View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Concordia University Research Repository

CLASSIFYING TABLET PC MODELS BASED ON USER PREFERENCES FROM ONLINE REVIEWS

KAMRUN NAHAR

A THESIS

IN

THE CONCORDIA INSTITUTE FOR INFORMATION SYSTEMS ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF APPLIED SCIENCE IN QUALITY SYSTEMS ENGINEERING

CONCORDIA UNIVERSITY

MONTRÉAL, QUÉBEC, CANADA

MAY 2012

© KAMRUN NAHAR, 2012

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Kamrun Nahar

Entitled: Classifying tablet PC models based on user preferences from online reviews

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science in Quality Systems Engineering

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Andrea Schiffauerova Chair

Dr. Amir Aghdam Examiner

Dr. Chun Wang Examiner

Dr. Benjamin Fung, Dr. Simon Li Supervisor

Approved by Dr. Chadi Assi
Chair of Department or Graduate Program Director

Dr. Robin A. L. Drew
Dean of Faculty

Date May 6, 2012

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Kamrun Nahar**

Entitled: **Classifying tablet PC models based on user preferences from online reviews**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science in Quality Systems Engineering

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Andrea Schiffauerova	Chair
Dr. Amir Aghdam	External Examiner
Dr. Chun Wang	CIISE Examiner
Dr. Benjamin Fung	Supervisor
Dr. Simon Li	Supervisor

Approved by ____

Chair of Department or Graduate Program Director

_______________ 20 ______ ______________________________

Dr. Robin A. L. Drew, Dean

Faculty of Engineering and Computer Science

Abstract

Classifying tablet PC models based on user preferences from online reviews

Kamrun Nahar

Online review sites are a good source of information for the manufacturers to understand the product market. Those sites allow users of the product to express their opinions about products which provide valuable information to other people. As these reviews are easily available and contain important information about the product and users, product designers can utilize those reviews for their new product design analysis. To be competitive the designer should consider the users preferences at the time of product designing and should offer product differentiation while offering a new product. Tablet PC is currently considered as a new class of product which needs to be well classified for the users. The history of portable computer tells that at first when portable computer arrived in the market it was also not well classified for different users. At first, almost all manufacturers had one line of portable computer in market which resembles the current time of tablet product. Motivated by the available online reviews by tablet users and the need of the tablet designers, we propose a method to extract interesting patterns from online reviews of tablet users. These extracted patterns can help the designers to understand the new product market of tablet PC to classify its model for different categories of users. We applied association rule mining technique on the online reviews to reveal interesting patterns between users and their preferred tablet features. For identifying this pattern we considered three categories of users: *personal*, *business* and *student* users. To examine the approach, the online reviews posted between April, 2010 and May, 2011 were collected. Then the resultant association rules between the users and tablet features are compared with the existing tablets in the market which supports this study.

Acknowledgments

I would like to thank the Almighty for His divine guidance and blessings that have made this thesis possible. I am grateful to my supervisors, Dr. Benjamin C. M. Fung and Dr. Simon Li, for their guidance and assistance throughout the duration of my research. Their knowledge, encouragement, guidance, and great efforts to explain things clearly and simply to me, have provided a good basis for this thesis work.

I also wish to thank all the faculty members and staff of the Concordia Institute for Information Systems Engineering (CIISE). I am grateful to the Faculty of Engineering and Computer Science, Concordia University for supporting this thesis work.

My special gratitude goes to my family. Without their encouragement and mental support it would have been impossible for me to continue and complete this long journey.

“Impossible is a word to be found only in the dictionary of fools.” -

Napoleon Bonaparte

To my loving *parents* and my supportive *husband*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is concerned with the tablet PC model classification for different categories of users. Tablet PC has been chosen for this work because presently it is considered as a new class of product. New products typically go through some development stages. For example, portable computers have evolved from a single model to different models for different categories of users. Tablet PC development should go through these stages. In the next section, we describe the background information related to tablet PC model classification.

1.1 Background

In the history of portable computers, we observe that at first there was only few models of portable computers in the market which were not classified for different categories of users. Based on the information from Hamm [21] and Wikipedia [1], the first portable computer IBM 5100, appeared in 1975. Then some other models from different manufacturers also appeared in early 1980s such as Osborne 1 (in 1981), GRiD Compass 1100 (in 1982) and Toshiba T11 (in 1985). These computers were considered to be a product mostly for users of specialized field applications like NASA, Air Forces and Military and their prices were too high for general people (e.g., about $2,000 to $8,000 USD).

By the pace of time manufacturers started to release different models of portable computers for different categories of users. By the end of the 1980s, portable computers were becoming popular among business people. Starting from 1990s, some famous lines of portable computers started to appear in the market such as Apple PowerBook 100 Series (in 1991) and IBM ThinkPad 700 Series (in 1992). Though the prices of the portable computers did not change much, their performance was getting closer to a typical desktop computer. In 2007, ASUS launched a new type of portable computer (now classified as netbook), Eee PC, which came with low price and light weight. This product has changed the market and made the other portable computers down their prices. As a result more users like journalists, accountants and sales representatives were added in the portable computer users list. Finally, a large number of student users were also included because of the low prices of portable computers. As the number of users were increasing due to the reduced prices, different types of portable computers were also well classified for general customers to choose in the market (e.g., desktop replacement, gaming, ultrabook, netbook, and etc). The development of portable computers with different user categories are shown in Figure 1. This graph is showing the three stages of acceptance of portable computers by different categories of users.

In this work we considered the development of tablet also resembles the development of portable computers. After the arrival of Apple iPad in 2010 many manufacturers started offering tablets in the market. Asus Eee Pad (in July, 2010), Samsung Galaxy Tab (in September, 2010), HP Slate 500 (in October, 2010), Motorola Xoom (in February, 2011) and many other tablets appeared after the release of Apple iPad. These tablets are usually designed for general public rather than specific user categories (e.g., business or student users). Many users have started using tablet at the early stage of tablet development as shown in Figure 2. This figure is showing the acceptance rate of tablet by the users in

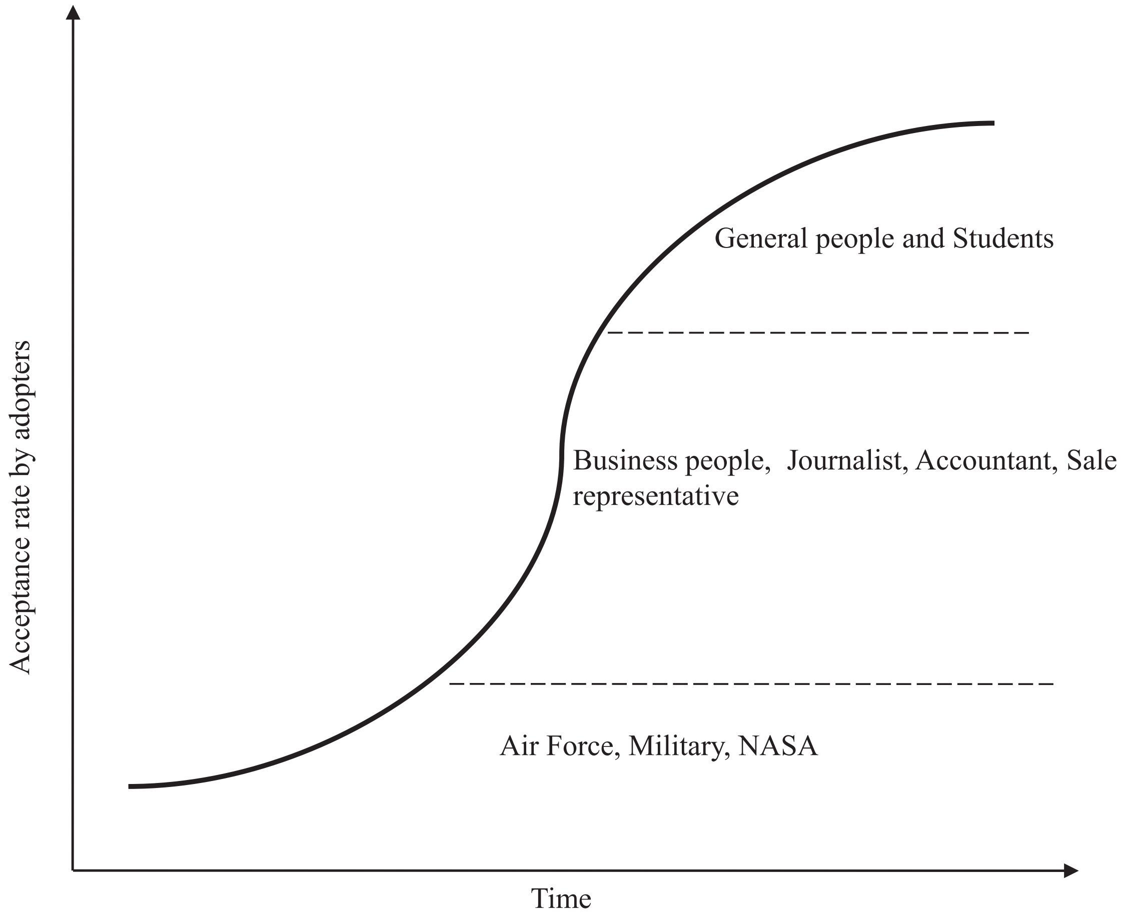


Figure 1: Development of portable computers

the early stage and the acceptance by time is expanding. However, when we examine the development history of portable computers (illustrated in Figure 1), we find that careful designs for specific users are important for future tablet development.

Though the tablet market is growing rapidly, it is still a new kind of product and manufacturers would have sufficient time to classify the tablet models for different categories of users (in the middle of 2011). Like the development of portable computers, tablet PC models also should be categorized for different categories of users. Thus the issue of this thesis is to help the tablet designers to classify the tablet PC models for different categories of users.

To classify the tablet PC models for different categories of users, we find that the user

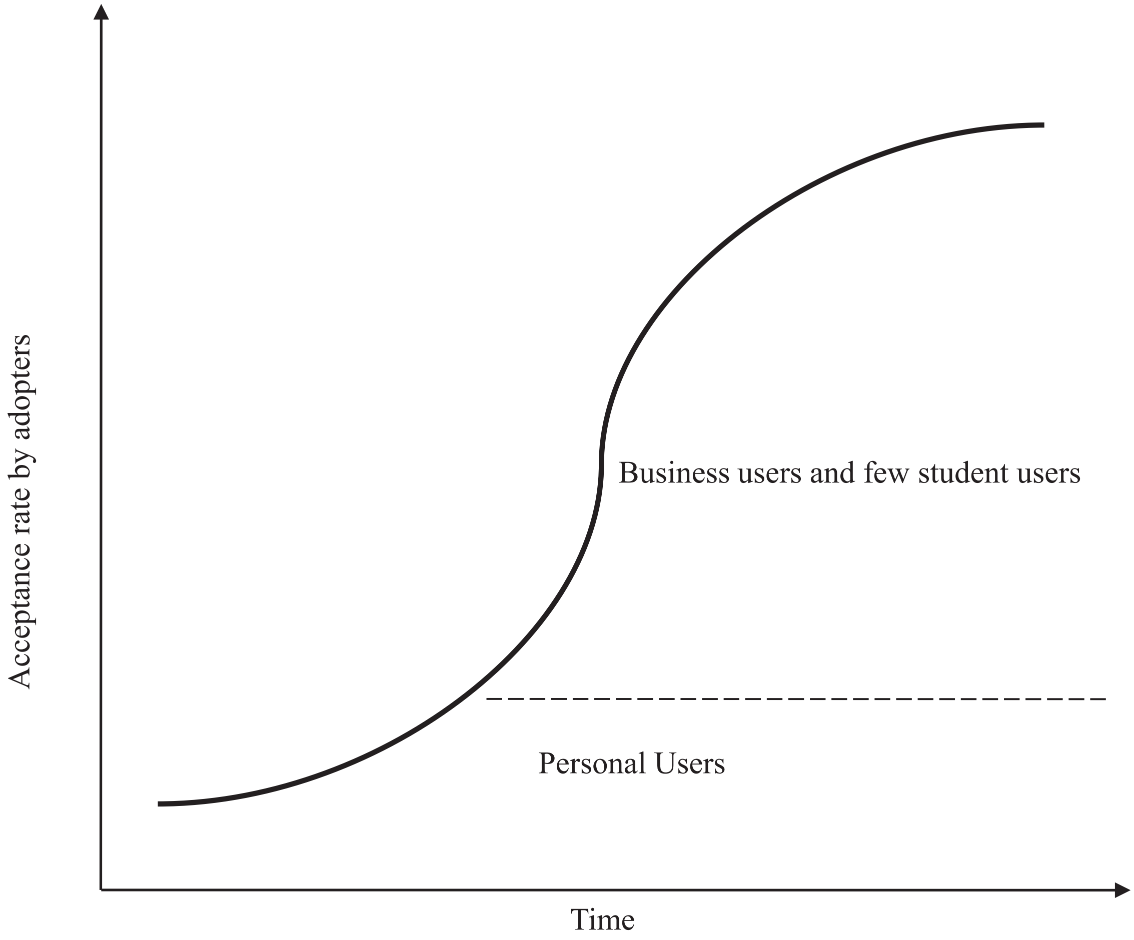

Acceptance rate by adopters
Business users and few student users
Personal Users
Time

Figure 2: Development of tablet PC

opinions posted in various websites contain valuable information about the users. In the net generation, people are willing to share their experiences and opinions about new products online [46]. As the number of online reviews can be numerous, we apply data mining techniques to analyze those reviews. We classified the tablet users in three categories, i.e., *personal*, *business* and *student* users. Then we try to identify any interesting patterns of these users with tablet features from the online reviews. We give an example of tablet PC user review from online to illustrate. Some portion of the review for Apple iPad (64 GB) looks like the following.

"Let me start off by saying why I bought this thing. I'm active duty military and I

deploy/travel very frequently. That being said, I am always on the look out for a device that will entertain me throughout LOTS of boredom/downtime. Whether it's playing a quick game, watching hours of movies and/or TV shows, or keeping the family up to date on Facebook when I find an internet connection. As for the device itself, I couldn't have been happier with my purchase. I went on a trip recently (within the States) and this is when I found out how long the battery will last. I spent about 6 hours on planes total and I only used about 20% of the battery. Considering I used to barely get through one movie on a fully charged laptop battery, that was awesome. When having to run through the airport trying to catch a connection flight, it's nice not to have to worry about stopping and trying to get that last minute charge on a device. This thing really will last all day. It's also nice to watch movies on the larger screen (as opposed to the iPod Touch or iPhone). It is nice to have the larger keyboard on screen."

From the above review, we identified the user category is *personal* as he is using it for entertainment purpose and the preferred feature set by the user is {*long battery life, screen size 9.7 inch, touch keyboard*}. An example of identified categories of users and their preferred features from 10 reviews are showed in Table 1.

r_i	**User Categories** $C(r_i)$	**Preferred Features** $F(r_i)$
r_1	{personal, student}	{long battery life, screen size 7inch, touch keyboard}
r_2	{personal}	{nice display, screen size 7inch, long battery life, fast processor}
r_3	{personal}	{fast processor, good graphics, long battery life, screen size 7inch, good touch screen}
r_4	{personal}	{good touch screen, long battery life, screen size 7inch}
r_5	{business, personal}	{beautiful design, nice display, good touch screen, long battery life, keyboard}
r_6	{student, business}	{long battery life, portable, nice display, beautiful design, front facing camera}
r_7	{personal}	{long battery life, screen size 7inch, beautiful design, front facing camera}
r_8	{personal}	{long battery life, screen size 10 inch, front facing camera, light weight}
r_9	{business}	{long battery life, portable, screen size 10 inch, front facing camera, light weight}
r_{10}	{student}	{long battery life, portable, beautiful design, light weight}

Table 1: Identified user categories and their preferred feature sets in 10 reviews

1.2 Motivation

This thesis work is to analyze the online reviews for classifying the tablet PC features for different categories of users and it is motivated by the following two observations:

The first observation is the need of the tablet designers to classify the tablet models for different users. After the arrival of Apple iPad in 2010, many tablets from different manufacturers have appeared in the market. Though the tablet market is expanding rapidly, it is noticeable that almost all the manufacturers offered a unique tablet model in the market. But to be competitive, they need to satisfy different categories of customers. So manufacturers have no other way except offering new tablet to compete in the tablet market. Thus the product designer must give time on analyzing the product design for meeting all users need.

The second observation is the availability of users opinions in online reviews which can be utilized to classify the tablet models. The online review is a large and easily accessible information source for both the users and product manufacturers. Many websites allow users to share their own experiences about tablet on those online reviews. These reviews contain important information for the manufacturers. These opinions of users can be used as Voice of Customer (VOC) by the tablet designer to understand the tablet market. VOC is the task of identifying customer needs, structuring customer needs, and providing priorities for customer needs [20]. VOC has been effective in helping many companies guide the development of product platform specifications and features [37]. For a company, it may be no longer necessary to conduct surveys, organize focus groups or employ external consultants in order to find consumer opinions about its products and those of its competitors because the user generated content on the Web can already give them such information [16]. The online reviews are useful and they should be considered as an alternative source of collecting VOC for several reasons:

- Online review contains the information regarding users preferences about product features and their purpose of usage.

- As the reviews are created by the user itself, they are bias free while traditional surveys are biased by the surveyors.
- Reviews are easily available, free and saves time while traditional surveys are costly and time consuming.

1.3 Objective

The main objective of this thesis is to provide a method to analyze the online reviews to aid the tablet PC feature classification for different categories of users. Our focus is on providing the method which is able to classify tablet features based on users categories from the data of online reviews. The method is also able to help the designers in tablet PC feature selection. As the reviews are made by interested customers, the scope of this research can only provide an alternative way for the tablet designers.

Our aim is to identify two types of interesting patterns from the online reviews, the first pattern is to classify the tablet PC models and the second pattern is to help in tablet PC feature selection. To meet the objective we can use data mining techniques. Employing the data mining technique on these reviews is an efficient approach that we believe will reveal the preferences of the users of different categories.

1.4 Contribution

In order to meet our objective, we have proposed a method for mining interesting patterns from online reviews of tablet PC. Such interesting patterns are valuable for the product designer to understand the feature classification of tablet PC for different categories of users. We have mined two kinds of patterns from the online review data. First one is the interesting patterns between the users and tablet PC features, and the second one is the interesting patterns among the features. Based on the interesting patterns we classify the tablet models for different categories of users. To see the effectiveness of our proposed method we also compared the result with the existing tablet models in the market.

Our approach has the following merits:

- We explore the feasibility of online reviews for determining the users preferences to categorize the tablet models.

- We employ the notions *frequent feature set* and *association rules* [4] to model the relationship among the features and user categories.

- We examine the use of online reviews in analyzing the associations between users and their preferable tablet features, and the associations among tablet features.

1.5 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 reviews the background knowledge and the previous work of text mining, opinion mining from online reviews and work of product information extraction from online review. Chapter 3 formally describes the problem, i.e., tablet PC model classification from online reviews for different categories of users . Chapter 4 describes the methodology of our problem solution. Chapter 5 shows the experimental result on user dataset and evaluates the experimental result with the existing product. Finally Chapter 6 concludes the thesis.

Chapter 2

Related Work and Background Knowledge

In this chapter, we describe the background information that is related to the online review mining for analyzing users preferences. In order to analyze the users preferences for a product, the users information required for the analysis must first be extracted from the reviews. Therefore, we discuss some of the text mining works which can be used to extract user information from the reviews. Some of the text mining works include text summarization, text classification and topic mining. In this chapter, we briefly explain some works on the text summarization, text classification and topic mining. In literature, review mining is studied under the topic of opinion mining. Therefore, we also discussed some works in the area of opinion mining. As we also employ data mining techniques in our work, we briefly describe the application of data mining techniques in product development and market analysis.

2.1 Text Mining

Text mining refers to the process of extracting useful information or knowledge from unstructured textual documents. Information can be extracted to derive summaries for the words contained in the documents [18,38,39] or to categorize textual documents [6,11,30] or to identify topic of textual documents [8,31]. Detailed description of the aforementioned areas are given in the following paragraphs.

Text summarization involves reducing a text document or a larger corpus of multiple documents into a concise collection of words or paragraph that conveys the main meaning of the text to the reader. Several techniques have been proposed for text summarization. Nahm and Mooney [39] proposed a text mining system DiscoTEX which has been applied to mine job postings and resumes posted to USENET news groups. The technique also has been applied to mine Amazon book description pages from the Web. Information extraction systems can be used to directly extract abstract knowledge from a text corpus, or to extract concrete data from a set of documents which can then be further analyzed with traditional data mining techniques to discover more general patterns [38]. An information extraction method based on Relational Markov networks has been developed by Mooney et al. [38] to directly extract text from unstructured documents. Interpretation and the development of new hypotheses from text documents are very important. Many works [18,42] addressed this problem. The approach by Plaisant et al. [42] has been applied in literature domain. Literary scholars could use a Naive Bayesian classifier to determine which letters of correspondence contained erotic content. It gave users some insights into the vocabulary used in the letters. Don et al. [18] proposed a system named FeatureLens that allows visual exploration of frequent text patterns in text collections. The concepts of frequent words, frequent expressions, and frequent closed itemsets of n-grams have been applied to guide the discovery process. By using FeatureLens users can find meaningful co-occurrences of text patterns by visualizing them within documents. This also permits users to identify the increasing, decreasing, and unexpected appearance of text patterns. Surveys on the web provide many questionnaire data about a product which may contain valuable information for making business decisions. Automatically mining and summarizing those open answers help the manufacturers to make decisions about the specification of the next version of their products. Li and Yamanishi [32] developed a text mining system that provides a method for analyzing open answers in questionnaire data.

Text categorization or text classification is the task of assigning a textual document to one or more classes or categories. There has been many works on automatic text categorization. Apte and Damerau [6] proposed a rule-based induction method for text categorization. They have adopted decision tree learning technique to learn the classifier. Then employed this classifier to identify the category of a given document. Chai et al. [11] explored the use of Bayesian classifier to classify text documents. Lam and Lai [30] proposed a different approach of text classification using meta-learning approach. Instead of applying a single method for all categories during classification, this new meta-learning approach can automatically recommend a suitable algorithm during training, from an algorithm pool. Zhang et al. [51] proposed a text classification technique to extract key phrases from web document.

The unsupervised topic identification or topic discovery is a technique of identifying topics of documents in text corpus by using content-based clustering. Clustering is used to identify important information from the documents without knowing any background knowledge. The documents of different clusters are dissimilar but the documents of a same cluster are similar. Document clustering helps to identify the topic of documents in collection of documents in several ways. A text mining tool has been designed by Larsen and Aone [31] to find relevant documents quickly. The top level of a cluster hierarchy summarizes the contents of a document collection, enabling users to selectively drill deeper to explore specific topics of interest without reading every document. The primary steps to generate hierarchy was the extraction of features from document and then clustering based on those features. Beil et al. [8] proposed a method of text clustering using the idea of frequent item set mining. This frequent item-based approach of clustering is able to reduce the large dimensionality of the document while clustering. Cutting et al. [14] proposed an approach of document clustering. They have introduced two algorithms for browsing a collection of documents conveniently. This is an iterative method where the system scatters the collection of documents into groups and provide short summaries of those groups to the user. Based on these summaries user can select one or more of the groups for further study. The system then apply clustering again to scatter the new sub collection into a small number of document groups. With each successive iteration the group becomes smaller and more detailed.

All the aforementioned works of text mining are to extract information from unstructured text documents. These text mining ideas can be used to extract product features and user categories from the online reviews. Extracting information from the online reviews are more challenging than extracting information from the traditional documents. Online reviews are not written in formal way which may contain spelling errors, special characters and symbols. For mining the knowledge from reviews the required information must first be extracted to make a structured dataset. The dataset is then utilized by using data mining techniques. For mining the knowledge from the dataset we have used association rules mining technique.

2.2 Opinion Mining

Opinion mining falls into the research area between data mining and text mining. The objective of opinion mining is extract the opinion of a writer (or a group of writers) in a particular subject. In the context of market analysis and product design, understanding the opinion of the current product users play an important role in the decision making process while designing the next version of a product [40]. Nowadays, product users often express their opinion about products and services through blogs, forums, and social networking sites. Consequently, these new media become a low-cost source of information for opinion collection.

Review mining [17, 24, 25] is an emerging research area in opinion mining that focuses on how to efficiently and effectively evaluate large volume of unstructured textual review data. Many works in opinion mining and review mining focus on feature-based sentiment classification. An increasingly popular trend in opinion mining is the combination of feature mining and sentiment identification techniques to extract consumer opinions in form of feature based summaries [10, 15, 43]. These methods first identify product features from the text, extract sentences which are the positive or negative comments about those features and then produce a summary of that product with the extracted features and their comments. In most of the works, the objective is to classify the reviews into positive or negative comments. Although these approaches can classify the opinion of a writer, they cannot help designer to get more in-depth information on particular features. In contrast, our work provides a methodology to help tablet designers better understand the correlation between product features and user categories.

2.2.1 Product Feature Extraction and Sentiment Analysis

Product feature extraction from online review is the extraction of feature set of a product about which reviewers have been commented on. Many works have been done to find product features that have been commented on the reviews. Natural language processor and Apriori algorithm have been used to identify the feature words of reviews [25]. The main idea is to find the features that appear explicitly as nouns or noun phrases in the reviews. To identify nouns/noun phrases from the reviews, they have used the part-of-speech tagging. Then they identified frequent item set of nouns in the reviews, which are likely to be product features. This work also introduced the idea of implicit feature mining. Hu et al. [24] also proposed techniques for identifying infrequent features using adjectives in reviews, which are considered as opinion words. Opinion words are the words which are used to say something positive or negative about a feature. This work also gives the summary of the reviews based on the extracted features and shows the negative and positive opinion about those features. A holistic lexicon-based approach was proposed by Ding et al. [17] which solved the problem of context dependent opinion word identification and improved the previous lexicon-based method [24]. Kim et al. [29] studied the problem of opinion summarization of online product reviews using association rules mining [5].

The objective of sentiment analysis is to identify the attitude of a writer towards a subject. Most works [41, 49, 52] in this area aim at classifying some textual data into either positive or negative opinions. For example, Zhuang et al. [52] applied a multi-knowledge based approach, which integrates WordNet, statistical analysis and movie knowledge for mining movie reviews. This work has been extended from [24] and used some grammatical rules to mine feature-opinion pairs. After mining feature-opinion pairs they have generated a summary about movie features and comments about those features. Zhang et al. [49] proposed a new feature mining method to improve the work by Qiu et al. [44]. This paper [49] proposed a method to extract features from corpora in different sizes while the previous work by Qiu et al. [44] focuses on mining features from medium size corpora. The feature mining method by Qiu et al. can not mine features from large and small size corpora.

All of the above works of opinion mining are to extract features and provide summary from the users reviews. These information can help the users to get idea about a specific product before purchasing it, while our work is to analyze the users preferences to help the product designers. In our work we mined the association rules between the user categories and users preferred feature sets of tablet PC and the association rules among the features only which are for classifying tablet PC models.

2.3 Extraction of Product Information Using Data Mining

The extraction of important information from large online data, is a powerful new technology with great potential to help companies focus on the most important information in their businesses. Data mining techniques predict future trends and behaviors, allowing businesses to take knowledge-driven decisions. Many researchers are working on this issues. A system to gather and annotate online discussions about different products has been proposed by Glance et al. [19]. The system is able to extract important information for the producers for their marketing analysis. This work used text classification and computational linguistics for analyzing the textual data. A system for predicting sales performance was applied by Liu et al. [36]. This work studied the problem of mining sentiment information from blogs and investigate ways to use that information to predict product sales performance. The main idea was to use the blog sentiments and revenue data by linear prediction to predict the future revenue. Li et al. [33] proposed a method to find important service aspects and to automatically generate customer service surveys through mining service reviews. This work used association rules mining to find frequent service aspect. Co-occurrence method and linear regression has been used to rank the candidate service aspects. In order to satisfy customer needs as well as to reduce supply chain complexity Kim et al. [28] applied a methodology to determine the appropriate product family size. They did an experiment with the data of mobile phone market to analyze the different groups of consumers' preferences about the mobile phone models. Historical data mining has been used to match customers' preferences and product characteristics.

There has been a considerable research on online review mining and product information mining. However there has been no studies on how to utilize online review information for new product design analysis. In our case we have utilized these online reviews for tablet PC feature classification using data mining technique. Most of the previous works of mining online reviews were on feature extraction and summarization of reviews about the product. There are some works on product development from survey data. Bae and Kim [7] proposed an approach for product development based on customer needs using data mining technique. For data collection they performed a survey with camera users and nonusers. Nonusers of a product can not be clear about their need while users can clearly recognize their real needs of product. Many people do not know what they want from a product before using it. The perception of the nature of a product's benefits can change as the product becomes more familiar [9]. Unlike their approach our research is to classify the new product, tablet features from online reviews. On those reviews the users clearly mentioned about their preferences of tablet PC features. These reviews are also less costly and unbiased than questionnaire survey.

2.4 Data Mining Techniques

Data Mining is the process of extracting useful knowledge from a large volume of data. Data mining can be performed on various types of data, such as relational data, transactional data, textual data, and different types information repositories, such as World Wide Web. There are many kinds of knowledge and patterns that can be discovered by data mining techniques to aid in the product development and marketing research. Product development and market analysis can be supported by different data mining techniques. Some of the data mining techniques include association rules mining, classification analysis, and cluster analysis. Description of the general application of these three techniques is given in the following paragraphs.

Association Rules Mining. The goal of association rules mining is to detect relationships or associations between items which exist together in a record. This technique has a wide range of application in the product development and market analysis. Many researchers [7, 34, 35, 47] applied association rules mining to extract customer knowledge or needs for product development. Liao and Chen [34] used association rules mining to extract marketing knowledge patterns for the electronic catalog marketing and sales management of a retailing mall in Taiwan. Bae and Kim [7] applied association rules mining on the customer data for new camera development based on customer needs. Market basket analysis is a typical example where association rules mining is widely used. Market basket analysis is a useful method of discovering customer purchasing patterns by extracting associations or co-occurrences from stores' transactional databases. The information obtained from the analysis can be used in forming marketing, sales, service, and operation strategies. Chen et al. [12] proposed a method for automatically extracting association rules in a multi-store environment. They have developed an Apriori-like algorithm to discover purchasing patterns for company with multiple stores.

Classification Analysis. Classification is the process of assigning labels to previously unseen data records based on the knowledge extracted from historical data. The goal of classification is to build a model for future prediction based on the predefined classes. Classification has numerous applications including credit approval, product marketing, and medical diagnosis. Credit scoring is a widely used technique that helps banks decide whether to grant credit to consumers who submit an application [26]. Huang et al. [26] proposed a credit classification model to evaluate the applicant's credit score. Hui and Jha [27] proposed a data mining approach for customer service support using classification analysis. In traditional customer service support of a manufacturing environment, a customer service database usually stores all the service information of faults and solutions. This service database is used to form a knowledge base to diagnose the faults.

Cluster Analysis. Clustering is the process of grouping objects of the similar kinds into respective categories. Clustering techniques have been applied in many research problems of product development and marketing analysis [3, 35, 50]. Zhang et al. [50] proposed a clustering-based market segmentation approach for product family positioning. The goal was to offer the product family to the targeted customer segment based on the customer requirement data. If the customers' choice do not match any product from the product family, they can pick a similar product of their choice from the product family. Agard and Kusiak [3] employed cluster analysis and association rules mining to derive product family requirements based on similar customer groups. In this paper, clustering is used to identify similar customers that share the same or highly similar behaviors, and then association rules mining has been used to identify the product requirements for a group. The idea behind clustering is that the customers from the same cluster share similar requirements. So it could be sensible to propose a specific product design for each cluster of customers. Laiao et al. [35] used association rules mining and cluster analysis to extract knowledge for tourism product development and customer relationship management. They have used cluster analysis to determine the cluster of tourism customers and used association rules mining to find the customers preferences about different tourism products. Knowledge extracted from this analysis can help upper management for planning and marketing the tourism products.

2.5 Summary

In this chapter, we have presented research works in the areas of text mining, opinion mining, product information extraction using data mining, and use of data mining in product design and market analysis. In the text mining, we focused on the text summarization, text categorization and topic mining works. For the works of opinion mining, we have focused on the review mining works. Review mining is a research area in opinion mining where most of the works of the researchers are to extract product features and reviewers' opinions about those features. We have also discussed works on extracting product information by data mining techniques. These works include extracting product information for market analysis, predicting sales performance, and developing product. The literature of data mining techniques for product design and market analysis include association rules mining, classification analysis, and cluster analysis techniques. In the following chapter, we will describe the problem of our work.

Chapter 3

Problem Description

Given a large collection of online reviews on a specific category of products, for example Tablet PC, a product designer wants to extract the opinions of product users from online reviews with the goal of identifying the desirable features for different categories of users and to design the specifications of the next generation of products based on the collected users' preferences. The challenge is how to efficiently and effectively extract the *combinations of features* preferred by different categories of users from the textual data in online reviews. The problem is formally defined as follows with the notation summarized in Table 2.

Let $F = \{f_1, \ldots, f_p\}$ be the set of possible *features* of a product. In this thesis, the term “feature" is broadly defined as any possible function of a product (e.g., *Wi-Fi*) or the specification of a physical item (e.g., *7-inch screen*). Let $C = \{c_1, \ldots, c_q\}$ be the set of possible user categories. For example, the users of computer notebook can be broadly classified into three categories $C = \{$*business*, *personal*, *student*$\}$.

In this thesis, we focus on the features in reviews with positive comments because understanding of users preferences about features are essential for the designer for designing new class of product. A product or service is designed effectively if company involves consumers in designing and encourage consumers to focus on what is wanted rather than what is not wanted [13]. For designing the product it is important to know what features are

Var	**Description**
F	Set of possible features of a product
$\{f_1, \ldots, f_p\}$	Elements of F
C	Set of possible user categories
$\{c_1, \ldots, c_q\}$	Elements of C
R	Set of online reviews
$\{r_1, \ldots, r_n\}$	Elements of R
$F(r_i)$	Set of preferred features in each review r_i
$C(r_i)$	Set of user categories in each review r_i
I	Feature set
k	Number of features in feature set
I_f	Frequent feature set
I_s	Subset of I_f
I_t	Antecedent of association rule
I_h	Consequent of association rule
sup	Support
$conf$	Confidence
min_sup	User specified support threshold
min_conf	User specified confidence threshold

Table 2: Descriptions of notations

preferable than what is not preferable by the users. We further assume that there exists some information retrieval preprocessing methods [17,24,25] to extract the positive features from online reviews. For user categories, we assume there is a classification method [6,22,30] to identify the user categories based on the online review contents. Below we formally define the representation of input online reviews, in which each review consists of a set of features and a set of user categories.

Definition 1 (Review)**.** Let $R = \{r_1, \ldots, r_n\}$ be a set of online reviews. Each *review* r_i is represented as a doublet, denoted by $\langle F(r_i), C(r_i)\rangle$, where $F(r_i) \subseteq F$ and $C(r_i) \subseteq C$. ■

For example, a review r_1 in Table 1 is represented as a set of features $F(r_1) = \{$*long battery life*, *screen size 7 inch*, *touch keyboard*$\}$ and $C(r_1) = \{$*personal*, *student*$\}$.

3.1 Frequent Feature Set

A product designer would like to identify the combinations of features that are frequently discussed together in online reviews. Such combinations may directly or indirectly reveal the preferences of the users of some particular categories; therefore, the extracted combinations may be useful for designing the next generation of products. The challenge is how to capture and model the combinations of features that are frequently discussed together in online reviews. In this thesis, we propose to employ the notions *frequent feature set* and *association rules* [4] to model the relationship among the features and user categories.

Let F be the universe of possible tablet PC features, I be a set of features called *feature set*, C be a set of possible user categories, where $I \subseteq F \cup C$. A feature set that contains k features is called a k*-feature set*. For example, the feature set $I = \{$*personal*, *student*, *long battery life*, *screen size 7 inch*, *touch keyboard*$\}$ is a 5-feature set. A review $r_i = \langle F(r_i), C(r_i) \rangle$ *contains* a feature set I if $I \subseteq F(r_i) \cup C(r_i)$. The *support* of a feature set is the percentage of reviews in R that contains the feature set. A feature set is a *frequent feature set* in a set of reviews R if the support of the feature set is greater than equal to a user-specified minimum support threshold.

Definition 2 (Frequent feature set)**.** *[5]* Let R be a collection of reviews, F be a set of possible features of a product, and C be the set of possible user categories. Let $I \subseteq F \cup C$ denote a feature set. The support of a feature set, denoted by $sup(I)$, is the percentage of reviews in R that contain the feature set I, i.e., $sup(I) = \frac{|R(I)|}{|R|}$, where $|R|$ is number of reviews in R and $|R(I)|$ is the number of reviews in R containing the feature set I. A feature set I is a *frequent feature set* in R if $support(I) \geq min\_sup$, where the minimum support threshold min_sup is a real number in an interval of $[0, 1]$. A frequent feature set with k features is called a *frequent* k*-feature set*. ■

Example 1 (Frequent feature set)**.** Consider Table 1. Suppose the user-specified threshold $min\_sup = 0.4$, which means that a feature set I is frequent if at least 4 out of the 10 reviews contain all features and categories in I. $\{$*nice display*$\}$ is not a frequent feature set because it has $sup($*nice display*$) = 3/10 = 0.3$, which is less than 0.4. $\{$*personal*$\}$ is a frequent 1-feature set because it has $sup($*personal*$) = 7/10 = 0.7$. $\{$*personal*, *screen size 7 inch*$\}$ is a frequent 2-feature set because it has support $5/10 = 0.5$. $\{$*personal*, *screen size 7 inch*, *battery life*$\}$ is frequent 3-feature set because it has support $5/10 = 0.5$. Example 3 will show how to efficiently compute all frequent feature sets from a large collection of reviews. ■

3.2 Association Rule

The discovery of associations or relationships among features (F) and user categories (C) may reveal some hidden patterns that are beyond the prior knowledge of product designers and market analysts and, therefore, may help them better shape the next generation of products. We capture such notion using association rules.

Definition 3 (Association rule)**.** *[5]* An association rule is denoted by $I_t \rightarrow I_h$, where $I_t \subset F \cup C$, $I_h \subset F \cup C$, $I_t \cap I_h = \emptyset$. The support of a rule is denoted by $sup(I_t \rightarrow I_h) = \frac{|R(I_t \cup I_h)|}{|R|}$, where $|R(I_t \cup I_h)|$ is the number of reviews in R containing both feature sets I_t and I_h. The *confidence* of a rule is denoted by $conf(I_t \rightarrow I_h) = \frac{|R(I_t \cup I_h)|}{|R(I_t)|}$, where $|R(I_t)|$ is the number of reviews in R containing feature set I_t. ■

In particular, designers and market analysts are interested in the association rules supported by some significant number of reviews. Also, the extracted association rules are useful only if they follow a specific patterns that can be utilized in market analysis and product design. Specifically, designers and market analysts are interested in associations between user categories and features, and associations among features. Thus, we define the notion of *interesting association rule* as follows:

Definition 4 (Interesting association rule)**.** An association rule $I_t \rightarrow I_h$ is *interesting* if

1. $sup(I_t \rightarrow I_h) \geq min\_sup$

2. $conf(I_t \rightarrow I_h) \geq min\_conf$

3. ($I_t \subseteq C$ and $I_h \subseteq F$) or ($I_t \subseteq F$ and $I_h \subseteq F$).

where min_sup and min_conf are user-specified minimum support threshold and minimum confidence threshold, respectively. ■

Example 2 (Interesting association rule)**.** Table 1 contains a frequent feature set $I_f =$ {*personal*, *screen size 7 inch*} whose support is 0.5. The non-empty subsets of I_f are {*personal*}, {*screen size 7 inch*}. The resulting association rules generated from I_f are:

- *personal* → *screen size 7 inch* $[conf = 71\%]$

- *screen size 7 inch* → *personal* $[conf = 100\%]$

If the user-specified minimum support min_sup is 40% and confidence threshold min_conf is 50%, then the above two association rules pass this threshold.

Using the rule constraint ($I_t \subseteq C$ and $I_h \subseteq F$) or ($I_t \subseteq F$ and $I_h \subseteq F$), the extracted interesting rule from the above rules is $personal \rightarrow screen\ size\ 7\ inch$. Section 4.3 will show how to generate interesting association rules from a large collection of reviews. ■

Given a collection of online reviews R on tablet PC, a minimum support threshold min_sup and a minimum confidence threshold min_conf, the *problem of interesting pattern mining for classifying the tablet features for different users* is to extract *all* interesting association rules given in Definition 4.

Thus the research issue is to classify tablet PC models and select features for designing tablet PC specification by using the interesting patterns $I_t \rightarrow I_h$, where ($I_t \subseteq C$ and $I_h \subseteq F$) or ($I_t \subseteq F$ and $I_h \subseteq F$) .

3.3 Summary

Tablet PC feature classification for different categories of users from online reviews is the problem of extracting all interesting association rules from the collection of online reviews on tablet PC. The extraction of association rules among tablet features and user categories can be utilized for tablet PC feature classification and tablet PC feature selection by the tablet designers. To address this issue, we have presented a method in the following chapter.

Chapter 4

Methodology

The general idea of our proposed method, depicted in Figure 3, can be summarized in four steps. The first step is to collect and interpret data to make the dataset for further processing. The second step uses the collected dataset as input to find the frequent feature sets using a data mining tool called *RapidMiner 5* [2]. Then the third step is to find two types of association rules, namely the association rules between the user category and their preferred features, and the association rules among features. The fourth step is to analyze these two forms of association rules to find out the useful information for product designers to classify the tablet PC models for different categories of users.

4.1 Data Collection and Pre-processing

Input data is a collection of online reviews about tablet PCs. In practice, the reviews can be collected from different product discussion websites, such as Amazon.com and CNet.com. Online reviews are in free-text format, so some preprocessing steps are required in order to transform the free-text into a format that can be processed by our proposed method.

Let $R = \{r_1, \ldots, r_n\}$ be a collection of reviews as described in Definition 1. First, we extract the features and user categories from the review, and transform the information

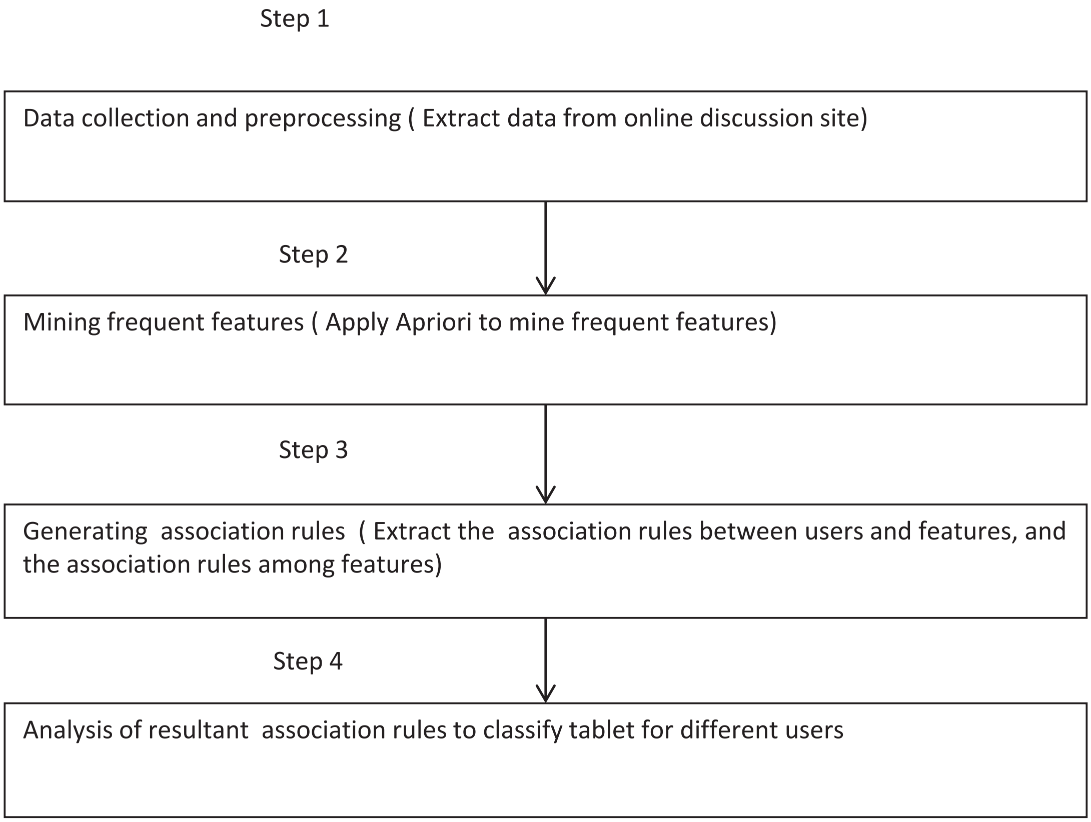
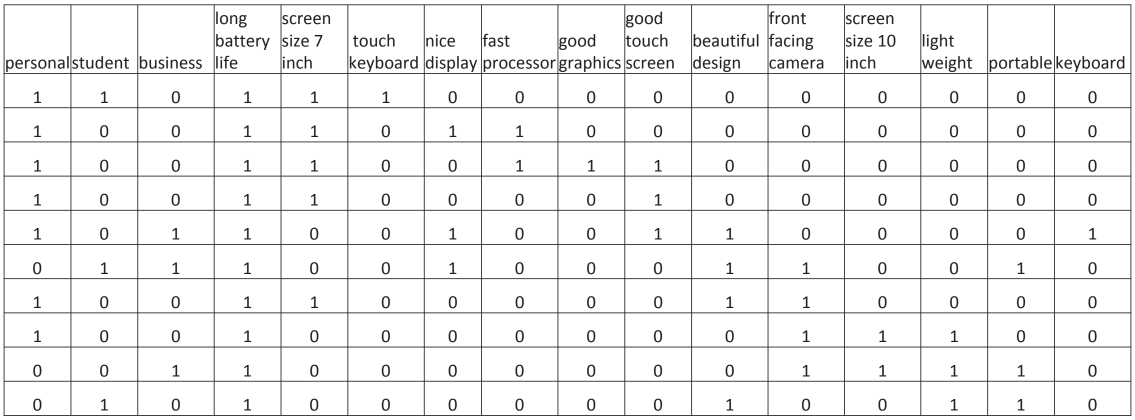

Step 1
Data collection and preprocessing (Extract data from online discussion site)
Step 2
Mining frequent features (Apply Apriori to mine frequent features)
Step 3
Generating association rules (Extract the association rules between users and features, and the association rules among features)
Step 4
Analysis of resultant association rules to classify tablet for different users

Figure 3: Research methodology

into a transaction data table. Each review is transformed into one transaction record in the table. Each transaction record consists of a set of feature items mentioned by the user in the review and the category of the user. Table 3 shows an example of a transaction data table of 10 reviews. To represent the presence of a user category and feature items in a review, the corresponding cells are assigned with the value of 1 and the rest of the feature items are given a value of 0. The corresponding dataset of the 10 reviews are provided in Figure 4. Note that a user may be classified into multiple categories.

r_i	**User Categories** $C(r_i)$	**Preferred Features** $F(r_i)$
r_1	{personal, student}	{long battery life, screen size 7inch, touch keyboard}
r_2	{personal}	{nice display, screen size 7inch, long battery life, fast processor}
r_3	{personal}	{fast processor, good graphics, long battery life, screen size 7inch, good touch screen}
r_4	{personal}	{good touch screen, long battery life, screen size 7inch}
r_5	{business, personal}	{beautiful design, nice display, good touch screen, long battery life, keyboard}
r_6	{student, business}	{long battery life, portable, nice display,beautiful design, front facing camera}
r_7	{personal}	{long battery life, screen size 7inch, beautiful design, front facing camera}
r_8	{personal}	{long battery life, screen size 10 inch, front facing camera, light weight}
r_9	{business}	{long battery life, portable, screen size 10 inch, front facing camera, light weight}
r_{10}	{student}	{long battery life, portable, beautiful design, light weight}

Table 3: Identified user categories and their preferred feature sets in 10 reviews



personal	student	business	long battery life	screen size 7 inch	touch keyboard	nice display	fast processor	good graphics	good touch screen	beautiful design	front facing camera	screen size 10 inch	light weight	portable	keyboard
1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0
1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0
1	0	0	1	1	0	0	1	1	1	0	0	0	0	0	0
1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	1	0	0	1	1	0	0	0	0	1
0	1	1	1	0	0	1	0	0	0	1	1	0	0	1	0
1	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	1	1	1	0	0
0	0	1	1	0	0	0	0	0	0	0	1	1	1	1	0
0	1	0	1	0	0	0	0	0	0	1	0	0	1	1	0

Figure 4: Dataset of 10 reviews

4.2 Mining Frequent Features

There are many data mining algorithms for extracting frequent features, for example, *Apriori* [5], *FP-growth* [23], and *ECLAT* [48]. To efficiently mine all frequent features from the reviews, we employ the Apriori algorithm [5], which is designed for extracting frequent patterns from transactional data. Below we provide an overview of Apriori algorithm which has been applied to various data mining tasks.

Let $F = \{f_1, \ldots, f_p\}$ be the set of possible *features* of a product and $C = \{c_1, \ldots, c_q\}$ be the set of possible user categories in a collection of reviews R. Each *review* $r_i \in R$ is represented as a doublet, denoted by $\langle F(r_i), C(r_i)\rangle$, where $F(r_i) \subseteq F$ and $C(r_i) \subseteq C$.

Refer to Table 2 for notations. Apriori is known as a level-wise search algorithm which uses an iterative approach to extract frequent $(k+1)$-feature sets based on frequent k-feature sets. First, the sets of frequent 1-feature are found by scanning each review $r_i \in R$, accumulating the support count of each 1-feature set I, and collecting the feature set I that has support $sup(I) \geq min\_sup$. The resulting frequent 1-feature sets are then used to identify the frequent 2-feature sets, which are then used to identify frequent 3-feature sets and so on. This process continues until no more frequent k-feature sets are found. Generation of the frequent $(k+1)$-feature set is based on the following Apriori property.

Property 4.2.1. ***(Apriori Property)*** *[5]. Suppose I_f is frequent feature set. If $I_s \subseteq I_f$, then I_s is also a frequent feature set because $sup(I_s) \geq sup(I_f) \geq min\_sup$.*

Suppose a feature set I is not frequent, i.e., $sup(I) < min\_sup$. Property 4.2.1 implies that any of its superset $I' \supseteq I$ is also not frequent because $sup(I') \leq sup(I) < min\_sup$.

The Apriori algorithm follows the above property 4.2.1 to generate the L_k sets of frequent feature sets from the set L_{k-1} by following a sequence of pruning and joining steps iteratively, until no more L_k sets can be found. The two steps of this algorithm are as follows:

The join step. To find L_k, a set of candidate C_k is generated by joining L_{k-1} with itself, i.e., $L_{k-1} \bowtie L_{k-1}$. Let l_1 and l_2 be feature sets in L_{k-1}. The join $L_{k-1} \bowtie L_{k-1}$ is performed, if first (k-2) features of L_{k-1} are in common. For example, two feature sets l_1 and l_2 of l_{k-1} are joinable if they satisfy $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \ldots \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$. The notation $l_i[j]$ refers to the jth item in l_i. Apriori assumes that features within a feature set are sorted in lexicographic order such that $l_1[k-1] < l_2[k-1]$ to ensure that no duplicates are generated. By joining l_1 and l_2 the resulting feature set we get is $l_1[1], l_1[2] \ldots, l_1[k-2], l_1[k-1], l_2[k-1]$.

The pruning step. The purpose of the joining step is to generate a list of feature sets that are frequent, based on the knowledge that they are constructed from feature sets that are frequent. C_k is a superset of frequent k-feature sets, L_k. The pruning step removes all non-frequent feature sets that occur in C_k. The resulting list after the pruning is L_k. To count the support of each candidate feature set in C_k, a scan of the dataset D is required, where the support count for each feature set l_i in C_k is calculated.

Algorithm 1 Frequent Feature Sets Mining

```
Input: A set of reviews data $R$.
Input: User-specified minimum support $min\_sup$.
Output: Sets of frequent features $L_1, \ldots, L_k$ with $sup(I_f)$.
Method:
L_1 = all frequent 1-feature sets in $R$;
for $(k = 2; L_{k-1} \neq \emptyset; k++)$ do
    $C_k = L_{k-1} \bowtie L_{k-1}$;
    for all feature sets $I \in C_k$ do
        if $\exists I_s \in I$ such that $I_s \notin L_{k-1}$ then
            $C_k = C_k - I$;
        end if
    end for
    $sup(I) = 0$ for every $I \in C_k$;
    for all reviews $r_i \in R$ do
        for all $I \in C_k$ do
            if $I \subseteq r_i$ then
                $sup(I) = sup(I) + 1$;
            end if
        end for
    end for
    $L_k = \{I \in C_k \mid sup(I) \geq min\_sup\}$;
end for
return $L_1, \ldots, L_k$ with $sup(I_f)$;
```

Algorithm 1 identifies all frequent feature sets by efficiently pruning all feature sets that are not frequent based on the Apriori property. Specifically, the algorithm finds the frequent k-feature sets from frequent $(k-1)$-feature sets based on the Apriori property. In the first iteration, the frequent 1-feature set, denoted by L_1, is found by scanning the reviews once and counting the support count for each feature. The support count of a feature set I, denoted by $sup(I)$, is the number of reviews containing I. The frequent 1-feature sets are then used to identify the candidate 2-feature sets, denoted by C_2. Then the algorithm scans the reviews once to count the support of each candidate feature set in C_2. All candidates that have support counts greater than or equal to min_sup are frequent 2-feature sets, denoted by L_2. The algorithm repeats the process of generating L_k from L_{k-1} and stops if L_{k-1} is empty.

The challenge is how to efficiently generate the candidate k-itemsets C_k from L_{k-1}. Two frequent $(k-1)$-feature sets are joinable to form a candidate k-feature set in Line 3 only if their first $(k-2)$-feature sets are identical. This process follows the Apriori property: a feature set I cannot be frequent if any of its subsets is not frequent. Thus, the only prospective frequent feature sets of size k are those that are generated by joining frequent $(k-1)$-feature set. Lines 4-8 describe the procedure of removing candidates that contain at least one non-frequent $(k-1)$-feature set. Lines 9-16 describe the procedure of scanning the reviews and obtaining the support count of each feature set I in C_k. If a review r_i contains a feature set I, $sup(I)$ is incremented by 1. If $sup(I_f)$ is larger than the user-specified minimum support threshold min_sup, then I_f is added to L_k, the frequent k-feature set with k elements. The algorithm terminates when the frequent L_k is empty, i.e., when none of the candidate feature set can pass the min_sup threshold. Finally, the algorithm returns all frequent feature sets with their support counts.

The following example shows how to use the Apriori algorithm to identify all frequent feature sets.

Example 3 (Frequent Feature Set Discovery). Consider a collection of reviews $R = \{r_1, \ldots, r_n\}$ where each review contains user categories $C(r_i)$ and their preferred tablet features $F(r_i)$. Each review r_i is represented as a doublet, denoted by $\langle F(r_i), C(r_i)\rangle$, where $F(r_i) \subseteq F$ and $C(r_i) \subseteq C$, as shown in Table 3. Suppose $min\_sup = 0.4$. The algorithm identifies frequent-1 feature set by scanning the R once. In this example the set of frequent 1-feature sets is:

$L_1 = \{\{\textit{personal}\}, \{\textit{long battery life}\}, \{\textit{screen size 7 inch}\}, \{\textit{front facing camera}\}\}$, with

$sup(\{\textit{personal}\}) = 7$

$sup(\{\textit{long battery life}\}) = 10$

$sup(\{\textit{screen size 7 inch}\}) = 5$

$sup(\{\textit{front facing camera}\}) = 4.$

Next the algorithm generates the set of candidate 2-feature sets by joining L_1 with itself, $L_1 \bowtie L_1$:

$C_2 = \{\{\textit{personal, long battery life}\}, \{\textit{personal, screen size 7 inch}\}, \{\textit{personal, front facing camera}\}, \{\textit{long battery life, screen size 7 inch}\}, \{\textit{long battery life, front facing camera}\}, \{\textit{screen size 7 inch, front facing camera}\}\}.$

Then the algorithm scans the reviews to obtain the frequent-2 feature sets, and determines:

$L_2 = \{\{\textit{ personal, long battery life}\}, \{\textit{personal, screen size 7 inch}\}, \{\textit{long battery life, screen size 7 inch}\}, \{\textit{long battery life, front facing camera}\}\}.$

Similarly, the algorithm performs $L_2 \bowtie L_2$ to generate the set of candidate 3-feature sets:

$C_3 = \{\{\textit{personal, long battery life, screen size 7 inch}\}, \{\textit{personal, long battery life, front facing camera}\}, \{\textit{long battery life, screen size 7 inch, front facing camera}\}\},$

Then the algorithm obtains the frequent-3 feature sets,

$L_3 = \{\textit{personal, long battery life, screen size 7 inch}\}.$

Finally, the algorithm terminates and returns the frequent feature sets L_1, L_2, and L_3 with $sup(I_f)$ for each $I_f \in L1 \cup L2 \cup L3$.

4.3 Generating Interesting Association Rules

Refer to Definitions 3 and 4 for the notions of association rules and interesting association rules. The following two steps generate all interesting association rules from the frequent feature sets.

- For each frequent feature set I_f, generate all nonempty subsets of I_f.

- For every nonempty subset I_s of I_f, generate the rule in the form of $I_s \rightarrow (I_f - I_s)$ if $\frac{sup(I_f)}{sup(I_s)} \geq min\_conf$, where min_conf is the minimum confidence threshold.

Table 3 contains a frequent feature set $I_f = \{$*personal*, *long battery life*, *screen size 7 inch*$\}$. The non-empty subsets of I_f are $\{\{$*personal, long battery life*$\}, \{$*long battery life, screen size 7 inch*$\}, \{$*personal, screen size 7 inch*$\}, \{$*personal*$\}, \{$*long battery life*$\}, \{$*screen size 7 inch*$\}\}$. The resulting association rules are:

- $long\ battery\ life \rightarrow personal[conf = 70\%]$

- $personal \rightarrow long\ battery\ life[conf = 100\%]$

- $personal \rightarrow screen\ size\ 7\ inch[conf = 71\%]$

- $long\ battery\ life \rightarrow screen\ size\ 7\ inch[conf = 50\%]$

- $screen\ size\ 7\ inch \rightarrow long\ battery\ life[conf = 100\%]$

- $screen\ size\ 7\ inch \rightarrow personal[conf = 100\%]$

- $personal \bigwedge long\ battery\ life \rightarrow screen\ size\ 7\ inch\ [conf = 71\%]$

- $long\ battery\ life \bigwedge screen\ size\ 7\ inch \rightarrow personal\ [conf = 100\%]$

- $personal \bigwedge screen\ size\ 7\ inch \rightarrow long\ battery\ life\ [conf = 100\%]$

- $personal \rightarrow long\ battery\ life \bigwedge screen\ size\ 7\ inch\ [conf = 71\%]$

- $long\ battery\ life \rightarrow screen\ size\ 7\ inch \bigwedge personal\ [conf = 50\%]$

- $screen\ size\ 7inch \bigwedge long\ battery\ life \rightarrow personal\ [conf = 100\%]$

If the user-specified confidence threshold min_conf is 60%, then all of the above rules are interesting except the fourth and eleventh rules with 50% confidence.

Suppose the user sets $min\_sup = 40\%$. The frequent feature set $I_f = \{$*long battery life, screen size 7 inch*$\}$. The resulting association rules from this frequent feature set are:

- *long battery life* → *screen size 7 inch* $[conf = 50\%]$
- *screen size 7 inch* → *long battery life* $[conf = 100\%]$

If the user-specified confidence threshold min_conf is 60%, then the only interesting rule is *screen size 7 inch* → *long battery life*.

4.4 Generating Constraint-based Association Rules

Not all interesting association rules are important for the designers of tablet PC. The extracted association rules are useful only if they follow specific types of patterns. Specifically, only the association rules between user categories and features and the association rules among features can be utilized in product design analysis. Thus, other rules that do not follow these patterns are removed.

The first type of constraint-based association rule is to discover the associations between different categories of users and their preferred feature sets, where the antecedent is a user category and the consequent is a set of features. In other words, the first type of rule satisfies the pattern $I_t \rightarrow I_h$, where $I_t \subseteq C$ and $I_h \subseteq F$, where C is the universe of possible user categories and F is the universe of possible features. A sample rule of this pattern is $personal \rightarrow long\ battery\ life \bigwedge screen\ size\ 7\ inch$. This rule shows that tablet for *personal users* should include features *long battery life* and *screen size 7 inch*. From this rule the designer may want to further investigate why personal user prefers *screen size 7 inch*, but not *screen size 10 inch*, and compare the feature preference of personal category users with the preference of other user categories. Different customers have different preferences for the features of tablet PC specifications. The result of association rules between the three user categories and their preferred feature sets are compared with each other for product segmentation.

The second type of constraint-based association rule is to discover the associations among features, where the antecedent and consequent are disjoint sets of features, where $I_t \subseteq F$, $I_h \subseteq F$, and $I_t \cap I_h = \emptyset$. A sample rule of this pattern is *screen size 7 inch* $\rightarrow$ *long battery life*. The rule suggests that *screen size 7 inch* and *long battery life* should be included together in a tablet design. The second type of rules can be utilized for feature selection at the time of designing the tablet PC specifications regardless of user categories. The feature items in a rule indicate that if the feature in the antecedent is included in the specification, then the feature in the consequent also should be included. This feature selection method will make the tablet PC specification more attractive to the users. More analysis on real-life online reviews will be given in the next chapter.

Misleading association rule. A rule is interesting if it passes the minimum support and minimum confidence thresholds; however, it does not necessarily imply that the antecedent and consequent of the rule have positive correlation. An interesting association rule that has no positive correlation is misleading. Suppose there are 10,000 personal users. 5,000 of them prefer *long battery life* and 8,000 of them prefer *screen size 7 inch*, and 3,000 of them prefer both *long battery life* and *screen size 7 inch*. Using $min\_sup = 20\%$ and $min\_conf = 50\%$, the following association rule is discovered:

prefer (*personal user, "long battery life"*) $\rightarrow$ *prefer* (*personal user, "screen size 7 inch*") $[sup = 30\%,\ conf = 60\%]$

The above rule is an interesting association rule with respect to the minimum support and minimum confidence. However, the above rule is misleading and negatively correlated because the ratio of personal users who prefer *screen size 7 inch* is 80% which is larger than 60%, implying that the personal users who prefer *long battery life* indeed has interest in *screen size 7 inch*. To identify and filter out this kind of misleading interesting rules, a correlation measure called *Lift* can be applied. *Lift* is a correlation measure that is used to find out the interestingness of the association rules. The lift between occurrence of I_t and I_h can be measured by Equation 1:

$$lift(I_t, I_h) = \frac{conf(I_t \rightarrow I_h)}{sup(I_h)} \tag{1}$$

If the lift value is greater than 1, then there is a positive correlation between I_t and I_h. If the lift value is 1, then I_t and I_h are independent and there is no correlation between them. If the lift value is below 1, then there is a negative relationship between I_t and I_h. Thus, the association rules with lift value less than 1 are removed.

For example, *Lift* (*personal* $\rightarrow$ *long battery life* $\bigwedge$ *screen size 7 inch*) $= 0.71/0.5 = 1.42$. The lift of this rule is greater than 1, so there is a positive correlation between the occurrence of $\{$*personal*$\}$ and $\{$*long battery life* $\bigwedge$ *screen size 7 inch*$\}$.

For the previously mentioned misleading association rule, *prefer* (*personal user, "long battery life"*) $\rightarrow$ *prefer* (*personal user, "screen size 7 inch*") $[sup = 30\%,\ conf = 60\%]$, the lift is $0.60/0.80 = 0.75$. The lift of this rule is less than 1, so the occurrence of *long battery life* and *screen size 7 inch* are negatively correlated.

4.5 Summary

In this chapter, we have presented a method to extract the interesting association rules from the online reviews. These rules can be utilized for tablet PC feature classification for different categories of users and for tablet PC feature selection. For extracting these association rules, the first step is to collect the product features and user categories from the online reviews. The following steps are to identify frequent feature set and generate interesting association rules from the frequent feature set. These interesting association rules help to identify the tablet PC feature classification. Based on this proposed method we have conducted an experiment, which will be discussed in the next chapter.

Chapter 5

Experimental Evaluation

Based on the proposed method in Chapter 4, we have conducted an experiment with the data of 304 online reviews of tablet PC users. The reviews were collected from online discussion site Amazon.com and Cnet.com. The objective of the experiment is to examine the effectiveness of using online reviews for tablet PC feature classification for different users. This chapter explains the observation from the experimental result. First, we provide an overview of the dataset and generation of interesting association rules from dataset, followed by the observation from association rules concerning the tablet PC model classification. As the tablet market is growing rapidly many new tablet models have been launched after the collected review postings. Thus, here we also want to compare the mined association rules with the existing tablet models to examine the proposed method.

5.1 Data Sets

At first the tablet reviews have been collected from Amazon.com and Cnet.com and 304 reviews have been collected within the time duration of April, 2010 and May, 2011. For this experiment three user categories have been fixed (*personal, student, business*) and 47 features of tablet PC have been used. We assumed that a review was given by any of the

User categories(C)	**Possible feature set for a tablet PC(F)**
personal, business, student	*bluetooth, long battery life, good browser, rear facing camera, front facing camera, customizable, card reader, beautiful design, nice display, ease of use, e-reader, facial recognition, flash, flip screen, gps, gorilla glass screen, graphics, hand writing recognition, HDMI, good keyboard, onscreen keyboard, multi tasking, multi touch, multimedia, OS Android, OS Windows, OSi, pen, phone, portable, fast processor, good speaker, storage, sd slot, screen resolution, screen size 7inch, screen size 10 inch, text to speech application to google maps, good touch screen, good touch pad, usb, voice dialing, video quality, video recording, light weight, Wi-Fi*

Table 4: Attributes of dataset

three user categories based on the users purpose of usage. *Personal* users are those users who are using the tablet only for their personal or home use (e.g., watching movie at home). *Personal* users do not use the tablet for professional purpose or study purpose. Users are called *business* users when in their postings it is found that they are telling about their usage for professional purposes (e.g., editing business report, using tablet in office presentation). *Student* users are those users who said about their usage of tablet for study purpose (e.g., taking notes in class).

The set of preferred features mentioned by the reviewer and his/her user category are extracted manually for each review. For the attributes of the dataset all possible features of tablet PC and user categories are listed in Table 4.

All the extracted preferred feature sets and user categories for all users are entered for making the dataset. Each row of the dataset represents the information about a tablet PC user's category and preferences of features. We assigned value of '1' for each of the preferred feature and corresponding user category in each row and for the other attributes we assigned value of '0'. In this experiment a data mining tool named RapidMiner 5 has been used to generate the association rules from this dataset.

5.2 Frequent Feature and Association Rule Generation

The dataset has been used to determine the association rules between user categories and feature set, and the association rules among the features. The association rules were mined through Apriori algorithm [5] using minimum support of 3% and minimum confidence of 10%. Generally for association rule mining the minimum support threshold of 1% to 10% is widely used. The resultant tables are composed of user categories and user's preferred feature sets. According to our pre-set rule constraint we extracted two types of association rules from the result. The first type of association rules are the rules with the user category in the antecedent and features in the consequent. The second type of association rules are the rules with features in both side of antecedent and consequent. After extracting these constraint-based rules the lift has been calculated to find out the correlation between the antecedent and consequent items. This lift helps to identify the rule interestingness. The interesting rules whose lift are greater than one are kept and others are removed. The first type of resultant association rules are divided into three types based on the three user categories.

5.3 Experimental Results

In this section we have an objective to justify the extracted two types of association rules which are association between *users* and *features*, and association among *features*. We tried to see if the extracted association rules between user categories and features ($user \rightarrow features$) represent the meaningful need of tablet PC models for different categories of users. In this section, we also analyze the resultant association rules among features ($features \rightarrow features$) to use those rules for the bundle of feature selection for tablet PC models.

5.3.1 Association Rules between User Categories and Tablet PC Features

Table 5, Table 6 and Table 7 show the result of association rules between three categories of tablet PC users and tablet PC features. These three tables are analyzed for getting the insights to classify tablet PC features for different categories of users.

No	Association Rule	Support	Confidence	Lift
1	*personal*→ *long battery life*	0.230263	0.29787234	1.040841
2	*personal*→ *good speaker*	0.125	0.16170213	1.170415
3	*personal*→ *light weight*	0.111842	0.14468085	1.157447
4	*personal*→ *beautiful design*	0.101974	0.13191489	1.055319
5	*personal*→ *OS Android*	0.095395	0.12340426	1.103379
6	*personal*→ *screen size 7 inch*	0.095395	0.12340426	1.250496
7	*personal*→ *fast processor, nice display*	0.095395	0.12340426	1.136815
8	*personal*→ *usb*	0.092105	0.11914894	1.065332
9	*personal*→ *front facing camera*	0.078947	0.102127	1.070579

Table 5: Association rules between personal category users and tablet PC features based on online review

Association between Personal Category Users and Tablet PC Features. Rules of the Table 5 represent the associations between $personal$ category users and their preferred feature set. From the mined result of the above table we can assume that the preferred feature set for *personal user* is {*long battery life, good speaker, light weight, beautiful design, OS android, screen size 7inch, fast processor, nice display, usb, front facing camera*}. The feature list of this set is sorted from higher preference to lower preference according to the confidence level of the association rules.

Association between Business Category Users and Tablet PC Features. Rules of the Table 6 represent the associations between *business* category users and their preferred feature set. The preferred feature set for the *business* category user is { *touch screen, keyboard, OS windows, portable, Wi-Fi*}. Here the feature list is sorted in the order of more preferred feature first and less preferred feature at last according to the confidence

No	**Association Rule**	**Support**	**Confidence**	**Lift**
1	*business→ good touch screen*	0.085526	0.34210526	1.333333
2	*business→ good keyboard*	0.065789	0.26315789	1.454545
3	*business→ OS Windows*	0.05	0.19736842	2.068966
4	*business→ portable*	0.04	0.14473684	1.073171
5	*business→ Wi-Fi*	0.04	0.14473684	1.1
6	*business→ good touch screen, good keyboard*	0.04	0.14473684	1.76

Table 6: Association rules between business category users and tablet PC features based on online review

level of the association rules.

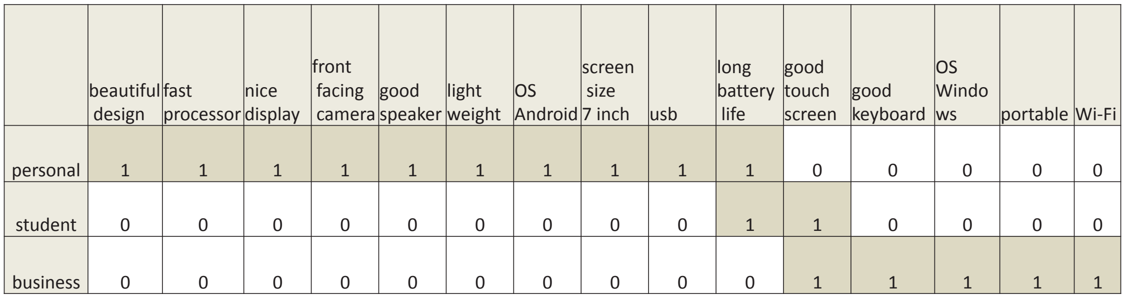
Association between Student Category Users and Tablet PC Features. The rules of the Table 7 represent the associations between *student* category users and their preferred features. It can be suggested that preferred feature set of tablet for *student* category user is {*long battery life, good touch screen*}.

No	Association Rule	Support	Confidence	Lift
1	*student→ long battery life*	0.032895	0.3030303	1.058865
2	*student→ good touch screen*	0.032895	0.3030303	1.181041

Table 7: Association rules between student category users and tablet PC features based on online review

Comparison of the Association Rules between Three User Categories and Tablet PC Features. The rules in Table 5, Table 6, Table 7 essentially indicate which tablet features are mentioned frequently by respective categories of users. To reveal the distribution of tablet features with user categories, a matrix is made in Figure 5. Particularly, the matrix's rows are labeled with user categories, and the matrix's columns are labeled with mined tablet features. The matrix entries show whether the tablet features are associated with the user categories (where '1' and '0' represent the presence and absence of associations, respectively).

From Figure 5, it is interesting to note that *personal* users and *business* users have strong association with two different tablet feature sets. However, these two categories of users both consider *portability* as a preferred attribute (i.e., *personal* user mentions *light weight* and *screen size 7 inch*, and *business* user mentions *portable*).



	beautiful design	fast processor	nice display	front facing camera	good speaker	light weight	OS Android	screen size 7 inch	usb	long battery life	good touch screen	good keyboard	OS Windows	portable	Wi-Fi
personal	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
student	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
business	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Figure 5: Association between user categories and tablet features

In view of their differences, we argue that *business* users treat tablets as devices supporting their office activities. Thus, how tablets can be integrated with their existing computing systems is the primary concern. This explains why *business* users prefer *OS Windows* though tablets running Windows are not readily available in the market yet. Also, the mentioned *Wi-Fi* feature should be used for connecting the company's system. Furthermore, the *business* users are interested in the input/output methods with the tablets as they expect a good experience with the features *touch screen* and *keyboard*.

In contrast, it seems that *personal* users focus more on the entertainment features. The features *fast processor*, *nice display* and *good speaker* are related to the experience of gaming, watching video and listening to music. Also, *personal* users frequently mention *beautiful design* in order to reflect their personal characters on their mobile gadgets. Furthermore, *personal* users expect certain openness with the tablets in terms of mentioning *OS Android* as an open-source system and *usb* as the standard interface for transferring files.

In our study, *student* users do not appear frequently in the review postings since the activities specific to *student* users (e.g., take notes, do homework, etc) are not often mentioned. From this observation we think that tablets are not yet considered as conventional device for student activities. Tablets can be argued as a popular device for students, as currently promoted by tablet manufacturers. In our study, the tablet usages by *student* users cannot be easily distinguished.

5.3.2 Association Rules among Tablet PC Features

In addition to the rules between tablet PC *user categories* and tablet*features*, 7 rules among *features* $(features \rightarrow features)$ are also generated, and they are listed in Table 8. In this section we analyze these rules to see whether these rules are meaningful for the tablet manufacturers for the feature selection of tablet PC models.

No	Association Rule	Support	Confidence	Lift
1	$good\ touch\ screen \rightarrow long\ battery\ life$	0.085526	0.333333	1.164750
2	$long\ battery\ life \rightarrow good\ touch\ screen$	0.085526	0.298850	1.164750
3	$long\ battery\ life \rightarrow screen\ size\ 10.1\ inch$	0.069078	0.2413793	1.157447
4	$front\ facing\ camera \rightarrow rear\ facing\ camera$	0.052631	0.551724	7.986863
5	$fast\ processor \rightarrow Wi - Fi$	0.05	0.178571	1.357142
6	$Wi - Fi \rightarrow nice\ display$	0.05	0.35	1.33
7	$nice\ display \rightarrow Wi - Fi$	0.05	0.175	1.33

Table 8: Association rules among tablet PC features based on online review

Tablet PC Feature Selection from Association Rules among Tablet PC Features. The resultant association rules between the tablet PC features can be utilized for the bundle of feature selection for tablet PC models. Table 8 shows the association rules among the features of tablet PC regardless of user categories.

From rules 1, 2 and 3 of Table 8 we see *long battery life* has a strong association with features *good touch screen* and *screen size 10.1 inch*. Every user loves to have a *long battery life* for their tablet PC while some of them also want to have a *good touch screen* in their tablet which are suggested by rules 1 and 2. Rule 3 of Table 8 suggests that *long battery life* and *screen size 10.1 inch* have strong association. Among these rules, it is observed that the feature *long battery life* is mentioned most often (regardless of user categories). It is not a surprising result as the battery performance is still difficult to improve due to the technological limitations. We can also assume that some users want *10.1 inch screen* for their tablet PC. So manufacturers can make *screen size 10.1 inch* tablet PC for some portion of the users.

From rule 4 of Table 8 we see that tablet users who want *front facing camera*, 55% of them also want *rear facing camera*, so the manufacturers should include both *front and rear facing cameras* together to make the tablet PC more attractive to users or they may exclude both cameras from the feature list which will be attractive for another portion of the users. In this way manufacturers can reduce their manufacturing cost and attract targeted customers.

Rule 5 of this table shows the users who want *fast processor*, 17% of them also want *Wi-Fi* as connectivity interface. According to rule 5, the suggestion is, if the tablet PC have a *fast processor* in the feature list then the manufacturer should include a *good connectivity interface* for that model of tablet PC. From rule 7 of Table 8 it can be suggested that if the tablet PC have a *good connectivity* then it should also have a *nice display*. Here we are assuming that *Wi-Fi* indicates about the *good connectivity* feature of tablet PC. The feature *Wi-Fi* is also frequently mentioned, and it is associated with *fast processor* and *nice display*. As the feature *Wi-Fi* is commonly available for tablets, the real concern of the customers who mention *Wi-Fi* should be about internet experience. At this point, the manufacturers should pay attention to the *speed* and *viewing* experience of internet usages.

So from the Table 8 we suggest manufacturers can use these rules to select the bundle of features for tablet PC which will be more attractive to customers.

5.4 Interpretation of Observation

Our research presents a method for knowledge discovery from online reviews of tablet PC users which can be utilized by the tablet designers to understand the tablet PC model classification for different users. We have three categories of tablet users which are *personal*, *business* and *student* users. On the online reviews all of these users discussed about their preferable tablet PC features. Applying the association rule mining on those online review data we found some interesting association rules between the users and tablet PC features. We also found some interesting association rules among tablet PC features. This section gives the suggestion of tablet model classification for different categories of users from the analysis of resultant association rules.

The rule for the *personal* user suggests that tablet users who are using it for personal purpose like the set of features {*long battery life*, *good speaker*, *light weight*, *beautiful design*, *OS android*, *screen size 7inch*, *fast processor*, *nice display*, *usb*, *front facing camera*}. This list of preferred features suggests *Android* as the operating system for *personal* user. As an OS Android is easy to maintain and there are many open source software available for the users. This might be tempting for the *personal* user. The feature screen size 7 inch is preferable because it is very convenient to hold and use at any place. *Battery life*, *speaker* and *display* should be good for watching movies. Users who use tablet for gaming need a *fast processor*. *USB port* is convenient for *personal* user to transfer their files like photos, videos etc. *Front facing camera* is important for the personal users because they like to take pictures with their tablet or sometimes they do video chat with it.

The resulting association rules for the *business* purpose user from our dataset is quite interesting. According to our experimental result the preferred feature set by the *business user* is {*touch screen*, *keyboard*, *OS windows*, *portable*, *Wi-Fi*}. Here we can see business users are more concerned about the *user interface* of the tablet PC. Generally business users do productive works in their tablets which need more convenient input and output methods. From this point of view *good touch screen* and *good keyboard* are important user interfaces for their purpose. *Wi-Fi* of the feature list indicates the importance of *connectivity interface* to the *business* users. To be connected with the business world *connectivity interfaces* should be in a good standard where users share their information by using *touch* and *keyboard interfaces* in a big network. The vast majority of businesses work in Outlook, Word, Excel, and PowerPoint, as well as with Microsoft Exchange. *Microsoft Windows 7* has all that and more. If we think about the *portability* it is important for the business people because they need to move with their tablet for their work, for example: meeting presentation.

The association rules among the tablet PC *features* suggest that a rule among *features* can be utilized to select bundle of *features* for tablet PC models. As those features are from users preference list, this feature selection will attract more customers.

In the association rules among *features* we see *long battery life* has strong association with *screen size 10.1 inch*. It indicates that *bigger screen* needs more *battery power* and some users also want both together. *Front facing camera* and *rear facing camera* are strongly associated, which suggests that the designers should include both in a tablet. People generally use *front facing camera* for video chatting to show their faces while chatting. By using the *rear facing camera* they will be able to show the other end of their face view which will increase the interaction between the chatter. *Rear facing camera* is also great as tablet PC is portable to go everywhere with the user so they can take informal pictures whenever they want. Some users also do not prefer any one of the cameras to have less costly tablet PC. In rule 5 of Table 8 we see *fast processor* and *Wi-Fi* have a strong association. Through *Wi-Fi* user connects to the internet where they watch online movies, play online games which need *fast processor*. These reasons may explain why the users want *fast processor* and *Wi-Fi* together. Rules 6 and 7 show *Wi-Fi* and *nice display* are strongly associated which indicate some users watch online videos by connecting to the internet through *Wi-Fi*. These users want a *nice display* for watching the video comfortably.

The key implications from the association rules between *users* and *features* are that at first, we suggest the tablet PC models for two user categories: *personal* users and *business* users. For *personal* users, the tablets should be designed with attractive exterior and optimized for entertainment purposes (i.e., games, music and video). For *business* user, the tablets should focus on the integration of office duties, with standard and high-quality input/output supports (i.e., touch screen for presentation and keyboard for productivity). As the tablet features for *student* users are not distinct enough, a customized tablet design for students may not be recommended in view of the development efforts for time being.

Some tablet features are identified as important for all user categories, including *portability*, *long battery life* and *internet experience* (associated with *Wi-Fi*). Also, if the manufacturers plan to remove *cameras* from the tablets for a budget design, it is suggested to remove both *front and rear facing cameras* rather than keeping one of them. Probably, keeping one camera would remind the customers the absence of another camera, leading to a poor perception of the product's quality.

5.5 Comparison with Existing Products

In this section we tried to assess our understanding of tablet PC feature classification for different users from the online reviews. The current stage of tablet PC is in its infancy which needs to be go further and well classified for the users. In our work we tried to get the insights about tablet PC feature classification from the users opinions and tried to compare our resultant association rules with some existing tablet models in the market.

The growth of tablet market officially started in 2010. Apple's launch of iPad started the revolution in tablet PC market. Many other manufacturers also started launching their tablets to catch the market. At present, the market of tablets is still evolving rapidly, and the manufacturers need to explore different things to increase their market share. ICD

Ultra came after iPad with *HD Flash video*, *multi touch screens*, *front-facing cameras*, *multi tasking* etc. Freescale's Smartbook with *7 inch touch screen*, *Android OS* , *Wi-Fi* and *bluetooth connectivity*, *camera* arrived at that time. Around that time Pegatron introduced Slate PC with *Windows 7*, *11.6 inch touch screen* etc. Dell streak also came with *5 inch touch screen*, *Android OS*, *Wi-Fi*, *bluetooth*, *GPS*, *camera*, *multi tasking* etc. Samsung Galaxy Tab appeared in the end of 2010 as a communication and entertainment tool with *3G connectivity*, *Wi-Fi*, *bluetooth*, *Android OS*.

From 2010 to now on tablet PC market is growing with new tablets. The market is hitting with tablets from manufacturers every month with new feature updates. Some tablet of recent time, 2011 include iPad2, Samsung Galaxy tab 8.9, Motorola Xoom, HP TouchPad, ASUS Eee Slate EP121, Lenovo IdeaPad A1, Lenovo ThinkPad and many more. Most of these tablets from the manufacturers are not targeted for a specific group of users. Various new tablet products have been launched after the postings collected in this study (i.e., May 2011). The most visible differentiating feature is the *screen size* (between 7 inches to 10 inches), and our study does not show a clear preference of the *screen size* with *user categories*. The mined rule $personal \rightarrow screen\ size\ 7\ inch$ is not very indicative for the tablet design.

Our understanding from this work suggests two models of tablet PC for two categories of users which are *personal user* and *business user*. Our suggestion from the experimental result resembles the current two models of Lenovo tablet. Lenovo has introduced IdeaPad Tablet A1, which is targeted to *personal* users and ThinkPad tablet which is targeted for *business* users. Notably, the release time of ThinkPad Tablet is after the time of review collection in our study. Lenovo introduced ThinkPad Tablet (released in July 2011) dedicated for business users. This product is featured with a digitalized pen and a keyboard accessory to support office tasks. This actual tablet development aligns with the design insights discussed in Section 5.4. We consider that this example supports our work of tablet model classification from online reviews using data mining technique. Also, we do not find tablet products that are dedicated for student users only, and this observation aligns with our discussion of the experimental results.

Feature list of personal user	**Feature list of IdeaPad Tablet A1**
long battery life	*7 hour battery life*
good speaker	*high definition audio*
light weight	*weight of 0.88 pound*
beautiful design	*beautiful design with four different color options*
OS Android	*OS Android 2.3*
screen size 7 inch	*7.0 inch SD-LED display*
fast processor	*Processor 1GHz*
nice display	*display with 1024-by-600-pixel resolution*
usb	*MicroUSB ports*
front facing camera	*0.3 megapixel front facing camera*

Table 9: Comparison between the feature list of *personal* user and feature list of Lenovo IdeaPad A1 tablet PC

The feature list for IdeaPad Tablet A1 is {*7 hour battery life*, *high definition audio*, *weight of 0.88 pound*, *beautiful design with four different color options*, *OS Android 2.3*, *7.0 inch SD-LED display*, *Processor 1GHz*, *display with 1024-by-600-pixel resolution*, *MicroUSB ports*, *Wi-Fi 802.11*, *Optional 3G version*, *Bluetooth*}. The comparison of this feature list with our resulting feature list for *personal* user is shown in Table 9. In Table 9 we see Lenovo IdeaPad has a *good battery life*, *light weight*, *Android OS*, *usb*, *front facing camera* which are similar to our feature list of *personal* users. Our list of *personal* user features also include *design* which means the look of the tablet and for design thinking Lenovo IdeaPad has been offered in four different colors. Lenovo IdeaPad A1 has *1 GHz processor* which is sufficient for *personal* users purpose.

The feature list for ThinkPad Tablet is {*HD multi touch screen with gorilla glass and pen input*, *keyboard folio*, *OS Android 3.1*, *weight starting at 1.6 lbs*, *8 hours battery life*, *Wi-Fi*, *3G*, *bluetooth*}. The comparison of our resulting feature list for *business* user with the feature list of ThinkPad tablet is shown in Table 10. In Table 10 we see the similarity

Feature list of business user	**Feature list of ThinkPad Tablet**
good touch screen	*HD multi touch screen with gorilla glass and pen input*
good keyboard	*keyboard folio*
OS Windows	*OS Android 3.1*
portable	*portable*
Wi-Fi	*good connectivity interfaces (Wi-Fi, bluetooth, GPS)*

Table 10: Comparison between the feature list of *business* user and feature list of Lenovo ThinkPad tablet PC

in *good touch screen* as ThinkPad has *Gorilla glass multi touch HD screen*. Our resultant feature list for business user has *good keyboard* while ThinkPad tablet came with a folio for keyboard. For differentiating the *keyboard* Lenovo offered this *keyboard* folio. *Wi-Fi* in our feature list indicate a *good connectivity* and ThinkPad also has *good connectivity interfaces* with *Wi-Fi*, *3G* and *bluetooth*. The *screen size* and *weight* of ThinkPad also indicate that it has good *portability*. Our feature list has a difference with ThinkPad for the *Android OS*. Our feature list contains *Windows* as *OS*. ThinkPad is the first announced *business* tablet by a manufacturer which does not necessarily mean that all other *business* tablet will also come in *Android*. Recently Microsoft announced their new *Windows 8 OS* for tablet. So there is probability that future *business* tablets will have *Windows* as *OS* because Windows is the most popular operating system among the users.

We are now in the early stage of next generation tablet PC models which will be well classified for the users. Thus it is important to select the best assortment of features which will attract more customers. For this purpose manufacturers can use the association rules among *features* of tablet PC which will help them in tablet PC feature selection. In our resultant association rules among *features* in Table 8 we see the rule *front facing camera*$\rightarrow$ *rear facing camera* [sup = 0.052631, conf = 0.551724], from this rule we are suggesting that when the manufacturers are including the *front facing camera* in the specification list of tablet PC model then they should also include *rear facing camera* or they should exclude both from the feature list to reduce the cost for some group of customers. Among the renowned tablet PC of 2011 Asus Eee Pad Transformer Prime, Asus Eee Pad Slider SL101, Samsung Galaxy Tab 8.9, Asus Eee Pad Transformer TF101, Samsung Galaxy Tab 10.1, Motorola Xoom 2, BlackBerry Playbook, all have both front facing and rear facing camera which support our suggestion of feature selection rule.

5.6 Summary

To examine the effectiveness of our proposed method in Chapter 4, we have conducted an experiment with the real tablet users' data from online reviews. The experimental result shows that the tablet features can be differentiated for different categories of users based on the extracted interesting association rules from the online reviews. We have also compared the experimental result with the existing models of tablet in the market and the comparison supports our study. In the next and last chapter of this thesis, the conclusion with some future works will be presented.

Chapter 6

Conclusion

This chapter concludes the thesis. First, we give a summary of this work, then we describe the research directions that can be conducted as future work.

6.1 Summary and Conclusions

In this thesis, we have proposed a method to utilize the online reviews for classifying the tablet PC features. The evolution of technology allowed users to share their experiences of product usage on the online discussion sites. Those reviews are easily available, less costly, free from bias and also time saving in contrast to traditional survey methods. As tablet PC is currently considered as a new class of product we have chosen tablet PC for our experiment to mine interesting pattern from the online reviews to classify the tablet PC models for different categories of users.

In our method we considered three categories of users: *personal*, *business* and *student*. We extracted users categories and their preferred feature sets from the online reviews and used these information as our input to determine the associations between the user categories and their preferred feature sets. Then we analyzed the output to classify the tablet

PC models for different categories of users. We also mined the associations among features which are analyzed as secondary rule to help the tablet design analysis regardless of user categories. After the analysis of our experimental result we have suggested two types of tablet models for *personal* users and *business* users. The most available tablets are for *personal* users who want the tablet features for entertainment purpose, while the *business* users are concerned to use the tablet in their office environment with better input methods, for example- *touch screen* and *keyboard*. We compared our result analysis with existing tablet of Lenovo in the market which support this study.

We believe that this problem will become increasingly important to the product designer as more people are using and providing their comments over online discussion sites. To the best of our knowledge there are no existing study that provide method to use online reviews for tablet model classification for different users. The effectiveness of our proposed method can be enhanced by the following future works.

6.2 Future Work

For future work, we identify several potential research directions.

Firstly, in this thesis we have used 304 user reviews for the experiments as the extraction was manual. So the extraction time was higher. Many works [25], [17], [45] have been done to extract product features from online reviews and there are many works on categorizing the user types based on the review texts [6], [30], [22]. As the online reviews are written in natural language, natural language processing in this context is still technically challenging. In future, it should be worth to integrate these two works in context of our problem to utilize an automated feature and user category extraction technique. It would be interesting to study how to extract the feature set for each user category more efficiently. This study will improve the method of mining online reviews to find the associations between user category and product features.

In our method, we limited the dataset with the preferred feature set by each user. The reason for that is because preferred features by the users are the most important specifications which they want most than the other features. However, it would be also interesting to consider the other features which they do not like. For example, if the user posts in the review that he or she does not like the *screen size 7 inch* then we should find out what size is preferable by this user from the other lines of his or her posting.

Bibliography

[1] Wikipedia, http://en.wikipedia.org/wiki/portable-computer.

[2] Rapidminer, http://rapid-i.com/content/view/181/190/. 2009.

[3] B. Agard and A. Kusiak. Data-mining-based methodology for the design of product families. *International Journal of Production Research*, 42(15):2955–2969, 2004.

[4] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22:207–216, 1993.

[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[6] C. Apte, F. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst.*, 12(3):233–251, 1994.

[7] J. K. Bae and J. Kim. Product development with data mining techniques: A case on design of digital camera. *Expert Syst. Appl.*, 38:9274–9280, 2011.

[8] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 436–442, New York, NY, USA, 2002. ACM.

[9] R. Brown. Managing the ’s’ curves of innovation. *Journal of Consumer Marketing*, pages 61–73, 1992.

[10] G. Carenini, R. T. Ng, and E. Zwart. Extracting knowledge from evaluative text. In *Proceedings of the 3rd international conference on Knowledge capture*, K-CAP ’05, pages 11–18, New York, NY, USA, 2005. ACM.

[11] K. M. A. Chai, H. L. Chieu, and H. T. Ng. Bayesian online classifiers for text classification and filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR 02, pages 97–104, New York, NY, USA, 2002. ACM.

[12] Y. Chen, K. Tang, R. Shen, and Y. Hu. Market basket analysis in a multiple store environment. *Decision Support Systems*, 40(2):339 – 354, 2005.

[13] S. Ciccantelli and J. Magidson. From experience: Consumer idealized design: Involving consumers in the product development process. *Journal of Product Innovation Management*, 10(4):341 – 347, 1993.

[14] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’92, pages 318–329, New York, NY, USA, 1992. ACM.

[15] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, WWW ’03, pages 519–528, New York, NY, USA, 2003. ACM.

[16] X. Ding. Opinion and entity mining on web content. Illinois, USA, 2010.

[17] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 231–240, New York, NY, USA, 2008. ACM.

[18] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 213–222, New York, NY, USA, 2007. ACM.

[19] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 419–428, New York, NY, USA, 2005. ACM.

[20] A. Griffin and J. R. Hauser. The voice of the customer. *Found. Trends Inf. Retr.*, 12:1–27, 1993.

[21] S. Hamm. *The Race for Perfect: Inside the Quest to Design the Ultimate Portable Computer*. McGraw-Hill, New York, 2009.

[22] E. Han, G. Karypis, and V. Kumar. Text categorization using weight adjusted k - nearest neighbor classification. In David Cheung, Graham Williams, and Qing Li, editors, *Advances in Knowledge Discovery and Data Mining*, volume 2035 of *Lecture Notes in Computer Science*, pages 53–65. Springer Berlin /Heidelberg, 2001.

[23] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29:1–12, 2000.

[24] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[25] M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence*, AAAI'04, pages 755–760. AAAI Press, 2004.

[26] C. Huang, M. Chen, and C. Wang. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4):847 – 856, 2007.

[27] S.C. Hui and G. Jha. Data mining for customer service support. *Information Management*, 38(1):1 – 13, 2000.

[28] T. Kim, G. E. Okudan, and M. Chiu. Product family design through customer perceived utility. In *Proceedings of the ASME 2010 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Montreal, Quebec, Canada, 2010. ASME.

[29] W. Y. Kim, J. S. Ryu, K. I. Kim, and U. M. Kim. A method for opinion mining of product reviews using association rules. In *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, ICIS '09, pages 270–274, New York, NY, USA, 2009. ACM.

[30] W. Lam and K. Lai. A meta-learning approach for text categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 303–309, New York, NY, USA, 2001. ACM.

[31] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 16–22, New York, NY, USA, 1999. ACM.

[32] H. Li and K. Yamanishi. Mining from open answers in questionnaire data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 443–449, New York, NY, USA, 2001. ACM.

[33] S. Li and Z. Chen. Exploiting web reviews for generating customer service surveys. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, pages 53–62, New York, NY, USA, 2010. ACM.

[34] S. Liao and Y. Chen. Mining customer knowledge for electronic catalog marketing. *Expert Systems with Applications*, 27(4):521 – 532, 2004.

[35] S. Liao, Y. Chen, and M. Deng. Mining customer knowledge for tourism new product development and customer relationship management. *Expert Systems with Applications*, 37(6):4212 – 4223, 2010.

[36] Y. Liu, X. Huang, A. An, and X. Yu. Arsa: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 607–614, New York, NY, USA, 2007. ACM.

[37] T. J. Marion and T. W. Simpson. Platform leveraging strategies and market segmentation. pages 73–90, 2006.

[38] R. J. Mooney and R. Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7:3–10, 2005.

[39] U. Y. Nahm. Text mining with information extraction. 2004.

[40] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, 2008.

[41] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[42] C. Plaisant, J. Rose, B. Yu, L. Auvil, M. G. Kirschenbaum, M. N. Smith, T. Clement, and G. Lord. Exploring erotics in emily dickinson's correspondence with text mining and visual interfaces. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '06, pages 141–150, New York, NY, USA, 2006. ACM.

[43] A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[44] G. Qiu, B. Liu, J. Bu, and C. Chen. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st international jont conference on Artifical intelligence*, IJCAI'09, pages 1199–1204, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[45] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin. Red opal: product-feature scoring from reviews. In *Proceedings of the 8th ACM conference on Electronic commerce*, EC '07, pages 182–191, New York, NY, USA, 2007. ACM.

[46] D. Tapscott and Williams. *Wikinomics: How Mass Collaboration Changes Everything*. Portfolio, New York, 2006.

[47] C. Tsai, P. Chang, and S. Wang. Applying association-rule techniques and artificial neural networks to product development. *Journal of the Chinese Institute of Industrial Engineers*, 20(2):101–112, 2003.

[48] M. J. Zaki. Scalable algorithms for association mining. *IEEE Trans. on Knowl. and Data Eng.*, 12:372–390, 2000.

[49] L. Zhang, B. Liu, S. H. Lim, and E. O'Brien-Strain. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1462–1470, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[50] Y. Zhang, J. Jiao, and Y. Ma. Market segmentation for product family positioning based on fuzzy clustering. *Journal of Engineering Design*, 18(3):227–241, 2007.

[51] Y. Zhang, N. Zincir, and E. Milios. Narrative text classification for automatic key phrase extraction in web document corpora. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, WIDM '05, pages 51–58, New York, NY, USA, 2005. ACM.

[52] L. Zhuang, F. Jing, and X. Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 43–50, New York, NY, USA, 2006. ACM.