Of Search and Semantics

Patrick Pantel

NSF Symposium on Semantic Knowledge
Discovery, Organization and Use
November 15, 2008
Vannaver Bush proposes to build a body of knowledge for all mankind: Memex

Memex, in the form of a desk, would instantly bring files and material on any subject to the operator’s fingertips.

Bush: "technical difficulties of all sorts have been ignored."

Semantics captured by an associative trail, personal comments and side trails

Gerald Salton, father of modern search: introduces concepts such as vector space model, Inverse Document Frequency (IDF), Term Frequency (TF), and relevancy feedback

1990: the first search engine Archie, from McGill, matches keyword queries against a database of Web filenames

Archie Query Form

Search for:
Yahoo! is founded around a taxonomical organization of the Web... manually

1994

Yahoo - A Guide to WWW

1994-1996

Key innovator:
Advanced search features + first SE to allow natural language queries

Key innovator:
Semantic models for ranking paid search results

Semantics and Query Logs:
Technologies developed/applied during the DARPA Tipster program and TREC make it commercial: spelling correction, query reformulations, also try,...
Future of search lies in a deep understanding and matching of information request behind user queries

Natural language questions answered by editors

Semantic repositories and user-annotated content grow rapidly

Semantic search engines emerge

Delicious social bookmarking

-7-

-8-
Semantics = 1$$, $$, $$, $$, $$}

2008
Aggregate star rating

Example review as summary

Current prices at merchant sites

Images, maps, specs, ...

---

Opportunities:
- Marriage of information extraction, content analysis, and query intent modeling
- User experience design
- Key Technologies
  - IE: Entity detection and salience; attribute detection
  - CA: Text classification, aboutness, information fusion
  - QIM: Entity detection, intent/task understanding
Task Modeling

Ana-Maria Popescu

• What is the user trying to do?
  – Find product reviews of the Nikon D300?
  – Buy a Nikon D300?
  – Find support for her camera?

• UED: enriching the search experience
  – How can we make use of this knowledge?

• Key Technologies
  – Entity detection
  – Document classification
  – Intent modeling / detection
Research Problem: Intent Synonymy

Is battery life a synonym of image quality?

How can we automatically discover intent synonyms?

Key Assumption: one intent per session
A searcher's intent remains the same within a search session

Intent Synonyms
Methodology very similar to constructing dictionary of distributionally similar words

Intent Discovery and Chaining via Clustering

<table>
<thead>
<tr>
<th>Review</th>
<th>Price</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>best thin</td>
<td>where can I get cheap</td>
<td>common problems</td>
</tr>
<tr>
<td>rate</td>
<td>black Friday sale</td>
<td>fault codes</td>
</tr>
<tr>
<td>compare kodak vs. canon</td>
<td>christmas sales</td>
<td>installing new</td>
</tr>
<tr>
<td>small portable</td>
<td>cheap</td>
<td>easy use</td>
</tr>
<tr>
<td>high zoom</td>
<td>buy now pay later</td>
<td>operating instructions</td>
</tr>
<tr>
<td>rate best</td>
<td>dell discount</td>
<td>won't heat</td>
</tr>
<tr>
<td>battery life</td>
<td>great deals</td>
<td>best schematics</td>
</tr>
<tr>
<td>user comments</td>
<td>overstock.com</td>
<td>keeps shutting</td>
</tr>
</tbody>
</table>
## Road Ahead

<table>
<thead>
<tr>
<th>Key Applications</th>
<th>Enabling Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart Snippets</td>
<td>Entity Detection</td>
</tr>
<tr>
<td>Task Modeling</td>
<td>Attribute Mining</td>
</tr>
<tr>
<td>Intent Modeling</td>
<td>Entailment</td>
</tr>
<tr>
<td>Search Assist</td>
<td>Distributional Similarity</td>
</tr>
<tr>
<td>Advertising Sciences</td>
<td>Intent Synonymy</td>
</tr>
<tr>
<td></td>
<td>Anaphora Resolution</td>
</tr>
<tr>
<td></td>
<td>Content Analysis</td>
</tr>
<tr>
<td></td>
<td>Text Classification</td>
</tr>
</tbody>
</table>
Dynamic Similarity Modeling
Eric Crestan and Vishnu Vyas

SeeLEx: Seed List Expansion

- SeeLEx, developed at Yahoo!, is a weakly supervised system for mining entity lists based on distributional similarity

Honda

Jaguar

Peugeot

Mercedes

Ford
SeeLEx: Seed List Expansion

- What is a caracal?
  - endangered carnivore fast hunted ...

- 25 -
Questions Asked and Conclusions

What is the effect of corpus size on expansion accuracy?

- Significant boost in performance

What is the effect of corpus quality on expansion accuracy?

- Significant boost in performance
- Does seed quality matter?
- Great variance in performance depending on set seed composition

How many seeds are on average optimal for expansion accuracy and are more seeds better than less?

- Somewhat surprising: 5-20 seeds in general is sufficient; more seeds gains little but don’t hurt

Gold Standard Sets


Average of 208 instances
Minimum of 11
Maximum of 1116
Total of 10,377 instances

50 sets extracted from Wikipedia (2007/12)
**Corpora**

Table: Corpora used to build our expansion models.

<table>
<thead>
<tr>
<th>CORPORA</th>
<th>UNIQUE SENTENCES (MILLIONS)</th>
<th>TOKENS (MILLIONS)</th>
<th>UNIQUE WORDS (MILLIONS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>k100</td>
<td>5,201</td>
<td>217,940</td>
<td>542</td>
</tr>
<tr>
<td>k020†</td>
<td>1040</td>
<td>43,588</td>
<td>108</td>
</tr>
<tr>
<td>k004†</td>
<td>208</td>
<td>8,717</td>
<td>22</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>30</td>
<td>721</td>
<td>34</td>
</tr>
</tbody>
</table>

† Estimated from k100 statistics.

---

**The Effect of:**

**Corpus Size and Corpus Quality**
**System and Corpora Analysis**
*(Precision vs. Recall)*

**Takeaway: Corpus Size Matters**
- \( k100 \) yields 13% higher \( R \)-precision than 1/5th its size
- \( k100 \) yields 53% higher \( R \)-precision than 1/25th its size

**Takeaway: Corpus Quality Matters**
- Wikipedia yields similar performance to a web corpus 60 times its size

**Opportunity: Model a more natural SeeLEx usage**
- What is the performance of the system on typical sets searched by editors?

---

**k100: List Effect**
*Precision vs. Recall*

**Opportunity: Bucket sets to find predictable behaviors**
- Our gold standard sets vary significantly in their expansion performance
- Look into predictability of open vs. closed class sets, large vs. small class sets, and types of sets such as locations, people, …
The Effect of:
Seed Selection

Takeaway: Seed set composition greatly affects performance
- Best performing seed set had 42% higher precision and 39% higher recall than the worst performing seed set

Opportunity: Reject seeds and/or ask for more
- We are investigating ways to automatically detect which seed elements are better than others in order to reduce the impact of seed selection effect
The Effect of:
Seed Size

Takeaway: Small numbers of seeds are sufficient to fully saturate the distributional similarity model.
- After 10-20 seeds, more seeds yield few new correct instances in the expansion.
Takeaway: Although *bad* seeds may be less desirable in a seed set than others, adding them does not seem to degrade performance.
- Percentage of errors does not increase with increased seed set size.

Questions Asked and Conclusions

What is the effect of corpus size on expansion accuracy?
- Significant boost in performance

What is the effect of corpus quality on expansion accuracy?
- Significant boost in performance

Does seed quality matter?
- Great variance in performance depending on set seed composition

How many seeds are on average optimal for expansion accuracy and are more seeds better than less?
- Somewhat surprising: 5-20 seeds in general is sufficient; more seeds gains little but don’t hurt
TODAY'S MENU

***

Thank You

***

mew patrickpaultelcom

Yahoo! Labs