Exploring Health-related Topics in Online Health Community using Cluster Analysis

Yingjie Lu
Antai College of Economics and Management
Shanghai Jiao Tong University
luyingjie982@163.com

Pengzhu Zhang
Antai College of Economics and Management
Shanghai Jiao Tong University
pzzhang@sjtu.edu.cn

Shasha Deng
Antai College of Economics and Management
Shanghai Jiao Tong University
dengshasha@sjtu.edu.cn

Abstract

Recently patients are increasingly turning to online health community to share their experiences and exchange healthcare knowledge. Exploring hot topics from online health community helps us better understand their needs and interests in health-related knowledge. However, statistical-based topic analysis employed in previous studies is becoming impractical to process the growing large-scale online data. Automatic topic analysis based on document clustering is an alternative approach but usually produce poor results as a result of lack of domain-specific knowledge. So this paper proposes a novel framework for health-related topic analysis using text clustering integrating medical domain-specific knowledge. Experiment results show that adding medical domain-specific features into feature set could achieve significantly better clustering performance than existing methods. In addition, further analysis reveals that there also exist some significant differences about hot topics among different kinds of disease discussion boards.

1. Introduction

Research on health communication has shown that patients are increasingly turning to the Internet for health information and support. A study of US-based cancer patients and their caregivers indicated that 80% of them were interested in health-related information on the Internet, while 65% expressed an interest in online support groups [23]. Especially in recent years, with the advent of some social media services, such as Wikipedia, Facebook, online forums and message boards, patients are more likely to obtain health information and share health experiences on these social media websites [19]. A recent survey shows that 80% of Internet users have looked online for information about any of 15 health topics such as a specific disease or treatment, 34% of them have read someone else’s commentary or experience about health or medical issues on online news group, website, or blog, and 24% of them have consulted online reviews of particular drugs or medical treatments [12].

There are some reasons why patients and their caregivers turn to Internet especially social media services for health information. (1)Patients feel that doctors are too busy to answer their questions [28], and many doctors just tell their patients basic medical information but are not willing to take much time to fully explain the details [10], supported by Tyson’s [27] argument there is a lack of attention to detail in the current doctor-patient relationship. (2)Internet enables patients to take a more active role in making decisions about their health through the use of social support and the ability to explore treatment options [13]. Especially for those patients with chronic diseases, they are more likely to search for online health information in order to be better informed about their illnesses [2]. (3)Convenience and anonymity is another important reason why patients turning to Internet [28]. Patients always expect to obtain health-related knowledge easily and quickly, and they feel not too embarrassed to ask health professionals online or comminute with online members about their conditions [1].

Although different types of social media applications can be used to offer health-related information, online health community is one of the most popular social media services, where patients and their caregivers can share their experiences and exchange their interesting information, meanwhile emotional support and encouragement offered by community members is also important for the patients suffering serious illness and help them cope with this fact a lot better than those who deal with serious diseases by themselves.

An thorough understanding of the interests, motivations and behaviors of these online health consumers could be important for many domains. For the websites providing health-related social media services, a better understanding of how people participated in the online discussions could assist the websites designers and developers to optimize the
human-computer interface, provide personalized tools and functions to facilitate patient engagement and improve the ease of use and social interaction. For the information analysts and researchers, characterizing patient online behaviors could assist them clearly summarize the present situation, reveal existent problem and plan the developmental direction of online health community. More importantly, the study is of great help to the end users themselves involved in online health community, especially for the newcomers. They might find it difficult to immediately know well about the new form of online communication, so health topic analysis enables them to get a sense of what online health community is, help them quickly find the issues they concerned about and make them more easily evolved in the online community and gain valuable information for their health self-education, self-caring and self-management. For these reasons, there are many studies on topic exploration of online health community using different research methods such as survey methods based on questionnaire and statistical content analysis. However, there exist some apparent limitations in the existing studies. So this studies proposed a novel framework of topic analysis based on clustering to explore health-related topics from online health community.

2. Literature Review

Recently many studies focus on online health community and explore health-related topics by using various methods, such as descriptive statistics, and qualitative analysis. Meanwhile, some topic analysis techniques are also widely used to process medical text including user-generated medical text from social media services.

This guideline provides the margins, placement, and print areas. If you hold it and your printed page up to the light, you can easily check your margins to see if your print area fits within the space allowed.

2.1. Topic exploration based on statistical analysis

The topics of online health community conducted by Finn [11] generally indicated two different realms of social support: informational support and emotional support. Informational support is defined as any interaction that deals with informational aid given to or sought from participants, such as ask for or provide information regarding the disease or symptoms, medication and treatments. Emotional support is defined as supportive communication where emotional themes are given and sought within the messages. Such as the expressions of venting or seek sympathetic encouragement, send positive energy, and show compassion or empathy.

In earlier studies, categories or themes of information shared on Internet medical support groups were determined according to the number of people who used the list, and how frequently they posted on it [20]. After that, many surveys were developed to evaluate the use of web-based medical informational resources from different user groups. Some adopted the methods of experimental study, questionnaires and interviews to make a statistics of interesting topics [3]. With the rapid development of some healthcare websites, some case studies were adopted to characterize the health-related messages on a targeted websites by extracting the online messages posted by community members and analyzing their preference to different medical information. Most of them focused on chronic diseases such as Parkinson's disease and common high-mortality diseases such as breast cancer [6][26]. Known from the previous researches, the most frequent themes patients concerned about were prevention and diagnosis, treatment, support, and long-term side effects of treatment. In addition, some interesting issues were uncovered that there existed significant difference about interesting topics from patients with different types of diseases.

2.2. Topic exploration based on text mining for medical text

Previous topic exploration based on text mining for medical text mainly focused on clinical narratives and medical literature. To find the interesting medical literature from MEDLINE biomedical literature database, Lin [22] designed an automatic document clustering framework to divide the retrieved literature into different topical groups and prioritized important literature in each group. To facilitate diabetic patients to find appropriate patient educational materials, Kandula [18] described a method for matching patient education material to patients clinic notes through the use of topic modeling so that relevant education articles could be recommended automatically to the patients. The study by Patterson [24] showed that document clustering is a feasible method to clearly differentiate the types of clinical narratives. The studies above mainly focused on professionally written medical text. The purpose of document clustering used in these medical text is to provide tailored presentation of the relevant medical text and facilitate the users to search. In recent year, however, some user-generated medical text occur in
many social media services including medical weblogs, health Q&A, online health community and so on. Some studies applied text mining techniques into these user-generated medical text to explore the topics that interest online health consumers. For example, Denecke [9] focused on medical weblogs and classified the topics medical weblogs into two types: informative and affective. Brody [5] used text classification based on LDA topic models to detect the salient aspects in online reviews of health professionals. Chen [8] performed cluster analysis to medical posts from three online health community and found the clusters fell into a set of common categories: generic, support, patient-centered, experiential knowledge, treatments or procedures, medications, and condition management. In his study, clustering method was proved available in identifying different types of topic information. However, these user-generated medical text from social media differ significantly from professionally written text, so general document clustering techniques do not produce good effects in distinguishing health-related topics as a result of lack of medical knowledge.

2.3. Topic analysis techniques

Topic analysis techniques can be roughly classified into two types: topic model approach [14] and clustering based approach [16]. Topic models are based upon the assumption that documents are a mixture of topics. Identifying topics is essentially estimating parameters related to the topic mixture and word distribution based on existing corpus using statistical methods. Some well-known topic models has been widely used such as Latent Dirichlet Allocation (LDA) [4] and Probabilistic Latent Semantic Analysis (PLSA) [17]. A clustering approach distinguishes topics based upon the similarity of the words used in documents and groups documents into topic clusters. Broadly, clustering can be classified into three categories: iterative distance-based clustering, hierarchical clustering and probabilistic clustering [29]. Distance-based clustering iteratively move instance from one cluster to another, starting from an initial partition, to optimize on certain criteria, including the classical K-means algorithm. Hierarchical clustering develops a tree structure for clusters and thus enables hierarchical topic representation. Probabilistic clustering models each instance as belonging to each cluster with a probability and resulting clusters are not exclusive of each other, such as EM clustering.

With the rapid development of online health community, many patients with various health conditions and their caregivers participate in online health community and publish their interesting messages related to symptoms, diagnosis, drug, treatment and other topics. It is very meaningful work to understand their interests and their online behaviors. However, there exist some apparent limitations in the previous studies of online health community.

(1) Survey analysis and statistical content analysis require the initial study design of coding schemes to determine topic categories, which may be not consistent with the actual topic categories. More importantly, survey research is always based on a sample of the population, and survey sample are usually self-selected from which the population characteristics cannot be inferred accurately [3]. Statistical content analysis is based on manual annotation, requiring much human effort, which is labor-intensive, expensive, time-consuming and often error-prone. Therefore, when faced with tremendous amounts of online health-related information from online community, it is becoming difficult to employ traditional statistical approaches to explore the health-related topics [21]. Recently text mining and machine learning provide alternative tools to dynamically process large amounts of data available online. However, few studies employ these automated methods to analyze these user-generated health information.

(2) Document clustering is considered to be an effective and practical method to explore theme categories of online health-related data automatically [8]. However, some studies of topic detection in biomedical text found that document clustering based on keyword extraction usually results in low clustering performance, while integrating domain-specific knowledge into textual feature representation could improve the clustering performance [30]. Since health-related messages posted in online community contain much medical knowledge, incorporating the medical domain-specific text features as additional features could enhance the performance of topic exploring significantly.

To address these gaps, we introduce medical domain-specific knowledge into topic clustering analysis to explore health-related topics from online health community and then conduct a case study in order to answer the following research questions:

RQ1: How medical domain-specific knowledge integrate into topic clustering system? Could improved topic analysis methods be used to distinguish different health-related topics more effectively?
RQ2: What can the uncovered topics tell us about the various aspects of online health-related information and what is the difference about interesting topics among different types of disease discussion boards?

4. System Design

In this study, we propose a framework integrating medical domain-specific knowledge to explore health-related topics from online health community. Our research framework consists of three steps: data collection and preprocessing, feature extraction, clustering and topic identification, as shown in Figure 1.

1) Data collection and preprocessing. In the data collection step, we focus on an online health community and choose some specific disease discussion boards as data source. By a Web crawler software Offline Explorer, we can get all the web pages in the discussion boards and then parsed the pages to extract available messages. Further, all the messages are split into sentences because many messages often contain more than one topic and sentence-level clustering could produce better results of topic analysis. Next, some noisy and unreliable data should be filtered by text preprocessing including: stop words removal, word stem, etc.

2) Feature extraction. Document clustering has been widely used to explore medical topics in biomedical informatics, where Vector Space Model (VSM) proposed by Salton [25] is considered to be effective to model text content. In vector space model, text is represented by a vector of terms which are typically key words and phrases. Therefore, we also choose these key terms as part of topic features. Additionally, as mentioned in section 2.1, health-related discussion board messages generally indicated two different realms of social support: informational support and emotional support. So sentiment features and domain-specific features should be incorporated as additional feature dimension to improve clustering performance. So in our studies, we incorporate the following three feature categories: keywords-based features, sentiment features and domain-specific features.

Sentiment features were often used to determine whether a text is objective or subjective and whether a subjective text contains positive or negative sentiments. In this study, sentiment features can effectively measure whether a health-related messages is informative topic or emotional topic. A text's overall sentiment is determined by the sentiments of a group of words or phrases appearing in the text. In order to get sentiment feature scores, we exploit SentiWordNet as lexical resource to compute sentiment polarity scores of terms in a text. SentiWordNet provides for each synset of WordNet a triple of polarity scores (positivity, negativity and objectivity). SentiWordNet has been created automatically by means of a combination of linguistic and statistic classifiers. Now SentiWordNet consists of around 207,000 word-sense pairs or 117,660 synsets and has been used as the lexicon in recent sentiment analysis studies. So our work refers to positivity, negativity and objectivity as sentiment features and incorporates them into health-related topic features.

Topic analysis model integrating domain-specific knowledge is known to significantly improve the performance of topic clustering than domain-independent topic model. So medical domain-specific features are introduced in this study as an additional feature dimension. In previous studies, UMLS Metathesaurus as the world's largest repository of biomedical concepts, are widely used to extract medical terminologies. It consists of 1.7 million biomedical concepts where each concept is assigned to at least one of the 134 semantic types. We just choose the health-related semantic types as domain-specific features, as listed in Table 1.
To score these semantic features, we need to extract health-related terminologies from the messages and compute the word frequency of the terminologies belonging to the same semantic types. MetaMap, a highly configurable program that maps biomedical text to concepts in the UMLS Metathesaurus, is used in this study to obtain these medical terminologies automatically.

(3) Clustering and topic identification. Features extracted from messages are quantified as feature vectors for the inputs of topic clustering. However, these vectors are characterized by high dimensionality, redundancy and high correlation among individual attributes, which is unfavorable for clustering [29]. A typical feature reduction technique in text mining—principal component analysis (PCA) can be used to deal with the problem. PCA assumes that the variance of all the attributes come from a few core factors, namely principal components. PCA estimates these principal components by calculating linear combinations of the attributes that have largest variances. Original attributes are substituted by principal components to reduce the high dimensionality. Interdependency between attributes are also reduced because principal components are orthogonal.

Many clustering methods have been used in previous studies. In this study, we choose EM clustering, a kind of probabilistic clustering approach modeling each instance as having a certain probability of belonging to each cluster. Most clustering algorithms require the expected number of clusters to be specified in advance, which is problematic since clustering is intended to be an unsupervised learning method, while EM clustering could evaluate various numbers of clusters and determine the optimal number of clusters by performing cross validation on different number of clusters. A more detailed description of the algorithm can be found in [29].

To better distinguish different clusters and understand the topics represented by the clusters, some key phrases were selected for topic identification. Key phrases contain medical terminologies extracted by using UMLS lexicon and some n-gram(uni-gram, bi-gram and tri-gram) words with high frequency. Key phrase extraction approach is devised similar to TF-IDF schema. Suppose all the resulted clusters are $C_1, C_2, \ldots, C_n$. For each n-gram terms $w$ in a cluster $C_i$, its score is calculated as

$$f(w; C_i) \cdot \log \frac{1}{\left| \left\{ C_j \mid f(w; C_j) \geq f(w; C_i), j = 1, 2, \ldots, N \right\} \right|}$$

where $f(w; C_i)$ is the frequency of $w$ in cluster $C_i$, and $\left| \{ C_j \mid f(w; C_j) \geq f(w; C_i), j = 1, 2, \ldots, N \} \right|$ is the total number of clusters with the frequency of term $w$ greater than or equal to the term $w$ frequency of the cluster being evaluated.

Key phrases with high scores will be ranked, and then combined with expert opinion, the topics represented by clustered messages can be identified.

5. Experiment

5.1. Research testbed and data collection

In this study, we choose Medhelp.org as our data source. Medhelp.org is one of the most popular online health communities. It consists of over 230 discussion boards covering different diseases. Community members are composed primarily of patients and their caregivers who seek health-related
information, and some individual health professionals who share their experiences and knowledge, in the meantime, a small number of doctors also get involved to offer their expertise and answer the questions from community members. Since its opening in 1994, nearly 3 million threads are posted in the community and the site attracts over 12 million visitors every month.

Next, some representative discussion boards need to be determined. Lung cancer and breast cancer are the most common cancer with high mortality and some studies show that both cancers are of the most common cancers Internet users searching for information about [7]. Diabetes, as one of the most common chronic disease, is also among the most frequently discussed diseases in online healthcare community. So this paper choose lung cancer, breast cancer and diabetes as research subjects and collect the messages posted in these three disease discussion boards(see Table 2). In order to evaluate our approach, we manually annotated the messages and classify them into different semantic categories based on their health-related topics.

5.2. Evaluation Metrics

The evaluation approach with reference to external criteria is used in this study to evaluate the results of the clustering [15]. Based on the pairwise comparison to the pre-specified category of the data set, three commonly used metrics Rand index, Jaccard coefficient and FM(Fowlkes and Mallows) index are utilized for performance evaluation. Let SS be the number of pairs of items belonging to the same cluster and category; SD the number of pairs belonging to the same cluster and different category; DS the number of pairs belonging to different cluster and the same category, and DD the number of pairs belonging to different category and cluster. SS and DD are "good choices", and DS, SD are "bad choices". The three metrics are defined as follows:

\[
\text{Rand statistic: } R = \frac{(SS + DD)}{(SS + SD + DS + DD)} \\
\text{Jaccard Coefficient: } J = \frac{SS}{(SS + SD + DS)} \\
\text{FM index: } FM = \frac{SS}{(\sqrt{SS} + SD) \cdot (\sqrt{SS} + DS)}
\]

5.3. Results

EM Clustering algorithm in our study was run using the Weka package, a popular suite of machine learning software written in Java, developed at the University of Waikato. After preprocessing and principal components analysis, EM clustering is implemented. Weka's implementation of EM provides an option to automatically determine the best number of clusters using 10-fold cross validation. First, the initial number of cluster was set to 1 and data set were divided into ten folds, then calculated the average log-likelihood by executing EM algorithm independently on every fold. As the number of cluster is incremented, the process is repeated until the average log-likelihood stops increasing.

We also verify that EM clustering is progressively finding a better fit for the data by checking that the log-likelihood after each iteration never increases. As can be seen in Figure 2, this value increases after each iteration until EM clustering reaches convergence after several iterations.

![Figure 2. Log-likelihood over different iterations for the experimental data from lung cancer discussion boards](image)

In order to make a better understanding of the topics represented by the clusters, we construct a set of category system and incorporate the clusters with similar semantic types into a more common category. For example, the cluster A related to the information about blood test or glucose test is labeled as laboratory procedure, while the cluster B related to the information about cat scan or MRI is labeled as diagnostic procedure. Both clusters are about the examination of disease, so they are classified into the same category Examination. Another example is that the cluster labeled as self-introductions including such a message I have been diagnosed with type 2 diabetes and the cluster labeled as see a doctor including the message I just visited my Doctor and the news was not good, involve no medical information, so they both are classified into the category Announcement.
Table 3. Key phrases extracted from lung cancer discussion boards

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Label</th>
<th>Key phrases</th>
<th>UMLS semantic types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Announcement</td>
<td>tumor, smoker, lobe, lung cancer, lymph nodes, was diagnosed with, family history, a heavy smoker, right upper lobe, in the lung thank you, hope this helps, be greatly appreciated, thanks so much, would be appreciated, stay positive, god bless, best regards, best wishes, all the best</td>
<td>n/a</td>
</tr>
<tr>
<td>2</td>
<td>Emotional</td>
<td>pain, symptoms, cough, breathless, chest pain, painful, shortness of breath, coughing up blood, short of breath, wheezing, nausea</td>
<td>sosy</td>
</tr>
<tr>
<td>3</td>
<td>Symptom</td>
<td>pneumonia, infection, tuberculosis, bronchitis, asthma, copd, pleural effusion, emphysema, atelectasis, collapsed lung</td>
<td>dsyn, patf</td>
</tr>
<tr>
<td>4</td>
<td>Complication</td>
<td>cat scan, biopsy, x-ray, pet scan, chest x-ray, scans, mri, bronchoscopy, imaging, biopsy needle</td>
<td>diap</td>
</tr>
<tr>
<td>5</td>
<td>Examination</td>
<td>chemo, radiation, chemotherapy, lobectomy, operation, therapy, surgery, removal, radiation therapy, wedge resection</td>
<td>topp</td>
</tr>
<tr>
<td>6</td>
<td>Procedure</td>
<td>silicas, tarceva, morphine, chantix, carboplatin, coumadin, alimta, advil, taxol, dilaudid</td>
<td>phsu</td>
</tr>
</tbody>
</table>

Table 4. Key phrases extracted from breast cancer discussion boards

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Label</th>
<th>Key phrases</th>
<th>UMLS semantic types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Announcement</td>
<td>family history, lymph nodes, worry about, my right breast, i was diagnosed, was diagnosed with, my left breast, had breast cancer thank you, god bless, hope this helps, keep us posted, god bless you, thanks so much, best wishes, good luck, all the best, i really appreciate</td>
<td>n/a</td>
</tr>
<tr>
<td>2</td>
<td>Emotional</td>
<td>biopsy, mammogram, ultrasound, mri, bi-rads, biopsy needle, core biopsy, cat scan, imaging, screening</td>
<td>diap, lbpr</td>
</tr>
<tr>
<td>3</td>
<td>Examination</td>
<td>chemo, radiation, mastectomy, lumpectomy, chemotherapy, implant, removal, operation, radiotherapy, surgical</td>
<td>topp</td>
</tr>
<tr>
<td>4</td>
<td>Procedure</td>
<td>tamoxifen, arimidex, femara, taxol, taxotere, effexor, carboplatin, raloxifene, valium, docetaxel</td>
<td>phsu</td>
</tr>
<tr>
<td>5</td>
<td>Symptom</td>
<td>infection, lymph edema, rash, fibrocystic breast, mastitis, idc, eczema, complex cyst, complex cysts, paget's disease, neuropathy, fibrocystic disease, fibrocystic breast disease</td>
<td>dsyn</td>
</tr>
</tbody>
</table>

Table 5. Key phrases extracted from diabetes discussion boards

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Label</th>
<th>Key phrases</th>
<th>UMLS semantic types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Announcement</td>
<td>was diagnosed, diabetic i, type 2, family history, diagnosed with diabetes, type 2 diabetic, type 1 diabetics, in my family, see a doctor, ask your doctor good luck, thank you, thanks again, greatly appreciated, hope this helps, all the best, thanks in advance, would be appreciated, be greatly appreciated, thanks so much, insulin, lantus, metformin, januvia, glucophage, actos, marihuana, avandia, glipizide, amaryl</td>
<td>n/a</td>
</tr>
<tr>
<td>2</td>
<td>Emotional</td>
<td>hypoglycaemia, low blood sugar, infection, dka, pcos, bgs, coma, kidney disease, obesity, diabetic neuropathy</td>
<td>dsyn, patf</td>
</tr>
<tr>
<td>3</td>
<td>Drug</td>
<td>pain, tired, thirsty, nausea, fatigue, tingling, frequent urination, hungry, sore, dizzy, itchy</td>
<td>sosy</td>
</tr>
<tr>
<td>4</td>
<td>Complication</td>
<td>blood test, fasting test, glucose test, fasting blood sugar, hemoglobin a1c test, glucose tolerance test, cat scan, gits, mri</td>
<td>lbpr, diap</td>
</tr>
<tr>
<td>5</td>
<td>Symptom</td>
<td>infusion, injection, transplant, therapy, dialysis, cde, rx, amputation, insulin injection, ect</td>
<td>topp</td>
</tr>
</tbody>
</table>

808
After incorporating the clusters with similar semantic types, main topic groups were identified and ranked for the three disease discussion boards, as shown in Table 3-5. The names of the topic groups are determined according to extracted key phrases. The cluster labeled as announcement mainly contains members’ self-introductions such as I have been diagnosed with type 2 diabetes and their experiences unrelated to diseases, such as we do and can live a normal life. Next, messages with emotional support frequently appeared in community forum, including the expression of sending others blessings and appreciating for the replies from other members. The remaining clusters obviously correspond to some specific UMLS semantic types, respectively, resulting from some UMLS semantic types selected as part of topic features. The clusters represented by semantic type sosy (Sign or Symptom) are assigned as symptom. The clusters represented by semantic types dsyn (Disease or Syndrome) and patf (Pathologic Function) are assigned as complication. The clusters represented by semantic types diap (Diagnostic Procedure) and lbpr (Laboratory Procedure) are assigned as procedure. The clusters represented by semantic type phsu (Pharmacologic Substance) are labeled as drug.

Table 6. Performance measures using different feature set

<table>
<thead>
<tr>
<th>Disease</th>
<th>Feature Set</th>
<th>Rand</th>
<th>Jaccard</th>
<th>FM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Cancer</td>
<td>F1</td>
<td>0.720</td>
<td>0.243</td>
<td>0.412</td>
</tr>
<tr>
<td></td>
<td>F1+F2</td>
<td>0.790</td>
<td>0.310</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td>F1+F2+F3</td>
<td>0.815</td>
<td>0.344</td>
<td>0.520</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>F1</td>
<td>0.704</td>
<td>0.201</td>
<td>0.351</td>
</tr>
<tr>
<td></td>
<td>F1+F2</td>
<td>0.780</td>
<td>0.301</td>
<td>0.476</td>
</tr>
<tr>
<td></td>
<td>F1+F2+F3</td>
<td>0.792</td>
<td>0.312</td>
<td>0.488</td>
</tr>
<tr>
<td>Diabetes</td>
<td>F1</td>
<td>0.715</td>
<td>0.230</td>
<td>0.394</td>
</tr>
<tr>
<td></td>
<td>F1+F2</td>
<td>0.784</td>
<td>0.258</td>
<td>0.415</td>
</tr>
<tr>
<td></td>
<td>F1+F2+F3</td>
<td>0.801</td>
<td>0.317</td>
<td>0.490</td>
</tr>
</tbody>
</table>

(Note: F1 are keywords-based features, F2 are domain-specific features and F3 are sentiment features)

To examine the effect of adding domain-specific features and sentiment features into feature set, we further evaluate the results of clustering based on different feature set. From Table 6 we can see all three types of measurement values increased as more types of features were incorporated. Specifically, feature set F1+F2 outperformed F1 significantly, indicating that incorporating the medical domain-specific text features as additional features could enhance the performance of topic exploring significantly. The highest measurement values were achieved on the selected feature set F1+F2+F3, indicating that improved topic analysis methods proposed in this study could distinguish different health-related topics more effectively, thus answering research question RQ1.

Figure 3. The distribution of informative and emotional topics

5.4. Analysis and Discussion

The remaining groups except announcement group and emotional support group, are known collectively as informative support groups. From the Figure 3, we can see that overall, there are more informative messages than emotional messages in the three disease discussion boards, indicating that there are more members who seek interesting health information than hope to get emotional support through online community. For different disease discussion boards, informative messages in diabetes discussion boards are less than in both cancer boards while emotional messages are more. One reason may be that diabetes as a chronic disease require long-term treatment, so the members with diabetes need more emotional support, while the members with breast cancer and lung cancer, are at higher risk of death, so they are more likely to seek informative messages helpful for their treatment.

Figure 4. The distribution of five informative topics

Lung Cancer ——— Breast Cancer —— Diabetes
To reveal the relationship between disease types and five informative topics, we used a radar diagram to visualize the results. As seen in Figure 4, the radar diagram indicates that for the patients with different types of diseases, their concern for various health-related topics were fairly different, thus answering research question RQ2. As for the symptom topic and examination topic, they are both hot topics related to disease diagnosis. Members in breast cancer discussion boards are more likely to talk about disease examination such as biopsy, mammogram, while in lung cancer discussion boards, members prefer to concern some symptoms such as chest pain and cough. One possible explanation is that breast tissue are on the body’s surface and people can find abnormality easily. In the meanwhile, there is no obvious other symptoms in the early stage of breast cancer supporting the diagnosis. So suspected breast cancer patients are more likely to go to hospital for clinical examination. Unlike breast cancer, lung cancer has some early symptoms like cough or wheezing, which is similar to some ailments such as fevers and bronchitis. The suspected lung cancer patients don't often think of these symptoms as signs of cancers, resulting in them seldom going to hospital for further examination. So they refer to online community to consult others about their conditions.

From the perspective of treatment, we see that the proportion distribution of drug and procedure topics are obviously distinguishable in three disease discussion boards. Drug topic in diabetes discussion boards has a significantly higher proportion than in the two cancer discussion boards, while procedure topic in diabetes discussion boards has a significantly lower proportion than in the two cancer discussion boards accordingly. The reason is both cancers are considered to be the diseases with high mortality, and are mainly treated with chemotherapy, radiation therapy and surgery. Procedure topic, therefore, is a major concern for cancer patients, while drugs are supplementary treatment for cancer patients, so drug topics take a relatively small proportion. In contrast, diabetes is a common chronic disease and drug treatment is a main treatment for diabetes. The patients with long-term use of anti-diabetic drug prefer to discuss the efficacy and side-effects of these drugs. So the drug topic has a big proportion in diabetes discussion boards.

The complication is another hot topic in online community. The distribution of complication topics help us understand that compared to lung cancer and breast cancer patients have fewer and less-severe complications.

6. Conclusions and future research

With the development of online health community, an thorough understanding of health-topic topics could be important for websites providers, information researchers and end users. In this paper, we proposed a framework based on clustering analysis technique to explore interesting health-related topics from online health community automatically instead of using statistical analysis employed in most previous studies. Integrating medical domain-specific knowledge, we construct a medical topic analysis model and the following case study validates the proposed framework as an effective approach to identify different health-related topics automatically. Several valuable conclusions can be drawn from the experimental results, including that there exists some significant differences about interesting topics among different kinds of disease discussion boards.

The paper also has some limitations that need to be considered further. First, the study has proved that additional medical domain-specific features could enhance the performance of topic exploring significantly, however, there exist some other features which could be used to further improve clustering performance. For example, the messages within a single thread most likely contain the same topics, so these structural features should be considered into feature set in the further research. Second, feature set containing a large number of features and the irrelevant ones could greatly impact clustering performance. However, traditional feature selection algorithms work only for supervised data where class information is available. We would need a lot more research about effective feature selection methods to improve clustering performance. Third, EM clustering is used in our study due to high clustering performance and automatically determining the number of clusters using cross validation. However, the rate of convergence of EM clustering algorithm could be slow and it may not be practical for very large numbers of data Because of the computational complexity. So in future research, we could try some more appropriate clustering approaches and compare their performance in our experiment data. Lastly, we didn't focus on the difference of online participants in this study. Actually, there are some different user groups involved in online health community except patients, such as caregivers and health professionals. Different health-related stakeholders have their respective interesting topic. We also plan to extend our study to find the difference in their needs and interests in the further research.
Acknowledgements

This research was supported by the National Natural Science Foundation of China under Grant 71171131.

7. References


