Automated Marketing Research Using Online Customer Reviews

Author(s): THOMAS Y. LEE and ERIC T. BRADLOW

Source: *Journal of Marketing Research*, October 2011, Vol. 48, No. 5 (October 2011), pp. 881-894

Published by: Sage Publications, Inc. on behalf of American Marketing Association

Stable URL: https://www.jstor.org/stable/23033526

THOMAS Y. LEE and ERIC T. BRADLOW*

Market structure analysis is a basic pillar of marketing research. Classic challenges in marketing such as pricing, campaign management, brand positioning, and new product development are rooted in an analysis of product substitutes and complements inferred from market structure. In this article, the authors present a method to support the analysis and visualization of market structure by automatically eliciting product attributes and brand's relative positions from online customer reviews. First, the method uncovers attributes and attribute dimensions using the "voice of the consumer," as reflected in customer reviews, rather than that of manufacturers. Second, the approach runs automatically. Third, the process supports rather than supplants managerial judgment by reinforcing or augmenting attributes and dimensions found through traditional surveys and focus groups. The authors test the approach on six years of customer reviews for digital cameras during a period of rapid market evolution. They analyze and visualize results in several ways, including comparisons with expert buying guides, a laboratory survey, and correspondence analysis of automatically discovered product attributes. The authors evaluate managerial insights drawn from the analysis with respect to proprietary market research reports from the same period analyzing digital imaging products.

Keywords: market structure analysis, online customer reviews, text mining

Automated Marketing Research Using Online Customer Reviews

Marketing research, the set of methods to collect and draw inferences from market-level customer and business information, has been the lifeblood of the field of marketing practice and the focus of much academic research in the past 30-plus years. Simply put, marketing research, the methods that surround it, and the inferences derived from it have put marketing as an academic discipline and as a functional area within the firm "on the map." From a practical perspective, this has brought forward the "stalwarts and toolbox" of the marketing researcher, including methods such as preference data collection using conjoint analysis (Green and Srinivasan 1978) inferring market structure through multidimensional scaling (Elrod 1988, 1991; Elrod et al. 2002), inferring market segments through clustering routines (DeSarbo, Howard, and Jedidi 1991), or simply understanding the sentiment and "voice of the consumer" (VOC; Griffin and Hauser 1993). Although these methods are here to stay, the radical changes resulting from the Internet and user-generated media promise to fundamentally alter the data and collection methods used to perform these methods.

In this study, we propose to harness the growing body of free, unsolicited, user-generated online content for automated market research. Specifically, we describe a novel text-mining algorithm for analyzing online customer reviews to facilitate the analysis of market structure in two ways. First, the VOC, as presented in user-generated com-

*Thomas Y. Lee is Visiting Assistant Professor, Operations and Information Technology Management, Haas School of Business, University of California, Berkeley, and Adjunct Assistant Professor, Operations and Information Management, the Wharton School, University of Pennsylvania (e-mail: thomasyl@upenn.edu). Eric T. Bradlow is the K.P. Chao Professor; Professor of Marketing, Statistics, and Education; Vice-Dean and Director, Wharton Doctoral Programs; and Co-Director, Wharton Customer Analytics Initiative, the Wharton School, University of Pennsylvania (e-mail: ebradlow@wharton.upenn.edu).The authors thank Esther Chen, Ellen Ngai, and Sojeong Hong for helping with the data coding and Steven O. Kimbrough, Balaji Padmanabhan, Yoram Wind, Paul E. Green, Abba M. Krieger, attendees of the Utah Winter Information Systems Conference and Florida Decision and Information Sciences Workshop, and the anonymous *JMR* reviewers for useful suggestions and comments. The authors also gratefully acknowledge Lyra Research Inc. and Future Image Inc. for providing results from the proprietary market research that they conducted during the study period. This work was partially supported by the Fishman Davidson Center for Service and Operations Management at the Wharton School of the University of Pennsylvania, the Wharton eBusiness Initiative, and the Wharton Customer Analytics Initiative. Wayne DeSarbo served as associate editor for this article.

ISSN: 0022-2437 (print), 1547-7193 (electronic)

Journal of Marketing Research
Vol. XLVIII (October 2011), 881–894

ments, provides a simple, principled approach to generating and selecting product attributes for market structure analysis. In contrast, traditional methods rely on a predefined set of product attributes (external analysis) or *ex post* interpretation of derived dimensions from consumer surveys (internal analysis). Second, the preponderance of opinion, as represented in the continuous stream of reviews over time, provides practical input to augment traditional approaches (e.g., surveys, focus groups) for conducting brand sentiment analysis and can be done (unlike traditional methods) continuously, automatically, inexpensively, and in real time.

Our focus on market structure analysis is not by chance but rather due to its centrality in marketing practice and its fit with text mining of user-generated content. Analysis of market structure is a key step in the design and development of new products as well as the repositioning of existing products (Urban and Hauser 1993). Market structure analysis describes the substitution and complementary relationships between the brands (alternatives) that define the market (Elrod et al. 2002). In addition to descriptive modeling, market structure analysis is used for predicting marketplace responses to changes such as pricing (Kamakura and Russell 1989), marketing strategy (Erdem and Keane 1996), product design, and new product introduction (Srivastava, Alpert, and Shocker 1984). Thus, if automated in a fast, inexpensive way (as described here), it can have a significant impact on marketing research and the decisions that emanate from it.

There is a long history of research in market structure analysis. Approaches vary by the type of data analyzed (e.g., panel-level scanner data, aggregate sales, consumer survey response) and by the analytic approach (Elrod et al. 2002). Regardless of an internal or external approach, with few exceptions, market structure analysis begins with a predetermined set of attributes and their underlying dimensions, the same set of survey or transaction sales data, and assumes that "all customers perceive all products the same way and differ only in their evaluation of product attributes" (Elrod et al. 2002, p. 229). Models that incorporate customer uncertainty about product attributes (Erdem and Keane 1996) serve to highlight the colloquial wisdom "garbage in, garbage out." Surprisingly, despite its importance, there is little extant research to guide attribute selection for these methods (Wittink, Krishnamurthi, and Nutter 1982). Although literature exists on the sensitivity of conjoint results to changing attribute selection, omitting an important attribute, level spacing, and so on (Green and Srinivasan 1978), little research exists on how to choose those attributes in the first place. We propose to fill this gap using automated analysis of online customer reviews. Specifically, in this study, we visualize market structure by applying correspondence analysis (CA) to product attributes mined from the VOC.

We note, however, that our study is certainly not the first to employ user-generated content or even specifically online reviews for the purposes of marketing action. The impact of customer reviews on consumer behavior has long been a source of study. A large body of work has explored how reviews reflect or shape a seller's reputation (Dellarocas 2003; Eliashberg and Shugan 1997; Ghose and Ipeirotis 2006). Other researchers have studied the implications of customer reviews for marketing strategy (Chen and Xie 2004). To stress its importance and ubiquitous nature, a 2009 conference cosponsored by the Marketing Science Institute and the Wharton Interactive Media Initiative had more than 50-plus applicants all doing work on user-generated content.

That said, there has been comparatively little work on what marketers might learn from customer reviews for purposes of studying market structure. Early work combining marketing and text mining focused on limited sets of attributes for purposes of analyzing price premiums associated with specific characteristics (Archak, Ghose, and Ipeirotis 2007; Ghose and Ipeirotis 2008). In contrast, our objective is to learn the full range of product attributes and attribute dimensions voiced in product reviews and to reflect that in a visualization of market structure that can be used for marketing actions. Work that analyzes online text to identify market competitors (Pant and Sheng 2009) focuses on the corporate level and relies on network linkages between web pages and online news. Social networks have also been used to segment customers for planning marketing campaigns, but they rely on directly observable relationships between people (e.g., friends and family calling plans) with no reference to user-generated comments (Hill, Provost, and Volinsky 2006). In this study, we focus on the product brand level using user-generated product reviews.

Most recently, researchers have applied natural language processing (NLP) to automatically extract and visualize direct comparative relationships between product brands from online blogs (Feldman et al. 2007, 2008). Our work is complementary in at least three ways. First, we present a simpler set of text-mining techniques that is less dependent on complex language processing (CLP) and hand-coded parsing rules, requires minimal human intervention (only in the analysis phase), and is better suited to customer reviews (Hu and Liu 2004; Lee 2005). Second, our focus is on learning attributes as well as their underlying dimensions and levels. Our goal in learning attributes, dimensions, and levels is not only to facilitate direct analysis but also to inform more traditional market structure and conjoint analysis methods. Third, we specifically highlight the VOC. Different customer segments, such as residential home users of personal computers versus hard-core gamers, may refer to the same product attribute(s) using different terminology (Randall, Terwiesch, and Ulrich 2007). These subtle differences in vocabulary may prove particularly useful in identifying unique submarkets (Urban, Johnson, and Hauser 1984).

Therefore, in summary, we present a novel combination of existing text-mining and marketing methodologies to expand the traditional scope of market structure analysis. We describe an automated process for identifying and analyzing online product reviews that is easily repeatable over reviews for both physical products and services and requires minimal human/managerial intervention. The process extends traditional market structure analysis in the following ways:

- It provides a principled approach for generating and selecting attributes for market structure analysis by identifying which attributes customers are commenting on. These include attributes that are not highlighted using more traditional methods for eliciting attribute and dimensions such as those used in generating expert buying guides.

- •It captures not only which attributes customers are speaking about but also subtle differences in vocabulary (West, Brown, and Hoch 1996) that may separate brands or identify unique submarkets. These differences manifest themselves in two ways: the granularity with which customers describe an attribute and the synonyms that customers might use.
- •It facilitates market structure analysis over time by enabling the periodic (re)estimation of market structure through discrete sampling of the continuous stream of reviews.
- •It enables more detailed analysis by exploiting the natural segregation of comments by sentiment polarity (pros versus cons) in user-generated online product reviews.
- •It has high face validity in attribute and dimension selection for practical significance because marketing managers are familiar with and have easy access to online reviews.

In the remainder of this article, we describe our methodology, apply our approach to six years of online product review data for digital cameras, and evaluate the approach in several ways, including comparisons with expert buying guides, a user survey, and proprietary market research reports.

METHODOLOGY

Figure 1 summarizes our approach. We review the entire process here in a high-level way because many of these techniques are new to the marketing audience; Web Appendix A (http://www.marketingpower.com/jmroct11) provides specific details. The process begins with a set of online reviews in a product category over a specified time frame. For example, here, we consider the reviews for all digital cameras available at Epinions.com between July 5, 2004, and January 1, 2008. Figure 1, Step 1, shows three reviews: one each for cameras manufactured by Olympus, Hewlett-Packard (HP), and Fuji. In Step 2, screen scraping software automatically extracts details from each review, including the brand and explicitly labeled lists of pros and cons. Our goal is to group all phrases discussing a common attribute into one or more clusters to reveal what customers are saying about the product space, the level of detail used to describe an attribute, and the specific word choices that customers make. While some review sites do not provide user-authored pro and con summaries (e.g. Amazon.com), many others, including Epinions.com, BizRate, and CNET, do (Hu and Liu 2004). Exploiting the structure provided by pro and con lists enables us to avoid numerous complexities and limitations of automated language processing of prose-like text. This enables us to "automate" our process relative to most extant research.

Then, we separated all the pros and cons into individual phrases, as depicted in Column 1 of the table in Figure 1, Step 3. Preprocessing transforms the phrases from Column 1 into a normalized form. Column 2 depicts one step of preprocessing, namely, the elimination of uninformative stop-words such as articles (*the*) and prepositions (*of*). Next, we rendered each phrase as a word vector. The matrix of word vectors is depicted in the remaining columns of the table in Step 3. Each row is the vector for one phrase. Each vector index represents a word, and the vector value records a weighted, normalized count of the number of times the indexed word appears in the corresponding phrase.

Phrases are automatically grouped together according to their similarity. Similarity is measured as the cosine angular distance between word vectors and is identical to the Hamming distance used in computer science–based research. Step 4 depicts the K-means clustering of the phrases from Step 1. Conceptually, we can think of each cluster as representing a different product attribute. The example shows clusters for "zoom," "battery life," "picture quality," and "memory."

While any number of clustering algorithms is acceptable, we selected K-means for its simplicity and its familiarity to both the text-mining and marketing communities. More complex topic-clustering algorithms such as probabilistic latent semantic analysis and latent Dirichlet analysis also require parameter estimation, and marketing approaches such as latent structure multidimensional scaling begin with a predefined set of product attributes. As we noted previously, the principal contribution of our work is the elicitation of product attributes directly from the customer.

In product design, a product architecture defines the hierarchical decomposition of a product into components (Ulrich and Eppinger 2003). In the same way, Step 5 depicts the hierarchical decomposition of a product attribute into its constituent dimensions (i.e., attribute levels, similar to conjoint analysis). In Step 5a, we show a conceptual decomposition of the digital camera product attribute "memory." In Step 5b, we show an actual decomposition using only the phrases from Step 1. The decomposition is treated as a linear programming assignment problem (Hillier and Lieberman 2001). The objective is to assign each word in the attribute cluster to an attribute dimension. Each word phrase defines a constraint on the assignment: Any two words that co-occur in the same phrase cannot be assigned to the same attribute dimension. Thus, we know that "smart" and "media" cannot appear as a value for the attribute dimension quantity (4, 8, or 16) or for the attribute dimension of memory unit (MB). Note that not all phrases include a word for every dimension. Intuitively, this is both reasonable and an important aspect of capturing the VOC. We want to know not only what customers say but also the level of detail with which they say it. For the algorithm, phrases that do not include a word for each attribute simply represent a smaller set of co-occurrence constraints than a phrase containing more words.

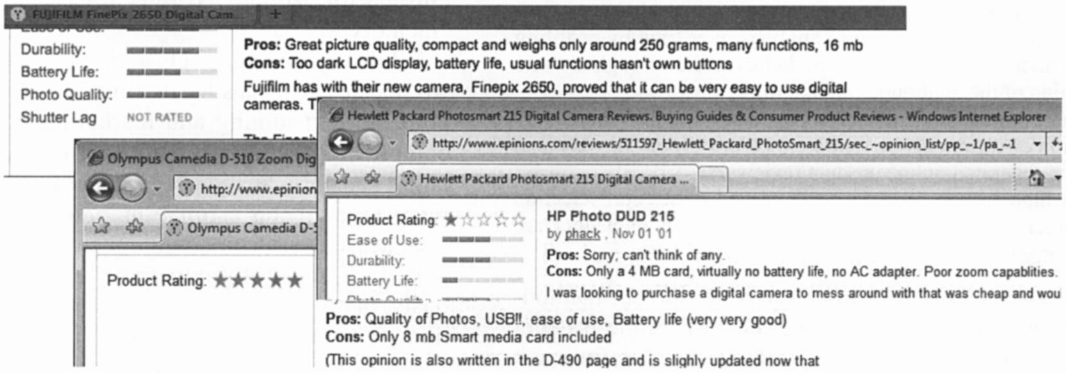
ANALYSIS AND EVALUATION

We evaluate the quality and efficacy of analyzing market structure using online product reviews in three ways: First, we ask whether any new information can be discovered from user-generated product reviews not otherwise obtained from existing methods. Second, we measure the importance of attributes discovered from within product reviews. Finally, we use CA to analyze the VOC and ask whether any meaningful managerial insight is gained by visualizing market structure using product reviews. The first two measures support the use of reviews as a complementary method for generating possible attributes for use in market structure analysis. The third measure supports reviews and their corresponding word counts as a complementary method for selecting attributes and deriving pairwise brand distances for visualizing market structure. In this section, we evaluate our approach using all three measures on an actual set of digital camera product reviews.

Figure 1

HEURISTIC DESCRIPTION OF TEXT PROCESSING OF USER-GENERATED PRO–CON LISTS

Step 1. Reviews

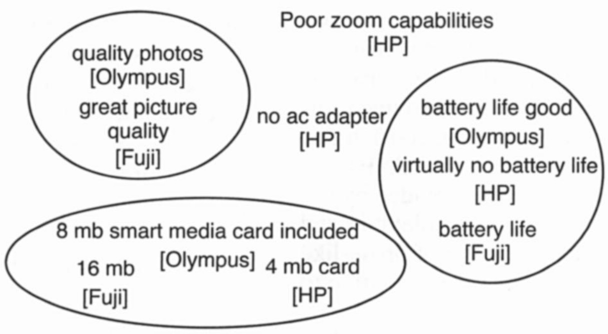


Step 2. Screen Scraping

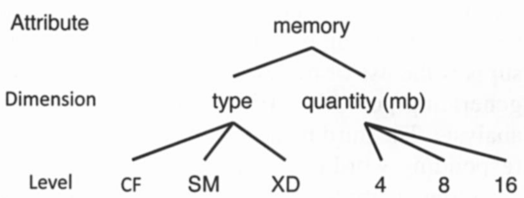
Olympus:	Quality of Photos, …, Battery life (very very good), Only 8 mb Smart media card included
HP:	…, Only a 4 mb card, virtually no battery life, no AC adapter, Poor zoom capabilities
Fuji:	Great picture quality, 16 mb, battery life, …

Step 3. Matrix of Word Vectors

Brand	Original phrase	Stop-words removed	only	life	mb	card	zoom	battery	...
Olympus	Quality of Photos	quality photos							
Olympus	Battery life (very very good)	battery life good		1				1	
Olympus	Only 8 mb Smart media card ...	8 mb smart media card inc	1		1	1			
HP	Only a 4MB card	4 mb card			1	1			
HP	Virtually no battery life	virtually battery life		1				1	
HP	No AC adapter	no ac adapter							
HP	Poor zoom capabilities	zoom capabilities					1		
Fuji	Great picture quality	great picture quality							
Fuji	16 mb	16 mb			1				
Fuji	battery life	battery life		1				1	



Step 4. Clustering Phrases



Step 5a. Attributes and Their Dimensions

D1	D2	D3	D4	D5	D6
8	mb	smart	media	card	included
4	mb			card	
16	mb				
8	mb			card	

Step 5b. Attributes and Their Dimensions in the Text

Digital Cameras

Our initial data set consists of 8226 online digital camera reviews downloaded from Epinions.com on July 5, 2004. The reviews span 575 products and product bundles that range in price from $45 to more than $1,000. Parsing the pro and con lists produces the aforementioned phrase × word matrix that is 14,081 phrases × 3364 words. We set k (the number of maximum clusters for K-means) at 50. There are several general, statistical approaches for initializing k including the Gap (Tibshirani and Walther 2005) and KL (Krzanowski and Lai 1988) statistics. Because of its computational simplicity, we plotted k versus KL in a range from 45 to 55 and found that k maximizes the value of KL at 50. Web Appendix A (http://www.marketingpower.com/jmroct11) provides processing details.

Given an initial set of 50 clusters (from K-means), our next step is to further filter the initial clusters into attribute dimensions. The CLP process produced 171 subclusters describing attributes and dimensions within the 50 initial clusters. Applying a chi-square threshold of .001 and further filtering the results using the Spearman rank test, r_s reduces those 171 subclusters to 99. Within each cluster, subclusters may represent noise from K-means or different ways in which customers express the dimensions of an attribute (e.g., the number of batteries, the type of batteries, battery life). To this point, the entire research process is fully automated, with no human intervention whatsoever. Finally, a manual reading reveals which of the remaining subclusters discuss a common product attribute and which are noise. (In the future, even this could be automated.) Our final reading identifies 39 clusters of product attributes (see Table 1). Although we might have expected 50 subclusters, one for each of the initial clusters, this is not the case. For some initial clusters, none of the subclusters pass the statistical filters. In other cases, multiple subclusters within a single cluster may indicate that the parent cluster does not cleanly distinguish a common product attribute.

To facilitate the presentation, we apply a naive convention for naming each cluster (a common practice in marketing studies): We scan the cluster for the most frequent word(s) in each cluster. Some resulting cluster names may only have meaning in the product context. Comments are inserted in parentheses to provide context to the automatically generated name as well as to indicate where certain product attributes are duplicated. Web Appendix B (http://www.marketingpower.com/jmroct11) presents a listing of automatically generated dimensions for each of the 39 attributes.

Generating Attributes from Product Reviews

The purpose of mining the text of user-generated online reviews is to complement rather than replace existing methods for analyzing market structure. As a measure of the value within the text, we ask whether reviews reveal product attributes not found using traditional measures, such as those used to create expert buying guides. As a metric, we compute precision (P) (Salton and McGill 1983) as the number of automatically generated attributes and dimensions also used by experts in published buying guides. Conversely, recall (R) counts the number of attributes and levels named in professional buying guides that are automatically discovered in the VOC. More formally, if X is the set of attributes from the VOC and Y is the set of set of attributes identified in professional guides, P and R are defined as follows: $P = |X \cap Y| / |X|$, and $R = |X \cap Y|/|Y|$.

However, exact matching is an aggressive standard in comparing hierarchies. For example, "optical zoom" is an attribute in some reference guides, but the automatic process identified "optical" as a dimension of "zoom." Borrowing from Popescu, Yates, and Etzioni (2004), we relax the definition of precision and recall so that specific terms match more general terms, provided that the more specific term appears as a dimension or level, and vice versa.

The columns of Table 2 correspond to each of ten expert guides aimed at the same consumer audience and available during the period covered by our reviews. For this study, the guides represent traditional methods for identifying meaningful product attributes. Epinions (A) comprises the attributes and levels by which customers can browse the digital camera product mix at Epinions (in 2005), and Epinions (B) represents a buying guide available on the Epinions website (in 2005). CR02–CR05 represent print buying guides for

Table 1

AUTOMATICALLY GENERATED PRODUCT ATTRIBUTES

LCD	Memory/Screen	Shutter (delay, lag)	Optical (zoom, viewfinder)	Floppy (storage media)	Support (service)	Shoot
Red eye	Lens (cap, quality, manufacturer)	Print (size, quality, output)	Slow (start-up, turn on, recovery)	Flash (memory card, photo)	Body (design, construction)	USB
Price	Picture (what, where)	Cover (lens, LCD, battery)	Feel (manufacturer, construction)	Battery (life, use, type)	Movie (audio, visual)	Size
Focus	Edit (in camera)	Disk	Instruction	Photo quality	Menu	Control
Features	Adapter (AC)	MB (memory)	Low light	Picture quality	Resolution	Zoom
Software	Image (quality)	Macro (lens)	Megapixel			

Table 2

INTERNAL CONSISTENCY: AVERAGE PRECISION AND RECALL BETWEEN ONE SOURCE AND ALL OTHERS

	Auto	*E(A)*	*DP*	*Mega*	*Biz*	*CNET*	*E(B)*	*CR02*	*CR03*	*CR04*	*CR05*	*Mean*
Precision	.37	.69	.33	.57	.41	.27	.48	.24	.26	.27	.37	.42
Recall	.72	.23	.55	.23	.48	.33	.46	.38	.37	.37	.50	.39

digital cameras from *Consumer Reports* for 2002–2005. The "Auto" column represents the attributes and dimensions derived automatically from online reviews. Row 1 indicates the average precision, and Row 2 indicates the average recall between the column source and all other reference sources.

Comparing the expert guides with one another reveals that there is no consensus among the experts on what the "important" attributes are. The absence of consistency indicates that analyzing online product reviews is not necessarily a redundant exercise; the VOC reveals interest in specific product attributes that are not identified by experts. If we observed high precision at any level of recall, that would indicate polling the VOC is redundant; that is, consumers do not reveal any information not already captured by existing methods. Low precision and high recall would indicate that, in their reviews, customers mention every possible attribute, which offers no discriminatory power. The data indicate that analyzing reviews yields high recall compared with expert attributes. This means that users report nearly all the attributes that experts do. In pairwise comparisons, the VOC has better recall than any other single expert. In addition, median precision suggests that consumers mention what the experts do but that consumers also mention some additional attributes.

The evidence suggests that reviews are neither a perfect reflection of the guides nor a mindless list of every conceivable attribute offering no value in discriminating among attributes. Product reviews do reveal information not used in the traditional methods. Therefore, the managerial question is not whether online product reviews provide information but rather what value that information provides. Do reviews include important "unseen" attributes? We address this issue next.

Importance of Attributes from Product Reviews

Through comparison with expert guides, we find that customers mention unseen attributes. To determine the significance, we conducted a laboratory survey in which participants evaluated the importance of different attributes for the purpose of purchasing a new digital camera. We find that automated analysis of online product reviews can support managerial decision making in at least two ways. First, our approach can identify significant attributes that experts otherwise overlook. Second, the reviews can serve as a filter for other attribute elicitation methods; attributes that experts and also customers identify may have more salience for purposes of product marketing and design.

Specifically, our survey, which took less than five minutes to complete, was administered as part of a 60- to 90-minute sequence of experiments, for which participants were compensated at the average rate of $10 per hour. Pretesting suggested no interference between our survey and the unrelated studies conducted during the same session. In total, 181 students at a large northeastern university participated in our web-based survey. After performing validity checks and eliminating response times, we obtained usable results for 164 participants.

We constructed a set of product attributes for testing by reconciling the 39 attributes shown in Table 1 with all the attributes identified in the ten reference buying guides listed in Table 2. After we eliminated duplicate attributes, we divided the resulting set of 55 attributes into overlapping thirds to reduce any individual respondent's burden (for the complete list, see Web Appendix C at http://www.marketingpower.com/jmroct11). A few attributes were repeated in each third as an additional validity check. Each participant viewed between 20 and 21 attributes rated their "familiarity with" and the "importance of" each attribute using a 1–7 scale. The specific questions were as follows:

> Imagine that you are about to buy a new digital camera. In the table below, we give a list of camera attributes. For each attribute, please answer two questions:
>
> First, from 1–7, please rate how familiar you are with the attribute. [1] means that you have no idea what this attribute is and a [7] means that you know exactly what this is.
>
> Second, from 1–7, please rate how much you care about this attribute when thinking about buying a new digital camera. [1] means that you do not care about this attribute at all and [7] means that this is critical. You would not think of buying without first asking about this attribute.

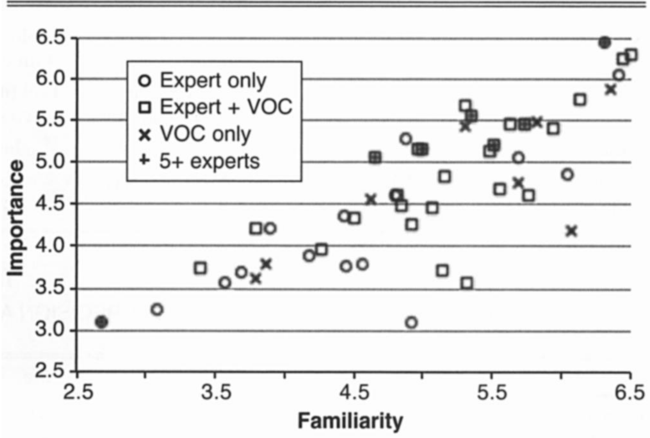
Participants were prompted to answer all questions. In particular, they were reminded to answer the second question for each attribute even if they answered [1] for the first question. In Part 2 of the survey, to understand the role that expertise might play, participants provided their self-assessed expertise on digital cameras (1 = "novice," and 7 = "expert").[1] Finally, participants provided a standard set of demographic variables such as age, gender, and education level. We used these variables as covariates to verify our main findings.

Figure 2 summarizes our main results, plotting the mean familiarity versus mean importance for each of the 55 attributes. We labeled attributes that appeared only in one or more buying guides "Expert Only," symbolized by circles. We labeled attributes that appear in at least one buying guide and also in our automated results "Expert + VOC,"

[1]Each subject also completed a six-item digital camera quiz after providing their self-assessment. The correlation was high (r = .4); thus, we use the self-assessment score in subsequent analyses.

Figure 2

USER ASSESSMENTS OF FAMILIARITY VERSUS IMPORTANCE



identified by squares. Finally, we labeled attributes that emerged only from our automated analysis of reviews "VOC Only," plotted as "X" symbols. As we would expect, the graph indicates a general trend upward and to the right. Users are more familiar with those product attributes that they tend to consider important. Similarly, if a user is unfamiliar with a particular attribute, they are unlikely to place a high value on that attribute. Web Appendix C (http://www.marketingpower.com/jmroct11) provides a complete table of attributes and their respective familiarity and importance means. In general, importance does not vary greatly by gender, expertise, or any other demographic variable. In several instances, whether the participant owned a digital camera does exhibit some significance. The significance of covariates should be checked separately in each applied domain.

Figure 2 suggests two significant managerial implications. First, there are eight attributes labeled VOC Only, and they tend toward the upper right-hand corner of the plot: (camera) size, body (design), (computer) download, feel (durability), instructions, LCD (brightness), shutter lag, and a cover (twist LCD for protection). The existence of product attributes that are both familiar and important to users suggests the value of processing online reviews to augment existing methods for identifying salient product attributes.

Moreover, comparing Expert Only with Expert + VOC in Figure 2 indicates that most high-importance, high-familiarity attributes appear as Expert + VOC, while the lower-left-hand region is populated primarily by Expert Only attributes. Comparing the Expert Only attributes with Expert + VOC attributes, we find a statistically significant difference ($p < .01$) in average familiarity (4.2–5.1) and in average importance (4.1–4.9). The difference ($p < .05$) between Expert Only and VOC Only is equally large and suggests a second managerial application of automated analysis: The VOC, as represented in online product reviews, can serve as a filter, highlighting meaningful product attributes.

The potential for applying VOC as a filter is also evident in the wide disagreement among the expert buying guides. There are only seven product attributes that appear in at least 50% (five) of the expert guides used in our evaluation. Labeled as "+" symbols in Figure 2, we observe that even when the experts agree, there is a wide variance in familiarity and importance that is mediated by the VOC.

At the same time, it is worth noting that our procedure misses some high-value attributes. In particular, two subclasses of attributes stand out. First, brand is a prominent missing attribute. Our approach relies on word frequencies. Specific references to any one brand (e.g., "Canon") may not appear with sufficient frequency to cluster as an attribute. We might instead map all explicit brand names to a single common word (e.g., "brand").

The second subclass of attributes missed by our approach is largely due to differences in how terms are classified. In Figure 2, the attribute ranked second in importance among Guide Only results is battery source. Although our automated approach captures characteristics of batteries and even battery type, these values are classified as properties of the "battery" attribute. Likewise, we distinguish between "shutter lag" and "shutter delay." Although the terms refer to the same physical characteristic, distinctions in vocabulary may reveal distinct submarkets. (We revisit this point further subsequently.) In this way, our analysis highlights differences between the VOC as represented in product reviews and that of the manufacturer and retailer's marketing literature.

Selecting Attributes and Visualizing Market Structure Using Product Reviews

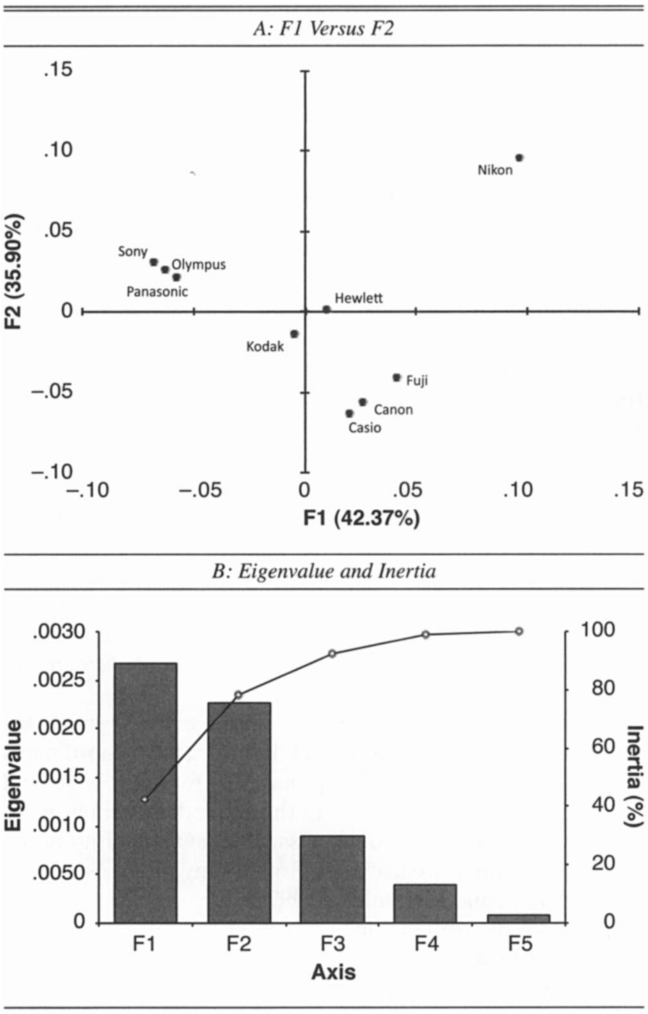
By comparing expert guides and consumer surveys, we demonstrate that customer reviews can complement existing methods for generating attributes used in market structure analysis. As a third measure of efficacy, we use attribute counts within customer reviews to select attributes, calculate brand distances, and visualize market structure. A product brand is associated with each online review. Using the automatically generated attribute clusters, we generate a brand by attribute matrix, counting the number of brand occurrences (number of phrases) for each attribute. We row (brand) normalize the matrix by the total number of phrases for that brand (Nagpaul 1999). Then, to turn this into a visual "map," we use CA, a technique for analyzing two-way, two-mode frequency data (Everitt and Dunn 2001), making it more appropriate for this task than the continuously scaled multidimensional scaling procedures commonly used for market structure maps. The CA approach is designed to help generate hypotheses by representing the data in a reduced space as determined by consulting the eigenvalues and the corresponding scree plot (Greenacre 1992).

To help interpret the dimensions in the reduced space, we use (stepwise) regression of each brand's (x, y) coordinates on the derived attributes. The probability for entry is .05 and probability for removal is .1, though other values were tested and yielded high robustness. To ensure stability of the regression results, we use a rotationally invariant, asymmetric CA (Bond and Michailidis 1997).

To make the dimensions both interpretable and actionable, we follow the "house of quality" (Hauser and Clausing 1988) in relating customer needs, as represented by user-generated reviews, to actionable manufacturer specifications. Specifically, product attributes elicited from online reviews include different granularities ("zoom" vs. "10× optical zoom") or reflect different levels of technical sophistication ("low-light settings" vs. "ISO settings"). Consulting the set of professional buying guides, we manually mapped our automatically generated attributes onto a coarser but actionable categorization of specifications. The visualizations are generated and interpreted using these meta-attributes (see Web Appendix D at http://www.marketingpower.com/jmroct11).

For attribute selection, Figure 3 depicts the brand map in two dimensions based on the initial set of digital camera reviews analyzed for this study alongside the scree plot of the eigenvalues and cumulative percentage of inertia. In this instance, the average percentage explained is 20%, indicating a clear separation between two and three dimensions. More generally, we recognize that the reliability of any such model is dependent on fit as represented by the percentage of inertia captured by those dimensions. At the limit, we imagine that marketing managers might use the two-dimensional figures as a point of departure for interpretation with respect to decision making. For example, we generated the additional F1–F3 and F2–F3 plots and found that the results are consistent with our two-dimensional interpretation.

Figure 3
MAPPING THE MARKET USING CUSTOMER REVIEWS



Regressing the CA coordinates (F1 and F2) on the meta-attributes defines F1 as a combination of comments about in-camera "navigation," "storage," and "video." We might characterize F1 as the first-order attributes that the lay consumer would notice: easy-to-use menus and navigation interfaces, the number of pictures available, and video capabilities. We note that video was only just emerging as a standard feature of digital cameras during the initial study period. In contrast, F2 is defined by attributes more relevant to technically sophisticated consumers: low-light or ISO controls and lens characteristics (e.g., name brand optics such as Zeiss). Price also emerges as an attribute of F2 but may reflect the correlation of price with higher-end products. We include a summary of the stepwise results for variable selection in Web Appendix D (http://www.marketingpower.com/jmroct11). An alternative approach to naming the dimensions is to regress on the actual, manufacturer-provided physical dimensions. However, this is exactly what makes our approach novel/different. Figure 3 is unique in that it reflects perceptions as revealed through automated analysis of customer reviews.

When interpreting CA maps, it is important to note that the dimensions only reflect relative relationships between brands and the underlying attributes. Being further from the origin does not necessarily imply higher term counts along one dimension or another. Rather, distance from the origin indicates greater deviation from the "average" brand.

We next assess the degree to which the brand map we derived from the VOC, as derived in our research, is consistent with empirical market share and documented brand strategies. Based on 2004 market share numbers from IDC, the top three manufacturers, in descending order, were Kodak, Sony, and Canon. All other brands had share percentages of 10% or less. The corresponding brand map reveals three distinct clusters, with Nikon as a notable outlier. In the mind of the customer, the secondary manufacturers all cluster by one of the three market leaders in a classic imitation strategy.

The positioning is also consistent with documented brand strategy, based on proprietary market research reports from the period. HP and Kodak explicitly attempted to differentiate themselves by emphasizing the camera-to-personal computer and camera-to-printer connection, linking them in the mind of the consumer (Worthington 2003). Among the best-known vendors, Nikon was the brand of professional photographers rather than lay consumers, late in its commitment to digital, and focused on the advanced amateur (Lyra Research 2003). Therefore, it is unsurprising that Nikon would be furthest from the origin (brand average).

At the same time, the brand map derived from the VOC reveals information that complements market share numbers and the proprietary market research reports. Traditional market research might cluster Fuji and Olympus on the basis of their shared commitment to and introduction of the xD-Picture Card memory format (Lyra Research 2003). Likewise, the "Four-Thirds" partnership between Olympus and Kodak to standardize image-sensor size and a removable lens mount might lead us to cluster those two brands (Lyra Research 2003). Instead, we observe that in the mind of the consumer, Olympus is more closely associated with market leader Sony and fellow follower Panasonic. Despite introducing a unique memory format with Olympus at the time, Fuji is paired with market leader Canon and Casio in the mind of the consumer.

Viewing market structure through product reviews enables three additional levels of analysis. First, automation enables managers to easily track market structure evolution by sampling reviews over time. Second, managers can drill into specific brand–attribute relationships by examining word choice: how consumers express themselves. Third, managers can gain additional insight into market structure by visualizing brand relationships by sentiment polarity rather than overall comments.

One advantage of analyzing market structure on the basis of user-generated product reviews is the ability to quickly repeat the process and analyze changes in market structure over time. To this end, we collected a parallel set of 5567 digital camera reviews from Epinions.com dated between January 1, 2005, and January 28, 2008. The new set of reviews revealed five new attributes replacing five previously formed clusters (see Table 3). Although it is difficult to discern the underlying causes, the changes have face validity. As the customer base becomes increasingly sophisticated and increasingly connected online, the need for instructions and support has shifted toward online self-service. Likewise, the ubiquity of personal computers and online photo

Table 3

CHANGES IN AUTOMATICALLY GENERATED ATTRIBUTES BEFORE AND AFTER 2005

Before 2005	Size	Support (service)	Feel (manufacturer)	Instruction	Edit (in camera)
After 2005	ISO	Modes	Accessories	Easy to use	White/color balance

management software may have shifted such functions away from the camera. In keeping with the theme of a more technically sophisticated audience, functions such as ISO settings, multiple-shot modes, and white/color balance are more significant or reflect active campaigns.

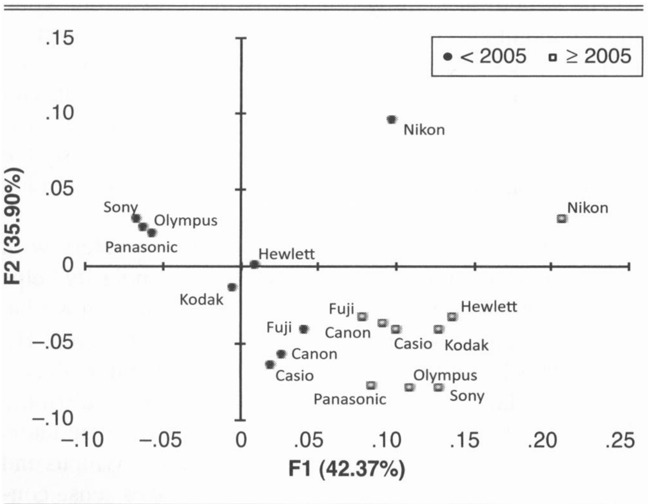
Although many attributes remain the same across time periods, changes in both customers and the marketplace may drive (or reflect) brand repositioning over time. To visualize these changes, we use the supplemental points method of CA (Nagpaul 1999). We construct a brand × attribute count matrix for Period 2; applying the transformation weights from the Period 1 decomposition, we map the Period 2 brand positions onto the Period 1 factors. Using supplemental points, we can trace market evolution between two (or more) periods on a common set of factors (axes). Figure 4 overlays Period 2 (year ≥ 2005) on the two-dimensional space defined for Period 1 (year < 2005).

Market share numbers for 2006 and 2007 published by the NPD Group and Hon Hai Worldwide show that Kodak, Sony, and Canon remained the three major manufacturers, though by 2007, Canon had emerged as the overall market leader, with Kodak falling to third. Consistent with market share, the Period 2 market structure continues to echo the overall positioning of a market leader with a cluster of imitators around each. However, in Period 2, the clusters exhibit a convergence in which even Nikon has now drawn closer to the three market leaders.

The timing of this convergence matches the competitive forecasts from proprietary market research for the digital imaging market (Lyra Research 2003) and is consistent with expectations for a rapidly evolving technology-driven marketplace. From Period 1 to Period 2, Nikon made a deliberate effort to extend its brand down-market to pursue the

Figure 4

EVOLUTION OF MARKET STRUCTURE WITH PERIOD 2 AS SUPPLEMENTARY POINTS



beginner photographer. Likewise, as the digital image market approached saturation, the remaining manufacturers were predicted to (and did) focus on expanding their product portfolios to cover all market segments (Lyra Research 2003). Successful new innovations were quickly adopted by competitors. Consider, for example, Nikon's early attempts to differentiate using swivel LCD screens, which was followed by the subsequent diffusion of flip or rotating LCDs into rival product lines.

In addition to observing market evolution, managers can drill into specific brand–attribute relationships by examining consumers' word choice. To visualize market structure, the VOC is mapped onto a set of product specifications (Hauser and Clausing 1988). Different users may refer to the same product attribute using different terminology such as "ISO settings" versus "photos in dim lighting." General terms such as "easy to use" need to be translated into actionable features such as "menus," "navigation," "auto settings," and so on. As noted previously, in this research, we followed the "house of quality" (Hauser and Clausing 1988) in manually rolling up different user terms into a common, actionable specification.

However, the manager may discover knowledge within specific consumer word choices. Focusing on F1 and F2 from our original market structure analysis, we analyze the degree of association between the three market leaders (Kodak, Canon, and Sony) and customer word choice on F1 and customer word choice on F2. Word choice in F1 is separated into two classes: the generic terms "ease of use" and "easy to use" and the specifications "navigation" "menus," and so on. Likewise, word choice in F2 is separated into two classes: lay terminology "low-light" photography, "dim-light," and so on, and advanced terminology "ISO settings," "ISO adjustment," and so on. Because we manually map different word choices onto a common factor, it is entirely possible that we are missing additional terminology. For example, consumers citing "ease of use" could also be implicitly referring to ergonomic factors such as "grip" or "lightweight." Our analysis here is intended to convey the potential of drilling down on word choice rather than provide an authoritative accounting for a specific product attribute.

A chi-square of 9.087 (p = .011) for F1and a chi-square of 21.409 (p < .0001) for F2 suggest a clear relationship between brand and word choice. We constructed the corresponding 2 × 2 contingency tables and report both the chi-square and the z-statistic from the log (odds ratio) in Tables 4 and 5.

We observe that of the big three players, Kodak is clearly the brand most associated with lay terminology such as "low-light photography" (F2) rather than "ISO" and generic words such as "ease of use" (F1). In contrast, Canon users clearly tend toward more sophisticated terms such as "ISO settings." These findings reinforce market research from the time period, which reports that gender, age, and education

Table 4
COMPARING CONSUMER WORD CHOICE (EASE OF USE V. NAVIGATION, MENUS) BY BRAND

	Kodak	*Canon*	*Sony*
Kodak			
Canon	5.568** (–2.335)		
Sony	9.117* (2.966)	1.266 (1.123)	

*Chi-square (z-statistic): $p < .005$.
**Chi-square (z-statistic): $p < .05$.

Table 5
COMPARING CONSUMER WORD CHOICE (LOW-LIGHT VS. ISO) BY BRAND

	Kodak	*Canon*	*Sony*
Kodak			
Canon	14.535* (–3.673)		
Sony	1.218 (1.213)	12.695* (–3.532)	

*Chi-square (z-statistic): $p < .0001$.

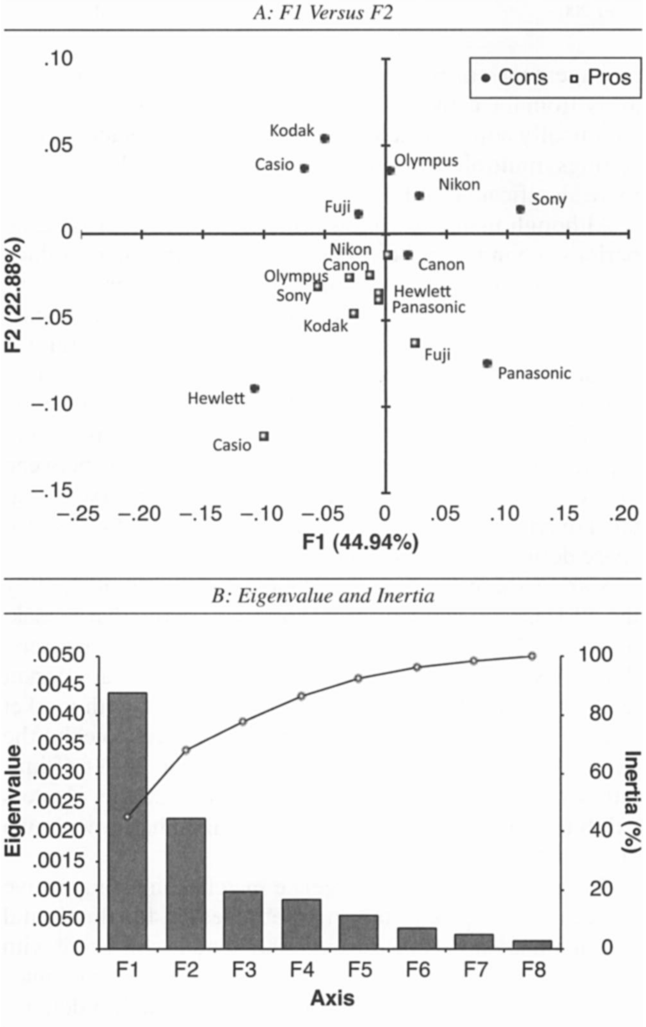
level all exhibit strong brand preferences for Kodak and Canon (Lyra Research 2005). For example, older consumers and those most likely to have stopped schooling after high school or a few years of college prefer Kodak. Canon users are typically younger and have at least a college degree.

Processing consumer reviews also enables a third level of market structure analysis. Because we explicitly use product reviews that separate user comments in structured pro–con lists, we can immediately identify the sentiment polarity. (Pro phrases convey positive polarity, and con phrases convey negative polarity.) Using the meta-attributes derived from clustering review phrases and rolled up to actionable specifications, we can label each meta-attribute on the basis of whether it includes only pro phrases, only con phrases, or both. In Figure 5, we construct a polarity space by mapping brands using con phrases and use supplemental points analysis to map brands using pro phrases onto the same dimensions. The scree plot shows a clear separation between F2 and F3, where average percentage explained is 12.5%. We constructed a similar space using pro phrases but only provide the con analysis here for space reasons.

The visualization characterizes the "must-have" product attributes as defined by Kano's needs hierarchy (Ulrich and Eppinger 2003). Must-haves define basic product requirements. Customers criticize products for their failure on must-haves; they do not praise products for must-haves because these attributes are expected. Consistent with this definition, the figure exhibits the expected brand differentiation based on cons. Positive comments mapped onto the same (F1, F2) dimensions reveal a single cluster of all brands, which share a limited number of pro remarks. Although attributes are not strictly "needs" in Kano's terms, we can think in terms of those attributes that support specific needs.

Using stepwise regression, F1 defines the minimum that *any* digital camera is expected to provide: picture quality, start-up time, lens cover, and shutter lag. These attributes certainly match our intuitive sense of must-haves for digital cameras. In contrast, F2 defines those must-have attributes that distinguish specific product segments within the market

Figure 5
MARKET STRUCTURE BY CONS WITH PROS AS SUPPLEMENTARY POINTS



space. Specifically, F2 (size, interchangeability of lenses, and the degree of in-camera programmability [e.g., menus, options]) separates the compact, "prosumer," and digital SLR product categories.

The structural map by polarity is most revealing when interpreted in the context of our two-period market structure map (Figure 4). Situated at the origin on both pro and con, Canon essentially defines the space. This is consistent with the Canon's position as a leader whose market share grew from Period 1 to Period 2 and is further evidenced by the Period 2 convergence of all groups toward the Canon brand cluster.

Market structure suggests three distinct clusters with leaders and imitators (Figure 4); focusing on polarity helps explains differences in market share. We know that Kodak and HP share a similar market strategy. However, HP clearly lags behind in sales; performance in the must-haves and the relative importance of the particular attribute dimensions help explain this difference. Likewise, imitators Fuji and Casio are closer to leader Canon than Olympus and Panasonic are to leader Sony. This again makes sense considering the overall market convergence toward the Canon, Fuji, and Casio cluster in Period 2.

Nikon, like Canon, is effectively neutral with respect to pros and cons. This is consistent with Nikon's strategy during the study period to produce excellent technology but to focus on a different customer base than the rest. Casio, whose market share numbers were too low to break out by IDC, NPD, or Hon Hai during the study period, is an outlier in the polarity maps.

DISCUSSION AND CONCLUSIONS

In this article, we present a system for automatically processing the text from online customer reviews. Phrases are parsed from the original text. The phrases are then normalized and clustered. A novel logical assignment approach exploits the structure of pro–con review summaries to further separate phrases into subclusters and then assigns individual words to unique categories. Notably, our work differs from sentiment-based strategies in that we do not rely on complex NLP and do not rely on the user to first provide a set of representative examples or keywords (Hu and Liu 2004; Turney 2002). We conclude by discussing the managerial implications of analyzing market structure based on the VOC, limitations of our current approach, and opportunities for further work.

Managerial Implications

The automated analysis of online product reviews promises to support rather than supplant managerial decision making both descriptively and prescriptively. First, analysis of online reviews can confirm the manager's marketing strategy at several levels. Describing market structure through the VOC reveals a brand's position compared with its peers, highlights salient product attributes, and reveals underlying segments. First, relative brand position can help confirm the success of a campaign's objectives. For example, consumer perceptions reflect the similarities in market strategy between HP and Kodak during the study period. Conversely, a brand's perception of itself may not match the consumer's, revealing a failure in strategy. For example, there is little evidence that Fuji or Olympus explicitly or deliberately engaged in a strategy of imitation.

Analysis of reviews also confirms which attributes are salient for describing consumer perception. For brands such as Nikon, which were known to focus on advanced consumers but are seeking to move down market (Lyra Research 2003), the market structure map reveals those product attributes that differentiate position in the minds of the consumer. Although HP and Kodak pursued similar printing and personal computer connection strategies, consumers equated those brands (and distinguished among others) using attributes such as menus, storage, and video.

Finally, the explicit word choices in reviews can confirm the success of targeted marketing and segmentation strategies. Proprietary market research has discovered distinct brand preferences along gender, age, and educational levels (Lyra Research 2005). Demographic variables are known to correlate with language, and we demonstrate that these same segments and brand preferences are revealed in the VOC (e.g., technical wording vs. layperson's speech). In this way, vocabulary may help identify unique submarkets (Urban, Johnson, and Hauser 1984).

In addition to describing the current state of the market to confirm existing strategies, managers can analyze product reviews in a prescriptive manner. Currently, managers use conjoint analysis, a form of external market structure analysis (Elrod et al. 2002), for new product introduction (Michalek, Feinberg, and Palambros 2005; Wittink and Cattin 1989), optimal product repositioning (Moore, Louviere, and Verma 1999) and pricing (Goldberg, Green, and Wind 1984), and segmenting customers (Green and Krieger 1991). However, conjoint is dependent on the representation of a customer's utility as an agglomeration of preferences for the underlying attribute levels that have been selected. This dependence holds regardless of either format (choice-based, ratings-based, ranking-based, constant sum, or self-explicated) or method for determining the profiles (Evgeniou, Boussios, and Zacharia 2005; Huber and Zwerina 1996; Moore, Gray-Lee, and Louviere1998; Toubia, Simester, and Hauser 2003).

Automated analysis of online product reviews can prescriptively assist in the design of conjoint studies in at least three ways. For attribute generation, customer reviews discover meaningful attributes not found by traditional means. For attribute selection, polarity analysis helps separate Kano's must-have requirements from linear satisfiers and delighters. Finally, analysis of reviews can help produce meaningful levels for conjoint study design.

Limitations

To generate initial clusters, the system requires phrases. Although Epinions' customer reviews provide lists of phrases, variations in human input are a source of noise. Moreover, we assume that a phrase represents a single concept and that individual words represent distinct levels of attribute dimensions. It is clear how this assumption creates difficulties with the natural language in reviews (e.g., people who write lists such as "digital and optical zoom").

One possible solution is to apply more sophisticated NLP techniques. Beginning with a representative set of meaningful words, there are ways of expanding the set of words, distinguishing between distinct attributes, and identifying relationships between attributes (Hearst 1992; Popescu, Yates, and Etzioni 2004). However, our pro–con summaries are simply lists of phrases with no associated linguistic context; Liu, Hu, and Cheng (2005) demonstrate that techniques that rely on representative examples perform markedly less well in the context of pro–con phrases. Our constrained optimization approach dispenses with the need for representative examples and knowledge of grammatical rules (Lee 2005). An entirely different approach that would preserve the unsupervised nature of our work is to attempt to identify phrases through frequent item-set analysis. Hu and Liu (2004) demonstrate that by tuning support and confidence thresholds, it is possible to discover whether a word pair represents one concept or two (e.g., "compact flash" versus "3× zoom").

We need the ability to assess the stability of our clusters and concomitant product features. One instance of stability is sensitivity to data sample size. Here, we relied on a large data set to yield the phrases from which we cluster and optimize the assignment. The large data set is also a boon because we can liberally discard phrases to minimize the effects of naive parsing. To measure the sensitivity to sample size, we would cross-validate on smaller sets of review samples. We can plot the trade-off between sample size and evaluation metrics to identify diminishing returns and attempt to estimate a minimal number of required reviews. Care needs to be taken to ensure sufficient heterogeneity in the sample selection with respect to different product features and the corresponding feature attributes.

Although our approach is not limited to evolving product domains, our dependence on sources that provide phraselike strings is a limitation. At least two factors ameliorate this limitation. First, there are other domains in which phraselike text strings apply rather than prose. Progress notes in medical records and online movie reviews (Eliashberg and Shugan 1997) are two such examples. Second, recognizing the current limitations of NLP tools, more online sources are soliciting customer feedback in the form of phrases rather than prose to facilitate automated processing (Google 2007).

Finally, the distinction between attributes, dimensions, and levels is imprecise. Conceptually, each attribute is parameterized by one or more dimensions, and each dimension is defined by the levels (the set of values) that a particular dimension may take. Operationally, attributes, dimensions, and levels are defined by the clustering. Attributes name the clusters from K-means. Each subcluster formed from the assignment algorithm represents one attribute dimension. The values within each subcluster define the dimension's levels. Knowing that the distinction among attributes, dimensions, and levels is noisy emphasizes the role of text processing as a decision aid rather than a substitute for traditional processes.

Further Work

Beyond the existing method, there are several ways we might enrich understanding of the VOC for market structure analysis. First, we have only scratched the surface of linguistically analyzing word choice. Recognizing that word choice can reflect distinct segments, we might attempt to simultaneously induce specific market segments and their corresponding language using strategies such as coclustering (Dhillon, Mallela, and Kumar 2002). A complementary strategy could seek to learn relationships between different word clusters. For example, different customers may address a concept using parallel categories: How does 32 MB of storage compare to a customer who has 130 images? We can apply concept clustering (Ganti, Gehrke, and Ramakrishnan 1999; Gibson, Kleinberg, and Raghavan 1998) in conjunction with our CLP approach to group words from parallel categories.

Structure exists not only at the market level but also at the individual consumer level (Elrod et al. 2002). A deeper understanding of the distinctions between segments requires understanding not only the differences between product attributes but also the differences in the underlying customer needs (Allenby et al. 2002; Srivastava, Alpert, and Shocker 1984; Yang, Allenby, and Fennell 2002). Unlike traditional data sources for market structure analysis, online product reviews include comments about not only what and how but also why. In the text of reviews, customers often relate why they purchased a product and what they use that product for. Research mining user needs from online reviews (Lee 2009) could be incorporated into market structure analysis.

We provide an approach for descriptive modeling of market structure based on the VOC, as captured in online product reviews. In this article, we applied the method to reviews of digital cameras representing a period of rapid evolution and compared our analysis with proprietary market research. Analyzing additional domains, such as mature markets and service goods, are beyond the scope of this research. However, our preliminary analyses of toaster ovens and hotel reviews suggest equal promise. Moreover, traditional approaches to market structure analysis produce models useful not only for describing existing markets but also for predictive purposes. Exploring the integration of the VOC and productive reviews into a predictive market structure analysis across multiple product domains is a great opportunity (Allenby et al. 2002; Elrod et al. 2002). We believe that our research offers an important first step.

REFERENCES

Allenby, Greg, Geraldine Fennell, Albert Bemmaor, Vijay Bhargava, Francois Christen, Jackie Dawley, et al. (2002), "Market Segmentation Research: Beyond Within and Across Group Differences," *Marketing Letters*, 13 (3), 233–44.

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis (2007), "Show Me the Money! Deriving the Pricing Power of Product Features by Mining Consumer Reviews," in *Proceedings of the ACM KDD 2007, San Jose, CA*. New York: Association for Computing Machinery.

Bond, Jason and George Michailidis (1997), "Interactive Correspondence Analysis in a Dynamic Object-Oriented Environment," *Journal of Statistical Software*, 2 (8), 1–31.

Chen, Yubo and Jinhong Xie (2004), "Online Consumer Review: A New Element of Marketing Communications Mix," (accessed June 27, 2011), [available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=618782]..

Dellarocas, Chrysanthos (2003), "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science* 49 (10), 1407–1424.

DeSarbo, Wayne S., Daniel J. Howard, and Kamel Jedidi (1991), "Multiclus: A New Method for Simultaneously Performing Multidimensional Scaling and Cluster Analysis," *Psychometrika*, 56 (1), 121–36.

Dhillon, Inderjit S., Subramanyam Mallela, and Rahul Kumar (2002), "Enhanced Word Clustering for Hierarchical Text Classification," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery, 191–200.

Eliashberg, Jehoshua and Steven M. Shugan (1997), "Film Critics: Influencers or Predictors?" *Journal of Marketing*, 61 (April), 68–78.

Elrod, Terry (1988), "Choice Map: Inferring a Product-Market Map from Panel Data," *Marketing Science*, 7 (1), 21–40.

——— (1991), "Internal Analysis of Market Structure: Recent Developments and Future Prospects," *Marketing Letters*, 2 (3), 253–66.

———, Gary J. Russell, Allan D. Shocker, Rick L. Andrews, Lynd Bacon, Barry L. Bayus, et al. (2002), "Inferring Market Structure from Customer Response to Competing and Complementary Products," *Marketing Letters*, 13 (3), 221–32.

Erdem, Tülin and Michael P. Keane (1996), "Decision-Making Under Uncertainty: Capturing Dynamic Choice Processes in Turbulent Consumer Good Markets," *Marketing Science* 15 (1), 1–20.

Everitt, Brian S. and Graham Dunn (2001), *Applied Multivariate Data Analysis*. New York: Oxford University Press.

Evgeniou, Theodoros, Constantinos Boussios, and Giorgos Zacharia (2005), "Generalized Robust Conjoint Estimation," *Marketing Science*, 24 (3), 415–29.

Feldman, Ronen, Moshe Fresko, Jacob Goldenberg, Oded Netzer, and Lyle H. Ungar (2007), "Extracting Product Comparisons from Discussion Boards," *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. New York: Association for Computing Machinery, 469–74.

——, ——, ——, ——, and —— (2008), "Using Text Mining to Analyze User Forums," in *Proceedings of the 2008 International Conference on Service Systems and Service Management*. New York: IEEE.

Ganti, Venkatesh, Johannes Gehrke, and Raghu Ramakrishnan (1999), *CACTUS—Clustering Categorical Data Using Summaries*. San Diego: ACM SIGKDD.

Ghose, Anindya and Panagiotis G. Ipeirotis (2006), "The Dimensions of Reputation in Electronic Markets," NYU Center for Digital Economy Research Working Paper No. CeDER-06-02, (accessed June 27, 2011), [available at SSRN: http://ssrn.com/abstract=885568].

—— and —— (2008), "Estimating the Socio-Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," NYU Stern Research Working Paper CeDER-08-06, Stern School of Business, New York University.

Gibson, D., J. Kleinberg, and P. Raghavan (1998), "Clustering Categorical Data: An Approach Based on Dynamical Systems," paper presented at 24th International Conference on Very Large Databases.

Goldberg, Stephen M., Paul E. Green, and Yoram Wind (1984), "Conjoint Analysis of Price Premiums for Hotel Amenities," *Journal of Business*, 57 (1), 111–32.

Google (2007), "About Google Base," (accessed July 6, 2011), [available at http://web.archive.org/web/20071120073851/http://base.google.com/support/bin/answer.py?answer=59260&topic=2904].

Green, Paul E. and Abba M. Krieger (1991), "Segmenting Markets with Conjoint Analysis," *Journal of Marketing*, 55 (October), 20–31.

—— and V. Srinivasan (1978), "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research* 5 (September), 103–123.

Greenacre, Michael J. (1992), "Correspondence Analysis in Medical Research," *Statistical Methods in Medical Research*, 1 (1), 97–117.

Griffin, Abbie and John R. Hauser (1993), "The Voice of the Customer," *Marketing Science* 12 (1), 1–27.

Hauser, John R. and Don Clausing (1988), "The House of Quality," *Harvard Business Review*, 3 (May/June), 63–73.

Hearst, Marti (1992), "Automatic Acquisition of Hyponyms from Large Text Corpora," in *COLING '92 Proceedings of the 14th Conference on Computational Linguistics*, Vol. 2. Stroudsburg, PA: Association for Computational Linguistics, 539–45.

Hill, Shawndra, Foster Provost, and Chris Volinsky (2006), "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks," *Statistical Science*, 21, 21.

Hillier, Frederick S. and Gerald J. Lieberman (2001), *Introduction to Operations Research*. New York: McGraw-Hill.

Hu, Minqing and Bing Liu (2004), "Mining and Summarizing Customer Reviews," in *KDD04*. Seattle: Association for Computing Machinery.

Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33 (August), 307–317.

Kamakura, Wagner A. and Gary J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research*, 26 (November), 87–96.

Krzanowski, Wojtek J. and Y.T. Lai (1988), "A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering," *Biometrics*, 44 (1), 23–34.

Lee, Thomas Y. (2005), "Ontology Induction for Mining Experiential Knowledge from Customer Reviews," presentation at Utah Winter Information Systems Conference, Salt Lake City, UT, (March 11–12).

—— (2009), "Automatically Learning User Needs from Online Reviews for New Product Design," in *AMCIS 2009 Proceedings*, (accessed June 29, 2011), [available at http://aisel.aisnet.org/amcis2009/22].

Liu, Bing, Minqing Hu, and Junsheng Cheng (2005), "Opinion Observer: Analyzing and Comparing Opinions on the Web," in *WWW '05: Proceedings of the 14th International Conference on World Wide Web*. New York: Association for Computing Machinery.

Lyra Research Inc. (2003), "The Digital Camera Market: Worldwide Trends and Forecast through 2006," research report.

—— (2005), "Inside the Mind of the Digital Photographer: A Complete Photo Behavior Profile by Gender, Age, Education and Income," research report.

Michalek, Jeremy J., Fred M. Feinberg, and Panos Y. Papalambros (2005), "Linking Marketing and Engineering Product Design Decisions via Analytical Target Cascading," *Journal of Product Innovation Management*, 22 (1), 42–62.

Moore, William L., Jason Gray-Lee, and Jordan J. Louviere (1998), "A Cross-Validity Comparison of Conjoint Analysis and Choice Models at Different Levels of Aggregation," *Marketing Letters*, 9 (2), 195–207.

——, Jordan J. Louviere, and Rohit Verma (1999), "Using Conjoint Analysis to Help Design Product Platforms," *Journal of Product Innovation Management*, 16 (1), 27–39.

Nagpaul, P.S. (1999), "Guide to Advanced Data Analysis Using IDAMS Software," UNESCO, (accessed June 29, 2011), [available at http://www.unesco.org/webworld/idams/advguide/TOC.htm].

Pant, Gautam and Olivia Sheng (2009), "Avoiding the Blind Spots: Competitor Identification Using Web Text and Linkage Structure," in *ICIS 2009 Proceedings*, (accessed June 29, 2011), [available at http://aisel.aisnet.org/icis2009/57].

Popescu, Ana-Maria, Alexander Yates, and Oren Etzioni (2004), "Class Extraction from the World Wide Web," in *Adaptive Text Extraction and Mining: Papers from the 2004 AAAI Workshop*, Ion Muslea, ed. Menlo Park, CA: American Association for Artificial Intelligence.

Randall, Taylor, Christian Terwiesch, and Karl T. Ulrich (2007), "User Design of Customized Products." *Marketing Science*, 26 (2), 268–80.

Salton, Gerald and Michael J. McGill (1983), *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Srivastava, Rajendra K., Mark I. Alpert, and Allan D. Shocker (1984), "A Customer-Oriented Approach for Determining Market Structures," *Journal of Marketing* 48 (Spring), 32–45.

Tibshirani, Robert and Guenther Walther (2005), "Cluster Validation by Prediction Strength," *Journal of Computational & Graphical Statistics*, 14 (3), 511–28.

Toubia, Olivier, Duncan I. Simester, and John R. Hauser (2003), "Fast Polyhederal Adaptive Conjoint Estimation," *Marketing Science*, 22 (3), 273–303.

Turney, Peter D. (2002), "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 417–24.

Ulrich, Karl T. and Steven D. Eppinger (2003), *Product Design and Development*. New York: McGraw-Hill.

Urban, Glen L. and John R. Hauser (1993), *Design and Marketing of New Products*. Upper Saddle River, NJ: Prentice Hall.

——, Philip L. Johnson, and John R. Hauser (1984), "Testing Competitive Market Structure," *Marketing Science*, 3 (2), 83–112.

West, Patricia M., Christina L. Brown, and Stephen J. Hoch (1996), "Consumption Vocabulary and Preference Formation," *Journal of Consumer Research*, 23 (2), 120–35.

Wittink, Dick R. and Philippe Cattin (1989), "Commercial Use of Conjoint Analysis: An Update." *Journal of Marketing*, 53 (July), 91–96.

——, Lakshman Krishnamurthi, and Julia B. Nutter (1982), "Comparing Derived Importance Weights Across Attributes," *Journal of Consumer Research*, 10 (1), 471–74.

Worthington, Paul (2003), "Digital Cameras: Strategies to Reach the Late Majority," Future Image Inc.

Yang, Sha, Greg M. Allenby, and Geraldine Fennell (2002), "Modeling Variation in Brand Preference: The Roles of Objective Environment and Motivating Conditions," *Marketing Science*, 21 (1), 14–31.