Mining Reviews for
Product Comparison and Recommendation

Dr. Minlie Huang

Dept. Computer Science, Tsinghua University
aihuang@tsinghua.edu.cn
Outline

• Background and Motivation
• System Architecture
• Product Comparison
  – Subjective perspective
  – Objective perspective
  – Overall perspective
• Product Recommendation
  – Evolution Tree
• Experiments
• System Demo
Background

• Web 2.0 times
  – Blogs, RSS, VSN, Forums, ...

• Ecommerce
  – Amazon, Alibaba, ...

• Express opinions, share opinions
  – Epinions: www.epinions.com
  – IMDB: movie reviews
  – Many forums associate with specific products
Motivation

• Product Reviews
  – Users may concern very specific features of a product, such as
    • The picture of a digital camera
    • The weight of a mobile phone
  – There is no comparison or recommendation system (based on reviews) for products.

• Users have to read through all the product reviews to:
  – catch contents with mentions of user experience
  – seek product comparison or recommendation suggestions
Our System

• Present comparison and recommendation at the feature level
  – For example:
    • “Whose screen is better, Sony T200 or Canon G9?”
    • ”What products can be recommended whose screen is as good as Sony T200’s?”
System Architecture

Figure: System Architecture
Feature Extraction

• N-Gram Feature (N<=3)
  – 1-Gram: “picture”, “screen”, “software”
  – 2-Gram: “memory card”, “picture quality”
  – 3-Gram: “red eye reduction”

• Feature Selection [2]
  – Frequently occurred N-Grams are selected as candidate feature words.
    • Noun phrase
    • The same NP-Chunker
    • Not containing stop-words
  – Various filtering methods are used on the candidates
    • Higher frequency
    • Higher Chi-square score between different product types
## Product Feature for “Digital Camera”

<table>
<thead>
<tr>
<th>[product]</th>
<th>[quality]</th>
<th>[display]</th>
<th>[shot]</th>
<th>[special]</th>
<th>[parts]</th>
<th>[service]</th>
</tr>
</thead>
<tbody>
<tr>
<td>camera</td>
<td>digit camera</td>
<td>canon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>size</td>
<td>pocket</td>
<td>canon camera</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>color</td>
<td></td>
<td>canon powershot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price</td>
<td>price rang</td>
<td>monei</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[quality]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>picture</td>
<td>pictur</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>video</td>
<td>video</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>display</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[shot]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zoom</td>
<td>optic zoom</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shutter</td>
<td>shutter speed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flash</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>focus</td>
<td></td>
<td>auto focu</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[special]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[parts]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>battery</td>
<td>batteri life</td>
<td>recharg batteaa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>batteri</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[service]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>customer service</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Mining Feature-Opinion Pairs

• Dependency Grammar Graph [3]
  – “Feature” keywords are used to retrieve sentences.
  – “Opinion” words have pos tags of “ADJ” “ADV” or “ADV ADJ”.
  – Finding dependency path from “Feature” to “Opinion” in the sentence.
Feature-level Product Assessment

• Using Sentiwordnet [4] to evaluate the sentimental strength of an “opinion” word.
  – Generated by WordNet [5].
  – Each word has a sentiment score (positive score, negative score)

• Evaluating the sentiment score of a “feature”
  – Integrating all the sentiment scores of “opinion” words related to “feature” by Feature-Opinion-Pairs (FOP).
  – Weighted voting method

\[
feature\_sentiment = \sum_{\text{opinion\_weight}_k \geq 2} \text{opinion\_weight}_k \times \text{opinion\_sentiment}_k
\]

*(opinion \_weight represents the frequency this FOP occurs)*
Product Comparison

• Subjective Perspective
  – Mining from reviews
  – Focusing on user experiences

• Objective Perspective
  – Incorporating technical details (physical parameters)
  – Reliable but not always comparable

• Overall Perspective
  – Rating products comprehensively according to review contents
Subjective Perspective

Compare various products based on “feature” sentiment.

- **rating**: Sony T200 < Canon G9 < Canon SD750, based on Amazon [6] ratings.
- **size**: T200 and SD750 are with slim vertical design, while SD750 is not.
- **price**: SD750 ($184.49) is so cheap compared to its relatively high performance. G9($614) is a new product, with 12.1MP and 6x image-stabilized optical zoom, whose price is worth its performance. T200 ($599.44) is too expensive even if people don’t care about the size.
- **picture, zoom, focus**: almost the same
- **display**: T200 and SD750 have bigger LCD screen, with clear display.
Objective Perspective

• Limitations of subjective perspective
  – Too subjective: sometimes make no sense.
  – Lack of user comments on some features.

• Incorporating technical details

<table>
<thead>
<tr>
<th></th>
<th>Sony T200</th>
<th>Canon G9</th>
<th>Canon SD500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture Revolution</td>
<td>8.1MP</td>
<td>12.1MP</td>
<td>7.1MP</td>
</tr>
<tr>
<td>Zoom</td>
<td>5</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>
Overall perspective

• Sentiment polarities in reviews
  – traditional definition
    • positive vs. negative
  • 5 star scheme
  – our definition
    • rating distribution
Analyze Content of Reviews

• Sentiment-orientation based approach
  – predefined “sentiment words”
  – disadvantages
    × Can not capture all expressions of subjectivity

• Our approach
  – all the content words
  – advantages
    ✓ weight sentiment ratings according to all contents
    ✓ capture general usage patterns
Word-Sentiment Model

- Generative Topic Model [7]

- Dirichlet distribution
- Z: topic variable
- W: word
- S: rating (1-5 star)
Generative Process

- Sample $\theta$ from Dirichlet distribution: $\theta \sim \text{Dir}(\alpha)$.
- For each word $w_n$, $n \in \{1, \cdots, N\}$
  - Sample topic factor $z_n$ from multinomial distribution:
    $z_n \sim \text{Mul}(\theta)$
  - Sample $w_n$ from multinomial distribution conditioned on $z_n$: $w_n \sim p(w_n \mid \beta, z_n)$
- For each sentiment rating $s_m$, $m \in \{1, \cdots, M\}$:
  - Sample indexing variable $y_m$ from a uniform distribution:
    $y_m \sim \text{Unif}(1, \cdots, N)$
  - Sample $s_m$ from multinomial distribution conditioned on $z_{y_m}$: $s_m \sim p(s_m \mid z_{y_m}, \beta)$.
Evolution Tree

• When we propose recommendation, we need to suggest the products with the same generation.
  – Generation can be defined by various standards.
    • Time?
  – But must be significant in the product development.
    • Mobile phone: GSM → 3G
    • Memory card: SDRAM → DDR → DDR2 → DDR3

• Evolution tree can provide a graphical representation
  – Informative
  – Intuitive
  – ...
Product Recommendation

• We construct a simple evolution tree by
  – Product selling period extracted from reviews
  – Product full name (rule based)
  – Important technical details:
    • Digital Camera: picture resolution, zoom

• Recommendation procedure
  – Select all the products who have the same generation with the current product.
  – Keep the winners in comparison to the current product.
  – Rank the recommended products.
Evolution Tree: an Example

Figure: Evolution tree of digital camera
Experiment Results

• Posterior Inference

\[ p(s \mid w, \alpha, \beta, \eta) \approx \sum_{z_n} q(z_n \mid \phi_n) p(s \mid \varepsilon, z_n) \]
## Experiment Results

- **Rating Classification**
  - Accuracy

- **Ranking**
  - Normalized Discounted Cumulative Gain

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy Comparison on 4-Stars Rating Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.35</td>
</tr>
<tr>
<td>KNN</td>
<td>0.4</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.45</td>
</tr>
<tr>
<td>WSModel</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Truncation Level</th>
<th>Ranking Comparison on mp3 Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 5</td>
<td>SVM, kNN, Naive Bayes, WSModel</td>
</tr>
<tr>
<td>Top 10</td>
<td>SVM, kNN, Naive Bayes, WSModel</td>
</tr>
<tr>
<td>Top 15</td>
<td>SVM, kNN, Naive Bayes, WSModel</td>
</tr>
<tr>
<td>Top 20</td>
<td>SVM, kNN, Naive Bayes, WSModel</td>
</tr>
</tbody>
</table>

### Graphs

- **NDCG Chart**
  - **Top 5, Top 10, Top 15, Top 20**
  - **Dataset Comparison**
  - **Algorithm Comparison** (SVM, kNN, Naive Bayes, WSModel)
Experiment Results

• Review Summary

  – By WSMModel

  – \( p(w | \alpha, \beta, \eta) \approx \sum_{z_n} q(z_n | \phi_n) p(w | \beta, z_n) \)

<table>
<thead>
<tr>
<th>Product</th>
<th>Description Words (TOP 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minox ecx sub-miniature 8x 11mm camera</td>
<td>small, size, pocket, light, carry, design, hand, easily, choice, reasonable, image,</td>
</tr>
<tr>
<td>Olympus stylus epic 35mm camera</td>
<td>lens, great, shoot, zoom, flash, small, film, buy, good, stylus, best</td>
</tr>
<tr>
<td>Polaroid silver blue joycam instant camera</td>
<td>polaroid, fun, film, instant, buy, easy, product, photo, expensive, color, price</td>
</tr>
</tbody>
</table>
Summary

• Comparing product from subjective (opinion) and objective perspectives
• Simulating the rating system from the overall perspective (the topic generative system)
• Recommending products by evolution tree
Demo System

**PCR System** is a review mining system aiming at implementing product comparison and recommendation on the feature level.

The system has the following functions:

- Querying a product's review summary on each feature. The reviews are from [Amazon](http://www.amazon.com) and partly from other websites. *Please tell me some user's comments on Canon G9's feature.*

- Comparing two products on the feature level, and telling customers such as: *Whose screen is better, Sony T200 or Canon G9.*

- Recommending similar products to the current one when given the feature standard. *What products can be recommended whose screen is as good as Sony T200?*

- Generating products' evolution tree.

**PCR System** has employed many innovative methods and algorithms, including natural language processing, Feature-Opinion-Pairs Extraction, feature sentiment summarization. For detailed descriptions of these algorithms, please [contact us](http://166.111.138.87:2008).

Website: [http://166.111.138.87:2008](http://166.111.138.87:2008)
References


Thank you!

Any questions?