Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)
© 2015 European Federation for Medical Informatics (EFMI).
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-512-8-572

Assisting e-patients in an *Ask the Doctor Service*

Amine ABDAOUI[a], Jérôme AZÉ[a], Sandra BRINGAY[a] and Pascal PONCELET[a]
[a] LIRMM, 860 St Priest Street, 34095 Montpellier, France

Abstract. *Ask the doctor services* are personalized forums allowing patients to ask questions directly to doctors. Usually, patients must choose the most appropriate category for their question among lots of categories to be redirected to the most relevant physician. However, manual selection is tedious and error prone activity. In this work we propose to assist the patients in this task by recommending a short list of most appropriate categories.

Keywords. Health forums, text categorization, recommender systems.

Introduction

Recently, specialised expert forums appeared allowing patients to directly ask questions to health experts. These forums, also named *ask the doctor services* (1), are often organized into lots of categories where patients can ask their questions and be redirected to medical specialists. Usually patients must choose the most appropriate category, among lots of categories, to have relevant answers to their questions. Unfortunately, such a task is tedious, time consuming but most importantly error-prone activity. Many users prefer choosing the category "Other" when they are uncertain or in a hurry. The objective of this work is to propose a tool that recommends a short list of categories according to the patient question and additional information that can be provided (e.g. title of the question, age, gender, etc.). The patient will still have the possibility of choosing another category that has not been recommended.

Text Mining techniques are increasingly applied to the huge amount of data available on health forums for different purposes, such as: identifying threads that need to be moderated (2), extracting adverse drug reactions (3), identifying emotions and sentiments (4), etc. In the case of the wildly studied text categorization techniques (5) on *Ask the doctor* forums, we found an interesting approach proposed in (6) for automatically classifying lay requests to medical experts. In our work, the task is slightly different. We aim at helping the patient by recommending a short list of most appropriate categories to his/her question. In its most common formulation, the recommendation problem is reduced to the problem of estimating ratings (scores) for the items that may interest a user (7). The items with the highest scores will be recommended. In the case of our work, the items are the health categories and the scores are estimated by combining the predictions of trained classification models. The models used in this work have been trained on a French Ask the doctor service. It is assumed that its questions have been correctly categorized.

The rest of the paper is organized as follows: Section 1 describes the corpus and the approach used to learn the classification models and compute the scores.

Experiments conducted are presented in Section 2. A discussion is proposed in Section 3. And finally, Section 4 concludes and gives our future works.

1. Methods

1.1. Corpus

6,140 questions have been collected from a French *ask the doctor service*[1]. This website allows patients to ask paying questions to health experts. In order to ask a new question, patients should choose the appropriate category among twenty proposed categories[2]. Usually, questions posted in the category "*Other*" are very heterogeneous and most of them should have been posted in another existing categories[3]. This observation supports our hypothesis that if the users are not assisted, they tend to choose *"Other*" rather than searching the real category. For our experiments, we removed questions posted in this category.

1.2. Learning the models

For each category, five classification models have been trained using the *Weka* API[4] (8). Each Model predicts whether a question belongs or not to the corresponding category. The classification models used are: SVM SMO, Naive Bayes, J48, Random Forest and JRip. A balanced data set has been created for each category in order to avoid overfitting. Each one contains all the questions posted in the given category, and the same number of questions taken from the rest of the categories by successive draws without replacements.

The features used in this classification process are the following: questions (unigrams + bigrams), length of the questions (number of words), title of the questions (unigrams + bigrams), user's gender, age, size and weight. All these features are given in our corpus except the gender of the user. Dictionaries of names have been used to predict the gender from the name given on the website. Before tokenizing questions into ngrams, a French stop word list has been used to remove noisy words. Then, ngrams that appears at least 2 times have been extracted. Each ngram computes its tf-idf score (term frequency * inverse document frequency) (9). Finally, a feature subset selection has been performed by computing the information gain (10) of each attribute with respect to the class attribute.

1.3. Recommendations based on scores

Let M be the set of the used classification models, $m \in M$ a given classification model, x the new question to classify and $Pm(c|x)$ the probability that x is assigned to the category c by the classification model m. Two scores have been used in order to detect the most appropriate topics according to a specific question. To compute each score, we combine the predictions (11) of the trained classification models learned for each category as described in the previous section. Other scores have been experimented but we present only those which are efficient and understandable.

1 www.masantenet.com [collected on: 18/02/2014]
2 www.lirmm.fr/~abdaoui/Annexe1.pdf
3 www.lirmm.fr/~abdaoui/Annexe2.pdf
4 Weka version 3.6.10 has been used in this work

1.3.1. Score 1 is a simple vote count of classification models that assigned the question x to the category c.

$$score1_c(x) = \sum_{m \in M} 1_{Pm(c|x) \geq 0.5} \qquad \in [0, |M|]$$

1.3.2. Score 2 considers both the number of models that agree for assigning the question x to the category c and the probabilities associated with these predictions. The number of algorithms is given more weight by using the exponential function.

$$score2_c(x) = \begin{cases} exp(score1_c(x)) \prod_{\substack{m \in M \\ Pm(c|x) \geq 0.5}} Pm(c|x) & , \ if\ score1_c(x) > 0 \\ 1 & , \ Otherwise \end{cases}$$

2. Results

2.1. Evaluating the learned models

Table 1 presents F1-scores obtained on the balanced set of each category by doing a 10-fold cross validation as well as the number of selected attributes. F1-score is the harmonic mean of precision[5] and recall[6], while n-fold cross validation is a validation technique that randomly partition the data set into n equal size subsets. A single subset is used for testing, while the remaining n-1 are used as training set. This process is repeated n times so that each of the n subsets is used as a testing set exactly once.

Table 1: F1-scores obtained by a 10-fold cross validation to test the classification models

Category	Number of attributes	SVM SMO	Naive Bayes	J48	Random Forest	JRip
Articulation	431	0.84	**0.85**	0.64	0.82	0.59
Cancer	286	**0.87**	0.79	0.76	0.8	0.67
Children	566	**0.9**	0.89	0.85	0.86	0.84
Dermatology	1,877	**0.91**	0.88	0.8	0.86	0.71
Drugs / Alcohol / Tobacco	127	0.85	**0.94**	0.67	0.8	0.72
Handicap	52	0.81	**0.91**	0.81	0.74	0.75
Heart	305	**0.88**	0.84	0.75	0.87	0.75
Infection / Virus	767	**0.88**	0.77	0.63	0.81	0.54
Liver / Digestion	836	**0.92**	0.82	0.69	0.82	0.67
Nutrition / Diabetes	523	**0.9**	0.85	0.7	0.81	0.62
Ophthalmology / Otolaryngology	622	**0.87**	**0.87**	0.74	0.81	0.68
Pain	1,991	**0.89**	0.81	0.78	0.82	0.8
Pneumology	357	0.91	**0.92**	0.85	0.87	0.78
Pregnancy / Gynecology	4,194	**0.9**	0.86	0.83	0.86	0.81
Prostate	38	0.88	**0.97**	0.75	0.91	0.72
Psychiatry / Neurology	647	**0.86**	0.8	0.61	0.75	0.55
Sexology	1,198	**0.89**	0.79	0.7	0.83	0.66
Sleep	242	**0.85**	0.84	0.75	0.81	0.67
Welfare systems / Safety at work	30	0.83	**0.87**	0.78	0.81	0.68

[5] The precision is the number of correct positive results divided by the number of all positive results.
[6] The recall is the number of correct positive results divided by the number of positive results that should have been returned.

2.2. Evaluating the recommendations based on each score

The learned models have used in the recommendation task according to each score. Table 2 presents accuracies obtained on each category using the two proposed scores when recommending $k \in [1, 5]$ categories. The accuracy computes the proportion of correct recommendations. In our case, a recommendation is considered to be correct if the real category of the question is present among the k categories recommended.

Table 2: Accuracies obtained on each category when recommending k=1, 2, 3, 4 and 5 categories.

k	1		2		3		4		5	
S1=score1, S2=score2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Articulation	0.28	0.41	0.64	0.78	0.84	0.86	0.87	0.93	0.94	0.95
Cancer	0.58	0.56	0.74	0.69	0.83	0.8	0.89	0.87	0.94	0.88
Children	0.74	0.77	0.83	0.84	0.87	0.88	0.92	0.92	0.96	0.96
Dermatology	0.63	0.72	0.82	0.86	0.9	0.91	0.93	0.93	0.96	0.95
Drugs / Alcohol / Tobacco	0.45	0.62	0.87	0.92	0.96	0.98	0.98	0.98	0.98	0.98
Handicap	0.5	0.69	0.81	0.92	0.88	0.92	0.88	0.92	0.92	0.92
Heart	0.31	0.66	0.63	0.77	0.72	0.81	0.81	0.86	0.83	0.9
Infection / Virus	0.44	0.32	0.58	0.51	0.72	0.64	0.8	0.73	0.86	0.8
Liver / Digestion	0.53	0.64	0.75	0.78	0.84	0.85	0.87	0.9	0.9	0.92
Nutrition / Diabetes	0.52	0.64	0.76	0.82	0.83	0.89	0.89	0.92	0.9	0.93
Ophthalmology / Otolaryngology	0.47	0.56	0.65	0.71	0.75	0.81	0.82	0.85	0.89	0.91
Pain	0.61	0.7	0.81	0.85	0.88	0.9	0.91	0.92	0.92	0.93
Pneumology	0.67	0.71	0.86	0.88	0.9	0.95	0.95	0.96	0.96	0.97
Pregnancy / Gynecology	0.73	0.8	0.85	0.86	0.89	0.9	0.93	0.93	0.95	0.95
Prostate	0.88	0.69	1	1	1	1	1	1	1	1
Psychiatry / Neurology	0.45	0.48	0.71	0.72	0.8	0.85	0.87	0.91	0.92	0.92
Sexology	0.64	0.53	0.84	0.79	0.9	0.87	0.94	0.91	0.96	0.94
Sleep	0.42	0.72	0.71	0.85	0.83	0.9	0.88	0.94	0.95	0.96
Welfare systems / Safety at work	0.62	0.58	0.81	0.69	0.96	0.92	0.96	0.96	0.96	0.96
Global	**0.6**	**0.65**	**0.78**	**0.8**	**0.86**	**0.87**	**0.9**	**0.91**	**0.93**	**0.93**

3. Discussion

F1-scores obtained using the cross validation are relatively high. SVM gives the highest F1-scores (between 0.81 and 0.92) while JRip gives the worst ones (between 0.54 and 0.84). Moreover, the categories with the highest number of questions give models with the highest number of attributes.

Regarding the performance of each score, we notice that the second score gives higher accuracies. The difference is noteworthy when recommending few categories and tends to decrease when recommending more categories. As expected, the accuracies increase with the increase of the number of recommended categories. Indeed, the more categories we recommend, the more likely we obtain the correct one. But at the same time, the more likely we predict false categories. For example, if we recommend two categories, at best we will have one correct recommendation and one wrong recommendation. While if we recommend three categories, at best we will have one correct recommendation and two wrong recommendations, etc. Therefore, our goal is to recommend a small number of categories (k) with an acceptable accuracy.

4. Conclusion

This paper addresses the recommendation of medical categories to the users of an *Ask the Doctor* website. The method can be easily reproduced on other forums. It is based on plain text categorization with unigram, bigram and four additional features after stop word removal, tf-idf scoring and information gain attribute selection. Five classifiers have been trained for each category in one-against-all mode, and two voting schemes have been tested to merge their predictions. Voting schemes should give better results since the sets of misclassified questions by the different classifiers would not necessarily overlap.

To improve this work, first, we plan to compare these results with the ones obtained by simply detecting medical concepts in the question and linking them to the categories. Then, it is assumed that the questions posted in all categories, except "Other", are correctly assigned. This assumption may be confirmed by a manual annotation or an automatic validation using search engines (for example we can cross the snippets returned by search engines when searching the category name with those present in the website). Moreover, the same data has been used for training the models and testing the recommendations, we are planning to manually annotate the questions posted in the category "Other" and use them as a testing set. Furthermore, adapting the number of recommendations based on the length of the question is worth study. In fact, we noticed that long questions seem to be better handled by this method than short ones[7]. Therefore, we can recommend more categories for short questions and less for long ones. Finally, the identity and the history of the user may be a feature to include. If a specific user asked many questions in the *Liver category*, he may have a liver disorder and may ask more questions in this category.

References

1. Umefjord G, Hamberg K, Malker H, Petersson G. The use of an Internet-based Ask the Doctor Service involving family physicians: evaluation by a web survey. Fam Pract. 2006;23(2):159–66.
2. Huh J, Yetisgen-Yildiz M, Pratt W. Text classification for assisting moderators in online health communities. J Biomed Inform. 2013;46(6):998–1005.
3. Segura-Bedmar I, de la Pena S, Martınez P. Extracting drug indications and adverse drug reactions from Spanish health social media. ACL 2014. 2014;98.
4. Yu B. The Emotional World of Health Online Communities. Proceedings of the 2011 iConference. New York, NY, USA: ACM; 2011. p. 806–7.
5. Yang Y, Liu X. A Re-examination of Text Categorization Methods. Proceedings of the ACM SIGIR Conference. New York, NY, USA: ACM; 1999. p. 42–9.
6. Himmel W, Reincke U, Michelmann HW. Text Mining and Natural Language Processing Approaches for Automatic Categorization of Lay Requests to Web-Based Expert Forums. J Med Internet Res. 2009;11(3):1–1.
7. Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Knowl Data Eng IEEE Trans On. 2005;17(6):734–49.
8. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explor Newsl. 2009;11(1):10–8.
9. Salton G. Developments in Automatic Text Retrieval. Science. 1991;253(5023):974–80.
10. Mitchell TM. Machine Learning. 1 edition. New York: McGraw-Hill; 1997.
11. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. IEEE Trans Pattern Anal Mach Intell. 1998;20(3):226–39.

7 www.lirmm.fr/~abdaoui/Annexe3.pdf