Understanding Participant Behavior Trajectories in Online Health Support Groups Using Automatic Extraction Methods

Miaomiao Wen
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
mwen@cs.cmu.edu

Carolyn Penstein Rosé
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
cprose@cs.cmu.edu

ABSTRACT

This paper presents an automatic analysis method that enables efficient examination of participant behavior trajectories in online communities. This method offers the opportunity to examine behavior over time at a level of granularity that has previously only been possible in small scale case study analyses, and thus complements both existing qualitative and quantitative methodologies. We provide an empirical validation of its performance. We then illustrate how this method offers insights into behavior patterns that enable avoiding faulty oversimplified assumptions about participation, such as that it follows a consistent trend over time. In particular, we use this method to investigate the connection between user behavior and distressful cancer events and demonstrate how this tool could assist in understanding participation trajectories in online medical support communities better so we are better able to design environments that meet the needs of participants.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Computer supported cooperative work.

Keywords

Online support groups, Cancer trajectory, Disease event, Natural language analysis

1. INTRODUCTION

The contribution of this paper is a new automatic analysis method that enables efficient examination of participant behavior trajectories in online communities. We demonstrate how it offers the opportunity to examine behavior over time at a level of granularity that has previously only been possible in small scale case study analyses. Using this tool we are able to offer new insights into the experiences of users in one of the largest online cancer support communities on the Internet. This insights offered by such a tool complement both existing qualitative and quantitative methodologies for studying community behavior patterns.

Online support groups provide a rich and valuable source of data related to chronic illness and the inner workings of social support. A growing number of people who suffer from chronic or life threatening diseases obtain valuable resources from online support groups, which are available anytime in the privacy of one's home [19]. These affordances of online support groups are particularly attractive in the case of stigmatizing illnesses such as AIDS, alcoholism, breast and prostate cancer, which are the topics of many popular online medical support communities [5]. In order to design such environments to maximize benefit to users, it is necessary to understand how the experiences of users of such environments unfold over time.

In this paper we seek to overcome some of the methodological limitations of current approaches to studying online support groups. Quantitative approaches to studying behavior in online communities abstract away from the details of individual users in order to reduce behavior to a small number of variables that may be related to one another statistically. Such a reduction is needed in order to understand the causal mechanisms at work. However, in order to do so in a valid way, it is important to avoid making assumptions that do not hold in practice.

Growing out of a tradition of analysis of threaded discussion forums that consist of a list of threads, each of which roughly corresponds to a topic of discussion, quantitative approaches to modeling participation in online communities typically model that participation in terms of frequency of types of contributions over time, with the idea of identifying increasing or decreasing trends and the reasons for these trends using linear modeling techniques. Our work challenges the underlying assumptions behind such approaches by demonstrating more of a periodicity in participation, centered on important cancer events. This is consistent with other work investigating the importance of key events in a patient's cancer history and their effect on behavior [20, 21].

One role for qualitative analyses of user behavior trajectories in mixed methods approaches is to offer insights that challenge overly simplistic assumptions about participation. However, even such detailed explorations are limited if they can only be conducted on a very small set of users. For example, a case study analysis of the complete posting his-

GROUP'12, October 27–31, 2012, Sanibel Island, Florida, USA.

tory of one participant in an online cancer support forum has been published in prior work [26]. That analysis suggests that frequency of participation was clearly correlated with stress-inducing events. But as a case study, the generalizability of the results is limited and restricted. Mapping how the themes of posts change as patients move from diagnosis, through treatment, and towards recovery or death at a grander scale, may provide valuable insights to inform ways to tailor future psychosocial and educational interventions in online medical support communities according to a patient's cancer event trajectory [25]. However, manually extracting such a cancer trajectory is effort-consuming and time consuming. The goal of our work is to automate such an analysis so that some of the benefits can be obtained without the time and effort. Using this tool, we work to contribute deeper insights towards understanding a patient's psychosocial reactions to important cancer events as they unfold within that patient's disease progression [26].

Motivated by earlier qualitative work [26], we use automatically extracted illness trajectories, and present an analysis of data from an active online community that provides support for a statistical connection between the pattern of participation in online discussion and stress-inducing events. We demonstrate that when the users are undergoing these stressful events, they post more than 2 times as often as in non-event months. The topic of their posts also varies according to the events. We also find that almost half of the long-term users began their participation in the online support community when they were facing some kind of stressful disease event, such as chemotherapy.

In addition to the methodological contribution and analysis, our work makes a technical contribution as well. Many practical applications in Natural Language Processing either require or would greatly benefit from the use of temporal information. For instance, question-answering and summarization systems demand accurate processing of temporal information in order to be useful for answering "when" questions and creating coherent summaries by temporally ordering information. Our technical approach involves development of effective temporal expression extraction in an informal writing genre that poses significantly different challenges than generes that have more typically been the focus of work on temporal expression extraction in the past.

In the remainder of this paper, we first review related work that provides the foundation for our investigation. We then describe the online community that provides the context for our work, as well as the data we extracted from it for our analysis. We then present how we automatically generate and visualize disease trajectories from a user's complete posting history. After validation of our visualization tool, we provide an analysis that illustrates the connection between posting behavior and the generated disease trajectory. The paper concludes with discussion and future work.

2. RELATED WORK

For decades, researchers have attempted to examine the well-being of women with breast cancer as it varies over time. Most of these studies are qualitative analyses and do not examine whether disease phase, such as progression of disease or disease recurrence, influence well-being [11]. On the other side of the spectrum, in quantitative analyses, individuals are frequently grouped together for analysis in ways that gloss over individidual differences between patients, including the specific issues they are dealing with at different times. Richer insights could be gained through analyses that consider the impact of important cancer events within a patient's trajectory.

Qualitative research studies exploring individuals' well-being across phases of disease report that distress peaks occur after diagnosis, during chemotherapy, at the conclusion of adjuvant therapy, 6 months - 1 year after mastectomy, when recurrence is diagnosed, and when the disease is declared terminal [10, 20, 21]. Researchers have constructed these cancer trajectories retrospectively from self-report questionnaires and interview data. Using such a methodology, researchers have identified distinct trajectories of mental and physical functioning over 4 years according to 363 patients' breast cancer experiences [23]. Disease trajectories for chronic illness have been defined more generally as the course of the illness over time as identified by eight disease phases: the period before the illness begins, the diagnostic period, crisis or life-threatening situation, acute illness, in which illness or complications require hospitalization, a stable phase, where illness is controlled, an unstable phase, in which illness is not controlled by a regimen, a progressive or deterioration phase, and dying [4]. This trajectory may inform the design of research about the experience of chronic illnesses such as breast cancer. Moreover, research using questionnaires and interviews do not tell us what we would see if we explored similar questions from the standpoint of what behavior looks like over time within these phases. However, these insights that are possible to glean from interviews or questionnaires are not readily accessible in the raw data traces of online communities that would allow us to go beyond self report and observe how participants respond to their cancer events in real time. The goal of pushing beyond what is possible with existing well established methodologies presents technical challenges, however. The boundaries between the phases identified here are not trivial to automatically extract from the posts.

As online support groups become more and more popular, there are more studies of online support groups [2]. Most qualitative evaluation research to date of online support groups has analyzed postings from a sample of users without relating their message content to the medical background of the patients or their disease trajectory [12, 22]. In one notable exception, researchers suggest that the pattern of online discussion group messages was clearly correlated with the stress-inducing events [26]. But as a case study, the generalizability of the results is limited and restricted. In our work, we want to do similar analysis quantitatively. We draw from prior work that offers the ability to automatically identify themes in discussion behavior in online groups. In particular, Wang and colleagues [24] have derived 20 topics from the forum posts using a technique referred to as Latent Dirichelet Allocation (LDA), which we describe later. This prior work reveals something of the distribution of topics discussed by cancer patients, but leaves open interesting questions about the topics patients talk about during specific periods related to important cancer events.

In order to construct cancer trajectories, we must associate events with points in time by extracting mentions of time points in posts. While much computational work has been done on temporal expression extraction, our own research differs from this previous work in several respects. Previous work has mainly focused on identifying the sepa-

rate timepoints of each event in news text, where multiple disparate events may be described [6, 13, 17]. Newswire text is the primary genre in that work, and that genre is known to include a lot of explicit temporal expressions, which are very different from our online forum corpus. Besides specific use of temporal expressions like "MM/DD/YYYY" or "Oct. 22nd, 2001", our system resolves generic time expressions, especially indexical expressions like "tomorrow", "next Tuesday", "two weeks after my diagnosis", etc., which designate times that are dependent on the time of the post or some referential time point. What is more, we also resolve self-contained time expressions that are special to each user like "at the age of 57", "Today is my third breast cancer anniversary", and "I am six months out of Chemotherapy". As the exact date of each post is known, these expressions could be utilized to infer the illness event times. The language style in online forums is highly informal. Our automatic analysis approach also considers forum specific jargon, slang and nicknames [16].

3. CONTEXT OF RESEARCH AND DATA SET

The data for our investigation was extracted from a large, online cancer support community operated by a nonprofit organization dedicated to providing the most reliable, complete, and up-to-date information about breast cancer. This organization also provides a variety of communication platforms, including discussion boards and chat rooms for patients, family members and caregivers so that all of these stakeholder communities are able to exchange support with each other. In particular, the discussion board platform is one of the most popular and active online breast cancer support groups on the Internet. It contains more than 90,000 registered members and 66 forums organized by disease stage (e.g., Stage IV and Metastatic Breast Cancer), treatment (e.g., Chemotherapy - Before, During and After), demographic group (e.g., Women 40-60ish with Breast Cancer) or entertainment (e.g., Humor and Games). In the forums, members can ask questions, share their stories, and read posts of others about how to deal with their disease. This discussion board platform is a rich environment for studying the dynamics of online support groups. We collected all of the public posts, users, and their profiles on the discussion board platform from the forum from October 2001 to January 2011. 31,307 users had at least one post.

4. SYSTEM DESCRIPTION

We aim to automatically generate and visualize cancer event trajectories of users. Figure 1 shows the configuration of our tool "Breast Cancer Trajectory".
Area 1. User ID
Area 2. Cancer event trajectory
Area 3. Events buttons
Area 4. Monthly post frequency

Use of the tool begins by inputting a forum user's ID in area 1 in Figure 1. The whole trajectory (area 2) begins with the month of the first post of this user and ends with the month of the last post. By pressing the event buttons in area 3, the corresponding event tag will appear in area 2, unless the date of this event is not retrievable for this user. The cancer event tags, "Diag" (Diagnosis), "Chemo" (Chemotherapy),"Rads" (Radiation therapy), "Mast" (Mastectomy), "Lump" (Lumpectomy), "Recon" (Reconstruction), "Recur" (Recurrence) and "Mets" (Metastasis) are located at the month of that event on the trajectory. When a user presses the "PostNum" button in area 3, then the bars in area 4 show the monthly posting frequency of the user. The height of the blue bar corresponds to the number of posts the user contributed to existing threads each month. The height of the pink bar corresponds to the number of thread starter posts the user initiated each month.

4.1 Automatic Cancer Trajectory Generation

Extracting cancer trajectories from a highly informal online forum is a non-trivial problem. Figure 2 shows the flow chart of event date extraction, which is the most challenging part of the process. A typical two-year frequent breast cancer forum user has 500-1000 posts, which contain 2000-5000 sentences. To reduce the search space, we first extract the sentences that may contain the temporal information of the disease events from the posts and then extract a date from these date sentences instead of directly from the complete raw posts. In Section 4.1.1, we present how we define and extract these "date sentences". The topics and contents of the messages in this forum are highly diverse. To reduce noise, we train a machine learning model to decide if the date in the date sentence is the date of the target event. Finally, we choose the most likely event date based on some intuitive rules of thumb.

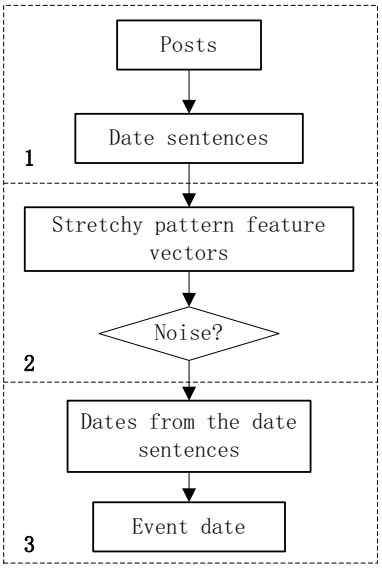


Figure 2: Event extraction flowchart.

4.1.1 Cancer Event Keywords

The cancer event is usually signaled by a set of keywords. In our experiment, we manually design an event keyword set for each cancer event. The keyword set includes the name of the event, abbreviations, aliases and other related words. For example, the Chemotherapy keyword set contains the common medical terms, such as AC (Adriamycin and Cytoxan). The Reconstruction keyword set contains the common surgery type names, such as DIEP (deep in-

Figure 1: Automatically-generated breast cancer trajectory of an example user.

ferior epigastric perforator flap breast reconstruction). By creating an analogous keyword set, our event date extraction method could be easily adapted to other datasets.

4.1.2 *Date Sentences Extraction*

We define a "date sentence" as a sentence that contains at least one time expression and at least one cancer event keyword. A user frequently shares the date of her disease event in close proximity to mention of that event by posting these "date sentences". For example, if we want to detect when a user was diagnosed of breast cancer, we will extract sentences like "I was <diagnosis keyword> on <temporal expression>" from her posts. Here the keyword set, <diagnosis keyword>, is {diagnosed, dx, dx., dx'd}, where "dx", "dx." and "dx'd" are the abbreviations of "diagnosed" that are often used by the breast cancer forum users.

In our technical approach, we recognize three types of time expressions: specific expressions, generic time expressions and self-contained time expressions (As illustrated below, the items in the brackets are optional.). Specific expressions could be resolved easily. Generic time expressions usually specify the length of interval between the time of the post and the time of the event. The date of the event could be calculated by subtracting the duration from the date of the post. For example, a date sentence is "I had my first radiation three months ago." Then her first radiation is three months before the post time. The most complicated cases are self-contained time expressions. Some of them could be resolved as generic time expressions. For example, "today is my first breast cancer anniversary" means that she was diagnosed one year ago. The other cases cannot be resolved without further information about the user. For example, a lot of users use their age as time references of the events. For example, "I was diagnosed at the age of 57". We obtain the age of users from their personal profile to handle these cases. Here are some examples of each type:

- Specific expressions
  I was diagnosed on Sep(tember|(.)) ((the) 8(th)(,)) (of) ((20)08|'08).
  I was diagnosed in (20)08|'08 Sep(tember) (8(th)).
  I was diagnosed on (0)9(/|−)((0)8)(/|−)(20)08.

- Generic time expressions
  $< 2008-10-20 >$ I was diagnosed in September.
  $< 2008-09-15 >$ I was diagnosed a week ago.
  $< 2008-09-15 >$ I was diagnosed last week.
  $< 2011-09-08 >$ Now I am three years from my diagnosis.
  $< 2011-09-08 >$ I have been a three years survivor since my first diagnosis.

- Self-contained time expressions
  I was diagnosed when I was 51 ((years|yr|yrs) old).
  I was diagnosed at the age of 51.

4.1.3 *Noise Reduction*

The temporal expression in the date sentences may not be the modifier of the target event, the second step of our cancer trajectory generation is to build a noise reduction machine learning model. The features are designed to capture both the characteristic of noisy date sentences and the style of "true" date sentences (i.e., date sentences that tell the date of the target event that has occurred to the user herself).

There are mainly four types of noisy date sentences. In online support groups, users not only tell stories about themselves, they also share other patients' stories (see example sentence (1) below). This kind of sentence usually includes the acquaintance's name or personal pronouns. They also discuss about or share breast cancer related news or results of published studies (example (2)). This kind of sentence usually includes keywords like "study" or "research". When talking about their own illness stories, they might be just concerned but have not actually experienced the event herself, like in examples (3) and (4). Neither of these sentences' authors had metastasis at the time of the post. When a sentence includes multiple disease keywords like in example (5), we have to decide which event the date expression modifies. In example (6), the date expression "Aug 2005" modifies a Reconstruction event but not a Mastectomy event. So before we extract the date from the date sentences, we must first judge if the topic of the sentence is actually the user herself or not, and whether she had this event already or is just concerned.

(1) A friend of mine started chemotherapy this week.
(2) In a recent retrospective study by md anderson reported

Dec 2008 at San Antonio, women who had her2+ cancer with negative nodes and tumors less than 1 cm had a 5 year recurrence of 23% and distant recurrence (mets) about 15%.
(3) I already freaked out and thought this was bc mets back in march when they told me I had thyroid nodules.
(4) I was diagnosed with stage II bc without metastasis in Aug..
(5) I had my mastectomy and later had reconstruction in Aug 2005.
(6) After my mastectomy and removal of 14 nodes on April 11th, my surgeon mentioned that if i got a cut or scrape on the affected side, i should go to the er.
(7) My mets to my bones and lymph nodes were found in Feb.

Users often tell the date of their cancer events with some detailed event-related information or description. This information is more highly individualized than what N-gram features that are typical of text extraction approaches can capture. For example, in example (6), besides the disease event keyword "mastectomy" and the date expression "April 11th", the user also tells how many lymph nodes are removed, which is an important feature of the mastectomy surgery. In example (7), besides the illness event keyword "mets" (an abbreviation of "metastasis") and the temporal expression "Feb", the user also tells her the metastasis sites. To capture these detailed but important features, we adopt a recently introduced method called "stretchy pattern" features instead of the commonly-used N-gram features. These stretchy pattern features can be extracted from text using a tool called LightSIDE [14];

The intuition behind this decision is that to better capture the wide variety of flexible and informal language found in social media, we need linguistic features that have strong expressive power and can be modeled with reasonably small amounts of training data. To this end, prior work has proposed the notion of a "stretchy pattern" to model stylistic variation in sociolects [9]. A stretchy pattern is defined as a sequence of word categories, some of which may be Gaps that are able to cover some number of symbols of any type. It is the Gap categories that make the patterns flexible. We designate every word instance by its word category label. A Gap is a special category. Compared to N-gram patterns, stretchy patterns allow longer linguistic patterns to be captured, and to do so in a flexible way. Using the appropriate word categories, stretchy patterns are applied here to classify if the temporal expression in the sentence is describing the the user's target event. For sentences like example (6), numbers are replaced by a word category <Number>. For sentences like example (7), a word category <BodyPart>, which includes "liver", "lung" and "brain", etc. can appear between <event keyword> and <temporal expression>. But if $k \in <event_1 keyword> \neq <event_2 keyword>$, then k should not appear between <event2 keyword> and <temporal expression>. If so, the temporal expression is more likely to be the date of $event_2$ but not $event_1$, like in example (5).

4.1.4 Rules of Thumb for Resolving Temporal Ambiguities

If there is more than one temporal expression in a sentence, then we intuitively choose the time expression that is the nearest to the event keyword. When more than one date is extracted for an event of a user, for example, $date_i$ is extracted from N_i sentences, then we choose the date(i) with the biggest N(i). The assumption is that the more frequently the user associates the event with a date, the more probably the date is the event date.

4.2 LDA Topic Modeling

To observe how topics of a user's posts can vary with her progression of disease events, a statistical topic modeling approach is used to identify topical themes in each message. In prior work [24], cancer-related dictionaries have been constructed using Latent Dirichlet Allocation (LDA). LDA is a statistical generative model that can be used to discover latent topics in documents as well as the words associated with each topic [3]. Wang and colleagues [24] first trained an LDA model using 30,000 breast cancer messages randomly selected from the entire dataset. Then 20 latent topics were derived from this document collection. For each topic, a topic dictionary consisted of 500 words that were determined to to strongly correlate with that topic. Table 1 shows sample vocabulary for each LDA topic dictionary. A complete list is provided in the online appendix [1]. With these 20 cancer-related topic dictionaries, the topic of each post is represented as a 20-dimension topic distribution vector. Each dimension of the vector calculates the frequency of words in a message matching its corresponding dictionary. For example, in the following post,

Girls please pray for me. I am so sick.
Susan

There are 10 words in this post. 3 words, "girls", "please" and "am", belong to the "Forum Communication" topic vocabulary, so the 4th dimension is 0.3. 4 words, "girls", "I", "am" and "sick", belong to the "Emotional reaction" topic vocabulary, so the 15th dimension is 0.3. Similarly, 3 words, "girls", "please" and "pray", belong to the "Spiritual" topic vocabulary, so the 17th dimension is 0.3.

5. TOOL VALIDATION

Before describing how to use our tool to uncover new knowledge about posting behavior and cancer histories, we first validate the accuracy of our cancer trajectory extraction system in this section.

5.1 Noise Reduction

We randomly choose 100 users and manually labeled all the date sentences in their posts as the training data. We use Bayesian logistic regression as our machine learning model. The stretchy pattern features are extracted using LightSIDE [14]. In our experiment, we used 16 manually-collected word categories when extracting stretchy pattern features. For example, <I> is the first person category. <prep> is the preposition category. <doctor> category contains the words that are used to refer to doctors. We used Weka [27], a machine learning toolkit, to build the regression models. We also experimented with an SVM classifier and found logistic regression to do slightly better. The 10-fold cross validation results are shown in Table 2. The results indicate that by using the stretchy pattern features, we could reliably remove noisy date sentences.

The top 10 stretchy pattern features for Metastasis events

[1]http://www.cs.cmu.edu/ yichiaw/Data/CSCW2012/CSCW2012-FeatureSet.htm

Table 1: Samples of Vocabulary in LDA Topic Dictionaries

LDA Topic	Sample Vocabulary
Pre-diagnosis	Told, appointment, wait, back
Treatment plan	Clinical, risk, medicine, therapy
Forum communication	Post, read, help, thread
Adjusting to diagnosis	Understand, trying, experience
Financial concerns	Insurance, plan, company, pay
Lymphedema	Arm, pain, swelling, fluid, area
Diet	Eat, weight, food, exercise, body
Family/Friends	Daughter, sister, wife
Positive life events	Love, nice, happy, enjoy, fun
Surgery	Breast, surgeon, mastectomy
Thoughts/Feelings	Think, remember, believe
Chemo/Radiation	Chemo, radiation, treatment
Family history	Mom, children, age, young
Emotional reaction	Better, lucky, scared
Tumor Treatment	Biopsy, nodes, positive, report
Spiritual	Love, god, prayer, bless, peace
Emotional support	Hope, hug, glad, sorry, best, luck
Routine/Schedule	Today, night, sleep, work
Hair loss/Appearance	Hair, wig, grow, head
Post-surgery problems	Pain, blood, tamoxifen, symptom

Table 2: 10-fold cross validation results.

Disease event	Number of sentences	Accuracy
Diagnosis	608	0.88
Lumpectomy	238	0.75
Mastectomy	530	0.81
Chemotherapy	544	0.80
Radiation	432	0.77
Reconstruction	263	0.81
Recurrence	345	0.87
Metastasis	164	0.82

are listed below. We can see that the stretchy patterns capture the form of the date expressions well. One example sentence is illustrated below to show how the stretchy patterns capture the structure of the true date sentences.

I was diagnosed in Feb. with soft tissue mets.
<diagnosis> <prep><T> [GAP] <K>
<T>[GAP]<BodyPart><K>

The top 10 Metastasis stretchy pattern features. <K> = <event keyword> = {metastasis, micrometastases, mets, metastasize, metastasises}, <T> = <temporal expression>:
<K> <prep> <T>
<BodyPart> <K> <T>
<K> [GAP] <prep> <T>
<T> [GAP] <BodyPart> <K>
<BodyPart> <K> <prep> <T>
<diagnosis> <K> <prep> <T>
<diagnosis> <prep> <T> [GAP] <K>
<BodyPart> <K> <prep> <T> <conj>
<prep> <BodyPart> <K> <prep> <T>
<BodyPart> <K> <prep> <T> <conj>

5.2 Event Date Extraction

For each user, we extract the year and the month of the following breast cancer events: Diagnosis, Lumpectomy, the beginning of Chemotherapy, the beginning of Radiation Therapy, Mastectomy, breast Reconstruction, Metastasis and cancer Recurrence. Notice that not all events occur for each individual. Also some users may have multiple rounds of Chemotherapy, Reconstruction, Metastasizes and Recurrences. Among all the 31,307 users who have at least one post, 7487 users are located with at least one event date. As there is no established baseline to compare with, we randomly choose 20 long-term users, who have been active on this forum for more than 2 years [16]. Then we manually extract the event dates from their posts and profile. The results are shown in Table 3. Except for Reconstruction, we see that we are able to reliably extract the date of these disease events. There are several reasons that Reconstruction date is hard to extract. One is that as reconstruction is an purely optional surgery, it is quite common for patients to change the surgery schedule. For example, one user posted "My reconstruction is scheduled 11/08". But in her later post, she postponed the surgery date. It is hard for the classifier to distinguish between the scheduled but changed surgery dates and the actual surgery dates. Another reason is that a major kind of breast reconstruction is done at the same time with the mastectomy surgery. So these users usually will not separately state the date of their reconstruction. Instead, they will post "I had mastectomy with immediate reconstruction". In this case, it is possible to extract the reconstruction time if the user had stated their mastectomy time. But there may be no date sentences extracted for the Reconstruction event directly.

Table 3: Event date extraction evaluation results

Disease event	Total	Correctly extracted
Diagnosis	20	16
Lumpectomy	11	8
Mastectomy	12	10
Chemotherapy	14	9
Radiation	12	8
Reconstruction	8	4
Recurrence	4	3
Metastasis	4	3

6. DISEASE EVENTS AND PATTERNS OF ONLINE FORUM POSTS

In order to find the relationship between posting behavior and the cancer events, we use our tool to investigate two questions in this section. One is what prompts participation in online heath support groups? Our hypothesis is that in the important event months, the patients are more distressed and crave more interaction. So they are more likely to join the community and increase participation in the forum discussions. The other question is what are the issues of the most interest during different event months? Our hypothesis is that users will be more interested in and post messages that are related to their ongoing cancer event.

6.1 Change of Message Frequencies Across the Cancer Trajectory

The distressful events prompt cancer patients' participation in online health support forums. Firstly, during event months, we see that users initiate and follow more posts.

We first calculate the number of monthly messages across the cancer trajectory. On average, a user initiates 0.34 threads month, and posts follow-up messages on existing threads 7.83 times per month. We define the month during which at least one event happened as an "event month". In event months, a user initiates 0.84 threads per month, and posts 14.47 follow-up posts on average. The message number peaks across the post trajectories usually is a signal of the user's cancer events (See the rectangles in Figure 3). When facing these distress-inducing events, people position themselves to receive more informational and emotional support.

Secondly, a large portion of the users join this community in the same month as one of their cancer events. Among the 7487 users who are located with at least one event date, 2145 users started using the forum in an event month (Table 4). Although it would be natural to imagine that users come to the community when they are diagnosed, we found that 1123 users posted their first post when they were starting chemotherapy, which was approximately twice the number that came in their diagnosis month. Typically, they update their chemotherapy treatment frequently to a thread to connects with the other women who started chemotherapy in the same month. Such threads include "2008 October Chemo Girls" and "Starting Chemo this October". Women receiving chemotherapy have reported increased levels of psychological distress, difficulties with psychosocial function [21] and increased level of uncertainty [11] when compared with women not receiving chemotherapy. Since chemotherapy typically takes several months, the patients who participate in such threads have time to get familiar with the subcommunity of users experiencing something similar.

Table 4: Number of users who join the community when they are undergoing a certain disease event.

Disease event	Number of users
Diagnosis	509
Lumpectomy	276
Mastectomy	340
Chemotherapy	1123
Radiation	351
Reconstruction	117
Recurrence	8
Metastasis	36
Total	2145

6.2 Topics of Messages Across the Cancer Trajectory

Although prior work has pointed out that people post more when they are experiencing some of these key cancer events [26], what has not been explored is what distribution of topics people are thinking about and posting about during those times include. Investigating this question offers a portal into their coping processes during these important periods of time. In this section, we shift to looking at the message topics as they vary across the cancer trajectory. We believe that users will engage in discussion topics that are relevant to their current cancer phase. We calculate the average topic percentage vector of each disease event as the average over all a user's posts during the month of the event. From Table 5, we can see that among all the events, the Chemotherapy event has the highest percentage on *Chemo/Radiation* topic, which validates our Chemotherapy date extraction. The Diagnosis event has the highest percentage on both *Pre-diagnosis* and *Adjusting to diagnosis* topics, which validates our Diagnosis date extraction. Both Reconstruction and Mastectomy events have high percentages on the *Surgery* topic as both these two events involve surgeries.

Our topic model analysis also provides some insights into what people are discussing about when facing these disease events. As metastasizing cancer is highly dangerous, the messages contain more Spiritual themes where people send prayers and blessings to each other during the month when Metastasis is found. The *Hair loss/appearance* topic is the most salient during the Chemotherapy month. This indicates that users are coping with the side effects of chemotherapy and adjusting to the illness during chemotherapy. For example,

Title: Hair Hair Hair - Another question
I know that there have been several threads on this, but I'm asking the question again: on average, when did your hair start growing back in after chemo?

6.3 Cancer Story Summarization and Cancer Trajectory

A majority of cancer patients find hearing and constructing illness stories useful for their own decision making [1, 8, 18, 20]. Users in the breast cancer forum search for or even directly ask for cancer stories that are similar to their own situation. Below is an example message showing this need. Currently there are few search facilities concerning illness stories on the Internet [18, 7]. Automatic cancer story summarization has tremendous potential to make cancer coping experiences accessible to patients within these online support communities, or as a biproduct of the data created within them.

Title: Boost my spirits...long term trip neg mets stories please
OK ladies.....I need a little boost here.
Most days I am doing fine with the mets Dx.
A few things still get me panicky.
One of them is the fact that almost all the long-term mets success stories I read are for hormone positive ladies.
I am a triple negative gal, and I really could use a few positive long term mets survivor stories.

Visualizing the events according to each user's cancer trajectory, as shown in Figure 3, could facilitate the summarization of the user's cancer story. In Figure 3(a), we could see that user 26326 joined this community one month before her mastectomy. She gradually got familiar with the community. She tended to post and respond to more and more posts. The number of her messages had a sharp increase when her metastasis was found (the red rectangle in Figure 3(a)). She was very frustrated and initiated a number of posts requesting both information and emotional support. Then she began chemotherapy for treating the metastasis. She posted less and less due to her worsening situation, and finally posted zero posts the month before her death. Her last message was posted by her husband after her death.

Table 5: Average topic percentage vector of disease event. The biggest percentage of each event is shown in bold. The biggest percentage of each topic is shown in italic.

LDA Topic	Avg.	Diag.	Chem.	Mets.	Reco.	Mast.	Recu.	Lump.
Pre-diagnosis	.171	***.199***	.195	.180	.191	.192	**.173**	.189
Treatment plan	.104	.124	.105	.110	.104	.104	*.144*	.114
Forum communication	.146	*.165*	.156	.158	.153	.154	.159	.152
Adjusting to diagnosis	.160	*.180*	.172	.168	.169	.172	.160	.165
Financial concerns	.113	*.117*	.114	.116	.114	.114	.117	.115
Lymphedema	.130	.137	.145	.129	.145	.146	*.153*	.138
Diet	.122	.116	*.131*	.118	.118	.120	.123	.119
Family/Friends	.147	.156	.142	*.160*	.146	.142	.135	.137
Positive life events	.119	.097	.102	.115	.107	.106	.102	.105
Surgery	.142	.167	.151	.133	***.209***	.193	.143	.159
Thoughts/Feelings	.138	.143	.140	.149	.144	.144	*.153*	.143
Chemo/Radiation	.111	.121	*.173*	.115	.118	.119	.117	.116
Family history	.154	*.170*	.164	.168	.159	.160	.153	.155
Emotional reaction	.184	.201	***.215***	.193	.199	**.200**	.172	**.190**
Tumor Treatment	.122	*.176*	.137	.128	.146	.159	.166	.167
Spiritual	.194	.195	.184	***.205***	.186	.181	.170	.174
Emotional support	.159	.159	.161	*.177*	.158	.154	.144	.145
Routine/Schedule	.161	.161	.193	.161	.171	.172	.141	.161
Hair loss/Appearance	.161	.161	*.189*	.161	.169	.167	.148	.161
Post-surgery problems	*.189*	.112	.128	.132	.115	.113	.129	.109

User 60351 in Figure 3(b) joined the community two months before diagnosis when her mammogram showed something unusual (the blue rectangle in Figure 3(c)). She began chemotherapy one month after diagnosis. Her post equency peaked in her first chemotherapy month (the red rectangle in Figure 3(b)), when she joined the March 2008 chemotherapy group and updated her situation to the board almost everyday. She had a lumpectomy but later found metastasis and soon began radiation therapy. The extracted cancer trajectory shows that she had both a mastectomy and reconstruction in June 2009, which is later than her last post (the black circle in Figure 3(b)). Actually, she planed to have mastectomy and reconstruction in June but failed to survive that far.

The death of members has a big influence on the online support group. Below is a message that shows the users' concern over the sudden death of community members. Our automatically generated cancer trajectory may potentially help monitor the death of the online support group members to anticipate where others may need additional support, and thus trigger interventions that might provide this needed support.

> *"I'll be lurking of course, but wanted to post because I know many of us can often just disappear, and I don't want to do that. My BFF will post when all is done."*

User 64251 shown in Figure 3(c) joined the community when she was first diagnosed (the blue rectangle in Figure 3(c)), followed by immediate lumpectomy and mastectomy. Later she joined a large number of threads about breast reconstruction. After her reconstruction was finished, she posted less and less posts and finally left the community. The decreasing trend of the monthly post frequency shows how a breast cancer patient gradually adjusts to this life-altering disease.

From these three example users, we can see how our interface could assist understanding (1) the different reasons of entering the community; (2) the association between peaks of the monthly post numbers and cancer events; (3) the unusual posting behavior of each individual user.

7. CONCLUSIONS AND CURRENT WORK

In this paper, we describe how we built machine learning models to reliably extract cancer event trajectories from messages in online breast cancer support groups. We examined the relationship between disease events and post patterns. The results demonstrate that both the frequency and the topic of messages were correlated with the distress-inducing cancer events. These events prompt cancer patients to begin or increase participation in the online health support groups. Our message topic analysis shows that users engage more in threads that are related to their ongoing current cancer event. Our visualization tool, "Breast Cancer Trajectory" indicates further application in cancer story summarization.

Much previous work has studied how breast cancer patients psychologically and physically adjust to breast cancer [15, 23]. These studies are done with a restricted number of participants. In contrst, the automatically extracted cancer trajectories will allow us to study how users adjust to this illness at a large scale. As the cancer events are tightly related to information and emotional support seeking, our work is potentially useful for online support group studies such as those published in related work [24].

There are several potential directions for improving the current interface. First, it is possible to extract from the posts whether the user was alive or not at the time of posting. Second, we can represent the message topic variation across the cancer trajectory. For example, we could visualize when the user talks about death-related topics as a way of understanding better how the experience of approaching death affects participation. During different cancer treatments, the cancer patients will develop close relationship

with several different doctors. There are many posts in the forum that talk about doctors. Since trust in doctor-patient relationships is an important factor in patient wellbeing, we are currently working on understanding patients' attitudes towards doctors and how they change over time in relation to important cancer events.

8. ACKNOWLEDGMENTS

We want to thank Dong Nguyen and Yi-chia Wang, who helped provide the data for this project. The research reported here was supported by National Science Foundation grant IIS-0968485.

9. REFERENCES

[1] K. Arthur. *The Illness Narratives. Suffering, Healing, and the Human Condition*. Basic Books, New York, 1988.

[2] A. Barak, M. Boniel-Nissim, and J. Suler. Fostering empowerment in online support groups. *Computers in Human Behavior*, 24(5):1867 – 1883, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] J. M. Corbin and A. Strauss. *In Chronic Illness Trajectory Framework: The Corbin and Strauss Nursing Model*. Elsevier, New York, 1992.

[5] K. P. Davison, J. W. Pennebaker, and S. S. Dickerson. Who talks? the social psychology of illness support groups. *American Psychologist*, 55(2):205 – 217, 2000.

[6] V. Eidelman. Inferring activity time in news through event modeling. In *Proc. ACL-HLT2008, Student Research Workshop*, pages 13–18, 2008.

[7] G. Eysenbach. The impact of the internet on cancer outcomes. *CA: A Cancer Journal for Clinicians*, 53(6):356–371, 2003.

[8] A. W. Frank. *The Wounded Storyteller. Body, Illness, and Ethics*. The University of Chicago Press, Ltd., London, 1995.

[9] P. Gianfortoni, D. Adamson, and C. P. Rosé. Modeling of stylistic variation in social media with stretchy patterns. In *Workshop on Modeling of Dialects and Language Varieties at EMNLP2011*, pages 49–59, 2011.

[10] M. Hanson Frost, V. J. Suman, T. A. Rummans, A. M. Dose, M. Taylor, P. Novotny, R. Johnson, and R. E. Evans. Physical, psychological and social well-being of women with breast cancer: the influence of disease phase. *Psycho-Oncology*, 9(3):221–231, 2000.

[11] B. A. Hilton. Getting back to normal: the family experience during early stage breast cancer. *Oncology Nursing Forum*, 23(4):605–614, 1996.

[12] K. Luker, K. Beaver, S. Leinster, and G. R. Owens. Information needs and sources of information for women with breast cancer: A follow-up study. *Journal of Advanced Nursing*, 23:487–495, 1996.

[13] I. Mani and G. Wilson. Robust temporal processing of news. In *Proc. of ACL-2000*, pages 69–76, 2000.

[14] E. Mayfield and C. P. Rosé. LightSIDE: Open Source Machine Learning for Text Accessible to Non-Experts. In *Invited chapter in the Handbook of Automated Essay Grading (in press)*, 2012.

[15] T. Morris, H. S. Greer, and P. White. Psychological and social adjustment to mastectomy: a two-year follow-up study. *Cancer*, 40(5):2381–2387, 1977.

[16] D. Nguyen and C. P. Rosé. Language use as a reflection of socialization in online communities. In *Workshop on Language in Social Media at ACL2011*, pages 76–85, 2011.

[17] T. Noro, T. Inui, H. Takamura, and M. Okumura. Time period identificaiton of events in text. In *Proc. COLING-ACL2006*, pages 1153–1160, 2006.

[18] R. I. Overberg, L. L. Alpay, J. Verhoef, and J. H. M. Zwetsloot-Schonk. Illness stories on the internet: what do breast cancer patients want at the end of treatment? *Psycho-Oncology*, 16(10):937–944, 2007.

[19] R. S. and C. Q. Internet community group participation: Psychosocial benefits for women with breast cancer. *Journal of Computer Mediated Communication*, 10(4), 2005.

[20] B. F. Sharf and M. L. Vanderford. *Illness narratives and the social construction of health*. Lawrence Erlbaum Associates, NJ, 2003.

[21] P. Trief and D.-S. M. Counseling needs of women with breast cancer: what the women tell us. *The Journal of Psychosocial Nursing and Mental Health Services*, 34(5):24–29, 1996.

[22] B. van der Molen. Relating information needs to the cancer experience: Themes from six cancer narratives. *European Journal of Cancer Care*, 9:48–54, 2000.

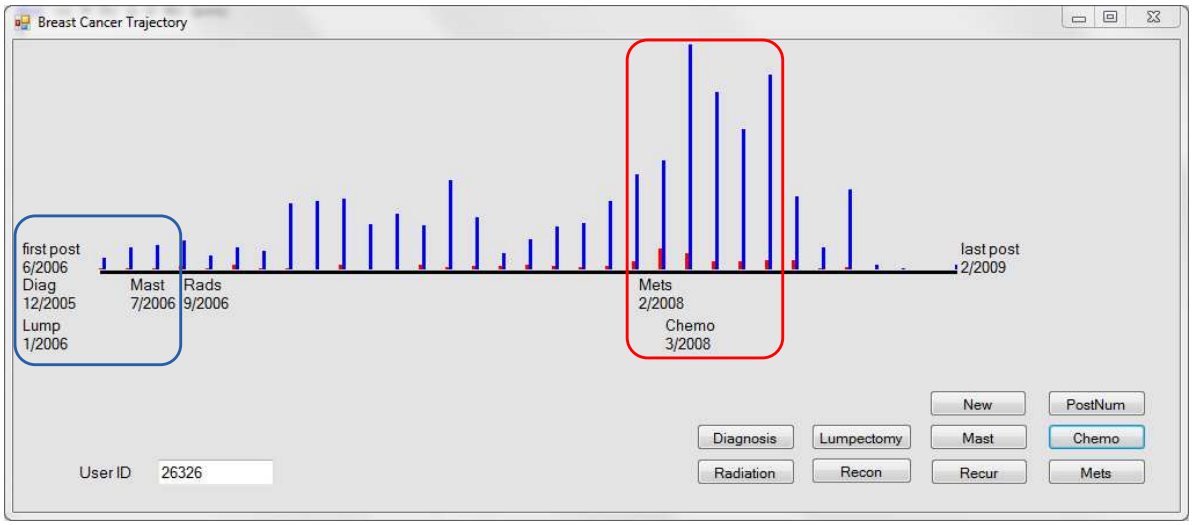
[23] S. H. Vicki, P. Snyder, and H. Seltman. Psychological and physical adjustment to breast cancer over 4 years: Identifying distinct trajectories of change. *Health Psychology*, 23(1):3–15, 2004.

[24] Y. Wang, R. Kraut, and J. Levine. To stay or leave? the relationship of emotional and informational support to commitment in online health support groups. In *ACM Conference on Computer Supported Cooperative Work*, pages 833–842, 2012.

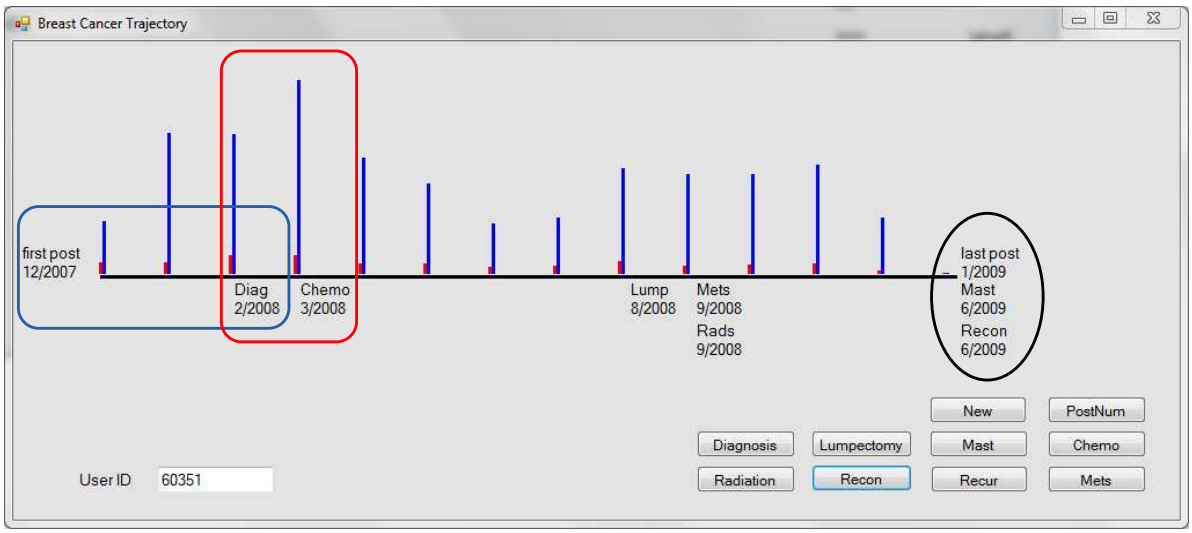
[25] K.-Y. Wen and D. H. Gustafson. Needs assessment for cancer patients and their families. *Health and Quality of Life Outcomes*, 2(11), 2004.

[26] K.-Y. Wen, F. McTavish, G. Kreps, M. Wise, and D. Gustafson. From diagnosis to death: A case study of coping with breast cancer as seen through online discussion group messages. *Journal of Computer-Mediated Communication*, 16:331–361, 2011.

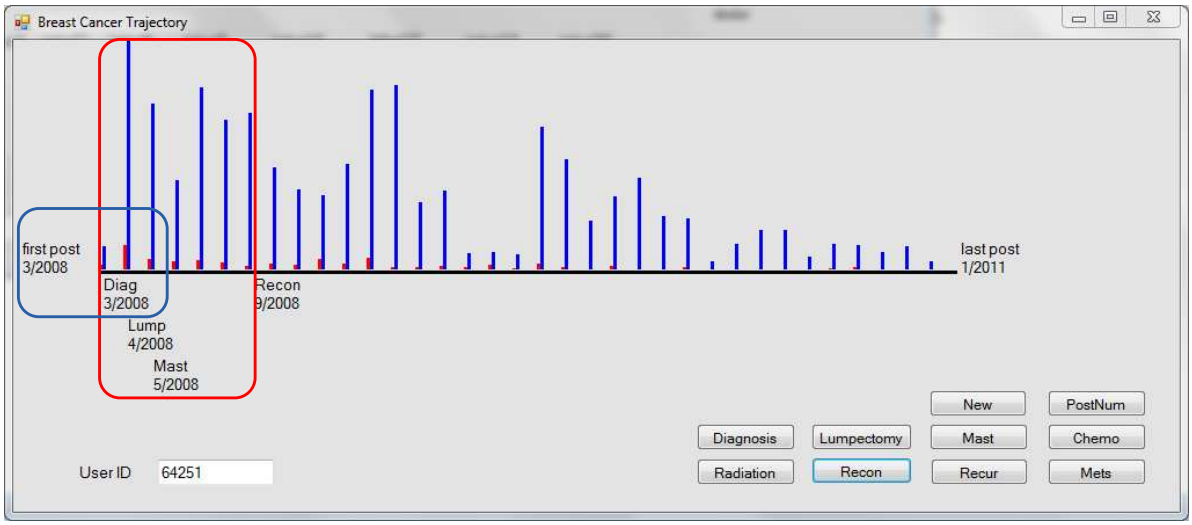
[27] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, second edition*. Elsevier, San Francisco, 2005.



(a)



(b)



(c)

Figure 3: Automatically-generated breast cancer trajectories of three users.