Ähnlichkeitssuchen

Phonetische und Distanzalgorithmen

Referent: Michael Zimmermann

Zimmermann@SZWeb.de

ts-projects GmbH

Organisation der Datenbankentwickler dbdev.org

Übersicht

- Motivation
- Problemstellung
- Phonetische Methoden
- Zeichenbasierte Methoden

Motivation

- Vermeidung von Mehrfacheingaben
- Auffinden erfaßter Dubletten
- Vereinheitlichung von Schreibungen
- Macht Spaß

Problemstellung

- Gleiche Inhalte kommen nicht zwingend in gleicher Gestalt
  - Verschiedene Schreibweisen
  - Rechtschreibfehler
  - Mißverständnisse
  - Tippfehler
- Wann sind verschiedene Strings ähnlich?

Ähnlichkeitsmaße 1

- Bei Zahlen
  - Absolut: Differenz
  - Relativ: Quotient
- Bei Strings
  - Katastrophe
  - Pragmatischer Ansatz: Was sind Fehlerquellen?

Ähnlichkeitsmaße 2

- Tippfehler: Tastaturdistanz
- Sprech-Hör-Fehler: phonetische Analyse
- Zeichenorientiert: String-Distanz, N-Gramm

Tastaturdistanz

- Zwei Zeichen ähneln sich um so mehr, je näher sie beieinanderliegen
- Dicke-Finger-Syndrom:
  - Sejr geeurtw Danen unf Herrenm
- Zehn-Finger-Syndrom:
  - Gtoftovj (Friedrich)

Phonetische Methoden

- Verschiedene Schreibungen mit gleicher oder ähnlicher Aussprache
- Versuch, diesen Prozeß umzukehren
- Exkurs Phonetik

Exkurs Phonetik 1

- Konsonanten sind eher Bedeutungsträger als Vokale
  - Rhbrbrkchn
  - aaeue
- Folgerung: Vokale ignorieren

Exkurs Phonetik 2

- Konsonanten unterscheiden sich nach
  - Artikulationsart
  - Artikulationsort
  - Nähe > Ähnlicher Klang

	bilabial		labio-dnt.		dental		alveolar		post-alv.		retroflex		palatal		velar		uvular		pharynga		glottal	
	stl.	sth.	stl.	sth.	stl.	sth.	stl.	sth.	stl.	sth.	stl.	sth.	stl.	sth.	stl.	sth.	stl.	sth.	stl.	sth.	stl.	sth.
Plosive	p	b					t	d			ʈ	ɖ	c	ɟ	k	g	q	ɢ			ʔ	
Nasale		m		ɱ				n				ɳ		ɲ		ŋ		ɴ				
Vibranten		ʙ						r										ʀ				
Taps/Flaps				ⱱ				ɾ				ɽ										
Frikative	ɸ	β	f	v	θ	ð	s	z	ʃ	ʒ	ʂ	ʐ	ç	ʝ	x	ɣ	χ	ʁ	ħ	ʕ	h	ɦ
laterale Frik.							ɬ	ɮ														
Approximanten				ʋ				ɹ				ɻ		j		ɰ						
laterale Appr.								l				ɭ		ʎ		ʟ						

Phonetische Ersetzung

- Funktion mit folgender Eigenschaft:
  - Argument: Natürlichsprachlicher String
  - Ergebnis: Grob vereinfachter Matchcode
  - **Nicht** Vergleich zweier Strings

Phonetische Ersetzung

- Graphem in Phonem umwandeln
  - Sprachabhängig
  - Kontextabhängig
- Irrelevante Laute eliminieren
- Ähnliche Phoneme zusammenfassen

Phonetische Ersetzung

- Beispiel
  - Chinachorknabe
  - ÇENEXERKNEPE
  - ÇNXRKNP
  - SNKRKNP
- Regeln:
  - Ch(**i)** > Ç; (**a)**ch > X i > E; n > N; a > E etc.
  - Alle E (Vokale) entfernen (optional)
  - Ç > S; X > K (optional)

Ergebnis verwerten

- Bei Datenerfassung Matchcode
  - … berechnen
  - … speichern
  - bei Bedarf auch mehrere
- Dadurch indizierte Suchvorgänge (Performance!)

Bekannte Verfahren

- Soundex
- Phonet
- Metaphone
- Caverphon
- Nysiis
- Kölner Phonetik

Soundex

Zeichen	Code
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

Kölner Phonetik

Zeichen	Kontext	Code
A, E, I, J, O, U, Y		0
H		-
B		1
P	nicht vor H	
D, T	nicht vor C, S, Z	2
F, V, W		3
P	vor H	
G, K, Q		4
C	im Anlaut vor A, H, K, L, O, Q, R, U, X	
	vor A, H, K, O, Q, U, X außer nach S, Z	
X	nicht nach C, K, Q	48
L		5
M, N		6
R		7
S, Z		8
C	nach S, Z	
	im Anlaut außer vor A, H, K, L, O, Q, R, U, X	
	nicht vor A, H, K, O, Q, U, X	
D, T	vor C, S, Z	
X	nach C, K, Q	

Bekannte Verfahren

- Prinzipiell gleiches Vorgehen
- Nur unterschiedliche Parameter
- Alle nicht völlig befriedigend
- Umsetzung i. d. R. hartcodiert

Vorstellung eigenes Verfahren

- Gleiches Prinzip, aber datenbankbasiert
- Regelkatalog beliebig erweiterbar
- Regelkatalog modifizierbar
- System ohne Programmierung lernfähig

Vorstellung eigenes Verfahren

- Verhaltensmodifikation über
  - Regelaktivierung/-deaktivierung
  - Regelpriorisierung
- Möglichkeit von Katalog-Templates
  - Für verschiedene Strenge
  - Für verschiedene Anwendungszwecke
  - Für verschiedene Sprachen
  - Zur Simulation der bekannten Verfahren

String-Distanz

- Levenshtein-Algorithmus
  - Vergleicht zwei Strings
  - Anzahl Operationen um von A nach B zu kommen
    - Einfügen
    - Löschen
    - Ersetzen
- Diverse Varianten, z. B. Damerau-Levenshtein, Weighted Levenshtein

N-Gramm-Methode

- Annahme: Je mehr möglichst lange Teilstrings identisch sind, desto ähnlicher sind zwei Strings
- Praktisch bewährt: Trigramme