DANDELON.COM - KOLLABORATIVE, MASCHINELLE CONTENT-ERSCHLIESSUNG UND SEMANTISCHES, MULTILINGUALES RETRIEVAL

MANFRED HAUER

ABSTRACT

In dandelon.com werden im Gegensatz zu den bisherigen Federated Search-Portal-Ansätzen die Titel neu mittels intelligentCAPTURE dezentral und kollaborativ erschlossen und inhaltlich stark erweitert. intelligentCAPTURE erschließt maschinell bisher Buchinhaltsverzeichnisse, Bücher, Klappentexte, Aufsätze und Websites, übernimmt bibliografische Daten aus Bibliotheken (XML, Z.39.50), von Verlagen (ONIX + Cover Pages), Zeitschriftenagenturen (Swets) und Buchhandel (SOAP) und exportiert maschinelle Indexate und aufbereitete Dokumente an die Bibliothekskataloge (MAB, MARC, XML) oder Dokumentationssysteme und an dandelon.com. Die Daten werden durch Scanning und OCR, durch Import von Dateien und Lookup auf Server und durch Web-Spidering/-Crawling gewonnen. Die Qualität der Suche in dandelon.com ist deutlich besser als in bisherigen Bibliothekssystemen oder deren Kopplungen durch Federated Search-Portale. Die semantische, multilinguale Suche mit derzeit 1,2 Millionen Fachbegriffen trägt zu den guten Suchergebnissen stark bei.

UNIQUE SELLING POINTS

1. dandelon.com ist ein internationales, offenes Portal für Bibliotheken und Informationslieferanten. Es überwindet damit die geografischen und hoheitlichen Grenzen der Verbundsysteme, da wissenschaftliches Denken sich keine Grenzen auferlegt.

2. IntelligentCAPTURE wird in Bibliotheken installiert, um dort dezentral gedruckte oder elektronisch verfügbare Inhalte zu akquirieren, aufzubereiten und zu indexieren. Dieser Ansatz erbringt eine tiefere und vollständigere Indexierung der Medien und damit eine wesentlich verbesserte Findbarkeit im lokalen Katalog und im kollaborativ produzierten „dandelon.com“ (1). Hier geht AGI wie Google oder Amazon weg von der bisherigen Bibliothekspraxis, der sehr knappen, sehr abstrakten, stark formalisierten Erschließung. Ca. 200 Bücher lassen sich so pro

Arbeitstag von Hilfskräften erschließen. Durchschnittlich werden 4,1 Seiten pro Buchinhaltsverzeichnis gescannt, die restliche Verarbeitung ist voll automatisiert. Im Gegensatz zu Google oder Amazon nutzt AGI hier eine anspruchsvolle computerlinguistisch-statistische Analyse der Dokumenttexte. Sie erkennt Morpheme, syntaktische Strukturen und semantische Beziehungen.

3. Alle Bibliotheken haben eine sehr einfach zu bedienende Export-Schnittstelle nach dandelon.com und nur eine Import- und eine Exportschnittelle zum jeweiligen lokalen Bibliothekssystem. Darüber hinaus haben sie für die Suche eine leicht zu installierende Verbindung vom Katalog nach dandelon.com und umgekehrt eine einfache Rückverlinkung von dandelon.com in Katalog durch eine eindeutige Medien-ID

4. dandelon.com hat folgende Retrievalfunktionen:
   - AND, OR, SENTENCE, PARAGRAPH, Klammerung
   - Trunkierung Links, Rechts, Mitte und Fuzzy-Search
   - Stemming und Grundformreduktion (deutsch)
   - Query-Expansion um Synonyme, Übersetzungen, Unterbegriffe
   - grafische Navigation in semantischen Strukturen (Thesaurus, Topic Maps, Ontology, Klassifikation, Taxonomie).
   - Die Feldsuche und die Termgewichtung sind für Benutzer in diesem Interface nicht unmittelbar verfügbar, sondern im Hintergrund implementiert.
   - Der Benutzer kann sich für die Suche sowohl seine Strategie aussuchen von hoher Precision bis hohem Recall.
   - Führt eine Einstiegsfrage nicht zu einem Ergebnis optimiert dandelon.com automatisch die Query solange, bis ein Ergebnis gefunden wird oder die Frage technisch nicht weiter optimiert werden kann. Derzeit sind hier bis 16 Einzelabfragen im Rahmen der automatischen Optimierung hintereinander geschaltet.

5. dandelon.com kann auf den Teilbereich einer Bibliothek suchen oder über den Bestand aller Bibliotheken gleichzeitig. Von dort kann direkt der zugehörige Datensatz in der jeweiligen Bibliothek geöffnet werden über die AGI Unique ID. Innerhalb von dandelon.com werden nach der Auswahl von Titeln aus der Suchergebnisanzeige bei Auswahl die verfügbaren PDFs oder gelinkten URLs unmittelbar geöffnet und zugleich die Metadaten dazu angezeigt.

6. Die Query-Expansion basiert auf derzeit 1,2 Millionen Termen in 16 Sprachen, die im Thesaurusentwicklungsprogramm IC INDEX gespeichert sind und dynamisch weiterentwickelt werden können. Die Mehrzahl der Thesauri

stammen aus dem öffentlichen und privatwirtschaftlichen Dokumentationswesen. Der Benutzer kann sich gezielt einzelne Thesauri für seine Recherche auswählen, um Ambiguitäten zu vermeiden, denn viele Terme werden je nach Wissensdomäne ganz anders inhaltlich belegt. Noch nicht verfügbar, aber geplant ist die Einschränkung auf Sprachen in der Suche und in der Anzeige.

7. Die Anzeige der Query-Expansion musste zum Schutz der Urheber der Thesauri deaktiviert werden. Eine umfangreiche grafische Navigation einzelner Begriffe ist aber separat verfügbar.

8. Die Suchergebnisse werden per default nach Ranking sortiert und durch Highlighting betont. Zusätzlich kann nach Jahr oder Autor sortiert werden.

   dandelon.com versteht sich als thematisch übergreifende Lösung, auf der gleichen Technik basiert das fachspezifische „Portal Informationswissenschaft"

9. dandelon.com kann prinzipiell mit der Funktion „Suche in Bibliothek fortsetzen" direkt andere Portalprogramme wie Electra, Metalib und andere ansprechen und umgekehrt sind für 2006 Webservice-Schnittstellen geplant, die wiederum dandelon.com integrieren können.

Der Ansatz unterscheidet sich von Query Broker/Federated Search-Systemen, denn es wird nur in einer technisch homogenen Datenbank (oder Datenbanken) mit allen Retrieval-Funktionalitäten gesucht, es bedeutet weit weniger unterschiedliche Formate, kaum Dubletten-Problem, kaum Updating-Problem bezüglich fremder Abfrageschnittstellen, da jede Bibliothek ihre Schnittstelle selbst einrichten und pflegen kann.

LITERATURVERZEICHNIS

Hauer, Manfred: Benchmarking Literatursuche 2005: Vergleich der Retrieval-Leistungen von Bibliothekskatalogen gegen erweiterte und neue Konzepte. ABI-Technik 2005, 25: S 295-301

Weitere Titel zu dandelon.com im Volltext im Portal Informationswissenschaft unter http://www.agi-imc.de/isearch/is_dgi.nsf unter Artikel.

ADRESSE DES AUTORS

Dipl. Inf.wiss. Manfred Hauer M.A.
AGI - Information Management Consultants
D-67433 Neustadt / Weinstrasse, Mandelring 238 b
E-Mail: Manfred.Hauer@agi-imc.de
Web: http://www.agi-imc.de