Characterizing Health-Related Community Question Answering

Alexander Beloborodov[1], Artem Kuznetsov[1], Pavel Braslavski[1,2]

[1] Ural Federal University, Institute of Mathematics and Computer Science, Russia
valan89@gmail.com
[2] Kontur Labs, Yekaterinburg, Russia
pb@kontur.ru

Abstract. Our ongoing project is aimed at improving information access to narrow-domain collections of questions and answers. This poster demonstrates how out-of-the-box tools and domain dictionaries can be applied to community question answering (CQA) content in health domain. This approach can be used to improve user interfaces and search over CQA data, as well as to evaluate content quality. The study is a first-time use of a sizable dataset from the Russian CQA site Otvety@Mail.Ru.

Keywords: community question answering, CQA, consumer health information, content analysis, latent Dirichlet allocation, LDA, Otvety@Mail.Ru

1 Introduction

According to a 2009 survey, 61% of American adults look for health information online [2]. A recent study reports that 55% of Russian adults do not go to the doctor if they are indisposed; in case of self-treatment 32% seek advice from friends and acquaintances or search information on the Web [4]. Community question answering (CQA) is one of the major destinations for health-related inquiries. Vast amounts of data collected by the CQA sites allow for re-using the "wisdom of crowds" [3].

Our study focuses on questions and answers on health and medicine. This topic is highly exemplary for CQA: search context (e.g. age, gender, or weight of the person the information is sought for) is important; ideally, the answerer has practical experience with the topic; users prefer a personalized answer. The quality of user-generated content (UGC) is essential for answers in the *Health* category.

Recent studies on health-related CQA data have relied on manual processing of small samples [5], [7]. An approach close to ours is described in [6]: topic modeling is applied to Twitter data in health domain. In our study we use latent Dirichlet allocation (LDA), domain dictionaries, and exploit question-answer structure of the pages to characterize the content. The approach can contribute to a better understanding and representation of CQA data, improved focused search and user interfaces, as well as content quality evaluation on a larger scale. The dataset used in the research comes from a popular Russian CQA site Otvety@Mail.Ru (http://otvet.mail.ru).

adfa, p. 1, 2011.

2 Data

Otvety@Mail.Ru is a Russian counterpart of Yahoo! Answers (http://answers.yahoo.com/) with similar rules and incentives. The site was launched in 2006 and has accumulated almost 80 million questions and more than 400 million answers by August 2012.[1] The most remarkable difference from Yahoo! Answers is the two-level directory used at Otvety@Mail.Ru. The users have to assign their questions to a second-level category using drop-down lists; no hints are provided.

Our data set contains all questions and corresponding answers from the *Health and Beauty* category from 1 April 2011 to 31 March 2012. The content is quite diverse, covering such subtopics as *Tanning*, *Manicure & Pedicure*, *Beauty Salons*, *Bath & Massage*, *Weight Correction*, etc. The total number of questions in the dataset is 313,101. We focus on the largest *Diseases and Medicine* subcategory that contains 95,002 (30.4%) questions. 133,163 unique users were active in the subcategory during the year (i.e. asked and answered questions). 74,760 (56.1%) of them have public profiles; age is indicated in 49.6% of public profiles, and location – in 44.2% cases (e.g. there are 3,004 users from Moscow region). 50.0% of public profiles are female, 33.5% – male, 16.4% – undefined.

3 Results

In this section we briefly describe two approaches we used for data processing: uncovering topics in the collection using LDA and detecting question type based on question-answer content.

We applied GibbsLDA++[2], an implementation of LDA, to discover topical structure of the collection. (In this case, a document refers to a concatenation of a question and all its answers.) We ran LDA with 100 topics and default parameters ($\alpha=0.5$; $\beta=0.1$). The most of resulting topics appeared quite meaningful. Out of 100 topics we discarded 29 topics represented by stop-words, digits, and general terms. Table 1 shows some valid topics.

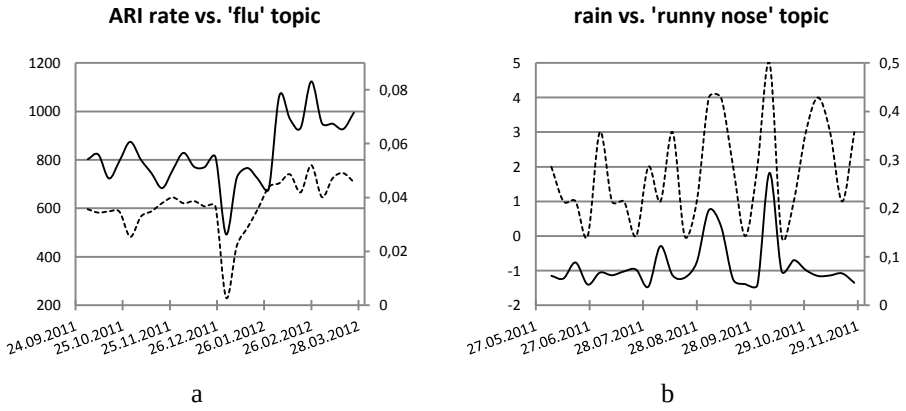
To validate the obtained distributions, we compared dynamics of some topics with infections outbreaks and weather conditions. Figure 2a shows weekly Acute Respiratory Infection (ARI) rates for Russia from WHO/Europe influenza surveillance[3] against the share of documents with a high probability of the "flu" topic (the first column in Table 1). Figure 2b juxtaposes the weekly share of "runny nose" threads (the second column in Table 1) started by Moscow inhabitants vs. rainy days count in Russia's capital[4]. The charts demonstrate an acceptable (given the data volume) association between the extracted topics and real-life events.

---

1 http://otvet.mail.ru/news/#hbd2012
2 http://gibbslda.sourceforge.net/
3 http://euroflu.org/
4 http://www.gismeteo.ru/

Table 1. Selected topics produced by LDA (top10 terms, originally in Russian)

Fever	nose	cough	hormone	cancer	liver
37	runny	lung	gland	tumor	gall
5	drop	bronchitis	endocrinologist	cell	diet
38	sinusitis	pneumonia	hormonal	stage	bladder
flu	ENT	asthma	organism	case	pancreatic
cold *n*	wash	dry	thyroid *n*	dangerous	organ
rise	breathe	phlegm	malfunction	oncology	ultrasonic
body	snivel	breathe	problem	location	pancreatitis
organism	mucosa	syrup	thyroid *adj*	mole	acute
high	sinus	breath	influence *v*	even	chronic



a b

Fig. 1. (a) ARI per 100,000 population (dashed line, left axis) vs. 'flu' topic (solid line, right axis); **(b)** rainy days in Moscow (dashed line, left axis) vs. 'runny nose' questions asked by the users from Moscow region (solid line, right axis)

Question and answer parts of the CQA pages allow us to detect different question types analogous to *evidence-directed* and *hypothesis-directed* queries in the Web health search [1]. For example, *hypothesis-directed* search intent can be associated with a template "disease in question – therapy in answers". To detect these questions we used a list of 1,049 diseases compiled from a reference book for medical assistants and the Russian State Register of Approved Drugs[5] (11,926 unique trade names effective September 2012). Since complex medicine names and diseases are often misspelled, we implemented a fuzzy search based on character trigrams with a subsequent Levenstein distance check with length-dependent threshold. 15,415 (16.2%) pages in the dataset contain at least one pair of this kind. Table 2 shows some disease-medicine pairs along with their frequencies.

[5] http://grls.rosminzdrav.ru/

Table 2. Sample disease-medicine pairs presented in the Otvety@Mail.Ru dataset (originally in Russian; asterisks designate drugs with the same active ingredients)

thrush		**angina**		**herpes**	
flucostat*	155	iodine	130	aciclovir***	307
candid**	92	chamomile	127	zovirax***	138
clotrimazole**	89	nitrofural	111	wax	95
fluconazole*	89	lugol	93	fenistil	41
diflucan*	77	salvia	70	valtrex	34

4 Conclusions and Future Work

Our study shows that even a "light" incorporation of domain semantics into CQA analysis can significantly improve understanding of the data. We plan to apply the tested approach to focused health search and representation of the collected data.

We also plan to develop and refine the proposed method. Our plan includes a large-scale quality evaluation of the health-related CQA data. To perform the evaluation, we will use disease classification along with the list of the drugs recommended for each disease. Another direction for future research is to investigate users' follow-up questions similarly to Web search query sessions.

Acknowledgements. We thank Mail.Ru and Maxim Babich personally for granting us access to the data.

References

1. Cartright, M.-A., White, R. W., Horvitz, E.: Intentions and Attention in Exploratory Health Search. In: Proceedings of SIGIR'11, pp. 65–74 (2011).
2. Fox, S., Jones, S.: The social life of health information, http://www.pewinternet.org/~/media/ Files/Reports/2009/PIP_Health_2009.pdf
3. Liu, Q., Agichtein, E., Dror, G., Gabrilovich, E., Maarek, Y., Pelleg, D., Szpektor, I.: Predicting Web Searcher Satisfaction with Existing Community-Based Answers. In: Proceedings of SIGIR'11, pp. 415-424 (2011)
4. O Samolechenii i Reklame Lekarstvennykh Preparatov (On Self-Treatment and Drug Advertising), 18.06.2012, http://fom.ru/obshchestvo/10489
5. Oh, S., Worrall, A., Yi, Y. J.: Quality Evaluation of Health Answers in Yahoo! Answers: A Comparison between Experts and Users. In: Proceedings of the American Society for Information Science and Technology, vol. 48(1), pp. 1–3 (2011)
6. Paul, M., Dredze, M.: You Are What You Tweet: Analyzing Twitter for Public Health. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 265–272 (2011)
7. Zhang, Y.: Contextualizing Consumer Health Information Searching: an Analysis of Questions in a Social Q&A Community. In: Proceedings of the 1st ACM International Health Informatics Symposium (IHI '10), pp. 210–219 (2010)