UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

CLAVES XML: UNA IMPLEMENTACIÓN DE ALGORITMOS DE IMPLICACIÓN Y VALIDACIÓN

Propuesta de Trabajo de Título para Ingeniero Civil en Informática

Nombre:	Emir Fernando Muñoz Jiménez
R.U.N.:	16.269.0302
Año Ingreso:	2009
Teléfono:	+569 8752 9608
E-mail:	emir.munozj@usach.cl
Profesores:	Dr. Mauricio Marín
	Dr. Flavio Ferrarotti

Lunes, 30 de Mayo de 2011

AGRADECIMIENTOS

A mi...

RESUMEN

La flexibilidad sintáctica, y el complejo anidamiento de los datos en una estructura tipo árbol dificulta expresar propiedades deseables de los datos XML, ofreciendo una capacidad limitada para expresar semántica. En esta tesis se presenta un estudio de las claves como restricciones de integridad sobre documentos XML, implementando algoritmos para los problemas de implicación y validación, con el fin de mostrar la factibilidad de usar las capacidades semánticas que éstas entregan, y que XML como modelo requiere.

Palabras Claves: XML; Claves XML; Implicación de claves; Validación de documentos XML; Cover no redundante

ABSTRACT

The syntactic flexibility and complex tree-like nested data make it challenging to express desirable properties of XML data, offering a limited capability to express semantic. In this thesis, we present a study of keys as integrity constraints on XML documents, implementing algorithms for implication and validation problems, with the aim of showing the factibility of using the semantic capabilities that keys gives and XML as a model requires.

Keywords: XML; XML keys; Key implication; XML document validation; Non-redundant cover

ÍNDICE DE CONTENIDOS

Índice de Figuras vii

Índice de Tablas viii

Índice de Algoritmos ix

1. Introducción 1

1.1. Antecedentes y motivación 1

1.2. Descripción del problema 2

1.3. Solución propuesta 3

1.4. Objetivos y alcance del proyecto 3

1.4.1. Objetivo general 3

1.4.2. Objetivos específicos 3

1.4.3. Alcances 3

1.5. Metodología y herramientas utilizadas 3

1.5.1. Metodología 3

1.5.2. Herramientas de desarrollo 3

1.6. Resultados Obtenidos 3

1.7. Organización del documento 3

2. Marco Teórico 4

2.1. Documentos XML 4

2.2. El modelo de árbol XML 4

Referencias 6

Apéndices **9**

A. Manual de Usuario **10**

A.1. Requerimientos 10

A.2. Instalación 10

ÍNDICE DE FIGURAS

2.1. *Modelo de árbol para un documento XML.* 5

ÍNDICE DE TABLAS

ÍNDICE DE ALGORITMOS

CAPÍTULO 1. INTRODUCCIÓN

1.1 ANTECEDENTES Y MOTIVACIÓN

El rápido desarrollo de la Web ha generado nuevos problemas y áreas de investigación en las Ciencias de la Computación e Informática. Junto a eso ha iniciado el desarrollo de (innumerables) nuevas tecnologías, y la evolución de otras. La comunicación e interoperabilidad de estas tecnologías es un tema crucial para mantener el desarrollo de la Web. En particular, XML (*eXtensible Markup Language* o Lenguaje extensible de marcado) ha surgido como un modelo de datos estándar para almacenar e intercambiar datos en la Web. Su rol en el intercambio de datos ha pasado de simplemente transmitir la estructura de los datos, a uno que también transmite su semántica (Benedikt et al., 2003; Davidson et al., 2007).

XML (Bray et al., 2006) es la propuesta del *World Wide Web Consortium* (W3C) como lenguaje de intercambio e interoperabilidad en la Web. Este lenguaje proporciona un alto grado de flexibilidad sintáctica, pero ofrece una capacidad limitada para expresar la semántica de los datos. Esta flexibilidad sintáctica, y el complejo anidamiento de los datos en una estructura tipo árbol, dificulta expresar propiedades deseables de los datos XML. En consecuencia, el estudio de restricciones de integridad ha sido reconocido como una de las áreas de investigación en XML más difíciles (Fan, 2005; Suciu, 2001; Vianu, 2003; Widom, 1999).

El estudio de restricciones de integridad ha sido reconocido como una de las áreas más difíciles de investigación en XML (Vianu, 2003). En el modelo relacional, las restricciones han sido estudiadas extensamente (Fagin & Vardi, 1984; Thalheim, 1991), y son esenciales para el diseño de esquemas, la optimización de consultas, y métodos eficientes de acceso y almacenamiento (Abiteboul et al., 1995). Varias clases de restricciones de integridad han sido definidas para XML, incluyendo claves (Buneman et al., 2002), restricciones de camino (Buneman et al., 2001, 2000), restricciones de inclusión (Fan & Libkin, 2002; Fan & Siméon, 2003), y dependencias funcionales

(Arenas & Libkin, 2004; Hartmann & Trinh, 2006; Vincent et al., 2004). Sin embargo, la mayoría de las clases de restricción, dada la compleja estructura de datos XML, resultan en problemas de decisión que son intratables, y es difícil encontrar clases de restricciones XML que sean naturales y útiles, y que puedan ser razonadas de manera eficiente (Fan, 2005; Fan & Siméon, 2003; Fan & Libkin, 2002; Suciu, 2001; Vianu, 2003; Arenas et al., 2002). Las principales candidatas de esas clases son las claves absolutas y relativas (Buneman et al., 2003, 2002) que son definidas en base a un modelo de árbol para XML como el propuesto por DOM (Apparao et al., 1998) y XPath (Clark & DeRose, 1999), de manera independiente a alguna especificación del tipo[1] de un documento XML como DTD[2] o XML *Schema* (Thompson et al., 2004). ...

1.2 DESCRIPCIÓN DEL PROBLEMA

La definición de claves XML es más compleja que en el modelo relacional, debido a la compleja estructura de árbol que poseen los documentos XML.

En este trabajo se plantea, en primer lugar, determinar la utilidad de trabajar en la práctica con claves XML utilizando un algoritmo para el problema de implicación. Definir ésta utilidad práctica permitirá avanzar en la aceptación de las claves como restricciones sobre XML por parte de los profesionales, considerando el poder expresivo que estas entregan a XML. En segundo lugar, existe la necesidad de un método que permita determinar la validez de un documento XML contra un conjunto predefinido de claves XML. A partir de los trabajos realizados en validación de documentos XML contra claves (Abrão et al., 2004; Bouchou et al., 2003; Chen et al., 2002; Liu et al., 2005, 2004), se plantea diseñar un algoritmo que permita validar documentos XML contra claves XML como las definidas en Buneman et al. (2003, 2002), las cuales consideran la igualdad en valor entre nodos elemento: si los subárboles que tienen por raíz a estos nodos, son isomorfos por algún isomorfismo que para cadenas de texto se corresponde con la función identidad.

Finalmente, considerando que la complejidad del algoritmo de validación depende en parte del tamaño del conjunto de claves, se investiga un método para obtener una optimización del proceso de validación de documentos XML contra claves, utilizando el algoritmo de implicación de

[1] Un *tipo* en XML es considerado como una gramática extendida libre de contexto, asociada a restricciones en la estructura de los elementos de un documento.

[2] *Document Type Definition* o Definición de tipo de un documento XML.

claves XML presentado por Hartmann & Link (2009).

1.3 SOLUCIÓN PROPUESTA

1.4 OBJETIVOS Y ALCANCE DEL PROYECTO

1.4.1 Objetivo general

1.4.2 Objetivos específicos

Para la consecución del objetivo general, se plantean las siguientes metas intermedias:

1. Estudiar la noción de claves XML y las axiomatizaciones existentes.
2.

1.4.3 Alcances

1.5 METODOLOGÍA Y HERRAMIENTAS UTILIZADAS

1.5.1 Metodología

1.5.2 Herramientas de desarrollo

1.6 RESULTADOS OBTENIDOS

1.7 ORGANIZACIÓN DEL DOCUMENTO

El presente trabajo está dividido en ocho capítulos considerando éste como el primero. En el Capítulo 2 se formalizan los fundamentos de documento XML, modelo de árbol XML, y lenguaje

de definición de expresiones de camino para definir claves XML. ...

CAPÍTULO 2. MARCO TEÓRICO

2.1 DOCUMENTOS XML

2.2 EL MODELO DE ÁRBOL XML

```
<?xml version="1.0" encoding="UTF-8"?
<libro>
  <titulo>El Juego</titulo>
  <capitulo id="cap1">
    <titulo>Parte 1</titulo>
  </capitulo>
  <capitulo id="cap2">
    <titulo>Parte 2</titulo>
  </capitulo>
</libro>
```

FIGURA 2.1: *Modelo de árbol para un documento XML.*

REFERENCIAS

Abiteboul, S., Hull, R., & Vianu, V. (1995). *Foundations of Databases: The Logical Level*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1st ed.

Abrão, M. A., Bouchou, B., Ferrari, M. H., Laurent, D., & Musicante, M. A. (2004). Incremental Constraint Checking for XML Documents. In Z. Bellahsène, T. Milo, M. Rys, D. Suciu, & R. Unland (Eds.) *Database and XML Technologies*, vol. 3186 of *Lecture Notes in Computer Science*, (pp. 358–379). Springer Berlin / Heidelberg. 10.1007/978-3-540-30081-6_9.
URL http://dx.doi.org/10.1007/978-3-540-30081-6_9

Apparao, V., Byrne, S., Champion, M., Isaacs, S., Hors, A. L., Nicol, G., Robie, J., Sharpe, P., Smith, B., Sorensen, J., Sutor, R., Whitmer, R., & Wilson, C. (1998). Document object model (DOM) level 1 specification. http://www.w3.org/TR/REC-DOM-Level-1/. Extraído el 19 de Octubre de 2010.

Arenas, M., Fan, W., & Libkin, L. (2002). What's Hard about XML Schema Constraints? In *Proceedings of the 13th International Conference on Database and Expert Systems Applications*, DEXA '02, (pp. 269–278). London, UK, UK: Springer-Verlag.
URL http://portal.acm.org/citation.cfm?id=648315.756182

Arenas, M., & Libkin, L. (2004). A normal form for XML documents. *ACM Trans. Database Syst.*, *29*, 195–232.

Benedikt, M., Chan, C.-Y., Fan, W., Freire, J., & Rastogi, R. (2003). Capturing both types and constraints in data integration. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, SIGMOD '03, (pp. 277–288). New York, NY, USA: ACM.
URL http://doi.acm.org/10.1145/872757.872792

Bouchou, B., Alves, M. H. F., & Musicante, M. A. (2003). Tree Automata to Verify XML Key Constraints. In V. Christophides, & J. Freire (Eds.) *WebDB*, (pp. 37–42).

Bray, T., Paoli, J., Sperberg-Queen, C., Maler, E., & Yergeau, F. (2006). eXtensible Markup Language (XML). `http://www.w3.org/TR/2006/REC-xml-20060816/`. Extraído el 21 de Marzo de 2010.

Buneman, P., Davidson, S. B., Fan, W., Hara, C. S., & Tan, W. C. (2002). Keys for XML. *Computer Networks*, *39*(5), 473–487.

Buneman, P., Davidson, S. B., Fan, W., Hara, C. S., & Tan, W. C. (2003). Reasoning about keys for XML. *Inf. Syst.*, *28*(8), 1037–1063.

Buneman, P., Fan, W., Siméon, J., & Weinstein, S. (2001). Constraints for Semi-structured Data and XML. *SIGMOD Record*, *30*(1), 47–45.

Buneman, P., Fan, W., & Weinstein, S. (2000). Path Constraints in Semistructured Databases. *J. Comput. Syst. Sci.*, *61*(2), 146–193.

Chen, Y., Davidson, S. B., & Zheng, Y. (2002). XKvalidator: A Constraint Validator For XML. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, (pp. 446–452). New York, NY, USA: ACM.
URL `http://doi.acm.org/10.1145/584792.584866`

Clark, J., & DeRose, S. (1999). XML Path Language (XPath). `http://www.w3.org/TR/xpath`. Extraído el 21 de Marzo de 2010.

Davidson, S., Fan, W., & Hara, C. (2007). Propagating XML constraints to relations. *J. Comput. Syst. Sci.*, *73*, 316–361.
URL `http://portal.acm.org/citation.cfm?id=1223810.1223864`

Fagin, R., & Vardi, M. Y. (1984). The Theory of Data Dependencies - An Overview. In *Proceedings of the 11th Colloquium on Automata, Languages and Programming*, (pp. 1–22). London, UK: Springer-Verlag.
URL `http://portal.acm.org/citation.cfm?id=646238.683349`

Fan, W. (2005). XML Constraints: Specification, Analysis, and Applications. In *DEXA Workshops*, (pp. 805–809). IEEE Computer Society.

Fan, W., & Libkin, L. (2002). On XML integrity constraints in the presence of DTDs. *J. ACM*, *49*(3), 368–406.

Fan, W., & Siméon, J. (2003). Integrity constraints for XML. *J. Comput. Syst. Sci.*, *66*(1), 254–291.

Hartmann, S., & Link, S. (2009). Efficient Reasoning about a Robust XML Key Fragment. *ACM Trans. Database Syst.*, *34*(2).

Hartmann, S., & Trinh, T. (2006). Axiomatising Functional Dependencies for XML with Frequencies. In J. Dix, & S. J. Hegner (Eds.) *FoIKS*, vol. 3861 of *Lecture Notes in Computer Science*, (pp. 159–178). Springer.

Liu, Y., Yang, D., Tang, S., Wang, T., & Gao, J. (2004). Extracting Key Value and Checking Structural Constraints for Validating XML Key Constraints. In Q. Li, G. Wang, & L. Feng (Eds.) *Advances inWeb-Age Information Management*, vol. 3129 of *Lecture Notes in Computer Science*, (pp. 399–408). Springer Berlin / Heidelberg. 10.1007/978-3-540-27772-9_40.
URL `http://dx.doi.org/10.1007/978-3-540-27772-9_40`

Liu, Y., Yang, D., Tang, S., Wang, T., & Gao, J. (2005). Validating key constraints over XML document using XPath and structure checking. *Future Generation Computer Systems*, *21*(4), 583–595. High-Speed Networks and Services for Data-Intensive Grids: the DataTAG Project.

Suciu, D. (2001). On database theory and XML. *SIGMOD Rec.*, *30*(3), 39–45.

Thalheim, B. (1991). *Dependencies in Relational Databases*. Teubner.

Thompson, H., Beech, D., Maloney, M., & Mendelsohn, N. (2004). XML Schema Part 1: Structures Second Edition. `http://www.w3.org/TR/2004/REC-xmlschema-1-20041028`. Extraído el 19 de Octubre de 2010.

Vianu, V. (2003). A Web odyssey: from codd to XML. vol. 32, (pp. 68–77).

Vincent, M. W., Liu, J., & Liu, C. (2004). Strong functional dependencies and their application to normal forms in XML. *ACM Trans. Database Syst.*, *29*(3), 445–462.

Widom, J. (1999). Data Management for XML: Research Directions. *IEEE Data Engineering Bulletin*, *22*, 44–52.

APÉNDICE A. MANUAL DE USUARIO

A.1 REQUERIMIENTOS

blablablabla....

A.2 INSTALACIÓN

blablablabla....

blablablabla....