Hierarchical Structure in Semantic Networks of Japanese Word Associations *

Maki Miyake[a], Terry Joyce[b], Jaeyoung Jung[c], and Hiroyuki Akama[c]

[a]Osaka University, 1-8 Machikaneyama-cho, Toyonaka-shi, Osaka, 560-0043, Japan
[b]Tama University, 802 Engyo, Fujisawa-shi, Kanagawa-ken, 252-0805, Japan
[c]Tokyo Institute of Technology, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan

mmiyake@lang.osaka-u.ac.jp
terry@tama.ac.jp
{catherina, akama}@dp.hum.titech.ac.jp

Abstract. This paper reports on the application of network analysis approaches to investigate the characteristics of graph representations of Japanese word associations. Two semantic networks are constructed from two separate Japanese word association databases. The basic statistical features of the networks indicate that they have scale-free and small-world properties and that they exhibit hierarchical organization. A graph clustering method is also applied to the networks with the objective of generating hierarchical structures within the semantic networks. The method is shown to be an efficient tool for analyzing large-scale structures within corpora. As a utilization of the network clustering results, we briefly introduce two web-based applications: the first is a search system that highlights various possible relations between words according to association type, while the second is to present the hierarchical architecture of a semantic network. The systems realize dynamic representations of network structures based on the relationships between words and concepts.

Keywords: Network analysis, Graph clustering, Japanese word associations.

1. Introduction

As an approach to deepening our understanding of lexical knowledge, many areas of cognitive science, including psychology and computational linguistics, are seeking to unravel the rich networks of associations that connect words together. Key methodologies for that enterprise are the techniques of graph representation and their analysis that allow us to discern the patterns of connectivity within large-scale resources of linguistic knowledge and to perceive the inherent relationships between words and word groups.

* This research has been supported by the 21st Century Center of Excellence Program "Framework for Systematization and Application of Large-scale Knowledge Resources". The authors would like to acknowledge here the generosity of the Center. The first and second authors have been supported by Grants-in-Aid for Scientific Research from the Japanese Society for the Promotion of Science (research project number 19700238 to first author and 18500200 to the second). In addition, the authors wish to express their thanks to Professor Shun Ishizaki for permission to use his Associative Concepts Dictionary in this study.

Although studies applying versions of the multidimensional space model, such as Latent Semantic Analysis (LSA) and multidimensional scaling, to the analysis of texts have been fairly fruitful, the methodologies of graph theory and network analysis are particularly suitable for elucidating the important characteristics of semantic networks.

Recently, a number of studies have applied graph theory approaches in investigating linguistic knowledge resources (Church and Hanks, 1990; Dorow, Widdows, Ling, Eckmann, Danilo and Moses, 2005; Steyvers and Tanenbaum 2005; van Dongen, 2000; Watts and Strogatz, 1998). For instance, Dorow, et al (2005) utilize two graph clustering techniques as methods of detecting lexical ambiguity and of acquiring semantic classes instead of word frequency based computations.

This paper applies graph theory and network analysis methods to the analysis of semantic network representations of Japanese word associations. After briefly outlining the two separate Japanese word association databases used—the Associative Concept Dictionary (Okamoto and Ishizaki, 2001) and the Japanese Word Association Database (Joyce, 2005, 2006, 2007)—the paper calculates some basic statistical features, such as degree distributions, clustering coefficients and the average clustering coefficient distribution for nodes with degrees. We also apply the recently developed Recurrent Markov Clustering (RMCL) algorithm (Jung, Miyake and Akama, 2006) which enhances the bottom-up classification method of the basic MCL algorithm by making it possible to adjust the proportion in cluster sizes. Given this greater control over cluster sizes, the RMCL clearly provides a very appealing approach to the automatic construction of condensed network representations, which, in turn, can facilitate the creation of hierarchically-organized semantic spaces as a way of visualizing large-scale linguistic knowledge resources.

2. Building Semantic Network Graphs of Japanese Word Associations

This section outlines the semantic network representations of the Japanese word association databases. Specifically, the section briefly describes two separate databases of Japanese word associations—the Associative Concept Dictionary (ACD) and the Japanese Word Association Database (JWAD)—and the semantic network representations created from them.

2.1.Existing word association norms

As frames of reference concerning the scales of the two Japanese word association databases, it worth noting that large-scale, comprehensive word association normative data has existed for some time for English. For example, Moss and Older (1996) collected between 40-50 responses for some 2,400 words of British English, while Nelson, McEvoy and Schreiber (1998) compiled perhaps the largest database of American English covering some 5,000 words with approximately 150 responses per item. Notwithstanding the early survey by Umemoto (1969), which gathered free associations from 1,000 university students for a very small set of 210 words, clearly there has been a serious lack of comparative databases of Japanese word associations. Both the ACD and the JWAD seek to redress this situation, especially the ongoing JWAD project which is committed to constructing a large-scale database for its current survey corpus of 5,000 basic Japanese kanji and words.

2.2.Associative Concept Dictionary

Okamoto and Ishizaki (2001) created the Associative Concept Dictionary (ACD), which is organized as a hierarchal structure of higher/lower level concepts. The data consists of 33,018 word association responses provided by 10 respondents according to specified response categories for 1,656 nouns. By excluding response words with a frequency of 1 and a clustering coefficient of 0, 9,373 words were selected for use in creating a semantic network representation.

2.3.Japanese Word Association Database

The Japanese Word Association Database is being constructed as part of a project to investigate lexical knowledge in Japanese by mapping out Japanese word associations (Joyce, 2005; 2006; 2007). While the particular task—specifying in advance the associative relationship for responses—employed in creating the ACD can arguably be justified in terms of constructing a dictionary of associated concepts, the data provides little insight into the rich and diverse nature of word associations. Accordingly, the JWAD employs the free word association task in collecting association responses. Also in contrast to the ACD, which only examined nouns, the JWAD is surveying words of all word classes. Version 1 of the JWAD consists of a random sample of 2,099 items from the survey corpus of 5,000 basic Japanese kanji and words that were presented to up to 50 respondents. For the JWAD network, only words with a frequency of 2 or more were selected, which resulted in set of 7,966 words to be clustered.

3. Analyses of the Network Structures

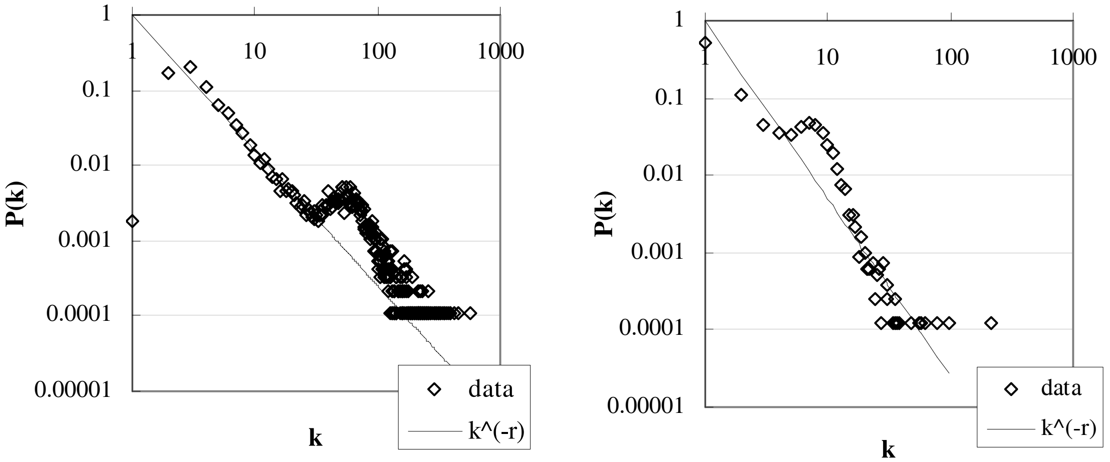
As already suggested, graph representations and the techniques of graph theory and network analysis are particularly promising techniques with which to examine the intricate patterns of connectivity within large-scale linguistic knowledge resources. For instance, Steyvers and Tenenbaum (2005) conducted a noteworthy study that examined the structural features of three semantic networks. By calculating a range of statistical features, including the average shortest paths, diameters, clustering coefficients, and degree distributions, they observed interesting similarities between the three networks in terms of their scale-free patterns of connectivity and small-world structures.

Following their basic approach, we analyze the characteristics of the two semantic network representations of Japanese word associations by calculating the statistical features of degree distribution and clustering coefficient—an index of the interconnectivity strength between neighboring nodes in a graph.

3.1.Degree distribution

From their computations of degree distributions, Balabasi and Albert (1999) suggest that the degree distribution, P(k), for scale-free network structures will correspond to a power law, which can be expressed as $P(k) \approx k^{-r}$.

Figure 1 presents degree distributions for word occurrences in the two semantic networks, which indicate that P(k) conforms to a power-law in both cases (with exponent values, r, of 1.8 for the ACD (panel a) and 2.3 for the JWAD (panel b). In the case of the ACD, the average degree value is 19.96 (0.2%) for the complete semantic network of 9,373 nodes, while the average degree value is 3.67 (0.05% for 7,966 nodes) in the JWAD's case. The results clearly indicate that the networks exhibit a pattern of sparse connectivity; in other words, that they possess the characteristics of a scale-free network.



(a). ACD

(b). JWAD

Figure 1: Degree distributions for the two semantic networks

3.2.Clustering coefficient

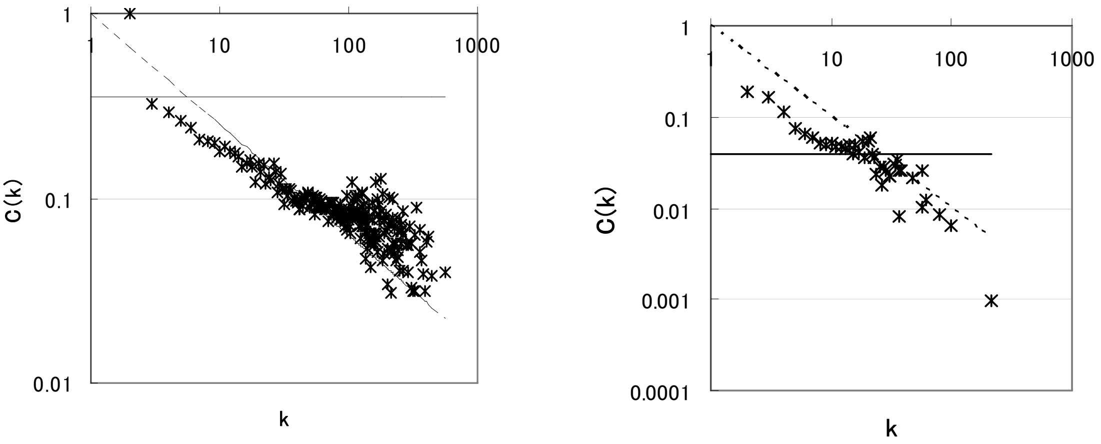
In their social network study investigating the probabilities that an acquaintance of an acquaintance is also an acquaintance of yours, Watts and Strogatz (1998) advocate the notion of clustering coefficient as an appropriate index of the degree of connections between nodes. In this study, we define the clustering coefficient of n nodes as:

$$C(n) = \frac{\text{number of links among n's neighbors}}{N(n) \times (N(n) - 1) / 2}$$

where N(n) represents the number of adjacent nodes. Accordingly, a clustering coefficient is a value between 0-1.

Moreoever, Ravasz and Barabasi (2003) introduce the notion of clustering coefficient dependence on node degree as an index of the hierarchical structures found in real networks—such as the WWW, the Actor Network based on the www.IMDB.com database—which are based on the hierarchical model of $C(k) \approx k^{-1}$ (Dorogovski, Goltsev, & Mendes, 2001). Specifically, the hierarchical nature of a network can be characterized by using the average clustering coefficient, C(k), of nodes with k degrees, which will follow a scaling law such as $C(k) \approx k^{-\beta}$, where β is defined as a hierarchical exponent.

Figure 2 presents results of scaling C(k) with k for (a) ACD and (b) JWAD. The dashed line in (a) has a slope of -1, while the fitting exponent, β, is 0.6 for JWAD. The solid lines correspond to the average clustering coefficient. In the case of the ACD, the average clustering coefficient is quite high at 0.35, which can be regarded as indicating the small-world property. In the case of the JWAD, the average clustering coefficient is 0.04, which indicates that the complete network basically consists of many star graphs connected together. As both networks conform well to a power law, we may conclude that both networks have intrinsic hierarchies.



(a). ACD

(b). JWAD

Figure 2: Clustering coefficient distributions for the two semantic networks

4. Graph Clustering: Recurrent Markov Clustering

4.1.Algorithm

Jung, et al. (2006) have recently proposed an improvement to Markov Clustering (MCL), called Recurrent Markov Clustering (RMCL), which provides for greater control over the sizes of clusters by making it possible to adjust graph granularity and, thus, the generality of concepts. MCL is an effective method for the detection of patterns and clusters within large and sparsely connected data structures. The first step in the MCL consists of sustaining a random walk across a graph by ‘expansions’. The recurrent process incorporates feedback about the states of overlapping clusters prior to the final MCL output stage. This reverse tracing procedure is a key feature of the RMCL making it possible to generate a virtual adjacency matrix for non-overlapping clusters based on the convergent state that emerges from the MCL process. The resultant condensed matrix provides a simpler graph that can highlight the conceptual structures that underlie similar words.

4.2.Results

The RMCL algorithm is realized as a series of calculations executed with gridMathematica. Taking the JWAD as an example of the calculation steps in the RMCL, Figure 3 presents the transition in cluster sizes as a function of the MCL process. Starting from the adjacency matrix for co-occurrences, the MCL process finally generated a nearly-idempotent stochastic matrix at the 19th clustering stage with 1,441 hard clusters, where the average number of cluster components is 5.6 with a standard deviation (SD) of 3.1. In contrast, the RMCL resulted in just 759 hard clusters with an average of 1.9 cluster components (SD = 1.5). Among the representative nodes for RMCL clusters, 1,176 nodes (83%) were found to be words that had been presented as stimulus words. Figure 4 presents MCL and RMCL cluster sizes for both the ACD and the JWAD, which illustrate the transitions occurring in downsizing the networks generated from graph clustering. Figure 5 plots the number of components for both MCL and RMCL clusters as a function of frequency. In the case of the ACD, the MCL resulted in 1,408 hard clusters (average cluster size = 6.7, SD = 8.6), while the RMCL resulted in 118 hard clusters, where the average number of cluster components was 11.9 with a rather high SD of 68.6.

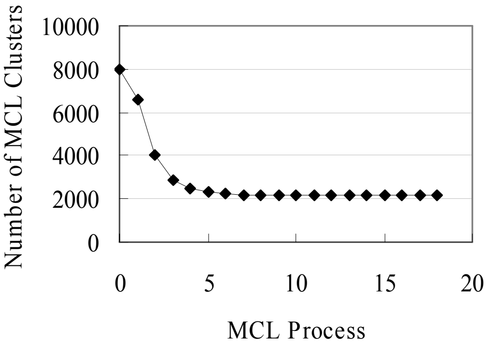


Figure 3: Cluster size transitions during MCL process

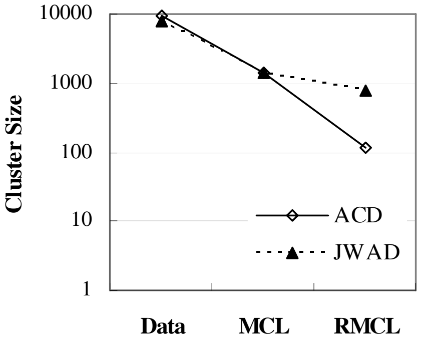
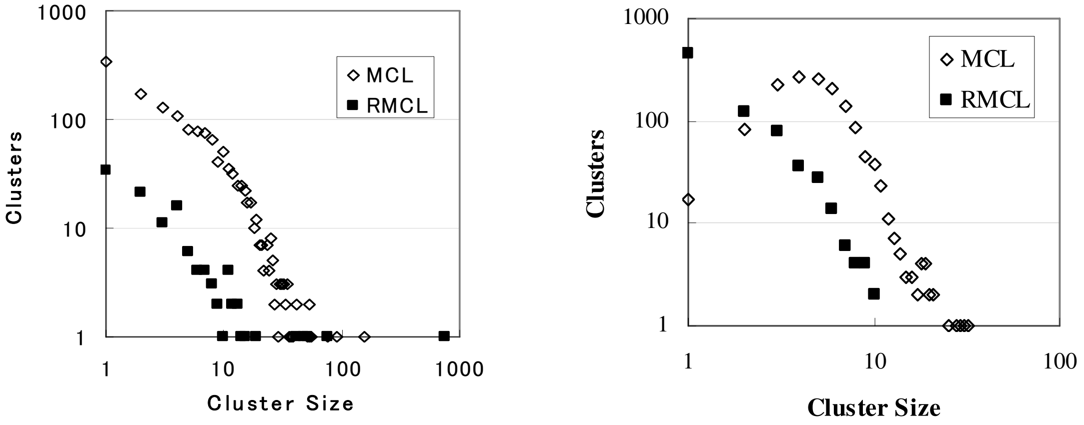


Figure 4: Cluster sizes for MCL and RMCL



(a). ACD

(b). JWAD

Figure 5: Component size distributions for both the MCL and the RMCL

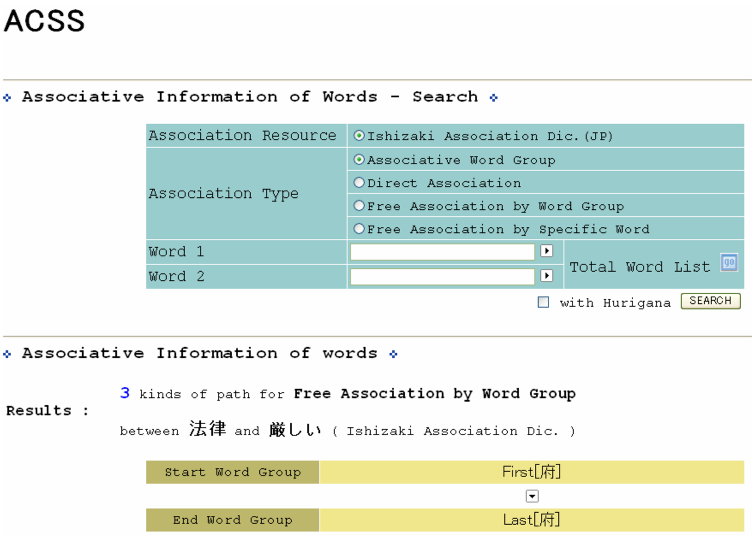
5. Applications of the RMCL

As Widdow, Cederberg, and Dorow (2002) astutely observe, graph visualization is a particularly powerful tool for representing the meanings of words and concepts. In order to utilize the MCL and RMCL clustering results of the networks, we have developed two web-based applications implemented by webMathmatica: the first is an 'Associative Composition Support System (ACSS)' to search for free association words according to different types of association information, while the second is 'RMCLnet' which elucidates the hierarchical architecture of large-scale networks.

5.1.The Associative Composition Support System

The free web-based ACSS proposed by Jung *et al.* (2006) seeks to promote associative thinking ability, and so, in turn, to foster language learning and creativity. ACSS is developed based on a database that makes it possible to retrieve three types of associative information such as word-based, concept-based and group-based associations. Such associative information is apparently sufficient to support system users in improving their associative thinking and creativity by encouraging them to move beyond literal, direct and superficial aspects to richer, freer, and more inspired conceptual associations. The variety of links between words can foster free, flexible, integrative, and imaginative thinking, while simultaneously encouraging users to discover the implicit relevance of words and even to occasionally fill in the semantic gaps between words with imaginative creations.

Figure 6 presents a screen shot of the main page for the ACSS system. Users can access the online system at http://atheneum.dp.hum.titech.ac.jp/semnet/ACSS/index.jsp. The entire interface on the user side is controlled by Javascript. When retrieval requirements are sent to the remote web server, search results are calculated in real-time by WebMathematica through the JSP and Mathematica kernel. The database was constructed in the form of a semantic network and is stored on the web server after calculating original Japanese word associations with GridMathematica. System users can input any two words to see three types of association information.



ACSS

❖ Associative Information of Words - Search ❖

Association Resource	● Ishizaki Association Dic. (JP)	
Association Type	● Associative Word Group	
	○ Direct Association	
	○ Free Association by Word Group	
	○ Free Association by Specific Word	
Word 1		Total Word List
Word 2		

□ with Hurigana

❖ Associative Information of words ❖

Results : 3 kinds of path for **Free Association by Word Group**

between 法律 and 厳しい (Ishizaki Association Dic.)

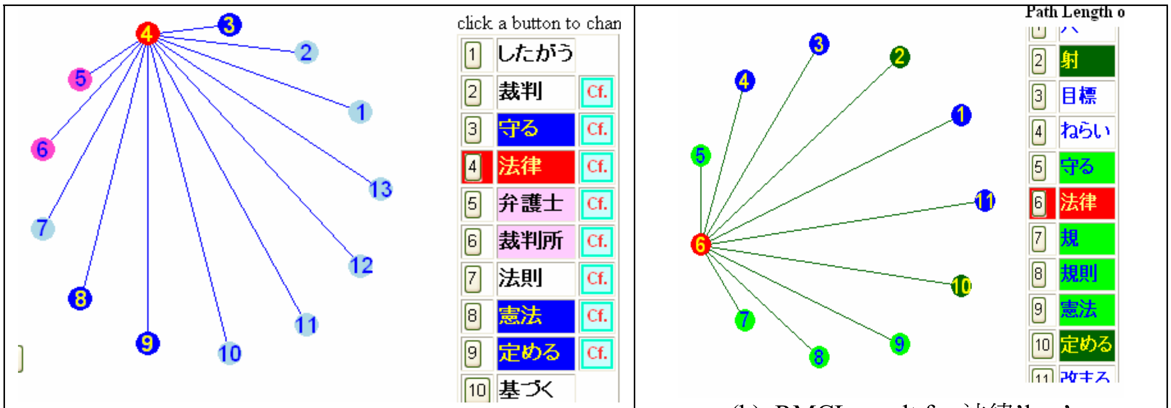
Start Word Group	First[府]
End Word Group	Last[府]

Figure 6: Screen shot of the GUI to the ACSS system

5.2.RMCLnet

Graph visualization of the semantic structures generated through MCL and RMCL clustering is implemented with webMathematica, employing basic techniques drawing on java servlet/JSP technology (Miyake, 2006). webMathematica can handle interactive calculations and visualization is realized by integrating Mathematica with a web server. The web server employs Apache2 as its http application server and Tomcat5 as a servlet/JSP engine. The URL for RMCLnet is http://perrier.dp.hum.titech.ac.jp/semnet/RmclNet/index.jsp.

Clustering results from both the MCL and RMCL processes can dynamically represent the relationships between words, with MCL components possibly corresponding to concepts (Figure 7). The implementation method is quite straightforward, as it is sufficient to simply store the multiple files that are created automatically when the RMCL process is executed. The system can simultaneously represent results for both the ACD and the JWAD, making it possible to examine the structural similarities and differences between the two semantic networks, which can yield interesting insights into the nature of word associations and how graph clustering functions.



click a button to chan

1	したがう	
2	裁判	Cf.
3	守る	Cf.
4	法律	Cf.
5	弁護士	Cf.
6	裁判所	Cf.
7	法則	Cf.
8	憲法	Cf.
9	定める	Cf.
10	基づく	

(a). MCL result for 法律 ’law’

Path Length o

[illegible]	[illegible]
2	射
3	目標
4	ねらい
5	守る
6	法律
7	規
8	規則
9	憲法
10	定める
11	[illegible]

(b). RMCL result for 法律’law’

Figure 7: Screen shot of the RMCLnet

6. Conclusions

In summary, this paper has reported on the application of graph clustering methodologies to the analysis of semantic network representations of Japanese word associations. After outlining two separate large-scale databases of Japanese word associations, the paper analyzed the characteristics of two semantic network representations of Japanese word associations. In addition to the calculation of degree distributions for the networks, which indicate that the networks are scale-free, average clustering coefficient distributions for nodes were found to conform to a power law, indicating that the networks have hierarchical organizations. Moreover, the ACD was found to have a high average clustering coefficient value, suggesting the small-world property, while the lower value for the JWAD network suggests it has less interconnectivity.

Finally, we briefly introduced two web-based applications as examples that utilize RMCL clustering results. The network representation application is useful in elucidating the structures within hierarchically-organized semantic spaces, which makes it an especially appealing approach to the visualization of large-scale linguistic knowledge resources.

References

Barabasi, A.L. and R. Albert. 1999. Emergence of scaling in random networks. *Science,* 286, 509-512.

Church, K.W. and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics,* 16, 22-29.

Dorogovtsev, S.N., A.V. Goltsev and J.F.F. Mendes. 2001. Pseudofractal Scall-free Web. *e-print cond-mat/0112143.*

Dorow, B., D. Widdows, K. Ling, J. Eckmann, D. Sergi and E. Moses. 2005. Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. *Proceeding of the Second Workshop of the Meaning Project.*

Jung, J., M. Miyake and H. Akama. 2006. Recurrent Markov Cluster (RMCL) Algorithm for the Refinement of the Semantic Network, *International Conference on Language Resources and Evaluation*. pp.1428-1432.

Jung, J., M. Miyake, N. Makoshi and H. Akama. 2006. Development of a Web-based Composition Support System: Using Graph Clustering Methodologies Applied to an Associative Concepts Dictionary. *Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies*, pp.431-435.

Joyce, T. 2005. Constructing a Large-scale Database of Japanese Word Associations. In K. Tamaoka, ed., *Corpus Studies on Japanese Kanji (Glottometrics 10)*. pp. 82-98. Hituzi Syobo and RAM-Verlag.

Joyce, T. 2006. Mapping Word Knowledge in Japanese: Constructing and Utilizing a Large-scale Database of Japanese Word Associations. *Proceedings of the Large-Scale Knowledge Resources Symposium*, pp.155-158.

Joyce, T. 2007. Mapping Word Knowledge in Japanese: Coding Japanese Word Associations. *Proceedings of the Large-Scale Knowledge Resources Symposium*, pp. 233-238.

Miyake, M. 2006. Implementing a Semantic Network of the Synoptic Gospels based on Graph Clustering, *IPSJ SIG Computers and the Humanities Symposium*, 161-165.

Moss, H. and Older L. 1996. *Birkbeck Word Association Norms*. Psychological Press.

Okamoto, J. and S. Ishizaki. 2001. Associative Concept Dictionary and its Comparison Electronic Concept Dictionaries. *PACLING2001,* 214-220.

Ravasz, E. and A.L. Barabasi. 2003. Hierarchical Organization in Complex Networks. *Physical Review E*, 67, 026112.

Umemoto, T. 1969. *Word Association Norms: Free Associations from 1,000 University Students.* (in Japanese). Tokyo Daigaku Shuppankai.

Steyvers, M. and J.B. Tenenbaum. 2005. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth, *Cognitive Science,* 29(1), 41-78.

van Dongen, S. 2000. *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht.

Vechthomova, O., D. Gfeller, J.-C. Chappelier and P. De Los Rios. 2005. Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. *International Symposium on Applied Stochastic Models and Data Analysis*, 106-113.

Watts, D. and S. Strogatz. 1998. Collective Dynamics of 'Small-world' Networks, *Nature*, 393, 440-442.

Widdows, D., S. Cederberg and B. Dorow. 2002. Visualisation Techniques for Analysing Meaning. *Proceeding of the Fifth International Conference on Text, Speech and Dialogue,* pp.107-115.