Segmenting and Characterizing Adopters of E-Books and Paper Books Based on Amazon Book Reviews

Lu Guan[1], Yafei Zhang[1,2], and Jonathan Zhu[1(✉)]

[1] Web Mining Lab, Department of Media and Communication, City University of Hong Kong, Kowloon, Hong Kong SAR, China
lguan3-c@my.cityu.edu.hk, j.zhu@cityu.edu.hk

[2] Key Laboratory of System Control and Information Processing, Ministry of Education of China, Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China
yflyzhang@sjtu.edu.cn

Abstract. Online product reviews through which consumers express their opinions and experiences with products are extremely valuable for both potential buyers to make informed purchase decisions and retailers to improve their products/services and adjust existing marketing strategies. One of the key challenges for mining product reviews is how to obtain a "ground truth" to guide the segmentation of reviewers properly. We propose a behavior-to-opinion approach, in which users are first categorized based on some unambiguous behavioral patterns (if available) and their online reviews are then classified to reveal unique and detailed characteristics of each user category. In this paper, we identify four categories of book consumers (i.e., kindle-only, print-only, print-to-kindle, and kindle-to-print) based on the long-term patterns of their review behavior. Their review posts are then clustered through *word2vec* and K-means, and four categories of adopters are matched with their concerned word topics. Finally, we find that print-only adopters show significantly different patterns on content-oriented topics as compared to other three groups. Kindle-to-print adopters pay more attention on portability whereas print-to-kindle adopters stress more on money and user experience. Taken together, our work indicates a diversity of characteristics among four categories of book reviewers.

Keywords: Text analytics · Behavioral patterns · E-books · Product reviews

1 Introduction

Online product reviews through which consumers express their opinions and experiences with products are extremely valuable for both potential buyers to make informed purchase decisions and retailers to improve their products/services and adjust existing marketing strategies. One of the key challenges

Y. Li et al. (Eds.): SMP 2016, CCIS 669, pp. 85–97, 2016.
DOI: 10.1007/978-981-10-2993-6_7

for mining product reviews is how to obtain a "ground truth" to guide the segmentation of reviewers properly. As a practical solution to the lack of ground truth knowledge for segmentation, we propose a "behavior-to-opinion" approach as outlined in Fig. 1.

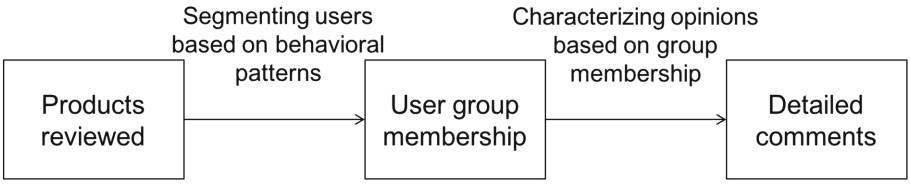

Fig. 1. The behavior-to-opinion approach

The approach involves two stages: (i) users are first divided into a set of groups based on certain behavioral patterns, and (ii) the group memberships are then used to identify opinion characteristics uniquely associated with each group. While the first stage is a usual supervised classification, the first stage is less straightforward conceptually and operationally. Ideally, there is ground-truth knowledge about the relevant group membership that is available in the data or from an external source. However, this is usually not the case. The group membership will have to be learned from the data. We believe that "near ground-truth" knowledge of group membership can be discovered from *unambiguous* and *repeated* behaviors inside product reviews. Amazon review data meet the requirements. First, each review post is dedicated to one and only one product (e.g., print books or kindle books), which ensures the unambiguity of the relevant behavior. Secondly, many reviewers publish multiple posts on the same product, which ensures the stability of the measured membership. Unambiguity and stability jointly provide the necessary face validity for the results of the first stage, which serves the basis for the quality of the second stage.

In the current study, over 7 million users posted reviews exclusively on print books, exclusively on kindle books, exclusively on print and then on kindle, or exclusively on kindle and then on print. We exclude all other users who commented only once on a product or several times in an alternate sequence, which reduces the sample size considerably but safeguards the unambiguity and minimal stability of the derived group membership.

Rogers [1] firstly proposed adopter segmentation in his book *Diffusion of innovations* in 1962 and categorized adopters into innovators, early adopters, early majority, late majority, and laggards. The criterion of this adopter categorization is based on the assumption of normal distribution of innovativeness and the five categories are divided by mean time of adoption plusorminus its one or two times standard deviations. Mahajan and others [2] then developed this adopter categorization using other established diffusion models to fit more products that may not follow normal distributions. Then Zhu and He [3] proposed a dynamic adopter categories including continuous adopters, discontinued

adopters, potential adopters, and continuous non-adopters and found distinctive characteristics for the four categories. Based on Zhu and He's categories, we revised the construct to fit our e-books and paper books adopter categories, including kindle-only (continuous e-book adopters), print-only (continuous e-book non-adopters), kindle-to-print (discontinued e-book adopters) and print-to-kindle (transitive e-book adopters). After segmenting the four categories of adopters, we established word vectors using *word2vec*, a deep learning approach for words embedding, and got 2000-dimensional vector representations of more than 30,000 words. Then, we employed K-means model to cluster words with similar meaning into the same cluster or topic. Finally, we conducted multinomial logistic regression analysis and detected features that discriminate the four categories.

2 Methods

2.1 Word Clustering Based on Deep Learning

Understanding the meaning of words or sentences is one of the core issues in natural language processing study. Traditional topic model methods, such as LSA (Latent Semantic Analysis) or LDA (Latent Dirichlet Allocation), are usually count-vector-based, and they care more about co-occurrence patterns of words but not their context. For example, for sentences like "*Tom loves Jessica*" and "*Jessica loves Tom*", traditional methods cannot figure out who loves who but treat these two sentences as the same instead. What's more, these traditional methods rely heavily on dimensionality reduction techniques, which may require more resources to handle on larger data. However, context-predicting models or neural language models stress more on contexts of words and thus can figure out semantic or syntactic relations deeper. Among these context-predicting models, *word2vec* is an efficient embedding method which can provide state-of-the-art results on a lot of natural language tasks [4]. Just as highlighted in Baroni's paper [5], "don't count, predict!", context-predicting methods generally outperform than count models. Furthermore, comparing *word2vec* with other neural-network-inspired word embedding models, e.g. *GloVe* [6], *shape word2vec* is more robust and scales nicely. In other words, although *word2vec* might not be the best approach for every task, it does not significantly underperform in a lot of scenarios [4].

The *word2vec* model and application by Mikolov and his colleagues [7] have attracted a great amount of attention since their release. *Word2vec* works in a way that is similar to deep learning approaches, but is computationally more efficient. It attempts to discover semantic relationships among words through word embeddings, a framework for vector representations of words. The vector representations of words learned by *word2vec* models have been shown to be efficient for learning high-quality vector representations of words from large amounts of unstructured text data, and proven to be useful in various NLP tasks.

Continuous bag-of-words model (CBOW) and Skip-gram model are two main techniques used in *word2vec* to build a neural network that maps words to real-number vectors, with the expectation that words with more similar meanings will be mapped to more similar vectors.

CBOW: Assuming word inputs to the model could be w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}, and the output will be w_i, where the subscripts from $i-2$ to $i+2$ indicate the index of words in order. Hence we can consider the task as "*predicting the word given its context*".

Skip-gram: While in this scenario, words input to the model is w_i, and the output could be w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}. So the task here is "*predicting the context given a word*". In addition, the window size of context is not limited to its immediate context, and training instances can be created by skipping a constant number of words corresponding to their contexts.

More generally, given word w and its contexts c, we can consider conditional probabilities $p(c|w)$ based on a set of Ω which denotes all word and context pairs derived from the text. Then the objective of the Skip-gram model is to set the parameters θ of $p(c|w;\theta)$ so as to maximize the probability:

$$\arg\max_{\theta} \prod_{(w,c)\in\Omega} p(c|w;\theta) \tag{1}$$

Following a neural-network approach and softmax function, we can obtain:

$$\arg\max_{\theta} \sum_{(w,c)\in\Omega} \log p(c|w;\theta) = \sum_{(w,c)\in\Omega} (\log e^{v_c \cdot v_w} - \log \sum_{c'} e^{v_{c'} \cdot v_w}) \tag{2}$$

where v_c and v_w are vector representations for c and w respectively. Therefore, finding the best parameters θ, which aim to maximize objective function (2), will result in good embedding of words.

However, it's computationally expensive to compute objective (2), therefore *word2vec* model also employs some other tricks, such as hierarchical softmax as well as negative sampling, to make the computation more tractable and efficient in real scenarios. Here we will not address about these methods any more due to space limitations (see Ref. [7] for more detail).

2.2 Logistic Regression

There are plenty of methods to investigate which words are crucial to determine the type of review texts. Among which logistic regression is a widely used method which measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a cumulative logistic function. The specific model employed by logistic regression, which distinguishes it from standard linear regression, is depicted as follows:

$$\ln \frac{p_i}{1-p_i} = \beta \cdot X_i = \beta_0 + \beta_1 \cdot x_{1,i} + \beta_2 \cdot x_{2,i} + \ldots + \beta_n \cdot x_{n,i} \tag{3}$$

where n represents the number of features derived from the i-th object X_i, $p_i \in [0, 1]$ means the probability of X_i, while $x_{n,i}$ indicates the n-th feature and β_n represents coefficient corresponding to $x_{n,i}$.

Here in our case, we can depict some typical properties of words as features, such as word topics or word vectors, and then employ logistic regression to find which features are crucial to characterize reviews or reviewers. Generally speaking, for a binary classification problem, features with positive coefficients and high level statistical significance contribute more to the positive category, while features with high level statistical significance but negative coefficients result in negative category.

3 Experiments and Results

3.1 Dataset

We use the Amazon product review data scrapped by McAuley [8], UCSD, with 142.8 million reviews spanning May 1996 - July 2014. Among 82.68 million Amazon reviews (with duplicate items removed), there are 19,446,034 reviews for paper books and 3,310,343 reviews for kindle e-books, with corresponding products number as 1,944,186 books and 435,370 e-books. As shown in Fig. 2, although the original Kindle was introduced in 2007, there appears e-books for sale and receiving reviews on Amazon.com starting from 1998. At that time, the number of e-book reviews per year remained less than four hundred, until 2007, began to dramatically increase.

Table 1. Descriptive statistics of four categories on number of reviewers, average number of reviews and average rating per reviewer

	Print-only	Kindle-only	Print-to-kindle	Kindle-to-print
Num. of reviewers	6,634,219	649,071	175,416	116,537
All category				
Ave. num. of reviews	5.01	2.87	6.58	5.97
Ave. rating	4.292	4.230	4.245	4.256
Prints				
Ave. num. of reviews	2.01	–	1.71	1.51
Ave. rating	4.348	–	4.357	4.405
Kindle				
Ave. num. of reviews	–	1.21	1.18	1.15
Ave. rating	–	4.243	4.167	4.148

– Indicates not applicable in that case

On average, paper book reviewers have the larger average number of reviews than e-book ones (2.60 and 2.23 respectively). One of the possible reasons is that paper book reviewers have much longer review history – 262.5 days on average, whereas e-book reviewers last 75.7 days averagely.

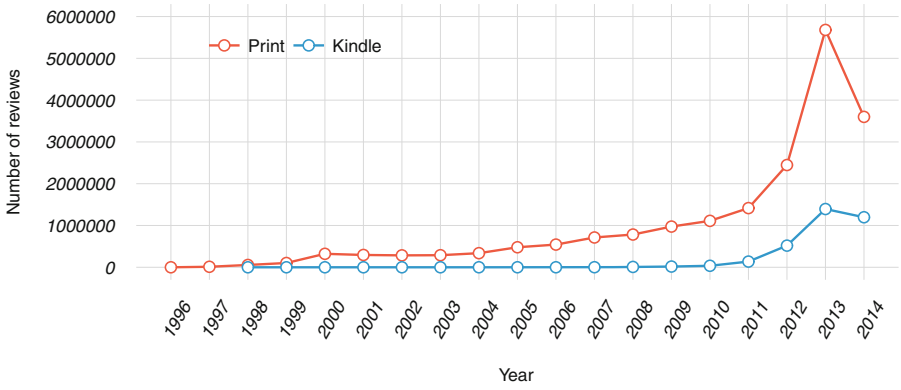

Fig. 2. Number of reviews for paper books and kindle e-books per year

3.2 Adopter Segmentation

As previously discussed, our adopter categories are revised into kindle-only (continuous e-book adopters), print-only (continuous e-book non-adopters), kindle-to-print (discontinued e-book adopters) and print-to-kindle (transitive e-book adopters). To obtain the categories, we firstly extract reviewers who have posted two or more reviews. Then for each reviewer, reviews are placed according to the publish time and each review are labeled as e-book or paper book according to the product it comments on. Reviewers who have only kindle e-book reviews are labeled as kindle-only and so do print-only reviewers. Reviewers who firstly review kindle e-books and then turn to paper books are labeled as kindle-to-print. Similarly, print-to-kindle reviewers refer to those who review paper books and then turn to kindle e-books. The other reviewers who repeatedly changing their review categories are labeled as mix reviewers. Here, we concern more of the difference between the former four kinds of reviewers. Table 1 basic descriptive statistics of four categories.

3.3 Words Clustering Based on Deep Learning Approaches

In order to dig up more discriminating characteristics between the four category adopters, we process the review text of each reviewer in the first step. We combine all the reviews of one reviewer into one paragraph and drop punctuations. Then strings are tokenized into words without stop words and words are further stemmed into their root forms.

Then we conduct *word2vec*, a word embedding model which employs deep learning approaches, to cluster words. *Word2vec* does not need labels in order to create meaningful representations, which is just to our tastes. In this scenario, we apply *word2vec* model to cluster words of reviews about books and kindle. With the aid of *word2vec*, a given word is represented by a vector with a reasonable dimension (tens to hundreds, according to the specification assigned to it).

In our case, we employ the raw text of more than 2 million reviews and train them to get vectors of 36,207 distinctive words with a dimension of 2,000 features. For example, given a word '*media*', it can be represented by a 2,000 dimensional vectors, say $v(\text{'}media\text{'}) = [v_1, v_2, \ldots, v_{2000}]$. Thus makes it very convenient to calculate similarity or distance measures for pairs of words.

After obtaining vectors for each word, we then employ K-means clustering method, which has been widely used, to allocate words into different groups, implying that similar words are more likely to be assigned to one group. In this way, words can be grouped into n (=724) different clusters. Examples of clustered words are shown in Appendix.[1]. For example, word '*Instagram*' and word '*Facebook*' are automatically group into the same cluster, which includes only the social applications.

3.4 Characterizing Four Categories of Adopters

To detect the characteristics of four category adopters, we have carried out multinominal logistic regression (MLR) using selected features from word clustering as independent variables and the membership of the four category labels as dependent variables. To construct features of each reviewer for MLR analysis, we randomly sample a set of 400,000 reviewers' review texts (100,000 for each category) and calculate each reviewer' cluster occurrence numbers according to his or her review text. Procedures to achieve this for one reviewer are described as follows: for a cluster with n words and each word's term frequency $tf_i, i = 1, 2, \ldots, n$, corresponding to this person's review corpus, the occurrence number of this cluster can be obtained through accumulation, $\sum_{i=1}^{n} tf_i$, and in the same way we can get all cluster numbers of this person. Repeating the steps for all 400,000 reviewers, and therefore all features required are obtained.

Although there are six-pair comparisons among the four adopters, MLR only conduct three pairs with one category as a baseline group. The results of the MLR analysis are shown in Table 2. Here category print-only is used as the baseline category and the MLR model compares print-to-kindle, kindle-to-print, and kindle-only adopters against print-only adopters respectively. The coefficients show the possibility of one individual belonging to one category against the baseline category (print-only adopters) and a negative coefficient for a comparison category represents the probability for one belonging to this comparison category is lower than that for the baseline category. To obtain the statistical significance of the six pairs, we conducted MLR model three times each with print-only, print-to-kindle and kindle-to-print as reference group separately. Full table is shown in Appendix.

The results of Table 2 can be interpreted as follows. Taking cluster 607 as an example, words mentioned about devices have a significant and strong impact in all six comparisons. Kindle-only adopters mention these words most, followed by kindle-to-print adopters, print-to-kindle, and print-only. Similarly, words on UI and interface are significant in all six comparisons and noted most frequently

[1] http://weblab.com.cityu.edu.hk/blog/wp-content/uploads/2016/09/Appendix.pdf.

Table 2. Multinomial logistic regression coefficients predicting four categories of adopters

Cluster id	Semantic topic of cluster	Print-to-kindle vs print-only	Kindle-to-print vs print-only	Kindle-only vs print-only
cluster607	devices	0.9369***	1.0544***	1.075***
cluster371	ui/interface	0.2724***	0.2533***	0.3428***
cluster272	light	0.0752***	0.1004***	0.0738***
cluster699	user experience	0.042***	0.037***	−0.1728***
cluster291	money	0.0289***	−0.0259***	−0.1018***
cluster167	misspelling	0.0004	−0.0306***	−0.1799***
cluster183	portable	−0.0548***	−0.0099	−0.115***
cluster564	format	−0.0809***	−0.124***	−0.2464***
cluster494	cookbook	−0.0273***	−0.0836***	−0.1529***
cluster275	law	−0.0278***	−0.0372***	−0.1917***
cluster289	school	−0.0962***	−0.1655***	−0.3383***
cluster216	clothing	−0.0846***	−0.1424***	−0.2412***
cluster197	language	−0.0524***	−0.1137***	−0.1588***
cluster48	comic books	−0.1091***	−0.1658***	−0.343***
cluster89	academic	−0.0756***	−0.1812***	−0.3197***
cluster62	programming	−0.0576***	−0.1943***	−0.2678***
cluster45	erotica	0.1559***	0.1439***	0.1496***
cluster19	social app	0.3529***	0.3385***	0.4064***
cluster104	personal health	0.01*	−0.0005	0.0441***

***, **, and * denote significance at 1, 5, and 10 %, respectively.

by kindle-only, followed by print-to-kindle, kindle-to-print and then print-only. Light is also significant, though somewhat weaker, in five of the six comparisons (except in kindle-to-print and print-to-kindle), and the coefficients shows that kindle-to-print and print-to-kindle adopters care more about light than the other two categories. Words of user experience, such as effortless, flawless, etc., are also significant in the prediction direction, with print-to-kindle adopters ranking first in attention. Topic of money significantly distinguishes the four categories, with print-to-kindle adopters as the most concerned group. Words about misspelling have a significant impact in five of the six comparisons (except in print-to-kindle vs print-only), showing that print-to-kindle and print-only adopters mention more about spelling mistakes than the other two categories. Cluster of format is significant in predicted directions of the four categories, with print-only adopters as the most concerned group. Topic of portability shows significance in five of the six comparisons (except in kindle-to-print vs print-only). Kindle-to-print and print-only adopters care more about books' portability topic than print-to-kindle and kindle-only adopters.

We also take some book content topics into considerations, such as cooking, law, school, academic research, language, programming, etc. Whereas for most of

the content clusters, print-only adopters concern more than the other three categories. Exceptions are clusters of social apps, erotica and personal health. Cluster of erotica is significant in three comparisons with print-only categories, which represents a lower possibility for print-only adopters to mention sex topic than the other three categories. The significant impacts of words about social apps are in five of the six comparisons (except kindle-only vs print-to-kindle), with large differences between print-only and the other three categories. Significance of cluster about personal health is in five comparisons (except in kindle-to-print vs print-only), showing that kindle-only and print-to-kindle adopters focus more on this topic than the other two categories.

Table 3. Concerned topics for four categories of adopters

Adopters	Function-oriented topics	Content-oriented topics
print-to-kindle	money, light, misspelling, user experience	erotica, social apps, personal health
kindle-to-print	light, portable	
kindle-to-print	devices, ui/interface	
print-only	portable, format, misspelling	cook, academic research, school, law, language, programming

We finally characterize four categories of adopters with function-oriented topics (topics related to interface, portability, etc., not the content of the books) and content-oriented topics. Table 3 matches the topics with the most concerned adopter categories. For topics of misspelling, user experience and portable, the most two related categories of each topic do not show significance in comparison, so the topics appear both in the two categories' lists.

4 Discussion

Comparing print-to-kindle and kindle-to-print adopters' concerned topics, we find that kindle-to-print adopters care more about price and spelling mistakes whereas kindle-to-print adopters pay more attention on portability. Parts of our results can be explained by the *Diffusion of Innovation Theory*. As Rogers [1] proposed that adopters evaluate an innovation on its relative advantage and relative advantage refers to perceived efficiency compared to the current ones. Here in our study, money, light, ui/interface represent the relative advantages of e-books compared to paper books and print-to-kindle adopters (transitive adopters) and kindle-only adopters (the continuous adopters) both show more concerns on these relative advantages. Whereas user experience including words such as effortless, flawless, etc., somehow, represent an individual's personal characteristics and self-efficacy. As Daugherty and others [9] demonstrated that users' adoption and usage of Web technologies, rely on their *'confidence in capability to handle the*

content online', defined as self-efficacy. So here in our study, kindle-to-print and print-to-kindle adopters (e-book adopters) concerns more about user experience than print-only (e-book non-adopters).

Back to the problem of the coexistence of e-book and paper book, in our study, we find that the overall book market is still paper-book-dominated, as when reviewers mention contents in most types of books, including cook, law, school, academic research, language, programming, etc., they are still most probably talking about the print ones. Same situation is also found in reviewer's rating preference, as shown in Table 1, people give much higher average rating on paper books rather than kindle e-books. So does it mean that after more than 20 years' rapid spread, e-book is still in a weak position in book market? The Book Reading 2016 report released by Pew Research Center in September 1, 2016 also supports our conclusion, where they found that 65 % of Americans had read a printed book in the last year, whereas only 28 % of them had read an e-book. Although the proportion of e-book is slowly increasing, the overall pattern tends to be stable. Perhaps that is what really happens in the coexistence of e-book and paper book in book market.

5 Related Work

Our work is related to the following:

Amazon book reviews. Traditionally, book reviews refer to reviews of new books, basically about the content of the books, which can help search for books best meeting personal needs [10]. With the emergence of the Internet, online book reviews are no longer limited to recently published books, but they are still recognized to have an intensive impact on consumers' purchase intension. In these circumstances, online book review data are mainly conducted and analyzed to attract consumers' interest and improve online bookstores' revenue. Chevalier and others [11] investigated the effect of online book review on sales volume at Amazon.com and Barnesandnoble.com and they found that improvements in book reviews can lead to increases in relative sales and negative comments have stronger impact than the positive ones. Also, online book reviews are also analyzed to other purposes, for example, consumers' reputations. David and others [12] scrapped review data from Amazon.com and investigated online consumers' review reputations especially on book review data. They found that hundreds of reviews on Amazon.com might be copies from one another and they finally proposed a framework to discuss the multi-tier online reputation economy. However, to the best of our knowledge, there are few studies conducting online book review data to investigate the adoption of e-book and paper book.

Studies on e-books and paper books. Previous studies investigated the relationship of digital book resources and paper books mainly from aspects of library book usage, academic research and education. Levine-Clark [13] conducted an online survey in University of Denver and found that humanists favor

paper books to digital resources at a higher rate than others and humanists care more about content, rather than e-book functions. Bierman and others [14] investigated e-book usage in pure and applied science with their online survey data and they stated an idea that e-books have a growing future in academia, whereas they found no significant difference between usage in pure and applied science. Stephens and others [15] compared the e-book and print collection usage on the Safari Books Online platform, which mainly includes programming and information technology books and found that the digital version received notably higher use than the paper books. Nicholas and others [16] conducted a survey on scholarly e-book usage and found that engineering scholars viewed digital books and resources more often than other subjects' scholars. Slater [17] managed to reason why e-books had not become the cornerstone of the academic library by reviewing previous studies before 2010. Some studies also explore the question why people prefer e-book or not. Knutson and others [18] reported students using e-textbooks for study and proponents in their interview mentioned that e-textbooks can save students' money, lighten backpacks and are convenient to update. Slater [17] mentioned in his study that although e-books can be easier to access initially, paper book can provide easier continual access than the digital versions. He also stated that a poorly designed and confusing interface may be a barrier for e-book users to get used to it.

Most of these studies investigated the difference of digital book resources and paper books from aspects of library book usage, academic research and education. Data used above were basically collected by survey, interviews or from library book collection usage statistics at that time. However, it remains some limitations when using library book collection database or surveying scholars and students to explore these questions. For example, most of the books collected in college libraries are divided by subjects, related to academic study, whereas some other types of common book, such as recipes, erotic books, etc., are rarely included. What's more, most of the digital resources scholars and students view in library collection are technical manuals, thesis and online journals, whereas the real sense of "e-books" are not in concerned and often told to buy them on online bookstores such as Amazon.com. In these circumstances, we believe that conclusions obtained from library and research aspects cannot represent the coexistence of e-books and paper books comprehensively. Further studies are still needed to explore this question under the overall dynamic book market.

Detecting user attributes based on text analysis. Previous studies show that text analysis can help detect users' attributes and profiles. Malouf [19] investigated 77,854 posts on political discussion sites to and classified posters with political orientation labels as left, right and others. They firstly identified texts of posters and then labeled them with the most frequent label in their texts. After combining other approaches as co-citation analysis, they finally get 68.48 % accuracy compared with posters' own descriptions. Pennacchiotti and others [20] inferred twitter user attributes from users' network structure and semantic contents. They employed sentiment analysis and topic models to detect Starbucks fans, user ethnicity, etc., and confirmed that text content can provide high value

in user classification. As far as we can see that previous studies generally classify reviewers into several categories based on their text analysis, whereas in our study we propose a new strategy to segment reviewers based on theory first and then characterize reviewer categories based on text analysis.

6 Conclusion and Future Work

In this paper, we conducted a data-driven analysis of the review patterns across four kinds of reviewers with the goal of characterizing reviewers and finding their potential tastes. We found that print-only adopters show significant different patterns with other three categories of e-book adopters in content-oriented topics, in the meantime, e-book adopters show similar concerns but a bit diversified attention to function-oriented topics as depicted in Table 3. When compared with other three types of reviewers, print-only adopters tend to mention words related to programming, language, law, etc., which may indicate that paper books are evolving to the forms as hand-books, tool books and textbooks and also suggest that print-only adopters have the potential demands for skill oriented types of books more than other themes. Therefore, we can recommend more books related to these topics to these reviewers afterwards.

In the case of transitive and discontinuous adopters (print-to-kindle and kindle-to-print), we found that these two categories of adopters both pay more attention to user experience than the other two adopters, which intuitively indicates that user experience may play an important role when people are changing their purchase decision, from books to kindle or from kindle to books. In addition, adopters in group of print-to-kindle care more about price and spelling mistakes whereas kindle-to-print adopters pay more attention on portability. This indicates that price may be one of the key factors that make book adopters divert attention from books to kindle. Therefore, it provides a potential strategy for book retailers to retain consumers as well as for e-book merchants to attract new buyers.

There are also many other interesting avenues to follow in our future work. For instance, do these segmentation and characterizing approaches work the same on other innovation and old technology groups' review, such as MP3 and CD, Amazon Instant video and DVD? Do people's review patterns remain consistently across different products categories? Specifically, for adopters here in kindle-to-print category, do they also pay more attention to price when purchasing other products such as electronics, clothing or movies? These questions remain to be answered by our future works. In addition, in our present work, we emphasis on part of speech of words whereas the sentiment they convey are not considered enough. For example, we know kindle-to-print adopters concern about portable topic, but do they in favor of e-books or paper books in the case of portability is still not so clear and this needs more comprehensive researches both semantically and sentimentally.

References

1. Rogers, E.M.: Diffusion of Innovations. S&S, New York (2003)
2. Mahajan, V., Muller, E., Srivastava, R.K.: Determination of adopter categories by using innovation diffusion models. J. Mark. Res. **27**, 37–50 (1990)
3. Zhu, J.J., He, Z.: Perceived characteristics, perceived needs, and perceived popularity adoption and use of the Internet in China. Commun. Res. **29**, 466–495 (2002)
4. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. In: TACL, vol. 3, pp. 211–225 (2015)
5. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: ACL (2014)
6. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of EMNLP, vol. 14, pp. 1532–4315 (2014)
7. Mikolov, T., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS (2013)
8. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of SIGIR, pp. 43–52 (2015)
9. Daugherty, T., Eastin, M., Gangadharbatla, H.: E-CRM: understanding Internet confidence and implications for customer relationship management. In: Advances in Electronic Marketing, pp. 67–82 (2005)
10. Lin, T.M., Luarn, P., Huang, Y.K.: Effect of Internet book reviews on purchase intention: a focus group study. J. Acad. Librarianship **31**(5), 461–468 (2005)
11. Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: online book reviews. J. Mark. Res. **43**(3), 345–354 (2006)
12. David, S., Pinch, T.J.: Six degrees of reputation: the use and abuse of online review and recommendation systems. Available at SSRN 857505 (2005)
13. Levine-Clark, M.: Electronic books and the humanities: a survey at the University of Denver. Collect. Build. **26**(1), 7–14 (2007)
14. Bierman, J., Ortega, L., Rupp-Serrano, K.: E-book usage in pure and applied sciences. Sci. Technol. Libr. **29**(1–2), 69–91 (2010)
15. Stephens, J., Melgoza, P., Wan, G.: Safari books online: currency, usage and book release policies of an e-book database. Collect. Build. **27**(1), 14–17 (2008)
16. Nicholas, D., Rowlands, I., Clark, D., Huntington, P., Jamali, H.R., Olle, C.: UK scholarly e-book usage: a landmark survey. ASLIB Proc. **60**(4), 311–334 (2008). Emerald Group Publishing Limited
17. Slater, R.: Why aren't e-books gaining more ground in academic libraries? E-book use and perceptions: a review of published literature and research. J. Web Librarianship **4**(4), 305–331 (2010)
18. Knutson, R., Fowler, G.A.: Book smarts? E-texts receive mixed reviews from students. Wall Street J. **20** (2009). http://www.wsj.com/articles/SB10001424052970203577304574277041750084938
19. Malouf, R., Mullen, T.: Taking sides: user classification for informal online political discourse. Int. Res. **18**(2), 177–190 (2008)
20. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to Twitter user classification. Proc. ICWSM **11**(1), 281–288 (2011)