Capturing the Structures in Association Knowledge: Application of Network Analyses to Large-Scale Databases of Japanese Word Associations

Terry Joyce[1] and Maki Miyake[2]

[1] School of Global Studies, Tama University,
802 Engyo, Fujisawa, Kanagawa, 252-0805, Japan
terry@tama.ac.jp
[2] Graduate School of Language and Culture, Osaka University,
1-8 Machikaneyama-cho, Toyonaka-shi, Osaka, 560-0043, Japan
mmiyake@lang.osaka-u.ac.jp

Abstract. Within the general enterprise of probing into the complexities of lexical knowledge, one particularly promising research focus is on word association knowledge. Given Deese's [1] and Cramer's [2] convictions that word association closely mirror the structured patterns of relations that exist among concepts, as largely echoed Hirst's [3] more recent comments about the close relationships between lexicons and ontologies, as well as Firth's [4] remarks about finding a word's meaning in the company it keeps, efforts to capture and unravel the rich networks of associations that connect words together are likely to yield interesting insights into the nature of lexical knowledge. Adopting such an approach, this paper applies a range of network analysis techniques in order to investigate the characteristics of network representations of word association knowledge in Japanese. Specifically, two separate association networks are constructed from two different large-scale databases of Japanese word associations: the Associative Concept Dictionary (ACD) by Okamoto and Ishizaki [5] and the Japanese Word Association Database (JWAD) by Joyce [6] [7] [8]. Results of basic statistical analyses of the association networks indicate that both are scale-free with small-world properties and that both exhibit hierarchical organization. As effective methods of discerning associative structures with networks, some graph clustering algorithms are also applied. In addition to the basic Markov Clustering algorithm proposed by van Dongen [9], the present study also employs a recently proposed combination of the enhanced Recurrent Markov Cluster algorithm (RMCL) [10] with an index of modularity [11]. Clustering results show that the RMCL and modularity combination provides effective control over cluster sizes. The results also demonstrate the effectiveness of graph clustering approaches to capturing the structures within large-scale association knowledge resources, such as the two constructed networks of Japanese word associations.

Keywords: association knowledge, lexical knowledge, network analyses, large-scale databases of Japanese word associations, Associative Concept Dictionary (ACL), Japanese Word Association Database (JWAD), association network representations, graph clustering, Markov clustering (MCL), recurrent Markov clustering (RMCL), modularity.

T. Tokunaga and A. Ortega (Eds.): LKR 2008, LNAI 4938, pp. 116–131, 2008.

1 Introduction

Reflecting the central importance of language as a key to exploring and understanding the intricacies of higher human cognitive functions, a great deal of research within the various disciplines of cognitive science, such as psychology, artificial intelligence, computational linguistics and natural language processing, has understandably sought to investigate the complex nature of lexical knowledge. Within this general enterprise, one particularly promising research direction is to try and capture the structures of word association knowledge. Consistent with both Firth's assertion [4] that a word's meaning resides in the company it keeps, as well as the notion proposed by Deese [1] and Cramer [2] that, as association is a basic mechanism of human cognition, word associations closely mirror the structured patterns of relations that exist among concepts, which is largely echoed in Hirst's observations about the close relationships between lexicons and ontologies [3], attempts to unravel the rich networks of associations that connect words together can undoubtedly provide important insights into the nature of lexical knowledge.

While a number of studies have reported reasonable successes in applying versions of the multidimensional space model, such as Latent Semantic Analysis (LSA) and multidimensional scaling, to the analysis of texts, the methodologies of graph theory and network analysis are especially suitable for discerning the patterns of connectivity within large-scale resources of association knowledge and for perceiving the inherent relationships between words and word groups. A number of studies have, for instance, recently applied graph theory approaches in investigating various aspects of linguistic knowledge resources [9] [12], such as employing graph clustering techniques in detecting lexical ambiguity and in acquiring semantic classes as alternatives to computational methods based on word frequencies [13].

Of greater relevance to the present study are the studies conducted by Steyvers, Shiffrin, and Nelson [14] and Steyvers and Tenenbaum [15] which both focus on word association knowledge. Specifically, both studies draw on the *University of South Florida Word Association, Rhyme, and Word Fragment Norms*, which includes one of the largest databases of word associations for American English compiled by Nelson, McEvoy, and Schreiber [16]. Steyvers and Tenenbaum [14], for instance, applied graph theory and network analysis techniques in order to examine the structural features of three semantic networks—one based on Nelson, et al [16], one based on WordNet [17], and one based on Roget's thesaurus [18]—and observed interesting similarities between the three networks in terms of their scale-free patterns of connectivity and small-world structures. In a similar vein, the present study applies a range of network analysis approaches in order to investigate the characteristics of graph representations of word association knowledge in Japanese. In particular, two semantic networks are constructed from two separate large-scale databases of Japanese word associations: namely, the Associative Concept Dictionary (ACD) compiled by Okamoto and Ishizaki [5] and the Japanese Word Association Database (JWAD), under ongoing construction by Joyce [6] [7] [8].

In addition to applying some basic statistical analyses to the semantic network representations constructed from the large-scale databases of Japanese word associations, this study also applies some graph clustering algorithms which are effective methods of capturing the associative structures present within large and

sparsely connected resources of linguistic data. In that context, the present study also compares the basic Markov clustering algorithm proposed by van Dongen [9] with a recently proposed combination of the enhanced Recurrent Markov Clustering (RMCL) algorithm developed by Jung, Miyake, and Akama [10] and Newman and Girvan's measure of modularity [11]. Although the basic Markov clustering algorithm is widely known to be an effective approach to graph clustering, it is also recognized to have an inherent problem relating to cluster sizes, for the algorithm tends to yield either an exceptionally large core cluster or many isolated clusters consisting of single words. The RMCL has been developed expressly to overcome the cluster size distribution problem by making it possible to adjust the proportion in cluster sizes. The combination of the RMCL graph clustering method and the modularity measurement provides even greater control over cluster sizes. As an extremely promising approach to graph clustering, this effective combination is being applied to the semantic network representations of Japanese word associations in order to automatically construct condensed network representations. One particularly attractive application for graph clustering techniques that are capable of controlling cluster sizes is in the construction of hierarchically-organized semantic spaces, which certainly represents an exciting approach to capturing the structures within large-scale association knowledge resources.

This paper applies a variety of graph theory and network analysis methods in analyzing the semantic network representations of large-scale Japanese word association databases. After briefly introducing in Section 2 the two Japanese word association databases, the ACD and the JWAD, which the semantic network representations analyzed in this study were constructed from, Section 3 presents the results from some basic statistical analyses of the network characteristics, such as degree distributions and average clustering coefficient distributions for nodes with degrees. Section 4 focuses on methods of graph clustering. Following short discussions of the relative merits of the MCL algorithm, the enhanced RMCL version and the combination of RMCL and modality, the graph clustering results for the two association network representations are presented. Section 5 provides a short introduction to the RMCLNet web application which makes the clustering results for the two Japanese word association networks publicly available. Finally, Section 6 summarizes the results from the various graph theory and network analysis methods applied in this study, and fleetingly mentions some interesting directions for future research in seeking to obtain further insights into the complex nature of association knowledge.

2 Network Representations of Japanese Word Associations

This section briefly introduces the Associative Concept Dictionary (ACD) [5] and the Japanese Word Association Database (JWAD) [6] [7] [8], which are both large-scale databases of Japanese word associations. The two network representations of word association knowledge constructed from the databases are analyzed in some detail in the subsequent sections.

Compared to the English language for which comprehensive word association normative data has existed for some time, large-scale databases of Japanese word

associations have only been developed over the last few years. Notable normative data for English includes the 40-50 responses for some 2,400 words of British English collected by Moss and Older [19] and, as noted earlier, the American English norms compiled by Nelson and his colleagues [16] which includes approximately 150 responses for a list of some 5,000 words. Although the early survey by Umemoto [20] gathered free associations from 1,000 university students, the very limited set of just 210 words only serves to highlight the serious lack of comparative databases of word associations for Japanese that has existed until relatively recently. While the ACD and the JWAD both represent substantial advances in redressing the situation, the ongoing JWAD project, in particular, is strongly committed to the construction of a very large-scale database of Japanese word associations, and seeks to eventually surpass the extensive American English norms [16] in both the size of its survey corpus and the levels of word association responses collected.

2.1 The Associative Concept Dictionary (ACL)

The ACD was created by Okamoto and Ishizaki [5] from word association data with the specific intention of building a dictionary stressing the hierarchal structures between certain types of higher and lower level concepts. The data consists of the 33,018 word association responses provided by 10 respondents according for 1,656 nouns. While arguably appropriate for its dictionary-building objectives, a major drawback with the ACD data is the fact that response category was specified as part of the word association experiment used in collecting the data. The participants were asked to respond to a presented stimulus word according to one of seven randomly presented categories (hypernym, hyponym, part/material, attribute, synonym, action and environment). Accordingly, the ACD data tells us very little about the wide range of associative relations that the free word association task taps into.

In constructing the semantic network representation of the ACD database, only response words with a response frequency of two or more were extracted. This resulted in a network graph consists of 8,951 words

2.2 The Japanese Word Association Database (JWAD)

Under ongoing construction, the JWAD is the core component in a project to investigate lexical knowledge in Japanese by mapping out Japanese word associations [6] [7] [8]. Version 1 of the JWAD consists of the word association responses to a list of 2,099 items which were presented to up to 50 respondents [21]. The list of 2,099 items was randomly selected from the initial project corpus of 5,000 basic Japanese kanji and words. In marked contrast to the ACD and its specification of categories to which associations should belong, the JWAD employs the free word association task in collecting association responses. Accordingly, the JWAD data more faithfully reflects the rich and diverse nature of word associations. Also, in sharp contrast to the ACD, which only collected associations for a set of nouns, the JWAD is surveying words belonging to all word classes.

Similar to the ACD network graph, in constructing the semantic network representation of the JWAD, only response words with a frequency of two or more were selected. In the case of the JWAD, this resulted in a network graph consisting of 8,970 words, so the two networks are of very similar sizes.

3 Analyses of the Association Network Structures

This section reports on initial comparisons of the ACD network and the JWAD network based on some basic statistical analyses of their network structures.

Graph representation and the techniques of graph theory and network analysis are particularly appropriate methods for examining the intricate patterns of connectivity that exist within large-scale linguistic knowledge resources. As discussed in Section 1, Steyvers and Tenenbaum [15] have illustrated the potential of such techniques in their noteworthy study that examined the structural features of three semantic networks. Based on their calculations of a range of statistical features, such as the average shortest paths, diameters, clustering coefficients, and degree distributions, they argued that the three networks exhibited similarities in terms of their scale-free patterns of connectivity and small-world structures. Following their basic similar approach, we analyze the structural characteristics of the two association networks by calculating the statistical features of degree distribution and clustering coefficient, which is an index of the interconnectivity strength between neighboring nodes in a graph.

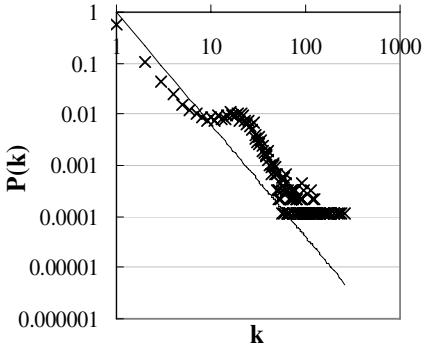
3.1 Degree Distributions

Based on their computations of degree distributions, Balabasi and Albert [22] argue that networks with scale-free structures have a degree distribution, P(k), that conforms to a power law, which can be expressed as follows:

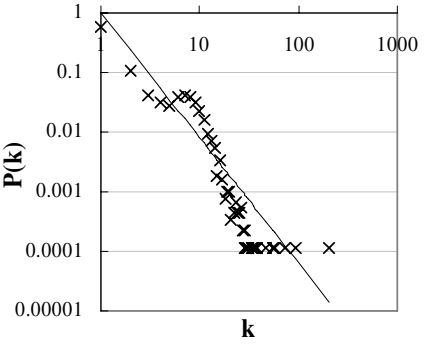
$$P(k) \approx k^{-r}$$

The results of analyzing degree distributions for the two association networks are presented in Figure 1. As the figure clearly shows, P(k) for both association networks conforms to a power law: the exponent value, r, is 2.2 for the ACD network (panel a) and 2.1 for the JWAD network (panel b).

For the ACD network, the average degree value is 7.0 (0.08%) for 8,951 nodes, while in the case of the JWAD network, the average degree value is 3.3 (0.03%) for



(a) ACD network



(b) JWAD network

Fig. 1. Degree distributions for the ACD network (panel A) and the JWAD network (panel B)

the 8,970 nodes. As these results clearly indicate that the networks exhibit a pattern of sparse connectivity, we may say that the two association networks both possess the characteristics of a scale-free network.

3.2 Clustering Coefficients

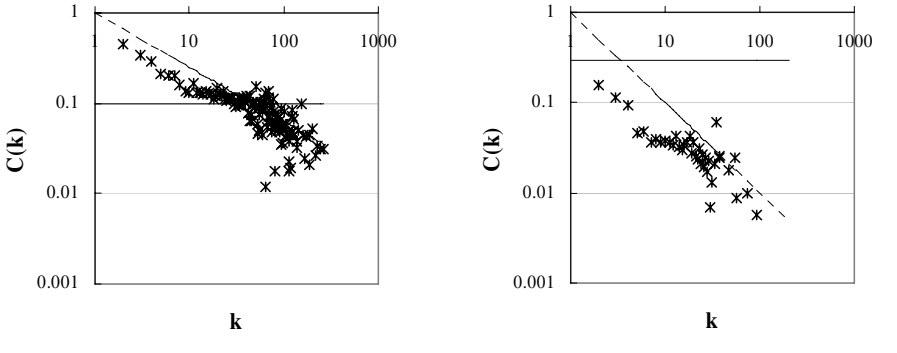
The association networks are next compared in terms of their clustering coefficients, which is an index of the interconnectivity strength between neighboring nodes in a graph. Watts and Strogatz [23] proposed the notion of clustering coefficient as an appropriate index of the degree of connections between nodes in their study of social networks that investigated the probabilities of an acquaintance of an acquaintance also being one of your acquaintances.

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. The equation yields a clustering coefficient value between 0-1; while a star-like sub-graph would have a clustering coefficient value of 0, a complete graph with all nodes connected would have clustering coefficient of 1.

Similarly, Ravasz and Barabasi [24] (2003) advocate the notion of clustering coefficient dependence on node degree, based on the hierarchical model of $C(k) \approx k^{-1}$ [25], as an index of the hierarchical structures encountered in real networks, such as the World Wide Web. Accordingly, the hierarchical nature of a network can be characterized 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 the hierarchical exponent. The results of scaling C(k) with k for the ACD network (panel a) and for the JWAD network (panel b) are presented in Figure 2.



(a) ACD network (b) JWAD network

Fig. 2. Clustering coefficient distributions for the ACD network (panel A) and the JWAD network (panel B)

The solid lines in the figure correspond to the average clustering coefficient. The ACD network has an average clustering coefficient of 0.1, while the value is 0.03 for the JWAD network. As both networks conform well to a power law, we may conclude that they both possess intrinsic hierarchies.

4 Graph Clustering

This section focuses on some graph clustering techniques and reports on the application of graph clustering to the two constructed association network representations based on the large-scale Japanese word association databases. Specifically, after considering the relative merits of the original MCL algorithm [9], the enhanced RMCL algorithm [10], and the combination of RMCL and modality [11] employed in the present study, we briefly present and discuss the results of applying these methods to the two association network representations.

4.1 Markov Clustering

Markov Clustering (MCL) is widely recognized as an effective method for detecting the patterns and clusters within large and sparsely connected data structures. The MCL algorithm is based on random walks across a graph, which, by utilizing the two simple algebraic operations of expansion and inflation, simulates the flow over a stochastic transition matrix in converging towards equilibrium states for the stochastic matrix. Of particular relevance to the present study is the fact that the inflation parameter, r, influences the clustering granularity of the process. In other words, if the value of r is set to be high, then the resultant clusters will tend to be small in size. While this parameter is typically set to be r = 2, a value of 1.6 has been taken as a reasonable value in creating a dictionary of French synonyms [26].

Although MCL is clearly an effective clustering technique, particularly for large-scale corpora [13] [14], the method, however, undeniably suffers from its lack of control over the distribution in cluster sizes that it generates. The MCL has a problematic tendency to either yield many isolated clusters that consist of just a single word or to yield an exceptionally large core cluster that effectively includes the majority of the graph nodes.

4.2 Recurrent Markov Clustering

In order to overcome this shortcoming with the MCL method, Jung, Miyake, and Akama [10] have recently proposed an improvement to the basic MCL method called Recurrent Markov Clustering (RMCL), which provides some control over cluster sizes by adjusting graph granularity. Basically, the recurrent process achieves this by incorporating feedback about the states of overlapping clusters prior to the final MCL output stage. As a key feature of the RMCL, the reverse tracing procedure makes it possible to generate a virtual adjacency matrix for non-overlapping clusters based on the convergent state resulting from the MCL process. The resultant condensed matrix provides a simpler graph, which can highlight the conceptual structures that underlie similar words.

4.3 Modularity

According to Newman and Girvan [11], modularity is a particularly useful index for assessing the quality of divisions within a network. The modularity Q value can highlight differences in edge distributions between a graph of meaningful partitions and a random graph under the same vertices conditions (in terms of numbers and sum of their degrees). The modularity index is defined as:

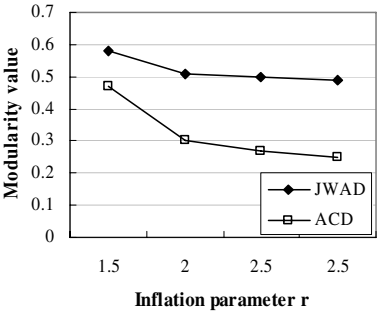
$$Q = \sum_i (e_{ii} - a_i^2)$$

where i is the number of cluster c_i, e_{ii} is the proportion of internal links in the whole graph and a_i is the expected proportion of c_i's edges calculated as the total number of degrees in c_i divided by the sum of degrees for the whole graph. In practice, high Q values are rare, with values generally falling within the range of about 0.3 to 0.7. The present study employs a combination of RMCL clustering algorithm with this modularity index in order to optimize the appropriate inflation parameter within the clustering stages of the RMCL process. The RMCL results reported in this paper are all based on the combination of the RMCL clustering method and modularity.

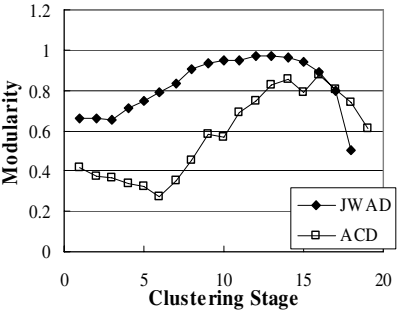
4.4 Clustering Results

The MCL and the RMCL algorithm were implemented as a series of calculations that are executed with gridMathematica. The MCL process generated a nearly-idempotent stochastic matrix at around the 20th clustering stage.

In terms of determining a reasonable value for the r parameter, while it is usual to identify local peaks in the Q value, as Figure 3(a), which plots modularity as a function of r, indicates there are no discernable no peaks in the Q value. Accordingly, the highest value of r equals 1.5 was taken for the inflation parameter. Plotting modularity as a function of the clustering stage, Figure 3(b) indicates that values of

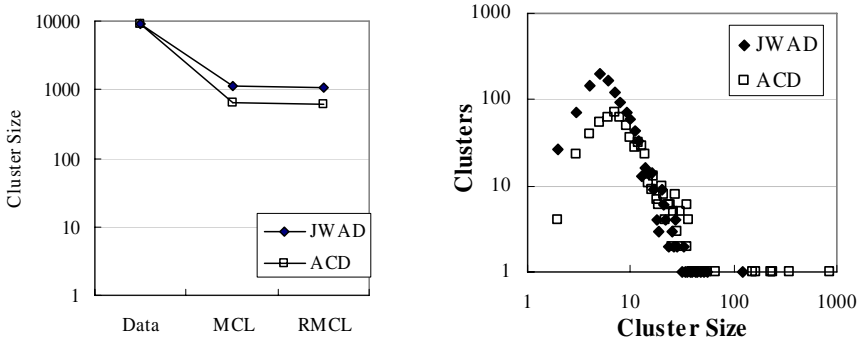


(a) Inflation parameter for MCL



(b) MCL clustering stage

Fig. 3. Basic clustering results, with panel a presenting modularity values as a function of r and panel b indicating modularity values as a function of the MCL clustering stage



(a) Cluster sizes for MCL and RMCL

(b) Distributions in cluster sizes of MCL

Fig. 4. Clustering results for MCL and RMCL, with panel a showing cluster sizes and panel b showing distributions for the MCL algorithm

Q value peaked at stage 14 in the case of the ACD network and at stage 12 for the JWAD network. Accordingly, these clustering stages were used in the RMCL process.

Figure 4(a) presents the MCL and the RMCL cluster sizes for both the ACD network and the JWAD network, illustrating the downsizing transitions that took place during the graph clustering process. Figure 4(b) plots the frequencies of cluster sizes for the results of MCL clustering. In the case of the ACD network, the MCL algorithm resulted in 642 hard clusters, with an average cluster size of 7.5 and an SD of 56.3, while the RMCL yielded 601 clusters, where the average number of cluster components was 1.1 with an SD of 0.42. In the case of the JWAD network, the MCL resulted in 1,144 hard clusters, with an average cluster size of 5.5 and an SD of 7.2, while the RMCL yielded 1,084 clusters, where the average number of cluster components was 1.1 with an SD of 0.28.

4.5 Discussion

In section 4.3, we presented the quantitative results of applying the MCL and the RMCL graph clustering algorithms to the two association networks in terms of the numbers of resultant clusters produced and the distributions in cluster sizes for each network by each method. In this section, we present a few of the clusters generated by the clustering methods in illustrating the potential of the clustering approach as an extremely useful tool for automatically identifying groups of related words and the relationships between the words within the groupings.

One objective of the project developing the JWAD is to utilize the database in the development of lexical association network maps that capture and highlight the association patterns that exist between Japanese words [6] [7] [8]. Essentially, a lexical association network map represents a set of forward associations elicited by a target word by more than two respondents (and the strengths of those associations), together with backward associations (both their numbers and associative strengths), as

well as the levels and strengths of associations between all members of an associate set [6]. While the lexical association network maps were first envisaged primarily at the single word level, the basic approach to mapping out associations can be extended to small domains and beyond, as the example in Figure 5 illustrates with a map building from and contrasting a small set of emotion words. Interestingly, this association map suggests that the positive emotion synonym words of しあわせ (happy) and 嬉しい (happy) have strong associations to a small set of other close synonyms, but that the negative emotion words of 寂しい (lonely) and 悲しい (sad) primarily elicit word association responses that can be regarded as having causal or resultant relationships. While the creation of such small domain association maps is likely to provide similarly interesting insights concerning association knowledge, the efforts required to manually identify and visualize even relatively small domains are not inconsequential. However, the clustering methods presented in this section represent a potentially very appealing way of automatically identifying and visualizing sets of related words as generated clusters.

Table 1 presents the word clusters for the target words of しあわせ (happy) and 寂しい (lonely) that were generated by the MCL algorithm for the JWAD network. Comparing the sets of associations for these two words in Figure 5 based on the JWAD with the word clusters in Table 1, clearly there are many words that are common to both. The additional words included in the MCL word clusters in Table 1

Table 1. Examples of clusters for the JWAD network generated by the MCL algorithm

手をたたこう (clap hands) 幸福 (happiness) しあわせ (happy)
怒 (anger) 嬉しい (happy) 歓喜 (delight) 喜 (joy) 喜び (joy) 喜ぶ (be glad) 喜寿 (77th birthday) 喜怒哀楽 (human emotions) 悲しむ (be sad) 大喜利 (final act in a *Rakugo* performance)
独り (alone) 一人 (alone; one person) さびしい (lonely)
寂しい (lonely) 悲しみ (sadness) 悲しい (be sad) 涙 (tears) 流す (shed)
負け (defeat) 涙 (tears) くやしい (regrettable)

Table 2. Examples of words in the ACD network clustered together by the MCL algorithm

結納 (engagement gift) 幸せ (happy) 入籍 (entry in family register) 式場 (ceremonial hall) 結婚 (marriage) 婚約 (engagement) 同棲 (cohabiting) 冠婚葬祭 (important ceremonial occasions)
貰う (receive) 嬉しい (happy) お駄賃 (tip) ありがたい (thanks) 褒美 (reward) 収入 (income) 小づかい (pocket money)
冬 (winter) 寒さ (coldness) 初冬 (early winter) 真冬 (midwinter) 寂しい (lonely) ウィンター (winter) 暖冬 (warm winter)
純粋 (pure) 分泌液 (secretion) 嬉し涙 (tears of joy) なみだ (tears) 溢れる (overflow) 悲しい事 (sad incident) 悔し涙 (vexation)
後悔 (regret) 反省 (reflection) 悔やむ (be sorry) 悔しさ (chagrin) 悔しい (regrettable)

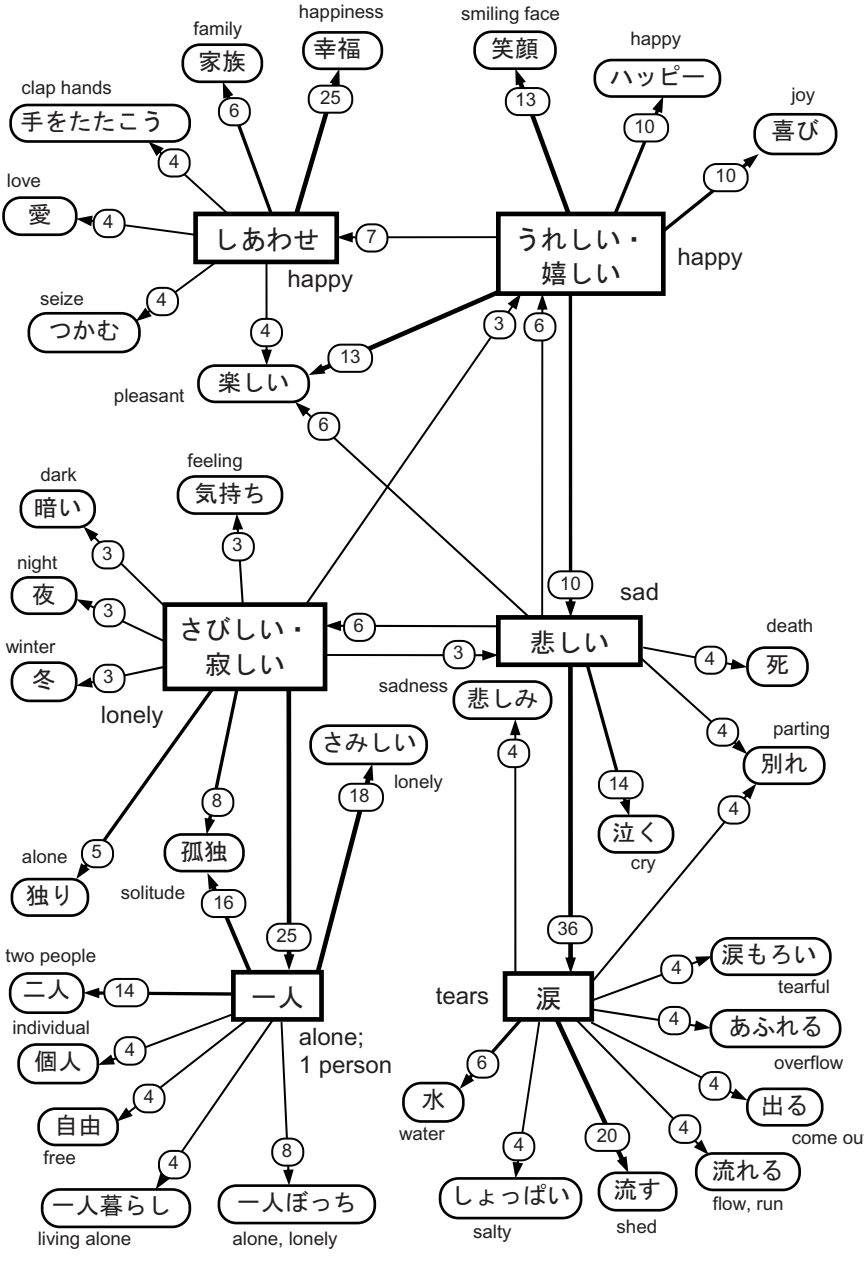


Fig. 5. Example of lexical association network map building from and contrasting a small set of emotion words within the JWAD. The numbers on the arrows indicate response frequencies as percentages of the respective association sets.

serve to demonstrate how the automatic clustering process can be a powerful technique for identifying more implicit, but nevertheless interesting patterns of association within collections of words that are mediated through indirect connections via closely related items.

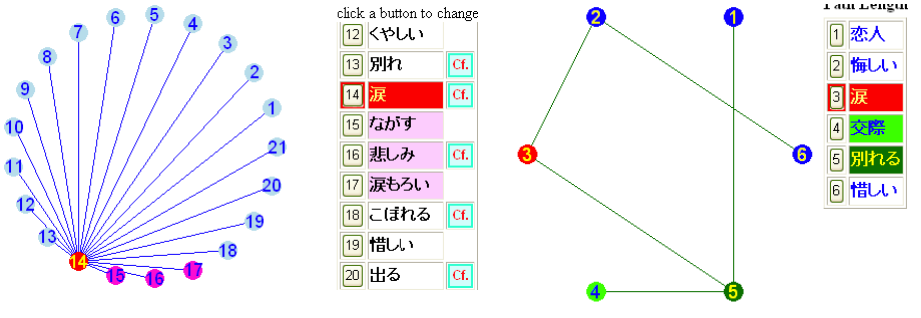
Similarly, Table 2 presents word clusters for the ACD network generated by the MCL algorithm, which illustrates how effective the clustering methods are in grouping together words that have a synonymous relationship.

5 RMCLNet

This section briefly introduces RMCLNet [26], which is a web application to make publicly available the clustering results for the ACD and the JWAD networks, in a spirit of seeking to foster a wider appreciation for the interesting contributions that investigations of word association knowledge can yield for our understandings of lexical knowledge in general.

As Widdow, Cederberg, and Dorow astutely observe [28], graph visualization is a particularly powerful tool for representing the meanings of words and concepts [24]. The graph visualization of the structures generated through both the MCL and the RMCL clustering methods is being implemented with webMathematica and utilizing some standard techniques of java servlet/JSP technology. Because webMathematica is capable of processing interactive calculations, the graph visualization is realized by integrating Mathematica with a web server that uses Apache2 as its http application server and Tomcat5 as its servlet/JSP engine.

The visualization system can highlight the relationships between words by dynamically presenting both MCL and RMCL clustering results for both the ACD and the JWAD networks, as the screen shots in Figure 6 illustrate. Implementation of the visualization system is relatively straightforward, basically only requiring storage of the multiple files that are automatically generated during execution of the RMCL algorithm. The principle feature of the system is that it is capable of simultaneously presenting clustering results for both the ACD and the JWAD networks, making it



(a) MCL result for 涙 (b) RMCL result for 涙

Fig. 6. Screen shots of RMCLNet, illustrating visualizations of MCL clustering results (panel a) and of RMCL clustering results (panel b) for the Japanese word 涙 'tears'

possible to compare the structural similarities and differences between the two association networks. Such comparisons can potentially provide useful hints for further investigations concerning the nature of word associations and graph clustering.

6 Conclusions

As a promising approach to capturing and unraveling the rich networks of associations that connect words together, this study has applied a range of network analysis techniques in order to investigate the characteristics of network representations of word association knowledge in Japanese. In particular, the study constructed and analyzed two separate Japanese association networks. One network was based on the Associative Concept Dictionary (ACD) by Okamoto and Ishizaki [5], while the other was based on the Japanese Word Association Database (JWAD) by Joyce [6] [7] [8]. The results of initial analyses of the two networks—focusing on degree distributions and average clustering coefficient distributions for nodes with degrees—revealed that the two networks both possess the characteristics of a scale-free network and that both possess intrinsic hierarchies.

The study also applied some graph clustering algorithms to the association networks. While graph clustering undoubtedly represents an effective approach to capturing the associative structures within large-scale knowledge resources, there are still some issues that warrant further investigation. One purpose of the present study has been to examine improvements to the basic MCL algorithm [9], by extending on the enhanced RMCL version [10]. In that context, this study applied a combination of RMCL graph clustering method and the modularity measurement as a means of achieving greater control over the sizes of clusters generated during the execution of the clustering algorithms. For both association networks, the combination of the RMCL algorithm with the modularity index resulted in fewer clusters.

This paper also illustrated the fact that clustering methods represent a potentially very appealing way of automatically identifying and visualizing sets of related words as generated clusters by looking at some of the clustered words generated by the MCL algorithm. The examples presented in Tables 1 and 2 suggest that automatic clustering techniques can be useful for identifying, beyond simply the direct association relationship, more implicit and indirect patterns of association within collections of words as mediated by closely related items, and for grouping together words that have synonymous relationships. The paper also briefly introduced the RMCLNet which is a web application specifically developed to make the clustering results for the ACD and the JWAD networks publicly available. It is hoped that further investigations into the rich structures of association knowledge by comparing the structural similarities and differences between the two association networks can provide useful hints concerning both the nature of word associations and graph clustering.

As alluded to at times in the discussions, much of the research outlined in this paper forms part of a larger ongoing research project that is seeking to capture the structures inherent within association knowledge. In concluding this paper, it is appropriate to acknowledge some limitations with the present study and to fleetingly sketch out some avenues to be explored in the future. One concern to note is that,

while the ACD database and Version 1 of the JWAD are of comparable sizes and both can be regarded as being reasonably large-scale, some characteristics of the present two semantic network representations of Japanese word associations may be reflecting characteristics of the foundational databases. As already noted, the ongoing JWAD project is committed to constructing a very large-scale database of Japanese word associations, and as the database expands with both more responses and more extensive lexical coverage and new versions of the JWAD are compiled, new versions of the JWAD semantic network will be constructed and analyzed in order to trace its growth and development.

While much of the discussions in section 4 focused on the important issue of developing and exercising some control over the sizes of clusters generated through graph clustering, the authors also recognize the need to evaluate generated clusters in terms of their semantic consistency. The presented examples of word clusters indicate that clustering methods can be effectively employed in automatically grouping together words related words based on associative relationships. However, essential tasks for our future research into the nature of association knowledge will be to develop a classification of elicited association responses in the JWAD in terms of their associative relationships to the target word and to apply the classification in evaluating the associative relationships between the components of generated clusters. While the manual inspection of generated clusters is undeniably very labor intensive, the work is likely to have interesting implications for the recent active development of various classification systems and taxonomies within thesauri and ontology research.

Finally, one direct extension of the present research will be the application of the MCL and the RMCL graph clustering methods to the dynamic visualization of the hierarchical structures within semantic spaces, as the schematic representation in Figure 7 illustrates. The combination of constructing large-scale semantic network representations of Japanese word associations, such as the JWAD network, and applying graph clustering techniques to the resultant network is undoubtedly a particularly promising approach to capturing, unraveling and comprehending the complex structural patterns within association knowledge.

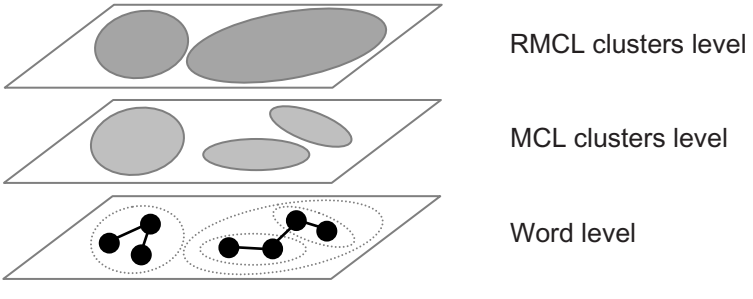


Fig. 7. Schematic representation of how the MCL and the RMCL graph clustering methods can be used in the creation of a hierarchically-structures semantic space based on an association network

Acknowledgments. 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 express their gratitude to Prof. Furui, Prof. Akama, Prof. Nishina, Prof. Tokosumi, and Ms. Jung. The authors have been supported by Grants-in-Aid for Scientific Research from the Japanese Society for the Promotion of Science: Research project 18500200 in the case of the first author and 19700238 in the case of the second author.

References

1. Deese, J.: The Structure of Associations in Language and Thought. The John Hopkins Press, Baltimore (1965)
2. Cramer, P.: Word Association. Academic Press, New York & London (1968)
3. Hirst, G.: Ontology and the Lexicon. In: Staab, S., Studer, R. (eds.) Handbook of Ontologies, pp. 209–229. Springer, Heidelberg (2004)
4. Firth, J.R.: Selected Papers of J. R. Firth 1952-1959. In: Palmer, F.R. (ed.), Longman, London (1957/1968)
5. Okamoto, J., Ishizaki, S.: Associative Concept Dictionary and its Comparison with Electronic Concept Dictionaries. In: PACLING 2001, pp. 214–220 (2001)
6. Joyce, T.: Constructing a Large-scale Database of Japanese Word Associations. In: Tamaoka, K. (ed.) Corpus Studies on Japanese Kanji (Glottometrics 10), pp. 82–98. Hituzi Syobo, Tokyo, Japan and RAM-Verlag, Lüdenschied, Germany (2005)
7. Joyce, T.: Mapping Word Knowledge in Japanese: Constructing and Utilizing a Large-scale Database of Japanese Word Associations. In: LKR 2006, pp. 155–158 (2006)
8. Joyce, T.: Mapping Word Knowledge in Japanese: Coding Japanese Word Associations. In: LKR 2007, pp. 233–238 (2007)
9. van Dongen, S.: Graph Clustering by Flow Simulation. Ph.D. thesis, University of Utrecht (2000)
10. Jung, J., Miyake, M., Akama, H.: Recurrent Markov Cluster (RMCL) Algorithm for the Refinement of the Semantic Network. In: LREC2006, pp. 1428–1432 (2006)
11. Newman, M.E., Girvan, M.: Finding and Evaluating Community Structure in Networks. Phys. Rev. E69, 026113 (2004)
12. Church, K.W., Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. Comp. Ling. 16, 22–29 (1990)
13. Dorow, B., et al.: Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. In: MEANING-2005 (2005)
14. Steyvers, M., Shiffrin, R.M., Nelson, D.L.: Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory. In: Healy, A.F. (ed.) Experimental Cognitive Psychology and its Applications (Decade of Behavior), Washington, DC, APA (2004)
15. Steyvers, M., Tenenbaum, J.B.: The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. Cog. Sci. 29, 41–78 (2005)
16. Nelson, D.L., McEvoy, C., Schreiber, T.A.: The University of South Florida Word Association, Rhyme, and Word Fragment Norms (1998), `http:// www.usf.edu/ FreeAssociation`
17. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)

18. Roget, P.M.: Roget's Thesaurus of English Words and Phrases (1991), http://www.gutenberg.org/etext/10681
19. Moss, H., Older, L.: Birkbeck Word Association Norms. Psychological Press, Hove (1996)
20. Umemoto, T.: Table of Association Norms: Based on the Free Associations of 1,000 University Students (in Japanese). Tokyo, Tokyo Daigaku Shuppankai (1969)
21. Version 1 of the JWAD, http://www.valdes.titech.ac.jp/~terry/jwad.html
22. Barabasi, A.L., Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509–512 (1999)
23. Watts, D., Strogatz, S.: Collective Dynamics of 'Small-world' Networks. Nature 393, 440–442 (1998)
24. Ravasz, E., Barabasi, A.L.: Hierarchical Organization in Complex Networks. Physical Rev. E 67, 26112 (2003)
25. Dorogovtsev, S.N., Goltsev, A.V., Mendes, J.F.F.: Pseudofractal Scale-free Web, e-Print Cond-Mat/0112143 (2001)
26. Vechthomova, O., et al.: Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. In: International Symposium on Applied Stochastic Models and Data Analysis, pp. 106–113 (2005)
27. RMCLNet, http://perrier.dp.hum.titech.ac.jp/semnet/RmclNet/index.jsp
28. Widdows, D., Cederberg, S., Dorow, B.: Visualisation Techniques for Analyzing Meaning. TSD5, 107–115 (2002)