Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2016 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 6-27-2016

UNDERSTANDING THE MASSIVE ONLINE REVIEWS: A NOVEL REPRESENTATIVE SUBSET EXTRACTION METHOD

Jin Zhang
Renmin University of China, zhangjin@rbs.org.cn

Baojun Ma
University of Posts and Telecommunications, mabaojun@bupt.edu.cn

Jilong Zhang
Renmin University of China, zhangjilong@ruc.edu.cn

Ming Ren
Renmin University of China, renm@ruc.edu.cn

Chong Ma
Renmin University of China, xbzhtx@ruc.edu.cn

UNDERSTANDING THE MASSIVE ONLINE REVIEWS: A NOVEL REPRESENTATIVE SUBSET EXTRACTION METHOD

Jin Zhang, School of Business, Renmin University of China, Beijing, China, zhangjin@rbs.org.cn

Baojun Ma (corresponding author), School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China, mabaojun@bupt.edu.cn

Jilong Zhang, School of Business, Renmin University of China, Beijing, China, zhangjilong@ruc.edu.cn

Ming Ren, School of Information Resource Management, Renmin University of China, Beijing, China, renm@ruc.edu.cn

Chong Ma, School of Business, Renmin University of China, Beijing, China, xbzhtx@ruc.edu.cn

Abstract

With the widespread use of e-commerce, explosive online reviews often get overwhelming to online consumers to read, leading to the depression of decision making in purchasing. To deal with the information overload problem caused by large-scale online reviews, this study focuses on the representative review extraction problem and formulates this problem as an optimization model with a submodularity property. Then, to analyze and draw topic information from the original review collection, the topic model of LDA is adopted in this study to semantically measure the similarity between each pair of reviews for the optimization model. Furthermore, a greedy extraction method named RR with a satisfactory error bound is proposed to extract a representative subset of reviews from the original review collection based on the optimization model. Experiments on real data and a user study are conducted in this study. The experimental results demonstrate that the proposed method is of high efficiency and scalability, and performs better than benchmark methods in terms of coverage, which proves that it can help online consumers better capture the main ideas of the whole set of original reviews in a limited time.

Keywords: Information overload, Representative reviews, Topic modelling, Coverage, Submodularity, Extraction method

1 INTRODUCTION

Recent years have witnessed the rapid development of e-commerce, where online reviews are playing an increasingly important role in consumers' decision making process. As a typical form of user generated content (UGC), online reviews are highly detailed and directly posted from past buyers (Bickart & Schindler 2001; Mudambi & Schuff 2010; Sun 2012). When consumers purchase products, select hotels, or pick books to read on various websites, they heavily depend on online reviews due to their comprehensiveness and objectivity. Success of major e-commerce sites, such as Amazon.com, are partly attributed to their voluminous and extensive customer reviews. Many online merchants enable, even encourage their past customers to express opinions on products they have bought. Websites like Epinions.com have established business that hosts online reviews to improve customer satisfaction and experiences.

However, the vast body of online reviews accumulating as time goes by can be overwhelming to consumers to read, leading to the problem of information overload (Lucian et al. 2007). For example, in Amazon.com, there are typically more than one thousand reviews for popular products. A large number of reviews are more likely to provide rich description on different aspects of a product, which is sometimes deemed desirable. However, abundant online reviews beyond consumers' information processing capability may be difficult for potential consumers to read. Consumers can hardly make a well-informed decision on whether to buy the product. Therefore, it is meaningful to extract a subset of reviews from the original reviews into a representative form that is easy for customers to absorb different content in a limited time.

In response to this problem, some studies have been proposed to extract a representative subset by laboriously elaborating on large-scale search results. Pan et al. (2005) put forward a greedy model to identify a representative subset from database query results in terms of high coverage and low redundancy. After that, Zhang et al. (2012) made attempts to improve the measurement of representativeness by proposing a transitive based method to extract a subset of tuples from the query results returned by database. Afterwards, a λ-representative method was proposed by Zhang et al. (2014) to extract a representative subset from web search results on search engines. These approaches especially provide users a small subset of tuples in databases or web pages returned by search engines. However, these studies are only proposed for processing database queries or search engine results, thus are limited in usage and application in online reviews, which are typical user generated content and crucial for customers' online shopping.

As for online reviews, a great number of efforts on review ranking have been made by scholars in recent years. A popular way is to present customers the top-*k* results by ranking the original online reviews based on some criteria, such as review time, review helpfulness, etc. (Kim et al. 2006; Zhang & Varadarajan 2006; Ghose & Ipeirotis 2007; Liu et al. 2008; Baccianella et al. 2009). The top-*k* results can alleviate the problem of information overload in online reviews to some extent. However, they have some aspects calling for improvement. In particular, they usually provide results of similar or even identical topics, and do not take into consideration the coverage over the whole body of original online reviews, with the result that some important information may be missed in the top-*k* results. For example, all the top-*k* reviews of a cell phone may be highly informative about its outstanding functions, but those reviews about its portability which are also important to consumers may be excluded from the highest ranking results. In addition to review ranking methods, a considerable amount of work with respect to review summarization has been proposed to provide statistical results (commonly in percentage) in terms of the number of reviews with positive opinions and negative opinions on each product feature, where feature extraction and sentiment analysis are used to identify product features and sentiment orientations regarding each feature (Hu & Liu 2004; Liu et al. 2005; Liu et al. 2007; Miao et al. 2010; Thet et al. 2010; Shimada et al. 2011). However, such summary results generated by these methods are regarded as lacking narrative and detailed

structure of natural reviews, leading to the result that such methods are not totally trusted by consumers in their decision process for purchasing (Archak et al. 2011).

To overcome the limitation of the review summarization methods, some research efforts in recent years have been focused on extracting individual online review to form a representative subset in terms of feature coverage (Lappas & Gunopulos 2010; Tsaparas et al. 2011). Formulations in these works are designed to select a subset of reviews covering various literal features mentioned in the original online reviews. However, there still lack in-depth investigations from the semantic perspective. Furthermore, the coverage performance of these methods needs to be further improved.

This study has designed a new model with respect to representative subset extraction from original online review collection. Moreover, a heuristic extraction method named *Representative Review*(*RR*) is proposed to find the solution for the model and extract a subset of reviews based on the similarity between each pair of reviews, which is calculated by using the topics drawn from the corpus of reviews. The method is with a satisfactory error bound based on the submodularity property of the designed model and can thus extract a subset with high coverage on different content in the collection of original reviews. Extensive experiments on real data and a user study are conducted to demonstrate the advantages of the proposed method over benchmark methods. By applying the proposed method, the extracted results are as informative as possible for a certain product on e-commerce. Therefore, consumers can benefit from the representative subset and quickly grasp the main ideas of the original online reviews.

2 LITERATURE REVIEW

This paper is to extract a representative subset of online reviews from a given online review corpus, and provide consumers with a informative subset that covers as much content as possible. This target is closely related to the research efforts in the fields of search result extraction, review summarization and review extraction.

2.1 Search Result Extraction

These approaches attempt to extract a subset from search results returned by databases or search engines. The extracted subset is expected to contain the content of the original search results. For example, given a large database, Pen et al. (2005) put forward a greedy model to identify a small subset from the perspective of information-theoretic measures to ensure its representativeness. Zhang et al. (2012) proposed a transitive based method to extract representative tuples from databases, and improved the representativeness measurement in terms of compactness and redundancy. They extracted the representative tuples from each class which was split from the original dataset to form a representative set. After that, Zhang et al. (2014) introduced the notion of λ-represent to extract a subset of search results with high coverage on the original search results and low redundancy in the extracted subset.

These methods can extract representative information from the original collection of search results. But they are mostly focused on database queries or search engine results, thus are limited in usage and application. The purpose of this paper is to extract a representative subset from the original online review corpus, which can be considered as the complementary to representative information extraction methods focusing on processing general search results.

2.2 Review Summarization

This paper also ties to the considerable work devoted to review summarization. Hu and Liu (2004) essentially introduced this problem. They extracted the features of electronic products (such as the picture quality and screen size of a digital camera), and analyzed whether each opinion sentence is

positive or negative. They created a statistical summary about the numbers of positive and negative opinions in the corpus. Liu et al. (2005) proposed a visual analysis method comparing customer opinions of different products. Miao et al. (2010) mined and integrated product features from multiple review sources. Shimada et al. (2011) used objective information of the products from other websites (i.e., Wikipedia) to enrich the review summary. In addition, Liu et al. (2007) explored ways to filter out low-quality reviews to prevent them from corrupting the summary results. Furthermore, in contrast to most review summarization on the sentence level, Thet et al. (2010) summarized opinions on the clause level, so different opinions on multiple aspects expressed in a sentence can be processed separately.

In review summarization, the given review corpus is processed to produce summaries which are statistical in nature. However, review summarization breaks coherence and integrity of the original review corpus. In order to maintain the narrative structure of reviews, the method proposed in this paper largely extended previous work by providing customers with user-friendly subsets of the original reviews.

2.3 Review Extraction

In order to obtain a small subset of reviews, review ranking is an important way by extracting the top-*k* best reviews which are scored according to certain criteria. A significant amount of work has been devoted to scoring different aspects of review quality. The methods proposed in (Kim et al. 2006; Zhang & Varadarajan 2006; Ghose & Ipeirotis 2007; Liu et al. 2008) focused on assessing user-supplied helpfulness votes presented on review-hosting websites (e.g., Amazon.com, Overstock.com). Later by, Baccianella et al. (2009) tackled the review ranking problem based on their textual content. In addition, Hu and Liu (2004) extended assessment of review helpfulness by exploiting contextual information about author identities and social networks to improve online review quality. In all these works, top-*k* highest-scoring reviews are extracted as a subset shown to customers, which contain reviews of similar content and may miss some important aspects of the whole review collection.

To overcome the limitation of review ranking methods, some efforts in recent years have been focused on review extraction to find a small set of reviews covering different content of the whole body of reviews. Lappas and Gunopulous (2010) presented a search framework which could extract a subset of reviews containing the features posted by customers. Tsaparas et al. (2011) proposed some heuristic methods to seek a subset of fixed size by giving an upper bound on the number of the selected reviews. In their work, the selected subset always contains at least one positive opinion and one negative opinion on each feature. In addition, Lappas et al. (2012) introduced and addressed a characteristic-review selection problem and provided a compact subset of reviews to reflect the ratios of positive and negative opinions over certain features from a given corpus. However, all these studies considerate the coverage problem and process online reviews at the feature-level granularity. They were proposed to cover the features mentioned in the original collection of reviews and lacked a deep understanding from the semantic perspective, such as the topics in the original reviews, with the result that the coverage performance of these methods still needs to be further improved. The limitation of existing studies thus motivates our study to a large extent in proposing a representative subset to semantically cover different content of original reviews with a better performance.

3 PROBLEM DEFINITION

This section introduces the problem definition of the representative subset extraction from online reviews, as well as the analysis of the proposed extraction model.

3.1 Representative Extraction Model

Given a product denoted as p on an e-commerce site, the corresponding collection of n online reviews with respect to the product is denoted as $R=\{r_1,r_2,\ldots r_n\}$, where r_i ($1 \leqslant i \leqslant n$) represents a single review posted by consumers. R are posted consumers who have purchased product p and depict various aspects of product p. If two reviews are similar to each other, they are considered to describe similar aspects of the product. For example, the following two sample reviews are posted by consumers of a restaurant.

r_1: *The food taste of the restaurant is great.*

r_2: *The food is tasty. We like it very much.*

Both the reviews r_1 and r_2 comment on the food taste of the restaurant, and they express a similar meaning. In other words, for consumers who plan to eat in the restaurant, they only need to browse one of them to obtain information about the food taste of the restaurant.

In this study, the similarity between each pair of reviews r_i and r_j is denoted as $sim(r_i, r_j)$. The value range of the similarity between reviews is [0,1]. If two reviews are identical to each other, the similarity between them is 1, whereas the similarity is 0. To semantically measure the similarity between each pair of reviews, the topic model LDA (Blei et al. 2003; Blei 2012) is adopted here to draw topic information hidden in the review collection. The similarities are calculated based on the topic information, which will be introduced in the following section in detail.

Based on the calculated similarity $sim(r_i, r_j)$, review r_i can be considered to represent r_j to a certain degree in terms of $sim(r_i, r_j)$ (Zhang et al. 2014). Therefore, in spirit of the research proposed by Ma et al. (2011), the representative review extraction problem can be formulated as follows.

Given a product p and the corresponding review collection $R=\{r_1,r_2,\ldots r_n\}$, the similarity between each two reviews in R (i.e., $sim(r_i, r_j)$), and an integer k (i.e., $1<k<n$) specified by users, the representative review extraction problem is to find a subset $R_{s*} \subseteq R$ and $|R_{s*}|=k$, such that,

$$R_{s*} = argmax_{R_s \subseteq R, |R_s|=k} Re(R_s, R) = argmax_{R_s \subseteq R\,|R_s|=k} \{\frac{\sum_{r_j \in R} max_{r_i \in R_s} \{sim(r_i, r_j)\}}{n}\} \quad (1)$$

In Equation (1), for a review r_j in R, it can be represented by the subset R_s with a degree of the maximum similarity between review r_i in R_s and r_j. This is a conservative estimate on the coverage degree between R_s and r_j. Then, the coverage degree of R_s on R can thus be calculated as the average coverage value of R_s on each single review in R. The target of the representative review extraction is to find a subset R_{s*} which possesses a maximum coverage degree on R. In other words, the subset R_{s*} can be considered to represent the whole review collection R by a maximum coverage degree. Consumers can directly get an overall picture by browsing the extracted subset R_{s*} in a short time.

3.2 Analysis of the Representative Extraction Model

The representative review extraction problem formulated as Equation (1) can be mapped to a typical Max Coverage Problem (Hochba 1997), which has been previously proved to be a NP-hard problem. Thus, it can be derived that the representative review extraction problem in Equation (1) is also a NP-hard problem. However, the objective function $Re(R_s, R)$ in Equation (1) possesses a desirable property of submodularity. Therefore, the greedy strategy can be leveraged to find the solution to the problem with a satisfactory error bound.

Concretely, a function with submodularity can be interpreted as the economic principle of diminishing marginal returns, i.e., the marginal benefit of adding a single review to a larger collection is less than that of adding it to a smaller collection of reviews. It can be proved that the objective function $Re(R_s, R)$ in Equation (1) is a submodular function, which is demonstrated as Proposition 1 in detail.

Proposition 1:*The objective function Re(R_s, R) of the representative review extraction problem in this paper is submodular*.

Proof: The submodular of *Re*(R_s, R) is equivalent to prove that the following inequality holds:

$$Re(R_{s2}, R) - Re(R_{s1}, R) \geq Re(R_{s2} + \{r\}, R) - Re(R_{s1} + \{r\}, R) \tag{2}$$

where $R_{s1} \subseteq R_{s2} \subseteq R$, and the review $r \in R/R_{s2}$.

Equation (2) can be converted as follows by equivalent transformation:

$$\begin{aligned} &\sum_{r_j \in R} \{max_{r_{i2} \in R_{s2}} sim(r_{i2}, r_j) - max_{r_{i1} \in R_{s1}} sim(r_{i1}, r_j)\} \\ &\geq \sum_{r_j \in R} \{max\{max_{r_{i2} \in R_{s2}} sim(r_{i2}, r_j), sim(r, r_j)\} - max\{max_{r_{i1} \in R_{s1}} sim(r_{i1}, r_j), sim(r, r_j)\}\} \end{aligned} \tag{3}$$

For each review r_j in R, we can considerate the following inequality in four separate cases.

$$\begin{aligned} &max_{r_{i2} \in R_{s2}} sim(r_{i2}, r_j) - max_{r_{i1} \in R_{s1}} sim(r_{i1}, r_j) \\ &\geq max\{max_{r_{i2} \in R_{s2}} sim(r_{i2}, r_j), sim(r, r_j)\} - max\{max_{r_{i1} \in R_{s1}} sim(r_{i1}, r_j), sim(r, r_j)\} \end{aligned} \tag{4}$$

Case 1: $sim(r, r_j) \geq max_{r_{i2} \in R_{s2}} sim(r_{i2}, r_j) \geq max_{r_{i1} \in R_{s1}} sim(r_{i1}, r_j)$. The right side of Equation (4) is 0. Because the left part is always a nonnegative number. Then Equation (4) holds.

Case 2: $max_{r_{i2} \in R_{s2}} sim(r_{i2}, r_j) \geq sim(r, r_j) \geq max_{r_{i1} \in R_{s1}} sim(r_{i1}, r_j)$. The right part of Equation (4) equals to $max_{r_{i2} \in R_{s2}} sim(r_{i2}, r_j) - sim(r, r_j)$. Because $max_{r_{i1} \in R_{s1}} \{sim(r_{i1}, r_j)\} \leq sim(r, r_j)$, Equation (4) also holds.

Case 3: $max_{r_{i2} \in R_{s2}} sim(r_{i2}, r_j) \geq max_{r_{i1} \in R_{s1}} sim(r_{i1}, r_j) \geq sim(r, r_j)$. The right part of Equation (4) equals to $max_{r_{i2} \in R_{s2}} sim(r_{i2}, r_j) - max_{r_{i1} \in R_{s1}} sim(r_{i1}, r_j)$. Thus, Equation (4) holds.

For each review r_j in R, we have that Equation (4) holds in all cases. Then, Equation (3) also holds, proving that the objective function *Re*(R_s, R) is submodular.

The submodularity of *Re*(R_s, R) in Proposition 1 leads to an appealing property, that is the greedy strategy can be used to find a solution with a satisfactory error bound. Concretely, let R_{opt} be the optimal subset of k reviews that maximizes *Re*(R_s, R) and R_{s*} be the subset of k reviews extracted by a greedy strategy, it can be derived that *Re*(R_{s*}, R) $\geqslant$(1-1/e) *Re*(R_{opt}, R) (Nemhauser et al. 1978). Thus, to find a satisfactory solution for the representative review extraction problem formulated as Equation (1), an extraction method with a greedy strategy is proposed in this study, the details of which will be introduced in the next section.

4 REPRESENTATIVE REVIEW EXTRACTION METHOD

This section introduces the representative review extraction method with a greedy strategy. Firstly, as illustrated in Section 3, the similarity between each pair of reviews can express the representativeness degree between reviews (Zhang et al. 2014). This study leverages topic information to semantically measure the similarity between reviews, which will be introduced in this section as well. Thereafter, based on the calculated similarity, a greedy method for representative review extraction is illustrated in this section.

4.1 Similarity between Reviews

In this study, the typical topic model of Latent Dirichlet Allocation (LDA) is adopted to identify latent topic information from the original review collection (Blei et al. 2003; Blei 2012). LDA is a probabilistic generative model with an unsupervised machine learning technique. It is widely used in general text processing and online review analyzing (Ma et al. 2015). In spirit of the assumptions in LDA, this study accordingly assumes that every review is a distribution over topics and every topic is a distribution over keywords. Specifically, given a product p, the corresponding review collection $R=\{r_1,r_2,\ldots r_n\}$ constitutes the training corpus. Thus, it can enable topic distillation through LDA modelling. Firstly, all reviews in R are processed into bags of keywords by word segmentation and stop word filtering. Each review r_i is interpreted as a multinomial distribution $Mult(\theta)$ over a series of topics $T=\{t_1,t_2,...,t_m\}$. And each topic $t \in T$ is assigned a multinomial distribution $Mult(\varphi)$ over all keywords $W=\{w\}$ in R. θ and φ represent two dirichlet distributions with hyper-parameters α and β respectively, and are denoted by $\theta \sim Dir(\theta|\alpha)$ and $\varphi \sim Dir(\varphi|\beta)$. For each keyword w in review r_i, sample a topic t from the multinomial distribution $Mult(\theta)$ with review r_i, and subsequently sample the observed keyword w from the multinomial distribution $Mult(\varphi)$ associated with the sampled topic t. Therefore, the generation probability of each keyword w in review r_i can be formulated as,

$$\begin{aligned} p(w \mid r_i) &= p(w \mid t, \beta)\, p(t \mid r_i, \alpha) \\ &= \int p(w \mid \varphi) Dir(\varphi \mid t, \beta) d\varphi \int p(t \mid \theta) Dir(\theta \mid r_i, \alpha) d\theta \end{aligned} \tag{5}$$

Repeating the sampling process above for all the keywords in the review r_i and the generation probability of the observed corpus R can be expressed as,

$$p(R) = \prod_{r_i \in R} \prod_{w \in r_i} p(w \mid r_i) \tag{6}$$

There are three parameters in the modelling process, i.e., θ, φ, and the number of topics m. To estimate the parameters θ and φ, the Gibbs sampling (Griffiths 2002), a fast and effective algorithm for approximate inference, is applied to infer the model. In addition, the indicator *perplexity* is used to determine the topic number m. The measurement of *perplexity* is used by convention in language modelling, and calculated as the inverse of the geometric mean per-keyword likelihood (Blei et al. 2003). It is used to measure the generalization performance with different topic numbers. Generally speaking, the lowest value of perplexity indicates the suitable parameter of the topic number m. Because online reviews are often regarded as short texts and contain limited topics about the corresponding product, without loss of generality, the *perplexity* measurement is utilized to determine the topic number m within the range of (0,10) in this study.

After parameter estimation, the latent topic information hidden in the review collection R can be obtained. Each review r_i is mapped into a topic distribution, denoted by $r_i=\{p_1, p_2, ..., p_m\}$, where p_q ($1 \leqslant q \leqslant m$) stands for the probability that the particular review r_i belongs to topic t_q. Topic distributions reflect semantic information of reviews. Based on the topic distributions, the similarity between reviews can be semantically calculated. Intuitively, if the topic distribution of two reviews are similar to each other, these two reviews can be considered as commenting on the similar aspects of the product or they can represent each other. Therefore, a commonly used measurement of topic distribution divergence, i.e., Jensen-Shannon (JS) divergence is used in this study to calculate the similarity between each pair of reviews in terms of evaluating the distance between two distributions (Endres & Schindelin 2003). Specifically, given two topic distributions P_i and P_j of two reviews r_i and r_j, the distance between them is calculated as follows:

$$JS(P_i \mid P_j) = \frac{1}{2} KL(P \,||\, M) + \frac{1}{2} KL(Q \,||\, M) \tag{7}$$

where M=1/2(P+Q) and KL is the well-known Kullback-Leibler divergence measurement (Kullback & Leibler 1951). For two reviews r_i and r_j, they are similar to each other when the corresponding JS distance is low. Therefore, in a given review collection R, the distance between each pair of reviews is linearly mapped to a similarity within the range of [0,1] by setting the largest distance as a similarity value of 0.0 and the lowest distance as a similarity value of 1.0. If the similarity is close to 1.0, it can be concluded that the two reviews are similar to each other and each review can represent the other well.

4.2 Greedy Extraction Method

As demonstrated in Section 3.2, Proposition 1 with respect to the submodularity of the objective function of $Re(R_s, R)$ leads to an appealing property, guaranteeing that a direct greedy method can be leveraged to find the solution to the representative review extraction problem formulated in Equation (1) with a satisfactory error bound of (1-1/e) compared to the optimal solution. Thus, a greedy extraction method named *Representative Reviews* (*RR*) is proposed in this study based on the calculated similarities between reviews in Section 4.1 to extract subset of reviews in terms of covering the whole review collection R

Concretely, at the first step, the representative subset R_s=∅. The *RR* method iterates through the review collection R to calculate the coverage value of $Re(R_s, R)$ when adding one single review to R_s. The review with the largest coverage value of $Re(R_s, R)$ is chosen as the first representative review and then is added to the subset of R_s. The chosen review is subsequently deleted from R.

At the second step, the *RR* method continues to iterate through R and chooses the second representative review with the largest coverage value of $Re(R_s, R)$ in the same way. The chosen review is subsequently added to R_s and deleted from R. The *RR* method continues to extract the other reviews from R in the same way until the number of reviews in R_s meets the requirement specified by users. The pseudo code of the *RR* method is listed in Table 1.

Algorithm *RR* (*Representative Reviews*)
Input: The review collection $R=\{r_1,r_2,...,r_n\}$. The number of representative k. Output: The extracted subset R_s. Begin: 1. R_s=∅; 2. Calculate-Sim(R); 3. while $\|R_s\|<k$ do 4. *maxvalue*=0.0; 5. for each review r_i in R do 6. *coverage*=Calculated-Degree ($R_e(R_s \cup \{r_i\}, R)$); 7. if (*coverage*>*maxvalue*) do 8. *maxvalue*=*coverage*; 9. *representative*=r_i; 10. end if 11. end for 12. $R_s=R_s \cup \{representative\}$; 13. $R=R\backslash\{representative\}$; 14.end while 15.Output(R_s);

Table 1. The Pseudo Code of the RR Method

In addition to a satisfactory error bound of (1-1/e), the *RR* method also possesses high computation efficiency in terms of a low polynomial complexity. As illustrated in Table 1, the Calculate-Sim in line 2 which calculates the similarity between each pair of reviews can be executed offline in advance.

Thus, there are two major loops for the online execution from lines 3-14. The outer iteration in line 3 is controlled by the parameter of k and the inner loop stating in line 5 is determined by the scale n of the review collection R. In each iteration of the inner loop, the core computation is the function of Calculated-Degree with a $O(kn)$ complexity. Therefore, the overall computation complexity is $O(k^2n^2)$. Since in most cases, the number of representative reviews is much smaller than the scale of the review collection R, i.e., $k<<n$, the dominating factor for the computation complexity is n, leading to a low polynomial complexity $O(n^2)$ for the *RR* method, which demonstrates the scalability of *RR* in dealing with large-scale reviews.

5 EXPERIMENTS

To comprehensively demonstrate the outperformance of the proposed method for representative review extraction, this section mainly introduces the experiments conducted on real online reviews as well as a user study to compare the proposed method *RR* with other existing review extraction methods in terms of effectiveness and efficiency.

5.1 Experiment Setup

The experiments of this study were conducted on Tmall (www.tmall.com) e-commerce platform, which is the largest B2C online shopping market in China (iResearch Consulting Group 2011) and the 17th most visited website all over the world according to Alexa (Alexa Internet 2015). The experiments were conducted by using a testing collection of 129 online products which were randomly chose and covered most categories of commodities in Tmall. For each product, all reviews were crawled to constitute a review collection.

In the experiments, three benchmark methods in a recent effort (Tsaparas et al. 2011) were involved, i.e., *Greey-U*, *Greedy-SU*, and *Greedy-GU*. All of these methods used the greedy strategy to extract subsets of reviews to cover all the features mentioned in the original review collection. Because the target of the proposed method *RR* is to extract a subset to cover the review collection, the coverage measure proposed by Ma & Wei (2012) was adopted in the experiment. It was specially designed to indicate the coverage degree of the small subset on the original large set on e-commerce platforms.

The experiment environment is a laptop with the operation system of 64-bit Windows 10 Professional, an Intel Core i7-3612QM CPU and 4 GB RAM, and all of the programs were implemented with the basic routines in Python.

5.2 Efficiency Experiments

On the practical e-commerce platforms, there are often hundreds of online reviews for a product. To test the efficiency of the proposed method in dealing with the review collection of that scale, the running time of *RR* on different scales is presented in this section. The testing data was generated by duplicating the review collection of a randomly chosen product to different scales of 100-1000. The proposed method *RR* was used to extract subsets of 10 reviews from the generated review collections with different scales. Figure 1 shows the running time of *RR*.

There are two major findings that can be derived from the testing results. Firstly, the trend of running times on different scales of review collections shows a low polynomial complexity, which is consistent with the theoretical analysis in Section 4.2. Secondly, the running times on the collections of 100 reviews and 1000 reviews are 0.15s and 15.56s respectively, indicating that the *RR* method can be directly used in practical online environment to provide representative subsets of reviews for users to quickly grasp main ideas in the original review collection. In addition, it is worth noting that the testing platform is just a laptop with limited computing capability. The execution efficiency can be dramatically improved by some powerful computing platforms, especially the parallel computing systems or cloud computing platforms.

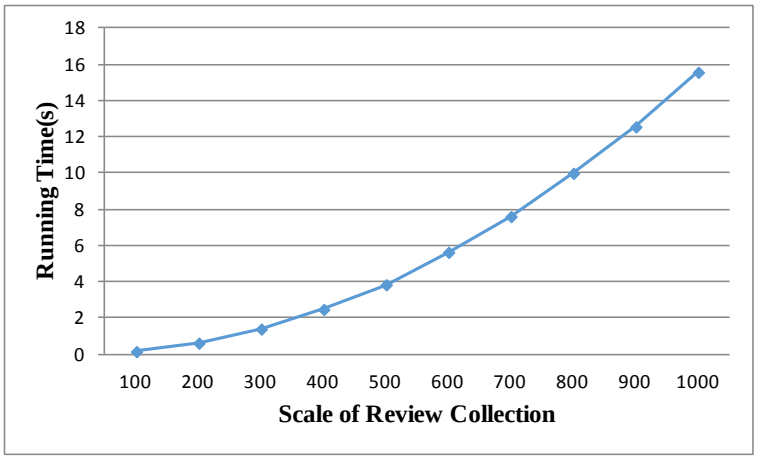

Figure 1. The Running Time of RR on Different Scales of Review Collection.

5.3 Effectiveness Experiments

The experiments with respect to effectiveness of *RR* mainly consists of two parts. The first part is about the coverage degrees of the extracted subsets with different numbers of reviews, aiming at providing an analysis of the coverage degree with the change of the extracted review number. Specifically, two products were randomly chosen from the 129 products. The *RR* method was used to extracted representative subsets with different scales for the two products and the corresponding coverage degrees were calculated. Figure 2 shows the coverage degrees for the two products.

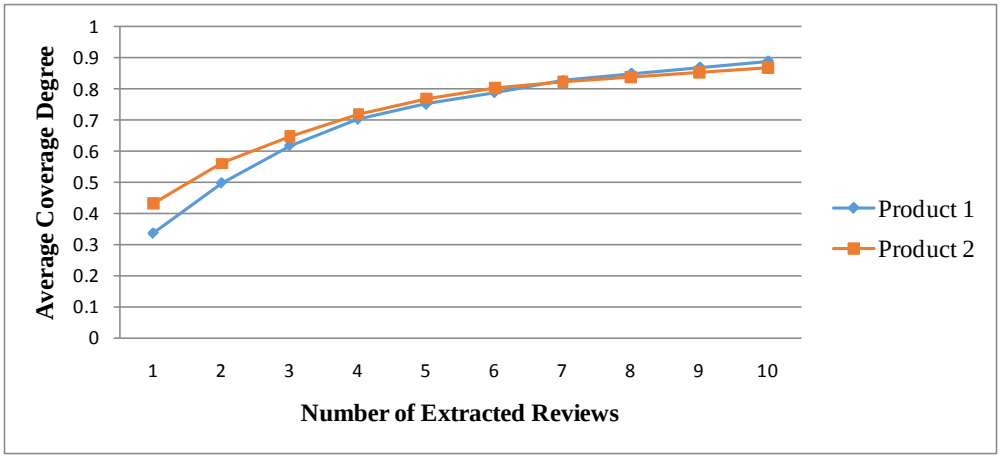

Figure 2. The Average Coverage Degrees of Extracted Reviews.

The testing results in Figure 2 reveal two properties of the *RR* method. The first is the coverage degree of the extracted subset increases with the number of extracted reviews becoming larger. This is consistent with the intuition. Given the original review collection, if the *RR* method is used to extract more reviews for consumers, they can obtain more content in the review collection. However, the increase of the coverage degree has a slower trend, which is the second property of the *RR* method. As illustrated in Figure 2, the increase trend slows down when the number of the extracted reviews exceeds 5 or 6. This property is consistent with the nature of the greedy *RR* method. In *RR*, the first few reviews are extracted due to their largest coverage degrees on the review collection. In other words, they can cover most content of the review collection. Therefore, the coverage degree that the following reviews contribute to the extracted subset becomes smaller. In addition, it is worth noting that since most consumers often comment on some main features of a particular product, there exist

some reviews expressing similar content in the review collection. Therefore, it can also be found in Figure 2 that the first extracted review can own a large coverage degree on the review collection.

In addition to the first part experiment focusing on the parameter of extracted review number in *RR*, the second part experiment was conducted on effectiveness comparison among *RR* and other benchmark methods, i.e., *Greedy-U*, *Greedy-SU*, and *Greedy-GU*. All the four methods were used to extracted subsets of reviews with the same size from the 129 products respectively. The size of the extracted subset in the experiment is 5, which is commonly adopted by recent related research efforts (Tsaparas et al. 2011; Lappas et al. 2012). The coverage degrees of the four methods are shown in Figure 3.

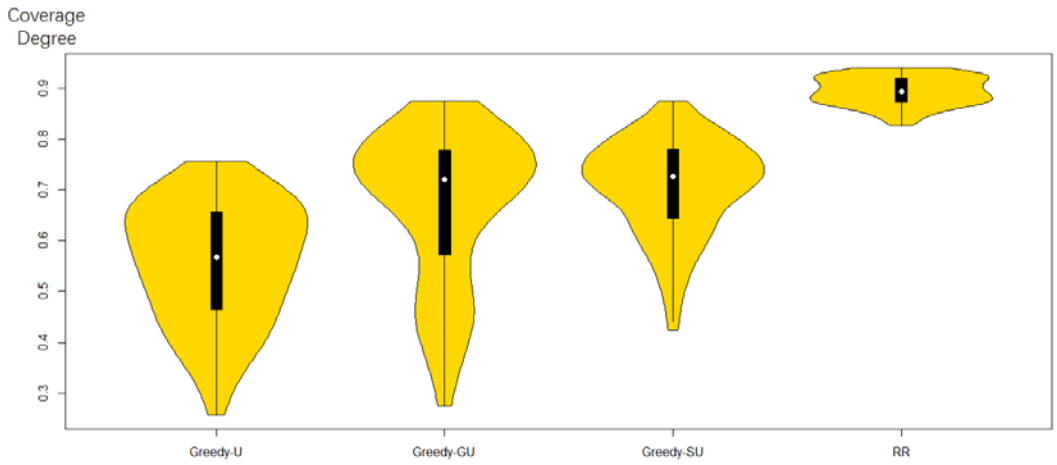

Figure 3. The Coverage Degrees of Four Extraction Methods.

In Figure 3, each while point represents the median of the 129 coverage values for each method, and the black line represents the interquartile range. The width represents the probability density of the data at the corresponding coverage value. For all the four methods, the maximum densities are normalized to the same width in Figure 3. It can be found from the testing results in Figure 3 that coverage degrees of the *RR* method are the largest for most of the testing products among all the four methods, demonstrating that the subset extracted by the *RR* method can cover more content for consumers to read. To further check the testing results, Table 2 shows the paired T-test statistical results.

Hypothesis	*T value*	*Significance*
Coverage Value of RR > Coverage Value of Greedy-U	29.786	***
Coverage Value of RR > Coverage Value of Greedy-GU	16.691	***
Coverage Value of RR > Coverage Value of Greedy-SU	21.049	***
Note: *: p<0.05, **: p<0.01, ***: p<0.001		

Table 2. Statistical Results of the Coverage Values With Respect to 129 Products

According to the statistical results shown in Table 2, the *RR* method performs significantly better than the other three benchmarks on coverage, revealing that it can be applied to extract a representative subset of reviews for consumers to quickly understand the whole picture of the original reviews.

5.4 User Study

To further demonstrate how useful the subset of reviews extracted by the proposed method is to actual users compared with the current state-of-the-art, a user study was performed in this section by

collecting human evaluations on the performances of different methods. Based on the Human Intelligence Tasks (HIT) conducted by Tsaparas et al. (2011), human evaluators were recruited to describe how informed they felt after reading different subsets of reviews.

In the user study, ten products were randomly chosen from the produce collection used in the effectiveness experiments to build the test data. To facilitate the user study, 30 reviews were randomly selected from all the original reviews for each chosen product. Moreover, the *RR* method and the other three benchmark methods were used to extract subsets of 5 representative reviews from 30 reviews of each product respectively. One typical test page showed the introduction to the product, the original 30 online reviews, and one subset of representative reviews. The subset was shown anonymously without specifying the name of extraction method to avoid bias.

32 undergraduate students in the field of management information systems were recruited in the user study, and all of them had sufficient knowledge and experience in online shopping. They were asked to read the test pages carefully. Then, they were required to provide scores for each review subset for the following two statements: "I. This review subset represents the entire set of reviews for the product well", "II. I feel well informed after reading this review subset". The scores range from 1(completely disagree) to 7(completely agree). Each evaluator was asked to complete the tasks for 1 or 2 products, resulting in a total of 38 accomplished tasks. The average scores provided by the evaluators with respect to the two statements are presented in Table 3.

Methods	*Statement I*	*Statement II*
Greedy-U	5.03	4.82
Greedy-GU	4.61	4.97
Greedy-SU	4.66	4.79
RR	5.55	5.16

Table 2. Human Evaluation Scores of the Four Extraction Methods

The testing results in Table 3 demonstrate some appealing findings. First of all, all the benchmark methods in the user study can provide a satisfactory subset of reviews for actual users in terms of covering the different content in the original collection of reviews, because all of the scores with respect to Statement I and II are more than 4.0. Secondly, compared with the other three existing methods, the results also indicate a human preference towards the representative subset provided by the proposed method *RR*. More importantly, these results are complementary to the effectiveness comparison experiments on real online reviews in Section 5.3. Although the methods of *Greey-U*, *Greedy-SU*, and *Greedy-GU* are proposed at the feature granularity which is different from the *RR* method to some extent, the results in Table 3 provide a fair comparison from the perspective of actual users who are the ultimate judges of all these methods. Concretely, on Statement I, the *RR* method has an average score of 5.55 which is more than the other three method, demonstrating that subsets extracted by *RR* are more informative. On Statement II, the *RR* method has an average score of 5.16, showing that human evaluators feel much more informed after reading review subsets extracted by *RR* and they can thus effectively make informed purchasing decisions.

6 CONCLUSIONS

It is deemed meaningful and desirable to extract a subset of reviews which can cover most content of the original review collection to facilitate the online shopping process for consumers. This study formulates the representative review extraction problem as an optimization model with an appealing submodularity property. In the model, the similarity between each pair of reviews is semantically calculated based on the divergences of topic distributions over the original review collection. Moreover, according to the model, a greedy extraction method named *RR* is proposed to extract a subset of representative reviews from the original review collection. Extensive experiments on real data from *Tmall* and a user study are conducted to demonstrate the outperformance of *RR* in terms of

efficiency and effectiveness over the benchmark methods, revealing the usability and applicability of the propose method in dealing with the information overload problem caused by large scale of online reviews.

The proposed method has good potential and can be applied as an online information processing tool on e-commerce platforms where consumers are unable to read the whole body of online reviews. When consumers have chosen a product they are interested in, they can firstly specify a number of representative reviews, and the *RR* method accordingly provide a subset of reviews for consumers. After reading the representative subset, consumers could quickly have definite ideas on whether to buy this product or switch to another one. The proposed method will help consumers tailor their shopping experience, and find exactly what they are looking for. Furthermore, it can also promote service quality and help to enhance customer satisfaction for e-commerce platforms.

Future studies can be focused on two limitations of this work. One is to take into account the redundancy among extracted reviews, which is also an important aspect of the extracted subset. The consideration of redundancy could be further leveraged as a possible way to automatically determine the number of the extracted reviews without user intervention. The other is to incorporate some other attributes of online reviews into the proposed model, such as the helpfulness values, the reputation of authors, and the temporal information, to widen the practicability of the current model.

Acknowledgments

The research of Jin Zhang Jilong Zhang, and Chong Ma was partially supported by the National Natural Science Foundation of China(71402186/71372044/71331007). The research of Baojun Ma was partially supported by the National Natural Science Foundation of China(71402007). The research of Ming Ren was partially supported by the National Natural Science Foundation of China (71302158).

References

Alexa Internet, Inc. (2015). http://www.alexa.com/siteinfo/tmall.com (accessed on June 1, 2015).

Archak, N., Ghose, A. and Ipeirotis, P.G. (2011). Deriving the pricing power of product features by mining consumer reviews. Management Science, 57(8), 1485-1509.

Baccianella, S., Esuli, A. and Sebastiani, F.(2009). Multi-facet rating of product reviews. Advances in Information Retrieval, pp.461-472, Berlin, Heidelberg, Springer.

Bickart B. and Schindler R.M. (2001) Internet forums as influential sources of consumer information. Journal of Interactive Marketing, 15(3), 31–40.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. IEEE Transactions on Information Theory, 49(7), 1858-1860.

Ghose, A. and Ipeirotis P.G. (2007). Designing novel review ranking systems: Predicting usefulness and impact of reviews. In Proceedings of the 9th International Conference on Electronic Commerce, pp.303-310, New York, ACM.

Griffiths, T. (2002). Gibbs sampling in the generative model of Latent Dirichlet Allocation, Technical Report, Stanford University.

Hochba, D. S. (1997). Approximation algorithms for NP-Hard problems. ACM SIGACT News, 28(2), 40-52.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.168-177, New York, ACM.

iResearch Consulting Group. (2011). iResearch China online shopping research report 2010-2011. http://www.iresearchchina.com/ reports/3772.html.

Kim, S.M., Pantel, P., Chklovski, T. and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp.423-430, Stroudsburg, PA, Association for Computational Linguistics.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 79-86.

Lappas, T., Crovella, M. and Terzi, E. (2012). Selecting a characteristic set of reviews. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 832-840, New York, ACM.

Lappas, T. and Gunopulos, D. (2010). Efficient confident search in large review corpora. In: Machine Learning and Knowledge Discovery in Databases, pp.195-210, Berlin, Springer.

Liu, B., Hu, M. and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th International Conference on World Wide Web, pp.342-351, Chiba, ACM.

Liu, J.J., Cao, Y.B., Lin, C.Y., Huang, Y.L. and Zhou, M. (2007). Low-quality product review detection in opinion summarization. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.334-342, Stroudsburg, Association for Computational Linguistics.

Liu, Y., Huang, X., An, A. and Yu, X. (2008). Modelling and predicting the helpfulness of online reviews. In Proceedings of 8th IEEE Conference on Data Mining, pp.443-452, Pisa, IEEE.

Lucian R., Moura, F.T., Durão, A.F. and Farias, S.A.D. (2007). Information overload on e-commerce. Ifip-the International Federation for Information Processing, pp.423-430, US, Springer.

Ma, B., & Wei, Q. (2012). Measuring the coverage and redundancy of information search services on e-commerce platforms. Electronic Commerce Research & Applications, 11(6), 560-569.

Ma, B., Wei, Q. and Chen, G.Q. (2011). A combined measure for representative information retrieval in enterprise information systems. Journal of Enterprise Information Management, 24(4), 310-321.

Ma, C.L., Wang, M. and Chen, X.W. (2015). Topic and sentiment unification maximum entropy model for online review analysis. In Proceedings of the 24th International Conference on World Wide Web. pp. 649-654, New York, ACM.

Miao, Q., Li, Q. and Zeng, D. (2010). Fine-grained opinion mining by integrating multiple review sources. Journal of the American Society for Information Science & Technology, 61(11): 2288-2299.

Mudambi, S.M. and Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.Com. MIS Quarterly, 34(1), 185-200.

Nemhauser, G. L., Wolsey, L. A. and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions. Mathematical Programming, 14(1), 265-294.

Pan, F., Wang, T. A. and Yang J. (2005). Finding representative set from massive data. In Proceedings of 5th IEEE International Conference on Data Mining, pp.338-345, Houston, IEEE.

Shimada, K., Tadano, R. and Endo, T. (2011). Multi-aspects review summarization with objective information. Procedia - Social and Behavioral Sciences, 27, 140–149.

Sun, M. (2012). How does the variance of product ratings matter? Management Science, 58(4), 696-707.

Thet, T.T., Na, J.C. and Khoo, C.S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 36: 823-848.

Tsaparas, P., Ntoulas, A. and Terzi, E. (2011). Selecting a comprehensive set of reviews. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.168-176, New York, ACM.

Zhang, J., Chen G.Q. and Tang X.H. (2012). Extracting representative information to enhance flexible data queries. IEEE Transactions on Neural Networks and Learning Systems, 23(6), 928-941.

Zhang, J., Wei, Q. and Chen, G. Q. (2014). A heuristic approach for λ-representative information retrieval from large-scale data. Information Sciences, 277: 825–841.

Zhang, Z. and Varadarajan, B. (2006). Utility scoring of product reviews. In Proceedings of the 15th Association for Computing Machinery International Conference on Information and Knowledge Management, pp.51-57, New York, ACM.