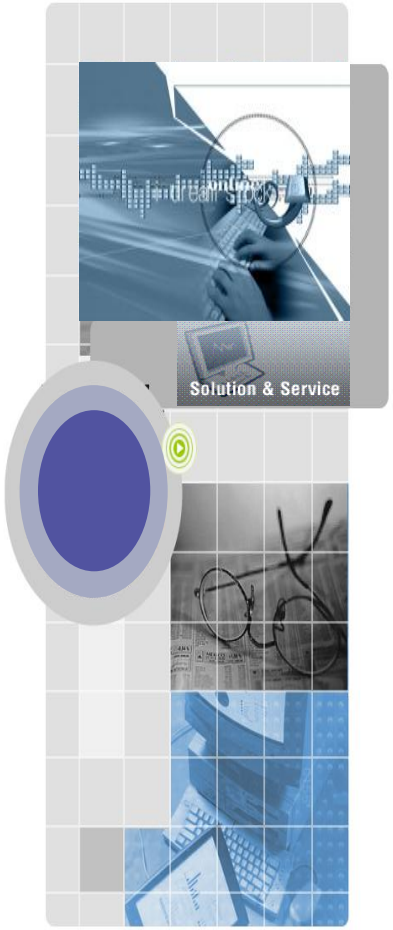
Text Mining Application to Internet Shopping Mall Customers' Reviews

Seok-Won Oh, Seohoon Jin
Korea University, Chungnam, Korea

contents



I. Introduction

I. Introduction

II. Text Mining

II. Text Mining

Text Mining

Statistical technique and machine learning algorithms are used for finding information from big sized text data

Application of TM

- Information Retrieval
- Natural Language Processing
- Information Extraction

Document classification

- Decision Tree
- Bayesian Classification
- Nearest Neighbor Classifier
- Support Vector Machine

Document clustering

- Hierarchical Cluster Analysis
- K-means Clustering
- **SOM(**Self Organizing Maps)
- Expectation-Maximization Clustering

1. structurizing

Term-document frequency matrix

	Doc 1	...	Doc j	...	Doc n
Term1	tf_{11}	...	tf_{1j}	...	tf_{1n}
⋮	⋮	⋱	⋮		⋮
Term i	tf_{i1}	...	tf_{ij}	...	tf_{in}
⋮	⋮		⋮	⋱	⋮
Term m	tf_{m1}	...	tf_{mj}	...	tf_{mn}

Term weighting : emphasize term-document relation

$$a_{ij} = L(i,j) \times G(i)$$

weight = local weight X global weight

frequency weight $L(i,j)$

Term-frequency	$L(i,j) = tf(i,j)$
Log transformation	$L(i,j) = \log_2(tf(i,j) + 1)$
Binary	$L(i,j) = 1, \quad tf(i,j) \geq 1$ $L(i,j) = 0, \quad tf(i,j) = 0$

term weight $G(i)$

Entropy	$G(i) = 1 + \sum_j \frac{(tf_{ij}/g_i)\log_2(tf_{ij}/g_i)}{\log_2(N)}$
GfIdf Global frequency Inverse document frequency	$G(i) = \frac{g_i}{d_i}$
Idf Inverse document frequency	$G(i) = \log_2(\frac{N}{d_i}) + 1$
Normal	$G(i) = \sqrt{\frac{1}{\sum_j L(i,j)^2}}$

2. Dimension reduction by singular value decomposition

$$A = U\Lambda V'$$

$A : m \times n \ matrix$

$U : m \times m \ matrix, rank = r, orthogonal$

$V : n \times n \ matrix, orthogonal$

$\Lambda : m \times n \ diagonal \ matrix$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq \lambda_{r+1} = \cdots = \lambda_{\min(m,n)} = 0$$

$U : m \times m$

$\Lambda : m \times n$

$V : n \times n$

$U : m \times r$

$\Lambda : r \times r$

$V : r \times n$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & 0 & \ldots & 0 \\ \vdots & 0 & \ddots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \lambda_m & \ldots & 0 \end{bmatrix} \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\lambda_i (i = 1,2,\cdots,\min(m,n))$$

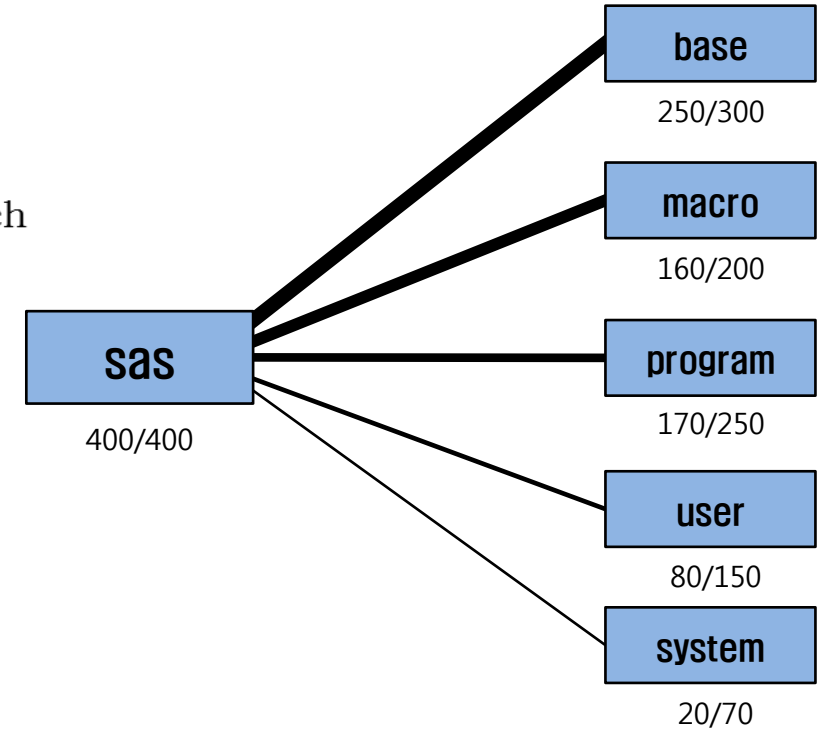
$$\boldsymbol{D_k = U\Lambda_k V'}$$

3. Clustering

4. Network Analysis(Concept Link)

Expectation-Maximization Clustering

- Step 1.
  - Obtain initial parameter estimates
- Step 2.
  - Compute the membership probability of x in each cluster k.
  - $P(G_j|x), \quad j = 1, \cdots, K$
  - Assign x to its maximum $P(G_j|x)$
- Step 3.
  - Update the mixture model parameters for each cluster
- Step 4.
  - Repeat Step 2 and Step 3 until converging



III. Practical application

1. data

1. Parsing and filtering
2. Singular value decomposition
3. Determine the number of clusters
4. Clustering and cluster profiling
5. Concept link

Customers' reviews of internet shopping mall

Best seller 100 products are considered initially

- lotion and cream are selected for analysis

2. Structurizing and filtering

Dictionary

Start List (Selecting Dictionary)	**Selecting Dictionary**
Stop List (Filtering Dictionary)	Filtering Dictionary

Part of speech filtering

Non-informative			**informative**		
Part of speech		filter	**Part of speech**		filter
Aux	auxiliary verb	Filtered	Abbr	abbreviation	Filtered
Conj	conjunction	Filtered	**Adj**	**adjective**	**Selected**
Det	determiner	Filtered	Adv	adverb	Filtered
Interj	interjection	Filtered	**Noun**	**Noun**	**Selected**
Part	particle	Filtered	Num	Number	Filtered
Prep	**preposition**	**Selected**	**Prop**	**proper noun**	**Selected**
Pron	pronoun	Filtered	**Verb**	**Verb**	**Selected**
			VerbAdj	**Verbal adjective**	**Selected**

synonym dictionary

Category	Term	Parent
Noun	Price	price
Noun	Value	
Noun	Cost	
Noun	delivery	delivery
Noun	shipping	
Noun	아가	아기
Noun	아이	
Noun	애기	
Noun	상품	제품
Noun	제조일자	제조일
Noun	제조날짜	
Adj	고맙다	감사하다
Adj	마르다	건조하다
Verb	소개하다	권하다

3. SVD

10 Maximum loaded terms					
SVD 1		SVD 2		SVD 3	
Term	Value	Term	Value	Term	Value
용량	0.6529	단점	0.8724	맘놓고	0.7184
바르다	0.6223	유지되다	0.8683	2010년	0.7148
크림	0.5881	질환	0.8669	바디클렌져	0.7024
많다	0.5115	외출	0.8648	벗기다	0.6889
촉촉하다	0.4980	탁월하다	0.8440	감사하다	0.6489
건조하다	0.4907	은은하다	0.7997	2011년	0.6218
발	0.4545	무취	0.7930	걱정없이	0.5790
저렴하다	0.4451	성분	0.7804	포장	0.5734
건조	0.4365	미끈거리다	0.7625	년	0.5440
바르다	0.4179	약하다	0.6962	깨지다	0.5354
10 minimum loaded terms					
이상하다	0.0976	크림통	-0.1671	상처	-0.1877
거치다	0.0943	피부가 보들보들	-0.1677	고민	-0.1896
요긴하다	0.0907	많다	-0.1684	증상	-0.1911
뽁뽁이	0.0900	간지럽다	-0.1745	심하다	-0.2027
건강	0.0864	촉촉하다	-0.1969	건조하다	-0.2131
남녀노소	0.0795	크림	-0.2028	겨울	-0.2158
교환하다	0.0776	울다	-0.2036	느낌	-0.2337
설명서	0.0670	부담스럽다	-0.2131	긁다	-0.2340
g	0.0548	듬뿍듬뿍	-0.2315	바르다	-0.2386
교환	0.0538	바르다	-0.2620	가볍다	-0.2445

SVD 1

+ ▪ volume, a lot of, moist, dry, cheap
+ ▪ exchange, g, manual, health, important

SVD 2

+ ▪ defect , disease, soft, scentless, slippery
− ▪ apply, generously, moist, tender

SVD 3

+ ▪ 2010, 2011, year, packing, unclothe
− ▪ scrape, winter, dry, symptom, abrasion

4. clustering

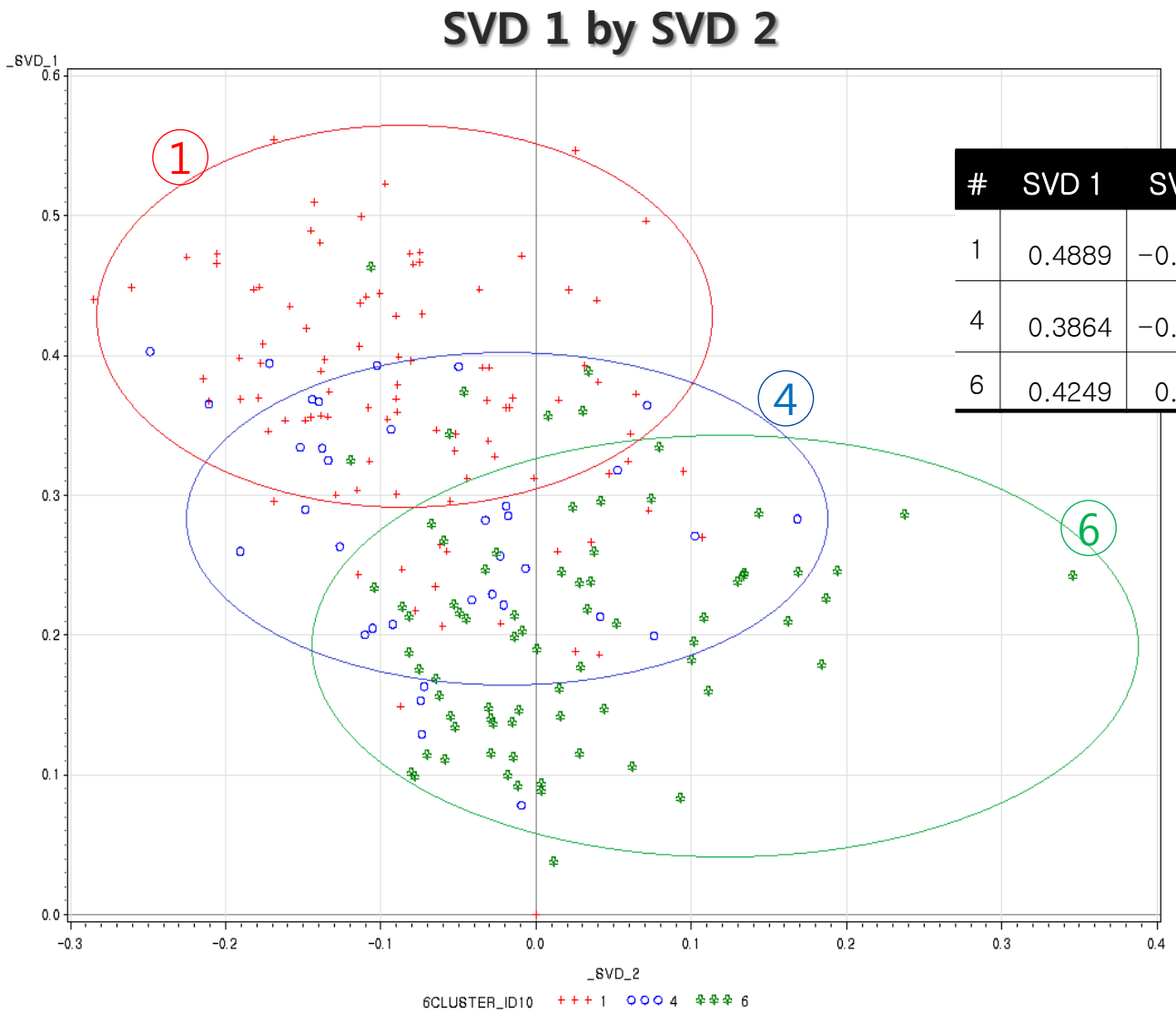
Criterion for determine number of clusters

K	RMS Std	Pseudo F	Hartigan
2	0.1314	0.423	-244.956
3	0.1369	0.158	-239.009
4	0.1307	2.397	-172.884
5	0.1281	5.652	-181.068
6	**0.1132**	**6.687**	**-209.477**
7	0.1240	3.364	11.238
8	0.1236	2.043	-179.220
9	0.1185	4.598	29.122
10	0.1155	3.066	-176.718

Cluster profiles

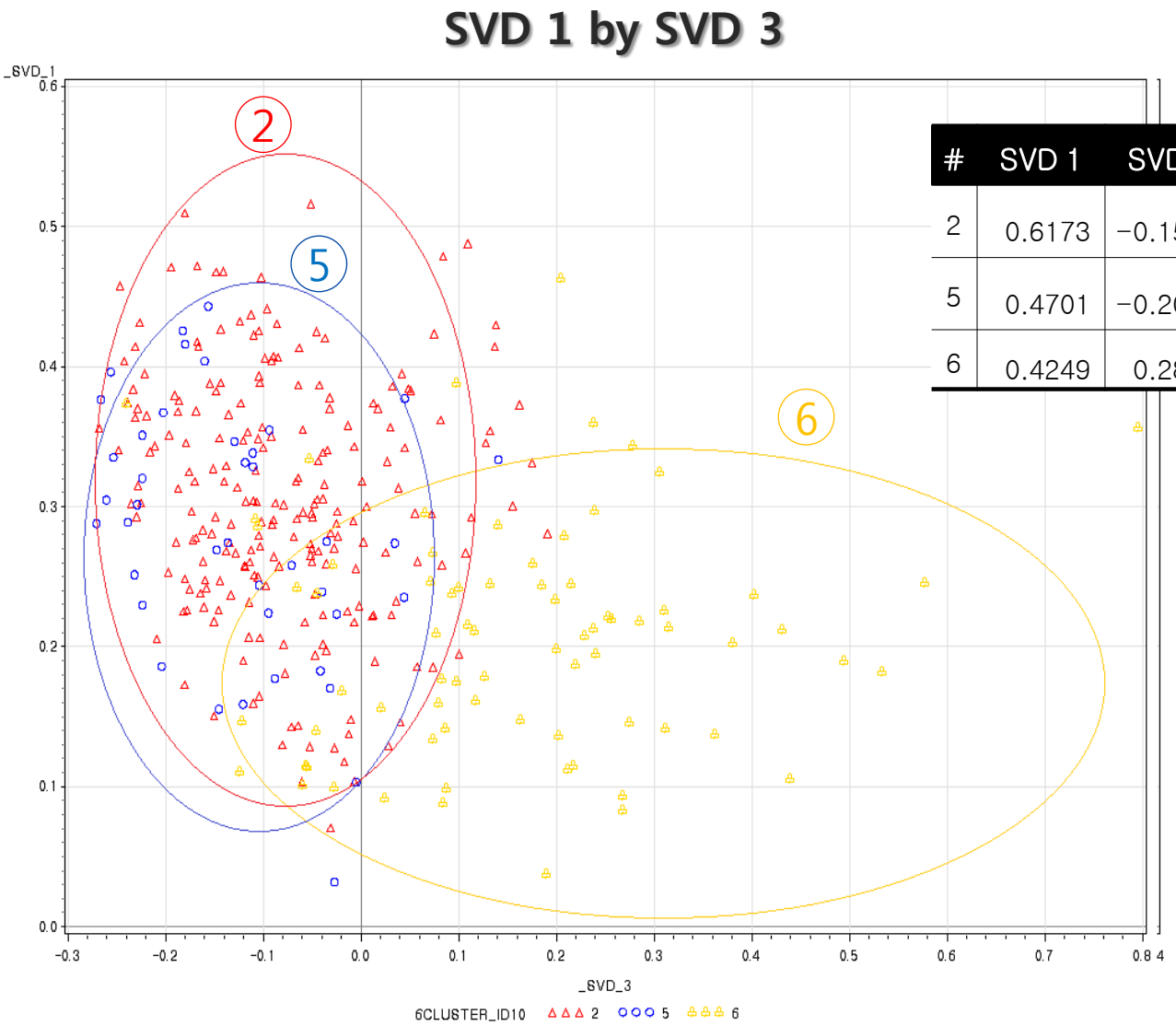
#	SVD 1	SVD 2	SVD 3	Freq	Descriptive Terms
1	0.4889	-0.0992	0.0933	131	용량, 많다, 저렴하다, 듬뿍듬뿍, 보습
2	0.6173	-0.1532	-0.1526	221	건조, 바르다, 크림, 피부, 얼굴
3	0.5065	0.0444	0.0250	99	빠르다, 저렴하다, 추천하다, 배송, 향기
4	0.3864	-0.0849	-0.1207	36	가격, 저렴하다, 배송, 향기, 아토피
5	0.4701	0.0930	-0.2035	42	느낌, 씻다, 거품, 클렌저, 겨울
6	0.4249	0.0237	0.2887	76	비닐, 포장, 년, 제조일, 용기

5. Cluster visualization by SVD



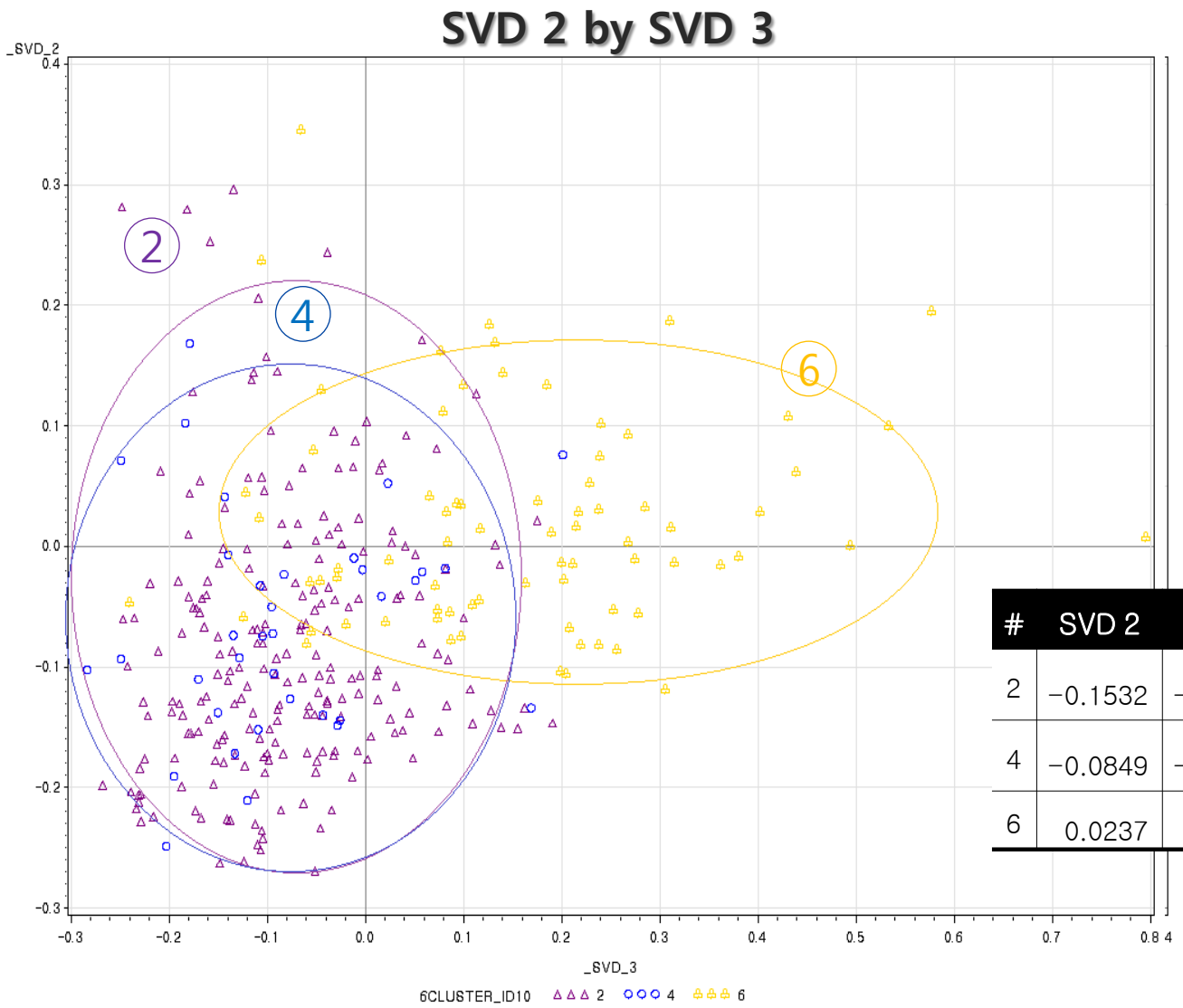
#	SVD 1	SVD 2	Descriptive Terms
1	0.4889	-0.0992	용량, 많다, 저렴하다, 듬뿍듬뿍, 보습
4	0.3864	-0.0849	가격, 저렴하다, 배송, 향기, 아토피
6	0.4249	0.0237	비닐, 포장, 년, 제조일, 용기

5. Cluster visualization by SVD



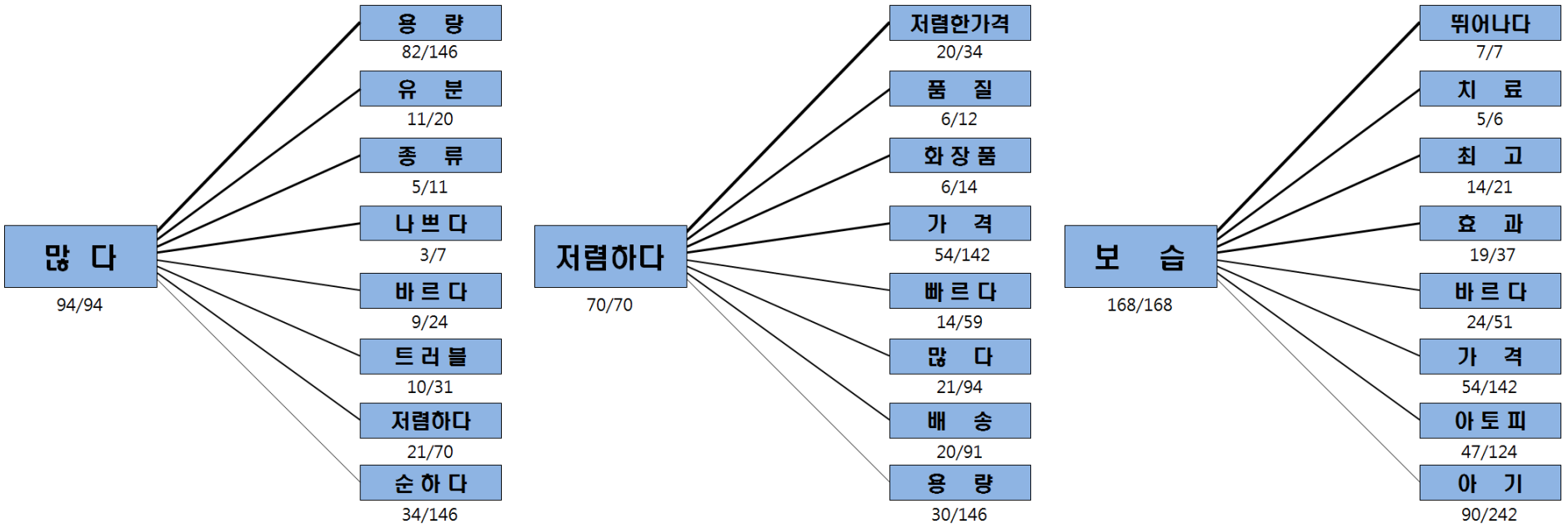
#	SVD 1	SVD 3	Descriptive Terms
2	0.6173	−0.1526	건조, 바르다, 크림, 피부, 얼굴
5	0.4701	−0.2035	느낌, 씻다, 거품, 클렌저, 겨울
6	0.4249	0.2887	비닐, 포장, 년, 제조일, 용기

5. Cluster visualization by SVD



#	SVD 2	SVD 3	Descriptive Terms
2	-0.1532	-0.1526	건조, 바르다, 크림, 피부, 얼굴
4	-0.0849	-0.1207	가격, 저렴하다, 배송, 향기, 아토피
6	0.0237	0.2887	비닐, 포장, 년, 제조일, 용기

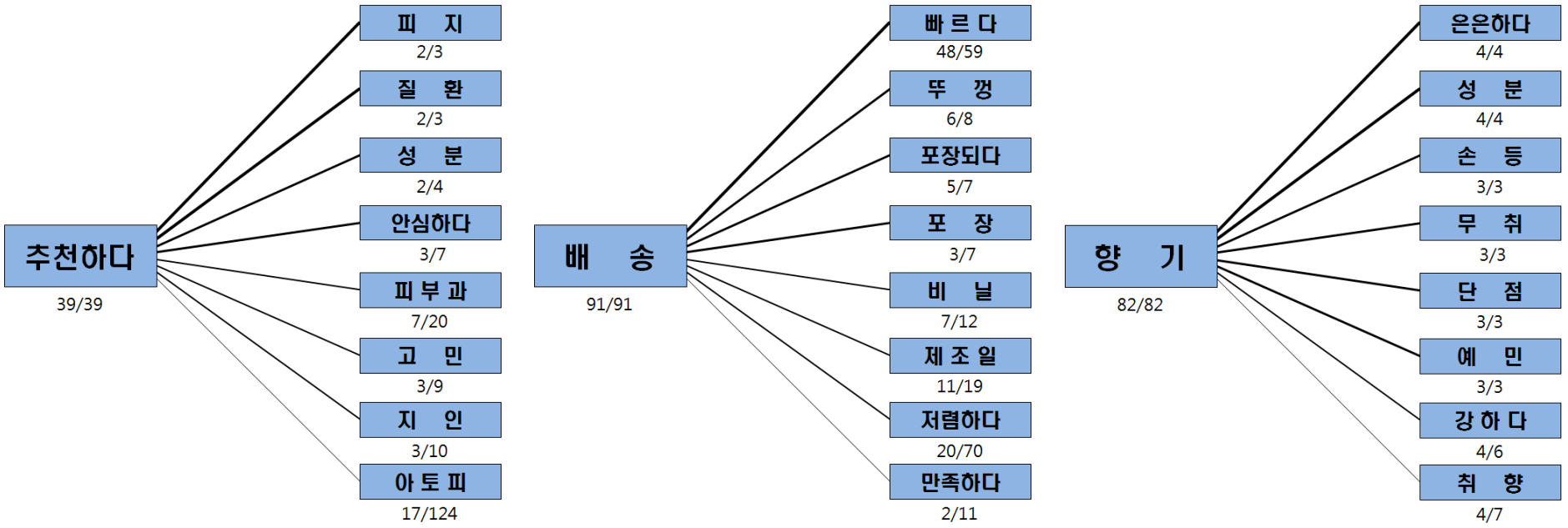
6. Concept link



cluster1 Large volume, cheap price, moisturizing, cure

cluster1 Customers' reviews which are about cheap price, excellent moisturizing, and curing an atopic dermatitis.

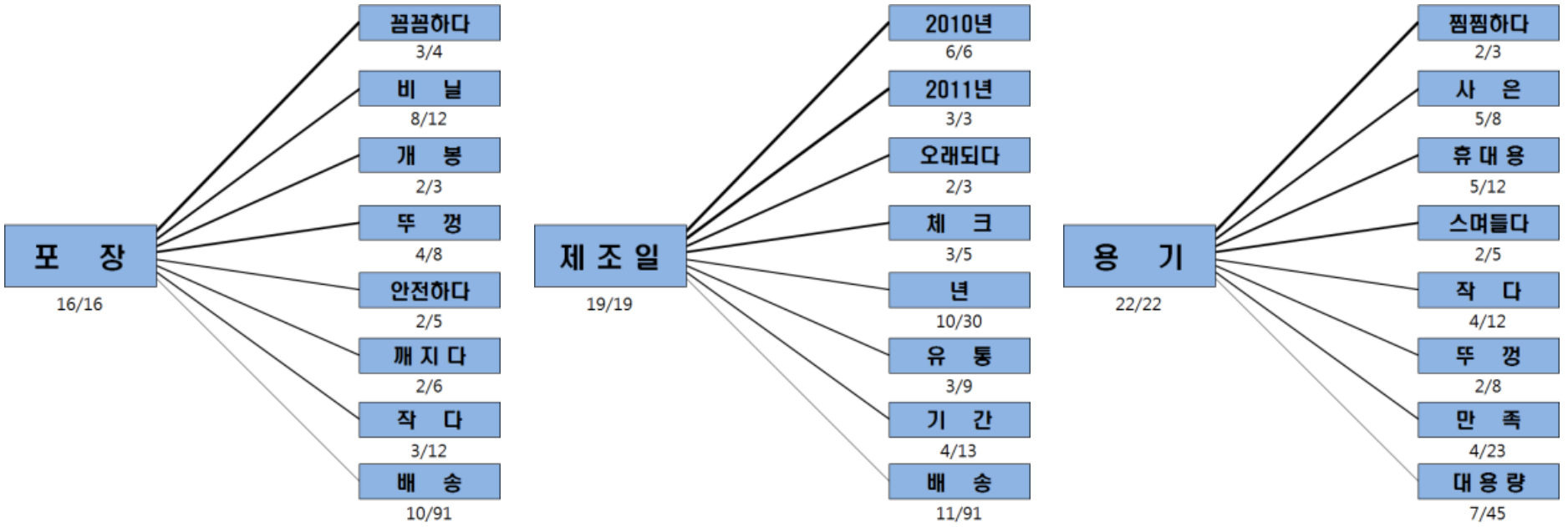
6. Concept link



cluster3 Quick delivery, cheap price, recommendation, scent

cluster3 Customers' reviews which are about recommendation for dermatitis, quick delivery, good packing, and scent.

6. Concept link



cluster6 Packing, expiration date, negative word

clsuter6 Customers' reviews which are about pacing problem, expired product, etc.
This cluster consists of negative documents.

7. Usage of results

Usage of Text mining results	
Customer management	▪Understanding customers' response about the products
marketing	▪Target marketing based on customers' preference
promotion	▪Finding strength and weakness about the products ▪Emphasize strength for selling promotion
Product development	▪Realize the weak points of the products ▪Re-design the products
Settle a grievance	▪Realize customers' complain ▪Reduce customers' complain by solving the problems

Thank you