

## **Text Mining zur Analyse der Anfragen an ein medizinisches Expertenforum**

Wolfgang Himmel  
Abteilung Allgemeinmedizin  
Georg-August-Universität  
Göttingen  
37073 Göttingen  
whimmel@gwdg.de

Frank Kroll  
Business Competence Center  
SAS Institute Heidelberg  
69118 Heidelberg  
Frank.Kroll@ger.sas.com

### **Abstract**

Die zunehmende elektronische Vernetzung führt zu einem steigenden Angebot an und Nutzung von Internet basierten medizinischen Expertenforen. Eine genauere Kenntnis des Inhalts der Anfragen – gerade auch in der Veränderung im Längsschnitt – könnte die Arbeit dieser Foren erleichtern, zugleich auf bisher vernachlässigte Informationswünsche hinweisen. Über 4000 Anfragen zum Thema „Unerfüllter Kinderwunsch“ an ein medizinisches Expertenforum wurden mit dem Text Miner von SAS untersucht. Die Analysemethoden wurden mehrfach variiert. Die ersten Cluster-Lösungen (mit konventioneller Stopp-Liste) spiegelten nur unzureichend die Informationsbedürfnisse der Besucher des Expertenforums wider. Eine Verfeinerung insbesondere der Liste von „Start“-Begriffen führte zu deutlich trennschärferen Clustern (z. B. ließen sich männerspezifische Themen, Anfragen zum optimalen Empfängniszeitpunkt bündeln). Durch Training und Validierung gelang nur bedingt eine Trennung und Klassifikation der Anfragen nach Struktur und Inhalt. Die Heterogenität des Inhalts und vor allem des Stils der Anfragen an das Expertenforum erschwert einfache Cluster-Lösungen mittels Text Mining. Größere Stichproben für Training und Validierung sowie ein mehr-dimensionales Cluster-Konzept dürften deutlich bessere Ergebnisse erbringen und damit die Nützlichkeit der Analysen (z. B. hinsichtlich des Informationsbedarfs im Gesundheitswesen) erhöhen.

**Keywords:** Text Mining, Internet, Expertenforum; Medizin, Kinderwunsch, Informationsbedarf text mining, internet, expert forum, medicine, wish for a child, information needs.

## Einleitung

Der folgende Beitrag beschreibt – im Sinne eines Arbeitsberichtes – verschiedene Methoden, mittels des SAS Text Miner die Anfragen an ein medizinisches Expertenforum im Internet zu klassifizieren. Einer solchen Analyse und Klassifikation liegt u.a. das Ziel zugrunde, Informationswünsche und (nicht erfüllte) Gesundheitsbedürfnisse von Patienten bzw. Laien zu erkennen und ggf. Abhilfe zu schaffen. Automatisierte, leistungsstarke Auswertungsroutinen sind dabei gefragt. Bevor das Projekt selbst skizziert wird, sollen theoretischer Hintergrund und Potential des Text Minings in Stichworten erläutert werden.

## 1 Theoretischer Hintergrund

SAS versteht unter Text Mining das Aufdecken und Verwenden von Wissen, das in einer Sammlung von Dokumenten existiert. Die einzelnen Analyseschritte, die im SAS Text Miner typischerweise nacheinander durchlaufen werden, sind:

- Einlesen der Dokumente
- „Linguistische“ Datenvorverarbeitung
- „Technische“ Datenvorverarbeitung
- Dimensionsreduktion
- Problemspezifische Mining-Analysen

Das **Einlesen der Dokumente** überführt die in textbasierten Formaten<sup>36</sup> vorliegende Information in ein SAS Data Set, in dem jedes Dokument als eine Beobachtung vorliegt.

Mit „**Linguistischer**“ **Datenverarbeitung** ist die erste Analysephase der Textinformation gemeint. So können u. a. mittels Stopp-, Start- und Synonymlisten die für die weitere Analyse zu verwendenden Begriffe gefiltert werden. Beispielsweise führt der Einsatz einer Stoppliste dazu, dass alle darin enthaltenen Begriffe von der weiteren Verarbeitung ausgeschlossen werden. Zusätzlich zu den Listen kann der Anwender auswählen, ob er abgeleitete Wortformen (z. B. Häuser; sprang) auf ihre Stammformen (Haus;

---

<sup>36</sup> Der Katalog einlesbarer Textformate umfasst ASCII, Word, PDF, HTML, Excel, Lotus, PowerPoint u. v. a.

springen) zurückführen und nur diese entsprechend weiterverarbeiten möchte. Weitere Möglichkeiten sind durch die Erkennung so genannter Entitäten (Orte, Personen, Organisationen, Produkte, etc.) oder die Identifikation aus mehreren Wörtern zusammengesetzter Ausdrücke (z. B. Herzlichen Dank, schöne Grüße) gegeben. Durch Einsatz dieser Vorverarbeitungsfunktionen können somit sprachliche Einheiten extrahiert und im Sinne des Analyseziels entsprechend weiterverarbeitet werden.

Der mit „**Technische**“ **Datenvorverarbeitung** umschriebene Analyseschritt beinhaltet die Gewichtung der auf Basis der „Linguistischen“ Datenverarbeitung generierten Häufigkeitstabelle (Term-by-Document-Matrix). Im Kern geht es darum, Begriffen, die in einer relativ kleinen Teilmenge der Dokumentensammlung häufig auftreten, ein höheres Gewicht zu verleihen, um dadurch den semantischen Inhalt dieser Dokumente entsprechend besser von den restlichen Dokumenten abgrenzen zu können. Im Gegensatz dazu bekommen Begriffe, die über die gesamte Dokumentensammlung relativ gleich verteilt sind, eine geringe Gewichtung, da sie vergleichsweise wenig zur Differenzierung der Dokumente beitragen.

Um die Bedeutung des nächsten Analyseschrittes einordnen zu können, sollte man sich vorab einmal klar machen, mit welchen Größenordnungen man es üblicherweise bei der o. g. Term-by-Document-Matrix zu tun hat: Schon bei recht überschaubaren Dokumentenanzahlen von einigen Hundert liegt die Anzahl der Begriffe oft im Bereich von Tausenden und mehr. Der im SAS Text Miner an dieser Stelle verfolgte Ansatz lautet deshalb **Dimensionsreduktion**; d. h., es werden Verfahren benötigt, um die in der Term-by-Document-Matrix enthaltene Information mit möglichst wenig Informationsverlust in einen niedrig(er) dimensionierten Raum zu transferieren. Hierzu stehen zwei Möglichkeiten zur Verfügung: Neben der „pragmatischen“ Verwendung der am höchsten gewichteten Begriffe („Roll-up terms“) bietet der SAS Text Miner mit der Singulärwert-Zerlegung ein leistungsfähiges mathematisches Verfahren der Matrizenfaktorisierung. Die als Ergebnis der Singulärwert-Zerlegung resultierenden numerischen Variablen („Columns“) stellen Linearkombinationen der zugrunde liegenden gewichteten Begriffshäufigkeiten dar.

Nach erfolgter Dimensionsreduktion steht dem Anwender Datenmaterial zur Verfügung, wie er es von „klassischen“ Data Mining-Anwendungen kennt. Deshalb kann nun in der mit „**Problemspezifische Mining-Analysen**“ genannten Phase das gesamte Spektrum der im Enterprise Miner zur Verfügung gestellten Verfahren (Clusterung, Regressionsanalyse, Neuronale Netze, Entscheidungsbäume, etc.) genutzt werden.

Zusammenfassend kann der SAS Text Miner als ein Werkzeug beschrieben werden, das es erlaubt, die bislang meistens vernachlässigte „Informationsquelle Text“ zu erschließen und den etablierten leistungsfähigen Data Mining-Analyseverfahren des SAS Enterprise Miner zugänglich zu machen.

## **2 Das „Expertenforum Kinderwunsch“**

Unter der Adresse [www.rund-ums-baby.de](http://www.rund-ums-baby.de) bieten mehrere sogenannte Expertenforen eine Beratung an. Die Experten arbeiten ohne Honorar und sind - je nach Thema des Forums - auf ihrem Gebiet ausgewiesen. Zum Thema „Kinderwunsch“ werden z.T. allgemeine Fragen über Schwangerschaft gestellt; genauso finden sich dort sehr spezielle Fragen über neuere Techniken der künstlichen Befruchtung, bis hin zur Übermittlung detaillierter Labordaten (z.B. eines Spermioogrammes) mit der Bitte um Beurteilung. Sämtliche Fragen an das Expertenforum wie auch die Antworten werden „ins Netz gestellt“; auch Kommentare von anderen Besuchern zu einer Frage oder der betreffenden Antwort werden dort veröffentlicht. Mittlerweile finden sich über 10.000 solcher Einträge; knapp die Hälfte davon sind Original-Anfragen.

## **3 Clusterung der Anfragen**

Ein erster Cluster-Versuch der über 4.000 Anfragen (Stand November 2000) erfolgte weitgehend mit den Vor-Einstellungen des Text Miners von SAS. Die Clusterung sollte auf Basis der Singulärwert-Zerlegung erfolgen. Um das Vokabular von Füllwörtern und weitgehend inhaltsleeren Begriffen zu reinigen, verwendeten wir die vorgegebene Stopp-Liste von SAS. Die Gewichtung der für die Clusterung verwendeten Begriffe erfolgte auf einer Log-Basis.

Abbildung 1 zeigt eine der Cluster-Lösungen mit diesem einfachen Modell. Anhand der „Deskriptoren“, die SAS zur Charakterisierung der Cluster anbietet, wird bereits deutlich, dass die Clusterung sich so gut wie kaum am fachlich-medizinischen Inhalt der Anfragen orientiert. Statt dessen erfolgt die Zuteilung der Anfragen zu der 10-Cluster-Lösung vorwiegend auf Basis der jeweils „stilistischen Eigenheiten“ der Fragensteller bzw. anhand einzelner herausragender Signalbegriffe (Mönch, Pfeffer, Präparat). Cluster 2, 9 und 10 enthalten als Deskriptoren zwar medizinische Begriffe, bleiben den-

noch aber völlig unspezifisch bezüglich des Inhalts der Anfrage. Selbst hier dominieren Brief-Floskeln.

## Clustering mit „konventioneller“ Stopp-Liste

CLUSTER	WEIGHT	FREQ	DESCRIPTION
1	0,01895649	105	andere, beantworten, bitte, einzig, experte, falsch, forum, frage, gynäkologe, gynäkologisch, info, kollege, lesen, mein, name
2	0,2009388	1113	behandlung, erfolg, fall, ivf, kind, wunsch, weit, schwangerschaft, diese, arzt, bei, geben, man, laß, sich
3	0,01299874	72	o.t, o.t., ot
4	0,11030872	611	dr., glück, hallo, halt, machen, mfg, stutterheim, v., viel, sollen, man, sein, ss, laß, kein
5	0,06391045	354	c.h., gruß c.h., mönch, pfeffer, präparat, gruß, hallo, beantworten, nehmen, frage, geben, sehen, daß, kollege, bei
6	0,10471204	580	dein, doch, forum, lieb, mal, mein, was, schwanger, dank, sagen, ganz, machen, lesen, nehmen, erst
7	0,06228561	345	antwort, bühn, dank, ehren, frage, für, herr, herzlich, mein, schnell, viel, von=das, lesen, dr., nochmals
8	0,0166095	92	
9	0,12041885	667	ats, blutung, erfolgen, gruß, sollen, zyklus, hormon, kind, sein, wunsch, schwanger, um, tag, kein, ss
10	0,25546127	1415	am, blutung, hormon, kb, lieb, tag, ss, gruß, zyklus, dann, vor, diese, schwangerschaft, um, kommen



www.allgemeinmedizin.med.uni-goettingen.de

www.sas.de



**Abb. 1:** Clustering mit Stopp-Liste

Es boten sich nun 2 Wege an: (1) eine Verfeinerung der Stopp-Liste oder das quasi gegenteilige Verfahren: die Arbeit mit einer eigenen Start-Liste. Hierbei werden nur Begriffe verwendet, die der Nutzer zulässt (und die daher im optimalen Fall dem Ziel der Auswertung entsprechen). Abbildung 2 zeigt ein Beispiel der Lösung auf Basis einer zu diesem Zeitpunkt schon mehrfach überarbeiteten Start-Liste. Einzelne Cluster (z.B. Cluster 1) werden zwar immer noch weitgehend auf Basis von Floskeln gebildet, dennoch lassen sich schon vergleichsweise gut Anfragen inhaltlich gegeneinander abgrenzen. Cluster 2 und 6 vereinen vorwiegend Anfragen zu den technisch orientierten Behandlungsmethoden (In-vitro-Fertilisation; Insemination; Cluster 2 bezieht sich dabei mehr auf Probleme von Frauen, Cluster 6 stärker auf Probleme von Männern). Auch Cluster 5 umfasst eine relativ klare Gruppe von Anfragen, die sich unter anderem auf den optimalen Zeitpunkt der Emp-

fängnis, Methoden der Temperaturmessung und Berechnung des Eisprungs und Unterstützung der Befruchtung und erster Schwangerschaftszeit durch Hormone beziehen. Aufgrund von Rechtschreibfehlern, unerwarteten Abkürzungen und einer nicht immer sofort möglichen Abgrenzung von sinnvollen und redundanten Begriffen dauerte es relativ lange, bis die Start-Liste zufriedenstellend war. Cluster 3 besteht z.B. noch aus einer unbefriedigenden Mischung von Namen („KB“ ist die Abkürzung eines Experten), Brieffloskeln und einigen Fachbegriffen.

### Clusterung mit „verfeinerter“ Start-Liste

CLUSTER	WEIGHT	FREQ	DESCRIPTION
1	0,1196967	663	antwort, antworten, beantworten, dank, experte, forum, frage, info, kollege, lesen, name, natur, nochmals, schnell, stellen
2	0,1440693	798	eileiter, gebären, ivf, klappen, mutter, schwanger, spiegelung, versuch, icsi, denken, chance, insemination, glück, problem, sagen
3	0,3596317	1992	ats, blutung, fa, früh, hormon, kb, kommen, lieb, nehmen, ss, test, progesteron, gelb, körper, wert
4	0,1141	632	ei, eizelle, kind, praxis, reifung, stock, wunsch, störung, ursache, follikel, hormon, stimulation, erhöhen, behandlung, ats
5	0,0868388	481	anstieg, eisprung, messen, stattfinden, tag, temperatur, zt, gelb, progesteron, test, spät, follikel, körper, lang, sehen
6	0,1754829	972	befruchtung, behandeln, behandlung, durchfahren, erfolg, icsi, insemination, ivf, mann, same, sperma, spermogramm, zelle, zentrum, fall



www.allgemeinmedizin.med.uni-goettingen.de

www.sas.de



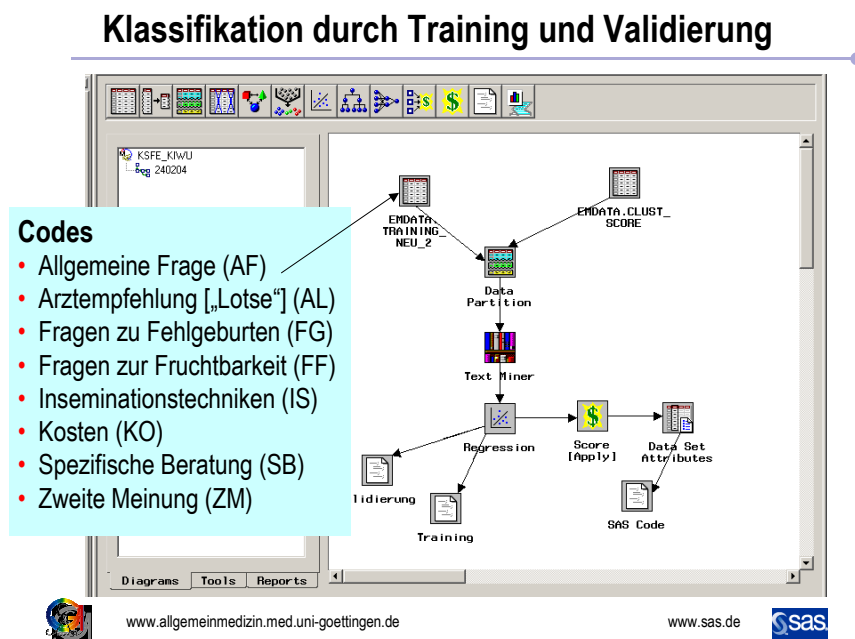
**Abb. 2:** Clusterung mit verfeinerter Start-Liste

In einer eigenen, handgestützten Klassifikation der Anfragen orientierten wir uns gleichermaßen an Inhalten und Struktur der Anfragen. Beispielsweise war es uns wichtig, Anfragen zu Behandlungskosten und -techniken (Inhaltsebene) und Anfragen zu allgemeinen Informationen, Befundauswertung oder Klinikempfehlungen (Strukturebene) zu unterscheiden. Um Clusterungen, präziser gesprochen: Klassifikationen in diese Richtung zu lenken, benutzten wir die Trainings- und Validierungs-Optionen des Text Miners.

Eine kleine Stichprobe von 200 Anfragen wurde „per Hand“ klassifiziert; die Codes sind aus Abbildung 3 (links) zu sehen. Die Stichprobe wurde dann

„partitioniert“ in einen Trainingssatz (70%) und einen Validierungssatz (30%) der Daten. Nach dem Training auf Basis unserer Codes wurden beide Datensätze vom Text Miner re-klassifiziert.

Erwartungsgemäß gelang die automatische Klassifikation der Trainingsdaten relativ gut. Abbildung 4 zeigt, wie die von uns klassifizierten Texte (Tabellenreihen) vom Text Miner



**Abb. 3:** Klassifikation durch Training und Validierung

klassifiziert wurden (Tabellenspalten). Die eingezeichnete Diagonale repräsentiert Treffer und liegt im allgemeinen zwischen 60 bis (häufig) 100%. Diese Re-Klassifizierung beruhte auf dem Verfahren der logistischen Regression.

Deutlich schlechter fiel die Re-Klassifikation der Validierungsdaten aus. Hier lag die Trefferquote nur noch bei 37 % in einzelnen Fällen wurden

unsere Klassifikationen der Anfragen (z.B. „Kosten“ oder „Zweite Meinung“) keinmal in der Re-Klassifikation erkannt (Daten nicht gezeigt).

Schließlich erlaubt das Training im Text Miner auch, gänzlich neue Daten (Anfragen) auf Basis der trainierten Ergebnisse zu klassifizieren („Scoring“). Dabei wird nicht nur eine Zuordnung einer Anfrage zu den von uns gebildeten Codes versucht, sondern für jeden Text (Anfrage) wird die Wahrscheinlichkeit angegeben, mit dem diese Anfrage einem der 8 Codes zugeordnet wird.

## Klassifikation durch Training und Validierung

### ► Trainingsdaten

The FREQ Procedure

Table of F\_code by I\_code

F\_code(Fron: code)      I\_code(Into: code)

Frequency Row Pct	AL	AL	FF	FG	IS	KD	SB	ZH	Total
AL	18 64.29	1 3.57	0 0.00	0 0.00	0 0.00	0 0.00	9 32.14	0 0.00	28
AL	1 12.50	5 62.50	0 0.00	0 0.00	0 0.00	0 0.00	2 25.00	0 0.00	8
FF	0 0.00	0 0.00	9 100.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	9
FG	0 0.00	0 0.00	0 0.00	6 100.00	0 0.00	0 0.00	0 0.00	0 0.00	6
IS	0 0.00	0 0.00	0 0.00	0 0.00	6 100.00	0 0.00	0 0.00	0 0.00	6
KD	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	6 100.00	0 0.00	0 0.00	6
SB	6 8.96	2 2.99	0 0.00	0 0.00	1 1.49	0 0.00	58 86.57	0 0.00	67
ZH	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	11 100.00	11
Total	25	8	9	6	7	6	69	11	141



Abb. 4: Klassifikation durch Training und Validierung: Trainingsdaten

Abbildungen 5 und 6 zeigen beispielhaft anhand von 6 Anfragen, in wie weit dieses Scoring gelang bzw. misslang. In den Abbildungen ist jeweils derjenige Code, der u. E. die Anfrage am besten repräsentiert mit einem \* gekennzeichnet.



## Prognose nach Training und Validierung

### ► Scoringdaten

### Code (p)

Sehr geehrter Herr Dr. Bühler, ich selbst bin bei der TKK freiwillig, mein Mann bei der Debeka privat versichert. Wie sieht hier die Kostenübernahme aus? Wäre es für mich günstiger, in die private KK zu wechseln, um im Falle eines Falles eine ...

Allgemeine Frage (.57)  
Spez. Beratung (.43)  
**Kosten (.00) \***

... Uns bleibt also nur ICSI. Wir wollen in den nächsten Tagen einen Antrag auf Kostenübernahme bei der Krankenkasse meines Mannes (private KK) stellen. Nun aber zu meiner Frage. Ich habe mal gehört, daß man bevor die Krankenkassen überhaupt irgendwelche Kosten übernehmen mindestens schon 2 Jahre "geübt haben". Ist dies korrekt??

Zweite Meinung (1.0)  
**Kosten (.00) \***

... wieder eine Blutung, die jetzt nur noch aus bräuner Schmiere besteht. Das hat mich erschreckt und ich bin zu meiner Frauenärztin gegangen. Die ließ den HCG-Wert im Blut testen und rief heute an, dass dieser minimal erhöht sei. (7), das könne aber auf eine Kreuzreaktion mit dem LH zurückzuführen sein. Sonst sei an einen sehr frühen Spontanabort zu denken. W Könnten Sie mir bitte erklären, was diese Kreuzreaktion bedeuten kann?

Allgemeine Frage (.42)  
Arzt-Empfehlung (.33)  
**Spez. Beratung (.24) \***





[www.allgemeinmedizin.med.uni-goettingen.de](http://www.allgemeinmedizin.med.uni-goettingen.de)

[www.sas.de](http://www.sas.de)



**Abb. 5:** Prognose nach Training und Validierung:  
Scoringdaten - Teil 1

Prognose nach Training und Validierung	
Scoringdaten	Code (p)
<p>... im letzten Jahr diagnostizierte mein FA (nach Monaten wo man nichts gesehen hatte aus dem US), man überwies mich in die Uniklinik zur PCo Diagnose, dort stellte man fest das sich der Verdacht auf ein PCo nicht erhärten lies (das soll mal jemand verstehen)... bin wieder zu meinem FA, er sagte mir jaja das ist das PCO, aha, auf meine Nachfrage ob ichs denn mit Mönchspfeffer versuchen sollte, sagte man mir nein das würde nicht wirken. ... Ist das wahr das man bei einem PCo keine Kindeer kriegt? Danke fürs zuhören.. Ciao Danny</p>	<p>Spez. Beratung (.61) Allgemeine Frage (.38) Arzt-Empfehlung (.01) <b>Zweite Meinung (.00) *</b></p>
<p>... Mein Mann hat beim Urologen zwei Spermogramme machen lassen im Abstand von zwei Wochen. Nach dem Ersten sagte der Urologe "unter 3 Millionen", er könne gar nicht sagen, wieviele Spermien. ... Ist es üblich, daß die Diagnose Oligo-Asthenzoospermie gestellt wird. ... Der niedergelassene Urologe rät zu einer Biopsie. ... Ich fand den Befundbericht mehr als knapp. Was meinen Sie dazu?</p>	<p><b>Zweite Meinung (1.0) *</b></p>
<p>... wir versuchen seit 1 Jahr ein Baby zu bekommen. 1. wo mißt man die Temperatur und 2. Woran merkt man das man schwanger ist? Muß man was beachten wenn man fruchtbar ist? Wann ist die beste Tageszeit um es zu versuchen?</p>	<p><b>Allgemeine Frage (.71) *</b> Spez. Beratung (.29)</p>
 <a href="http://www.allgemeinmedizin.med.uni-goettingen.de">www.allgemeinmedizin.med.uni-goettingen.de</a>	<a href="http://www.sas.de">www.sas.de</a> 

**Abb. 6:** Prognose nach Training und Validierung:  
Scoringdaten – Teil 2

Ohne im Detail hierauf einzugehen, wird deutlich, dass der Inhalt einzelner Anfragen z.T. gar nicht erkannt und fehl-klassifiziert wurde (z.B. die ersten beiden Zitate). Andere Anfragen (Zitat 3) wurden zwar nicht optimal klassifiziert, aber zumindest wäre der von uns als sinnvoll erachtete Code mit einer gewissen Wahrscheinlichkeit erkannt worden. Etwa die Hälfte der Anfragen der Scoring-Daten wurde nach unserer groben Einschätzung richtig bzw. einigermaßen befriedigend prognostiziert.

## Diskussion

Folgende Erfahrungen und Schlussfolgerungen ergeben sich aus unserer bisherigen Analyse der Anfragen an ein medizinisches Expertenforum mittels SAS Text Miner:

1. Die mitgelieferte Stopp-Liste ist möglicherweise für normale schriftliche Dokumente geeignet, keinesfalls für das Vokabular spezieller Bereiche mit typischen Besonderheiten (z.B. Anrede- und Grußformeln). Genau diese (inhaltsleeren) Floskeln und Eigentümlichkeiten bekommen dann ein falsches Gewicht.
2. Die Verwendung von Start-Listen (in Zusammenhang mit Synonyma-Listen) führt zu ersten brauchbaren Ergebnissen einer Klassifikation der vorliegenden Anfragen; die Pflege und Verfeinerung dieser Start-Liste erfordert viel Zeit, ist aber lohnenswert.
3. Die erzielten Cluster-Lösungen lassen vergleichsweise gut Fragen nach speziellen Behandlungstechniken bzw. Fragen zum optimalen Empfängniszeitpunkt und zur Unterstützung der Schwangerschaft erkennen.
4. Die Verfeinerung der Klassifikation durch Trainings- und Validierungsdaten ist bisher nur teilweise gelungen. Der Umfang der von uns als Stichprobe gewählten Daten war wahrscheinlich zu gering bzw. war die Zahl der Klassen (Codes) zu groß.
5. Die Zuordnung der Anfragen in dem Scoring-Datensatz in Form von relativen Wahrscheinlichkeiten zu den einzelnen Codes war ebenfalls nur teilweise erfolgreich, ist aber insoweit von Vorteil, als Zuordnungen mit geringen Wahrscheinlichkeiten beachtet werden können.
6. Als nächste Aufgabe soll ein mehr-dimensionaler Cluster-Versuch die Möglichkeit eröffnen, einzelne Anfragen mehreren Codes gleichzeitig (jeweils mit hoher Wahrscheinlichkeit) zuzuordnen. Hier sehen wir das Potential, komplexere Texte gleichzeitig unter inhaltlichen und strukturellen Aspekten zu klassifizieren.
7. Damit wären die Voraussetzungen deutlich besser, Anfragen (im Sinne von Informationsbedürfnissen) sowohl nach ihrer Thematik als auch nach ihrer Relevanz und Dringlichkeit zu bewerten.

## **Literatur**

- [1] Baker L, Wagner TH, Singer S, Bundor MK. Use of the internet and email for health care information. JAMA 2003; 289: 2400-2406
- [2] Berry M, Dumais S, Letsche T. Computational methods for intelligent information access, in: Proceedings of Supercomputing'95, San Diego, CA, December 1995.  
<http://citeseer.ist.psu.edu/berry95computational.html>
- [3] Eysenbach G, Dieppen TL. Patients looking for information on the internet and seeking teleadvice. Motivation, expectations, and misconceptions as expressed in emails sent to physicians. Arch Dermatol 1999; 135: 151-156
- [4] SAS Institute Inc. (2002), Getting Started with SAS Text Miner Software, Release 8.2, Cary, NC: SAS Institute Inc.
- [5] SAS Institute Inc. (2001), SAS Text Miner - Distilling Textual Data for Competitive Business Advantage, White Paper, Cary, NC: SAS Institute Inc.