

# Das Indexierungssystem AUTINDEX

Bärbel Ripplinger

IAI  
Martin-Luther-Str. 14  
66111 Saarbrücken  
babs@iai.uni-sb.de

31. Januar 2001

Mit der wachsenden Zahl elektronisch verfügbarer Dokumente stehen Informationsanbieter wie beispielsweise bibliographische Datenbanken, Bibliotheken, Verlage etc. aber auch Internet-Suchmaschinen vor der Aufgabe diese Informationsflut in kürzester Zeit zu verarbeiten. Die Voraussetzung fuer das Wiederfinden der Dokumente ist einmal das Indizieren der Dokumente und zum zweiten ein Retrievalsystem. Zumeist wird das Indexieren manuell von einer Vielzahl von menschlichen Indexierern geleistet, zunehmend jedoch auch von Softwaresystemen. Am IAI wurde in den letzten Jahren das System AUTINDEX entwickelt, das eine generische Lösung zur automatischen Indexierung und Klassifizierung von Dokumenten in verschiedenen Sprachen anbietet.

AUTINDEX benützt als Grundlage umfangreiche linguistische Verfahren, wobei das Kernsystem aus einer morpho-syntaktischen Komponente (Mpro) besteht. Diese Analyse weist jedem Wort im Dokument neben grammatikalischen Informationen wie der Wortklasse auch semantische Merkmale zu. Für deutsche Texte wird eine Kompositaanalyse durchgeführt, so dass auch Informationen über die möglichen Zerlegungen von zusammengesetzten Wörtern vorhanden sind. In einem zweiten Schritt werden die bedeutungstragenden Wörter (Grundformen von Nomen und Verben) gesammelt und aus der ihnen zugewiesenen semantischen Information die häufigsten semantischen Klassen ermittelt. Danach werden alle Wörter, die diesen Klassen zugeordnet sind, gesammelt und gewichtet. Das Gewicht berechnet sich dabei aus der Frequenz des Wortes und der semantischen Klasse. Zur Berechnung der Frequenz betrachtet der Gewichtungsalgorithmus nicht nur komplette Wörter sondern auch die Bestandteile von Komposita. In einem dritten Schritt wird zusätzlich ein 'Shallow Parsing' durchgeführt zur Erkennung von Mehrwortlexemen und deren syntaktischen Varianten, insbesondere werden hier Nominalphrasen betrachtet. Mehrwortlexeme, die den in Schritt 2 ermittelten semantischen Klassen zugeordnet werden können, bilden zusammen mit den in Schritt 2 ermittelten Einwort-Termini die Menge der sogenannten *Schlüsselwörter*. Soll die Indexierung anhand eines Thesaurus erfolgen, werden jetzt die Schlüsselwörtern gegen den entsprechenden Thesaurus abgeglichen und so die Menge der *Deskriptoren* ermittelt. Die Zahl der einem Dokument zugewiesenen Deskriptoren kann über einen Parameter vom Benutzer bestimmt werden.

Zur Klassifikation eines Dokuments wird die entsprechende Information, die bestimmten Begriffen im Lexikon oder im Thesaurus mitgegeben ist, statistisch ausgewertet. Die am häufigsten vorkommenden Klassen werden dem Dokument als Klassifikation zugewiesen. Auch hier

kann der Benutzer die maximale Zahl der Klassifikatoren durch eine Parameter festlegen. Existiert kein anwendungsspezifisches Schema, dann wird der ebenfalls im Lexikon vorhandene NACE-Kode zur Klassifizierung verwendet.

AUTINDEX arbeitet jedoch nicht nur monolingual, sondern kann auch eine multilinguale Indexierung durchführen, dabei wird der Text in der Sprache, in der er vorliegt, indexiert, d.h. die Menge der Schlüsselwörter und der Deskriptoren ermittelt. Diese werden mithilfe von Transferwörterbüchern in die gewünschte Zielsprache übersetzt, wobei hier die Klassifikationsinformation als zusätzliche Wissensquelle zum Einsatz kommt. Diese multilinguale Indexierung wird zur Zeit innerhalb des von der EU im IST-Programm geförderten Projekt BINDEK weiterentwickelt. In BINDEK steht die bilinguale Indexierung und Klassifizierung von deutschen und englischen Dokumenten im Vordergrund.

AUTINDEX hat den Vorteil, dass es alle Arten von Dokumenten indexieren und klassifizieren kann. Die Einbeziehung eines Thesaurus und/oder eines Klassifikationsschematas ist optional und erfolgt anwenderspezifisch.

## **System-Daten**

### **Unterstützte Sprachen:**

**Monolingual:** Deutsch, Englisch,

**Bilingual:** Deutsch → Englisch, Englisch → Deutsch

**Interface:** Web-Interface, Windows-Interface (wird zur Zeit entwickelt)

**Unterstützte Plattformen:** Unix, Linux, Windows, NT

### **Platzbedarf:**

**Monolingual:** ca. 6-7 MB pro Sprache

**Bilingual:** zusätzlich 25 MB (maximal) pro Sprachpaar

zusätzlich der Platzbedarf für den Thesaurus.

**Hauptspeicher:** mind. 64 MB