Phonetische Algorithmen

Michael Fuchs

Lazaruskonferenz 2023

23.09.2022

Motivation

- Support-Hotline im Geschäftskundenbereich
- **Frage niemals nach der Kundennummer!**
- Regelmäßige Anrufer:
  - Wiepprecht
  - Wietbrecht
  - Lietprecht

Definition

Was ist ein phonetischer Algorithmus?

- Kodierverfahren für Worte (sprachl.)
- phonetische Repräsentation des Wortes
- dient als Grundlage für phonetische Suchen,
  d.h. Suche nach ähnlich klingenden Worten
- hauptsächlicher Einsatz bei Namen (Personen, Straßen, Worten)

bekannte phonetische Algorithmen

- **Soundex**
- Caverphone
- **Kölner Phonetik**
- Metaphone
- Porter-Stemmer-Algorithmus
- Miracode
- ...

Soundex

Soundex

- entwickelt von Robert C. Russell & Margaret King Odell
- patentiert im Jahr 1918
  - manuelle Berechnung notwendig
- ursprünglicher Einsatz:
  Indizierung der Familiennamen im United States Census
- weit verbreitete Integration
  - Oracle
  - MySQL
  - Freepascal
- vermutlich der am häufigsten angewendete phonetische Algorithmus

Soundex: Funktionsweise

- kopiere den ersten Buchstaben des Wortes
- kodiere die restlichen Buchstaben nach folgender Tabelle und füge sie an

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

- kürze das kodierte Wort auf vier Zeichen
- ist die Länge des Wortes kleiner als vier, fülle mit 0 auf

Soundex: Funktionsweise

Beispiele:

- Freepascal
  → F
  → F61224
  → F612
- Lee
  → L
  → L
  → L000

Nutzung in Freepascal

Sourcecode:

```
program TestSoundex;

uses
  StrUtils;

var
  EncodedWord: String;

begin
  EncodedWord := Soundex('Rodgers');
  WriteLn(EncodedWord);
end.
```

Nutzung in Freepascal

Ausgabe:

```
R326
```

Soundex: Probleme



Soundex: Probleme

- nur auf englische Sprache zugeschnitten

Brathering

Brathering

Bressering

Soundex: Probleme

- nur auf englische Sprache zugeschnitten
- nur grobe Analyse
  - “Britney Spears”→ B635 S162
  - “bewährten Superzicke”→ B635 S162
- ungenau
  - “Rogers”→ R262
  - “Rodgers”→ R326

Kölner Phonetik

Kölner Phonetik

- auch bekannt als “Kölner Verfahren”
- entwickelt 1969 von Hans Joachim Postel (BfV)
- selten in Standard-Bibliotheken oder Datenbanken integriert
- Implementierungen in vielen Programmiersprachen verfügbar

Kölner Phonetik: Funktionsweise

- codiere jeden Buchstaben von links nach recht gehende entsprechend der Umwandlungstabelle
- entferne alle mehrmals nebeneinander vorkommenden Ziffern
- entferne alle 0 außer am Wortanfang

Kölner Phonetik: Ersetzungstabelle

Buchstabe	Kontext	Code
A, E, I, J, O, U, Y, Ä, Ö, Ü		0
H		-
B		1
P	*nicht vor H*	1
D, T	*nicht vor C, S, Z*	2
F, V, W		3
P	*vor H*	3
G, K, Q		4
C	*im Anlaut vor A, H, K, L, O, Q, R, U, X*	4
C	*vor A, H, K, O, Q, U, X außer nach S, Z*	4
...	...	...

Kölner Phonetik: Ersetzungstabelle

...	...	...
X	*nicht nach C, K, Q*	48
L		5
M, N		6
R		7
S, Z		8
C	*nach S, Z*	8
C	*im Anlaut außer vor A, H, K, L, O, Q, R, U, X*	8
C	*nicht vor A, H, K, O, Q, U, X*	8
D, T	*vor C, S, Z*	8
X	*nach C, K, Q*	8

Kölner Phonetik: Funktionsweise

Beispiele:

- Freepascal
  → 370010805
  → 37010805
  → 37185

Kölner Phonetik: Nutzung in Freepascal

Implementation in Freepascal

- https://ypa-software.de/pages/de/libraries/koelner-phonetics.php
- LGPL mit Linking Exception
- Beispiele, Beschreibungen und Bibliotheken im SVN

Kölner Phonetik: Nutzung in Freepascal

Sourcecode:

```
program TestKoelnerPhonetik;

uses
  KoelnerPhonetics;

var
  EncodedWord: String;
  Encoder: TKoelnerPhonetics;

begin
  Encoder := TKoelnerPhonetics.Create;
  EncodedWord := Encoder.KoelnerPhonetic('Stefanie');
  WriteLn(EncodedWord);
  Encoder.Free;
end.
```

Kölner Phonetik: Nutzung in Freepascal

Ausgabe:

```
8236
```

Kölner Phonetik: Probleme

- stark auf die deutsche Sprache zugeschnitten
- Vokale im Wort werden nicht berücksichtigt
  - “Rhön” → 67
  - “Rohan”→ 67
- End-Vokale werden nicht berücksichtigt
  - “Stefanie”→ 8236
  - “Stephan”→ 8236

Kölner Phonetik: Vocal Removing

- zusätzliche Eigenschaft: **VocalRemoving**
- vrStandard - Standardverhalten
- vrNoTailRemoving - abschließende 0 wird nicht entfernt
- vrNoRemoving - keine 0 wird entfernt, außer doppelte Vorkommen

	vrStandard	vrNoTailRemoving	vrNoRemoving
Stephan	8236	8236	820306
Stefanie	8236	82360	8203060

Einsatz von phonetischen Algorithmen

Einsatzoptionen in DB-Systemen

- On-the-fly-Berechnung in SQL Abfrage
  - `SELECT * FROM Customers`
    `WHERE Soundex(Lastname) = Soundex('Meier');`
  - DB-System muss das unterstützen
    - Soundex in MySQL vorhanden
    - Kölner Phonetik müsste per UDF nachgerüstet werden
- vorberechnete Werte in DB
  - Worte werden bei INSERT/UPDATE codiert und in separate Spalten geschrieben
  - `SELECT * FROM Customers`
    `WHERE LastnameSoundex = Soundex('Meier');`
  - Codierung kann auch außerhalb Datenbank (im Programmcode) erfolgen
  - mehr Speicherplatz pro Datensatz notwendig

Vorberechnung vs. On-fly-Berechnung

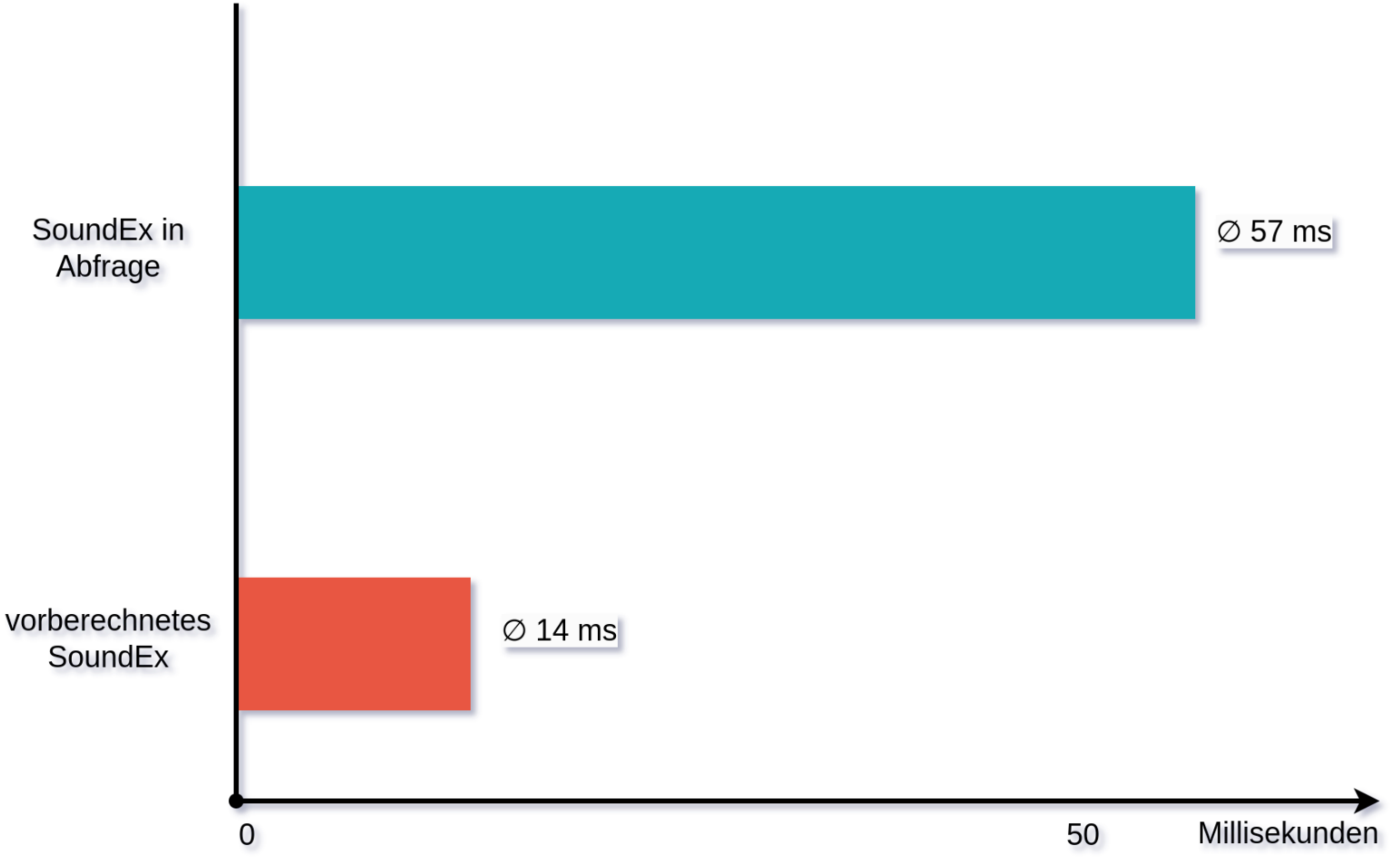


Abbildung: Soundex-Abfrage in MySQL, 1.000.000 Datensätze

Fragen

