AUTOMATISCHE INHALTSERSCHLIESSUNG STRUKTURIERTER DOKUMENTE

Projekt an der FH Darmstadt
Fachbereich lud und Informatik
unter Beteiligung der G MD (IPSI) und der THD

Es wird zunehmend erkannt, daß es für Unternehmen von strategischer Bedeutung sein kann, Dokumente nicht nur als fortlaufenden Text zu betrachten, sondern als Objekte mit einer wohldefinierten logischen Struktur. Darüber hinaus ist es häufig für eine qualitativ befriedigende Lösung von Informationsproblemen notwendig oder wünschenswert, Dokumente auch inhaltlich aufzubereiten und im Hinblick auf ein gezielt es Auffinden mit Informationen anzureichern. Traditionell geht es dabei um die manuelle Zuordnung von Klassifikationscodes oder Deskriptoren - eine Lösung, die vielfach aus wirtschaftlichen oder grundsätzlichen Gründen nicht ernsthaft in Frage kommt. Darüber hinaus stellt sich im Kontext neuerer Systemarchitekturen (Hypertext) das Problem der inhaltlichen Erschließung in weitaus komplexerer Weise. 1

Im Rahmen des Projektes der Fachhochschule Darmstadt (93 - 95) unter Beteiligung des IPSI-Institutes der GMD und der TH Darmstadt geht es darum, ein Werkzeug zu konzipieren, zu implementieren und anhand zweier konkreter Anwendungsfälle zu erproben, mit dem sich Aufgaben der inhaltlichen Erschließung spezifizieren und automatisieren lassen.

Die Arbeiten basieren auf der Konzeption des AIR/X-Indexierungssystems, den in der GMD entwickelten Werk zeugen SFK (Smalltalk Frame Kit) und Dream (Parser zur Konvertierung nach SGML).

Der Schritt vom "reinen" Textdokument zu einer strukturell reichhaltigen Repräsentation als SGML-Text über die Zwischenschritte *Document Type Definition (DTD)* und *Document Structure Definition* unter Verwendung des Dream-Parsers ist mittlerweile gut untersucht und im Griff. Gegenwärtig wird damit begonnen, die entwickelten Werkzeuge zur morphologischen und syntaktischen Analyse an die Randbedingungen des SGML-Formates anzupassen. Über die Effektivität spezieller Regelklassen für eine weitergehende inhaltliche Klassifikation liegen erste Ergebnisse vor. Es ist mittlerweile eine PC-Datenbank aufgebaut (und dokumentiert), in der manuell klassifizierte Textstellen mit ihrem linken und rechten Kontext abgelegt sind. In dieser Datenbank werden die Ergebnisse der Anwendung inhaltlicher Regeln abgespeichert und ausgewertet.

Die entscheidenden Implementierungsarbeiten, deren Ergebnis ein vielseitig einsetzbares Werkzeug zur automatischen inhaltlichen Erschließung sein wird, sollen 1994 abgeschlossen werden.

1 Hier geht es darum, innerhalb eines Dokumentes oder zwischen Dokumenten einzelne Dokumentteile inhaltlich zu verknüpfen. Neu ist das Problem zu entscheiden, wie denn überhaupt Quelle und Ziel einer solchen Verknüpfung im Einzelfall zu definieren sind: als Nominalphrase, als Satz, als Abschnitt, **...**

Kontakt: *G. Knorz,* FHD