

Selección de contenidos a partir de un corpus enriquecido y especializado para la generación personalizada de documentos

Guillermo Barrutieta Anduiza

Departamento de Informática
Mondragon Unibertsitatea
Loramendi, 4 - 20500 Mondragón
gbarrutieta@eps.mondragon.edu

Resumen: Tesis doctoral en Informática realizada en la Universidad de Deusto por Guillermo Barrutieta Anduiza bajo la dirección de JosuKa Díaz Labrador y Joseba Abaitua Odriozola (Univ. de Deusto). La defensa de la tesis tuvo lugar ante el tribunal formado por los doctores Mikel L. Forcada Zubizarreta (Univ. de Alicante), Ana García Serrano (Univ. Politécnica de Madrid), Xavier Gómez Guinovart (Univ. de Vigo), Anselmo del Moral Bueno y Javier Oliver Bernal (Univ. de Deusto) el 29 de abril de 2004. La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad.

Palabras clave: Filtros de contenido, perfiles de usuario, corpus, RST, XML, XSL y DTD.

Abstract: PhD thesis in Computer Science written by Guillermo Barrutieta Anduiza at Univesidad de Deusto under the supervision of JosuKa Díaz Labrador and Joseba Abaitua Odriozola (Univ. de Deusto). The thesis defense (viva voce) took place before the commitee formed by the doctors Mikel L. Forcada Zubizarreta (Univ. de Alicante), Ana García Serrano (Univ. Politécnica de Madrid), Xavier Gómez Guinovart (Universidad de Vigo), Anselmo del Moral Bueno and Javier Oliver Bernal (Univ. de Deusto) on the 29th of April 2004. The grade obtained was Sobresaliente Cum Laude.

Keywords: Content filters, user profiles, corpora, RST, XML, XSL and DTD.

1. Introducción

El objetivo principal de la tesis era proporcionar un modelo computacional basado en tecnología XML y *web* que implementase la propuesta de solución algorítmica para el aprovechamiento de un corpus paralelo multilingüe destinado a la generación de vistas de documentos que se adecuasen a las necesidades de los usuarios. Este objetivo principal suponía la formalización de dos cuestiones troncales: el modelo lingüístico contenido en el corpus y el perfil de usuario.

2. Generación personalizada de documentos

La tesis abordó la problemática asociada a la generación de documentos y aportó como solución novedosa *un algoritmo de selección de contenido* que permite la generación personalizada de documentos a partir de la *formalización del documento maestro*.

El enfoque utilizado para la representación computacional del documento

maestro es la variante de la teoría de la estructura retórica del discurso (RST) de grano grueso que explicita la organización básica de los segmentos de texto candidatos a formar parte de la vista del documento. La explicitación de la estructura del discurso enriquece el documento y se materializa computacionalmente utilizando *metadatos*, representados mediante etiquetas XML.

Por otra parte, el planteamiento utilizado para resolver la implementación del algoritmo se basa en *filtros en cascada* que seleccionan y determinan de forma coordinada los segmentos de texto que formarán la vista del documento a generar para el usuario final. Estos filtros, desarrollados en XSL, junto con el *mecanismo de activación* Javascript, implementan las *reglas de selección de contenido*. En estas reglas recae la responsabilidad de decidir qué segmentos de texto, o más concretamente, qué tipos de satélites RST, se deben elegir para que se satisfagan los aspectos de usuario requeridos.

Tal y como muestra la figura 1, el documento maestro consta de n segmentos de texto. El tamaño y estructura de estos segmentos viene determinada por la *variante RST de grano grueso*. Por acción de los filtros y de acuerdo a los valores de aspectos de usuario dados, algunos de esos segmentos se eliminan y otros no. Como resultado del proceso de filtrado se obtiene un nuevo documento, denominado *vista de documento maestro*, que contiene m segmentos. Formalmente, n es un número mayor que m donde los segmentos i , j hasta m pertenecen al conjunto de segmentos originales del documento maestro. Esta selección de segmentos la realizan los filtros en virtud de lo expresado en las reglas de selección de contenido, de acuerdo al perfil de usuario dado.

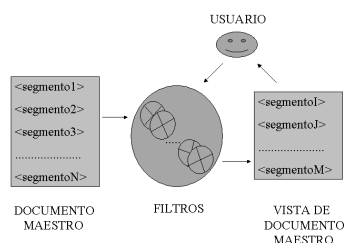


Figura 1: Proceso de filtrado

La contribución principal de esta tesis es el modelo computacional basado en un algoritmo de selección de contenido, que ha sido implementado en dos capas: los filtros, por un lado, y el mecanismo de activación, por otro.

- La selección de contenidos del documento maestro se lleva a cabo mediante la intervención de los *filtros XSL en cascada* que, en sucesivas fases, y junto con el mecanismo de activación Javascript, implementan las reglas de selección de contenido. Los filtros actúan sobre el documento indistintamente del idioma en el que estén escritos. Todos los valores dados por los aspectos de usuario se interrelacionan de forma combinada, y se ha demostrado que esta estrategia es eficaz tanto para la implementación de las reglas de selección de contenido como para su posterior modificación y ajuste. Esta estrategia se presentó en el *Second workshop on NLP and XML* (sobre las aplicaciones de XML en el procesamiento del lenguaje natural) que tuvo lugar en el

marco del *Congreso COLING 2002* celebrado en la ciudad taiwanesa de Taipei. También se presentó en el *XVIII Congreso de la SEPLN* de 2002, celebrado en Valladolid.

- La segunda contribución importante de la presente tesis es la modelización del corpus de documentos maestro. La *variante RST de grano grueso en XML* permite la explicitación de la estructura del discurso en etiquetas XML, de acuerdo a una DTD. La variante RST de grano grueso, al segmentar el texto en fragmentos de al menos una frase, proporciona un corpus isomórfico de documentos maestro en inglés, castellano y euskera. Este isomorfismo evita que el aprovechamiento de este corpus por parte del algoritmo de selección de contenido tenga que estar sujeto a las particularidades de cada idioma. Este logro de la tesis se presentó en el *MT Summit VIII* que tuvo lugar en la ciudad de Santiago de Compostela en 2001.
- Finalmente, otras contribuciones relacionadas con la ciencia de la computación son la *modelización del usuario* y las *reglas de selección de contenido*. El modelo de usuario de la tesis se presentó en el *VIII Simposio Internacional de Comunicación Social* que tuvo lugar en Santiago de Cuba, y las reglas de selección de contenido en la *SEPLN 2002* y el *workshop* sobre NLP y XML a los ya se ha hecho referencia anteriormente.

En resumen, a la vista de los resultados presentados en la tesis se puede concluir que la estrategia para la generación de documentos es una contribución novedosa y ha demostrado ser eficaz, plausible y adecuada para el aprovechamiento de los documentos maestro etiquetados en XML de acuerdo a la variante RST de grano grueso. El modelo computacional aportado es también un planteamiento tecnológico que permite la implementación de las reglas de selección de contenido y el modelo de usuario.