Understanding Consumer Preferences---Eliciting Topics from Online Q&A Community

(Guan Liu, School of Management, Jinan University, liuguan@jnu.edu.cn
Ying Wei*, School of Management, Jinan University, yingwei@jnu.edu.cn
Feng Li, South China University of Technology, fenglee@scut.edu.cn

ABSTRACT
Online Q&A community provides a platform to create and share knowledge by posting and answering questions. Soliciting information from Q&A thus offers an alternative approach for firms to understand their consumers. This paper proposes an innovative approach to extract consumer preferences based on the online Q&A information. We develop a framework to conduct the analysis and employ Latent Dirichlet Allocation (LDA) algorithm to distill and cluster topics. Based on Zhihu, the most popular online Q&A community in China, we collect almost 50000 answers under the discussion topics of “iPhone 7” and “iPhone X”. We find that our approach can effectively extract and rank consumer preferences to the product. In addition, we find those preferences are inter-related.

Keywords: Consumer preference, online Q&A community, topic model, word co-occurrence.

*Corresponding author

INTRODUCTION
In the past few years online Q&A communities have grown explosively into important knowledge-sharing platforms, for example, Quora and StackOverflow in the United States, and Zhihu in China. Online Q&A community provides a platform to create and share knowledge by posting and answering questions. Topics in the platforms cover a wide range of subjects via the form of questions, opinions, experiences and comments, etc. By posting (or searching) questions and collecting answers, users can quickly learn and adopt the domain knowledge relating to their concerns, most of which are first-hand answers from domain experts (Wang et al., 2013).

On the flip side, soliciting information from Q&A in the platform also provides an alternative approach for firms to understand their consumers. Understanding consumer preferences is important to a firm. It helps to determine a firm’s business strategy and identify its competitive position. However, to understand consumer is not easy although firms employ different approaches to do so. For example, a consistent narrative was that Apple Inc. was struggling to compete with local brands and that its flagship iPhone X model was too expensive for Chinese consumers (Ben, 2018). In contrast, according to the most recent report of Apple, in the year of 2018 Q2 has reached a 21% growth in Great China, which is the strongest growth rate in 10 quarters.

This paper proposes an innovative approach to understand consumer preferences based on the online Q&A information. Unlike the short-text posted in micro-blog (Li & Du, 2014), the answers in online Q&A community are in general longer than micro-blog. This thus calls for new approaches to handle the user-generated contents (UGC) for topic extraction. Our paper proposes a framework to conduct the text analysis and employ LDA algorithm (Blei, Ng, & Jordan, 2003), a natural language processing algorithm, to distill and cluster topics from the online Q&A information. LDA is an unsupervised machine learning algorithm, based on “bag of words” pattern, which uses the co-occurrence to measure the global relationships between words (Weng et al., 2010). LDA is widely used to discover hidden topics from text corpus and generates document-topic (documents are modeled as a mixture of topics) relationship, as well as topic-word (a topic is contributed by words) relationship. By LDA, it can automatically identify topics in which platform users take interests based on the contents they generate.

Based on the most popular online Q&A community in China, Zhihu, we first crawl the textual content of answers under the platform topics of “iPhone 7” and “iPhone X”. Then, after data preprocessing we use LDA algorithm to analyze the texts and extract the clustered topics. Our findings are summarized below. First, we find that LDA approach can effectively extract consumer concerns fitting the product’s physical attributes. With Zhihu’s online Q&A information, the clustering topics can be labelled as “camera”, “screen”, “charging”, “handset”, etc. In addition, LDA also discloses that some non-physical attributes, for example, “life style”, “brand reputation”, contribute to the concerns of the consumers. Second, by LDA we find that those topics are not equally important and consumers have different preferences to the attributes. This thus helps firms to understand the rank of the concerns and disclose the contributions of each attribute to the adoption of the product. In addition, we find that those concerned topics are inter-related. For example, topic “display” is related to topic “experience” considering both topics contain the keyword of “screen”.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents the framework of the methodology. Section 4 depicts the dataset and the procedure to conduct the analysis. Section 5 reports the findings and section 6 concludes the whole paper.

RELATED LITERATURE

Studies relating to this paper can be divided into three categories: understanding consumer preferences, text-based analysis of user-generated content (UGC) and topic extraction models.

Understanding consumer perceptions and preferences about products is important in designing goods and service products, making communication strategies and brand decisions. One of the most popular methods to understand consumer preferences is conjoint analysis (Green & Srinivasan, 1978). Green and Srinivasan (1990) introduce a model to evaluate customer preferences by using party-worth utilities of products’ attribute levels. Traditionally, firms try to understand consumer needs by questionnaires and real sales data, and marketers can decide which attributes of products are important and which one should be chose (Louviere & Hout, 1988). Therefore, it is useful for marketers to improve product configurations, predict market response to new products, and diagnose competitive strength (North & Vos, 2002).

Text mining in UGC was largely ignored in business research decades ago because of the lack of practical tools to analyze those unstructured data (Lee, 2007; Archaik, Ghose, & Ipeirotis, 2007). However, due to the exponential growth of textual data, text-based analysis of UGC has attracted considerable attention in recent years. It has been widely applied to discover key product attributes (Lee & Bradlow, 2011), to develop product recommendations (Ghose, Ipeirotis, & Li, 2012), to predict the sales of products (Dellarocas, Zhang, & Awad, 2007), as well as to determine market structure (Netzer et al., 2012). A large number of studies use econometrics and psychometric tools to conduct the research (Dellarocas, Zhang, & Awad, 2007; Decker & Trusov, 2010).

Another closely related research is studies on how to extract topics from UGC. According to Tirunillai and Tellis (2014), the specific words in UGC are only indicators of latent and priori unknown topics, which are defined by a cluster of words with a high frequency or probability of usage. This is the core idea of latent topic modeling in LDA (Blei, Ng, & Jordan, 2003) and the author–topic model (Rosen-Zvi et al., 2012). LDA is one kind of the Probabilistic Topic Model (Steyvers, Griffiths, & Dennis, 2006), which is a widely used technique for text analysis. LDA combines those similar words into a cluster based on the frequency of occurrence of words in each document. It is built on a three-layer Bayesian network to calculate the topic probability distribution of the words. According to “bag of words” pattern, the algorithm consists of the following assumptions, (a) given a corpus, (b) built a dictionary comprised by unique tokens appeared in this corpus at least once, (c) each document is represented as a probability distribution over some topics, and (d) each word in the document is generated by a specific topic (Weng et al., 2010).

There has been an increasing trend in applying text mining and topic extraction techniques to business decisions in recent years. For example, Lee et al. automatically derive market structure from consumer reviews collected from epinions.com using an optimization approach (Lee & Bradlow, 2011); Ghose and Ipeirotis (2011) combine multi-attribute choice models with multiple text analysis methods to examine the influence of product reviews on consumers’ product choice decisions. Netzer et al. (2012) define similarity between products based on exploiting the co-occurrence of words and semantic network analysis. Büschken and Allenby (2016) propose an LDA model that uses the sentence structure of reviews to improve prediction of consumer ratings. Our work contributes to the literature by proposing an innovative approach to understand consumer preferences with online Q&A information, based on an LDA model.

THE FRAMEWORK

In this section, we first make a brief description of Zhihu.com, China’s largest online Q&A community. We then propose a two-step framework to conduct text analysis and topic extraction, namely, (1) information preparation and (2) topic Modeling. Step one collects the UGC data from Zhihu platform and transforms the unstructured texts to a structured form that can be analyzed in the next stage. In step two, the structured data are input to the classic topic model to identify and cluster the topics through the Latent Dirchlet Allocation Algorithm. Figure 1 depicts an illustration.

Introduction of Zhihu.com

Zhihu is the most popular online Q&A community in China, where users can ask and answer questions and comment on or vote for existing answers. Platform users can search questions by keywords, or find their interests or relevant knowledge by topics and users. Zhihu also allows users to follow each other and share their knowledge via social connections. As in Twitter, users can follow anyone without explicit permission. If user A have followed user B, B’s actions, including new questions, answers, comments and votes, will push to A’s timeline stream. In this way, A is B’s follower or fans and B is A’s followee or friend.

Step one. Information preparation

As mentioned, the framework consists of two steps. The first step, so-called information preparation, collects UGC data from Zhihu platform and conducts data preprocessing to transform the unstructured texts to texts with structured form, as shown in Figure 1.
Figure 1: The two-step framework to conduct text analysis and topic extraction

Data collection
In Zhihu, each discussion topic generated by the platform has its own homepage, which includes all related questions and answers. Figure 2 presents a snapshot of a Zhihu topic’s homepage. A page consists of five parts: (1) introduction, (2) number of followers and questions, (3) active users, (4) topic tags and information of questions and (5) answers under this topic. Under this topic homepage, all questions and answers could be gathered by the URL API of Zhihu.com with a web-based crawler tool. We develop a tool with python (version 3.6) to crawl the data under the topics’ homepages.

Figure 2: The homepage of topic “iPhone X” in Zhihu.com.

The crawling process is also shown in Figure 1. We first crawl all the questions’ information under the topic. Each question contains information such as author profile, title, content, followers count, comments count, answers count, and creating time, etc.. Then we collect all the answers’ information to each question, including answer contents, author profile, voter count, comment count, etc. Considering our purpose is not to match the answers to the questions but to understand topics of user interests, we aggregate all the answers delivered by an individual user into a text document following Weng et al. (2010). Thus, a document essentially corresponds to an individual user’s generated content.
Data preprocessing

With each individual user’s document prepared, the first step of data preprocessing is to transform the unstructured text documents into a structured form that is amenable to analysis (Mankad et al., 2016). We employ a popular transformation approach with the “bag of words” representation of text, where the set of documents to be analyzed (usually referred to as a “corpus”) is represented by a document-term matrix (Kosala & Blockeel, 2000). This document-term matrix is a structured table with two dimensions, the column for each term and the row for each document in the corpus. Each element in the matrix is a count of the number of times the corresponding word appears in each document.

Thus, the analyses based on text mining require special preprocessing of the corpus, where the general objective is to emphasize meaningful words by removing uninformative ones and to keep the number of unique terms that appear in the corpus from becoming extremely large (Mankad et al., 2016). We conduct the data preprocessing of Chinese words into seven steps, as indicated in Figure 3: (1) removing all the html tags and urls in documents; (2) customizing dictionary and stop words (very common words); (3) detecting new words and word segmentation; (4) transforming all English text into lowercase; (5) gathering all answers posted by the same author into one document; (6) removing exceptional words that occur either too frequently or very rarely; (7) establishing a unique token dictionary of corpus where each word corresponds to a unique index number; and finally (8) transforming all the documents to a document-term matrix.

Step two: Topic modelling

With the structured texts as input, step two aims to generate clustering topics. Topic modeling is a type of text-mining algorithms that can be used to discover latent topics automatically. A simple model of the topic model is the LDA model. One important parameter of LDA model is the number of latent topics that appear across multiple documents. It determines the data granularity of result by the clustering words that appeared in each document.

In this framework, the LDA model plays a pivotal role due to the core idea behind this method is that latent and priori unknown topics are defined by a cluster of words with a high frequency or probability of usage. In addition, each document is regarded as a mixture of various topics where each document is considered to have a set of topics that are assigned via LDA. Consequently, there are two probability distributions generated by LDA. The “document-topic” distribution describes the probability of topics occurring in a given document. Since all documents share the same topic set, which is determined by the number of latent topics, and each document has a different probabilistic mixture of those topics (Mankad et al., 2016). The “topic-word” distribution describes the probability of words occurring in a given topic. An underlying assumption is that documents covers only a small set of topics and that topics use only a small set of words frequently. LDA assumes a generative process infer the latent topics in a corpus as indicated in the following steps: (a) choosing the document by a prior parameter “alpha”, picking a topic from its “document-topic” distribution, (b) associated with the selected topic, sampling a word by a hyper-parameter “beta” from the “topic-word” distribution, (c) repeating the process for all the words in corpus until the sampling is convergent. The two hyper-parameters, alpha and beta, are priori parameters which could be inferred by experience or existing work. As alpha gets smaller, the document is close to one topic. Similarly, as beta gets smaller, all words tend to belonging to the same topic.

DATA AND METHODS

Data Collection

We developed a python (version 3.6) project to implement the procedure. Some python libraries were called in the program, as demonstrated below. The web crawler was created by Scrapy 1.4, a fast and powerful scraping and web crawling framework in python. Gensim was a python library for document indexing and topic modelling. We used Matplotlib 2.2.3 to deliver the figures of...
the result, which was a Python 2D plotting library. All the data were stored in Mongodb 3.2, a document database that can be used to flexibly store the JSON data, considering the crawling results from API of Zhihu.com were JSON-type documents.

Table 1: The number of answers and users from dataset.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>iPhone 7</th>
<th>iPhone X</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of answers</td>
<td>16240</td>
<td>35809</td>
<td>52409</td>
</tr>
<tr>
<td>Number of users</td>
<td>11493</td>
<td>25587</td>
<td>35888</td>
</tr>
</tbody>
</table>

We used an open application-programming interface (API) provided by Zhihu.com to crawl data. We collected all the answers under the platform topic of “iPhone 7” in the period between June 10, 2016 and July 21, 2017, and the topic of “iPhone X” in the period between July 21, 2017 and April 2, 2018. The number of answers and users are listed in Table 1. Figure 4 plots the number of answers dynamically changes as time proceeds monthly. We implemented the preprocessing steps as described in section “The framework”.

After data preprocessing we then obtained a refined subset of all the answers, as a document-term matrix. Only answers containing more than 15 terms were retained and constructed as a document-term vector. Table 2 exhibits a processing flow table of the refining process. We further refined the dataset by first gathering all content posted by a user into a document, and then removing exceptional words that occur too frequently or too rarely. Although after data preprocessing the number of answers has been reduced from 52409 to 23794, the remaining data is more informative and abundant enough to conduct LDA analysis.

Table 2: The processing flow of dataset.

<table>
<thead>
<tr>
<th>Processing sequence</th>
<th>Before</th>
<th>Preprocessed</th>
<th>Aggregated</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of answers</td>
<td>52409</td>
<td>23794</td>
<td>19512</td>
<td>19512</td>
</tr>
<tr>
<td>Number of authors</td>
<td>35888</td>
<td>17062</td>
<td>17062</td>
<td>17062</td>
</tr>
<tr>
<td>Total terms</td>
<td>1863409</td>
<td>1863409</td>
<td>1111374</td>
<td></td>
</tr>
<tr>
<td>Average number of</td>
<td>78.4</td>
<td>95.5</td>
<td>56.96</td>
<td></td>
</tr>
<tr>
<td>terms per answer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of Dictionary</td>
<td>69542</td>
<td>69542</td>
<td>7426</td>
<td></td>
</tr>
</tbody>
</table>

**Topic modelling analysis**

We used a python library, called “Gensim”, to implement the LDA algorithm and applied Gibbs sampling to compute the whole data. The LDA model is conditioned on three parameters: two hyper-parameters alpha and beta, and the number of topics, T. In order to perform topic modeling effectively, we set the parameter values following (Griffiths & Steyvers, 2004), who suggested to fix alpha
to 50/T, beta to 0.1 and change the number of topics in a range from 8 to 20. Here we set the number of topics to be 15. Table 3 lists part of the results, that is, the top 20 high-frequency words in six clustered topics among the fifteen topics.

To have an overview of the output clustering topics, we used “Gephi (0.91)” to visualize the topics and their high-frequency words’ network in Figure 5. As shown in the figure, some high-frequency words contribute to more than one topic and thus topics are connected due to the common words.

Table 3: High-frequency words in each topic.

<table>
<thead>
<tr>
<th>Term</th>
<th>Topic 2</th>
<th>Topic 4</th>
<th>Topic 11</th>
<th>Topic 12</th>
<th>Topic 13</th>
<th>Topic 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Advertising</td>
<td>Charging</td>
<td>Unlock</td>
<td>Headset</td>
<td>Feature</td>
<td>Screen</td>
<td></td>
</tr>
<tr>
<td>2  Shoot</td>
<td>Wireless</td>
<td>Key</td>
<td>airpods</td>
<td>lens</td>
<td>iPhone x</td>
<td></td>
</tr>
<tr>
<td>3  Shootiing</td>
<td>Battery</td>
<td>home</td>
<td>interface</td>
<td>Effect</td>
<td>Full screen</td>
<td></td>
</tr>
<tr>
<td>4  Three minutes</td>
<td>Use</td>
<td>Face id</td>
<td>Sound quality</td>
<td>inside</td>
<td>Bangs</td>
<td></td>
</tr>
<tr>
<td>5  Movie</td>
<td>Customer</td>
<td>Identification</td>
<td>cancel</td>
<td>camera</td>
<td>Design</td>
<td></td>
</tr>
<tr>
<td>6  iPhone x</td>
<td>Chip</td>
<td>Screen</td>
<td>Bluetooth headset</td>
<td>photos</td>
<td>iPhone x</td>
<td></td>
</tr>
<tr>
<td>7  Chen Kexin</td>
<td>Support</td>
<td>Fingerprint</td>
<td>bluetooth</td>
<td>use</td>
<td>Technique</td>
<td></td>
</tr>
<tr>
<td>8  Story</td>
<td>Fast charge</td>
<td>Operation</td>
<td>Listen to music</td>
<td>plus</td>
<td>Face id</td>
<td></td>
</tr>
<tr>
<td>9  Video</td>
<td>Usage</td>
<td>Feature</td>
<td>Listen</td>
<td>advance</td>
<td>Screen</td>
<td></td>
</tr>
<tr>
<td>10 Lens</td>
<td>Device</td>
<td>Need</td>
<td>Wireless headset</td>
<td>shooting</td>
<td>Oled</td>
<td></td>
</tr>
<tr>
<td>11 Short movie</td>
<td>Hour</td>
<td>Password</td>
<td>connect</td>
<td>camera</td>
<td>Camera</td>
<td></td>
</tr>
<tr>
<td>12 SLR</td>
<td>Need</td>
<td>Face</td>
<td>take</td>
<td>device</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>13 Director</td>
<td>Wire</td>
<td>Interface</td>
<td>music</td>
<td>support</td>
<td>Frame</td>
<td></td>
</tr>
<tr>
<td>14 Children</td>
<td>Charger</td>
<td>Touch id</td>
<td>sound</td>
<td>ar</td>
<td>Question</td>
<td></td>
</tr>
<tr>
<td>15 Profession</td>
<td>Technique</td>
<td>application</td>
<td>case</td>
<td>pixes</td>
<td>Display</td>
<td></td>
</tr>
<tr>
<td>16 Mother</td>
<td>Interface</td>
<td>app</td>
<td>Headset plug</td>
<td>game</td>
<td>Fingerprint</td>
<td></td>
</tr>
<tr>
<td>17 Inside</td>
<td>Standard</td>
<td>Can not</td>
<td>Easy</td>
<td>video</td>
<td>Glass</td>
<td></td>
</tr>
<tr>
<td>18 Film</td>
<td>Endurance</td>
<td>id</td>
<td>Inside</td>
<td>both</td>
<td>Influence</td>
<td></td>
</tr>
<tr>
<td>19 Photography</td>
<td>Ad</td>
<td>setting</td>
<td>lightning</td>
<td>technique</td>
<td>Face identification</td>
<td></td>
</tr>
<tr>
<td>20 Device</td>
<td>Question</td>
<td>use</td>
<td>charging</td>
<td>different</td>
<td>Samsung</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: The high-frequency words’ network based on LDA topics.
FINDINGS AND RESULTS

We summarize our findings and interpret the results as follows. First, based on the list of high-frequency words in Table 3, we categorize them into different topics about iPhone product and interpret them as customer preferences as follows.

- **Topic 2 (Advertising).** The high-frequency words such as “shoot(拍摄),” “lens(镜头),” “advertising(广告),” “three minutes(三分钟)”, and “Chen Kexin(陈可辛)” are associated with a popular three-minute video taken by Chen Kexin, a famous Chinese movie director, with his brand-new iPhone; thus, it could be labeled as “Advertising”.
- **Topic 4 (Charging).** The high-frequency words of this topic include “charging(充电),” “battery(电池),” “Fast charge(快充),” and “charger(充电器)” and associated with the charging feature of iPhone. It means that some customers are concerned about the charging feature of mobile phone.
- **Topic 11 (Unlock).** The high-frequency words of this topic include “Unlock(解锁),” “Home,” “Face id,” and “fingerprint(指纹)” and associated with the unlock feature of iPhone.
- **Topic 12 (Headset).** The high-frequency words of this topic include “Headset(耳机),” “airpods,” “Sound quality(音质),” and “Bluetooth headset(蓝牙耳机)” and associated with the headset feature of iPhone.
- **Topic 13 (Camera).** The high-frequency words of this topic include “lens(镜头),” “camera(摄像机),” “photos(照片),” and “shooting(拍摄)” and associated with the camera feature of iPhone.
- **Topic 14 (Screen).** The high-frequency words include “Screen(屏幕),” “Oled,” “Face identification,” “Frame(边框),” “Glass(玻璃)” and “Bangs(刘海)” and are associated with the screen feature of iPhone. It means that some customers are concerned about the screen of mobile phone.

According to the analysis, we find that the classified labels (topics) match iPhone’s product features to a great extent. Therefore, we can claim that the topic model can effectively understand customer preferences or customer concerns. In addition, we also find other non-physical attributes, for example, “advertising”, are also parts of the consumer concerns. This is evident that marketing campaigns can effectively increase firms’ brand reputation and promote their products.

To further understand the importance of each topic, we follow the work of Sievert and Shirley (2014) and call the python library “pyldavis” to calculate the ranking of each topic. The output is listed in Table 4, at a descending order. Figure 6 visualizes the results, where the area of each circle is proportional to the relative prevalence of the topic in the corpus. The larger the circle is displayed in the left panel, the higher frequency the topic has, meaning the topic is more important. For each topic selected, pyldavis can also output its top 30 most relevant terms. As depicted in the right hand side of Figure 6, the top 30 high-frequency words for topic 8 are listed and ranked, which constitutes the proportion of total tokens at about 6.3%. Therefore, our second finding is that we can prioritize the topics and find out the importance of consumer concerns. This is important to a firm to understand which concerns contribute more to the adoption of the product.

<table>
<thead>
<tr>
<th>Topic</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>12.6%</td>
<td>12%</td>
<td>12%</td>
<td>8.7%</td>
<td>8%</td>
<td>7.5%</td>
<td>6.6%</td>
<td>6%</td>
<td>5.8%</td>
<td>4.8%</td>
<td>4.7%</td>
<td>4.5%</td>
<td>3.2%</td>
<td>2.2%</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

Another way to conceptualize these topics is to find out the relation between topics by the relevant terms. As shown in Figure 4, group 1 includes words of “lens(镜头),” “shooting(拍摄),” and “video(视频)” and connects two clusters, “Topic 2” and “Topic 13”. Topic 2 labelled with “advertising” and Topic 13 labelled with “Camera” before. Therefore, we can easily find that both two topics have three relevant terms and are predominantly related to the event using iPhone to shoot the three-minute advertising. Meanwhile, group 2 is comprised with two words, “interface(接口)” and “charging(充电)” and connects “Topic4” and “Topic 12”. It means that customers usually put these two attributes together to discuss. Because of the innovation design of iphone7, the device has only one interface for the charging and headset. In group 3, three words, “face id,” “screen(屏幕),” and “fingerprint(指纹)”, has been detected as a bridge between “Topic 11” labelled with “Unlock” and “Topic 14” labelled with “Screen”. “Face ID”, the innovation feature of iPhone X, is a new technology created by Apple company that can use face to unlock iPhone. Therefore, it is an effective approach to find the similar topics using the high-frequency words’ network.

CONCLUSION AND DISCUSSION

In this study, we propose an innovative approach to understand consumer preferences based on the online information of Q&A platform. We propose a two-step framework to conduct text analysis and topic extraction: (1) information preparation, and (2) topic Modeling. Step one collects the UGC data from the Q&A platform and transforms the unstructured texts to a structured form that can be analyzed in the next stage. In step two, the structured data are input to the classic topic model to identify and cluster the topics through the Latent Dirchlet Allocation Algorithm. Based on Zhihu, the most popular online Q&A community in China, we collect

The 18th International Conference on Electronic Business, Guilin, China, December 2-6, 2018

Liu, Wei & Li

696
almost 50000 answers under the discussion topics of “iPhone 7” and “iPhone X”. We find that our approach can effectively extract and rank Chinese consumer’ preferences to the product. In addition, we find consumers preferences are inter-related.

Understanding consumer preferences is fundamental work. Based on our research, extensions could be varied. To name a few, currently we only use UGC to analyze consumers preferences, an extension with platform users’ social connections would be an interesting exploration; with consumer preferences ranking orders, it also helps firm to identify which attributes to upgrade with limited budget concern.

ACKNOWLEDGMENT

This work is partially supported by the Natural Science Foundation of China [Grant no. 71572070]; the Fundamental Research Funds for the Central Universities [Jinan University, Grant no. 15JNQM003].

REFERENCES


