2015.11 산업공학특론 발표자료

Mining Marketing Meaning from Online Chatter:
Strategic Brand Analysis of Big Data using Latent Dirichlet Allocation

Tirunillai, Seshadri, and Gerard J. Tellis. *Journal of Marketing Research* 51.4 (2014): 463-479.

서울대학교 산업공학과

양호성, 조성준

hoseong@dm.snu.ac.kr, zoon@snu.ac.kr

0 Content

Extract dimensions of quality, valence, validity, importance, optimality, heterogeneity and dynamics of those dimensions Using LDA

0 용어 정리

Term	definition
UGC	User-generated content (Social media, product review, blog)
Valence	Expression of positive vs negative performance on a dimension or attribute and is termed "sentiment" in text-mining research.
Dimensions of quality	Latent dimensions, variables that consumers may not explicitly mention but capture or represent a large number of attributes (e.g. "Performance" dimension – attributes (the speed, power, or multitasking capabilities of a computer)
Vertically differentiated dimensions	characteristics on which all consumers agree that more is better (e.g., reliability)
Horizontally differentiated	taste dimensions on which consumers might disagree (e.g., aesthetics)

1

Introduction

The quality of a product or service is an important determinant of consumer satisfaction, brand performance

With advances in online media and technologies, customers share opinions about products

Surveys / interviews	UGC (User Generated Content)
– Limited samples – Administered periodically	– Product reviews, bulletin boards, social networks – Spontaneous, passionate, widely available, low cost, easily accessible, temporally disaggregate (days, hours, minutes), live – Based on numerous customers

1

Quality dimensions

Quality is a multi-dimensional construct.

User generated content provides a rich source of data to extract the dimensions of quality

- Traditional : obtain the latent dimensions of quality through consumer survey
- Latent dimensions are variables that consumers may not explicitly mention but represent a large number of attributes

- Examples of Mobile Phone

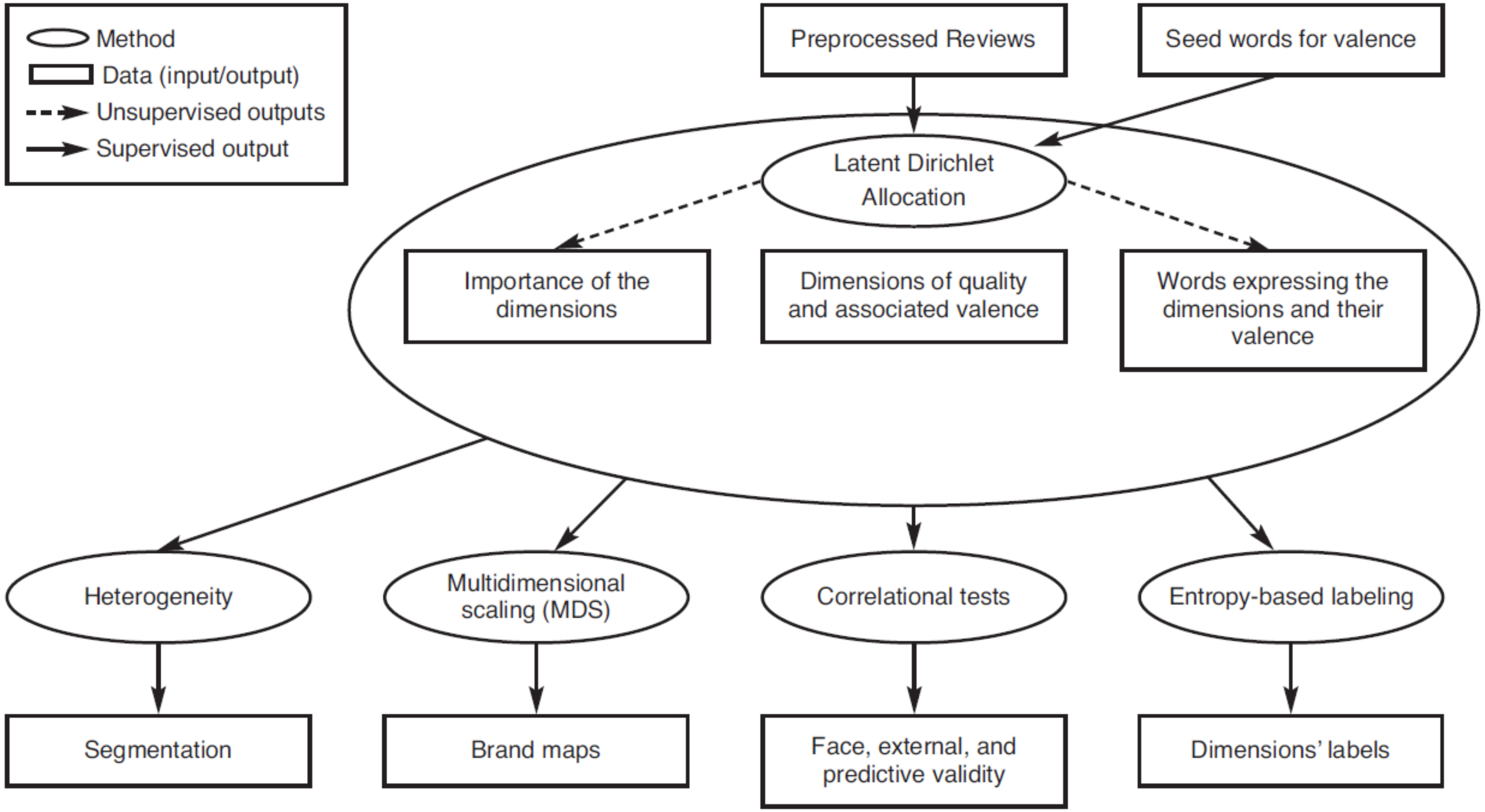
Latent Dimensions	Attributes
Performance	Speed
Portability	Smooth, Handy
Compatibility	Universial, accessible

1 Introduction

This study suggests a unified framework

UGC provides a rich source of data to extract the dimensions of quality

1) Extract valence with dimensions
2) Identify an optimum number of dimensions.
3) Label the dimensions
4) Assess the heterogeneity of the dimensions
5) Position brands on the dimensions
6) Analyze the dynamics of dimensions and brand positions over time



1 The benefits of LDA & Advantages

The benefits of LDA

- Benefits of LDA
  1) It allows for exploration of dynamics over time
  2) It allows for computation of the importance of the extracted dimensions
  3) We can use the results of LDA for further analysis to offer rich managerial insights
     - dimensions' importance over time
     - Heterogeneity
     - perceptual maps of competing brands & dynamics of these maps

- Advantages relative to previous methods
  1) Using unsupervised methods that involve little human intervention
  2) extracts valence without requiring client or rater inputs. -> minimal bias or errors

2

Methods

Sampling → Preprocessing → *Dimension and Valence Extraction*

dataset

1) 350,000 customer reviews
2) five markets, 17 Brands

Markets	Brands
personal computing	Hewlett-Packard [HP], Dell
cellular phones	Motorola, Nokia, Research in Motion Limited [RIM], Palm
footwear	Skechers USA, Timberland Company, Nike
toys	Mattel, Hasbro, LeapFrog
data storage	Seagate Technology, Western Digital Corporation, SanDisk

2 Methods

prepocessing

Sampling → **Preprocessing** → *Dimension and Valence Extraction*

1) Eliminate non-English characters and words (HTML tags, URLs, numbers, punctuation..)
2) Break into sentences
3) Part of Speech tagging to retain only adjectives, nouns, adverbs
4) Replace common negatives of words (e.g. “hardly”, “no” -> “not”)
5) Stemming (“likable, liked, liking” -> like)
6) Remove stopping words (“when, the, and, is”)
7) Remove all the words that do not appear in at least 2%

2

Methods

Sampling → Preprocessing → *Dimension and Valence Extraction*

Dimension and Valence Extraction is the primary contribution of this study.

Dimension and Valence Extraction

Challenges → intuition → Specification → estimation → labeling

- The problems of extracting dimensions of quality from reviews in traditional methods (e.g. PCA)

1) Customers use their <u>own words</u> to describe the quality of the attributes -> *curse of dimension*
2) Customers express opinions on only those dimensions that are salient to <u>their experience</u> -> *sparse representation*
3) Valence and adjectives are *context specific*. ("small" in laptop, memory)

➡ So, We introduced probabilistic topic model LDA

In LDA, topics = dimensions of product quality expressed by consumers

2 Methods

Sampling → Preprocessing → *Dimension and Valence Extraction*

Dimension and Valence Extraction is the primary contribution of this study.

Dimension and Valence Extraction

Challenges → intuition → Specification → estimation → labeling

1) Consumers express one or more dimensions of quality that they believe are worthy
2) Words that describe a dimension will co-occur across the reviews

→ Co-occurrence, statistical distribution helps us capture the latent dimension

LDA model	Human
(prior on distribution) uncovers the distribution of the latent dimensions	Choose theme
Draws of the words as a multinomial choice	Choosing the words
Compute the conditional distribution of the latent variables (dimensions) given the observed variables (words in review)	Given words in review, infer dimensions of quality

2 Methods

Sampling → Preprocessing → *Dimension and Valence Extraction*

Dimension and Valence Extraction is the primary contribution of this study.

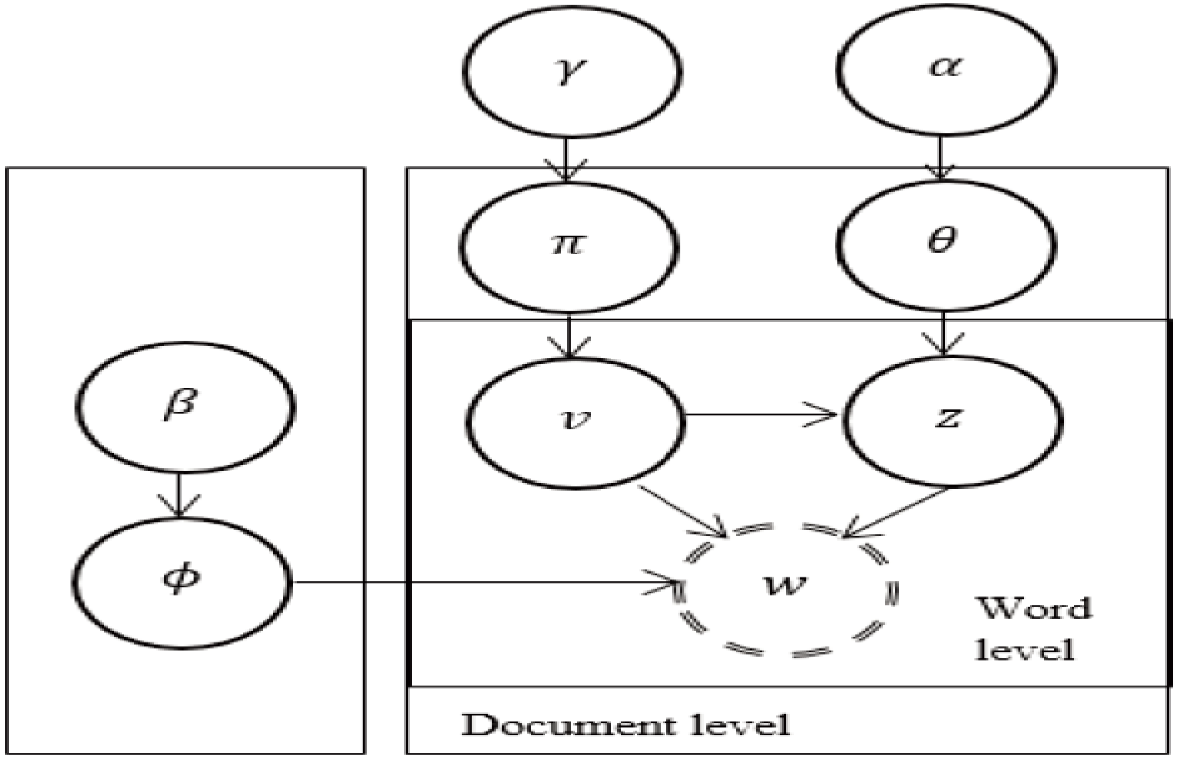
Dimension and Valence Extraction

Challenges → intuition → Specification → estimation → labeling

w : vectors of all words
z : vectors of all dimensions
v : vectors of all valence

Φ : multinomial distribution of dimension
π : proportion of valence in the review
θ : dimension's importance

α : hyperparameters on θ
β : hyperparameters on Φ
γ : hyperparameters on π



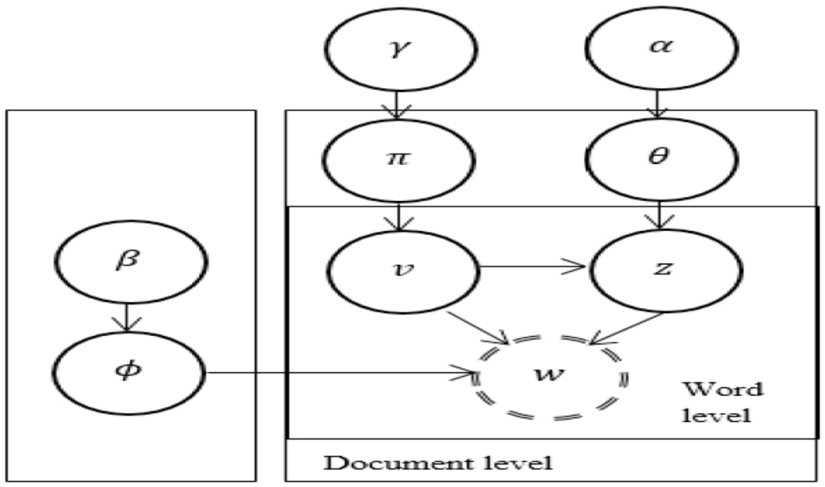
2 Methods

Sampling → Preprocessing → *Dimension and Valence Extraction*

Likelihood functions of the generative model that can be used to derive the posterior

Dimension and Valence Extraction

Challenges → intuition → Specification → estimation → labeling



w : words z : dimensions v : valence
Φ : multinomial dist of dimension
π : proportion of valence
θ : dimension's importance

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\pi}, \mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$$

$$= \prod_{n=1}^{N} p(\mathbf{w_n} \mid \mathbf{z_n}, \mathbf{v_n}, \boldsymbol{\phi}) p(\mathbf{z_n} \mid \boldsymbol{\theta}, \mathbf{v_n}) p(\mathbf{v_n} \mid \boldsymbol{\pi}) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) p(\boldsymbol{\pi} \mid \boldsymbol{\gamma}) p(\boldsymbol{\phi} \mid \boldsymbol{\beta}).$$

$$p(\mathbf{w} \mid \alpha, \beta, \gamma) = \iiint p(\phi \mid \beta) p(\theta \mid \alpha) p(\pi \mid \gamma) \times \prod_{n=1}^{N} p(\mathbf{w_n} \mid \phi_{z,v}) p(\mathbf{z_n} \mid \theta) p(\mathbf{v_n} \mid \pi) \, d\phi \, d\theta \, d\pi.$$

$$= \iiint p(\phi \mid \beta) p(\theta \mid \alpha) p(\pi \mid \gamma) \times \prod_{n=1}^{N} p(\mathbf{w_n} \mid \theta, \pi, \phi) \, d\phi \, d\theta \, d\pi.$$

Dimension and Valence Extraction is the primary contribution of this study.

Dimension and Valence Extraction

Challenges → intuition → Specification → estimation → labeling

$$P(\mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{v} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{\mathrm{p}(\mathrm{z}, \mathrm{w}, \phi, \theta, \pi, \mathrm{v}, \alpha, \beta, \gamma)}{\mathrm{p}(\mathrm{w} \mid \phi, \theta, \pi, \alpha, \beta, \gamma)}.$$

joint probability distribution of all the variables

marginal probability distribution
(the probability of observing the review corpus given any parameters of the latent model)

- Infer the distribution of the latent dimensions in a review(θ)
  & distribution of the words in a dimension(Φ)
- Directly estimating Φ, θ can be unreliable

θ, Φ, π are estimated by Gibbs sampling

2 Methods

Sampling → Preprocessing → *Dimension and Valence Extraction*

Identify the valence of the words associated with the dimension in conjunction with the identification of dimension

Dimension and Valence Extraction

Challenges → intuition → Specification → estimation → labeling

1) Use an initial set of seed words
   unambiguously positive or negative (good, great), (bad, horrible)

1) Probability of the valence of the newly encountered words -> probability of their co-occurrence with the initial seed word

2) newly classified words -> appended to the list of the seed words -> next iteration

3) repeated until the entire vocabulary of words in the reviews is classified on the basis of the valence

2

Methods

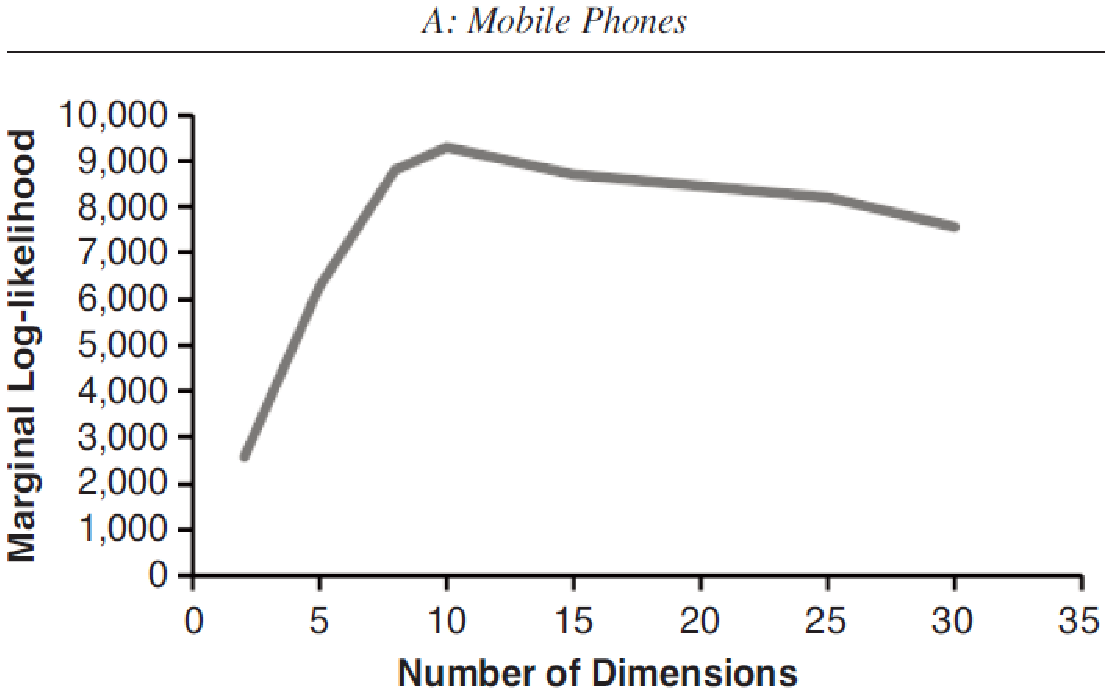
Sampling → Preprocessing → *Dimension and Valence Extraction*

Dimension and Valence Extraction is the primary contribution of this study.

Dimension and Valence Extraction

Challenges → intuition → Specification → estimation → labeling

- Selection of optimal number of dimensions
  - Marginal log-likelihood with 5 fold cross-validation
  - Harmonic mean estimator
  - First extracting two dimensions and then increase the number of dimensions until the log-likelihood reaches a *maximum*



2 Methods

Sampling → Preprocessing → *Dimension and Valence Extraction*

Dimension and Valence Extraction is the primary contribution of this study.

Dimension and Valence Extraction

Challenges → intuition → Specification → estimation → labeling

1) Select words that better distinguish the reviews associated with that dimension
2) Assign a label to the given dimension

- Mutual information
  reduction in the amount of *uncertainty* associated with a dimension due to a given word
  (MI ↑ -> greater contribution)

- MI is measured by *entropy*

$$E(k) = -\sum_{\ell=0}^{1} P(\eta=\ell)\log_2 P(\eta=\ell).$$

P : 무작위로 선택한 review 가 topic k 에 의해 생성되었을 확률
η : review discusses the k-th dimension

$$MI(k|w) = E(k) - E(k|w) \geq 0 \ \forall (k, w).$$

E(k) : entropy of given dimension
(모든 reviews 가 하나의 dimension에서 생성되면 최소값)

3 Result – Extracted dimensions of quality

Table 1 lists the words with highest MI score relating to each dimension. These words help label dimension.

Table 1

DIMENSIONS OF QUALITY FOR MOTOROLA (MOBILE PHONES, QUARTER 4, 2008)

Instability (Negative)	*Portability (Positive)*	*Receptivity*[a] *(Positive)*	*Compatibility (Positive)*	*Discomfort*[b] *(Negative)*	*Secondary Features (Positive)*
Unstable	Smooth	Dependable	Universal	Cramp	Feature
Error	Handy	Reception	Expandable	Big	App
Crash	Portable	Sharp	Supported	Layout	Card
Freeze	Small	Quick	Compatible	Finger	Camera
Reboot	Compact	Crisp	Accessible	Heavy	Wi-Fi

[a]Refers to mobile phone signal.
[b]Refers to discomfort regarding the mobile phone's physical layout.
Notes: The table shows the words with the top MI scores.

- Limitations : for some dimensions, labeling may not convey the words' meaning in its entirety
- delete dimensions about the retailer

4 Validation

Check the validity

2 Human Raters

Randomly selected 100 reviews -> read -> select set of dimensions and associated valence
Fleiss' kappa coefficient 이용 human 과 model 의 agreement 계산

Consumer Reports

Consumer reports : magazine that evaluate brands deemed important by expert testers of the products

Assess the overlap of the dimensions extracted from the automated analysis with that of the dimensions used for rating the brands in the markets

4 Validation

Face validity with human raters

Fleiss' kappa κ : measures the interrater agreement

Average κ = 0.59

Market-specific	κ
Mobile Phones	60%
Computers	62%
data storage	57%
Toys	53%
footwear	61%

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

$$\bar{P}_e = \sum_{k=1}^{K} p_k^2$$

p_k: proportion of reviews that the raters assigned to a given dimension

Kappa Statistic	*Strength of Agreement*
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

Landis, J. Richard, and Gary G. Koch. "The measurement of observer agreement for categorical data." *biometrics* (1977): 159-174.

4 Validation

External validity with Consumer Reports

Use Jaccard coefficient to test the degree of overlap between the dimensions evaluated in *Consumer Reports* and *automated analysis*

Jaccard coefficient

$$JC = \frac{\left|N\left(Dim_{lda} \cap Dim_{CR}\right)\right|}{\left|N\left(Dim_{lda} \cup Dim_{CR}\right)\right|}.$$

Markets	Average JC
Mobile Phones	0.65
Computers	0.72
data storage	0.81

Table 3

COMPARISON OF *CONSUMER REPORTS* AND AUTOMATED ANALYSIS

A: Mobile Phones, 2009		
Dimension	*Automated Method*	Consumer Reports
Ease of use (e.g., voice commands, navigation)	✓	✓
Performance (e.g., voice clarity, sensitivity)	✓	✓
Messaging	✓	X
Exhaustibility (battery)	✓	X
Layout discomfort	✓	X
Secondary features (e.g., camera, music player)	✓	✓
Compatibility (e.g., Bluetooth, headphones)	✓	✓

4 Heterogeneity of dimensions

Introduce Herfindahl index which is a measure of the size of firms in relation to the industry and an indicator of the amount of competition among them

Herfindahl index↑ -> generally competition ↓

Herfindahl index $H = \sum_{i=1}^{n} \alpha^2$, $\alpha = \frac{\text{Total number of reviews citing the dimension}}{\text{Total number of reviews of the brand}}$.

-> inverse measure of the diversity or heterogeneity of the dimensions by reviewers
즉, H 값이 크면 diversity 가 적다. 특정 dimension 에 집중되어 있다.

Vertically differentiated markets

H relatively high

Horizontally differentiated markets

H relatively low

Mobile phone, computers have objective dimensions

little heterogeneity across dimensions

Market, Brand	*Herfindahl Index of Concentration*	*Heterogeneity in Dimensions*	*Instability of Herfindahl Index over Time (%)*
Mobile Phones			
Nokia	45.78	Low	3.3
RIM	54.12	Low	3.5
Palm	43.58	Low	2.3
Motorola	48.18	Low	2.1
Computers			
Dell	24.80	Low	1.4
HP	31.68	Low	2.7
Toys			
Hasbro	12.82	Moderate	4.9
Mattel	11.64	High	5.4
LeapFrog	13.58	High	7.6
Footwear			
Timberland	25.74	Moderate	5.1
Skechers	21.52	Moderate	7.4
Nike	23.82	Moderate	8.9
Data Storage			
Seagate	52.44	Moderate	4.8
Western Digital	44.86	Low	3.6
Sandisk	61.02	Low	3.8

4 Stability of Heterogeneity of Dimensions over Time

Calculate the percentage instability ini Herfindahl index of the dimension over time

In Table 5, do not suggest a significant change in the dimension over the time

$$V_t = \Delta H_t + 2\left[H_{t-1} - \rho\sigma_t\sigma_{t-1} - \frac{1}{n}\right],$$

H_t: Herfindahl index at time t (week)

ρ : correlation between % share of consumers citing the dimension

σ_t : standard deviation of shares of dimensions at t

n : total number of dimensions at t

Vertically differentiated markets — 1~4 %, stable over time

Horizontally differentiated markets — 4~8 %, unstable

Table 5

SPLIT-SAMPLE TEST FOR ROBUSTNESS OF THE STABILITY OF THE DIMENSIONS

Market, Brand	*Instability of Herfindahl Index over Time (%) Sample 2005–2007*	*Instability of Herfindahl Index over Time (%) Sample 2008–2009*
Mobile Phones		
Nokia	3.1	3.5
RIM	3.4	3.7
Palm	2.4	2.6
Motorola	1.8	2.4
Computers		
Dell	1.5	1.8
HP	2.8	2.5
Toys		
Hasbro	5.1	4.7
Mattel	5.2	5.4
LeapFrog	7.8	7.5
Footwear		
Timberland	5.3	5.0
Skechers	7.6	7.3
Nike	8.6	8.5
Data Storage		
Seagate	4.6	5.1
Western Digital	3.8	3.2
Sandisk	3.6	3.9

5 Brand Mapping

Positioning of competing brands in a market on the basis of their location in space defined b the key dimensions

- Distance measure : Hellinger distance

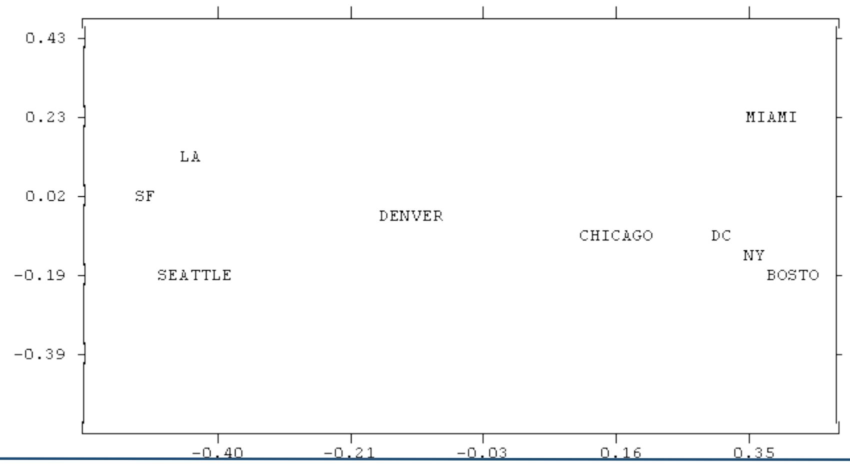
$$f\left(\theta_k^a, \theta_k^b\right) = \left[\frac{1}{2}\int\left(\sqrt{\frac{dA}{dx}} - \sqrt{\frac{dB}{dx}}\right)^2 dx\right]^{\frac{1}{2}}.$$ (continuous)

$$f\left(\theta_k^a, \theta_k^b\right) = \left[\frac{1}{2}\sum_k\left(\sqrt{\theta_k^a} - \sqrt{\theta_k^b}\right)^2\right]^{\frac{1}{2}}.$$ (discrete)

Derive *similarity matrix* using Helliger distance

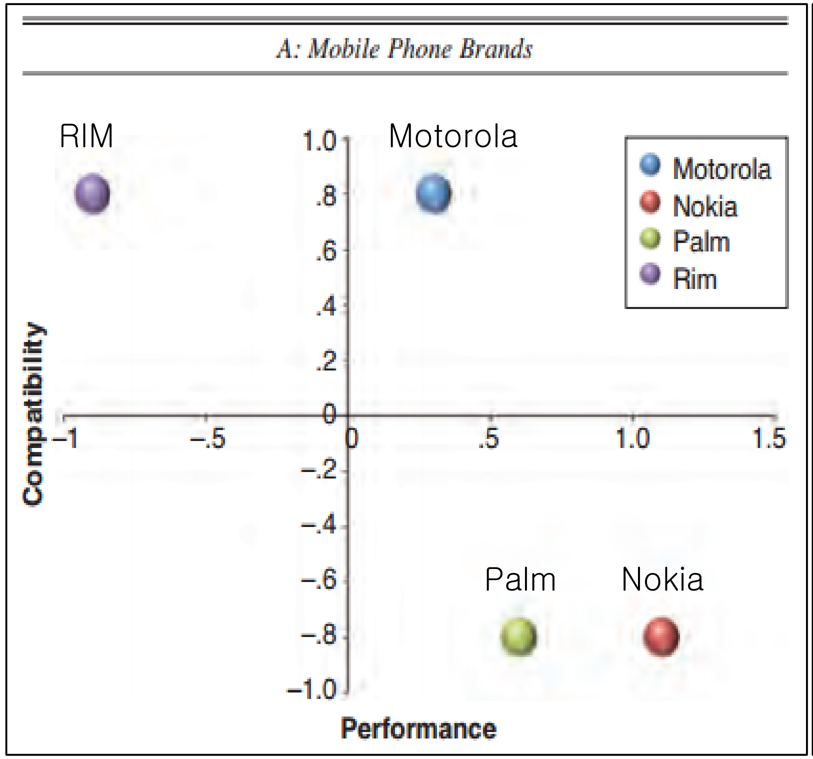
Visualize using Multidimensional scaling(MDS)

		1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0

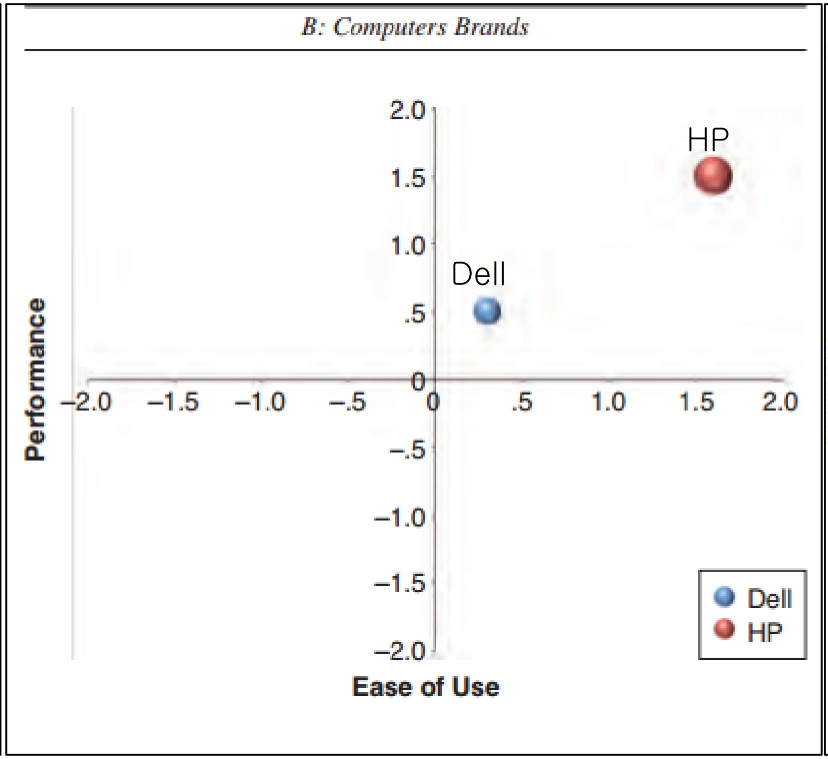


5 Static Brand Mapping

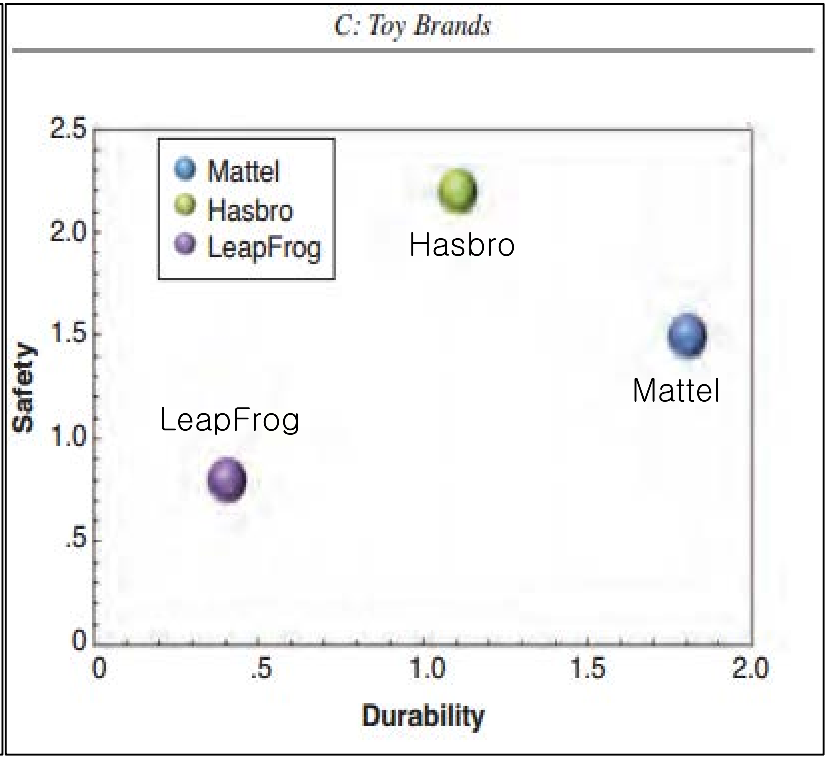
Choose top 2 dimensions on the basis of the frequency of occurrence of these dimensions across all the reviews



Performance : Motorola > Palm



Performance & Ease of use : HP > Dell

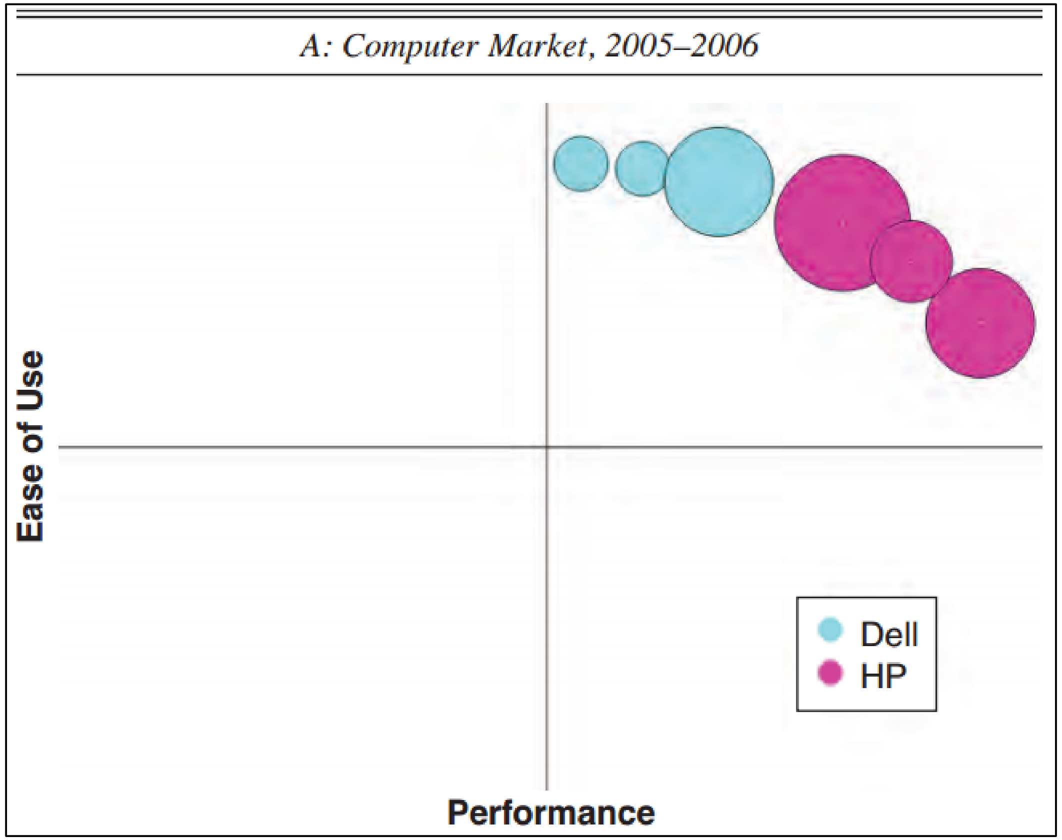


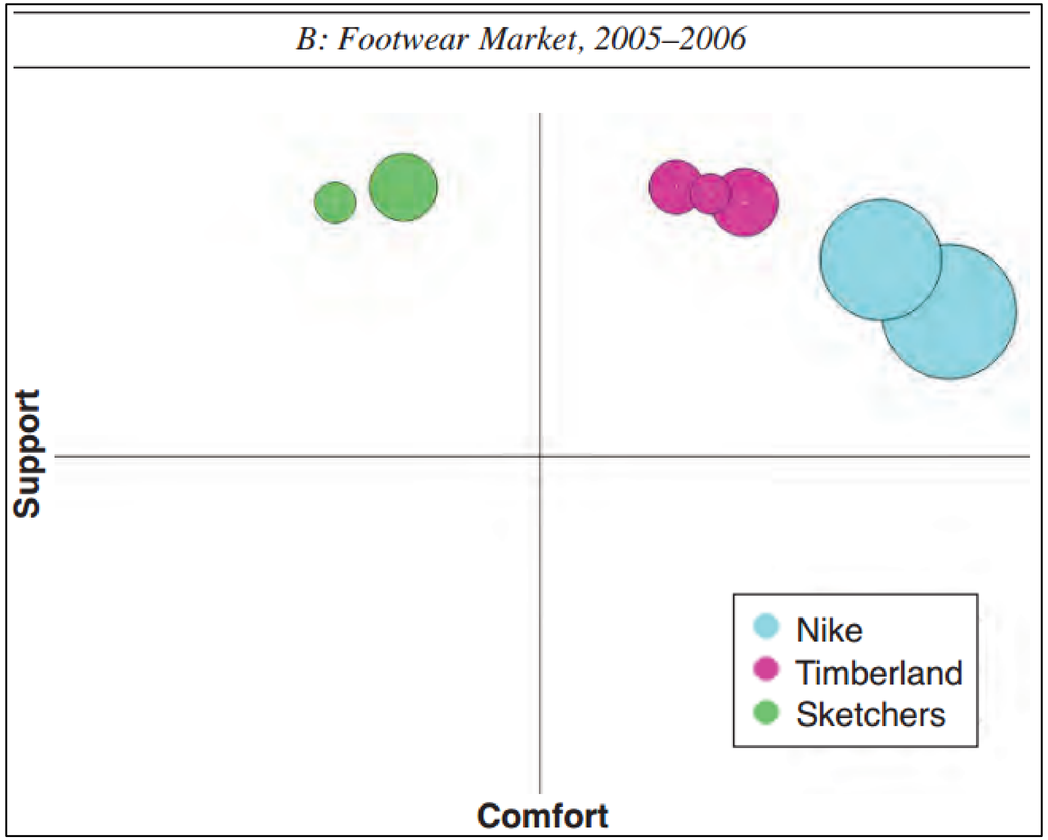
Durability & Safety : Mattel > LeapFrog

5 Within Brand Segmentation

Segment consumers on the basis of the proportion of words they allocate to the various dimensions of quality in their review

Size : volume of reviews citing these dimensions





5 Dynamic brand mapping

Dell's position on the ease of use is more unstable and changes rapidly over the time period.

- Define the estimated probability that a dimension k occurred in the review d in time period t

$$\hat{p}(k|t=\tau) = \sum_{d|t_d=\tau} \hat{p}(k|d)\hat{p}(d|t=\tau).$$

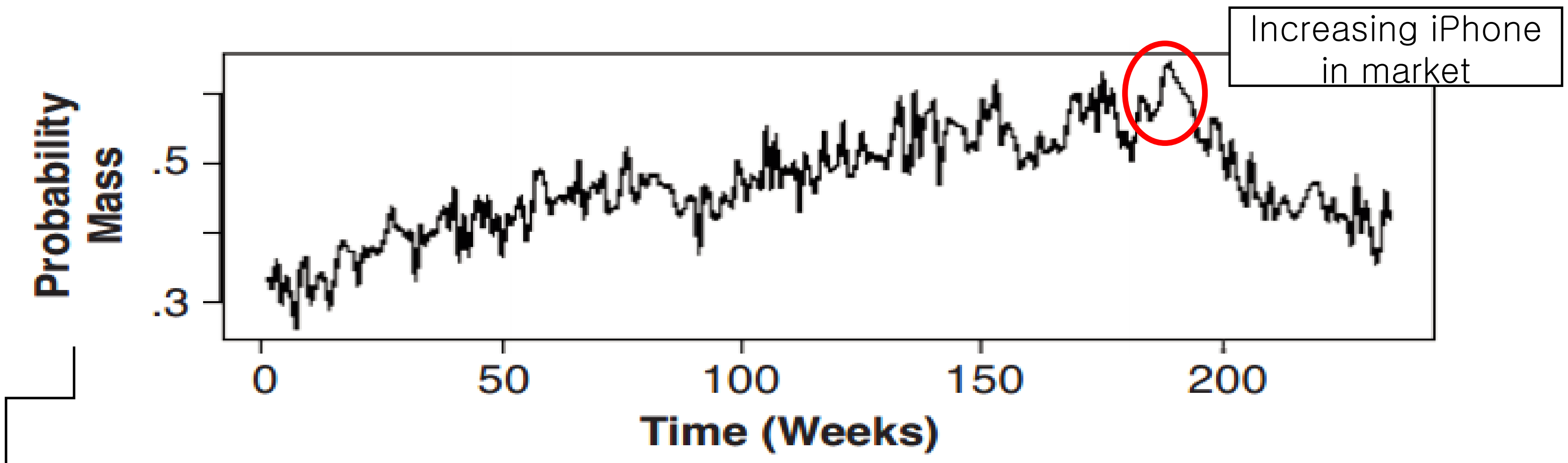
- 전체 review 에서 추출한 dimension 을 prior 로 사용

5 Dynamics of Dimensions

For the dimensions of quality that vary over time, we can obtain more insights.
In this case, trajectory seems to be related to the entry and exit of other brands

VARIATION IN THE EASE OF USE DIMENSION FOR THE MOBILE PHONE MARKET (BLACKBERRY)

A: Probability Mass Associated with the Ease of Use Dimension



$$\hat{p}(k|t=\tau) = \sum_{d|t_d=\tau} \hat{p}(k|d)\hat{p}(d|t=\tau).$$

5 Implications and Future work

- Implications
  1) It enables managers to ascertain the valence, labels, validity, importance, dynamics, and heterogeneity of latent dimensions of quality
  2) It enables managers to observe how brands compete on multidimensional space.
  3) It enables managers to track how this competition varies over time in great detail. (weekly level)

- Limitations and Future Research
  1) Computationally intensive
  2) This study focus only on product reviews
  3) LDA model is sensitive to hyperparameter of the Bayesian priors
  4) We neither include marketing mix variables nor study their impact on the brands or dimensions
  5) We do not analyze rare or infrequent words

Brand2Vec

*

Brand2vec Versus This study

- 각 브랜드를 단 하나의 ‘차원‘으로만 생각한 것의 한계
- 기존의 Brand2Vec 방법에 Sentiment, Dimensions of Quality 등의 정보를 추가적으로 extract 할 수 있는 방법에 대한 고민
- 이 연구에서 사용한 Herfindahl index, Mutual Information, Helliger distance 등의 적용
- 시간에 따른 변화