AUTOMATIC SUMMARIZATION OF ONLINE CUSTOMER REVIEWS

Jiaming Zhan¹, Han Tong Loh¹ and Ying Liu²
¹Department of Mechanical Engineering, National University of Singapore, Singapore
²Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

Keywords: e-Commerce, customer reviews, multi-document summarization, web mining.

Abstract: Online customer reviews offer valuable information for merchants and potential shoppers in e-Commerce and e-Business. However, even for a single product, the number of reviews often amounts to hundreds or thousands. Thus, summarization of multiple reviews is helpful to extract the important issues that merchants and customers are concerned about. Existing methods of multi-document summarization divide documents into non-overlapping clusters first and then summarize each cluster of documents individually with the assumption that each cluster discusses a single topic. When applied to summarize customer reviews, it is however difficult to determine the number of clusters without the prior domain knowledge, and moreover, topics often overlap with each other in a collection of customer reviews. In this paper, we propose a summarization approach based on the topical structure of multiple customer reviews. Instead of clustering and summarization, our approach extracts topics from a collection of reviews and further ranks the topics based on their frequency. The summary is then generated according to the ranked topics. The evaluation results showed that our approach outperformed the baseline summarization systems, i.e. Copernic summarizer and clustering-summarization, in terms of users’ responsiveness.

1 INTRODUCTION

Nowadays, with the rapid development of e-Commerce and e-Business, it is common that products are sold on the websites such as Amazon.com. Customers are invited to write reviews to share their experiences, comments and recommendations with respect to different products. Also, in modern enterprises, a lot of emails are received from customers every day regarding products and services. These product reviews are valuable for designers and manufacturers to keep track of customers’ feedback and make improvements on their products or services. Moreover, the reviews posted on the World Wide Web (WWW) offer recommendations to potential buyers for their decision making. However, the number of reviews can grow very quickly and it is time-consuming to read through all of them manually. For example, there are hundreds of reviews posted on the web for some popular products in Amazon.com; and thousands of customer emails may be received by the manufacturer regarding one particular product.

Some work has been reported dealing with the vast amount of customer reviews (Hu & Liu, 2004; Popescu & Etzioni, 2005; Turney, 2001). All these work focused on opinion mining which was to discover the reviewers’ orientations, whether positive or negative, regarding various features of a product, e.g. weight of a laptop and picture quality of a digital camera. However, we noticed that although some comments regarding product features could not be labelled as positive or negative, they were still valuable. For example, the following two sentences are extracted from the customer reviews of mobile phone Nokia 6610 from Hu’s corpus (Hu & Liu, 2004):

#1: The phone’s sound quality is great.
#2: The most important thing for me is sound quality.

Both sentences discuss the product feature sound quality. Unlike the first sentence, the second one does not offer any orientation, either positive or
negative, regarding the specific phone Nokia 6610, yet it does provide valuable information for designers and manufacturers about what mobile phone consumers are really concerned about. Such neutral comments and suggestions are currently not considered in the method of opinion mining.

Moreover, opinion mining focuses mainly on product features which can not cover all significant issues in customer reviews. Figure 1 shows some sentences extracted from the customer reviews of Nokia 6610. These sentences all discuss flip phone and they reveal the different perspectives from customers about flip phone. Some customers also elaborate on the reasons for their choices. This information is believed to be valuable for designers and manufacturers. However, in the method of opinion mining, such important issues were not pointed out because flip phone is not an explicit product feature of Nokia 6610.

2 AUTOMATIC TEXT SUMMARIZATION

During the last decade, there has been much interest with automatic text summarization due to the explosive growth of electronic documents online (Barzilay & Elhadad, 1997; Gong & Liu, 2001; Hovy & Lin, 1997; Yeh et al., 2005). There are also some initial web applications. For example, Google provides a short summary for each retrieved document in the form of scraps related to the query words. Another example is NewsInEssence (http://www.newsinessence.com/) which is able to summarize news articles from various sources.

There are two major groups of automatic summarization approaches: statistical methods and linguistic methods. Statistical methods are widely used because of their robustness and independency of document genre. The first implementation can be traced back to Luhn’s work (Luhn, 1958) in which the author developed a method based on frequency of words. Subsequent researchers extended Luhn’s work to deal with more features in addition to frequent words, e.g. title and heading words (Edmundson, 1969), sentence position (Hovy & Lin, 1997), indicator phrases (Hovy & Lin, 1997), sentence length (Kupiec et al., 1995), etc. Linguistic methods present a different way for summarization. The typical methods include discourse structure (Mann & Thompson, 1988; Marcu, 1999) and lexical chains (Barzilay & Elhadad, 1997).

Recently, as an outcome of the capability to collect large sets of documents online, there is an increasing demand for MDS. Instead of focusing only on single document, MDS is performed to deal with multiple related documents (Mani & Bloedorn, 1999; McKeown & Radev, 1995), e.g. news articles regarding an event from various sources. The most popular MDS approach is clustering-summarization (Boros et al., 2001; Maña-López, 2004; Radev et al., 2004). The approach of clustering-summarization first separates a collection of documents into several non-overlapping groups of documents or sentences. Summarization is then performed separately within each group. There are two limitations to the clustering-summarization approach when applied to the domain of customer reviews:

- The number of clusters is difficult to determine
without prior knowledge regarding the collection of reviews. Inappropriately choosing this number will inevitably introduce noisy information and reduce effectiveness.

In clustering-summarization, the document set is split into non-overlapping clusters and each cluster is assumed to discuss one topic. However, in a collection of reviews, topics often overlap with each other and are not perfectly distributed in the non-overlapping clusters of documents. Each topic is associated with various reviews. Likewise, each review in the collection possibly discusses several topics instead of only one because customers usually comment on various aspects of product rather than focus on one aspect. These two limitations of the clustering-summarization method are tackled in our approach based on topical structure.

3 SUMMARIZATION BASED ON TOPICAL STRUCTURE

Based on analysis of various text corpora including DUC (http://duc.nist.gov/) and Hu’s corpus (Hu & Liu, 2004), we observed that in a document collection, topics often overlapped with each other and are not perfectly distributed in the non-overlapping clusters. As shown in Figure 2 which lists some topics in the review collection of Nokia 6610 and review IDs with respect to these topics, review 18 has comments regarding all the topics and some other reviews are also associated with multiple topics. The approach of clustering-summarization is not suitable in this situation since clustering this collection into non-overlapping groups will cut off the relationship among reviews.

- Sound quality 8, 13, 18, 20, 27, 33, 34, 40
- Battery life 2, 5, 10, 13, 17, 18, 26, 28, 29, 30, 37
- Flip phone 4, 18, 26, 33
- Nokia phone 1, 2, 16, 17, 18, 31, 37
- Samsung phone 18, 40
- ...

Figure 2: Some topics from the review collection of Nokia 6610.

We propose a summarization approach based on the topical structure demonstrated in Figure 2. The framework of our approach is shown in Figure 3. Detailed steps are given as follows.

3.1 Pre-processing

The summarization process starts with a collection of customer reviews as the input. These reviews are collected from WWW or retrieved from Intranet, e.g. all customer emails regarding a product. Pre-processing steps are first applied to the reviews, including stop words removal and term stemming (Porter, 1980). The purpose of these steps is to reduce the noise in the following processes.

3.2 Topic Identification

The key step of our framework is to identify topics in the review collection and generate the topical structure based on these topics. Some work of topic identification has been reported in previous literature. The typical method is text segmentation, which is to segment the text by similarity of adjacent passages and detect the boundary of topics (Choi, 2000; Hearst, 1997; Moens & De Busser, 2001; Ponte & Croft, 1997). This method works well for
single text. For multiple texts, however, it is hard to find such straightforward boundaries.

Our process of topic identification is based on frequent word sequences (FSs) (Liu, 2005) and equivalence classes (Ahonen, 1999). A FS is a sequence of words that appears in at least \( \sigma \) documents in a document collection (\( \sigma \) is the threshold for supporting documents). Algorithm 1 demonstrates the process to extract all the FSs in a document collection. The process starts with collecting all the frequent word pairs, i.e. FSs with length two. These FSs are then expanded with one more word and therefore form a set of word sequences with length three. All the FSs with length three are then expanded. This process is iteratively performed until there is no FS left for expansion. The threshold for supporting documents is chosen according to the size of the review collection. For a large collection, a high threshold may be chosen to let more important concepts to surface. For a small collection, say 20 reviews, a low threshold is considered to reduce noisy information.

**Algorithm 1:** Discovery of all FSs in a review collection

```
//Input: D: a set of pre-processed reviews
\( \sigma \): frequency threshold
//Output: Fs: a set of FSs
//Initial phase: collecting all frequent pairs
1 For all the reviews \( d \in D \)
2 Collect all the ordered pairs and occurrence information in \( d \)
3 Seq = all the ordered word pairs that are frequent in \( D \)
//Discovery phase: building longer FSs
4 \( k = 2 \)
5 Fs = \{ Seq \}
6 While \( Seq \neq \) 
7 For all phrases \( s \in Seq \)
8 Let \( l \) be the length of the sequence \( s \)
9 Find all the sequences \( s' \)
10 For all \( s' \)
11 If \( s' \) is frequent
12 \( S = S \cup \{ s' \} \)
13 Fs = Fs \cup S
14 Seq = Seq \cup S
15 \( k = k + 1 \)
16 Return Fs
```

FSs can be further pruned and grouped into equivalence classes according to their cooccurrences with each other. The equivalence classes are generated in the following way. Let \( A \) and \( B \) be two FSs. The equivalence class of \( A \), \( Eq_A \), contains the set of FSs that cooccur with \( A \) in almost the same set of reviews, as given by a confidence parameter. \( Det_A \) is the set of FSs that are determined by \( A \), and is required in deciding which FSs belong in \( Eq_A \). For \( A \) and \( B \), if:

\[
\frac{\text{frequency}(A, B \text{ cooccur})}{\text{frequency}(A)} \geq \text{confidence}
\]  

we add \( B \) to the set \( Det_A \); \( A \) itself is also included in \( Det_A \). Other FSs are tested in the same manner, and will be added to \( Det_A \) if they satisfy the above criterion. \( Eq_A \) is thus made up of all FSs \( X \) such that \( Det_X \subset Det_A \).

A FS or an equivalence class is considered as the representative of one topic in a review collection. In the following experiments, we intend to compare the performance between FSs and equivalence classes as topics. Topics are ranked based on their scores. The score of a FS is calculated in the form of Equation 2. The score of an equivalence class equals to the average scores of its FSs.

\[
\text{score} = f \cdot \log \frac{N + 1}{n} \cdot \log (l + 2)
\]

where \( f \) is the frequency of the FS in the whole review collection, \( N \) is the total number of reviews, \( n \) is the number of reviews in which the FS occurs, \( l \) is the length of the FS.

### 3.3 Candidate Sentence Extraction

For each topic in a collection, all relevant sentences are extracted and added into a pool as candidate segments of final summary until the expected summary length is reached. Each sentence will be accompanied by a label including its source review ID. The method of Maximal Marginal Relevance (MMR) is implemented to reduce the redundancy in the sentence selection process (Carbonell & Goldstein, 1998). MMR intends to balance the tradeoff between the centrality of a sentence with respect to the topic (the first part in Equation 3) and its novelty compared to the sentences already selected in the summary (the second part in Equation 3), i.e. to maximize the marginal relevance in the following form:

\[
MR(s) = \Delta \text{Sim}(s, D) - (1 - \lambda) \max_{s \in \mathcal{S}} \text{Sim}(s, s')
\]
where $s_i$ is a candidate sentence, $D$ is the set of relevant sentences to a particular topic, $S$ is the set of sentences already included in the summary, $\lambda$ is the redundancy parameter ranging from 0 to 1. With regard to $Sim$, we adopt a cosine similarity measure between sentence vectors. Each element of a sentence vector represents the weight of a word-stem in a document after removing stop words.

### 3.4 Post-processing and Final Presentation

The final step is to regenerate sentences from the candidate sentences and present the summary output to users.

![Summary (Nokia 6610, 41 reviews)](image)

**Summary (Nokia 6610, 41 reviews)**

- **Sound quality** (8, 13, 18, 20, 27, 33, 34, 40)
- **Battery life** (2, 5, 10, 13, 17, 19, 26, 28, 29, 30, 37)
- **Flip phone** (4, 15, 26, 33)
  - Wish it were a flip phone [3]
  - As much as I like Nokia phones the flip phones are much better because a) you won’t scratch your screen b) you don’t need to lock your phone all the time to prevent accidentally hitting the keys. Personally I like the Samsung phones better because I found myself liking the flip phones so much more. [18]
  - My last two phones were all flip phones, and I was beginning to tire of them. Nokia was my first non-flip phone, and I’m glad I decided to go with them. [16]
  - This is probably your best bet if you are looking for a phone in this price range, or like me, do not have the patience to deal with annoying flip phones. [33]
- **Nokia phone** (1, 2, 16, 17, 18, 31, 37)
  - Samsung phone (19, 40)

Figure 4: Summarization output for the review collection of Nokia 6610.

Figure 4 shows an example of the summary presented to readers. Topics are ranked according to their saliency in the review collection. Reviews relevant to each topic have been identified and hyperlinked, with their IDs included in the parenthesis following the topical phrase, to make it easy for users to browse the details of each review article. If users are interested in a particular topic, they can click the unfolding button prior to the topical phrase to expand this topic and the detailed information will then be presented. In Figure 4, the topic *flip phone* is unfolded and all the relevant sentences to this topic are displayed along with reviews’ IDs.

### 4 Evaluation of Summarization Results

We compared our summarization approach with the baseline summarization systems of Copernic summarizer (http://www.copernic.com) and clustering-summarization. Copernic summarizer is a commercial summarization software using undisclosed statistical and linguistic algorithms. The method of clustering-summarization is a popular method for MDS, especially in the context of information retrieval system (Maña-López, 2004; Roussinov & Chen, 2001). In clustering-summarization, a document collection is separated into non-overlapping clusters and summarization is then performed in each cluster.

#### 4.1 Experimental Data Sets and Parameter Setting

The data sets used in our experiments included five sets from Hu’s corpus (Hu & Liu, 2004) and three sets from Amazon.com. For each review collection, summaries were generated using Copernic summarizer, clustering-summarization method and our approach based on topical structure. These document sets were normal-sized with 40 to 100 documents per set. Therefore, we extracted FSs with at least three supporting documents in our approach. The confidence level for equivalence classes was set to 0.9 and redundancy parameter $\lambda$ in candidate sentence selection was set to 0.5. Since the document sets in our experiments were normal-sized, the clustering number in clustering-summarization method was set to five. The clustering algorithm in our experiments was implemented in Cluto (Karypis, 2002).

The compression ratio of summarization was set to 10%, i.e., the length ratio of summary to original text was 10%. The summary generated by Copernic was a set of ranked sentences. The summary generated by clustering-summarization was divided into clusters, as shown in Figure 5 (only three clusters are shown here).
4.2 Intrinsic Evaluation and Extrinsic Evaluation

The methods of summarization evaluation can be classified into intrinsic method and extrinsic method. Intrinsic method compares candidate summaries with reference summaries (Jing et al., 1998). Reference summaries are usually generated manually and are therefore biased by human authors. Extrinsic method requires no reference summary and is task-oriented or user-oriented (Maña-López, 2004; Tombros & Sanderson, 1998). In our case, since it is hard to define an ideal reference summary to fulfill the diverse information requirements of different users, extrinsic evaluation is more suitable.

We evaluated summarization performance according to users’ responsiveness. Human assessors were required to give a score for each summary based on its structure and coverage of important topics in the review collection. The score was an integer between 1 and 5, with 1 being least responsive and 5 being most responsive. In order to reduce bias in the evaluation, three human assessors from different background joined the scoring process. For one collection, all the peer summaries were evaluated by the same human assessor so that the hypothesis testing (paired t-test) could be performed to compare the peer summaries.

4.3 Evaluation Results

Table 1 shows the average responsiveness scores of Copernic summarizer, clustering-summarization method and our approach based on all the review collections. Table 2 presents the results of paired t-test between our approach (using FSs as topics) and other methods.

It can be found that the approach based on topical structure performed the best amongst all the peer methods (Table 1 & 2), because this approach better represents the internal structure of a review collection than clustering-summarization. We also analyzed the clustering quality in the clustering-summarization method. Table 3 shows the intra-cluster similarity and inter-cluster similarity for the review collection Nokia 6610. As can be seen, there was not much difference between intra-cluster similarity and inter-cluster similarity, especially for cluster 4 and 5 which were the two major clusters in the collection. This implies that the review collections are difficult to be clustered into non-overlapping clusters.

As shown in Table 1 & 2, we found that using FSs as topics was significantly better than equivalence classes with the p-value of 0.0008 in paired t-test. Review writers usually write in an arbitrary style and cover different topics in a review rather than focus on only one topic. Therefore, using equivalence classes might introduce much noisy information, since equivalence classes are grouping topics based on their cooccurrences. Copernic summarizer performed worse than other summarization methods. The possible reason is that Copernic summarizer does not take into account the case of MDS and treats all sentences from a review collection as the same in the pool of candidate segments for summarization.
Table 1: Average responsiveness scores.

<table>
<thead>
<tr>
<th>Method</th>
<th>Responsiveness score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copernic summarizer</td>
<td>1.1</td>
</tr>
<tr>
<td>Clustering-summarization</td>
<td>2.3</td>
</tr>
<tr>
<td>Topical structure-based summarization</td>
<td></td>
</tr>
<tr>
<td>FSs</td>
<td>4.3</td>
</tr>
<tr>
<td>Equivalence classes</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Table 2: Hypothesis testing (paired t-test).

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalence classes vs. Frequent word Sequences (FSs)</td>
<td>$2.26 \times 10^{-5}$</td>
</tr>
<tr>
<td>Equivalence classes vs. Clustering-summarization</td>
<td>$2.43 \times 10^{-4}$</td>
</tr>
<tr>
<td>Frequent word Sequences (FSs) vs. Equivalence classes</td>
<td>$7.68 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

5 CONCLUSION

Summarization of online customer reviews is a process to transfer reviews from unstructured free texts to a structured or semi-structured summary which can reveal the commonalities and links among reviews. The automation of this process, in the context of e-Commerce and e-Business, should be able to assist potential consumers in seeking information and to facilitate knowledge management in enterprises as well.

We proposed an approach to automatically summarize multiple customer reviews based on topical structure. Based on the observation that topics often overlap with each other in a collection of reviews, we extracted topics across reviews, instead of dividing reviews into several non-overlapping clusters. Evaluation results demonstrated that our approach achieved better summarization performance and users’ satisfaction compared to the baseline systems of Copernic summarizer and clustering-summarization method. Moreover, this approach is able to address different concerns from potential consumers, distributors and manufacturers. Potential consumers usually concentrate on the positive or negative comments given by other consumers. Designers and manufacturers, on the other hand, may be more concerned about the overall important issues and the reasons why customers are favoring or criticizing their products.

The emergence of Blogs and e-Opinion portals has offered customers novel platforms to exchange their experiences, comments and recommendations. Reviews for a particular product may be obtained from various sources in different writing styles. How to integrate information from different sources will be the focus in our future work.

REFERENCES


