Data Modeling for Content-Based Support Environment (C-BASE): Application on Epilepsy Data Mining*

Mohammad-Reza Siadat[a,b], Hamid Soltanian-Zadeh[b,c],Farshad Fotouhi[d], Ameen Eetemadi[b,d], and Kost Elisevich[e]

[a]*Department of Computer Science and Engineering, Oakland University, Rochester, MI 48309, USA*
[b]*Radiology Image Analysis Lab., Henry Ford Health System, Detroit, MI 48202, USA*
[c]*Control and Intelligent Proc. Center of Excellence, Dept. of Elec. & Comp. Eng., Univ. of Tehran, Tehran 14395-515, Iran*
[d]*Department of Computer Science, Wayne State University, Detroit, MI 48202, USA*
[e]*Department of Neurosurgery, Henry Ford Health System, Detroit, MI 48202, USA*

Abstract

Data Modeling is an essential first step for data preparation in any data mining procedure. Conventional entity-relational (E-R) data modeling is lossy, irreproducible, and time-consuming especially when dealing with unstructured image data associated with complex systems like the human brain. We propose a methodological framework for more objective E-R data modeling by eliminating the structured content-dependent metadata associated with the unstructured data. The proposed method is applied to epilepsy-related image data and a system called the human brain image database system (HBIDS) is developed accordingly. Supported with navigation, segmentation, data fusion, and feature extraction modules, HBIDS provides a content-based support environment (C-BASE). Such an environment potentially provides an unlimited (ad hoc) query support with a reproducible and efficient database schema. Switching between different modalities of data, while confining the feature extractors within the object(s) of interest, HBIDS yields anatomically specific query results. The price of such scheme is large storage requirements and relatively high computational cost. Examples of navigation through unstructured image data and content-based retrieval are presented in this paper. The results show the potential of HBIDS in content-based data management for decision support systems in real life medical applications.

1. Introduction

The first step in any data mining procedure is data preparation [1], which has to be built upon a database system. Entity-relational databases are the most common and well-established type of database that offers a wide range of desired features. On the other hand, conventional entity-relational (E-R) data modeling for systems as complex as the human brain and their multimodality unstructured data does not lead to a lossless, feasible, and reproducible result [2]. A lossless data model supports all future ad hoc queries and a feasible data model allows doable database implementation and data entry. A reproducible data model promotes inter- and intra-institutional collaborative work. Unstructured data typically comprise about 85% of an organization's data [3], e.g., audio and video clips, body of an email, human brain images, and segmented models of anatomical structures.

Traditionally, modeling unstructured data (e.g., images) leads to certain structured metadata [4,5] (e.g., volume of an anatomical structure in the human brain) as a set of entities and attributes. The lossy nature of such a modeling scheme makes it impossible to answer questions about features that are not part of the database schema [6]. This impedes unrestricted retrieval support from arbitrary aspects of the unstructured data that characterizes a lossy data model.

It is usually very difficult, if not impossible, to enumerate all features that one can extract from an unstructured piece of data. On the other hand, there is a trade-off between the number of attributes in a data model and its feasibility, e.g., more attributes imply data entry at higher cost. Therefore, even if one could list all features, it might not be still feasible to include all those features in the data model. Therefore, feasibility is an important limitation when dealing with unstructured data.

Knowledge engineers usually design data models in consultation with experts of related fields. Since this is a subjective procedure, the data model is not theoretically reproducible. To reduce this limitation, several practical guidelines have been recommended in textbooks for E-R data modeling [7]. However, when modeling very complex and unstructured data like that of the human brain, the above guidelines are not effective [8]. This is due to the complexity of the system and diverse backgrounds of the consulting experts. In short, conventional E-R data modeling for complex systems with unstructured data presents lossyness, infeasibility, and high degrees of intra- and inter-subject variability. We have observed all these limitations in modeling the data associated with temporal lobe epilepsy.

Kirlangic et al [8] have developed a database system for objective therapy planning and evaluation in epilepsy. They have implemented this system for structuring and managing the associated data for different treatment modalities available for epilepsy. The focus of their work is the electroencephalogram (EEG). They use quantitative EEG (QEEG) measures to lessen the subjectivity of the outcome of the EEG reading. The QEEG as well as the electrode position and timing comprise the neuroprofile as a structured set of

*This work is partially supported by NIH Grant R01 EB002450.

possible quantitative measures managed in their database system. Barb et al [6], have studied the well established approaches to content management and image retrieval. They have concluded that most of these approaches lack the flexibility of sharing both explicit and tacit knowledge involved in the decision-making processes. They propose a framework using semantic methods to describe visual abnormalities, offering a solution for tacit knowledge elicitation and exchange in the medical domain. To find related functional neuroimaging experiments, Nielsen et al [9] propose a content-based image retrieval technique. Although frameworks and approaches proposed in the above literature and elsewhere [10-11] contribute a lot to the field of decision support systems in medicine and biomedical databases, they do not directly tackle the problem of unstructured image data modeling and content-based data management.

In this paper, we propose a data model for temporal lobe epilepsy that excludes the content-dependent structured metadata from the process of E-R data modeling. The rationale for this is that the content-dependent structured metadata modeling can only be manifested through a countless number of attributes, e.g., the following features of an anatomical structure in the brain (i.e., hippocampus): volume, surface, curvature, standard deviation of curvature, average intensity, standard deviation of intensities, etc. On the other hand, there is usually a limited number of unstructured metadata pertaining to a piece of unstructured raw data, e.g., limited number of structures in the human brain. The latter is true with content-independent structured metadata as well, e.g., voxel-size of an imaging study. Coupled with a method to navigate through unstructured data and a set of information extraction and fusion procedures, this scheme provides a content-based support environment. Through its query module, this environment engages appropriate feature extraction procedures confined within the brain's structures of interest to retrieve information regarding any arbitrary aspect of the data. This can be indefinitely expanded and, therefore, provide an unlimited query support. Note that with increasing costs of storage and processing power, the storage of the entire raw data and their analyses "on-demand" becomes increasingly feasible.

2. Method

We have coined the phrase "Content-Based Support Environment (C-BASE)" for systems built upon databases with unparseable and unstructured raw data, which support these features:

1. Navigation through the raw data
2. Segmentation of raw data into meaningful objects and episodes
3. Fusion of several modalities of raw data

Switching between different modalities of data while focusing the feature extractors within the object(s) of interest potentially yields descriptive, indicative, and distinctive features of the raw data. Such an environment eliminates the need for the modeling of the content-independent structured metadata, which makes the entire procedure of E-R data modeling more objective and robust. In a nutshell, this summarizes our proposed approach to the problem stated in the previous Section. In the following sections, details of the proposed content-based support environment system for temporal lobe epilepsy will be described.

2.1. Content-based support environment for temporal lobe epilepsy

To make the contents of the image data accessible to future arbitrary queries, we propose a multimodality image database, which manages the raw data, supported by navigational guidance, segmentation, data fusion (registration), and feature extraction modules. In our specific application, the knowledge-based anatomical landmark localization (K-BALL) method provides the required navigation to browse through unstructured and unparseable image data [12]. Segmentation and registration modules also benefit from the navigational guidance as K-BALL provides the initial model for segmentation and it guides the registration process. The navigation feature is the most important requirement of C-BASE. These modules and their interactions are shown in Fig. 1 in gray. Further details about the modules presented in this figure can be found in our previous publication [13].

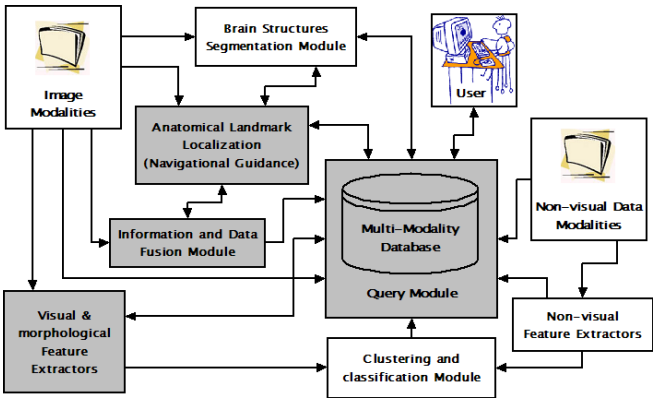

Fig. 1. Content-based support environment (C-BASE) modular architecture for temporal lobe epilepsy. Modules discussed in this paper are in gray.

2.2. E-R data modeling for temporal lobe epilepsy

In our application, the goal is to describe the unstructured raw image data so that all aspects of the data, which are of interest, can be queried and mined based on their content in the future. We distinguish between the set of metadata (ε^{C} : content-dependent metadata) that describes the contents of the raw data, and the set of metadata (ε^{NC} : content-independent metadata) that treats the raw data as black boxes and describes them regardless of their contents. Examples of the former and latter cases are the volume of the left hippocampus and the date an imaging study has been performed, respectively. We also distinguish between the sets of structured (ε_{SQL}) and

unstructured (ε_{NSQL}) metadata. We consider any data that cannot be *directly* used in a basic SQL statement as unstructured, e.g., audio and video clips, body of an email, human brain images, and segmented models of brain structures. We propose to exclude the content-dependent structured metadata in the schema. In other words, the set of metadata that are included in our data model is: $\varepsilon_{SQL}^{NC} \bigcup \varepsilon_{NSQL}^{NC} \bigcup \varepsilon_{NSQL}^{C}$. Table I shows two examples of several levels of data and metadata dealt with in this application. The ones that are included in the E-R data modeling are in gray. This table shows that we segment and store anatomical structures (e.g., hippocampus) in the database, however, the content of the segmented model (e.g., average curvature) will not be included in the database as this piece of information is content-dependent and structured (ε_{SQL}^{C}). One can simply imagine that such features are almost countless, and therefore, it is not worth including them in the database schema. Similar to the segmented models, the registration information will be part of the database schema since it is in ε_{NSQL}^{C} . An alternative method would be to store all the extracted features in an all purpose attribute (APA) as part of a table with a one-to-many relationship to the table that contains the ε_{NSQL}^{C} item. We do not discuss the former method in this paper.

Table I. Examples of data and metadata levels dealt with in the proposed E-R data modeling.

Level	Example I	Example II
Data	Image data, e.g., MRI	Image Data, e.g., MRI and SPECT
Metadata	Segmented structure, e.g., hippocampus model	Registration transformation
Meta-metadata	Features of the segmented structure, e.g., average curvature, ...	
…		

2.3. Content-based retrieval

When querying unparseable raw data stored in a database, one of the following scenarios can be followed:

1. Calculating the quantitative measures of interest using stand-alone software to produce structured data. Adding new tables and attributes to the existing database schema to keep track of the calculated structured data.
2. Integrating the feature extraction routines that calculate the quantitative measures of interest as functions or operators into the query module and make them available within the SQL code.

The advantages of the first scenario are: a) It does not need anything but the requirements of the conventional database management systems; and, b) The calculated quantitative measures will be permanently stored in the database and can be retrieved quickly in the future. The disadvantages of the first scenario are: a) The data model of the database constantly changes; b) Managing such variable data model will be difficult, as each user may add new items to query the raw data from their own standpoint, which can produce an endless number of tables and attributes; and, c) The end user of such a system needs to have a high level of expertise in database management systems to add new tables and attributes with correct relationships and within the right tables, respectively. Note that the end users are usually experts in biological and medical fields with limited knowledge of database management systems (DBMS). In addition, as it has been discussed before in great detail, this is a subjective matter and it is almost impossible to resolve disagreements between different users. On the other hand, the advantages of the second scenario are: a) All features supported by DBMS will continue to be supported by this scheme (e.g., security and privacy at the database level) since the function can encapsulate values that the user does not have the privilege to access or modify; b) All features supported by SQL will be available to the end user; c) It will have the capability to offer a unified integrated interface for query composition module; and, d) There will be no need to change the data model, eliminating the headache of managing a database with variable schema. The disadvantages of the second scenario is that it demands higher expertise in programming and more time and effort in integrating programs that calculate quantitative values into a unified system. Note that only the developers are supposed to meet this requirement and not the end users. The first scenario is more appealing if there are only a few new fields to be added throughout the life expectancy of the database. This may imply that in this situation, the data is intrinsically less unstructured and less unparseable. Therefore, as we move towards non-conventional applications that use more unstructured data, e.g., brain image databases, the advantages of the first scenario fade out and the second scenario will be more appealing.

Fig. 2 shows the scheme that we have used to implement the second scenario. The PL/SQL (procedural language/SQL) communicates with functions available in dynamic link library (DLL) files or PL/SQL functions when each function extracts a feature of interest from the raw data. PL/SQL is a procedural extension of the Oracle-SQL that offers language constructs similar to those in imperative programming languages. The DC_DLL function performs the deformable surface model segmentation (for technical details see [13]) and VOL_DLL calculates the volume of the segmented model. This scheme allows adding an unlimited number of DLL files (and functions) to extract any arbitrary feature from the raw data. In some cases, the binary images need to be passed to the DLL function as its parameter. Since the PL/SQL cannot pass the binary data, we need a mediator to retrieve the required data from the database and make it available to the DLL function through the OCI (Oracle Call Interface). When the extracted feature is in ε_{NSQL}^{C}, we store the result back into the database using the PL/SQL. The above scheme allows the users to take advantage of the functionalities available in the DLL files in

their SQL code. We have also designed and implemented a web-based query composition interface to facilitate the access and retrieval of the data.

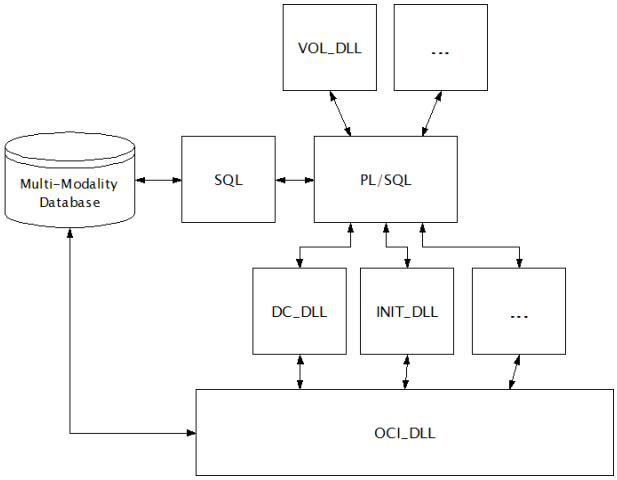

Fig. 2. The design that enables the proposed system to perform content-based query and retrieval from within the SQL shell.

3. Results

The proposed system (called human brain image database system or HBIDS) has been designed and implemented as a 3-tier web-based system, with an Oracle database as tier 1 (database server), Apache server as tier 2 (web server), and Tomcat as tier 3 (application server). The interface is implemented using a java server page (JSP) and Java applet. We have currently designed and implemented 50 forms by which users can logon into the system, insert, update, delete, and query the data. The forms of the HBIDS interface are categorized into six interface categories: 1) personal, 2) medical, 3) image, 4) base tables, 5) query composition, and 6) administration tool. Primary results of this project are reported in [13]. Here, we only present the specific results directly related to the Method Section of this paper.

3.1. E-R data modeling

The conceptual entity-relationship diagram (E-RD) of the proposed database is shown in Fig. 3. The entities and relationships of interest to this paper are in gray. An "epileptic patient" is a "person." An "epileptic patient" has "image data" and, normal and abnormal brain tissues or structures. "MRI" and "SPECT" are "image data" and "img slice" belongs to either a SPECT or MRI study. Normal and abnormal tissues may be segmented ("seg. on") on an MRI study. "MRI" and "SPECT" may be registered on ("reg. on") MRI. The "seg. on" and "reg. on" relationships have the method attribute as well as the data related to the segmented model and registration information, respectively. When implementing the schema, we created a table (IMG_DATA) allowing many image modalities to be stored for each patient with unique identifiers {MRN, Modality, DofIMG}, where MRN (Medical Record Number) uniquely describes an epileptic patient and, DofIMG and Modality refer to the date of imaging and modality of the image, respectively. The common attributes of each image modality (voxel size, resolution, etc) are kept in the IMG_DATA table. The latter pieces of data, DofIMG, voxel size, and resolution all belong to $\mathcal{E}_{SQL}^{NC}$. This table breaks down into MRI and SPECT tables. The main reason of this sub-typing comes from the fact that SPECT requires attributes which are not applicable to MRI (e.g., Time of injection begins (TofI) and completion of flush (CofF), etc, which are all in $\mathcal{E}_{SQL}^{NC}$).

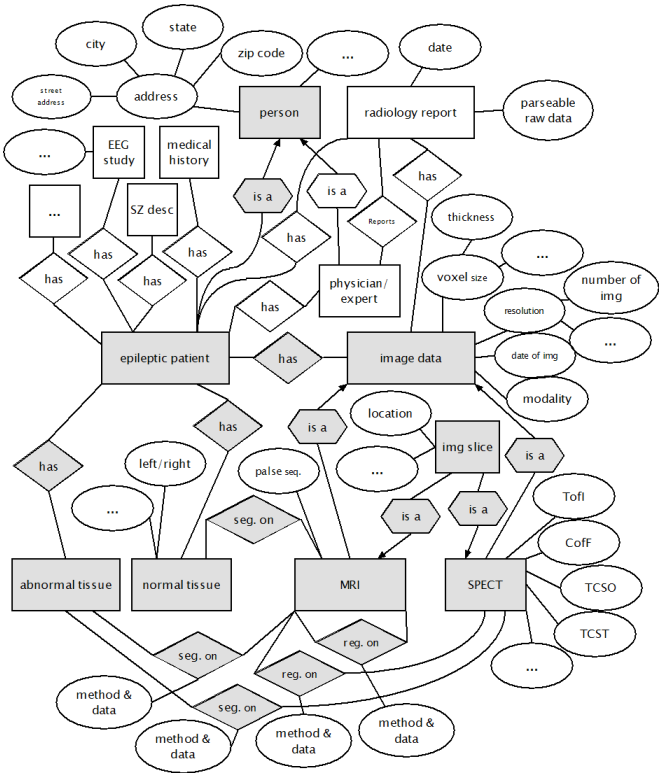

Fig. 3. Conceptual entity relational diagram describing a major part of the multimodality image database backbone.

In our implementation of the schema, the "NORMAL_TISSUES" table keeps track of the "normal tissue" entity. This table descends from the MRI table since we segment the structures of interest, e.g., hippocampus, on T1-weighted MRI. The segmented model is a content-dependent metadata; however, since it is unstructured, we accept it as an entity of the schema, "NORMAL_TISSUES $\in \mathcal{E}_{NSQL}^{C}$." The same statement is true with regard to the registration information. Therefore, we include the registration information as part of the data model in the "SPECT" and "MRI" tables. The registration information includes the base image modality information from which the segmented model is supposed to be retrieved. The "NORMAL_TISSUES" table includes foreign keys from the MRI table, the structure name, and a flag, "LeftRight," indicating whether the segmented structure belongs to the left or right hemisphere or is a shared component by both (e.g., corpus callosum). The flag is

required to distinguish symmetrically placed structures in the brain such as the hippocampus.

3.2. Navigation through unstructured data (images)

The K-BALL method has been applied for hippocampus localization. The search areas have been determined using desired and undesired statistical distributions, on a training set consisting of six epileptic patients. The T1-weighted MR images are searched for landmarks of the lateral ventricles, hippocampus, and insular cortex. Finally, using a triangulation method, an initial polygon is constructed from qualified landmarks and stored in the database as a member of $\mathcal{E}_{NSQL}^{C}$.

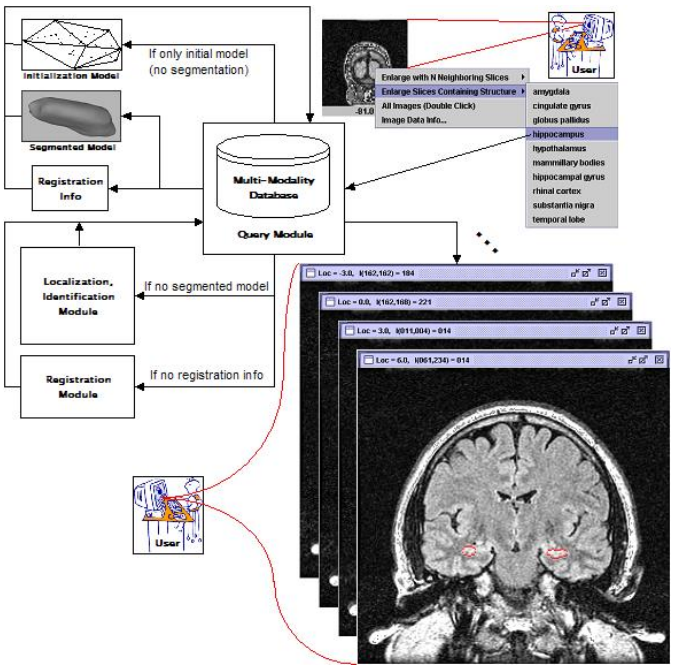

Fig. 4. Details of a navigation example through unstructured image data. Examples of downloading and displaying slices containing the hippocampus. Cross sections of the hippocampus are also displayed.

Fig. 4 shows the "behind the scene" mechanisms of navigating through image data in which the user is interested in seeing the slices containing the hippocampus in a FLAIR modality. Right-clicking on a thumbnail, the user can choose the "Enlarge Slices Containing Structure" option. The user can then specify which structure is of interest. Using the hippocampus as our example (Fig. 4), the query module looks for a model of the hippocampus segmented on the displayed image modality (i.e., FLAIR). If the model does not exist, the query module looks for the registration information. If the registration information is available as well as a segmented model for the hippocampus associated with the base modality, then the query module transfers the model to the displayed image space (FLAIR). If there is registration information but no segmented model associated with the base modality, the query module invokes the navigation module on the base modality to determine the initial model determined for the hippocampus. Since the intention is simply to navigate through image space without establishing quantitative measures, the query module does not invoke the deformable model to segment the hippocampus at this point. The initial model becomes available within a very short interval (i.e., 1-2 s for a dataset of 128 images of 256×256 pixels on a PC with Pentium III CPU @ 800 MHz). The registration information is then used to transfer the model to the FLAIR image space. Note that we prefer to perform initialization and segmentation of the hippocampus on T1-weighted MRI since the anatomical boundaries of this structure are best presented by this modality. If the registration information is not available, the query module invokes the mutual information (registration) routine. As soon as the segmented model or the initial model becomes available in the FLAIR image space, the extent of the model, consisting of vertices with maximum and minimum locations (i.e., *z*-values), are determined. Using the above locations, the slices containing the hippocampus are retrieved and displayed as shown in Fig. 4 on the lower-right side.

3.3. Content-based retrieval

As an example of content-based retrieval, we retrieve two groups of patients: 1) Patients with a pre-surgical hippocampus ipsilateral to the operated hemisphere with volumes smaller than that of the other side and with successful surgery outcomes. 2) All patients with successful outcome who have had a hippocampal resection. We wish to calculate the ratio of the number of patients in the first group divided by that of the second group. This provides some insight as to the extent that the volume of the hippocampus is predictive of successful outcome. Surgeries with postoperative Engel class I outcome are considered successful. The above ratio partially shows the sensitivity of basing lateralization on hippocampal volumes.

The "behind the scene" workflow of the above query is shown in Fig. 5. The volumeCalc function within PL/SQL checks to verify whether the segmented model exists. If the model is available, then the VOL_DLL is invoked and the model is passed to it for volume calculation. If the hippocampus is not segmented but is initialized, the initialization will be passed to DC_DLL to segment this structure. If neither segmentation nor initialization of this structure is available, K-BALL followed by DC_DLL will be evoked to compute the segmented model. As soon as the segmented model becomes available, VOL_DLL calculates the volume. We pass the required parameters (i.e., `MRN`, `MODALITY`, `DOFIMG`, `STRUCT_NAME`, `LEFT_RIGHT`) to DC_DLL and it retrieves the images and initial segmentation from the database. The same parameters are passed to INIT_DLL as well.

We may consider a Gaussian probability density function (pdf) for the hippocampal volumes and compute the pdf of the left to right volumes ratios. Fig. 6 illustrates the pdfs of the ratios of the left to right volumes of the hippocampus for patients with successful outcomes with left (solid line) and right (dash line) hippocampal resections, respectively. One can now estimate the best decision threshold for lateralization

purposes. This can be done by minimizing the following error function:

$$\varepsilon(threshold) = \int_{threshold}^{\infty} pdf_{left\_lobectomy} dx + \int_{-\infty}^{threshold} pdf_{right\_lobectomy} dx$$

The dashdot curve in Fig. 6 shows the above error function, where it becomes a minimum at $x_{\min} = 0.975,\ \varepsilon_{\min} = 0.4478$.

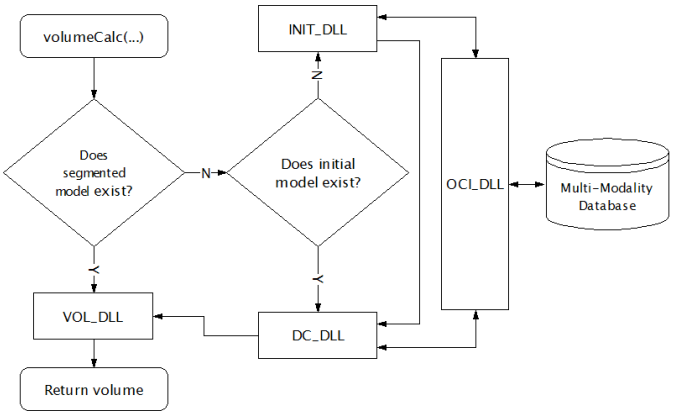

Fig. 5. "Behind the scene" workflow of the volume query.

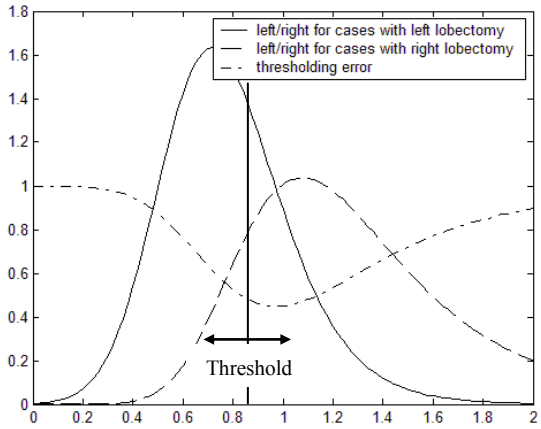

Fig. 6. Probability density functions estimated for ratios of the left to right hippocampal volumes for patients with successful outcomes with left (solid line) and right (dash line) hippocampal resections, respectively. Dash-dot curve shows $\varepsilon(threshold)$.

4. Conclusions

We have proposed a methodological framework for unstructured data management by introducing the notion of content-dependent (ε^{C}), content-independent (ε^{NC}), structured (ε_{SQL}), and unstructured (ε_{NSQL}) metadata and their inclusion or exclusion in the process of E-R data modeling. Excluding ε_{SQL}^{C} from the data modeling procedure reduces the limitations of the conventional E-R data modeling for unstructured data, i.e., lossyness, infeasibility, and irreproducibility. Married with navigation, segmentation, fusion, and feature extraction methods, the proposed data management scheme provides a content-based support environment (C-BASE). This environment facilitates unlimited query support. Switching between different modalities of data, while confined within the object(s) of interest, it yields anatomically specific query results. All these come at the cost of high computation and large memory space as well as expertise to integrate navigation, fusion, segmentation, and feature extraction routines into the proposed system during the development phase. The HBIDS, with all the features it offers, can potentially lessen the need for invasive procedures (e.g., phase II study) that are involved in temporal lobe epilepsy surgery candidacy determination.

References

[1] J. Caverlee, L. Liu, "QA-Pagelet: Data Preparation Techniques for Large-Scale Data Analysis of the Deep Web," IEEE Trans. on Knowledge and Data Engineering, vol. 17, no. 9, 2005.

[2] http://en.wikipedia.org/wiki/Data_modeling.

[3] D. Robb, "Getting the Bigger Picture: Dealing with Unstructured Data," in Storage features (http://www.enterpriseitplanet.com/storage/features/article.php/3407161), 2004.

[4] D. Marco, Building and Managing the Meta Data Repository: A Full Lifecycle Guide: Wiley, 2000.

[5] T. Eiter, J. J. Lu, T. Lukasiewicz, and V. S. Subrahmanian, "Probabilistic Object Bases," ACM Transactions on Database Systems, vol. 26, pp. 264-312, 2001.

[6] A. S. Barb, C.-R. Shyu, and Y. Sethi, "Knowledge Representation and Sharing Using Visual Semantic Modeling for Diagnostic Medical Image Databases," IEEE Trans Inf Technol Biomed, vol. 9, pp. 538-553, 2006.

[7] R. Elmasri and S. B. Navathe, Fundamentals of Database Systems, 3rd ed: Addison-Wesley Longman 2000.

[8] M. E. Kirlangic, J. Holetschek, C. Krause, and G. Ivanova, "A database for therapy evaluation in neurological disorders: application in epilepsy," IEEE Trans Inf Technol Biomed, vol. 8, pp. 321-32, 2004.

[9] F. A. Nielsen and L. K. Hansen, "Finding Related Functional Neuroimaging Volumes," Artificial Intelligence in Medicine, vol. 30, pp. 141-151, 2004.

[10] D. S. Obrosky, S. M. Edick, and M. J. Fine, "The Emergency Department Triage of Community Acquired Pneumonia Project Data and Documentation Systems: A Model for Multicenter Clinical Trials," IEEE Trans. on Information Technology in Biomedicine, vol. 10, pp. 377-384, 2006.

[11] A. G. Anagnostakis, M. Tzima, G. C. Sakellaris, D. I. Fotiadis, and A. C. Likas, "Semantics-Based Information Modeling for the Health-Care Administration Sector: The Citation Platform," IEEE Trans. on Information Technology in Biomedicine, vol. 9, pp. 239-247, 2005.

[12] M.-R. Siadat, H. Soltanian-Zadeh, K. Elisevich, F. Fotouhi, "Knowledge-Based Localization of Hippocampus in Human Brain MRI," *Journal of Computers in Biology and Medicine*, in press 2007.

[13] M.-R. Siadat, H. Soltanian-Zadeh, F. Fotouhi, and K. Elisevich, "A Content-Based Image Database System for Epilepsy," Computer Methods and Programs in Biomedicine, vol. 79, pp. 209-226, 2005.