A Layered Approach to Semantic Similarity Analysis of XML Schemas

Jaewook Kim[1,2], Yun Peng[1], Serm Kulvatunyou[2], Nenad Ivezic[2], and Albert Jones[2]
[1]*Department of Computer Science and Electrical Engineering*
University of Maryland, Baltimore County
[2]*National Institute of Standards and Technology*
{jaewook, nivezic, jonesa}@nist.gov, ypeng@umbc.edu, and kbserm@psualum.com

Abstract

One of the most critical steps to integrating heterogeneous e-Business applications using different XML schemas is schema mapping, which is known to be costly and error-prone. Past research on schema mapping has not fully utilized semantic information in the XML schemas. In this paper, we propose a semantic similarity analysis approach to facilitate XML schema mapping, merging and reuse. Several key innovations are introduced to better utilize available semantic information. These innovations, including: 1) a layered semantic structure of XML schema, 2) layered specific similarity measures using an information content based approach, and 3) a scheme for integrating similarities at all layers. Experimental results using two different schemas from an real world application demonstrate that the proposed approach is valuable for addressing difficulties in XML schema mapping.

Keywords: XML Schema, e-Business Integration, Schema Mapping, Similarity Measure, Information Content

1. Introduction

Schema mapping, merging, and reuse are critical steps in integrating independently developed, heterogeneous e-business applications, either within or across enterprises. Typically, manual mapping is very labor-intensive, costly and error-prone [1]. Many schema mapping methods have been proposed [2], but they often fail to thoroughly analyze and fully utilize semantic information in the XML schemas. In this paper, we introduce a semantic similarity analysis approach aimed at facilitating XML schema mapping and reuse. We created a prototype system to validate this approach with real world application data. Several key innovations are introduced, including: 1) a layered semantic structure of XML schema, 2) layered specific similarity measures using an information content based approach, and 3) a scheme for integrating similarities of all layers.

Our approach focuses on recommending a set of data elements in the target schema as likely mapping/merging candidates for each element in the source schema based on their semantic similarities. Various similarity measures, including those we developed earlier [3], are used to measure different aspects of the semantic distance between pairs of data elements.

A series of computer experiments have been conducted using the schemas from two different workgroups at the Automotive Industry Action Group (AIAG) [4] to validate the approach and assess its performance. The experiments produced encouraging results, and several directions were suggested for further performance improvement.

The rest of the paper is organized as follows. Section 2 provides the background of this research, including a brief review of selected existing similarity metrics and an introduction to the real world integration data used in the experiments. Detailed descriptions of the proposed approach are given in Sections 3 and 4. Section 5 reports the computer experiments and results. Finally Section 6 concludes with directions for future research.

2. Background

The common approach to integrating heterogeneous e-business applications is to provide interfaces (also known as adapters) that translate data from native specifications to an interlingua (also known as a merged or standard business document (SBD) specification) whose structure and semantics are agreed upon and understood by all parties involved. The difficulty associated with this approach is that neither the standard nor the context of the integration is well documented. Additionally, relevant techniques such as semantic markup using domain ontologies have not yet reached a level of industrial maturity. Instead, practitioners in industrial integration increasingly rely on expressing the SBD specification in the form of XML schemas. An example of such a standard is BODs (Busi-

ness Object Documents) developed by an OAG (Open Application Group) [5].

The semantics of XML Schema-based SBD specifications are not formally defined but implicitly embedded in the meanings of English words or phrases appearing in the names of the schemas' components and fields as well as in associated descriptions. Precise understanding of these descriptions is difficult because of, among other things, the lack of clearly documented common approaches to associate and specify descriptions. For these reasons, it is very costly for experts to identify the reusable standard components that can be shared by other schemas and to understand how to use them. As a result, users often end up creating new standard components or customizing a standard by adding new, typically duplicating or overlapping components rather than attempting to reuse existing ones [6, 7]. This phenomenon also occurs in large organizations where multiple groups concurrently make SBD specifications (we witness this in the AIAG consortium). The result is a proliferation of standards, many of which have duplicate or overlapping semantics.

In this paper, we use three key terms for XML schema integration: mapping, merging, and reuse. They refer to three closely related but different integration tasks. *Mapping* is a task in which one attempts to populate information in one format into another format. *Reusing* is a task in which one looks for integration specifications to use in an integration project. *Merging*, perhaps the most time-consuming task of the three, is an attempt to combine two or more specifications into a single one. All of the three tasks rely on identifying semantically similar data elements between two schemas.

2.1. Similarity measures and related works

Various approaches have recently been developed to help schema mapping, merging, and reuse between two schemas. Most of these approaches first attempt to identify semantic relations between the elements of the two schemas. The simplest approach to semantic similarity is a linguistic-based metric that computes similarity between names or descriptions of two elements by using string matching [8]. There are a variety of string matching algorithms such as the widely used Jaccard [9] and cosine similarity [10, 11] measures. Others have proposed methods based on a linguistic taxonomy [12] such as WordNet [13], from which one can obtain more accurate and less ambiguous semantics for words in the element names.

Structural similarity measures, such as those based on path length between two entities in a taxonomy, fail to recognize the different importance individual entities and relations have and the different roles they play in semantic analysis and measurement. The information content (IC) based metric was proposed to address this problem [14, 15]. This approach measures the similarity between two entities (e.g., two words, two objects, two structures) x and y based on how much information is needed to describe the commonality between them (e.g., the features or hypernyms that two words share). The more specific the common(x, y), the more similar x and y will be. According to information theory, more information is needed for describing more specific objects, and the degree of specificity can be measured by their information content.

This approach was first applied to the semantic relatedness of word senses in WordNet [15]. It defines common(x, y) as their most specific hypernyms C, and the similarity is given as

$$Sim(x, y) = I(C) = -\log P(C) \quad (1)$$

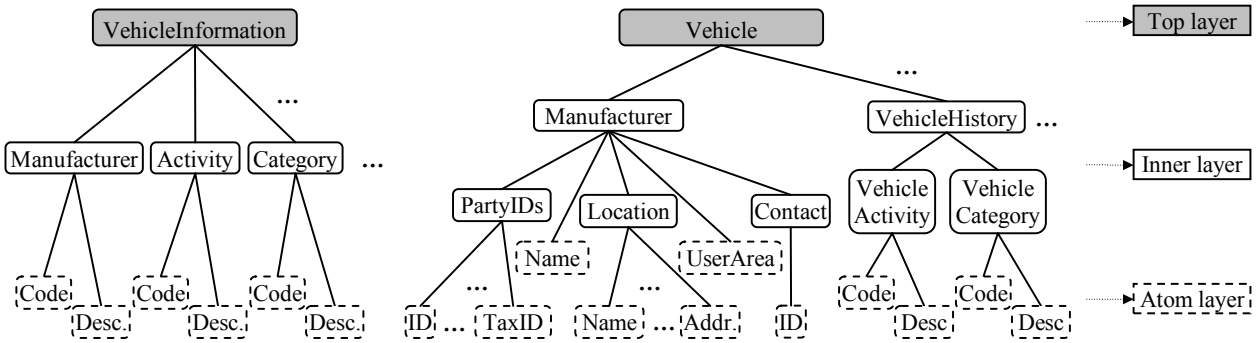
where $I(C)$ is the information content of C, and $P(C)$ were calculated as word frequencies in a corpus.

Research in [16] compares the differences between the IC and structural approaches in measuring similarity between elements in a single XML schema. It shows that better results can be achieved by combining the two approaches.

Each of the existing similarity metrics has its strengths and weaknesses. More importantly, each typically makes use of only part of the available semantic information. In contrast, in this paper we propose an innovative approach that employs a variety of similarity metrics, including lexical, taxonomical, and information content based, in a coherent and justifiable manner.

2.2. Real world data for validation experiments

To test and evaluate the proposed approach, we obtained schemas and manual mapping data from two different workgroups at the Automotive Industry Action Group (AIAG). The AIAG Resource schema and the Truck and Heavy Equipment (T&HE) schema were used as the target and source, respectively. Both schemas are based on the OAG schema [5] and have overlapping concepts. However, they define some elements quite differently. For example, as can be seen in Figure 1 below, both AIAG schema's "Vehicle" and T&HE schema's "VehicleInformation" are intended to describe the same object, but they have different labels (names) and different data structures.



(a) T&HE's VehicleInformation (b) AIAG's Vehicle

Figure 1. Three layers of XML schema

There are a total of 139 global (top) elements defined in the T&HE schema that need to be mapped into the set of 145 global elements of the AIAG schema. Thus, the semantic distances of 139 x 145 (~ 20,000) pairs of elements need to be examined. Roughly 140 human hours were spent mapping 49 top elements in T&HE to those in the AIAG schema. A substantial amount of time is further required to merge at the message level. This is an indication that manual mapping is very time consuming.

3. Layered semantic structure of XML schemas

An XML Schema defines a set of global elements, each of which can be represented as a tree with a set of linked nodes. Each node in a tree has zero or more child nodes. We can classify the nodes into three types: 1) the root; 2) the leaves; and 3) the intermediate nodes (those with both parent and children). We call leaf nodes "atoms" since they are the smallest units and cannot be further divided.

Each tree can be divided into three layers: 1) the *top layer* (containing the root of the tree), 2) the atom layer (containing leaf nodes), and 3) the inner layer (containing intermediate nodes). Note that some trees may have empty inner layers, while others may have only one node, which is considered to be in both top and atom layer.

Each layer typically captures the semantics of a global element from different perspectives. A top layer node through its label and namespace specifies the data object the global element is intended to describe. Nodes in the atom layer indicate the atomic elements. They include, for example, XML schema attributes, simpleType, and simpleContent the designer felt were necessary to describe the global element. The inner layer provides the structural information of the global element by specifying how the atomic elements are grouped into intermediate nodes and, eventually, into the global element (the root). The linguistic information in the labels of both atomic and intermediate nodes may also help to qualify the semantics of the global element.

Consider the two global elements defined in the T&HE and AIAG schemas in Figure 1. The labels in their top layer nodes indicate that both of them are intended to represent the same "vehicle" object. However, the designers differ in their thinking about what atomic elements are needed (see their different atom layers) and how they should be organized (see their different inner layers). In fact, the VehicleInformation in the T&HE schema has 12 intermediate nodes and 198 atoms, while the numbers for the Vehicle in the AIAG schema are 81 and 972, respectively. On the other hand, the same set of ingredients (atoms) can produce elements of different semantics depending on how they are cooked (structured) or packaged (what the top layer node is). For example, several party elements (CustomerParty, DealerParty, and SellingParty) all contain the same atoms and intermediates, but they are intended for semantically different data objects.

4. Similarity measures

The complex relationship between nodes at different layers requires layer specific semantic analysis tools and a mechanism to combine these layer-wise similarities. For this reason, we developed two similarity measures. The first one, called *atom level similarity*, measures the similarity between two atom layers of two elements. The second one, called *label similarity*, measures the similarity between the labels (names). This measure can compare two top layers (each contains a single label) as well as two inner layers (each contains a set of labels). These two measures and how to combine them are described next.

4.1. Atom level similarity

Not every atom is equal in determining semantic similarity. Two elements sharing an atom that is widely used in many elements is not as strong an indication of similarity as sharing an atom that is rarely used [14, 15]. To account for the degree of importance of individual atoms,

we developed an IC based measure for atom layer similarity. Specifically, let $A(x)$ and $A(y)$ denote the sets of atoms of global elements x and y, respectively. Then, the atom level similarity between x and y is defined as

$$Sim_A(A(x), A(y)) = \frac{2 \cdot \Sigma_{c_i \in A(x) \cap A(y)} \cdot I(c_i)}{\Sigma_{c_i \in A(x)} I(c_i) + \Sigma_{c_j \in A(y)} I(c_j)} \quad (2)$$

The probability of each atom is taken as its frequency using the corpus formed by all labels in both T&HE and AIAG. The atom statistics are given in the table below.

Table 1. Statistics of atoms in the two schemas

	AIAG Schema		T&HE Schema	
Total # of atoms	67688		53812	
# of distinct atoms	793		825	
	non-OAG	OAG	non-OAG	OAG
	90	703	119	706

Eq. (2) is based on the assumption that the source and target schemas share a significant number of atoms. This is the case for the AIAG and T&HE schemas (as shown in Table 1; more than 70% of atoms in the two schemas are defined in the OAG schema). Therefore we simply treat two atoms as either completely similar (with similarity score 1) if they have the same label and completely dissimilar (score 0) if they do not. Eq. (2) can be generalized to work in situations where similarity scores between many atom pairs are between 0 and 1. Details of one such measure can be found in [3].

4.2. Label similarity

The label or name x of a node is a word or concatenation of words (or their abbreviations). Before similarity can be compared, a pre-process called "label normalization" is conducted to obtain full words from the concatenations and abbreviations, denoted as $L(x)$. For example, L(VehicleInformation) = {vehicle, information}. To better ascertain the semantics of these words and to deal with the problem of synonyms, we expand each word by its description in WordNet, which consists of definitions of all synonyms, denoted as $d(x_i \in L(x))$.

The descriptions of all the words in $L(x)$ are then put together under two constraints to form a vector of words, $W(x)$. First, for a fair comparison, $W(x)$ should be independent of the lengths of descriptions from the WordNet, which vary greatly from word to word. To achieve this, we require that all $W(x)$ be normalized to a given length, say G words. From the statistics collected on the number of words in the labels and the lengths of word descriptions, we have $\Sigma_i length(d(x_i \in L(x)))$ ranging from 50 to 100 in the experiments. We therefore set G = 500.

Secondly, words in $L(x)$ are not equally important in defining x's semantics (for example, "vehicle" is certainly more important than "information" in the label "VehicleInformation"). Semantic analysis that uses advanced techniques, such as noun phrase analysis from natural language processing is complex and time consuming. Instead, we measure the importance of each word x_i by its information content $I(x_i)$ and require that the vector $W(x)$ be formed in such a way that the number of words from description $d(x_i)$ is proportional to $I(x_i)$.

For example, suppose the vector length G = 10; I(vehicle)/I(information) = 4; and descriptions d(vehicle) = (a b c d) and d(information) = (r s t). To satisfy both constraints, we would have

W(VehicleInformation) = (a b c d a b c d r s)

where d(vehicle) is duplicated and d(information) truncated.

Finally, the similarity of labels x and y is measured by the cosine of the two vectors $W(x)$ and $W(y)$ [10].

The procedure for label similarity is outlined below: For labels x and y:

1) Normalize x and y to obtain full words $L(x)$ and $L(y)$;
2) Calculate the semantic weight of each word $L(x)$ and $L(y)$ by

$$w_{IC}(x_i) = \frac{I(x_i)}{\Sigma_{x_k \in L(x)} I(x_k)}, \; w_{IC}(y_j) = \frac{I(y_j)}{\Sigma_{y_k \in L(y)} I(y_k)} \quad (3)$$

where $I(x_i) = -\log P(x_i)$, and $P(x_i)$ and $P(y_j)$ are taken as their frequencies in their respective schema;
3) Obtain from the WordNet the description of each word in $L(x)$ and $L(y)$, remove most of the stop words from the descriptions [17], make each description a set of words of size $G * w_{IC}(x_i)$ by duplicating or truncating the description, and take a union (keeping all duplicates) of all these sets to form $W(x)$ and $W(y)$;
4) Measure Sim(x, y) by *cosine*($W(x)$, $W(y)$):

$$Sim_T(x, y) = \frac{W(x)W(y)}{|W(x)||W(y)|} = \frac{\sum_{i \subset W(x) \cap W(y)} f_x(i) f_y(i)}{\left(\sum_{i \subset W(x)} f_x(i)^2 \right)^{1/2} \left(\sum_{j \subset W(y)} f_y(j)^2 \right)^{1/2}} \quad (4)$$

where $f_x(i)$ is the frequency of the term 'i' in $W(x)$.

Label similarity for intermediate nodes is measured in the same way and denoted as $Sim_I(x, y)$. In this case, x (and y) is the union of labels of all intermediate nodes.

4.3. Combined similarity score

Several approaches for combining individual similarity measures (Sim_A, Sim_T, Sim_I) have been used in experiments, including: average(a, b, c), max(a, b, c), additive $(1 - (1 - a)(1 - b)(1 - c))$, and weighted sum. The weighted sum seemed to work the best in the experiments:

$$Sim(x, y) = w_A Sim_A + w_T Sim_T + w_I Sim_I \quad (5)$$

where $w_A + w_T + w_I = 1$.

Among other things, this combination scheme allows us to adjust the weights to best reflect the importance of measures at individual layers. The weights can be obtained from the domain experts or learned from human semantic mapping data.

5. Experiments and results

A prototype system is implemented. The system not only computes Sim_A, Sim_I, and Sim_T as given in Eqs. (2) and (4) but also supports several combination rules, including Eq. (5). A series of experiments has been conducted in the prototype system with varying parameters. In these experiments, the 49 manual mappings produced by human integrators are used as the basis to evaluate the performance of the system. For each of the 49 T&HE global elements, the system recommends five most similar AIAG elements. We evaluate performance using a set rather than a single recommendation because the objective is not to fully-automate the process but rather to assist the human expert. A recommendation is considered a match if it contains the manual mapping. Results from using various similarity measures, individual and combined, were obtained and reported in the table below.

Table 2. Experiment results

Similarity measure	# of matches
Sim_T	35
Sim_I	8
Sim_A	22
Sim_T or Sim_I	35
Weighted sum	31

As shown in rows 2 and 3, intermediate-level and atom-level measures by themselves generate poor results (with 8 and 22 matches, respectively). This is because, as discussed earlier, the same set of atoms and intermediates can be used to produce semantically different elements (just like the same ingredients can be made into several kinds of dishes).

The overall performance is mixed. The weighted sum leads to match rate of 63% (31 out of the 49 manual mappings). The combination weights are currently predetermined according to the ratio of the number of matches in each individual measure. This result is certainly very encouraging considering how difficult the problem is even for experienced integrators. However, a detailed examination of the results reveals that 13 manual mappings obtained by human integrators did not appear in any of the recommendations using either individual or combined similarity measures. This calls for further investigation.

Another phenomenon to be noted is evidence suggesting that more weight should be given to the label similarities (top and inner layers). First, only one of the 22 matches found using atom-level similarity was not found by either of the two label-similarity measures. Second, the highest number of matches found by individual measure was using the top-layer measure. Lastly, the cosine method, which uses the combined top and intermediate labels, found 35 matches (4 of them are different from those obtained using the weighted sum combination).

6. Conclusions and plan for future research

In this paper, we propose an innovative semantic similarity analysis approach for XML schemas that exploits semantic information embedded in XML schemas beyond existing methods. This is done by dividing data elements into layers and measuring semantic similarity using layer specific metrics. We also implemented a prototype system to evaluate the proposed approach. This system recommends for each element in a source XML schema a set of mapping candidates in a target schema based on the semantic similarity measures between the elements in these two schemas. The proposed approach and the prototype system have the potential to provide valuable assistance to the human integrators for the problem of XML schema mapping, merging and reuse.

A series of experiments were conducted with encouraging results. The system found a match to the human experts' mapping in 31 of 49 cases in a real world application. The experiments also revealed that the problem is much more complicated than we initially thought. One observation is that the similarity scores vary greatly among the manual mappings (ranging from 0 to 1). This calls for further examination of similarity measures and the way they are combined and for exploring more elaborated mapping procedures. The following immediate steps are planned for future research.

1) Automatically determine the combination weights. Some machine learning techniques are under consideration, including regression and neural networks.
2) Increase the use of structural information. Our experiments show that labels at higher levels are more important than at lower ones. There is also evidence that the atom layer becomes more important when an element's structure is shallow. How to better incorporate the structural information into the semantic analysis will be investigated. In addition to the structural information, utilization of other features of the XML schema, such as cardinality and data type, will also be investigated.

3) Explore an iterative mapping procedure. The hypothesis is that the similarity measures for complex, difficult, or ambiguous elements will become more accurate when more mappings for other easier elements are established with each iteration. For example, atoms defined in the T&HE schema (not in the OAG schema) are currently considered to have zero similarity with any atoms in the AIAG schema. This will be rectified if we map them first, and atom-level similarity for other elements in the subsequent iterations will be improved.

Without proper tools, a harmonized international library of integration specifications such as that envisioned by the UN/CEFACT TBG17 [18] is far-fetched. The number of data elements to harmonize can grow to hundreds of thousands, taking years, if possible at all, to yield usable integration results. The work discussed in this paper shows promise to assist experts in accomplishing integration tasks more efficiently.

Acknowledgements

This work was supported in part by NIST award 60NANB6D6206.

Disclaimer

Certain commercial software products are identified in this paper. These products were used only for demonstration purposes. This use does not imply approval or endorsement by NIST, nor does it imply that these products are necessarily the best available for the purpose.

References

[1] E. Rahm and P.A. Bernstein, "A Survey of Approaches to Automatic Schema Matching", *VLDB Journal*, volume 10, issue 4, 2001, pp. 334-350.

[2] P. Shvaiko and J. Euzenat, "A Survey of Schema-Based Matching Approaches", *Journal on Data Semantics IV*, LNCS 3730, 2005, pp. 146-171

[3] Y. Peng, "On Semantic Similarity Measures", Technical Report from Syllogism.Com to NIST, 2006.

[4] Automotive Industry Action Group (AIAG) Website, http://www.aiag.org

[5] The Open Application Group, "Open Application Group Integration Specification", version 8.0. 2002.

[6] N. Anicic, N. Ivezic, and A.T. Jones, "An Architecture for Semantic Enterprise Application Integration Standards", in Proceedings of the 1st Conference on Interoperability of Enterprise Software and Applications, Geneva Switzerland, 2005.

[7] B. Kulvatunyou, N. Ivezic, and A.T. Jones, "Content-Level Conformance Testing: An Information Mapping Case Study", in Proceedings of TestCom 2005, Montreal, Canada, May 31 - June 2, 2005, pp. 349-364.

[8] H. H. Do and E. Rahm, "COMA - A System for Flexible Combination of Schema Matching Approaches", in Proceedings of the Very Large Data Bases Conference (VLDB), 2001, pp 610–621.

[9] Jaccard similarity, http://www.dcs.shef.ac.uk/~sam/stringmetrics.html#jaccard

[10] Cosine similarity, http://www.dcs.shef.ac.uk/~sam/stringmetrics.html#cosine

[11] B. Jeong, B. Kulvatunyou, N. Ivezic, H. Cho, and A.T. Jones, "Enhance reuse of standard e-business XML schema documents", in Proceedings of international workshop on contexts and ontology: theory, practice and application (C&O'05) in the 20th national conference on artificial intelligence (AAAI'05), 2005.

[12] D. Yang and D.M.W. Powers, "Measuring Semantic Similarity in the Taxonomy of WordNet", in the 28th Australasian Computer Science Conference (ACSC2005), Newcastle, Australia, 2005, pp. 315-322.

[13] WordNet, http://wordnet.princeton.edu/

[14] D. Lin, "An Information-Theoretic Definition of Similarity", in Proceedings of International Conference on Machine Learning, Madison, Wisconsin, July, 1998.

[15] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", in Proceedings of the 14th International Joint Conference on AI, Montreal, CA, 1995, pp. 448-453.

[16] A. Formica, "Similarity of XML-Schema elements: a structural and information content approach", *The Computer Journal*, volume 51, issue 2, 2008, pp. 240-254.

[17] English Stopwords List Website, http://www.ranks.nl/tools/stopwords.html

[18] UN/CEFACT TBG17 Harmonisation workgroup http://www.uncefactforum.org/TBG/TBG17/tbg17.htm.