3ra Conferencia Internacional de Biblioteca Digital y Educación a Distancia

Expandiendo las Fronteras del Conocimiento

Mérida – Venezuela • 12 y 13 de noviembre 2009

PLATAFORMA DE RECOLECCIÓN EN FUENTES HETEROGÉNEAS DE LA WEB Y SU APLICACIÓN A LA ORGANIZACIÓN DE UN REPOSITORIO SEMÁNTICO EN SEDICI: PRELIMINARES

Marisa Raquel De Giusti, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires y Proyecto de Enlace de Bibliotecas, marisa.degiusti@sedici.unlp.edu.ar

Ariel Sobrado, Proyecto de Enlace de Bibliotecas, asobrado@sedici.unlp.edu.ar

Agustín Vosou, Proyecto de Enlace de Bibliotecas, agustinvosou@sedici.unlp.edu.ar

Gonzalo Luján Villarreal, Consejo Nacional de Investigaciones Técnicas y Científicas y Proyecto de Enlace de Bibliotecas gonetil@sedici.unlp.edu.ar

Resumen

Se presenta una plataforma de recolección destinada a relacionar y unificar información disponible en distintos lugares de la Web _que siguen diferentes convenciones_ para crear un repositorio temático que puedan navegar los usuarios. La plataforma será usada en el Servicio de Difusión de la Creación Intelectual (SeDiCI) y utiliza de manera combinada ontologías y tesauros para brindar información mejor clasificada.

Actualmente, la información está diseminada en recursos de la Web y los motores de búsqueda tradicionales le devuelven al usuario listas rankeadas sin proveer ninguna relación semántica entre documentos. Los usuarios pasan gran cantidad de tiempo para vincular unos documentos con otros y saber cuáles atacan el dominio completo del problema; recién al localizar las semejanzas y las diferencias entre fragmentos de información éstas se trasladan a su trabajo y sirven para la creación de nuevo conocimiento.

La plataforma propuesta separa los módulos de funcionamiento de los diferentes dominios de interés (temas) para permitir su utilización en distintas áreas de conocimiento. El desarrollo incluye un agente que recorre las URLs almacenadas en una base de datos, un módulo capaz de reconocer las páginas marcadas, interpretar las etiquetas y proveer las reglas para extraer la información y guardarla en un fichero RDF; tras esta etapa se aplica una homogeneización y la información así transformada se clasifica en función de una ontología de dominio.

La plataforma vuelve más eficientes los procesos de extracción automática y búsqueda de información en fuentes heterogéneas que representan los mismos conceptos siguiendo distintas convenciones.

Palabras Clave: SeDiCI, repositorio semántico, ontologías y tesauros.

PLATFORM FOR COLLECTION FROM HETEROGENEOUS WEB SOURCES AND ITS APPLICATION TO A SEMANTIC REPOSITORY ORGANIZATION AT SEDICI: PRELIMINARIES

Marisa Raquel De Giusti, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires y Proyecto de Enlace de Bibliotecas, marisa.degiusti@sedici.unlp.edu.ar
Ariel Sobrado, Proyecto de Enlace de Bibliotecas, asobrado@sedici.unlp.edu.ar
Agustín Vosou, Proyecto de Enlace de Bibliotecas, agustinvosou@sedici.unlp.edu.ar
Gonzalo Luján Villarreal, Consejo Nacional de Investigaciones Técnicas y Científicas y Proyecto de Enlace de Bibliotecas gonetil@sedici.unlp.edu.ar

Abstract

Presentation of a web collection platform designed to relate and unify information available on different standard web sources with a view to creating a user-browseable thematic repository. The platform will be used at the Servicio de Difusión de la Creación Intelectual (SeDiCI) [Intellectual Creation Diffusion Service] combined with ontologies and thesaurus to provide improved data sorting.

Data is currently spread on web resources and traditional search engines return ranked lists with no semantic relation among documents. Users have to spend a great deal of time relating documents and trying to figure out which ones fully address the issue domain. It is only after locating similarities and differences that information fragments are applied to the user's work, enabling knowledge creation.

The proposed platform sorts out the different theme domain functioning modules to allow their use in various knowledge areas. Development includes an agent that searches data base stored URLs, capable of identifying bookmarked pages, interpreting labels and providing rules for extracting information and storing it in a RDF data file. After this stage, homogenization is applied and transformed information is sorted out according to domain ontologies.

The platform allows for more efficient automatic extraction processes and information search among heterogeneous sources that represent the same concepts using different standards.

Key words: SeDiCI, semantic repository, ontology and thesaurus.