Paper 210-2008

Using Text Mining to Classify Lay Requests to a Medical Expert Forum and to Prepare Semiautomatic Answers

Wolfgang Himmel, Department of General Practice/Family Medicine, University of Göttingen, Germany
Ulrich Reincke, Competence Center Enterprise Intelligence, SAS Institute, Heidelberg, Germany
Hans Wilhelm Michelmann, Department of Obstetrics and Gynecology, University of Göttingen, Germany

ABSTRACT

We developed a scoring procedure to automatically classify lay requests to an internet medical forum about involuntary childlessness. The requests should be classified according to their subject matter (32 categories) and the sender's expectation (6 categories). Building upon this procedure, the experts' answers to former requests could be the basis of an automatic answer for new incoming requests, finding their "nearest neighbors". Our text mining approach comprised the following steps: a large start list of relevant words and the calculation of the Cramer's V statistic for the association between relevant words and the 38 categories. We trained logistic regression models with high precision and recall. We then simulated a scenario in which a subset of requests (n=50) served as 'new' requests. To find the most nearest neighbors, we applied a formula, which gave high weight to singular value decompositions (SVDs) but also considered the automatically classified subject matter of this 'new' request and—to a lesser degree—the sender's expectation. If we can implement this procedure into the real life of a health internet forum, the workload of medical experts could be lightened, visitors to a forum could receive a more timely answer to their question and they could become aware of former questions and answers that were rather similar to the one they had sent.

INTRODUCTION

Both healthy and sick people increasingly use electronic media to obtain medical information and advice (Umefjord 2007). One important application is web-based "ask the doctor" services. People may send requests to these services to better understand their disease, to learn about new therapies or to ask for a second opinion (Widman 1997, Eysenbach 1999).

To facilitate the work of medical experts in web forums, it would be helpful to classify visitors' requests automatically. The classification of medical requests - or any other documents - can refer to (1) their subject matter, e.g. the drug treatment of high blood pressure or the pros and cons of screening for colon cancer and (2) the sender's expectation, e.g. to receive a commentary on the current treatment (second opinion), to get general information about a certain disease or biological processes, or where to seek adequate medical help.

Text mining has been successfully applied for automatic classifications of large volumes of documents (Weiss et al.), for example in classifying consumer complaints. The automatic classification of medical requests is more difficult. Since such requests come from lay people, they are often very long and unstructured; at the same time, many of these requests ask for the same information although they tell a different story. Moreover, lay people often mix personal experiences with laboratory data or medical issues with the need for psychological help.

We have finished an initial trial to automatically classify these requests using standard text-mining software and are developing further strategies to refine this process (Himmel et al. 2007). We use a large sample of requests in the section "Wish for a Child" on the German website www.rund-ums-baby.de, which provides information for parents and potential parents.

In this preparatory work, we developed a scoring procedure that calculated for any medical request the probability to belong to a certain subject matter (e.g. hormones or insemination) and to belong to a respective expectation of the sender (e.g. general information or interpretation of laboratory data). We will first describe this procedure and then present a method how to find the "nearest neighbors" for the requests studied. If this procedure leads to satisfying results, it will be possible to classify new incoming requests to the expert forum and to look for the nearest neighbors of this request in the large stock of the expert forum. These nearest neighbors and the experts' former answers to these requests could be the basis of an automatic answer.

METHODS

The analysis is based on a random sample of requests (n=988) from the German website www.rund-ums-baby.de. According to the above-mentioned 2 dimensions (subject matter and expectations), these requests were first manually classified to provide a basis for training and validation. The categories for the subject matter dimension were mutually exclusive; the categories for the expectation dimension permitted overlapping (see Box 1 for the 38 categories).

Dimensions	Validation Data	
Category	**Precision** (%) *	**Recall** (%) *
Medical category		
Abortion	91	100
Abrasion	100	100
Clomifen	100	100
Oviduct	100	100
Examination of the oviduct	100	100
Ovulation	90	86
Endometriosis	75	100
Nourishment	100	100
Luteal phase defects	88	100
Sexual intercourse	100	100
Habitual abortion	100	100
Hormones	78	78
Insemination	100	100
IVF	81	88
Cost	100	100
Cryo transfer	100	75
Medical drugs	92	100
Multiples	100	100
Menstruation	90	100
Naturopathy	90	100

Dimensions	Validation Data	
Category	**Precision** (%) *	**Recall** (%) *
Medical category		
PCO	100	100
Birth control pill	100	100
Problems Sexual intercourse	100	100
Thyroid gland	100	100
Pregnancy worries	100	92
Pregnancy symptoms	100	100
Pregnancy test	88	88
Semen analysis	88	93
Stimulation	63	100
Intermenstrual bleeding	100	100
Cycle	80	86
Cysts	100	100
Expectations**		
General information	92	84
Actual treatment	85	72
Results	86	82
Feelings	100	61
Interpretation	78	69
Possibilities	82	81

* Model (selection criterion): To calculate recall and precision, we first chose the best model according to the following selection criteria: *Akaike's Information Criterion*, *Schwarz Baysian Criterion*, *cross validation misclassification of the training data*, *cross validation error of the training data*.

** multiple categories possible

Box 1: Goodness of automatic classification

A TEXT MINING APPROACH TO CLASSIFY THE DOCUMENTS

To automatically classify the 988 requests sent to the expert forum, we developed an individual text mining approach comprising the following steps: (1) A large start list of possibly relevant words (2) A word count of all words from the start list appearing in all requests of each category. (3) Calculation of the average Cramer's V statistic for the significant word associations in each category, principle component analysis, and singular value decomposition (SVD) with the SAS Text Miner (Albright 2006, Reincke 2003). (4) On the basis for these 3 types of input variables, we trained logistic regression models and built a meta-model for the automatic classification. (5) Calculation of recall and precision for a subset of validation data.

SCORING OF DOCUMENTS FOR A SEMI-AUTOMATIC ANSWERING

A scoring procedure should calculate for all 988 requests the probability to fall in any of the 38 classifications of the subject matter and expectation dimension. We then determined “nearest neighbors” to these requests, according to the 2 dimensions, and additionally considered the SVD as a distance measure. We simulated a scenario in which a subset of requests (n=50) served as ‘new’ requests. For each of these requests, 3 similar former requests should be found as “nearest neighbors”, if possible. The experts’ former answers to these requests should be displayed as preliminary answers to the ‘new’ request.

As a crude measure of the quality of our procedure, we calculated how many of the 3 displayed requests and their respective expert answers corresponded to the ‘new’ request.

RESULTS

TEXT MINING APPROACH

The start list of words for the sample of 988 requests consisted of about 11,000 terms, resulting in a total of about 4,100 synonyms, which we called ‘parents’ (Table 1 in Box 2). Terms and parents were treated as binary variables (0/1) and we could determine whether a term or parent appeared in a document or not (Table 2 in Box 2). This large table was then transposed in a transaction table (Table 3 in Box 2), showing whether a document fell into a certain category (Box 2 shows an example for category 37) or not. This transaction table also shows whether a certain parent (Box 2, parent 4) was present in this document or not.

Table 1 Synonym List

Term	Parent
abartig	abartig
abbauen	abbauen
abbaut	abbauen
abgebaut	abbauen
abblutet	abbluten
abbort	abort

Table 2 Binary Indicator Variables Based on Parents

Document ID	ABARTIG(p1) abartig(t1)	ABBAUEN(p2) abbauen(t2)	ABBAUEN(p2) abbaut(t3)	ABBAUEN(p2) abgebaut(t4)
330	0	0	1	0
333	0	0	0	0
336	0	0	0	0
338	0	0	0	0
353	0	0	0	0
355	0	0	0	1

Table 3 Transposing Table 2 into a Transaction Table

target_id	target	Document ID	parent_id	parent
37	1	299	4	0
37	1	296	4	1
37	1	293	4	0
37	1	286	4	0
37	1	290	4	0
37	1	289	4	0
37	0	292	4	0

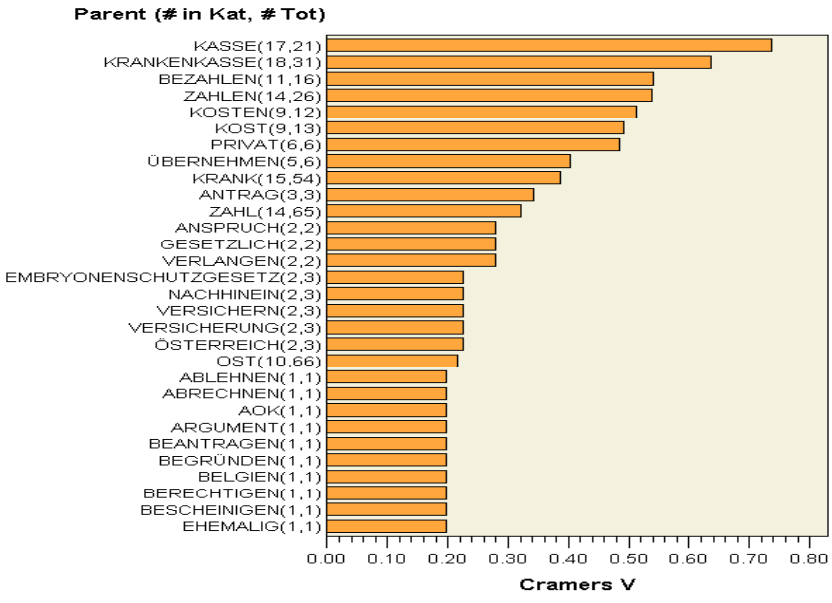
Box 2: Terms, Parents and Categories Appearing in Each Document in Binary Presentation

We calculated the average Cramer’s V statistic for each parent with each of the 38 categories. Significance levels for including a parent’s Cramer’s V statistic was alternatively set at 1%, 2%, 5%, 10%, 20%, 30% and 40%. We summed for each category all Cramer’s V coefficients over the significant words. About 4 million chi^2 tests were necessary for this task and were performed automatically with one call of the SAS procedure *proc freq*.

```
proc freq noprint data=trans ;
tables parent*target
    /ChiSq measures outpct
    expected nowarn out=freq_count ;
output out=ChiSq pchi measures cramv ;
by parent_id target_id;
run;
```

Box 3 shows one example for the association of the most significant parents with the category “treatment costs”. This category comprised requests about costs for insemination and microinjection techniques (ICSI). The corresponding SAS statement is as follows:

```
proc iml ;
use trn_p;read all into T
use sco_p;read all into S
use tassoc;read all into A
C=T*A
create KV trn form C;
quit;run;
```



Box 3: Target Association

For further data processing, we used 2 more techniques for dimension reduction: (1) principle component analysis (PCA) to reduce the seven indicator variables of varying significance levels per category into five orthogonal dimensions and (2) SVD on the basis of the standard settings of the SAS TM Software. These steps resulted in three sorts of input variables:

1. 38 * 7 Cramer’s V indicators
2. 38 * 5 PCAs per code
3. ≈ 500 SVDs.

By means of logical combinations of input variables with different selection criteria, a huge number (< 1,700) of models were trained. For the different categories, different models with different input variables proved best. To give just two examples:

(1) As can be seen for the categories “oviduct” and “examination of the oviduct” in Box 4, some of the predictive words are similar in both categories so that—in this case—only the summary of the Cramer’s statistic as an indicator variable or—in other cases—PCA or SVDs guaranteed high precision and recall.

Word	Frequency; n (%)		Cramer's V	p
	In category	In other categories		
category "oviduct"				
tube	16 (100)	60 (6)	.44	> 0.001
fallopian tube	16 (100)	60 (6)	.44	> 0.001
removed	8 (50)	16 (2)	.40	> 0.001
exception	2 (13)	0 (0)	.35	> 0.001
away	8 (50)	35 (4)	.29	> 0.001
link	7 (44)	28 (3)	.28	> 0.001
category "examination of the oviduct"				
tube	15 (79)	61 (6)	.37	> 0.001
fallopian tube	15 (79)	61 (6)	.37	> 0.001
laparoscopy	11 (58)	35 (4)	.35	> 0.001
endoscopy	12 (63)	43 (4)	.35	> 0.001
x-ray	3 (16)	1 (0)	.34	> 0.001
angiography	2 (11)	0 (0)	.32	> 0.001

Box 4: Most predictive words for the categories "oviduct" and "examination of the oviduct"

(2) For some categories, the input variables also included variables from other categories, most often with a negative sign. For example, the meta-model for "pregnancy test" included a sample of words (as an indicator variable) predictive for the category "menstruation" with a negative sign. This means that a lack of words predictive for "menstruation" was a strong indicator for the category "pregnancy test".

On the basis of the different input variables, we trained regression models and combined them to meta-models in order to classify a training sample of requests. We calculated a probability for each request to belong to any of the 38 categories and set a cut-off of 50%. Applying these models to a subset of requests (validation data), we achieved high rates of precision and recall, especially for the subject matter dimension; in the expectations dimension, recall was in some instances not quite as satisfying (Box 1).

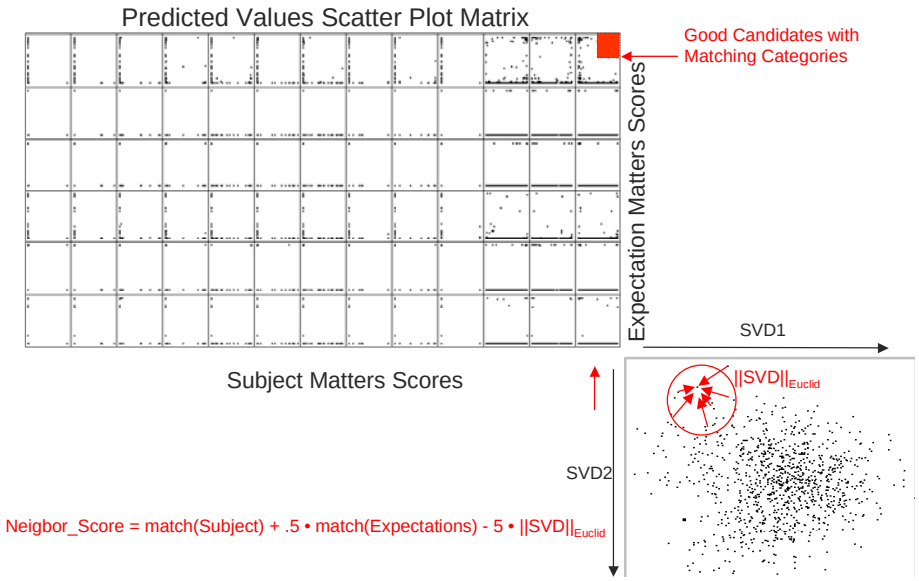
NEAREST NEIGHBORS

The good rate of classification of a validation sample of requests motivated us to experiment with a semi-automatic answering procedure. At the moment this procedure is, of course, only being tested in the form of a simulation.

For this simulation, we restricted ourselves to 2 methods of dimension reduction: Cramer's V indicator variables on the basis of the chi^2 statistic and a SVD with SAS Text Miner. A sub-sample of 50 randomly selected requests were considered to be new incoming requests to the expert forum. Out of the more than 6,000 requests from the whole sample of the website we tried to find matches for these 50 requests. Each of these new requests and its three matches should correspond in the subject matter and expectation dimension. Moreover, their first 10 SVDs should be rather close to each other. To find the most nearest neighbors we applied the following formula:

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

meaning that we considered the subject matter (_s) dimension with a factor of 1, the expectation dimension (_e) with a factor of 0.5 and weighed the SVDs with the factor 5. This procedure is graphically displayed in Box 5.



Box 5: Nearest neighbor method for automatic answers

In the following, we present 2 examples for a realisation of our scoring procedure. The first example (Box 6) is an excellent match. The 3 nearest neighbors were nearly perfectly in line with the simulated new request. Subject matter classification and expectation classification were identical and, obviously, the SVD match led to a high agreement between the 4 requests so that the experts' 'former' answers will meet the need of the sender of the 'new' request.

Excerpts from 'new' and former requests

'New' request	I would like to get pregnant again but my whole endocrine system is out of balance and extremely irregular. This may be due to the fact that I am still lactating. Shall I go to a center of reproductive medicine or shall I wait and see?
Automatic classification	*Cycle (99%), Treatment opportunities (98%)*
Neighbor 1	My cycle is extremely irregular and I try to find out my day of ovulation but, so far, without any success. Shall I see my gynecologist?
Automatic classification	*Cycle (100%), General information (99%), Treatment opportunities (90%)*
Expert Answer to Neighbor 1	I suggest seeing your gynecologist to check your endocrine status and to monitor your cycle. **(Appropriate!)**

Box 6: Nearest neighbor: a good example

The second example (Box 7) is far from good. Obviously, the new request was not adequately classified. The woman sending this request talked about ovulation - and this was the result of the automatic classification, but her real concerns had to do with pregnancy symptoms. Consequently, all 3 matches deal with pregnancy symptoms but will not address the woman's problem and the experts' answer won't help her.

Excerpts from 'new' and former requests	
'New' request	Who can tell me if I am pregnant? After ovulation, my morning temperature has increased constantly to 37^0C for 5 days now.
Automatic classification	*Ovulation (82%), General information (99%)*
Neighbor 1	For quite a long time, I have been trying to get pregnant and have measured my temperature daily. But it differs constantly. Why?
Automatic classification	*Ovulation (100%), General information (94%), Treatment opportunities (84%)*
Expert Answer to Neighbor 1	Sometimes it is because of your mode of measurement or your thermometer is not working properly. **(Not appropriate!)**

Box 6: Nearest neighbor: a poor example

Considering all 50 requests of the sub-sample, we found a good fit with three adequate matches in 29% of the cases, in 39 cases (78%) at least one match met the sender's subject matter and expectation and its sublime contents. However, in 11 requests (22%) none of the matches corresponded to the sender's information need; in these cases, none of the experts' former answers would have been helpful for the sender.

DISCUSSION

A combination of different methods to automatically classify requests to a medical expert forum, according to their subject matter and the sender's expectations, yielded rates of precision and recall above 80% in nearly all categories. This was a rather satisfying basis to implement a semi-automatic answering procedure for new requests. As simulation model was successful for about 80% of 'new' requests.

Several factors may have contributed to this result:

1. A high input of expert knowledge to build a large and meaningful start list of words for our chi^2 statistics.
2. The combination of different text mining strategies, including meta-models for regression, obviously met with the different text nature of our categories.
3. In some instances, the meta-models not only considered words or PCAs significant for a category but also the lack of certain words is a (negative) predictor for a category.
4. Although we had a lot of subject categories (n=32) they are still to simplistic to find nearest neighbors; SVDs were a good means to find, within a given category, those requests that are very close to each other.

One important limitation must be mentioned: although matches to a new request had to correspond with respect to the subject matter and the expectation and should be close to each other with regard to the SVDs, this does not protect us against mismatches due to false classifications. In this case, the experts' answers from former requests cannot meet the sender's information needs on principle. If a semi-automatic answering of medical requests should become reality, the possibility of such mistakes should be clearly mentioned. But even then, visitors to an expert health forum will be disappointed if they do not receive a more adequate and individual answer in due time.

In conclusion, a text mining strategy as presented in our paper may be helpful for health politicians and researchers to identify *in vivo* health needs and information needs of the public in different medical areas. It may lighten the workload of experts and help visitors receive a more timely answer, because similar patient requests and their corresponding answers could be collated, even before the expert himself/herself replies.

REFERENCES

Albright R. Taming Text with the SVD. 2004.
Available at: ftp://ftp.sas.com/techsup/download/EMiner/TamingTextwiththeSVD.pdf

Eysenbach G, Diepgen TL. Patients looking for information of the Internet and seeking teleadvice: motivation, expectations, and misconceptions as expressed in e-mails sent to physicians. Arch Dermatol. 1999;135:151-6.

Himmel, W, Reincke, U, Michelmann HW. Semi-automatic answering of health questions to a medical internet forum using text mining [in German]. In: Muche R, Bödeker R-H (eds.). KSFE 2007 - Proceedings der 11. Konferenz der SAS-Anwender in Forschung und Entwicklung (KSFE); March 1-2, 2007, Ulm University. Aachen: Shaker (ISBN 978-3-8322-6680-6) 2007; 113-22.

Reincke U. Profiling and classification of scientific documents with SAS Text Miner. Paper presented at the third "Knowledge Discovery" workshop in October 2003 in Karlsruhe, Germany. Available at: http://km.aifb.uni-karlsruhe.de/ws/LLWA/akkd/8.pdf

Umefjord G, Sandström H, Malker H, Petersson G. Medical text-based consultations on the Internet: A 4-year study. Int J Med Inform. 2007; doi:10.1016/ijmedinf.2007.01.009

Weiss SM, Indurkhya N, Zhang T, Damerau FJ. Text Mining: Predictive Methods for analyzing unstructured information. New York: Springer; 2005.

Widman LE, Tong DA. Requests for medical advice from patients and families to health care providers who publish on the World Wide Web. Arch Intern Med. 1997;157:209-12

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Professor Wolfgang Himmel
Dept. General Practice / Family Medicine
University Medical Center Göttingen
Georg-August-University
Humboldtallee 38
37073 Göttingen
Tel # 00 49. (0) 551. 39-2648
Fax # 00 49. (0) 551. 39-9530
whimmel@gwdg.de