Mining and Summarizing Customer Reviews

Minqing Hu and Bing Liu

Data Mining and Exploration

Instructor: Dr. Amos Storkey

Students:

Marios Bartzokas

Joern Wuebker

Elisavet Koutsoudaki

Problem

- Products sold online receive numerous customer reviews
- Problem:
  - Reviews may be long
  - The may have only a few sentences containing opinions of the product

-> Difficult for a potential customer to get an enlightening idea by reading the reviews – biased view

-> Difficult for the manufacturer to keep track of the customer opinions
  - the same product may be sold in many sites
  - a manufacturer produces many kinds of products

Need for mining and summarizing all the customer reviews of a product

Mining customer reviews

- **Traditional text summarization**
  - Select a subset of the original text and rewrite it so as to capture the basic points
  - Free text documents summaries

- **Approach in the paper**
  - Mine only the specific product features on which customers have expressed their opinion
  - Capture only negative or positive opinion
  - Structured summaries (*feature-based summaries*)

Steps of the mining task

1. Identify features of the products that customers have expressed their opinions on. (make use of both data mining and natural language processing techniques)

2. For each feature, identify review opinion sentences and decide whether they are positive or negative.

3. Produce a summary using the discovered information

Steps of the mining task

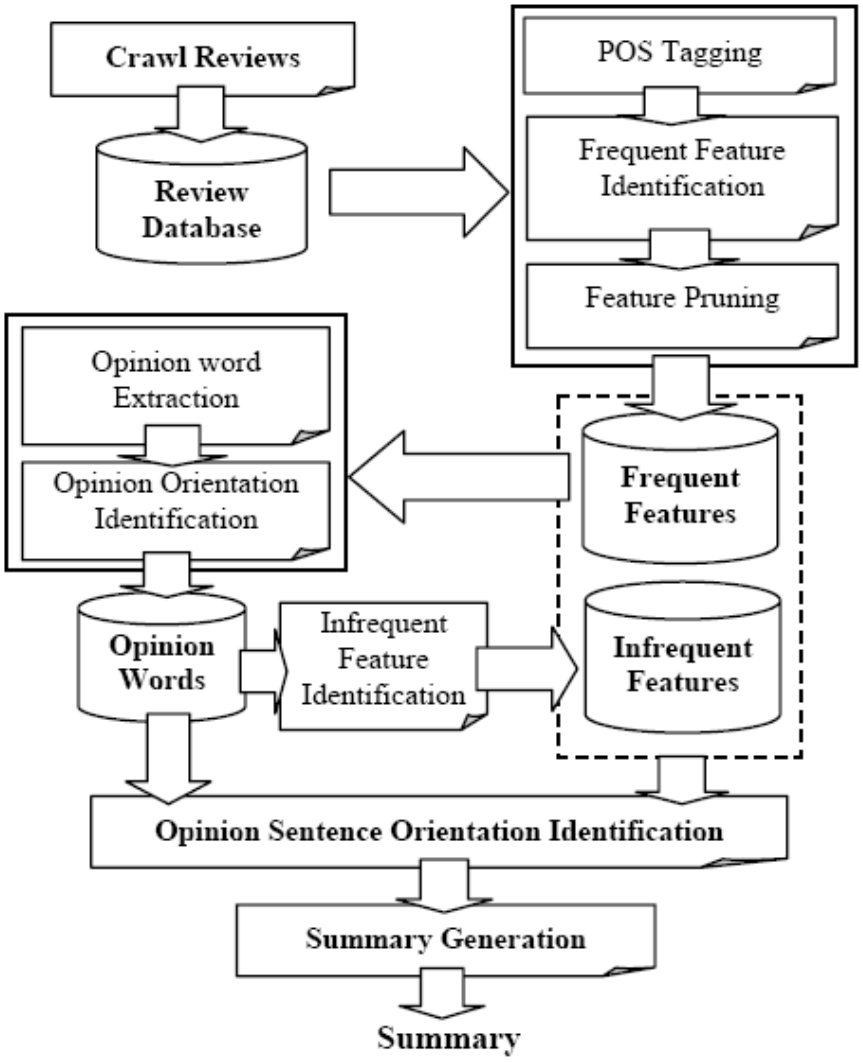
- Feature based summary example: digital camera

*Digital_camera_*1:
Feature: **picture quality**
Positive: 253
<individual review sentences>
Negative: 6
<individual review sentences>
Feature: **size**
Positive: 134
<individual review sentences>
Negative: 10
<individual review sentences>
…

Deciding feature orientation

- Identify a set of adjective words (*opinion words*)
  - Use a natural language processing method
- For each opinion word determine its semantic orientation; positive or negative
  - A bootstrapping technique using WordNet is proposed
- Decide the opinion orientation of each sentence
  - An efficient algorithm is proposed

The opinion summarization system



Opinion summarization system

- Input:
  - A product name
  - An entry web page for all the reviews of the product
- Output:
  - Summary
- Process:
  1. Download reviews and store in the database
  2. Find frequent features
  3. Extract opinion words using these frequent features
  4. Identify semantic orientations of the opinion words using WordNet.
  5. Find infrequent features, using the extracted opinion words.
  6. Identify the orientation of each opinion sentence
  7. Produce the final summary

(1) Part of speech tagging (POS)

- Split input review into sentences and produce part of speech tag for each word

- Use of NLProcessor linguistic parser
  - generates XML output

```
<S> <NG><W C='PRP' L='SS' T='w' S='Y'> I </W> </NG>
<VG> <W C='VBP'> am </W><W C='RB'> absolutely
</W></VG> <W C='IN'> in </W> <NG> <W C='NN'> awe
</W> </NG> <W C='IN'> of </W> <NG> <W C='DT'> this
</W> <W C='NN'> camera </W></NG><W C='.'> .
</W></S>
```

- Save output in database

- Transaction file
  - Noun and noun phrases of each sentence
  - Removing stopwords, stemming, fuzzy matching

(2) Frequent features identification (I)

- Produces features on which many people have expressed their opinions on
- Difficulty in natural language understanding -> difficult to deal with
- *Cases supported: only when feature words are present*
  - *Yes : "The pictures are very clear"*
  - *No: "While light, it will not easily fit in pockets"* *size*
- *Use association mining (association rules)*
  - Itemset: set of words or phrases in sentences
  - Why? Words describing word features converge among reviews -> frequent itemsets likely to be product features

(2) Frequent features identification (II)

- *CBA association miner*
  - Apriori algorithm, with 1% minimum support
  - Generate candidate frequent itemsets

- Compactness pruning
  - Remove candidate frequent itemsets whose words do not appear together in a specific order
  - Why? Less likely to be meaningful phrases

- Redundancy pruning
  - Prune based on min p-support (set to 3% in paper)
  - p-support: the number of sentences that a feature appears in as a noun or a noun phrase, while these sentences must contain no other features being a superset of this feature

Opinion Word Extraction

- **Goal**: predict orientation of opinion sentences
- Opinion sentence: expresses **subjective opinion**
- Presence of **adjectives** indicates a subjective sentence
  - => adjectives as opinion words
- ***Limit extraction to sentences containing product features***

Opinion Word Extraction

- **Effective opinion** for a frequent feature:
  - The nearest adjective

- Example:
  - *The strap is horrible and gets in the way of parts of the camera you need access to.*

Orientation Identification

- **Goal**: identify **semantic orientation** of opinion words
  - Possible orientations: *positive* and *negative*
- Utilise adjective **synonym and antonym set** in WordNet
  - Organised in bipolar clusters:

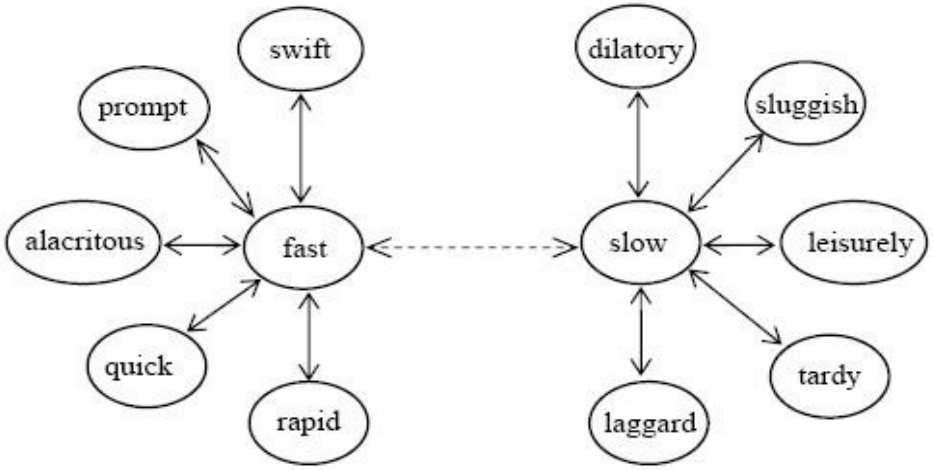


Figure 4: Bipolar adjective structure,
(→ = similarity; ⇢ = antonymy)

Orientation Identification

- Synonyms have the same, antonyms opposite orientations
- Predict orientation by **searching** sets of **synonyms** and **antonyms** for words of known orientation
- => **seed adjectives** needed
  - Example: *great, fantastic, nice, bad, dull*
- *Adjectives are added to the seed list once orientation has been predicted*
- *Iterate over list of adjectives until no new predictions*
- *Discard any left over adjectives (user notified)*

Orientation Identification

- Algorithm:

```
while(word had been added to seed list)
    for(all words w in adjective list)
        search synonym and antonym sets of all
                words in seed list
        if(found): assign orientation
    end for
end while
```

Infrequent Feature Identification

- Infrequent feature: feature only few people comment on
- **Utilise opinion words**
- Example:
  - *The pictures are absolutely amazing.*
  - *The software that comes with it is amazing.*
- => *different features, same opinion word*

Infrequent Feature Identification

- *Algorithm:*
  - *For all sentences **containing opinion words but no frequent features**:*
    - *Add nearest noun/noun phrase to feature list*
- *Problem: method finds irrelevant features*
- *Low support in the data*
  - *=> low rating*
  - *=> shows up at the bottom of summary*

Predicting Orientation of Opinion Sentences

- Sentence orientation is determined by **predominant opinion word orientation** in the sentence (case 1)
- If it is a tie, number of **effective opinions** counts (case 2)
- If **negation word** nearby, twist word orientation
- 'but' clause indicates sentimental change
  - Use effective opinion inside the clause, otherwise opposite orientation to the main sentence.

Predicting Orientation of Opinion Sentences

- Two main cases are dealt with:
  - The user **likes/dislikes most or all of the features** in a sentence
    - Example: *Overall this is a good camera with a really good picture clarity & an exceptional close-up shooting capability.*
  - *The user likes/dislikes most of the features in one sentence, but **equal number of pos. and neg. opinion words***
    - *Example: The auto and manual along with movie modes are very easy to use, but the software is not intuitive.*
- *Otherwise **same orientation** as the sentence before*

Summary Generation

Feature: **picture**

Positive: 12

- Overall this is a good camera with a really good picture clarity.
- The pictures are absolutely amazing - the camera captures the minutest of details.
- After nearly 800 pictures I have found that this camera takes incredible pictures.

…

Negative: 2

- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

...

Experimental Evaluation
Feature-Based Summarization System

1. The effectiveness of feature extraction.

2. The effectiveness of opinion sentence extraction.

3. The accuracy of orientation prediction of opinion sentences.

Experimental Evaluation
Datasets

Amazon.com and C|net.com (large # of reviews):

- 2 Digital cameras
- 1 DVD player
- 1 mp3 player
- 1 Cellular phone

Information apart from the title and the text was discarded

- e.g. date, ratings, location.

Experimental Evaluation
Procedure (I)

- ∀ ***product download first 100 reviews***
- Remove html tags
- NLProcessor generated the part-of-speech tags
- FBS system is applied

Experimental Evaluation
Procedure (II)

- Manually read the reviews
- Tag features
- Identify opinions orientation
- ∀ product produce manually feature list
- Compare FBS system results with manual results

Experimental Evaluation Issues

- Feature tagging: explicit or implicit in a sentence

  FBS system will not recognize the implicit feature opinion sentences

- Opinion subjectivity => two taggers used for those cases

  e.g. “The picture quality is ok, but it could have been better”

Experimental Evaluation
Feature Extraction

Recall and precision at each step of feature generation

Product name	No. of manual features	Frequent features (association mining)		Compactness pruning		P-support pruning		Infrequent feature identification	
		Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Digital camera1	79	0.671	0.552	0.658	0.634	0.658	0.825	0.822	0.747
Digital camera2	96	0.594	0.594	0.594	0.679	0.594	0.781	0.792	0.710
Cellular phone	67	0.731	0.563	0.716	0.676	0.716	0.828	0.761	0.718
Mp3 player	57	0.652	0.573	0.652	0.683	0.652	0.754	0.818	0.692
DVD player	49	0.754	0.531	0.754	0.634	0.754	0.765	0.797	0.743
Average	**69**	**0.68**	**0.56**	**0.67**	**0.66**	**0.67**	**0.79**	**0.80**	**0.72**

Experimental Evaluation
Compare with FASTR (Christian Jacquemin)

FASTR: term extraction & indexing system

Recall and precision of FASTR

	Recall	Precision	No. terms
Digital camera1	0.1898	0.0313	479
Digital camera2	0.1875	0.0442	407
Cellular phone	0.1493	0.0275	364
Mp3 player	0.1403	0.0214	374
DVD player	0.1633	0.0305	262
Average	**0.1660**	**0.0309**	**377.2**

Experimental Evaluation
Sentence Extraction & Orientation

Results of opinion sentence extraction and sentence orientation prediction

Product name	Opinion sentence extraction		Sentence orientation accuracy
	Recall	Precision	
Digital camera1	0.719	0.643	0.927
Digital camera2	0.634	0.554	0.946
Cellular phone	0.675	0.815	0.764
Mp3 player	0.784	0.589	0.842
DVD player	0.653	0.607	0.730
Average	**0.693**	**0.642**	**0.842**

Experimental Evaluation Limitations

- Opinion sentences that need pronoun resolution are not encountered

  e.g. "it is quiet but powerful"

- Verbs & nouns are not used for sentence orientation

  e.g. "I like the feeling of the camera"

  "I highly recommend the camera"

- Opinions strength is not studied

Summary

- Task: Feature-Based Summarization of reviews
- Methods

  1) Identify features

     Part-of-Speech tagging
     Frequent features identification
     Infrequent feature extraction

  2) Identify review opinion sentences and their orientation.

     Opinion words extraction
     Orientation identification for opinion words
     Prediction of the orientation of opinion sentences

  3) Summary generation

- Evaluation

Questions?