Data-Brain driven systematic human brain data analysis: A case study in numerical inductive reasoning centric investigation

Action editor: Rajiv Khosla
Jianhui Chen a, Ning Zhong a,b,⇑, Peipeng Liang a,c

a International WIC Institute, Beijing University of Technology, Beijing 100124, China
b Department of Life Science and Informatics, Maebashi Institute of Technology, Maebashi-City 371-0816, Japan
c Department of Radiology, Xuanwu Hospital, Capital Medical University, Beijing 100053, China

Received 10 June 2010; accepted 31 December 2010
Available online 1 February 2011

Abstract

As a crucial step in understanding human intelligence, Brain Informatics (BI) focuses on thinking centric investigations of human cognitive functions with respect to multiple activated brain areas and neurobiological processes for a given task. Although it has been recognized that systematic human brain data analysis is an important issue of BI methodology, the existing expert-driven multi-aspect data analysis excessively depends on individual capabilities and cannot be widely adopted in BI community. In this paper, we propose a Data-Brain driven approach for systematic brain data analysis, which is implemented by using the Data-Brain, Data-Brain based BI provenances and Global Learning Scheme for BI. Furthermore, a human numerical inductive reasoning centric investigation is described to demonstrate significance and usefulness of the proposed approach. Such a Data-Brain driven approach reduces the dependency on individual capabilities and provides a practical way for realizing the systematic human brain data analysis of BI methodology.

© 2011 Elsevier B.V. All rights reserved.

Keywords: Data-Brain; Systematic human brain data analysis; Provenance; Brain Informatics

1. Introduction

The capabilities of human intelligence can be broadly divided into two main aspects: perception and thinking. As an emerging interdisciplinary field of brain science and information science, Brain Informatics (BI) (Zhong, 2006) focuses on the thinking centric investigations of human cognitive functions, which systematically study human thinking oriented higher cognitive functions, such as reasoning, problem-solving, learning, computation, from both macro and micro points of view by coopera-

⇑ Corresponding author. Addresses: Department of Life Science and Informatics, Maebashi Institute of Technology, 460-1 Kamisadori-Cho, Maebashi-City 371-0816, Japan, and the International WIC Institute Beijing University of Technology, Beijing 100124, China. Tel./fax: +81 27 265 7366.
E-mail address: zhong@maebashi-it.ac.jp (N. Zhong).
ment design also makes it necessary and possible to analyze brain data coming from multiple cognitive experiments for the research of a special cognitive function. Although our previous studies Motomura, Hara, Zhong, and Lu (2008), Zhong and Motomura (2009) have proposed a POM centric multi-aspect brain data analysis approach to realize systematic brain data analysis, the expert-driven approach needs investigators to hold all of domain and data related knowledge. This is difficult for most of ordinary investigators because a holistic multi-aspect brain data analysis should integrate various data and analyses.

In order to offset the above deficiency of the expert-driven approach, BI needs the supporting of advanced IT technologies. At present, many brain databases (Cocosco, Kollokian, Kwan, & Evans, 1997; Van Horn et al., 2001; Hunter et al., 2005) have been constructed to effectively store and share multiple levels of brain data. Some distributed analytical platforms of brain data, such as LONI pipeline,1 also provide the supporting for the integration of analytical methods. However, these existing information systems cannot effectively support the systematic human brain data analysis of BI methodology. Using those brain databases, the data selection still depends on the knowledge owned by investigators because those brain databases mainly focus on the description of experiments and data processing, and neglect the relationships among different experiments and data processing. The reason is that their data are mainly coming from isolated experiment researches and difficult to be described synthetically. Using those distributed analytical platforms, the creation of workflows still adopts an expert-driven approach because those analytical platforms mainly focus on the description and performance of analytical workflows. Hence, BI needs to develop a new approach for systematic brain data analysis by using various IT technologies.

This paper proposes a Data-Brain driven approach to support the implementation and popularization of systematic brain data analysis. The domain-driven conceptual model of brain data, i.e., Data-Brain (Chen & Zhong, 2008, 2009), Data-Brain based BI provenances, and Data-Brain based Global Learning Scheme for BI (GLS-BI) are used to perform a Data-Brain driven systematic human brain data analysis. Furthermore, a numerical inductive reasoning centric investigation is described to demonstrate significance and usefulness of the proposed approach. As a case study of domain-driven data mining in brain science, such a Data-Brain driven approach provides a practical way to realize the systematic human brain data analysis of BI methodology.

2. Methods

2.1. Data-Brain

The Data-Brain is a conceptual model of brain data, which represents functional relationships among multiple human brain data sources, with respect to all major aspects and capabilities of human information processing system, for systematic investigation and understanding of human intelligence (Chen & Zhong, 2008, 2009).

In our previous study (Chen & Zhong, 2008), a multi-view and multi-dimension framework of Data-Brain has been proposed. As shown in the top of Fig. 1, it includes various conceptual views which illustrate thinking centric investigations of BI from different perspectives based on functional relationships among human cognitive functions. Each concept in conceptual views represents a cognitive function or interesting characteristic focused by BI. Because of the limitation of space, Fig. 1 only gives two conceptual views: “Reasoning centric” conceptual view and “Computation centric” conceptual view, which describe systematic BI investigation from reasoning centric

1 http://pipeline.loni.ucla.edu.
and computation centric perspectives, respectively. For extracting these conceptual views, the Data-Brain includes its own four dimensions, namely function dimension, data dimension, experiment dimension, and analysis dimension, as shown in the bottom of Fig. 1. By a BI methodology based ontological modeling approach (Chen & Zhong, 2009), these four dimensions can be constructed to model the four issues of systematic BI methodology as stated in Section 1. Fig. 2 is a fragment of the function dimension in which human cognitive functions are classified based on the perspectives of BI investigations. The functional relationships among cognitive functions are also described according to some proofs obtained by BI investigations.

In our studies, the OWL-DL\(^2\) is used for Data-Brain construction. The formal Data-Brain provides a global framework to integrate multi-aspect domain knowledge coming from systematic BI research. This makes it possible to develop Data-Brain based systems for realizing the Data-Brain driven systematic human brain data analysis.

2.2. BI provenances

At present, BI focuses on two kinds of important brain data, ERP (event-related potential) data and fMRI (functional magnetic resonance imaging) data. Because both of them are unstructured data, the construction of semantic metadata is a key issue for the development of research supporting systems. The metadata describing the origin and subsequent processing of biological images is often referred to as “provenance” (Simmhan, Plale, & Gannon, 2005). Similarly, we call “BI Provenances”, including data provenances and analysis provenances, which are the metadata describing the origin and subsequent processing of various human brain data in systematic BI research. The data provenance describes the origin of BI data and the analysis provenance describes the processing undergone by BI data.

Though the existing common provenance models, such as OPM (Moreau et al., 2007) and provenir (Sahoo, Barga, Goldstein, & Sheth, 2008), provide effective conceptual frameworks to capture key information related to the origin and subsequent processing of biological images. In order to develop supporting systems for systematic BI research, especially systematic brain data analysis, new BI provenance models still need to be developed for the construction of various data provenances and analysis provenances.

By the BI methodology based ontological modeling process, the Data-Brain and its own domain ontologies form a knowledge base of brain data which provides a general BI provenance model to guide the construction of BI provenances for the multi-aspect requirements coming from systematic BI research. Thus, using the Data-Brain, various BI provenances can be constructed by a Data-Brain based approach:

- collecting experimental data and related information for identifying the key concepts based on the Data-Brain,
- extracting a view from the Data-Brain to get the provenance model, and
- constructing the BI provenance by creating and integrating instances based on the obtained provenance model.

By such a Data-Brain based approach, the obtained BI provenences not only include data information needed by systematic BI research but also become a bridge which connects the Data-Brain and heterogeneous brain data to form a brain data and knowledge base, as shown in Fig. 3. Such a brain data and knowledge base can provide the important data and knowledge to support systematic BI research, especially the systematic human brain data analysis.

The more descriptions about data provenances and analysis provenances will be given in the next subsections.

2.3. BI data provenances

A BI data provenance is a metadata set that describes the BI data origin by multi-aspect experimental information, including subjects information, how experimental data of subjects were collected, what instrument was used, etc. As stated above, the BI provenance can be constructed by a Data-Brain based approach.

For example, our previous study (Liang et al., 2006) introduced a group of fMRI experiments of human numerical inductive reasoning. 30 paid undergraduate or graduate male students from the Beijing University of Technology participated in the experiment\(^3\). One kind of intelligence test problems was adapted as the experimental tasks. Its basic element is a reverse triangle in which the three numbers located in three different positions may constitute a calculation rule. Addition and subtraction are used to form the calculation rule for induction tasks. Two kinds of tasks were designed, including reversed trian-

\(^2\) http://www.w3.org/TR/owl-features/.

\(^3\) Because some data were identified as invalid data, only 20 subjects were reported in the literature.
gle induction tasks and reversed triangle computing tasks, as shown in Fig. 4. Computing tasks were designed as the baseline of induction tasks. Aiming at this experiment, a BI data provenance can be constructed to describe experimental data by the following Data-Brain based approach.

Firstly, as shown in Fig. 5, a group of data and information are collected from various research supporting systems. They include a group of experimental materials, a group of equipment parameters, thirty experimental records which describe the information about experimental processes, thirty subject records, and thirty fMRI data sets in which each data set includes a series of whole-brain Blood Oxygenation Level-dependent (BOLD) images with the EPI (echo planar imaging) sequence. Based on the Data-Brain, a group of key concepts can be identified from these data and information, such as “MRI”, “College-Student”.

Secondly, extracting these concepts and other related concepts, as well as relations among them, by a traversal in the ontological Data-Brain, an ontological view (Noy & Musen, 2004) can be obtained. After some necessary revisions, including merging subclasses, removing reduplicate relations, etc., a data provenance model can be created, as shown in Fig. 6. In this model, the fMRI experiments are general-
ized as an “Experimental-Group” which includes a group of “fMRI-Experiment” and uses “MRI” as the measuring means. Each “fMRI-Experiment” is involved with a “College-Student” and produces a fMRI data set which includes a series of whole-brain BOLD images with the “EPI-Sequence” format. In this “Experimental-Group”, two kinds of “Reversed-Triangle-Task”, “Reversed-Triangle-Induction-Task” and “Reversed-Triangle-Computing-Task”, are adopted and reflect two important human cognitive functions “Numeric-Induction” and “Computation”, respectively. These tasks are realized by a group of “Experimental-Materials”.

Lastly, as shown in Fig. 7, a BI data provenance can be constructed by creating and integrating instances of related concepts and relations based on the obtained provenance model. In our studies, the RDF\(^4\) is used to construct BI data provenances.

---
\(^4\) [http://www.w3.org/TR/rdf-concepts/](http://www.w3.org/TR/rdf-concepts/)
2.4. BI analysis provenances

A BI analysis provenance is a metadata set that describes what processing in a brain data set has been carried out, including what analytic tasks were performed, what experimental data were used, what data features were extracted, and so on. Similar to the data provenance, it can also be constructed based on the Data-Brain.

For example, in our previous study (Liang et al., 2006), the fMRI data sets obtained by the above fMRI experiments of the reversed triangle inductive reasoning were analyzed using Statistical Parametric Mapping v.2 (SPM2). The whole of analytical process includes multiple sub-steps, such as slice timing, realign, etc. Aiming at this data analysis, a BI analysis provenance can be constructed as follows.

Firstly, as shown in Fig. 8, a group of data and information are collected from research supporting systems. They include an analytical record which describes the basic information of analytical process, twenty valid fMRI data sets, a software and a group of parameters, and a group of found brain activations. Based on the Data-Brain, key concepts can be identified from these data and information, such as “SPM”, “Activation”.

Secondly, as shown in Fig. 9, an analysis provenance model can be obtained by extracting and revising the onto-

![Fig. 8. The data and information coming from a SPM2 based fMRI data analysis.](image1)

![Fig. 9. An analysis provenance model for the SPM2 based fMRI data analysis.](image2)
logical view of Data-Brain. In this model, the SPM2 based analysis is generalized as an analytical process “Finding-Activations-by-SPM”. It uses “SPM” to perform the “Analytic-Task” which includes two important sub-tasks, “Pre-processing-fMRI-Data” and “Identifying-Activations-in-fMRI-Data”. Its input data are a group of fMRI data sets with the “EPI-Sequence” format. Its results are a group of “Activation” in which each brain activation is produced by “Reversed-Triangle-Induction-Task” and reflects a special “Information-Processing-Course”. The locations of brain activations are described by two types of “Brain-Area”, “Anatomical-Area” and “Brodmann-Area”.

Lastly, the corresponding RDF based BI analysis provenance can be constructed by the creation and integration of instances, as shown in Fig. 10.

2.5. Global Learning Scheme for BI

In order to offset the disadvantages of the existing expert-driven multi-aspect brain data analysis, the Global Learning Scheme for BI (GLS-BI) (Zhong & Chen, 2010) was designed to perform a Data-Brain driven mining process planning for multi-aspect brain data analysis. It is an extension of the Global Learning Scheme (GLS) (Brazdil, Giraud-Carrier, Soares, & Vilalta, 2009; Zhong, Liu, & Ohsuga, 2001) to model the process of multi-aspect brain data analysis as an organized society of autonomous knowledge discovery agents for dynamically organizing and managing the brain data mining process based on the Data-Brain.

In the GLS-BI, semantic Web Service technologies are used to wrap physical or virtual data and analysis resources in the BI community as various data and analysis agents, whose descriptions are annotated by the Data-Brain and BI provenances. According to different analysis purposes, these data and analysis agents can be integrated as various mining workflows by the multi-aspect mining process planning. As shown in Fig. 11, the whole of planning process includes the four steps: purpose definition, Data-Brain driven data selection, Data-Brain driven analysis agent discovery, and workflow extraction, where only steps 1 and 2 need users’ participation. In purpose definition, users need to define the analysis purpose by choosing an objective cognitive function and multiple types of data features. In Data-Brain driven data selection, users need to choose fit rules for defining “similar” data features according to different requirements.

Based on our previous studies (Motomura et al., 2008; Zhong & Motomura, 2009), a multi-aspect brain data analysis can be generalized as “aiming at an objective cognitive function (e.g. induction) to investigate its information processing course by extracting and analyzing various spatio-temporal features (e.g. activations in brain) on related data sets”. Because the thinking centric investigations are complex and tightly related to each other, the “related” means various pertinences between data sets and objective cognitive functions. Besides the “matching” data, i.e., the data coming from the experimental researches of objective cognitive functions, the “similar” data, i.e., the data coming from the experimental researches of similar cognitive

<table>
<thead>
<tr>
<th>Concepts in Data-Brain</th>
<th>Instances of Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytic-Process</td>
<td>ad:Analytic-Process</td>
</tr>
<tr>
<td>ERP-Data-Analysis</td>
<td>ad:ERP-Data-Analysis</td>
</tr>
<tr>
<td>MRI-Data-Analysis</td>
<td>ad:MRIData-Analysis</td>
</tr>
<tr>
<td>Software</td>
<td>ad:Software</td>
</tr>
<tr>
<td>Brain-Vision-Analyzer</td>
<td>ad:Brain-Vision-Analyzer</td>
</tr>
<tr>
<td>SPM</td>
<td>ad:SPM</td>
</tr>
<tr>
<td>Analytic-Task</td>
<td>ad:Analytic-Task</td>
</tr>
<tr>
<td>Data-Preprocessing</td>
<td>ad:Data-Preprocessing</td>
</tr>
<tr>
<td>Preprocessing-fMRI-Data</td>
<td>ad:Preprocessing-fMRI-Data</td>
</tr>
<tr>
<td>Data-Feature-Extraction</td>
<td>ad:Feature-Extraction</td>
</tr>
<tr>
<td>Identifying-Activations-in-fMRI-Data</td>
<td>ad:Identifying-Activations-in-fMRI-Data</td>
</tr>
<tr>
<td>Literature</td>
<td>ad:literature</td>
</tr>
<tr>
<td>Experimental-Task</td>
<td>ad:Experimental-Task</td>
</tr>
<tr>
<td>Numerical-Task</td>
<td>ad:Numerical-Task</td>
</tr>
<tr>
<td>Bio-Data</td>
<td>ad:Bio-Data</td>
</tr>
<tr>
<td>Original-fMRI-Data</td>
<td>dd:Original-fMRI-Data</td>
</tr>
<tr>
<td>BOLD-Image</td>
<td>dd:BOLD-Image</td>
</tr>
<tr>
<td>EPI-Sequence</td>
<td>dd:EPI-Sequence</td>
</tr>
<tr>
<td>Temporal-Feature</td>
<td>dd:Temporal-Feature</td>
</tr>
<tr>
<td>Activation</td>
<td>dd:Activation</td>
</tr>
<tr>
<td>Information-Processing</td>
<td>dd:Information-Processing</td>
</tr>
<tr>
<td>Course</td>
<td>fd:Information-Processing</td>
</tr>
<tr>
<td>Domain Concepts</td>
<td>dc:DomainConcepts</td>
</tr>
<tr>
<td>Anatomical-Area</td>
<td>dc:Anatomical-Area</td>
</tr>
<tr>
<td>Left-Superior-Frontal-Gyrus</td>
<td>dc:Left-Superior-Frontal-Gyrus</td>
</tr>
</tbody>
</table>

Fig. 10. Analysis provenance construction based on the Data-Brain.
functions, need to be comparatively analyzed to discover the key spatio-temporal features in the whole of information processing course. The “functionally related” data, i.e., the data coming from the experimental researches of functionally related cognitive functions, also need to be synthetically analyzed to understand the found key spatio-temporal features in depth. Hence, the GLS-BI provides the following three types of workflows:

- An equivalent workflow represents a possible mining process which aims at the experimental data “matching” to the objective cognitive function of multi-aspect data analysis. The “matching” does not mean that the data is just corresponding to the objective cognitive function. The data corresponding to direct or indirect sub-functions of objective cognitive function are also regarded as “matching” data. For example, when the “reasoning” is selected as an objective cognitive function, both the data corresponding to its direct sub-functions, such as “induction” and “deduction”, and the data corresponding to its indirect sub-functions, such as “numerical induction” and “sentential induction”, are also the “matching” data.

- A related workflow represents a possible mining process which aims at the experimental data “functionally related” to the “matching” data. As stated above, the thinking centric investigations are complex and involved with multiple inter-related cognitive functions for a given task. Thus, it is necessary to synthetically analyze the data coming from “functionally related” experimental tasks to understand the found key spatio-temporal features in depth. The “functionally related” means that these experimental tasks elicit similar spatio-temporal features in neural structures and mechanisms. Furthermore, because spatio-temporal features can be reflected by the extracted data features, the experimental tasks from whose data the “similar” data features are
extracted can be regarded as the “functionally related” experimental tasks. Their data are just the “functionally related” data. For example, our previous study shows that the ERP component “posterior LPC”, extracted from the experimental data of numerical induction tasks (Liang et al., 2007), is similar to the classical ERP component “P300” extracted from the experimental data of mental arithmetic tasks (Iguchi & Hashimoto, 2000). Thus, these two experimental tasks are “functionally related”. When the “numerical induction” is selected as an objective cognitive function, the data of the numerical induction tasks stated in Liang et al. (2007) are the “matching” data, and the data of the mental arithmetic tasks stated in Iguchi and Hashimoto (2000) are the corresponding “functionally related” data.

3. Results

The Data-Brain, BI provenances and the GLS-BI make it possible to perform a Data-Brain driven systematic human brain data analysis. In this section, a numerical inductive reasoning centric case study will be introduced to demonstrate such a Data-Brain driven approach. It integrates a series of induction centric data and analyses in BI community for understanding the principles and mechanisms of human numerical inductive reasoning related functions in depth.

3.1. Inductive reasoning centric data and analyses

Inductive reasoning is an important human higher-level cognitive function. It has been studied by various domains, including logic, cognitive psychology, and artificial intelligence (AI), etc. For disclosing the information processing mechanism of inductive reasoning in the brain, a series of inductive reasoning related cognitive experiments have been carried out in BI studies. As shown in Table 1, they include nine groups of experiments in which each group

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>EG2</td>
<td>The fMRI experiments of sentential induction strength judgment (thirty subjects)</td>
</tr>
<tr>
<td>EG3</td>
<td>The fMRI experiments of sentential induction with multi-level preconditions (twenty-two subjects)</td>
</tr>
<tr>
<td>EG5</td>
<td>The fMRI experiments of trained number series completion induction (thirteen subjects)</td>
</tr>
<tr>
<td>EG6</td>
<td>The fMRI experiments of figural induction (sixteen subjects)</td>
</tr>
<tr>
<td>EG7-1</td>
<td>The ERP experiments of reversed triangle induction (eleven subjects)</td>
</tr>
<tr>
<td>EG7-2</td>
<td>The fMRI experiments of reversed triangle induction (thirty subjects)</td>
</tr>
<tr>
<td>EG8</td>
<td>The ERP experiments of simultaneously presented reversed triangle induction (fourteen subjects)</td>
</tr>
<tr>
<td>EG9</td>
<td>The ERP experiments of sentential induction strength judgment (fourteen subjects)</td>
</tr>
<tr>
<td>EG10</td>
<td>The ERP supplementary experiments of sentential induction strength judgment (sixteen subjects)</td>
</tr>
</tbody>
</table>

Fig. 12. A series of graphical user interfaces for data provenance construction.
includes multiple experimental tasks. All of experimental
data were stored into a brain data center. The correspond-
ing data provenances were also created by using a series of
graphical user interfaces (GUIs) shown in Fig. 12. Based
on these data and data provenances, twenty-four data
agents were constructed in the GLS-BI. Each data agent
is corresponding to a data set obtained by an experimental
task.

Furthermore, based on our previous studies (Liang
et al., 2006; Motomura et al., 2008; Yang et al., 2010), a
group of BI analysis provenances were created by using
the GUIs shown in Fig. 13. Ten analysis agents were also
constructed in the GLS-BI. For simplifying descriptions
of workflows, each analysis agent is corresponding to not
an atomic data operation but a group of data operations
whose results can be used for multiple analytical methods.

These inductive reasoning centric data, data/analysis
provenances, data/analysis agents provide data, knowledge
and analysis bases for performing a numerical inductive
reasoning centric systematic brain data analysis.

3.2. Numerical inductive reasoning centric systematic brain
data analysis

Similar to the multi-aspect brain data analysis, a numer-
icial inductive reasoning centric systematic brain data anal-
ysis can be presented as “aiming at the objective cognitive
function numerical induction to investigate its information
processing course by extracting and analyzing various spa-
tio-temporal features on related data sets”. Based on the
Data-Brain, BI provenances and the GLS-BI, such a sys-
tematic brain data analysis can be implemented by the fol-
lowing four steps:

- the GLS-BI based multi-aspect mining process planning,
- the equivalent workflow based typical analysis,
- the reference workflow integrated exploratory analysis, and
- the related workflow integrated specific analysis.

The more descriptions will be given in the next
subsections.

3.3. The GLS-BI based multi-aspect mining process planning

The GLS-BI based multi-aspect mining process plan-
ing is to generate all of possible mining workflows using
the GLS-BI. As stated above, this process includes four
steps in which the steps 1 and 2 need users’ participation.
In this case, for purpose definition, the objective cognitive
function “Numerical-Induction” and five types of data fea-
tures, “Topography-from-ERP-Data”, “ERP-Component”,”
Peculiarity-in-the-Amplitude-of-ERP- Waveform”, “Pecu-
liarity-in-the-Latent-Time-of-ERP- Waveform”, and “Acti-
vation”, were chosen. For Data-Brain driven data selection,
a rule “all of activations, which are located in a same

Fig. 13. A series of graphical user interfaces for analysis provenance construction.
anatomical brain area, are the similar activations” was chosen to define the “similar” data features. This rule can be represented as the following custom Jena rule:\footnote{http://jena.sourceforge.net/inference/}:

\[
\begin{align*}
&(?x \text{ rdf:type} \text{ dd:Activation}), \\
&(?y \text{ rdf:type} \text{ dd:Activation}), \\
&(?xa \text{ rdf:type} \text{ rd:Anatomical-Area}), \\
&(?ya \text{ rdf:type} \text{ rd:Anatomical-Area}), \\
&(?x \text{ db:located-in} ?xa),(?y \text{ db:located-in} ?ya), \\
&(?xa \text{ rdf:type} \text{ ?at}), (?ya \text{ rdf:type} \text{ ?at}), \\
&\text{notEqual}(?x,?y), \\
&(?at \text{ rdfs:subClassOf} \text{ rd:Anatomical-Area}) \\
&\rightarrow (?x \text{ dd:similar-to} ?y) 
\end{align*}
\]

By such a GLS-BI based multi-aspect mining process planning, three types of workflows can be obtained, including 14 equivalent workflows, 24 reference workflows, and 26 related workflows, as shown in Fig. 14.

3.4. The equivalent workflow based typical analysis

Each equivalent workflow represents a possible mining process which aims at the experimental data “matching” to the “Numerical-Induction”. It mainly applies an analytical tool or algorithm on a data set coming from an experimental group. This type of mining process is the most typical mode of brain data analysis. For example, the equivalent workflow $\text{wf}_2$ shown in Fig. 15 represents a SPM2 based mining process for the fMRI data set obtained by the reversed triangle induction tasks. Such a mining process has been adopted in our previous study (Liang et al., 2006). As shown in Table 2, various brain activations were gotten.

Although an equivalent workflow represents a mining process based on an analytical method and a data set, this does not mean only to apply an analytical method on a data set. As shown in Fig. 15, the equivalent workflow $\text{wf}_1$ represents that a NOC-SVM based method (Yang et al., 2010) can also be applied on the fMRI data set of the reversed triangle induction tasks for finding brain activations.
Fig. 15. Two mining workflows for the fMRI data set obtained by the reversed triangle inductive reasoning tasks. The Data Agent 12 represents the fMRI data set obtained by the reversed triangle inductive reasoning tasks (Liang et al., 2006). The analysis agents 8 and 9 represent the fMRI data analysis method used in (Liang et al., 2006). The Analysis Agent 8 represents a series of data preprocessing operations using the software SPM2, including slice timing, realign, etc. The Analysis Agent 9 represents a series of data operations for finding activations using the SPM2, including model specification, data specification, etc. The Analysis Agent 10 represents the NOC-SVM based method for finding activations (Yang et al., 2010).

Table 2
The results of workflow WF2.

<table>
<thead>
<tr>
<th>Location</th>
<th>Cluster</th>
<th>BA</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left superior frontal gyrus</td>
<td>263</td>
<td>6</td>
<td>-4</td>
<td>6</td>
<td>64</td>
<td>5.48</td>
</tr>
<tr>
<td>Right medial frontal gyrus</td>
<td>263</td>
<td>6</td>
<td>4</td>
<td>10</td>
<td>48</td>
<td>5.01</td>
</tr>
<tr>
<td>Left middle frontal gyrus</td>
<td>101</td>
<td>6</td>
<td>-26</td>
<td>2</td>
<td>62</td>
<td>5.23</td>
</tr>
<tr>
<td>Right middle frontal gyrus</td>
<td>101</td>
<td>6</td>
<td>-36</td>
<td>4</td>
<td>54</td>
<td>3.84</td>
</tr>
<tr>
<td>Right medial frontal gyrus</td>
<td>93</td>
<td>6</td>
<td>30</td>
<td>8</td>
<td>64</td>
<td>4.75</td>
</tr>
<tr>
<td>Right medial frontal gyrus</td>
<td>93</td>
<td>6</td>
<td>44</td>
<td>10</td>
<td>52</td>
<td>4.53</td>
</tr>
<tr>
<td>Parietal lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left precuneus</td>
<td>1534</td>
<td>7</td>
<td>-26</td>
<td>-72</td>
<td>32</td>
<td>6.50</td>
</tr>
<tr>
<td>Left inferior parietal lobule</td>
<td>1534</td>
<td>7</td>
<td>-16</td>
<td>-70</td>
<td>48</td>
<td>5.66</td>
</tr>
<tr>
<td>Right inferior parietal lobule</td>
<td>76</td>
<td>40</td>
<td>-42</td>
<td>-42</td>
<td>40</td>
<td>4.26</td>
</tr>
<tr>
<td>Right superior parietal lobule</td>
<td>928</td>
<td>40</td>
<td>34</td>
<td>-52</td>
<td>34</td>
<td>6.16</td>
</tr>
<tr>
<td>Occipital lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left fusiform gyrus</td>
<td>1204</td>
<td>18</td>
<td>-20</td>
<td>-90</td>
<td>-24</td>
<td>7.97</td>
</tr>
<tr>
<td>Left inferior occipital gyrus</td>
<td>1204</td>
<td>18</td>
<td>-34</td>
<td>-90</td>
<td>-14</td>
<td>7.37</td>
</tr>
<tr>
<td>Right middle occipital gyrus</td>
<td>1325</td>
<td>18</td>
<td>22</td>
<td>-86</td>
<td>-14</td>
<td>7.60</td>
</tr>
<tr>
<td>Right inferior occipital gyrus</td>
<td>1325</td>
<td>18</td>
<td>34</td>
<td>-84</td>
<td>-18</td>
<td>7.04</td>
</tr>
</tbody>
</table>
The equivalent workflow based analysis is the traditional and typical analysis. It is the base of Data-Brain driven systematic human brain data analysis. On the one hand, various brain data can be initially understood using different analytical methods or tools. On the other hand, the obtained data features can be used for the further analysis.

3.5. The reference workflow integrated exploratory analysis

Each reference workflow represents a possible mining process which aims at the data sets “similar” to the “Numerical-Induction”, i.e., the data sets coming from sentential induction tasks, figural induction tasks, and other types of induction tasks.

The reference workflow integrated exploratory analysis is just to integrate analyzed results of multiple equivalent workflows and reference workflows to perform an exploratory analysis, i.e., synthetically analyze the objective cognitive function with some “similar” cognitive functions to find the key spatio-temporal features in the whole of information processing course of the objective cognitive function.

For example, the activation likelihood estimation (ALE) method (Turkeltaub, Eden, Jones, & Zeffiro, 2002) can be used to implement an exploratory analysis. We integrated the activations obtained by the equivalent workflows wf_2, wf_16, wf_18, i.e., a group of SPM2 based mining processes for the data of numeric induction tasks, and the reference workflows wf_32, wf_34, wf_36, wf_38, wf_40, wf_42, i.e., a group of SPM2 based mining processes for the data of sentential and figural induction tasks, to perform an ALE based meta-analysis for finding the common activated brain regions during the information processing course of induction. In order to improve the reliability of analysis, activations stated in some related induction studies (Goel, Gold, Kapur, & Houle, 1997; Goel & Dolan, 2000, 2004; Christoff & Prabhakaran, 2001) were also integrated into this meta-analysis. Fig. 16 describes the distribution of activations which were used for the ALE based meta-analysis.

The software GingerALE v1.1 (Laird et al., 2005) was used for this meta-analysis. Finally, twenty-six activated brain regions shown in Table 3 were identified. The results disclose that the common activated brain regions of induction include left middle frontal gyrus (BA 6, 8, 10, 11), left inferior frontal gyrus (BA 9, 47), right superior frontal gyrus (BA 6), right middle frontal gyrus (BA 46), right inferior frontal gyrus (BA 9, 47), left insula (BA 13), right inferior temporal gyrus (BA 20), bilateral superior parietal lobule (BA 7), bilateral occipital gyrus (BA 18, 19), left putamen, left caudate head, left ventral lateral nucleus, and right hippocampus. Such broad activations show that, as a higher-level cognitive function, induction has a complex neural mechanism and is completed by the cooperation of multiple brain regions.

The reference workflow integrated analysis is an exploratory analysis. Guiding by equivalent workflows and reference workflows, the results of various typical analysis are
integrated to find the key spatio-temporal features in the whole of information processing course of the objective cognitive function. On the one hand, such an exploratory analysis can deepen the understanding of objective cognitive function. On the other hand, analyzed results can be regarded as valuable prior knowledge for the further specific analysis.

3.6. The related workflow integrated specific analysis

Each related workflow represents a possible mining process which aims at the data sets obtained by the experimental tasks “functionally related” to the “Numerical-Induction”. As stated above, the “functionally related” means that these experimental tasks elicit “similar” spatio-temporal features, i.e., the extracted data features, with one or several experimental tasks of “Numerical-Induction”.

The related workflow integrated specific analysis is just to integrate analyzed results of equivalent workflows, reference workflows and related workflows to perform a specific analysis, i.e., synthetically analyze the objective cognitive function with other “functionally related” cognitive functions to understand the found key spatio-temporal features in depth.

The above exploratory analysis discloses that induction is involved with frontal lobe, temporal lobe, parietal lobe, occipital lobe and subcortical structures. Based on this prior knowledge, the specific analyses pointed to these brain regions can be performed to understand the numerical induction in depth.

For example, because many previous induction studies (Girelli, Semenza, & Delazer, 2004; Goel et al., 1997; Goel & Dolan, 2004) focused on the activations on DLPFC (BA 46, 9), we chose DLPFC as the object of specific analysis and analyzed its functions based on domain knowledge. All of activations obtained by equivalent workflows, reference workflows and related workflows were integrated for this analysis. According to domain knowledge, the DLPFC includes BA46 and a part of BA9. Thus, the functional analysis of DLPFC can be completed by performing a SPARQL query on the brain data and knowledge base. The following is the corresponding query expression:

(3) ?DataFeatureURI db:produced-by ?ExpTaskURI.
(7) ?BrainAreaURI2 rdf:type rd:Brodmann-Area.
(8) (?BrainAreaURI2 rdf:type rd:BA46.)
(9) Union {?BrainAreaURI1 rdf:type rd:Middle-Frontal-Gyrus.} ) ORDER BY ASC(?ExpTask)

Table 4 gives the results of query, which show that the DLPFC is involved with not only the numerical induction task NST04, but also the figural induction tasks FT01 and FT02. This means that it cannot be regarded as the special brain region of numerical induction.

Table 3
Partial brain regions identified by the ALE based meta-analysis.

<table>
<thead>
<tr>
<th>Location</th>
<th>BA</th>
<th>Cluster</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Middle-Frontal-Gyrus</td>
<td>11</td>
<td>1952*</td>
<td>−24</td>
<td>40</td>
<td>−8</td>
<td>0.0167</td>
</tr>
<tr>
<td>Left Inferior Frontal Gyrus</td>
<td>47</td>
<td></td>
<td>−24</td>
<td>34</td>
<td>−8</td>
<td>0.0153</td>
</tr>
<tr>
<td>Left Middle-Frontal-Gyrus</td>
<td>9</td>
<td>832</td>
<td>−42</td>
<td>10</td>
<td>28</td>
<td>0.0153</td>
</tr>
<tr>
<td>Left Medial Frontal Gyrus</td>
<td>10</td>
<td>200</td>
<td>−36</td>
<td>52</td>
<td>8</td>
<td>0.0118</td>
</tr>
<tr>
<td>Left Inferior Frontal Gyrus</td>
<td>6</td>
<td>152</td>
<td>−8</td>
<td>−26</td>
<td>54</td>
<td>0.0102</td>
</tr>
<tr>
<td>Right Inferior Frontal Gyrus</td>
<td>47</td>
<td>1168</td>
<td>30</td>
<td>28</td>
<td>−2</td>
<td>0.0184</td>
</tr>
<tr>
<td>Right Superior Frontal Gyrus</td>
<td>6</td>
<td>1264</td>
<td>2</td>
<td>14</td>
<td>48</td>
<td>0.0158</td>
</tr>
<tr>
<td>Right Middle-Frontal-Gyrus</td>
<td>46</td>
<td>840</td>
<td>42</td>
<td>32</td>
<td>16</td>
<td>0.0152</td>
</tr>
<tr>
<td>Right Inferior Frontal Gyrus</td>
<td>9</td>
<td>456</td>
<td>48</td>
<td>12</td>
<td>24</td>
<td>0.0141</td>
</tr>
<tr>
<td>Right Middle-Frontal-Gyrus</td>
<td>6</td>
<td>416</td>
<td>30</td>
<td>−2</td>
<td>50</td>
<td>0.0142</td>
</tr>
<tr>
<td>Temporal lobe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Insula</td>
<td>13</td>
<td>1952*</td>
<td>−30</td>
<td>24</td>
<td>0</td>
<td>0.0127</td>
</tr>
</tbody>
</table>

Table 4
The results of query Q1.

<table>
<thead>
<tr>
<th>ExpTask</th>
<th>CognitiveFunction</th>
<th>DataFeatureURI</th>
</tr>
</thead>
<tbody>
<tr>
<td>NST04</td>
<td>… /Numerical-Induction/NI03</td>
<td>… /A36</td>
</tr>
<tr>
<td>FT01</td>
<td>… /Figural-Induction/FI01</td>
<td>… /A101</td>
</tr>
<tr>
<td>FT02</td>
<td>… /Figural-Induction/FI02</td>
<td>… /A87</td>
</tr>
<tr>
<td>FT02</td>
<td>… /Figural-Induction/FI02</td>
<td>… /A128</td>
</tr>
<tr>
<td>FT02</td>
<td>… /Figural-Induction/FI02</td>
<td>… /A88</td>
</tr>
</tbody>
</table>

6 http://www.w3.org/TR/rdf-sparql-query/.
The related workflow integrated analysis is a specific analysis. Integrating three types of workflows and their results, multiple key spatio-temporal features of information processing course can be chosen and analyzed based on the prior analyses and knowledge. This can help investigators understand the objective cognitive function in depth for getting new cognitive models.

4. Discussion

The above scenario illustrates a case study on the Data-Brain driven systematic human brain data analysis. During such a domain-driven data mining process, the Data-Brain, BI provenances and the GLS-BI guided a systematic analysis process to integrate various data and analyses for disclosing the information processing course of numerical inductive reasoning step by step. This is a Data-Brain driven approach. On the one hand, the formal Data-Brain and Data-Brain based metadata and information systems simplify and standardize the analytical process for systematic human brain data analysis, which needs a large amount of domain and data related knowledge. The investigators who lack enough domain and data related knowledge can also easily apply multiple analytical methods to systematically analyze various brain data. This is a practical way to realize the systematic human brain data analysis of BI methodology. On the other hand, the exploratory analysis and specific analysis integrate “similar” and “functionally related” data to investigate a special objective cognitive function. This means that a data set can be utilized for multiple purposes. The valuable data resources can be adequately utilized.

Over the last decade, the fast developing of brain science has led to a vast increase of brain data from across a variety of spatial and temporal scales. Large amounts of brain databases have been constructed, such as Olfactory Receptor DataBase, BrainWeb, and fMRI Data Center. In recent years, long-term community-oriented databasing of brain data has gotten attentions (Van Horn & Toga, 2009). Many researches begin to focus on how to effectively and adequately utilize the existing brain data and methods for a special research community.

The goal of Neuroscience Information Framework (NIF) is to develop an inventory of information and other resources within a framework that enables neuroscientists to identify resources relevant to their research needs. This framework focuses on concept based queries (spanning multiple levels of biological organizations and functions) within and across the diverse types of information inventories. BrainMap is an online database of published functional neuroimaging experiments with coordinate-based activation locations in Talairach space. Its goal is to provide a vehicle to share methods and results of studies in specific research domains. A software suite is also provided for the ALE based meta-analyses of similar studies on BrainMap. The LONI Pipeline is a graphical environment for the construction, validation and execution of advanced neuroimaging data analysis protocols. Its goal is to provide an efficient distributed computing platform to combine diverse data, software tools and network infrastructures for the large-scale and distributed neuroimaging data analysis.

All of above projects focus their efforts on simplifying the finding and integration of brain data and analyses by advanced IT technologies. However, for specific domain problems, it is an important issue to develop various methodologies for guiding the utilities of information systems. The ALE meta-analysis workflow (Laird et al., 2009) is a valuable work on this issue. This paper also represents a case study in this issue, which aims at the systematic human brain data analysis, an important domain problem in BI. Furthermore, as a domain-driven conceptual model of brain data, the Data-Brain makes the proposed approach practical and easy to be popularized.

5. Conclusions

For offsetting the disadvantages of the existing expert-driven multi-aspect data analysis, this paper proposed a Data-Brain driven approach for systematic human brain data analysis by using the Data-Brain, BI provenances and the GLS-BI. A numerical inductive reasoning centric investigation has been described to demonstrate the whole analytical process. As a case study of domain-driven data mining in brain science, this Data-Brain driven approach standardizes the analytical process and reduces the dependency on individual capabilities. It provides a practical way to realize the systematic human brain data analysis of BI methodology.

Our studies only obtained some preliminary results. The future work mainly includes the following two aspects:

- Performing the further analysis based on public brain databases: Our present studies mainly focus on the data and analyses in BI community. However, the limited data and analysis resources cannot meet the requirements of a holistic systematic brain data analysis. Thus, it is necessary to use the experimental data and analyzed results stored in public brain databases. For example, the activation information in BrainMap can be used to perform the exploratory analysis for getting more reliable results.

- Embedding the GLS-BI into a mature distributed computing environment: The current GLS-BI only provides graphical descriptions of mining workflows for guiding the systematic brain data analysis. Investigators still need to manually access each data/analysis agent and perform the data analysis step by step. In order to improve the automatic degree of analytical processes,

---

7 http://senselab.med.yale.edu/ORDB/default.asp.
9 http://www.fmridc.org/l/fmridc.
the GLS-BI needs to be embedded into a mature distributed computing environment, such as the LONI pipeline processing environment, for constructing a global platform of systematic human brain data analysis.

Acknowledgments

The work is partially supported by the China Scholarship Council (CSC) (File No. 2009654018), National Natural Science Foundation of China (No: 60905027), and Support Center for Advanced Telecommunications Technology Research, Foundation (SCAT), Japan.

References


