SKIL
Studentenkonferenz
Informatik Leipzig
02. Dezember 2011

Entwicklung von IR-Algorithmen zur automatischen Bewertung von Krankenversicherungstarifen

Stefan Veit
HTWK Leipzig
s.veit86@gmx.de

HTWK Leipzig
HfTL
UNIVERSITÄT LEIPZIG

Agenda

(1) Einleitung und Motivation

(2) Grundlagen

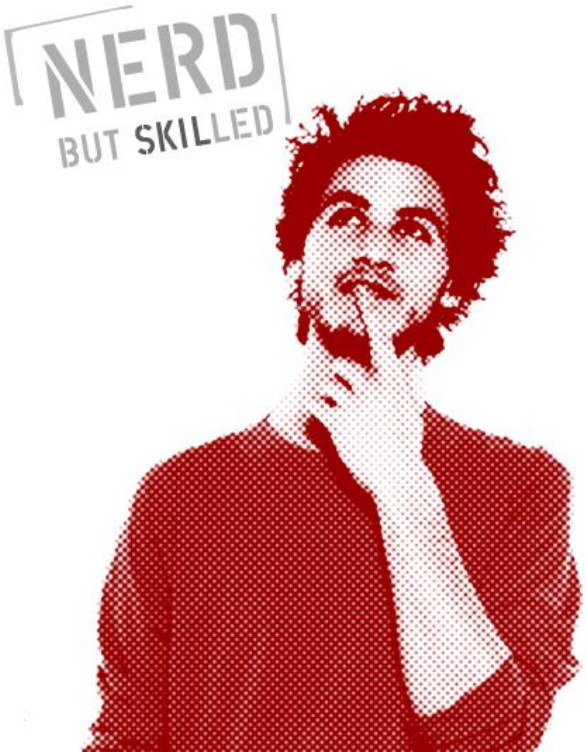
- Problembeschreibung
- Struktur von Versicherungswerken
- Verwendete Techniken

(3) Problemlösung

- Umsetzung von Analyseverfahren
- Ablauf des Bewertungsprozesses

(4) Ergebnisse und Zusammenfassung

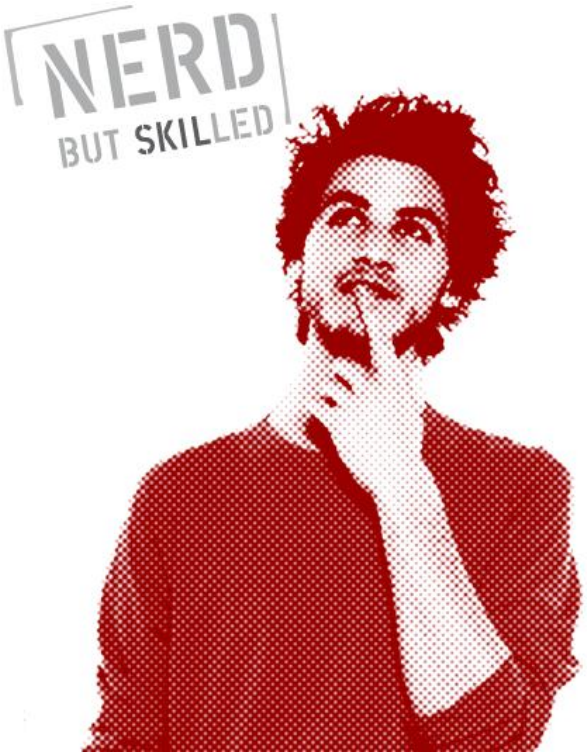
EINLEITUNG UND MOTIVATION



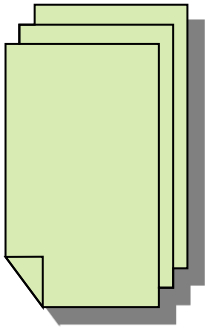
Einleitung und Motivation

- Im Bereich privater Krankenversicherungen Vielzahl von Gesellschaften und angebotenen Tarifen
- Individuell abgestimmte, tagesaktuelle Vergleiche von unabhängigen Versicherungsmaklern für Kunden
- Vom Makler verwendete Software enthält ausreichende Anzahl an Versicherungen sowie Informationen über Leistungs- und Prämienstruktur
- Notwendige Informationen werden von Drittanbietern geliefert, die diese manuell aus den Versicherungswerken extrahieren
- Manueller Analyseprozess gestaltet sich zeit- und kostenintensiv

GRUNDLAGEN



Problembeschreibung



Allgemeine Versicherungsbedingungen 2008 (AVB 2008)

für die Krankheitskosten- und Krankenhaustagegeldversicherung

„Welche Hilfsmittel werden erstattet?“

Tarif 190/1 Erstattungsfähig sind die Kosten für: Arznei- und Verbandmittel, Heilmittel, Hilfsmittel (ohne Sehhilfen) bis zu einer Selbstbeteiligung von 200 EUR je Versicherungsjahr bei Erwachsenen bzw. 100 EUR bei Kindern und Jugendlichen (zu 85%). Darüber hinaus zu 100%

- Als Hilfsmittel gelten ausschließlich Bandagen, Bruchbänder, Fußeinlagen, Gummistrümpfe, Hör- und Stützapparate, handgetriebene Krankenfahrstühle, Kunstglieder, Leibbinden, Sprechgeräte (elektronischer Kehlkopf).

Hilfsmittel	
Kostenübernahme von Hilfsmitteln	Ja
Offener Hilfsmittelkatalog	Nein
Brillen	Ja
Kontaktlinsen	Ja
Krankenfahrstuhl	Ja
Blindenhund	Ja
Blindenstock	Ja
Kunstauge	Ja
Brustprothesen	Ja
Beinprothesen	Ja
Armprothesen	Ja
Sanitäre Verbrauchsartikel	Ja
Orthopädische Schuhe	Ja

Problembeschreibung (Fortsetzung)

- Ziel ist die Umsetzung einer automatisierten Bewertung der Versicherungstarife
- Umsetzung im Rahmen des Projektes anhand eines Beispielszenarios
  - Analyse von 3 ausgewählten Tarifen (Ergebnisse der manuellen Analyse bekannt + manuelle „Arbeitsschritte“ dokumentiert)
  - Modellierung der Tarifstruktur
  - Bewertung bestehender Algorithmen, ggf. Anpassung dieser Methoden an die KV-Domäne
  - Generierung von Resultaten als Wahrheitswerte (Ja/Nein-Liste) bzw. als numerischer Wert (Höhe der erstatteten Leistung)
  - Aufbau einer Wissensbasis, Prozess durch Expertenwissen unterstützen

Struktur von Versicherungswerken

- Meist einheitlicher Aufbau der Dokumente
- Informationen als fortlaufender Text, Einteilung nach Leistungskategorien (allgemeine Bedingungen, Krankenhaustagegeld, Pflegeversicherung, usw.)
- Leistungsbeschreibungen orientieren sich an Musterkatalog des Verbands privater Krankenversicherungen
  - Unverbindliche Empfehlung für Gestaltung der Versicherungstexte
  - Notwendiger und vorgeschriebener, minimaler Versicherungsschutz
  - Vereinfacht Sammlung von Hintergrundwissen, da meist einheitliche Formulierung
- Wissensrepräsentation in Listen/ Aufzählungen bzw. Tabellenform

Verwendete Techniken

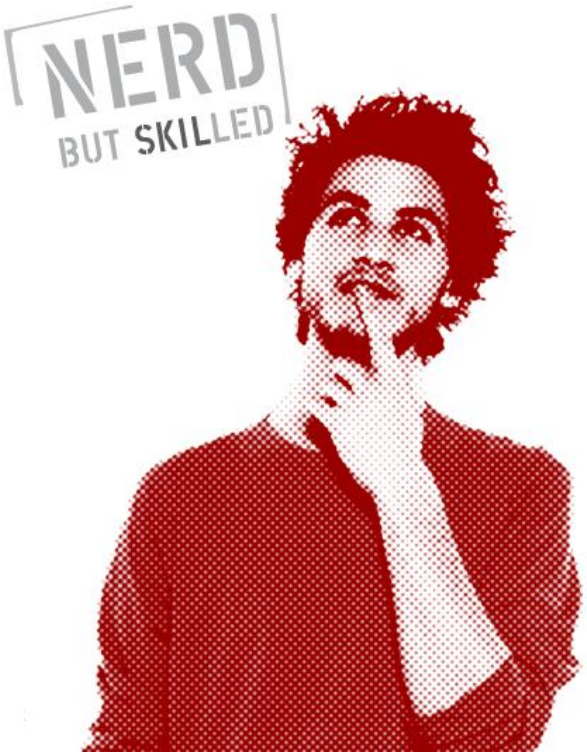
- Text Mining
  - Computergestütztes Verfahren zur semantischen Analyse von Texten [HWQ06]
  - Strukturierung von Texten, Aufdecken von Zusammenhängen in und zwischen Texten
  - Vorverarbeitung bzw. „Bereinigung“ von Texten für Projektumsetzung relevant

- Informationsextraktion
  - Suche und Extraktion relevanter Informationen innerhalb eines Textes [Cun05]
  - Gewinnung von Wissen für ein Fachgebiet aus unstrukturierten Objekten

Verwendete Techniken (Fortsetzung)

- Information Retrieval
  - Extraktion von Informationen aus Texten mittels Computerunterstützung, welche durch eine Suchanfrage angefordert werden
  - Identifizierung relevanter Textabschnitte bzw. Schlagwörter, welche Problemstellung lösen
  - Boolesches Retrieval
    - Anfragen werden mithilfe der booleschen Operatoren aufgebaut
    - Alle Begriffe der Suchanfrage müssen gefunden werden – keine Teiltreffer
  - Vektorraum Retrieval
    - Suchanfrage und Texte werden in Vektor transferiert
    - Ähnlichkeit der Vektoren als Maß für Relevanz

PROBLEMLÖSUNG



Umsetzung von Analyseverfahren

- Umsetzung von 4 Algorithmen – davon 3 bestehende Ansätze, eine „Eigenentwicklung“
- Anwendung von auf Schlag- bzw. Suchwörter aufbauende Regeln

Boolesche Analyse

- Verknüpfung von Suchwörtern mit Operatoren AND und NOT
- Zuweisung von festgelegten Ergebniswerten zu den Regeln
- Regel feuert nur, wenn kompletter Ausdruck zutrifft (keine „Teiltreffer“)
- Wenn keine passende Regel gefunden, keine Aussage möglich

Umsetzung von Analyseverfahren (Fortsetzung)

Erweiterte Boolesche Analyse mit Wortgewichtung

- Erweiterung um Gewichtungen/ Prioritäten für Suchwörter
- Nicht der ganze Ausdruck wird aufgebaut, jeder Teilausdruck für sich wird untersucht
- Bei Übereinstimmung: festgelegter Wert wird auf ein Gesamtgewicht aufaddiert
- Höhe des Gesamtgewicht entspricht Einordnung in Ergebnisbereich

Analyse mit regulären Ausdrücken

- Aufbau regulärer Ausdrücke aus Suchwörtern
- Nicht nur Wortvorkommen, auch struktureller Zusammenhang kann abgebildet werden (Wortreihenfolge)

Umsetzung von Analyseverfahren (Fortsetzung)

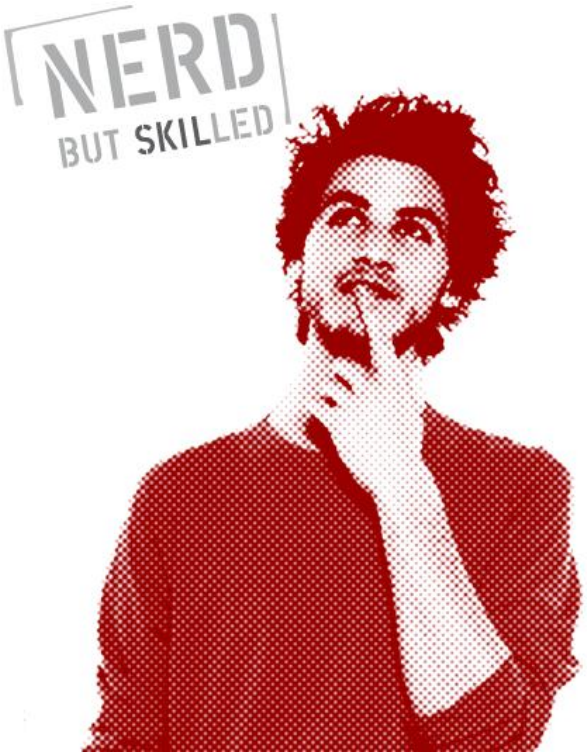
Listenerkennung

- Erkennung von Listen- und Aufzählungsstrukturen
- Erster Schritt: Überprüfung auf Liste (Merkmale wie unvollständige Sätze, Aufzählungssymbole)
- Zweiter Schritt: Einteilung in Positiv- und Negativliste (Aufzählung bzw. Ausschluss erstattungsfähiger Leistungen)
- Durchsuchen der Liste nach Schlagwörtern

Ablauf des Bewertungsprozesses

1. Auswahl der abzufragenden Leistung
2. Eingabe Textfragment
3. Textvorverarbeitung
    1. Auftrennung in Wortgruppen, Wörter oder Sätze (Tokenisierung)
    2. Herausfiltern von Stoppwörtern
    3. Wortstammbildung
4. Anwendung Analyseverfahren
    1. Anwendung der verfahrensspezifischen Regeln
    2. Bestimmung der „passenden“ Regel unter Berücksichtigung von Prioritäten
    3. Zuordnung der Ergebnisse

ERGEBNISSE UND ZUSAMMENFASSUNG



Ergebnisse des Beispielszenario

- Untersuchung „bekannter“ und „unbekannter“ Textpassagen
- Jeweils 19 Fragestellungen für 6 Tarife wurden analysiert
- Gegenüberstellung der automatisch ermittelten Ergebnisse mit denen der manuellen Analyse (=Referenzwert):
  - Übereinstimmung der Ergebnisse („Erfolgsfall“)
  - Keine Übereinstimmung, automatische Analyse findet kein Ergebnis
  - Keine Übereinstimmung, automatische Analyse liefert falsches Ergebnis

Ergebnisse des Beispielszenario (Fortsetzung)

Tarife	Erfolgsquote gesamt	Erfolgsquote Verfahren 1	Erfolgsquote Verfahren 2	Erfolgsquote Verfahren 3	Erfolgsquote Verfahren 4
BT1	84,2 %	73,7 %	52,6 %	73,7 %	5,3 %
BT2	89,4 %	89,4 %	42,1 %	89,4 %	15,8 %
BT3	94,7 %	94,7 %	47,4 %	94,7 %	0 %
UT1	78,9 %	63,2 %	36,8 %	36,8 %	10,6 %
UT2	81,3 %	68,2 %	26,3 %	52,6 %	0 %
UT3	69,2 %	53,8 %	7,7 %	15,4 %	15,4 %
	82,9 %	**73,8 %**	**35,5 %**	**60,4 %**	**7,9 %**

Verfahren 1… Boolesche Analyse
Verfahren 2… Boolesche Analyse mit Wortgewichtung
Verfahren 3… reguläre Ausdrücke
Verfahren 4… Listenerkennung

Fazit und Ausblick

- Algorithmen zur domänenspezifischen Problemlösung konnten gefunden und umgesetzt werden
- Verfahren unterscheiden sich in Genauigkeit, als Gesamtheit jedoch zufriedenstellende Werte
- Abbildung des manuellen Analyseprozesses ohne fachspezifisches Wissen oft problematisch
- Verbesserung im Bereich der initialen Informationsgewinnung mithilfe statistischer Verfahren

Quellen

- [Cun05] Hamish Cunningham: *Information Extraction, Automatic in Encyclopedia of Language and Linguistics*, 2nd Edition, Elsevier, 2005, Oxford
- [HWQ06] Gerhard Heyer, Uwe Quasthoff und Thomas Wittig. *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse.* W3l, 2006.

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT