A Unified Graph Model for Chinese Product Review Summarization Using Richer Information

He Huang
School of Software
Tsinghua University
Beijing, 100084, China
huanghe09@mails.tsinghua.edu.cn

Chunping Li
School of Software
Tsinghua University
Beijing, 100084, China
cli@tsinghua.edu.cn

ABSTRACT

With e-commerce growing rapidly, online product reviews open amounts of studies of extracting useful information from numerous reviews. How to generate informative and concise summaries from reviews automatically has become a critical issue. In this paper, we present a novel unified graph model, composited information graph (CIG), to represent reviews with lexical, topic and together with sentiment information. Based on the model, we propose an automatic approach to address this issue. We use probabilistic methods to model the lexical, topic and sentiment information separately, associate with the discovered information in the CIG model, and generate summaries with a HITS-like algorithm called Mix-HITS considering both the *Representativeness* and *Proportion Approximation*. The experiments demonstrate that our method has improved performance over LexRank and ClusterHITS with Chinese and English datasets. Experimental results show that the proposed approach helps to build an effective way towards both the overall and contrastive summarization.

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.

Keywords

Review Summarization, Product Facet Detection, Sentiment Classification, Graph Ranking.

1. INTRODUCTION

With the wide spread of Internet and rapid growth of e-commerce, more and more people are getting used to shopping online. It is now a common practice for online merchants to provide a product review section to customers after purchase. Besides, there are many professional forums where many users discuss specified products online. Posting the reviews and opinions on the websites and forums becomes more and more popular. As a result, the number of reviews grows rapidly. Some popular products can even receive thousands of reviews in one day. This rich text information is of great reference value to both customers and manufacturers. The potential customers can choose more suitable products by referring to previous reviews. The manufacturers can develop and improve the products for real market by analyzing the advantages and disadvantages of their own products and monitor the competitors' products after receiving users' feedbacks.

However, it is difficult to use the Chinese product reviews posted by common Internet users directly. The reviews usually have the following characteristics:

(1) The amount of reviews is extremely large and increasing.

(2) Reviews are plain but short and informal Chinese text.

(3) There are many similar sentences among different reviews.

(4) Reviews are usually about several product facets, i.e., topics which are pre-defined by website or interested by users.

(5) Many sentences in reviews are opinionated.

Users may get an incomplete or biased view if they only read a few reviews. With the amount increasing, it is impossible to keep tracking and analyzing manually. How to automatically extract information from numerous online reviews becomes a critical issue.

Inspired by this, we study the problem of using summarization technology, combined with topic and sentiment analysis to address this issue. A summary is a text that is produced from one or more texts, which conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that [1]. Thus it can help people better understand the reviews. In addition, with the topic and sentiment information, we can obtain contrastive viewpoint on different product facets.

Considering some important sentences contains neither topic nor sentiment information, meanwhile a summary of contrastive viewpoints on specified product facet is more intuitive. Referring to Kim's research [2], we set the goal to generate two different types of summaries: the overall and the contrastive summaries.

Hu has described the product review summarization task [3]. In our research, given a set of customer reviews of a particular product, the task involves three subtasks:

(1) Identify facets of the product that customers have expressed their opinions on;

(2) Identify review sentences that give positive or negative opinions;

WISDOM'12, August 12 2012, Beijing, China.

(3) Produce both an overall and several contrastive summaries using the discovered information.

Let us demonstrate this case with an example. Assume that we are analyzing the reviews on a product, and we will obtain the summaries as Table 1 shows.

Table 1. Summaries of Product Reviews

<table>
<tr><th colspan="5">Overall</th></tr>
<tr><td colspan="5">Sentence 1
Sentence 2
Sentence 3
……</td></tr>
<tr><th colspan="2">Facet 1</th><th colspan="2">Facet 2</th><th>…</th></tr>
<tr><th>Positive</th><th>Negative</th><th>Positive</th><th>Negative</th><td rowspan="2">…</td></tr>
<tr><td>Sentence 1
Sentence 2
Sentence 3
……</td><td>Sentence 1
Sentence 2
Sentence 3
……</td><td>Sentence 1
Sentence 2
Sentence 3
……</td><td>Sentence 1
Sentence 2
Sentence 3
……</td></tr>
</table>

With the result, users can understand the overall and multiple aspects of the product better.

In Xu's point of view [4], a good summary should consider the representativeness and diversity besides aspect-relevance and sentiment intensity. In our work, we consider the *Representativeness* and *Proportion Approximation* as the requirements of a good summary. Representativeness means that the summary contains important lexical, topic and sentiment information. The meaning of proportion approximation here has two folds: 1) information diversity; 2) information proportion approximation, i.e., maintaining the proportions of topics and sentiments while compressing the reviews.

In this paper, we propose a novel approach to generate both the overall and contrastive summaries from multiple Chinese product reviews in a unified graph model, which will be explained in details in the following sections.

2. RELATED WORK

The product review summarization is close to opinion summarization. It's an extension of multi-document summarization (MDS) with other techniques such as topic modeling and sentiment analysis in the field of opinion mining.

Researches on summarization already have a longer history. The earliest work on summarization focused on scientific and technical documents. The following researches concentrated on many other domains, especially on newswire data. MDS has gained interest since mid-1990s, mostly on news articles. Several online news clustering systems were driven by research on it, such as Google News, Columbia NewsBlaster, and News In Essence, etc.

Existing summarization techniques mainly fall in two categories: extractive summarization and abstractive summarization. Most researches are in the extractive framework. It concerns with what the summary content should be by extracting the most important sentences.

A number of interesting approaches are graph-based. They build a graph to represent the text by establishing connections between text entities with meaningful relations, and then use link analysis algorithms to rank the sentences by simulating a "voting" between the vertices. LexRank [5] is a representative of these approaches. It measured content overlap between sentences using cosine similarity based on TFIDF and established links, and then utilized PageRank to rank the sentences. Wan improved the ranking algorithm by differentiating intra-document links and inter-document links [6]. He further proposed a manifold-ranking method to make uniform use of sentence-to-sentence and sentence-to-topic relationships [7]. In order to incorporate the document and cluster information, a conditional Markov Random Walk model and ClusterHITS model were proposed based on the two-layer link graph and the bipartite link graph [8, 9].

Another issue is the evaluation. Evaluating a summary is a difficult task since there doesn't exist an ideal summary for a given document set. Lin [10] introduced a set of recall based metrics called ROUGE that have become standards of automatic evaluation of summaries. And he further proposed an information-theoretic measure between distributions called JS divergence for evaluation [11].

In recent years, some researchers have focused on the opinion summarization [3, 12, 13]. Liu [12] presented a system called Opinion Observer to help analysts process the reviews. Liu [14] presented a system called CRO (Chinese Review Observer) for online Chinese product review structurization. It could collect reviews about a user-specified product through Internet, extracted opinions and product features from the reviews, identified implicit and synonym features, and conducted polarity analysis for each feature-opinion pair. The final result was visualized to tabular format and chart format. CRO considered the topic and sentiment information but didn't give a more detailed summary to represent the whole review set.

Most researches use topic and sentiment information as independent features for selecting sentences on a single facet. More recently, Titov [15], Lin [16] and Jo [17] studied to improve the analysis by combining topic and sentiment information using variations of topic models. These models can be used for summarization. However, they didn't discuss further. Xu [4] focused on the summarization in his research. He performed a Markov Random Walk in an aspect-sentiment graph to generate summaries.

Some other scholars also focused on contradiction detection. Lerman [18] proposed a summarization approach to model the difference between products. Kim [2] and Paul [19] researched on contrastive summarization of a single product.

Previous evaluation methods of review summarization are quite different from each other. Hu [3] used precision and recall metrics to evaluate the extracted sentences. Recently, Paul [19] and Xu [4] used ROUGE to evaluate the summaries.

Researches on Chinese opinion summarization started late. The main approaches were the same to the English tasks. However, the absence of unified evaluation resources and metrics in Chinese and the immature Chinese NLP technologies constrained the development of Chinese opinion summarization.

The closest work to ours is perhaps that of Xu and Paul. Xu proposed an aspect-sensitive Markov Random Walk Model using bag-of-words feature sets [4] and promoted the quality of overall summary. Paul used a topic-aspect model using lexical and syntactic features to incorporate the topic and viewpoint information, and proposed an improved comparative LexRank

[19], and generated contrastive summaries in both macro and micro levels. How our research differs from theirs is:

(1) We mainly focus on the issue of incorporating lexical, topic, sentiment and document information in a unified graph model.

(2) We study the problem of generating both of the overall and contrastive summaries in a hub-authority mutual reinforcement way.

3. THE PROPOSED APPROACH

3.1 Composited Information Graph Model

3.1.1 Main Idea

Most existed approaches make use of different information separately. Some state-of-the-art approaches intend to incorporate the topic and sentiment information in a unified topic model. However, it's a difficult task to design proper features.

We believe that the integration of richer information at the summarization stage will be effective. Thus, we introduce a novel unified graph model called composited information graph (CIG) to represent the reviews. The CIG model contains two different but relevant graphs, i.e., the basic graph and the mix graph.

3.1.2 Basic Graph

The basic graph is a widely used graph representation of sentences based on the lexical similarity.

Given a review sentence set *R*, let G_{basic}=(*V*, *E*) be a graph to represent the lexical relationships between sentences in *R*. *V* is the set of vertices and each vertex v_i in *V* represents a sentence in *R*. *E* is the set of edges, which is a subset of *V*×*V*.

This graph is closely related to the Markov Random Walk Model. By simulating a "voting" between the vertices, the saliency score of a sentence can be determined by received votes and the scores of the vertices casting these votes.

Figure 1 gives an example of the basic graph.

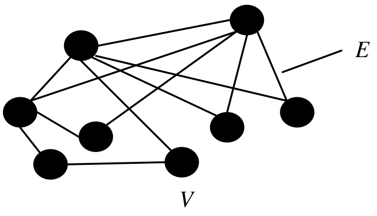

Figure 1. Basic Graph

3.1.3 Mix Graph

Apparently, in the basic graph we use only lexical information to obtain important scores of sentences on a sentence-level. We treat the sentences uniformly and ignore what they are about, where they are from.

Assuming that the topics and polarities are not equally important and can influence the importance analysis of sentences, to gain the representativeness and proportion approximation, we improve Wan's approaches [8, 9] and use the higher-level information in the CIG since both the topic cluster and sentiment polarity are the subset of sentences by different partitions. Furthermore, co-occurrence in the same document is an important mutual reinforcement. Moreover, some conjunctions may bring a significant shift of saliency among neighboring sentences. For example, the sentences following "不过 (however)" are more important than the preceding ones. Thus, we exploit additional links to reflect the mutual relationships between sentences in the same review. The following three cases are mainly considered:

(1) Co-occurrence. The sentences in the same review are associated to each other by the links with weights attached. The weights are decreasing while the distance is enlarging since the same sentence pair will have more influence on each other when they are closer.

(2) Shift. The links are bidirectional and with two weights. On most of the occasions the two weights are the same. However, if a conjunction that brings a shift of saliency occurs, several more important sentences in a window of the specific size will be marked. The links associated to these sentences will have a higher in-weight and a lower out-weight.

(3) Consensus. A higher weight will be assigned if the linked sentences express a similar opinion on the same product facet.

To sum up, we propose a mix graph model.

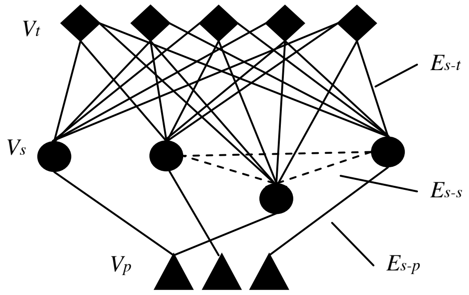

Figure 2. Mix Graph

As shown in Figure 2, the mix graph is denoted as G_{mix}=<V_s, V_t, V_p, $E_{s\text{-}t}$, $E_{s\text{-}p}$, $E_{s\text{-}s}$>, where V_s=*V* is the set of sentences (i.e. sentence authorities), V_t is the set of topics and V_p is the set of polarities (i.e. topic and sentiment hubs), $E_{s\text{-}t}=\{e_{ij}|v_i\in V_s, v_j\in V_t\}$, $E_{s\text{-}p}=\{e_{ij}|v_i\in V_s, v_j\in V_p\}$ corresponds to the correlations between sentences and topic and sentiment information, $E_{s\text{-}s}=\{e_{ij}|v_i, v_j\in V_s\}$ corresponds to the relations among sentences in the same review. Thus $|V_t|$ is equal to the number of topics corresponding to the product facets, and $|V_p|$ is equal to 3 since we only consider the positive, negative and neutral polarities.

3.2 Overview on Summarization

Our approach based on the CIG is mainly conducted in three steps. First of all, we split the review texts which are crawled from Internet into sentences, and extract specified features by analyzing the sentences in lexical and syntactic levels. After that, we model the lexical similarity, product facet and sentiment separately using the extracted features. Finally, we generate summaries by ranking the sentences and removing the redundancy in the CIG which incorporates the lexical, topic, sentiment and document information. Figure 3 gives the overall framework.

3.3 Chinese Processing

We can't model the crawled reviews immediately. The Chinese NLP technologies are required to transfer the reviews into the feature set as the input of probabilistic models. In our approach, each product review is first split into short sentences by commas.

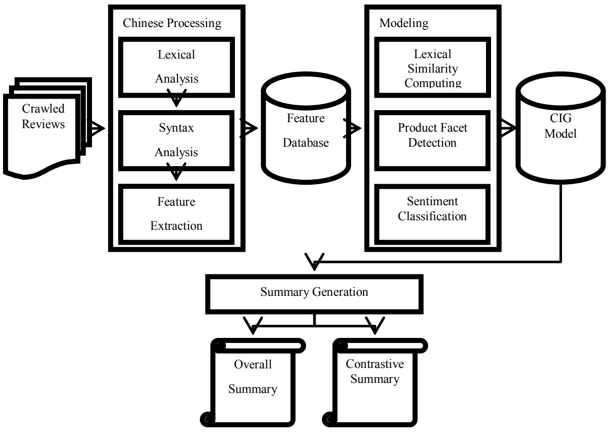

Figure 3. Framework for Chinese Product Review Summarization

3.3.1 Lexical Analysis

Since there are no separators between words in Chinese texts, we employ the ICTCLAS (http://ictclas.org/), which returns the word and POS tag sequences together.

3.3.2 Syntax Analysis

We input the word and POS tag sequences to the Stanford Parser (http://nlp.stanford.edu/software/) and obtain typed dependencies that represent the sentence with the highest probability. For this, we transfer the POS of ICTCLAS to corresponding tags according to a pre-defined mapping table.

3.3.3 Feature Extraction

From parsed results of sentences, we extract Bag-of-Words and dependency tree-based features. Besides, we preserve the document information and some contextual lexical information such as conjunctions for building the graph.

3.4 Review Modeling

We model the lexical, topic and sentiment information separately since we believe that different features should be used in different models. It is proved that the syntactic features can improve the accuracy of sentiment models. But for a topic model, word features are preferred since the word distribution representation is more intuitive.

3.4.1 Lexical Similarity

A review set can be viewed as a basic graph of sentences that are related to each other. The relations can be modeled as the similarity between sentences.

To define the similarity, we follow [5] and use the bag-of-words model with stop word removal. Then we use vector space model to represent each sentence as an *N*-dimensional vector, where *N* is the number of words in the vocabulary. The value of each dimension in the vector is the TFIDF of the corresponding word. The similarity between two sentences can be computed with the formula below.

$$sim(x,y)=\frac{\sum_{w\in x,y}\mathrm{tf}_{w,x}\cdot\mathrm{tf}_{w,y}\cdot(\mathrm{idf}_w)^2}{\sqrt{\sum_{x_i\in x}(\mathrm{tf}_{x_i,x}\cdot\mathrm{idf}_{x_i})^2}\cdot\sqrt{\sum_{y_i\in y}(\mathrm{tf}_{y_i,y}\cdot\mathrm{idf}_{y_i})^2}} \qquad (1)$$

where $\mathrm{tf}_{w,s}$ is the term frequency of the word *w* in the sentence *s* and idf_w is the inverse document frequency of *w*. Specially, we let *sim*(*x,x*)=0 to avoid self-related.

The review set then could be modeled as a cosine similarity matrix, where an entry is the similarity between the sentence pair. This model can measure the lexical content overlap.

3.4.2 Product Facet Detection

The product facets are usually abstract coarse-grained concepts. For example, "外观(exterior)" is a facet of notebook. But in the reviews, people often use words like "烤漆(paint)", "外壳(shell)" instead of the word "外观". Recently many studies represented topics with multinomial distribution of words. We employ the Gibbs-LDA (http://gibbslda.sourceforge.net/) based on the Latent Dirichlet Allocation (LDA) [20], to represent product facets using the distribution of words. We use the bag-of-words model without stop word removal to represent the review set, and give the topic number, and then put it into the Gibbs-LDA to model the topics in an unsupervised way. After modeling, we will obtain the most likely words and the topic-sentence distributions. By human judgment, we assign each topic a concept name.

However, the topics extracted in the automatic way are not reliable, since it's hard to align them to the users' interested topics and the accuracy is low.

3.4.3 Sentiment Classification

We assume that a sentence has three types of polarity, i.e., positive, negative and neutral. And we perform the sentiment classification in a supervised way by employing the libSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). Sun manually labeled 5244 reviews of notebooks with polarities to train a sentiment classification model using a tree kernel with sub-tree features [21]. In the classification stage, the model will output a value which indicates the probability. A sentence is attached with a label according to the interval where its value belongs to.

The experimental result on the ThinkPad data set which is employed in this paper shows the F-measure of the classifier is over 0.93.

3.5 Building the CIG Model

With the lexical, topic and sentiment information of sentences, we start to build the CIG model.

3.5.1 Basic Graph

With the lexical similarity matrix, we build the basic graph, G_{basic}=(*V*, *E*). An edge e_{ij} is added to *E* only if the similarity between sentences v_i and v_j ($i\neq j$) exceeds a threshold which is 0.0001 empirically. A transition probability weight w_{ij} is assigned to e_{ij} which could be computed using the lexical similarity as follows.

$$w_{ij}=\begin{cases}\dfrac{sim(i,j)}{\sum_{k\in V-\{i\}} sim(i,k)}, & \text{if } \sum sim\neq 0;\\ 0, & \text{otherwise.}\end{cases} \qquad (2)$$

3.5.2 Mix Graph

We use the topic, sentiment and document information as weights assigned to the mix graph, G_{mix}=<V_s, V_t, V_p, $E_{s\text{-}t}$, $E_{s\text{-}p}$, $E_{s\text{-}s}$>. Different weights are assigned to all of the vertices and edges, denoted as W_{mix}=<$w_{v\text{-}s}$, $w_{v\text{-}t}$, $w_{v\text{-}p}$, $w_{e\text{-}st}$, $w_{e\text{-}sp}$, $w_{e\text{-}ss}$>.

The weight of a vertex w_v indicates the vertex's importance in the belonging set. For each vertex v in V_t and V_p, a free weight is equally assigned as default which is computed with $w_{v\text{-}ti}$=1/|V_t| and $w_{v\text{-}pi}$=1/|V_p|. For each vertex $v\text{-}s_i$ in V_s, a normalized saliency score will be assigned to $w_{v\text{-}si}$ after G_{basic} has been analyzed.

Each vertex in V_s connects to all the vertices in V_t and one vertex in V_p, since the product facet model outputs a topic-sentence distribution while the sentiment model outputs only one polarity with a score for a sentence. The weights which are assigned to the edges in $E_{s\text{-}t}$ and $E_{s\text{-}p}$ denote the strength of the relationship between the sentence and the topic and sentiment information. For each edge *e-st*$_{ij}$ *in* $E_{s\text{-}t}$, the weight $w_{e\text{-}stij}$ is given by the probability distribution value of topic v_j on sentence v_i. For each edge *e-sp*$_{ij}$ *in* $E_{s\text{-}p}$, the weight $w_{e\text{-}spij}$ is given by the absolute value of the score of the sentence v_i for polarity v_j.

For each edge *e-ss*$_{ij}$ *in* $E_{s\text{-}s}$, the weights $w_{e\text{-}ssij}$ are computed with the formula below.

$$w_{e-ssij} = (basevalue)^c \cdot m \cdot d^{D_{ij}-1} \tag{3}$$

where c is the consensus degree, m is the magnification factor, d is the damping factor and D_{ij} is the distance between sentence v_i and v_j. Especially, for sentences from different reviews, D=∞. In this research, we set the *basevalue* to 0.25, d to *0.5*, and set c and m as below.

$$c = \begin{cases} 0.5, & \text{if } i \text{ and } j \text{ gain a consensus;} \\ 1.0, & \text{otherwise.} \end{cases} \tag{4}$$
$$m = \begin{cases} 0.5, & \text{if } i \text{ is more important than } j; \\ 2.0, & \text{if } i \text{ is less important than } j; \\ 1.0, & \text{otherwise.} \end{cases}$$

The maximum value of $w_{e\text{-}ss}$ will be 1.

3.6 Summary Generation

The summarization is based on the CIG model, using link analysis algorithms.

3.6.1 Graph Ranking

We propose the Mix-HITS, a HITS-like algorithm, to rank the sentences. In the HITS model [22], the hubs and authorities exhibit what could be called a mutually reinforcing relationship. In the mix graph, a good hub is a topic or sentiment polarity that points to many good authorities; a good authority is a sentence that is pointed to by many good hubs. In particular, we treat topic and sentiment polarity as similar hubs but separately.

The ranking is a two-stage ranking procedure.

3.6.1.1 LexRank in Basic Graph

LexRank [5] is used to decide the saliency of a sentence in the basic graph. A stochastic matrix M is used to describe G_{basic} with each entry $M[i, j]$ corresponding to the transition probability w_{ij}. And we replace the rows with all zero by a smoothing vector with all elements equaling to 1/|V|. So we can treat it as a Markov chain in the random walk way.

The saliency scores for sentences then can be determined in a recursive form as follows.

$$score(i) = \frac{1-d}{|V|} + d \cdot \sum_{j \in adj[i]}^{n} M[j,i] \cdot score(j) \tag{5}$$

where d is the damping factor and is set to 0.85 empirically.

Before we assign the saliency scores to the vertices in V_s of the mix graph, the scores are normalized by the formula below.

$$w_{v-si} = \frac{score(i)}{\max_{j \in V}(score(j))} \tag{6}$$

3.6.1.2 Mix-HITS Rank in Mix Graph

By now we have incorporated the lexical, topic, sentiment and document information in the mix graph. The authority scores of sentences, the hub scores of topics and sentiment polarities can be computed recursively, as Figure 4 shows.

```
1.Let $\vec{s}^{\,1} = \vec{e}$, $\vec{t}^{\,1} = \vec{e}$, $\vec{p}^{\,1} = \vec{e}$.
2. Repeat the following calculations $k$ times,
    $\vec{s}^{\,n+1} = L_t\overline{\mathrm{F}}(\vec{t}^{\,n}, \overrightarrow{w_t}) + L_p\overline{\mathrm{F}}(\vec{p}^{\,n}, \overrightarrow{w_p}) + L_s^T\overline{\mathrm{F}}(\vec{s}^{\,n}, \overrightarrow{w_s})$,
    $\vec{t}^{\,n+1} = L_t^T\overline{\mathrm{F}}(\vec{s}^{\,n} + L_s^T\overline{\mathrm{F}}(\vec{s}^{\,n}, \overrightarrow{w_s}), \overrightarrow{w_s})$,
    $\vec{p}^{\,n+1} = L_p^T\overline{\mathrm{F}}(\vec{s}^{\,n} + L_s^T\overline{\mathrm{F}}(\vec{s}^{\,n}, \overrightarrow{w_s}), \overrightarrow{w_s})$,
    $\vec{s}^{\,n+1} = \vec{s}^{\,n+1} / \| \vec{s}^{\,n+1} \|$,
    $\vec{t}^{\,n+1} = \vec{t}^{\,n+1} / \| \vec{t}^{\,n+1} \|$,
    $\vec{p}^{\,n+1} = \vec{p}^{\,n+1} / \| \vec{p}^{\,n+1} \|$,
    $n = n+1$,
   if $\| \vec{s}^{\,n} - \vec{s}^{\,n-1} \| < \varepsilon$ and $\| \vec{t}^{\,n} - \vec{t}^{\,n-1} \| < \varepsilon$ and $\| \vec{p}^{\,n} - \vec{p}^{\,n-1} \| < \varepsilon$,
      goto step 3.
3.Return $\vec{s}^{\,n}, \vec{t}^{\,n}, \vec{p}^{\,n}$.
```

Figure 4. Mix-HITS Algorithm

Where $L_{s|Vs| \times |Vs|}$, $L_{t|Vs| \times |Vt|}$ and $L_{p|Vs| \times |Vp|}$ denote the adjacency matrices of G_{mix} and are defined as follows.

$$L_s[i,j] = \begin{cases} w_{e-ssij}, & \text{if } (i,j) \in E_{s-s}; \\ 0, & \text{otherwise.} \end{cases}$$
$$L_t[i,j] = \begin{cases} w_{e-stij}, & \text{if } (i,j) \in E_{s-t}; \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$
$$L_p[i,j] = \begin{cases} w_{e-spij}, & \text{if } (i,j) \in E_{s-p}; \\ 0, & \text{otherwise.} \end{cases}$$

$\overrightarrow{w_s}$ =[$w_{v\text{-}si}$] $_{|Vs|\times 1}$, $\overrightarrow{w_t}$ =[$w_{v\text{-}ti}$] $_{|Vt|\times 1}$ and $\overrightarrow{w_p}$ =[$w_{v\text{-}pi}$] $_{|Vp|\times 1}$ denote the weights of vertices. $\vec{s}$=[$s\_score(s_i)$] $_{|Vs|\times 1}$、 $\vec{t}$ =[$t\_score(t_i)$] $_{|Vt|\times 1}$、 $\vec{p}$ =[$p\_score(p_i)$] $_{|Vp|\times 1}$ are the vectors of the authority scores of sentences, the hub scores of topics and sentiment polarities. k is the maximum number of iterations, and $\overline{\mathrm{F}}$ is defined as below.

$$\overline{\mathrm{F}}(\vec{v_n}, \vec{w_n}) = [v_1w_1, v_2w_2, \ldots, v_nw_n]^T \tag{8}$$

All of the initial scores are set to 1 and the iteration is used to update the scores until convergence. To guarantee the convergence of iterative form, we normalize the scores to maintain the invariant that their squares sum to 1.The convergence is achieved when the difference between the scores at two successive iterations is below a threshold ε which is 0.0001 empirically.

We select the sentences with larger scores as better authorities.

3.6.2 Redundancy Removal

The top ranked sentences using richer information have already satisfied the requirement of representativeness. Moreover, to gain diversity, we propose a redundancy removal algorithm which can choose sentences and reduce redundancy effectively.

```
1. Let $S=\varnothing$, $R=V$,where $V$ is the sentence set of the reviews
with $s\_scores$.
2. Sort $R$ by the $s\_scores$ in a descending order.
3. Repeat the following steps $S$ reaches the pre-defined length or
$R==\varnothing$.
    i) Choose the highest ranked sentence $s_i$,
        $S=S\bigcup\{s_i\}, R=R-\{s_i\}$.
    ii) For each $s_j$ in $R$, If $sim(i,j)>\lambda$ , $R=R-\{s_j\}$.
4. Return $S$ as the summary.
```

Figure 5. Redundancy Removal Algorithm

Where λ is a free threshold and set to 0.3 as default.

3.6.3 Parameters Tuning

We predict the different values of the weights w_{v-t}, w_{v-p} will bring interesting results such as a focus on a particular topic or polarity. In other words, the hub weights may influence the information proportion. An effective tuning strategy is a critical issue to improve the performance.

Moreover, we use a free threshold λ in the redundancy removal to find a balance between the representativeness and proportion approximation.

We will tune these parameters in the experiment.

3.6.4 Overall Summary

After Mix-HITS ranking, we obtain the scores of the hubs, too. The scores somehow represent the relative importance of topics and polarities. To gain proportion approximation, we give the hubs new weights according to the hub scores, and re-rank to generate an overall summary.

3.6.5 Contrastive Summary

In our model, by increasing a single polarity's weight, we can easily obtain a set of the biased opinions. We then do the same to the other polarity to obtain the set of the opposite opinions. We integrate the two sets to generate a macro contrastive summary is then generated.

By performing this procedure as well as tuning a topic's weight and removing irrelevant top sentences on other topics, we will obtain a micro contrastive summary on the facet.

4. EXPERIMENT AND EVALUATION

4.1 Data Set and Standard Summaries

We employ both the Chinese and English data sets for experiment. The Chinese product reviews on ThinkPad notebook are crawled from www.it168.com. We assign each sentence a unique facet and a sentiment polarity (+/-). There are 7 facets such as **Service**, **Quality**, **Price**, **Battery**, **Sound**, **Cooling**, and **None**. The data set contains 554 reviews, 1143 sentences.

We use Canon G3 and Nokia 6610 data sets of the publicly available English reviews from [3] to perform the English task. We preprocess the reviews by stemming and removing the stop words, and use the first topic and sentiment information of each sentence. Canon G3 contains 45 reviews, 597 sentences. Nokia 6610 contains 40 reviews, 546 sentences.

We manually generate overall summaries at different length of 20, 50, 80 sentences for Chinese data sets and summaries at the length of 50 sentences for English data sets as the standard summaries. We choose the sentence as a unit instead of words for information proportion evaluation. We ensure the summaries contain no redundant and unimportant information and have a similar topic and sentiment ratio to the whole data set.

We focus on the evaluation of the overall summarization. We only validate the contrastive summarization since there is no general acknowledged evaluation metrics for it.

4.2 Baselines

We choose LexRank and ClusterHITS as baselines.

LexRank: We have performed LexRank on the basic graph. We sort the sentences by the saliency scores, and remove the redundancy using the algorithm in Figure 5.

ClusterHITS: Another baseline is ClusterHITS [9], which use only topic and sentiment information.

Note that these baselines only generate overall summaries.

4.3 Metrics

We consider both the representativeness and the proportion approximation, which require minimum important information loss and a similar information proportion.

We use the ROUGE evaluation metric with a brevity penalty [10] to measure the representativeness.

$$ROUGE-N = BP\cdot\frac{\sum\limits_{S\in StdSum}\ \sum\limits_{n-gram\in S} Count_{match}(n-gram)}{\sum\limits_{S\in StdSum}\ \sum\limits_{n-gram\in S} Count(n-gram)} \tag{9}$$

$$BP=\frac{\text{length of standard summary}}{\text{length of candidate summary}} \tag{10}$$

We show two of the ROUGE metrics in the experimental results: ROUGE-1 (unigram-based) and ROUGE-2 (bigram-based).

And we use the Euclidean distance score (ED) between the information proportions to measure the approximation. Specially, we normalize the proportion using Z-SCORE first to put the data on the same scale. The formulas are given below.

$$ED=\sqrt{\sum_{i=1}^{|X|}(x'_i-y'_i)^2} \tag{11}$$

$$x'=\frac{x-\mu}{\sigma} \tag{12}$$

where $X=[x_1, x_2, \ldots, x_n]$ ($n=|V_t||V_p|$) is a proportion of different polarities on different topics, μ is the mean of x and σ is the standard deviation.

Higher ROUGE and lower ED scores will indicate a better performance.

4.4 Results and Analysis

We perform the Chinese task of overall summarization using LexRank, ClusterHITS, Mix-HITS^D with default weights and Mix-HITS^T with tuned weights. Table 2 shows the comparisons among these approaches with different metrics.

Table 2. Performance Comparison of Different Approaches

Approach	ROUGE-1	ROUGE-2	ED	Word Num
20 Sentences (228 words)				
LexRank	0.4487	**0.2518**	2.8165	269
ClusterHITS	0.3078	0.0925	2.1381	256
Mix-HITS^D	0.3975	0.1901	**1.3339**	291
Mix-HITS^T	**0.4760**	0.1997	2.1253	243
50 Sentences (577 words)				
LexRank	0.5027	0.3301	1.7182	678
ClusterHITS	0.4351	0.1454	2.7511	599
Mix-HITS^D	0.5261	0.3784	1.6419	652
Mix-HITS^T	**0.5621**	**0.3726**	**1.3130**	646
80 Sentences (987 words)				
LexRank	0.5822	0.4366	2.2252	1101
ClusterHITS	0.5819	0.2766	2.7899	968
Mix-HITS^D	0.6358	0.4954	1.9019	1063
Mix-HITS^T	**0.7241**	**0.5590**	**1.7726**	1063

Mix-HITS^T performs the best in all. Mix-HITS^D also gets better overall scores than LexRank. ClusterHITS which doesn't use lexical information performs the worst.

The result tells us that the lexical information is the most important information to gain both representativeness and proportion approximation, while adding topic, sentiment and document information is an effective way to improve. The comparison of ED scores proves our prediction that the hub information influences the information proportion.

Table 3 gives the overall summary of ThinkPad generated by Mix-HITS^T.

Table 3. Overall Summary of ThinkPad

1. Quality[+] 自带的系统很稳定，外观做的很不错，性价比比较高，散热做的不错。(Stable original system, good appearance, good price and nice cooling.)
2. Quality[+] 做工不错，钢琴烤漆，配置也不错，性价比好，散热好不错。(Decent workmanship with piano paint, good configuration, good price and good cooling.)
3. Service[-] 而且这样还不给换。(Not allowed to return under this condition.)
4. Battery[-] 电池不怎么好。(Poor battery.)
5. Quality[+] 外观不错散热不错加根一G内存在换成XP系统速度还可以。(Good appearance and good cooling. The performance could be good if 1G ram is added or the XP system is installed.)
6. Cooling[-] 就个散热好没别的。(Nothing but cooling is bad.)
7. Service[-] 我们希望联想会做得更好。(We hope that Lenovo can do better.)
8. Quality[-] 显卡太差了，我看这个还不如苹果的集成显卡，声卡我怀疑都不行。(The graphic card is unacceptable and I think it may be worse than the Macbook's. I even doubt the quality of the sound card.)
9. Quality[+] 键盘感觉很好。(The keyboard has a good texture.)
10. Quality[+] 总体上还是不错的、若不是拿来玩游戏的话，还是款不错的本本。(The overall quality of this notebook is decent, however, it is not suitable for playing games.)

The top ten sentences of the overall summary cover five different facets except **Sound** and **None**. And they are all opinionated and different from each other. We think it meets the requirements of Representativeness and Proportion Approximation.

We then tune the redundancy threshold λ from 0 to 0.5 and run the 80 sentence task with LexRank and Mix-HITS^T. Figure 6 and 7 show the ROUGE-1 and ED scores.

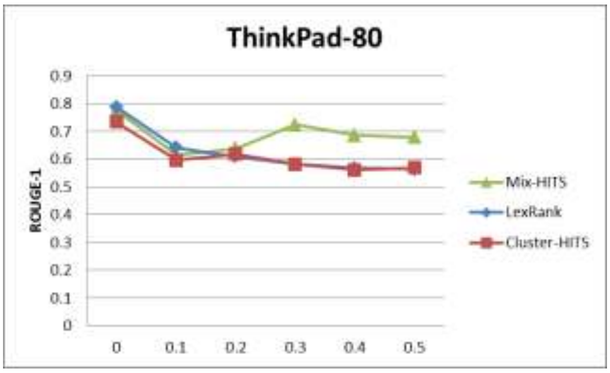

Figure 6. ROUGE-1 scores with different λ

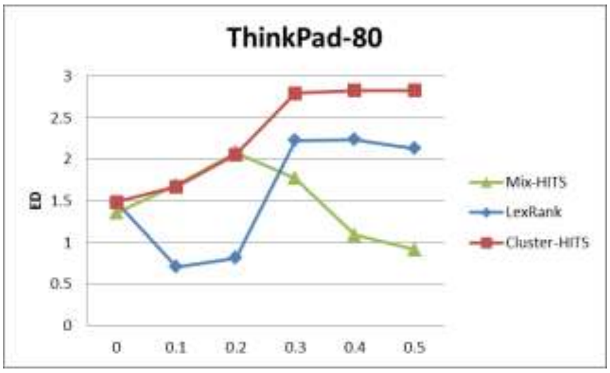

Figure 7. ED scores with different λ

The result shows that, for Mix-Rank, a higher λ which indicates a highlight on the diversity brings not only a better proportion approximation but also a stable representativeness. However, the parameter is ineffective for LexRank and ClusterHITS which get higher ED scores when λ is higher. Similar results are also obtained in 20 and 50 tasks.

In the English task, we use LexRank and Mix-HITS^D for comparison. Table 4 shows the result.

Table 4. Performance Comparison on English Reviews

Approach	ROUGE-1	ROUGE-2	ED	Word Num
Canon G3 (158 words)				
LexRank	0.5685	0.4868	3.7520	182
Mix-HITS^D	**0.9751**	**0.9132**	**0.6773**	152
Nokia 6610 (190 words)				
LexRank	0.7276	**0.6256**	**3.3799**	165
Mix-HITS^D	**0.7904**	0.4907	7.0081	125

Table 5 gives the overall summary of Canon G3 generated by Mix-HITS^D.

Table 5. Overall Summary of Canon G3

1. camera[+3][u]and so far , i 've been very pleased .
2. camera[+3][p]i highly recommend it .
3. picture[+3]the highest optical zoom pictures are perfect .
4. camera[+2]i love this camera .
5. camera[+2]this is a great camera for you !
6. camera[+3]it is a fantastic camera and well worth the price .
7. image quality[+3]the image quality is excellent .
8. camera[+2][p]overall i 'm happy with my toy .
9. camera[+2]am i ever glad that i decided on this camera !
10. camera[+3]this is my first digital camera and i could n't be happier .

The summary is not so diverse due to the facts that there are too many sentences about the general facet **camera.** The weight tuning is also a difficult task because there are too many unimportant facets. However, the comparison result on the English data set still suggests the potential of Mix-HITS.

In another task, we generate contrastive summaries using the approach in subsection 3.6.5. We give a micro result generated on ThinkPad's Quality as an example in Table 6.

Table 6. Contrastive Summary of ThinkPad's Quality

Positive	Negative
1. 做工不错，钢琴烤漆，配置也不错，性价比好，散热好不错。(Decent workmanship with piano paint, good configuration, good price and good cooling.)	1. 显卡太低，显示屏色彩也不太好，摄像头不过关，偏红。(The graphic card is low, the color of screen is not good and the camera is unacceptable.)
2. 自带的系统很稳定，外观做得很不错，性价比比较高，散热做的不错。(Stable original system, good appearance, good price and nice cooling.)	2. 显卡太差了，我看这个还不如苹果的集成显卡，声卡我怀疑都不行。(The graphic card is so bad and I think it may be worse than the Macbook's. I even doubt the quality of the sound card.)
3. 外观不错散热不错加根一 G 内存在换成 XP 系统速度还可以。(Good appearance and good cooling. The performance could be good only if 1G ram is added or the XP system is installed.)	3. 系统装的自带软件太多，每次开机都有点慢。(The system is slow due to the excessive pre-installed applications.)
4. 手感舒适，操作感觉很好。(Good texture and easy to operate.)	4. 钢琴烤漆容易脏，不过贴个膜就好，键盘比较小，有些键不好按。(The piano paint gets dirty easily without membrane. The keyboard is small and difficult to press.)
5. 外观稳重，大气，外壳及键盘手感好，LED 屏幕色彩好。(Decent appearance with fine shell, keyboard and superior LED screen.)	5. 不喜欢 Vista 系统，感觉慢。(The Vista system is slow, I don't like it.)

We also generate a macro contrastive summary of Canon G3, Table 7 gives the result.

Table 7. Contrastive Summary of Canon G3

Positive	Negative
1. picture[+3]the highest optical zoom pictures are perfect .	1. design[-3]this camera has a major design flaw .
2. image quality[+3]the image quality is excellent .	2. viewfinder[-2]* lens visible in optical viewfinder .
3. camera[+3]and so far , i 've been very pleased .	3. viewfinder[-2]you can see the lens barrel in the view-finder .
4. picture quality[+2]i love the quality of the pictures .	4. focus[-2], shoot[-2]it feels slow to focus , and unbearably slow to shoot .
5. camera[+3]i highly recommend it .	5. picture[-3]got way too many blurry pictures .

Apparently, the sentences in the result are informative. That's because the opposite opinions on the same sub-aspect reinforce each other during the generation and get higher saliency scores. This kind of summary is useful for both customers and manufacturers.

5. CONCLUSIONS

In this paper, we propose a unified graph-based model, CIG, to generate two different types of summaries by incorporating the lexical, topic, sentiment and document information. Furthermore, we propose an automatic approach based on the CIG model to summarize multiple Chinese product reviews, by using the topic model, sentiment model, graph model and other Chinese NLP techniques. The model achieved good performance in the experiments, and accomplished both the overall and contrastive summarization tasks.

In our future work, we plan to improve our approach further by focusing on the problems mentioned before, such as, how to find the optimal weights without human intervention, and how to design proper features for both the topic and sentiment analysis.

6. REFERENCES

[1] Radev, D. R., Hovy, E. and McKeown, K. 2002. Introduction to the special issue on summarization. *Computational Linguistics*. 28(4), 399-408.

[2] Kim, H.D. and Zhai, C.X. 2009. Generating comparative summaries of contradictory opinions in text. In *Proceedings of CIKM'09*. 385–394.

[3] Hu, M.Q. and Liu, B. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of KDD'04*. 168-177.

[4] Xu, X.K., Meng, T. and Cheng, X.Q. 2011. Aspect-based Extractive Summarization of Online Reviews. In *Proceedings of SAC'11*. 968-975.

[5] Erkan, G. and Radev, D.R. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*. 2004(22), 457-479.

[6] Wan, X.J. and Yang, J.W. 2006. Improved affinity graph based multi-document summarization. In *Proceedings of NAACL-HLT'06*. 181-184.

[7] Wan, X.J., Yang, J.W. and Xiao, J.G. 2007. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of IJCAI'07*. 2903-2908.

[8] Wan, X.J. 2008. An exploration of document impact on graph-based multi-document summarization. In *Proceedings of EMNLP'08*. 755-762.

[9] Wan, X.J. and Yang, J.W. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of SIGIR'08*. 299-306.

[10] Lin, C.Y. and Hovy, E. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of NAACL-HLT'03*. 71-78.

[11] Lin, C. Y., Cao, G. H., Gao, J. F. and Nie, J. Y. 2006. An Information-Theoretic Approach to Automatic Evaluation of Summaries. In *Proceedings of NAACL-HLT'06*. 463-470.

[12] Liu, B., Hu, M.Q., and Cheng, J.S. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of WWW'05*. 342–351.

[13] Hu, M.Q. and Liu, B. 2006. Opinion extraction and summarization on the Web. In *Proceedings of AAAI'06*. 1621-1624.

[14] Liu, H.Y., Yang, H., Li, W.B., et.al. 2008. CRO: a System for Online Review Structurization. In *Proceedings of KDD'08*. 1085-1088.

[15] Titov, I. and McDonald, R. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-HLT'08*. 308–316.

[16] Lin, C. H. and He, Y. L. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM'09*.375-384.

[17] Jo, Y. and Oh, A. 2011. Aspect and Sentiment Unification Model for Online Review Analysis. In *Proceedings of WSDM'11*. 815-824.

[18] Lerman, K. and McDonald, R. 2009. Contrastive summarization: an experiment with consumer reviews. In *Proceedings of NAACL'09*. 113-116.

[19] Paul, M.J., Zhai, C.X. and Girju, R. 2010. Summarizing Contrastive Viewpoints In Opinionated Text. In *Proceedings of EMNLP'10*. 65-75.

[20] Blei, D., Ng, A. and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*. 3(5), 993-1022.

[21] Sun, Z. and Li, C.P. 2011. *A Dependency Tree Kernel for Sentiment Classification*. Technical Report. Tsinghua University.

[22] Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*. 46(5), 604-632.