Internet-Suchmaschinen
Praktische Übung

Vu Tran

Mai 2015 — Juli 2015

0 Einführung

0.1 Ziel der Veranstaltung

Die Teilnehmer des IR-Praktikums sollen lernen, die in der Vorlesung vorgestellten Konzepte an einem praktischen Beispiel anzuwenden.

0.2 Ablauf

Das Praktikum ist in drei Blöcke aufgeteilt.

Im ersten Block wird der Suchserver Solr installiert und damit eine Kollektion indexiert.

Im zweiten Block wird anhand der indexierten Kollektion eine Evaluation durchgeführt.

Im dritten Block wird das Suchsystem um einen Webcrawler erweitert zu einer Internet-Suchmaschine.

0.3 Arbeitsweise

Die Bearbeitung der Aufgaben sowie die Abnahme erfolgt in Einzelarbeit oder in Gruppen zu zwei Personen.

Bei einigen Aufgaben muss Programmcode geschrieben werden. Dabei sind prinzipiell alle sinnvollen[1] Sprachen zugelassen, insbesondere Java, Python und Ruby, Shellskripte, aber auch Textbearbeitungssprachen wie `sed` und `awk`. Wichtig ist, dass der Code funktioniert, bei der Abnahme erklärt werden kann und dass wesentliche Schritte programmatisch durchgeführt werden. Wenn Du Programmbibliotheken kennst, die Dir die Arbeit erleichtern: nutze sie! Im Zweifelsfall bitte vorher nachfragen.

0.4 Abnahme

Abschluss des Praktikums ist die Abnahme des Suchsystems im Rahmen einer Vorführung. Dazu muss jeder Teilnehmer bzw. jede Gruppe einen Abnahmetermin im Zeitraum 20.7.2015–21.7.2015 vereinbaren. Die Abnahme dauert ca. 20 Minuten.

Bei der Abnahme werden die Lösungen der einzelnen Blöcke betrachtet, die zur Lösung erstellten Tools und das jeweils verwendete Schema. Dazu kann ein eigener Rechner mitgebracht werden. Falls das nicht möglich ist, kann die Abnahme als auch die Bearbeitung auch auf einem Rechner im DB-Pool (Raum LF 230) erfolgen. Studierende, die die Aufgaben gerne im DB-Pool bearbeiten möchten, melden sich bitte frühzeitig per E-Mail zur Vereinbarung von Terminen und Accounts.

[1] Nicht zugelassen sind esoterische Sprachen wie Whitespace, Brainfuck und Shakespeare

1 Block 1: Wochen 1 und 2

1.1 Aufgabe 1: Solr

(a) Installiere Solr 5.1 direkt von der Apache-Seite[2] in ein beliebiges leeres Verzeichnis auf Deinem Rechner.

(b) Arbeite den Tutorial[3] durch. Die Dokumentation zu Solr findest Du im Wiki[4] und im Reference Guide[5].

1.2 Aufgabe 2: CLEF-Kollektion indexieren

Die CLEF-Kollektion enthält Dokumente aus zwei Zeitungen, Suchaufgaben („Topics“) in der Datei `Top-de03.txt` und Relevanzbewertungen in der Datei `qrels_DE`.

(a) Nachdem Du Dich in Solr eingearbeitet hast, lösche die während des Tutorials indexierten Dokumete im Solr-Index.

(b) Ändere die Datei `conf/schema.xml` so, dass die Felder der CLEF-Kollektion gespeichert werden können. Die Felder, die in beiden Zeitungen vorhanden sind, sind DOCNO, DOCID, DATE, HEADLINE, BYLINE und TEXT. *Verwende sinnvolle Feldtypen, nachdem Du die zu indexierenden Daten analysiert hast.*

(c) Indexiere die CLEF-Kollektion[6]. Für das Indexieren werden nur die Dokumente benötigt. Die Hauptaufgabe dabei besteht darin, die in SGML vorliegenden Dokumente in ein für Solr geeignetes Format aufzubereiten. Sinnvoll ist, erst einige wenige Dokumente zu indexieren und das Indexierungs-Schema mit Suchanfragen darauf zu überprüfen, ob die erwarteten Dokumente gefunden werden. Wenn die Suchanfragen nicht zufriedenstellend beantwortet werden, wird das Schema aktualisiert, die bereits indexierten Dokumente gelöscht und von vorne begonnen. Wenn das Schema zufriedenstellend ist, können dann die restlichen Dokumente indexiert werden.

2 Block 2: Wochen 3 und 4

2.1 Aufgabe 3: CLEF-Evaluation durchführen

In diesem Block wird eine Retrieval-Evaluation durchgeführt. Die dazu notwendigen Dokumente sind bereits in Block 1 dem Index hinzugefügt worden.

[2] `http://lucene.apache.org/solr/`
[3] `http://lucene.apache.org/solr/quickstart.html`
[4] `http://wiki.apache.org/solr/`
[5] `https://www.apache.org/dyn/closer.cgi/lucene/solr/ref-guide/apache-solr-ref-guide-5.1.pdf`
[6] `ftp://ftp.is.inf.uni-due.de/pub/ismp_ss14/fr_rundschau.zip`

Die Idee bei der Evaluation ist, die Suchaufgaben aus der Topics-Datei von Solr durchführen zu lassen und die Ergebnislisten anhand der vorhandenen Relevanzbewertungen in der Datei `qrels_DE`[7] zu prüfen.

Um aus der Topics-Datei[8] Suchanfragen zu generieren, kann man ein Skript schreiben oder die Suchanfragen intellektuell erzeugen und so speichern, dass sie der Topic-ID zugeordnet werden können. Das ist für die Berechnung von Precision und Recall anhand der Relevanzbeurteilungen notwendig.

Das Berechnen der Retrieval-Maße kannst Du komplett selbst programmieren, aber gerne auch das Programm `TRECEval`[9] dazu verwenden. Das muss vor der Verwendung vermutlich erst kompiliert werden. Dazu ist ein C-Compiler und das `make`-Tool notwendig. Eine Beschreibung von TRECEval liegt dem ZIP-File als Word-Dokument bei.

Die Suchanfragen können am einfachsten über das Admin-Oberfläche ausgeführt werden, da hier die zurückgegebenen Felder angegeben werden können und der Score als Eingabe für TRECEval notwendig ist.

(a) Führe eine Retrieval-Evaluation mit den gegebenen Daten durch und ermittle die Mean Average Precision.

(b) Optional: Wenn Du Deine Evaluation durchgeführt hast, hast Du ein Maß für die Güte des Systems. Experimentiere mit unterschiedlichen Indexierungsvarianten und beobachte, wie sich die Güte dadurch verändert. Kandidaten für Änderungen sind z.B. die Analyzer, Tokenizer und Filter in den Feld-Definitionen in `schema.xml`.

3 Block 3: Wochen 5 und 6 (BAI)

3.1 Aufgabe 4: Implementierung eines Webcrawlers

Du hast nun mit Solr im Default-Core des Beispiel-Servers eine Kollektion indexiert. In diesem Block geht es darum, Webdokumente im Solr-Index zu speichern. Um die Webdokumente von den CLEF-Dokumenten zu trennen, sollte als erstes ein neuer Core erzeugt werden.

Als Startwert für Deinen neuen Core, `web`, kannst Du das Verzeichnis `conf` inkl. Inhalt aus dem Beispiel-Verzeichnis in das neue Verzeichnis `web` kopieren.

Nun kannst Du den Solr-Server neu starten.

3.2 Aufgabe 5: Webcrawler bauen

Implementiere einen Webcrawler, der, ausgehend von einer Start-URL, Webdokumente findet und in den Solr-Index schreibt. Achte dabei auf die Vermeidung von

[7] `ftp://ftp.is.inf.uni-due.de/pub/ismp_ss14/qrels_DE`
[8] `ftp://ftp.is.inf.uni-due.de/pub/ismp_ss14/Top-de03.txt`
[9] `ftp://ftp.is.inf.uni-due.de/pub/ismp_ss14/TrecEval.zip`

Zyklen beim Crawlen. Die in den gefundenen Dokumenten enthaltenen URLs sollen dann rekursiv weiterverfolgt werden. Informationen über die Architektur von Websuchmaschinen finden sich in den Einführungsfolien zur Vorlesung. Bei dieser Aufgabe sind Programmbibliotheken zugelassen, aber keine fertigen Crawler- Implementierungen wie z.B. Nutch.

Ein Benutzer-Interface für die Suchmaschine muss **NICHT** implementiert werden – wir benutzen dazu einfach das Solr-Suchformular.

4 Block 3: Wochen 5 und 6 (Komedia)

4.1 Aufgabe 4: User Interface für Solr

Erstellt ein User Interface für den Solr-Index aus Block 1. Dazu soll das Projekt `ajax-solr`[10] benutzt werden. Die Dokumentation zu diesem Projekt und ein Tutorial ist auf der Webseite des Projekts zu finden.

[10] `https://github.com/evolvingweb/ajax-solr`