Bachelorarbeit

„Erkennung von Fließtext in PDF-Dokumenten“

- 16.08.2016

- David Spisla

Albert-Ludwigs-Universität Freiburg

Technische Fakultät

Institut für Informatik

Gliederung

- Motivation
- Schwierigkeiten bei der Fließtextextraktion
- Lösungsansatz
- Empirische Analyse
- Vorführung des Programmes
- Ausblick

Motivation

- Aufbereitung und Transfer von Informationen (z.B. für sehbehinderte Menschen)
- Semantische Suche
- Dokumentenverwaltung
- Vorlesesysteme

Schwierigkeiten der Fließtextextraktion

- PDF ein rein layoutbasiertes Format

  → keine Information über strukturelle Zusammenhänge

- Entfernung irrelevanter Elemente

  (Bilder, Tabellen, Überschriften, Fußnoten, Referenzen, Code, Formeln, Kopf- und Fußzeilen, Bildunterschriften, etc.)

Lösungsansatz

Einsatz von Maschinellem Lernen

- Gewinnung von empirischem Wissen
- Erkennung von wiederkehrenden Mustern anhand von Trainingsdaten
- Anschließend Klassifikation von unbekannten Daten

Lösungsansatz

Warum Maschinelles Lernen?

- kein statisches Regelwerk, System reagiert flexibel auf abweichende Muster
- Anpassung der Trainingsdaten jederzeit möglich
- Moderne Rechner erlauben Verarbeitung hoher Trainingsmengen

Lösungsansatz

Voraussetzungen für den Erfolg des Verfahrens

- Sorgfältige Aufbereitung und Auswahl von Trainingsdaten
- Vermeiden eines „overfitting“

  (zu viele Features, zu hohe Anzahl an Trainingsschritten)

Lösungsansatz

Aufbereitung der Trainingsdaten mithilfe des Systems „icecite“

- Extraktion von Informationen zu Paragraphen und Zeilen eines PDFs
- Grundlage zur Erstellung der Featurevektoren zur Fließtexterkennung und Erkennung zusammengehörender Absätze

Lösungsansatz

Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is

Lösungsansatz

- PDF-Parser zur Erstellung einer tsv-Datei

Eigenschaft	Text	Seite	minX	minY	maxX	maxY	häufigste Schriftart	häufigste Schriftgröße	häufigste Schriftfarbe	(...)	Rolle
paragraph	Reducing quasi-ergodicity (...)	1	126.0	626.0	484.2	663.4	font-33	17.2	color-0		title
line	Reducing quasi-ergodicity (...)	1	126.0	647.9	484.2	663.4	font-33	17.2	color-0		title
line	by Tsallis Monte Carlo (...)	1	188.9	626.0	424.0	641.5	font-33	17.2	color-0		title
paragraph	Masao Iwamatsu†∗and Yutaka Okabe† ∗Department of (...)	1	134.6	526.9	475.6	609.1	font-32	12.0	color-0		unknown
line	Masao Iwamatsu†∗and Yutaka Okabe†	1	208.1	598.9	401.6	609.1	font-32	12.0	color-0		unknown

Lösungsansatz

- TeX-Parser zur Erstellung einer txt-Datei

Rolle	Startzeile	Endzeile	Koordinaten im TEX_FILE	Text
title	5	6	(1;125,99;529,35;484,25;663,38]), (1;125,79;180,31;386, 83; 187,98])	Reducing quasi-ergodicity (...)
authors	10	10	(1;[125,79;460,26;457,18;505,53])	Masao Iwamatsu[formula]
abstract	24	31	(1;[153,07;375,11;457,18;465,72])	A new Monte Carlo scheme based on the system of (...)
heading	43	43	(2;[125,79;695,17;241,09;705,13])	Introduction
text	45	59	(2;[125,79;521,77;484,45;678,39])	The ergodic hypothesis is fundamental to statistical mechanics. This (...)

Lösungsansatz

Warum der Einsatz von 2 Parsern?

→ Trainings- und Testdaten benötigen einen Ergebnisvektor d.h. jeder Paragraph braucht eine klare Klassifizierung (Fließtext / kein Fließtext). Diese Klassifizierung wird durch Koordinatenvergleich ermöglicht. Der TeX-Parser erkennt zuverlässiger Fließtext.

→ Supervised Learning

→ Optimierung der Trainingsdaten
langwierigste Phase in der Systementwicklung

Lösungsansatz

2 On the equivalence principle

New

According to Einstein's equivalence principle, see [20], gravity can be *locally simulated* in a gravity-free region of spacetime by going over from the Cartesian coordinates, anchored in an inertial frame of reference (including an intertial clock) and used in (1), to arbitrary curvilinear coordinates yielding a non-inertial frame in general, as in (3). In this context, the metric g_{ij}, occurring in (2) and in the semicolons of (3), is understood as a flat metric in curvilinear coordinates. Thus, the minimal coupling can be interpreted, in a first step, just as a coordinate transformation from Cartesian to curvilinear coordinates. And, moreover, it *identifies* the metric as the gravitational potential.

New

On the other hand, let us assume that we are in a region *with* gravity and (2) and (3) are valid together with the Einstein equation for the metric. Then, also according to Einstein's equivalence principle, we must be able to pick suitable coordinates such that locally the equations look like in special relativity in Cartesian coordinates. In Riemannian geometry, the local coordinates are called Riemannian normal (hence geodesic) coordinates at one point P, if the Christoffel symbols

$$\Gamma_{ij}{}^{k} := \frac{1}{2}\, g^{kl}\, (g_{il,j} + g_{jl,i} - g_{ij,l}) \tag{6}$$

Old

vanish at P and and the metric becomes Minkowskian:

$$\Gamma_{ij}{}^{k}|_{\mathrm{P}} \stackrel{*}{=} 0\,, \qquad g_{ij}|_{\mathrm{P}} \stackrel{*}{=} \mathrm{diag}\,(+1,-1,-1,-1)\,. \tag{7}$$

Old

Accordingly, the semicolon becomes a comma and the metric in (2), at one given point, looks flat.

New

Still, the curvature is non-vanishing, of course: $R_{ijk}{}^{l}|_{\mathrm{P}} \neq 0$. The equations look flat since they contain only *first* derivatives. If they contained second derivatives, then the semicolons goes to comma rule and its reverse would *not* work since on that level not only the Christoffels enter but potentially also the curvature which, in contrast to the Christoffels, is a tensor and cannot be nullified by means of a suitable choice of coordinates. For that reason, the minimal coupling

Lösungsansatz

Featurevektor zur Erkennung von Fließtext

häufigsteSchriftart	häufigsteSchriftgröße	häufigsteTextbreite	enthältSchlüsselwort	istBodyText	istReferenz	ErgebnisKlasse
0	0	1	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	0	0	0	0
1	1	0	0	0	0	0
0	0	0	0	0	0	0
1	1	1	0	1	0	1

Lösungsansatz

Featurevektor zur Erkennung von zusammengehörenden Absätzen

TSV_PARA	Einrückung	nachÜberschrift	GroßbuchstabeUndPunkt	ErgebnisKlasse
8	0	1	1	0
9	1	0	1	0
10	1	0	1	0
12	0	1	0	0
14	0	0	0	1
16	0	0	0	1
18	0	0	0	1
21	0	0	0	1
24	0	0	0	1

Lösungsansatz

- Die Daten beider Featurevektoren werden in je ein csv-File geschrieben
- Zu jedem PDF werden die entsprechenden csv-Files erstellt.
- Qualität der Featurevektoren wird in einer empirischen Analyse evaluiert

Empirische Analyse

- 200 verschiedene PDFs aus dem Bereich der MINT-Fächer
- Anwendung der Trainings- und Testdaten auf 5 verschiedene Lernverfahren:

  Logistic Regression, Naive Bayes, Decision Trees, k-Nearest-Neighbors, SVM
- 2 Varianten:

  → je 100 PDF als Training und 100 als Test

  → Alle 200 PDF als Training und Test

Empirische Analyse

Analysewerte und Wahrheitsmatrix der Fließtexterkennung mit Bernoulli Naive Bayes

class	precision	recall	f1-score	support
0.0	0.99 / 0.99	0.97 / 0.97	0.98 / 0.98	7996 / 15853
1.0	0.95 / 0.95	0.98 / 0.98	0.97 / 0.97	4717 / 9675
avg / total	0.98 / 0.97	0.98 / 0.97	0.98 / 0.97	12713 / 25528

	0 zugewiesen	1 zugewiesen
0 tatsächlich	7776 / 15381	220 / 472
1 tatsächlich	84 / 214	4633 / 9461

Empirische Analyse

Analysewerte und Wahrheitsmatrix der Erkennung zusammengehörender Absätze mit Bernoulli Naive Bayes

class	precision	recall	f1-score	support
0.0	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	2851 / 5642
1.0	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1866 / 4029
avg / total	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	4717 / 9671

	0 zugewiesen	1 zugewiesen
0 tatsächlich	2851 / 5642	0 / 0
1 tatsächlich	0 / 0	1866 / 4029

Vorführung des Programmes

Ausführungsschritte:

- Erstellung des tsv-File aus einem PDF

  (mit „icecite“)
- Erstellung eines csv-File aus den Daten des tsv-File
- Einlesen der csv-Files für die Trainingsphase
- Anwendung des trainierten Lernmodells auf die csv-Files des PDFs
- Erstellung einer txt-Datei mit den zusammengehörenden Fließtextparagraphen
- Optional: Erstellung eines markierten PDF

Ausblick

- Austesten der optimalen Trainingsmenge
- Systematische Sichtung der Trainings- und Test-PDFs (Verbesserung des Parsers und des Algorithmus zur Featurevektorenerstellung)
- Austesten einer Gewichtung einzelner Features
- Eines neues Feature zur Erkennung zusammengehörender Absätze
- Überlegungen zum Einsatz von semantischen Techniken bzw. Fehlertoleranzen

Danke für Ihre Aufmerksamkeit!