compute query similarity by comparing query distributions,

\[
sim(q_s, q_s') = \sum_{q_t} p(q_t|q_s) \log \frac{p(q_t|q_s)}{p(q_t|q_s')}\]
### Session Logs

[Cucerzan and Brill 2006]

| bank of america | wells fargo, american express, capital one, washington mutual, bank one, wachovia, sprint pcs, citibank, providian, mbna |
**Session Logs**

[Cucerzan and Brill 2006]

| united airlines | american airlines, delta airlines, southwest airlines, northwest airlines, continental airlines, expedia, travelocity, us airways, orbitz, alaska airlines |
### Session Logs

[Cucerzan and Brill 2006]

| spiderman          | batman, hulk, superman, kirsten dunst, disney, harry potter, x-men, marvel comics, power rangers, barbie |
Other Uses of Query and Click Logs

Evaluation

- Many reformulations $\rightarrow$ system failure
- Many clicks on different results $\rightarrow$ multiple intents (click entropy)
  - is this always true?
- No click $\rightarrow$ system failure (abandonment)
  - is this always true?
Other Uses of Query and Click Logs

Information Extraction

- Queries often include both entity names as well as entity attributes.
  - $[\text{macbook} \mid \text{firewire}]$
    - entity
    - attribute
  - $[\text{lenovo} \mid \text{firewire}]$
    - entity
    - attribute

- Frequently co-occurring query terms suggest possible extractable attributes.
  - Don’t need to manually specify a taxonomy.
Other Uses of Query and Click Logs
Health Informatics

- Users often query about health symptoms before visiting a physician.
- **Application**: detect flu trends by looking for queries associated with flu symptoms.
- **Application**: detect drug side-effects by looking for patients searching for symptoms.
Other Uses of Query and Click Logs

Flu Trends

Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.

Summary

- Query logs provide a rich source of query expansion data.
- Successful, general purpose information retrieval systems resulted a rich source of query information.
  - **caveat**: need to properly instrument logging.
  - **caveat**: data is not public; need to have users!
- Search engines can be seen as very large sensors for what users are interested in at most times of day.
  - can do a lot more than query expansion!