Exploration of a text collection and identification of topics by clustering

Antoine Naud[1,2] and Shiro Usui[1]

[1] RIKEN Brain Science Institute
2-1 Hirosawa, Wako City, 351-0198 Saitama, Japan.
naud@brain.riken.jp, usuishiro@riken.jp
[2] Department of Informatics, N. Copernicus University
ul. Grudziadzka 5, 87-100 Torun, Poland.

Abstract. An application of cluster analysis to identify topics in a collection of posters abstracts from the Society for Neuroscience (SfN) Annual Meeting in 2006 is presented. The topics were identified by selecting from the abstracts belonging to each cluster the terms with the highest scores using different ranking schemes. The ranking scheme based on log-entropy showed better performance in this task than other more classical TFIDF schemes. An evaluation of the extracted topics was performed by comparison with previously defined thematic categories for which titles are available, and after assigning each cluster to one dominant category. The results show that repeated bisecting k-means performs better than standard k-means.

1 Introduction

An increasing amount of published documents like research papers, computer programs, analyzed data or related references are gathered in databases or repositories in order to enable quick access to literature from a given field of research. The development of such databases in the field of neuroscience is a major goal in neuroinformatics [1]. The resulting large amounts of documents give rise to the need for tools that automatically organize them into indexing structures. These structures may fasten the retrieval for searched information as well as provide an overview of a corpus and help navigation. A subsequent task is the organization of the keywords in a structure reflecting the semantic contents of the documents. To this purpose, the general structure of a documents collection can be detected by clustering the documents into groups covering similar topics. This work is devoted to the analysis of the posters presented at the Annual Meeting of the Society for Neuroscience (SfN) in 2006. SfN is, with more than $37,500$ members, the world's largest organization of scientists devoted to the study of neuroscience and the brain science. Its Annual Meeting is the largest event in neuroscience. The primary goal of this work was the automatic discovery of topics covered in poster sessions, on the basis of the posters abstracts and titles. Another potential application is the automatic partitioning into sessions of the posters submitted to future SfN Annual Meetings.

2 Construction of the Vector Space Model

The most widely used approach in Natural Language Processing is the *vector space model*. In this model, a set of terms $\mathcal{T}$ is first built by extracting all words occurring in a collection of documents $\mathcal{D}$, followed by stop words removal and stemming steps [2]. The number of occurrences of each term in each document (usually called *frequency*) is counted and denoted f_{ij}. Then a frequency matrix $\mathbf{F}$ is built with the $\{f_{ij}\}$ as entries. As we will cluster documents in this work, it is more convenient to build $\mathbf{F}$ as a $[documents \times terms]$ matrix, where each document is a row vector in the space of all terms, called the *term space* later on. Depending on the purpose of the application, terms occurring too often or very seldom can also be discarded. When the number of documents N in the collection is in the range of a few thousands, the number of extracted terms M is often larger than a few tens of thousands, leading to very high dimensional space for the documents. In order to remove less semantically significant terms and also to enable further processing, it is necessary to reduce the term space dimension by selecting a smaller subset of terms, usually using a ranking of the terms according to their Document Frequency (DF). In general, we are interested in selecting the terms that best represent the semantic content of the documents. This intuitive feature is however very difficult to catch only by statistical means. In the present application, the terms were extracted from the posters' abstracts and titles. The preprocessing scheme and extraction of candidate terms was the same as in [3]. From the abstracts and titles of the $N = 12844$ posters, we obtained directly $M = 40767$ terms, which is a too large value to allow further processing. 3 term spaces were built by selecting terms occurring in at least $2, 13$ and 45 documents for the following reasons: (a) selecting terms with $DF \geq 2$ allows to decrease the term space size roughly by a factor of two, leading to $M = 19794$ terms; (b) selecting terms with $DF \geq 13$ leads to $M = 6127$, this is the maximal size allowing the application of Matlab's kmeans function in section 5, (c) selecting terms with $DF \geq 45$ decreases again by two the number of terms, ending up with $M = 3006$ terms. Only unigrams (single words) were considered for the terms in this preliminary study.

3 Exploratory analysis of existing categories

The posters abstracts and titles were extracted from a CD-ROM distributed to all the participants of the Annual Meeting. Four types of categories are provided by the Meeting's organizing committee: *theme*, *subtheme*, *topic* and *session*, and a name is given to each category. Each of the 12844 posters for which an abstract and a title (called hereafter *documents*) were available was also assigned by the organizers to one poster session, one topic, subtheme and theme. A summary of basic statistics of this collection of documents is given in Table 1. The purpose of this analysis is to check whether the various originally defined groupings of posters into categories can be observed in the term spaces that we defined in the previous section.

Table 1: Summary data of the Society for Neuroscience 2006 Annual Meeting

1	Number of themes	7
2	Number of subthemes	71
3	Number of topics	415
4	Number of poster sessions	650
5	Number of poster abstracts	12844
6	Number of words / abstract (average)	278
7	Number of extracted terms	40767

3.1 Average cosine measures between documents

The frequency matrix $\mathbf{F}$ is a sparse contingency table where each row represents one document, and the similarity of two documents can be evaluated by the cosine of the angle between the two document vectors. In order to balance the frequencies of terms occurring in long abstracts with respect to terms occurring in shorter abstracts, a normalization of the rows of matrix $\mathbf{F}$ is performed after the term weighting (see [4] for a review of weighting schemes). The cosine between 2 vectors in the high-dimensional term space is defined as

$$cos(\mathbf{d}_1, \mathbf{d}_2) = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{\|\mathbf{d}_1\| \|\mathbf{d}_2\|}, \tag{1}$$

where $\cdot$ is the dot product. As vectors $\{\mathbf{d}_i\}$ are of unit length, expression (1) simplifies to the dot product. The mean cosine for all pairs of documents within each category is a measure of how dense are the categories in the term space. Similarly, for each category, the mean of the cosines between each document in the category and all the documents in all other categories measures to which extend this category is separated from the others. The averages of these two means for all the categories were computed efficiently using the centroid vectors of each category, as described in [5]. The results are presented in Figure 1. Note that the cosine function is a similarity measure (i.e. the more similar two documents are, the higher is their cosine) and not a distance (or dissimilarity). The average cosines within categories are clearly higher than between categories in each term space, especially for the *topic* and *session* categories, which indicates that these categories are also well defined in the 3 term spaces. The above two average cosines among categories are equivalent to clusters' *cohesion* and *separation*, some internal measures of clusters validity presented e.g. in [6].

3.2 MDS layouts of the original categories

As it was seen above, the differences in average cosines between and within categories are larger for *topic* and *session* categories, which indicates that those categories are better separated in the terms space. This can be confirmed by visualizing the different categories. To this purpose, we processed the data as follows:

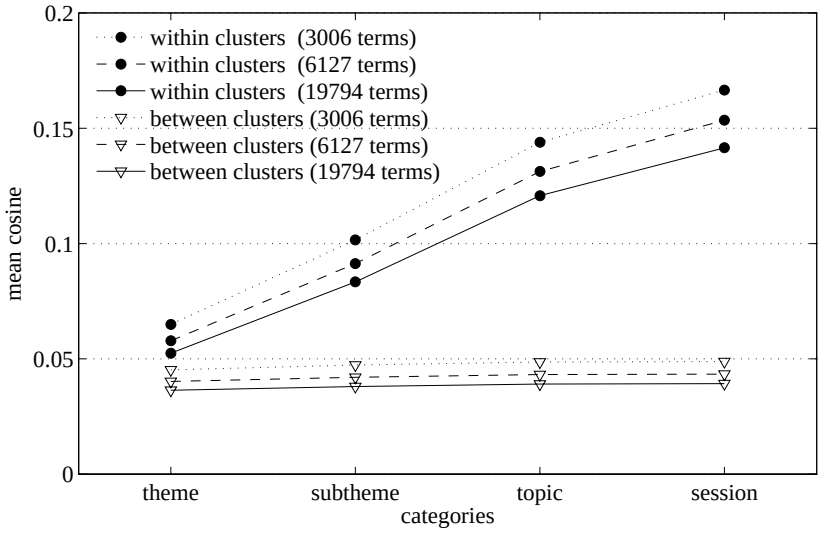


Fig. 1: Mean cosines among original categories in the 3006, 6127 and 19794 term spaces.

1. Build a similarity matrix C with mean cosines between categories as entry and mean cosines within categories on its diagonal,
2. Compute a dissimilarity matrix $D = -log(C)$, in order to have squared distance measures instead of similarities,
3. Map the categories using multidimensional scaling (MDS) [7] or Spherical Embedding algorithm [8] (using the dissimilarity matrix D as input distances) into a 2-D or 3-D space. [1]
4. Plot the 2-dimensional layout of categories, marked according to the dominant theme, that is the theme, which has the largest number (majority) of abstracts among all the abstracts belonging to that category.

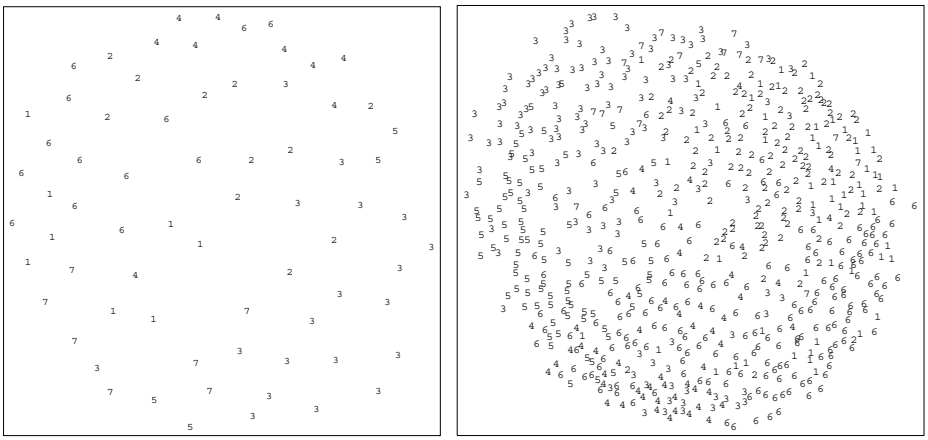
The layouts resulting from least squares MDS mapping of 2 types of categories (*subtheme* and *session*) are presented in Figure 2. We observe that the items of these 2 types of categories are mapped in good agreement with the *theme* categories because their marks are clustered. This also confirms the conclusion of section 3.1.

4 Identification of documents subsets

4.1 Proposed approach for topic identification

We assume that documents belonging to a given category refer to a common topic. The topics of the categories are naturally best described by their given

[1] MDS was used rather than PCA because the feature matrix $\mathbf{F}$ is too large to allow its direct decomposition by the classical (non-sparse) versions of PCA calculations.



(a) 71 *subtheme* categories (b) 650 *session* categories

Fig. 2: MDS layouts of original categories in the 3006 terms space. The different numbers represent the dominant themes in each category.

titles, so we just wanted to check to what extend are we able to retrieve these titles. The topic of a set of documents was identified by extracting the most important terms occurring in these documents. To this purpose, 3 ranking schemes were used: a) the Document Frequency (denoted hereafter DF), b) the Term Frequency-Inverse Document Frequency, or TF-IDF (hereafter TI), c) the Log-Entropy (hereafter LE). They are defined for each term $t_j, j = 1, ..., M$ as follows:

$$\begin{aligned} DF\,(t_j) &= \sum_{i=1}^{N} \chi\,(f_{ij}), \quad \text{with } \chi(t) = 1 \text{ if } t > 0 \text{ and } \chi(0) = 0 \\ TI\,(t_j) &= \sum_{i=1}^{N} f_{ij}\ .\ \log\left(\frac{N}{\sum_{i=1}^{N}\chi(f_{ij})}\right), \\ LE\,(t_j) &= \sum_{i=1}^{N} \log\,(1 + f_{ij})\ .\ \left(1 + \sum_{i=1}^{N} \frac{p_{ij}\log p_{ij}}{\log N}\right), \quad \text{with } p_{ij} = f_{ij} / \sum_{i=1}^{N} f_{ij} \end{aligned} \tag{2}$$

For each type of category, the top 20 terms were selected using the 3 rankings defined above, in the 3 term spaces built in section 2. The numbers of terms (among the top 20 ranked or all the terms) matching after stemming one term of the category title were counted. Table 2 presents the results. We get naturally the best possible results when taking all the terms (NO ranking) extracted from the abstracts. We can see that the log-entropy ranking (LE) performs the best among the 3 rankings, with an average retrieval score of 54.0 % (against 53.1 % for DF and 13.0 % for TI). Another result is that there is no significant decrease of performance when the term space size k decreases, which means that the strategy based on Document Frequency for building the terms space is sensible.

Table 2: Numbers of retrieved terms of the categories titles among the top 20 terms using different rankings (*TI, DF, LE*) or among all terms (*NO* ranking). The percentages in parenthesis are calculated wrt the numbers of title terms in the fourth column.

M	Category titles			Top 20 terms rankings						All terms	
	name	(# cat.)	# terms	*DF*	(%)	*TI*	(%)	*LE*	(%)	*NO*	(%)
3006	*theme*	(7)	16	3	(18.8)	3	(18.8)	**4**	(25.0)	15	(93.7)
	subtheme	(71)	168	87	(51.8)	43	(25.6)	**88**	(52.4)	151	(89.9)
	topic	(415)	1111	606	(54.5)	163	(14.7)	**610**	(54.9)	976	(87.8)
	session	(650)	2191	1138	(51.9)	289	(13.2)	**1163**	(53.1)	1883	(85.9)
6127	*theme*	(7)	16	3	(18.8)	3	(18.8)	**4**	(25.0)	15	(93.7)
	subtheme	(71)	168	89	(53.0)	40	(23.8)	**90**	(53.6)	158	(94.0)
	topic	(415)	1111	615	(55.4)	154	(13.9)	**619**	(55.7)	1022	(92.0)
	session	(650)	2191	1152	(52.6)	256	(11.7)	**1179**	(53.8)	1968	(89.8)
19794	*theme*	(7)	16	3	(18.8)	3	(18.8)	**4**	(25.0)	15	(93.7)
	subtheme	(71)	168	89	(53.0)	39	(23.2)	**90**	(53.6)	160	(95.2)
	topic	(415)	1111	615	(55.4)	141	(12.7)	**617**	(55.5)	1041	(93.7)
	session	(650)	2191	1153	(52.6)	222	(10.1)	**1179**	(53.8)	2000	(91.3)

4.2 Identified topics for the original categories

Table 3 presents a list of the 10 first session titles for which all title terms are among the top 20 log-entropy ranked terms, extracted from the posters' titles and abstracts belonging to this session. There were 130 entirely retrieved titles among the 650 sessions.

5 Clustering of the abstracts and evaluation

5.1 Clustering experiments

The primary rationale for clustering the abstracts is to try to build the different thematic categories in an automatic manner. For this reason, and to allow a comparison with the original categories, the abstracts were clustered into k clusters, for $k = 7, 71, 415$ and 650. Among the numerous existing clustering algorithms, we chose k-means for this analysis, because it was reported to perform well on documents [5]. K-means was used in two versions: (i) standard (naive) k-means and (ii) bisecting k-means (or *repeated bisections*) introduced in [5]. The k-means algorithm has been successfully applied to cluster large collections of documents as it scales relatively well with the space dimensionality, especially when the cosine similarity is used and the vectors are normalized [9], in the so-called *spherical k-means*. Matlab `kmeans` function with cosine distance measure was used as spherical k-means, and the repeated bisections k-means used was the `vcluster` function (with default parameters) from CLUTO clustering package [10]. In a purpose of comparing these two versions of k-means clustering, the clusters resulting from both functions have been evaluated by comparison

Table 3: 10 session titles with the selected terms in the 3006 term space. Boldface terms matched one title word after stop word removal and stemming. Title words like *and, other, neural* or numbers are in the stop list.

Session title	Top 20 terms (log-entropy ranking)
Serotonin Receptors I	**receptors** HT proteins rats functional agonist signals antagonist **serotonin** regulation Inhibition drugs brain dose injecting pathway assay coupled OH DPAT
Ion Channels: Trafficking and Other	**channel** proteins membrane subunits functional **ions** regulation interaction voltage form hippocampal domains gating dendritic cultured potentials conductance local surface **trafficking**
Dopamine Transporters I	DAT **transport dopamine** DA regulation proteins uptake functional phosphorylated surface terminal interaction synaptic internal site Inhibition cocaine membrane trafficking kinase
Short-Term Plasticity	synaptic potentials synapse presynaptic action depolarized recordings **short** release **term** regulation **plasticity** Layer cortical Inhibition vesicle trains amplitude form transmission
LTD I	**LTD** receptors synaptic depressant mGluRs hippocampal long term CA1 proteins form plasticity stimulation synapse NMDAR AMPA glutamate DHPG required AMPAR
Neural Oscillators	membrane potentials intrinsic **oscillation** spike models dynamics depolarized recordings properties voltage hyperpolarizing channel synaptic mV clamp conductance thresholding slowing low
Retina I	**retinal** light photoreceptors functional visual recordings mice bipolar rods proteins processes cones Dark synapse determined membrane receptors degeneration rats synaptic
Retina II	**retinal** ganglion receptors functional RGCs light pathway ON Layer visual recordings dendritic stimulus properties signals mice stimulation modulation field photoreceptors
Eye Movements: Saccades	**saccadic eye** monkey stimulus fixating visual **movements** error direct anti located field instructed pro cue reaction SC points signals Inhibition
Trigeminal Processing	**trigeminal** rats pain injecting receptors nociception regions modulation behavioral stimulation chronic central ganglion formalin nucleus **processes** hyperalgesia sensitive sensory spinal

with previously defined classes, namely the thematic categories provided by the meeting's organizers. We used the following external measures of clusters validity: *purity*, *entropy*, *F-measure* and *Mutual Information*, as proposed in [11]. These measures assess to which extend two objects from the same class (category) are in the same cluster and vice-versa. Table 4 summarizes the evaluation of clusters obtained by standard and repeated bisections k-means in 3006 and 6127 term spaces, clustering in the 19794 term space was not performed due to excessive memory requirements. It can be observed that repeated bisecting k-means algorithm performs better in terms of Entropy and Mutual Information, whereas spherical k-means is better in terms of Purity and F-measure. Relying primarily on Mutual Information, which is a theoretically well founded and unbiased measure, we conclude that our experiments confirm that repeated bisection performs better than spherical k-means, as reported in [5]. For both of the applied clustering techniques, the quality of the clusters increases with a decreasing k, indicating that categories *theme* and *subtheme* correspond in

Table 4: External measures of cluster validity for the clusterings obtained from spherical k-means and repeated bisecting k-means. An up arrow ↑ (resp. down arrow ↓) below the measure name indicates that a higher (resp. lower) value means a better clustering. Boldface entries identify the best result according to each measure, for each (M, k) pair.

clustering algorithm	M	k	Purity ↑	Entropy ↓	F-measure ↑	Mut. Inf. ↓
spherical k-means	3006	7	**0.543**	0.344	**0.486**	0.251
		71	**0.441**	0.510	**0.359**	0.404
		415	**0.285**	0.608	**0.253**	0.559
		650	**0.240**	0.641	**0.242**	0.635
	6127	7	**0.565**	0.363	**0.517**	0.270
		71	**0.448**	0.512	**0.363**	0.407
		415	**0.299**	0.617	**0.266**	0.568
		650	**0.255**	0.648	**0.252**	0.642
repeated bisecting k-means	3006	7	0.505	**0.300**	0.427	**0.207**
		71	0.380	**0.459**	0.302	**0.353**
		415	0.248	**0.578**	0.216	**0.528**
		650	0.206	**0.612**	0.207	**0.606**
	6127	7	0.507	**0.301**	0.434	**0.208**
		71	0.384	**0.464**	0.298	**0.359**
		415	0.253	**0.581**	0.219	**0.532**
		650	0.210	**0.615**	0.209	**0.609**

these term spaces to real clusters in a better way than *topic* and *session* categories. The results are slightly better in the 6127 term space in terms of Purity, whereas the 3006 term space performs better in terms of Entropy and Mutual Information, this last term space having a lower amount of 'noisy' terms.

5.2 Identification of topics for the clusters

Once we have performed the clustering of the documents, we extracted terms from the abstracts of each obtained cluster in a similar manner as in section 4.1, in order to identify the topics covered by the clusters. We selected again the top 20 terms according to a log-entropy ranking of the terms occurring in the cluster's documents. Finally, we assigned each cluster to one original category, in order to check the selected terms against the category's title (for $k = 7$ clusters, we assigned each cluster to one of the 7 themes, for $k = 71$, we assigned to one of the 71 subthemes, and so on...). The assignment was done to the *dominant* category: For all the documents in a cluster, the original categories of the documents were counted (we built the histogram of the categories) and the cluster was assigned to the category for which the number of documents was the largest. The top 10 terms, according to the LE ranking, were selected in the 3006 and 6127 term spaces. The numbers of retrieved title terms of the assigned categories is expectedly lower than for the original categories (we select only 10

terms instead of 20 and we don't use the original categories defined by human experts), but still satisfying with an average of 32.1% retrieved title terms in the 3006 term space, and 34.0% in the 6127 terms space. This demonstrates that the k-means approach is well suited to this practical application. As an illustration, a list of top 10 terms for 10 clusters (for which all the assigned title's terms were retrieved) obtained by repeated bisections with $k = 415$ is presented in Table 5. Boldface terms matched, after stemming, one word from the assigned category title.

Table 5: Selected terms identifying topics of 10 clusters among the 66 category titles entirely retrieved (out of the 415 *topic* categories) in the 3006 terms space.

Assigned title	Top 20 terms (log-entropy ranking)
Maternal behavior	**maternal behavioral** pups rats care offspring lactate mothers mice receptors
Opioid receptors	morphine **opioid receptors** tolerance rats mice analgesia injecting analgesic dose
Motor unit	muscle contract Forced **motor** isometric voluntary **unit** EMG rate variables
Aggression	**aggression** behavioral social mice Intruder receptors brain models rats Resident
Alcohol	ethanol rats **alcohol** intake consumption receptors drinking behavioral water dose
Metabotropic glutamate receptors	mGluRs **receptors glutamate metabotropic** III rats synaptic mGluR5 synapse regulation
Reward	NAc rats accumbens nucleus behavioral DA **reward** drugs dopamine shell
Cocaine	**cocaine** drugs exposure rats receptors brain behavioral abstinence withdrawal regions
Transplantation	grafting rats **transplants** axonal regenerate cord nerves Survival spinal injury
Parkinson's disease Models	MPTP mice **Parkinson disease models** PD DA dopamine dopaminergic striatal

6 Conclusions

This preliminary analysis of abstracts of posters presented at SfN 2006 Annual Meeting shows that the original thematic categories are to some extend separated in the term spaces extracted from posters abstracts and titles: it was possible to extract from the documents 54.0% of all the titles words of these categories. The log-entropy ranking scheme performed better than TF-IDF or DF rankings. A clustering of the abstracts using two versions of k-means algorithm resulted in clusters of higher average quality for repeated bisections in terms of Entropy and Mutual Information. An identification of topics, performed by selection of terms from the abstracts was also performed. Each of the obtained clusters was assigned to one original thematic categories by choosing the category with the majority of abstracts. These clusters were also evaluated in terms of their capacity to retrieve their assigned category titles. The achieved performance is satisfying as compared to the retrieval rates for original categories. The results

can be further improved, e.g. by applying more elaborate methods for the selection of relevant terms, in particular by using bigrams. By construction, k-means algorithms assume that the clusters are spherical and of similar densities, which might be untrue in the case of documents. An effort towards finding clustering techniques that are better suited to documents collections is noticeable in the literature, among others based on Nonnegative Matrix Factorization. A comparison of these techniques with the approach adopted in the present research is envisaged.

Acknowledgments

The authors wish to thank to the Society for Neuroscience for granting them the use of SfN 2006 Annual Meeting abstracts. Discussion with N. Ueda from NTT-CS, Kyoto, Japan and a collaboration on the Vector Space Model construction with T. Taniguchi from IVIS Inc., Tokyo, Japan are gratefully acknowledged.

References

1. Usui, S.: Visiome: Neuroinformatics Research in Vision Project. Neural Networks, **16** (2003) 1293–1300
2. Porter, M.: An algorithm for suffix stripping. Program, 14(**3**) (1980) 130–137
3. Usui, S., Palmes, P., Nagata, K., Taniguchi, T., Ueda, N.: Keyword Extraction, Ranking, and Organization for the Neuroinformatics Platform. Bio Systems, **88** (2007) 334–342
4. Kolda, T. G.: Limited-memory matrix methods with applications. University of Maryland, CS-TR-3806, chap. 7, (1997) 59–78
5. Steinbach, M., Karypis, G., Kumar, V.: A comparison of documents clustering techniques. In KDD Workshop on Text Mining (2000)
6. Tan, P. N., Steinbach, M., Kumar, V.: Introduction to datamining. Addison-Wesley (2006)
7. Groenen, P.: Modern multidimensional scaling: Theory and Applications. Springer Series in Statistics, Springer (1996)
8. Saito, K., Iwata, T., Ueda, N.: Visualization of Bipartite Graph by Spherical Embedding. JNNS (in Japanese) (2004)
9. Dhillon, I. S., Modha, D. S.: Concept decomposition for large sparse text data using clustering. Machine Learning, Issue 1/2, **42** (2001) 143–175
10. CLUTO, Karypis, G., et al.: University of Minnesota, (2003) available at: `http://glaros.dtc.umn.edu/gkhome/views/cluto`
11. Strehl, A., Ghosh, J., Mooney, R.: Impact of similarity measures on Web-page clustering, In Proc. AAAI Workshop on AI for Web Search (AAAI 2000), Austin, AAAI-MIT Press, (2000) 58–64