Does Everybody Lie? Characterizing Answerers in Health-Related CQA

Alexander Beloborodov, Pavel Braslavski
Ural Federal University
Yekaterinburg, Russia
xander-beloborodov, pbras@yandex.ru

***Abstract*—The study described in the paper aims at multi-faceted characterization of active community question answering (CQA) users who provide answers to health-related questions. The study employs various research techniques – both qualitative (surveys) and quantitative. With two online surveys we get insights into 1) perception of online health-related information and its use by patients by medical professionals and 2) motivation of most active CQA answerers, a significant share of which apparently constitute users with medical education. In the second series of experiments we apply topic modeling to a yearly collection of questions and answers from a popular Russian CQA service in order to find users focused on a particular topic. Further, we attempt to find users with professional medical background based on the lexis of their answers. The obtained results provide a better understanding of motivation and background of CQA users and can be used for the improvement of CQA services, as well as for solving problems such as CQA content quality evaluation, expert search, and question routing, etc.**

I. Introduction

In the winter of 2016 a New York-based health insurance company launched an advertising campaign depicting an orc with a laptop and citing "Ask a doctor instead of randos on the internet". With this ad the company promoted its mobile app allowing users to consult a doctor and get prescriptions remotely, as well as reminded of untrustworthy and anonymous health-related information available on the web.

Nevertheless, the Web has become an important information source of health information for laypeople. According to research conducted in 2012 by the Pew Research, 72% of the US internet users looked online for health information within the past 12 months (http://www.pewinternet.org/fact-sheets/health-fact-sheet). These figures are lower in Russia, but still substantial and growing: in 2014, 21% of Russian population searched for health-related information online (http://healthcare.ipsos-comcon.ru/files/synovatecomcon_healthindex-daydzhest19.pdf – in Russian).

There are various destinations for online health information seekers. Major search engines like `Google`, `Bing`, or `Yandex` still remain the primary entry points, but in response to health-related queries they usually lead to specialized resources such as `WebMD` or `patientslikeme`. Another popular destination is community question answering (CQA) platforms such as `Yahoo!Answers` or `WikiAnswers`. These sites allow users to post their questions on an arbitrary topic, describe question's context, and get timely answers, which can be especially crucial in case of health-related questions. CQA is a good complement to web search that allows for a more detailed description of information need, delivers more social and personalized search experience, and suits users with low search engine proficiency. CQA services have accumulated a large amount of human knowledge in form of questions and answers, but content quality can be the major issue, which is critical in case of medical information.

In our previous studies we investigated topical characteristics of health-related CQA content [1] and semi-automatic quality evaluation of the data [2]. In the current study we concentrate on the users providing answers to health-related questions on CQA services. We aim at characterizing users answering activity, finding out motivation and background of the most active answerers, their topical preferences and specialization, as well as attempting to automatically detect the groups of most professional answerers based on their lexis.

In our work we use a collection of questions and answers in the health-related categories of a popular Russian CQA service Otvet@Mail.Ru (https://otvet.mail.ru/) – 227,828 questions and corresponding answers provided by 127,602 unique users during the year of 2012. The majority of users (54%) gave only one answer during the year; in our analysis, we focus on the most active users.

The novelty of our study is that we characterize users who answered health-related community questions from various perspectives using different research methods – both qualitative and quantitative. Thus, we conducted two surveys – among doctors and active answerers. In the first case, we wanted to find out how medical professionals perceive online health information and the way it is used by patients. In the second case, we aimed at exploring the motivation of the most active users and their attitude towards evidence-based medicine. In addition, we applied topic modeling to the dataset in order to define the topical spectrum of answers and find 'monothematic' users. (In case of Otvet@Mail.Ru a special technique must be applied, since in contrast to Yahoo! Answers, for example, the platform has a very coarse two-level category structure.) Finally, we attempted to find CQA users with professional medical background based on the lexis of their answers.

The obtained results show that the professional medical community is in general skeptical about freely available health-related information on the web, fearing that it could do more harm than good. The main motivations of active users in CQA Health category are altruistic: the users basically want to help others and share knowledge. The topic modeling allows finding

active answerers focused on a particular subject. A list of specific medical terms can be a good supplement for finding experts in health domain. In general, the results allow for a more comprehensive and multifaceted understanding of CQA users that provide answers in the health-related categories. The results can be used for improvement of CQA services and community modeling, as well as for such applied tasks as expert search, question routing, and content quality evaluation in social media.

II. Related work

CQA gained a great deal of attention in research community in the last decade. There is a series of studies investigating the quality of CQA data [3], [4] and possibility of re-using CQA data for Web search and question answering [5], [6]. Researchers propose a comprehensive approach for automatic evaluation of answer quality based on a wide variety of content, user and interaction features. Moreover, researchers propose different notions of content quality from different perspectives. Thus, researchers distinguish between askers [5], [7], [4] and external perception of answer quality [3]. A cognate area is community questions classification in regard of informational and conversational intent [8], [9].

In context of our research the most relevant studies are those shifting attention from content quality or CQA data recycling to the answering users. Answerer's motivation is considered in [10]. The authors specify a commitment to a social role and a social gratification among main reasons to answer (these findings are confirmed in our work). Dearman et al. [11] complement this research by finding out why CQA users do not answer questions. Pelleg et al. [12] analyze sensitive information like body measurements or income shared on CQA sites. They state that CQA users tend to be open even in case of sensitive topics both asking questions and providing answers.

Expert search among CQA users is also a task close to our study: researchers aim at finding good answer providers instead of finding good answers themselves [13]. Expert users provide technically correct and reliable content and could be recommended for newly arrived questions, which can increase the potential of a Q&A community [14], [15].

Dedicated studies on health-related CQA content are relatively few and rely mostly on manual processing. Zhang's study [16] described linguistic features, users' motivations and question types, as well as temporal, emotional and cognitive aspects of a sample of about 270 questions in *Health* category of Yahoo!Answers. Kim et al.[17] semi-automatically assessed around 5,400 questions and answers on H1N1 influenza strain posted on Yahoo!Answers. The authors identified major subtopics in H1N1 questions, types of resources askers and answerers referred to, and medical concepts mentioned in the data. In our previous work [1] we use topic modeling and domain dictionaries to learn more about medical CQA content and to match dynamics of automatically extracted topics (such as *acute respiratory infection*) with real-world events. In the current study we exploit those methods in a deeper and more detailed manner. Works [18], [19] discuss medical CQA quality issues and evaluate health-related answers from different angles: the authors compare quality perception by different groups - questioners, health reference librarians, and nurses. Authors in [2] approach a similar problem, but try to conduct evaluation in a semi-automatic way. Sillence et al. [20] focus on users' Q&A in medical domain and reveal a paradox: laypeople like other patients' experience shared online, but they do not fully trust the information they find there.

The current study aims at finding users with professional medical background, which can be seen as a kind of profession-based search. A similar task has been investigated in two recent studies: researchers try to identify scientists [21] and journalists [22] on Twitter based on an initial set of known professionals. However, in our case the data is much sparser: we do not dispose of profile information, as well as a reliable and sizeable seed set. Twitter users are much more active, and, which is most important, there are no strong ties between CQA users in contrast to following, re-tweeting, and mentions on Twitter.

III. Data and users

Otvety@Mail.Ru is a Russian counterpart of Yahoo!Answers with similar rules and incentives. The initial data set used in the study contained 11,170,398 questions and corresponding answers (4.85 answers per question) provided by 2,690,358 unique users during 2012. The service has a two-level directory with about 30 top-level categories and about 200 subcategories. The users assign their questions to a second-level category using drop-down lists. We focused on four second-level categories: *Diseases and Medicines*, *Doctors, Clinics, and Insurance*, *Doctors answers*, and *Kids Health* that contain 227,828 questions and corresponding answers (authored by 225,427 unique users) in total. Hereafter we will refer to the union of these four categories as *Medicine and Health* category.

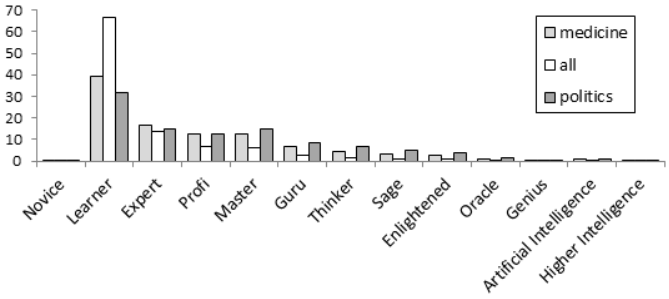

Fig. 1. Answerer percentage distributions over user levels in the whole data set, *Medicine and Health* and *Politics* categories

Each data item (question or answer) is provided with additional meta-information, for instance, timestamp and 'best' indicator showing an answer won in user polling. Each user is characterized by several scores or badges, e.g. rating, level, and *best answer ratio (BAR)* (percentage of 'best' answers given by a user among all her/his answers). User level is a string assignment based on her/his Q&A activity and rating. Fig. 1 shows possible user levels ordered by complexity and compares user percentage distributions over the levels in the whole data set and two categories: *Medicine and Health* and *Politics*.

To find out what distinguishes user posting behavior in the medical domain from one in other domains we chose the

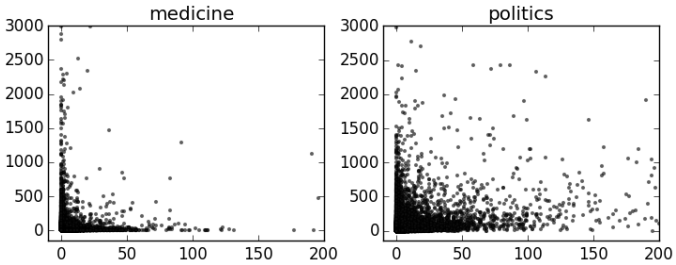

Fig. 2. User activity in *Medicine and Health* (left—225,427 points) and *Politics* (right—104,370 points) categories. Each point denotes one user

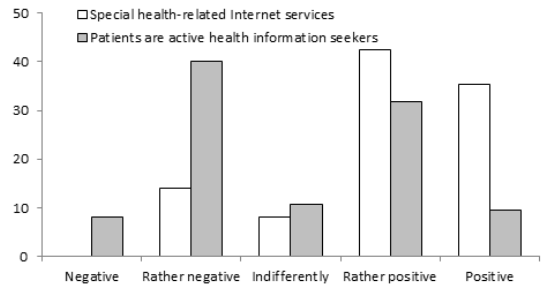

Fig. 3. Doctors' attitude to Internet medicine

Politics category as of comparable-size by amount of questions and user level distributions (see Fig. 1). Table I shows that *Medicine and Health* category has much more users (225,427 vs. 104,370) and is significantly smaller by other parameters. It means that a large number of the medical categories people basically ask and a relatively small amount of people answer. *Politics* category has answerers prevalence instead. In the Fig. 2 one can see that *Medicine and Health* users are much less scattered over questioning/answering activity area. Using terminology introduced in [9] we can suppose that the medical category users tend to ask informational questions while the *Politics* users are engaged in conversations. So, in the Fig. 2(left) one can recognize two distinctive groups of medical domain users: those who substantially asks (horizontal scatter) and those who substantially answers (vertical scatter). We are focusing particularly on the second group of users as potentially experienced in medical topics.

	Medicine and Health	Politics
Number of questions	227 828	170 202
Answers per question	4.13	8.57
Q&A users	225 427	104 370
Questions per 'asker'	1.83	5.01
Answers per 'answerer'	7.38	16.72

TABLE I. STATISTICS COMPARISON BETWEEN THE *Medicine and Health* AND *Politics* CATEGORIES HAVING COMPARABLE QUESTION COUNTS

IV. RESULTS

A. What doctors think, why users answer

As a preliminary step, we carried out a survey to figure out how medical professionals in Russia perceive online health information and the patients practice to search medical information online alongside with consulting a doctor. We published a questionnaire on a Russian professional social network "Doktor na rabote" (*Doctor at work*) (http://www.doktornarabote.ru). The network has a multi-step registration procedure that ensures only authorized medical professionals can register with the social network; the site claims to have more than 400K registered users. We received 85 responses from a wide diversity of specialists (internists and surgeons are most frequent specializations of respondents, accounting for 13% each) and professional experience (under 5 years – 30.6%, 5–10 years – 25.9%, 10–20 years – 17.6%, more than 20 years – 25.9%). 54.1% of respondents work at ambulant clinics, 35.3% at hospitals, and 16.5% – in research and education.

Most doctors (78.8%) are aware of online forums and Q&A sites on health and medicine. However, the majority do not answer health-related questions online, 28% of respondents do it only occasionally. Among those who answer health-related questions online 42.2% are motivated by the desire *to help people*; 35.6% are led by *sharing experience and knowledge* (respondents had multiple-choice answers to select from). In general, the perception of online information is mixed: the respondents agree that there is both trustful and doubtful information on the web, and only a professional is able to interpret and evaluate it. Fig. 3 depicts most interesting survey results: while most doctors commend specialized web resources, opinions about patients searching medical information online are divided: 41.2% *positive* and *rather positive than negative* vs. 48.2% *negative* and *rather negative than positive*.

Along with surveying the physicians we examined the CQA users' opinion about their trust in medical information propagated through the Internet. We sent an email with the invitation to participate in an online survey to the 800 most active users in medical categories. 172 (21.5%) users responded to the request. The questionnaire contained an optional field for email that was filled by 87 (50.6%) respondents, which allowed us to match them to Otvety@Mail.Ru users. A surprisingly high share of active answerers who filled out the form had medical education – 83 (48%), 63 (37%) claimed to have university degree in Medicine.

Despite the fact that the survey was conducted among active users, 56 people (33%) reported that they visit Otvety@Mail.Ru only once a week; 60% of all respondents answer questions related to their professional experience. The main goal of the questionnaire was to get insight into motivation of most active answerers in Health and Medicine categories. Fig. 4 shows the distribution of selected answers.

B. Topics, answerers are focused on

Although Otvety@Mail.Ru has its own predefined set of categories these categories are quite coarse: the *Medicine and Health* consists of only four Otvety@Mail.Ru categories. We exploit topic modeling approach to learn hidden topics in the *Medicine and Health* category and investigate the finer-grained structure of the collection and hence – of the users.

To discover topical structure of the collection we used a Latent Dirichlet Allocation (LDA) approach [23]. In our case, a 'document' refers to a concatenation of all answers of a particular user in the *Medicine and Health* category. All terms in the collection are lemmatized by the Snowball algorithm,

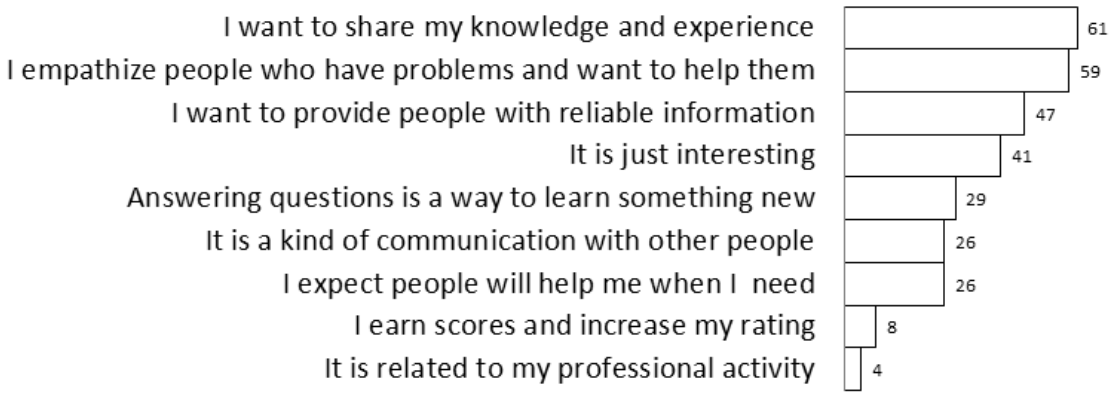

Fig. 4. CQA active users opinions on the question: *What does motivate you to answer questions on the service?* (The users have multiple-choice answers to select from.)

then the 100 most frequent words and words with counts of 10 and below are eliminated. After all preprocessing steps the collection contained about 30,000 unique terms. As discussed in section III, medical professionals are likely to be among the active answerers rather than askers. As the topical modeling approach is based on text analysis we consider only users having at least 50 words in their *Medicine and Health* answers. There are 16,124 users out of 127,602 unique users in *Medicine and Health* category (12.6%) satisfying this condition. We applied GibbsLDA++, an implementation of LDA, with 100 topics and default parameters $\alpha = 0.5$ and $\beta = 0.1$ [24]. The most of resulting topics appeared quite meaningful.

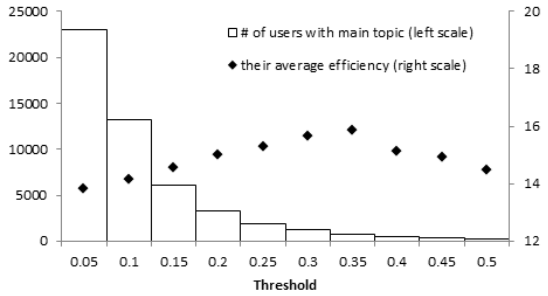

Fig. 5. The users count and their average efficiency (i.e. BAR) depending on the different *focus* thresholds

Having topic model built on the corpus of 'documents' representing users we can infer the user's topics of interest. It seems natural that people are usually focusing on fields of their expertise. Therefore we hypothesize that user having one topic dominant over other topics is likely to be an expert in that topic. To formalize it lets define a user *focus* as the highest probability in her/his topic distribution: $focus(user) = argmax(P(topic|user))$. If this metric for a user is more than some predefined threshold we can consider that user as a person focused in one thematic field i.e. 'monothematic' user. For example, $focus(user) = 0.8$ means that 80% of the user's attention is focused on one particular topic and other 20% are scattered over 99 remaining topics. We have fitted the threshold to extract 'monothematic' users out of 16,124. Fig. 5 shows the number of users extracted and their average *BAR* score depending on the threshold. With the threshold higher than 0.5 extracted users amount becomes insignificant to analyze (less than 200). One can notice that average user BAR reaches a maximum value with the threshold equal to 0.35 and the corresponding user set still has meaningful size 795. Therefore our following discussion considers only this set of 'monothematic' people.

pregnancy, gynecologist, menstruation, test, pills	12	**27.07**
baby, month, breast-feed, milk, pediatrician	2	**23.67**
diabetes, gland, sugar, hormone, normal	9	**22.90**
tablespoon, juice(of the plant), herb, glass, leaf	**124**	15.22
provisions, diet, vegetables, meat, fish	**41**	18.36
brain, abnormality, head, condition, disorder	**32**	15.70

TABLE II. TOP-3 TOPICS BY THE AVERAGE USER BAR (UPPER HALF) AND BY THE AMOUNT OF USERS *focused* ON THEM (BOTTOM HALF)

Medical topics where a user could naturally advise a treatment or medication was of our particular interest. For example, pregnancy, influenza, or traditional remedies topics are the case, but symptoms or clinics are not. Out of 100 topics we got 57 ones of our interest. Having 795 focused users one could investigate which of their main topics are most popular and which of the topics have most successful answerers in terms of a user *BAR*. Table II shows Top-3 topics of both kinds.

C. Medical terms as a cue of a medical professional

The important feature of the asking questions about people' health is the trust in the medical information user aims to get. Answering the medical questions the CQA users do not certify themselves as physicians. So what can an asker use to estimate the veracity of answers she or he gets? In this section we propose a method which helps to distinguish the people, answering like medical professionals, among the CQA answerers. The method of finding journalists on Twitter which has been introduced in [22] is aimed to address the similar problem but it is not suitable for the CQA data because there are no connections between the CQA users by the service design (as the Twitter users' following or retweeting).

People discussing questions related to their profession are often mentioning the field slang. So one more way to find medical professionals is to detect special medical terms in a user talk. First we need to compile a dictionary of special medical words which are normally unknown to a layperson but

every doctor have to know them. The idea is to extract CQA answerers actively mentioning terms from the dictionary. We use two evaluation metrics to test our approach: an average user BAR and a percentage of doctors among users extracted. To do this, we manually gathered 133 users who reported themselves as doctors from our data set and joined them with surveyed CQA users which identified themselves as having medical degree (section IV-A) — 216 users at all.

1) Mining self-reported doctors: Answering question a CQA user is able to specify his/her answer source in a plain text field. Inside *Medicine and Health* category people usually specify a medical handbook name or link to a Web page (usually a Wikipedia page). Besides that users can place there a kind of personal information like *My own experience* or *I have asked my mom*. A particular case is when a user profession is mentioned in the field. We manually processed the list of most frequent answer sources of *Medicine and Health* answerers and extracted medical profession mentions, for example *Im pediatrician myself*; *I have a medical degree*; *Doctor*; *I am a psychotherapist*; *I am a retired physician*. It gave us 133 users who reported themselves as a medical specialists at least once. As the answer source field is not required to fill in, giving wrong information does not make sense, therefore we perceive these people as medical professionals and use them to partially evaluate our approach.

2) Medical terms dictionary: We built a list of medical terms in two steps: 1) medical domain word list mining; 2) cleaning the list out of common items. The starting point for the mining step was data gathered from the Registry of Medicine (http://rlsnet.ru) and the State Register of Approved Drugs (http://grls.rosminzdrav.ru/). We used disease and medicine dictionaries compiled from these resources in our previous study [2]. In addition, the Paramedic Guide [25] was taken because it contains information on the modern clinical medicine: disease descriptions, symptoms, diagnostic methods, and treatments. Since the information gathered was full of common words we used the Russian National Corpus (http://ruscorpora.ru/en/index.html) frequency dictionary (words with counts above 3) to clean it. Along with common words like *disease, cough, pill, advice* there were some diseases and medications removed, for example, *runny nose, allergy, hydrogen peroxide, aspirin*. Finally, the list contained 3,844 highly specialized medical terms.

3) Method discussion: We consider a *Medicine and Health* category's user as a medical professional if her/his answers contain more distinct medical terms than some predefined threshold. As the threshold increases from 5 to 20 the extracted set size decreases accordingly from 1,646 to 322 out of 127,602 answerers in the category which is demonstrated in the Fig. 6. Also the figure shows the relation between the medical professionals extracted and the focused users: as the threshold grows, the percentage of 'monothematic' users (black points) increases as well. People mentioning many special medical terms and concentrated on a small number of topics are presumably medical professionals (or are similar to them). So, these two weak independent features in conjunction could serve a good indicator of a medical professional.

The evaluation of the approach is demonstrated in the Fig. 6 (white points): similarly to the 'monothematic' people, the percentage of the self-reported doctors grows with the

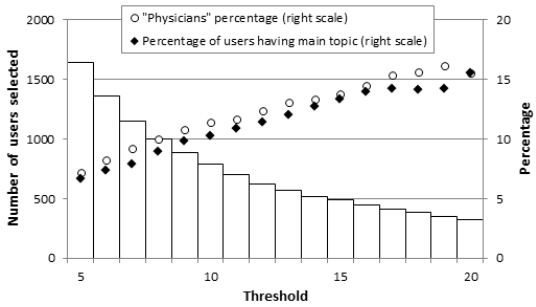

Fig. 6. Medical professionals extraction by means of the medical terms dictionary. Bars (left scale) denote the amounts of people extracted

threshold increase. Second part of evaluation can be seen in the Fig. 7 (black diamonds): the average BAR value of extracted medical professionals. Although it shows a slight increase, the growth rate is negligible. This small but stable increase could be explained by the fact that even small threshold value extracts users of an acceptable level.

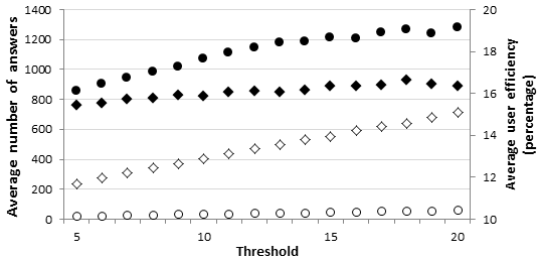

Fig. 7. The user BAR (black points; right scale) and the number of answers (white points; left scale) averaged by the users extracted. The users have been extracting from the whole data set (circles) and from the *Medicine and Health* category (diamonds)

An interesting feature of the special medical terms dictionary is that we do not need to be limited only *Medicine and Health* category: the dictionary imposes natural limits on an answerer topics. Thus we tried to extract medical professionals from the whole data set and to compare the average BAR between corresponding sets: extracted from the whole corpus and from the *Medicine and Health* category (black circles and diamonds in the Fig. 7). Size of a set extracted from the whole data was in average 1.77 times the size of a set extracted from *Medicine and Health* category. One can see that the average BAR of sets extracted from the whole collection outperforms the metric for sets extracted from *Medicine and Health* category. Such a behavior looks strange because we expect users answering in medical categories to be more specialized (and therefore to have higher BAR) than all users over the whole data. So, we hypothesized that it could be a user BAR measure drawback. Recall that a user BAR is a percentage of 'best' user answers among all her/his answers. As user have only one answer and this answer was chosen as 'best', user BAR value would be equal to 100%. At the same time, a user who gave 100 answers needs to have all 100 answers chosen as 'best' to retain 100% BAR. So, the BAR value of a user having a significant number of answers costs more than the value of a user answered only once. In the Fig. 7 (white circles and diamonds) we show an average number of answers among the users in an extracted set. One

can see that users extracted from *Medicine and Health* category (white diamonds) have significantly more answers than users extracted from the whole data set, therefore *Medicine and Health* answerers average BAR is more 'valuable'.

Although the method looks promising it still has some limitations. As the whole Medicine field has many subdivisions it is a difficult task to ensure a high coverage of the medical terms dictionary. Thus the method ignores medical professionals using medical terms which are not belonging to the dictionary. One more limitation is our hypothesis weakness. We assumed that people discussing questions related to their profession tend to mention terms which are special in the field. This assumption may be questioned because people communicating in the Web are free to choose their lexis. Some doctors in an effort to convey information to people far from medicine, deliberately do not use professional terms in their answers. Such users could potentially be even better than the users we extract, as their answers are clear to a layperson, but the method is not suitable to find them.

V. Conclusion

Analysis of the activity and behavior of CQA users in the *Medicine and Health* category shows that the domain is mostly informational, with clear distinction of askers' and answerers' groups. Despite the fact that medical professionals in Russia are rather skeptical about online health forums, the share of users with professional background among the active answerers is surprisingly high. The motivation of the latter are mainly altruistic: they are interested in helping others and share their knowledge and experience.

We conducted experiments on finding users focused on specific medical subjects and users with some evidence of professional medical background. The results demonstrate that the use of medical vocabulary correlates with higher ratio of 'best' answers and self-reported doctors, though this approach cannot deliver a high recall. Topic modeling allows finding a groups of users, who answer predominantly on a certain topic (not necessarily using narrow-domain terminology). The proposed methods – topic modeling and professional lexicons – can be applied to an expert/professional search task, help model community in and work towards CQA service improvement and ensuring content quality.

VI. Acknowledgement

This research was supported by the Russian Foundation for Basic Research, project #14-07-00589 "Data Analysis and User Modelling in Narrow-Domain Social Media". We also appreciate Mail.Ru for granting us access to the data, "Doctor na rabote" service for publishing the survey on their site and medical professionals who volunteered to answer the survey questions.

References

[1] A. Beloborodov, A. Kuznetsov, P. Braslavski, "Characterizing health-related community question answering", *in Proc. ECIR'13 Conf.*, 2013, pp. 680-683.

[2] A. Beloborodov, P. Braslavski, M. Driker, "Towards automatic evaluation of health-related CQA data", *in Proc. CLEF'14 Conf.*, 2014, pp. 7-18.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, "Finding high-quality content in social media", *in Proc. WSDM'2008 Conf.*, 2008, pp. 183-194.

[4] C. Shah, J. Pomerantz, "Evaluating and predicting answer quality in community QA", *in Proc. SIGIR'2010 Conf.*, 2010, pp. 411-418.

[5] Q. Liu, E. Agichtein, G. Dror, E. Gabrilovich, Y. Maarek, D. Pelleg, I. Szpektor, "Predicting web searcher satisfaction with existing community-based answers", *in Proc. SIGIR'2011 Conf.*, 2011, pp. 415-424.

[6] M. Surdeanu, M. Ciaramita, H. Zaragoza, "Learning to rank answers to non-factoid questions from web collections", *Computational Linguistics*, vol. 37, 2001, pp. 351-383.

[7] E. Agichtein, Y. Liu, J. Bian, "Modeling information-seeker satisfaction in community question answering", *in TKDD*, vol. 3, 2009, pp. 10.

[8] E. Mendes Rodrigues, N. Milic-Frayling, "Socializing or knowledge sharing?: characterizing social intent in community question answering", *in Proc. CIKM'2009 Conf.*, 2009, pp. 1127-1136.

[9] F. Harper, D. Moy, J. Konstan, "Facts or friends?: distinguishing informational and conversational questions in social Q&A sites", *in Proc. CHI'2009 Conf.*, 2009, pp. 759-768.

[10] D. Raban, F. Harper, "Motivations for answering questions online", *New media and innovative technologies*, vol. 73, 2008.

[11] D. Dearman, K. Truong, "Why users of Yahoo! Answers do not answer questions", *in Proc. CHI'2010 Conf.*, 2010, pp. 329-332.

[12] D. Pelleg, E. Yom-Tov, Y. Maarek, "Can you believe an anonymous contributor? On truthfulness in Yahoo! Answers", *PASSAT/SocialCom*, 2012, pp. 411-420.

[13] G. Zhou, S. Lai, K. Liu, J. Zhao, "Topic-sensitive probabilistic model for expert finding in question answer communities", *in Proc. CIKM'2012 Conf.*, 2012, pp. 1662-1666.

[14] A. Pal, J. Konstan, "Expert identification in community question answering: exploring question selection bias", *in Proc. CIKM'2010 Conf.*, 2010, pp. 1505-1508.

[15] J. Guo, S. Xu, S. Bao, Y. Yu, "Tapping on the potential of Q&A community by recommending answer providers", *in Proc. CIKM'2008 Conf.*, 2008, pp. 921-930.

[16] Y. Zhang, "Contextualizing consumer health information searching: an analysis of questions in a social Q&A community", *in Proc. IHI'2010 Conf.*, 2010, pp. 210-219.

[17] S. Kim, T. Pinkerton, N. Ganesh, "Assessment of H1N1 questions and answers posted on the Web", *American Journal of Infection Control*, vol. 40, 2012, pp. 211-217.

[18] S. Oh, "The Characteristics and Motivations of Health Answerers for Sharing Information, Knowledge, and Experiences in Online Environments", *JASIST*, vol. 63, 2012, pp. 543-557.

[19] S. Oh, A. Worrall, "Health answer quality evaluation by librarians, nurses, and users in social Q&A", *Library & Information Science Research*, vol. 35, 2013, pp. 288-298.

[20] E. Sillence, C. Hardy, P. Briggs, "Why don't we trust health websites that help us help each other?: An analysis of online peer-to-peer healthcare", *in Proc. WebSci'13 Conf.*, 2013, pp. 396-404.

[21] A.T. Hadgu, R. Jäschke, "Identifying and analyzing researchers on Twitter", *in Proc. WebSci'14 Conf.*, 2014, pp. 23-32.

[22] M. Bagdouri, D. Oard, "Profession-based person search in microblogs: using seed sets to find journalists", *in Proc. CIKM'2015 Conf.*, 2015, pp. 593-602.

[23] D. Blei, A. Ng, M. Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, 2003, pp. 993-1022.

[24] X. Phan, C. Nguyen, "GibbsLDA++: AC/C++ implementation of latent Dirichlet allocation (LDA)", *Tech. rep.*, 2007.

[25] G.Yu. Lazareva, *Spravochnik feldshera (Paramedical Guide)*, MintRight Inc, 2009.