Product Review Summarization from a Deeper Perspective

Duy Khang Ly
National University of Singapore
Computing 1,
13 Computing Drive
Singapore 117417
ldkhang@gmail.com

Kazunari Sugiyama
National University of Singapore
Computing 1,
13 Computing Drive
Singapore 117417
sugiyama@comp.nus.edu.sg

Ziheng Lin
National University of Singapore
Computing 1,
13 Computing Drive
Singapore 117417
linzihen@comp.nus.edu.sg

Min-Yen Kan
National University of Singapore
Computing 1,
13 Computing Drive
Singapore 117417
kanmy@comp.nus.edu.sg

ABSTRACT

With product reviews growing in depth and becoming more numerous, it is growing challenge to acquire a comprehensive understanding of their contents, for both customers and product manufacturers. We built a system that automatically summarizes a large collection of product reviews to generate a concise summary. Importantly, our system not only extracts the review sentiments but also the underlying justification for their opinion. We solve this problem through a novel application of clustering and validate our approach through an empirical study, obtaining good performance as judged by F-measure (the harmonic mean of purity and inverse purity).

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Abstracting methods; I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithms, Experimentation, Languages, Performance

Keywords

Sentiment Analysis, Summarization, Clustering

1. INTRODUCTION

Product reviews have become an important source of information, not only for customers to find opinions about products and voice their comments, but also for producers to understand the feedback on their products. In digital libraries, catalogs have also integrated review content both from curated sources as well as from their patrons. However, this wealth of information also makes it unwieldy; sense making on such a large collection is difficult at best on products with thousands of reviews. At this scale, users and manufacturers are unlikely to read all product reviews, however insightful. To address these issues, we build a product review summarization system that achieves the following two important goals: (1) to efficiently identify topics and subtopics in the reviews (*product facet identification*), and (2) to summarize the corresponding opinions into a coherent summary to users (*summarization*).

Unlike previous approaches, our summary captures opinions from different dimensions of the product. More importantly, it allows a user to quickly see how the reviewers feel about the product, yet equip him with sufficiently detailed information.

2. RELATED WORK

We briefly review the two pertinent areas of sentiment analysis and summarization to lay the groundwork for how our approach differs from convention.

Research on sentiment analysis examines the detection of subjectivity and opinion, and measuring its polarity (positive or negative) and its intensity, in text spans as small as individual words up to as large as entire documents. At the word level, Hu and Liu [6] utilized WordNet [10] to grow a initial seed list of known orientation adjectives into a larger list that covers all the remaining adjectives in WordNet. At the sentence level, Kim and Hovy [7] aggregated the polarity of each individual adjective or sentimental word that appeared in the sentence itself. Their subsequent work introduced additional sentence-surface features (*e.g.*, counts of positive/negative adjectives in a target sentence, or in a sentence window around a target sentence), used in a supervising a learned model for detection [8].

We observe that finding the sentiment polarity of the sentence is insufficient in product reviews. It is necessary to identify the internal semantics of the opinion, as it may describe particular facets of the target product in the review. For example, *battery life*, *lens*, *flash system*, and *price* would be examples of facets that could be discussed in the product category of cameras. In order to address this problem, Ding *et al.* [3] proposed a system that further incorporated a set of complex, carefully-built grammar rules between adjacent sentence construction as well as neighboring facets, together with a collection of comprehensive polarity-annotated lists of idioms, nouns, verbs, adjectives and adverbs.

While the work on sentiment analysis discussed above perform well at delimiting and extracting user opinions in reviews, they do not aggregate these opinions together. We pursue this goal through the use of text summarization. In fact, summarization researchers have examined opinion summarization, even at the facet level. Hu and Liu [6], as well as Popescu and Etzioni [11] attached sentence-

a. Lens

(+): 57 sentences

1. The lens feels very solid!
2. I have taken a whole bunch of excellent pictures with this lens.

. . .

(−): 15 sentences

1. I am not satisfied with the included lens kit.
2. The lens cap is very loose and comes off very easily!

. . .

(a) Output of summary produced by existing systems.

a. Lens

(+) The lens feels very solid! (+10 similar)
(−) I think the lens is not worth it, it's a bit too fragile. (+2 similar)

(+) I have taken a lot of excellent pictures with this lens. (+7 similar)
(−) Don't buy this lens, I always get my pictures blurred. (+0 similar)

. . .

(b) Output of desirable summary that our proposed system aims at.

Figure 1: Comparison of summmaries obtained from (a) existing, and (b) our proposed systems.

level statistics, *i.e.*, the number of positive/negative sentences to facets. Subsequently, Liu *et al.* [9] extended the single facet-driven summary into comparative-based summary among products in the same category, where the orientation of all shared facets are plotted together with their number of supporting sentences for visualization. These works bind opinion polarity detection with individual sentences, but only address the opinions' content minimally. As the input is a set of reviews, multi-document summarization approaches that address content issues are relevant. In this area, Radev and McKeown [12] summarized news from different sources by generating summaries using a template based natural language generation approach, using key information extracted from each source. Their extraction approach allowed values to be extracted only once, preventing redundancy in the output. In another approach, maximum marginal relevance (MMR) [2] creates summaries by choosing new sentences iteratively. It downweights potential summary sentences by the amount of overlap with existing summary sentences.

These multidocument systems all work on general or news domains, and have not been geared towards opinion summarization. This is where we seek to make a contribution, as to the best of our knowledge, no system combines sentiment analysis with multidocument summarization to generate a product review summary.

3. PROPOSED METHOD

We first examine the output of a representative existing product review summarization – Hu and Liu's system [5] – to justify our proposed approach to discover the underlying reasons for users' opinions, As shown in Figure 1, both summaries have their structure based on product facets, in which the facet *Lens* is shown. However, the summary in Figure 1(a) does not attempt to organize the positive or negative sentences beyond their polarity and users will still need to read through the sentences to uncover the actual reasons that justify the positive or negative sentiment. To address this, we propose to generate the summary in Figure 1(b) which further provides a representative reason for the sentiment and clusters other, similar reasons to remove redundancy.

Figure 2 shows an overview of our product review summarization system, which consists of two main components: (1) product facet identification, and subsequent (2) summarization. We describe the two components in turn.

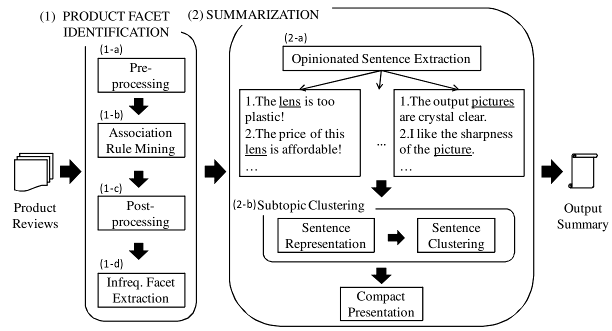

Figure 2: System overview.

(1) Product Facet Identification. To identify candidate facets, we first preprocess the input set of reviews, tagging part-of-speech, stemming and assigning syntactic roles (Figure 2 (1-a)). We utilize Stanford Part-Of-Speech (POS) Tagger[1]. The tagger generally performs fairly well for nouns and noun phrases (the important classes for facet identification), even with the oddly-structured sentences in the reviews. Stopwords are removed and noun[2] are further stemmed[3].

The resulting list of nouns contain facets but also many extraneous, regular nouns. In the existing Hu and Liu system [5], the equivalent list is not filtered further. In contrast, we introduce the use of syntactic role information within a sentence to distinguish genuine facets from noise. We deploy the Stanford Dependency Parser[4] to detect the role of each noun, and discard nouns that do not play a subject or object role. Our method delivers a larger proportion of legitimate nouns to the final two downstream steps.

We use association rule mining [1] (Figure 2 (1-b)) to identify frequent explicit product facets. We run only the first phase of the Apriori method to obtain the set of frequent itemsets (product facets), and concurrently obtain their ranking from their support values. This ranking is a key piece of evidence for use in the summarization module in the second half of the pipeline. We attempt to post-process to remove irrelevant facets incorrectly detected by association rule mining by employing two commonsense heuristics (Figure 2 (1-c)):

· *Usefulness Pruning* targets the removal of meaningless single-word facets. For example, in camera reviews, *life* by itself is not a useful facet, while *battery life* is a meaningful facet. We compute the pure support of a facet f, defined as the number of sentences that f appears alone, without being subsumed by any other facet. When the support number is below a predefined threshold, we drop the single word noun, as it can be described its superset.

· *Compactness Pruning* targets the removal of redundant phrasal facet. For example, *photo pixel*, *sample image* can be replaced by *pixel* and *image*. For each word in a candidate phrase, we compute the ratio of support between the phrase and the individual word. If any of a phrase's ratios is lower than a predefined threshold, we drop it.

Association mining does not discover facets that are infrequent due to their low support. We use a two-step propagation method to try to recover them. We first compile the list of opinion words that modify frequent facets, and then in any sentence that does not con-

[1]http://nlp.stanford.edu/software/tagger.shtml

[2]We use "nouns" to refer to both "nouns" and "noun phrases".

[3]Using Porter stemmer: http://www.tartarus.org/~martin/PorterStemmer/

[4]http://nlp.stanford.edu/software/lex-parser.shtml

tain a facet but does contain an compiled opinion word, the nearest modified noun is included as a facet.

(2) Summarization. For each identify facets, summarization associates it with relevant opinion sentences and selects a representative to be shown for each (positive/negative) polarity. We restrict our algorithm to extract only opinionated sentences from the reviews, as we are only concerned on the users' opinions (Figure 2 (2-a)). We perform sentiment analysis based on Ding *et al.*'s method [3], assigning a polarity score per sentence, calculated as the summed polarity of its constituent words. In this approach, words have polarity if they are on a seed list of known-polarity adjectives, or are connected to a seed list word through synonymous/antonymous relationships.

We calculate content-based pairwise similarities between all resulting opinion sentences, and then cluster them. To compare performance, we tried both hierarchical groupwise-average clustering and the non-hierarchical exchange method [13] (Figure 2 (2-b)). We partition each facet's sentence cluster into a positive part and a negative part, using the sentences' individual polarity score.

The final task is to select the most representative sentence for each partition, which needs to cover as much information in the other sentences. We equate coverage with similarity and choose the partition's centroid sentence that satisfies:

$$argmax(\sum_{s_j \in P-s_i} sim(s_i, s_j)). \qquad (1)$$

This centroid sentence is displayed to users as the exemplar for the facet-polarity combination. In the display, we also include the number of other sentences in the partition (e.g., "+2 similar").

4. EXPERIMENTS

To benchmark our approach, we use publicly available sets of reviews for 3 products (camera, phone, and DVD player) from [5]. The numbers of sentences for each of the products, camera, phone, and DVD player are 160, 139, 111, respectively. In evaluating the product facet identification component, we employ standard precision and recall measures. In evaluating our summarization component, we needed to prepare our own labeled data, consisting of sentences being partitioned into subtopics for a set of 22 facets extracted from the 3 products. The inter-annotator agreement between two annotators was 85%. The final extraction of the data for evaluation that reached both annotators' consensus was 90%. Performance is measured using purity, inverse purity, and F_1-measure (the harmonic mean of purity and inverse purity, weighted equally), widely used for evaluating clustering measures [4].

(1) Results for Product Facet Identification. Tables 1 and 2 compare the results of our implemented version of Hu and Liu's system [5], and the results when we integrate information syntactic roles into the decision, respectively. Table 1 shows that our system can achieve the results reported in [5]. We observe that our system identifies most of the common facets such as: *battery*, *picture*, *lens* for the camera, *signal*, *headset* for the phone and *remote control*, *format* for the DVD player. Table 2 shows that we observe an improvement in precision compared with Table 1 as more noise has been filtered away by the incorporation of syntactic role information. For example, in *Camera*, while the precision in infrequent facet extraction in Table 1 achieves 0.747, the precision in infrequent facet extraction in Table 2 achieves 0.842, showing a significant 0.095 absolute improvement.

(2) Results for Summarization. Table 3 shows the results for the summarization component. We first note that the DVD's facet of $Price$ contains only one cluster. Looking deeper, we find our input reviews for this facet only express opinions about the player's affordability. In such single cluster cases, our system does not improve over the current state-of-the-art. On the other hand, facets having a lot of subtopics (*e.g.*, $Lens$ in $Camera$ (7 subtopics), LCD in $Camera$ (6 subtopics), *etc.*) exhibit many different properties (the size, ease of use, price for the $lens$, or the resolution, material, color for LCD), and users discuss freely on any of these subtopics. In such cases, our system is most beneficial in aligning like-themed comment with each other.

Interestingly, the number of subtopics varies not only from facet to facet, but also from product to product. In our data, the product $Camera$ shows the greatest number ($\sim$ 5 on average), while DVD shows the lowest ($\sim$ 2 on average). This shows that the facets that belong to $Camera$ usually have richer properties, compared with those belonging to the DVD product, which has a impact on the performance of our clustering algorithm.

We compare the performance of our algorithms with a baseline which randomly assigns sentences to clusters. For both the random baseline and the stochastic non-hierarchical clustering approach, we report the average performance over 200 trials. We see that the overall performance of both clustering systems betters the random baseline significantly. On the other hand, we observe small difference in average performance between hierarchical and non-hierarchical approach; although non-hierarchical approach tends to perform better when the number of subtopics is large (*e.g.*, LCD and $Megapixels$ in Camera, $Service$ in Phone), it fares less well on facets where the number of subtopics is small (*e.g.*, $Service$ in DVD). We think that the flat clustering may be less sensitive to larger number of subtopics, as every move or swap operation directly affects the objective function. However, in cases with only a few subtopics, its move and swap operations may result in local minima and cause termination quickly, whereas the hierarchical approach which uses average-link distance may maintain a better balance between clusters.

We have shown that both hierarchical and non-hierarchical clustering outperform the baseline of random clustering in all of three products. However, we observe that this margin decreases when the number of subtopics is reduced. Our further examination shows that this is expected by chance, as with fewer subtopics, the random guess will be correct a larger percentage of time.

5. CONCLUSION

In this work, we have proposed a system that can summarize product reviews, and further organizes the reviews into a structured, extractive summary. A key insight of our work is that product reviews need to be organized further than just at the facet level as even individual facets often consist of subtopics. Our system's summaries go deeper in organizing its summary by aligning user's opinions about different subtopics of a product's facets. In the first component that identifies product facets, we demonstrated that performance can be improved by utilizing syntactic role information within a sentence. In the second summarization component, we employed two clustering methods to identify these subclusters, and further extract a representative compact sentence examplifying sentiment. From our experiments, we conclude that both clustering methods are effective but that a hybrid combination may yield better performance.

Several extensions from our current system are possible. Different brand names that belong to a particular product class (*e.g.*, Nikon, Canon (Camera); Pioneer (DVD); iPod (Music Player), *etc.*), or product/manufacturer names of the accessories that go together with the main product (*e.g.*, Kingston (memory card for cameras), Nvidia (graphic card for computers), *etc.*), are all treated as genuine facets in the annotation from the dataset. However, in most

Table 1: Performance of the product facet identification component – Hu and Liu [5].

Data	Number of manually extracted facets	Association mining		Post processing		Infrequent facet	
		Recall	Precision	Recall	Precision	Recall	Precision
Camera	79	0.671	0.552	0.658	0.825	0.822	0.747
Phone	67	0.731	0.563	0.716	0.828	0.761	0.718
DVD	49	0.754	0.531	0.754	0.765	0.797	0.793
Average	65	0.719	0.549	0.709	0.806	0.793	0.753

Table 2: Performance of "product facet identification" component – Hu and Liu [5] + syntactic role.

Data	Number of manually extracted facets	Association mining		Post processing		Infrequent facet	
		Recall	Precision	Recall	Precision	Recall	Precision
Camera	79	0.671	0.646	0.658	0.894	0.822	0.842
Phone	67	0.731	0.648	0.716	0.903	0.761	0.769
DVD	49	0.754	0.610	0.754	0.818	0.797	0.867
Average	65	0.719	**0.634**	0.709	**0.872**	0.793	**0.826**

Table 3: Summarization component performance.

Data	Facet	Number of manually defined clusters	**Hierarchical clustering**			**Non-hierarchical clustering**			**Random clustering**		
			$Purity$	$I\text{-}Purity$	F_1	$Purity$	$I\text{-}Purity$	F_1	$Purity$	$I\text{-}Purity$	F_1
Camera	Battery	4	0.864	0.591	0.702	0.864	0.636	**0.733**	0.864	0.455	0.596
	Memory	3	0.643	1.000	**0.783**	0.643	0.786	0.707	0.500	0.643	0.563
	Flash	4	0.556	0.722	0.628	0.667	0.722	**0.693**	0.500	0.611	0.550
	LCD	6	0.478	0.826	0.606	0.565	1.000	**0.722**	0.348	0.739	0.473
	Lens	7	0.792	1.000	**0.884**	0.792	1.000	**0.884**	0.500	0.667	0.571
	Megapixels	5	0.621	0.483	0.543	0.724	0.552	**0.626**	0.552	0.414	0.473
	Mode	6	0.813	1.000	**0.897**	0.813	1.000	**0.897**	0.500	0.625	0.556
	Shutter	6	0.643	0.929	**0.760**	0.643	0.929	**0.760**	0.429	0.786	0.555
	Average	5.13	0.676	0.819	0.725	0.714	0.828	**0.753**	0.524	0.617	0.542
Phone	Battery	3	0.824	0.765	**0.793**	0.765	0.706	0.734	0.706	0.588	0.642
	Camera	3	0.727	0.636	**0.679**	0.727	0.636	**0.679**	0.727	0.545	0.623
	Headset	4	0.467	0.733	**0.570**	0.400	0.600	0.480	0.400	0.667	0.500
	Radio	3	0.737	0.737	**0.737**	0.737	0.737	**0.737**	0.737	0.579	0.648
	Service	5	0.438	0.875	0.583	0.563	1.000	**0.720**	0.375	0.625	0.469
	Signal	3	0.824	0.941	**0.878**	0.824	0.765	0.793	0.824	0.588	0.686
	Size	3	0.760	0.680	0.718	0.920	0.680	**0.782**	0.720	0.520	0.604
	Speaker	4	0.684	0.895	**0.775**	0.684	0.789	0.733	0.684	0.632	0.657
	Average	3.50	0.682	0.783	0.717	0.702	0.739	**0.722**	0.647	0.593	0.604
DVD	Price	1	1.000	0.714	0.833	1.000	0.762	**0.865**	1.000	0.524	0.688
	Remote	4	0.625	0.750	**0.682**	0.563	0.750	0.643	0.500	0.688	0.579
	Format	1	1.000	0.714	**0.833**	1.000	0.571	0.727	1.000	0.500	0.667
	Design	1	1.000	1.000	**1.000**	1.000	1.000	**1.000**	1.000	1.000	1.000
	Service	1	1.000	0.739	**0.850**	1.000	0.522	0.686	1.000	0.522	0.686
	Picture	4	0.800	0.850	**0.824**	0.800	0.850	**0.824**	0.450	0.500	0.474
	Average	2.00	0.904	0.795	**0.837**	0.894	0.743	0.791	0.825	0.622	0.682

cases, they appear together with some other facets when comparison is made between that product and its competitors (*e.g.*, "My Canon camera has longer battery life than Nikon"). In certain cases, such entities are better linked to as a separate resource and excluded from the current product's summarization. We leave this as a challenge to future work to build a module that recognizes these proper names and processes them appropriately.

6. REFERENCES

[1] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. of 20th International Conference on Very Large Data Bases (VLDB'94)*, pages 487–499, 1994.

[2] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 335–336, 1998.

[3] X. Ding, B. Liu, and P. S. Yu. A Holistic Lexicon-based Approach to Opinion Mining. In *Proc. of the International Conference on Web Search and Web Data Mining (WSDM'08)*, pages 231–240, 2008.

[4] A. Hotho, A. Nürnberger, and G. Paaß. A Brief Survey of Text Mining. *GLDV-Journal for Computational Linguistics and Language Technology*, 20(1):19–62, 2005.

[5] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, pages 168–177, 2004.

[6] M. Hu and B. Liu. Mining Opinion Features in Customer Reviews. In *Proc. of the 19th National Conference on Artificial Intelligence (AAAI-2004)*, pages 755–760, 2004.

[7] S.-M. Kim and E. Hovy. Determining the Sentiment of Opinions. In *Proc. of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1367–1374, 2004.

[8] S.-M. Kim and E. Hovy. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proc. of the Workshop on Sentiment and Subjectivity in Text co-located with the 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, pages 1–8, 2006.

[9] B. Liu, M. Hu, and J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proc. of the 14th International World Wide Web Conference (WWW2005)*, pages 342–351, 2005.

[10] G. A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[11] A.-M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews. In *Proc. of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 339–346, 2005.

[12] D. R. Radev and K. R. McKeown. Generating Natural Language Summaries from Multiple On-line Sources. *Computational Linguistics*, 24(3):470–500, 1998.

[13] H. Spath. *The Cluster Dissection and Analysis Theory FORTRAN Programs Examples*. Prentice-Hall, Inc., 1985.