Vergleich von Tarifmodellen unterschiedlicher Versicherungsgesellschaften am Beispiel der Behandlung von Berufen

Daniel Bärthel, Thomas Kudraß

Fakultät Informatik, Mathematik, Naturwissenschaften
Hochschule für Technik, Wirtschaft und Kultur Leipzig
Postfach 30 11 66
04251 Leipzig
dbaerthe@imn.htwk-leipzig.de
kudrass@imn.htwk-leipzig.de

Abstract: Dieses Papier beschreibt als ein Fallbeispiel einen Lösungsansatz zur automatischen Verknüpfung heterogener Berufslisten unterschiedlicher Versicherer. Dabei dienen die Berufsbezeichnungen als Schlüssel für semantische Äquivalenzen, weisen jedoch häufig unterschiedliche Schreibweisen auf. Die benutzten Berufslisten werden von jeweils unterschiedlichen Versicherungsgesellschaften verwaltet und für die Berechnung von Versicherungsprämien genutzt, die in einer Maklersoftware verglichen werden sollen. Um die Behandlung heterogener Berufslisten zu ermöglichen, kommen Techniken aus Text Mining und Computerlinguistik zum Einsatz.

1 Problembeschreibung

Versicherungsmakler bieten im Unterschied zu Versicherungsvertretern ihren Kunden eine Vielzahl von Versicherungsprodukten unterschiedlicher Anbieter an. Dabei setzen sie zur Beratung der Interessenten Software ein, die einen exakten und tagesaktuellen Vergleich der Versicherungstarife der unterschiedlichen Anbieter ermöglichen soll. Um zu einer für den Kunden optimalen Entscheidung zu kommen, muss der Makler mit dieser Software die Leistungen und Prämien unterschiedlicher Versicherungsprodukte vergleichen können. Eine neue EU-Vermittlerrichtlinie [EU02] sieht dabei vor, dass ab 2008 bei jedem Kunden eine transparente Beratung nachweisbar optimal abgestimmt auf dessen Bedürfnisse durchgeführt werden soll. Der Makler ist verpflichtet, seiner Empfehlung eine hinreichende Zahl von auf dem Markt angebotenen Versicherungsverträgen, die von verschiedenen Versicherungsunternehmen kommen müssen, zu Grunde zu legen.

Es gibt zurzeit viele Anbieter von Vergleichsprogrammen für einzelne Versicherungsarten wie Lebensversicherung und Berufsunfähigkeit, z. B. MORGEN & MORGEN oder Ino24. Die dort angebotenen Lösungen ermöglichen dem Makler jedoch keine exakte Bestimmung des Preises, um dem Kunden damit ein verbindliches Angebot zu unterbreiten, das dem nach der VVG-Reform [VVG07] vorgeschriebenen Antragsmodell gerecht wird. Aktueller Stand der Technik ist, dass die Versicherungsgesellschaften eigene Software und Webportale anbieten, mit denen Makler Tarife berechnen und Anträge erstellen können. Die von den Gesellschaften zur Verfügung gestellten Rechenkerne werden in der vom Softwarehaus Inveda.net entwickelten Maklersoftware unter einer einheitlichen Oberfläche integriert. Allerdings sind die Rechenkerne der Versicherungen nicht transparent und arbeiten mit unterschiedlichen Variablen. Um den gesetzlichen Vorgaben zu entsprechen, müsste ein Makler, sämtliche Tools (on- und offline) parallel bedienen und mit unterschiedlichsten Daten versorgen. Im Rahmen eines Kooperationsprojekts zwischen der HTWK Leipzig und der Inveda.net GmbH wird an einer Optimierungslösung auf Basis von Zwischenspeichern und Heuristiken gearbeitet, um den immensen Rechenaufwand beim Tarifvergleich zu reduzieren.

Die Probleme der Integration unterschiedlicher Tarifmodelle sollen am Beispiel des Parameters *Beruf* nachfolgend behandelt werden. Der vom Versicherungskunden ausgeübte Beruf geht als Risikofaktor in die Prämienberechnung bei der Lebensversicherung ein. Die Berufsbezeichnungen werden von den einzelnen Gesellschaften vorgegeben und sind nicht einheitlich. Versicherer A benutzt zum Beispiel *Imker* als Berufsbezeichnung, Versicherer B hingegen verwendet *Bienenzüchter*. Versicherer C hat *Bienenvater* als Schlüsselwort eingeführt. Hierbei wird ersichtlich, dass ein und derselbe Beruf durch mehrere unterschiedliche Synonyme referenziert wird. Deshalb muss hier eine Abbildung aus einer einheitlichen Berufsliste in die jeweilige Berufsliste des Versicherers gefunden werden. Diese Zuordnung sollte weitgehend automatisch erfolgen. Die Zuordnung von Berufen weist Parallelen auf mit dem Bestimmen von korrespondierenden Schemaelementen [CRF03, RB01]. Ebenso lassen sich Methoden der Duplikaterkennung nutzen [LN07], wobei nicht Tupel, sondern nur einfache Attributwerte analysiert werden sollen.

Das Papier hat folgenden Aufbau: In Abschnitt 2 wird die Grundlage des Mapping-Problems beschrieben und mit Beispielen verdeutlicht. Im 3. Abschnitt wird kurz auf die verwendeten Techniken und theoretischen Grundlagen eingegangen und die Lösungsidee skizziert. Abschnitt 4 zeigt den prinzipiellen Ablauf einer Berufsanfrage mit Synonymauflösung. Abschnitt 5 geht auf Performance-Aspekte ein. Mit einem Fazit schließt Abschnitt 6 das Papier ab.

2 Heterogenität der Berufslisten

Durch die wettbewerbsbedingt forcierte Intransparenz der Berechnungsvorschriften der Versicherer und durch die nicht standardisiert gewachsenen Systeme unterscheiden sich auch die benutzten Berufslisten in Format und Inhalt gravierend.

Der Umfang der Listen beträgt ca. 1000 bis 20000 Einträge. Die starke Abweichung ist auf die Speicherung von Synonymen und auf die Benutzung von Berufen obsoleten Charakters zurückzuführen.

2.1 Beispieltabellen für Berufslisten

Jede Versicherungsgesellschaft verwaltet ihre Berufslisten in unterschiedlichen Datenstrukturen. Neben der Berufsbezeichnung können in einer Tabelle andere Attribute stehen (vgl. Tab. 1), oder es handelt sich im einfachsten Fall nur um einspaltige Tabellen mit den Berufsbezeichnungen, die in unterschiedlicher Form erfasst sein können (mehr dazu in den Abschnitten 2.2 und 2.3)

Die folgenden Beispiele dienen als Anschauungsmaterial um die unterschiedlichen Strukturen der verwendeten Berufslisten zu verdeutlichen.

10	Abbrucharbeiter	472001	-	-	0,08	50	50	D		0	C

Tabelle 1: Tabellenstruktur von Versicherer C

2	Abbruchfacharbeiter	Abbruchfacharbeiterin

Tabelle 2: Tabellenstruktur von Versicherer D

5	Abbrucharbeiter/in	4727	1
30	Abfertigungsbeamt(er/in) (Zoll)	7621	1

Tabelle 3: Tabellenstruktur von Versicherer E

Versicherer A
Zylinderhutmacher
Zylindermacher (Hohlglasmacher)
Justizvollstreckungsobersekretaer

Versicherer B
Zylinderhutmacher/In
Hohlglasmacher/In
Justizvollzugsbeamtin/Justizvollzugsbeamter

Tabelle 4: Beispiele für Heterogenität der Berufslisten von zwei Versicherern

Akademiedirektor an Hoeheren Fachschulen und Akademien
Akademisch Gepruefter Uebersetzer

Tabelle 5: Beispiele für Komposita

2.2 Unterschiedliche Behandlung des Genus von Berufen

Die Speicherung geschlechtsabhängiger Berufsbezeichnungen wird sehr unterschiedlich von den Versicherern behandelt. Zum Einen werden die weiblichen Formen der Berufe in separaten Listen oder in Spalten in der gleichen Liste (Tabelle 2) abgelegt, oder gänzlich außer Acht gelassen. Zum Anderen wird zur Unterscheidung von weiblichen und männlichen Berufsbezeichnungen der Suffix *In* benutzt und dieser im selbigen Feld in Form von *Imker/Imkerin* oder *Imker/In* gespeichert. Eine Ausnahme bilden hierbei die weiblichen Berufe, die durch Suffixersetzung gebildet werden. In Tabelle 3 stehen die jeweiligen Suffixe für die Bildung des Genus in Klammern.

Da eine Regelbildung für die getrennte Verarbeitung des Genus zu aufwendig ist, wird sich nur auf die männliche Form bezogen. Die weiblichen Berufe bzw. deren Merkmale werden durch eine einfache Vorverarbeitung entfernt. Lediglich bei der Sucheingabe des Berufes muss beachtet werden, dass die etwaige weibliche Form umgewandelt wird.

2.3 Speicherung von Kohyponymen

Einige Versicherer benutzen Unterformen (Kohyponyme), die die Berufe kategorisieren. Wie man am Beispiel aus Tabelle 4 erkennen kann, benutzt Versicherer A Subkategorien (in Klammern), da die Eindeutigkeit der Zugehörigkeit zur Berufsbranche durch die Berufsbezeichnung *Zylindermacher* nicht gegeben ist. *Zylinderhutmacher* ist diesbezüglich das Hyperonym (Überbegriff). Versicherer B benutzt nur die eindeutigen Bezeichnungen. Einige der Versicherer pflegen diesen speziellen Beruf nicht. Diese greifen stattdessen auf die Hyperonyme wie *Anformer, Arrangeur, Ausschneider, Bridierer, Bimser* oder *Hutmacher* zurück. Unterkategorien werden zum Teil auch durch direkte Risikobeschreibungen abgebildet. Zum Beispiel *Hilfsarbeiter (mit erhöhten Unfallgefahren)* und *Hilfsarbeiter (ohne erhöhte Unfallgefahr)*.

Soll nun ein Beruf ausgewählt werden, ist es aufgrund dieser Eigenschaften möglich, dass für einige Versicherer keine Auswahl getroffen werden kann. Die Auflösung muss durch den Bearbeiter per manuellen Vergleich getroffen werden.

3 Verwendete Techniken

3.1 Approximate String Matching (Unscharfe Suche)

Um etwaige Fehleingaben oder Flexive in die direkte Suche einbeziehen zu können, kann man auf unscharfe Suchmethoden zurückgreifen. N-Gramme, Lose Diagramme, die Editier- und die Hamming-Distanz sind hierbei Methoden der Ähnlichkeitsbestimmung [Na01]. Die perfekte Methode für alle String-Matching-Probleme existiert nicht. Die Wahl des Algorithmus ist stark abhängig vom gewünschten Ergebnis und der Anwendung.

Dementsprechend ist eine Vorverarbeitung der Berufe in Betracht zu ziehen, findet aber aktuell noch nicht statt. Tabelle 4 und Tabelle 5 zeigen das eigentliche Problem. Ohne Vorverarbeitung wird es schwierig etwaige Benutzereingaben durch den Makler welche im vorliegenden Fall zum Beispiel „*Dozent*“ sein könnte, mit den in den Listen gespeicherten Berufen unscharf zu vergleichen. Der augenscheinlich passende Beruf „*Akademiedozent an Hoeheren Fachschulen und Akademien“* würde aufgrund der Unähnlichkeit (z.B. Editierabstand) somit nicht ausgewählt werden. Jedoch ist durch Auftrennen der Zeichenkette (Tokenizing) in einzelne Token durch Bildung von n-Grammen (z.B. mit n=3) die eigentliche Berufsbezeichnung separierbar [BN05].

Die für die Suche in den Berufslisten benutzte Methode ist die Levenshtein-Distanz [Le65, MVM09]. Sie zählt zu den Editierdistanzen, auch Editierabstand genannt. Hierbei wird die Ähnlichkeit anhand des Maßes bestimmt, wie viele Editier-operationen (Einfügen, Ersetzen, Löschen) benötigt werden, um aus dem Quellstring den Zielstring herzuleiten. Umso geringer die Anzahl der Operationen ist, desto ähnlicher sind sich die Strings. Jedoch bleibt die bei Buchstabenverschiebungen entstandene Verschiebungsdistanz unberücksichtigt. Eine Erweiterung dieser Distanz ist die sogenannte Damerau-Levenshtein-Distanz. Diese erweitert die Funktionalität mit Berücksichtigung von Vertauschungen. Die Levenshtein-Distanz benötigt für diese Korrektur zwei Operationen, die Erweiterung hingegen nur eine. Eine speicher-optimale Variante ist der sogenannte Hirschberg- oder der Needleman-Wunsch-Algorithmus [La09].

3.2 Synonymsuche mit fokussierten Webcrawlern (data mining) und SOAP

Falls die indizierte Suche und die Fuzzy-Suche für das Suchwort negativ ausfallen, wird versucht, aus den angegebenen Synonymquellen ein bedeutungsgleiches Wort (Beruf) zu ermitteln. Zu benutzende Quellen sind zum Beispiel:

http://www.wie-sagt-man-noch.de
http://synonyme.woxikon.de/
http://wortschatz.uni-leipzig.de/

Wenn die Quelle über eine SOAP Schnittstelle verfügt, wird diese direkt benutzt, wenn dies nicht der Fall ist, werden die benötigten Daten mithilfe eines Crawlers extrahiert. Um die Güte der Ergebnisse für die Synonymtreffer zu erhöhen, kann man die Resultate der unterschiedlichen Quellen vergleichen und bei Übereinstimmung(en) das ermittelte Synonym mit einem zusätzlichen Gewicht versehen. Hierbei kann die Anzahl der Übereinstimmungen als Maßzahl dienen. Auf Basis dieser Maßzahl kann die Ergebnisliste für die Nachverarbeitung absteigend sortiert werden. Muss eine automatische Selektion für das entsprechende Synonym benutzt werden, ist dies für Fehlerminimierung unabdingbar.

Um die Latenzzeit der Anfrage zu verringern, werden mithilfe der vorliegenden Berufslisten die möglichen Synonyme aus den Quellen zwischengespeichert. Zur Speicherung bieten sich Datenbanktabellen mit Indizierung an. Weiterhin kann man mit User Defined Functions (UDF) Stringvergleichsalgorithmen direkt implementieren Das in dieser Arbeit benutzte RDBMS ist MySQL. In Systemen wie Oracle ist diese Funktionalität schon von Haus aus vorhanden [Ora10].

Der Synonymvergleich wird als einzige Möglichkeit betrachtet, automatische Zuordnungen zwischen heterogenen Wertelisten durchzuführen. Um eine vollständige Zuordnung zu gewährleisten, muss die Synonymauswahl im initialen Zustand des Systems, jedoch nur für die Fehlzuweisungen, manuell erfolgen.

4 Algorithmus für die Ermittlung eines äquivalenten Berufs

In der folgenden Abbildung wird ein vereinfachter Überblick gegeben über die automatische Auflösung eines Berufes über alle vorhandenen Berufslisten.

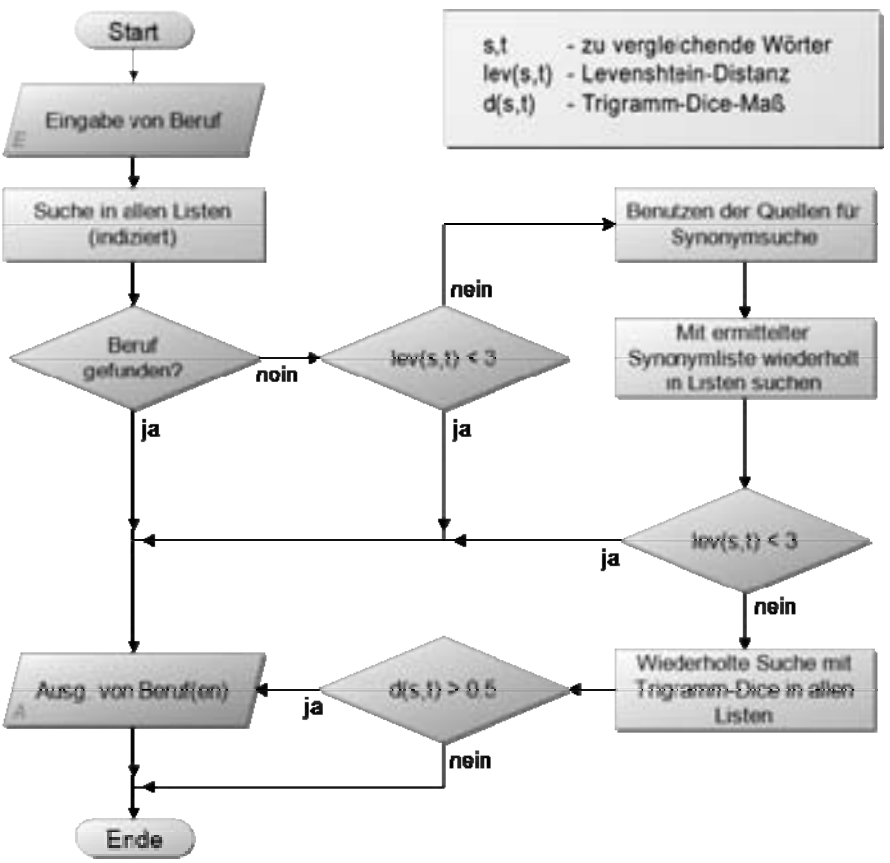

Abbildung 1: Ablauf der Auflösung eines Berufes über alle Berufslisten

Nach der Eingabe des gewünschten Berufes beginnt die Suche in allen vorhandenen Berufslisten. Danach erfolgt die Auswertung der ermittelten Berufe. Sollte hierbei der ermittelte Editierabstand den definierten Grenzwert überschreiten oder für den Beruf keine Übereinstimmung gefunden werden, wird eine Auflösung über die Synonymfindung versucht. Der Grenzwert für den Editierabstand wurde experimentell mit dem Korpus aus der Fusionierung aller Berufslisten zu einer Wortliste ermittelt. Problematisch ist diese statische Festlegung bei Wortlängen unter sechs Zeichen. Um dem entgegenzuwirken, kann man den Grenzwert dynamisch aus der Wortlänge berechnen. Je länger ein Wort ist, desto mehr Fehler sind möglich, um dennoch eine Ähnlichkeit zu erhalten. Dies sollte jedoch mit einem oberen Grenzwert versehen werden.

Für die Synonymfindung werden die vordefinierten Quellen benutzt. Wurden Synonyme gefunden, wird mit diesen erneut in den Listen gesucht. Die Zuweisung erfolgt durch die größtmögliche ermittelte Ähnlichkeit zwischen Synonym und dem jeweiligen Beruf.

Für jeden Versicherer liegt hierbei eine Liste vor. Da die Listen unterschiedliche Formate aufweisen (siehe Abschnitt 2.1) und für die Weiterverarbeitung lediglich der Beruf von Belang ist, werden die Listen naiv analysiert, die Position des Berufs (Spalte) automatisch ermittelt und interne Hashtabellen erzeugt. Um etwaige übereinstimmende Berufe von der Synonymsuche auszuschließen, wird der angegebene Beruf per Hash gesucht. Die Listen mit Treffern werden von der weiteren Verarbeitung ausgeschlossen.

Die verbliebenen Listen werden nun mithilfe der unscharfen Suche wiederholt durchsucht, um mit dem angegebenen Editierabstand syntaktisch ähnliche Berufe zu ermitteln. Sollte dieser Vergleich jedoch zu keiner Auflösung des Berufes über die verbliebenen Listen führen, werden anschließend mithilfe der Quellen zu diesem Beruf Synonyme gesucht. Mit den ermittelten Synonymen wird nun erneut in den Listen nach Übereinstimmungen gesucht.

Sollte diese Methode auch fehlschlagen, kann der gewünschte Beruf mindestens einem Versicherer nicht zugeordnet werden und das Mapping muss per Hand erfolgen. Dies bedeutet, dass der Makler den gesuchten Beruf manuell den in den Listen sich befindlichen Berufen zuweisen muss. Alle anderen gefundenen Übereinstimmungen werden dem Versicherer zugeordnet ausgegeben.

5 Performancebetrachtung und Evaluierung der Ergebnisse

Für die Durchführung der Tests wurde ein Intel©Celeron®M mit 1.5GHz und 2 GByte RAM benutzt. Als OS wurde Windows XP Professional® eingesetzt.

Die Reaktion des Systems auf eine Berufsanfrage ist stark abhängig von den benutzten Listen (Mächtigkeit), da die unscharfe Suche linear über alle Listen ausgeführt wird. Jedoch benötigt das System nur 1-2 Sekunden für die Suche in einer Liste mit 20000 Berufen.

Die Verarbeitungszeit der WWW-Quellen ist vernachlässigbar, da es sich um rein textuelle Requests handelt, jedoch ist diese abhängig von der Erreichbarkeit der Server. Da es sich um ein Einzelplatzsystem handelt, welches von einem Makler bedient wird, ist die Leistung als ausreichend zu bezeichnen.

Um die Güte des Verfahrens zu bewerten, wird dessen Effektivität über die Maße *precision* und *recall* ausgedrückt. Hierbei sind die *relevanten Treffer* die korrekten Zuordnungen eines gesuchten Berufes je Liste. Die *Gesamtzahl der Treffer* umfasst die Falschzuweisungen durch Ähnlichkeitsfehler oder Synonymfehlzuweisungen und die *relevanten Treffer*. Als Testkorpus wird eine zufällige Auswahl von 1500 Berufen aus der Liste eines Versicherers 5 (V5) benutzt. Die Berechnungen der Maßzahlen stammen aus dem Bereich des Information Retrieval [BR99]:

$$precision = \frac{relevante\_Treffer}{Gesamtzahl\_der\_Treffer} \qquad recall = \frac{relevante\_Treffer}{Umfang\_Testkorpus}$$

$$f-measure = \frac{2 \times recall \times precision}{recall + precision}$$

	Berufe	Idx+	Sim+	Syn+	$prec_{Syn+}$	$recall_{Syn+}$	f-measure	Ø sek/req
V1	1665	919	986	1228	0.98	0,80	0.88	1.630
V2	2013	1118	1208	1425	0.92	0,87	0.89	2.325
V3	1154	503	565	842	0.95	0.53	0.64	1.502
V4	2249	756	812	1023	0.82	0.56	0.67	2.418

Tabelle 6: Testergebnisse

Tabelle 6 zeigt die Ergebnisse der Evaluierung. Es wurden fünf Testreihen mit unterschiedlichen Korpora erzeugt. Idx+, Sim+ und Syn+ bezeichnen die Trefferanzahl des jeweiligen Verarbeitungsschrittes, wie Indexsuche, Ähnlichkeitssuche und Synonymsuche. Die Auswertung zeigt, dass die Heterogenität der Wertelisten sich vor allem negativ auswirkt, wenn die Berufe mithilfe von komplexeren Nomenkomposita kategorisiert werden. Die Ergebnisse von Versicherer V3 und V4 (siehe Tab. 4) zeigen die Auswirkung. Die hohe Precision weist darauf hin, dass die gefundenen Berufe den tatsächlichen Berufen entsprechen, der relativ niedrige Recall besagt jedoch auch, dass oftmals keine Zuweisung in der jeweiligen Liste getroffen werden konnte. Die Performance bezieht sich auf den Durchschnitt aller Requests für die jeweilige Liste des Versicherers.

6 Fazit und Ausblick

Es wurde ein Ansatz zur automatischen Lösung von semantischen Wortäquivalenzproblemen beschrieben. Dieser ist jedoch stark abhängig von den benutzten Synonymquellen. Die Güte dieser Quellen zeigt sich deutlich in der Qualität der erkannten Wortabhängigkeiten und Synonyme.

Aufgrund der noch fehlenden Vorverarbeitung für Problemfälle (vgl. Tab. 4 und 5) ist die Qualität der Ergebnisse für diese Suchreihen nicht zufriedenstellend, denn es findet keine Auflösung statt. Performancetechnisch gibt es keine Unzulänglichkeiten, auch wenn es noch Potential für Verbesserungen in Hinblick auf Datenstrukturen und Datenbanken gibt. Insgesamt kann jedoch gesagt werden, dass dieser Ansatz eine Möglichkeit der Lösung darstellt in Bezug auf Funktionalität und Performance. Im Verlauf der weiteren Entwicklung werden zusätzliche Vorverarbeitungsschritte hinzugefügt, um die eben angesprochene Problematik adäquat zu lösen. Außerdem könnten die Ergebnisse der Levenshtein-Distanz mit denen der Losen Diagramme kombiniert werden, um deren Eigenschaften der besseren Toleranz gegenüber Blockumstellung und Buchstabenverdrehern zu nutzen.

Danksagung

Das Projekt wurde unterstützt von der Sächsischen Aufbaubank im Rahmen des Förderprogramms „Innovative technologieorientierte Verbundprojekte auf dem Gebiet der Zukunftstechnologien im Freistaat Sachsen“ (Projekt-Nr. 13004/2189).

Literaturverzeichnis

[BN05] Bilke, A.; Naumann, F.: Schema Matching Using Duplicates. In: Proc. of the International Conference on Data Engineering (ICDE), Tokyo, 2005.

[BR99] Baeza-Yates, R.A.; Ribeiro-Neto, B.A.: Modern Information Retrieval. Addison-Wesley, 1999.

[CRF03] Cohen, W.W.; Ravikumar, P.; Fienberg, S.E.: A comparison of String Distance Metrics for Name-Matching Tasks, In: Proc. of IJCAI-03 Workshop on Information Integration, S. 73-78, 2003.

[EU02] Richtlinie 2002/92 des Europäischen Parlaments und des Rates vom 9. Dezember. 2002 über Versicherungsvermittlung, http://www.beratungsprotokoll.de/service/rili_2002/vermittlerrichtlinie.php.

[La09] Lang, H.W.: Needleman-Wunsch-Algorithmus. FH Flensburg, http://www.inf.fh-flensburg.de/lang/bioinformatik/needleman-wunsch.htm, 11.12.2009.

[Le65] Vladimir I. Levenshtein: Binary codes capable of correcting spurious insertions and deletion of ones. Problems of Information Transmission 1, S. 8-17, 1965.

[LN07] Leser, U.; Naumann, F.: Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. dpunkt Verlag, 2006.

[MVM09] Miller, F.P.; Vandome, A.F.; McBrewster, J.: Levenshtein-Distanz. Alphascript Publishing, 2009.

[Na01] Navarro, G.: A guided tour to approximate string matching, ACM Computing Surveys 33(19), S.31-88, 2001.

[Ora10] "Unscharfe Suche" in Datenbeständen vom 05. Mai 2010, Oracle Deutschland, http://www.oracle.com/global/de/community/tipps/adressen_text/index.html.

[RB01] Rahm, E.; Bernstein, P.: A survey of approaches to automatic schema matching. VLDB Journal 10(4), S. 335 – 340, 2001.

[VVG07] Gesetz zur Reform des Versicherungsvertragsrechtes vom 23.11.2007. In Bundesgesetzblatt Jahrgang 2007 Teil I Nr. 59 vom 29.11.2007; S. 2631-2678. http://bundesrecht.juris.de/vvg_2008/index.html.