

Konzeption und Implementierung eines Systems zur automatischen Analyse und Klassifikation von Berichten in Nachrichtenportalen.

Conception and implementation of a system for the automatic analysis and classification of articles in newsportals

Tobias Arens

Bachelor-Abschlussarbeit

Betreuer: Karl Hans Blaesius

Trier, 13.12.2016

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

1	Einleitung und Problemstellung	1
2	Weitere Kapitel	2
3	LaTeX-Bausteine	3
	3.1 Abschnitt	3
	3.1.1 Unterabschnitt	3
	3.2 Abbildungen und Tabellen	4
	3.3 Mathematische Formel	4
	3.4 Sätze, Lemmas und Definitionen	5
	3.5 Fußnoten	6
	3.6 Literaturverweise	6
4	Beispiel-Kapitel	7
	4.1 Warum existieren unterschiedliche Konsistenzmodelle?	
	4.2 Klassifizierung eines Konsistenzmodells	8
	4.3 Linearisierbarkeit (atomic consistency)	8
5	Zusammenfassung und Ausblick	10
Li	teraturverzeichnis	11
Sa	chverzeichnis	12
$\mathbf{G}\mathbf{l}$	lossar	13
\mathbf{Er}	klärung der Kandidatin / des Kandidaten	14

Abbil	ldung	raver	zeic	hnis
ADDI	լասուբ	Sver	Zeic.	111112

3.1 Bezeichnung der Abbildung	4
-------------------------------	---

Tabellenverzeichnis

3.1	Bezeichnung der Tabelle	4
4.1	Linearisierbarkeit ist erfüllt	9
4.2	Linearisierbarkeit ist verletzt, sequentielle Konsistenz ist erfüllt	9
4.3	Linearisierbarkeit und sequentielle Konsistenz sind verletzt	9

Einleitung und Problemstellung

Zur Informationsbeschaffung in der heutigen Zeit gibt es zahlreiche Möglichkeiten. So gibt es hunderte Online-Nachrichtenportale, Zeitschriften, Zeitungen und Magazine aus denen wertvolle Informationen gewonnen werden kann.

Diese Diversität hat viele positive Eigenschaften, wie das vorliegen verschiedener Standpunkte und die Beleuchtung von Ereignissen und Nachrichten aus vielen Blickrichtungen. Die dadurch erreichte Abdeckung ist begrüßenswert schafft allerdings einige Probleme. Es ist umständlich bei Interesse an einem bestimmten Thema passende Artikel zu finden da dafür eine Möglichkigkeit einer großflächigen Abdeckung verschiedener Portale fehlt.

Um Artikel thematisch zuordnen zu können ist eine automatisierte Klassifizierung sinnvoll, die mithilfe vorab definierter Klassen neue Informationsquellen zielsicher in ebenjene Abbilden kann. Dies ermöglicht eine zielgerichtete, breitgefächerte Suche nach themenbezogenen Informationen.

In dem Rahmen dieser Arbeit wird eine Anwendung entwickelt, die dieses Problem lösen soll. Dafür sind allerdings mehrere Schritte notwendig: Es muss eine Schnittstelle entwickelt werden, die Nachrichtenartikel verschiedener Quellen auslesen kann um diese für die weitere Verarbeitung nutzen zu können. Die so gewonnenen Daten müssen mit entsprechenden Metriken bewertet und mithilfe eines Klassifizierungsalgorithmus korrekt zugeordnet werden können. Um dies zu ermöglichen soll ein möglichst simpler Algorithmus gefunden werden, der eine möglichst hohe Genauigkeit der Zuordnung erreicht. Damit die Anwendung bedienbar ist, wird eine graphische Oberfläche entwickelt, die eine Kategorisierung aktueller Nachrichten erlaubt und einen Überblick über die zugrunde liegenden Klassifikationsdaten ermöglicht.

Die Arbeit beinhaltet Techniken aus dem Bereich des Text-Mining, welches sich mit Möglichkeiten der Gewinnung von neuen Informationen aus schwachstrukturierten Texten befasst und hat einige Überschneidungen mit der Information Retrieval, das sich damit befasst bestehende Informationen aufzufinden.

Weitere Kapitel

Die Gliederung hängt natürlich vom Thema und von der Lösungsstrategie ab. Als nützliche Anhaltspunkte können die Entwicklungsstufen oder - schritte z.B. der Softwareentwicklung betrachtet werden. Nützliche Gesichtspunkte erhält und erkennt man, wenn man sich

- in die Rolle des Lesers oder
- in die Rolle des Entwicklers, der die Arbeit z.B. fortsetzen, ergänzen oder pflegen soll.

versetzt. In der Regel wird vorausgesetzt, dass die Leser einen fachlichen Hintergrund haben - z.B. Informatik studiert haben. D.h. nur in besonderen, abgesprochenen Fällen schreibt man in populärer Sprache, so dass auch Nicht-Fachleute die Ausarbeitung prinzipiell lesen und verstehen können.

Die äußere Gestaltung der Ausarbeitung hinsichtlich Abschnittformate, Abbildungen, mathematische Formeln usw. wird in Kapitel 3 kurz dargestellt.

LaTeX-Bausteine

Der Text wird in bis zu drei Ebenen gegliedert:

```
    Kapitel (\chapter{Kapitel}),
```

- 2. Unterkapitel (\section{Abschnitt}) und
- 3. Unterunterkapitel (\subsection{Unterabschnitte}).

3.1 Abschnitt

Text der Gliederungsebene 2.

3.1.1 Unterabschnitt

```
1 Prozess 2:
2
3 Acquire();
4 b:= 1;
5 Release();
6 ...
7 Acquire();
8 if (a == 0)
9 {
10 c:= 5;
11 d:= b;
12 }
13 Release();
```

Größere Code-Fragmente sollten im Anhang eingefügt werden.

3.2 Abbildungen und Tabellen

Abbildung und Tabellen werden zentriert eingefügt. Grundsätzlich sollen sie erst dann erscheinen, nach dem sie im Text angesprochen wurden (siehe Abb. 3.1). Abbildungen und Tabellen (siehe Tabelle 3.1) können im (fließenden) Text (here), am Seitenanfang (top), am Seitenende (bottom) oder auch gesammelt auf einer nachfolgenden Seite (page) oder auch ganz am Ende der Ausarbeitung erscheinen. Letzteres sollte man nur dann wählen, wenn die Bilder günstig zusammen zu betrachten sind und die Ausarbeitung nicht zu lang ist (< 20 Seiten).

Prozesse	$\mathbf{Zeit} \rightarrow$	
$\overline{P_1}$	W(x)1	
P_2	W(x)2	
P_3	R(x)2 $R(x)$	1
P_4	R(x)2 R(x)	1

Tabelle 3.1. Bezeichnung der Tabelle

3.3 Mathematische Formel

Mathematische Formeln bzw. Formulierungen können sowohl im laufenden Text (z.B. $y=x^2$) oder abgesetzt und zentriert im Text erscheinen. Gleichungen sollten für Referenzierungen nummeriert werden (siehe Formel 3.1).

$$e_i = \sum_{i=1}^n w_i x_i \tag{3.1}$$

Entscheidungsformel:

$$\psi(t) = \begin{cases} 1 & 0 <= t < \frac{1}{2} \\ -1 & \frac{1}{2} <= t < 1 \\ 0 & sonst \end{cases}$$
 (3.2)

Matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$
(3.3)

Vektor:

$$\overline{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \tag{3.4}$$

3.4 Sätze, Lemmas und Definitionen

Sätze, Lemmas, Definitionen, Beweise, Beispiele können in speziell dafür vorgesehenen Umgebungen erstellt werden.

Definition 3.1. (Optimierungsproblem)

Ein Optimierungsproblem \mathcal{P} ist festgelegt durch ein Tupel $(I_{\mathcal{P}}, sol_{\mathcal{P}}, m_{\mathcal{P}}, goal)$ wobei gilt

- 1. $I_{\mathcal{P}}$ ist die Menge der Instanzen,
- 2. $sol_{\mathcal{P}}: I_{\mathcal{P}} \longmapsto \mathbb{P}(S_{\mathcal{P}})$ ist eine Funktion, die jeder Instanz $x \in I_{\mathcal{P}}$ eine Menge zulässiger Lösungen zuweist,
- 3. $m_{\mathcal{P}}: I_{\mathcal{P}} \times S_{\mathcal{P}} \longmapsto \mathbb{N}$ ist eine Funktion, die jedem Paar (x, y(x)) mit $x \in I_{\mathcal{P}}$ und $y(x) \in sol_{\mathcal{P}}(x)$ eine Zahl $m_{\mathcal{P}}(x, y(x)) \in \mathbb{N}$ zuordnet $(= Ma\beta \text{ für die Lösung } y(x) \text{ der Instanz } x)$, und
- $4. \ qoal \in \{min, max\}.$

Beispiel 3.2. MINIMUM TRAVELING SALESMAN (MIN-TSP)

- $I_{MIN-TSP} =_{def}$ s.o., ebenso $S_{MIN-TSP}$
- $sol_{MIN-TSP}(m,D) =_{def} S_{MIN-TSP} \cap \mathbb{N}^m$
- $m_{MIN-TSP}((m,D),(c_1,\ldots,c_m)) = \underset{def}{def} \sum_{i=1}^{m-1} D(c_i,c_{i+1}) + D(c_m,c_1)$
- $goal_{MIN-TSP} =_{def} min$

Satz 3.3. Sei \mathcal{P} ein NP-hartes Optimierungsproblem. Wenn $\mathcal{P} \in PO$, dann ist P = NP.

Beweis. Um zu zeigen, dass $\mathbf{P} = \mathbf{NP}$ gilt, genügt es wegen Satz A.30 zu zeigen, dass ein einziges \mathbf{NP} -vollständiges Problem in \mathbf{P} liegt. Sei also \mathcal{P}' ein beliebiges \mathbf{NP} -vollständiges Problem.

Weil \mathcal{P} nach Voraussetzung **NP**-hart ist, gilt insbesondere $\mathcal{P}' \leq_T \mathcal{P}_C$. Sei R der zugehörige Polynomialzeit-Algorithmus dieser Turing-Reduktion. Weiter ist $\mathcal{P} \in$

3.6 Literaturverweise 6

PO vorausgesetzt, etwa vermöge eines Polynomialzeit-Algorithmus A. Aus den beiden Polynomialzeit-Algorithmen R und A erhält man nun leicht einen effizienten Algorithmus für \mathcal{P}' : Ersetzt man in R das Orakel durch A, ergibt dies insgesamt eine polynomielle Laufzeit. \square

Lemma 3.4. Aus PO = NPO folgt P = NP.

Beweis. Es genügt zu zeigen, dass unter der angegeben Voraussetzung KNAP-SACK $\in \mathbf{P}$ ist.

Nach Voraussetung ist MAXIMUM KNAPSACK \in **PO**, d.h. die Berechnung von $m^*(x)$ für jede Instanz x ist in Polynomialzeit möglich. Um KNAPSACK bei Eingabe (x, k) zu entscheiden, müssen wir nur noch $m^*(x) \geq k$ prüfen. Ist das der Fall, geben wir 1, sonst 0 aus. Dies bleibt insgesamt ein Polynomialzeit-Algorithmus.

3.5 Fußnoten

In einer Fußnote können ergänzende Informationen¹ angegeben werden. Außerdem kann eine Fußnote auch Links enthalten. Wird in der Arbeit eine Software (zum Beispiel Java-API²) eingesetzt, so kann die Quelle, die diese Software zur Verfügung stellt in der Fußnote angegeben werden.

3.6 Literaturverweise

Alle benutzte Literatur wird im Literaturverzeichnis angegeben³. Alle angegebene Literatur sollte mindestens einmal im Text referenziert werden[CDK02].

¹ Informationen die für die Arbeit zweitrangig sind, jedoch für den Leser interessant sein könnten.

 $^{^2}$ http://java.sun.com/

³ Dazu wird ein sogennanter bib-File, literatur.bib verwendet.

Beispiel-Kapitel

In diesem Kapitel wird beschrieben, warum es unterschiedliche Konsistenzmodelle gibt. Außerdem werden die Unterschiede zwischen strengen Konsistenzmodellen (Linearisierbarkeit, sequentielle Konsistenz) und schwachen Konsistenzmodellen (schwache Konsistenz, Freigabekonsistenz) erläutert. Es wird geklärt, was Strenge und Kosten (billig, teuer) in Zusammenhang mit Konsistenzmodellen bedeuten.

4.1 Warum existieren unterschiedliche Konsistenzmodelle?

Laut [Mal97] sind mit der Replikation von Daten immer zwei gegensätzliche Ziele verbunden: die Erhöhung der Verfügbarkeit und die Sicherung der Konsistenz der Daten. Die Form der Konsistenzsicherung bestimmt dabei, inwiefern das eine Kriterium erfüllt und das andere dementsprechend nicht erfüllt ist (Trade-off zwischen Verfügbarkeit und der Konsistenz der Daten). Stark konsistente Daten sind stabil, das heißt, falls mehrere Kopien der Daten existieren, dürfen keine Abweichungen auftreten. Die Verfügbarkeit der Daten ist hier jedoch stark eingeschränkt. Je schwächer die Konsistenz wird, desto mehr Abweichungen können zwischen verschiedenen Kopien einer Datei auftreten, wobei die Konsistenz nur an bestimmten Synchronisationspunkten gewährleistet wird. Dafür steigt aber die Verfügbarkeit der Daten, weil sie sich leichter replizieren lassen.

Nach [Mos93] kann die Performanzsteigerung der schwächeren Konsistenzmodelle wegen der Optimierung (Pufferung, Code-Scheduling, Pipelines) 10-40 Prozent betragen. Wenn man bedenkt, dass mit der Nutzung der vorhandenen Synchronisierungsmechanismen schwächere Konsistenzmodelle den Anforderungen der strengen Konsistenz genügen, stellt sich der höhere programmiertechnischer Aufwand bei der Implementierung der schwächeren Konsistenzmodelle als ihr einziges Manko dar.

In [Che85] ist beschrieben, wie man sich Formen von DSM vorstellen könnte, für die ein beachtliches Maß an Inkonsistenz akzeptabel wäre. Beispielsweise könnte DSM verwendet werden, um die Auslastung von Computern in einem Netzwerk zu speichern, so dass Clients für die Ausführung ihrer Applikationen die am wenigsten ausgelasteten Computer auswählen können. Weil die Informationen dieser Art innerhalb kürzester Zeit ungenau werden können (und durch die Verwendung der

veralteten Daten keine großen Nachteile entstehen können), wäre es vergebliche Mühe, sie ständig für alle Computer im System konsistent zu halten [CDK02]. Die meisten Applikationen stellen jedoch strengere Konsistenzanforderungen.

4.2 Klassifizierung eines Konsistenzmodells

Die zentrale Frage, die für die Klassifizierung (streng oder schwach) eines Konsistenzmodells von Bedeutung ist [CDK02]: wenn ein Lesezugriff auf eine Speicherposition erfolgt, welche Werte von Schreibzugriffen auf diese Position sollen dann dem Lesevorgang bereitgestellt werden? Die Antwort für das schwächste Konsistenzmodell lautet: von jedem Schreibvorgang, der vor dem Lesen erfolgt ist, oder in der "nahen" Zukunft, innerhalb des definierten Betrachtungsraums, erfolgten wird. Also irgendein Wert, der vor oder nach dem Lesen geschrieben wurde.

Für das strengste Konsistenzmodell, Linearisierbarkeit (atomic consistency), stehen alle geschriebenen Werte allen Prozessoren sofort zur Verfügung: eine Lese-Operation gibt den aktuellsten Wert zurück, der geschrieben wurde, bevor das Lesen stattfand. Diese Definition ist aber in zweierlei Hinsicht problematisch. Erstens treten weder Schreib- noch Lese-Operationen zu genau einem Zeitpunkt auf, deshalb ist die Bedeutung von "aktuellsten" nicht immer klar. Zweitens ist es nicht immer möglich, genau festzustellen, ob ein Ereignis vor einem anderen stattgefunden hat, da es Begrenzungen dafür gibt, wie genau Uhren in einem verteilten System synchronisiert werden können.

Nachfolgend werden einige Konsistenzmodelle absteigend nach ihrer Strenge vorgestellt. Zuvor müssen wir allerdings klären, wie die Lese- und Schreibe-Operationen in dieser Ausarbeitung dargestellt werden.

Sei x eine Speicherposition, dann können Instanzen dieser Operationen wie folgt ausgedrückt werden:

- R(x)a eine Lese-Operation, die den Wert a von der Position x liest.
- W(x)b eine Schreib-Operation, die den Wert b an der Position x speichert.

4.3 Linearisierbarkeit (atomic consistency)

Die Linearisierbarkeit im Zusammenhang mit DSM kann wie folgt definiert werden:

- Die verzahnte Operationsabfolge findet so statt: wenn R(x)a in der Folge vorkommt, dann ist die letzte Schreib-Operation, die vor ihr in der verzahnten Abfolge auftritt, W(x)a, oder es tritt keine Schreib-Operation vor ihr auf und a ist der Anfangswert von x. Das bedeutet, dass eine Variable nur durch eine Schreib-Operation geändert werden kann.
- Die Reihenfolge der Operationen in der Verzahnung ist konsistent zu den <u>Echtzeiten</u>, zu denen die Operationen bei der tatsächlichen Ausführung aufgetreten sind.

Prozesse	$ \mathbf{Zeit} ightarrow$			
$\overline{P_1}$	W(x)1		W(y)2	
P_2		R(x)1		R(y)2

Tabelle 4.1. Linearisierbarkeit ist erfüllt

Die Bedeutung dieser Definition kann an folgendem Beispiel (Tabelle 4.1) nachvollzogen werden. Es sei angenommen, dass alle Werte mit 0 vorinitialisiert sind.

Hier sind beide Bedingungen erfüllt, da die Lese-Operationen den zuletzt geschriebenen Wert zurückliefern. Interessanter ist es, zu sehen, wann die Linearisierbarkeit verletzt ist.

Prozesse	$\mathbf{Zeit} ightarrow$		
$\overline{P_1}$	W(x)1	W(x)2	
P_2			R(x)0 R(x)2

Tabelle 4.2. Linearisierbarkeit ist verletzt, sequentielle Konsistenz ist erfüllt.

In diesem Beispiel (Tabelle 4.2) ist die Echtzeit-Anforderung verletzt, da der Prozess P_2 immer noch den alten Wert liest, obwohl er von Prozess P_1 bereits geändert wurde. Diese Ausführung wäre aber sequentiell konsistent (siehe kommender Abschnitt), da es eine Verzahnung der Operationen gibt, die diese Werte liefern könnte (R(x)0, W(x)1, W(x)2, R(y)2). Würde man beide Lese-Operationen des 2. Prozesses vertauschen, wie in der Tabelle 4.3 dargestellt, so wäre keine sinnvolle Verzahnung mehr möglich.

Prozesse	$ \mathbf{Zeit} ightarrow$		
$\overline{P_1}$	W(x)1	W(x)2	
P_2			R(x)2 R(x)0

Tabelle 4.3. Linearisierbarkeit und sequentielle Konsistenz sind verletzt.

In diesem Beispiel sind beide Bedingungen verletzt. Selbst wenn die Echtzeit, zu der die Operationen stattgefunden haben, ignoriert wird, gibt es keine Verzahnung einzelner Operationen, die der Definition entsprechen würde.

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.

Sachverzeichnis

Abbildung, 4 Abschnitt, 3

Beispiel, 5 Beweis, 5

Definition, 5

Echtzeiten, 8

Formel, 4

Freigabekonsistenz, 7

Inkonsistenz, 7

Kapitel, 3 Konsistenz, 7 schwach, 7

 $Konsistenz modelle,\ 7$

Lemma, 5

Linearisierbarkeit, 7, 8

Literatur, 6

Matrix, 5

Operation Lese, 8 Schreib, 8 Optimierung, 7

Quelltext, 3

Replikation, 7

Satz, 5 schwach, 8 sequentiell Konsistenz, 7 streng, 8

Tabelle, 4

Unterabschnitt, 3

Vektor, 5

Verfügbarkeit, 7

Glossar

DisASTer DisASTer (Distributed Algorithms Simulation Terrain),

A platform for the Implementation of Distributed Algo-

rithms

DSM Distributed Shared Memory

AC Linearisierbarkeit (atomic consistency)

SC Sequentielle Konsistenz (sequential consistency)

WC Schwache Konsistenz (weak consistency)
RC Freigabekonsistenz (release consistency)

Erklärung der Kandidatin / des Kandidaten

Die Arbeit habe ich selbstständig verfasst und keine anderen als die angegebe- nen 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:
Oatum Unterschrift der Kandidatin / des Kandidaten