

"M-CAST - Wielojęzyczny System Agregacji Informacji oparty na Wyszukiwarce TRUST"

(projekt programu eContent nr 22249)

Borys Czerniejewski

Biuro Współpracy Międzynarodowej, Infovide S.A., Warszawa

Celem projektu jest stworzenie wielojęzycznego systemu, który umożliwi dostawcom informacji przeszukiwanie i integracje zasobów wielkich kolekcji danych tekstowych i multimedialnych, takich jak biblioteki internetowe, zasoby wydawnictw, agencji prasowych i naukowe bazy danych.

Wielojęzyczny system agregacji informacji (M-CAST), poprzez agregacje rozproszonych danych, dostępnych w różnych formatach, pozwoli na tworzenie nowoczesnych bibliotek cyfrowych. System zostanie przetestowany w dwóch bibliotekach, w których stworzone zostaną multimedialne portale agregujące informacje (M-CAP), bazujące na wcześniej stworzonych systemach i portalach. Będą one umożliwiały znalezienie w wielkich kolekcjach wielojęzycznych danych odpowiedzi na pytania zadawane w jednym z sześciu języków naturalnych. Warstwa prezentacji występująca w tych portalach będzie obsługiwała dane multimedialne i umożliwi prezentacje cyfrowych obrazów starodruków, oryginałów dokumentów, zapisów muzycznych, fotografii, plików wideo itp. Oczywiście indeksowane będą mogły być tylko tekstowe opisy tych obiektów.

Projekt M-CAST będzie wykorzystywał wyniki projektu TRUST - Multilingual Semantic and Cognitive Search Engine for Text Retrieval Using Semantic Technologies (Wielojęzyczny mechanizm semantyczny i kognitywny do wyszukiwania tekstów z wykorzystaniem technik semantycznych - IST-1999-56416) dofinansowanego w ramach piątego Programu Ramowego Badan, Rozwoju i Prezentacji Unii Europejskiej (www.trustsemantics.com). Wyszukiwarka TRUST, umożliwiająca wyszukiwanie w czterech językach (francuskim, polskim, portugalskim i włoskim) dostępna jest dotychczas w jedностanowiskowej wersji na komputery osobiste typu PC. W ramach projektu zostanie ona przeniesiona na serwery pracujące pod kontrola systemu operacyjnego Unix lub Windows. Zasoby językowe zgromadzone w ramach projektu TRUST zostaną zaktualizowane i wzbogacone.

Wykorzystywana dotychczas taksonomia języka zostanie dostosowana do uniwersalnej klasyfikacji dziesiętnej (UDC) wykorzystywanej przez biblioteki na całym świecie do tworzenia katalogów tematycznych. System zostanie wzbogacony o dwa dodatkowe języki: angielski, dodany przez jednego z partnerów projektu TRUST oraz czeski, w ramach projektu M-CAST.

System agregacji informacji M-CAST będzie centralnym elementem portalu agregującego MCAP, który zostanie zbudowany zgodnie z metodyka budowy opartych na wiedzy systemów zarządzania informacją, stworzona przez Infovide S.A. w ramach innego projektu dofinansowanego z 5. Programu Ramowego - ICONS - Intelligent Content Management System - IST-2001-32429 (www.icons.rodan.pl). Portale zostaną wdrożone i przetestowane w dwóch bibliotekach publicznych: w Polskiej Bibliotece Internetowej (www.pbi.edu.pl), której operatorem jest Wojewódzka Biblioteka Publiczna - Książnica Kopernikańska w Toruniu oraz w Bibliotece Narodowej Republiki Czeskiej (www.nkp.cz). Informacje będą mogły być wyszukiwane w wielojęzycznych, cyfrowych zasobach bibliotek za pomocą zapytań sformułowanych w języku naturalnym. Odpowiedzią będą fragmenty tekstów zawierające odpowiedź na pytanie.

Po zakończeniu projektu powstanie produkt ułatwiający zarządzanie wielojęzyczną wiedzą zgromadzona w dużych zbiorach danych tekstowych. Przewiduje się, że znajdzie on zastosowanie w bibliotekach (wyszukiwanie informacji, katalogi tematyczne), zarządzaniu zbiorami (pozyskiwanie i agregacja danych, statystyki wykorzystania, poprawa jakości danych), bibliograficznych bazach danych, serwisach informacyjnych (selektywne upowszechnianie informacji, personalizacja, odkrywanie wiedzy) oraz w semantycznych sieciach danych.