Chapter 1. Introduction

1.1. Research Motivation

In the past decades, World Wide Web (WWW or Web) has changed everything, especially business computing. More and more enterprises adopt Internet/Intranet business model, and large amount of business information are exchanged over the Internet everyday. But data sources on the Web are often distributed and heterogeneous. Enterprises have to spend more time and efforts searching the data they need. Recently, electronic business (EB), enterprise information integration (EII), and enterprise application integration (EAI) become popular within the enterprise gradually. Enterprise application must interact with disparate data sources, including databases, file systems, Web pages, and other applications. Heterogeneity and interoperability become one of the key issues in enterprise information extraction and integration. A solution to this heterogeneous information integration problem is that providing a uniform access to data obtainable from different sources in EB environment.

Extensible Markup Language (XML) and XQuery have become the standard data exchange format and query language respectively. Therefore, most researches have adopted them as standard input and output to integrate heterogeneous data sources. Using XML can resolve the problem that different data sources store their data in different structures. But XML shows some limitations on the semantic heterogeneity resolution. Because XML tags are human-readable, not machine-readable, the meaning of the information interchanged cannot be understood across different systems.

Ontology provides much richer modeling means with classes and properties organized into is-a hierarchy and enriched with axioms and relations processable with inference. Using ontology for the explication of implicit and hidden knowledge is a possible approach to overcome the problem of semantic heterogeneity. However, in method, there is still no adequately systematic benchmark method for “quantitative” performance measurement and evaluation on heterogeneous information integration.

1.2. Research Problem

Information integration issue has arisen in 1980s. Today, with the popularity of Internet/Intranet, it becomes a hot information technology (IT) topic in EB field. There have been research projects applying XML and ontology as mediated techniques to consolidate heterogeneous information. In order to measure and evaluate the mechanism for heterogeneous information integration, a benchmark approach to these new techniques is needed.

There are separate benchmarks on XML, relational database, object-oriented database, and Web server. But they are independent and use a predetermined set of test database and test query. When the domain or application changes, they are not reproducible. Domain dependency is a core issue in current benchmarks. Heterogeneous information integration needs to incorporate XML and ontology. However, there is still no adequate benchmark developed for such integration in EB environment. Information integration methods can be classified into global-as-view (GAV) defining the global schema as a view over the local schemas and local-as-view (LAV) defining the local sources as views over the global schema (Manolescu, Florescu, & Kossmann, 2001). GAV approach is chosen over LAV in this research, because query on the global schema transformed into queries on the local sources is

better fit in EB environment.

1.3. Research Objective

Developing a benchmark requires the definition of a workload model. A workload is the core of a benchmark. In this research, we propose a workload model that incorporates XML and ontology in heterogeneous information integration under EB environment. It evaluates the XML processing performance of heterogeneous information integration systems. In order to capture the semantic aspect of them, the workload model is also designed to evaluate whether the ontology can represent the real meaning of heterogeneous information.

This workload model is designed to be domain independent and generic-construct-based. It is hard to apply a domain-specific benchmark to different application domains. The generic model describes the data structure and usage of the system that do not tie with a predetermined scenario. The workload model is developed with intent to meet the desirable characteristics of a good benchmark. First, the workload model can scale with the complexity of data and operation. Second, the workload model adopts open standards, such as generic constructs in relational model, object model, Web page, XML, and ontology. This makes the workload model to be portable. Third, the workload model is simple to understand and implement because of generation process is automated.

1.4. Research Flow

In this research, we study heterogeneous information integration, XML, XQuery, and ontology. Benchmarks on XML and ontology are reviewed simultaneously. After

that, we analyze the XML-specific and ontology-specific requirements systematically in order to stimulate a new design. The workload model development consists of three components: data model, operation model, and control model. The data model depicts the structures of data to be tested. The operation model describes the generic functionalities provided by heterogeneous information integration. The control model defines the variables to setup the test and experiment environment. We implement a prototype on the basis of the research workload model to prove the feasibility.

The research flow of this research is shown in Figure 1.1 as follows.

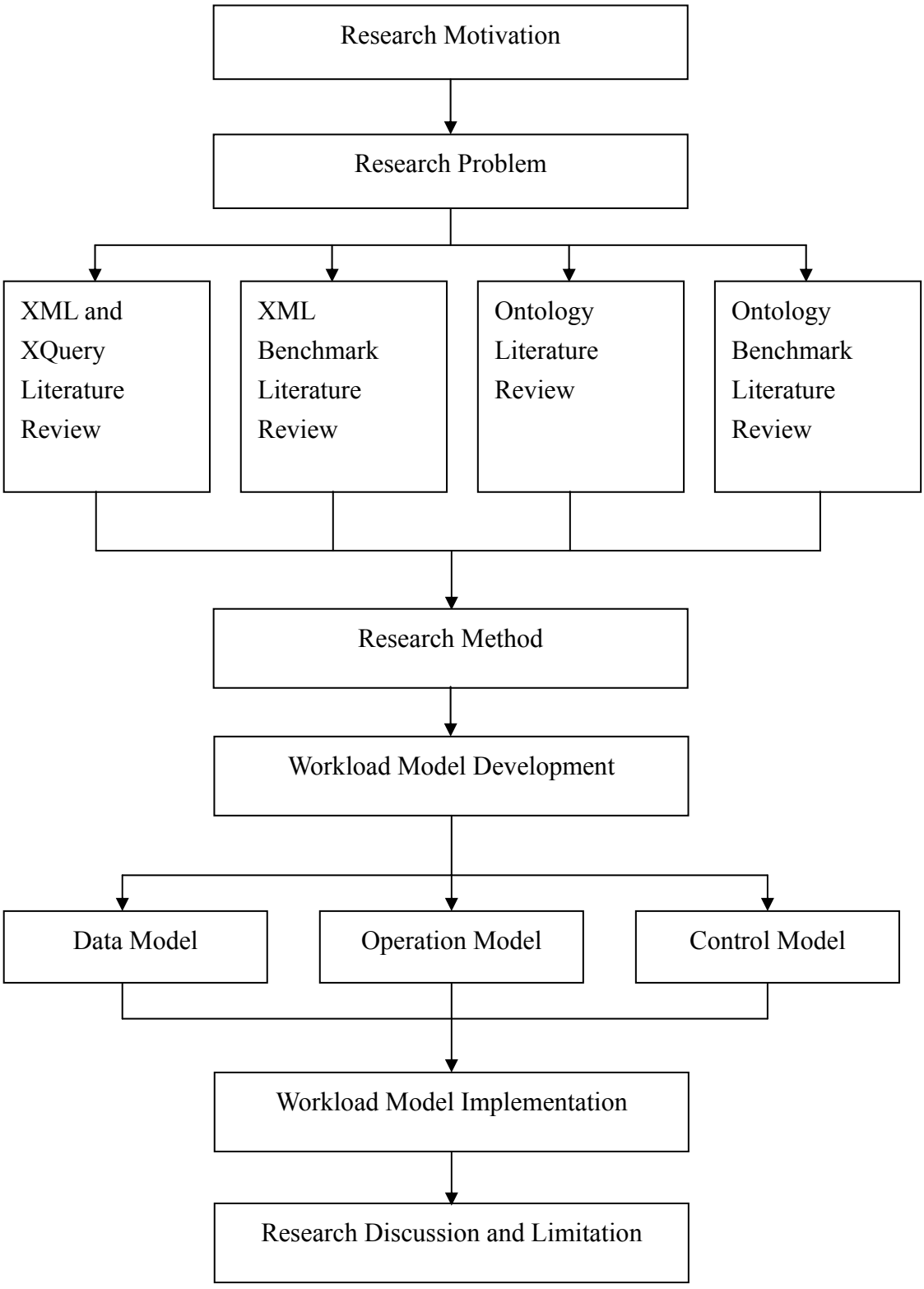

Figure 1.1: Research Flow

1.5. Organization of Thesis

The remainder of the thesis is organized as follows:

- In chapter two, we review and discuss related XML and ontology benchmark works.
- Chapter three presents the model of this research on benchmarking the heterogeneous information integration systems.
- In chapter four, we implement the research model of this research on a heterogeneous information integration prototype platform.
- Chapter five discusses the research results and limitations.
- Chapter six concludes with a summary and possible extension of this research.