Score and display the features of opinion mining

Luole Qi
Supervisor: Dr. Li Chen

Department of Computer Science
HongKong Baptist University

The Postgraduate Research Symposium, 2010

Outline

1. Introduction
   - What’s opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

Outline

1. Introduction
   - What’s opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

- Two main types of textual information
  - Facts and opinions
  - Much of the existing research on text information processing has been (almost exclusively) focused on factual information.
- Opinions
  - One can express personal experiences and opinions on almost anything, at review sites, forums, discussion groups, blogs... (called the user generated content.)
  - The usual format of Opinion is review.
- Definition of Opinion Mining :
  opinion mining aims to extract attributes and components of the object from comments and to determine whether the comments are positive, negative or neutral.
  E.g. The image (feature) is incredible (opinion words)(positive).

Outline

1. Introduction
   - What’s opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

Introduction

Why we need opinion mining ?

- Many e-commerce websites allow users to write reviews
  - (such as Amazon.com, Yahoo !Shopping, etc...)
- Whenever you need to make a decision, you may want some opinions from others,
  - It's hard to find the result by traditional search engine ! (e.g. CANON VS NIKON)
- Going over all reviews costs much time.
- good to both customers and manufacturers !

Outline

1. Introduction
   - What’s opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

- How to extract features of a product from reviews ?
- How to determine whether the opinions on the features are positive, negative or neutral ?
- How to summarize the result of opining mining ?

Feature-based opinion mining and summarization

(Hu and Liu, KDD-04)

- They use Association rule mining to find out the features.
- Their mining method is based on Apriori algorithm.
- They determine the polarity of opinion words using wordnet.
- They score the features and visualize the result.

Red Opal : product-feature scoring from reviews

(Christopher Scaffidi et. al, EC-07)

- They use a language model approach with the assumption that product features are mentioned more often in a product review than they are mentioned in generic English to find out the features.
- They use lemma-frequency data derived from a 100 million-word corpus of spoken and written conversational English.
- They score each product on each feature of products.
- Their system is fully automatic and not restricted to specific product categories

Other related works

- Mining tag clouds and emoticons behind community feedback(WWW'08 Kavita A. Ganesan, Neelakantan Sundaresan)
- A multimedia interface for facilitating comparisons of opinions(IUI'09 Giuseppe Carenini, Lucas Rizoli)
- Movie Review Mining and Summarization(CIKM'06 Zhuang, L., Jing, F)
- Using PMI, syntactic relations and other attributes with SVM (EMNLP'04 Mullen and Collier)
- Comparing supervised and unsupervised methods(HICSS'05 Chaovalit and Zhou)
- Many others...

Outline

1 Introduction
- What’s opinion mining ?
- Why we need opinion mining ?
- Research questions

2 Related Work

3 Proposed Method
- Review Format
- The flow of our method
- Features extraction
- Opinion word extraction
- Scoring features
- Tagcloud Generation

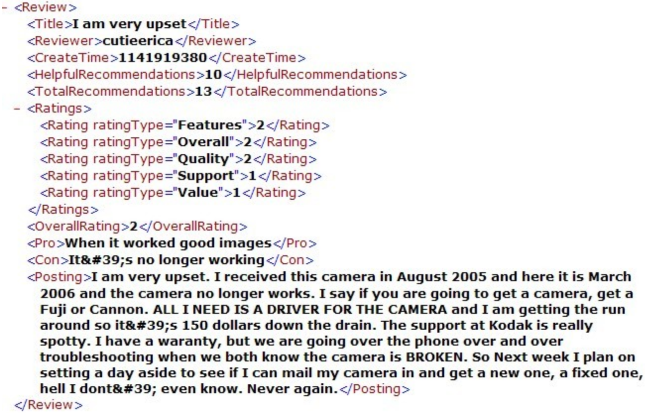
4 Conclusion
- Our contributions
- Future Works

Three basic review formats

- Format 1 - Pros, Cons and detailed review :The reviewer is asked to describe Pros and Cons separately and also write a detailed review.
  (e.g. Epinions.com, Yahoo !Shopping)
- Format 2 - Pros and Cons :The reviewer is asked to describe Pros and Cons separately.
  (e.g. Cnet.com)
- Format 3 - free format :The reviewer can write freely, i.e., no separation of Pros and Cons.
  (e.g. Amazon.com)

Yahoo Shopping reviews

- We use Yahoo !shopping reviews as our dataset(XML File)
- It’s a semi-structures format



```
- <Review>
    <Title>I am very upset</Title>
    <Reviewer>cutieerica</Reviewer>
    <CreateTime>1141919380</CreateTime>
    <HelpfulRecommendations>10</HelpfulRecommendations>
    <TotalRecommendations>13</TotalRecommendations>
  - <Ratings>
      <Rating ratingType="Features">2</Rating>
      <Rating ratingType="Overall">2</Rating>
      <Rating ratingType="Quality">2</Rating>
      <Rating ratingType="Support">1</Rating>
      <Rating ratingType="Value">1</Rating>
    </Ratings>
    <OverallRating>2</OverallRating>
    <Pro>When it worked good images</Pro>
    <Con>It&#39;s no longer working</Con>
    <Posting>I am very upset. I received this camera in August 2005 and here it is March
      2006 and the camera no longer works. I say if you are going to get a camera, get a
      Fuji or Cannon. ALL I NEED IS A DRIVER FOR THE CAMERA and I am getting the run
      around so it&#39;s 150 dollars down the drain. The support at Kodak is really
      spotty. I have a waranty, but we are going over the phone over and over
      troubleshooting when we both know the camera is BROKEN. So Next week I plan on
      setting a day aside to see if I can mail my camera in and get a new one, a fixed one,
      hell I dont&#39; even know. Never again.</Posting>
  </Review>
```

Outline

1. Introduction
   - What's opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

Flow chart

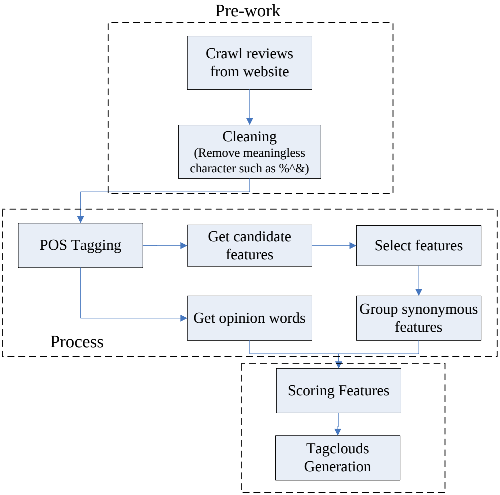


FIG.: the whole process of our algorithm

Outline

1. Introduction
   - What's opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

Features extraction

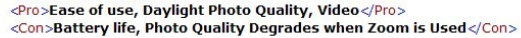
P-O-S Tagging

- Obeservation : Nearly all the features are noun words or phrases. (Liu and Hu 2004 KDD)
- The task of POS tagging is the process of marking up the words in a text (corpus) as corresponding to a particular part-of-speech, such as noun and verb
- For example, "The image is incredible." from one review.
  =>
  (DT The) (NN image) (VBZ is) (JJ incredible) (. .)

Features extraction

Select features

- Focus on pros and cons parts !



<Pro>Ease of use, Daylight Photo Quality, Video</Pro>
<Con>Battery life, Photo Quality Degrades when Zoom is Used</Con>

  - simple sentences or some words segments
  - kind of like the summary of post part by users
  - the opinion polarity is known
- The words tagged with (NN) and (NNS) are selected as features
- If there are less than three consecutive noun words, they will be seen as a noun phrase as a whole.
  (E.g. (NN Battery) (NN life) => Battery life.

Outline

1. Introduction
   - What’s opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

Opinion word extraction

- Definition (Word Step) :The distance between any two words. The step of two consecutive words is 1.
- For each feature, any adjective word within 3 steps will be selected as opinion words.
- if there is a colon, stop or any other punctuations within 3 steps which is more near the feature word, the adjective words can not be seen as an opinion word
  (The image is good, nice high ISO.)

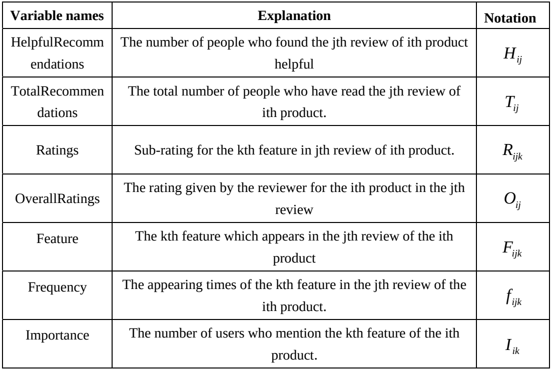
Outline

1. Introduction
   - What's opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

Scoring features

Definition of some variables

- We score features based on the semi-structured review from yahoo shopping.



Variable names	Explanation	Notation
HelpfulRecommendations	The number of people who found the jth review of ith product helpful	H_{ij}
TotalRecommendations	The total number of people who have read the jth review of ith product.	T_{ij}
Ratings	Sub-rating for the kth feature in jth review of ith product.	R_{ijk}
OverallRatings	The rating given by the reviewer for the ith product in the jth review	O_{ij}
Feature	The kth feature which appears in the jth review of the ith product	F_{ijk}
Frequency	The appearing times of the kth feature in the jth review of the ith product.	f_{ijk}
Importance	The number of users who mention the kth feature of the ith product.	I_{ik}

Scoring features

- We consider all the factors into our score formula.
- We consider the impact of the HelpfulRecommendations and TotalRecommendations.

$$\boxed{\prod_{j=1}^{N_i}[(\ln(f_{ijk})\cdot(1+\frac{H_{ij}}{T_{ij}})]}$$

- We consider the impact of overall rating value and sub-rating values.

Impact of Overall rating	Impact of Sub-ratings
$\frac{O_{ij}}{\frac{\sum_{j=1}^{N_i}O_{ij}}{N_i}}=\frac{N_iO_{ij}}{\sum_{j=1}^{N_i}O_{ij}}$	$\frac{R_{ijk}}{\frac{\sum_{j=1}^{N_i}R_{ijk}}{N_i}}=\frac{N_iR_{ijk}}{\sum_{j=1}^{N_i}R_{ijk}}$

- The final scoring formula :

$$\boxed{F_{ijk}=(1+\frac{I_{ik}}{N_i})\cdot\prod_{j=1}^{N_i}[(\frac{N_iR_{ijk}}{\sum_{j=1}^{N_i}R_{ijk}})\cdot(1+\frac{f_{ijk}}{\max(f_{ijk})})\cdot(1+\frac{H_{ij}}{T_{ij}})\cdot\frac{N_iO_{ij}}{\sum_{j=1}^{N_i}O_{ij}}]}$$

Outline

1 Introduction
- What's opinion mining ?
- Why we need opinion mining ?
- Research questions

2 Related Work

3 Proposed Method
- Review Format
- The flow of our method
- Features extraction
- Opinion word extraction
- Scoring features
- Tagcloud Generation

4 Conclusion
- Our contributions
- Future Works

Generation method

- Normalize the score of each feature to the range from the smallest fontsize to the biggest fontsize.

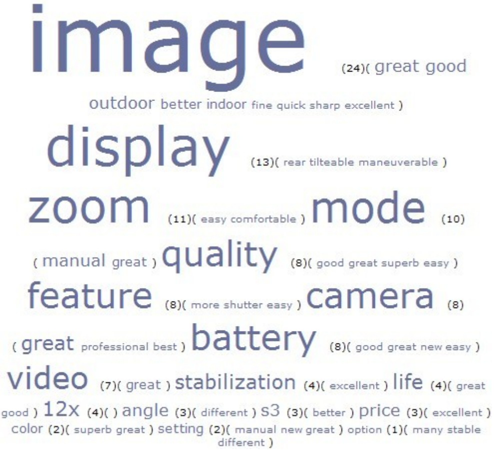


image (24)(great good
outdoor better indoor fine quick sharp excellent)
display (13)(rear tilteable maneuverable)
zoom (11)(easy comfortable) mode (10)
(manual great) quality (8)(good great superb easy)
feature (8)(more shutter easy) camera (8)
(great professional best) battery (8)(good great new easy)
video (7)(great) stabilization (4)(excellent) life (4)(great
good) 12x (4)( ) angle (3)(different) S3 (3)(better) price (3)(excellent)
color (2)(superb great) setting (2)(manual new great) option (1)(many stable
different)

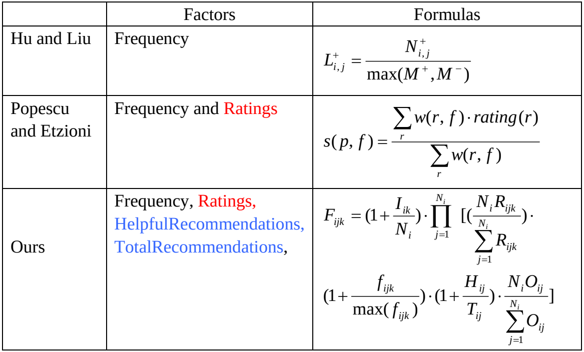
Outline

1. Introduction
   - What's opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

Contribution 1

Comparison of different approaches

- We score features considering more factors based on the semi-structured review from yahoo shopping.

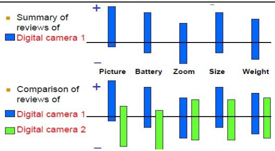
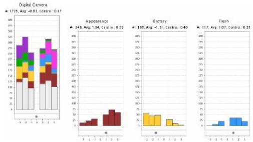
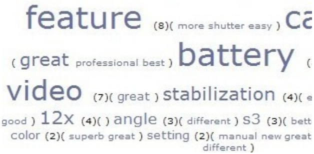


	Factors	Formulas
Hu and Liu	Frequency	$L^{+}_{i,j} = \frac{N^{+}_{i,j}}{\max(M^{+}, M^{-})}$
Popescu and Etzioni	Frequency and Ratings	$s(p, f) = \frac{\sum_{r} w(r, f) \cdot rating(r)}{\sum_{r} w(r, f)}$
Ours	Frequency, Ratings, HelpfulRecommendations, TotalRecommendations,	$F_{ijk} = (1 + \frac{I_{ik}}{N_i}) \cdot \prod_{j=1}^{N_i} [(\frac{N_i R_{ijk}}{\sum_{j=1}^{N_i} R_{ijk}}) \cdot (1 + \frac{f_{ijk}}{\max(f_{ijk})}) \cdot (1 + \frac{H_{ij}}{T_{ij}}) \cdot \frac{N_i O_{ij}}{\sum_{j=1}^{N_i} O_{ij}}]$

Contribution 2

Comparison of different interfaces

- We use a more straightforward way to make the summary based on the scores of features.

	Summary method
Hu and Liu	
Giuseppe Carenini and Lucas Rizoli	
Ours	 feature (8)(more shutter easy) ca (great professional best) battery (video (7)(great) stabilization (4)(e good) 12x (4)( ) angle (3)(different) S3 (3)(bett color (2)(superb great) setting (2)(manual new great different)

Outline

1. Introduction
   - What’s opinion mining ?
   - Why we need opinion mining ?
   - Research questions
2. Related Work
3. Proposed Method
   - Review Format
   - The flow of our method
   - Features extraction
   - Opinion word extraction
   - Scoring features
   - Tagcloud Generation
4. Conclusion
   - Our contributions
   - Future Works

Future Works

- Evaluate our systems based on the criteria adopted by Hu and Liu.
- Conduct an user study to evaluate our system and investigate the users' needs.

- Thank you !