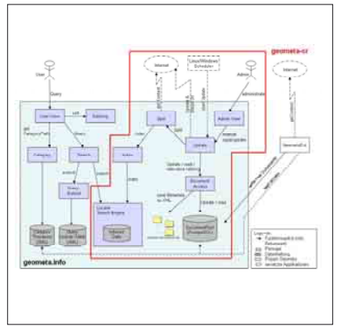
Adrian Höhn
Thomas Schwyn

geometa-cr

Content Refresher für die Suchmaschine geometa.info

Diplomanden	Adrian Höhn, Thomas Schwyn
Examinator	Prof. Stefan F. Keller
Experte	Claude Eisenhut, Eisenhut Informatik AG, Burgdorf
Themengebiet	Suchmaschinen, Relevance Ranking und Web-GUI



Übersicht Softwarekomponenten geometa-cr

Aufgabenstellung: Suchmaschinen gehören zum Fachbereich der Informationssysteme. Geometa.info ist eine Spezial-Suchmaschine, die seit 2003 an der HSR betrieben wird. Sie dient dem Suchen von kartografisch aufbereiteten Geoinformationen, wie beispielsweise Online-Karten und Stadtpläne oder thematische Karten aus Deutschland, Österreich und der Schweiz. Die Webdokumente und Webdienste werden zurzeit noch von Hand verwaltet - unter anderem auch, um die Qualität zu garantieren. Einmal eingetragene Seiten werden nicht mehr überprüft oder neu indexiert.

Für das periodische Aktualisieren der Inhalte und Dienste ist die Updater- oder Refresher-Komponente verantwortlich. Sie soll zudem die Webdokumente und Webdienste auf Verfügbarkeit prüfen und das Relevance Ranking anpassen.
Die Anwender der Ergebnisse der Arbeit sind Administratoren der Suchmaschine geometa.info.

Ziel der Arbeit: Die Updater-/Refresher-Komponente wird periodisch von einem Task-Scheduler aufgerufen. Es gibt eine Datenschnittstelle zur Webcrawler-Datenbank (GeometaBot-DB). Aus dieser werden URLs gelesen, um sie zu prüfen und gegebenenfalls ihren Inhalt zu aktualisieren, bzw. neu indexieren zu lassen. Beim Aktualisieren soll das Relevance Ranking aller betroffenen URLs angepasst werden.
Die Verwaltung der Konfiguration und des Task-Schedulers soll im Rahmen dieses Projekts implementiert werden (JSP). Das Resultat der Arbeit soll dokumentiert und in die Versuchsplattform auf geometa.hsr.ch integriert werden (offiziell ist die Suchmaschine über www.geometa.info erreichbar).
Dienste und GIS-Formate werden direkt indexiert. Vom GeometaBot neu gefundene Dokumente vom Typ HTML, PDF und Webmap-Links müssen über die aktualisierte Administrations-Komponente verwaltet werden. Weitere Programmschnittstellen bestehen zum RSS-Format (z.B. www.geowebforum.ch) sowie zum Speichern von Geometadaten-Dateien (Realisierung des Metadaten-Protokolls mit www.geoshop.ch mit GM03). Eine Integration von GM03-Dateien (Import) sowie einer Zusatzdienstleistung zur Messung der Qualität von Geodiensten ist wünschenswert.

Lösung: Die Updater-/Refresher-Komponente wurde in Java implementiert, das dazugehörende Web-GUI mit JSP. So können die beiden Komponenten nahtlos ineinander greifen.
Beim Relevance Ranking fiel die Wahl auf den PicASHOW-Algorithmus. Dieser berechnet die Relevanz einer URL alleine aufgrund von Inlinks und lässt Outlinks aus der Bewertung weg. Dies ist notwendig, da Webdienste und Multimedia-Daten keine Outlinks aufweisen und zudem praktisch, weil damit ein einheitliches Relevance Ranking sichergestellt werden kann.

Übersicht Softwarekomponenten geometa-cr

Weitere Infos:
Detaillierte Informationen zu den einzelnen Funktionen der Suchmaschine findet man unter www.geometa.info/about.jsp
Die Suchmaschine selbst ist unter www.geometa.info erreichbar.