An Ontology and Brain Model-based Semantic Discovery and Visualization System

Xia Lin, Mi Zhang, Yue Shang, Yuan An
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
{xlin, mz349, ys439, ya45}@drexel.edu

Abstract— **A system was built to let the user click on a location of a brain model to explore neuroscience ontology terms related to the location. The user can explore further the related terms and related documents through a visualization interface and learn new concept or document relationships derived from the ontology and annotated document collections. This paper discusses the semantic technologies used to build the system and introduces various features of the visualization interfaces. It concludes that semantic technologies can be integrated with visual brain models and ontologies to support visual semantic exploration and discovery.**

Keywords-Information visualization; Brain models for information retrieval; Semantic browsing; Neuroscience ontology; Visual interface design

I. INTRODUCTION

Ontology, modeling, and visualization are three knowledge representation techniques for neuroscience research. Their methodologies are related and complement to each other, but the connections among them are not so obvious. Ontology, as a formal language, seeks to explicitly define concepts and concept relationships for the purposes of concept retrieval, semantic concept lookup, concept linking, and semantic inferences [1][2]. Through the standardized classes, instances, and relationships, ontology helps facilitate data interoperability and provides linkages between research data and literature. The Neuroscience Information Framework Project [3] represents a good example of how comprehensive ontologies can unite a domain's literature, data, and research projects. Modeling, while mostly theory or data driven, seeks to represent complex natural systems (such as the neural system or brain) through mathematical or graphical models in order to reveal the most relevant relationships of the underlying data [4]. It can also help to define concept and concept relationships. Various brain map projects such as Talairach Atlas [5] and Allen Brain Atlas [6] are good examples of using models' layers, locations, and views to unify concepts, data, functions, and other relevant information [7]-[9]. Both ontology and modeling will be more effective if modern visualization techniques can be applied to them. Like ontology and modeling, visualization makes implicit relationships explicit. It takes advantages of human visual capability to allow users to explore and understand large amount of data and make visual inferences among the data [10]. It facilitates interaction with data and allows researchers to select different views or to zoom in to specific locations to explore data or concepts and their relationships.

To experiment how to bring ontology, modeling, and visualization together for interactive concept exploration and semantic discovery, a collaborative project funded by the U.S. National Science Foundation and several major industrial partners was carried out in the Center on Visual and Decisions Informatics (CVDI) in our university. In the following, we present a semantic discovery and visualization system we are implementing and discuss how a brain-model and ontologies have enhanced functionality of the system.

The rest of the paper is organized as follows. Section II provides an overview of the system with detailed descriptions of each system component. Section III discusses semantic technologies used to build the system. Section IV shows and discusses several visual interfaces of the system, and finally, Section V provides a summary of the project.

II. SYSTEM OVERVIEW

The main goal of the project is to create an innovative system for information retrieval and semantic discovery on neuroscience literature. With permission and support from the publisher Elsevier, we downloaded about 1 million documents related to neuroscience from hundreds of Elsevier's journals and books for this experimental system. Each of the documents includes full text of the documents and the metadata created by the publisher, both in XML format.

Through initial requirements analysis and discussion with neuroscientists, we recognized that ontologies, brain models, and visualizations should be the key considerations for the systems. The Neuroscience Information Framework (NIF) Standard Ontology was chosen for the project as it is considered both an ontology and an extended framework for concept-based indexing and retrieval [11]. The ontology composes of a collection of Web Ontology Language (OWL) modules covering distinct domains of bio-medical areas such as anatomy, molecule, disease, and organism, etc. The ontology can be downloaded as OWL files and can be imported into Protégé, the popular ontology creation and editing tool [12].

While the ontology is well constructed and easy to download, the challenge we faced is how to annotate the very large neuroscience document collection (about 1 million documents) with this very large ontology (more than 108k classes). This is a scale-up issue of annotation. Thus the first component of the system we developed is an annotation tool called SemIntegrator, which can be used either through the APIs or through the Protégé interface.

We also learn that domain experts often need to work with both a comprehensive ontology and a specialized ontology most relevant to their own specialties. Thus, we include several specialized ontologies in the system also. Linking concepts in multiple ontologies, however, is another challenge we faced. The second component of the system is an ontology linking tool and a faceted-based interface that integrates two or more ontologies for searching, browsing, and exploration. A unique feature of the interface is to let the user search in one ontology and browse in another.

The third component of the system is an ontology-based visualization and exploration interface where one can click on a specific location of the brain map to find relevant terms from two or more ontologies and then click on the terms for searching and browsing. Details of the interface will be described in the later sections.

III. System Technology

The core technology of our system includes semantic annotation, semantic integration, and semantic visualization, as shown in Figure 1.

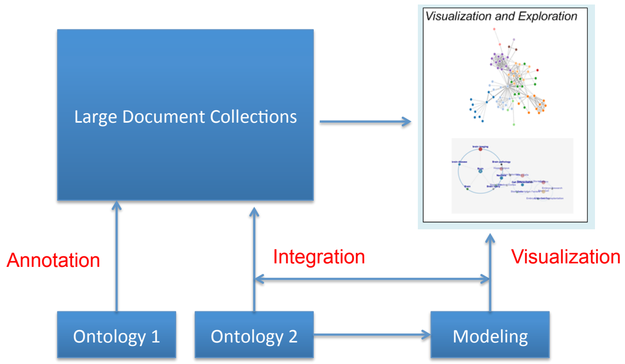


Figure 1. Overview of the semantic technologies used for the system.

A. Semantic Annotation

Our focus on semantic annotation is to apply information extraction techniques to identify in the documents all occurrences of ontology concepts and enrich the metadata of the documents with the identified concepts. An annotation component of our system was developed for this purpose. The component was first developed as a plugin of Protégé and then as a stand-alone Web service made available through Application Programming Interfaces (API). Protégé is an open source toolkit that can be used to build, alter and search ontologies [13]. It is extensible and offers API for researchers to work with ontologies in OWL format.

The annotation process we implemented involves several open-source packages (Fig. 2). First, Protégé is used to parse the OWL-formatted ontology files. Lingpipe [14] is then used to implement the term matching. The matching terms are saved into Trie, also called “prefix tree”, which is a data structure used to improve search efficiency [15]. Chunk is an interface in Lingpipe that specifies a slice of a character sequence and used to match the article with terms in Trie. Levenshtein distance [16] is used here to calculate the similarity of two strings. The Levenshtein distance between two strings is the minimum number of single-character operation (three kinds of operation: insertion, deletion and substitution) that can change on string to the other.

Using SemIntegrator, we were able to process the whole Elsevier neuroscience document collection and, on average, 72 concepts are annotated for each document. In addition, a set of Java/Java-script modules was created to bridge those open-source packages and make them work together for the multiple ontologies and the document collections.

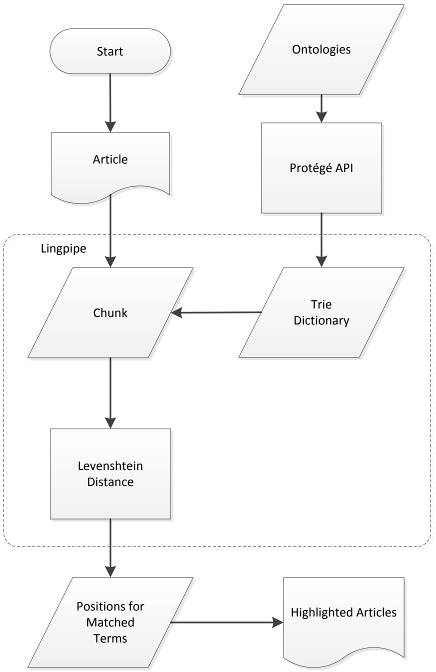
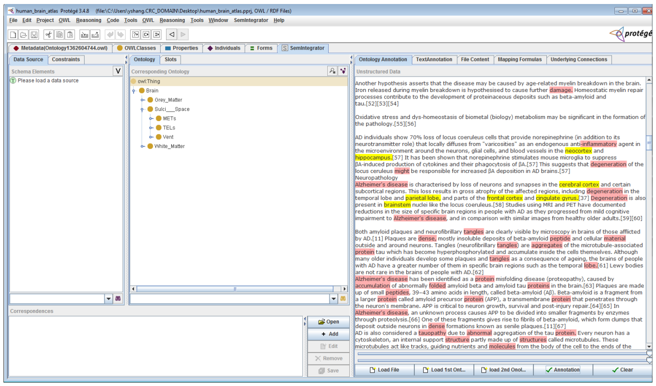


Figure 2. The overall process of semantic annotation.

B. Semantic Integration

Linking concepts in one ontology to concepts in another ontology is semantic integration [17]. In this project, we use SemIntegrator created in this project to annotate multiple ontologies to the same collection, and then use the collection as the bridge to link two or more ontologies. Here, we use an example to illustrate how it works.

Say a researcher is interested in exploring associations between human brain dysfunctions and brain structures. There are well-developed ontologies on each side, the Allen Brain Atlas Ontology [6] for brain structures and the NIF-Dysfunction ontology for brain dysfunctions. But, there are no direct associations between concepts in the two ontologies even though they are clearly related and the concepts often appear in the same documents. Using SemIntegrator, we can annotate documents with both ontologies and highlight the annotated concepts from each ontology with a different color (Fig. 3). This allows the reader to make associations among the concepts from different ontologies. From the highlighted result, the reader can quickly scan the article by reading terms highlighted in one color, say the term "Alzheimer's Disease," and find the correlated brain regions highlighted in a different color, such as the terms "degeneration in parietal lobe," "frontal cortex" and "cingulate gyrus" that show potential associations with "Alzheimer's Disease." The reader can further explore the unfamiliar terms such as "cingulate gyrus" through our visualization interface to find the correlation of this concept with other entities.


human_brain_atlas Protégé 3.4.8 (file:\C:\Users\yshang.CRC_DOMAIN\Desktop\human_brain_atlas.pprj, OWL / RDF Files)
File Edit Project OWL Reasoning Code Tools OWL Reasoning Tools Window SemIntegrator Help
protégé
Metadata(Ontology1362604744.owl) OWLClasses Properties Individuals Forms SemIntegrator
Data Source Constraints
Schema Elements
Ontology Slots
Corresponding Ontology
Ontology Annotation TextAnnotation File Content Mapping Formulas Underlying Connections
Unstructured Data
Correspondences
Open
Add
Edit
Remove
Save
Load File
Load 1st Ont...
load 2nd Onol...
Annotation
Clear

Please load a data source

owl:Thing
- Brain
  - Grey_Matter
  - Sulci___Space
    - METs
    - TELs
    - Vent
  - White_Matter

Another hypothesis asserts that the disease may be caused by age-related myelin breakdown in the brain. Iron released during myelin breakdown is hypothesised to cause further damage. Homeostatic myelin repair processes contribute to the development of proteinaceous deposits such as beta-amyloid and tau.[52][53][54]

Oxidative stress and dys-homeostasis of biometal (biology) metabolism may be significant in the formation of the pathology.[55][56]

AD individuals show 70% loss of locus coeruleus cells that provide norepinephrine (in addition to its neurotransmitter role) that locally diffuses from "varicosities" as an endogenous anti-inflammatory agent in the microenvironment around the neurons, glial cells, and blood vessels in the neocortex and hippocampus.[57] It has been shown that norepinephrine stimulates mouse microglia to suppress βA-induced production of cytokines and their phagocytosis of βA.[57] This suggests that degeneration of the locus ceruleus might be responsible for increased βA deposition in AD brains.[57]
Neuropathology
Alzheimer's disease is characterised by loss of neurons and synapses in the cerebral cortex and certain subcortical regions. This loss results in gross atrophy of the affected regions, including degeneration in the temporal lobe and parietal lobe, and parts of the frontal cortex and cingulate gyrus.[37] Degeneration is also present in brainstem nuclei like the locus coeruleus.[58] Studies using MRI and PET have documented reductions in the size of specific brain regions in people with AD as they progressed from mild cognitive impairment to Alzheimer's disease, and in comparison with similar images from healthy older adults.[59][60]

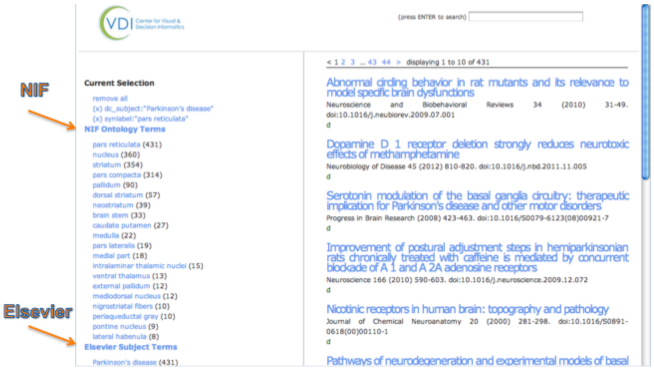
Both amyloid plaques and neurofibrillary tangles are clearly visible by microscopy in brains of those afflicted by AD.[11] Plaques are dense, mostly insoluble deposits of beta-amyloid peptide and cellular material outside and around neurons. Tangles (neurofibrillary tangles) are aggregates of the microtubule-associated protein tau which has become hyperphosphorylated and accumulate inside the cells themselves. Although many older individuals develop some plaques and tangles as a consequence of ageing, the brains of people with AD have a greater number of them in specific brain regions such as the temporal lobe.[61] Lewy bodies are not rare in the brains of people with AD.[62]
Alzheimer's disease has been identified as a protein misfolding disease (proteopathy), caused by accumulation of abnormally folded amyloid beta and amyloid tau proteins in the brain.[63] Plaques are made up of small peptides, 39–43 amino acids in length, called beta-amyloid (Aβ). Beta-amyloid is a fragment from a larger protein called amyloid precursor protein (APP), a transmembrane protein that penetrates through the neuron's membrane. APP is critical to neuron growth, survival and post-injury repair.[64][65] In Alzheimer's disease, an unknown process causes APP to be divided into smaller fragments by enzymes through proteolysis.[66] One of these fragments gives rise to fibrils of beta-amyloid, which form clumps that deposit outside neurons in dense formations known as senile plaques.[11][67]
AD is also considered a tauopathy due to abnormal aggregation of the tau protein. Every neuron has a cytoskeleton, an internal support structure partly made up of structures called microtubules. These microtubules act like tracks, guiding nutrients and molecules from the body of the cell to the ends of the

Figure 3. The SemIntegrator interface with annotated concepts from two ontologies highlighted in different colors.

Another way to realize semantic integration is through search engines. Our system uses the Apache Solr indexing service [18] to index the document collection after the metadata have been enriched with multiple ontologies. Since Solr is a facet-based indexing, each ontology could be treated as a facet and the user has the option of searching by a particular ontology and displaying terms of other ontologies. This creates a very useful function of linking related concepts from multiple ontologies and using them for searching and browsing. Fig. 4 shows an example of the Solr-based interface where both NIF ontology terms and the original Elsevier indexing terms related to the query are shown. The user can click on either type of term to narrow down search results or search in one type of the terms and browse through documents that have been annotated by another type of the terms.

C. Semantic Visualization

Visualization may be applied to neuroscience research in many different ways [10][19]. For this system, our focus is on semantic concept visualization, or visualizing knowledge structure of ontologies and document collections through meaningful concept displays [20]. One of the main advantages of ontologies is the rich concept relationships existed within the ontologies and the annotated document collections. Some of those relationships are explicitly defined. Some can be derived from their semantic relationships or co-occurrence relationships. Some can only be discovered through computational learning algorithms. These relationships essentially form a knowledge structure that can assist users in navigating and exploring both the conceptual space and the document space of the domain, particularly if the knowledge structure can be visualized in an interactive visual interface. In our system, we have implemented such an interactive interface with learning and visualization algorithms such as PFNet, D3.js, Gephi, and Sigma.js [21]-[24]. Details of the interface are described in the next section.


NIF
Elsevier
VDI Center for Visual & Decision Informatics
(press ENTER to search)
< 1 2 3 ... 43 44 > displaying 1 to 10 of 431

Current Selection
remove all
(x) dc_subject:"Parkinson's disease"
(x) synlabel:"pars reticulata"

NIF Ontology Terms

pars reticulata	(431)
nucleus	(360)
striatum	(354)
pars compacta	(314)
pallidum	(90)
dorsal striatum	(57)
neostriatum	(39)
brain stem	(33)
caudate putamen	(27)
medulla	(22)
pars lateralis	(19)
medial part	(18)
intralaminar thalamic nuclei	(15)
ventral thalamus	(13)
external pallidum	(12)
mediodorsal nucleus	(12)
nigrostriatal fibers	(10)
periaqueductal gray	(10)
pontine nucleus	(9)
lateral habenula	(8)

Elsevier Subject Terms

Parkinson's disease	(431)

Abnormal circling behavior in rat mutants and its relevance to model specific brain dysfunctions
Neuroscience and Biobehavioral Reviews 34 (2010) 31-49. doi:10.1016/j.neubiorev.2009.07.001
d

Dopamine D 1 receptor deletion strongly reduces neurotoxic effects of methamphetamine
Neurobiology of Disease 45 (2012) 810-820. doi:10.1016/j.nbd.2011.11.005
d

Serotonin modulation of the basal ganglia circuitry: therapeutic implication for Parkinson's disease and other motor disorders
Progress in Brain Research (2008) 423-463. doi:10.1016/S0079-6123(08)00921-7
d

Improvement of postural adjustment steps in hemiparkinsonian rats chronically treated with caffeine is mediated by concurrent blockade of A 1 and A 2A adenosine receptors
Neuroscience 166 (2010) 590-603. doi:10.1016/j.neuroscience.2009.12.072
d

Nicotinic receptors in human brain: topography and pathology
Journal of Chemical Neuroanatomy 20 (2000) 281-298. doi:10.1016/S0891-0618(00)00110-1
d

Pathways of neurodegeneration and experimental models of basal

Figure 4. The search interface of the system that allows searching in one ontology and browsing in another.

IV. THE VISUALIZATION INTERFACE

The visualization interface is implemented to assist users in exploring semantic relationships among the concepts and let users follow the relationships for document browsing, exploration and discovery. Figure 5 (at the end of the paper) shows an example of the visualization interface.

The interface is divided into three main parts. The top-left is an interactive brain map, which consists of six areas of the human brain, including frontal lobe, parietal lobe, temporal lobe, occipital lobe, cerebellum and brain stem. Each of the brain area is filled with a different color. The user can zoom in or out the map by clicking on a specific brain area of the map. When a specific area of the map is clicked, a list of concept, function and dysfunction terms associated with that area will be shown in the bottom-left of the interface. This helps the user to quickly find concepts and functions related to a certain brain location.

When the user scans through the list of concepts, functions or dysfunctions, they can choose to explore any of them by clicking on a term to bring up either a hierarchical or associative concept display on the right-hand side of the interface. The hierarchical display visualizes the hierarchical structure of the ontology. The user can follow the visual display to see a concept's parent, children, or sibling terms.

Each of the terms on the visual display is also clickable – the user can click on any of them to expand the structure or show a new hierarchical structure.

The associative concept display shows concept relationships not based on the ontology itself but on the annotated document collections. On the display, the size of a node is decided by the frequency of the concept occurred in all the articles of the collection, and the color of the node indicates which brain part the concept is related to. The links and distance of nodes are based on the co-occurrence of the annotated terms in the whole corpus. Through Solr indexing, extensive computation was done to calculate the co-occurrence of any two concepts of the ontology in the collections of a million documents. When the user clicks on a term, the system will first select the top twenty concepts that have the highest co-occurrence frequencies with the term, and then generate a co-occurrence matrix of 20 by 20 for these 20 terms. The Pathfinder algorithm [21] is then applied to the matrix to generate a meaningful display of semantic relationships of the concepts.

Through the hierarchical and associative displays, the user will be able to explore concept relationships both in the ontology and the collections, and use them complementarily for their semantic exploration and discovery. Moreover, during the user's interaction with the visual interface, a search query is automatically generated and updated, and the number of search results is shown (on the bottom-left corner). The user can click on the results to retrieve relevant documents any time during the interactive exploration.

When the user moves from the concept space to the document space, he or she can also browse the document cluster map where each node is a document and each link indicates sharing of concept terms (Fig. 6). The cluster map was first generated using Gephi [23] and then used sigmajs [24] to provide interactive functions. By interacting with both the concept maps and document maps, the user can explore semantic relationships of terms and documents at both the global level (with all the documents) and at the detailed and focused levels.

V. Summary

In this article, we presented a semantic discovery and visualization system that has two distinct features. One is the capability of annotating full text documents with concepts from multiple ontologies. The other is the set of visual interactive functions that link ontology concepts to a brain model for browsing and exploration. The system has showed promising results. The next step for us is to test and evaluate the system while continuously improving the implementation of the system. Through this paper, we hope to bring the attention of the research community to the central idea of the system: using ontologies, modeling, and visualization together to support semantic exploration and discovery.

Acknowledgment

This research was conducted at the Center for Visual and Decision Informatics (CVDI). The support of the Center's industrial members was gratefully acknowledged. Thanks in particular are due to the industrial member Elsevier who provides the data and feedback for this project. This research is also partially funded by the NSF IIP 1160960.

References

[1] S. D. Larson and M. E. Martone, "Ontologies for neuroscience: What are they and what are they good for?" Frontiers in Neuroscience, 3(1), 60-67, 2009.

[2] Y. Le Franc, et al. "Computational neuroscience ontology: A new tool to provide semantic meaning to your models," BMC Neuroscience, 13(suppl 1), 2012, p. 149.

[3] NIF. Neuroscience Information Framework. http://neuinfo.org [retrieved: April 5, 2014].

[4] D. Sterratt, B. Graham, A. Gillies, and D. Willshaw, Principles of Computational Modelings in Neuroscience, Cambridge University Press, 2011.

[5] BrainMap. http://brainmap.org. [retrieved: April 5, 2014].

[6] Allen Brain Atlas. http://www.brain-map.org/ [retrieved: April 5, 2014].

[7] A. P. Alivisatos, M. Chun, G. M. Church, R. J. Greenspan, M. L. Roukes, and R. Yuste, "The Brain Activity Map Project and the Challenge of Functional Connectomics," Neuron, vol. 74 (6), June 2012 , pp. 970-974.

[8] A. R. Laird, J. L. Lancaster, and P. T. Fox, "BrainMap," Neuroinformatics, 3(1), March, 2005, pp.65-77.

[9] A. R. Laird, et al., "The BrainMap strategy for standardization, sharing, and meta-analysis of neuroimaging data," BMC Research Notes, 2011, 4:349.

[10] Emilio Badoer, (ed.) Visualization techniques: From Immunohistochemistry to Magnetic Resonance Imaging (Neuromethods). Humana Press, 2012.

[11] D. Gardner, H. Akil, G.A. Ascoli, and D. M. Bowden, "The Neuroscience Information Framework: A Data and Knowledge Environment for Neuroscience," *Neuroinform*, 2008, pp. 149–160.

[12] D. Rubin, N. Noy, and M. Musen, "Protégé: A Tool for managing and using terminology in radiology," *Journal of Digital Imaging*, 2007, pp. 34-46.

[13] Protégé, "Protégé: A open source ontology editor and knowledge-base framework," Available at: http://protégé.stanford.edu [retrieved: April 5, 2014].

[14] Alias-I, "LingPipe 4.1.0." http://alias-i.com/lingpipe [retrieved: April 5, 2014].

[15] J. Bentley and R. Sedgewick, "Ternary search trees," Dr. Dobb's Journal, April, 1998.

[16] D. Hirschberg, "Serial computations of Levenshtein distances." In A. Apostolico and Z. Galil (eds.), *Pattern matching algorithms*, pp. 123 - 141. Oxford University Press, 1997.

[17] N. F. Noy, "Semantic integration: a survey of ontology-based approaches," *SIGMOD Rec.* 33(4), 2004, pp. 65-70.

[18] Apache Solr, "Apache Solr: open source enterprise search platform," Available at: http://lucene.apache.org/solr/ [retrieved: April 5, 2014].

[19] P. Mutton and J. Golbeck, "Visualization of semantic metadata and ontologies," Proceedings of the *Seventh International Conference on Information Visualization," 2007,* pp.300 – 305.

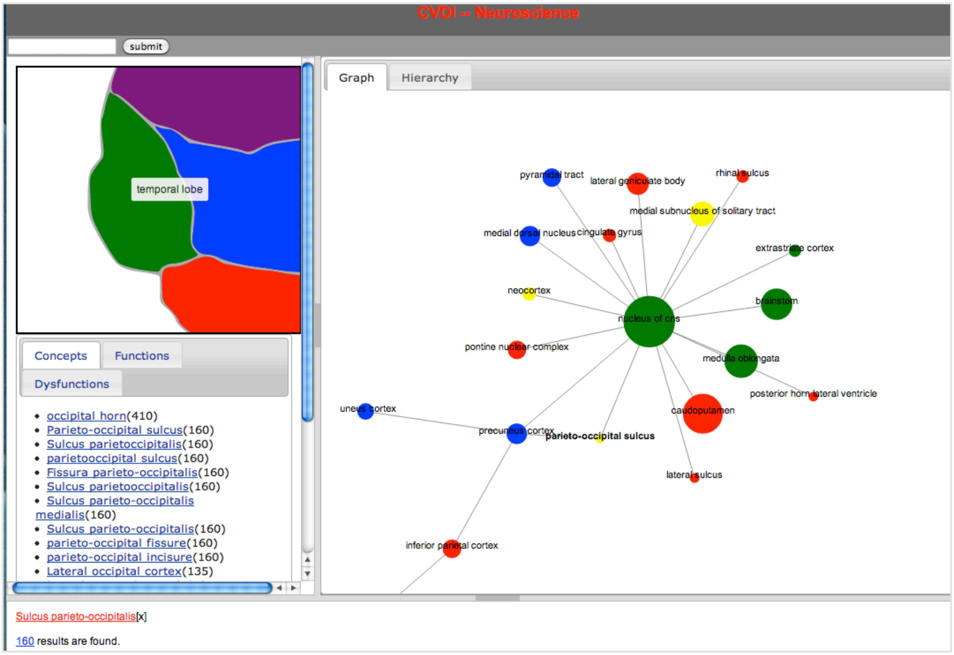
[20] X. Lin and J. Ahn, "Challenges of knowledge structure visualization," Classification & Visualization: Interfaces to Knowledge, Proceedings of the International UDC Seminar, October 2013, pp 73-87.

[21] R. W. Schvaneveldt, Pathfinder Associative Networks. Westport, CT, US:Ablex Publishing, 1990.

[22] M. Bostock, V. Ogievesky, and J. Heer, "D3: Data-driven documents," IEEE Transaction on Visualization and Computer Graphics, 17(12), Dec. 2011, pp. 2301-2309.

[23] Gephi, "Gephi, an open source graph visualization and manipulation software," Available at: https://gephi.org/ [retrieved: April 5, 2014].

[24] SIGMAJS. "Sigma.js: an open-source lightweight javaScript library," Available at: http://sigmajs.org/ [retrieved: April 5, 2014].



- occipital horn(410)
- Parieto-occipital sulcus(160)
- Sulcus parietoccipitalis(160)
- parietooccipital sulcus(160)
- Fissura parieto-occipitalis(160)
- Sulcus parietooccipitalis(160)
- Sulcus parieto-occipitalis medialis(160)
- Sulcus parieto-occipitalis(160)
- parieto-occipital fissure(160)
- parieto-occipital incisure(160)
- Lateral occipital cortex(135)

Sulcus parieto-occipitalis[x]

160 results are found.

Figure 5. An example of the model-based interface and associative concept maps.

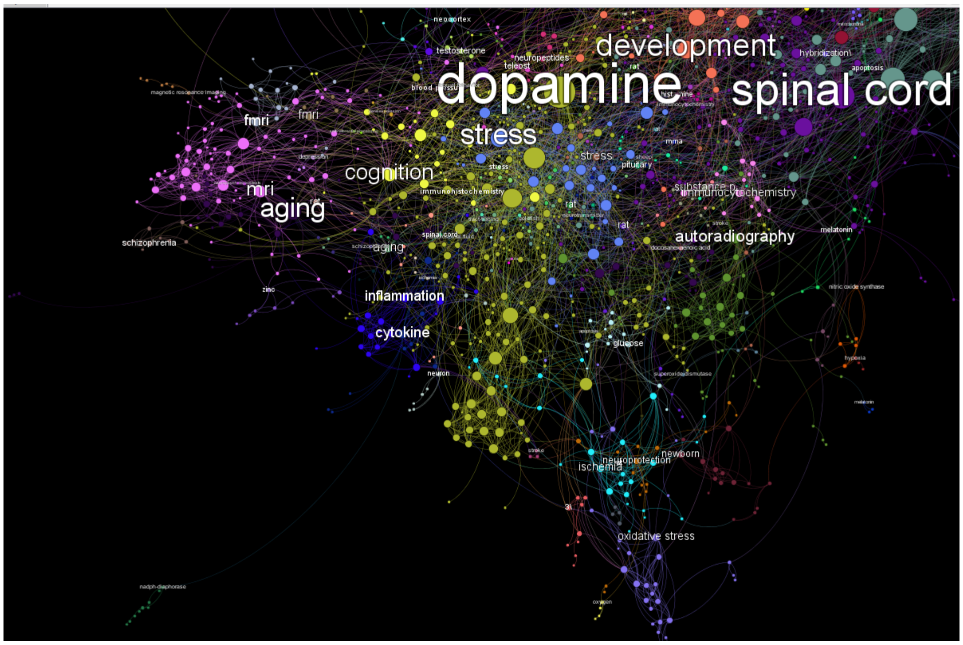


Figure 6. An example of the grobal mapping based on the relationships of the annotated concept terms over the whole collection.