Heterogeneous Product Description in Electronic Commerce

Wee Keong Ng*
and
Guanghao Yan†
and
Ee-Peng Lim*

The heterogeneity of product descriptions is an impediment to successful interactions among buyers and vendors in electronic commerce. There is no uniform description even for the same product type among different vendors. In electronic commerce activities involving interactions among different vendors (business-to-business) or between a buyer and vendors (consumer-to-business), a common ontology for products is critical. There are two approaches to resolve the information heterogeneity problem in electronic commerce: standardization and integration. This short paper presents an overview of these approaches and shows that even with standardization efforts, some form of integration would still be required due to the potential multiplicity of standards and their degree of acceptance.

Additional Key Words and Phrases: schema integration, information exchange, electronic commerce

1. INTRODUCTION

It has been pointed out by a survey on strategic directions in electronic commerce [1] that electronic commerce encompasses many issues such as acquiring and storing information, finding and filtering information, securing information, auditing access, cost management and financial instruments, and so on. Among these issues, *finding and filtering information* is of essential importance to a successful electronic system where consumers need online facilities to help them retrieve information and locate resources that match their expectations and desires. Specifically, consumers would like to find products and services at low costs using languages and termi-

Address:
*School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore
†Computer Science Department, State University of New York at Stony Brook, Stony Brook, NY 11794-4400, USA

nologies they are familiar.

However, the rich and diverse descriptions that vendors use to describe their products increases the difficulty of locating products and services accurately and efficiently. Although many Web search engines exist, it is still difficult for human users to locate a vendor selling a certain product or to compare among different vendors as most search engines are keyword-based. In fact, one of the biggest challenges for electronic commerce today is to create mechanisms to allow buyers to locate products and services with specific characteristics and to allow vendors to locate potential buyers with specific traits [1].

The heterogeneity of product information is a critical impediment to efficient business information exchange. There is no uniform description for each product type among vendors. In electronic commerce activities involving interactions among different vendors (business-to-business model) or between one buyer and multiple vendors (consumer-to-business model), a common ontology for the products is critical.

There are two general approaches to resolve the problem of information heterogeneity: standardization and integration. In the standardization approach, a common vocabulary and common protocol are drafted to be unanimously supported and adopted by all parties involved in a business exchange. This is a common approach seen in the industry. In the integration approach, mappings are found semantic components so that differences can be resolved. As we shall see, a proliferation of standards could bring us back to square one in the quest to resolve the information heterogeneity problem. In the next two sections, we introduce the standardization and integration approaches respectively.

2. STANDARDIZATION EFFORTS IN ELECTRONIC COMMERCE

In the electronic commerce industry, solutions for electronic commerce activities are generally based on standards. Usually there are two prerequisites for efficient electronic commerce information exchange; a common vocabulary and a common protocol. A protocol defines the rules of information exchange between parties engaged in an electronic commerce activity. Much effort has been expended to provide related standards for these two factors [2; 5; 6; 9; 10; 11]. For example, `Ontology.org` is an organization devoted to developing industry specific XML DTDs and thus to solve the vocabulary problem. The ICE (Information and Context Exchange) protocol [11] provides a solution for the protocol problem by managing and automating the establishment of syndication relationships, data transfer, and results analysis. In addition, the eCo Framework Project [6] by CommerceNet has also addressed some of the heterogeneity issues. They created a base set of common terms and mappings among existing terms for electronic commerce specifications. The eCo working group considers a list of related specifications among which the RosettaNet Specification [9] and the Common Business Library (CBL) [2] will be briefly discussed next.

2.1 UN/SPSC

The UN/SPSC (United Nations Standard Product and Services Codes) [5; 10] is an open, global standard that provides a logical framework for classifying products and services of all kinds throughout the world. It is structured as a five-level hierarchy.

At each level, a two-character numerical value classifies each item more specifically. For example, 'leased addressing machines' are assigned UN/SPSC number 44-10-21-05-10. The first two-character number '44' means 'Office equipment, accessories and supplies'; the second two-character number '10' means 'Office machines, and their supplies and accessories'; the third two-character number '21' means 'Mail machines'; the forth two-character number '05' means 'Addressing machines'; the last two-character number '10' means 'Addressing machines, leased'. Any modification of these five two-character numbers results in a different product.

The UN/SPSC is an invaluable tool for doing business globally although it has not addressed product attribute issues. Its hierarchical structure ensures that a company finds a meaningful level of product analysis conveniently. Its unique coding scheme makes it suitable for multi-language uses. UN/SPSC is used widely in business, especially in electronic commerce system. For example, Commerce One's [4] Commerce Chain Solution [3] and Ariba.com Network have adopted it in their work on product content management.

2.2 RosettaNet Specification

RosettaNet [9] creates 'property' definitions for various entities in electronic commerce, such as property definitions for a certain product and its properties. For example, 'Modem' is a property (or an attribute) for computers (Figure 1). Once these property definitions are completed, they will be distributed to some standards maintenance organizations that will enumerate possible values for those properties. Then, property definitions as well as their values are distributed to companies in the industry supply chain as standards for business information format, say for product descriptions.

Let us take a look at how one property of a product is defined by RosettaNet. We consider the definition of the 'Central Processor Unit' property for a laptop given by the RosettaNet Laptop Technical Specification. There are several fields for the property 'Central Processor Unit': 'Property Name', 'Synonym', 'Property Definition', 'Dictionary References', 'Where Used', 'Property Type' and so on. Moreover, some of these fields contain sub-fields. For example, 'Property Name' has 'Abbreviation' and 'Acronym' as its sub-fields. All these fields serve as metadata for the product property (attribute).

2.3 Common Business Library (CBL)

The Common Business Library (CBL) by Veo Systems is a set of building blocks with common semantics and syntax to ensure interoperability among XML applications. CBL consists of information models for generic business concepts including:

—business description primitives like companies, services, and products;
—business forms like catalogs, purchase orders, and invoices;
—standard measurements, date and time, locations, classification codes.

CBL consists of an extensible, public set of XML DTDs and modules. These building blocks can be assembled to create complete XML documents representing a business interaction such as a purchase order or an inventory stock query. Where possible, CBL takes advantage of other standards using, for example, relevant ISO standards for dates, currencies, and names. CBL is closely related to the work of

RosettaNet, and the property definitions given by RosettaNet can be referenced by CBL to compose DTDs and modules for various electronic commerce transactions, including product descriptions.

To use CBL, an organization starts by creating a CBL document describing its offers and services. Then, it integrates a CBL system with its back-end system by writing custom code that interprets information between the CBL format and the organization's previous format. It is like building a 'wrapper' for back-end systems by using CBL blocks. After that, organizations interact on the basis of CBL semantics and syntax.

2.4 Summary

Most industrial solutions for product information heterogeneity are based on standardization. As with most standards, it will be some time before electronic commerce standards are widely used. Currently, standards for electronic commerce transactions are far from mature. In addition, it is expected that there would be a multiplicity of standards in the future given the concurrent efforts among different organizations. Hence, it is conceivable that some form of integration would still be required for the various standards.

3. THE INTEGRATION APPROACH

In this section, we introduce the integration problem and identify characteristics and issues of the problem. Some preliminary definitions are in order: We use a set of attributes to describe a product. For example, we may use (*Celine Dion*, *Falling into you*, *Sony*) to represent the singer's name, the title and the company of a music CD respectively. In relational database terminology, a set of column names of a relation is called the *schema* of that relation. Correspondingly, we may call the set of attribute names of a product the schema of that product. In this way, the schema of music CDs is (*artist*, *album*, *company*).

Different vendors may differ in the way they describe their products. They may adopt different sets of attributes or vocabularies to describe the same product. For example, (*year*, *classification*, *singer*, *title*, *company*) may be another schema for music CDs. We call such a vendor-specific schema a *local* product schema.

A *global* product schema is a uniform interface for a product based on which heterogeneous product information can be exchanged correctly and efficiently. The interface functions like a common ontology for vendors of the same product. In general, a uniform product interface is desired in any electronic commerce systems that manipulate heterogeneous product information, especially in agent-based electronic commerce systems where transactions are automated.

A product schema in its simplest form is a *flat* set of attributes. However, the general schema model of a product is a *rooted tree* in which the root denotes the product and tree nodes denote product attributes. With the advent of XML (eXtensible Markup Language), the trend is to adopt tree-structured product schemas because DTDs (Document Type Definition) that define the structures of XML documents are tree-like structures.

Figure 1 shows a tree product schema for laptop computers extracted from one of the representative product description samples provided by Veo Systems. We observe the following characteristics in a typical tree-structured product schema:

—*Shallowness*: A product schema tree is usually shallow (say of at most two or three levels) because the relationship among product attributes is generally of the simple 'belongs-to' kind.

—*Bushiness*: The number of attributes at one level can be very large, especially at the second level of the product schema tree because there are many attributes to describe a product and most of them hold a 'many-to-one' relationship to the product (i.e., the root of the tree).

The relationships among product attributes in tree schemas are no longer flat; some attributes are siblings, some are parents and children. In addition to the semantic heterogeneity among individual product attributes, the heterogeneity of attribute positions in different tree schemas is also an important issue.

Product schema integration is essentially a process of building mappings among product attributes from different product descriptions. As in other schema integration problems (such as database schema integration), heterogeneity among local product schemas can be classified into two categories, namely *naming conflicts* and *missing attributes*. Naming conflicts include *synonyms*, words similar in meaning but different in spelling, and *homonyms*, words similar in spelling but different in meaning in different contexts. For example, 'album' and 'title' are synonyms in the local product schemas of music CDs.

In addition, some product attributes used by one vendor may not be used by another. This results in missing product attributes. For example, some vendors may use 'chassis' as an attribute to describe a PC while others may not.

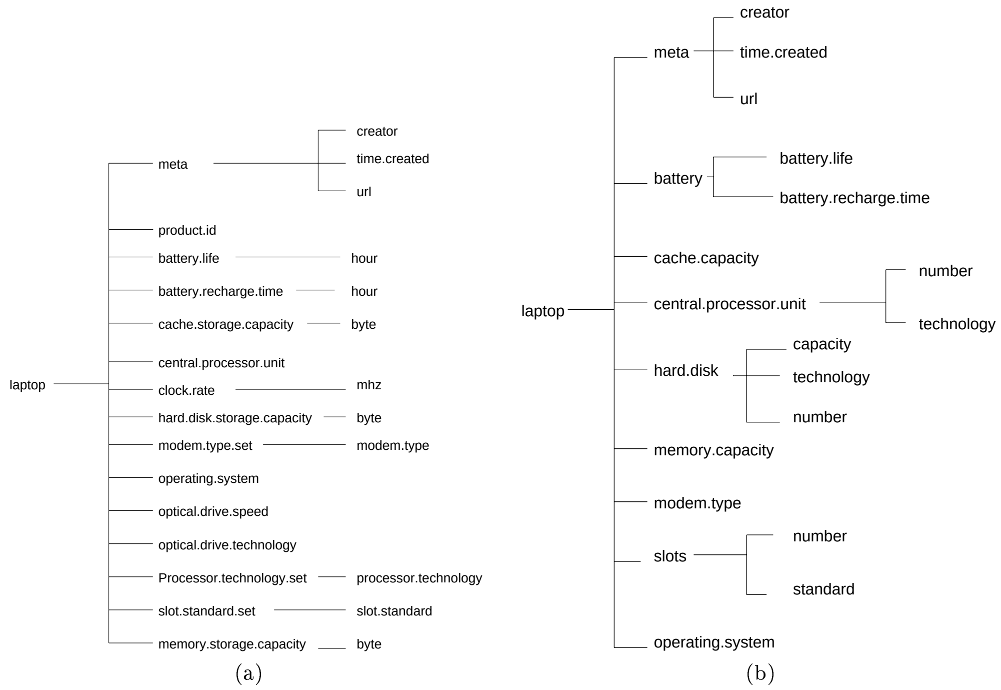
When product schemas are multi-level trees, naming conflicts and missing attributes have additional properties. We generalize the following properties by comparing the laptop computer schemas in Figures 1(a) and 1(b):

—*Partial attribute names*: For example, to describe the storage capacity of the hard disk of a laptop, one may use 'hard.disk.storage.capacity' (which is a full name, see Figure 1(a)) but another one may use 'capacity' (which is a partial name) as a sub-attribute of 'hard.disk' (see Figure 1(b)). Partial attribute names should be expanded to full attribute names when the product schema tree is flattened into one-level.

—*Single-child product attributes*: In Figure 1(a), the attribute 'clock.rate' has only one child 'mhz', which is the measurement for 'clock.rate'. However, in Figure 1(b), attribute 'clock.rate' is a leaf node. In fact, the only child of a product attribute is not a product attribute; it is either a measurement or value for its parent node. Thus, the only child of each single-child attribute should be removed before any structure transformation takes place.

It has been pointed out by D. Florescu *et al.* [7] that web data integration has to deal with large and evolving number of web sources with little metadata about the characteristics of the sources but a high degree of source autonomy. Specifically, product schema integration has the following characteristics:

—*Limited knowledge of local schemas*: Since product information is proprietary, we may only obtain product schemas without further information about attribute domains or data types from the vendors' web pages. Thus, conventional schema integration methods built on the availability of attribute domain information are



(a) (b)

Fig. 1. Two tree product schemas for laptop computers.

no longer applicable. This presents additional difficulties in understanding the semantics of local product schemas.

—*Large number of local schemas*: The number of different vendors even for the same product can be large. In this situation, human intervention is hardly feasible. A low-cost, scalable and fully automated solution is therefore required.

—*Fast local schema evolution*: Whenever new features of a product are added or old features of a product are removed, the local schema of that product must be updated. For example, newer versions of a multimedia PC product includes additional peripherals which extends the local product schema. This gives rise to the problem of dynamic maintenance of the consistency and integrity of an integrated, global schema.

Product schema integration in the context of electronic commerce differs from the related problem of schema integration in database systems. Although there are many existing methodologies for schema integration in multi-database systems such as the use of knowledge bases, neural networks or manual normalization before integration, they are not applicable to product schema integration in lieu of the characteristics above. More importantly, the automation of product schema integration is an essential requirement. The large number of local schemas to be integrated and the frequent updates of product schemas make it impossible for manual schema integration. In short, a simple, scalable and fully automated schema integration technique at the *attribute name* level should be found.

4. SUMMARY

Product description heterogeneity is an inherent problem in electronic commerce due to the autonomy of vendors in describing their product. Although there are concurrent efforts in the industry to standardize product descriptions, the degree of acceptance and the possible multiplicity of standards remains an issue impeding the progress of standardization itself. An alternative and complementary approach is to develop techniques for product schema integration. Although much work has been done in multi-database schema integration, integration presents a somewhat different problem in the electronic commerce context due to special characteristics. With the widespread adoption of XML in the future, some form of integration would still be needed as vendors retain the freedom to define their description vocabulary.

REFERENCES

[1] N. Adam and Y. Yesha and others. Strategic Directions in Electronic Commerce and Digital Libraries: Towards a Digital Agora. *ACM Computing Surveys*, 28(4):818–835, December 1996.

[2] Common Business Library (CBL). http://www.veosystems.com/xml/cbl/cbl.html.

[3] The Commerce Chain Solution. http://www.commerceone.com/solutions.

[4] Commerce One. http://www.commerceone.com.

[5] UN/SPSC in D & B. http://www.dnb.com/unspsc.

[6] The eCo Framework Project. http://www.commerce.net/projects/currentprojects/eco.

[7] D. Florescu and A. Levy and A. Mendelzon. Database Techniques for the World-Wide Web: A Survey. *ACM Special Interest Group on Management of Data Record (SIGMOD Record'98)*, 27(3), September 1998.

[8] Open Buying on the Internet (OBI). http://www.openbuy.org.

[9] RosettaNet. http://www.rosettanet.org.

[10] United Nations Standard Product and Services Codes. http://www.unspsc.org.

[11] Neil Webber, Conleth O'Conell, Bruce Hunt, et al., editors. *The Information and Content Exchange (ICE) Protocol.* World Wide Web Consortium Recommendation, October 1998. http://www.w3.org/TR/NOTE-ice.