Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

3D-SE Viewer: A Text Mining Tool based on Bipartite Graph Visualization

Shiro Usui, Antoine Naud, Naonori Ueda, Tatsuki Taniguchi

***Abstract*—A new interactive visualization tool is proposed for textual data mining based on bipartite graph visualization. Applications to three text datasets are presented to show the capability of this interactive tool to visualize complex relational information between two sets of items by embedding their graph in a 3-dimensional space. Information extracted from texts, such as keywords, indexing terms or topics are visualized to allow interactive browsing of a field of research featured by keywords, topics or research teams. This 3-D visualization tool conveys more information than planar or linear displays of graphs.**

I. 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 [6]. The numbers of occurrences of each term in each document (usually called *frequency*) are counted and denoted f_{ij}. A matrix $\mathbf{F}$ is built, with one row for each term, 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 $\mathbf{F}$, 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 ($M \sim 1000$). Different term weighting schemes can be found in the Information Retrieval literature, which catch different desired properties for the terms (see [2] for a review and comparison of term weighting schemes). The most popular one is probably *TF.IDF*, the *term frequency inverse document frequency* which has been used in this work. The matrix of frequencies $\mathbf{F}$ 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 $\mathbf{F}$ are then converted to binary entries to build a second *occurrence* matrix $\mathbf{O}$ ($o_{ij} = sgn(f_{ij})$), 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.

II. 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 III, 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 to display them.

A. Formal definition of a bipartite graph

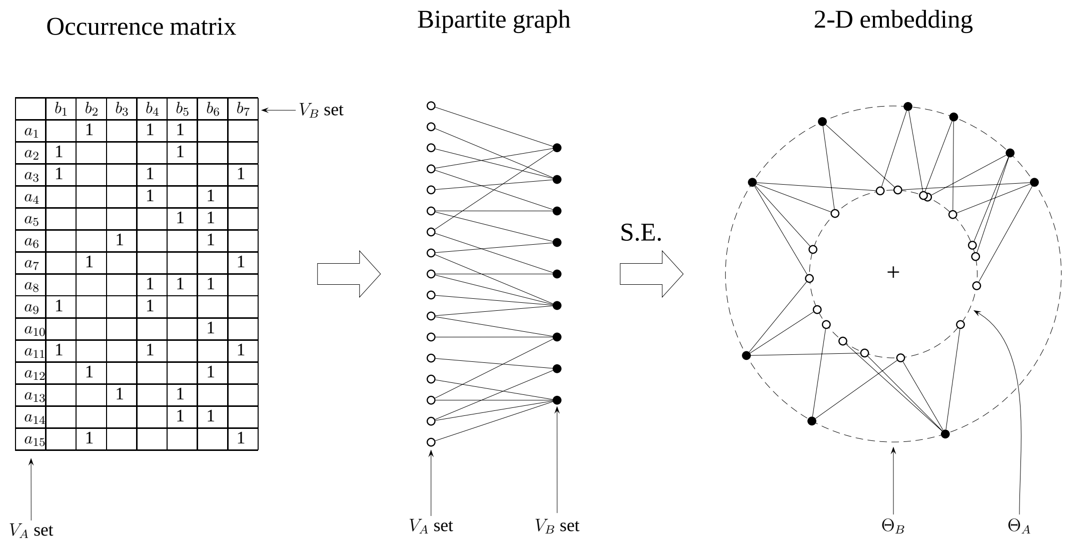
A graph G is defined as $G := \{V, E\}$, where V is the set of vertices or nodes and E the set of edges. Graph G is undirected if the pairs in E are unordered. An undirected graph G is called a bipartite graph if there exist a partition of the vertex set $V = V_A \cup V_B$, so that there is no edge in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

B. The Spherical Embedding algorithm

The Spherical Embedding (SE) algorithm [8] 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.[1] 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. 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

Shiro Usui and Antoine Naud are with the Laboratory for Neuroinformatics, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako City, Saitama 351-0198, JAPAN, Email: usuishiro@riken.jp, naud@brain.riken.jp, Naonori Ueda is with the NTT Communication Science Laboratories, Kyoto, JAPAN, and Tatsuki Taniguchi is with IVIS Corp., JAPAN.

[1] 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.



	b_1	b_2	b_3	b_4	b_5	b_6	b_7
a_1		1		1	1		
a_2	1				1		
a_3	1			1			1
a_4				1		1	
a_5					1	1	
a_6			1			1	
a_7		1					1
a_8				1	1	1	
a_9	1			1			
a_{10}						1	
a_{11}	1			1			1
a_{12}		1				1	
a_{13}			1		1		
a_{14}					1	1	
a_{15}		1					1

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

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

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

where $a_{ij} = +1$ if nodes i and j are connected and -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.

C. 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 [5] 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, 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. A different approach is proposed by Zheng et al. [12], 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. [1], is another interesting approach. Hong et al. proposed [3] 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, so that nodes are displayed together with lines representing the edges. In our approach, the focus is set primarily on the visualization of the graph's vertices; whereas its edges can be displayed on demand by selecting interactively the corresponding nodes.

D. The **3D-SE viewer** *visualization tool*

The 3D-SE viewer [2] 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, 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.

III. 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

[2] **3D-SE viewer** ©BSI-NI and NTT-CS Labs

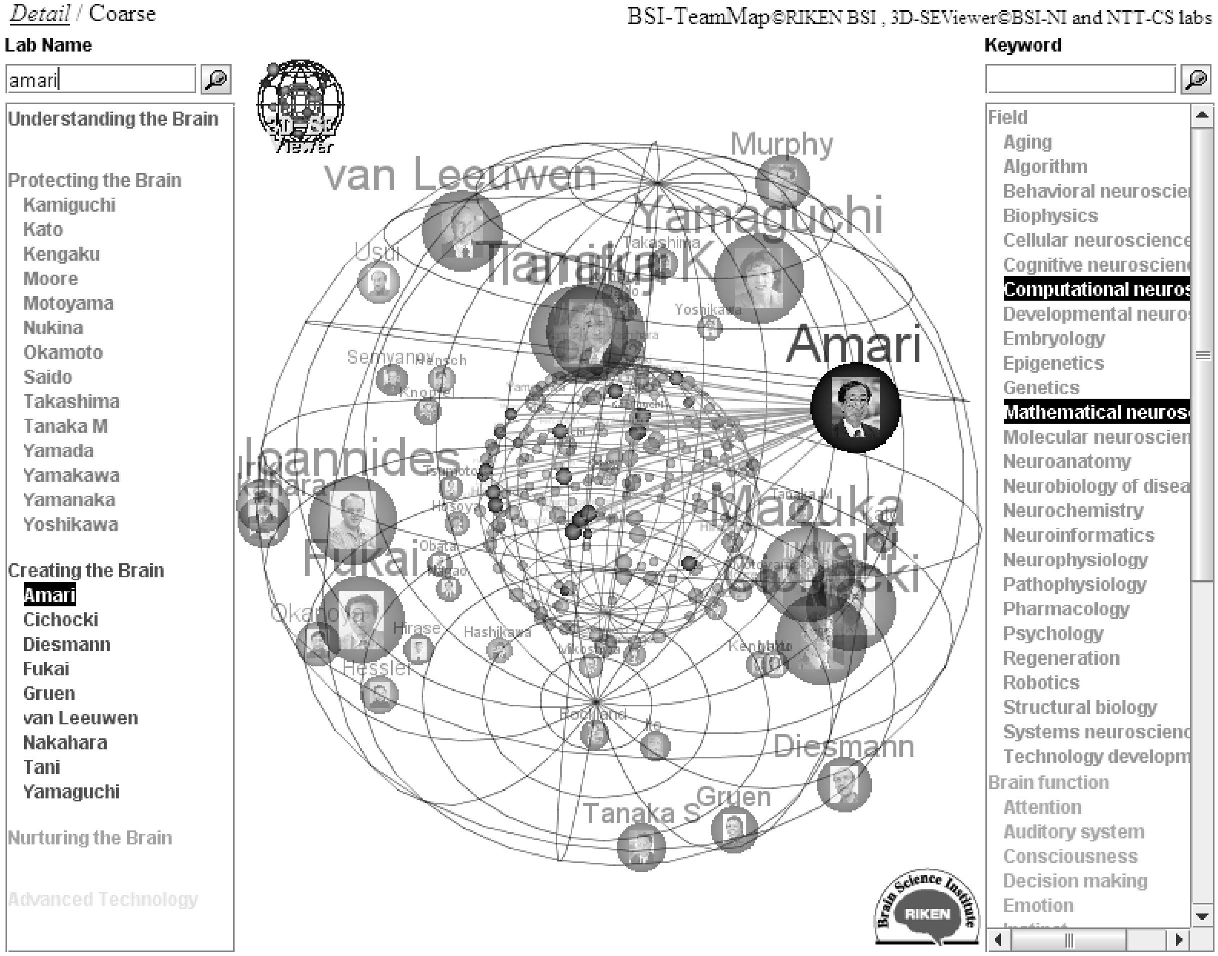


Fig. 2. BSI-Team map: 3D-SE viewer-based visualization of RIKEN Brain Science Institute research teams. The nodes are colored according to the teams Units listed in the left panel. It can be seen that nodes with the same color appear in the same region on the outer sphere.

also in their sparseness ratio (percentage of empty cells in occurrence matrix) defined as $S = |E|/(|V_A||V_B|)$.

A. *The* BSI-Team Map

The first application of 3D-SE viewer was the visualization of the structure of a research organization: the RIKEN Brain Science Institute (BSI), by showing relationships between its laboratories and research units. The resulting tool is accessible on the Internet, it allows visitors to see at a glance all the teams of BSI, as well as search facilities and direct access to a chosen team. Such a representation convey much more information on inter-team similarities than a simple list of names or a planar graph does: In 3-dimensions, there is larger degree of freedom to position the teams in a way that reflects similarities of interests. In order to feature research interests of the different units of the Institute, a questionnaire has been sent to the 53 research team leaders. Based on their answers, a list of 123 keywords has been established for the whole Institute. This list was then sent back to team leaders who were asked select keywords that best correspond to their teams research interests, and distinguish keywords of primary and secondary interest. Gathering finally the answers, a table was formed with the keywords on rows, teams on columns and numbers in entries (1 or 2 when the keyword was selected as primary or secondary keywords, and 0 otherwise). From this ”frequency” matrix, an binary occurrence matrix was derived (sparseness: 85.12%) and the bipartite graph was build and visualized using the 3D-SE viewer. The inner sphere represents the keywords and the outer sphere contains teams. The resulting interactive exploration tool was named BSI-Team Map[3] and it is accessible from RIKEN BSI's main webpage (http://www.brain.riken.go.jp/english/teammap/index.html). Since its public accessibility in December 2006, some positive comments was 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' webpages in Japan has been recently reproached [4]. We agree that visibility could be improved by visualizing topics, for instance extracted from the teams Web pages describing research activities, topic oriented structure being more accessible than people oriented. Figure 2 illustrates an example of team search: the

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

team leader's name amari was entered in the search field on top of left panel. The found name is enlightened in the team list, the view centered on this point and the links to the keywords of this team were drawn. Then a single click on the team's name shows directly this team's webpage. Similarly, when entering a keyword in the search field in the top of the right panel, the links from this keyword to all the teams related to it will be shown.

B. The Visiome Platform

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) [9] [11] 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. A tool for automatic extraction of the keywords from papers titles and abstracts was developed [10]. 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. Selection of keywords for each content was also performed by human experts in the field of vision science. The bipartite graph contains a set of 3002 contents linked to a set of 432 index terms. The sparseness of occurrence matrix is 99.23%. The density of the graph can be estimated by the number of edges per index terms, which ranges from 0 (5 index terms have no content linked to them) to 295, with a mean value of 22.21 and a total of 9946 edges. Similarly, the number of edges per content ranges from 1 to 31 index terms, with an average value of 3.31. In the Visiome platform, the 3002 contents are filed into 8 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 documents. When clicking the name just above a node in the 3-D display, the corresponding Visiome web page is opened: If an index term node name is clicked on the outer sphere, the Visiome page of this index term is shown allowing to view and access all individual documents 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. The found node Retina is connected to a large number of keywords as indicated by the links, which shows that the position of this node on the outer sphere is reliable. Experts in this field can evaluate how close from a semantic point of view the neighboring nodes are to this keyword.

C. 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 were extracted from the XML file that is created during installation of the 2006 Neuroscience Meeting Planner software. The sessions are organized into 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*. A summary of some basic figures concerning the 2006 SfN Annual Meeting is presented in Table I. Although we were interested in extracting information from posters titles and abstracts, the figures in Table I concern both posters and slides for which a title and an abstract were available. The purpose of this last application of 3D-SE viewer is the visualization of the different sessions, in order to see how they organize on the basis of their similarities. 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 I and ranked according to their *TFIDF* values. The top 20 words for each session were selected, and gathered into one set of all words for

TABLE I
SOCIETY FOR NEUROSCIENCE 2006 ANNUAL MEETING SUMMARY DATA

1	Number of themes	8
2	Number of subthemes	71
3	Number of topics	415
4	Number of poster sessions	650
5	Number of posters	12844
6	Number of slides sessions	111
7	Number of slides	1236
8	Total (Posters + Slides)	14080

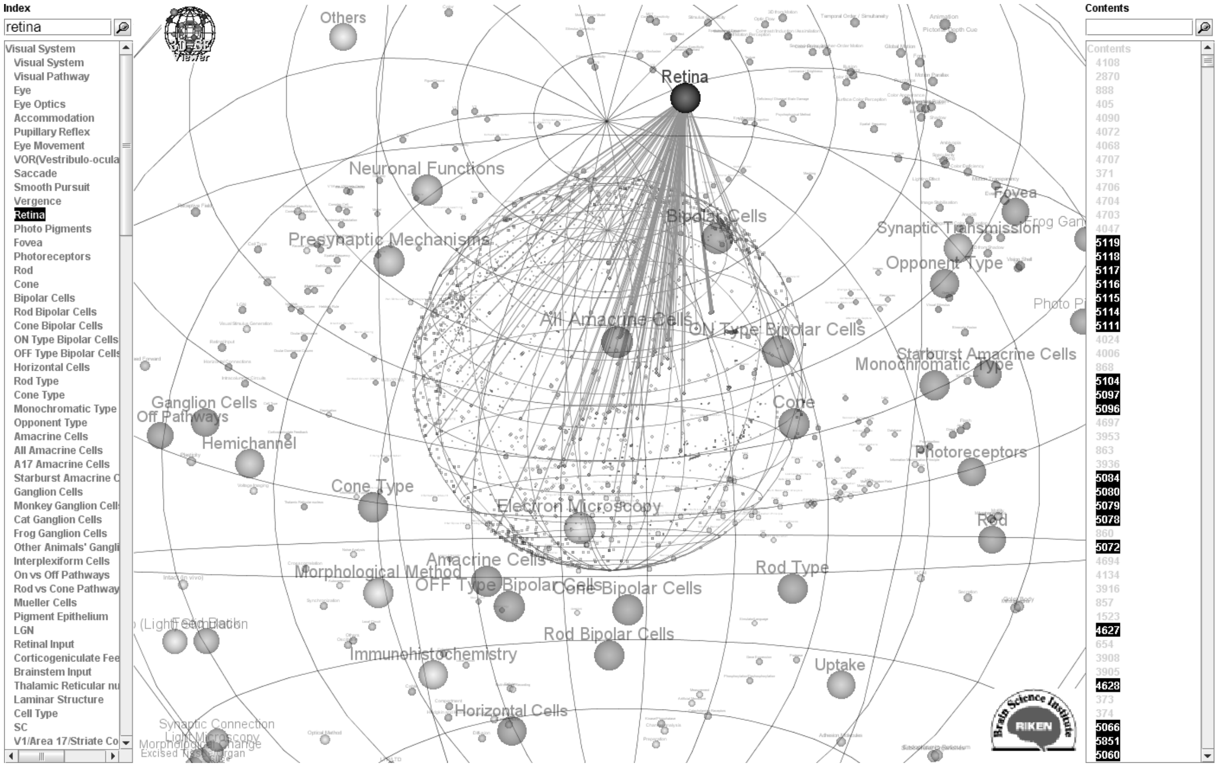


Fig. 3. Visiome Platform index keywords: 3D-SE viewer-based visualization with focus on selected Retina keyword.

all sessions (many words were common to several sessions, so the final set has 2164 words). Finally, a sessions × words occurrence matrix was build (sparseness: 97.50%) 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.

IV. Main Results

The 3D-SE viewer shows an interesting capability of visualizing bipartite graphs on two spheres. From 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 observed also several times 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 number of edges connected to each vertex should not vary too much 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 should be analyzed more completely. Another advantage of the 3D-SE viewer is its fairly competitive time complexity, which allows to obtain the visualization of several thousands of nodes in a few seconds. 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.

V. Conclusions

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 a particular knowledge domain. The results are encouraging and applications to larger datasets can be considered. The 3D-SE viewer can be used e.g. for other research institutes, showing for example a higher level of RIKEN's organizational structure. Applications to the visualization of n-partite graphs for $n > 2$ can also be considered.

Acknowledgment

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 III-C.

References

[1] E. Di Giacomo, L. Grilli and G. Liotta, "Drawing Bipartite Graphs on Two Curves," 14th International Symposium on Graph Drawing, Universität Karlsruhe, Sep. 18-20, 2006.

[2] T. Gibson Kolda, "Limited-memory matrix methods with applications," PhD thesis, Dept of Computer Science, University of Maryland, 1997.

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

[4] M. Ito and T. Wiesel, "Cultural differences reduce Japanese researchers' visibility on the Web," Nature, 444, p. 817, Dec. 14, 2006.

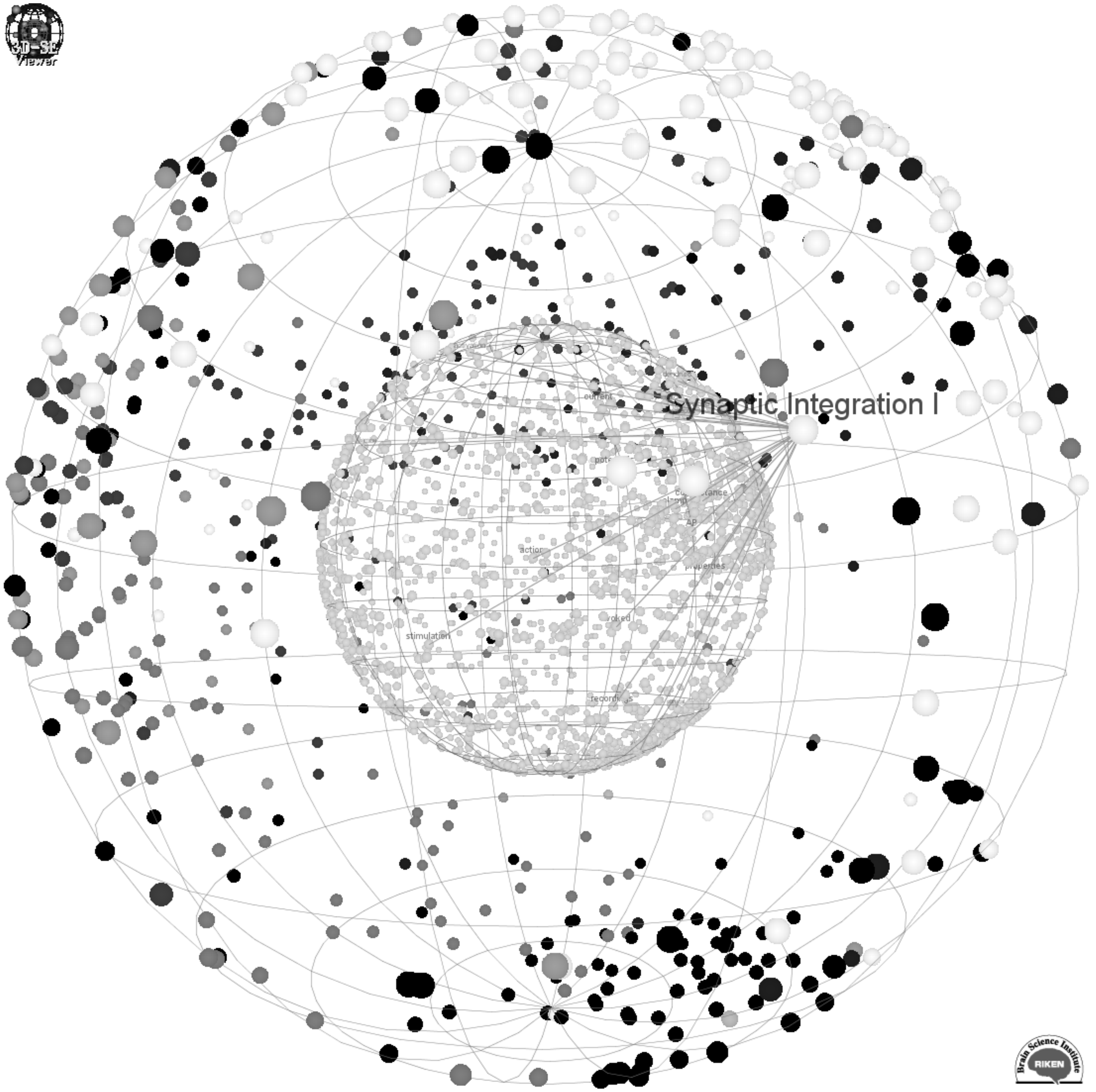


Fig. 4. Society for Neuroscience 2006 Annual Meeting: a view of 650 poster sessions (outer sphere) and 2164 extracted terms (inner sphere). The nodes on the outer sphere are colored according to the session's dominant theme. It can be seen that the nodes cluster in accordance with their theme.

[5] K. Misue, "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, 2006.

[6] M.F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, 3, pp. 130–137, July 1980.

[7] G. Salton and M.J. McGill, Introduction to Modern Retrieval, McGraw-Hill Book Company, 1983.

[8] K. Saito(P), T. Iwata and N. Ueda, "Visualization of Bipartite Graph by Spherical Embedding," JNNS 2004 (in Japanese).

[9] S. Usui, "Visiome: Neuroinformatics Research in Vision Project," Neural Networks, vol. 16, pp. 1293–1300, 2003.

[10] S. Usui, P. Palmes, K. Nagata, T. Taniguchi, N. Ueda, "Keyword Extraction, Ranking, and Organization for the Neuroinformatics Platform," Bio Systems, vol. 88, pp. 334–342, 2007.

[11] S. Usui, "A Neuroinformatics Platform for Vision Science: Visiome," IJCNN 2007 (companion paper at this conference).

[12] L. Zheng, L. Song, P. Eades, "Crossing Minimization Problems of Drawing Bipartite Graphs in Two Clusters," Proc. of the Asia-Pacific Symposium on Information Visualization, CRPIT, vol. 45, pp. 33–38, 2005.