Exploration of a collection of documents in neuroscience and extraction of topics by clustering

Antoine Naud and Shiro Usui
Laboratory for Neuroinformatics, RIKEN Brain Science Institute
2-1 Hirosawa, Wako-shi, 351-0198 Saitama, Japan
naud@brain.riken.jp, usuishiro@riken.jp

Abstract

This paper presents an exploratory analysis of the neuroscience knowledge domain, and an application of cluster analysis to identify topics in neuroscience. A collection of posters abstracts from the Society for Neuroscience (SfN) Annual Meeting in 2006 is first explored by viewing existing topics and poster sessions using the 3D-SE viewer interactive tool and multidimensional scaling. In a second part, topics are determined by clustering the abstracts and selecting in each cluster the 10 terms with highest Document Frequency or Log-Entropy scores. Extracted topics are evaluated by comparison to the titles of thematic categories defined by human experts. Several Term spaces in the Vector Space Model were built on the basis of (a) a set of terms extracted from poster abstracts and titles, (b) a set of free keywords assigned to the posters by their authors. The ensuing Term Spaces are compared from the point of view of retrieving the genuine categories titles.

Key words: neuroinformatics, bipartite graph, document clustering, text mining, knowledge domain visualization

1. Introduction

The rapid growth of the amount of published documents like research papers, computer programs, analyzed data or related references gathered in databases or repositories lead to an urgent need for tools facilitating quick access to literature from a given field of research. In order to face this growing demand, an important purpose of neuroinformatics is the development of visualization tools for databases in the field of neuroscience (Usui, 2007). Another useful approach is the automatic creation of indexing structures enabling the organization of documents hierarchically. These structures may help the user in his search for information, as well as they fasten the retrieval of relevant documents and provide ways to overview a corpus that can help navigation. In databases dedicated to a broad field of research such as neuroscience, it is necessary to build a structure of keywords reflecting the semantic contents of the documents. For this purpose, we propose to detect the general structure of a collection of documents through a clustering of the documents into groups covering similar topics. This work is devoted to the analysis of a collection of 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. This study focuses on the automatic extraction of topics covered by posters based on clustering. The topics are featured using (a) the most frequent terms extracted from poster abstracts and titles, and (b) the keywords assigned to posters by their authors. A comparison of the capability of the ensuing Term

Spaces to retrieve the genuine categories defined by human experts is investigated. A possible practical application of this work is the automatic grouping of posters or other presentations into sessions for future SfN Annual Meetings.

2. Exploratory analysis of original categories

Four types of categories are provided by the organizers of the Meeting, namely the *theme*, *subtheme*, *topic* and *session* types that are used to build a tree structure with research subjects. The *theme*-type categories (called hereafter simply *themes*) are the most general ones and placed on top of this hierarchy. Each *theme* is subdivided into a number of *subthemes*, and similarly, each *subtheme* is subdivided into different *topics*. An excerpt of the list of category titles structured in 3 levels is presented in Table 1. Among all the 12856 posters existing on the CD, we selected the 12844 posters for which both an abstract and a title were given. Each retained poster (called hereafter *document*) is assigned by a committee member of SfN Annual Meeting to one poster session and is featured by a topic, a subtheme and a theme. On the basis of these assignments of the posters, we determined for each category of type subtheme, topic and session the *dominant theme* by looking at the theme of all the posters in a category and checking which theme has the largest number of posters. The dominant themes are used to color the category markers on the displays. From the assignments of the 12844 posters, lists of 7 themes, 71 subthemes, 415 topics and 650 sessions were built. We are primarily interested in the visualization of the above categories in order to provide an overview of the field and check whether the ensuing groupings of posters into categories are homogeneous and naturally cluster in the Term Spaces defined in the following section 2.1. Two visualization techniques were used: 3D-SE viewer and multidimensional scaling, so that the particular advantages of each approach could be exploited.

2.1. *The construction of Term Spaces*

The *Vector Space Model* (Salton et al., 1975) is the most widely used approach in Natural Language Processing. 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 (Porter, 1980). 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 $\{f_{ij}\}$ in entries, as a [$terms \times documents$] matrix or as a [$documents \times terms$] matrix, where each document is a row vector in the space of all terms occurring in documents. This space of all terms is called *Term Space* in the present paper. Depending on the size of the Term Space, terms occurring too often or very seldom in documents can 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 in the range of tens of thousands, leading to very high dimensional Term Spaces. In order to reduce the Term Space dimensionality, it is necessary to remove less semantically significant terms by keeping only a subset of the extracted terms, which was done using a ranking of the terms according to their Document Frequency scores (denoted DF hereafter). 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 means of statistics. Two different sources of information from which words were extracted to build the Term Spaces are presented here below. Generated Term Spaces, identified hereafter by their dimension M, and the basic features of the corresponding data matrices are summarized in Table 2.

2.1.1. *Terms extracted from the posters' abstracts and titles*

The posters abstracts and titles were extracted from a CD-ROM distributed to all the participants of the Annual Meeting. Terms originating from title were given equal weight to terms extracted from the abstracts, although higher weighting for title terms is sometimes used (e.g. frequencies of title terms can be doubled to reflect the higher semantic importance of titles). Using the same preprocessing scheme and extraction of candidate terms as in Usui et al. (2007), a number $M = 40767$ of terms were extracted directly from the abstracts and titles of the $N = 12844$ posters. The number of terms in each document varies from 61 to 456, with an average of 278.86 terms per document. This space is much too large to allow further processing. A smaller Term Space was built by selecting terms occurring in at least 45 documents ($DF \geq 45$), in order to reduce the Term Space size to $M = 3006$ terms. For the sake of simplicity, only unigrams (single words) were

Table 1
The hierarchical structure of research areas in neuroscience is reflected by the categories' titles (selected categories: all themes, subthemes in theme A and topics in subtheme A1). Each category is identified by a short label (e.g. A or A1) and a full title (e.g. Development or Neurogenesis and Gliogenesis).

Themes and Subthemes of theme A	*Topics in subtheme A1*
A. Development	
A1. Neurogenesis and Gliogenesis	
A2. Axonal and Dendritic Development	A1a. Neural induction and patterning
A3. Synaptogenesis and Activity-Dependent Development	A1b. Neural stem cells: Basic biology
A4. Developmental Cell Death	A1c. Neural stem cells: Clinical applications
A5. Development of Motor Systems	A1d. Neural stem cells: Neurogenesis after birth
A6. Development of Sensory and Limbic Systems	A1e. Proliferation
A7. Transplantation and Regeneration	A1f. Cell migration
A8. Evolution of Development	A1g. Cell lineage and cell fate specification
B. Neural Excitability, Synapses, and Glia: Cellular Mechanisms	A1h. Neuronal differentiation: Autonomic and sensory neurons
C. Sensory and Motor Systems	A1i. Neuronal differentiation: Central neurons
D. Homeostatic and Neuroendocrine Systems	A1j. Glial differentiation
E. Cognition and Behavior	A1k. Neuron glia interactions
F. Disorders of the Nervous System	
G. Techniques in Neuroscience	
H. History and Teaching of Neuroscience	

considered as terms in this study.

2.1.2. *Free keywords provided by the posters authors*

Free keywords were also extracted from the Annual Meeting's CD where 5 separate XML tags are given. A total of 12695 posters were assigned from 1 to 5 such keywords, with an average of 4.26 keywords per poster. After basic data cleaning (correction of misspelling and other typos in keywords) and simple stemming (elimination of plurals), a set of 10022 keywords was established. This excessively high dimensionality of the Term Space was reduced to the $M = 3560$ keywords assigned to two or more posters ($DF \geq 2$).

2.2. *Visualization of categories by* 3D-SE viewer

The 3D-SE viewer [1] visualization tool is based on Spherical Embedding (Saito et al., 2004), an algorithm designed for the visualization of bipartite graphs. In order to build an interactive tool usable on web pages, the 3D-SE viewer has been implemented as a Java applet (Usui, 2007), which has been successfully applied to the visualization of documents and concepts (Naud et al., 2007a). The sparse term frequency matrix $\mathbf{F}$ may be conveniently viewed as a bipartite graph $G = \{V_A \cup V_B, E\}$ in which the sets of vertices V_A and V_B contain e.g. terms and documents, and the set of edges E is build from the occurrences of terms in documents. The visualized items are represented on two concentric spheres embedded in a 3-D Euclidean space, for instance terms are mapped on the inner sphere and documents on the outer sphere. This interactive tool allows the user to modify the viewpoint by rotating the spheres around their center, zooming in or out, or centering the view on selected nodes, and allows to hyperlink the nodes to other web pages. The lists of visualized items are displayed in panels on both sides of the central view. 3D-SE viewer was used to visualize some of the genuine categories, namely topics and sessions as sums of their respective documents, providing an general overview of neuroscience on the outer sphere and access to terms or keywords on the inner sphere. Figure 1 presents an overview of the 415 topics in the space of 3006 terms extracted from abstracts. Groupings of topics according to the main themes are clearly visible. Figure 2 presents a view of the 650 poster sessions in the space of 3560 free keywords, with a focus on the *Neuroinformatics* poster session.

2.3. *Visualization of categories by multidimensional scaling*

Multidimensional scaling (MDS) (Borg and Groenen, 2005) is a classical family of techniques used for the visualization of multidimensional data. Least-squares MDS is based on the minimization of a Stress function involving the differences between Euclidean distances in the high dimensional space and the target 2-D or 3-D space. MDS is preferred here to a PCA-based dimensionality reduction because the feature matrix $\mathbf{F}$ is too large to allow its direct decomposition by the classical (non-sparse) versions of PCA. The previously defined Term Spaces being still very high-dimensional (with several thousands of dimensions) and data being very sparse, a direct application of MDS is not possible

[1] 3D-SE viewer ©BSI NI lab. and NTT-CS.

Table 2
Term Spaces built for the representation of posters. nnz is the number of non-zero elements in matrix $\mathbf{F}$, S is the sparseness of $\mathbf{F}$ defined as $S = 1 - nnz/(M \cdot N)$. Term frequency matrices are usually very sparse, typically $S = 99\%$, the extracted data are even more sparse than this in the free keywords case.

#	*source of terms*	*selection*	*# documents* N	*# terms* M	*nnz*	*sparseness* S (%)
1.	abstract and title	*no selection*	12844	40767	1008321	99.81
2.	abstract and title	$DF \geq 45$	12844	3006	857839	97.78
3.	free keywords	*no selection*	12695	10022	54376	99.96
4.	free keywords	$DF \geq 2$	12695	3560	47914	99.89

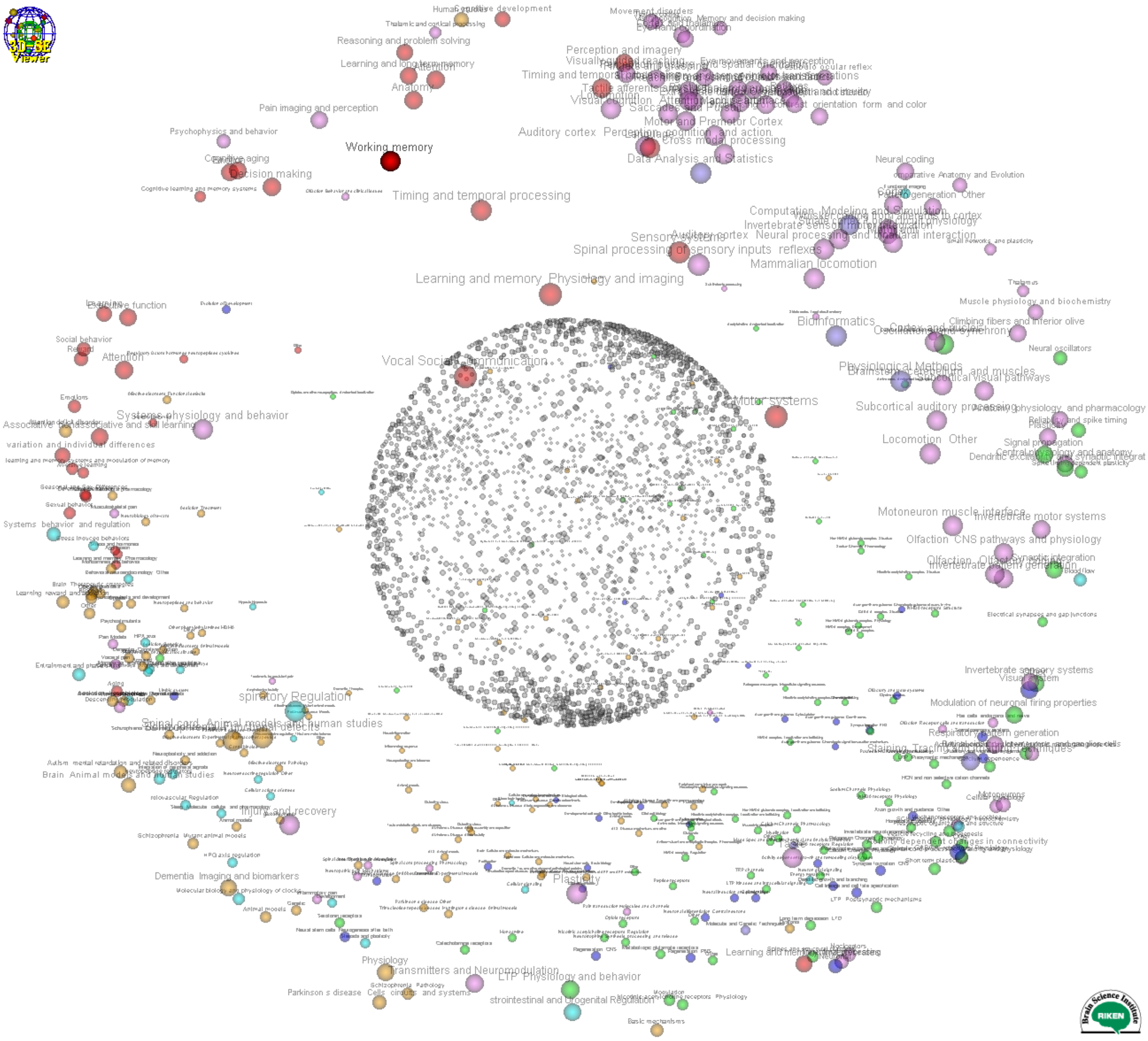

Fig. 1. 3D-SE viewer: an overview of the 415 topics in the space of 3006 terms extracted from abstracts. The 7 main themes are displayed in distinct areas.

due to the curse of dimensionality causing distances to become meaningless. In order to reduce this effect, a similarity matrix based on average cosine measures between categories is first computed, this matrix is then transformed into a dissimilarity matrix and used as input to the MDS algorithm.

2.3.1. *Average cosine measures between categories*

The frequency matrix $\mathbf{F}$ is a sparse contingency table where each row represents one document, and

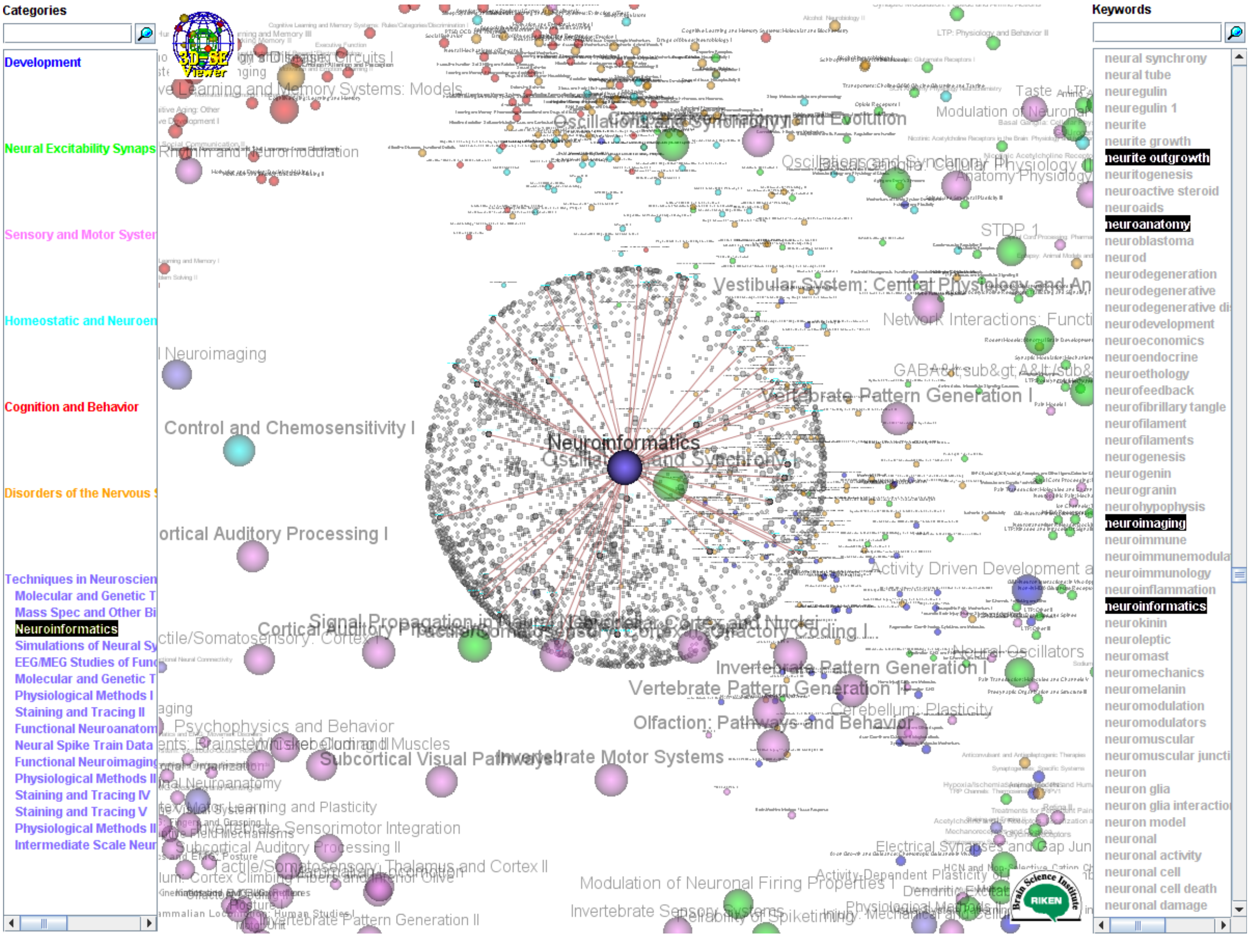

Fig. 2. 3D-SE viewer: a view of the 650 poster sessions in the space of 3560 free keywords, with a focus on the *Neuroinformatics* poster session.

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 Kolda (1997) for a review of weighting schemes). The cosine between 2 vectors in the high-dimensional Term Space is defined as

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

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 in the two reduced Term Spaces (3006 and 3560) using the centroid vectors of each category, as described in Steinbach et al. (2000). The resulting means are presented in Figure 3. 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 studied Term Spaces. The average cosine between categories in the free keywords space are significantly lower, which is due to the higher sparseness of data in this Term Space. The above two average cosines among categories are equivalent to clusters' *cohesion* and *separation*, some internal measures of clusters validity presented e.g. in Tan et al. (2006).

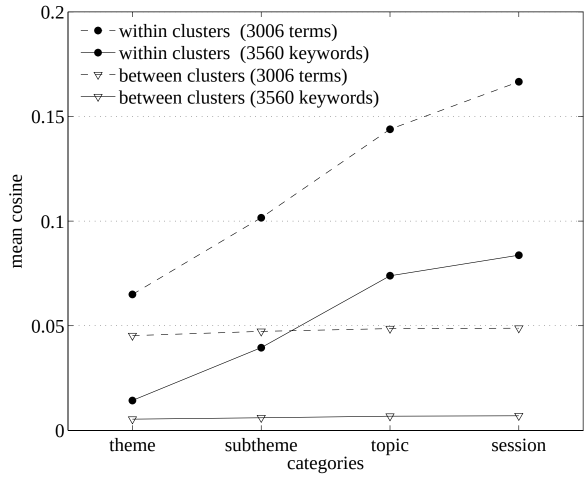

Fig. 3. Mean cosines among documents in the original categories in the 3006 and 3560 Term Spaces.

2.3.2. *Proposed scheme for the visualization of categories*

As illustrated in Section 2.3.1, the different average cosines between and within categories are larger for *topic* and *session* categories, indicating that these categories are better separated in our Terms Spaces. This can be confirm by visualizing the different categories. To this purpose, we processed the data as follows:

(i) Build a similarity matrix C with mean cosines between categories as entry and mean cosines within categories on its diagonal,

(ii) Compute a dissimilarity matrix $D = -log(C)$, in order to obtain distance-like measures instead of similarities,

(iii) Map the categories into a 2-D or 3-D space using MDS using the dissimilarity matrix D as input distances,

(iv) Plot the 2-dimensional layout of categories, marked according to the dominant theme.

Figure 4 (and Figure 5) presents the layout of 2 types of the 71 *subthemes* (and respectively 650 *sessions*) resulting from least squares MDS mapping. We observe that the items of these 2 types of categories are mapped in good agreement with the *theme* categories because their marks are grouped according to their *theme* color. The almost uniform distribution of nodes in the target space is also remarkable and suggests a good separation in the input high dimensional space, although no clear demarcation is visible between the areas occupied by the different themes.

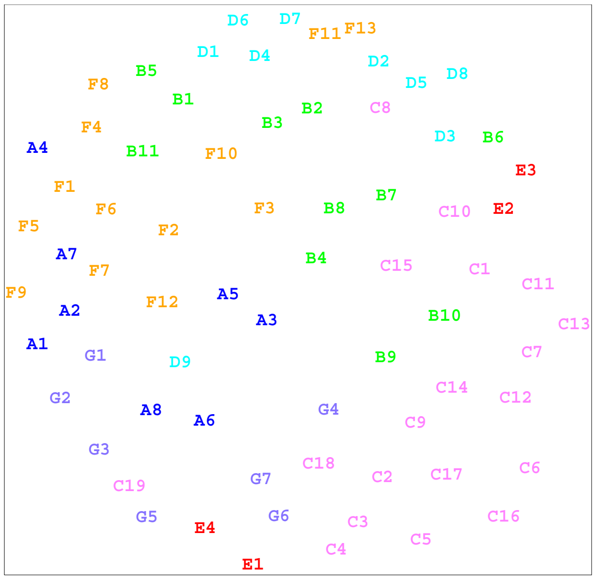

Fig. 4. MDS visualization: 2D layouts of 71 *subtheme* categories in the 3006 Term Space. Each *subtheme* is marked using its short label colored according to its dominant theme.

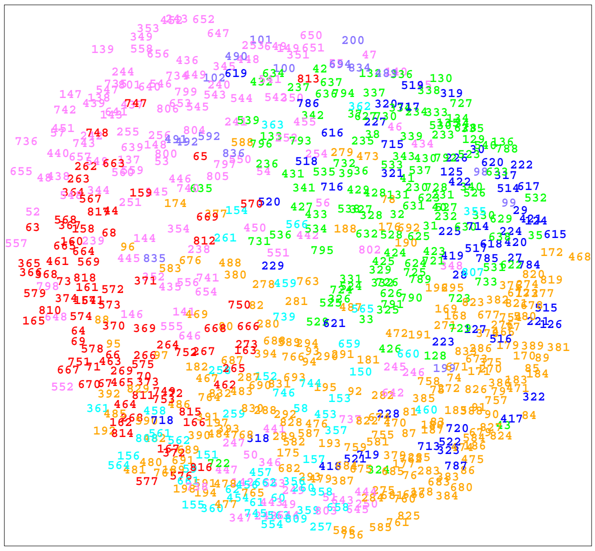

Fig. 5. MDS visualization: 2D layouts of 650 *session* categories in the 3006 Term Space. Each *session* is marked using its identification number colored according to its dominant theme.

3. Identification of topics by document clustering

3.1. *Recent trends in document clustering*

Document clustering has drawn the interested of researchers in Natural Language Processing for more than two decades. Some recent trends in this area are briefly outlined in this section. Document clustering is a task that has received much attention in recent years due to the rapid growth of documents available on the Web. The newly developed clustering techniques exploit naturally the graph formed by hyperlinks connecting documents to each other. Another recent active area of research is clustering of documents enriched with ontologies (Yoo et al., 2006), in which similarities between documents incorporate inter-concepts semantic relationships in a given knowledge domain captured by the appropriate ontology. Both hierarchical/agglomerative clustering (Zhao et al., 2005) and partitional clustering (mainly based on k-means) (Dhillon et al., 2000) have been successfully applied to this task. Co-clustering refers to a more recent approach in which both words and documents are clustered at the same time (Dhillon, 2001). The clusters may be disjoint as in information-theoretic co-clustering (Dhillon et al., 2003), or overlapping using probabilistic modeling as proposed in (Banerjee et al., 2005). Non-negative Matrix Factorization (NMF) is another successful approach in document clustering, being based on a decomposition of the frequency matrix into a product of two non-negative matrices (Xu et al., 2003).

3.2. *Proposed approach for topic identification*

It is assumed that documents belonging to a given subset of documents (cluster or category) refer to a common topic. The topics of the existing categories are naturally best described by the titles their are given, and our aim is to check to what extend it is possible to retrieve these titles. The topic(s) covered by a cluster of documents can be identified by a list of the most meaningful terms occurring in these documents. To this purpose, these terms were ranked according to a specific score and the top 10 terms were retained to describe the topic. Several ranking schemes for selecting terms have been tested in Naud et al. (2007b). The two best performing rankings were applied in this study, namely Document Frequency (DF, the same as used to reduce the Term Space dimensions in Section 2.1) and Log-Entropy (denoted hereafter LE). They are defined for each term $t_j, j = 1, ..., M$ as follows:

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

For each type of category, the top 10 terms were selected using these 2 rankings, in the 4 Term Spaces defined in section 2.1. The numbers of terms (among the top 10 ranked or among all the terms) exactly matching after stemming one term of the category title were counted, they are presented in Table 3. We get naturally the best results when taking all the terms in the Term Space (*NO* column), and LE ranking performs always better than DF. Another result is that there is no dramatic decrease of performance when the Term Space size is decreased by a factor of order of 10 (40767/3006), which means that the DF-based strategy for building the terms space is sensible. In the 40767 Term Space, the 6.68% of unretrieved title words is mostly due to misspelled words in the abstracts. The performance is lower for the two Term Spaces based on keywords, this result is due to the fact that free keywords are often very specialized terms, and hence not suitable for being part of a category title.

3.3. *Identification of topics in the original categories*

Table 4 presents a list of 10 session titles for which all the words were among the top 10 LE-ranked 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 *Roman Numbers* are in the stop list. These titles were entirely retrieved, as 90 other session titles out of the 650 sessions.

In order to illustrate the kind of difficulties arising in the keywords Term Spaces, a list of 15 *subtheme* category titles together with the top 10 LE-ranked keywords selected from the 10022 Term Space is shown in Table 5. Titles like *Data Analysis and Statistics* difficult to retrieve because they involve

Table 3
Numbers of retrieved terms of the categories titles among the terms from the original categories documents in different Term Spaces. The top 10 terms using DF and LE rankings or without ranking (among all 3006 terms) are compared. The percentages in parenthesis are calculated wrt the numbers of title terms in the fourth column.

M	Category titles			Term ranking		All terms
	name	(# cat.)	# terms	DF (%)	LE (%)	NO (%)
40767	*theme*	(7)	16	3 (18.75)	2 (12.50)	15 (93.7)
	subtheme	(71)	168	75 (44.64)	75 (44.64)	164 (97.6)
	topic	(415)	1111	523 (47.07)	522 (46.98)	1051 (94.6)
	session	(650)	2191	984 (44.91)	998 (45.55)	2023 (92.2)
3006	*theme*	(7)	16	3 (18.75)	2 (12.50)	15 (93.7)
	subtheme	(71)	168	74 (44.05)	74 (44.05)	151 (89.9)
	topic	(415)	1111	519 (46.71)	519 (46.71)	976 (87.8)
	session	(650)	2191	973 (44.41)	988 (45.09)	1883 (85.9)
10022	*theme*	(7)	16	3 (18.75)	3 (18.75)	13 (81.2)
	subtheme	(71)	168	72 (42.86)	72 (42.86)	145 (86.3)
	topic	(415)	1111	343 (30.87)	343 (30.87)	887 (79.8)
	session	(650)	2191	587 (26.79)	587 (26.79)	1788 (81.6)
3560	*theme*	(7)	16	3 (18.75)	3 (18.75)	12 (75.0)
	subtheme	(71)	168	72 (42.86)	72 (42.86)	130 (77.4)
	topic	(415)	1111	342 (30.78)	342 (30.78)	817 (73.5)
	session	(650)	2191	590 (26.93)	590 (26.93)	1662 (75.9)

Table 4
Identification of topics in the original categories: A list of 10 session titles together with the top 10 LE-ranked terms from the original categories' documents, in the 3006 Term Space.

Session title	Top 10 terms (LE ranking)
Cognitive Aging: Other	**age** adult older **cognitive** processes functional regions participated decline young
Entrainment and Phase Shifts	light SCN **phase** circadian **entrainment** clock rhythms **shift** cycling Dark
Eye Movements: Saccades	**saccadic eye** monkey stimulus fixating visual **movements** error direct anti
Inflammatory Pain II	**pain** rats injecting **inflammatory** behavioral CFA models inflammation receptors nociception
Language I	processes area left semantic word **language** temporal speech stimuli regions
Parkinson's Disease: Other I	proteins PD **disease Parkinson** kinase mutation functional DA gene stress
Retina I	**retinal** light photoreceptors functional visual recordings mice bipolar rods proteins
Retina II	**retinal** ganglion receptors functional RGCs light pathway ON Layer visual
Sexual Differentiation	sex brain **sexual** receptors behavioral rats **differential** hormone area dimorphic
Taste	**taste** rats receptors stimuli recordings sucrose nucleus stimulation processes information

very general concepts usually not mentioned in the specialized papers abstracts.

3.4. *Clustering experiments*

The primary rationale for clustering the abstracts is to build the different thematic categories in an automatic manner. For this reason, and to allow a comparison with the original categories, the documents were clustered into k clusters, successively with $k = 7, 71, 415$ and 650. The clustering algorithm used in this purpose is the *repeated bisecting k-means* as it was reported to perform well on documents (Steinbach et al., 2000) (Naud et al., 2007b). The `vcluster` function (with default parameters `'rb'`) from CLUTO clustering package (Karypis at al., 2003) was used to perform the calculations of repeated bisecting k-means. Table 6 presents the numbers of retrieved terms of the categories titles among the terms from the clustered documents. The first column specifies the Term Space in which documents were clustered and from which terms were selected to describe the clusters' topics, in order to enable a fair comparison of the two Term Spaces. From the results presented in Tables 3 and 6, the following observations are made: 1) The "title retrieval" performances of clusters are generally lower than using the original categories, which is not surprising considering that human experts shaping the categories had more knowledge about neuroscience than is captured by the abstracts, but k-means still performed relatively well with an average rate of 31.0% against 37.1% for the original categories in the same two Term Spaces. 2) The Term Space based on abstracts lead to better results than based on the keywords, which confirms the result expressed in Section 3.3 that keywords are unlikely to appear in titles of categories.

3.5. *Identification of topics for the clusters*

Once the documents clustered, we proceeded in a similar manner as in section 3.2 in order to identify the topics covered by the documents in the found clusters. We selected again the top 10 terms among the cluster's documents according to LE ranking in

Table 5
15 *subtheme* titles with the top 10 *LE*-ranked keywords selected in the 10022 Term Space. Boldface keywords matched one title word after stop word removal and stemming. Italic titles were entirely retrieved.

subtheme title	Top 10 keywords (*LE* ranking)
Biological Rhythms and Sleep	**'sleep'** 'circadian rhythm' 'circadian' 'suprachiasmatic nucleus' 'eeg' 'sleep deprivation' 'electrophysiology' 'entrainment' 'hypocretin' 'orexin'
Brain Blood Flow, Metabolism, and Homeostasis	'blood brain barrier' 'cerebral blood flow' **'metabolism'** 'optical imaging' 'permeability' 'vascular' 'blood flow' 'energy metabolism' 'hippocampus' 'barrel'
Chemical Senses	'olfaction' 'olfactory bulb' 'electrophysiology' 'glomerulus' 'oscillation' 'coding' 'gustatory' 'taste' 'brainstem' 'odor'
Data Analysis and Statistics	'brain imaging' 'fmri' 'human' 'modeling' 'cerebral cortex' 'functional mri' 'behavior' 'eeg' 'electrophysiology' 'erp'
Demyelinating Disorders	'multiple sclerosis' **'demyelination'** 'oligodendrocyte' 'inflammation' 'myelin' 'animal model' 'microglia' 'cytokine' 'eae' 'growth factor'
Ion Channels	'potassium channel' 'calcium channel' **'ion channel'** 'sodium channel' 'hippocampus' 'patch clamp' 'excitability' 'pain' 'electrophysiology' 'calcium'
Ligand Gated Ion Channels	'glutamate receptor' 'nicotinic receptor' 'patch clamp' 'electrophysiology' 'ion channel' 'hippocampus' 'nmda receptor' 'gaba receptor' 'glutamate' 'trafficking'
Network Interactions	'hippocampus' **'network'** 'synchrony' 'oscillation' 'interneuron' 'rat' 'synchronization' 'cortex' 'epilepsy' 'modeling'
Neurogenesis and Gliogenesis	**'neurogenesis'** 'neural stem cell' 'development' 'differentiation' 'hippocampus' 'proliferation' 'stem cell' 'brdu' 'migration' 'cell cycle'
Neurotransmitter Release	'synaptic vesicle' 'exocytosis' 'synaptic transmission' 'presynaptic' 'endocytosis' 'hippocampal neuron' 'calcium' 'drosophila' 'gabaergic' **'neurotransmitter release'**
Pattern Generation and Locomotion	**'locomotion'** 'central pattern generator' 'spinal cord' 'cpg' 'serotonin' 'motor control' 'human' 'rhythm' 'invertebrate' 'neuromodulation'
Physiological Methods	'electrophysiology' 'eeg' 'behavior' 'patch clamp' 'in vitro' 'in vivo' 'ischemia' 'parkinson's disease' 'stroke' 'voltage clamp'
Synaptic Transmission	**'synaptic transmission'** 'synapse' 'hippocampus' 'presynaptic' 'gaba' 'glutamate' 'dendrite' 'interneuron' 'neurotransmitter release' 'exocytosis'
Tactile/Somatosensory	'somatosensory cortex' **'tactile'** 'barrel' **'somatosensory'** 'vibrissa' 'whisker' 'cortex' 'rat' 'thalamocortical' 'sensorimotor'
Visuomotor Processing	'motor control' 'sensorimotor' 'reaching' 'eye movement' 'saccade' 'parietal cortex' 'vision' 'visual perception' 'motor learning' 'spatial memory'

Table 6
Numbers of retrieved terms of the categories titles among the top 10 terms in *LE* ranking from the clustered documents. The percentages are ratios of numbers of found terms over the numbers of terms existing in titles of the assigned categories to the clusters.

M	k	title terms (*LE* ranking)		
		existing	*found*	(%)
3006	7	16	2	(12.50)
	71	184	46	(25.00)
	415	1051	362	(34.44)
	650	2186	679	(31.06)
3560	7	17	2	(11.76)
	71	194	53	(27.32)
	415	1111	188	(16.92)
	650	2203	312	(14.16)

Table 7
Top 10 terms identifying the topics of 10 clusters obtained from repeated bisecting k-means, among the 66 titles entirely retrieved (out of the 415 *topic* titles) in the 3006 Term Space.

Assigned title	Top 10 terms (*LE* 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

two Term Spaces. Finally, each cluster was assigned to one original category, in order to check the selected terms against the category's title. In a clustering of the documents into $k = 7$ clusters (respectively $k = 71, 415, 650$), each cluster was assigned to the *dominant category* among the 7 themes (resp. $k = 71$ subthemes, 415 topics, 650 sessions) as follows: The original categories of all the documents in a cluster were counted (making a histogram of the categories), then 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 3560 Term Spaces. As an illustration, a list of 10 *topic* titles for which all the terms were retrieved in the top 10 terms of their assigned clusters (obtained from repeated bisecting k-means with $k = 415$) is presented in Table 7. Boldface terms matched, after stemming, one word from the assigned category title.

4. Conclusions

An exploratory analysis of a collection of posters presented at SfN Annual Meeting in 2006 has been performed using the 3D-SE viewer Java applet and multidimensional scaling. The original thematic categories are displayed in distinct areas. Several Term Spaces based on posters abstracts and titles, and on free keywords were constructed and used successfully (to some extent) to retrieve the titles of original categories defined by human experts. Term Spaces based on abstracts performed better in this task than those based on free keywords. A clustering of the abstracts using repeated bisecting k-means was performed, followed by an identification of the topics covered by the documents of the resulting clusters. Each cluster was assigned to one of the original thematic categories by choosing the category with the majority of documents, and was evaluated in terms of its capacity to retrieve its assigned category title. The achieved performance is satisfying as compared to the retrieval rates for original categories. We believe that these results can be further improved: 1) by applying more elaborate methods for the selection of relevant terms, in particular by extracting N-grams ($N = 2, 3$) from abstracts, 2) by reducing further the Term Space dimensionality using e.g. Latent Semantic Analysis (Deerwester et al., 1990). Using both the terms extracted from posters abstracts and the free keywords together in one Term Space should also improve performance. K-means algorithm assumes that the clusters are spherical and of similar densities, which might be untrue in the case of documents. Other clustering techniques, among others based on Nonnegative Matrix Factorization, may be also evaluated and compared to the approach adopted in the present research.

Acknowledgments

The authors wish to thank to the Society for Neuroscience for granting them the use of SfN 2006 Annual Meeting abstracts. Discussions 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

Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., Mooney, R. J. (2005). Model-based overlapping clustering. SIGKDD international conference on Knowledge discovery in data mining, 532–537.

Borg, I., Groenen, P. J. F. (2005). Modern multidimensional scaling: Theory and Applications. 2nd edition. Springer Series in Statistics, Springer.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990) Indexing by Latent Semantic Analysis, JASIST, vol. 41, nr. 6, 391–407.

Dhillon, I. S., Modha, D. S. (2001). Concept decomposition for large sparse text data using clustering. Machine Learning, Issue 1/2, 42, 143–175.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. Knowledge Discovery and Data Mining 2001, San Francisco, California, USA, 269–274.

Dhillon, I. S., Mallela, S., Modha, D.S. (2003). Information-Theoretic Co-Clustering. Proceedings of ACM SIGKDD 2003, 89–98.

CLUTO, Karypis, G., et al. (2003). University of Minnesota, available at: `http://glaros.dtc.umn.edu/gkhome/views/cluto`.

Kolda, T. G. (1997). Limited-memory matrix methods with applications. University of Maryland, CS-TR-3806, chap. 7, 59–78.

Naud, A., Usui, S., Ueda, N., Taniguchi, T. (2007a). Visualization of documents and concepts in neuroinformatics with the 3D-SE Viewer. Frontiers in Neuroscience (Frontiers in Neuroinformatics), Volume 1, Issue 1, Nov. 2007.

Naud, A., Usui, S. (2007b). Exploration of a text collection and identification of topics by clustering. In H. Yin et al. (Eds.): IDEAL 2007, Birmingham, UK, LNCS 4881, 115–124.

Porter, M. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

Saito, K., Iwata, T., Ueda, N. (2004). Visualization of Bipartite Graph by Spherical Embedding. Proc. of the 2004 Annual Conference of JNNS (in Japanese).

Salton, G., Wong, A., and Yang, C. S. (1975). A Vector Space Model for Automatic Indexing, Communications of the ACM, vol. 18, nr. 11, 613-620.

Steinbach, M., Karypis, G., Kumar, V. (2000). A comparison of documents clustering techniques. In KDD Workshop on Text Mining.

Strehl, A., Ghosh, J., Mooney, R. (2000). Impact of similarity measures on Web-page clustering, In Proc. AAAI Workshop on AI for Web Search (AAAI 2000), Austin, AAAI-MIT Press, 58–64.

Tan, P. N., Steinbach, M., Kumar, V. (2006). Introduction to datamining. Addison-Wesley.

Usui, S., Naud, A., Ueda, N., Taniguchi, T. (2007). 3D-SE Viewer: A Text Mining Tool based on Bipartite Graph Visualization, IJCNN 2007, Orlando, Florida, USA, paper # 1303.

Usui, S., Palmes, P., Nagata, K., Taniguchi, T., Ueda, N. (2007). Keyword Extraction, Ranking, and Organization for the Neuroinformatics Platform. Bio Systems, 88, 334–342.

Xu, W., Liu, X., Gong, Y. (2003). Document clustering based on non-negative matrix factorization. Proceedings of ACM SIGIR 2003, Toronto, Canada, 267–273.

Yoo, I., Hu, X., Song, I-Y. (2006). Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering. Proceedings of the 12th ACM SIGKDD 2006, 791–796.

Zhao, Y., Karypis, G., Fayyad, U. M. (2005) Hierarchical Clustering Algorithms for Document Datasets. Data Min. Knowl. Discov, vol. 10, nr. 2, 141–168.