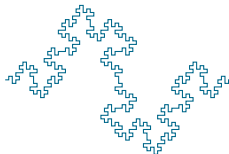


Review Mining

Automatically Assessing Review Helpfulness

Sanae Sato
Haotian He



April 22, 2014

OVERVIEW

OVERVIEW

INTRODUCTION

The Issue

Goals for This Issue

METHODOLOGY

Define Helpfulness

Ranking

Features

Evaluation

RESULTS

Results

Summary

THE ISSUE

- ▶ Online reviews vary in quality
- ▶ Current ranking of reviews is only by their recency or product rating, other than assessing relevance according to their text reviews
- ▶ "Helpfulness" is very relevant information which directly affects customers' decision making, but the challenge is that it's also hard to define and measure what exactly it is

GOALS FOR THIS ISSUE

- ▶ A system for automatically ranking reviews according to helpfulness
- ▶ An analysis of different classes of features most important to capture review helpfulness (structural, lexical, syntactic, semantic, and meta-data)

DEFINE HELPFULNESS

Formally, given a set of reviews R for a particular product, our task is to rank the reviews according to their *helpfulness*. They define a review *helpfulness* function h , as:

$$h(r \in R) = \frac{\text{rating}_+(r)}{\text{rating}_+(r) + \text{rating}_-(r)}$$

Data: Amazon.com reviews for particular electronics products obtained by using Amazon Web Services API.

APPROACH

Ranking System

SVM regression model and RBF kernel to estimate function h .

Why choose SVM regression, rather than SVM ranking?

APPROACH

Choose Features

What features may affect the assessment of review helpfulness?

APPROACH

Features

Feature Class: Structural Feature

- ▶ Length (LEN)
- ▶ Sentential (SEN)
- ▶ HTML (HTM)

APPROACH

Features

Feature Class: Lexical Feature

- ▶ Unigram (UGR)
- ▶ Bigram (BGR)

APPROACH

Feature Extraction

Feature Class: Syntactic Feature

- ▶ Syntax (SYN)

APPROACH

Features

Feature Class: Semantic Feature

- ▶ Product-Feature (PRF)
- ▶ General-Inquirer (GIW)

APPROACH

Features

Feature Class: Meta-data Feature

- ▶ Stars (STR/STR1/STR2)

APPROACH

Feature Extraction

For LEN/SEN/UGR/BGR/SYN:

- ▶ Minipar dependency parser (Lin 1994)
- ▶ Parser tokenization
- ▶ Sentence Breaker
- ▶ Syntactic categorization

APPROACH

Feature Extraction

For PRF:

- ▶ Developed an automatic way of mining reference to product features
- ▶ Basic approach: turn user generated pros/cons list found in Epinions.com into a feature list based on the assumption that pros/cons list tend to contain references to the product features that are important
- ▶ number of unique Product-Feature

APPROACH

Feature Extraction

For GIW:

- ▶ Extract sentiment words using General Inquirer Dictionary

APPROACH

Feature Extraction

For STR:

- ▶ Directly created from the star rating

APPROACH

Evaluation

- ▶ Gold Standard: Labeled dataset $\{review, h(review)\}$ for supervised machine learning
- ▶ Spearman correlation coefficient
- ▶ Person coefficient

RESULTS

Table 1. Sample of 4 out of 43 reviews for the *iPod Photo 20GB* product from Amazon.com along with their ratings as well as their helpfulness ranks (from both the gold standard from Amazon.com and the SVM prediction of our best performing system described in Section 5.2).

REVIEW TITLE	HELPFUL VOTES	UNHELPFUL VOTES	RANK(h)	
			GOLD STANDARD	SVM PREDICTION
"iPod Moves to All-color Line-up"	215	11	7	1
"iPod: It's NOT Music to My Ears"	11	13	25	30
"The best thing I ever bought"	22	32	26	27
"VERY disappointing"	1	18	40	40

RESULTS

Table 3. Evaluation of the feature combinations that make up our best performing system (in bold), for ranking reviews of Amazon.com *MP3 Players* and *Digital Cameras* according to *helpfulness*.

FEATURE COMBINATIONS	MP3 PLAYERS		DIGITAL CAMERAS	
	SPEARMAN [†]	PEARSON [†]	SPEARMAN [†]	PEARSON [†]
LEN	0.575 ± 0.037	0.391 ± 0.038	0.521 ± 0.029	0.357 ± 0.029
UGR	0.593 ± 0.036	0.398 ± 0.038	0.499 ± 0.025	0.328 ± 0.029
STR1	0.589 ± 0.034	0.326 ± 0.038	0.507 ± 0.029	0.266 ± 0.030
UGR+STR1	0.644 ± 0.033	0.436 ± 0.038	0.490 ± 0.032	0.324 ± 0.032
LEN+UGR	0.582 ± 0.036	0.401 ± 0.038	0.553 ± 0.028	0.394 ± 0.029
LEN+STR1	0.652 ± 0.033	0.470 ± 0.038	0.577 ± 0.029	0.423 ± 0.031
LEN+UGR+STR1	0.656 ± 0.033	0.476 ± 0.038	0.595 ± 0.028	0.442 ± 0.031

LEN=*Length*; UGR=*Unigram*; STR=*Stars*

[†]95% confidence bounds are calculated using 10-fold cross-validation.

RESULTS

FEATURE COMBINATIONS	MP3 PLAYERS		DIGITAL CAMERAS	
	SPEARMAN [†]	PEARSON [†]	SPEARMAN [†]	PEARSON [†]
UGR	0.593 ± 0.036	0.398 ± 0.038	0.499 ± 0.025	0.328 ± 0.029
BGR	0.499 ± 0.040	0.293 ± 0.038	0.434 ± 0.032	0.242 ± 0.029
PRF	0.591 ± 0.037	0.400 ± 0.039	0.527 ± 0.030	0.316 ± 0.028
GIW	0.571 ± 0.036	0.381 ± 0.038	0.524 ± 0.030	0.333 ± 0.028
UGR+PRF	0.570 ± 0.037	0.375 ± 0.038	0.546 ± 0.029	0.348 ± 0.028
UGR+GIW	0.554 ± 0.037	0.358 ± 0.038	0.568 ± 0.031	0.324 ± 0.029
STR1	0.589 ± 0.034	0.326 ± 0.038	0.507 ± 0.029	0.266 ± 0.030
STR2	0.556 ± 0.032	0.303 ± 0.038	0.504 ± 0.027	0.229 ± 0.027
LEN+UGR+STR1	0.656 ± 0.033	0.476 ± 0.038	0.595 ± 0.028	0.442 ± 0.031
LEN+UGR+STR1+SEN	0.653 ± 0.033	0.470 ± 0.038	0.599 ± 0.028	0.448 ± 0.030
LEN+UGR+STR1+HTM	0.640 ± 0.035	0.459 ± 0.039	0.594 ± 0.028	0.442 ± 0.031
LEN+UGR+STR1+SYN	0.645 ± 0.034	0.469 ± 0.039	0.595 ± 0.028	0.447 ± 0.030
LEN+UGR+STR1+SEN+HTM+SYN	0.631 ± 0.035	0.453 ± 0.039	0.600 ± 0.028	0.452 ± 0.030
LEN+UGR+STR1+SEN+HTM+SYN+PRF+GIW	0.601 ± 0.035	0.396 ± 0.038	0.604 ± 0.027	0.460 ± 0.030

LEN=Length; SEN=Sentential; HTM=HTML; UGR=Unigram; BGR=Bigram;

SYN=Syntax; PRF=Product-Feature; GIW=General-Inquirer; STR=Stars

[†]95% confidence bounds are calculated using 10-fold cross-validation.

SUMMARY

- ▶ A system for automatically ranking reviews according to helpfulness

They successfully assessed helpfulness and ranking reviews according to it. SVM regression suits and works well to learn the helpfulness function for their purpose. Compared with Gold Standard, the results shows a good match, as Spearman correlation coefficient scores of 0.656 (MP3) and 0.604 (digital cameras) against the gold standard.

SUMMARY

- ▶ An analysis of different classes of features most important to capture review helpfulness (structural, lexical, syntactic, semantic, and meta-data)

The top three significant features:

- ▶ Length of the review
- ▶ Unigram (UGR)
- ▶ Product rating

Semantic/sentiment features were subsumed by the simple unigram features. Structural feature except length and syntactic feature had no significant impact.