메타데이터 레지스트리를 위한 메타데이터 교환 프로토콜의 질의 효율성 평가

정동원*

Query Efficiency Evaluation of the Metadata Exchanging Protocol for Metadata Registries

Dongwon Jeong

Abstract

이 논문에서는 메타데이터 레지스트리간의 메타데이터 교환을 위해 제안된 교환 프로토콜의 장점을 명시적으로 보이기 위한 시뮬레이션 결과에 대하여 기술한다. 기존 접근 방법들은 교환 메커니즘의 높은 복잡도, 지역 메타데이터 관리 시스템으로의 종속성, 새로운 메타데이터 관리 시스템의 추가를 위한 높은 비용 등의 문제점을 지닌다. 이를 해결하기 메타데이터 교환 프로토콜이 제안되었다. 그러나 지금까지의 연구에서는 정성적으로만 그 장점을 보였다. 이 논문의 목적은 정략적 관점에서 제안된 프로토콜의 장점을 기술하고자 함에 있다. 특히 이 논문에서는 질의 효율이 가장 중요한 문제로서 다른 요인에 의한 결과에도 영향을 주기 때문에 질의 효율성 문제에 초점을 둔다. 이를 위해 평가 항목과 평가 모델에 대해 정의한다. 질의 효율성 평가를 위한 요인 중에서 대상 메타데이터의 개수, 각 메타데이터를 구성하는 컴포넌트의 개수 및 각 컴포넌트들을 구성하는 속성들의 개수를 주 파라미터로 이용하였다. 이에 따라 시뮬레이션은 크게 세 가지 유형을 지니게 된다. 정량적 평가를 통해 교환 프로토콜이 요구되는 이유와 그 정당성을 명시적으로 보인다.

Key Words: Exchanging protocol, Metadata registry, Query efficiency, Metadata exchaging

* 군산대학교

1. Introduction

A metadata exchanging protocol was developed to enhance the interoperability between metadata registries [1,2,3]. The metadata registry is one of the core components of ISO/IEC 11179, the international standard for high quality metadata exchanging and sharing [4,5,6]. This standard provides many merits for the interoperability. Therefore, in many/various application fields, metadata registries have been built to achieve the interoperability in their application domain [7,8,9,10,11,12,13,14]. These metadata registries are managing by each metadata registry management system. Most of the systems has been developed based on the Web.

The existing metadata registries have been developed for each application domain. Thus, the interoperability issue between the metadata registries should be solved to achieve the global data exchange and sharing. To do this issue, one-to-one mapping must be done because this standard provides no exchanging protocol (ex., like SQL for the relational databases). This one-to-one mapping approach, called adapter-based approach in this paper, has several problems as follow: (1)Complexity of distributed querying; (2)High query modeling cost; (3)Dependency of query description; (4)Complicated mechanism and high cost for exchanging.

A metadata exchanging protocol was developed to solve these issues [1]. This protocol is the SQL-based query language for consistent access of a variety of metadata registries. It provides a standardized access method to exchange and share metadata between distributed metadata registries. In a word, this approach has many advantages in aspects of query modeling simplicity, independency of query description, lower query modeling cost, lower distributed query processing, consistency of access method, etc.

However, In the research [1], the paper just showed its merits in some qualitative views not quantitative views, so its good points do not be identified. In this paper, simulations are illustrated to show the effectiveness of the metadata exchanging protocol-based approach. First, this paper presents a set of notations and symbols for description.

Furthermore, several simulation models are shown to compare the metadata exchanging protocol-based approach and the previous approach. Finally, this paper shows and discusses the evaluation results using the defined simulation models.

Additionally, this paper just focuses on the query efficiency issue. The query efficiency is one of the most important factors that determine the performance of metadata exchanging systems. this paper will not introduce the details of the concept of the metadata registry and the metadata exchanging protocol because of space restriction. However, you can see the details through the given references including the international standard ISO/IEC11179.

2. Simulation Model

Comparative items and evaluation models are defined to quantitatively prove the advantages of the metadata exchanging protocol. This section describes preliminary constraints, notations, and symbols including the items and models.

2.1 Definition of Comparative Items

Table 1 shows the parameters and key notations for our simulation. In the simulation, the first major factor *nMDR* is the number metadata registries to exchange metadata. Each metadata registry consists of components such as data element, data element concept, object class, etc [4]. Also, each component is composed of many attributes to describe its properties. For example, the data element has many attributes such as name, description, definition, max_length, data type, etc. Therefore, *nCom*, the number of the component of the metadata registry and *nAttrib*, the number of the attributes of each component are key simulation factors. In our simulation, we assume every metadata registry has the same number of the components and also every component

A user query for retrieving metadata from several metadata registries has logically semantic-abstracted components and attributes. They are mapped to actual components and attributes in local metadata registries. The written user query has *nSemCom* and *nSemAttrib* (the number of semantic components and the number semantic attributes respectively). The semantic component and semantic attribute is mapped to one or more actual components and attributes (*nActCom* and *nActAttrib*) in the local metadata registries.

The analysis time θ to find proper actual components and attributes is depend on many factors and is different according to situations. We assume the time is uniform and this paper uses the value created by a random number generator.

Finally, we must consider the selection time to choose appropriate actual components and attributes through the analyzing. The time is depend on the

Table 1. Summary of notations and symbols

Notations & Symbols	Description
nMDR	The number of metadata registries
nCom	The number of components of which a metadata regisry consists
nAttrib	The number of attributes per each component
nSemCom	The number of semantic components for the user query writing
nSemAttrib	The number of semantic attributes for the user query writing
nActCom	Actual components in local metadata registries
nActAttrib	Actual attributes in local metadata registries
θ	Analysis time semantically
SelCom	Time for selecting proper components
SelAttrib	Time for selecting proper attributes

number of actual components and attributes.

This paper handles two comparative items which are defined to explicitly show the metadata exchanging protocol's advantages. The items include query modeling cost and processor development cost (system integrating cost). Because the predominance of the remainder has been qualitatively shown with query examples in [1], this paper focuses on the items. In other words, the remainder can be recognized in a logical and conceptual aspect, but the items directly affect the performance of the physical development.

2.2 Simulation Model for Query Efficiency

In the previous approach, as many queries as the number of metadata registries have to be written because metadata registries respectively use different structures and there is no standardized method to exchange information between them. However, in the case of using the metadata exchanging protocol, only one query statement is necessary to accomplish it.

The most similar approach in other fields is the standard query language for the relational databases SQL. Every relational database use the query language as the standard protocol. Therefore, we can access all databases in a unified manner. The proposed protocol is very similar to the query language for relational databases. The proposed protocol has been developed to provide a standardized access method for various metadata registries

Fig. 1 illustrates the models to evaluate the query modeling cost. For the simulation according to the models, several assumptions are needed and predefined in section 2.1.

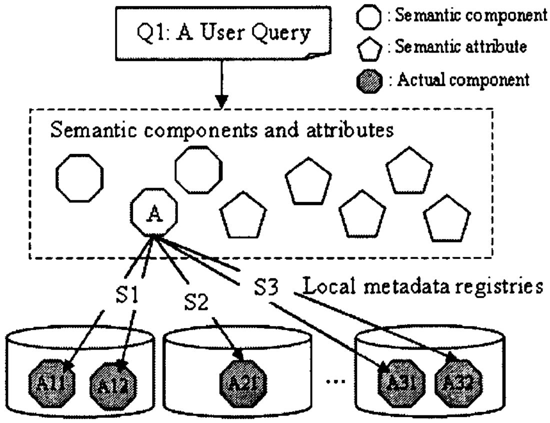

Fig. 1. Evaluation model for comparing query modeling cost

In the previous approach, there is no unified method to access metadata registries. Thus, query statements as many as metadata registries are required to generate an integrated result. In Fig. 1, the previous approach needs three query statements, *S1, S2, S3* for the metadata registries.

By contrast, the protocol-based approach requires only one query statement to gather data from the metadata registries. Therefore, the modeling cost of this approach is independent of the number of metadata registries.

In the previous approach, one semantic component can be mapped to one or more actual components. For one semantic component, we can get appropriate components by analyzing every actual component. Therefore, θ_{PRE}, the calculation formula for estimating the query modeling cost of the previous approach, is as follows:

$$\Theta_{PRE} = nMDR \times \left(\sum_{i=1}^{nSemCom} \left(T_i + \sum_{k=1}^{nActCom} T_{ik} \right) + \sum_{j=1}^{nSemAttrib} \left(T_j + \sum_{m=1}^{nActAttrib} T_{jm} \right) \right)$$

By contrast, for the metadata exchanging protocol-based approach, only one query statement is written to gather data from the metadata registries. Therefore, the modeling cost of this approach is independent of the number of metadata registries and its formula Θ is as follows:

$$\Theta = \sum_{i=1}^{nSemCom} \left(T_i + \sum_{k=1}^{nActCom} T_{ik} \right) + \sum_{j=1}^{nSemAttrib} \left(T_j + \sum_{m=1}^{nActAttrib} T_{jm} \right)$$

3. Simulation Results

In this section, the simulation results are described based on the simulation models referred in the previous section.

3.1 Model Setting and Evaluation Factors

In section 2.1, we assume several factors are uniform. This paper consider three factors to simulate the proposed approach. That is, we use the three factors as simulation parameters. The parameters include the number of metadata registries, the number of components, and the number of attributes. The query writing time (modeling time) depends on the number of metadata registries, so this parameter is variable.

The schema structures of metadata registries can be diversified according to design, but the number of components must be unified because the metadata registry is a part of the ISO/IEC 11179 standard. As referred, we assume that *nCom* and *nAttrib* are uniform.

Different user queries are written with different components and attributes. Sometimes, one same query might be written with different components and attributes. The number of components and attributes (exactly semantic components and attributes) is the key factor for simulation. Hence, *nSemCom* and *nSemAttrib* are variable.

The simulation on the query modeling cost has three types according to the evaluation factors as follow: (1)The first type is that : the number of MDRs gradually increases; The second type is that the number of semantic components to be written increases in steps; (3) The final type is that the number of semantic attributes to be described increases by degrees.

3.2 Simulation result considering *nMDR*

Fig. 2 shows the evaluation result considering the number of metadata registries.. In this case, a variable factor is the number of metadata registries. The value domain of *nMDR* is {1, 2, 3, 4, 5}. The other factors have the fixed value. As shown in this result, the metadata exchanging protocol-based approach is proved to be more efficient against the previous approach. In case the number of metadata registry is one, then the results of both of them are similar. However, as the number of metadata registries increases, the difference between them also increases. *nMDR* is 2, then the difference is about 2. When the number of metadata registries is 5, New : Previous = 21.05 : 112.69 (i.e., the difference is about 1:5)

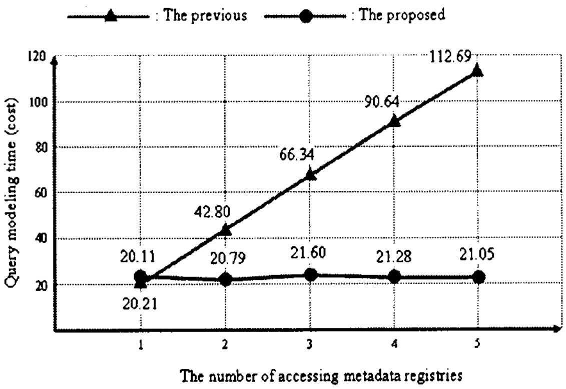

Fig. 2. Simulation result with the number of metadata registries

3.3 The Result of Type II

Type II is about the simulation with the number of components which are included in a metadata registry. In this simulation, the variable factor is the number of semantic components. Fig. 3 illustrates the simulation result considering the number of components. Its value domain for the simulation is {1, 2, 3, 4, 5}. As the number increases, the modeling cost difference between the two approaches continues to grow up. When the number of components is 2, the difference between the two is 27.17 (44.74-17.57). The number of components is 5, then the difference between the approaches is about 40 and reaches to about two times of the modeling time that the metadata exchanging protocol-based approach requires in the same situation.

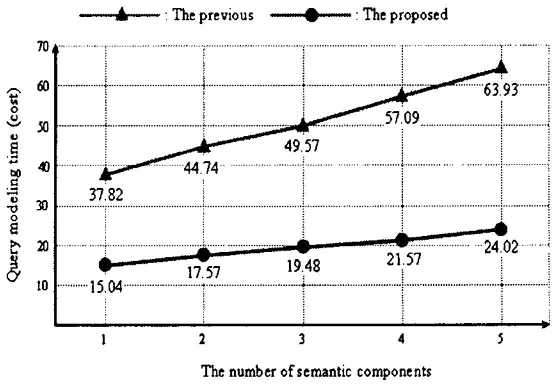

Fig. 3. Simulation result according to the number of components

3.4 The Simulation Result considering The Number of Attributes

In the final simulation, the variable factor is the number of semantic attributes. The parameter values include {1, 3, 5, 7, 9, 10}. Fig. 4 shows the simulation result considering the semantic attributes. In this result, as the number of semantic attributes increases, the modeling time difference between the two approaches grows up.

As we can see through the three simulation results in Figure 2, 3, and 4, in the aspect of the query modeling, the previous approach requires much more time than the metadata exchanging protocol-based approach.

Consequently, it can be explicitly seen that the metadata exchanging protocol-based approach is more efficient than the previous approach.

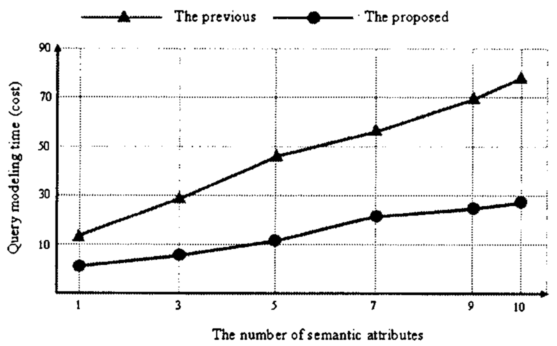

Fig. 4. Simulation result considering the semantic attributes

4. Conclusion

This paper illustrated the evaluation results to explicitly show the advantages of the metadata exchanging protocol. It is referred in [11] that the protocol is superior to the existing approach in the following aspects: simplicity of query modeling, simplicity of the exchanging mechanism, ease of use, independent description for distributed querying, low cost of system development, and so on. Most of advantages were qualitatively shown with query examples. However, quantitative evaluations are required to show its merits completely and definitely.

This paper aims at tackling with the issue and defining two comparative items: query modeling cost and system development cost. Evaluation models are defined and calculation models for them are proposed. The metadata exchanging protocol-approach is more efficient than the previous approach in all of the evaluations. Consequently, this paper clearly proved that the metadata exchanging protocol is an effective sharing method between metadata registries.

In this paper, we assumed that analysis time is uniform. However, the analysis time is different depending on the human ability. Further work on the experiment needs to consider the human factor.

References

[1] D. Jeong, Y.-G. Kim, and D.-K. Baik, "SQL/MDR : Query Language for Consistent Access of Metadata Registries based on SQL3," Springer Verlag, Lecture Notes in Computer Science (LNCS), Vol. LNCS 3129, pp. 239-248, July 2004.

[2] D. Shin, D. Jeong, Y.-G. Kim, S.-H. Park, and D.-K. Baik, "A Query Language for Consistent Access of Metadata Registries," Korea Information Science Society, Journal of KISS: Database, Vol. 31, No. 6, pp. 609-623, December 2004.

[3] D. Jeong and D.-K. Baik, "Incremental data integration based on hierarchical metadata registry with data visibility," Elsevier Science, Information Sciences, Vol. 162, pp. 147-181, June 2004.

[4] ISO/IEC JTC 1/SC 32, "ISO/IEC 11179: Information Technology–Metadata registries–Part 1~6," 2003.

[5] D. Jeong, S.-H. Park, and D.-K. Baik, "A Practical Approach: Localization-based Global Metadata Regis-try for Progressive Data Integration," World Scientific, Journal of Information & Knowledge Management, Vol. 2, No. 4, pp. 391-402, December 2003.

[6] Environmental Protection Agency (EPA), "Environmental Data Registry," 2004.

[7] Environmental Protection Agency, "Data Standards Publications and Guidances," 2003, http://oaspub.epa.gov/edr/.
[8] Australian Institute of Health and Welfare, "Australian National Health Information Knowledgebase," 2004, http://www.aihw.gov.au/.
[9] Australian National Health Data Committee, "National Health Data Dictionary," 2003.
[10] U.S. Department of Transportation, "U.S. Intelligent Transportation System," 2004, http://www.dot.gov/.
[11] ITS Architecture Development Team, "ITS Logical Architecture- Vol. I, Vol. II: Process Specifications. Vol. III: Data Dictionary," 2003, http://www.itsa.org/.
[12] Korea Institute of Science and Technology Information, "A study on the development of standardization and management model for science and technology information," Korea Institute of Science and Technology Information, Research Report, 2002.
[13] Korea Institute of Science and Technology Information, "A study on the standardization of metadata management for science and technology information," Korea Institute of Science and Technology Information, Research Report, October 2003.
[14] Electronics and Telecommunication Research Institute, "Research on the Registration and Search System of Component," Electronics and Telecommunication Research Institute, Research Report, 2000.
[15] D. Jeong, D. Shin, E. Jeong, Y.-G. Kim, J. Lee, and D.-K. Baik, "Metadata Registry Management System for Science and Technology Information," Korea Information Science Society, Journal of KISS: Computing Practices, Vol. 10, No. 6, pp. 472-483, December 2004.

주 작 성 자 : 정 동 원
논문투고일 : 2005. 04. 26
논문심사일 : 2005. 05. 03(1차), 2005. 05. 09(2차), 2005. 06. 14(3차)
심사판정일 : 2005. 06. 14

● 저자소개 ●

Dongwon Jeong (정동원)

1997 Kunsan National University, Computer Science, B.S.

1999 Chungbuk National University, Computer Science, M.S.

2004 Korea University, Computer Science, Ph.D.

1999~2000. Full-time Instructor, Advanced Institute of Information Technology, Korea.

2000~2001. Senior Researcher, Jigunet Corporation, Korea.

2002~2005. Researcher, Lime Media Technologies Co. Ltd., Korea.

2004. Research Assistant Professor, Research Institute of Information and Communication Technology, Korea University.

2002~Present. Committee Member, Telecommunications Technology Association, Korea.

2005~Present. Program Committee, SERA2005.

2005~Present. Professor, Dept. of Informatics & Statistics, Kunsan National University.

Research Interests: The metadata registry-based integration and application, mobile agent security, XML-to-RDB conversion and hybrid security, ubiquitous computing & semantic maintenance, and ontology conversion.