Visualization of documents and concepts in neuroinformatics with the 3D-SE Viewer.

Antoine Naud
Laboratory for Neuroinformatics, RIKEN Brain Science Institute
2-1 Hirosawa, Wako City, Saitama 351-0198, JAPAN
naud@brain.riken.jp
Department of Informatics
ul. Grudziadzka 5, Nicolaus Copernicus University
87-100 Torun, POLAND
naud@is.umk.pl

Shiro Usui*
Laboratory for Neuroinformatics, RIKEN Brain Science Institute
2-1 Hirosawa, Wako City, Saitama 351-0198, JAPAN
usuishiro@riken.jp

Naonori Ueda
NTT Communication Science Laboratories
Kyoto, JAPAN
ueda@cslab.kecl.ntt.co.jp

Tatsuki Taniguchi
IVIS, Inc.
Tokyo, JAPAN.
taniguti@ivis.co.jp

Abstract

A new interactive visualization tool is proposed for mining text data from various fields of neuroscience. Applications to several text datasets are presented to demonstrate the capability of the proposed interactive tool to visualize complex relationships between pairs of lexical entities (with some semantic contents), such as terms, keywords, posters or papers abstracts. Implemented as a Java applet, this tool is based on the Spherical Embedding algorithm, which was designed for the visualization of bipartite graphs. Items such as words and documents are linked on the basis of occurrence relationships, which can be represented in a bipartite graph. These items are visualized by embedding the vertices of the bipartite graph on spheres in a 3 dimensional space. The main advantage of the proposed visualization tool is that 3 dimensional layouts can convey more information than planar or linear displays of items or graphs. Different kinds of information extracted from texts, such as keywords, indexing terms or topics are visualized, allowing interactive browsing of various fields of research featured by keywords, topics or research teams.

Keywords

neuroinformatics, data visualization, bipartite graph, 3D-SE viewer, text mining, keyword extraction.

1. Introduction

Very often when dealing with textual data, people are interested in the relationships between entities belonging to two distinct categories: e.g. relationships between words and documents, between topics and documents or between authors and documents. 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 words from a collection of documents $\mathcal{D}$ followed by stop words removal and stemming (Porter, 1980). The numbers of occurrences of each term in each document (usually called *frequency*) are counted and denoted f_{ij}. A matrix **F** is built, with one row for each term and one column for each document, and with the frequencies f_{ij} as entries. When the number of documents N in the collection is in the range of a few thousands (as it is in the examples presented below), the number of terms extracted is often larger than a few tens of thousands, leading to very high dimensional space for the documents. In order enable further processing of matrix **F**, we reduce its size by selecting a number M of terms using a ranking scheme. This is done by ranking the terms according to a term weighting scheme and retaining the top M terms ($M \sim 1000$). Several term weighting schemes have been defined in the Information Retrieval literature, catching different desired properties for the terms, see (Kolda, 1997) for a review and comparison of term weighting schemes. The most popular one is probably *TF.IDF* (*term frequency inverse document frequency*) which has been used in this work. The matrix of frequencies **F** is usually sparse because most of the terms occur only in a few documents. In this case, it is convenient to regard the data as a graph which vertices represent both terms and documents, and each edge connects one term to one document if the term occurs at least once in the document. The frequencies in matrix **F** are then converted to binary entries to build a second *occurrence* matrix **O** ($o_{ij} = sgn(f_{ij})$), from which a *bipartite graph* is defined. In such a graph, each term is connected to all the documents in which it occurs, and each document is connected to all the terms from set $\mathcal{T}$ it contains, but there is no connection between terms, nor between documents.

2. Bipartite graph visualization

The purpose of bipartite graph visualization is to display simultaneously two types of relationships: the similarities existing between items within each of two subsets, on the basis of the relationships defined by the graph edges. In the terms and documents application introduced above, we are interested in seeing the similarities between terms, as well as similarities between documents, based on the occurrences of terms in the documents. In each of the applications presented in section 3, the most important information is the configuration of the graph's vertices. The edges of the graph are not of primary interest here, they are not visualized by default, although the 3D-SE viewer allows displaying them.

2.1. Formal definition of a bipartite graph

A graph G is defined as $G := \{ V, E \}$, where V is the set of vertices or nodes and E is the set of edges. The graph G is undirected if the pairs in E are unordered. An undirected graph G is called a bipartite graph if there exist a partition of the vertex set $V = V_A \cup V_B$, so that there is no edge in E connecting V_A to V_B.

2.2. The Spherical Embedding algorithm

The Spherical Embedding (SE) algorithm (Saito, 2004) was primarily designed for the visualization of bipartite graphs. The items of the two subsets V_A and V_B are represented as nodes positioned on 2 concentric spheres in a 3-dimensional Euclidean space. The number of dimensions of the embedding space was set to 3 because more information can be visualized in 3-D than in 2-D. Items from subset V_A are mapped on the inner sphere θ_A (with radius $r_A = 1$), whereas items from V_B are mapped on the outer sphere θ_B (with radius $r_B = 2$). Items positions are defined in such a way that similar items in V_A are close to each other on θ_A, and similar items in V_B are close to each other on θ_B. Figure 1 illustrates the process of bipartite graph construction and visualization in a 2-dimensional space using the Spherical Embedding algorithm.

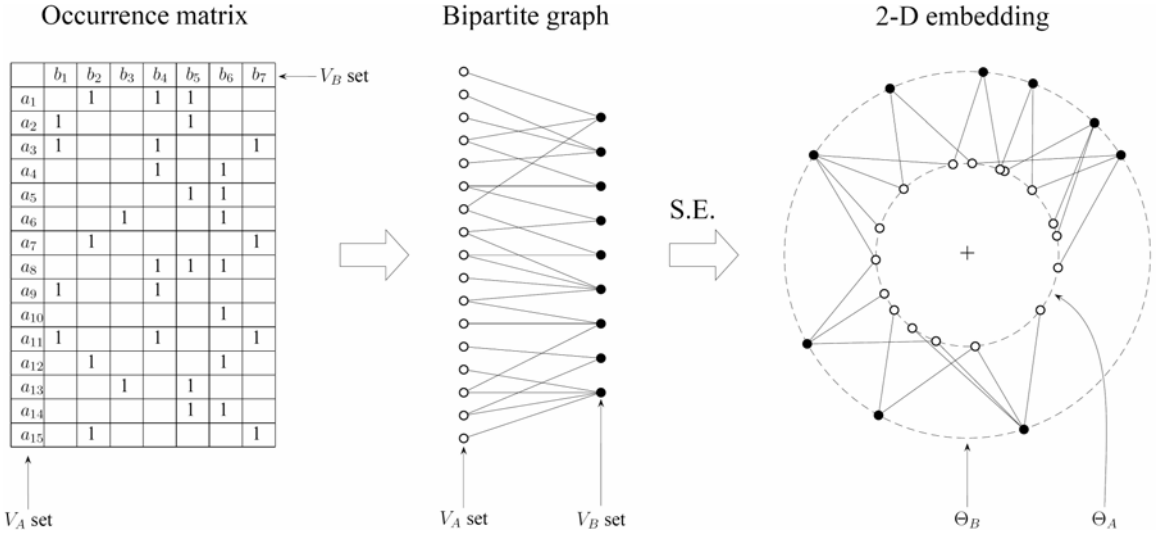

Figure 1: Visualization process: from the binary *occurrence* matrix O to the bipartite graph and its visualization using the Spherical Embedding algorithm.

To achieve this goal, we minimize a sum over the edges in E of the Euclidean distance between the corresponding nodes in the 3-D space. The minimization is performed through a gradient descent procedure, under constraints requiring that the points lie on the two spheres. This constrained optimization problem is converted to an unconstrained one using some sufficient statistics results. The whole process amounts to minimizing the sum of Euclidean distances between pairs of points along all the edges in E, that is find the nodes coordinates $\{ \mathbf{x}_i \}$ subject to $\mathbf{x}_i{}^T \mathbf{x}_i = r_i^2$ that minimize

$$\mathbf{E} = \frac{1}{2} \sum_{i=1}^{M+N-1} \sum_{j=i+1}^{M+N} w_{ij} \left(a_{ij} r_i r_j - \mathbf{x}_i{}^T \mathbf{x}_j \right)^2 \tag{1}$$

where $a_{ij} = +1$ if nodes i and j are connected and $a_{ij} = -1$ otherwise, and $r_i = r_A$ (resp. r_B) for nodes from subset V_A (resp. V_B). The $\{ w_{ij} \}$ are weights that can be used to give more emphasis on pairs of nodes belonging to E.

2.3. Related approaches

Although graph drawing is a very active field of research, very few work exist on the visualization of bipartite graphs. An interesting method called Anchor Maps (Misue, 2006) has been proposed recently. It provides a visualization of the graph in a 2-dimensional space, proceeding in two steps: the items of the first subset of vertices V_A are plotted on a circle at equal intervals, after which the vertices of the second subset V_B are added to the plot by allocating them with respect to the vertices of V_A using a spring embedding with restrictions technique, which ensures a minimization of the total length of edges and the number of crossings. A different approach is proposed by Zheng et al. (Zheng, 2005), in which a layout of points on two parallel planes is sought for, such that a view in three dimensions from which the number of observed crossings will be minimal. Drawing the vertices on planar curves, as proposed by Di Giacomo et al. (DiGiacomo, 2006), is another interesting approach. Hong et al. proposed (Hong, 2005) a layered drawing of bipartite graphs in $2^{1/2}$ dimensions, that is the vertices are allocated on two surfaces embedded in a three dimensional Euclidean space. In all these approaches, the ultimate goal is the visualization of the graph itself, that is, nodes are displayed together with lines representing the edges. In our approach, the focus is set primarily on the visualization of the vertices of the graph; although edges can be interactively displayed on user's request by selecting the corresponding node(s).

2.4. The 3D-SE viewer visualization tool

The 3D-SE viewer [1] visualization tool has been designed and developed for the general purpose of bipartite graphs visualization. In order to build an interactive tool available on web pages, it has been implemented as a Java applet. The visualized items are represented as colored nodes with labels, embedded in a 3-D Euclidean space. Their positions are first calculated by the SE algorithm, and then they are viewed on a pseudo 3-D layout implemented using standard Java graphics context Java.awt.graphics. Interactively, the viewpoint can be modified by the user (rotation of the spheres around their center, zooming in or out, translation of the center). The nodes of subset V_A (respectively V_B) are displayed on the inner sphere θ_A (resp. θ_B) and they are listed in the list panel on the right (resp. left) side of the central view. A node can be selected by clicking it directly in the central view or in the side panel (several nodes can be selected by pressing the Shift or Controls key while clicking nodes). When a node is selected, all the edges connected to it are displayed, and the nodes from these edges second end are also selected. Finally, one can search a node by entering a search phrase matching its name in one of the two search text fields on top of the listing panels displayed on both sides of the central view.

3. Applications

Three different data sets have been used to test visually the performance of the 3D-SE viewer tool. These data sets are all based on relationships between words and other entities (research teams, documents or conference sessions), expressed in a bipartite graph. The data sets differ in size (numbers of nodes in each subset) that is, $|V_A|$ and $|V_B|$ and also in their sparseness ratio (percentage of empty cells in occurrence matrix) defined as $S = 1 - |E|/(|V_A||V_B|)$. Table 1 below summarizes the 3 data sets visualized in this section:

3D-SE viewer	*inner sphere*	*outer sphere*	*origin of links*	*sparseness ratio (%)*
bipartite graph	$\|V_A\|$	$\|V_B\|$	$\|E\|$	$S = 1 - \|E\| / \|V_A\|\|V_B\|$
BSITeam map	teams	keywords	answers to a questionnaire	
	53	175	2823	69.56
Visiome platform keywords	indexing keywords	contents	by authors of contents	
	432	3002	9946	99.23
SfN'06 poster sessions	sessions	terms	occurrences (20/session)	
	650	2164	13000	99.08

Table 1: Basic figures of the 3 data sets used in application of the 3D-SE viewer.

3.1. The BSI-Team Map

The first application of 3D-SE viewer was the visualization of the structure of a research center: the Brain Science Institute (BSI) in RIKEN (Wako, Japan). The purpose of this visualization was to exhibit the relationships and similarities of interests between laboratories and research units in BSI. The resulting interactive exploration tool was named BSI-Team Map [2] and it is accessible on the Internet on RIKEN BSI's main webpage (http://www.brain.riken.go.jp/english/teammap/index.html). This tool allows visitors to see at a glance all the teams of BSI, as well as search facilities and provides direct access to a chosen team. Since its public accessibility in December 2006, some positive comments were received about this tool's capability to show how people interact (personal friend communication). This tool is an effort to make the structure of a research structure more accessible and understandable to the international community, although a lack of international visibility of scientists' web pages in Japan has been recently

[1] **3D-SE viewer** ©BSI-NI and NTT-CS Labs

[2] **BSI-Team Map** ©RIKEN Brain Science Institute

reproached (Ito, 2006). We agree that visibility could be improved by visualizing topics, for instance extracted from the teams' Web pages where research activities are described; a topic oriented structure being more generally accessible than a people oriented one. The proposed representation conveys more information on inter-team similarities than a simple list of names or a planar graph would do: In 3 dimensions, there is one more degree of freedom to position the teams in a way that reflects similarities of interests between entities. In order to feature research interests of the different research units in BSI, a questionnaire has been sent to the 53 research team leaders. Based on their answers, a common list of 175 keywords has been established for the whole institute. This list was then sent back to team leaders who were asked to select the keywords that best correspond to their team's research interests, and to distinguish between keywords of primary and secondary interest. After collecting all the final answers, a table was formed with the keywords on rows, teams on columns and numbers in entries: 1 or 2 whether the keyword was selected as of primary or secondary interest and 0 otherwise. From this "interest" sparse matrix, a binary occurrence matrix was derived (replacing the twos by ones); the corresponding bipartite graph was build and visualized using the 3D-SE viewer. The inner sphere represents the keywords and the outer sphere contains research teams, represented by the team leader's name. Figure 2 illustrates an example of team search: the team leader's name Amari was entered in the search field on top of left panel. The found name is selected in the list of team names, the view was automatically centered on this node and the links to the keywords of research interests for this team are displayed. Then a single click on the team's name will display the Web page of this team. Similarly, when entering a keyword in the search field of the right panel, the view will be centered on the found node(s) and the links from the keyword(s) to all the teams having some research interest it will be shown.

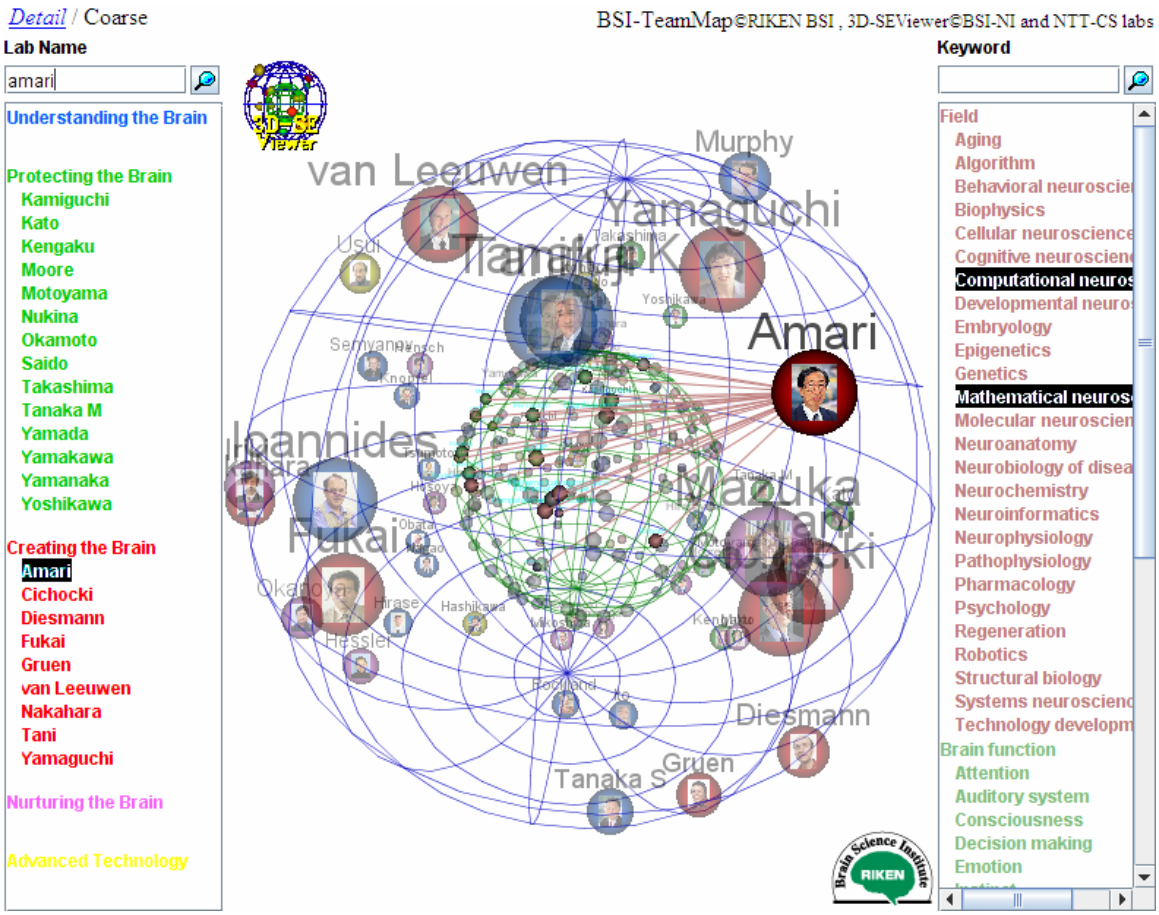

Figure 2: BSI-Team map: 3D-SE viewer-based visualization of RIKEN Brain Science Institute research teams. The nodes are colored according to the 5 team Units listed in the left panel. It can be seen that nodes with the same color appear in the neighboring regions on the outer sphere.

3.2. The Visiome Platform index keywords

Understanding the brain as a system requires worldwide collaboration of scientists specializing in different areas of brain science. This issue confronting many areas of research and much more compounded in the fields of brain research, prompted for the development of a field called Neuroinformatics (NI). Its main goal is to help brain scientists handle the analysis, modeling, simulation, and management of the information resource before, during, and after the conduction of research. The Neuroinformatics Platforms such as Visiome (http://platform.visiome.neuroinf.jp) (Usui, 2003) aim to address these issues by providing portal sites to different fields of brain research, such as in NeuroInformatics Japan Center (NIJC). One vital component of the Neuroinformatics platform is the index tree which is used to organize the electronic materials (digital contents) submitted by the contributors. Automating the keyword index extraction is necessary to support the evolution of the platform in operation and for the establishment of new platforms (Usui, 2007). The 3D-SE viewer was used here to visualize both the indexing keywords and the documents of different types contained in the database, called *contents*. In a first application, a manually established list of indexing terms was used. A selection of keywords for each content was also performed by human experts in the field of vision science. The resulting bipartite graph is made of 3434 vertices split into a set of 3002 contents and a set of 432 index terms, connected by a total of 9946 edges. In the Visiome platform, the contents are filed into 8 main categories: Visual System, Visual Stimulus, Basic Neuroscience, Tools & Techniques, Models & Theory, Applications, Links, Binders, these categories are used to color the nodes on the display. Besides the selection and search possibilities that were described earlier, the 3D-SE viewer allows to access directly other documents by following hyperlinks. When the node of an index term on the outer sphere is clicked by the user, the corresponding page of Visiome database is displayed. This page lists all the registered contents linked to this index keyword, allowing then to access each individual document related to this term. Similarly, if a content node is clicked on the inner sphere, then the Visiome page of this content is shown, providing access to the details of the corresponding document. Figure 3 illustrates a view zoomed into the area of searched term retina. As can be seen, the node of the found index keyword Retina is connected to a large number of documents sketched by the red links, which is an indication that the position of this node on the outer sphere is quite reliable. Experts in this field can evaluate how close from a semantic point of view the neighboring nodes are to this keyword.

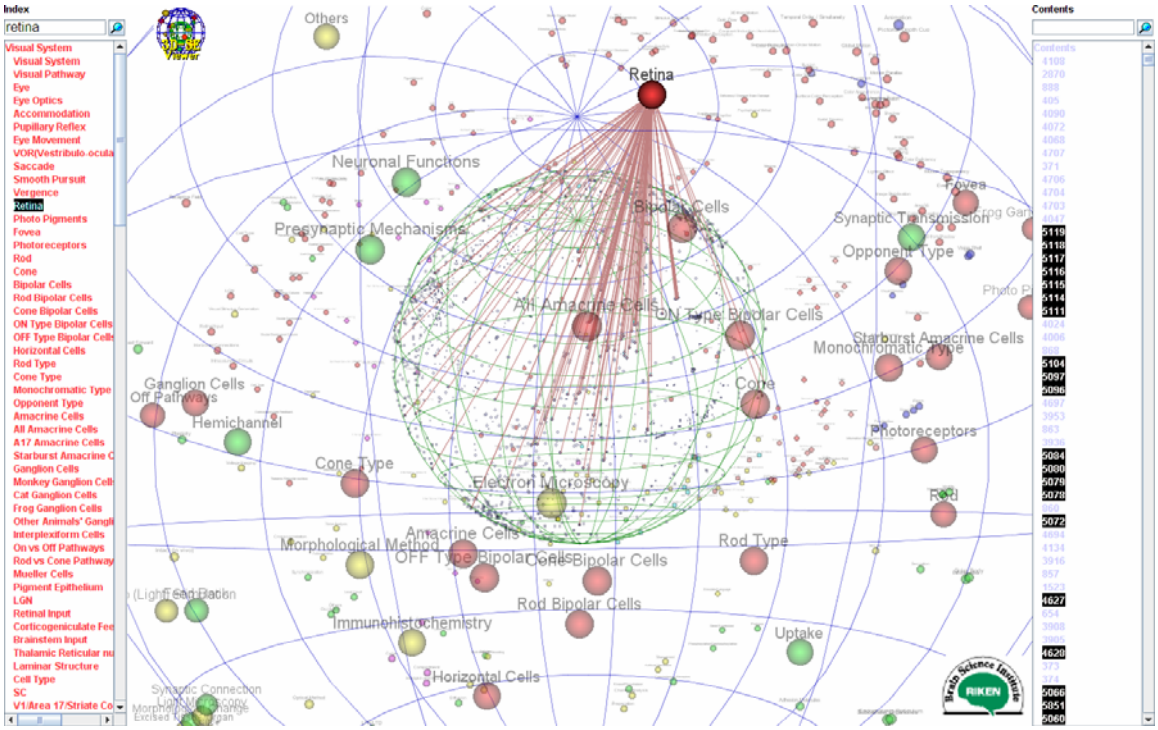

Figure 3: Visualization of Visiome Platform index keywords and contents. The searched term retina found in the index keywords list is selected, showing simultaneously all the links to related contents.

3.3. Society for Neuroscience Annual Meeting abstracts

The Society for Neuroscience (SfN) is, with more than 37,500 members, the world's largest organization of scientists devoted to the study of the brain. Its Annual Meeting is the largest event in neuroscience, which gathered in 2005 nearly 35,000 scientific and nonscientific attendees. Among other scientific events, 650 poster sessions allow researchers to communicate results of their last research. Since the year 2000, abstracts of all posters presented at the Annual Meeting are available online at the SfN Website (http://www.sfn.org), as well as on a CD distributed to participants. The data presented here were extracted from an XML file that is created on disk during installation of the 2006 Neuroscience Meeting Planner software. Several types of sessions such as poster sessions, slide sessions, etc. are defined and each session is assigned to one *theme* among the following 8 main themes in neuroscience: (A) Development, (B) Neural Excitability Synapses and Glia: Cellular Mechanisms, (C) Sensory and Motor Systems, (D) Homeostatic and Neuroendocrine Systems, (E) Cognition and Behavior, (F) Disorders of the Nervous System, (G) Techniques in Neuroscience, (H) History and Teaching of Neuroscience. Each theme is subdivided into *subthemes*, and each subtheme is divided into *topics*. In this preliminary analysis, we focused on poster sessions only, extracting information from posters titles and abstracts. Altogether, there were 12844 posters for which a title and an abstract were available. These posters were presented in 650 poster sessions, and assigned to one among 415 topics, 71 subthemes and 7 themes. The purpose of this application of the 3D-SE viewer is the visualization of the different poster sessions, in order to see how they organize on the basis of the similarities of the posters they listed. Such a display could be usable in future SfN Meetings for attendees to help them plan an itinerary as a path connecting items on the sphere where *themes*, *subthemes* and *topics* would be represented. In this preliminary work, we wanted to visualize only the poster sessions on the basis of their relationships to posters abstracts and titles. In this purpose, for each session, words were extracted from titles and abstracts of the session's posters (from 15 to 30 posters per session) in the same manner as described in section **1.** and ranked according to their *TF.IDF* values. The top 20 words for each session were selected, and gathered into one set of all words for all sessions (many words were common to several sessions, so the final set has 2164 words). Finally, a [sessions $\times$ words] occurrence matrix was build and the ensuing bipartite graph connecting words to sessions was visualized using the 3D-SE viewer applet. Figure 4 represents the 650 poster sessions visualized on the basis of the terms extracted from posters abstracts and titles.

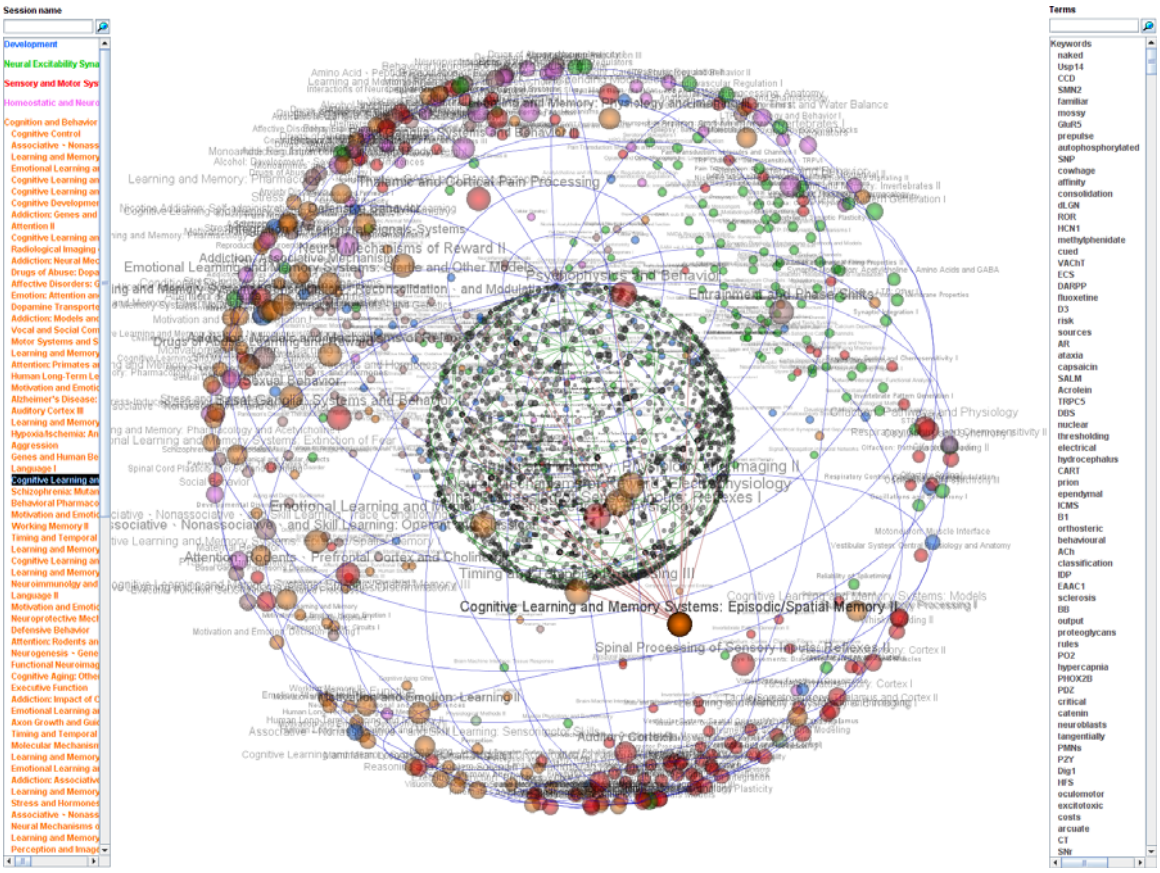

Figure 4: Society for Neuroscience 2006 Annual Meeting: a view of 650 poster sessions (outer sphere) and 2164 extracted terms (inner sphere). The nodes on the outer sphere are colored according to the session's dominant theme. It can be seen that the nodes form some clusters according to their theme.

4. Discussion and perspectives

The 3D-SE viewer is a very attractive tool visualizing bipartite graphs on two spheres. The presented applications of 3D-SE viewer show that it is a useful tool for the visualization of data such as research teams of a large research institution, terms indexing documents in a neuroscience database or poster sessions from the neuroscience knowledge domain. Another advantage of the 3D-SE viewer is its fairly competitive time complexity, which allows obtaining the visualization of several thousands of nodes in a few seconds. An online update of the layouts, needed as documents and keywords will increase in time, should be possible by optimizing further the SE algorithm. From our numerous experiments conducted with various datasets, it has been observed that best visual effects are obtained when the bipartite graph is balanced, that is the numbers of items in each of the 2 subsets is of the same range. We also observed sometimes that the nodes are more uniformly allocated on the two spheres in cases when the graph's edges are themselves more uniformly distributed over the graph's vertices. This means that the degree of each vertex (the number of edges connected to it) should not vary too importantly among the different vertices; otherwise we may observe some empty areas on both spheres. This "hole effect" is probably related to the graph density properties, and it is under analysis. These results of the application of 3D-SE viewer to such data are very preliminary and further research in this area will be conducted in the near future. The results are encouraging and applications to larger datasets can be considered, although in such cases the interactivity can be slowed down by the Java applet technology. The 3D-SE viewer can be used e.g. for larger research institutes, showing for example a higher level of the organizational structure of RIKEN. Applications to the visualization of n-partite graphs for $n > 2$ can also be implemented as layouts of n concentric spheres.

Acknowledgments

The authors would like to thank W. Duch for helpful comments and discussions. The NeuroInformatics Committee of the Society for Neuroscience is gratefully acknowledged for providing and granting the use of the abstracts visualized in section 3.3.

References

Di Giacomo, E., Grilli, L., and Liotta, G. (2006). Drawing Bipartite Graphs on Two Curves, 14–th International Symposium on Graph Drawing, Universität Karlsruhe.

Gibson Kolda, T. (1997). Limited-memory matrix methods with applications, PhD thesis, Dept of Computer Science, University of Maryland.

Hong, S., Nikolov, N. (2005). Layered Drawings of Directed Graphs in Three Dimensions, Proceedings of the Asia-Pacific Symposium on Information Visualization, CRPIT, vol. 45, pp. 69–74.

Ito, M., and Wiesel, T. (2006). Cultural differences reduce Japanese researchers' visibility on the Web, Nature *444*, p. 817.

Misue, K. (2006). Drawing Bipartite Graphs as Anchored Maps, In Proc. Asia Pacific Symposium on Information Visualisation (APVIS2006), Tokyo, Japan. CRPIT, 60. Misue, K., Sugiyama, K. and Tanaka, J., Eds., ACS., pp. 169–177.

Porter, M.F. (1980). An Algorithm for Suffix Stripping, Program *14*, 3, 130–137.

Saito, K., Iwata, T., and Ueda, N. (2004). Visualization of Bipartite Graph by Spherical Embedding, JNNS.

Salton, G., and McGill, M.J. (1983). Introduction to Modern Retrieval (McGraw-Hill Book Company).

Usui, S. (2003). Visiome: Neuroinformatics Research in Vision Project, Neural Networks *16*, 1293–1300.

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

Zheng, L., Song, L., and Eades, P. (2005). Crossing Minimization Problems of Drawing Bipartite Graphs in Two Clusters, CRPIT, vol. 45, pp. 33–38.