ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Semesterarbeit for one student

Diff and XML-Diff Algorithms

Contact: paul.sevinc@inf.ethz.ch

Introduction & Project Objectives

Tracking changes in XML documents can be of relevance to security. For example, when authors cannot repudiate their changes, they may be more reluctant to mount a social-engineering attack such as phishing by changing the URL of the log-in page of a banking application in an otherwise authentic electronic brochure:

- <li href="http://www.niceandh**o**nestbank.ch/">N & H Bank</li>
- <li href="http://www.niceandh**o**nestbank.ch/">Nice & Honest Bank</li>

XML documents are text based, so one could compare XML documents by comparing them line by line. XML documents are also tree structured, so one could compare them node by node. One approach may be better in terms of execution speed of the computer while another may be better in terms of ease of spotting (security!) relevant differences for a human user.

In the example above, one might think that the lines differ only in the text if the line as a whole was flagged as having changed when in fact both the attribute node (lower-case letter o to digit zero) and the text node ("N & H" to "Nice & Honest") have been changed.

The objectives of this project are to survey diff and XML-diff algorithms and to discuss their differences with the help of insightful examples.

Work Plan

1. Search, collect, and study relevant publications (cf. References).
2. Define a set of abstract comparison criteria and a set of concrete examples.
3. Apply the criteria to diff and XML-diff algorithms and illustrate their differences with the examples.
4. Present the findings in a talk (in German or English) and in a report (in German or in English).

Prerequisites

Working knowledge of Java development and the Extensible Markup Language (XML).

Supervision

Prof. Dr. David Basin and Paul E. Sevinç (paul.sevinc@inf.ethz.ch).

References

1. R. Chavez. “Exploring the problems involved in comparing XML.”
   Available at <http://www.oreillynet.com/pub/wlg/3861>.
2. Ercato Project. “XOp.”
   Home page at <http://www.living-pages.de/de/projects/xop/index.html>.
3. Free Software Foundation. “Diffutils.”
   Home page at <http://www.gnu.org/software/diffutils/diffutils.html>.
4. C.E. Perez. “Open Source XML Diff Written in Java.”
   Available at <http://www.manageability.org/blog/stuff/open-source-xml-diff-in-java>.
5. J. Udell. “Structured change detection.”
   Available at
   <http://www.infoworld.com/article/04/02/27/09OPstrategic_1.html>.
6. G. von Walter. “XML operations: Adding and subtracting XML Documents.”
   Available at
   <http://www.netobjectdays.org/node04/de/Conf/publish/talks.html#operations:_Adding_a.t>.