Ein Performanz-Vergleich zwischen der Kölner und der von Reth-Schek Phonetik

Tobias Bachteler und Rainer Schnell
Zentrum für quantitative Methoden und Surveyforschung
Universität Konstanz

13. März 2006

1 Einleitung

Für den Leistungsvergleich zwischen der Kölner Phonetik (Postel 1969) und der „von Reth-Schek Phonetik“ (von Reth und Schek 1977) wird ermittelt, wie erfolgreich beide Phonetiken bei der Zuordnung von fehlerhaften zu fehlerfreien Nachnamen sind. Zur besseren Einordnung wird der Vergleich auf den exakten Abgleich der Namen und den Abgleich anhand von Soundex (Knuth 1988) erweitert.

Phonetiken transformieren zwei Zeichenketten zunächst jeweils in einen phonetischen Code. Falls die phonetischen Codes übereinstimmen, wird dem Paar als Ähnlichkeitswert die Zahl 1 zugewiesen, sonst die Zahl 0. Die zugrunde liegende Vorstellung von Ähnlichkeit ist, dass zwei Zeichenketten sich dann ähnlich sind, falls sie denselben phonetischen Code aufweisen.

Der hier dargestellte Vergleich wurde anhand eines im Rahmen des DFG-Projektes „Filemerge mit fehlerhaften Schlüsseln“ (Schnell, Bachteler und Bender 2003; 2004) erstellten Testdatenkorpus und unter Verwendung des im selben Projekt entwickelten Record-Linkage Programms „Merge-Toolbox“ (Schnell, Bachteler und Reiher 2005) vorgenommen.

2 Testdaten

Um die Testdaten zu gewinnen wurden aus einem Einwohnermeldeamtsregister 1500 verschiedene Nachnamen zufällig ausgewählt und je 100 Namen auf 15 Tonbänder gesprochen. Je ein Tonband wurde im Sommersemester 2004 einer von insgesamt 15 Gruppen von Studenten der Universität Konstanz vorgespielt. Die Gruppen 1-10 haben die gehörten Namen in Schreibschrift, die Gruppen 11-13 in Blockbuchstaben

niedergeschrieben. Alle niedergeschriebenen Namen wurden dann von einer Sekretärin maschienenlesbar abgetippt. Die Gruppen 14 und 15 haben dagegen die gehörten Namen direkt in einen Rechner eingegeben. 288 Studenten nahmen an diesem Experiment teil, insgesamt liegen 27 301 potentiell fehlerbehaftete Nachnamen vor. Um die verschiedenen Stringähnlichkeitsfunktionen zu vergleichen, wurden jeweils die Ähnlichkeitswerte dieser 27 301 Namen mit allen 1500 fehlerfreien Nachnamen berechnet. Zu jedem der fehlerbehafteten Namen gehört einer und nur einer der 1500 fehlerfreien Namen. Der Vergleich gibt wieder, wie erfolgreich die verschiedenen Stringähnlichkeitsfunktionen die 1500 Originalnamen den fehlerbehafteten zugeordnet haben.

3 Vergleichsmethoden

Die hier betrachteten Stringähnlichkeitsfunktionen weisen nur zwei mögliche Funktionswerte auf: 1 oder 0. Als „positiv“ gilt ein Namenspaar, wenn die betrachtete Funktion den Ähnlichkeitswert 1, als „negativ“ wenn sie den Ähnlichkeitswert 0 zuweist. Namenspaare, die wirklich zusammen gehören, werden als „*matches*“ bezeichnet. Im gegenteiligen Fall handelt es sich um „*non matches*“. Daraus ergeben sich vier Möglichkeiten für jedes Namenspaar:

	match	*non match*
positive	richtig positiv (RP)	falsch positv (FP)
negative	falsch negativ (FN)	richtig negativ (RN)

Der Vergleich der Ähnlichkeitsfunktionen erfolgt anhand der für die Bewertung der Effektivität von Retrieval-Systemen meist verwendeten Kennwerte „Recall“ und „Precision“ (Salton und McGill 1987; Baeza-Yates und Ribeiro-Neto 2003). Der „F-Score“ (van Rijsbergen 1979) kombiniert Precision und Recall in einem Gütekriterium als harmonisches Mittel aus diesen. Die Gütekriterien sind wie folgt definiert:

$$Recall = \frac{\sum RP}{(\sum RP + \sum FN)}$$

$$Precision = \frac{\sum RP}{(\sum RP + \sum FP)}$$

$$F - Score = 2\frac{(Precision \times Recall)}{(Precision + Recall)}$$

4 Ergebnisse

Tabelle 1 zeigt die Ergebnisse der einzelnen Funktionen in Zahlen. Die im Vergleich zu Reth-Schek und dem exakten Abgleich mehr gefundenen „Richtig Positiven“ der Kölner Phonetik und von Soundex gehen mit einer wesentlich höheren Zahl an „Falsch Positiven“ einhergeht.

Die Kölner Phonetik findet zwar 4729 *matches* mehr als Reth-Schek, zeigt dabei jedoch 42583 mehr der *non matches* als positiv an.

In Tabelle 2 sind die Kennwerte *Recall*, *Precision* und *F-Score* für die Vergleichsfunktionen verzeichnet. Die Kölner Phonetik weist zwar einen höheren *Recall*-Wert als Reth-Schek auf, aber gleichzeitig einen wesentlich niedrigeren *Precision*-Wert. Der *F-Score* als Maß des Ausgleichs zwischen diesen Kennwerten weist für Reth-Schek einen etwas höheren Wert aus. Schließt man die exakten Übereinstimmungen für die Berechnung der Kennwerte aus, so ergeben sich für die Phonetiken die in Tabelle 3 gezeigten Ergebnisse. Die Rangfolge der Funktionen nach dem *F-Score* ändert sich dadurch nicht. Die Kölner Phonetik findet im Vergleich zu Reth-Schek in jedem

Tabelle 1: *Vergleichsergebnisse: Zahlen*

Funktion	*RP*	*FP*	*FN*	*RN*
Exakter Abgleich	12320	9409	14981	40914790
Kölner Phonetik	19213	55400	8088	40868799
Reth-Schek	14484	12429	12817	40911770
Soundex	18409	52352	8892	40871847

Tabelle 2: *Vergleichsergebnisse: Kennwerte*

Funktion	*Recall*	*Precision*	*F-Score*
Exakter Abgleich	0.451	0.567	0.503
Kölner Phonetik	0.704	0.258	0.377
Reth-Schek	0.531	0.538	0.534
Soundex	0.674	0.260	0.375

Tabelle 3: *Vergleichsergebnisse: Kennwerte ohne exakte Übereinstimmungen*

Funktion	*Recall*	*Precision*	*F-Score*
Kölner Phonetik	0.460	0.130	0.203
Reth-Schek	0.144	0.417	0.215
Soundex	0.406	0.124	0.190

Fall mehr „Richtig Positive“ auf, wobei diesem Vorteil eine wesentlich höhere Zahl an „Falsch Positiven“ entgegensteht. Im Rahmen eines Record-Linkage Verfahrens ist sowohl für Block- als auch für Matchvariablen ein möglichst guter Abgleich zwischen diesen beiden Eigenschaften zu beachten, wobei bei Blockvariablen die Eigenschaft des hohen *Recalls*, für Matchvariablen die der hohen *Precision* eher wichtig sein sollte. Reth-Schek zeigt sich als die sehr viel restriktivere Phonetik, bei der zugunsten einer hohen *Precision* ein vergleichsweise niedriger *Recall* erzielt wird.

5 Literatur

Baeza-Yates, R. A. und Ribeiro-Neto, B. (2003). Modern Information Retrieval. New York, NY: ACM Press.

Knuth, D. E. (1998). The Art of Computer Programming, Vol. 3: Sorting and Searching. 2. Aufl., Reading/Mass.: Addison-Wesley.

Postel, H. J. (1969). Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. IBM-Nachrichten, 19, 925-931.

von Reth, H.-P., und Schek, H.-J. (1977). Eine Zugriffsmethode für die phonetische Ähnlichkeitssuche (Technical Report No. 77.03.002). Heidelberg: IBM Scientific Center.

Rijsbergen, C. J. van (1979). Information Retrieval. 2. ed., London: Butterworths.

Salton, G. und M. J. McGill (1987). Information Retrieval. Grundlegendes für Informationswissenschaftler. Hamburg: McGraw-Hill.

Schnell, R., Bachteler, T., und Bender, S. (2003). Record Linkage Using Error Prone Strings. Proceedings of the joint statistical meeting, S. 3713-3717, American Statistical Association.

Schnell, R., Bachteler, T. und Bender, S. (2004): A Toolbox for Record Linkage. Austrian Journal of Statistics, 33(1-2), 125-133.

Schnell, R., Bachteler, T. und Reiher, J. (2005): MTB: Ein Record-Linkage-Programm für die empirische Sozialwissenschaft. ZA-Information, 56, 93-103.