Ägyptischer Bruch? Knowledge Discovery in Bibliotheken

Dipl.-Inf.wiss. Manfred Hauer M.A.
AGI-Information Management Consultants
D-67544, Neustadt/Weinstraße, Deutschland
Manfred.Hauer@agi-imc.de

Abstract

AGI designed an solution for improving searchabilty of books in libraries. Table of contents are scanned, converted to text via OCR and indexed automatically with a linguistic approach. Relevance given to indexing terms is combined with the ranking technology of GTR form IBM. Indexing as well as searching is using thesauri – in the retrieval interface a graphical visualization of the semantic structures is supported. Query expansion (thesaurus, machine translation) can be used internally as well as externally in library catalogs or Google.

1 Einführung

Das Internet bildet trotz seiner Größe nur einen Teil des verfügbaren Weltwissens ab. Bibliotheken erfüllten traditionell diese Funktion, fallen aber in der Wahrnehmung zunehmend zurück. Ihre Inhaltserschließung ist gegenüber der Volltexterschließung im Internet mager - obgleich sie auch ihre Vorzüge hat: intellektuell kontrolliert und sprachlich standardisiert.

2 Datenaufbereitung

Durch Scanning von Texten innerhalb von **intelligentCAPTURE** oder außerhalb werden Images generiert oder angeliefert, die automatisch durch das OCR-Programm Adobe Acrobat Capture 3.0 - voll integriert in intelligentCAPTURE - in einen Textfile und ein suchbares PDF-Image konvertiert werden. Die jeweilige Textdatei wird je nach Anwenderprofil, Dokumenttyp (meist Inhaltsverzeichnisse oder Aufsätze) und Sprache und weiterer Anforderungen unterschiedlich gefiltert. Zugehörige Workflows sind frei definierbar (verteilte Umgebung und Multithreading). Alternativ zu gescannten Images ist ein Textstrukturerkennungs- und Extraktionsverfahren für "normale", "gedruckte" PDF-Dateien mit Spaltensatz (typischerweise eJournals) für Acrobat 6.0 in Arbeit.
Der über OCR oder Extraktion akquirierte Text eignet sich für die maschinelle Indexierung. XML- und andere Textimportformate sind alternativ möglich.

3 Inhaltserschließung

Innerhalb von intelligentCAPTURE läuft die Indexierung. Mit morphosyntaktischen Verfahren, entwickelt am IAI in Saarbrücken, werden Worte, Sätze, Phrasen in den Texten analysiert, semantische Klassen zugeordnet, gewichtet und mit Thesauri verglichen. Die Produktbezeichnung für die maschinelle Indexierung "CAI-Engine" steht für Computer Aided Indexing. Relevante Begriffe werden gewichtet und jeweils in der Grundform ausgegeben. Geografische Begriffe, Namen von Personen, Produkten oder Organisationen werden durch Wortlisten und Algorithmen gefunden. In der Regel erfolgt eine monolinguale Erschießung, also Beschreibung in der gleichen Sprache. Eine bilinguale Version ist im Einsatz für den Bereich Technik.

4 Information Retrieval

Fertig indexierte und evtl. human kontrollierten Daten und PDF-Dateien werden in Bibliotheksmanagement-Systeme übertragen und sind über den jeweiligen OPAC (Online Public Access Catalog) suchbar. Meist sind hier die Retrieval-Funktionen nicht sehr weit fortgeschritten. Deshalb wurde **intelligentSEARCH** als zentrale Lotus Domino-Datenbank für alle Bibliotheken mit der Retrieval-Engine (GTR) ergänzt. Diese ist öffentlich und soll mittelfristig alle neuen Sachbuchtitel der angeschlossenen Bibliotheken enthalten. Die Oberfläche strebt eine Ähnlichkeit zu Google an (Vertrautheit). Die Termweights der CAI-Engine werden je nach Dokumentgröße zur Multiplikation relevanter Terme in einem unsichtbaren Feld genutzt. Die N-Gram-Engine GTR gibt somit den Dokumenten einen höheren Ranking-Wert und nutzt zusätzlich Häufungen gleicher Terme im PDF-Volltext. IDF, um hochfrequente Terme wieder zu relativieren, folgt bald. Sortierungen der Ergebnismengen u.a. nach Hauptschlagworten haben einen Effekt wie mathematische Clustering-Verfahren. Nur die "Top 10" Schlagworte werden gezeigt. Query-Expansion über mehrere Thesauri (derzeit 120.000 Terme mit gerichteten und benannten Kanten, mehrsprachig) ist per Default zugeschaltet und auch grafisch navigierbar. Diese Visualisierung basiert Flash und dem Thesaurus-Programm IC INDEX. Eine Ad-hoc-Übersetzung von Queries durch Linguatec eTS bzw. IBM Translation Server ist optional.

References

http://www.agi-imc.de/intelligentSEARCH.nsf
http://www.agi-imc.de - siehe Publikationen