Adaptive automatische Indexierung für komplexe Dokumente

Knorz, Gerhard/Arz, Johannes/; Rostek, Lothar; Steffen, Jan/ /
Fachhochschule Darmstadt; 2GMD /IPSI Darmstadt

Einleitung

Wenn das World Wide Web im Bewußtsein vieler Menschen bereits nachhaltig etwas verändert hat, dann gehört sicher eine neue Einschätzung von "Dokumentenretrieval" dazu: Die Web-Nutzer können nun ganz praktisch erleben, daß der prinzipielle technische Zugang zu einer Information wenig nützt, wenn es nicht gelingt, das Gesuchte von vielen mehr oder wenig ähnlichen, aber letztlich irrelevanten Dokumenten zu separieren. Ein prinzipieller Ansatz zur Verbesserung der Retrievalergebnisse besteht darin, bereits beim Input die Dokumente um inhaltliche Kennzeichnungen anzureichern, die ein gezieltes Wiederauffinden vorbereiten sollen. Der Einsatz professioneller Arbeitsleistung für diesen Zweck der "Indexierens" wird in vielen Fällen nicht zu leisten sein. Der Appell an die Autoren läßt flächendeckend kaum auf Erfolg hoffen. Der Einsatz automatischer Verfahren zur inhaltlichen Erschließung verspricht deshalb in vielen Anwendungsfällen eine vertretbare Lösung.

In diesem Kontext wurde seit 1993 im Rahmen verschiedener Diplomarbeiten der FH Darmstadt gemeinsam mit dem GMD-Institut für Integrierte Informations- und Publikationssysteme (IPSI) daran gearbeitet, ein gleichermaßen anspruchsvolles und praxiserprobtes Verfahren ((Lustig 86, Lück et al. 92)) so in eine portable Software umzusetzen, daß es den Bedingungen gegenwärtiger Anwendungsszenarien entspricht.

Zum Wandel des Dokumentenbegriffs

Die "natürliche" und traditionelle Einheit für das Schreiben, Speichern und Suchen von Informationen ist das *Dokument.* Vor zehn Jahren noch galt für elektronische Dokumente folgender Steckbrief: *Ausschließlich Text, wissenschaftlichtechnische Fachsprache, vermutlich nur Kurzfassung eines Originals, einfache*

Verwaltungsstruktur (Titel, Autor, Quelle, ...), im Textbereich fast strukturlos. Die enorme Entwicklung graphischer Oberflächen, objektorientierter Dokumentarchitekturen und Werkzeuge, billiger Massenspeicher und flächendeckender Netzwerke haben den Dokumentenbegriff grundlegend verändert: *Dokumente enthalten neben Text auch strukturierte Daten, Graphiken, Bilder, evtl. sogar Audio- und Video-Daten. Die Inhalte sind heterogen und keinesfalls nur wissenschaftlich-technischer Art. Die Dokumente haben eine reichhaltige Struktur bis hin zu Hyperlinks im Dokument und über Dokumentgrenzen hinweg.*

Für Unternehmen, die Information und Dokumente als Ressource betrachten, haben *Dokumentenstandards* eine strategische Bedeutung erlangt. SGML (Standard Generalized Markup Language) hat überall dort, wo es um elektronisches Publizieren geht, Verbreitung gefunden und wurde durch HTML, die SGML-basierte Grundlage für Dokumente im World Wide Web, weithin populär. Die Strukturierung mittels SGML erlaubt es, mit Dokumenten so flexibel umzugehen, wie es Datenbanken mit Daten ermöglichen.

Automatische Inhaltserschließung

Es ist Stand der Technik, bei der Suche nach Dokumenten Bedingungen zu formulieren, die sich auf das Vorkommen einzelner Wortformen im Dokument beziehen:

> *"Suche alle Dokumente, in denen* Indexierung *und* (automatische *oder* manuelle) *vorkommen* ".

Man kann dies so auffassen, daß damit jedes Wort im Text zum suchbaren *Stichwort* wird und man es damit bereits mit der einfachsten Form von automatischen Indexierung zu tun hat (Knorz 91). Computerlinguistische Indexierungsansätze lösen das Problem, daß der Computer zunächst stets nach Wortformen und nicht nach Wörtern sucht: Die Ähnlichkeit zwischen *Wald* und *Wall* ist für den Rechner größer als zwischen *Wald* und *Wälder*. Statistische Indexierungsverfahren streben an, die Relevanz von Dokumenten für eine Frage auf der Basis von Worthäufigkeitsverteilungen abzuschätzen (Knorz 93).

Mehr als zehn Jahre hindurch hat sich das Indexierungsverfahren AIRlPHYS bei etwa 10.000 Dokumenten! Monat in täglicher Praxis (Lück et al.) bewährt. Es arbeitet mit statistischen Lernverfahren, mit denen es an manuell indexierten Dokumenten trainiert wurde und versucht abzuschätzen, mit welchen Begriffen ein Indexierungsexperte ein Dokument indexieren würde. Nach Aussagen des Betreibers handelt es sich dabei um die weltweit größte Anwendung eines ambitionierten automatischen Indexierungsystems. Die Nachteile dieses Indexierungssystems AIRlPHYS für englischsprachige Physikabstracts liegen in seiner Histo-

rie begründet: Es ist ein umfangreiches Softwaresystem für eine Großrechneranlage, dessen Entwicklungsteam an der TH Darmstadt längst nicht mehr existiert. Eine institutionalisierte Systempflege existiert nicht. Eine Anpassung an erweiterte Anwendungsbereiche und ein Einsatz durch Dritte sind damit praktisch ausgeschlossen.

Der Projektansatz

Mit dem 1993 - 95 an der FH Darmstadt geförderten Projekt *Automatische Indexierung strukturierter Dokumente* sollte der AIR-Indexierungsansatz auf neue Anwendungsbereiche und Dokumenttypen übertragen werden. Auf Basis von Smalltalk wurde ein flexibles Werkzeug für Einrichtung, Optimierung und prototypischen Betrieb eines automatischen Indexierungssystems konzipiert und entwickelt, das auf PC- und Workstation-Umgebungen erprobt werden kann. Dazu wurden folgende Komponenten realisiert:

- *Wörterbuchentwicklung.* Manuell indexierte Dokumente müssen mit dem Ziel analysiert werden, statistische Zusammenhänge zwischen auftretenden Textformulierungen und relevanten (manuell zugeteilten) Deskriptoren zu ermitteln. Deskriptoren können Klassifikationsnotationen oder Thesaurusbegriffe sein, mit Einschränkungen auch freie Schlagwörter.
- *Dokumentenanalyse.* Dokumente müssen dahingehend untersucht werden, welche indexierungsrelevanten Wörter oder Phrases im Dokument in welcher Form und in welchem Kontext auftauchen. Dazu benötigt der Rechner Kenntnisse insbesondere über morphologische und syntaktische Gesetzmäßigkeiten (siehe (Dostal 93, Digula 93, Matousova 93)), sowie Wissen über die logische Struktur von Dokumenten (siehe (Spengler 93)). Anschließend muß der erhobene Befund mithilfe des Wörterbuchs mit relevanten Deskriptoren in Verbindung gebracht werden.
- *Indexierungsfunktion.* Die Indexierungsfunktion hat die Aufgabe abzuschätzen, mit welcher Sicherheit ein Deskriptor auf der Basis der verfügbaren Informationen dem Dokument zugeteilt werden kann (oder aber abgelehnt werden muß). Die Implementierung muß zwei Modi unterscheiden: Zunächst müssen die Parameter der Indexierungsfunktion an Lerndaten eingestellt (optimiert) werden. Anschließend wird die optimierte Indexierungsfunktion zum Indexieren eingesetzt.

Der Indexierungsansatz setzt also voraus, daß zur Entwicklung und Optimierung des Systems eine manuelle Inhaltserschließung entweder bereits verfügbar ist oder aber als Bestandteil der Entwicklungsarbeiten angefertigt wird.

Stand und Ausblick

Am Anfang der Projektarbeiten standen exemplarische Machbarkeitsstudien im Rahmen von Diplomarbeiten «Spengler 93, Klinger 94)), die mit sehr gutem Erfolg und vielversprechenden Ergebnissen abgeschlossen wurden. Die zweite Arbeit «Klinger 94)) hat den Fachbereich IuD als beste studentische Arbeit auf dem Internationalen Symposium für Informationswissenschaft in Graz (ISI'94) vertreten. Auf der Implementierungsseite wurden zunächst Diplomarbeiten mit computerlinguistischer Orientierung vergeben und durch ein studentisches Smalltalk-Projekt zur Entwicklung weitergehend integrierter Systemteile ergänzt. Der gegenwärtige Stand wurde durch eine Diplomarbeit erarbeitet (Steffen 96), deren Ergebnis ein in allen Komponenten prinzipiell arbeitsfähiger Prototyp war. Dieser Prototyp stellt eine Ausgangsplattform für Evaluierungen und weitere Entwicklungen dar.

Um eine Einschätzung von der Tragfähigkeit des Ansatzes für Aufgaben des vorgesehenen Typus zu erhalten, wurde eine erste Sondierung mit deutschsprachigen dpa- Texten im Rahmen des Verbundprojektes CLIP-ING (Projektbeteiligte sind neben der GMD auch dpa, STEP und IPTC) unternommen. Die Indexierungsaufgabe besteht darin, die Agenturmeldungen (das Spektrum reicht von Absätzen bis zu mehrseitigen Meldungen) mit Schlagwörtern aus einem vorgegebenen kleinen Vokabular (ca. 40 Begriffe) zu belegen. Dies wird gegenwärtig manuell bewerkstelligt und dient u. a. als Grundlage für die dpa-Profildienste und die Suche in den Agenturdatenbanken. Beispiele für die Indexierungsbegriffe sind etwa *Inland, Außenpolitik, Wirtschaft, Sport, Agrar; Persönliches, ...*

Zur Behandlung der Morphologie wird die finnische Software *Gertwal* eingebunden, die bei der Morpholympics, einem deutschen computerlinguistischen Wettbewerb der Gesellschaft für Linguistische Datenverarbeitung (GLDV), den ersten Platz belegt hat. Zur Wörterbuchentwicklung und zur Optimierung der Indexierung wurden ca. 500 Dokumente verwendet. Von einer ernstzunehmenden Evaluierung kann man nicht sprechen, aber das Ergebnis der Sondierung sind die auch in anderen Fällen so gerne zitierten "ermutigenden Resultate": Eine manuelle Sichtung der Indexierungsergebnisse ergibt das Bild, daß so gut wie ohne Ausnahme die manuell vergebenen Indexierungsbegriffe den mit höchstem Gewicht automatisch zugeteilten Deskriptoren entsprechen.

Mit dieser ersten "Ermutigung" wird nun mit einer kleinen, aber vom Testdesign her einwandfreien Evaluierung begonnen. Auf die Ergebnisse sind nicht nur Fachhochschule und Projektbeteiligte gespannt, sondern all diejenigen, die nach neuen und bezahlbaren Wegen für verbesserte Informationsdienstleistungen suchen.

Literatur

Digula, M.Z.: Automatische Inhaltserschließung von Freitexten auf der Basis einer Dependenzanalyse von Nominalphrasen. Fachhochschule Darmstadt, Fachbereich Informatik: Diplomarbeit, 1993.

Dostal, P.: Morphologische Analyse des Deutschen, ein lexikonbasierter Ansatz. Fachhochschule Darmstadt, Fachbereich Informatik: Diplomarbeit, 1993.

Knorz, G.: Automated Input into Databases: OCR and Automated Cataloguing. S. 7-1 7-10 in: Proc. of AGARDffIP '91, Madrid (AGARD-CPP-505), 1991

Knorz, G.: Automatische Indexierung. In: Hennings/KnorzlManecke/Reinicke/Schwandt: Wissensrepräsentation und Information Retrieval. Universität Potsdam, Informationswissenschaft, Modellversuch BETID, Lehrmaterialien Nr. 3, S. 138-198, 1993.

Klinger, K.-H.: Automatische Inhaltserschließung einer Volltextdatenbank. Machbarkeitsstudie am Beispiel der FAZ. Fachhochschule Darmstadt, Fachbereich Information und Dokumentation: Diplomarbeit, 1994.

Lück, W.; Rittberger, W.; Schwantner, M.: Der Einsatz des Automatischen Indexierungs- und Retrievalsystems (AIR) im Fachinformationszentrum Karlsruhe, S. 141-170 in: Kuhlen, R. (Hr.): Experimentelles und praktisches Information Retrieval. Konstanz: Universitätsverlag Konstanz, 1992.

Lustig, G. (Hr.): Automatische Indexierung zwischen Forschung und Anwendung. Hildes heim: Olms, 1986

Matousova, R.: Morphologische Analyse des Deutschen, ein regelbasierter Ansatz. Fach hochschule Darmstadt, Fachbereich Informatik: Diplomarbeit, 1993.

Spengler, M.: Dokumentanalyse amerikanischer Finanzberichte und SGML-Konvertierung. Machbarkeitsstudie auf der Basis des DREAM-Parsers. Fachhochschule Darmstadt, Fachbereich Information und Dokumentation: Diplomarbeit, 1993.

Steffen, J.: Verbesserung des Darmstädter Indexierungsansatzes. Fachhochschule Darm stadt, Fachbereich Informatik: Diplomarbeit, 1996.