Department of Social Systems and Management
Discussion Paper Series

No.1205

Visual Inspections of the Semantic Interface on the Web:

A Comparison of Top Cosmetic Brands by SVD and Network

by
Noriyuki Matsuda
Department of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennou-dai, Tsukuba, 305-8573, Japan
email: mazda@sk.tsukuba.ac.jp
Akiko Machida
Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennou-dai, Tsukuba, 305-8573, Japan
Kei Mizuno
Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennou-dai, Tsukuba, 305-8573, Japan

May, 2008

A revised version will appear in
Proceedings of IADIS Computer Graphics and Visualization (*CGV*);
Amsterdam, Netherlands, July, 25-27, 2008

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Discussion Paper Series, No. 1205, Department of Social Systems and Management, University of Tsukuba, 2008

Visual Inspections of the Semantic Interface on the Web: A Comparison of Top Cosmetic Brands by SVD and Network

Noriyuki Matsuda
Departmentl of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennou-dai, Tsukuba, 305-8573, Japan
email: mazda@sk.tsukuba.ac.jp

Akiko Machida
Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennou-dai, Tsukuba, 305-8573, Japan

Kei Mizuno
Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennou-dai, Tsukuba, 305-8573, Japan

ABSTRACT

On the premise that web presentations of consumer products serve as an important interface between businesses and consumers, we analyzed verbal information on the webs of the top two skincare brands in the Japanese market. *Term* by *context* matrices created from the documents were subjected to SVD to construct component based graphs. In addition to descriptive inspections of the networks, core nodes and their neighborhoods were extracted. The brands showed both contrasts and commonalities. On one hand, the distinctiveness of the components was higher in ELIXIR than in SEKKISEI. On the other hand, the latter showed crystal core-ness and full extent of the core neighborhood as compared with the former. It was noteworthy that the top-most cores of the two brands pertained to moisturizing ingredients. Probably, this reflected the shared business awareness of the central concern about skincare among the Japanese consumers. Our plan of further investigations were stated in conclusion.

KEYWORDS

semantic interface, graph and network, centrality, core, neighborhood

1. INTRODUCTION

Manufacturers today actively employ web sites to appeal to consumers. Of particular interest in terms of affective information design (e.g., Matsuda, 1997) is the presentation of their products in nicely phrased terms regarding benefits, ingredients, prices and so forth. Consumers also actively use these sites in search for convincing information compatible their various concerns. In this vein, the web presentations of consumer goods serve as semantic interface between business and consumer interests.

Among others, cosmetics provide interesting interactions of multiple issues, ranging from brand-images, health and beauty, safety as well as regulations about products and labeling. Though each product is small in size, cosmetics are seldom used in isolation but are likely to be consumed in a set or series particularly among women. Hence, not only brands of product families but also corporate brands bear significance both to manufacturers and consumers. Indeed, Matsuda and Namatame (1995) found hierarchically related brand images about skincare goods among young Japanese women to whom the product category bears special significance not only as an entry to cosmetics but as a continuous basis for makeup in the ensuing life.

Therefore, the present exploratory study will focus on the top skincare brands in the Japanese market in view of the drug legislation and other related regulations that permit displays of approved medical effects for cosmetics-drugs, leaving displays of the pertaining ingredients as optional owing to rigorous testing in advance (MHLW, 2007). In contrast, the labeling of ordinary cosmetics should carry no medical effect but should list up major ingredients. In short, designing of cosmetics labels on the web as well as packages is almost an art under various constraints.

Our study was inspired by the development of LSA (Latent Semantic Analysis) and LSI (Latent Semantic Indexing) by Deerwester, Dumais et al (1990), Landauer, Foltz and Laham (1998) and Berry (1999) aimed at analyzing verbal structures underlying huge texts by SVD (Singular Value Decomposition). As a point of departure from LSA/LSI, we will also employ the recent techniques of graph/network analysis to gain rich quantitative and visual insights.

1.1 SVD (Singular Value Decomposition)

SVD is a powerful tool in linear algebra that decomposes a rectangular matrix into the product form as $\boldsymbol{A}_{mxn}=\boldsymbol{U}_{mxn}\boldsymbol{S}_{nxn}\ \boldsymbol{V}_{nxn}{}^{\mathrm{T}}$ where $\boldsymbol{U}$ and $\boldsymbol{V}$ are ortho-normal bases, and $\boldsymbol{S}$ is a diagonal matrix containing singular values ordered from the largest to the smallest. By the nature of these matrices, we obtain the following important relationships among $\boldsymbol{A}$, the vectors of $\boldsymbol{U}$ and $\boldsymbol{V}$, and the singular values σ_i**:**

$$\boldsymbol{A}\boldsymbol{v}_i=\sigma_i\boldsymbol{u}_i \text{ and } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}_i=\sigma_i\boldsymbol{v}_i$$

The major merit of SVD accrues from the estimation of the original $\boldsymbol{A}$ by a truncated matrix $\boldsymbol{A}^*$ by restricting to, for instance, the largest k singular values, i.e., $\boldsymbol{A}_{\mathrm{mxn}}{}^* = \boldsymbol{U}_{mxk}\boldsymbol{S}_{kxk}\ \boldsymbol{V}_{kxn}{}^{\mathrm{T}}$. Given the sufficiency of k, $\boldsymbol{A}^*$ is known to be a good estimate of $\boldsymbol{A}$ devoid of noise. Of equal interest is the principal-component type treatment of $\sigma_i\boldsymbol{u}_i$ to be conducted in the present work, leaving the former approach to our companion work (Matsuda et al, 2008).

After LSA/LSI approach, we will adjust the frequency matrix $\boldsymbol{A}$ (*term* x *context*) by tf-idf (see Dumais, 1991 for various techniques), where a *term* is a word/symbol or their combinations and a *context* is an entity in which the term occurs. Terms on higher and lower ends of the components, i.e., $\boldsymbol{u}$, will be selected for the ensuing graph/network analysis.

1.2 Classification of the Text Information

Usually, a family of cosmetic items are sold under the family brand often accompanied by the corporate-level brand. An extensive family of skincare items consists of lotion, emulsion, cream, foam and a mask some of which may carry item brands as well. Their presentations on the web like other advertisement contain an attractive slogan and other information such as ingredients, major benefits, price, effective use and so forth. Ideally they appear in separate phrases, sentences and paragraphs, but they are actually often mixed. Being an art as a whole, poetic statements are blended with non-poetic ones in various syntactical forms.

Hence, the texts will be segmented into constituent terms that may be words, particles, symbols or their combinations on the basis of the field-expertise knowledge. Also, the contexts within which they appear are to be classified into a slogan, ingredients, benefits and miscellaneous others depending on the principal tone.

1.3 The Focus of Network Analysis

Interests in the identification of central nodes and communities (see, Freeman, 1979) are growing among network researchers besides the development of layout algorithms for visualization. Our analysis is in line with this trend, but with the principal aim at finding out core nodes with high centralities and their neighborhoods for which four centrality measures were used—betweenness, closeness, degree and PageRank (Brin and Page, 1998), that are briefly explained in the Appendix.

Core nodes are defined here as those with highest scores in all or in the majority of centrality indices. A preliminary analysis of the networks indicated no need to divide them into clusters/communities in contrast to our companion work (Matsuda, Machida and Mizuno, 2008). Therefore, core nodes and their neighborhoods were to be identified directly from the entire networks.

2. METHOD

Data--The text data were collected, during the first week of April, '07, from the web sites of the two top skincare brands in the Japanese market: SEKKISEI and ELIXIR of KOSÉ and SHISEIDO, respectively. Total 245 and 204 terms were extracted across 12 and 10 items from the respective brands. The contexts were created as the direct product of [item]x[slogan, ingredient, benefit, miscellaneous]. Deleting 8 empty contexts from the SEKKISEI data, we obtained the same number of contexts for each brand, i.e., 40. Thus constructed frequency matrix ***A***'s (term by context) were adjusted by tf-idf prior to SVD. All the procedures, except for the matrix construction, were run by software package *R*.

In accord with the principle of parsimony, we first selected the *k*-largest singular values whose squared cumulative proportions exceeded 80%-- $\sum_i^k \sigma_i^2 / \sum_i^{40} \sigma_i^2 \geq .80$; *k*=15, 18 for SEKKISEI and ELIXIR, respectively. Then, a new matrix ***B*** consisted of column vectors $\sigma_i \boldsymbol{u}_i$ for each brand: i.e.,

$$\boldsymbol{B}=[\sigma_1\boldsymbol{u}_1\ \ \sigma_2\boldsymbol{u}_2 \ldots \sigma_k\boldsymbol{u}_k] \qquad (k=15 \text{ or } 18)$$

The rows of ***B*** pertain to the terms, while the columns contain "loadings" on the components in rough analogy to principal component analysis. For the subsequent graph inspection, new matrices ***B**** were derived from respective ***B*** by selecting terms whose loadings fell in the higher or lower ends of components. The cut-point was set at 20% of the column length on both ends, i.e., 40% of the terms per component.

Network representation--In the subsequent analysis, the relations between the terms selected from ***B**** and the components were represented in networks comprised of the term-nodes and the component-nodes akin to latent factors in factor analysis (see Figure 1). The links were weighted by $\sigma_i \boldsymbol{u}_i$.

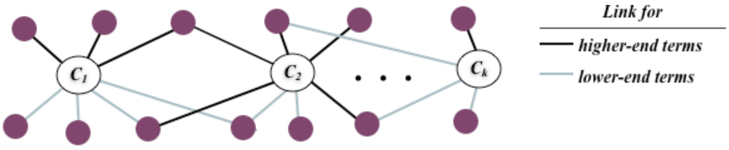

Figure 1. A schematic graph comprised of term- (in gray) and component-nodes (in white)

Elimination of isolated term-nodes (7 and 6) that had no links resulted in the networks of 237 and 197 term nodes in addition to 15 and 18 component nodes for SEKKISEI and ELIXIR, respectively.

Network indexes—Our main overall indexes were size, in terms of the number of nodes, and the path lengths. Centrality of nodes to be used as a basis of coreness were measured by four indexes—betweenness, closeness, degree and PageRank (see the Appendix for a brief explanation.).

3. RESUTLS

The results were presented in this section in order of SEKKISEI and ELIXIR unless otherwise noted. For the network analysis and layout, we employed the *igraph* library of the statistical package *R*. Concerning the layout, Kamada-Kawai's (1991) and Reingold-Tilford's (1981) algorithm were used in consideration of the ease of visual inspections as compared to other alternatives. Nevertheless, the produced layouts should be topologically interpreted due to the randomly generated initial states. Finally, the components-nodes will be abbreviated as components hereinafter.

3.1 Descriptive inspections

3.1.1 Overall patterns

The networks for the two brands are shown in Figure 2. The number of term-nodes were 237 and 197 connected to 15 and 18 respective components by 1470 and 1476 links. Selection of 40% of the terms per component could have produced more links for ELIXIR across relatively large number of components. Perhaps the result might be attributable to the relative concentration of single-link nodes on the first component in it. Before proceeding to the examination of degrees, we report the results regarding path-length in terms of geodesic distance.

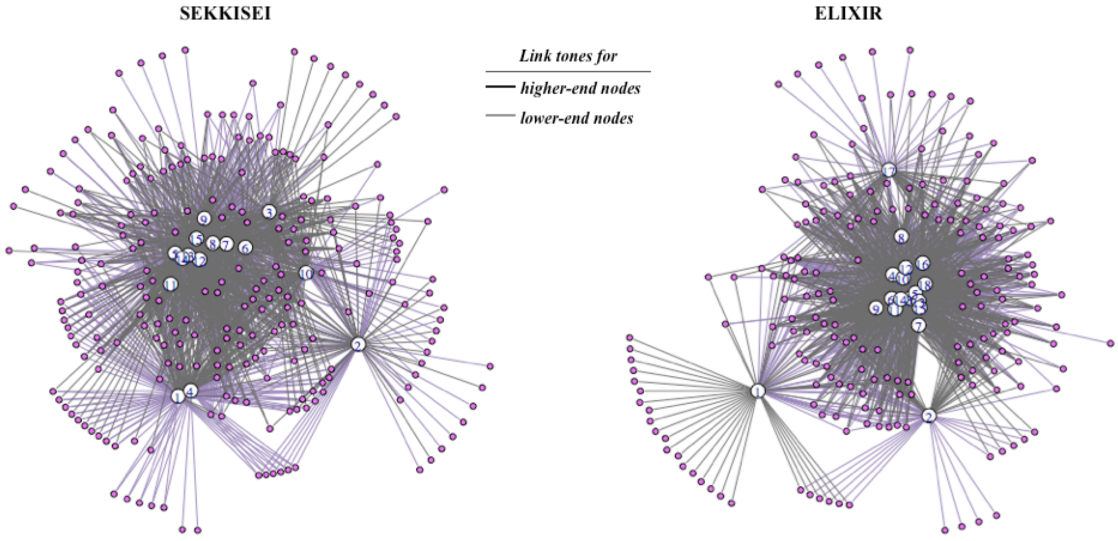

Figure 2. Networks consisted of non-isolated nodes mediated by components (white circles): Layout by Kamada-Kawai's algorithm (1989)

Every node in the networks was reachable from all other nodes in average 2.43 and 2.45 steps, ranging from 2 (i.e., shortest-paths) to 4 (farthest-paths) steps in both networks. This might account for the heavy clouds near the middle of the networks as seen in Figure 2. This is expected to affect the subsequent identification of core neighborhoods.

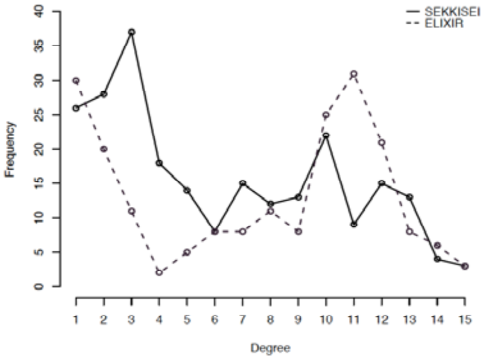

Figure 3. Frequency of degrees by brand

The degrees associated with the term-nodes were distributed asymmetrically in both networks with the maximum of 15 as shown in Figure 3. While the distribution in SEKKISEI was uni-modal (at 3), that in ELIXIR were bimodal (at 1 and 11). The mode-ness (i.e., the relative frequency) were nearly equal: .156 (at 1) in the former, and 1.52 (at 1) and .157 (.157) in the latter.

The two brands differed in the frequency of the nodes, with degree 1, connected to single components. that would indicate the distinctiveness of the components. In SEKKISEI, there were 26 such nodes for which 7 components were involved, whereas in ELIXIR only 4 components were involved in 30 such nodes. Namely, the difference implied relatively distinctiveness of the components in the latter. The tendency was succinct on the first component. Component 1 of ELIXIR had 63.3% share that was greater than the sum of shares of components 1 and 2 of SEKKISEI (30.7% each).

3.1.2 Centrality indices

Four popular central indices—betweenness, closeness, degree and PageRank--were adjusted by dividing them by the respective maximum values for the ease of comparisons. Shown in Figures 4 are their distributions by brand arranged in descending order with respect to betweenness.

The distributions were fairly concordant for both brands except for the steepness of the curves to be measured later by Gini coefficient. The Pearson's correlation coefficient ranged from .839 (betweenness—closeness) to .998 (PageRank—degree) in SEKKISEI, and from .756 (betweenness--degree) to .998 (PageRank—degree) in ELIXIR.

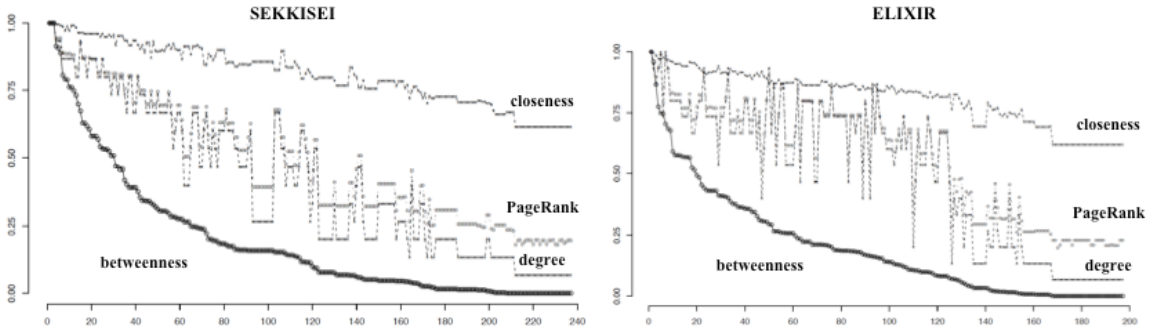

Figure 4. Adjusted centralities rearranged in descending order with respect to betweenness by brand

Gini coefficients indicated slightly greater unevenness of individual distributions in SEKKISEI as compared to ELIXIR: (.604, 0.079, .373, .265) and (.559, .075, .331, .235), respectively, in order of betweenness, closeness, degree and PageRank.

3.2 Core nodes and their neighborhoods

3.2.1 Identification of core nodes

A node was judged as a core of the network if its centrality ranked highest on more than one indices. The task was straightforward in the case of SEKKISSEI due to the consistent rankings of three terms on all the centrality indices (see Table 1). However, rankings about ELIXIR required somewhat lenient rule. Among other alternatives, we selected the terms which ranked first or second on at least two indices. Thus, the core-ness of the terms were weak relative to that of SEKKISEI in this regard.

Table 1. Ranking of the core nodes by brand

Brand	Core node	betweenness	closeness	degree	PageRank
SEKKISEI	A: (moisturizing)[*1]	1	1	1	1
	B: moist	2	2	2	2
	C: realize	3	3	3	3
ELIXIR	A: ubiqinon[*2]	1	1	2	1
	B: resilient	5	5	1	2
	C: care	2	2	4	4

Note: 1) The parenthesized labeling for ingredients comply with the regulation.
2) A co-enzyme generally known as CoQ10.

3.2.2 Core neighborhoods

We first created individual neighborhoods of each core by collecting the term-nodes connected to it by two links via intermediate components. Then, they were combined by set operations (union/intersection), if appropriate.

The individual core neighborhoods in SEKKISEI were maximally extensive in that they were all identical with the total network. That is, every node was reachable from the cores within two links via components. Hence, no set operation was necessary. In contrast, the core neighborhoods were less extensive in ELIXIR with a slight variation in size between 206 to 211. Besides, not all the components were included in them.

However, the variation was minor such that the union neighborhood was identical with the total network. The intersection neighborhood comprised of 191 term-nodes and 8 components. The cores A, B and C marked in the total networks (see Figure 5) in which the term-nodes and components excluded from the intersection were drawn in small circles.

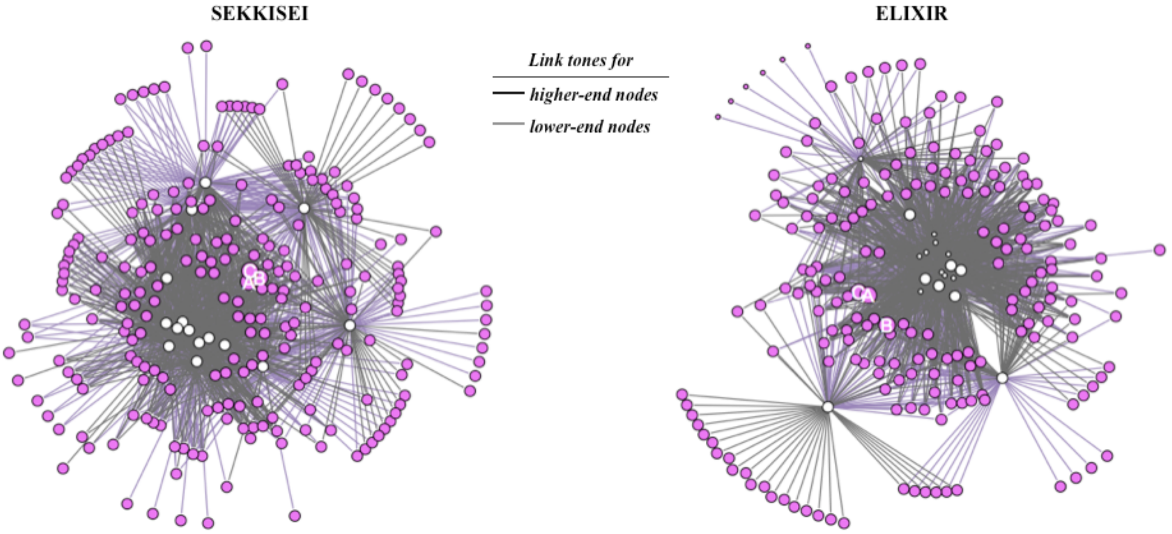

Figure 5. The core neighborhood highlighted in the total network of each brand:

Layout by Kamada-Kawai's algorithm (1989)

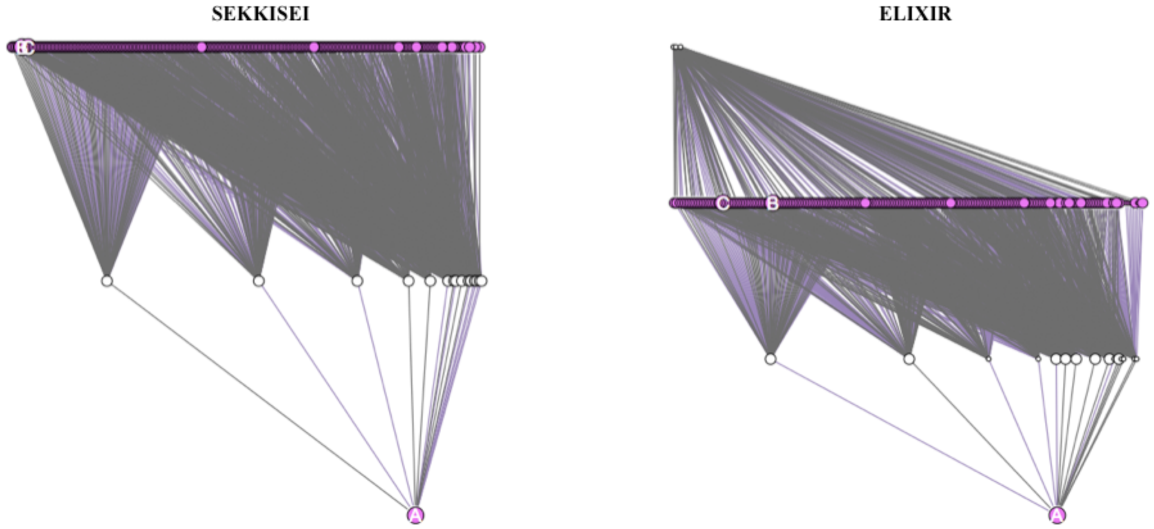

Figure 6. The tree-structured version of the network by brand:

Layout by Reingold-Tilford's algorithm (1981)

The tree-like layout, in Figure 6, with the topmost node A at the roots revealed an interesting contrast between the brands. On one hand, the topmost node A of SEKKISEI was directly linked to all the components through which every other term-node was immediately reachable. On the other hand, node A of ElIXIR had no links to three components, located in the top layer, which fell out of the intersection neighborhood. Among 15 components directly connected to A, seven were excluded from the intersection.

4. DISCUSSION AND CONCLUSION

Recent progress in graph analysis and visualization techniques enabled us to explore the semantic interface on the web pertaining to the top Japanese skincare brands. We have treated the term "semantics" in the dual sense—one denoting the conventional relationships between words/terms and their meanings, and, the other denoting more broad ones resulting from their interactions. The latter is of particular importance to the texts, notes and other verbal information on the web which is likely to be processed neither sequentially nor consecutively by readers. Hyperlinks among texts and sites enhance the tendency. Hence, the present application of SVD to the *term* by *context* matrix prior to graph analysis seems to a practically viable way to handle the semantics in the dual sense. Ideally, one could incorporate audio and/or visual information in the semantic studies.

Although the brands did not radically differ from each other, interesting contrasts emerged through our analysis such as the distinctiveness of the first component of ELIXIR in terms of concentration of the single-degree term nodes. A close examination is needed to see whether these terms are interpretable as a group without regard to others. Equally noteworthy is the high connectivity of the SEKKISEI network, most clearly evidenced by the identicalness of the core neighborhoods and the total network. This raises an intriguing question whether such tight network will contribute to the informativeness of the web presentation as a whole.

It is our hope that readers will gain deeper insights The value of the present work must be evaluated Moreover, SVD enables us to shed light on the problem in two different approaches as mentioned earlier in section 1.1—one via matrix truncation as done in our companion work (Matsuda, Machida and Mizuno, 2008), and, the other, component-based treatment as done in the present work. The cores and their neighborhoods extracted in our two studies differed in terms of "embedded-ness" in the initial networks and

The two studies differed produced by both union and intersection operations of individual core neighborhoods. This is one of the most striking differences between the two works. It also revealed both contrasts and commonalities between the two brands. We plan to integrate the findings of the two studies to enrich our knowledge.

One of the principal reasons for applying skincare is to moisturize skin with little health risk. Many consumers, particularly in the Japanese market, consumers are often concerned about the ingredients, even though they may lack sufficient physio-chemical knowledge. Also, the governmental regulations require manufactures to make them known to the public. Therefore, it was not surprising to find that the topmost cores of the two brands pertained to moisturizing ingredients, though the core of SEKKISEI was the category label under which specific ingredients were listed, that of ELIXIR referred to a specific one. Another principal concern among Japanese women is plump appearance in face. Actually, the term "plump" was identified as the second-most core in both brands in our companion study (Matsuda, Machida and Mizuno, 2008) which employed the truncated matrices obtained by SVD.

We hope that our approach is valuable to web-designers who need to examine whether the essential information is both explicitly and implicitly organized as planned across related pages. Market analysts may also benefit from explicating marketing intentions of competing businesses in lieu of conventional content analysis. In addition to the business sides, consumers can have alternative means to learn about the products they are concerned with, given the development of user-friendly tool.

We conclude the paper with a critical remark on our own work. Our attempts of incorporating link weights were not fully satisfactory in index computations and layouts. Improvements in this regard are strongly desired. .

REFERENCES

Journal

Berry, M.W., Drmac, Z., & Jessup, E.R., 1999. Matrices, Vector Spaces, and Information Retrieval. *SIAM Review*, 41:335–362. doi: 10.1137/S0036144598347035.

Brin, S., & Page, L., 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, Vol. 30, pp.107–117.

Clauset, A., Newman, M.E.J., & Moor, C., 2004. Finding Community Structure in Very Large Networks. *Physical Review, E* 70, 066111.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., & Harshman, R.A., 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, Vol. 41, pp.391-407.

Dumais, S.T., 1991. Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments and Computers*, Vol.23(2), pp.229-236.

Duarte, M.C., Santos, J.B., & Melo, L.C., 1999. Comparisons of Similarity Coefficients based on RAPD Markers in the Common Bean. *Genetics and Molecular Biology*, Vol. 22(3), pp. 427-432.

Freeman, L.C., 1979. Centrality in Social Networks. *Social Networks*, Vol.1, pp.215-239.

Fruchterman, T.M.J., & Reigngold, E.M., 1991. Graph Drawing by Force-Directed Placement. *Software--Practice and Experience*, Vol.21(11), pp.1129—1164.

Kamada, T., & Kawai, S., 1989. An Algorithm for Drawing General Undirected Graphs. *Information Processing Letters*, 31(1): 7-15.

Landauer, T.K., Foltz, P.W., & Laham, D., 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, Vol.25, pp.259-284.

Matsuda, N., 1997. *Kansei Information Design (in Japanese)*. Tokyo: Ohm.

Matsuda, N., & Namatame, M., 1995. Interactive Measurement of Hierarchically Related Consumers' Images. *Behaviormetrika*, Vol.22, pp.129-143.

Matsuda, N., Machida, A., & Mizuno, K., 2008. Structural Comparisons of the Semantic Interface of the Top Cosmetic Brands on the Web by Network Analysis. In *Proceedings of IADIS Interfaces and Human Computer Interaction* 2008 (*IHCI*); Amsterdam, Netherlands, July, 25-27, 2008.

Newman, M.E.J., & Girvan, M., 2004. Finding and Evaluating Community Structure in Networks. *Physical Review E* 69(2), 026113.

Pons, P., & Latapy, M., 2005. Computing Communities in Large Networks Using Random Walks (long version). *Physics and Society,* arXiv:physics/0512106v1 [physics.soc-ph]

Reingold, E., & Tilford, J., 1981. Tidier Drawing of Trees. *IEEE Trans. on Software Engineering*, SE-7(2):223–228.

Web sites

ELIXIR, SHISEIDO: http://www.shiseido.co.jp/eis/index.htm

MHLW. (2007). *Drug Legislation*. Ministry of Health, Labour and Welfare, Japan: http://wwwhourei.mhlw.go.jp/hourei/index.html

SEKKISEI, KOSÉ: http://www.sekkisei.com/jp/

Software

R Development Core Team. 2007. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Csardi, G. 2006. *IGraph Library*. http://cneurocvs.rmki.kfki.hu/igraph/.

APPENDIX

Centrality measures--*Degree* measures the centrality of a node in terms of the number of other nodes directly connected to it, free from the size of a network. Both *closeness* and *betweenness* are based on the geodesics, the shortest paths, among pairs of nodes. The former is the inverse of the sum of geodesics of a given node to all other nodes. To adjust for the size effect, it is multiplied by the size minus 1. The latter is based on the probabilistic notion that that the centrality of a given node increases as a function of the times it falls on the geodesics of pairs of all other nodes. The index is calculated by the ratio of the sum of the probability over all the pairs, excluding the given node, to the maximum value for the given size of the network.

The three indexes coincide in the case of the central node of a wheel- or a star-like network in that the central one a) has a greater number of links than others, b) is located at the minimum distance from all others, and c) is maximally close to others (Freeman, 1979). Note that closeness is not applicable, in its crude form, to the network comprised of separate groups of nodes.

What makes *PageRank* (Brin and Page, 1998) distinct from others is its recursive nature. That is, the importance of a node depends on the importance of nodes connected to it., and the importance of these nodes further depend on nodes connected to them. Although the algorithm was originally designed for directed graphs, it can be applied to the undirected one by treating links as bidirectional.