Informatik
Hauptcampus

H O C H
S C H U L E
T R I E R

Konzeption und Realisierung eines Systems zur Informationssuche in einem Dokumentenarchiv basierend auf Textinhalt und Metadaten.

Conception and Realization of an Information Retrival System for a Document Archive based on Text Content and Metadata

Annika Kremer

Bachelor-Abschlussarbeit

Betreuer: Prof. Dr. Karl Hans Bläsius

Trier, Abgabedatum

Vorwort

Ein Vorwort ist nicht unbedingt nötig. Falls Sie ein Vorwort schreiben, so ist dies der Platz, um z.B. die Firma vorzustellen, in der diese Arbeit entstanden ist, oder einigen Leuten zu danken, die in irgendeiner Form positiv zur Entstehung dieser Arbeit beigetragen haben. Auf keinen Fall sollten Sie im Vorwort die Aufgabenstellung näher erläutern oder vertieft auf technische Sachverhalte eingehen.

Kurzfassung

In der Kurzfassung soll in kurzer und prägnanter Weise der wesentliche Inhalt der Arbeit beschrieben werden. Dazu zählen vor allem eine kurze Aufgabenbeschreibung, der Lösungsansatz sowie die wesentlichen Ergebnisse der Arbeit. Ein häufiger Fehler für die Kurzfassung ist, dass lediglich die Aufgabenbeschreibung (d.h. das Problem) in Kurzform vorgelegt wird. Die Kurzfassung soll aber die gesamte Arbeit widerspiegeln. Deshalb sind vor allem die erzielten Ergebnisse darzustellen. Die Kurzfassung soll etwa eine halbe bis ganze DIN-A4-Seite umfassen.

Hinweis: Schreiben Sie die Kurzfassung am Ende der Arbeit, denn eventuell ist Ihnen beim Schreiben erst vollends klar geworden, was das Wesentliche der Arbeit ist bzw. welche Schwerpunkte Sie bei der Arbeit gesetzt haben. Andernfalls laufen Sie Gefahr, dass die Kurzfassung nicht zum Rest der Arbeit passt.

The same in english.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1

Einleitung und Problemstellung

1.1 Einleitung

Tag für Tag treffen neue E-Mail Nachrichten mit den unterschiedlichsten Inhalten im Posteingang ein, weshalb die Anzahl gesendeter und empfangener Mails schnell unübersichtlich groß wird.
Will der Nutzer eine Weile später auf eine bestimmte Nachricht erneut zugreifen, steht er oftmals vor der Problematik, mit der manuellen Suche aufgrund der schieren Datenmenge überfordert zu sein.
Besonders schwierig wird es für ihn, wenn er nach bestimmten Inhalten sucht, sich jedoch weder an das genaue Datum noch an den Absender erinnert. An dieser Stelle hilft dem Nutzer ein Suchprogramm, mit dem er selbstgewählte Suchkriterien beliebig logisch verknüpfen kann.
Weiß er beispielsweise noch, dass die Nachricht im Mai ankam und eine wichtige Adresse beinhaltet, könnte seine Suchanfrage „Datum = Mai und Freitext = Adresse “lauten.
Die Ausgabe des Programms besteht aus einer Liste damit für ihn in Frage kommender Dokumente, was die Suche auf einen kleinen Kreis relevanter Treffer einschränkt.
Die Suche basierend auf Freitext sowie Metadaten beschränkt sich nicht nur auf E-Mails, sondern lässt sich auch auf andere Formen von Dokumentenarchiven anwenden.

Begonnen werden soll mit einer Einleitung zum Thema, also Hintergrund und Ziel erläutert werden.

Weiterhin wird das vorliegende Problem diskutiert: Was ist zu lösen, warum ist es wichtig, dass man dieses Problem löst und welche Lösungsansätze gibt es bereits. Der Bezug auf vorhandene oder eben bisher fehlende Lösungen begründet auch die Intention und Bedeutung dieser Arbeit. Dies können allgemeine Gesichtspunkte sein: Man liefert einen Beitrag für ein generelles Problem oder man hat eine spezielle Systemumgebung oder ein spezielles Produkt (z.B. in einem Unternehmen), woraus sich dieses noch zu lösende Problem ergibt.

Im weiteren Verlauf wird die Problemstellung konkret dargestellt: Was ist spezifisch zu lösen? Welche Randbedingungen sind gegeben und was ist die Zielsetzung?

Letztere soll das beschreiben, was man mit dieser Arbeit (mindestens) erreichen möchte.

1.2 Problemstellung

Ziel der Arbeit ist die Konzeption und Realisierung eines Systems zur Informationssuche (engl. Information Retrieval System) in einem Dokumentenarchiv, wobei die Dokumente von teilweise strukturierter Natur sind.
Dies bedeutet, dass sie sowohl gewöhnlichen Freitext als auch Metadaten enthalten. Der Nutzer soll spezifizieren können, in welchen Metadaten er suchen möchte, zudem soll die Freitextsuche auswählbar sein. Alle Suchanfragen sollen hierbei beliebig mit den booleschen Operatoren *AND* (engl. und) sowie *OR* (engl. oder) verknüpfbar sein.
Hauptanwendungszweck des Systems sind E-Mail-Archive wie Posteingang und Postausgang, allerdings soll das System so flexibel sein, dass es auch auf andere teilweise strukturierte Dokumentenarchive anwendbar ist.

1.3 Teilprobleme

Aus der Aufgabenstellung ergeben sich die im folgenden beschriebenen Teilprobleme.

1.3.1 Dynamisches einlesen der Metadaten

Die genauen Metadaten sind, da das System flexibel sein soll, vor dem Ausführen des Systems noch nicht bekannt. Demnach muss das IR-System die Namen der Metadaten beim Starten des Programms dynamisch einlesen und diese dem Nutzer auf der grafischen Oberfläche anzeigen.

1.3.2 Unterscheidung Metadaten und Freitext

Damit das System die Metadaten dynamisch einlesen kann, müssen die folgenden Punkte erfüllt sein:

1. Das System muss zwischen Metadaten und Freitext unterscheiden können.
2. Die Namen der Metadaten, welche im Folgenden mit „Keywords “ bezeichnet werden, müssen vom Inhalt der Metadaten abgegrenzt werden.
3. Der Inhalt kann unterschiedlichen Datentyps sein, z.B. String oder Liste, weshalb dieser bestimmt werden muss.

1.3.3 Metadatensuche

Es muss erkannt werden, welche Metadaten der Nutzer ausgewählt hat und in genau diesen muss, unter Berücksichtigung des jeweiligen Datentyps der Inhalte, gesucht werden.
Im Gegensatz zur Freitextsuche muss zu jedem Dokument geprüft werden, ob das entsprechende Schlüsselwort überhaupt auftritt, bevor in diesem gesucht wird.

1.3.4 Freitextsuche

Bei der Freitextsuche ist die Wortzahl weitaus größer als bei der Metadatensuche. Daraus resultieren zwei Probleme:

1. Wie kann effizient in großen Wortmengen gesucht werden?
2. Wie kann die Suche bei begrenztem Speicher bewältigt werden?

Zudem stellt sich die Frage nach einem geeigneten Verfahren, das bei längeren Anfragen auch teilweise passende Ergebnisse liefert und messen kann, wie gut die erzielten Treffer zur Nutzeranfrage passen.

1.3.5 Verknüpfung mit UND/ODER

Alle Anfragen sollen beliebig logisch verknüpfbar sein. Dies beinhaltet die folgenden Problemstellungen:

1. Innere Schachtelung der Suchanfrage:
   - Die Resultate zu den verschiedenen ausgewählten Keywords einer Anfrage müssen miteinander verknüpft werden.
   - Die Resultate der Freitextsuche,sofern in der Anfrage ausgewählt, müssen mit den Ergebnissen der Medatensuche verknpüft werden.
   - Es sollt möglich sein, die Anfrage mit dem logischen Operator NOT (engl. nicht) zu negieren.
2. Äußere Schachtelung der Suchanfrage:
   - Mehrere Anfragen sollen miteinander logisch verknüpfbar sein.
   - Hierbei sollen sowohl AND als auch OR möglich sein.

1.3.6 Benutzeroberfläche

2

Information Retrieval

Dieses Kapitel soll dem Leser einen Überblick darüber vermitteln, wofür der Begriff „Information Retrieval “steht.

2.1 Bedeutung

Der englische Begriff „Information Retrieval “ lässt sich mit „Informationsrückgewinnung “ ins Deutsche übersetzen.

2.1.1 Retrieval

„Retrieval“wurde bewusst mit „Rückgewinnung “übersetzt, da beim Information Retrieval keine neuen Informationen gewonnen, sondern bereits existierende aufgefunden werden.

2.1.2 Information

Die Bedeutung des Begriffsbestandteils Information “erweist sich als weitaus schwieriger darzustellen: Was ist eine Information?
Eine einheitliche Definition hierzu lässt sich nicht finden, zumal es zahlreiche Betrachtungsweisen aus unterschiedlichen Disziplinen gibt. Für diese Arbeit ist lediglich die Perspektive der Disziplin Informatik relevant, darum wurde hieraus ein Definitionsansatz gewählt.

Information

2.2 Information Retrieval

Definition 2.1. *(Information Retrieval)*
Mit Information Retrieval, kurz IR, wird das Auffinden von in unstrukturierter Form vorliegender und ein Informationsbedürfnis befriedigender Materialien innerhalb großer Sammlungen bezeichnet.

Mit unstrukturierten Materialien sind hierbei meist Dokumente in Textform gemeint. Üblicherweise liegen die Sammlungen auf dem Computer gespeichert vor ([CDM08], S.1).

3

Boolean Retrieval

3.0.1 Matrix

3.0.2 Posting List

3.0.3 Anfragenverknüpfung

4

Das Vektorraum-Modell

4.1 Funktionsprinzip

4.2 Term Frequency

4.3 Document Frequency

4.4 Inverted Document Frequency

4.5 $Tf \times idf$ Weighting

4.5.1 Formeln

4.6 Ähnlichkeitsfunktion

4.6.1 Euklidische Distanz

4.6.2 Cosine Similarity

4.6.3 Alternativen

5

Bewertung eines Information Retrieval Systems

5.1 Precisison

5.2 Recall

6

Implementierung

7

Die Benutzeroberfläche

8

Zusammenfassung und Ausblick

In diesem Kapitel soll die Arbeit noch einmal kurz zusammengefasst werden. Insbesondere sollen die wesentlichen Ergebnisse Ihrer Arbeit herausgehoben werden. Erfahrungen, die z.B. Benutzer mit der Mensch-Maschine-Schnittstelle gemacht haben oder Ergebnisse von Leistungsmessungen sollen an dieser Stelle präsentiert werden. Sie können in diesem Kapitel auch die Ergebnisse oder das Arbeitsumfeld Ihrer Arbeit kritisch bewerten. Wünschenswerte Erweiterungen sollen als Hinweise auf weiterführende Arbeiten erwähnt werden.

Literaturverzeichnis

CDM08. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: *Introduction to Information Retrieval*. Cambridge University Press, 2008.

A

Glossar

DisASTer	DisASTer (Distributed Algorithms Simulation Terrain), A platform for the Implementation of Distributed Algorithms
DSM	Distributed Shared Memory
AC	Linearisierbarkeit (atomic consistency)
SC	Sequentielle Konsistenz (sequential consistency)
WC	Schwache Konsistenz (weak consistency)
RC	Freigabekonsistenz (release consistency)

B

Erklärung der Kandidatin / des Kandidaten

☐ Die Arbeit habe ich selbstständig verfasst und keine anderen als die angegebenen Quellen- und Hilfsmittel verwendet.

☐ Die Arbeit wurde als Gruppenarbeit angefertigt. Meine eigene Leistung ist ...

Diesen Teil habe ich selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet.

Namen der Mitverfasser: ...

Datum	Unterschrift der Kandidatin / des Kandidaten