

Using Lexical Semantic Analysis to Derive Online Brand Positions: An Application to Retail Marketing Research

Praveen Aggarwal^{a,*}, Rajiv Vaidyanathan^{b,1}, Alladi Venkatesh^{c,2}

^a Labovitz School of Business & Economics, University of Minnesota Duluth, 385A LSBE, 1318 Kirby Drive, Duluth, MN 55812, United States

^b Labovitz School of Business & Economics, University of Minnesota Duluth, 385B LSBE, 1318 Kirby Drive, Duluth, MN 55812, United States

^c CRITO (Center for Research on Information Technology and Organizations) at the University of California, Irvine, CA 92697, United States

Abstract

This paper provides an innovative approach to brand tracking in the context of online retail shopping by deriving meaning from the vast amount of information stored in online search engine databases. The method draws upon research in lexical text analysis and computational linguistics to gain insights into the structural schema of online brand positions. The paper proposes a simple-to-use method that managers can utilize to assess their brand's positioning relative to that of their competitors' in the online environment.

© 2009 New York University. Published by Elsevier Inc. All rights reserved.

Keywords: User-generated content; Lexical semantics; Brand personality; Positioning

Recent developments in e-Commerce indicate a phenomenal growth in Internet retailing (Weathers, Sharma, and Wood 2007; Yadav and Varadarajan 2005). One area of increasing attention among retailing researchers and strategists is online branding and customer behavior (Ailawadi and Keller 2004). In this context, a typical question the brand manager might ask is, "What meanings do people associate with our brand?" As it is becoming increasingly difficult to exclusively claim an attribute or quality assertion for one brand ("Volvo makes safe cars") in a market crowded with a multitude of brands, the best that brand managers can do is to put some distance between their brands and those of their nearest competitors for such claims. From a managerial perspective, it would be helpful if researchers could provide an efficient and easy-to-use method for ascertaining a brand's association with key descriptors that could also be used to gauge brands' relative performance on those descriptors. This is a critical issue in online retailing and branding research.

The objective of this paper is to propose a simple, non-intrusive way of discovering brand–descriptor associations by

exploiting the vast and continuously expanding information databases of online search engines such as Google. Using such associations, we propose a method based on lexical semantic analysis for comparing a brand's positioning against its competitors, assessing differences in how the brand is perceived, and drawing inferences about a brand's personality. This method permits us to obtain a summary assessment of the online representation of a brand, based on a set of managerially relevant descriptors. The basic premise of this paper is that there is valuable information contained in these brand–descriptor associations that can be efficiently mined using simple, semantic search-algorithms.

Background and problem statement

If managers do a simple keyword search to assess what the online world says about their brands, they will usually get back thousands of hits from a search engine. They will then face the onerous task of sifting through the voluminous data generated by those hits. Currently, the usefulness of the hit information is limited by the fact that the listings contain no summary information about the content. Nor do the listings contain information on the "semantic" properties of hits containing textual data. This raises the question of how to derive meaning from text-oriented data, especially as it relates to a neutral target such as a brand name.

* Corresponding author. Tel.: +1 218 726 8971.

E-mail addresses: paggarwa@d.umn.edu (P. Aggarwal),
rvaidyan@d.umn.edu (R. Vaidyanathan), avenkate@uci.edu (A. Venkatesh).

¹ Tel.: +1 218 726 6817.

² Tel.: +1 949 824 6625.

To obtain managerially relevant insights through web searches, we draw upon research in lexical text analysis. We propose some critical indices that can be used by managers to glimpse the structural schema of their brands in the online world by examining links between brand names and key adjectives/descriptors in these massive online databases. We propose some techniques and measures for drawing on data stored in search engines. We also discuss three studies that report “real” applications of these techniques to generate information of interest to marketers.

A recent development motivating studies of this kind is the notion of a semantic web (Berners-Lee, Hendler, and Lassila 2001; Leuf 2006). Until recently, the architecture of the web had been viewed in syntactic terms; the web is examined merely as a repository of information without interpretation or meaning attribution. The focus of search system development has been primarily on the efficient storage and retrieval of information rather than on recognizing the meaning of content. A semantic vision of the web allows for machine-based inferences, and implies that information can be processed and understood by computers, which leads to “lexical semantic analysis.” In our research, we draw on the insight that the co-occurrence of adjectives (e.g., “reliable”) and nouns (e.g., “Sony”) is a strong indicator of subjectivity (Hatzivassiloglou and McKeown 1997). Subjectivity refers to the aspect of a language used to communicate an evaluation or an opinion (Banfield 1982; Wiebe 1994). When such co-occurrences are compounded over a vast amount of textual data, reliable inferences about a brand’s online position can be obtained.

To exploit the semantic potential of the web, this paper draws on theoretical developments in computational linguistics, lexical semantic analysis, and statistics. We provide an easy to use, comprehensive methodology for studying brand associations, and for drawing meaningful conclusions from the hit counts provided by popular search engines such as Google. This methodology will allow retailers and marketers to effectively use the vast dynamic data stored in search engine databases to develop online brand strategies.

Lexical semantics—some initial ideas

Typing the name of a brand or a travel destination into a search engine will often provide thousands of hits, but the searcher does not gain information on whether these sites have positive, negative, or neutral things to say about the brand or destination (Turney 2002). To get a better insight into the meaning in the text, researchers in the fields of artificial intelligence and natural language processing are beginning to focus on evaluating the content of large text-based corpora.

One of the early approaches to textual data in the field of marketing is content analysis (Kassarjian 1977). Content analysis has been used to analyze advertisements and textually based promotional materials (Arnold, Kozinets, and Handelman 2001; Voss and Seiders 2003), to determine knowledge structures of salespeople (Sharma, Levy, and Kumar 2000), and to understand communication through home shopping networks (Warden et al. 2008). Content analysis has not been embraced in a signif-

icant way by marketers to the extent it has been in fields such as communication research and journalism. Presumably marketing research has been deterred by problems associated with sampling and measurement, as well as the reliability and validity of content categories. Also, over the years, there has been a greater push toward research driven by numeric data. In addition, content analysis requires manual coding of data from diverse sources, and this task can be quite prohibitive.

Another recent online research method gaining attention is “netnography” (Kozinets 2002). Marketing scholars have studied online reviews to uncover the dimensions of online service quality (Yang and Fang 2004), as well as online conversations as word-of-mouth communication (Godes and Mayzlin 2004). A common theme running across these studies is that authors typically restrict their data to a small subset of available discourses (typically a few hundred pages) in order to keep the process of data parsing and analysis manageable. While manual analysis of these pages results in a rich and in-depth investigation of the restricted dataset, the obvious compromise involved in this approach is the exclusion of a vast majority of available pages.

Recent developments in computerized semantic analysis have opened new possibilities in text-data analysis, and are enabling researchers to overcome the shortcomings of traditional content analysis and netnographic approaches. With developments in machine intelligence and semantic analysis software, there has been a greater scope for exploring web-based information using new techniques. It is in this spirit that our paper examines the semantic properties of brand positions.

Lexical semantics—the method

For the purpose of our analysis, we utilize two key properties of textual data: (1) consumers attach meanings to words and (2) meanings are inherent in the text, or more specifically in the adjectival expressions used by the author of the text. Understanding the valence or semantic orientation of a word (Hatzivassiloglou and McKeown 1997) can help describe its associated noun. That is, we can infer the evaluative nature of a sentence describing a noun by examining the association between the noun and the adjectives modifying that noun.

Descriptors which identify the evaluative nature of brand-related text have another use. Prior research has demonstrated a positive relationship between the presence of adjectives and the subjectivity of the sentence (e.g., Bruce and Wiebe 1999; Hatzivassiloglou and Wiebe 2000). Within the context of brand information available on the web, we can infer the overall evaluative nature of text-based, brand-related information by examining the semantic orientation of the adjectives or descriptors used in association with the brand. The goal is to get a sense of the evaluation of brands in subjective sentences (as opposed to objective sentences which merely state non-evaluative facts) by examining the association between the brand and various carefully selected adjectives or descriptors.

Researchers have inferred the semantic orientation of phrases extracted from an online database using a pointwise mutual information (PMI) algorithm. PMI essentially takes the difference between the number of associations between the target and

selected positive words and the number of associations between the target and selected negative words. Turney (2002) developed such an algorithm to classify reviews as positive or negative, based on the semantic orientation of the phrases used in the reviews. For each review, he determined the semantic orientation of phrases by examining their association with terms “excellent” and “poor” (for positive and negative orientation, respectively). Based on the average semantic orientation of all adjectives and adverbs in a given review, the algorithm assigned a “recommended” or “not recommended” label to the review. The algorithm was applied to classify 410 online reviews from Epinions (for automobiles, banks, movies, and travel destinations), and achieved an accuracy of 74 percent.

PMI is based on the premise that the strength of semantic association between two words (points) can be measured by the level of statistical dependence (mutual information) of the words. Statistical dependence has been defined in terms of their co-occurrence in a set of text documents and is defined as:

$$PMI(a, b) = \log_2 \left[\frac{p(a \& b)}{p(a)p(b)} \right]$$

where $p(a \& b)$ is the probability that the two words, a and b , co-occur in the text, $p(a)$ is the probability of a 's occurrence in the text, and $p(b)$ is the probability of b 's occurrence in the text.

If the two words are statistically independent, then the ratio will be equal to one and its log will be zero. However, if the two words are strongly associated with each other, then the log of the ratio is positive and indicates the amount of information we acquire about the presence of one word given the presence of the other word. This co-occurrence essentially builds on the idea that “a word is characterized by the company it keeps” (Firth 1957, as noted by Turney, 2001). Church and Hanks (1989) note that linguistic researchers frequently classify words based not only on their meaning but also “on the basis of their co-occurrence with other words” (p. 76).

Co-occurrence of brands and descriptors

In our applications, word probability $p(x)$ will be calculated as the number of documents in an online database that have the word x mentioned in them, divided by the total number of documents relevant to that problem. Let's say we are interested in examining the associations between automobile brands (Toyota, Honda, BMW, etc.) and certain descriptor variables (reliable, efficient, expensive, etc.). In this scenario, one would calculate the PMI of “Toyota” and “reliable” as follows:

$$PMI = \log_2 \left(\frac{p(\text{Toyota}\&\text{Reliable})}{p(\text{Toyota}) \times p(\text{Reliable})} \right)$$

Here, $p(\text{Toyota})$ is calculated as the number of hits for “Toyota” – $f(\text{Toyota})$ – divided by the total number of pages for “automobiles” – $f(\text{auto})$. Similarly, other probabilities are calculated by estimating the number of documents with a given word(s) divided by the total number of documents that contain the word

“automobile.” Hence,

$$PMI = \log_2 \left(\frac{f(\text{Toyota}\&\text{Reliable})/f(\text{auto})}{(f(\text{Toyota})/f(\text{auto})) \times (f(\text{Reliable})/f(\text{auto}))} \right)$$

or

$$PMI = \log_2 \left(\frac{f(\text{Toyota}\&\text{Reliable})}{f(\text{Toyota})} \times \frac{f(\text{auto})}{f(\text{Reliable})} \right)$$

Notice here that the ratio $(f(\text{auto})/f(\text{Reliable}))$ is the same for all brands. Also, given that the log function is monotonically increasing, we can drop it without affecting the end result. Thus, the modified PMI (PMI_{mod}) in the context of inter-brand comparisons can be stated as follows:

$$PMI = \left(\frac{f(\text{Toyota}\&\text{Reliable})}{f(\text{Toyota})} \right)$$

In the empirical studies reported below, we make use of this ratio to compare brands and draw conclusions about brand positions in the online environment. Specifically, we examine the occurrence of certain adjectives and descriptors in the context of particular brands to draw conclusions about a brand's online persona. The underlying assumption of this methodology draws upon the work of Bruce and Wiebe (1999) and Hatzivassiloglou and Wiebe (2000), which assumes that if a particular descriptor or adjective (for instance, “lightweight”) appears frequently in association with a particular brand (for instance, Samsonite luggage), then one can tentatively infer a positive relationship between the two. Of course, the assessment of the relationship will not be as objective or “clean” as it would be if one were to actually read the entire text to come to a definite conclusion about the intended meaning. However, if the descriptor and the target brand appear together in hundreds or thousands of documents, one's confidence in the association between the two is higher. Although there is no reliable measure of the amount of consumer-generated content on the web, the consensus is that the consumer-generated component of the web content is very large and growing (OECD 2007). However, since the WWW includes a large amount of content generated both by consumers and marketers, the analysis of the relationships reflects both consumer- and marketer-generated representations of brands.

Data source

We used searches of Google's database of web content to obtain data on the number of hits for various brand–descriptor combinations. In order to access the Google database easily for repeated queries, we used a beta version of Google's application program interface (API) in the Fall of 2004. APIs provide a common set of standards by which organizations allow users to access specific program functionality without exposing the operation of their proprietary software. The interface allows commands (written in a language of the user's choice) to access specific aspects of the organization's software, and returns uniquely formatted results. In the case of Google, the API allowed us access to their search database. The API describes both how the request query should be formatted and also how to interpret the results returned by Google. We used Visual Basic.

NET search queries to extract data of interest from the results returned by Google (number of hits for pages containing specific terms). The API provides an automated way of getting the results of repeated searches of the database. While we collected data using Google's original API, the API has been repeatedly revised by Google, Inc. The contribution here is the use of algorithms based on computational linguistics literature to make inferences about the meaning in the textual content on the web. The algorithms provided here can be used to write programs using any API to gather data. For example, Yahoo! provides their own API to access the content in their database.

Google uses unique identifiers to manage the hits returned by the database. For example, the "+" sign before a word tells Google to only return those pages in the results that contain the word. Similarly, the "-" sign before a word only returns pages that do not contain the word following the sign. For example, the search query "+Toyota -Reliable" will only return those pages that contain the word "Toyota" but do not contain the word "reliable." It is important to note that we were only interested in the number of page hits returned for word combinations. We could have used the Google.com home page to individually query each combination of words in which we were interested, since Google.com does provide the number of hits for any combination of words used in a query on the home page. This would have been painstaking because each query would have had to be typed in separately, and the number of hits reported by Google would have had to be manually entered into a spreadsheet. However, the use of a computer program and Google's API allowed us to automate the process of making numerous queries to the database by grabbing the search terms from a spreadsheet, capturing the returned information, performing minor calculations (e.g., percentages) on the results, and formatting it in a manner that made data analysis easy. The software allowed us to send Google hundreds of requests (differently formatted query sets), and record the results in a time span of a few minutes. Additional information about using the current version of Google's APIs to access the database can be found at <http://code.google.com/apis/gdata>. Information on Yahoo! Inc.'s search APIs can be found at <http://developer.yahoo.com/search/>.

Study 1: Brand positioning analysis

Despite the central role of positioning in marketing, there is surprisingly limited empirical work on consumer-derived typologies that reflect brand positioning strategies (Blankson and Kalafatis 2004). Branding is particularly important in the retail industry given the industry's highly competitive nature (Ailawadi and Keller 2004) and the fact that many retail chains are introducing private brands to compete with national brands (Bellizzi et al. 1981). Brand positioning relates to how brands are perceived by consumers, relative to the competition. Conceptually, this requires an understanding of how brands become part of the consideration set based on general brand category associations or descriptors representative of an exemplar brand, and then how consumers differentiate between the brands within their consideration set so that a brand has a unique image (Punj

and Moon 2002). Marketers attempt to define the position of their brand using *positioning statements* that implicitly identify not only the key associations that are relevant to brand comparisons, but also the key differentiating factors that set the brand apart from its competition. Brands are generally positioned as having more good attributes and fewer bad attributes than competing offerings (Henderson, Iacobucci, and Calder 2002). Combined with the research discussed earlier on the importance of adjectives in defining the semantic orientation of a target, this would suggest that brand positioning statements offer a valuable source of adjectives or descriptors that define the *intended position* of a brand.

Given the importance of understanding how consumers perceive brands relative to one another, it becomes important for managers to continually monitor their brands and ensure that this consumer-based positioning is consistent with the marketer-intended positioning. The most widely used tool for positioning analysis is the *perceptual map*. In perceptual mapping, brands within a product category are plotted on a multidimensional space for key dimensions that differentiate the brands (Shocker and Srinivasan 1979). These maps can be used to identify new product opportunities, verify managers' views on competitive structure and positioning, identify competitors, and study reputation or image (Lilien and Rangaswamy 2002). Although multidimensional scaling has traditionally been used to develop positioning maps from similarity data, the research on semantic orientation and adjective association suggests that it may also be a fertile source of information on online representations of brands on key dimensions.

Building on the research in lexical semantic analysis and the use of word associations to identify semantic orientations, we develop a model to create brand positioning maps based on an analysis of web pages containing brand-related information and descriptors that define brand positions in the offline world. We used custom-coded applets that built upon the APIs provided by Google to analyze the data in all 8+ billion web pages in their search engine database. We assume an *association* between a brand and a descriptor if the search engine returns a "hit" when we query the search engine for both the brand and that descriptor. While the descriptors are bound to be associated with multiple brands, if the descriptors truly associate with the brand, we would expect to see *relative differences* in the extent of association between each brand and the descriptors. We next present the model underlying our analysis, and then provide a concrete example of the application of this model to an analysis of the positioning of Procter & Gamble (P&G) detergents.

Conceptual model

Assume that we are interested in determining how a set of competing brands fares on a list of descriptors that are relevant for the brand's product category. As stated above, we observe association by counting the number of pages that include both the brand's name as well as the descriptors under consideration. So, for example, if there were three competing brands of chocolates and we wanted to determine how they were viewed on the descriptor "bitterness," we count the number of page hits for

each brand that also contained the word “bitter.” Thus, if 70 percent of all pages that included brand A also included the word “bitter,” compared to 20 percent and 25 percent for brands B and C, one may conclude that the brand A is associated with bitterness more often in the online environment compared to brands B or C. In order to avoid including those pages that have the brand name but may still be totally unrelated (e.g., a web page that talks about planet Mars may have nothing to do with Mars chocolate), we can use a product category filter (e.g., “chocolate”) to limit search to only those pages that are product category related. Note that it is still possible to capture pages that go past the filter and may have nothing to do with the brand. For example, a page may record views on the harmful impact of soaps and detergents on our water resources, and also mention how some of the pollutants may get carried to the ocean through tides. Such a page would get past the “detergent” filter, will contain the brand “Tide,” and still not be brand related. While it would be ideal to remove all such pages from our analysis, such removal would be impractical. Given the vast amount of information and the number of pages available on the web for nationally marketed brands, even without removing such pages it is still meaningful to draw broad generalizations based on page hits containing the targeted word associations. While some of the pages included in the analysis will be “false” hits, the overall patterns should still hold.

Starting with the PMI ratio described earlier, we extend the model to include multiple attributes for each brand and then examine the relative differences between brands to draw conclusions of interest to retailers and marketers.

Assume that there are k brands that can be evaluated on y attributes. For a given brand, we first count pages that contain the brand name (after placing the product category filter). Next we count the number of pages with the brand–adjective association. These are the subset of pages that have the brand name and also contain the targeted adjective. For each association of brand i on attribute j , we can define a “raw score” as the percentage of web pages that contain both brand i and the attribute j . Let us denote the raw scores (in percentage terms) with s_{ij} . We can calculate these raw scores for all combinations of brand names and attributes ($k \times y$ number of raw scores). In order to compare how different brands fare against each other on a given attribute, we first create an index that stretches out the spread to a uniform scale (100 points). Thus, the brands at the two extremes serve as anchors that are 100 points apart, and every other brand falls somewhere in between these two anchors. This modification allows us to observe inter-brand distances using a uniform scale. We do this by applying the following modification to obtain x_{ij} , the positioning score:

$$x_{ij} = \frac{s_{ij} - \bar{s}_j}{s_{j \max} - s_{j \min}}, \tag{1}$$

where \bar{s}_j is the average of raw scores for attribute j .

Note that the above modification will create both positive and negative indices. Now, in order to allow comparison of this modified score across attributes, we apply another scale modification that shifts the 100 point spread in such a way that the anchors move to +50 and –50 at the two extremes, without altering the

inter-brand distances. This “centers” the data onto a common scale. We apply the following modification to get $x_{ij \text{ mod}}$, the modified positioning score (mps):

$$x_{ij \text{ mod}} = x_{ij} - \left(\frac{x_{j \min} + x_{j \max}}{2} \right) \tag{2}$$

These modified positioning scores can be used to create a perceptual map with the dimensions of interest. The map will show, on a standardized scale, how the brands are perceived relative to each other on different attributes.

While the modifications noted above stretch (or compress) the range between different brands for a given attribute to a standardized 100 points, it is important to note that the raw score range does contain useful information. The spread of a range indicates how far apart the brands are perceived as being from each other on that attribute. For example, if the range is narrow, the brands are fairly comparable to each other on that attribute. Thus, the raw range spread is an indicator of an attribute’s diagnosticity, or how well the attribute differentiates one brand from others. Thus, we define diagnosticity, D_j , as follows:

$$D_j = s_{j \max} - s_{j \min} \tag{3}$$

In order to determine a brand’s uniqueness on a given attribute, we calculate its distance from the brand nearest to it on that attribute. We calculate I_{ij} , the attribute-level isolation index of brand i on attribute j , by computing the difference between the modified positioning score for a brand and that of the brand nearest to it. Since we are indifferent to the direction in which the nearest brand lies, we ignore the signs on the difference. Essentially, this index shows how *isolated* a brand is on a certain attribute j . If there are any other brands close to brand i on attribute j , it will have a low isolation index (I_{ij}):

$$I_{ij} = \text{Abs}(x_{ij \text{ mod}} - x_{ij \text{ mod}(\text{near})}), \tag{4}$$

where $x_{ij \text{ mod}(\text{near})}$ is the mps of the brand closest to brand i in either direction. Note that a high I_{ij} indicates that the brand i has a relatively uncluttered (isolated) position on attribute j . In other words, the brand is perceived as distinct from competing brands on that attribute.

Application

We now apply the above approach to a real-life example. P&G offers a number of detergent brands. It also goes to great lengths to position these brands differently. Using the approach outlined in the previous section, we examine how successful P&G is in achieving distinct positions for its detergent brands.

Step 1: Generating brand descriptors

In order to generate a list of brand attributes/benefits on which to compare competing brands, we first identified positioning statements for the seven brands included in this study. These positioning statements, derived from published and company sources, are listed in Table 1. The last column in this table indicates the adjectives/descriptors derived from these statements that were subsequently used in our analysis. Where we identi-

Table 1
Positioning statements and brand adjectives.

Brand	Positioning statement	Derived adjectives
Bold	For clean, soft, fresh-smelling clothes, with a built-in fabric softener	Fresh
Cheer	The color guard	Color, guard
Dreft	Helps remove tough baby stains, pediatrician recommended	Baby
Era	Built-in stain remover, the power tool for stains	Power
Ivory Snow	Gently cleans fine washables, 99.4 percent pure	Gentle
Oxydol	Stain-seeking bleach	Bleach
Tide	Fights tough stains, keeps clothes looking like new	Tough

fied more than one term, we used the underlined descriptor to keep the analysis simpler.

Step 2: Getting page hits for each brand

After determining the adjectives/descriptors for inclusion in this study, we determined the number of page hits for each of the seven brands. In order to limit our search to the pages that were more likely to be related to detergents, we used a page filter to pick out only those pages that contained the term “detergent.” Using the filtered pages, we next took a raw count of the numbers of pages that had a mention of each of the brands. These numbers are reported in Table 2.

Step 3: Determining brand rankings for each descriptor

Once we had the page count for each brand, we determined the percentage of pages for a given brand that had a mention of each of the seven adjectives/descriptors used in the study. This gave us our raw count, s_{ij} . Next, we were interested in determining if the positioning descriptor intended for a particular brand was more likely to be associated with that brand. For example, we wanted to determine whether or not the term “gentle” was more often associated with “Ivory Snow” (after controlling for the number of page hits for Ivory Snow). The results of this analysis are reported in Table 3. The last column in Table 3 refers to the rank order of a brand for a given descriptor. For example, if Ivory Snow had the highest percentage of hits with the word “gentle” among the seven detergent brands, its rank for “gentle” would be #1. As noted in Table 3, all seven brands do reasonably well in terms of their respective positioning adjectives (as indi-

Table 2
Page hits for each brand.

Rank	Brand name	# of pages
1	Tide	10,800
2	Era	8,060
3	Bold	6,520
4	Cheer	3,230
5	Dreft	1,300
6	Ivory Snow	971
7	Oxydol	477

Table 3
Brand association with positioning adjective.

Brand	Descriptor	Ranking
Bold	Fresh	2
Cheer	Color/guard	1
Dreft	Baby	1
Era	Power	2
Ivory Snow	Gentle	2
Oxydol	Bleach	1
Tide	Tough	3

Table 4
Brands' performance on positioning adjective “Baby”.

Rank	Name	Raw percent (s_{ij})	Positioning score (x_{ij})	MPS ($x_{ij\ mod}$)
1	Dreft	59.5	52	50
2	Ivory Snow	55.5	42	40
3	Cheer	55.1	41	39
4	Oxydol	31.2	-19	-21
5	Bold	27.1	-29	-32
6	Era	22.7	-40	-43
7	Tide	19.8	-48	-50

cated by their high rank on those adjectives). Please note that we looked at both “color” and “guard” for Cheer (ranked #1 for both terms), but will be using only “color” descriptor for the rest of our analysis.

Step 4: Determining positioning scores for each brand/descriptor combination

In order to determine the relative positioning of brands on a given attribute, we calculate the positioning scores of brands on each of the seven attributes. To illustrate the steps involved in these calculations, let's examine the relative brand positions on “baby” and “power.” We begin with looking at the percentage of pages for each brand that included the word “baby.” We do the same for the descriptor “power.” These percentages are recorded in the third column in Tables 4 and 5. Column 4 lists the positioning scores obtained using the formula in Eq. (1). The modified positioning scores are calculated by applying the scale shift adjustment outlined in Eq. (2).

The modified positioning scores for the remaining attributes are given in Table 6.

Table 5
Brands' performance on positioning adjective “Power”.

Rank	Name	Raw percent (s_{ij})	Positioning score (x_{ij})	MPS ($x_{ij\ mod}$)
1	Cheer	51.1	53	50
2	Era	44.8	34	31
3	Bold	40.4	21	18
4	Oxydol	31.9	-4	-7
5	Tide	30.0	-10	-13
6	Dreft	17.6	-47	-49
7	Ivory Snow	17.4	-47	-50

Table 6
Modified positioning scores.

Name	Fresh		Bleach		Gentle		Color		Tough	
	Raw percent (s_{ij})	MPS (x_{ijmod})	Raw percent (s_{ij})	MPS (x_{ijmod})	Raw percent (s_{ij})	MPS (x_{ijmod})	Raw percent (s_{ij})	MPS (x_{ijmod})	Raw percent (s_{ij})	MPS (x_{ijmod})
Bold	29.4	-21	16.4	-45	17.0	10	34.8	-8	13.9	-18
Cheer	46.4	50	35.0	24	23.5	50	48.6	50	24.2	50
Dreft	27.5	-30	35.7	27	15.9	3	38.9	9	9.0	-50
Era	25.9	-36	15.0	-50	12.8	-17	32.9	-16	17.9	8
Iv. Snow	29.4	-22	37.5	33	20.2	29	40.4	16	9.9	-44
Oxydol	22.6	-50	41.9	50	7.5	-50	28.5	-34	9.0	-50
Tide	23.4	-47	16.9	-43	11.3	-26	24.7	-50	14.0	-17

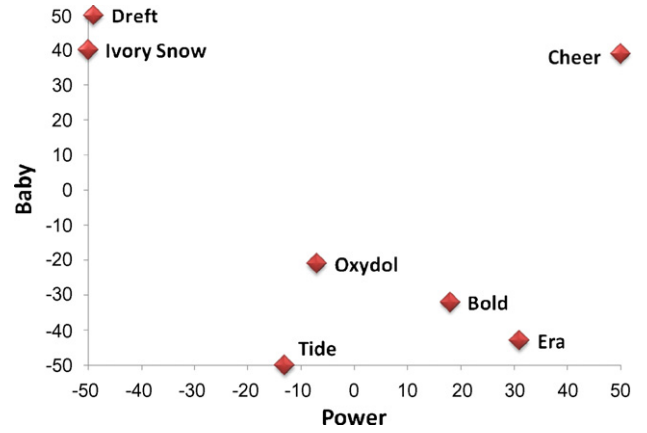


Fig. 1. Positioning Map derived from Google Analysis.

Step 5: Drawing a perceptual map

Once we have the modified positioning scores, we can plot them on a graph whose axes go from -50 to +50. We have plotted such a graph using “baby” and “power” dimensions in Fig. 1. On the basis of the way different brands plot on these two dimensions, we may conclude that for most brands (except Cheer), there seems to be an inverse relationship between the two dimensions. Brands that score high on the “baby” dimension fare poorly on the “power” dimension. Similar graphs can be plotted using other combinations of the descriptors used in this study.

Step 6: Determining the diagnosticity of attributes

While the modified positioning scores and perceptual maps show the relative performance of each brand on different attributes, they do not directly address the ability of a given attribute to discriminate among brands. To determine that, we calculate the diagnosticity index using the formula developed in (3) above. The results are reported in Table 7.

Thus, from Table 7, one can conclude that brands differ significantly from each other on attributes “baby” and “power,” whereas attributes “gentle” and “tough” have the least diagnostic power.

Step 7: Calculating the isolation index

The final step in the process of determining the online positioning of brands is to calculate an isolation index for brands based on individual attributes. We do so using the formula developed in Eq. (4). The results are reported in Table 8.

As can be seen from Table 8, certain brands stand out on certain criteria as distinct from other brands. For example, on “freshness,” Cheer is clearly distinct from every other brand. Similarly, Era, and to a lesser extent, Cheer, occupy a relatively uncluttered spot on the “toughness” dimension.

Table 7
Diagnosticity index.

	Fresh	Bleach	Power	Gentle	Color	Tough	Baby
Diagnosticity	23.8	26.9	33.7	15.9	23.9	15.2	39.7

Table 8
Isolation index.

Brand	Fresh	Bleach	Power	Gentle	Color	Tough	Baby
Bold	1	2	13	7	8	1	9
Cheer	71	3	19	21	34	42	1
Dreft	6	3	1	7	7	0	10
Era	6	5	13	9	8	42	7
Ivory Snow	1	6	1	19	7	6	1
Oxydol	3	17	6	24	16	0	11
Tide	3	2	6	9	16	1	7

The above illustration using P&G detergent brands demonstrates one way of using search engine data to infer brand associations and brand positioning. In the case of P&G, we discovered that the brands are relatively well established in their respective positioning domains (as indicated by their rankings on positioning adjectives). We also derived a perceptual map using indices developed in this paper to understand the relative placement of brands on different dimensions. Finally, we looked at indices (diagnosticity and isolation) that can be used as surrogates for a dimension's effectiveness at segregating brands and a brand's relative positioning on a given dimension.

Study 1 discussion

Study 1 was an initial attempt at mining a search engine database for information that can be used to understand consumer perceptions of brands. When people participate in an online setting, they often do so to express opinions. In the context of product-related postings, these opinions are often expressed in terms of product attributes and performance measures. Therefore, such online discussions present a valuable source of qualitative and narrative assessments of products and brands. In the past, marketers have generally ignored this information because it is voluminous and does not lend itself to quantification very easily. While it is true that doing text analysis of tens of thousands of documents can be a complicated and onerous task, one can use a short-cut if one knows what to look for in those documents. If one has a set of keywords that are central to answering a marketing question, one does not have to scan each and every word in a document. Instead, one can focus on those keywords and associations to derive meaningful conclusions. Study 1 presented one such example where the keywords to be used in the analysis were available from another context. It is important to point out that since this study used hit rates of pages with brand information, it included content generated both by consumers and marketers. However, a method could be developed to restrict the analysis to pages with primarily consumer-generated content (e.g., only blogs or review sites).

By examining associations between brands and carefully selected adjectives/descriptors, one can go beyond merely counting text content to uncover the meaning of the content. The content reveals relationships that have the potential to inform marketing decisions. In the context of brand positioning, one can identify the keywords that represent a brand's position, and

this opens up the possibility of identifying interesting and meaningful relationships between brands and descriptors. Of course, one does not have to restrict the analysis to a brand's stated position descriptors. One can, for example, use brand personality descriptors to discover a brand's online persona. This was the focus of Study 2.

Study 2: Online brand personality assessment

It is now common to refer to brands as having their own personalities (Aaker 1997; Fournier 1998). Aaker (1997) defined brand personality as "the set of human characteristics associated with a brand" (p. 347). Through her empirical work, she identified five distinct dimensions of brand personality: sincerity, excitement, competence, sophistication, and ruggedness. She further broke down these five dimensions into 15 "facets" that were themselves composed of 42 "traits." In Study 2, we attempt to link brands to personality traits using the Google search engine database and the PMI index discussed earlier. The 42 brand personality traits (mostly adjectives) provide us with a rich list of descriptors that we can use to assess a brand's online personality.

Although research papers have been published on brand personality, there have been few empirical studies that have actually attempted to determine a given brand's personality. Given the richness of data available on brands in the online environment, it should be possible to map a brand's personality. Such an assessment will, of course, depict how a brand's persona is reflected primarily on the Internet. To the extent that people talk about a brand online based on their experiences with the brand in the offline environment, it is reasonable to propose that the online brand personality is closely related to the actual market persona of that brand.

We chose to evaluate the online brand personalities of two perfumes: Stetson and White Diamonds. We chose perfumes as our test category because there are very few tangible differences in product attributes. Most differences are based on image or brand personality attributes carefully crafted by marketers to position brands in a competitive marketplace. The two brands in our analysis were chosen for the stark differences they portray in their personalities. Stetson is generally positioned as the "legendary fragrance of the American West," "a rich blend of rugged woods and spice," and "for men and women who live with confidence, independence and pride" (from Stetson's corporate web site at www.stetsonshop.com). On the other hand, White Diamonds, with Elizabeth Taylor as its spokesperson and the tagline "I never forget a woman in diamonds," is positioned more as an upscale, prestige scent (Klepacki 2004). Given their positioning differences, we would expect to see significant differences in the brand personalities of these perfumes. On Aaker's five dimensions, Stetson is likely to perform better on ruggedness, competence, and sincerity, while White Diamonds should do better on sophistication and excitement. In order to do our analysis, we first collected brand association data (PMI_{mod}) for all 42 traits listed in Aaker (1997) for both brands. The filter used for this study was "fragrance OR perfume." The total number of estimated hits for the two brands was comparable: 17,600

Table 9
PMI_{mod} scores for Stetson and White Diamonds.

Brand personality traits	Stetson, percent	White Diamonds, percent	Personality facets	Stetson, percent	White Diamonds, percent
Down-To-Earth	2.53	0.20	Down-To-Earth	1.85	0.15
Family-Oriented	0.10	0.03			
Small-Town	2.91	0.22			
Honest	5.04	1.11	Honest	5.38	2.62
Sincere	0.99	0.53			
Real	10.11	6.22			
Wholesome	0.64	0.25	Wholesome	14.04	16.02
Original	27.44	31.79			
Cheerful	1.49	0.63	Cheerful	1.88	1.80
Sentimental	1.28	0.30			
Friendly	2.88	4.48			
Daring	3.13	1.41	Daring	4.84	4.72
Trendy	3.50	2.91			
Exciting	7.90	9.85			
Spirited	4.47	1.36	Spirited	14.29	21.20
Cool	27.95	56.63			
Young	10.45	5.61			
Imaginative	0.55	0.18	Imaginative	5.65	4.56
Unique	10.74	8.93			
Up-To-Date	3.81	3.01	Up-To-Date	4.51	3.52
Independent	4.79	3.27			
Contemporary	4.93	4.28			
Reliable	4.30	2.63	Reliable	3.10	2.94
Hardworking	0.24	0.08			
Secure	4.76	6.12			
Intelligent	4.18	3.66	Intelligent	20.11	5.51
Technical	3.99	3.19			
Corporate	52.16	9.69			
Successful	5.51	3.96	Successful	5.50	3.60
Leader	7.27	4.98			
Confident	3.73	1.85			
Upper Class	0.17	0.07	Upper Class	1.36	2.25
Glamorous	2.90	6.48			
Good Looking	1	0.21			
Charming	5.41	3.57	Charming	10.06	18.11
Feminine	16.99	44.85			
Smooth	7.78	5.92			
Outdoorsy	1.13	0.44	Outdoorsy	8.22	9.01
Masculine	13.30	22.40			
Western	10.23	4.20			
Tough	5.52	2.48	Tough	6.80	4.07
Rugged	8.07	5.66			

for Stetson and 19,600 for White Diamonds. Based on Aaker's conceptual model, we calculated the PMI_{mod} scores for each of the 42 traits and averaged these to find the PMI_{mod} scores for each of the 15 facets of personality. Trait-level and facet-level PMI_{mod} scores are reported in Table 9. Finally, the 15 facets were averaged to arrive at the five personality dimensions for both brands. The results for two brands are reported in Table 10.

The results seem consistent with expectations. Compared to White Diamonds, Stetson seems to portray a very different brand personality. Stetson is perceived as more sincere, competent, and rugged, whereas White Diamonds is viewed as more exciting and sophisticated. This is consistent with the positioning of these brands as well as their images in the "real" world. Although some of the differences in percentages are not large, they reflect differences across the hundreds of thousands of web pages in the

Google database so that small differences in percentage reflect a large number of pages. Still, it is clear that some differences in the positioning are stronger than others. The brands vary a great deal on "competence" and "sophistication" and very little on "sincerity." Also note that we do not present any statistical tests

Table 10
Brand personality scores for Stetson and White Diamonds.

Personality dimension	Stetson, percent	White Diamonds, percent
Sincerity	5.79	5.15
Excitement	7.32	8.50
Competence	9.57	4.02
Sophistication	5.71	10.18
Ruggedness	7.51	6.54

Note: The higher the score, the more often the brand is associated with the traits for a given dimension.

for the observed differences because the results are obtained for the entire “universe” of data instead of for a sample from that universe.

It is also worth noting that on some of the individual traits, the percentages seem contrary to expectations. For example, on the “masculine” trait, White Diamonds seems to have a much higher percentage of hits than Stetson. As we average across all the traits, Stetson appears to be more “rugged” than White Diamonds but it is worth noting that on some individual traits, the percentages seem contrary to expectations. This could be because of genuinely poor positioning or because of the page-level analysis which does not distinguish between “masculine” and “not at all masculine.” The results do, however, point to the importance of drawing conclusions based on multiple descriptors.

Study 2 discussion

Study 2 demonstrates another application of delving into massive online textual databases to derive insights that are relevant to marketing decision-makers. It shows that an analysis of hit ratios can be used to derive meaning from the textual content of these databases. The analysis of hit ratios using the PMI_{mod} algorithm seems to highlight brand personalities that match expectations based on the long-term positioning efforts of the brands’ marketers. Although online brand–descriptor relationships do not necessarily have to match “real world” relationships, such consistency does contribute to the face validity of this analysis. Also note that one does not have to compare two or more brands to derive meaningful conclusions pertaining to a brand personality. One can, instead, do a longitudinal analysis of a brand’s personality to examine changes (intended or unintended) in a brand’s personality over time. Using a before-and-after design, one can also examine the effect of an advertising campaign on a brand’s performance on a particular attribute. For example, suppose the cruise industry suffers from an image of offering vacation alternatives that are generally considered to be very expensive. One of the bigger players in the industry (say Carnival) launches an advertising campaign to dispel the myth that cruise vacations are more expensive than land vacations. One can do an analysis similar to the one reported in Study 2 to infer the effectiveness of this campaign. Carnival can determine associations between its brand and descriptors such as “affordable,” “reasonable,” and “inexpensive” before the ad campaign starts and do another study after the campaign ends and compare the results.

By now, it is evident that the quality of our analysis depends on the choice of descriptors included in the study. While the choice may be simple and straightforward in many cases (such as those reported in Studies 1 and 2), it may not be so easy and direct in others. In order to make our approach less sensitive to word choice, we propose that single descriptors be replaced by a set of synonyms for that descriptor and an average of all synonyms be used for analysis. Study 3 reports an application illustrating this approach, and an attempt to validate the results using a published, external comparison.

Study 3: Attribute-level brand comparisons

In study 3, we compare two brands on a given attribute. Specifically, we compare brand pairs on “reliability” for several home appliances. Brand pairs were chosen based on the results reported in Consumer Reports’ *2004 Buying Guide*. Consumer Reports’ results were based on a national survey of tens of thousands of responses it received for its annual questionnaire about products bought in the previous five years or so (Consumer Reports 2004). Thus, the Consumer Reports listing of brand reliabilities is based on a long-term assessment of consumers’ actual experiences with brands, and should be somewhat stable over time. Of course, it is possible that the online content may be more reflective of the current situation than long-term performance, but there should be a relatively clear-cut difference between the most and least reliable brands based on Consumer Reports’ long-term assessment of actual consumer experiences. To counter the effects of the time delay and possible shifts in brand positions, we chose brand pairs that exhibited significant contrast on reliability.

The product categories included in this study were gas ranges, dishwashers, digital cameras, microwave ovens, televisions, vacuum cleaners, refrigerators, and camcorders. We expanded our analysis to include multiple adjectives to describe the factor of interest (reliability). The list of adjectives for “reliability” was developed using a thesaurus and WordNet (Fellbaum 1998) and is included in Table 11. We also introduced a refinement in Study 3 by adding another filter to our search. In addition to the product category filter (such as “refrigerator”), we excluded pages that had both the brands mentioned together in a document. Thus, the number of page hits in the denominator reflected only those pages that were relevant to the product category, had a mention of the focus brand (and not the comparison brand), and had a given descriptor. The results are reported in Table 11.

Results for each product category are organized in three columns. The first column reports the PMI_{mod} for the fourteen synonyms of the term “reliable” for the brand that performs better on reliability as per the Consumer Reports survey. The second column reports the same for the brand that ranks lower on reliability. The third column reports the inter-brand difference. A positive difference indicates that our results are consistent with those reported by Consumer Reports. We also provide a summary statistic for each product category by adding up the differences on all fourteen descriptors. Of the eight categories included in our study, our results are consistent with Consumer Reports results for seven categories. A closer examination of the television category (for which our results were opposite to those reported in CR) revealed that the term “television” was used in such a diverse set of conversations that our filter was not effective at “cleaning” the obtained hits, and limiting the search findings to pages of most relevance to our analysis.

Study 3 discussion

The results of this study lend further credence to the value of the proposed method of analyzing large text-based corpora. By analyzing the co-occurrence of brands with descriptors that

Table 11
Brand reliability comparisons: PMI_{mod} scores for home appliances.

Descriptor	Gas range			Microwave ovens			Refrigerators		
	Hotpoint (A), percent	Amana (B), percent	(A – B), percent	Amana (A), percent	Sharp (B), percent	(A – B), percent	Kenmore (A), percent	Maytag (B), percent	(A – B), percent
Authentic	1.12	5.81	-4.69	14.20	2.67	11.53	4.19	4.36	-0.17
Certain	2.85	4.32	-1.47	9.33	9.84	-0.51	7.63	4.57	3.06
Consistent	0.78	1.42	-0.64	4.34	3.89	0.45	3.13	2.23	0.90
Dependable	1.19	6.88	-5.69	14.05	1.36	12.69	4.67	2.81	1.86
Durable	1.88	7.25	-5.37	14.80	4.05	10.75	3.73	3.47	0.26
Genuine	2.68	1.85	0.83	4.12	3.04	1.08	4.42	2.76	1.66
Reliable	15.93	7.83	8.10	17.46	6.94	10.52	9.06	4.24	4.82
Sound	20.18	5.95	14.23	18.57	17.87	0.70	12.99	7.98	5.01
Steady	2.38	1.13	1.25	2	3.95	-1.95	2.58	1.77	0.81
Straight	12.21	4.29	7.92	5.95	8.06	-2.11	6.47	3.27	3.20
Sure	15.40	4.20	11.20	13.33	18.04	-4.71	13.31	12.10	1.21
Tested	3.34	2.17	1.17	4.61	5.36	-0.75	6.79	3.70	3.09
Tried	2.04	1.87	0.17	3.60	8.40	-4.80	4.11	2.55	1.56
True	6.96	5.48	1.48	10.50	11.21	-0.71	14.57	6.54	8.03
			28.49			32.18			35.30
Descriptor	Digital camera			Television sets			Camcorders		
	Sony (A), percent	Epson (B), percent	(A – B), percent	Sanyo (A), percent	RCA (B), percent	(A – B), percent	Sony (A), percent	JVC (B), percent	(A – B), percent
Authentic	0.67	0.43	0.24	0.57	0.90	-0.33	0.29	0.59	-0.30
Certain	4.20	2.66	1.54	3.26	4.96	-1.70	1.37	0.99	0.38
Consistent	0.78	0.89	-0.11	0.99	1.32	-0.33	0.29	0.43	-0.14
Dependable	0.27	0.38	-0.11	0.29	0.35	-0.06	0.16	0.28	-0.12
Durable	1.53	1.12	0.41	1.55	1.29	0.26	0.81	0.62	0.19
Genuine	1.90	2.12	-0.22	1.85	1.47	0.38	1.02	0.70	0.32
Reliable	2.78	2.38	0.40	2.15	2.71	-0.56	1.46	1.11	0.35
Sound	42.81	10.57	32.24	34.71	37.13	-2.42	12.29	7.88	4.41
Steady	1.35	0.54	0.81	0.97	1.51	-0.54	0.94	0.46	0.48
Straight	3.23	1.85	1.38	2.98	4.25	-1.27	1.07	0.79	0.28
Sure	13.11	5.02	8.09	10.26	10.34	-0.08	3.47	1.81	1.66
Tested	2.50	2.02	0.48	1.70	3.77	-2.07	1.18	0.90	0.28
Tried	3.43	2.27	1.16	1.76	4.09	-2.33	1.23	0.85	0.38
True	10.60	4.94	5.66	6.77	10.58	-3.81	2.89	1.21	1.68
			51.97			-14.86			9.85
Descriptor	Washing machines			Vacuum cleaners					
	Kenmore (A), percent	Maytag (B), percent	(A – B), percent	Kenmore (A), percent	Dirt Devil (B), percent	(A – B), percent			
Authentic	18.78	2.44	16.34	6.54	3.85	2.69			
Certain	25.40	16.64	8.76	22.36	13.34	9.02			
Consistent	4.24	3.54	0.70	9.84	13.10	-3.26			
Dependable	16.19	13.54	2.65	23.46	4.68	18.78			
Durable	4.46	10.61	-6.15	15.59	20.37	-4.78			
Genuine	14.39	3.79	10.60	21.18	25.48	-4.30			
Reliable	21.37	9.75	11.62	25.75	14.65	11.10			
Sound	52.59	29.77	22.82	33.54	17.65	15.89			
Steady	7.41	3.36	4.05	8.74	8.48	0.26			
Straight	28.49	12.85	15.64	15.91	21.76	-5.85			
Sure	39.06	17.07	21.99	34.72	26.47	8.25			
Tested	9.21	7.46	1.75	18.03	20.78	-2.75			
Tried	17.99	7.99	10.00	23.07	15.51	7.56			
True	29.93	13.46	16.47	39.37	26.04	13.33			
			137.24			65.94			

Note: A positive score for (A – B) indicates that brand A is considered as more reliable than brand B. Consumer Reports lists brand A as more reliable than B in all cases.

have a defined semantic orientation, we can better understand the semantic orientation of the textual content surrounding brands in massive online databases of text-based content. One can decrease dependence on a specific descriptor term by replacing it with a set of synonyms that people might use in place of the descriptor, as they express their opinions and experiences in their online postings. While replacing a single descriptor term by a set of synonyms greatly increases the computational requirements of the analysis, automated software queries of the database can keep the task manageable. Another advantage of using a set of descriptors instead of a single descriptor is that some descriptors seem to be more popularly used for certain product categories than others. For example, descriptors such as “dependable” and “consistent” were heavily used for products such as vacuum cleaners and sparsely used for digital cameras and camcorders. By using multiple descriptors, one can capture the broader contextual domain of the core attribute.

The results of Study 3 demonstrate that the pattern of hit ratios for brand reliability is consistent with objective data on consumer experiences collected from an independent source (Consumer Reports). Still, there is reason to be somewhat cautious in using these results, and the potential for significant research and additional refinement of these techniques remains.

Discussion

This paper demonstrates that massive search engine databases like Google can be successfully mined to provide information of interest to online retailers. The three studies reported here are based on lexical semantic analysis which yields valuable insights into online brand representations that are of considerable value for making strategic marketing decisions.

Our approach offers a promising potential for analytical examination of a freely available database. Because the web provides an impressive record of brand-related communications and is dominated by user-generated content, it can be used by marketers to get a deeper understanding of consumer relationships with brands and how consumers evaluate brands.

We show that by examining associations between brands and carefully selected descriptors, one can go beyond merely counting text content to uncovering the relational meaning of the content. The content reveals relationships that can serve as supplemental evidence that informs marketing decision-making, especially in contexts involving comparisons.

In the three studies reported in our paper, we validated our results by comparing them to the “real world.” For the brand positioning study, we examined the positioning of laundry detergents with respect to the image promoted by the company. For brand personality analysis, we compared our findings to our general understanding of the brands’ persona in real life. In the final study, we compared our findings to the results reported in Consumer Reports. While our results were validated by such external comparisons, we believe that it is not always necessary to make such comparisons. If a brand’s online persona fails to match its real-life counterpart, it is still useful information for the manager of that brand. In fact, we are likely to observe a divergence between the two every time a company is ineffective

in positioning its brand. Thus, what gets said online is interesting and relevant, in and of itself. By this argument, the fact that online relationships did not show Sanyo TVs to be as “reliable” as RCA TVs is not a “negative” result as much as an insight into online positions worthy of additional research.

Our analysis was based on document count—every document that had a mention of the target words (e.g. “reliable”) counted as one hit. We used document count instead of frequency count as the unit of our analysis for two reasons. First, none of the search engines commonly available today allow for a frequency count *within* a document. Thus, we were limited by the search features offered by Google. Second, and more importantly, we also believe that document count is a better measure than frequency count. To the extent that different people write different web pages, a frequency count will give more weight to the opinions of an individual who repeats himself more often in a review. Document count has the benefit of smoothing out such effects. Also, research done by Pang, Lee, and Vaithyanathan (2002) shows that the frequency count method does not yield any performance improvement over the document count method.

Our method does not account for negative qualifiers (called “negation tags” in the literature). Thus, for example, our algorithm does not discriminate between “reliable” and “not reliable.” Appearance of either will count as a hit, despite their opposite meaning. Also, our approach does not account for the polarity of a statement. (Polarity refers to the strength of an argument.) For example, “very reliable” is a stronger statement than “reliable” or “somewhat reliable.” While it makes intuitive sense that accounting for these variations will improve performance, previous research has shown that the effects are marginal. Pang et al. (2002) show that the inclusion of negative qualifiers leads to a marginal improvement in an algorithm’s performance. Interestingly, Dave, Lawrence, and Pennock (2003) reported a slight decline in performance after inclusion of negative qualifiers. Similarly, assigning different weights to different qualifier terms has proven to be largely useless in previous research (Dave et al. 2003).

The approach outlined in this paper is a starting point for a stream of research on using web-based information to monitor and evaluate brands and products. Methods similar to those described in this paper can be developed to classify product reviews as positive or negative. Marketing managers and consumers can access the “wisdom of the web” to get an overall evaluation of products and brands on any number of dimensions. Managers may also be able to monitor their online brand positions by evaluating the strength and tone of online content related to their brands. Given the manner in which web-based data are classified, it may be possible to make comparisons of brand positions across countries and continents. The dynamic nature of the web also allows for the longitudinal tracking of brand positions over time. It may be useful for managers to track shifts in brand positions to evaluate strategic changes and repositioning needs.

In the last decade, we have seen several attempts at identifying the dimensions of service quality as it pertains to online environments. Cox and Dale (2001) noted that the traditional service quality dimensions such as competence, courtesy, and

cleanliness may not be applicable to online retailers. Madu and Madu (2002) observed that additional dimensions such as security and system integrity are integral to online service quality. Online retailers now face a variety of multidimensional service quality frameworks offered by different researchers. Zeithaml, Parasuraman, and Malhotra (2001) have identified eleven dimensions of online service quality, Wolfinbarger and Gilly (2003) proposed four dimensions, and Yang and Jun (2002) uncovered six service quality dimensions. After a comprehensive review of extant literature, Zeithaml, Parasuraman, and Malhotra (2002) identified seven dimensions for e-service quality—reliability, privacy, efficiency, fulfillment, responsiveness, contact, and compensation. The method proposed in this paper can easily be adapted to shed more light on this issue by conducting a semantic analysis of descriptors most frequently employed in discussions of online service quality.

Google has not revealed its protocol for searching its database to retrieve the number of hits. If Google were to modify its search engine logic, then longitudinal comparisons could become more difficult. One possible way of overcoming this problem is to download the first few thousand hits for each search and then conduct the analysis locally without using Google APIs. While this task could be onerous, it is likely to yield more reliable results.

The analysis can go beyond an examination of product reviews that are selectively placed next to purchase information (e.g., Chatterjee 2001), which, although valuable, are limited in scope. The insights from such an analysis of meaning in online databases can be used to potentially develop more efficient product recommender systems. Given recent research that product recommenders are particularly effective in online retailing (Senecal and Nantel 2004), the potential for these techniques to enhance online retailing is significant.

Conclusions and limitations

The untamed jungle of the textual content on the web, coupled with the ease and frequency with which people can add content and the automated web crawlers that are commonplace today, provides a rich and detailed source of content for consumer researchers looking to better understand consumers' relationships with products and brands. We present a significant first step at making sense of this content for marketing analysis, while encouraging and inviting other scholars to generate a stream of research to make this analysis more sophisticated.

Of course, there is a need for caution in using insights developed from search engine database analysis. The indiscriminate indexing of sites makes it harder to separate the wheat from the chaff. It is important for researchers to carefully develop models that are able to discriminate between the information of value and the "error" in the data. The signal-to-noise ratio in such databases is likely to be low. The user of the methods described in our studies must have a hypothesis or a research question to begin with ("Is Volvo seen as a safer car brand compared to Toyota?"). These methods are not suited for exploratory analysis where a user may want to develop a brand profile without first identifying the dimensions of that profile. For this reason,

the descriptors used to label a dimension ought to be carefully selected. Where possible, such descriptors should be supplemented with their synonyms to capture the full spectrum of a descriptor's meaning and usage. Related to this is the significant issue of measure refinement. Typically, search engines return "hits" as the output from any query. In this study, we used hit information and filter variables to draw conclusions about brand positions, and to assign meaning to the hit information. However, raw "hits" are generally a crude measure of semantic content. Additional research is underway to refine the analysis by using the "brand-descriptor-relationship" algorithms to parse through pages of content, and to identify links between brands and descriptors at the individual sentence level. This will allow us to focus on the consumer-generated content of blogs or product review sites, as well as to draw more meaningful conclusions on consumer perceptions of brands in the online space.

Finally, the textual corpora continue to change rapidly with the addition of hundreds of thousands of pages on a daily basis. Thus, any brand persona generated through the methods discussed in this paper will change and evolve over time. As search engines add more web pages to their database, the search results will change as well. We believe it will be interesting for marketers to do longitudinal tracking of their brands to capture the spirit of the evolving online dialogue about their brands.

Acknowledgements

The authors would like to thank Aleksey Cherfas for his significant assistance with the programming for this project. We also gratefully acknowledge the financial support of the UMD Chancellor's Grant program and additional support from CRITO/Project POINT.

References

- Aaker, Jennifer L. (1997), "Dimensions of Brand Personality," *Journal of Marketing Research*, 34 (3), 347–56.
- Ailawadi, Kusum L. and Kevin Lane Keller (2004), "Understanding Retail Branding: Conceptual Insights and Research Priorities," *Journal of Retailing*, 80 (4), 331–42.
- Arnold, Stephen J., Robert V. Kozinets and Jay M. Handelman (2001), "Added Hometown Ideology and Retailer Legitimation: The Institutional Semiotics of Wal-Mart Flyers," *Journal of Retailing*, 77 (2), 243–71.
- Banfield, Ann (1982), "Unspeakable Sentences," Boston, MA: Routledge and Kegan Paul.
- Berners-Lee, Tim, James Hendler and Ora Lassila (2001), "The Semantic Web," *Scientific American*, 284 (5), 34–43.
- Bellizzi, Joseph A., Harry F. Krueckeberg, John R. Hamilton and Warren S. Martin (1981), "Consumer Perceptions of National, Private, and Generic Brands," *Journal of Retailing*, 57 (4), 56–71.
- Blankson, Charles and Stavros P. Kalafatis (2004), "The Development and Validation of a Scale Measuring Consumer/Customer-Derived Generic Typology of Positioning Strategies," *Journal of Marketing Management*, 20 (1–2), 5–43.
- Bruce, Rebecca F. and Janyce M. Wiebe (1999), "Recognizing Subjectivity: A Case Study of Manual Tagging," *Natural Language Engineering*, 5, 187–205.
- Chatterjee, Patrali (2001), "Online Reviews: So Consumers Use Them?," *Advances in Consumer Research*, 28 (1), 129–33.

- Church, Kenneth W. and Patrick Hanks (1989), "Word Association Norms, Mutual Information and Lexicography," in *Proceedings of the 27th Annual Conference of the ACL*, New Brunswick, NJ: ACL, 76–83.
- Consumer Reports (2004), "Buying Guide 2004," Yonkers, NY: Consumers Union.
- Cox, J. and Barrie G. Dale (2001), "Service Quality and E-Commerce: An Exploratory Analysis," *Managing Service Quality*, 11 (2), 121–3.
- Dave, Kushal, Steve Lawrence and David M. Pennock (2003), "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," in *Proceedings of the 12th international conference on World Wide Web*.
- Fellbaum, Christiane D. (1998), "WordNet: An Electronic Lexical Database," Cambridge, MA: MIT Press.
- Firth, Raymond (1957), "A Note on Descent Groups in Polynesia," *Man*, 57, 4–8.
- Fournier, Susan (1998), "Consumers and Their Brands: Developing Relationship Theory in Consumer Research," *Journal of Consumer Research*, 24 (March), 343–7.
- Godes, David and Mayzlin Dina (2004), "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science*, 23 (4), 545–60.
- Hatzivassiloglou, Vasileios and Kathleen R. McKeown (1997), "Predicting the Semantic Orientation of Adjectives," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL*, New Brunswick, NJ: Association for Computational Linguistics, 174–81.
- Hatzivassiloglou, Vasileios and Janyce M. Wiebe (2000), "Effects of Adjective Orientation and Gradability on Sentence Subjectivity," in *Proceedings of 18th International Conference on Computational Linguistics*, New Brunswick, NJ: Association for Computational Linguistics.
- Henderson, Geraldine R., Dawn Iacobucci and Bobby J. Calder (2002), "Using Network Analysis to Understand Brands," *Advances in Consumer Research*, 29 (1), 397–405.
- Kassarjian, Harold H. (1977), "Content Analysis in Consumer Research," *Journal of Consumer Research*, 4 (1), 8–18.
- Klepachi, Laura (2004), "Diamonds for the Masses," *WWD: Women's Wear Daily* 5/21/2004, 187 (107), 8.
- Kozinets, Robert V. (2002), "The Field Behind the Screen: Using Netnography for Marketing Research in Online Communications," *Journal of Marketing Research*, 39 (1), 61–72.
- Leuf, Bo (2006), "The Semantic Web: Crafting Infrastructures for Agency," John Wiley & Sons.
- Lilien, Gary L. and Arvind Rangaswamy (2002), "Marketing Engineering: Computer-Assisted Marketing Analysis & Planning, 2/e," Upper Saddle River, NJ: Prentice Hall, Inc.
- Madu, Christian N. and Assumpta A. Madu (2002), "Dimensions of E-quality," *International Journal of Quality & Reliability Management*, 19 (3), 246–58.
- OECD (2007), "Participative Web: User-Created Content," Report of the Organisation for Economic Co-operation and Development. Retrieved from <http://www.oecd.org/dataoecd/57/14/38393115.pdf> on July 5, 2008.
- Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan (2002), "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 79–86.
- Punj, Girish and Junyeon Moon (April 2002), "Positioning Options for Achieving Brand Association: A Psychological Categorization Framework," *Journal of Business Research*, 55 (4), 275–83.
- Senecal, Sylvain and Jacques Nantel (2004), "The Influence of Online Product Recommendation on Consumers' Online Choices," *Journal of Retailing*, 80 (2), 159–6.
- Sharma, Arun, Michael Levy and Ajith Kumar (2000), "Added Knowledge Structures and Retail Sales Performance: An Empirical Examination," *Journal of Retailing*, 76 (1), 53–69.
- Shocker, Allan D. and V. Srinivasan (1979), "Multiattribute Approaches for Product Concept Evaluation and Generation: A Critical Review," *Journal of Marketing Research*, 16 (2), 159–80.
- Turney, Peter D. (2001), "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," in *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, 491–502.
- (2002), "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, 417–24.
- Voss, Glen. and Kathleen B. Seiders (2003), "Exploring the Effect of Retail Sector and Firm Characteristics on Retail Price Promotion Strategy," *Journal of Retailing*, 79 (1), 37–52.
- Warden, Clyde A., Stephen Chi-Tsun Huang, Tsung-Chi Liu and Wann-Yih Wu (2008), "Global Media, Local Metaphor: Television Shopping and Marketing-As-Relationship in America, Japan, and Taiwan," *Journal of Retailing*, 84 (1), 119–2.
- Weathers, Danny, Subhash Sharma and Stacey Wood (2007), "Effects of Online Communication Practices on Consumer Perceptions of Performance Uncertainty for Search and Experience Goods," *Journal of Retailing*, 83 (4), 393–401.
- Wiebe, Janyce (1994), "Tracking Point of View in Narrative," *Computational Linguistics*, 20 (2), 233–87.
- Wolfenbarger, Mary F. and Mary C. Gilly (2003), "eTailQ: Dimensionalizing, Measuring and Predicting E-Tail Quality," *Journal of Retailing*, 79 (3), 183–98.
- Yadav, Manjit S. and P. Rajan Varadarajan (2005), "Understanding Product Migration to the Electronic Marketplace: A Conceptual Framework," *Journal of Retailing*, 81 (2), 125–40.
- Yang, Zhilin and Minjoon Jun (2002), "Consumer Perception of E-Service Quality: From Internet Purchaser and Non-Purchaser Perspectives," *Journal of Business Strategies*, 19 (1), 19–41.
- Yang, Zhilin and Xiang Fang (2004), "Online Service Quality Dimensions and Their Relationships With Satisfaction," *International Journal of Service Industry Management*, 15 (3), 302–26.
- Zeithaml, Valarie A., A. Parasuraman, and Arvind Malhotra (2001), "A Conceptual Framework for Understanding E-Service Quality: Implications for Future Research and Managerial Practice," *MSI Working Paper Series, No. 00-115*. Cambridge, MA, 1–49.
- , A. Parasuraman and Arvind Malhotra (2002), "Service Quality Delivery through Web Sites: A Critical Review of Extant Knowledge," *Journal of the Academy of Marketing Science*, 30 (4), 362–75.