In: Bekavac, Bernard; Herget, Josef; Rittberger, Marc (Hg.): Informationen zwischen Kultur und Marktwirtschaft. Proceedings des 9. InternationalenSymposiums fur Informationswissenschaft (ISI 2004), Chur, 6.-8.Oktober2004. Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 465 – 468

Dandelon.com – ein internationales Wissenschaftsportal mit morphosyntaktischer Indexierung und semantischem Retrieval

Manfred Hauer

AGI - Information Management Consultants & FHS Burgenland,
FB Informationsberufe
Manfred.Hauer@agi-imc.de, http://www.dandelon.com

Zusammenfassung

Web-OPACs mit Human-Indexierung fallen Retrieval-Tests deutlich hinter maschinelle Verfahren zurück. intelligentCAPTURE saugt Content über Scanning, File-Import und Web-Spidering ein und indexiert nach morphosyntaktischen und semantischen Verfahren. Neben Bibliothekssystemen übernimmt dandelon.com den Content und die Indexate. Dandelon.com ist öffentlich und kostenlos zugänglich für Endbenutzer, Austauschzentrale und Katalogerweiterung für angeschlossene Bibliotheken. Die Kosten sind gegenüber der Humanerschließung wesentlich geringer bei zugleich deutlich höherem Wirkungsgrad in der Recherche.

Abstract

In a benchmarking between human indexing in library catalogs and machine indexing in the open service dandelon.com the machine approach succeeded. It is based on intelligentCAPTURE, an program capturing content via scanning, file import or web spidering. Text is indexed by a build in morphosyntactical, semantical engine. Results are pasted to library catalogs as well as to dandelon.com. It is a free of charge service and exchange center for linked libraries. Cost are much below human indexing approach but with much better retrieval results.

Ca. 90 % aller Bibliotheksrecherchen durch Benutzer sind Themenrecherchen. Ein Anteil dieser Recherchen bringt kein Ergebnis. Die Gründe hierfür wurden immer wieder untersucht: Plural- anstelle Singularformen, zu spezifische Suchbegriffe, Schreib- oder Bedienungsfehler. Ungenügend untersucht sind Recherchen, die nicht mit einer Ausleihe enden, denn auch dann kann in vielen Fällen von einem Retrieval-Mangel ausgegangen werden. Schließlich: Von den ausgeliehenen Büchern werden nach Einschätzung vieler Bibliothekare 80 % nicht weiter als bis zum

Inhaltsverzeichnis gelesen (außer in Präsenzbibliotheken) - und erst nach Wochen zurückgegeben – sind somit für andere nicht zeitnah ausleihbar, totes Kapital. Die Effizienz heutiger Bibliotheken ist suboptimal: hohe Kosten bei sehr niedrigem Ertrag. Studenten und Wissenschaftler reagieren darauf, sie ignorierend zunehmend die Institution Bibliothek. Bibliotheken (als Funktion) sind jedoch unverzichtbar für die wissenschaftliche Kommunikation. Deshalb geht es darum, Wege zu finden und auch zu beschreiten, welche die Schätze von Bibliotheken (als Institution) effizienter an die Zielgruppe bringen. Der Einsatz von neuem Content, neuen Erschließungsmethoden, Information Retrieval-Technologie und internationale Vernetzung sind Ansätze dazu.

1 Bibliometrischer Retrieval-Test: OPAC gegen dandelon.com

Im Rahmen einer Lehrveranstaltung „Information Retrieval in Bibliotheken" mit Studenten der FH Darmstadt wurden empirisch die Wirkungen von der Content-Generierung und Indexierungsmaschine intelligentCAPTURE an der Vorarlberger Landesbibliothek und in den Daten der Vorarlberger Landesbibliothek unter dandelon.com in einem ersten Experiment evaluiert. Weitere klassische Universitätskataloge, Verbundkataloge und KVK wurden im Vergleich dazu genutzt. Normale OPACs auch von den Verbundkatalogen fielen bei den Recherchen fast immer durch, leicht besser schneidet Bielefeld ab, weil durch die technische Aufbereitung der Katalogdaten durch das Retrieval-System FAST Ähnlichkeit und Ranking möglich sind. Doch auch dort erwies sich die klassische, aber doch sehr kurze Indexierung als problematisch. Eine etwas breitere Human-Indexierung an der Vorarlberger Landesbibliothek wirkte sich deutlich positiver aus, noch wirksamer sind die zusätzlichen maschinell generierten Indexate, die in ALEPH ergänzt wurden. Doch der reiche Content, die Indexierung und die automatische Thesaurusunterstützung in dandelon.com bringen relevante Titel auf die erste Seite der Suchergebnisse. Ein sehr klares Ergebnis zugunsten maschineller Indexierung und semantischen Retrievals.

Dieser kleine Test zeigt schon im Ansatz, dass auch bei sehr genauer Kenntnis eines Thema eine systematische treffsichere Suche in OPACs bis heute nicht möglich ist. Die Indexierungssprachen sind zu grob, der Indexierungsumfang zu eng, die Menge der indexierten Titel zu niedrig, Volltexte nicht suchbar. Obwohl alle Studenten vier oder mehr Semester Informations- und Bibliothekswissenschaften hinter sich hatten und durchaus

geschickt und professionell recherchierten, konnte man im Ansatz nicht mehr aus diesen klassischen Katalogen herausholen.

Diese Erkenntnis ist für Insider nicht wirklich neu. Nur bislang konnte man sich hinter dem allgemeinen (aus informationswissenschaftlicher Sicht veralteten) „State-of-the-Art" gut verstecken. Doch warum ist dandelon.com wesentlich effizienter als alle getesteten OPACs?

2 Vom Projekt zum Produkt intelligentCAPTURE

Wegen dieses Mangels hat die Vorarlberger Landesbibliothek zusammen mit AGI - Information Management Consultants – ein Spin-off der Informationswissenschaft in Konstanz 1983 - ein Projekt vor zwei Jahren begonnen, das über den Begriff „Kataloganreicherung" deutlich hinausgeht. Daraus entstand das Produkt intelligentCAPTURE, vor einem Jahr das Produkt intelligentSEARCH und seit Frühjahr 2004 das internationale wissenschaftliche Portal dandelon.com. Hinter diesem internationalen wissenschaftlichen Portal steckt intelligentSEARCH und dahinter das Produktionssystem intelligentCAPTURE, dahinter linguistische Verfahren (CAI-Engine) und Thesauri (IC INDEX).

intelligentCAPTURE versteht sich als „Durchlauferhitzer", um Content „heiß" zu machen und an beliebige Zielsysteme zu übergeben. Über 20.000 Bücher bzw. deren Inhaltsverzeichnisse sind in dandelon.com (August 2004) mittels Scanning, OCR und PDF-Konvertierung aufbereitet und der Text des jeweiligen Dokumentes maschinell mit der integrierten CAI-Engine (Computer Aided Indexing) inhaltlich indexiert worden. Durch morphosyntaktische, semantische, heuristische und statistische Verfahren der Textanalyse werden inhaltsbeschreibende Metadaten zu den jeweiligen Dokumenten ergänzt. intelligentCAPTURE übergibt diese Metadaten an Bibliotheksmanagementsysteme, die meist auf relationalen Datenbank-Management-Systemen beruhen. Dadurch kann auch in diesen OPACs besser recherchiert werden und es können die Inhaltsverzeichnisse von Büchern, Texte von Aufsätzen angezeigt werden. Mittlerweile hat intelligentCAPTURE nicht nur einen Scan/OCR-Workflow, es erkennt auch automatisch ob Image-Dateien, PDF-Images oder eigentliche PDFs als Dateien angeliefert werden. Jeweils wird der Text richtig extrahiert - auch bei mehrspaltigen Texten. Bei der Textextraktion können einzelne Textbereiche wie Zusammenfassungen oder Dokumentenschlüssel wie die internationale DOI erkannt und extrahiert werden. Über eine URL-Eingabe und Spider-Settings können Seiten einmalig oder periodisch akquiriert werden. Das geht auch mit Listen: So wird die

Zeitschriftenliste mit Links auf Artikel von eJournals von Swets Blackwell wöchentlich automatisch ausgewertet: die Artikel im Internet gespidert, indexiert und das Indexat in dandelon.com publiziert. Das Spidering von Open Archives steht auf der Agenda. Ähnliches ist mit Forschungsinstituten ab Herbst im Einsatz.

intelligentCAPTURE hat eine CAI-Version für allgemeine Bibliotheken und kann ebenso für spezielle Domänen/Themenfelder aufbereitet werden. Für Medizin, Technik, Wirtschaft stehen spezifische CAI-Versionen bereit – andere Domänen sind in Arbeit. Bei der Texterkennung fallen laufend neue Begriffe auf, die statistisch ausgewertet und intellektuell die jeweilige semantische Basis wie Klassifikationen, Thesauri, Topic Maps, semantischen Netze integriert werden können. Diese Netze werden in IC INDEX gepflegt und direkt an die CAI-Engine exportiert.

3 Dandelon.com als Portal und Verteilzentrum

intelligentSEARCH basiert auf IBM Lotus Notes & Domino und nutzt die dort integrierte GTR-Engine (Global Text Retrieval). Sie ist eine n-Gram-Engine mit Stemming, Fuzzy-Search, Feldsuche, numerische Suche, Datumssuche, Termweighting, boolesche Operatoren, Abstandsoperatoren und kann über mehrere Datenbanken gleichzeitig suchen. intelligentSEARCH erweiteret die GTR-Funktionen. So sind bislang über 360.000 Fachbegriffe aus verschiedenen Themenfeldern in Form von semantischen Netzstrukturen zur Suche parallel geschaltet. Daraus resultiert eine automatische Erweiterung und teilweise auch Übersetzung der Anfrage. Optional kann jeder Benutzer diese „Topic Maps" über Flash-Visualisierung sehen, darin navigieren und für die Suche auswählen. Eine Plural- bzw. Wortformkonvertierung der Query auf die Grundform ist für Deutsch integriert. Stemming für andere Sprachen.

Dandelon.com als Service für Endbenutzer basiert auf intelligentSEARCH und ist zugleich internationales Austauschzentrum für Inhalte, die andere Bibliotheken schon erschlossen haben. Jede angeschlossene Bibliothek kann dandelon.com als zusätzliches Front-End zum eigenen Katalog nutzen. Alternativ kann intelligentSEARCH auch für andere Portale genutzt werden - das „Portal Informationswissenschaft“ ist ein erstes Beispiel (www.dgi-info.de).