Department of Social Systems and Management
Discussion Paper Series

No.1204

Structural Comparisons of the Semantic Interface of the Top Cosmetic Brands on the Web by Network Analysis

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 Interfaces and Human Computer Interaction 2008 (*IHCI*);
Amsterdam, Netherlands, July, 25-27, 2008

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Discussion Paper Series, No. 1204, Department of Social Systems and Management, University of Tsukuba, 2008

STRUCTURAL COMPARISONS OF THE SEMANTIC INTERFACE OF THE TOP COSMETIC BRANDS ON THE WEB BY NETWORK ANALYSIS

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

ABSTRACT

Web presentations of consumer products serve as an interface between businesses and consumers. The present paper focused on the verbal information on the webs of the top two skincare brands in Japan. *Term* by *context* matrices created from the texts were truncated by SVD to construct graphs that represented high similarities among terms. In pursuit of essential terms and the networks surrounding them, the primary clusters were decomposed in to communities in which the core nodes and their neighborhoods were identified. Though the brands differed greatly in concentrations of clusters, communities and neighborhoods, they showed similarities in the properties of cores. Of particular interest was the core term "plump" that appeared in both brands. Probably, this reflected the shared business awareness of the central concern about skincare among the Japanese consumers. In lieu of conclusions, our plan of further inquiries were stated.

KEYWORDS

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

1. INTRODUCTION

As vital part of marketing, manufactures present their products on the web to appeal to consumers in nicely phrased terms, often accompanied by audio-visual images. Those web pages serve as the semantic interface between consumers who seek for convincing information compatible their various concerns and businesses which translate their interests and products into expressions suitable to web.

Among others, cosmetics provide interesting cases in terms of interactions of multiple concerns, ranging from health to beauty 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 products among young Japanese women.

The present study, as an exploratory investigation of the semantic web-interface, will focus on the skincare category in view of the fact that it is an entry to cosmetics and remains for a long time to be essential by itself and as a precondition to makeup. Moreover, labeling of skincare products that are also drugs (e.g., FDA, 2007) are regulated differently from ordinary ones. It naturally constrains their web presentations.

In the Japanese market, the drug legislation and other related regulations permit displays of approved medical effects for cosmetics-drugs, leaving displays of the pertaining ingredients as optional owing to

rigorous testing in advance. In contrast, the labeling of ordinary cosmetics should carry no medical effect but should list up major ingredients (MHLW, 2007). In short, designing of cosmetics labels on the web as well as packages is almost an art under multiple constraints.

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

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. Owing to 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 s_i**:**

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

Interpretations according to these equations allow principal component type analysis as carried out by our companion study (Matsuda, Machida and Mizuno, 2008). Equally interesting is the approach based on the truncation of matrix $\boldsymbol{A}$. By restricting to, for instance, largest k singular values, one may compute a truncated matrix $\boldsymbol{A}_{\mathrm{mxn}}$* from $\boldsymbol{U}_{mxk}\boldsymbol{S}_{kxk}\ \boldsymbol{V}_{kxn}^{\mathrm{T}}$. Given the sufficiency of k, $\boldsymbol{A}$* is known to be a good estimate of $\boldsymbol{A}$ devoid of noise.

In the LSA/LSI approaches, a frequency matrix $\boldsymbol{A}$ (*term* x *context*) is subjected to SVD, where a *term* is a word or a combination of words and a *context* is an entity in which the term occurs. Similarities of terms (or contexts), for instance, are measured by cosines between the corresponding rows (or columns) of $\boldsymbol{A}$*. There are various techniques to adjust the original frequencies prior to SVD such as tf-idf (see, Dumais, 1991).

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 nicely blended with non-poetic ones in various syntactical forms.

Therefore, the texts were segmented into constituent terms that may be words, particles, or their combinations on the basis of the field-expertise knowledge. Also, the contexts within which they appeared were 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 are growing among network researchers, besides the development of layout algorithms for visualization. Our analysis is in line with this trend, but it sets the principal aim at finding out core nodes and their neighborhoods in large networks where the coreness is defined as the high centralities accompanied by the agreement among the indexes, i.e., degree, PageRank (Brin and Page, 1998), closeness and betweenness (see the Appendix for the explanation).

Since the cores should be located in the main part of a given network, we will first select the largest subnetwork, and then, further dived them into various size of communities. The cores are to be found in the largest community. According to Freeman (1979), a cluster refers to a group of connected nodes among which every node is reachable from other member(s) but not from outside. A community is a group of nodes within which nodes are more closely related than to those outside the group. Hence, a cluster is an extreme form of a community.

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 $\boldsymbol{A}$'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 order to truncate the matrices, 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, the new matrices $\boldsymbol{A}^*$'s were derived from the results of SVD as

$$\boldsymbol{A}^*_{m\times 40} = \boldsymbol{U}_{m\times k}\boldsymbol{S}_{k\times k}\boldsymbol{V}^{\mathrm{T}}_{k\times 40}$$

where *m*=245 and *k*=15 for SEKKISEI, and *m*=204 and *k*=18 for ELIXIR.

Network representation--In the subsequent network analysis, terms and their mutual relationships were represented by nodes and links. To explore core patterns, we chose substantial relationships in strength as measured by Pearson's correlation coefficients, $.8 \leq |r| \leq 1$. Isolated nodes with no links were to be omitted from the networks.

Network indexes—Our main overall indexes were size and density defined as the number of nodes and the ratio of the number of links to the maximally possible number of links, respectively. Centrality of nodes were measured by four indexes—degree, PageRank, closeness and betweenness (see the Appendix for brief explanations.).

3. RESUTLS

For the sake of consistency, the results were presented in order of SEKKISEI and ELIXIR in this section unless otherwise noted. For the network analysis and layout, we employed the *igraph* library for the statistical package *R*. As for the layout, Fruchterman-Reingold's (1991) algorithm was used in consideration of the ease of visual inspection of the networks as compared to other alternatives. Since it starts with the randomly generated state, the produced layout should be topologically interpreted.

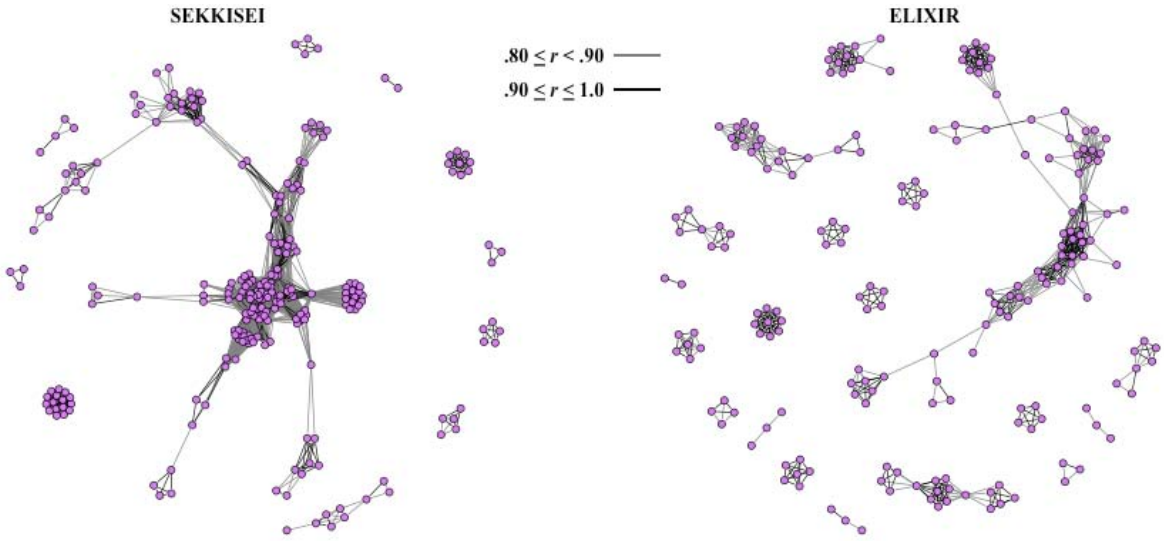

Figure 1. Initial networks comprised of non-isolated nodes by brand

3.1 Communities in the primary clusters

The selection of highly correlated links yielded networks of 243 and 196 nodes with 2,473 and 646 links that amounted to the densities .084 and .034 for SEKISEI and ELIXIR (see Figure 3), after the removal of two and eight isolated nodes from the respective data. The correlations were all positive (i.e., $r \geq .8$).

3.1.1 Primary clusters

The networks of the two brands were comprised of 11 and 19 separate clusters with varying size and density. The nodes of SEKKISEI were more concentrated in the primary cluster in comparison to ELIXIR. More specifically, the largest cluster of the former contained 183 nodes (75.3% of all) with density .136, whereas that of the latter was smaller in size (74 nodes--37.8% of all) and density (.124). It required six more clusters for the latter to accumulate the comparable share of nodes (74.5%).

The size of the second largest clusters of the two brands were 16 (6.6%) and 18 (9.2%). Because of the large differences between the primary and secondary clusters, only the primary ones of them were subjected to the subsequent analysis that proceeded in two stages—community decomposition and core neighborhood identification.

3.1.2 Community detection in the primary clusters

In order to facilitate configurative comparisons with the initial networks, we employed the ege-betweenness community algorithm (Newman and Girvan, 2004) according to which links connecting relatively dense groups of nodes would be likely to have high edge-betweenness as all the shortest paths from nodes in one group to another should traverse through them. After removing an edge with the highest edge-betweenness score, it computes the score and repeats the process until it reaches a specified condition. Therefore, the removal of an edge may or may not produce separate communities.

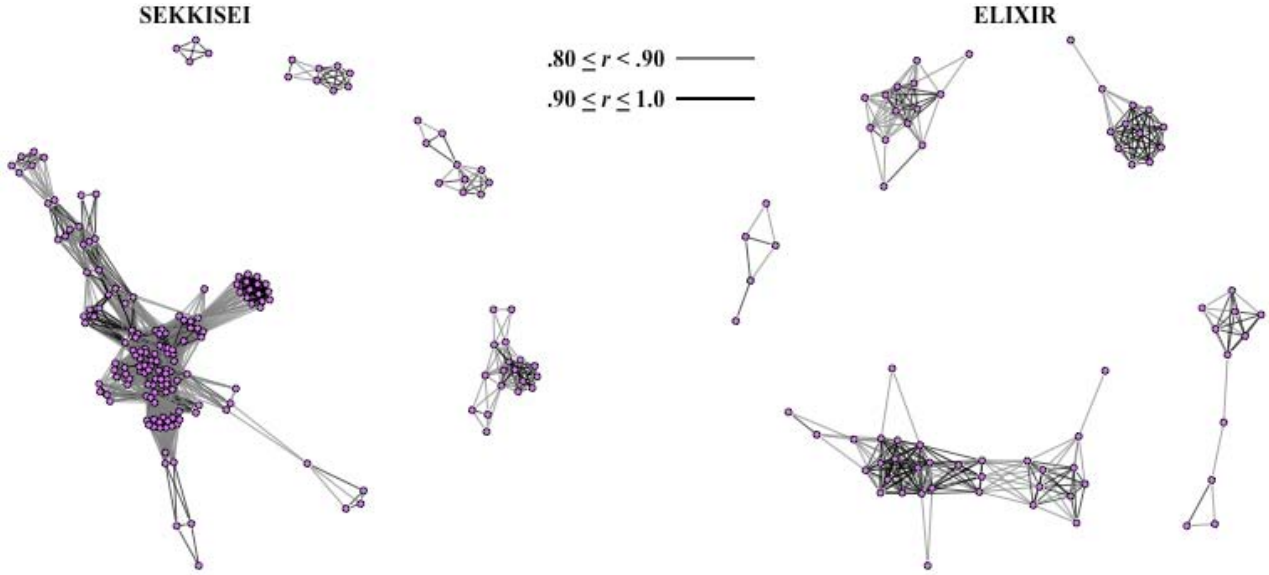

Figure 2. Five communities identified in the primary cluster by brand

In the absence of well-agreed termination criteria, we continued the process while the largest community maintained observable properties of the primary clusters to secure interpretability. Shown in Figure 2 are the five communities of each brand identified after 13 and 17 steps of edge removal.

Like in the primary cluster, the nodes were highly concentrated on the chief community in SEKKISEI consisted of 78.1% (n=143) of nodes followed by smaller communities whose relative proportions were 9.8,

5.5, 4.4 and 2.2%, whereas they were much less concentrated in ELIXIR as reflected in the relative proportions—43.2 (n=32), 18.9 (14), 16.2 (12), 14.9 (11) and 6.8% (5) in decreasing order.

In addition, the brands differed in the relationships between community size and density. While the density increased as the size decreased in SEKKISEI—.210, .497, .511, .714 and 1.00 in descending order of the community size, there was no such clear-cut tendency in ELIXIR— .345, .648, .773, .455 and .660 in the same order. The chief communities were subjected to the ensuing analysis.

3.2 Core nodes and their neighborhoods

3.2.1 Identification of core nodes

The centrality indexes of the chief communities were highly correlated among degree, PageRank and closeness in both brands as reflected in the Pearson's and Spearman's coefficients, i.e., $.803 \leq r \leq .983$ and $.830 \leq \rho \leq .955$. Agreement of betweenness with the other indexes in the Pearson's coefficients were small to moderate in SEKKISEI—i.e., $.382 \leq r \leq .537$, but moderate in ELIXIR-—i.e., $.523 \leq r \leq .587$. The Spearman's coefficients pertaining to betweenness were moderate in both brands, ranging from .635 to .743, and from .550 to .702.

Table 1. The centrality ranks of the core nodes of the chief communities by brand

SEKKISEI					ELIXIR				
Core node	degree	PageR	clsn's	btwn's	Core node	degree	PageR	clsn's	btwn's
A: non-, un-	1	1	1	1	A: supple	1	1	1	1
B: plump	2.5	2	2.5	13	B: plump	2	2	2	4
C: fragrance	2.5	3	2.5	18	C: condition	3.5	4	3	9
D: retain	4	4	4.5	7					

Notes: a) Decimals resulted from ties in ranking; b) "PageR", "clsn's" and "btwn's" denote PageRank, closeness and betweeness, respectively.

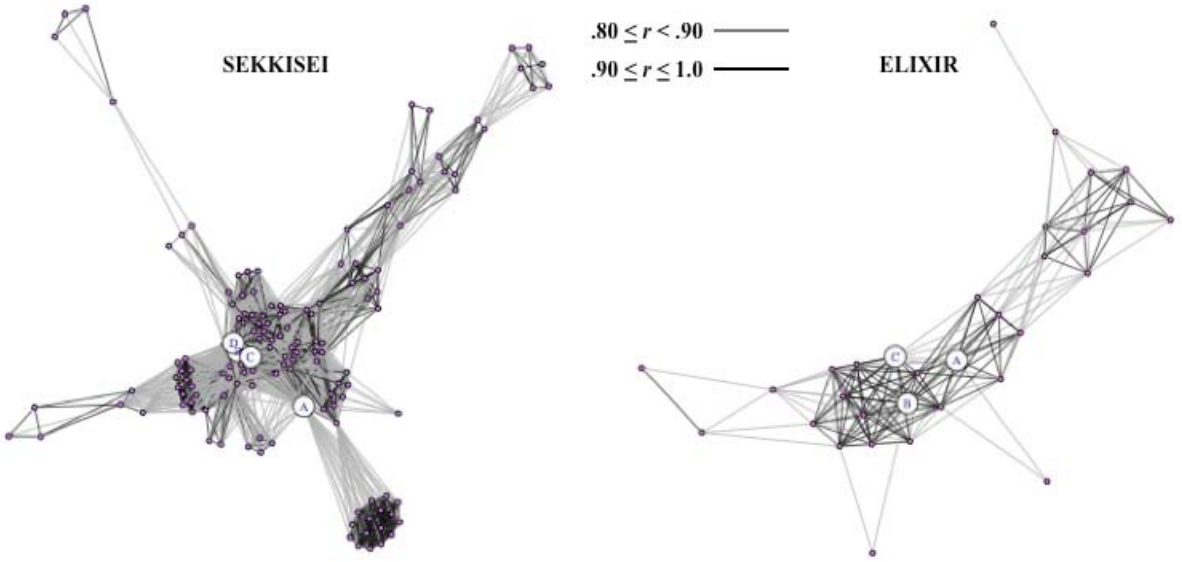

Figure 3. Core nodes in the chief community by brand
Note: See Table 1 for the terms of the core nodes shown in alphabets.

We selected the core nodes on the basis of high rankings as well as agreement among indexes within the range of .5 or 1. The rankings shown in Table 1 were identical or nearly identical except for betweenness. Nevertheless, the top node of each brand was exactly the same across four indexes. Interestingly, the second-

top node termed "plump" appeared in both brands. See Figure 3 for the locations of the core nodes in the chief community by brand.

3.2.1 Core neighborhoods

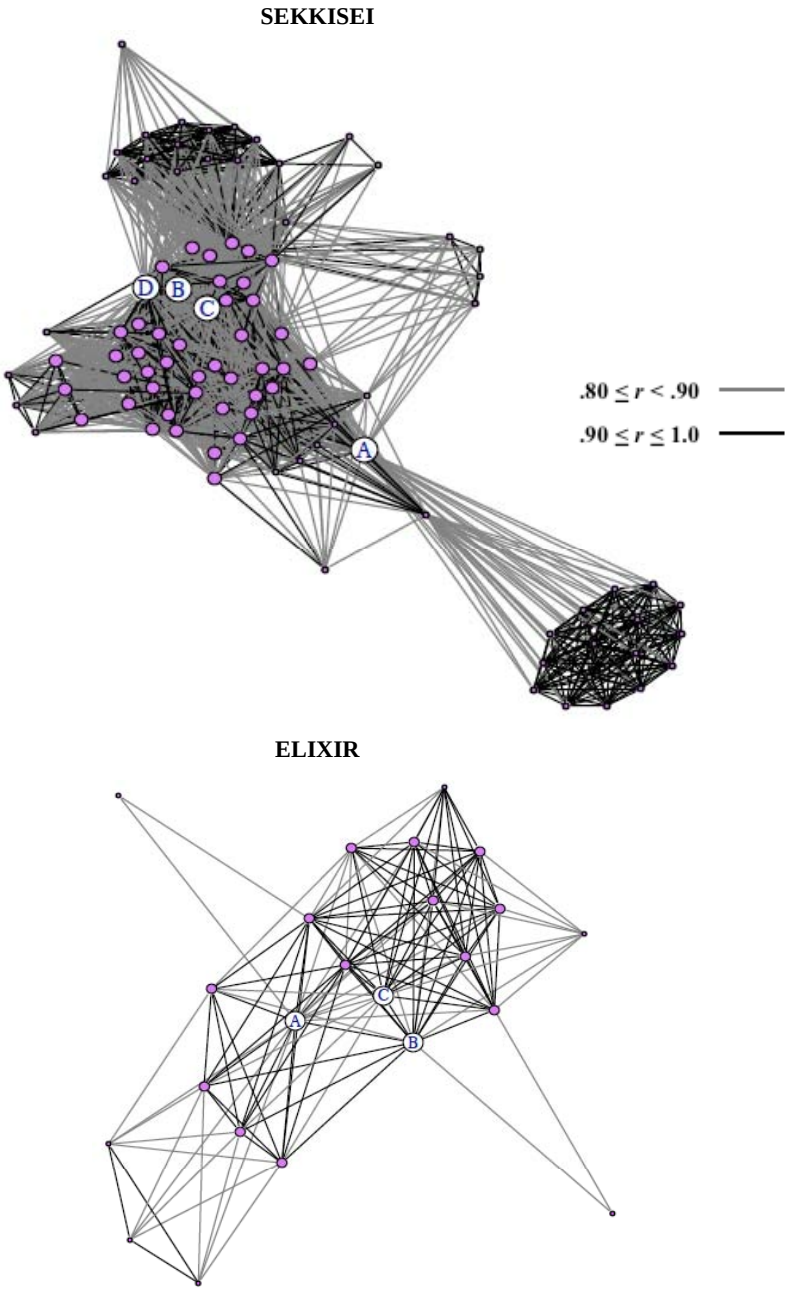

Figure 4. The combined neighborhoods produced by union of those around the individual cores

Note: Large nodes belong to the intersection of individual neighborhoods.

For both brands, all the cores were closely connected to each other, being the members of the other's neighborhood. This led us to form broad and narrow neighborhoods by taking the union and the intersection of the individual neighborhoods around the cores. The size (and, the density) of the union and the

intersection neighborhoods were 96 (.381) and 47 (.873) for SEKKISEI in contrast to 23 (.522) and 16 (.775) for ELIXIR. The greater reduction in size (51.0 vs. 30.4%) between the brands were accompanied by larger increase in density (129.1 vs. 48.5%), reflecting branched out as opposed to meshed structures of the chief communities in which the neighborhoods are embedded. Shown in Figure 4 are the combined neighborhoods produced by union of those around the individual cores with emphasis on the nodes comprising the intersection.

4. DISCUSSION AND CONCLUSION

Recent progress in graph analysis and visualization techniques enabled us to explore the semantic interface on the web designed for the top Japanese skincare brands. We have treated the term "semantics" in the dual sense—one denoting the conventional relationships between individual 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. 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 this dual sense. Ideally, one could incorporate audio and/or visual information in the semantic studies.

Our pursuit of cores and their neighborhoods revealed both sharp contrast and commonalities between the networks of the two brands. The most noteworthy contrasts were found in the extent and manner of concentration of nodes in clusters, communities and neighborhoods, whereas the basic properties of the cores were shared by both brands. It is noteworthy that the core term "plump" appeared in both brands, probably, stemming from the shared business awareness of the consumers' central concern about skincare. Had we investigated non-central clusters and non-primary communities, other patterns of interest could have emerged.

Our approach is expectedly valuable to web-designers who need to examine whether the essential information is both explicitly and implicitly organized as planned across related pages, for instance, by examining the cores and their neighborhoods. Besides, they can learn from focus group discussions about the informativeness of noteworthy sub-graphs, be they core neighborhoods, cliques or clusters. Market analysts may also benefit from explicating marketing intentions of competing businesses in lieu of conventional content analysis. In addition, consumers can have alternative means for learning about the products they are interested in, given the development of user-friendly tool to manipulate graphs.

In our companion study (Matsuda, Machida and Mizuno, 2008), we extracted different kinds of cores from the networks created from $s_i \boldsymbol{u}_i$ relationships and found that the top most cores of the two brands both pertained to moisturizing ingredient(s) that would realize the major physiological effect. The two studies together have shed light on the semantic interface from different angles.

Since no wok can exhaust possibilities, we need to conclude with the remarks on our plan of further exploration: a) analysis of non-central but cohesive communities that may bear significance of their own, b) testing other similarity measures (see Duarte, 1999), c) testing other community-search algorithms among which the walktrap (Pons and Latapy, 2005) and fast-greedy optimization of modularity (Clauset, Newman and Moore, 2004) methods are of particular interest, and, finally, d) integration of the findings produced by the two types of applications of SVD.

REFERENCES

Journal

Berry MW., 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.

Freeeman, 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.

Landauer, T.K., Foltz, P.W., & Laham, D. 1998. Introduction to latent semantic analysis. *Discourse Processes*, Vol.25, pp.259-284.

Matsuda, N., and Namatame, M. 1995. Interactive measurement of hierarchically related consumers' images. *Behaviormetrika*, Vol.22, pp.129-143.

Matsuda, N., Machida, A., & Mizuno, K. 2008. Visual inspections of the semantic interface on the web: A comparison of top cosmetic brands by SVD and network. In *Proceedings of IADIS Interfaces and Human Computer Interaction* 2008 (*IHCI*); Amsterdam, Netherlands, July, 25-27, 2008. A long version is available as *Discussion Paper*, No. 1205, Department of Social Systems and Management, University of Tsukuba.

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]

Web sites

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

FDA. (2007). *Cosmetics*. Center for Food Safety and Applied Nutrition. U.S. Food and Drug Administration.: http://www.cfsan.fda.gov/~dms/cos-toc.html

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.