Making Sense of a Healthcare Forum – Smart keyword and user navigation graphs

Akhil Kumar
Penn State University
University Park, Pennsylvania, USA.
akhil@psu.edu

Kang Zhao
University of Iowa
Iowa City, Iowa, USA.
kangzhao7@gmail.com

Abstract

We describe a visual navigation interface that can organize the relevant information in a forum systematically and provide better pathways for accessing content of a medical forum. The interface consists of a smart keyword navigation graph and a smart user navigation graph. These graphs show the frequently occurring keywords in the forum posts and the users who make the posts, along with the links among them. On clicking a node or a link, one can access the top N posts related to the keywords or users. The paper describes our methodology and its implementation for the posts on an IBS forum.

Keywords: healthcare forums, IBS forum, smart navigation, access paths, keywords, users

Introduction

Consumers are increasing their use of social media to keep abreast of personal health and well-being issues. In addition to web portals (e.g., WebMD.com and MedicineNet.com), which serve as repositories of expert derived health information (Baker et al. 2003; Sillence et al. 2007), consumers are turning in large numbers to medical forums (e.g., DiabetesForum.com and DiabetesDaily.com for diabetes management, IBSGroup.org for IBS, etc.), where communities of people with shared medical issues exchange knowledge and experiences with their illnesses (Tustin 2010). On a large forum with several hundred thousand posts, it is very hard for users to find the most relevant posts just by using the search functions of the forum. Hence, there is a need for smarter visual navigation interfaces that can organize the relevant information more systematically and provide better pathways for accessing content.

Most healthcare forums consist of various topics or threads, where each thread starts with a new post by a user on a certain topic and consists of a series of replies from other users that result from the initial post and lead to a discussion on the topic of the initial post. In this paper we describe a general methodology for developing navigation graphs from a healthcare forum. The navigation graph shows the most frequently occurring keywords related to patient symptoms, diagnostics, diagnoses, treatments, diets and emotions. A user can glance at these keywords and search for relevant posts by clicking on keywords and also on links between pairs of keywords. Further, she may follow several links along a chain of related keywords to discover new relationships between keywords that she did not know existed. In doing so she may find new treatments for a symptom, or the effect of a certain diet on her condition. In addition to building a keyword navigation graph, we also developed a user navigation graph. This graph allows patients to find posts of active users and see how they are connected with other users in a social network, where a link between two user nodes indicates that the users have interacted on the same thread.

To illustrate our approach, we analyzed nearly half a million posts on a forum for a chronic disease called irritable bowel syndrome (IBS) and showed how keyword and user navigation graphs can make it easier for users to find relevant information pertaining to their condition. We decided to focus on IBS because it is a chronic syndrome that affects 25 to 55 million people in the US alone. The medical science is poorly understood, and hence patients rely even more on healthcare forums to seek advice from one another.

Currently, the main way for users to search forums is through a keyword search. While useful, it does not facilitate navigation of a forum. Users have to repeatedly type keywords to get answers to queries. Our goal in this paper is to make the data in a forum easier to navigate through a simple visual interface. Thus, at a glance they can see the frequently occurring keywords and also the relationships among them. A seed keyword can serve as a starting point and the graph can suggest directions for further exploration of the forum in a more efficient and effective way. We briefly describe the conceptual model of a forum.
next, and then turn to describe the keyword and user navigation maps at length in subsequent sections. Later we discuss related work and our plans for developing our vision further before concluding the paper.

**Conceptual Model of a Forum**

A forum can be modeled by a UML diagram as shown in Fig. 1. The main classes in this diagram are Thread, User, Post and Keywords. The keywords are categorized into standard medical subclasses: **symptoms, diagnostics, diagnoses, treatments, diets** and **emotions**. A forum user may place one or more posts in the forum. The post may start a new topic or thread, or it may add to the discussion in a current thread. A post usually refers to keywords that are relevant to the forum. Thus a forum about IBS has keywords like pain, appetite, nausea, diarrhea, colitis, colonoscopy, Gas-X, etc. that are used frequently in the posts. Along with the post id, text and timestamp, the forum keeps information on the userId’s of the users, the date they joined, their city and country, the user title and the user group. For each thread, there is a ThreadID, ThreadURL, date of creation, replies and number of views.

![Figure 1. A UML class diagram for a healthcare forum](image)

The UML class diagram can be stored as relational tables in a database which is populated with actual posts from a forum. The data in the tables is queried and analyzed for creating forum navigation graphs.

**Smart Keyword Navigation Graph**

A keyword navigation graph KNG(N,L) is a graph of the N most frequently occurring keywords in a forum as nodes, and the L most frequent links among them. Some of the top keywords we found in the forum are shown by category in Table 1. The steps in making a KNG graph are:

1. \( K = \) List all (keyword, frequency) pairs for medical terms in a forum
2. \( K_1 = \) Sort \( K \) in descending order by frequency
3. \( K_2 = \) Select the top N keywords from list \( K_1 \)
4. \( P = \) For all pairs of keywords in \( K_2 \), find the frequency of posts with both keywords present in them
5. \( P_1 = \) Sort them in descending order by frequency
6. \( P_2 = \) Select the top L pairs from list \( P_1 \)
7. Make a graph \( G \) of the keywords in list \( K_2 \) as nodes,
8. Connect the nodes in \( G \) with links for every entry in list \( P_2 \)

We extracted all the posts from the general discussion forum at the IBSGroup.org website (with permission) over a period from 1998 to 2012. There were 466,710 posts for 59,822 threads. Then we created the forum navigation graph shown in Figure 2. In this graph we included the top 135 links which...
appeared in approximately 400 or more posts (about .1% of all posts). The highest frequency link is (IBS, pain) with a frequency of 60, 252. The number of keywords is 63. So, this is a KNG(63, 135) graph. The values of N and L are chosen by judgment for now, but more formal methods will be developed later.

Table 1. Top frequent words by category and frequencies in the IBS general discussion forum

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequent words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptoms</td>
<td>Pain (117,118), gas (88,456), colon (51,968), constipation (310,52), bloating (258,65), spasm (22,047)</td>
</tr>
<tr>
<td>Diagnoses</td>
<td>IBS (258,138), urine (39, 128), diarrhea (33,667), bacteria (21,511), intestine (17,386), colitis (7,278)</td>
</tr>
<tr>
<td>Diagnostics</td>
<td>Colonoscopy (18, 713), hypnotherapy (9109), blood test (7979), ultrasound (4697), cbt (4131)</td>
</tr>
<tr>
<td>Diet</td>
<td>Fiber (27,946), lactose (10,890), coffee (8,733), alcohol (7,885), yogurt (5,270), smoking (3,745)</td>
</tr>
<tr>
<td>Treatment</td>
<td>Immodium (24,368), calcium (24,001), probiotic (17,027), antibiotic (15,797), lotronex (14,118)</td>
</tr>
<tr>
<td>Emotions</td>
<td>Feeling (50,269), anxiety (37,830), depression (15,019), agoraphobia (367)</td>
</tr>
</tbody>
</table>

This graph was made clickable by invoking a PHP script on the click of a node or a link, such that the script would in turn invoke a query to a MySQL database to return the top M relevant posts related to the (pair of) words for the node or link. This graph is accessible on the web. Tables 2 and 3 show results for sample node and link queries. These queries run very quickly in real-time, typically in 1 to 2 seconds.
In addition to accessing information for single and pairs of keywords, the graph visual interface facilitates navigation. Thus, a patient may follow a link from ‘vomit’ (a symptom) to ‘anxiety’ (an emotion related to the symptom). The graph shows that there are links between anxiety and pain, nausea, diarrhea, etc. If a patient pursued the link to diarrhea (a diagnosis), there is a further link to Imodium (a treatment for Diarrhea). Thus by following this series of links, a patient can understand that: vomiting is a symptom and diarrhea is a diagnosis/condition associated with the emotion of anxiety; and, Imodium is a treatment for diarrhea. In general, a patient can follow a path along a series of such links that have a cause effect relationship between the nodes on either side. Often a patient may discover an unexpected or counterintuitive relationship, such as the three-way relationship among coffee, spasm and bacteria.

Smart User Navigation Graph

A user navigation graph is a derived social network among users of a forum. When two users post in the same thread, it creates a connection between them. If they do so frequently, it suggests that they have an interest in similar kinds of issues and topics in the forum. A user navigation graph UNG(L) is a graph of the L most frequently occurring links among users in a forum, and the N users among whom the links occur are represented as nodes. The steps in making such a graph are as follows:

1. $T_1 =$ Run a SQL query to extract a table $(U_1, U_2, \text{link weight})$ in descending order by link weight, where link weight is the number of different threads where users $U_1$ and $U_2$ have posted together.

2. $T_2 =$ Select the top L rows in table $T_1$.

3. Make a graph for $T_2$, where users are nodes and a link is shown for a pair of users in each row.

We analyzed the data for the IBS Group discussion forum. There were 19,982 users who contributed the 466,710 posts. Altogether there were more than 4.5 million links between users. However, we focused only on the top 1000 links (spanning 285 nodes) with a link weight of more than 50 (the highest link weight was 1241). These links reflect active users who interact frequently with each other. Preliminary social network analysis on these top 1000 links showed that the subgraph had characteristics of a scale free network with the average degree of 7 and network diameter of 6. It is sparse with a density of 0.025, and the node degree ranges between 1 and 139.

A user navigation graph for only the top 50 links is shown in Figure 3, and it is accessible on the web. There are 29 users in this graph. The nodes and edges of this graph were also made clickable as before. By clicking on a node we can see the top 20 threads (by number of views) where the user has participated. Similarly, by clicking a link one can find the top 20 threads in which both the users on either side of the link have participated. Sample results for these queries are shown in Tables 4 and 5.
The purpose of such navigation is that if a patient finds a certain user’s post interesting, she may wish to see other popular posts of that user. Further, she may navigate to posts of other users who post frequently along with this user.

Discussion and Related work

The need for better navigation methods is prompted by the growing use and perceived value of the Internet, especially of online health communities or support groups for health information. According to a study by the Pew Research Center, 80% of adult Internet users in the U.S. use the Internet for health-related purposes (Fox 2011). Often health professionals fail to meet their patients’ needs fully. Further, having access to information on the internet empowers patients and gives them confidence to ask questions of their providers (Eysenbach 2003). A more recent study (Tustin 2010) of cancer patients showed that in a sample of 178 cancer listserv users, “35% chose the Internet as their preferred source of health information compared with 19% who named their oncologist.” Social media have become increasingly popular as sources of public health information in recent years since they offer the
advantages of low costs, rapid transmission through a wide community, and ease of user interaction (Vance et al. 2009). A study in 2009 revealed that about 5% of all Internet users in the U.S. participated in online support groups (Chou et al. 2009). It is reported that, “patients with diabetes, their family members and their friends use Facebook to share personal clinical information, to request disease-specific guidance and feedback, and to receive emotional support” (Greene et al. 2011).

Despite its many advantages, the use of social media for health information also poses challenges of information overload, poorly organized information, medical jargon, and, simply, difficulties in searching (Cline and Haynes 2001). Our research tries to address this problem by making the navigation of healthcare forums more efficient through the keyword and user navigation graphs. It creates a “big picture” view of possible symptoms, diagnoses and treatments of a disease, and thus helps patients make sense of the gigantic amount of information in an online forum.

Another feature of our keyword navigation graph is that it is generated in a bottom-up way from user-generated contents in a forum. This leads to two advantages compared to semantic networks of medical terms from medical thesauruses (such as UMLS). First, many of the words and phrases in our network are essentially folk vocabularies that are used by forum users. In contrast to professional medical terms, the common terms used in forums are easier to understand and relate to (Smith and Wicks 2008). Second, the navigation graph can provide ideas for query expansion to improve the retrieval of health information. Research has shown that when people search for complicated health information, they tend to use simplistic queries due to their limited knowledge of the medical vocabulary (Zeng et al. 2006). Sometimes these terms do not accurately reflect their information needs and do not form effective queries (Zeng et al. 2004). Thus query expansion techniques, such as those based on term co-occurrence (Minker et al. 1972) and semantic knowledge from medical thesauruses (Göbel et al. 2001; Zeng et al. 2006), have been adopted to improve the quality of information search (Zhang et al. 2008).

We have several ideas for developing our preliminary work further. First, the initial set of “seed” or tier 1 keywords on our navigation map is not exhaustive. A simple extension is to maintain a list of tier 2 keywords in the database. When results from a query are shown, any tier 2 word that appears in the thread title or body can be made clickable just like tier 1 words. Second, our current rankings are based mainly on views. We would like to personalize the ranking algorithm based on past browsing history of a user as well as any personal information (e.g. age, gender, race, symptoms, lab values, etc.) they may offer about their condition in their online profiles. Third, we would like to use such additional information to pose queries to users as pop-ups by including such information in the search terms. Fourth, we can apply topic modeling algorithms to identify broader topics and suggest topically related web queries to users (Li, et al. 2013). Here we can even benefit from previous studies in information retrieval that used language models to match queries and documents (Ponte and Croft 1998). We also hope to leverage methods based on the thread structure (Duan and Zhai 2011; Seo et al. 2009) for health information retrieval from online forums. After implementing some of these ideas in our prototype we intend to conduct a validation study to assess the value provided to patients through better navigation interfaces.

Conclusions

As patients start to depend more on social media and healthcare forums it is becoming more important to help them navigate these resources in better ways than simply by keyword searches. In particular these new methods should be more visual, intuitive and interactive. It is also important that these interfaces help patients personalize the interface by maintaining patient navigation history and other information.

In this paper we described a novel approach for facilitating such navigation by means of keyword and user navigation graphs. These graphs allow patients to click on keywords (or users), and links connecting them, and thus find the most relevant posts related to them.

We illustrated our ideas by analyzing nearly half a million posts on about 60,000 threads or topics from a forum on IBS. Our prototype demo navigation graphs are accessible on the web. This forum was chosen because IBS is a chronic syndrome for which the medical knowledge is still primitive and hence patients rely even more on interaction with other patients in similar conditions as themselves. As this work progresses we hope to develop the navigation features further and make this tool more interactive by giving users the chance to personalize their information and assisting them in asking the right kinds of questions for their condition.
Acknowledgment: We thank Mr. Jeffrey Roberts for giving permission to analyze the IBSGroup forum.

References


