A New Approach to Merging Structured XML Files

Hung Dinh[1]

***Abstract*—Collaborative work with keeping data up-to-date becomes imperative.To support collaborative problem needs to solve how to editparallel copies of a document and merge the copies into a single document. Many methods such as UNIX diff3, XmlDiff, DeltaXML and 3DM are published to support for merging data from changes. But there are also some problems about semantic in the XML tag suggestions, creating optimal differential sets. The paper uses a new approach of merging 3-way structuredxml files by a base file with three input XML files. Merging process is based on the determination of the structure mapping between the branched files through the base file. To increase the accuracy of the mapping, the paper proposes to refine this mapping by applying WordNet dictionary for XML tag element. The search for this mapping is a necessary step before the merging processoccurring. In addition, the paper's approach allows the identification and treatment of the conflicts that arise in the XML data merging process at different levels. The solution handlingthe conflicts is the key to achieving effective consolidation.**

***Index Terms*— XML, Three-Way Merge, Merging Structured Documents,**

I. Introduction

One basis for computer supporting collaborative work is how to edit parallel copies of a document, and the reintegration of copies into a single document containing the changes.The 3-way merging is a technique that can work integrated changes into a document in the case of many amending independent copies made.

UNIX diff3 is front-end of the *difftool*.It provides support for the consolidation and differentiate text files. Suppose we have a basic document B, it has two branches B_1 and B_2. Diff3 performs consolidation by generating different sets $B\Theta B_1$ and assembling into B_2, or by assembling on B_1 with different sets $B\Theta B_2$.

IBM's Alphaworks also developed a "merging and differentiate XML" tool (XmlDiff) [3]. This tool compares two XML input file, and display differences. Users will then be able to interact merge the two versions by choosing this branch or other branches in different locations.

Merging toolsin synchronize process asDeltaXML [7] do not consideratemoving operation due to manipulate based on analyzing the changes in the 40000 XML files on the Web showingmovingoperation not typically occurring. The author argues that the move operation is capable of applying the XML documents rather than XML data, where the structure tends more fixed.

3DM [8] is a merging tool handling structured data merging problems with good extension mechanism. However the tools not to take advantage of the XML tag suggestions. They based solely on the structure to perform the merging. The conflict handling of tools is quite subjective and unfriendly. For example, with conflicts between two updated versions of the same data, the tool selects automatically by the identified priority (for either version derivatives). The tool generating differential sets and assembling tree uses direct mapping between two trees limited to insert and copy operation leading not to not create optimal differential sets.

The paper proposes the new approach of the structured 3-way merging with three input XML files T_B, T_1 and T_2 where T_B is the base file and the files T_1, T_2 are branched files containing changes compared to the base file T_B. The synchronization will base on the transmission of a number of intermediate files to change sources. The intermediate files allow transmitting only difference parts but the whole file. Thus, it saves bandwidth transmission. In addition, the approach suggested by the paper allows the identification and treatment of the conflicts that arise in the merging process at different levels. The solution handling conflicts is the key to achieve effective merging process.

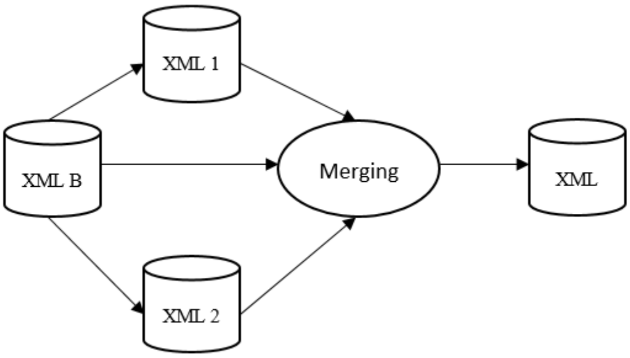


Fig. 1. The 3-way merging method

When the remote computer (client) must be synchronized in low bandwidth environments to host (server),merging system can be used to transmit different information (called delta) and to summarize in the XML file in synchronization handling process.

1) Creating delta: the client generates a delta file to save all changes (T_B-T_1 = Δ_1).

2) Sending delta to server: delta file (usually small) is sent to the server.

3) Re-creating the data of the client (file T_1): server will recreate the status data of client by assembling Δ_1 and T_B (T^B + Δ_1 = T_1).

4) Merging: The 3-way merging task will be performed at the server to aggregate the editing done at the client and at the

Manuscript received May, 2015.

[1] ***Hung Dinh**, Information Technology Department, Ho Chi Minh City University of Foreign Languages and Information Technology, Ho Chi Minh City, Viet Nam, +84903789108.*

server (T_B, T_1, and T2).

5) Creating file Δ_M: server will create a new delta file from the merging files T_M and T_B (T_B-T_M = Δ_M).

6) The file Δ_M is sent to the client and reassembled with T_B ($T_B + \Delta_M = T_M$).

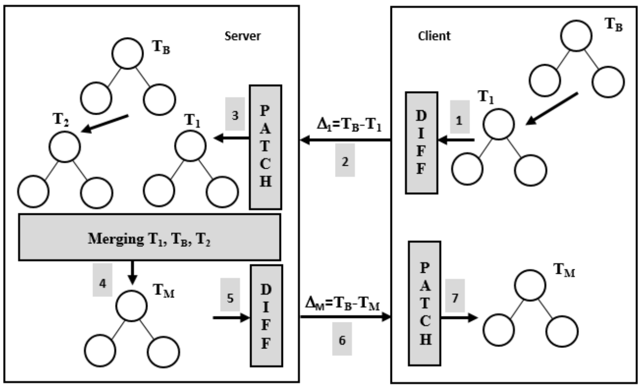


Fig. 2. Synchronization data by the method of 3-way merging

In the paper, the 3-way structured xml-file merging problem is solved by dividing into two specific problems.

Supposing T_1 and T_2 are two ordered trees derived from tree T_B. We will analyze and design a tool that can:

1) Implement the 3-way merging following the structure of tree T_1, T_2 and T_B with to detect, to describe the conflicts occurring in merging process. It is called tree merging problem.

2) Generate difference set between the two trees T_1 and T_2 as an editing script. After that, using different setand information of tree T_1 is to get back tree T_1, and T_2. It is called the problem of differentiating and assembly tree.

II. TREE MAPPING

A. Determination of the same content by q-gram

Q-gram distance between two sequences is defined [5]:

- Given $\sum$the finite character set, $\sum^*$is the set of all sequences generated from $\sum$ and $\sum_q$ are all sequences of length q generated from $\sum$. A q-gram is a sequencev $= a_1 a_2 \dots a_q \in \sum_q$.
- Given q-gram v and $x = a_1 a_2 \dots a_n$ is a sequence in $\sum^*$, if $v = a_i a_{i+1} \dots a_{i+q-1} \subset x \ with\ i\ any$, then v appears in x. We denote the number of occurrences of v in x with$G_q(x)[v]$. Q-gram profile of x is vector$G_q(x) = (G(x)[v]), v \in \sum_q$.
- Q-gram distance between x and y is the normalization 1 of $G_q(x) - G_q(y)$: $D_q(x, y) = \sum_{v \in \sum_q} |G_q(x)[v] - G_q(y)[v]|$

B. Determination of semantic similarity by WordNet

WordNet [2] contains standard information found in the dictionary and thesaurus. A concept often used in the WordNet is *hypernym*[6]. The *hypernym* of a term is a more general term consistent with remarks "_____ *is a kind of* ____". For example, dog is a canine, so the canine is a *hypernym* of dog. A canine is a mammal, therefore is a *hypernym* of canine mammal. This continues until reaching a certain peak terms, very extensive; in this case, the top *hypernym* of dog and canine is entity. These terms are related to each other through *hypernymy* concept, forming a *hypernym* tree. The Opposite relations are called hyponymy.

The distance between two nodes: Given node A and B, the distance between A and B is denoted $Distance(A, B)$ calculated using the formula:

$$Distance(A, A) = 0$$

$$Distance(A, B) = \begin{cases} 0\ if\ Similar(A, B) \\ Distance(A, C) + Distance(B, C) \\ with\ C = NearestAncestor(A, B) \\ if\ not\ (Similar(A, B)) \end{cases}$$

Currently, Vietnamese WordNet is not built, and the Vietnamese WordNet realization is outside the scope of the paper. To illustrate the ability to use Vietnamese WordNet in the future, we will build the thesaurus (and antonyms) Vietnamese.We use Vietnamese dictionary to measure two synonyms (antonym) in the merging process.

C. Search algorithm for mapping seedlings

We will seek the greatest possible seedlings of T_B and T_1 mapping together with greedy heuristic algorithm.

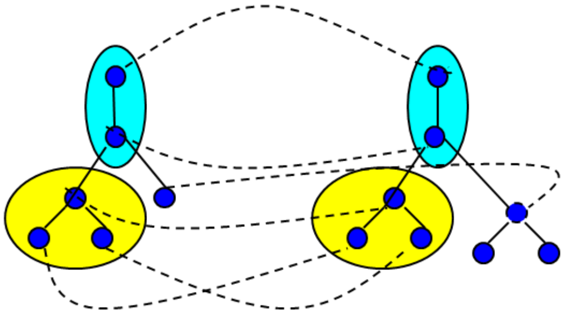


Fig. 3. Mapping between two trees

The mapping between two trees is to find the corresponding seedlings in two trees. The corresponding (mapping) is defined as follows:

- Accurate mapping between two seedlings: The contents of the root node and the child node as well as their order must be identical. Generally, two trees can pile up together.
- Homologous structures mapping: The content root node can vary but the content as well as the order of the child nodes must be identical. So, the strategies with similar mapping will depend on accepting deviations between two root nodes together. Because of these factors, we will interfere with 2 measuring deviations using 2 methods: thesaurus, q-gram.

Here is an illustration of the above methods:

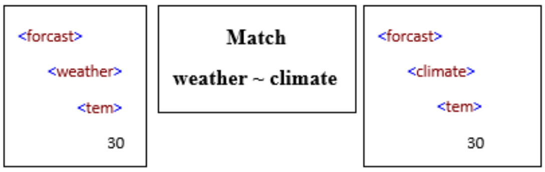


Fig. 4. Mapping using thesaurus

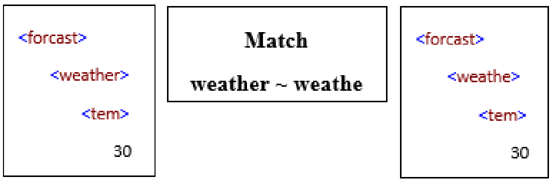


Fig. 5. Mapping using q-gram

III. PROPOSED METHOD

A. *The 3-way merging structured xml file algorithm*

Merging process includes three stages:

- Phase 1: Generating merging list of each tree seedlings based on T_B. In this phase, change characteristics of tree seedlings compared with T_B will be marked by two operations: suspending the new added nodes and locking the moved node.
- Phase 2: Creating merging pairs. Input of this phase is twomerging list from stage 1: two seedlings T_1 and T_2. Output is a list of node pairs will be implemented merging at the later stage. All changes characteristic of two seedlings that we identified in Phase 1 will be reserved during this period.
- Phase 3: Merging pairs taken from the list in stage 2. We will perform merging each pair in this list. The result will be a new node in the merging tree.

Handling conflict updates/updates:

- When conflictsoccurring in tag name of element node will be treated with three options: automatically select by T_1, choose by users, select by measuring the semantics distance of the tag name using WordNet.
- When conflicts occurring in text node will be treated with three options: automatically select by T_1, choose by users, and allow editing the content of the text node.

To increase the effectiveness of tools, conflict solving processes in update/update text node will be divided into two phases: calculate the difference between T_B and T_1 and between T_B and T_2, selection and editing.

The distinctness of the two text node content is given by the LCS calculation algorithm [4].

B. *Generating differential set between two trees and assembling trees*

To export to differential set of two XML files T_1 and T_2, we use the algorithmabout creating editing script [1] to obtainminimum differential set but not focused on differential set be encrypted for friendly user. Our purpose is differential set ability to assembling with T_2 to generating T_1.

To describe the differences between two XML files (two ordered trees), we use the concept of editing script with minimal cost. The minimum cost editing scripting of the two treesis defined by using the insert, delete node and move sub-trees operation as basic editing operations. We divided into two sub problem:(1) finding a mapping between objects in two files, (2) finding the editing script.

If the objectshave unique identifiers, the first problem is simplified and we can use this property to increase processing speed. We will use the results of the mapping in the tree mapping stage (just use the exact mapping) to perform the calculation editing script.

Recurrence of T_1 and T_2 from the script is applicable only operations listed in the editing script with T_1.

C. *Architecture overview*

Fig. 6 presents the overall architecture of XML file merging tool. The tool reads information from two or three input XML files and analyzes XML files into the internal tree structure. The node in the tree structure will then be mapped together.

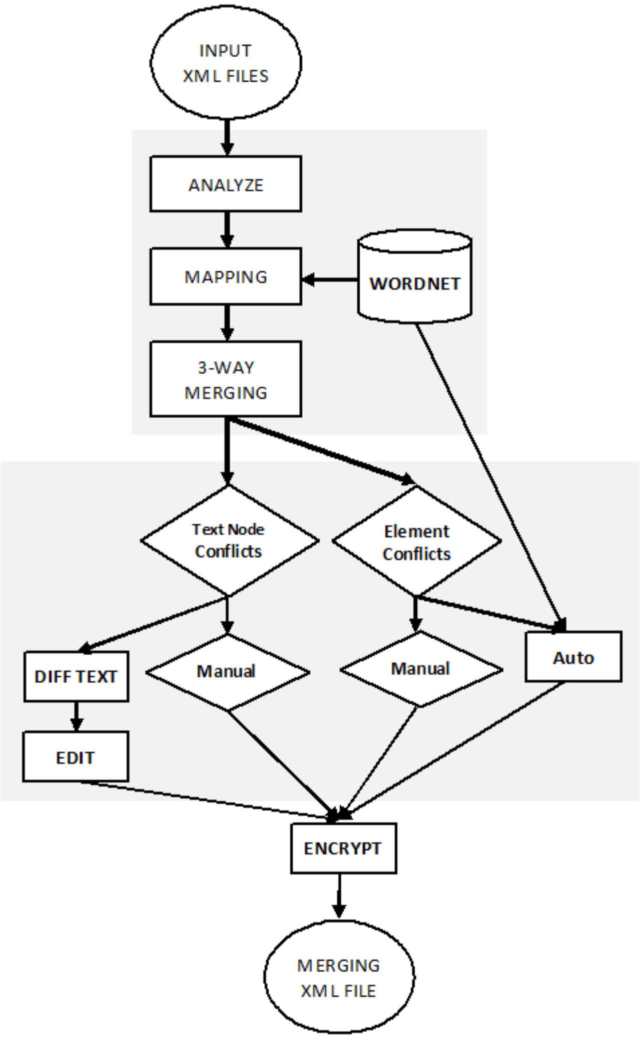
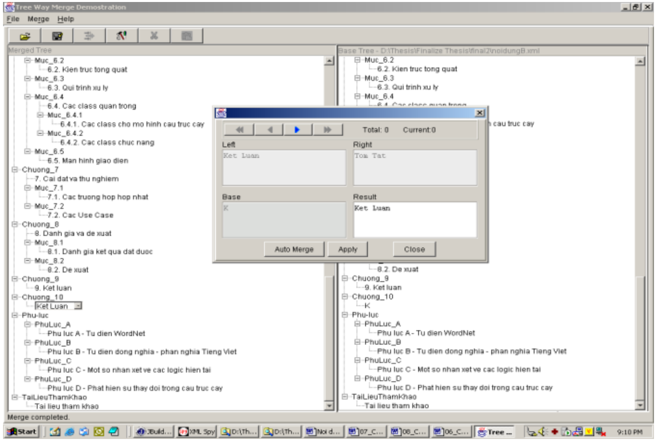


Fig. 6. Architecture of XML file merging tool

The strength of this architecture is to allow interaction with the user. When merging process occurred conflicts, the selection can be performed automatically, but also allows the user to decide. Tool is designed intuitive and easy to use. Information input is standard XML files.

Tool generates differential set not require a certain knowledge. Moreover, the requirement is to create minimum differential set and to "assemble" the tree as the original tree without reviewing or searching a good mapping in the sense of merging problem.



Total: 0 Current:0

Left: Ket Luan

Right: Tom Tat

Base: K

Result: Ket Luan

Merge completed.

Fig. 7. GUI of XML file merging tool

IV. CONCLUSION AND FUTURE WORK

Besides refining fuzzy mapping, based on merging process on the minimal support, the paper also installed test systems by measuring semantic WordNet based prolog database of Princeton University converted to more appropriate mechanism to facilitate the process between two nodes semantic measure aimed initially mapped handling process and conflicts subtly and usefully; so we assess that future mobile devices with preinstalled WordNet is a very common and it only occupies negligible space (as well as the processing time) (approximately <5MB). Besides, in order to demonstrate supporting multilingual system issues, the paper also installed test thesaurus-antonym Vietnamese (because currently there is not Vietnamese WordNet).

To handle conflicts, the paper has distinguishes between TagName conflicts and Text Node conflicts. For TagName conflict, because of suggestion of the tagname, we can use WordNet to select automatically. For TextNode conflict, through LCS algorithm, users can identify the difference between the conflict contents, select and edit directly on the TextNode collision.

We have used encryption of differential set as anediting script to allow minimizing differences set, but do not use this script to merge. With this approach, the paper has to overcome two drawbacks in the process of merging (to avoid the fixed link problem) and in the process making a difference (to minimize the difference set).

These issues need to be resolved next to perfect the tool include: improve 3-way merging algorithm in the case of structure conflict, applying new algorithms to create differential set as small as possible, demonstrating the editing script and XML files as user friendly, Handling DTD (Document Type Definitions) where two files with the same structure but DTD different need to be aware, data processing in XML file format CDATA, handling tag name with long names (to process language), considering the semantics of Text Node.

REFERENCES

[1] Chawathe S. S, Rajaraman A, Garcia-Molina H. and Widom J., “Change detection in hierarchically structured information”, In *Proceedings of the 1996 ACM SIGMOD International Conference* on Management of Data, 1996

[2] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller, “Introduction to WordNet: An On-line Lexical Database**”,** (Revised August 1993)

[3] IBM Alphaworks, “XML diff and merge tool home page”, http://www.alphaworks.ibm.com/tech/xmldiffmerge

[4] J.W.Hunt, “An Algorithm for Difference File Comparison”, Department of Electrical, M. D. McIlroy,Bell Laboratories, Murray Hill, New Jersey 07974 - Bell Laboratories Computing Science Technical, Report #41, dated July 1976.

[5] Stefan Kurtz, “Foundations of Sequence Analysis**”,**Lecture notes for a coursein the Winter Semester 2000/2001 July 18, 2002

[6] http://www.cogsci.princeton.edu/~wn/ - Five Papers on WordNet

[7] http://www.deltaxml.com - DeltaXML Project [referenced 08 Feb 2015]

[8] Tancred Lindholm, “A Three-way Merge for XML Documents”, ACM Symposium on Document Engineering - ACM Press, October 2004.

Hung Dinh received the Master degree in Computer Science from the Ho Chi Minh City University of Science, Ho Chi Minh City, Vietnam, in 2003.

In 2008, he has joined the Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology, Ho Chi Minh City, Vietnam, where he is currently a Lecturer. His research interests include database systems, ontology, and parallel programming.