

Semiautomatisierte Beantwortung von Gesundheitsfragen an ein medizinisches Internet-Expertenforum durch Text Mining

Wolfgang Himmel
Abteilung Allgemeinmedizin
Georg-August-Universität
Humboldtallee 38
37073 Göttingen
whimmel@gwdg.de

Ulrich Reincke
Competence Center Analytical
Solution
SAS Institute Heidelberg
69118 Heidelberg
ulrich.reincke@ger.sas.com

Hans Wilhelm Michelmann
Abteilung Gynäkologie und Geburtshilfe
Georg-August-Universität
Robert-Koch-Str. 40
37075 Göttingen
hwmichel@med.uni-goettingen.de

Zusammenfassung

Zunehmend häufiger schicken Menschen Fragen zu Gesundheitsproblemen an medizinische Expertenforen im Internet. Wir entwickelten ein mehrgleisiges Verfahren der Klassifizierung komplexer Gesundheitsanfragen, um eine z. T. automatisierte und kontextsensitive Beantwortung von Anfragen an ein medizinisches Expertenforum zu ermöglichen. Zunächst klassifizierten wir eine Trainings-Stichprobe von Anfragen an ein Kinderwunsch-Forum in zwei Dimensionen (32 Codes für medizinische Themen, z. B. „Schwangerschaftssymptome“; 6 Codes für Erwartungen, z. B. „emotionale Unterstützung“). Auf Basis einer umfangreichen und selbstentwickelten „Startliste“ haben wir die relative Häufigkeit jedes Wortes pro Code automatisch ausgezählt, Indikatorvariablen anhand Cramer’s V gebildet und – durch eine Singulärwertzerlegung (singular value decomposition [SVD] – die Dimensionen reduziert. Durch ein Scoring (nach Themen- und Erwartungs-Codes sowie SVD-Distanzen) wurden zu jeder Anfrage „nearest neighbors“ errechnet. Das gelang bei über 30 % der mehr als 10.000 Anfragen. In 4 von 5 Fällen war das Ergebnis inhaltlich akzeptabel. Zu diesem Erfolg dürften mehrere Faktoren beigetragen haben: (1) die expertenbasierte Startliste, (2) die Klassifikation sowohl nach Themen als auch nach Erwartungen, (3) die zusätzliche Analyse der Anfragen nach Singulärwerten zur Verfeinerung der Inhaltsanalyse und (4) eine experimentell unterschiedliche Gewichtung der einzelnen Text Mining-Strategien zur Erhöhung der Trefferquote für „nearest neighbors“.

Schlüsselwörter: text mining, natural language processing, consumer health informatics, internet, infertility, Sprachverarbeitung, automatische Email-Beantwortung, IT im Gesundheitswesen, Kinderwunsch

1 Einleitung

Zunehmend häufiger nutzen gesunde und kranke Menschen elektronische Medien, um medizinische Informationen und Beratung zu erhalten. Zum Beispiel kann man Fragen zu Gesundheitsproblemen an verschiedene Expertenforen im Internet (sog. „Ask-the-doctor services“) schicken [1,2].

Aus Gründen einer schnelleren Beantwortung der Anfragen und auch zur Entlastung der Experten wäre eine halb-automatisierte Beantwortung solcher Anfragen wünschenswert. Dies könnte im Sinne von „Frequently Asked Questions“ (FAQs) geschehen; anspruchsvoller und „kundenorientierter“ wäre es, wenn eine neue Anfrage auf ihre Ähnlichkeit zu früheren Anfragen untersucht und die Antworten der Experten auf diese ähnlichen Anfragen angezeigt würden.

Eine solche halb-automatische Beantwortung setzt eine Klassifikation der neuen und bisherigen Anfragen voraus, die vor verschiedenen Schwierigkeiten steht: (1) medizinische Anfragen sind teilweise sehr umfangreich und komplex; (2) sie sind typischerweise eine Mischung aus einem oder mehreren medizinischen Problemen bzw. Themen und einer psychologischen Erwartung. Ein Text Mining-Verfahren, das diese verschiedenen Aspekte berücksichtigt, könnte mit hoher Zuverlässigkeit eine neue Frage ähnlichen Anfragen aus der Vergangenheit zuordnen und dem Fragesteller vorab die damaligen Antworten der Experten zur Verfügung stellen.

In diesem Beitrag möchten wir ein mehrgleisiges, expertenkontrolliertes Verfahren der Klassifizierung komplexer Texte vorstellen. Auf dieser Grundlage sollte geprüft werden, ob eine z. T. automatisierte und kontextsensitive Beantwortung von Anfragen an ein medizinisches Expertenforum möglich ist. Außerdem sollte gezeigt werden, dass man mittlerweile mit SAS Seiten sehr individuell formatieren (gestalten) kann. Ziel war es, das „Antwort-System“ in Form einer nutzerfreundlichen Navigation – mit klarem „Internet-Layout“ – zu gestalten.

2 Methoden

Eine umfangreiche Stichprobe von Anfragen an ein Kinderwunsch-Expertenforum wurde per Hand klassifiziert (Trainings-Stichprobe). Sämtliche bisherige Anfragen an das Forum wurden dann automatisch klassifiziert und zu jeder Anfrage - soweit möglich - mehrere „nearest neighbors“ gesucht.

2.1 Das Expertenforum Kinderwunsch

Unter der Adresse www.rund-ums-baby.de bieten mehrere sog. 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. B. allgemeine Fragen über Schwangerschaft gestellt; genauso finden sich dort sehr spezielle Fragen über neuere Techniken der künstlichen Befruchtung bis hin zur Beurteilung detaillierter Labordaten (z. B. eines Spermioogrammes). Sämtliche Fragen an das Expertenforum als auch die Antworten werden „ins Netz gestellt“. Kommentare von anderen Besuchern zu

einer Frage oder der betreffenden Antwort werden dort ebenfalls veröffentlicht. Mittlerweile finden sich über 10.000 Original-Anfragen unter dieser Adresse.

2.2 Klassifikation

Eine Trainings-Stichprobe von $n=987$ aus den gesamten Anfragen wurde in zwei Dimensionen klassifiziert; die erste Dimension umfasste 32 Codes für medizinische Themen (z. B. „Schwangerschaftssymptome“); die zweite Dimension umfasste 6 Codes für Erwartungen (z. B. „emotionale Unterstützung“; oder „zweite Meinung“). Während die meisten Anfragen relativ klar in eine der 32 Medizin-Codes fielen, ließen sich viele Anfragen mehreren Erwartungs-Codes zuordnen (z. B. der gleichzeitige Wunsch nach einer Beurteilung von Labordaten und der Frage, ob man die Arztpraxis wechseln solle). Tabelle 1 (nächste Seite) zeigt die 38 Klassifikationscodes.

2.3 Automatische Klassifikation

Auf Basis der Trainings-Stichprobe erstellten wir eine umfangreiche „Startliste“. Sie enthielt alle Wörter, die wir für die Klassifikation nach Themen bzw. Erwartungen für inhaltlich bedeutsam hielten. Gleichzeitig entwickelten wir diese Startliste zu einer Synonyma- bzw. Stammform-Liste. Alle Wörter wurden dabei entweder auf ihre Stammformen zurückgeführt (Hormone = Hormon etc.), ähnliche Wörter auf einen gemeinsamen Begriff (z. B. ICSI oder In-vitro-Fertilisation = IVF). Die Startliste umfasste über 10.000 verschiedene Wörter (terms), die dann zu ca. 4.000 Stammformen bzw. Synonyma (parents) reduziert wurden. Redundante oder nichts sagende Wörter (z. B. „Hallo Doktor“) wurden auf diese Weise zugleich entfernt.

Für jeden Term der Startliste wurde die relative Häufigkeit pro Code ausgezählt und als statistisches Maß Cramer's V berechnet. Für jeden Code wurden dann als Indikatorvariablen Cramer's Vs summiert. Durch eine Hauptkomponentenanalyse und – als paralleles Vorgehen – durch eine Singulärwertzerlegung (singular value decomposition [SVD] gem. den Standard-Einstellungen des SAS TextMiners) sollten die Dimensionen reduziert werden [3-5].

2.4 „Nearest Neighbors“

Ein Scoring errechnete dann für jede Anfrage die Wahrscheinlichkeit, dass sie in eine der 38 Klassifikationen fällt. Bei einem Cut off von $>70\%$ (für die Themen-Codes) bzw. $>80\%$ (für die Erwartungs-Codes) wurden für jede Anfrage diejenigen anderen Anfragen gesucht, die ebenfalls mit über 70% bzw. über 80% Wahrscheinlichkeit in denselben Themen- bzw. Erwartungscode fielen. Die SVDs wurden als zusätzliches Distanzmaß herangezogen. Anfragen, die in dieselbe Codierung fielen und auf Basis der SVDs nicht weiter als 0,2 Einheiten entfernt lagen, galten als „nearest neighbors“.

Technisch sollte das Verfahren so ausgelegt sein, dass die jeweils 3 „nearest neighbors“ bei einer (simulierten) neuen Anfrage dem Fragesteller angezeigt werden, gleichzeitig mit den damals von den Experten gegebenen Antworten.

Tabelle 1: Klassifikationsgüte für die Trainings-Stichprobe

Code-Klasse	Anfragen [n]	Precision [%]	Recall [%]
Medizinische Codes			
(1) Abort	40	83	88
(2) Ausschabung	13	91	77
(3) Clomifen	26	86	96
(4) Eileiter	16	94	100
(5) Eileiterprüfung	19	76	100
(6) Eisprung	81	97	91
(7) Endometriose	11	73	73
(8) Ernährung	9	82	100
(9) Gelbkörperschwäche	25	80	80
(10) Geschlechtsverkehr	14	50	86
(11) Habituelles Abort	17	85	100
(12) Hormone	36	97	86
(13) Insemination	29	96	90
(14) Ivf	102	97	97
(15) Kosten	25	89	96
(16) Kryotransfer	13	88	54
(17) Medikamente	47	100	98
(18) Mehrlinge	7	86	86
(19) Menstruation	35	91	91
(20) Naturheilkunde	33	91	88
(21) PCO	27	86	93
(22) Pille	23	87	87
(23) Probleme Geschlechtsverkehr	5	100	100
(24) Schilddrüse	13	68	100
(25) Schwangerschaftssorgen	49	94	92
(26) Schwangerschaftssymptome	36	89	92
(27) Schwangerschaftstest	30	82	90
(28) Spermogramm	57	92	95
(29) Stimulation	40	87	85
(30) Zwischenblutung	14	100	93
(31) Zyklus	79	94	95
(32) Zysten	16	57	100
Summe:	987	89	92
Erwartungs-Codes*			
(1) Allgemeine Fragen	533	95	94
(2) Derzeitige Behandlung	331	90	87
(3) Ergebnisse	310	92	89
(4) Gefühle	90	97	93
(5) Interpretation	242	91	85
(6) Möglichkeiten	351	88	83
Summe		92	89

*Mehrfachklassifikationen möglich

3 Ergebnisse

Die Ergebnisse der automatischen Klassifikation für die Trainings-Stichprobe zeigt Tabelle 1. Für die Erwartungs-Codes lagen Recall (Sensitivität) und Precision (positiver prädiktiver Wert) im Durchschnitt bei 89 % und 92 %; für die Themen-Codes ebenfalls um 90 %. Bei mehreren Codes lagen Precision oder Recall sogar bei 100%.

Diese Klassifikationsergebnisse betrachteten wir als ausreichend. Sämtliche Anfragen ($n > 10.000$) des Forums wurden nun nach Thema und Erwartung - auf Basis der 38 Codes - klassifiziert. Soweit die Wahrscheinlichkeit einer Anfrage, zu einem Themen- bzw. Erwartungs-Code zu gehören, bei über 70 % bzw. 80 % lag, wurde sie dieser Anfrage zugeordnet (mit der Chance einer Mehrfach-Klassifikation).

Gemäß unserer Kriterien zur Auswahl von „nearest neighbors“ (vgl. oben 2.4) wurden neben der χ^2 -Statistik zur Klassifikation auch Singulärwertdimensionen mit dem SAS Text Miner extrahiert und für die jeweils ersten 10 SVDs einer Anfrage die Entfernung zu den 10 ersten SVDs der anderen Anfragen berechnet. Die Gewichte für die drei Informationen pro Anfrage wurden so gesetzt, dass die thematische Dimension (t) mit dem Faktor 1, die Erwartungsdimension (e) mit dem Faktor 0,5 und die Distanz der SVDs (soweit sie $>0,2$ war) mit dem Faktor 5 in die Berechnung der „nearest neighbors“ eingingen:

```
Score = match_t + .5 * match_e - 5 * dist_svd;
```

Bereits aufgrund der oben genannten Kriterien für die Klassifikation und Extraktion von SVDs fielen über 6.000 der 10.000 Dokumente aus der weiteren Kalkulation, sodass für diese Dokumente a priori keine Matches gefunden werden konnten.

Abbildung 1 und Abbildung 2 zeigen zwei Beispiele für eine zufällig gewählte Anfrage (= simulierte „Neu“-Anfrage) und deren drei „nearest neighbors“. Jeweils für alle Texte sind die Wahrscheinlichkeiten für Themen- und Erwartungs-Codes angegeben. Das erste Beispiel wäre eine optimale Realisation. Die Anfrage und die drei Matches sind richtig klassifiziert und - offensichtlich aufgrund der SVD-Distanz - liegen die konkreten Fragen inhaltlich relativ eng beieinander. Damit würden für die neue Fragestellerin, die offensichtlich schon länger auf Antwort wartete, die Antworten auf die drei früheren Anfragen mit Sicherheit hilfreich sein.

Weniger bzw. nicht gelungen ist die Lösung im 2. Beispiel. Hier wurde die Original-Anfrage nicht ganz korrekt klassifiziert. Es wird bei diesem Beispiel zwar von „Eisprung“ gesprochen (und auch so automatisch klassifiziert), aber es geht eigentlich um „Schwangerschaftssymptome“. Alle drei Matches behandeln deshalb – logischerweise und fälschlicherweise – ein Problem mit dem Eisprung und treffen nicht das inhaltliche Problem der 1. Fragestellerin.

Wir haben bisher nur in einer kleinen Stichprobe von 50 „Neu“-Anfragen die ersten drei „nearest neighbors“ berechnet, selektiert und auf ihre „Passung“ untersucht. Gemäß unserer Zielsetzung, dass mindestens ein „nearest neighbor“ zur „Neu“-Anfrage passen soll, galt dies für 78 % der insgesamt 50 Anfragen. Bei immerhin 29 % passten alle drei Matches gut zur „neuen“ Anfrage. In 22 % allerdings erfüllte keiner der Matches das Kriterium (und wäre somit unbefriedigend für die Fragestellerin gewesen).

„Neue“ Anfrage als simuliertes Beispiel

... Ich habe bereits eine Tochter, die ich durch eine Hormonbehandlung (...) bekommen habe. Mein Zyklus ist sehr unregelmäßig und seit 2 Monaten stelle ich fest, dass ich noch Milch habe, obwohl ich schon seit 10 Monate abgestillt habe. Hatte auch schon Abtilltabletten bekommen. Wäre es sinnvoll wieder mit meinem Kinderwunsch in eine KiWu-Praxis zu gehen oder sollte ich noch abwarten? Ich wäre sehr froh, wenn sie mir antworten könnten, denn ich hatte schon 2 x auf Antwort gehofft...

Automatisches Scoring: Zyklus: 89 %, Möglichkeiten: 98 %, Allgemeine Fragen: 74 %

... Mein Zyklus ist nach wie vor unregelmäßig und schwankt zwischen 35 und wie dieses mal 22 Tagen. Ich versuche es seit 2 Monaten mit Persona weiss aber nicht genau ob das in meinem Fall überhaupt was bringt denn es wurde auch noch kein Eisprung angezeigt. ... wüßte ich gerne Näheres betreffend Möglichkeiten meines Frauenarztes und wann wir uns an diesen wenden können.

Scoring: Zyklus: 100 %, Allgemeine Fragen: 99 %, Möglichkeiten: 90 %

Bei unserem ersten Kind war ich nach absetzen der Pille innerhalb von 3 Monaten schwanger! ...Nun kommt noch hinzu, das seit einiger Zeit meine Regel etwas unregelmäßiger kommt, als ich es gewohnt bin. Normalerweise kam diese stets auf den Tag genau, selbst nach absetzen der Pille letzten Jahres. Mein Frauenarzt sagt immer nur, ich sei doch gesund, forscht aber nicht nach den möglichen Ursachen .. Was können wir tun oder sollen wir tun...

Scoring: Interpretation von Ergebnissen: 99 %, Möglichkeiten: 93 %

... habe vor 4 Monaten abgestillt und hatte zuvor genau 6 Monate lang gestillt. Nun kam meine Regel auch wieder, aber seit dem habe ich zwei Zyklen gehabt mit je 45 Tagen! Nun bin ich schon wieder dem 30 Tag nah und würde mich eigentlich freuen, wenn ich einen ES wieder "normal" bekommen würde.

Scoring: Zyklus : 98 %, Möglichkeiten: 81 %

Abbildung 1: Beispiele für gelungene Matches

„Neue“ Anfrage als simuliertes Beispiel

Hallo Ihr Lieben, kann mir jemand sagen, ab wann man sicher sein kann, dass es geklappt hat? Ich habe die Morgentemperatur gemessen und die war bis zum Eisprung konstant 36,8°C. Dann auf 37,0°C erhöht und seit dem (5 Tage) konstant hoch. Kann ich mir zum zweiten Kind gratulieren?

Automatisches Scoring: Eisprung: 82 %, Allgemeine Fragen: 97 %

... wir probieren schon seit längerer Zeit schwanger zu werden. Jetzt muß ich jeden Tag die Temperatur messen... ein heilloses Durcheinander. Teilweise liegen sie bei weit unter 36°C und sind gar nicht mehr eintragbar in dem Vordruck. Woran kann das liegen. Vielleicht streß ich mich auch zu sehr in das Ganze ein...

Scoring: Eisprung: 100 %, Allgemeine Fragen: 98 %, Möglichkeiten: 84 %

... Wieso passiert es, daß manche Frauen keinen Eisprung haben? Wie kommt so etwas (nicht) zustande?

Scoring: Eisprung: 91 %, Allgemeine Fragen: 94 %

... Laut Persona hatte ich am 15. oder 16. Tag meinen Eisprung. Aber meine Temperatur ist nicht nennenswert hoch gegangen und heute auch schon wieder runter. Woran merke ich denn, ob ich gar keinen Eisprung hatte? Bekomme ich meine Tage auch, wenn ich keinen Eisprung hatte? Und wenn ja, woran könnte es liegen, dass ich keinen Eisprung habe? Was kann ich dagegen tun?

Scoring: Eisprung: 100 %, Allgemeine Fragen: 98 %

Abbildung 2: Beispiele für nicht-gelungenes Matching

Abschließend möchten wir zeigen, dass man – ohne die SAS-Umgebung verlassen zu müssen – die Fragen an das Internet-Expertenforum in freier Gestaltung darstellen kann und zugleich die als „nearest neighbors“ gefundenen früheren Anfragen (und deren Antworten) durch einfaches Navigieren („Anklicken“) aufrufen kann. Abbildung 3 zeigt ein solches Layout. Im oberen Teil ist eine Anfrage mit ihrer Klassifikation (links und rechts oben) sowie die Antwort des Internet-Arzt zu sehen. Darunter die ähnlichen Fragen, die man durch Mausklick (einschließlich ihrer Antworten) aufrufen könnte. Abbildungen 4 und 5 zeigen ausschnittsweise, wie man solche Berichte in SAS erstellt, u. a. mit der Prozedur „proc template“ (Abbildung 4). Außerdem gestattet SAS mittlerweile sog. inline-Formatierungen, sodass im laufenden Text Schriftwechsel etc. möglich sind (Abbildung 5). Bis zu welchem Niveau das (relativ einfach) möglich ist, zeigt das Beispiel von Morgan [6].

Eisprung:96	Derzeitige:88
< Frage >	
<p>Sehr geehrte Frau Dr. med. Czeromin, ich hätte gerne ihr Meinung dazu. Meine Tochter übten wir 14 Monate, davon 3 mit Clomifen bis es klappte. Ich hatte damals aber immer einen ES und sehr regelmäßigen Zyklus von 31 Tagen. Mein Sohn war ein Unfall, einmal nicht aufgepasst, sofort ss. Nun üben wir seit 5 Monaten für das Dritte. Als ich die Pille absetzte hatte ich meine Tage, danach einen Zyklus von 72 Tagen. Danach hatte ich einen Zyklus ohne ES, laut LH-Test und Zervixschleim und MUMU beobachtung und wieder einen 75 Tage Zyklus wobei er dann die Mens eingeleitet wurde und ich bekam Clomifen. Hormonwerte waren okay. Nach der Mens nahm ich zw. 5-9 Tag Clomi und am 10ZT hatte ich einen US wo rechts ein Folikel von 18mm und li ein kleiner gesehen wurde. Der FA meinte am 12ZT müsste mein ES erfolgen. Ich machte nun bis zum 20ZT die Lh-tests alle negativ und auch laut Temp. messen kein ES. Hatte am 15ZT einen spinnbaren, glasigen ZS und Mumu weit oben aber zu. Hier meine Temp-Kurve: http://kurve.kleinersonnenschein.de/kurve_show.php?id_ist_gleich_28571&data_ist_gleich_1133422288 Nun zu den Fragen: Was wäre jetzt empfehlenswert, da ja ein Folikel herangereift ist, aber nicht gesprungen? Was könnte die Ursache dafür sein? Muss meine Mens wieder ausgelöst werden, da wohl ein ES erfolgt ist? Vielen Dank für Ihre Hilfe. LG Johanna (5. Dezember 2005, 0:05 Uhr)</p>	
Antwort von Dr. Czeromin	
<p>Liebe Johanna, leider kann ich Ihre BTK nicht öffnen. Genauer als die BTK wäre aber eine Hormonuntersuchung bei Ihrem Frauenarzt ca. acht tage nach angenommenem Eisprung. Ist Progesteron nachweisbar, war ein Eisprung und die Blutung kommt von alleine, wenn Sie nicht schwanger sind. Ist kein Progesteron nachweisbar, ist dieser Zyklus ohne Eisprung geblieben und die Blutung könnte mit Med. ausgelöst werden. Sinnvollerweise kann im Clomifenfolgezyklus bei entsprechender Follikelreifung der Eisprung mit einer Eisprungaushlösenden Spritze ausgelöst (5000 IE HCG) werden. Ihnen alles Gute! Ute Czeromin</p>	
Ähnliche Fragen	
<p>Eisprung:92 , Derzeitige Behandlung:92 , Dr. Tandler-Schneider Schwangerschaftssymptome:79 Eisprung:78 , Allgemeine Fragen:98 Derzeitige Behandlung:94 , Dr. Böhm Eisprung:100 Medikamente:76 , Ergebnisse:99 Derzeitige Behandlung:81 Inhalte:65 , Dr. v. Stutterheim Eisprung:93 , Ergebnisse:95 Derzeitige Behandlung:83 , Dr. Bühler Eisprung:100 , Derzeitige Behandlung:85 Allgemeine Fragen:77 Inhalte:53 , Dr. Happel Hormone:99 Eisprung:82 , Derzeitige Behandlung:99 Ergebnisse:94 Inhalte:59 , Dr. Böhm</p>	

Abbildung 3: Frei formatierte Berichte in SAS

```
PROC TEMPLATE;  
DEFINE STYLE pdfbiblio; PARENT=styles.printer; Modifikation des „printer“  
Style Templates  
REPLACE BODY FROM DOCUMENT /  
LEFTMARGIN=.75in      RIGHTMARGIN=.75in  
TOPMARGIN=.75in       BOTTOMMARGIN=.75in;  
STYLE SYSTEMTITLE / FOREGROUND=black FONT_FACE=arial FONT_SIZE=4;  
STYLE SYSTEMFOOTER / FOREGROUND=black FONT_FACE=arial FONT_SIZE=2;  
/* Trick: Unterdrücken der Tabellenränder weisse Farbe  
auf weissem Hintergrund*/  
STYLE TABLE / CELLPADDING=1.5      CELLSPACING=1.5  
                BORDERCOLOR=white   BORDERWIDTH=0;  
STYLE HEADER / FOREGROUND=black      BORDERCOLOR=white  
                FONT_FACE=arial      FONT_SIZE=3.75;  
STYLE PAGENO / FONT_FACE=arial      FONT_SIZE=2;  
END; Unterdrücken der Tabellenränder  
RUN;  
OPTIONS CENTER NODATE; Output zentrieren und Seitenzahlen unterdrücken
```

Abbildung 4: Erstellung von Reports durch „proc template“

```
data test.kopf(keep=I S ); set test.temp;  
length i s $150.;;  
i="  ||cats( Inhalt_n_1)||" style='text-decoration:none;'>  
  ||cats( Inhalt_l_1, ':',put( Inhalt_p_1,3.), "</a></span>");  
  
Logischer Link  
(nicht sichtbar)  
für URL → Eisprung:96 ←  
  
Neighbor="

<a href='http://localhost/kiwu/'||cats(Nr_)||".html '  
  style='text-decoration: none;'" ||cats(Inhalt_)||" </a>,  
  <i> "||cats(Struktur )||" </i>,  
  <b> "||cats(expert)||" </b> </p>";  
Inline Formatierung Kursiv  
Inline Formatierung Fett

Zyklus:98 , Möglichkeiten:81 , Dr. Bühler
```

Abbildung 5: Erstellung von Reports mit html-Links und inline-Formatierungen

4 Diskussion

Eine automatische Klassifikation auch komplexer und mehrdimensionaler Antworten ist möglich, sodass ein individuell angepasstes Text Mining-Verfahren den Arbeitsaufwand z. B. von medizinischen Experten reduzieren kann. Zur Lösung dieser Aufgabe haben vermutlich mehrere Faktoren beigetragen: (1) die umfangreiche Start- bzw. Synonyma-Liste, (2) die Klassifikation sowohl nach Themen als auch nach Erwartungen, sodass eine Klassifikation der Anfragen möglich wurde, (3) die Analyse der Anfragen nach Singulärwerten, sodass innerhalb eines Themas bzw. einer Erwartung sich noch feinere Details einer Anfrage erschlossen und schließlich (4) die Möglichkeit, durch Gewichtung der einzelnen Text Mining-Strategien die Trefferquote für „nearest neighbors“ zu erhöhen. Einschränkungen und Grenzen unseres Ansatzes sollen im Folgenden genannt werden: Lediglich einen Teil der Anfragen haben wir überhaupt nur dann für eine halb-automatische Beantwortung zugelassen, wenn das Scoring eine hohe Sicherheit der Klassifikation anzeigte bzw. der Abstand der Singulärwerte nicht zu groß war. Das würde im praktischen Internet-Betrieb für dieses Beispiel bedeuten, dass über die Hälfte der Anfragen keine automatisch identifizierbaren „Nachbarn“ hat. Diese Fragesteller müssten mit der ausschließlich individuellen Beantwortung durch Experten vorlieb nehmen (und ggf. weiterhin länger auf Antwort warten).

Wir sind uns bewusst, dass beim derzeitigen Verfahren einige der „Nachbarn“ nicht unbedingt dem Inhalt der Neu-Anfrage entsprechen. Allerdings zeigt sich in der Arbeit des Expertenforums, dass häufig eine Erst-Anfrage weitere Anfragen nach sich zieht und diese dann tatsächlich sehr häufig im Bereich der Nachbarn liegen. Dann wären die Matches gewissermaßen vorweggenommene Antworten auf eine noch nicht gestellte Frage.

Einzelne Inhalts-Codes erwiesen sich als problematisch, insbesondere der Code „Medikamente“. Für die Trainings-Stichprobe waren zwar Precision und Recall des Codes „Medikamente“ sehr gut; bei den neu klassifizierten Anfragen dagegen fiel die Qualität deutlich ab. Hier ist das Auffinden signifikanter Wörter für die Codierung selbst bei einer großen Trainings-Stichprobe von ca. 1.000 Anfragen nicht unbedingt gewährleistet. Zum Beispiel wurden neue Medikamente in den weiteren Anfragen durch die inhaltliche Klassifikation und durch die SVDs nicht zufriedenstellend erkannt, so dass auch das „Matching“ von „nearest neighbors“ nicht ausreichend gelang. Hier dürfte die thematische Kategorie „Medikamente“ nicht günstig gewählt worden sein, da sie letztlich keinen „inhaltlichen Mittelpunkt“ hat. Insoweit verweist das Text Mining-Verfahren auf einen logischen Fehler *unserer* Klassifikation.

Die guten Klassifikationsergebnisse für Precision und Recall haben damit zu tun, dass wir uns zunächst auf eine Trainings-Stichprobe beschränkt haben und die Klassifikation noch nicht an einer neuen Stichprobe überprüften.

Aus diesen Beschränkungen ergibt sich Optimierungsbedarf für eine halb-automatische Beantwortung medizinischer Anfragen an ein Expertenforum. Schon jetzt aber konnten wir in einem Ausblick zeigen, wie sich „Handling“ und „Präsentation“ einer semi-automatischen Beantwortung von medizinischen Anfragen gestalten lässt und - daran anschließend - auch ein laufend aktualisierter Katalog für FAQs aufgebaut werden kann.

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
Available at: <http://jama.ama-assn.org/cgi/reprint/289/18/2400.pdf>
- [2] Himmel W, Meyer J, Kochen MM, Michelmann HW. Information needs and visitors' experience of an Internet expert forum on infertility. J Med Internet Res 7 (2005) e20
Available at: <http://www.jmir.org/2005/2/e20/>
- [3] Himmel W, Reincke U, Michelmann HW. Combining text mining strategies to classify requests about involuntary childlessness to an internet medical expert forum. ECML/PKDD 2006 Workshop on Practical Data Mining: Applications, Experiences and Challenges. Berlin 2006.
Available at: <http://www.ecmlpkdd2006.org/ws-pdmaec.pdf>
- [4] Jeske D, Liu R. Mining massive text data and developing tracing statistics. In: Banks D, House L, McMorris FR, Arabie P, Gaul W, editors. Classification, clustering and data mining applications. Berlin: Springer; 2004: 495-510
- [5] Reincke U. Profiling and classification of scientific documents with SAS Text Miner.
Available at <http://km.aifb.uni-karlsruhe.de/ws/LLWA/akkd/8.pdf>
- [6] Morgan D. Inline Formatting + Long Character Variables = "I Didn't Know You Could Do That with SAS®"! SUGI 31 Paper 051-31
Available at: <http://www2.sas.com/proceedings/sugi31/051-31.pdf>