Dokumentenmanagement und Groupware-Systeme

5. Suchen nach Informationen – Teil 1

Sommersemester 2007

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Institut für Angewandte Informatik und Formale Beschreibungsverfahren AIFB

Lehrstuhl für
Betriebliche Informationssysteme

Dr. Stefan Klink

Inhalt

Suchen nach Informationen – Phonetische Suche

1 Motivation

2 Idee von Davidson

3 Soundex

4 Phonex

1. Phonetische Suche

Schreibweise eines Namens/Wortes ist unklar.

Der Name wurde z.B. mündlich (Telefon) übermittelt.

Trotzdem soll er bei einer Suchanfrage gefunden werden.

Problem:

Das Eingeben aller denkbarer Schreibweisen ist mühselig (und in vielen Fällen nicht praktikabel).

1 Beispiel – Buchdatenbank

Szenarium:

Suche nach einem Autor

- Name darf falsch eingetippt werden
- Name kann aber auch falsch in der Datenbank gespeichert sein
- Verschiedene Varianten in der DB sind auch denkbar (inkonsistente Datensätze)

Ansatz:

Wenn kein Name syntaktisch identisch ist,
wird nach ähnlich klingenden Namen gesucht,
z.B.: Maier, Mayer, Meier, Meyer

1. Phonetische Suche

Ziel:

Auffinden von ähnlich klingenden Worten

Methode:

Indexierung des **phonetischen** Gehalts von Suchbegriffen

2. Idee von Davidson

Leon Davidson (1962):
„Retrieval of Misspelled Names in an Airline Passenger Record System“

Ignoriert werden:

- Vokale
- H, W, V
- Doppelte Konsonanten (der zweite wird gelöscht)

Ausnahme:

- Der erste Buchstabe bleibt unverändert

2. Idee von Davidson – Beispiele

Wait	WT
Weight	WGT
Knight	KNGT
Night	NGT
Nite	NT
Gnome	GNM
Noam	NM
Rees	RS
Reece	RC

3. Soundex

- Ursprünglich 1910 (!) entwickelt für eine Volkszählung in den USA.
  Hieß damals noch „***Miracode***“.
- 1918 patentiert zum Auffinden von Namen in Telefonverzeichnissen.

Idee:

Buchstaben werden nach phonetischen Gesichtspunkten in Gruppen eingeteilt.
Die Gruppen werden durch Zahlen repräsentiert

3. Soundex

Algorithmus:

1. Eliminierung aller nichtalphabetischen Zeichen (', -, blanc, etc.)
2. Alle Zeichen in Großschreibung
3. Initialisierung des Resultats mit dem ersten Buchstaben
4. Entfernung der nichtausgesprochenen Konsonanten H und W
5. Zeichenweise Ersetzung ab zweitem Buchstaben durch Ziffern:

3. Soundex

zu 5):

Labiale (Lippenlaute):

B, F, P, V $\rightarrow$ 1

Gutturale (Gaumen-, Kehllaute) und Sibilanten (Zisch-, Reibelaute):

C, G, J, K, Q, S, X, Z $\rightarrow$ 2

Dentale (Zahnlaute): D, T $\rightarrow$ 3

lange Liquiden (Fließlaute): L $\rightarrow$ 4

Nasale (Nasenlaute): M, N $\rightarrow$ 5

kurze Liquiden (Fließlaute): R $\rightarrow$ 6

3. Soundex

6. Kombination von zwei oder mehr identischen Ziffern zu einer und damit
   - LL $\rightarrow$ 4
   - SC $\rightarrow$ 2
   - MN $\rightarrow$ 5
   - ...

7. Entfernung der verbleibenden Buchstaben (Vokale A, E, I, O, U, Y)

8. Hinzufügung der ersten drei Ziffern zum Resultat. (Resultatslänge: 1 bis 4 Zeichen)

3. Soundex – Beispiele

McCloud	M243
MacCloud	M243
McLoud	M243
McLeod	M243
M'Cloud	M243

Smith	S53
Schmid	S53
Smid	S53
Smyth	S53
Schmidt	S53

Ng	N2
Eng	E52
Ing	I52

Rauchers	R262
Rogers	R262
Rodgers	R326
Rutgers	R326

Widerhold	W364
Weiderhold	W364
Widderholt	W364
Wiederholt	W364
Wiederhout	W363

3. Soundex – Fazit

Soundex ist etwas besser als Davidson
Aber noch weit vom Ziel entfernt

Probleme (u.a.):

- Keine Berücksichtigung von Ausspracheregeln
- Bleibender erster Buchstabe problematisch: Knight/Night
- Falsche Treffer: Powers, Pierce, Price, Perez, Park
- Gerade in den USA:
  viele Namen aus anderen Ländern
  → Regeln passen nur bedingt

4. Phonix – kurzer Einblick

- PHONIX: „PHONetic IndeX“

- ***Motivation***:
  Die Ergebnisse von Soundex durch
  Berücksichtigung von Ausspracheregeln verbessern.

  → Ähnlich zu Soundex, aber ***wesentlich komplexer***.

- Erste Implementierung Anfang der 80er Jahre in (Universitäts-) Bibliotheken

- Gadd (1988): „Fisching fore werds“

4. Phonix vs. Soundex

Was bleibt gleich?

- Ersetzen von Buchstaben durch Zahlen
- Ignorieren von Vokalen, W H Y, doppelten Konsonanten
- Erster Buchstabe bleibt erhalten

Was ist neu ?

- All dies geschieht ***NACHDEM*** einige Ersetzungsregeln angewendet wurden
- 8 statt 6 Buchstabenklassen
- Der erste Buchstabe wird (bis auf Ausnahmen) auch in den Zahlencode übernommen

4. Phonix – Algorithmus

Algorithm:

1. Perform phonetic substitutions (see table later on)
   - only the specified characters are dropped, e.g. the 'v' or vowel is not dropped in the substitution "N for NP if NPv";
   - the parameters are applied in the specified order;
   - process all occurences of one substitution before proceeding to the next substitution parameter;
   - the result of the substitution may create new target strings for substitution by subsequent parameters.
2. Retain the first character for the retrieval code.
3. Replace by 'v' if A, E, I, O, U.
4. Where names end in ES, drop the E.
5. Append an E where names end in A, I, O, U or Y.

4. Phonix – Algorithmus

6. Drop the last character regardless.
7. Drop the new last character if not A, E, I, O, U or Y
8. Repeat step 7 until a vowel (including Y) is found.
   This results is a word or name without its ending-sound
9. Strip all occurences of A, E, I, O, U, Y, H, W.
10. Remove one of all duplicate successive consonants.
11. Replace all consonants by their numeric value (see table later on)
12. Prefix the retrieval code with the retained first character (may be a 'v' [lowercase – see above]).
13. Repeat steps 9, 10 and 11 on the chracters removed as stripped ending-sounds.

4. Phonix – Algorithmus

Ending-sound algorithm:

1. If the ending sound values of an entered name and a retrieved name are the same, the retrieved name is a LIKELY candidate.
2. If an entered name has ending-sound value, and the retrieved name does not, then the retrieved name is a LEAST-LIKELY candidate.
3. If the two ending-sound values are the same for the length of the shorter, and the difference in length between the two ending-sound is one digit only, then the retrieved name is a LESS-LIKELY candidate.
4. All other cases result in LEAST-LIKELY candidates.

Sub	Start	Middle	End
G	DG	DG	DG
KO	CO	CO	CO
KA	CA	CA	CA
KU	CU	CU	CU
SI	CY	CY	CY
SI	CI	CI	CI
SE	CE	CE	CE
KL	CL if CLv		
K	CK	CK	CK
K			GC
K			JC
KR	CHR if CHRv		
KR	CR if CRv		
R	WR		
NK	NC	NC	NC
KT	CT	CT	CT
F	PH	PH	PH
AR	AA	AA	AA
SH	SCH	SCH	SCH

Phonetic Substitutions (1)

Sub	Start	Middle	End
TL	BTL	BTL	BTL
T	GHT	GHT	GHT
ARF	AUGH	AUGH	AUGH
LD		LJ if LJv	
LOW	LOUGH	LOUGH	LOUGH
KW	Q		
N	KN		
N			GN
N	GHN	GHN	GHN
N			GNE
NE	GHNE	GHNE	GHNE
NS			GNES
N	GN		
N		GN if GNc	GN if GNc
S	PS		
T	PT		
C	CZ		
Z		WZ if vWZ	
CH		CZ	

Phonetic Substitutions (2)

Sub	Start	Middle	End
LSH	LZ	LZ	LZ
RSH	RZ	RZ	RZ
S		Z if Zv	
TS	ZZ	ZZ	ZZ
TS		Z if cZ	
REW	HROUGH	HROUGH	HROUGH
OF	OUGH	OUGH	OUGH
KW		vQv	
Y		J if vJv	
Y	YJ if YJv		
G	GH		
E			GH if vGH
S	CY		
NKS	NX	NX	NX
F	PF		
T			DT
TIL			TL
DIL			DL
ITH	YTH	YTH	YTH

Phonetic Substitutions (3)

Phonetic Substitutions (4)

Sub	Start	Middle	End
CH	TJ if TJv		
CH	TSJ if TSJv		
T	TS if TSv		
CHE	TCH	TCH	TCH
VSKIE		WSK if vWSK	WSK if vWSK
N	MN if MNv		
N	PN if PNv		
SL		STIL if vSTL	STIL if vSTL
ENT			TNT
OH			EAUX
ECS	EXCI	EXCI	EXCI
ECS	X	X	X
ND			NED
DR	JR	JR	JR
EA			EE
S	ZS	ZS	ZS
AH		R if vRc	R if vRc
AH		HR if vHRc	HR if vHRc
AH			HR if vHR

Phonetic Substitutions (5)

Sub	Start	Middle	End
AR			RE
AH			R if vR
LE	LLE	LLE	LLE
ILE			LE if cLE
ILES			LES if cLES
null			E
S			ES
AS			SS if vSS
M			MB if vMB
MPS	MPTS	MPTS	MPTS
MS	MPS	MPS	MPS
MT	MPT	MPT	MPT

4. Phonix

Tabelle of numeric character values:

B, P	→ 1	M, N	→ 5
C, G, J, K, Q	→ 2	R	→ 6
D, T	→ 3	F, V	→ 7
L	→ 4	S, X, Z	→ 8

Key length:

7 significant consonants

Key format:

Annnnnn, where n represents a numerical character value in the range from 1 to 8.

Word	Soundex	Phonix
KNIGHT	K523	N53
NIGHT	N230	N53
NITE	N300	N53

WRITE	W630	R63
WRIGHT	W623	R63
RITE	R300	R63
WHITE	W300	W3
WEIGHT	W230	W3

YAEGER	Y230	v2
YOGA	Y200	v2
EAGER	E230	v2
AUGER	A230	v2

Phonix vs. Soundex

4. Phonix – Fazit

Deutliche Verbesserung gegenüber Soundex

Jedoch:

- wurde nie ausführlich von Gadd evaluiert
- Sehr geringe Verbreitung gegenüber Soundex

4. Fazit – Phonetische Suche

Weder Soundex noch Phonix sind perfekt

Es handelt sich um ein ***prinzipielles Problem***:

- Komplexes Regelwerk, das alle Feinheiten der Aussprache einfängt, existiert nicht
- Aufwand für das Regelschreiben sehr groß
- Aussprache kann sich verändern
- Verschiedene Sprechweisen können existieren

Dennoch:

für Standardanwendungen lassen sich brauchbare Ergebnisse erzielen

Literatur – Phonetische Suche

„Das Soundex System“
http://www.uni-oldenburg.de/nausa/soundex.htm

„Phonet – Doppelgänger gesucht“
http://www.heise.de/ct/ftp/99/25/252

W. B. Frakes, R. Baeza-Yates: Information Retrieval:
Data Structures and Algorithms, chap. 4, Englewood Cliffs, NJ:
Prentice Hall, 1992.