Listening to the online “Voice of the Consumer”

Every year, American companies spend more than $10 billion dollars seeking to learn what consumers think about their products or how they rate against their competitors. The techniques they typically use have guided the field of market research for the last three decades – methods such as focus groups or detailed customer surveys. These practices still dominate even as thousands of the most engaged and avid customers – larger than a focus group by a factor of thousands – have been taking to the Internet daily in growing numbers for nearly 15 years to hold, free-ranging and unsolicited discussions of what they like about consumer brands, what they dislike, or to compare similar products.

Car aficionados, for example, flock to popular consumer discussion boards such as the automotive website Edmunds.com, where a discussion on power-steering issues with the Toyota Corolla draws more than 400 posts and the Nissan Altima board features conversations on everything from leasing issues to engine fires. Today, researchers are developing programs that can sift through this voluminous data and identify trends; their findings show that sometimes online conversations can not only validate the key qualities of a consumer brands – marking a Toyota Camry as “inexpensive,” for example – but discover attributes that marketing executives did not know about, such as buyers singling out a Nissan Sentra as a “college car.” At the Wharton Customer Analytics Initiative, or WCAI, which supported this new research, these promising findings are a giant step toward an ambitious goal of helping American businesses reduce the amount now spent on market research by as much as one-half, through mining the vast pool of free consumer opinion online.

These vast Internet conversations are increasingly becoming what Eric T. Bradlow, a WCAI co-director, calls simply: “The Voice of the Consumer.” Most experts agree that these large online consumer discussions boards, as well as a growing array of product rankings, ratings and review websites, carry the potential to revolutionize the world of market research, offering businesses a massive and free data base of what consumers think about their products.

Academic experts in the field of marketing research experts have known about the vast potential of mining Internet discussions for roughly a decade, but the practical obstacles in actually extracting this data and making sense out of it have loomed

---

1 At the time of this research project, the Wharton Customer Analytics Initiative was known as the Wharton Interactive Media Initiative.
equally large. A team of top researchers led by Oded Netzer of the Columbia Business School summarized the dilemma quite succinctly in a recent paper – “Mine Your Own Business: Market Structure Surveillance Through Text Mining” – published by the Marketing Science Institute. They wrote: “Consumer-generated content on the Web is both a blessing and a curse.” The sheer size of the data pool can be overwhelming, and the freewheeling nature of the consumer chat raises problems with spelling and grammar, in addition to interpretation.

Despite these obstacles, successful web scraping of online consumer discussions has become a kind of Holy Grail for marketing researchers. The Wharton Customer Analytics Initiative, or WCAI, which is pioneering the use of individual-level customer data to drive better business decisions, has taken on a leading role in this effort. In 2009, more than 100 researchers working on issues related to user-generated content attended a WCAI conference on the Wharton campus to share and advance knowledge in this key area. This WCAI Insights paper features two recent reports, including one co-authored by WCAI co-director Bradlow, that showcase new web scraping techniques that both validate the reliability of information from online discussion board and its potential as a marketing research tool.

Large online consumer discussions boards carry the potential to revolutionize the world of market research, offering businesses a massive and free data base of what consumers think about their products.

**Do marketers ask the right questions?** U.S. companies alone currently spend more than $10 billion annually on traditional marketing research that aims primarily to learn what attributes of a product are most important to potential customers or how their product is perceived in the marketplace. The rapid growth of this field began in the 1980s and was based on academic advances in the study of conjoint analysis, which today remains the popular method for marketing experts – a practice developed at Wharton by marketing professor Paul Green – to determine which product attributes matter most to customers. Over the last 30 years, the main techniques for gathering this information – in-depth customer surveys as well as more open-ended focus groups that pay pre-selected consumers to take part in a monitored discussion – have taken deep root in the field of marketing. This feedback has become critical for consumer-oriented businesses in developing new products or features to meet consumer demand or in crafting marketing campaigns that emphasize the attributes of a product believed most likely to motivate would-be buyers.

But the rapid growth of the Internet as a powerful tool for consumers to share product opinions has also made experts in marketing research more aware of several flaws that are built into the leading conjoint analysis techniques. For one thing, the process of identifying the specific product attributes that will be highlighted in a consumer survey, typically guided by company marketing managers, might altogether ignore completely different issues that are being raised by the rank-and-file consumers.

---

2 At the time of the User-Generated Content conference, the Wharton Customer Analytics Initiative was known as the Wharton Interactive Media Initiative.
In addition, experts note, the selected focus groups might not always reach the most passionate and engaged consumers who are voluntarily discussing products and brands on the Internet.

Meanwhile, both the volume of online consumer discussions and the availability and variety of such forums has grown rapidly since Internet use began becoming prevalent in the mid-to-late 1990s, and in recent years has spiked exponentially. A 2010 paper by University of Maryland marketing professor Michael Trusov found that 24 percent of online shoppers pay attention to consumer review sites. At the same time, customers are increasingly less likely to take part in the traditional tools of conjoint analysis, with the percentage who decline to take part in traditional surveys rising every year. Online, the new windfall of millions of product reviewers takes advantage of a variety of different kinds of forums; including some geared specifically toward product ratings but also open-ended product discussion on popular sites such as Epinions.com, which has hosted customer-generated product reviews since 1999. This significant shift away from survey-and-focus-group participation and into this huge and growing pool of free, accessible and open-ended discussion is creating what WCAI’s Bradlow calls “the automatic voice of the consumer – what people are talking about naturally.”

**Finding the “unforeseen attributes.”**
For several years, Bradlow and co-researcher Thomas Y. Lee – an assistant professor of operations and information management at the Wharton School – have focused on developing an automated system that would allow marketing executives to automatically sift through these thousands of online consumer conversations and systematically identify and then quantify the most important product qualities or map out the kind of brand comparisons that customers are making.

“Can we infer the market structure – the brands and attributes that people are talking about?” Bradlow asked. “Can we determine which brands are simultaneously compared to each other?” One of their key goals has been developing a method that will identify what Bradlow identifies as “unseen attributes of a product” – that is, issues that buyers are discussing which executives back at the headquarters were not even aware of.

But the initial wave of research into extracting useful marketing data from Internet discussion boards tended to focus as much on the sizable obstacles as the benefits. Columbia’s Netzer, in his paper, cites a 2006 report that aimed to use reviews from moviegoers on Yahoo.com message boards as a tool to predict box office sales and which concluded that trying mechanically to make sense of as many as 12,000 posted comments was “an extremely tedious task.”

**Learning how consumers speak.** Most of the obstacles encountered by researchers are the result of the same factor that makes these consumer conversations
so potentially valuable: their raw and democratic nature. Additionally, comments are not only informal but often ungrammatical, raising additional challenges. As noted by Bradlow and Lee in their recent paper entitled “Automated Marketing Research Using Online Customer Reviews,” a successful automated system of data collection needs to account for these stark differences in language, vocabulary and other ways that different consumers discuss products online. A useful system, they write, must identify “not only what attributes customers are speaking about but also how they speak about it.”

Thus, the most up-to-date research in successfully culling this type of language increasingly involves collaboration between marketing experts and highly skilled computer scientists. In designing his program, Bradlow decided to avoid product rating sites – where typically customers assign stars or other numerical rankings – and focus on what the researchers called “Pro” and “Con” sites like Epinions.com where customers raise positive and negative product qualities. The results are mapped to show both the relationships between competing brands as well as the frequency that consumers discuss key attributes.

Bradlow’s most recent research looked at online discussion boards for digital cameras – a high-tech product with highly opinionated customers and a mix of professionals and amateurs – but also comment boards about hotel service in several major U.S. cities as well as customer forums on a more mundane consumer products, toaster ovens. Even with toaster ovens – whose customers posted on topics such as “browning” or “ease of access to interior” – Bradlow reported finding evolution in the ongoing discussions of brands, perhaps reflecting the introduction of new product features.

In the camera conversations, the automated web-scraping methods and mapping used by the researchers showed, for example, a larger than expected number of buyers was raising issues with the Casio brand and its start-up speed. On the other hand, they also found that some digital-camera attributes singled out by marketing managers in more traditional conjoint analysis – such as “body construction” or “power adaptors” – were not major topics of interest to consumers. Bradlow’s automated web-scraping method also showed that camera discussions changed over a period of time, with newer discussions of topics such as “ISO settings” and “multiple shot modes” suggesting a market that was becoming more technically sophisticated.

“A critical goal in developing the technique involved understanding the specific phrases that consumers use and then creating a series of “word vectors” to help extract data from online forums.”

“Fly on the wall” of the Web. Netzer’s research team – which also included Ronen Feldman and Jacob Goldenberg from Hebrew University – also sought to develop an advanced method for mining online consumer discussions and quantifying the results – rooted in understanding of the semantics of Internet forums. The specific system developed by this team is called CARE, which stands for CRF Assisted Relation Extraction; essentially it combines the successful data-extraction technique of condition
random field analysis, or CRF, with rules crafted by humans based on studying the language of online forums. As with Bradlow’s research, the Netzer team found that less-structured consumer forums provided a richer vein of conversations and topics to mine for consumer insights.

Netzer said the goal of the research was to better understand customer thinking without conducting interviews. “We’re actually a fly on the wall of the Web – we never asked a single question,” he said. Netzer’s team focused on two diverse categories of consumer conservations – automobile consumer forums and an Internet discussion site devoted to diabetes drugs. In the case of the car conversations, the CARE method enabled the team to develop an elaborate map demonstrating showing which vehicles that consumers are most likely to compare to each other; it shows compact cars like the Geo and Toyota Echo clustered at one end and luxury brands such as Lexus on the other.

Perhaps more importantly, the Netzer-led research developed data that closely linked certain attributes with specific car brands. Some of those associations, the paper noted, might not be a surprise to company executives. The Toyota Corolla, for example, was closely linked in the research with terms like “valuable,” “markup,” “inexpensive,” and “investment,” which is in line with the way that vehicle is marketed to would-be buyers. However, the data also showed that the Nissan Sentra had an unexpected association with the word “college.” The authors note that this suggests the Sentra is perceived as “a college car,” adding: “This market segment may not be obvious to Nissan through a simple demographic analysis of Sentra’s buyers because the buyers may be parents of college students.”

The voice of the consumer, interpreted. The diabetes-drug forum studied by the Netzer team also proved a valuable source of material for the authors, in part because of the passion of the conversations between patients or family members affected by the disease. The researchers found that. Most importantly, such discoveries supported the broader theory put forward by leaders in Internet marketing research that consumer forums can identify major issues that firms were not even aware of.

Because their web scraping techniques are so new, both the Bradlow and Netzer teams conducted tests to confirm the reliability of the data. In the Bradlow paper, the automated data was compared to traditional conjoint analysis methods and showed a high correlation in results; similarly, the Netzer team read back on 500 selected review and found the automated CARE program was roughly 90 percent reliable in summarizing the message from the online authors, At the same time, the researchers have identified several areas involving the dynamics of online discussion that require additional research. For example, other academic researchers are currently looking at the way that the interaction between consumers online can alter the course of the conversation over a period of time, in ways that affect the data.
However, the researchers also assert their recent advances in web scraping techniques are already creating a valuable pool of consumer knowledge that traditional methods such as focus groups were ignoring. The ability to use Internet forums to discover what Bradlow calls “unseen attributes” of brands and their products will mean that businesses can identify and solve once unnoticed problems, add new features that are in demand from everyday consumers, or – as in the case of the college-friendly Nissan Sentra – design marketing campaigns for a newly discovered niche audience. WCAI remains actively engaged in sponsoring a new round of research that will bring its empirical perspective to the next generation of marketing solutions, so that American businesses can reap the full benefit of this valuable pool on information for decades to come.