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

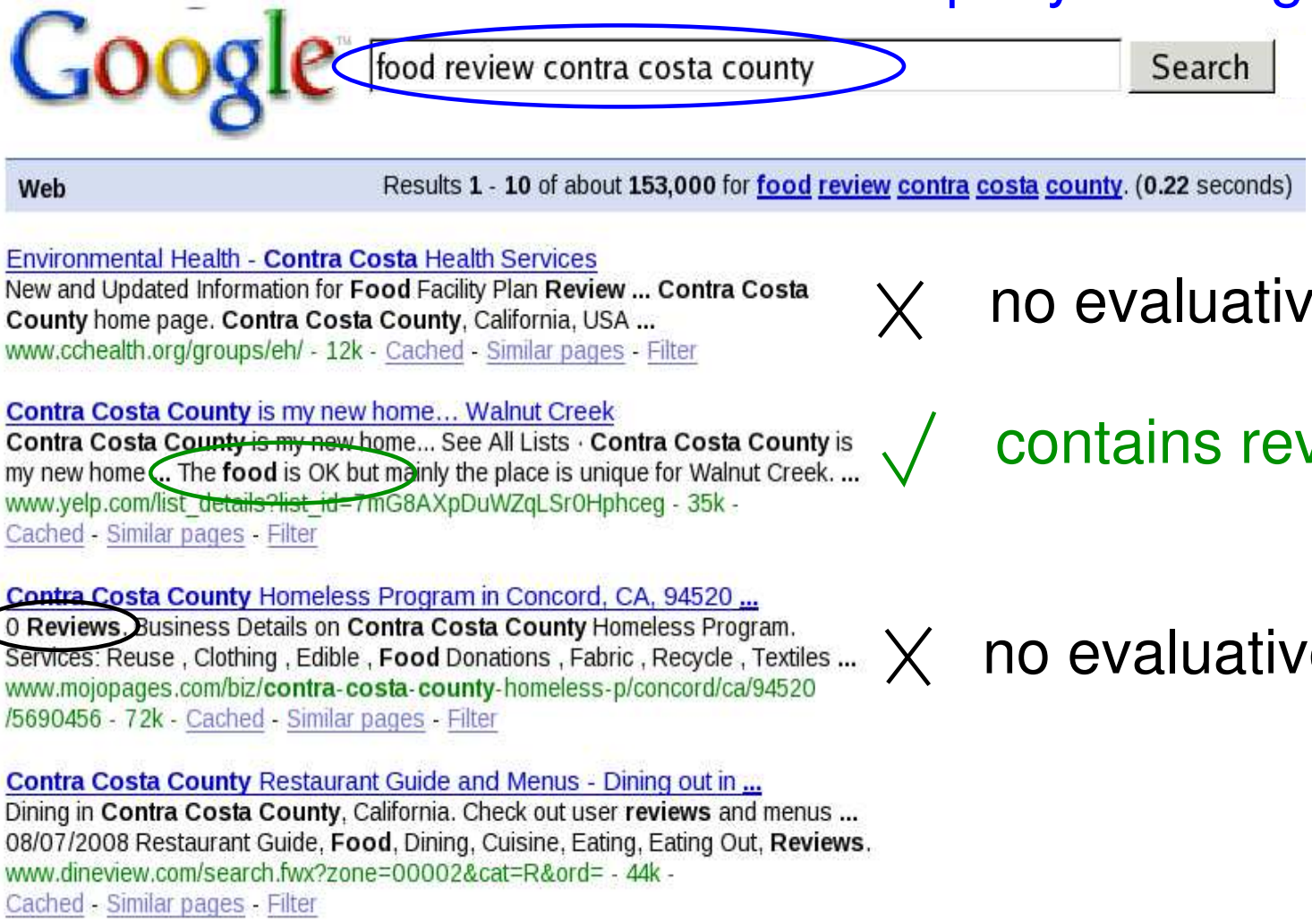
There never was in the world two opinions alike, no more than two hairs, or two grains; the most universal quality is diversity.

— Montaigne, *Essays*

Where an opinion is general, it is usually correct.

— Austen, *Mansfield Park*

real user query seeking reviews



✗ no evaluative text

✓ contains review

✗ no evaluative text

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 $\approx$ document d's terms are relatively rare within search set ss.

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

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

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 ≤ 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.

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

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.

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

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?