IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4180

Web-Based Information Sourcing: Mechanism and Verification

Binshan Lin
Dept of Management and Marketing, Louisiana State University in Shreveport, Tel: (318) 797-5025

ABSTRACT

More and more approaches for representing information on the web are present, such as HTML, web database application, and XML documents. Information on the web is often very important for individuals and enterprises. However, the integration of various types of information is still a problem, since some information is structured and full of semantic, whereas others are semi-structured or document-centric. In this paper, a mechanism for heterogeneous information integration using the notion of EER conceptual data model is investigated so that the semantic and structure (i.e., data model) of not only database systems but also other web information resources can be discovered and integrated into a universal structure.

INTRODUCTION

Since the heterogeneous WWW information integration and exchange is the key point of EC, information integration on WWW is very important. It has evolved from a traditional multi-database architecture to a new framework dealing with a variety of information available in diverse formats and structures, since information on the web is diversified not only with structured but also with semi-structured data [3]. With corporate data already available in the database management systems, the requirement for connectivity between database and WWW information resources is therefore apparent. However, the integration and management of the various distributed types of information are still a problem. Various ways for representing information on the web are present. Information exchange and integration with different information standards is still a problem: Some approaches are structured and easy to be integrated, such as Active Server Pages (ASP) or other web-based database accessing technologies; some others, like extensible Markup Language (XML) documents, are semi-structured and only partial information can be integrated. There are still some other information formats, such as Hypertext Markup Language (HTML), which are document-centric. The integration of semi-structured information is harder since such resources provide virtually no information for view integration. Furthermore, semantic on each information resource on the web is unknown for external users and systems. Information integration means schema (includes information structure and its semantic) integration among information resources. Since there is less semantic declaration in web information standard, it is hard to manipulate such integration process automatically. Manual semantic mapping process by the Database Administrator (DBA) is often required. Due to the dynamics and complexity of information on the web, it is not enough for such manual operation. Hence, it is necessary to reduce the reliance on a DBA.

In this paper, a semi-automatic mechanism for web information integration using Enhanced Entity Relationship (EER) Model will be investigated, so that the structured part of different type of information can be integrated within a universal view (i.e. schema) and users can view and understand distributed web information easier.

RELATED DATABSES TECHNOLOGIES

As the popularity of heterogeneous database systems, the concept of multi-database systems have been proposed to provide a uniform environment in which users can access data from heterogeneous component databases by using a single data definition and manipulation language [9]. A multi-database system is a federation of independently developed component databases. It provides a homogenizing layer on top of these component databases to give the users a feeling of the homogeneous system. An important element of the homogenizing layer is the multi-database system schema that integrates the schemas of component databases. Schematic and data heterogeneity is the problem in building a multi-database system.

Database schema integration is therefore increasingly important in designing multi-database systems. It defines a global unified schema from several existing schemas in a distributed multi-database system. Some studies have focused on the methodology to integrate distributed or cooperated database schemas. For example, a two-phase methodology to integrate the independent, local and logical database schemas into a global database schema for a multi-database system by using EER Model is proposed in [7, 9]. Furthermore, the schema integration methodology has been verified by information capacity in [7].

Stanford's Object Exchange Model (OEM) is treated as a general model for representing database and web data structures (i.e., structured or semi-structured information). It is a self-describing way of representing meta-data on the web. It is so flexible since [8, 11]:

- It does not require the pre-definition of classes or types. Arbitrary structures with arbitrary attribute names can be included in OEM structures. This enables it to more directly represent the irregular structures found within and among Web resources;
- Not support encapsulation. Applications can directly access the OEM structures, and
- It does not support object behavior. No object methods necessary for OEM nodes.

The OEM effectively defines global models for a federated or distributed information resources, where the federated components might include structured or semi-structured information resources, such as the HTML web resources and XML web resources. There are some studies, which have focused on web information structure manipulation. For example, in [11], the authors try to discover a "typical" collection of objects in several semi-structured information resource denoted in OEM diagrams whose semantics are known (i.e., Discover structured parts of information resources whose semantic and structure are known). It is useful in some applications, such as document clustering and information index construction.

Besides HTML, XML is another way for information representation on web. It is also a kind of meta-languages. It has some advantages for information representation, such as straightforwardness in using over the internet, support a wide variety of applications, easy to write programs which process XML documents, and easy to create document, etc. As such, XML is apparently a good candidate for data exchange [6, 10].

Much research, such as [10], has focused on using XML as a common information exchange format. Some studies also indicate that its hierarchical structure and user-defined tags can be used to express structured and semi-structured data [2].

The concept of web database applications, such as Active Server Pages (ASP) or Common Gateway Interface (CGI), is popular in application servers of multi-tier environment. It is a server-side script environment for creating interactive pages may contain HTML tags, text, and script commands and can build an interactive web pages or an entire web application with an HTML user interface [4]. In our study, such applications are also considered as a kind of information resource. They should also be integrated.

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

UNIVERSAL SEMANTIC REPRESENTATION

EER Model is one kind of high-level data model, or conceptual and semantic data model. It has the following characteristics for semantic representation [5]:

- **Expressiveness:** It can distinguish different types of data, relationships, and constraints.
- **Simplicity:** Everyone can understand and use its concepts.
- **Minimization:** There must be minimal number of basic concepts that are distinct and non-overlapping in meaning.
- **Formality:** So that the model concepts can be defined accurately and unambiguously.

The EER Model is widely adapted for database semantic representation in most database software. It can express more semantic than other logical data model, such as relational data model. It is also widely used in database schema integration [9]. It is adequate for semantic representation of information resource on the web in this research with the same reason.

INTEGRATION MECHNISM

To translate the structure of source data model into destination EER structure, a three-step method is used for data model translation:

- **Entity Translation:** To translate available corresponding components in the source data model into entities in EER structure.
- **Relationship Translation:** To translate available corresponding components in the source data model into relationships with some cardinality in EER structure.
- **Attribute Translation:** To translate available corresponding components in the source data model into attributes of some entity in EER structure.

We express traditional HTML-based web information resources in OEM model, and structured information resources in EER model. It is necessary to translate the OEM model to EER model. In our method, we find the super set (i.e., union) among different expressions of one object. We use the notion of Brand-First-Search (BFS) to visit each simple path in the tree-like OEM graph and cluster edges with the same label.

The major difference between HTML-based web information and XML-based information is that there are additional schema definitions, DTD, within XML information. It is necessary to translate DTD into EER model in order to integrate two different types of information resource. However, the limitation for such translation is that there is weak notion of atomic type in DTD: Some data types lacked, such as integers or dates, which are common in traditional database schema. Only #PCDATA (i.e., Strings) atomic data type is present in DTD [1]. In our method, the data type should be translated according to the content of XML information. Figure 8 through Figure 10 show our three-rounds algorithm to translate the structure of XML information resource into EER structure. Notice that in our study, we assume that XML information on the web is valid. In other words, a DTD is present for structure definition of the XML information resource.

VERFICATION

In this section, we discuss the correctness of our mechanism by verifying the correctness of each phase. Our mechanism is correct if and only if each phase in our mechanism is correct. In Re-engineering Phase, the correctness of our data model translation methods will be proved by the concept of "*information capacity*" [12].

Configuration Phase is actually a definition and collection of user requirement and domain knowledge. In this phase, user can define:

- *Requirement for information;*
- *Domain Knowledge between Semantic and Information Characteristics, and*
- *Range of information resource.*

Such configuration meets the user's requirement, since it defined manually by the users themselves. Besides some user interactive operations, resource classification is automatically detected and done in this phase. In our study, resource classification provides a guideline to analyze different information resources with different strategies since various types of information are presented on the web.

Information Capacity equivalence and dominance [7, 12] is used as a basis for verifying the transformed schemas without information loss for schema integration and translation. There are five classifications of information capacity [12]: Functional, Injective, Total, Surjective, and Bijection Relation. In our methodologies, correctness of each translation rule will be certificated if and only if the rule is at least a total relation (i.e., an information capacity preserving mapping.) since there may be semantic loss when translating among different data models.

Integration phase is responsible for semantic prediction and schema integration. As for semantic prediction, we use the decision tree for domain knowledge expression, since it is adequate and powerful for rule definition, classification, and prediction process in Step 5. In our study, we apply and extend the three-step method discussed in [7] as our schema integration method. Correctness of this method has also already been proved in [7].

CONCLUSION

We have introduced in this paper our mechanism for heterogeneous information integration using EER Diagram. Since EER Diagram is a common conceptual data model, it is an appropriate solution for representation of different kinds of information. As such, in our mechanism, EER Diagram is applied to represent and integrate schemas of different kinds of information. Since there is a universal view of different heterogeneous information, enterprises will have a better way for information integration and standardization. Users can therefore understand and analyze information in a more convenient and broad way.

REFERENCES

Available Upon Request