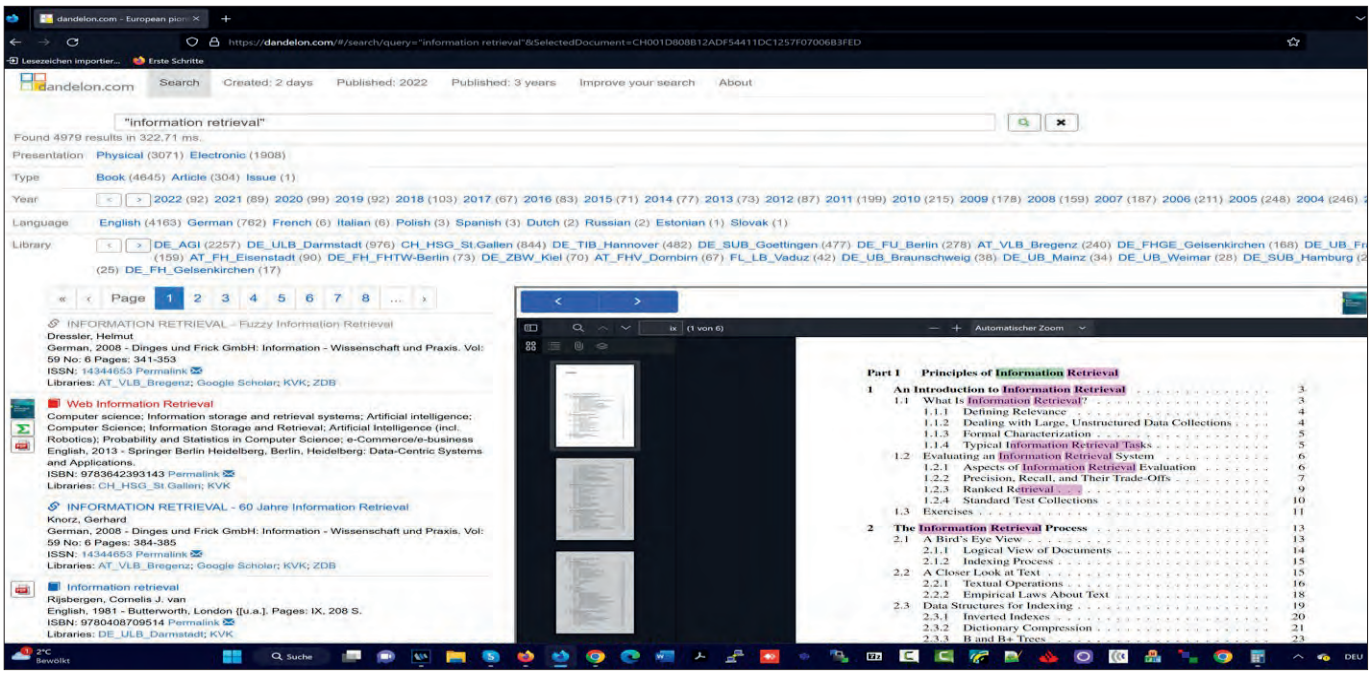
20 Jahre Community für maschinelle Inhaltserschließung in Bibliotheken

Dandelon wird 20 Jahre

Manfred Hauer

Bei der Entstehung der Datenbanktechnologien führte Information Retrieval lange ein Schattendasein, ging es doch primär in der frühen IT um Lagerverwaltung und Buchhaltung – und Briefe schreiben. Die Entwickler von Bibliotheksverwaltungsprogrammen entschieden sich folgerichtig für diese Datenbanktechnologien, vorwiegend relationale Datenbank-Management-Systeme – mit kurzen, fixen Feldlängen. Gegenüber Karteikarten war das ein großer Sprung – die Logik dahinter blieb jedoch ähnlich: Records/Karteikarten und Tabellen/Sortierreihenfolgen. Information Retrieval, die Suche in wenig strukturierten Texten, seit den 50er Jahren bekannt und früh in der Dokumentation von meist Aufsätzen genutzt, erfuhr erst mit dem Einstieg und Erfolg von Google und deren zusätzlichem Page-Rank-Algorithmus (vergleichbar den Ranking-Ansätzen des Science Citations Indexes von Garfield, der Gewichtung von Autoren nach Zitationshäufigkeit, ab 1964 digital) den nötigen Take-off, also zwischen etwa 1996 bis 2000. Teil des Information Retrieval ist stets die Textanalyse, mindestens Stemming, die Grundformreduktion, die in der englischen Sprache leicht zu implementieren ist, bereitet im Deutschen erhebliche Schwierigkeiten.

AGI, das Unternehmen des Autors, nutzte seit 1983 Information-Retrieval-Programme, Thesaurus-Programme zur intellektuellen Terminologiekontrolle und Programme zur maschinellen Textanalyse. Die Vorarlberger Landesbibliothek war ein Pionier bei der Nutzung von Bibliotheksverwaltungsprogrammen und 2001 hatten Bibliothekar Karl Rädler und der Autor die Idee, Inhaltsverzeichnisse zu scannen und daraus mit linguistischer Textanalyse die relevanten Worte zu normalisieren, zu extrahieren und zu gewichten. In einer relationalen Datenbank-Umgebung, konkret im Bibliotheksmanagement-System ALEPH, entstanden Suchergebnisse auf Basis von 10-15 % des Textes, das war einer Volltextrecherche ohne Gewichtung ähnlich. Da die Recherche innerhalb des Produktionssystems intelligentCAPTURE, es basiert auf Notes mit integrierter Retrieval Engine, hinsichtlich des Rankings der Suchergebnisse besser war und weil zwei Bibliotheken in der Nachbarschaft (Landesbibliothek Liechtenstein und Universität St. Gallen) und eine erste in Berlin (HTW) auch in die Produktion eingestiegen waren, kam die Idee des automatisierten Datenaustauschs auf. Die Eingabe des Barcodes zog sofort aus dem Bibliothekssystem die bibliografischen Angaben und nutzte daraus die ISBN zur Suche



in der gemeinsamen Verbunddatenbank, lieferte die vorhandenen Inhaltsverzeichnisse und die Textdatei für die lokale Inhaltsanalyse. Die Analyse und die Art und Menge der übernommenen maschinell generierten Deskriptoren bestimmt jede Bibliothek selbst.

Der Nutzen war für die Bibliothekarinnen/Bibliothekare und Leserinnen/Leser in den produzierenden Bibliotheken offensichtlich: Titel wurden jetzt gefunden, die vorher zwar relevant, aber mit verbaler Recherche kaum auffindbar waren. Die Inhaltsverzeichnisse konnte man sofort am Bildschirm browsen, sparte sich das Suchen der Bücher in den Regalen oder gar Anforderungen aus dem geschlossenen Magazin.

Der Autor öffnete diese dandelon.com gemeinsame Verbunddatenbank von Anfang für die akademisch interessierte Öffentlichkeit im Internet. Im heißen Sommer 2003 wurde diese Lösung erstmals auf der IFLA in Berlin gezeigt. 2004 folgte eine eigene Domäne „dandelon.com". „dandelon" ist eine Wortschöpfung in Anlehnung an den englischen Löwenzahn (dandelion), wie die Löwenzahn-Schirmchen den Samen so soll dandelon.com das medial gespeicherte Wissen in alle Welt tragen und neues wachsen lassen.

2005 schloss die Leitung der Verbundzentrale des GBV einen Kooperations- und Hostingvertrag zur Nutzung der Inhaltsverzeichnisse (PDF-Dateien). Bislang sind über 3 Millionen Inhaltsverzeichnisse zwischen Florenz und Trondheim aus über 80 Sprachen mit intelligentCAPTURE produziert worden. Nur der primär akademische Anteil landete in dandelon.com. 2022 kamen 64.015 Inhaltsverzeichnisse hinzu, von Print- und E-Books, Tendenz leicht wachsend. Bei E-Books wird die Springer-Kollektion komplett erfasst, kleinere Verlage oder digitalisierte Medien fremder Institutionen/Bibliotheken werden manuell pro Titel erfasst – es sind nur 2 Eingaben nötig, Systemnummer und ein PDF des Inhaltsverzeichnisses in einem Browser-Fensterchen.

Aufsätze

Current Contents (Titel von Zeitschriftenaufsätzen) kam schon 2004 hinzu – seit vier Jahren werden die bibliothekarischen Angaben in intelligentCAPTURE maschinell im gescannten Inhaltsverzeichnis der Hefte erkannt – eine Kombination mehrerer KI-Methoden. Es braucht kein weiteres Training oder Musterauswahl durch die Bibliotheken. Vier Bibliotheken liefern diese Inhalte an den K10plus (UBs in Kiel, Hamburg, Bonn und IAI Berlin). Über 1,3 Mio Aufsatztitel bis hin zum Volltext landeten bislang in Katalogen und dandelon.com.

Inhaltserschließung

Um den technischen Restriktionen relationaler Bibliotheksmanagement-Systeme entgegen zu kommen, lieferte AGI mit intelligentCAPTURE ca. 10 Prozent der Worte aus dem Inhaltsverzeichnis, alle normalisiert auf ihre Grundform (Häusern --> Haus), gewichtete, kontrollierte und freie Deskriptoren, aussagestarke Komposita und Phrasen sowie geografische Angaben und Namensartiges. Die Textbasis kam via Scanning und OCR zustande – seit acht Jahren wird bei nicht-deutschen Texten jeweils ins Deutsche übersetzt. Nach dieser Textinhaltsanalyse können die Indexate und Titel in weitere Zielsprachen übersetzt werden, meist Englisch. Im Ibero-Amerikanischen Institut in Berlin entstehen aus 20 Quellsprachen jeweils sechs Suchsprachen. Verfügbar sind bislang 135 Sprachen über die integrierte Google Translation API. All dies verlangt von den Bibliotheksmitarbeitern nicht mehr als die Eingabe von Barcode oder Systemnummer/PPN und in der ersten Bibliothek das Scannen der internen Titelseite und der 2-3 Seiten des Inhaltsver-

zeichnisses. Bei der hochautomatisierten Erfassung von einzelnen Aufsätzen oder Kapiteln braucht es mehr Kontrolle und Mitwirkung, dennoch ist es wohl die derzeit schnellste Erfassungsmethode – und Abstracts, Volltexte, DOIs lassen sich leicht ergänzen und der Inhalt maschinell erschließen.

Vergleicht man die intellektuelle Erschließung über Normdaten, sie sind zu nicht geringem Anteil identisch oder Varianten der Titelwörter – außer in engagierten (Spezial-) bibliotheken –, dann bietet dieser Ansatz sofort Zugang auch bei neuer Fachterminologie und sehr vielen durch die GND nicht repräsentierten Fachbegriffen sprachübergreifend. Allein durch die Volltext-Indexierung der Inhaltsverzeichnisse wird dies nicht erreicht. Im Zeitalter der Discovery Engines als User-Frontend liefert dieser Ansatz zusätzliche Gewichtung und höhere Precision – solange man nicht zu allgemein oder im Sinne des Retrieval Systems ungeschickt fragt.

Suchapplikation

Die derzeit dritte Generation von dandelon.com basiert auf der Retrieval Engine ElasticSearch und HCL Domino-Datenbanken, etwa eine Sekunde nach dem Hochladen sind die neuen Titel suchbar.

Dandelon.com orientiert sich im Bildschirmdesign an Google Maps. Der erste Treffer wird sofort angezeigt, nicht nur die Liste möglicher Treffer. Es zeigt auf einem FullScreen-Computer-Monitor die Elemente Query, Facetten, Ergebnisliste und das gewählte Dokument gleichzeitig an. In Smartphones muss zwischen Liste und Anzeige gewechselt werden. Die Suchterme werden im Dokument farblich hervorgehoben (das kann leider nur Mozilla Firefox mit Adobe Acrobat), so dass die Relevanz schnell erkennbar wird. Es kann direkt zwischen den Dokumenten geblättert werden, ohne die Suchergebnisliste zu nutzen. Dandelon.com betont den Dialog mit dem Benutzer und platziert deshalb – nicht zu übersehen – die Facetten direkt unter die Suchzeile. Neben inhaltlichen Aspekten ist die Eingrenzung auf die gewünschte Bibliothek zwecks schneller Verfügbarkeit nützlich. Die Nutzerinnen und Nutzer kommen häufig aus anderen Kontinenten, ihnen hilft der Button „KVK“ mit internationalen Bestandnachweisen oder bei Aufsätzen die Weiterleitung zu „Google Scholar“ oder zur „ZDB“. Den Permalink kann man kopieren oder via E-Mail-Link versenden. Wo Abstracts/Summaries verfügbar sind, zeigt das mathematische Summenzeichen den Text in einem zusätzlichen Fenster. Ähnlich die Cover Pages, deren Thumbnails vergrößert werden können.

Power User können unter „Improve your search“ die Features von ElasticSearch einsehen, um über Stringsuche, Trunkierung, Term Boosting, Klammern, Boolesche Operatoren die Suche zu optimieren.

Unter dem Label „Discovery Engine“ entstanden etwa zeitgleich zu dandelon.com andere Retrieval-Lösungen, Googles „Google Books“ sowie von Bibliothekssystemanbietern (ExLibris, OCLC), Content Anbietern (EBSCO) und Bibliotheksorganisationen (GBV und BSZ – K10plus), die nicht nur auf die gescannten Inhaltsverzeichnisse aufsetzen, sondern vor allem auf von den Bibliotheken nicht verwaltetem Content von Verlagen und Open-Access-Repositorien, also ein Entdecken (Discovery) zusätzlicher Wissensressourcen. Dandelon.com ist einer der Pioniere – seine Inhalte sind häufig früher recherchierbar als in anderen Quellen.

Manfred Hauer
hat einen Master in Soziologie und Politikwissenschaft und ein Diplom in Informationswissenschaft, Universität Konstanz. 1983 gründete er AGI – Information Management Consultants, Inhaber von dandelon.com. Das Unternehmen startete mit Informationsvermittlung aus Online-Datenbanken, ergänzte um Software-Vertrieb und entwickelt seit 1994 Software für Information Center in Unternehmen und seit 2004 primär für wissenschaftliche Bibliotheken.
https://agi-imc.de,
manfred.hauer@agi-imc.de