Hierarchical Organization of Unstructured Consumer Reviews

Jianxing Yu, Zheng-Jun Zha, Meng Wang, Tat-Seng Chua
School of Computing
National University of Singapore
{jianxing, zhazj, wangm, chuats}@comp.nus.edu.sg

ABSTRACT

In this paper, we propose to organize the aspects of a specific product into a hierarchy by simultaneously taking advantages of domain structure knowledge as well as consumer reviews. Based on the derived hierarchy, we generate a hierarchical organization of the consumer reviews based on various aspects of the product, and aggregate consumer opinions on the aspects. With such hierarchical organization, people can easily grasp the overview of consumer reviews and opinions on various aspects, as well as seek consumer reviews and opinions on any specific aspect by navigating through the hierarchy. We conduct evaluation on two product review data sets: Liu et al.'s data set containing 314 reviews for five products [2], and our review corpus which is collected from forum Web sites containing 60,786 reviews for five popular products. The experimental results demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Mining

General Terms

Algorithms, Experimentation

Keywords

Product Aspect Hierarchy, Consumer Review Organization

1. INTRODUCTION

With the rapid development of e-commerce, huge collections of consumer reviews containing opinions are now available on the Web. These reviews have become an important resource for both consumers and firms. Consumers commonly seek quality information from online reviews before making purchase decisions. Accordingly, many firms take online reviews into consideration in production development, marketing etc. However, the online reviews of a specific product are usually unstructured and containing hundreds of aspects of the product. It is impractical for people to grasp the overview of consumer reviews and opinions on various aspects of the product from such enormous reviews. It is also inefficient for people to aggregate and browse the reviews and opinions on a specific aspect. To address the above problems, in this paper, we propose to organize the various aspects of a product into an aspect hierarchy, based on which we generate a hierarchical organization of consumer reviews and opinions on these aspects. Given a product,

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

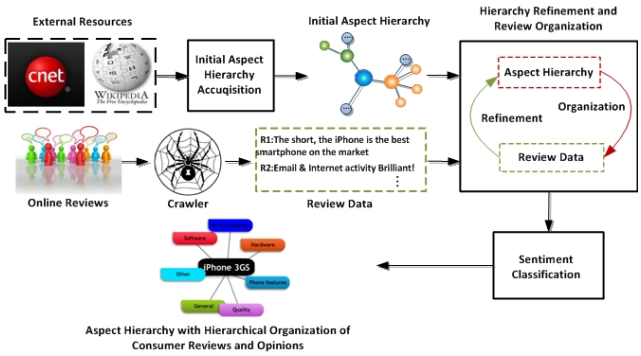

Figure 1: Flowchart of Our Approach

we first automatically construct an initial aspect hierarchy based on the domain structure knowledge mined from product selling Web sites, such as CNet.com. We then iteratively cluster the consumer reviews into corresponding aspects in the hierarchy, and refine the hierarchy by taking advantage of the characteristics of review clusters. Afterwards, we employ sentiment classification technique to determine consumer opinions on the aspects. As a result, we generate an aspect hierarchy, as well as the hierarchical organization of the consumer reviews and opinions on the aspects. With such hierarchical organization, people can easily grasp the overview of consumer reviews and opinions on various aspects of a product, as well as navigate through the hierarchy to explore consumer reviews and opinions on a specific aspect.

2. APPROACH

As illustrated in Figure 1, our approach contains three main components: (a) initial aspect hierarchy acquisition, (b) hierarchy refinement and review organization, and (c) sentiment classification.

2.1 Initial Aspect Hierarchy Acquisition

Domain structure knowledge of product aspects is now available on the Web. For example, there are more than 248,474 product specifications in the product site CNet.com [1]. The specification of the product provides coarse-grained parent-child relationships between some aspects of the product. This structure knowledge is useful to help organize product aspects into a hierarchy. Here we employ the approach proposed in [5] to automatically acquire an initial aspect hierarchy from the product specifications.

2.2 Hierarchy Refinement and Review Organization

In this component, we leverage the initial aspect hierarchy to assist aspect hierarchy generation and review organization. The ini-

tial aspect hierarchy usually does not fit the review data well. For example, some aspects that are commented in the reviews may not be covered in the hierarchy, while some aspects in the hierarchy may not be users' interests in the reviews. Therefore, we propose to refine the initial hierarchy by taking into consideration the review data. By iteratively organizing the reviews into the hierarchy and refining the hierarchy, we will obtain a refined hierarchy which can fit the review data well and is able to generate better review organization.

In particular, in the stage of review organization, we utilize the noun terms and sentiment terms as the discriminative features to represent the reviews into vectors. Similarly, we represent the aspect terms and related aspect description terms in the node as vectors and view them as the initial centroids of the nodes. We cluster the reviews into relevant aspect nodes in the hierarchy based on cosine similarity. Since the reviews may contain multiple aspects, we segment them into several parts by using the approach in [6]. Each of the segmented reviews covers at most one aspect and belongs to the unique node in the hierarchy. In the hierarchy refinement stage, there are three manipulations: node pruning, node splitting, and node assignment handling. In node pruning, we delete the aspects which are not talked about in the reviews. In node splitting, we split the node whose review cluster is not homogeneous into sub-clusters, and derive the frequent noun phases from the sub-clusters as their aspects. In node assignment handling, we create an "*Other*" node as the descendent of the root to cache the reviews which are dissimilar to all the nodes. We iteratively conduct review organization and hierarchy refinement. As a result, a refined aspect hierarchy can be obtained, and the reviews with the same aspects, which are either explicit or implicit, will be clustered into the corresponding aspect nodes in the hierarchy. Thus, we can identify the implicit as well as the explicit aspects of the reviews, while traditional frequency-based or rule-based methods [2] fail to extract the implicit aspects.

2.3 Sentiment Classification

We leverage the *Pros* and *Cons* reviews on the Web as the training data to build a *SVM* sentiment classifier. We utilize the unigram as feature and represent the reviews into vectors with Boolean weighting. As the review may contain some non-sentiment sentences, in the preprocessing part, we use the sentiment lexicon provided by *MPQA* project [3] to filter out those not contained the sentiment terms and view them as the non-sentiment sentences.

3. EXPERIMENTS

Product Name	*Review#*	*Sentence#*
Canon Powershot G3 (Canon G3)	45	597
Nikon coolpix 4300 (Nikon 4300)	34	346
Nokia 6610	41	740
Creative Nomad Jukebox Zen Xtra (Nomad)	95	1,716
Apex AD2600 DVD player (Apex AD2600)	99	546
iPhone 3GS	12,096	43,031
Nokia N95	15,939	44,379
Nokia 5800 XpressMusic (Nokia 5800)	28,129	75,001
BlackBerry Bold 9700 (BlackBerry)	4,070	11,008
Apple MacBook Pro (MacBook Pro)	552	4,221

Table 1: Statistic of the data sets. The top five product review corpus are Liu et al. data set and the other five product review corpus are our data collection. Sentence# denotes the number of sentences in the reviews

We conduct evaluation on two product review data sets: Liu et al.'s data set [2], and our review corpus collected from four forum Web sites including cnet.com, viewpoints.com, gsmarena.com and reevoo.com. Liu et al.'s data set contains 314 reviews for five products, while our review collection contains 60,786 reviews for another five popular products (see Table 1). Figure 2 shows the performance comparisons of various methods on aspect identification, including Hu's method [2], Wu's method [4], aspect identification based on initial aspect hierarchy, as well as our approach. F_1-measure is adopted as the performance metric. We can see that our approach outperforms the other methods on all the ten product review corpus. For sentiment classification, our *SVM* classifier achieves the average F_1-measure of 0.834 on Liu et al's data set and 0.830 on our data set. Figure 3 illustrates the aspect hierarchy of "*iPhone 3GS*" generated by our approach. While Figure 3 (a) shows the first layer of the aspect hierarchy, Figure 3 (b) illustrates the sub-hierarchy of the aspect "*software*." With such hierarchical organization, we can easily obtain the overview of consumer reviews and opinions on various aspects of the product, as well as navigate through the hierarchy to explore consumer reviews and opinions on a specific aspect.

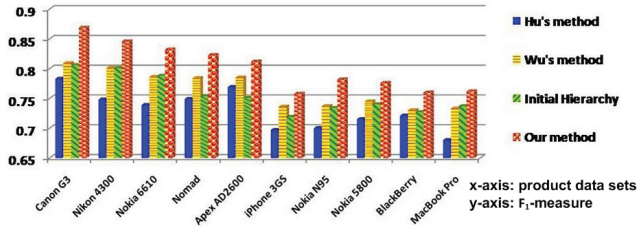

Figure 2: Performance comparisons on aspect identification

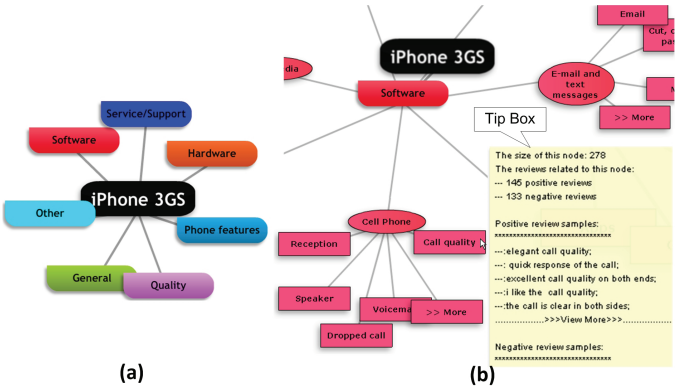

Figure 3: Sample of aspect hierarchy and hierarchical review organization

4. REFERENCES

[1] J. Beckham. The cnet e-commerce data set. *Technical Reports*, 2005.
[2] M. Hu and B. Liu. Mining and summarizing customer reviews. *KDD*, 2004.
[3] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. *HLT/EMNLP*, 2005.
[4] Y. Wu, Q. Zhang, X. Huang, and L. Wu. Phrase dependency parsing for opinion mining. *ACL*, 2009.
[5] S. Ye and T.-S. Chua. Learning object models from semi-structured web documents. *IEEE TKDE*, 2006.
[6] J. Zhu, H. Wang, and B. Tsou. Aspect-based sentence segmentation for sentiment summarization. *TSA*, 2009.