Technology
Arts Sciences
TH Köln

Institut für Informationswissenschaft

Fakultät für Informations- und Kommunikationswissenschaften

Prof. Dr. Klaus Lepsky
klaus.lepsky@th-koeln.de

Wahlpflichtveranstaltung – Information Retrieval und Automatisches Indexieren

Tutorial „Automatisches Indexieren“

9. April 2018

Zusammenfassung

Dieses Skript enthält die Aufgabenstellung für den Programmteil *Automatisches Indexieren* der Wahlpflichtveranstaltung „Information Retrieval und Automatisches Indexieren“. Teilziele der Aufgabenstellung sind der Aufbau einer Dokumentkollektion mit *Midos* und deren automatische Indexierung mit *Lingo*.

1 Einführung

Als Indexierung (inhaltliche Erschließung) werden Methoden und Verfahren bezeichnet, die der Zuordnung von Indextermen (Indexaten, Erschließungsmerkmalen) zu Dokumenten (dokumentarischen Bezugseinheiten) dienen. Ziel der Indexierung ist es, über die Indexterme ein gezieltes Auffinden der Dokumente zu ermöglichen. Indexterme können inhaltsbeschreibende Merkmale wie Notationen oder Deskriptoren, kontrollierte oder freie Schlagwörter, aber auch Stichwörter sein. Sie können intellektuell und/oder automatisch gewonnen werden.

Es werden drei mögliche Arten der Durchführung der Indexierung unterschieden: intellektuell, computergestützt (halbautomatisch) und automatisch. Während bei der intellektuellen Indexierung die geeigneten Indexterme als Ergebnis einer intellektuellen Inhaltsanalyse bestimmt und zugeteilt werden, findet eine solche bei der automatischen Indexierung gar nicht statt. Dabei werden durch die eingesetzten Algorithmen Indexterme automatisch aus dem Dokumenttext ermittelt. Bei der computergestützten Indexierung werden diese dem Indexierer vorgeschlagen, bei der automatischen Indexierung werden sie dem Dokument direkt zugeteilt.

Im Rahmen dieses Tutorials wird für eine vorgegebene Dokumentkollektion eine linguistisch basierte automatische Indexierung durchgeführt. Eingesetzt werden dabei die Testkollektion *GIRT*[1] mit deutschsprachigen bibliografischen Referenzdaten aus dem Bereich der Sozialwissenschaften, die linguistisch und statistisch arbeitende Software *Lingo* für die automatische Indexierung der Kollektion und das Datenbanksystem *Midos* als Container für die Dokumentkollektion und als Tool für die Unterstützung des Indexierungs-Workflows.

2 Aufbau der Dokumentkollektion

In diesem Teil des Arbeitsprogramms wird eine eigene Dokumentkollektion bibliografischer Referenzdaten als *Midos*-Datenbank aufgebaut. Die Daten der *GIRT*-Testkollektion sind bereits im *Midos*-Speicherformat aufbereitet und können als zip-Archiv heruntergeladen werden: girt-midos.zip.

Die Archivdatei enthält die *.dbm-Datei und das dazugehörige Datenschema (MISCHABS). Nach dem Entpacken kann die Datenbank über „Datei öffnen" in *Midos* geöffnet werden.

Analysieren Sie den Inhalt der Datenbank und die Felder der Datenbeschreibung. Machen Sie sich insb. ein Bild von der vorhandenen intellektuellen Erschließung.

3 Automatisches Indexieren

Für das Verständnis der Thematik „Automatisches Indexieren" und die Arbeit mit *Lingo* ist die intensive Lektüre des Kapitels 5 des Lehrbuchs „Informationserschließung

[1] Vgl. **kluck_girt-testdatenbank_2004**.

und Automatisches Indexieren“[2] unerlässlich. Dort sind auch die einzelnen Schritte des Workflows für die automatische Indexierung detailliert beschrieben.[3]

3.1 Installation von *Lingo*

Auf den Laborrechnern ist *Lingo* bereits installiert. Es ist dann nur noch nötig, eine sog. „Arbeitsumgebung“ einzurichten. Für die völlige Neuinstallation auf einem eigenen Rechner folgen Sie bitte dieser Anleitung auf der *Lingo*-Webseite: lex-lingo.de. Bei vorinstalliertem *Lingo* genügen die beiden folgenden Schritte für das Einrichten der Arbeitsumgebung:

- Einrichten eines Verzeichnisses `/lingo-work` (= *Lingo*-Arbeitsumgebung):
  `lingoctl demo lingo-work`
- Testlauf der *Lingo*-Arbeitsumgebung:
  `lingo -c lingo.cfg txt/artikel.txt`

3.2 Export der Daten aus *Midos* für die automatische Indexierung

Lingo benötigt die Daten für die Indexierung in einem vorgegebenen Format (dem sog. *LIR*-Format), das in Kapitel 5 auf S. 296 dokumentiert ist. Für einen Export in diesem Format wird ein *Midos*-Ausgabeformat erstellt, das *nur* die zu indexierenden Kategorien der Datensätze enthält. Dies sollten ausschließlich Kategorien mit einem *Bezug zum Dokumenteninhalt* sein.

Es ist möglich und sinnvoll, verschiedene Kategorieninhalte auf getrennte Indexierungsläufe zu verteilen, um die jeweils erzielten Ergebnisse später miteinander vergleichen zu können. So wären z. B. getrennte Indexierungsläufe für „Titel“, „Titel und Abstract“, „Abstract und Schlagwörter“ etc. denkbar.

Falls auch unterschiedliche Funktionalitäten von *Lingo* miteinander verglichen werden sollen, müssten auch dafür getrennte Indexierungsläufe durchgeführt werden (allerdings keine weiteren spezifischen Datenexporte). Das könnten z. B. sein: „Grundformerkennung“, „Grundformerkennung und Kompositumzerlegung“, „Grundformerkennung, Kompositumzerlegung und Mehrworterkennung“ etc.

Der Export der Daten aus *Midos* kann z. B. (es gibt mehrere Wege) über den Dialog „Ausgabe“, „Datei“ mit den Optionen „Text (Ausgabeformat)“ und „UTF-8“ erfolgen.

[2] **godert__informationserschliesung__2012**, Kapitel 5: http://link.springer.com/chapter/10.1007/978-3-642-23513-9_5.

[3] Bitte verwenden Sie **nicht** die Befehlsaufrufe für *Lingo* aus dem Kapitel 5. *Lingo* ist inzwischen weiter entwickelt worden und dessen Bedienung hat sich dadurch teilweise geändert.

3.3 Indexierung mit *Lingo*

Es ist zweckmäßig, die exportierte Datei in das Verzeichnis `/txt` der *Lingo*-Arbeitsumgebung zu legen. Die Indexierung lässt sich dann von `/lingo-work` aus mit folgendem Befehl starten:

- Indexierungslauf mit *Lingo*:
  `lingo -c lir.cfg txt/export.txt`
  („export.txt“ durch den Namen der eigenen Export-Datei aus *Midos* ersetzen)

Die Konfigurationsdateien „lir.cfg“ und „de.lang“ sind ggf. zuvor hinsichtlich der gewünschten Funktionalität anzupassen.[4]

3.4 Analyse, Verbesserung und Nutzung der Indexierungsergebnisse

- Analysieren Sie die jeweils erzielten Indexierungsergebnisse hinsichtlich ihres funktionalen Umfangs. Ändern Sie ggf. die Konfigurationsdateien, um Einfluss auf die Ergebnisse zu nehmen.
- Überlegen Sie, welche Vergleiche in Bezug auf die Funktionalität der Indexierung und/oder der zu indexierenden Kategorieninhalte sinnvoll sein könnten.
- Schaffen Sie für diese Vergleiche die benötigte Datenbasis. Erstellen Sie ggf. eine *Midos*-Retrievalanwendung, um vergleichende Suchen durchführen zu können.

[4] Für eine Nachnutzung der Indexierungsergebnisse in *Solr* ist die Sortierfunktion in der „lir.cfg“ zuvor zu deaktivieren: `- vector_filter: { in: res, lexicals: 's', sort: false }`.