Using very simple statistics for review search: An exploration

(Or, surprising results in sentiment analysis for a very knowledge-lean approach)

Bo Pang, Yahoo! Research
Lillian Lee, Cornell University
There never was in the world two opinions alike, no more than two hairs, or two grains; the most universal quality is diversity.
— Montaigne, Essays

Where an opinion is general, it is usually correct.
— Austen, Mansfield Park
real user query seeking reviews

- food review contra costa county

- Contra Costa County
  - Environmental Health - Contra Costa Health Services
    - New and Updated Information for Food Facility Plan Review ... Contra Costa County home page. Contra Costa County, California, USA ...
    - www.cchealth.org/groups/eh/ - 12k - Cached - Similar pages - Filter
  
  Contra Costa County is my new home... Walnut Creek
  - Contra Costa County is my new home... See All Lists - Contra Costa County is my new home... The food is OK but mainly the place is unique for Walnut Creek ...
    - www.yelp.com/list_details/list_id=7MgG8AXpDuWZqLSr0HPCeG - 35k - Cached - Similar pages - Filter
  
  Contra Costa County Homeless Program in Concord, CA, 94520 ...
  - 0 Reviews - Business Details on Contra Costa County Homeless Program.
    - Services: Reuse, Clothing, Edible, Food Donations, Fabric, Recycle, Textiles ...
      - www.mojopages.com/biz/contra-costa-county-homeless-p/concord/ca/94520/5690455 - 72k - Cached - Similar pages - Filter
  
  Contra Costa County Restaurant Guide and Menus - Dining out in ...
  - Dining in Contra Costa County, California. Check out user reviews and menus ...
      - www.dineview.com/search.fwx?zone=00002&cat=R&ord=-44k - Cached - Similar pages - Filter
Approach used by most TREC Blog-track systems:
[Overviews: Ounis et al. 2006, Macdonald et al. 2007]

Stage 1: Perform topic-based retrieval

Stage 2: Re-rank results for subjectivity using pre-compiled lexicons or labeled training data

Q: Can we re-rank without either resource?

- Intellectually interesting
- Could enhance domain independence
  (query topics vary wildly)
Supporting hypothesis (H1):

1. Assume Stage 1 \( \Rightarrow \) the retrieved documents (\( = \) the search set) are all relevant to the query topic.

2. Opinions and their expressions differ; objective documents discuss the same aspects of the query topic.

This suggests re-ranking the search set by idiosyncrasy, which requires no information sources outside the search set (test data) itself.
Data

We used real user queries from an online query log. [the KDDCup 2005 data, www.sigkdd.org/kdd2005/dkkcup/KDDCUPData.zip]

- We selected among those with the word “review” or “reviews” (indicative of review search).

Search sets = top 20 Yahoo! search results per query.

12 annotators in total assigned subjective/objective labels to the documents in 69 search sets.

Corpus available soon at www.cs.cornell.edu/home/llee/data/search-subj.html
Asides on annotation:

- Annotators performed 4-way doc. classification:
  (1) single review; (2) multiple reviews; (3) subjective/objective mixture; (4) objective or “sales pitch” (not a useful review).
  ▶ (1)-(3) were collapsed into “subjective”.

- Avg. pairwise agreement per search set: 88%; minimum agreement: 75%; avg. Kappa: .73

Search-set documents were presented in random order; the annotators were all tech-savvy frequent Web searchers; almost everyone had 2 search sets in common with another annotator; detailed instructions and an example were provided; etc.
Instantiation of Hypothesis 1 (H1):

*Idiosyncratic* ≈ document $d$’s terms are relatively rare within search set $ss$.

So, define a term $t$’s *rarity* as its within-$ss$ IDF:

$$\text{Rarity}_{ss}(t) \overset{def}{=} \frac{1}{\# \{t \text{ occurs in } ss\}},$$

and the *idiosyncrasy* of $d$ as the average rarity of its 100 terms most commonly occurring in $ss$.

We use terms that are $ss$-frequent to focus on topic-relevant terms and to avoid noise (e.g., many mentions of site-specific info). Stopwords, plus words with $ss$-doc-frequency $\leq 3$ for fair baseline comparison, are also deleted. Variant definitions yield qualitatively similar results.
Comparison algorithms

- The original Yahoo! ranking
- **Percentage of adjectives**
  - Simple form of pre-compiled subjectivity lexicon
    - [Hatzivassiloglou & Wiebe '00, Wiebe et al. '04]
- **OpinionFinder** [Riloff & Wiebe '03, Wiebe & Riloff '05]
  - State-of-the-art system using pre-compiled knowledge sources and trained classifiers
  - Applied independently on TREC Blog data by He et al. ['08]
Results (1): High idiosyncrasy does about the same as adjective percentage, worse than OpinionFinder.

All outperform the initial search-engine ranking.

<table>
<thead>
<tr>
<th></th>
<th>p@1</th>
<th>p@2</th>
<th>p@3</th>
<th>p@4</th>
<th>p@5</th>
<th>p@10</th>
<th>p@S</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search engine</td>
<td>.536</td>
<td>.543</td>
<td>.541</td>
<td>.554</td>
<td>.554</td>
<td>.528</td>
<td>.538</td>
<td>.612</td>
</tr>
<tr>
<td>% of adjectives</td>
<td>.710</td>
<td>.703</td>
<td>.696</td>
<td>.681</td>
<td>.678</td>
<td>.625</td>
<td>.633</td>
<td>.715</td>
</tr>
<tr>
<td>OpinionFinder</td>
<td>.754</td>
<td>.717</td>
<td>.729</td>
<td>.725</td>
<td>.733</td>
<td>.675</td>
<td>.690</td>
<td>.768</td>
</tr>
<tr>
<td>High idio.</td>
<td>.739</td>
<td>.703</td>
<td>.696</td>
<td>.696</td>
<td>.678</td>
<td>.606</td>
<td>.633</td>
<td>.711</td>
</tr>
</tbody>
</table>

p@n: precision at n; S: # of subjective documents.

Our intuition failed us ...
Competing hypothesis (H2): Reviews on the same topic tend to all discuss (the same) salient attributes, even if they evaluate these attributes differently.

This suggests re-ranking the search set, lowest idiosyncrasy first.
Results (2): **Low idiosyncrasy** is very competitive with OpinionFinder.

<table>
<thead>
<tr>
<th></th>
<th>p@1</th>
<th>p@2</th>
<th>p@3</th>
<th>p@4</th>
<th>p@5</th>
<th>p@10</th>
<th>p@S</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search engine</td>
<td>.536</td>
<td>.543</td>
<td>.541</td>
<td>.554</td>
<td>.554</td>
<td>.528</td>
<td>.538</td>
<td>.612</td>
</tr>
<tr>
<td>% of adjectives</td>
<td>.710</td>
<td>.703</td>
<td>.696</td>
<td>.681</td>
<td>.678</td>
<td>.625</td>
<td>.633</td>
<td>.715</td>
</tr>
<tr>
<td>OpinionFinder</td>
<td><strong>.754</strong></td>
<td><strong>.717</strong></td>
<td><strong>.729</strong></td>
<td><strong>.725</strong></td>
<td><strong>.733</strong></td>
<td><strong>.675</strong></td>
<td><strong>.690</strong></td>
<td><strong>.768</strong></td>
</tr>
<tr>
<td>LOW idio.</td>
<td><strong>.754</strong></td>
<td><strong>.783</strong></td>
<td><strong>.768</strong></td>
<td><strong>.739</strong></td>
<td><strong>.716</strong></td>
<td><strong>.630</strong></td>
<td><strong>.665</strong></td>
<td><strong>.743</strong></td>
</tr>
</tbody>
</table>

Only the p@10 difference between OF and low idio. is statistically significant idiosyncrasy (paired t-test, .05 level). Different parameter settings for low idiosyncrasy yield p@1 as high as .797.
The performance of OpinionFinder, which has access to training data and pre-compiled lexicons, can be matched using search-set statistics alone.

Next steps:

- Parameter selection?
- Combine with OpinionFinder?
- Comparison to pseudo-feedback?