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 Textinhalten sucht, sich jedoch weder an das genaue Datum noch an den Absender erinnert. An dieser Stelle hilft dem Nutzer ein Suchprogramm, mit dem er 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. Das Suchprogramm präsentiert dann eine Liste für ihn in Frage kommender Dokumente, was seine Suche auf einen kleinen Kreis relevanter Treffer einschränkt.

Hier rumlabern von wegen in EMails sucht man manchmal nach iwas, unübersichtlich, bla blubb

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

Teilweise strukturiert, teilweise Freitext, man will zeugs finden -¿ relativ allgemein halten

1.3 Teilprobleme

1.3.1 Unterscheidung zwischen Metadaten und Freitext

1.3.2 Metadatensuche

1.3.3 Freitextsuche

1.3.4 Verknüpfung mit UND/ODER

1.3.5 Benutzeroberfläche

2

Information Retrieval

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

CDK02. Coulouris, George, Jean Dollimore und Tim Kindberg: *Verteilte Systeme: Konzepte und Design*. Addison-Wesley-Verlag, 2002.

Che85. Cheriton, David R.: *Preliminary Thoughts on Problem-oriented Shared Memory: A Decentralized Approach to Distributed Systems*, 1985.

Mal97. Malte, Peter: *Replikation in Mobil Computing*. Seminar No 31/1997, Institut für Telematik der Universität Karlsruhe, Karsruhe, 1997. `http://www.ubka.uni-karlsruhe.de/cgi-bin/psview?document=/ira/1997/31`.

Mos93. Mosberger, David: *Memory Consistency Models*. Technical Report 93/11, University of Arizona, November 1993.

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