Analysis of user activities on popular medical forums

M V Kamalov\textsuperscript{1,2}, V Y Dobrynin\textsuperscript{1}, Y E Balykina\textsuperscript{1} and R S Martynov\textsuperscript{1,2}

\textsuperscript{1} Applied Math and Control Processes Department, Saint Petersburg State University, Saint Petersburg 198504, Russia
\textsuperscript{2} EPAM systems, Saint Petersburg 196084, Russia

E-mail: mkamalovv@gmail.com

Abstract. The paper is devoted to detailed investigation of users’ behavior and level of expertise on online medical forums. Two popular forums were analyzed in terms of presence of experts who answer health related questions and participate in discussions. This study provides insight into the quality of medical information that one can get from the web resources, and also illustrates relationship between approved medical experts and popular authors of the considered forums. During experiments several machine learning and natural language processing methods were evaluated against to available web content to get further understanding of structure and distribution of information about medicine available online nowadays. As a result of this study the hypothesis of existing correlation between approved medical experts and popular authors has been rejected.

1. Introduction

Evaluating quality of available information was always a hard to solve problem, especially nowadays, when the data grows with incredible velocity. Content, published on the web by users, always requires verification, especially if it considers such a field as healthcare. Experts in this area could provide an incredibly useful information, but there are also people with lower level of expertise, whose incorrect interpretation or formulation of some concept could lead to a real damage to someone’s health. Moreover, an instrument which would allow to divide experts from others could be useful for research purposes, such as finding special patterns in language, sentiment analysis, etc.

There were several approaches to similar problems. Researches around the world studied users’ behavior online in terms of social grouping \cite{1}, community-based methods \cite{2} and clicks and ratings in search engines and web stores. Another subject was analysis of interconnections between unstructured data on the web (discussion forums and well-known sources of information) \cite{3}. Also, there are studies in health-related information found on the Internet: log \cite{4} and search queries analysis aimed at evaluating users’ preferences and expertise \cite{5}. By the way existing researches which focused on analysing abstracts of medical studies from the perspective of evidence based medicine \cite{6, 7}. This article focused on quality evaluation of conducted medical studies by analysing text abstracts. But most of them provides either too general information without deep insights to specific semantic areas or operates on highly structured data (e.g. NCBI PubMed database). This paper presents deep analysis of medicine-related content based on unstructured information from online forums.
2. Data description
We used content from two popular health-related forums: doctorslounge.com and patient.info. We extracted the data using web crawling instruments, saving posts text, information about authors and hierarchical structure (forum categories) in terms of each forum labeling system. The overall data is shown in Table 1.

|                        | doctorslounge.com | patient.info |
|------------------------|-------------------|--------------|
| Total amount of data   | 143921            | 1476990      |
| (number of posts)      |                   |              |
| Number of top level    | 22                | 31           |
| categories             |                   |              |
| Number of second level | 95                | 325          |
| categories             |                   |              |
| Number of specific     | 25843             | 157754       |
| topics (“branches” of  |                   |              |
| discussion)            |                   |              |

Further exploration was based on the fact that the doctorslounge.com forum has labels for experts in medicine area – it is a highly moderated forum with authentication procedure, while patient.info provides only information about “popularity” of users. That means, that in patients.info forum there is a system of levels of expertise (there are six of them), which is based on users’ activity and social approval. Based on that data we formulated the following research questions:

**RQ1:** Is there a correlation between level of user expertise from patients.info and medical doctors from doctorslounge.com?

**RQ3:** Does the forum internal structure affect expert distribution and the amount of produced content?

To answer these questions, we first looked at the existing distribution of users with respect to their labels (figure 1, figure 2). In the doctorslounge.com data there were more specific labels for experts (such as “Paramedic” and “Psychiatrist”), so we merged them into one label “Medical Doctor”, which was already presented. In the patients.info data users were divided into 6 levels which meant rank of expertise and popularity of author on this forum. Each level has its own identification labels: level 1 – beginner, level 2 – Contributor, level 3 – Regular, level 4 – Accomplished, level 5 – Guru, level 6 – Oracle.

![Figure 1](attachment:image1.png)  
**Figure 1.** Users’ post count by label on doctorslounge.com data

![Figure 2](attachment:image2.png)  
**Figure 2.** Users’ post count by label on patient.info data
Based on this data we focused on the main hypothesis about difference between groups of users was that they language, i.e. words and collocations, would differ a lot. To test this, we first built a classification model based on doctorslounge.com data.

The whole text corpus (posts, titles) was preprocessed for evaluation: all words were transferred to lower case, stemmed using Porter stemmer, and all stop words were removed. Then a TF-IDF representation of processed words and user labels as target values was used as an input to classification algorithm (considering sparsity of data we used logistic regression). Quality of the model was checked on 10-fold cross validation (figure 3) in terms of F1-score and AUC-ROC – two well-known metrics used in classification tasks. As one can see, it fitted existing data.

Then the predictions were made on the patient.info data, preprocessed in the same way. After that we had labels for each posts determining whether it was written by an expert (or “Medical doctor” in terms of doctorslounge.com labeling) or not.

![Performance of logistic regression](image)

**Figure 3.** Performance metrics of classification model

3. **Labels relationship**

To answer the RQ1, we studied the distribution of newly labeled experts over existing level system (figure 4). As one can see, the distribution is almost uniform in a sense that each level contains almost equal proportion of experts (numbers inside bars). That means that there is only a small correlation between real expertise and social-based leveling in the data.

![Experts’ distribution over levels in patient.info data](image)

**Figure 4.** Experts’ distribution over levels in patient.info data
Other approach was to compare the vocabulary of experts and less experienced users in order to find out, whether there is any strong dependency. To do that we built a distribution of words from each class over user labels. These distributions allow us to measure the distance between them in terms of Jensen-Shannon divergence, which corresponds to similarity between vocabularies and overall language use in classes. Jensen-Shannon divergence is defined as follows:

\[
\text{JSD}(P \parallel Q) = 0.5 \times \text{D}(P \parallel M) + 0.5 \times \text{D}(Q \parallel M),
\]

where \(M = 0.5 \times (P + Q)\) and \(\text{D}(PM) = \sum_i p(i) \log \frac{p(i)}{q(i)}\) – the Kullback-Leibler divergence.

We used a distribution of words over “Medical Doctor” label \(P_{MD}\) based on doctorslounge.com data as a base distribution and measured distance from it to all others. Results could be found in Table 2. We also calculated the divergence value between \(P_{MD}\) and distributions of words of users that our model labeled as experts as an additional sanity check for classification algorithm (if our classification is correct, the divergence value between those two distributions must be relatively small).

| Label corresponding to distribution \(P_l\) | \(\text{JSD}(P_l \parallel P_{MD})\) |
|------------------------------------------|----------------------------------|
| Guest                                   | 0.1109                           |
| Level 1                                  | 0.1362                           |
| Level 2                                  | 0.1354                           |
| Level 3                                  | 0.1397                           |
| Level 4                                  | 0.1408                           |
| Level 5                                  | 0.1419                           |
| Level 6                                  | 0.1251                           |
| Predicted expert                         | 0.0956                           |

The values are very close to each other for the level-based labeling system (except for Level 6), so we can not rely on that if we want to divide users by their expertise. The values also suggest that there is no strong correlation between vocabularies of users at each level and experts.

To further investigate possible dependency we decided to look at experts distributions over existing top 5 patiens.info forum categories which contain more classified users as medical doctors than others (figure 5).

![Figure 5. Medical Doctors over patiens.info category](image-url)
Based on the results presented in figure 5 we have not found a category which contains more classified medical doctors than others. It means that distribution of medical doctors over forum category looks like uniform. Thus, these results can not completely answer the RQ2 question. However more detailed representation of classified medical doctors over levels of expertise is presented in figure 6 and 7 for categories “infections” and “injury”. This representation allow us to notice that the level 6 contain more doctors than others.

![Figure 6. Medical doctors present over levels in “Infections” category](image1)

![Figure 7. Medical doctors present over levels in “Injury” category](image2)

However forum category does not completely answer at the questions RQ1, RQ2. Those result led us to use topic modelling approach.

4. Approach based on topic modelling

The topic modelling approach available to extract topics based on the post semantics. These topics describe medical areas in health-related forums. For this approach we applied Latent Dirichlet Allocation (LDA) algorithm which has been approved as the best solver for topic modelling task.

At the first step of topic modelling we preprocessed data for the LDA training and testing set. The pre-processing consist of tokenization, cast to lower case, stemming by Porters’ stemmer, and removing stop words and words which contain less than 3 letters and higher than 10 letters. After that we divided posts for training and testing sets: – 600000 posts for train; – 200000 posts for test. Based on the grid search techniques and using perplexity as an evaluation measure we started the process of topic number selection. The result of grid search is presented in figure 8.

![Figure 8. Perplexity over topics](image3)
Based on the obtained results we can see that optimal amount of topics equals 270. Thus, we get representation of doctors over most popular topics. Under the most popular topics we mean topics which contain a lot of user posts. From this point we got doctor distribution presented in figure 9. All topics in figure 9 is represented as a numerical index (id) which has been obtained from gensim lda\(^1\). This topics has been selected as containing more Medical Doctors than others.

![Figure 9. Doctor distribution over topics](image)

Each topic index gets its own representation in terms of words distributions. In table 3 we get definition for each topic as words distribution.

**Table 3** Most probably 5 words for topics

| Topic index | Most probably words |
|-------------|---------------------|
| 27          | examine; awar; stool; symptom; asthma |
| 52          | infect; sinu; chronic; color; treatment |
| 67          | disea; cell; allergi; histori; lead |
| 125         | women; abnorm; vagin; determin; symptom |
| 144         | evalu; bone; best; pain; wish |
| 146         | surgeon; opinion; second; schedul; pictur |
| 269         | oral; lesion; amanda; blockag; transmit |

Additionally we prepare more detailed representation of classified medical doctors over levels of expertise is presented in figure 10, 11 for topics 144 and 146.

\(^1\) [https://radimrehurek.com/gensim/index.html](https://radimrehurek.com/gensim/index.html)
This results allow us to notice that the level 6 contain more doctors than others and topic modelling approach can extract semantic categories which has highest probability of a post written by a Medical Doctors. In our case we found topic 146 (surgeon; opinion; second; schedul; pictur) which get more the 90 present users classified as Medicals Doctors.

5. Conclusion
In this article, we presented in-depth analysis of health-related information found on two major forums: doctorslounge.com and patient.info. We carefully studied the level of expertise of users on both of these forums. As the result, we have strong evidence that there is no significant correlation between real expertise and the level of “popularity”, which is based on analysis of forums internal hierarchical structure, experts vocabulary and language use.

We also proposed an approach to identify experts based on the topic modelling. Analysis of topic distribution allowed us to find semantic categories which has the highest probability of a post written by an experts.

For further research it is planned to append in topic modelling information from Medical subject headings (MeSH). We suppose that this structure could increase performance of extracting topics from post and reject uninformative topics. Additionally we start looking methods for extracting special behavioral patterns in content provided by medical experts.

References
[1] Xiaolin Shi, Jun Zhu, Rui Cai and Lei Zhang 2009 User grouping behavior in online forums Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining (KDD 2009)
[2] Rowe M, Fernandez M and Alani H 2013 Modelling and analysis of user behaviour in online communities STCSN E-letter 1(2)
[3] Pendry L and Salvatore J 2015 Individual and social benefits of online discussion forums Computers in human behavior p 211–220
[4] Dogan R, Murray G, Névèol A, and Lu Z 2009 Understanding PubMed® user search behavior through log analysis Database (Oxford)
[5] Illhoi Y and Mosa A 2015 Analysis of PubMed user sessions using a full-day PubMed query log: a comparison of experienced and nonexperienced PubMed users JMI Med Inform
[6] Kolbin A, Kamalov M, Dobrynin V, Balykina Y and Verbitskaya E 2016 Development of a Search Engine For The Medline Database With Search Results Ranking From the Perspective Of Evidence-Based Medicine Value in Health 19(7) A369
[7] Kamalov M et al 2015 Improving data retrieval quality: evidence based medicine perspective International Journal of Risk & Safety in Medicine 27 s1 p 106–107