Amazon Product Reviews

Categorization and Visualization

Ha Nguyen – Huy Nguyen
Computer science department
Introduction

• Customer reviews in online retailing:
  • Different from traditional brick-and-mortar stores
  • Insights about the product
  • Decide to buy or not

• Problems:
  • Read all reviews: exhausting task
  • Read only top reviews: not complete picture
  • Review sentiment may be not sentence sentiment
Solution

- Visualization system:
  - Extract & Visualize ALL reviews by topics
  - Topics are learn from data
  - Indicate topic-sentence sentiment
  - Feed back to improve the system
  - Deployed as a plugin of browser
Great Camera
★★★★★ 2008/01/03
5/6 people find this helpful

[-] These are not point & shoot cameras even though they do have auto functions. [+]

Samsung GX-10
★★★★★ 2007/12/28
3/5 people find this helpful

[+] Samsung GX-10 is an excellent DSLR (if not one of the best) on the market, technologically and value-wise. [+]

Based on Pentax technology, this modified DSLR is full of technical options only available in higher-priced cameras.
Related works

• Opinion space
  • Review as a point in space, good overview of product
  • Our hierarchical solution is for an easier way to browse reviews

• Review summarization
  • Feature-based, e.g. camera = picture quality + screen + battery …
  • Word counting, e.g. pair of adjective-noun
  • Our topic-based categorization is expanded for out-of-product attributes, e.g. customer service

• Opinion mining
  • Feature-based opinion mining, count +/- keywords for product features
  • Our approach is simpler in the sense that using only +/- ngrams
Work flow

• Review retrieval:
  • Perl scripts to crawl product pages
  • Amazon web service

• Features learning from training dataset:
  • Star rating (review sentiment), helpfulness ratio
  • Learn review topics
  • Learn positive / negative ngrams

• Review computation
  • Infer topics of new review
  • For each topic extract the most representative sentences
  • Compute sentiment score of those sentences
Components

• Visualization
  • Bubble Tree Library (JavaScript, JQuery)
  • Interactive visualization of hierarchical data

• Opinion mining
  • Retrieve 1-2star and 4-5star reviews to build training data
  • Learn positive and negative ngram with sentiment score

• Topic modeling
  • Run LDA on ~500 products to learn 20 topics
  • Record top 100 words of each topic for topic labeling
  • Record word probabilities given topic to compute topic score for sentence
Opinion Mining

- Collect 1, 2, 3-grams from two sets of review
  - 1, 2-star reviews for the set of negative ngrams
  - 4, 5-star reviews for the set of positive ngrams
  - In each set, sort ngrams by frequency
  - For each individual ngram, compute difference of ranks on 2 lists
    - Sentiment_score = rank_difference \times frequency\_in\_list
      - Each individual ngrams has 2 scores

- Aggregating ngrams’ sentiment_score for sentiment score of sentence
Topic Modeling

- Run Latent Dirichlet Allocation (LDA) on reviews of 500 products with 20 topics (> 33.5K reviews)
  - Each topic, record top-100 words with their probabilities
  - Each top word is represented by a vector of 20 probabilities
  - Probabilities < mean are set to 0
  - Final list of triples <word, topic, prob.>

- Compute 20 topic-scores for each review sentence
  - Return the topic has the largest score
  - Segment review into topic-sentences
Data and user study

- Camera products on Amazon
- Conduct user study to evaluate
  - Advantage of hierarchical representation over linear method
  - Accuracy of topic modeling and opinion mining
- Compare to Amazon traditional review list
- Different versions of review browsing
  - Categorize on star-rating
  - Categorize on topic then star-rating
- Sentence sentiment
  - Indicate or not
- Measure on user satisfaction
Conclusion

• Product review exploring
  • Learn topics and sentiments from data
  • Infer topic score and sentiment score for each review sentences
  • For HCI course, NLP techniques play supporting role

• Main purpose:
  • Product review summarizing -> an application for product study, and market research,
  • Exploit all possible information

• How to make it a feasible NLP project
  • Topic labeling
  • Standard evaluation for inferred topic/sentiment at sentence level
Thank You!
Demo