DOI: http://dx.doi.org/10.14236/ewic/FDIA2015.18

Whether a CQA User is a Medical Professional? Work in Progress

Alexander Beloborodov
Ural Federal University
620002, 19 Mira Street, Ekaterinburg, Russia
xander-beloborodov@ya.ru

This work-in-progress aims to address the problem of detecting whether a CQA user is a medical professional or not. Proposed approach is based on the technique of learning from positive and unlabeled examples. A few classification features and a simple evaluation methodology are presented.

Keywords: User Expertise, Medical IR, One-class SVMs, Community Question Answering, CQA

1. INTRODUCTION

The paper describes an ongoing research project investigating Community Question Answering (CQA) users answering questions in the domain of people's health. CQA is a service where people are able to ask any question and answer questions asked earlier. There is a wide variety of question topics. Our research project is focused on the people's health topic. It is very important for user, asking question about his/her health, to know the level of expertise of an answering user. We investigate a problem of detecting whether an answering user is a medical professional or not.

The paper is organized as follows: section 2 discusses relevant previous work; the data and resources on which experiments conducted are described in section 3; the prospective method and some features are described in section 4. Finally, in section 5 a few challenges for future work are discussed.

2. RELATED WORK

The tasks of expert search and classification are investigated in a number of works.

In [Liu et al., 2005] the authors focus on automatically finding experts in an open-domain community-based QA service. They cast the expert finding problem as an IR problem where the given question can be viewed as query and the expert profiles can be viewed as documents. Such documents are ranked using language models. While considering primarily language aspects this approach are not taking in account other useful and interesting types of features (for example, statistical or semantic features).

A few different types of features are used in [Pennacchiotti & Popescu, 2011]. The work addresses the task of Twitter user classification by leveraging observable information such as the user behaviour and the linguistic content of the user's Twitter feed. The authors provide an in-depth analysis of the relative value of feature classes and show experimentally that content features are in general highly valuable in classification tasks.

The task of learning from positive and unlabelled examples is considered in [Manevitz & Yousef, 2002]. In [Zhang & Zuo, 2008] a good implementation called One-Class SVMs is suggested and described in detail. This implementation is chosen for our experiments. In addition, the authors argue that the absence of negative information entails a price, and one should not expect as good results as when they are available. That is why it is so important to find a way to form a sample of negative examples.

3. DATA & RESOURCES

The research project is focused on Russian language Community Question Answering (CQA) Otvety@Mail.Ru (Otvety means answers in Russian). All questions in the service are divided into predefined set of categories. We study all questions related to people health with corresponding answers from 4 categories: *Diseases and Medicines; Doctors, Clinics, and Insurance; Doctors' answers; and Kids' Health* in the timespan from 1 April 2011 to 31 December 2012 – 227,828 questions at all.

For evaluation purposes the online survey among most active users is conducted. The respondents answered a few questions about their professional skills and motivation to answer questions. Every respondent was able to specify his/her email by which he/she could be identified in the service. Among other questions, people were asked also a question:

Is your profession related to the people's health?

An invitation to participate was sent to about 700 users which are most active in the categories. 171 users participated in the survey, 54 of them specified their emails, 26 respondents answered that their profession is related to people's health, and 28 respondents answered that their profession is not related to people's health.

4. PROPOSED APPROACH

4.1 Training set

Proposing approach is an automatic classification using supervised machine learning that needs sufficiently large training set. Unfortunately, online survey results do not provide us with data set of suitable size. Therefore training set is collected by other means while the online survey results are supposed to be a test set.

CQA answerer has an option to specify his/her answer source. The answer source field has an unstructured plain text format. Users in the health-related categories are often filling the answer source field with phrases such as "I'm physician myself!" or "My medical degree". From 1867 unique answer sources 182 were selected as items certifying that answer author has profession related to people's health. Assuming that source fields contain true statements we have 263 users related to people's health to form positive samples for the training set.

4.2 Proposed method

Although the answer source field often contains medical profession mentions, there are no mentions of other professions the field, so we cannot form a negative samples subset in such a way. Therefore, at this stage, the technique of learning from positive and unlabelled examples is used. One-class SVM with non-linear kernel RBF is used as the technique implementation.

Initial experiments show that the algorithm is often classifying users who specified his/her profession as not related to people's health as positive cases (26 from 28).

4.3 Features

Three types of features are selected for the classification task.

4.3.1. Statistical features
This group of features usually characterizes texts in general. Some examples are mean answer length (in words), an amount of unique words in all user answers, an amount of unique words divided by a whole amount of words answerer mentioned. Additionally we slightly exploit CQA structure to enrich feature set with an amount of user answers divided by an amount of his/her questions inside one particular category.

4.3.2. Linguistic features
After manual investigation of surveyed user answers we assume that speech of a person who is a prospective physician is less emotional than speech of a CQA user in general. According to this assumption we exploit following features: the amount of all general punctuation marks, the amount of "emotional" marks ('!', '?', '(', ')'), the amount of repeating marks ("…..", "!?!?!?!?", ")))))"), fully uppercased words ("EXAMPLE") abuse.

4.3.3. Semantic features
The most interesting features are semantic. Assuming the fact, that a vocabulary of user with medical degree is full of specific terms, a medical domain term dictionary is collected. Feature is an amount of specific terms in user answers.

Many physicians even answering in the CQA do not believe that answer could replace regular real doctor visits. Therefore a hypothesis that users with medical degree more often route a questioner to a real physician is formulated. To verify it two sets of words collected: words analogous to word "physician" and modal words. Examples of "physician"-like words: *doctor, specialist, ambulance, clinic, dentist,* etc. Examples of modal words are *consult, go, see, necessarily, need,* etc. An event when "physician"-like word and modal word appeared in the distance 2 or less from each other are recognizing as recommendation to visit a doctor:

Q: *Help me please. Inexplicable rash on the body! [Photo attached]*

A: *It looks like herpes. It is contagious and can be transmitted from animals. Urgently need to* ***[see a doctor]***

Traditional medicine and ICD-10 drug mentions are serving as features as well.

5. CHALLENGES

There are some challenging questions that we faced during the research project.

The current evaluation methodology is supposed to test the method performance using the online survey results as a gold standard. This approach has a number of drawbacks:

a) We have only 26 respondents answered that their profession is related to people's health, and 28 respondents answered that their profession is not related to people's health. Such amount seems too small to make reliable conclusions but it still useful on the initial stage of the research.

b) A manual investigation shows that answerers, who specified his/her profession as not related to people's health, are often giving answers of a good quality. This can make it difficult to understand who can be considered as a medical professional.

Other important features that need to be integrated into the evaluation process are readability and understandability of expert answers. As shown in [Zuccon & Koopman, 2014], the understandability is considering as a critical issue for supporting online consumer health search because consumers may not benefit from health information that is not provided in an understandable way; and the provision of unreliable medical condition or treatment may led to negative health outcomes.

As follows from [Zhang & Zuo, 2008], with the absence of negative information we should not expect as good results as when negative examples are available. At this point building such a training set is a challenge as well.

One more method drawback is a lack of user-level features, such as the user score, age, gender, and so on.

6. ACKNOWLEDGMENTS

This work was supported in part by the ELIAS (Evaluating Information Access Systems), an ESF Research Networking Programme, and in part by the Russian Foundation for Basic Research, project #14-07-00589 "Data Analysis and User Modelling in Narrow-Domain Social Media".

The author also would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the work.

7. REFERENCES

Liu, X., Croft, W. B., & Koll, M. (2005). Finding experts in community-based question-answering services. Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 315-316). ACM.

Pennacchiotti, M., & Popescu, A. M. (2011). A Machine Learning Approach to Twitter User Classification. ICWSM, 11, 281-288.

Zhang, B., & Zuo, W. (2008). Learning from positive and unlabelled examples: A survey. Information Processing (ISIP), 2008 International Symposiums, 650-654. IEEE.

Manevitz, L. M., & Yousef, M. (2002). One-class SVMs for document classification. the Journal of machine Learning research, 2, 139-154.

Zuccon, G. & Koopman, B. (2014). Integrating understandability in the evaluation of consumer health search engines. Medical Information Retrieval (MedIR) Workshop, 11 July 2014, Gold Coast, Australia.