16

Spatial Clustering Technique for Data Mining

Yuichi Yaguchi, Takashi Wagatsuma and Ryuichi Oka
The University of Aizu
Japan

1. Introduction

For mining features from the social web, analysis of the shape, detection of network topology and corresponding special meanings and also clustering of data become tools, because the information obtained by these tools can create useful data behind the social web by revealing its relationships and the relative positions of data. For example, if we want to understand the effect of someone's statement on others, it is necessary to analyze the total interaction between all data elements and evaluate the focused data that results from the interactions. Otherwise, the precise effect of the data cannot be obtained. Thus, the effect becomes a special feature of the organized data, which is represented by a suitable form in which interaction works well. The feature, which is included by social web and it is effect someone's statement, may be the shape of a network or the particular location of data or a cluster.

So far, most conventional representations of the data structure of the social web use networks, because all objects are typically described by the relations of pairs of objects. The weak aspect of network representation is the scalability problem when we deal with huge numbers of objects on the Web. It is becoming standard to analyze or mine data from networks in the social web with hundreds of millions of items.

Complex network analysis mainly focuses on the shape or clustering coefficients of the whole network, and the aspects and attributes of the network are also studied using semistructured data-mining techniques. These methods use the whole network and data directly, but they have high computational costs for scanning all objects in the network.

For that reason, the network node relocation problem is important for solving these social-web data-mining problems. If we can relocate objects in the network into a new space in which it is easier to understand some aspects or attributes, we can more easily show or extract the features of shapes or clusters in that space, and network visualization becomes a space-relocation problem.

Nonmetric multidimensional scaling (MDS) is a well-known technique for solving new-space relocation problems of networks. Kruskal (1964) showed how to relocate an object into n-dimensional space using interobject similarity or dissimilarity. Komazawa & Hayashi (1982) solved Kruskal's MDS as an eigenvalue problem, which is called quantification method IV (Q-IV). However, these techniques have limitations for cluster objects because the stress, which is the attraction or repulsive force between two objects, is expressed by a linear formula. Thus, these methods can relocate exact positions of objects into a space but it is difficult to translate clusters into that space.

This chapter introduces a novel technique called Associated Keyword Space (ASKS) for the space-relocation problem, which can create clusters from object correlations. ASKS is based on

Q-IV but it uses a nonlinear distance measure, space uniformalization, to preserve average and variance in the new space, sparse matrix calculations to reduce calculation costs and memory usage, and iterative calculation to improve clustering ability. This method allows objects to be extracted into strict clusters and finds novel knowledge about the shape of the whole network, and also finds partial attributes. The method also allows construction of multimedia retrieval systems that combine all media types into one space.
Section 2 surveys social-web data-mining techniques, especially clustering of network-structured data. In Section 3, we review spatial clustering techniques such as Q-IV and ASKS. Section 4 shows the results of a comparison of Q-IV and ASKS, and also shows the clustering performance between ASKS and the *K*-nearest neighbor technique in a network. Section 5 explains an example application utilizing ASKS. Finally, we summarize this chapter in Section 6.

2. Related work

2.1 Shape of the network

Data-mining techniques for network-like relational data structures have been studied intensively recently. Examining the shape of a network or determining a clustering coefficient for each object is an important topic for complex networks (Boccaletti et al. (2006)), because these properties indicate clear features of whole or partially structured networks. Watts & Strogatz (1998) explained that human relationships exhibit a small-world phenomenon, and Albert & Barabási (2002) showed that the link structure of web documents has the scale-free property. These factors, the small-world phenomenon, which has $\log n$ of radius of n objects in the network, and the scale-free property, which has a power-law distribution of the rate number of degree, are found in many real network-like data such as protein networks (Jeong et al. (2001)), metabolic networks (Jeong et al. (2000)), routing networks (Chen et al. (2004)), costar networks (Yan & Assimakopoulos (2009)), and coauthor networks (Barabási & Crandall (2003)). The clustering coefficient (Soffer & Vázquez (2005)) is another measure of network shape and of the local density around an object in a network. Although the clustering of coefficients can extract "how much an object is included in a big cluster", it is not able to identify actual objects that are included in a cluster. Thus, to extract objects into a cluster, the nearest-neighbor technique can be applied to extract objects into the cluster (Wang et al. (2008)), but it is difficult to check the actual cluster size. Hierarchical clustering is another useful technique (Boccaletti et al. (2006)), but it is still difficult to find the density of a cluster.

2.2 Web mining categorization

Web mining applications can be categorized into the following three groups.

1. Web content mining retrieves useful information by performing text mining.
2. Web structure mining discovers communities and the relevance of pages based on hyperlink structures.
3. Web usage mining analyzes user access patterns from access logs and click histories.

An excellent review of Web mining can be found in Kosala & Blockeel (2000).
In terms of the above categorization, we have developed an algorithm for Web content mining Yaguchi et al. (2006); Ohnishi et al. (2006). This tool helps a user discover text information by displaying the hyperlink structure between related Web pages. The following subsection gives a summary of related work on Web content and structure mining methods.

2.3 Web data mining

Many schemes have used hyperlink structures to extract valuable information from the Web Carrière & Kazman (1997); Kleinberg (1999); Pirolli et al. (1996); Spertus (1997).

Dean et al. introduced two algorithms to identify related Web pages: one derived from the HITS algorithm Kleinberg (1999) and the other based on cocitation relationships. To increase accuracy, the HITS algorithm has been combined with content information Bharat & Henzinger (1998); Chakrabarti et al. (1999); Modha & Spangler (2000).

He et al. proposed a method to retrieve pages related to a query given by a user that grouped pages into distinct topics He et al. (2001). In the process, they introduced similarity metrics based on text information, hyperlink structure, and cocitation relationships.

Moise et al. treated the problem of how to find related pages effectively (Moise et al. (2003)). They proposed three approaches: hyperlink-based, content-based, and hybrid approaches. They developed an algorithm and showed that it outperformed conventional algorithms in the precision of its retrieved results.

In general, related Web pages are densely connected to each other by hyperlinks, and graph mining approaches can be used to discover such clusters of related Web pages, which are called "Web communities." Recent approaches to the discovery of Web communities are described in (Murata (2003)), and the requirements for graph mining algorithms suitable for the discovery of Web communities are also discussed.

Youssefi et al. applied data mining and information visualization techniques to Web domains, aiming to benefit from the combined power of human visual perception and computing ability (Youssefi et al. (2004)).

Liu et al. modeled a Web site's content structure in terms of its topic hierarchy by utilizing three types of information associated with a Web site: hyperlink structure, directory structure, and Web page content (Liu & Yang (2005)).

3. Spatial clustering

3.1 Nonmetric multidimensional scaling

The problem of creating a new N-dimensional space using the correspondence of pairs of objects is the same as the nonmetric multidimensional scaling (MDS) problem. The metric MDS was first proposed in Young and Householder's study (Young & Householder (1938)), where numerical affinity values were used, and the nonmetric MDS was also presented using only orders of affinities (Shepard (1972); Kruskal (1964)). We describe brief definition for nonmetric MDS of Kruskal's approach.

In the study of nonmetric MDS, let N denote the dimension of the space in which objects are allocated, and let each object be numbered i and its location be denoted by x_i. The similarity or dissimilarity (nonnegative value) between objects i and j is defined by δ_i and the Euclidean distance between them is defined as $d_{ij} = -(x_j - x_i)^2$. Now, object x_i is given a more suitable position $\hat{x}_i$ as a next state by utilizing δ_i, and the new distance between objects i and j is also set as $\hat{d}_{ij} = -(\hat{x}_j - \hat{x}_i)^2$. Then, the stress S can be defined as:

$$S = \sqrt{\frac{\sum_{i<j} d_{ij}^2}{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}}. \tag{1}$$

Finally, the goal of nonmetric MDS is able to express the following equation:

$$\min_{\text{all n-dimensional configurations}} \sqrt{\frac{\sum_{i<j} d_{ij}^2}{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}}. \tag{2}$$

3.2 Quantification method IV

Komazawa & Hayashi (1982) solved the nonmetric MDS problem as an eigenvalue problem. Let M_{ij} denote the nonnegative value of the affinity measure between object i and j, and M_{ij} becomes bigger as the objects i and j become more similar. The location of object i is denoted by x_i in the N-dimensional space, and if two objects, i and j, are more similar, x_i and x_j are closer; if they are more dissimilar, the distance between them is larger. Practically, this problem is defined as the maximization of the following function ϕ:

$$\phi = \sum_{i=1}^{n} \sum_{j=1}^{n} -M_{ij} d_{ij} \to \max \tag{3}$$

$$d_{ij} = |x_i - x_j|^2. \tag{4}$$

Hence,

$$\phi = -\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} |x_i - x_j|^2 = -\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} (|x_i|^2 - 2x_i x_j + |x_j|^2) \tag{5}$$

$$= 2\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} x_i x_j - \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} |x_i|^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} |x_j|^2 \tag{6}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} (M_{ij} + M_{ji}) x_i x_j - \sum_{i=1}^{n} |x_i|^2 \sum_{j=1}^{n} (M_{ij} + M_{ji}) x_{ij} = x_{ij} \tag{7}$$

Let a_{ij} be:

$$a_{ij} = M_{ij} + M_{ji}. \tag{8}$$

Then:

$$\phi = 2\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j - \sum_{i=1}^{n} |x_i|^2 \sum_{j=1}^{n} a_{ij}. \tag{9}$$

If we eliminate $a_i i$ from this equation, then:

$$\phi = 2\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} a_{ij} x_i x_j - \sum_{i=1}^{n} |x_i|^2 \sum_{j=1, j\neq i}^{n} a_{ij} = \mathbf{x}' \mathbf{B} \mathbf{x} \tag{10}$$

$$B = \begin{pmatrix} -\sum_{j=1, j\neq 1}^{n} a_{1j} & a_{12} & \cdots & a_{1n} \\ a_{21} & -\sum_{j=1, j\neq 2}^{n} a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & -\sum_{j=1, j\neq n}^{n} a_{nj} \end{pmatrix}, x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \tag{11}$$

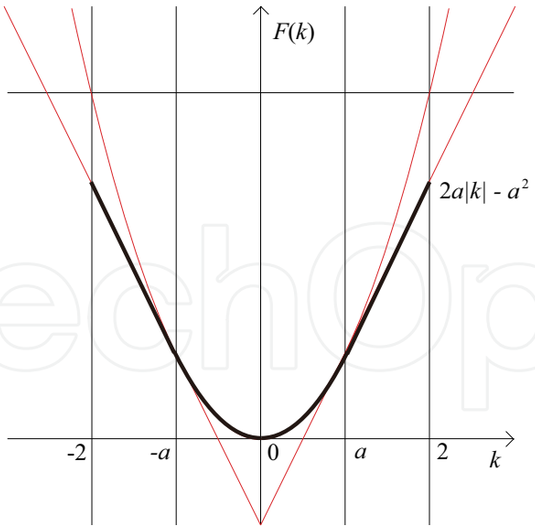


Fig. 1. Nonlinear function used in ASKS.

Maximizing $\mathbf{x}'\mathbf{B}\mathbf{x}$ under the condition $\mathbf{x}'\mathbf{x} = const$, requires solving equation (3):

$$\phi^* = \mathbf{x}'\mathbf{B}\mathbf{x} - \lambda\mathbf{x}'\mathbf{x} - \mathbf{c} \tag{12}$$

$$\frac{\partial\phi^*}{\partial x} = \mathbf{B}\mathbf{x} - \lambda\mathbf{I}\mathbf{x} = \mathbf{0} \tag{13}$$

. Finally, equation (3) becomes the following equation:

$$(\mathbf{B} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \tag{14}$$

This eigenvalue problem can be solved more quickly if matrix $\mathbf{B}$ is sparse. However, this method requires all N eigenvalues to be positive. Normally, to ensure eigenvalues are positive, a sufficiently large value must be subtracted from all elements of $\mathbf{B}$. Thus, the calculation time and memory requirement becomes $O(N^2)$ in many cases.

3.3 Associated keyword space (ASKS)

ASKS is a nonlinear version of MDS and is effective for noisy data Takahashi & Oka (2001). This section explains ASKS and describes how to calculate it.
Let N denote the spatial dimension of an allocated object. Each object is indexed by i and its location is defined by x_i. The distance is measured by the formula F:

$$d_{ij} = -F(x_j - x_i). \tag{15}$$

F has a parameter a and is defined as:

$$F(k) = \begin{cases} |k|^2 & (|k| < a) \\ 2a|k| - a^2 & (|k| \geq a). \end{cases} \tag{16}$$

Figure 1 shows a plot of this function.
Three types of constraints on the distribution of objects are specified to decide the amount of space to be allocated to similar objects in distinguishable clusters:

1. make the original point the center of gravity for the objects;

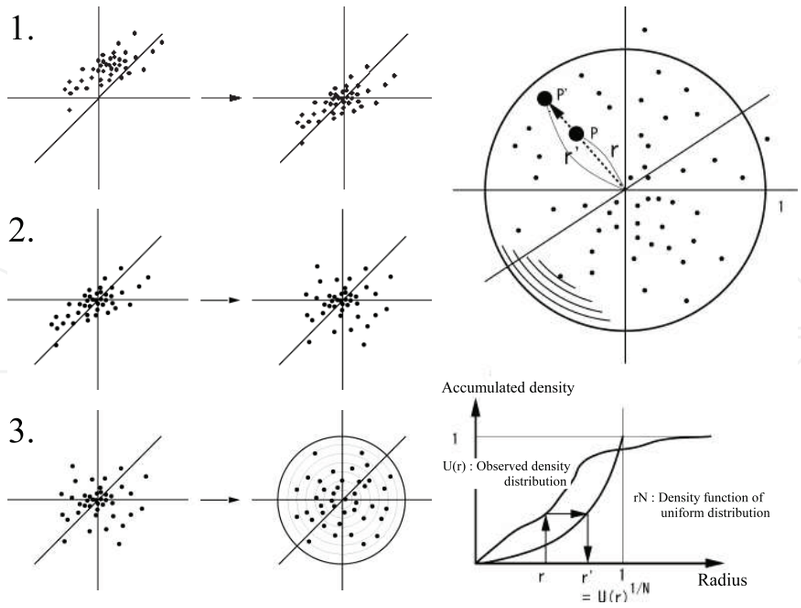


Fig. 2. Uniformalization types used in ASKS.

2. obtain covariance matrices such that dispersion in any direction creates the same value; and
3. uniformalize the objects in a radially from origin.

Figure 2 shows the method for uniformalization in the super-sphere. Uniformalization is useful for clustering noisy data that otherwise tend to distribute connections too evenly across the data.

3.4 Iterative solution of nonlinear optimization

The criterion function of ASKS is:

$$J(x_1, x_2, \ldots, x_n) = \sum_i \sum_j \{-M_{ij} F(x_j - x_i)\} \rightarrow \max \tag{17}$$

M_{ij} is an affinity (a nonnegative value) between objects i and j. It is calculated from the co-occurrence of objects i and j. The partial derivative of J with respect to x_i gives the formula for determining the values of x_i that maximize J:

$$\frac{\partial}{\partial x_i} \sum_i \sum_j \{-M_{ij} F(x_j - x_i)\} \equiv 0, \tag{18}$$

$$\sum_j M_{ij} F'(x_j - x_i) \equiv 0. \tag{19}$$

The derivative of F is:

$$F'(k) = \begin{cases} 2k & (|k| < a) \\ 2a\frac{k}{|k|} & (|k| \geq a), \end{cases} \tag{20}$$

and parameter a is junction of linear and non-linear distance measure for controlling density. Next, define D by:

$$D(k) = \begin{cases} 2 & (|k| < a) \\ \frac{2a}{|k|} & (|k| \geq a), \end{cases} , \tag{21}$$

from which we derive the expression:

$$F'(x_j - x_i) = D(x_j - x_i)(x_j - x_i). \tag{22}$$

The following iterative computation converges to the solution x_i.

$$x_i^{t+1} = \frac{\sum_j M_{ij} D(x_j^{(t)} - x_i^{(t)}) x_j^{(t)}}{\sum_j M_{ij} D(x_j^{(t)} - x_i^{(t)})} \tag{23}$$

The three constraints must be enforced at each step of the iterative computation for all variables x_i $(i = 1,2,\ldots,n)$.

4. Experiment

4.1 Comparison of Q-IV and ASKS

The effectiveness of ASKS is shown by comparing its performance with that of Q-IV.

Assume that 1,000,000 objects are to be clustered into C categories of 100, 1000, or 10,000 objects. We generated a set of affinity data between objects $M_{ij}(1 \leq i \leq C, 1 \leq j \leq C)$, where each M_{ij} took a value of 1 if objects i and j belonged to the same category, and 0 otherwise. We counted the numbers for the first case (N_i) and the second case (N_o), and then we defined R_i as the sum of the affinities in a class for the first case and R_o as the sum of the affinities between classes for the second case. If objects i and j belonged to the same category, then M_{ij} was set to $M_{ij} = 1$ with a probability of R_i/N_i, and the other values of M_{ij} were set to $M_{ij} = 0$. In the same way, if objects i and j belonged to different categories, the value of M_{ij} was set to $M_{ij} = 1$ according to R_o/N_o. The ratio of R_o/R_i expresses the level of noise, where a value of zero denoted no noise and larger values (which could be > 1.0) denoted a high level of noise. Both methods were applied to the case of 1000 categories. The Q-IV method is characterized by linear optimization and standard distributions of the various noise levels. The clustering results for the Q-IV approach are shown in Figure 3, where a subset of 20,000 objects belonging to 20 categories is plotted to aid visualization. The ASKS method is characterized by nonlinear optimization and a uniform distribution of the various noise levels. The results for the ASKS method under the same conditions are shown in Figure 4. These results show that the ASKS technique is superior to the Q-IV approach because ASKS can gather objects belonging to the same category into a more compact space and can distinguish categories at higher noise values.

To give a comparison numerically, we measured the ratio of the Standard Distribution (SD) in the associated spaces. The parameter S_i is the sum of the SD of objects i and j that belong to the same category, and S_o is the same sum when the objects are in different categories. An ideal MDS system would gather objects of the same category into a single point, causing the value $S_i = 0$. Therefore, we can compare the effectiveness of the above methods in terms of the ratio S_i/S_o.

Experiments were performed using a range of noise levels ($0.01 \leq R_i/R_o \leq 100.0$) and various numbers of categories. Figure 5(a) shows the results for 100,000 objects in 50 categories for the

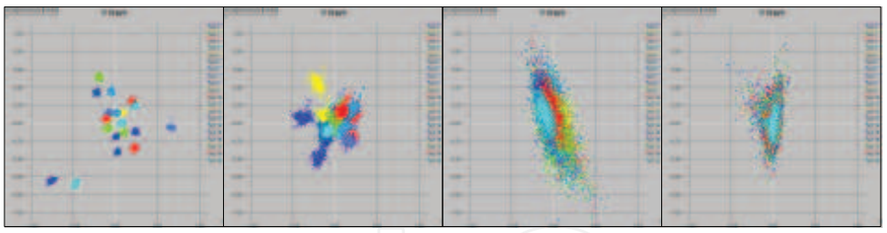


Fig. 3. Allocation of items by Q-IV. Noise level (Ro/Ri) [left = 0.01, 0.1, 1.0, right = 100.0].

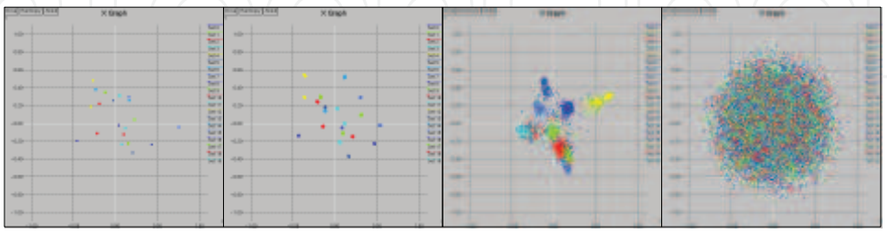


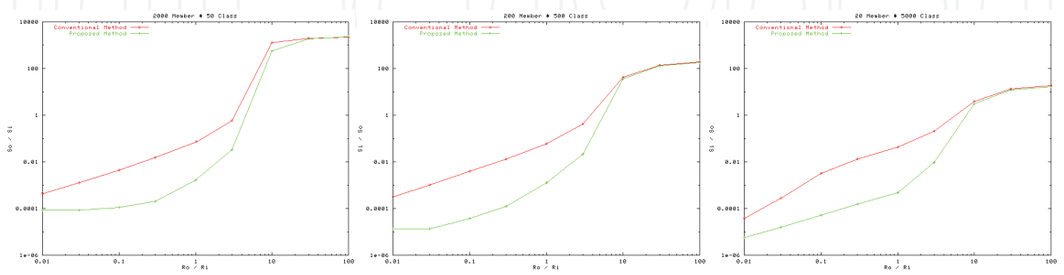
Fig. 4. Allocation of items by ASKS. Noise level (Ro/Ri) [left = 0.01, 0.1, 1.0, right = 100.0].

same conditions as those shown in Figures 3 and 4. Figure 5(b) shows the results for 100,000 objects with 500 categories, and Figure 5(c) shows the results for 5000 categories.

Another experiment was also performed to show the effect of parameter a in equation (20). If $a = 2$, then the function of ASKS is same as Q-IV without uniformalization. Thus, we can call this case as uniformalized Q-IV. Now, we set 100,000 samples, which belong to 1000 classes, into three-dimensional space. Figure 6 shows a comparison study on noise robustness between uniformalized Q-IV and ASKS with $a = 0.2$, and the number of iteration is set to 200. From this figure, Q-IV could not discriminate the classes when ratio $Ro/Ri = 0.1$ but ASKS still easily finds the clusters.

Figure 7 explains the effect of parameter a which is the junction of the group of linear and non-linear distance functions. In this figure, if parameter a is getting smaller, then each cluster becomes tighter but the speed of convergence is slower.

To check the dense of clustering, we separate the clustering space into $20 \times 20 \times 20$ boxes and we count the number of objects in each box. Figure 8 shows that the result, which is indicated by the red circled area, is perfectly clustered one or several groups, because each class in the dataset consists of 100 elements, and we can distinctively see in the graph where a box has more than 100 elements. Q-IV was unable to cluster these objects when $a = 0.01$ and $a = 0.1$, but ASKS was able to perform that clearly.



(a) 2000 Members * 50 Classes (b) 200 Members * 500 classes (c) 20 Members * 5000 classes

Fig. 5. Relationship between R_o/R_i and S_o/S_i using 100,000 samples: (a) 50, (b) 500, and (c) 5000 classes (for 2000, 200, and 20 samples/class, respectively.) For the larger noise levels ($R_o/R_i > 10$), there is little difference in efficiency between the conventional method (upper line) and the proposed method (lower line).

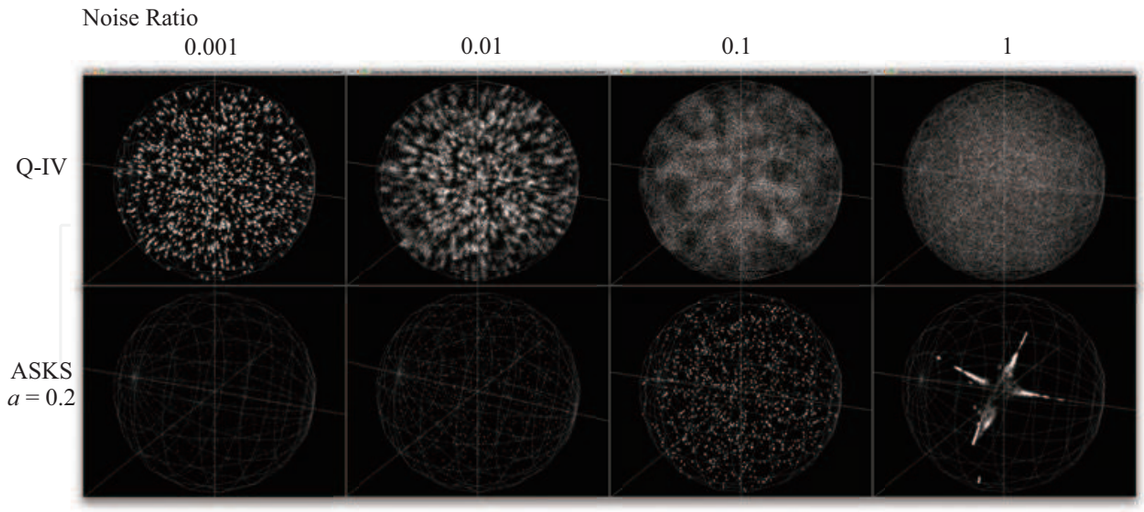


Fig. 6. Comparison study on noise robustness between uniformalized Q-IV and ASKS with $a = 0.2$.

5. Application examples

5.1 Text retrieval system

Takahashi & Oka (2001) constructed a text retrieval system using ASKS. From this study, they planned to search similar Japanese documents from fj news group which belongs to a news system on the Internet. It gathered 3.7 million articles from 1985 to 2000, and the number of words was approximately 520,000. The result of ASKS clustering shows that the study was able to find the associated word such as the word "Tabasco" and "Hot cod ovum" can be found around the word "Mustard" in the space which has same property "Hot", or "Rice", "Laver", "Soybean paste soup" and "Egg" also can be found around "Soybean paste", which are usually appeared in Japanese breakfast(figure 9).

5.2 Multimedia clustering

Wagatsuma et al. (2009) also constructed Web mining system using ASKS was performed as follows.

1. Create an affinity matrix for each of several media-content items and merge these matrices.
2. Create 3D coordinates and allocates each item (e.g., URL or text) by using ASKS.
3. Analyze the associated space.

In this experiment, Web pages were crawled from the page "Office of Prime Minister of Japan" [1] to a maximum hyperlink depth of four and with no restriction on URL domains. A total of 1371 pages were collected, with included words of 6948 types, and images of 579 types. Textual information was analyzed by MeCab [2], an open-source Japanese morphological analyzer. This study used three types of media, namely Web page hyperlinks, text, and image data.

[1] http://www.kantei.go.jp/
[2] http://mecab.sourceforge.net/

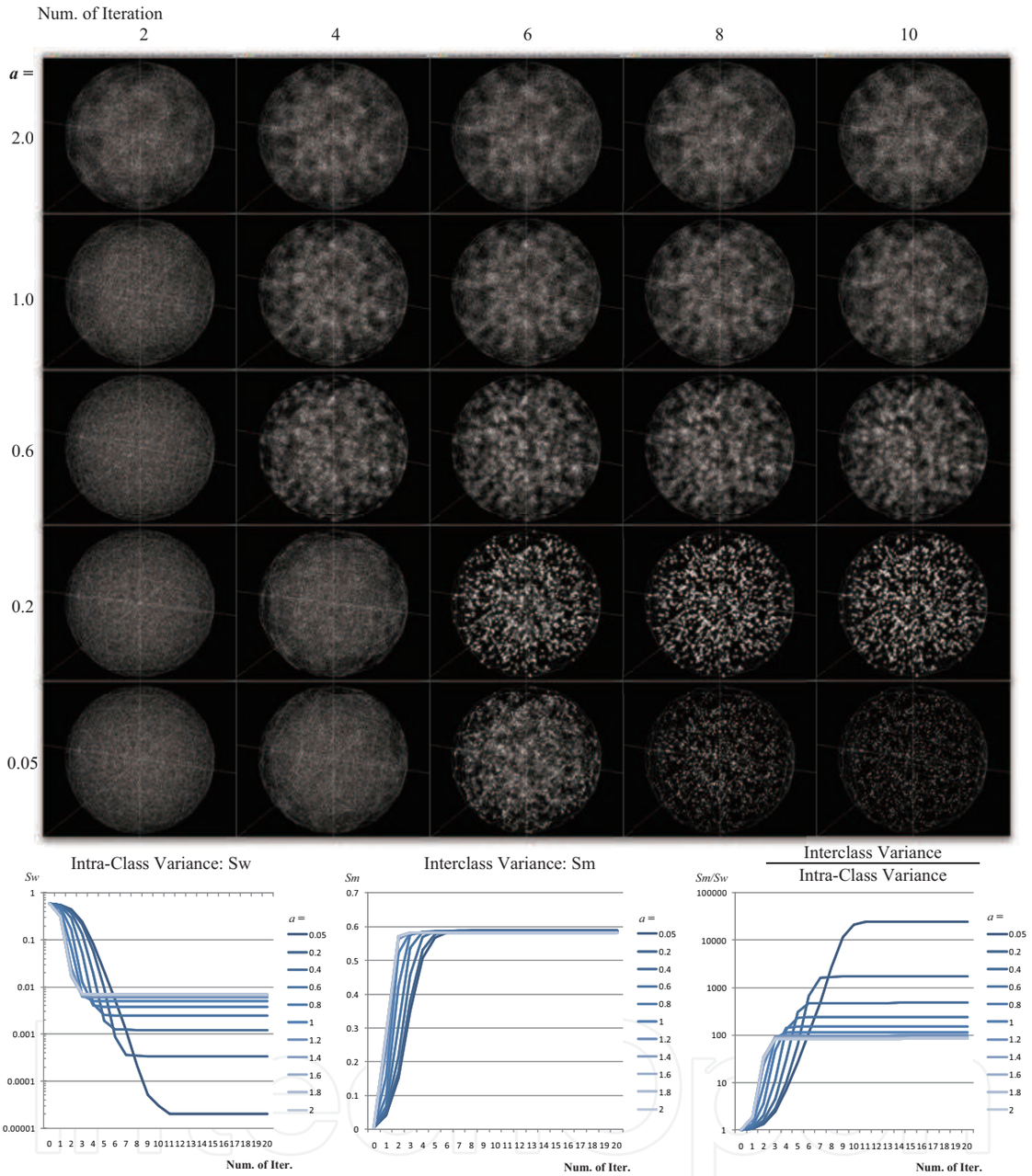


Fig. 7. Comparison study on the effect of parameter *a* to capability of clustering

5.2.1 Calculation of the affinity matrix

In this experiment, the affinity information could be specified in terms of six matrices (see Figure 10). This study defined the meaning of semantic similarity for each affinity matrix as follows.

1. *Web page hyperlink structure (page vs. page)*
   Increase affinity by 1 when there is a hyperlink from a page to the other page.

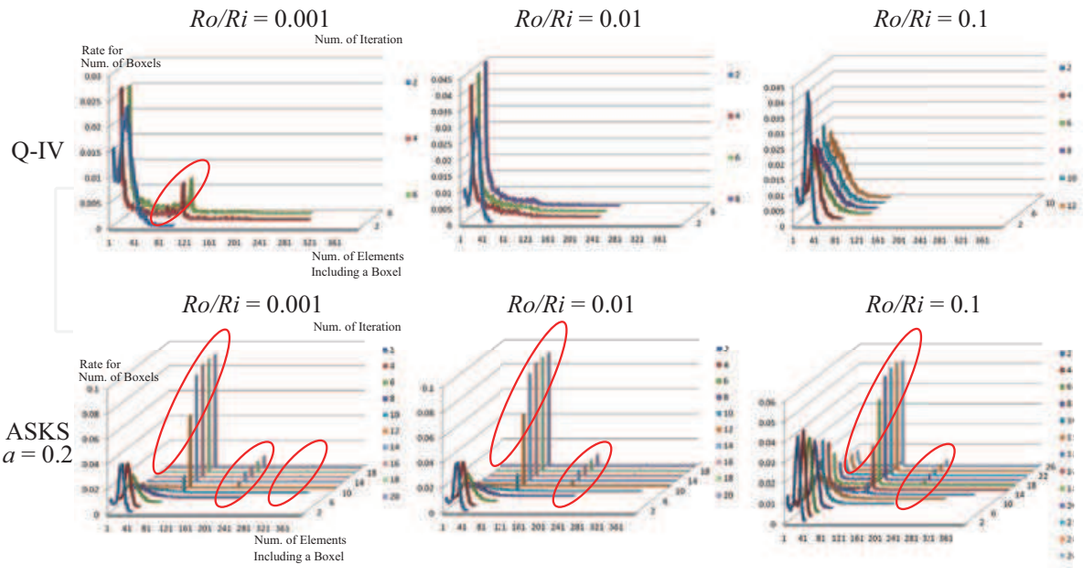


Fig. 8. Comparison study on clustering ability in 3D space: We set $20 \times 20 \times 20$ small boxes into 3D affinity space, and count the number of boxes which have the same the number of objects inside.

2. *Word co-occurrence in a sentence (word vs. word)*
   If a word appears in a sentence with other words, then their affinity is calculated according to the interword distances. If word i and word j appear in a sentence, the distance d_{ij} is specified as 1 plus the number of words appearing between them. Then the affinity of the two words is defined as:

$$d_{ij} = 1 - \frac{d_{ij} - 1}{L},$$

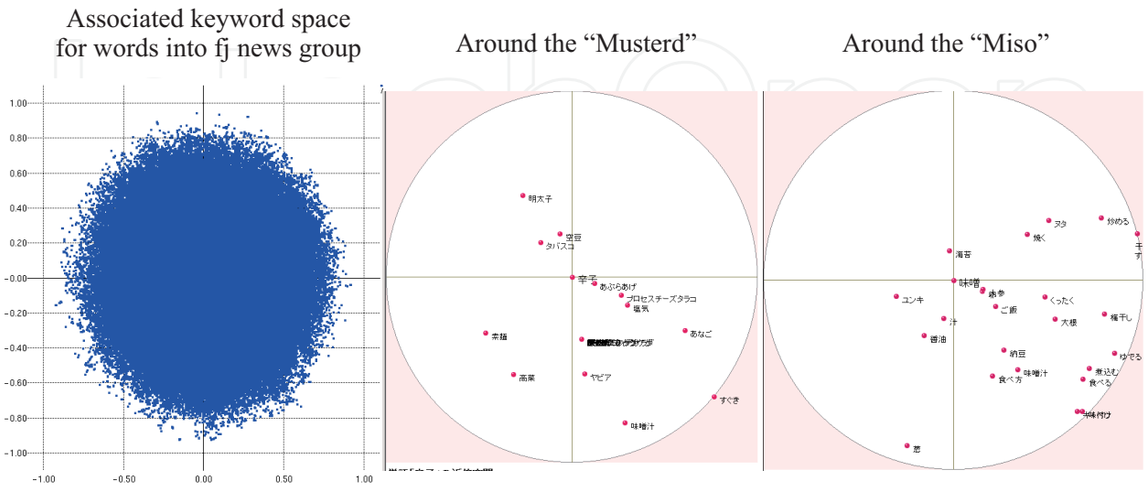


Fig. 9. ASKS in text retrieval system.

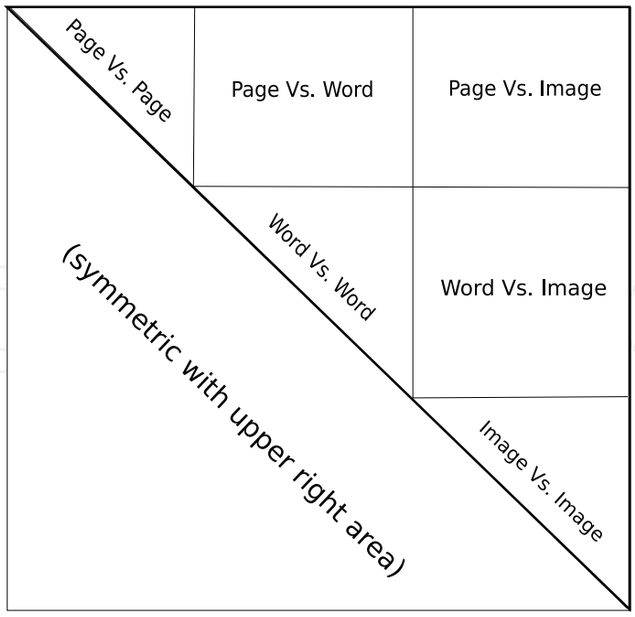


Fig. 10. The affinity matrix represents the presence of semantic similarity between types of media or content (Web page hyperlinks, text, and images.) This affinity matrix is created by merging six affinity matrices for the separate types.

where $L(=10)$ is the maximum allowed distance between two words. This definition was developed in Ohnishi et al. (2006).

3. *Similarity between images (image vs. image)*
   All of the images used in a Web page have a mutual affinity. This affinity is most frequently calculated in terms of the distances of the correlation of their color histograms. To calculate the affinity between image i and image j, with histograms Hi and Hj, their distance d_{ij} is defined as:
   $$d_{ij} = \frac{\langle H_i, H_j \rangle}{(\|H_i\| \cdot \|H_j\|)}$$
   .
   This study uses the binarized values:
   $$d_{ij} = \begin{cases} 1 & if\ d_{ij} \geq 0.5, \\ 0 & otherwise. \end{cases}$$
4. *Word occurrence in a Web page (page vs. word)*
   If a word appears in a certain page, then the affinity between them is calculated using the Term Frequency—Inverse Document Frequency (TF-IDF).
5. *Image occurrence in a Web page (page vs. image)*
   If an image appears in a certain page, then the affinity between them is set to 1.
6. *Image occurrence with word (word vs. image)*
   If an image has a word defined by an alt tag, then the affinity between the image and the alt word is available.

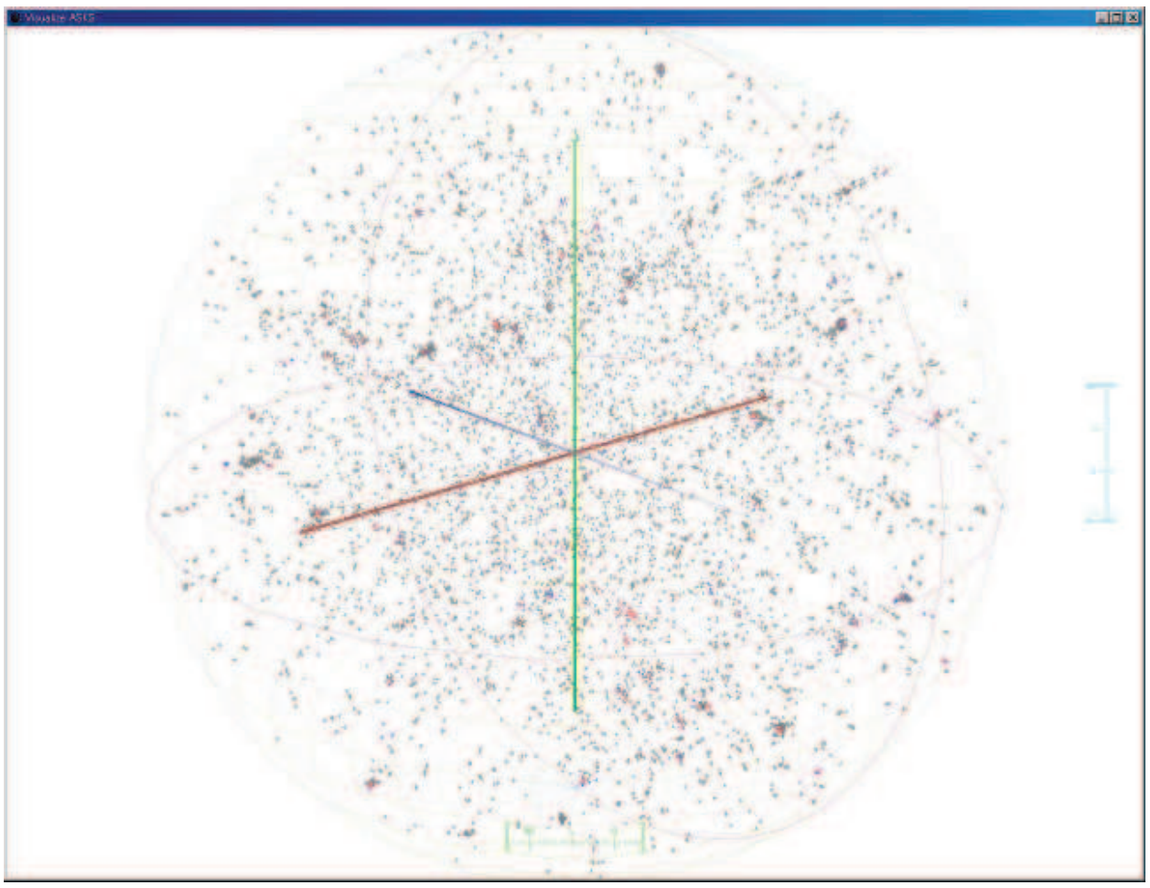


Fig. 11. Visualized associated space with merged affinity matrix. Each allocated node expresses a Web page, a word, or an image. This study can find several clusters in this associated space.

5.3 Merging the affinity matrices

After all six affinity matrices are created, they are simply concatenated into one matrix (see Figure 10). This merged affinity matrix represents the semantic similarities within the various types of media or content.

5.3.1 Visualization of the associated space

This study has developed software to visualize and analyze the 3D-associated space generated by the affinity matrix, called Visualize ASKS. It allows users to recognize the correlations between items more intuitively. The study also found several clusters in the associated space of our example (see Figure 11).

5.3.2 Cluster investigation

This study targeted one cluster constructed from neighboring items to analyze the features of the allocation in the association space generated by the affinity matrix involving several media. This study also selected one word within the cluster (the name of a previous Prime Minister of Japan "Junichiro Koizumi") as the source word, and analyzed the space within a 0.1 radius of this word. Note that the association space has a radius of 1.0.
The target area included the following three elements.

Fig. 12. Images gathered in the target area. There is little semantic similarity among them. These were all of the images in the Web pages reached by a few hyperlink steps from the seed page.

– *Pages:* A large number of Web page nodes existed in the target area, but semantically dissimilar pages were also mixed in with these pages.
– *Words:* Examples of several words in the target area (translated from Japanese into English) were

  cabinet official, prime minister, ministry, media person, interview, talk, cabinet secretariat, safety, and government.

  Many words linked to politics, the economy, and the names of the previous Prime Minister of Japan were gathered in the target area.
– *Images:* Three images were gathered in the target area, as shown in Figure 12. The first image appeared in a Web page referring to the Japanese governmental problem expressed the word "kidnapping" in Japanese [3]. The second image was used in the home page of the "Prime Minister of Japan and His Cabinet" [4]. The third image is a facial portrait image of the previous Prime Minister of Japan, "Junichiro Koizumi", found in the Web page "Introducing Previous Prime Ministers of Japan" [5].
  These images do not have high mutual semantic similarity scores, as calculated by our definition in Section refsubsec:definition. These were the only images in the Web pages reached by a few hyperlink steps from the seed page.

A noteworthy feature is that both of the items linked strongly to each other are found within the target area. However, many Web pages with semantically dissimilar information are also included. This Web page cluster was constructed from Web pages reached by a few hyperlink steps from the seed page.

5.3.3 Image allocation investigation

This study investigated the features of a collection of images having semantic similarity, being facial portraits of the previous Prime Minister of Japan (see Figure 13). These images were allocated to clusters in the associated space as shown in Figure 14.

[3] http://www.rachi.go.jp/
[4] http://www.kantei.go.jp/foreign/index-e.html
[5] http://www.kantei.go.jp/jp/koizumisouri/index.html

Fig. 13. Target images of the previous Prime Minister of Japan. These images have high mutual semantic similarity.

Fig. 14. Target images allocated to clusters. Images allocated to one cluster usually have similar domain names.

Images were allocated to several detached clusters, although they all had high mutual affinity values. From an analysis of the information about nodes around each image, we found that images allocated to the same cluster often have similar domain names. However, a few pairs of images in the same cluster have high affinities but different domain names. We therefore conclude that the allocation of image nodes is affected by other information.

6. Conclusion

We have introduced a novel spatial clustering technique that is called ASKS. ASKS can relocate objects into a new n-dimensional space from network structured data. Comparing ASKS with Q-IV, it improves the performance of clustering, and it can find actual clusters of objects and retrieve similar objects that are not related by an object of query, and it can be used in a multimedia retrieval system that combines words, Web pages and images.
We plan to pursue the following developments in future work. We expect that the visualized space used in this research will resemble existing relation graphs, which can be described by rubbery models or which may be easier to understand. Therefore, we should compare the visualization in this research with existing relationship graphs. Then there is the progression to categorization using clustering methods with visualized associated spaces to investigate the meaning of each category. In addition, if we apply categories, it may be possible to build a search system using the categorized information provided.

7. References

Albert, R. & Barabási, A. (2002). Statistical mechanics of complex networks, *Reviews of modern physics* 74(1): 47–97.

Barabási, A. & Crandall, R. (2003). Linked: The new science of networks, *American journal of Physics* 71: 409.

Bharat, K. & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment, *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 104–111.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. (2006). Complex networks: Structure and dynamics, *Physics Reports* 424(4-5): 175–308.

Carrière, S. & Kazman, R. (1997). WebQuery: Searching and visualizing the Web through connectivity, *Computer Networks and ISDN Systems* 29(8-13): 1257–1267.

Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D. & Kleinberg, J. (1999). Mining the Web's link structure, *Computer* 32(8): 60–67.

Chen, J., Gupta, D., Vishwanath, K., Snoeren, A. & Vahdat, A. (2004). Routing in an Internet-scale network emulator, *The IEEE Computer Society's 12th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems, 2004.(MASCOTS 2004). Proceedings*, pp. 275–283.

He, X., Ding, C., Zha, H. & Simon, H. (2001). Automatic topic identification using webpage clustering, *Proceedings of the 2001 IEEE international conference on data mining*, IEEE Computer Society, pp. 195–202.

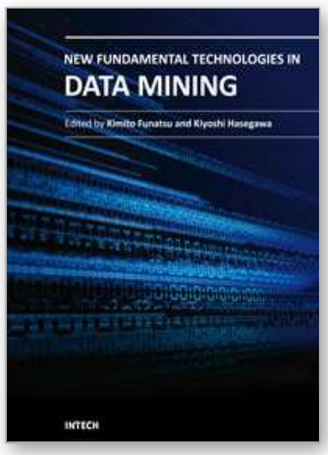
Jeong, H., Mason, S., Barabási, A. & Oltvai, Z. (2001). Lethality and centrality in protein networks, *Nature* 411(6833): 41–42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. & Barabási, A. (2000). The large-scale organization of metabolic networks, *Nature* 407(6804): 651–654.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment, *Journal of the ACM*

(JACM) 46(5): 604–632.
Komazawa, T. & Hayashi, C. (1982). Quantification Theory and Data Processing, *Tokyo: Asakura-shoten* .
Kosala, R. & Blockeel, H. (2000). Web mining research: A survey, *ACM SIGKDD Explorations Newsletter* 2(1): 1–15.
Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* 29(1): 1–27.
Liu, N. & Yang, C. (2005). Mining web site's topic hierarchy, *Special interest tracks and posters of the 14th international conference on World Wide Web*, ACM, pp. 980–981.
Modha, D. S. & Spangler, W. S. (2000). Clustering hypertext with applications to web searching, *Proceedings of the eleventh ACM on Hypertext and hypermedia*, ACM, pp. 143–152.
Moise, G., Sander, J. & Rafiei, D. (2003). Focused co-citation: Improving the retrieval of related pages on the web, *Proceedings of the 12th International world wide web Conference (Budapest, Hungary, 2003)* .
Murata, T. (2003). Visualizing the structure of web communities based on data acquired from a search engine, *IEEE transactions on industrial electronics* 50(5): 860–866.
Ohnishi, H., Yaguchi, Y., Yamaki, K., Oka, R. & Naruse, K. (2006). Word space : A new approach to describe word meanings, *IEICE technical report. Data engineering* 106(149): 149–154.
Pirolli, P., Pitkow, J. & Rao, R. (1996). Silk from a sow's ear: extracting usable structures from the Web, *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, ACM, p. 125.
Shepard, R. (1972). Multidimensional scaling: Theory and applications in the behavioral sciences, Seminar Press New York.
Soffer, S. & Vázquez, A. (2005). Network clustering coefficient without degree-correlation biases, *Physical Review E* 71(5): 57101.
Spertus, E. (1997). ParaSite: Mining structural information on the Web, *Computer Networks and ISDN Systems* 29(8-13): 1205–1215.
Takahashi, H. & Oka, R. (2001). Self-organization an associated keyword space for text retrieval, *WMSCI2010, World Multi-Conference on Systemics, Cybernetics and Informatics* pp. 302–307.
Wagatsuma, T., Yaguchi, Y. & Oka, R. (2009). Cross-media data mining using associated keyword space, *10th IEEE International Conference on Computer and Information Technology (CIT10)* 2: 289–294.
Wang, C., Au, K., Chan, C., Lau, H. & Szeto, K. (2008). Detecting Hierarchical Organization in Complex Networks by Nearest Neighbor Correlation, *Nature Inspired Cooperative Strategies for Optimization (NICSO 2007)* pp. 487–494.
Watts, D. & Strogatz, S. (1998). Collective dynamics of "small-world" networks, *Nature* 393(6684): 440–442.
Yaguchi, Y., Ohnishi, H., Mori, S., Naruse, K., Oka, R. & Takahashi, H. (2006). A mining method for linkedweb pages using associated keyword space, *IEEE/IPSJ International Symposium on Applications and the Internet (SAINT'06)* pp. 268–276.
Yan, J. & Assimakopoulos, D. (2009). The small-world and scale-free structure of an internet technological community, *International Journal of Information Technology and Management* 8(1): 33–49.
Young, G. & Householder, A. (1938). Discussion of a set of points in terms of their mutual

distances, *Psychometrika* 3(1): 19–22.
Youssefi, A., Duke, D. & Zaki, M. (2004). Visual web mining, *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, ACM, pp. 394–395.



New Fundamental Technologies in Data Mining

Edited by Prof. Kimito Funatsu

ISBN 978-953-307-547-1
Hard cover, 584 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by "Data Mining" address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Yuichi Yaguchi, Takashi Wagatsuma and Ryuichi Oka (2011). Spatial Clustering Technique for Data Mining, New Fundamental Technologies in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-547-1, InTech, Available from: http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/spatial-clustering-technique-for-data-mining

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen