Different Strokes of Different Folks

Searching for Health Narratives in Weblogs

Andrew S. Gordon, Christopher Wienberg
Institute for Creative Technologies
University of Southern California
Los Angeles, CA USA
gordon@ict.usc.edu, cwienberg@ict.usc.edu

Sara Owsley Sood
Computer Science Department
Pomona College
Claremont, CA USA
sara@cs.pomona.edu

***Abstract*—The utility of storytelling in the interaction between healthcare providers and patients is now firmly established, but the potential use of large-scale story collections for health-related inquiry has not yet been explored. In particular, the enormous scale of storytelling in personal weblogs offers investigators in health-related fields new opportunities to study the behavior and beliefs of diverse patient populations outside of clinical settings. In this paper we address the technical challenges in identifying personal stories about specific health issues from corpora of millions of weblog posts. We describe a novel infrastructure for collecting and indexing the stories posted each day to English-language weblogs, coupled with user interfaces designed to support targeted searches of these collections. We evaluate the effectiveness of this search technology in an effort to identify hundreds of first person and third person accounts of strokes, for the purpose of studying gender differences in the way that these health emergencies are described. Results indicate that the use of relevance feedback significantly improves the effectiveness of the search. We conclude with a discussion of sample biases that are inherent in weblog storytelling and heightened by our approach, and propose ways to mitigate these biases.**

Keywords- weblogs, storytelling, health, information retrieval

I. INTRODUCTION

Within health and healthcare research, qualitative methods are routinely used to investigate illness and other health-related issues from the perspective of the patient and their family. An important line of qualitative health research applies narrative analysis methods to the stories that these individuals share about their health-related experiences. These stories have been a focus of interest in narrative medicine, which Charon (2006) defines as "medicine practiced with these skills of recognizing, absorbing, interpreting, and being moved by the stories of illness." Stories of illness have been gathered directly from the patient population during interactions with health professionals, but how these narratives have been used to inform/improve healthcare varies. Methodological and ethical issues in health-related narrative analysis have been vigorously debated, e.g. in response to Atkinson's (1997) criticism of the view that narratives offer analysts privileged access to personal experience (Thomas, 2010).

Before these methodological and ethical debates could be resolved, however, the rise of weblogs and social media opened up an entirely new set of concerns and opportunities. Now millions of people routinely publish intimate details of their lives to the Internet, including stories of illness in themselves, their family, and friends. Weblogs may yield hundreds or thousands of stories about very specific health issues written by people of increasingly diverse backgrounds in diverse healthcare contexts. Conceivably, investigators could study a wide range of health-related issues using the methods of narrative analysis without ever interacting with the patient population directly.

In this paper, we focus on a practical problem: How can investigators identify suitably large numbers of personal stories about specific health issues in public weblog posts? We developed a technical solution to this problem, where tens of millions of personal stories are collected from public weblogs, indexed using a high-performance text search engine, and made accessible in a retrieval interface tailored to support this search task. We evaluated the effectiveness of this solution given a challenging search task. A small team of investigators used our search technologies to identify hundreds of personal accounts of having a stroke, in an effort to begin a new investigation of gender differences in the way that people describe this health emergency. At the completion of their search, we evaluated the contribution of relevance feedback in improving the precision and recall performance of their queries, and identified the parameters that would optimize similar queries in future searches. In the sections that follow, we review related research on health narratives in weblogs, describe our own development efforts and evaluation, and conclude with a discussion of how to use health narratives as evidence in health investigations.

II. HEALTH NARRATIVES IN WEBLOGS

Popular political websites are most associated with weblogs as a genre of social media, but a more typical weblog takes the form of a personal journal, read by a small number of friends and family (Munson & Resnik, 2011). Recent research has attempted to uncover what drives bloggers to share day-to-day experiences. Hollenbaugh (2011) surveyed personal bloggers, finding boredom and social connection as motives, which is consistent with motives found in studies of Internet use, more broadly. Some bloggers crave attention, while others use their personal blogs as online journals for archiving their thoughts and experiences. Other bloggers share information in hopes of helping others in some way.

Health-related blogs make up a large portion of personal blogs, either indirectly by addressing health among other topics, or directly by chronicling experiences with a specific condition or disease (Adams, 2010). Recent studies have attempted to catalogue the topics and perspectives of a sample of health-related blogs to further understand the nature of health-related blogging. In their health-related blog dataset, Miller and Pole (2010) found that 54.3% are writing as medical professionals, 37.7% are patients and 8.0% are caregivers; 42.6% of all blogs in their dataset chronicle experiences with a specific condition. Miller and Pole categorically labeled these conditions and found mental health as the most frequently occurring category, followed by reproduction, chronic disease, physical disability, HIV/AIDS, and cancer. They also found majority female health-bloggers, contrasting with the overall blogosphere.

Miller and Pole describe a "purposive-snowball sampling approach" to gathering their corpus of health-related weblogs, identifying initial blogs through aggregators and keyword searches via publicly available weblog search engines and expanding outward through blogrolls. This method produced a reasonably large dataset (951 health-related blogs), but faces a number of limitations. This process is biased toward the identification of popular blogs, which they note has great potential to misrepresent topics in the blogosphere. In particular, their dataset may disproportionately represent more authoritative or popular blogs written by physicians. In contrast, weblogs written by patients and caregivers are harder to discover, as their smaller readership yields fewer links to weblog posts. This is a major barrier to researchers wishing to identify blogs written by patients and caregivers.

Adams (2010), calls for future research to improve our understanding of health-related weblogs from the patient perspective, and to investigate "under which circumstances individuals do or do not choose to use this broadcast format" to expose "how individuals use public applications to resolve communication difficulties." These communication difficulties are outlined by Pratt et al. (2006), who note the increasingly significant responsibility of the patient to communicate health information among many different health care providers, insurance, and their employer, friends and family, while maintaining privacy (as they chose) and integrity of the information. This is clearly a motivating factor for some health-related bloggers—particularly parents, caregivers and the patients themselves—who often write to share experiences and progress with friends and family.

III. Search Infrastructure

There are several Internet search providers that focus on providing access to Weblog content, e.g. Google Blog Search and Technorati.com. While possible to conduct searches for specific classes of health-related stories using existing tools, there are two aspects of the task that are poorly suited to these search services: genre filtering and event retrieval. Genre filtering involves removing retrieved posts not written in the personal story genre. Gordon and Swanson (2009) estimate that only 4.8% of randomly sampled non-spam English-language weblog posts can be characterized as personal stories, defined as non-fiction narrative discourse describing a specific series of events in the past, spanning minutes, hours, or days, where the storyteller or close associate is a participant. As a consequence, the large majority of search results for a given health issue are not suitable, leaving the investigator with the burden of assessing appropriateness of the genre. Event retrieval is to identify only those stories in the collection where the events described in the narration are of interest to the investigator. Existing search services are particularly well suited for retrieval tasks where relevance is determined by the existence of proper names (e.g. "Mayo Clinic") or technical terminology (e.g. "myocardial infarction"). However, events of interest to investigators often do not have clear lexical identifiers (e.g. the event of consulting with a physician in their offices) and are narrated using layman's vocabulary with a high degree of polysemy (e.g. "heart attack", "stroke"). Search results lack precision when query terms lean toward blogger's vocabulary and lack recall when they lean toward technical terms. Thus, investigators face too many non-relevant search results to consider or too few to answer research questions.

Our solution was to develop a new weblog search infrastructure engineered to support the retrieval of English-language personal stories of specific events. Our aim was to minimize the time an investigator spent reading and rejecting blog posts that were not personal stories relevant to the issue of interest. As a source of weblog data, we partnered with Spinn3r.com, a commercial weblog aggregator that provides a spam-filtered, language-sorted feed of weblog data to corporate clients and to researchers as part of a special program for academics. In 2010 and 2011, we downloaded and processed over one billion English-language weblog posts using a software pipeline that ran successfully for 634 out of 730 days.

A. Genre Filtering

For genre filtering, we used an automated story filter developed by Gordon and Swanson (2009), which classifies the content extracts provided by Spinn3r.com as either story or non-story content using supervised machine learning techniques (precision=.66, recall=.48). We applied this text classifier to each of the spam-filtered, English-language posts downloaded from Spinn3r.com in 2010 and 2011, labeling 17.4 million posts as personal stories.

B. Duplicate Post Removal

Critical to any system that collects and indexes content is the ability to identify duplicate items. Duplicate blog posts exist in our dataset for many reasons. Some blogs are syndicated to many sites; many are reproduced to generate ad revenue. Others are duplicated because the author made an (often minor) update to a previous post, resulting in the appearance of an updated version of the post in our content feed (from Spinn3r.com). In the case of the former, the content of two blog posts are often identical, though not always. For the latter, we find many posts that are nearly identical (near duplicates).

To identify duplicate and near duplicate blog posts, we employ two hashing algorithms. The first, the md5 message digest algorithm, generates 128 bit hashes of each blog post, which are matched to identify exact duplicates. The second, Charikar's simhash algorithm (Charikar, 2002), casts the net

more broadly, also identifying near duplicate blog posts. Unlike other hashing techniques, the simhash algorithm generates hashes in such a way that similar documents produce similar hashes (differing in a small number of bits). This allows one to compare hashes in order to detect near duplicates, in which the number of identical bits in two hashes exceeds a similarity threshold.

While relatively common solutions exist for detecting duplicates, they often face complexity issues, as each item must be compared to all other items. As our dataset contains approximately 17.4 million stories (and continues to grow by 30,000 stories each day), this is an issue of both time and space. While the simhash algorithm is accurate for detecting near duplicates, it is also beneficial in that past work has concluded that 64 bit simhash representations suffice for most tasks (Manku, 2007). Representing each story as 64 bits allows for hashes to be stored in memory, enabling fast comparisons. Using md5 and simhash representations, we identified 2,757,326 (15.9%) duplicates and near duplicates in our corpus of 17.4 million stories.

C. *Event Retrieval Using "Boring Stories"*

To support the retrieval of relevant health narratives, we drew inspiration from the story-retrieval approach of Gordon and Swanson (2008). In their work, stories of specific activity contexts are retrieved by ranking their similarity to "boring story" queries, i.e. fictional past-tense narration of prototypical examples of the desired events. This simplifies the search task by encouraging users to produce queries that are similar to relevant results at a lexical level, but requires a scalable information retrieval platform that supports searches based on document similarity, e.g. using the vector-space model. We selected the Terrier Information Retrieval Platform from the University of Glasgow (Ounis et al., 2007), using their default divergence from randomness retrieval model.

We authored a web-based user interface to allow teams of investigators to collaboratively search for relevant weblog stories. Users begin by authoring a single "boring story" query for their topic of interest, then rate the relevance of stories in a ranked list returned by the search engine. This interface displayed the top ten search results in a manner similar to commercial web search providers, ordered by their relevance score and presented with hyperlinks to the weblog posts. Users annotate each item in the list as relevant or not relevant by selecting from buttons displayed along with each item in the list. In addition, users are given the option of skipping an item, removing it from further consideration. This option was made available primarily to remove content that appears due to failures in spam filtering (by Spinn3r.com), genre filtering, or duplicate detection. After judging items on a results page, the user reissues the query and is presented with a new set of stories that have not yet been judged. A search is complete when the set of relevant items is suitably large for the subsequent narrative analysis task.

D. *Relevance Feedback Using the Rocchio Algorithm*

The use of "boring story" queries proposed by Gordon and Swanson (2008) provides greater precision than can be achieved using a handful of ambiguous keywords. A common technique for increasing recall performance of queries is the use of relevance feedback. As a user provides additional input about the relevance of documents, we refine the search query for the collection using a standard query refinement algorithm (Rocchio, 1971). Initially, the user provides a query—in the form of a "boring story"—and searches the system. When the user annotates a document, we update the query by combining the original query's terms and the terms of the annotated document in a weighted combination. The Rocchio parameters were set to $\alpha=2$ for the weight of the original query terms, $\beta=1$ for terms in relevant items, and $\gamma=1$ for the negative weight given to terms in non-relevant items. This weighting allows terms in non-relevant items to fully discount those that appear in relevant items, and effectively considers the original query terms as being twice as important as any other. We discuss the tuning of these parameters for optimal performance in the results section of this paper.

Typically in relevance feedback applications the user is provided several documents to mark as relevant, and then the user's annotations are used to construct an expanded query, with no subsequent stages of relevance feedback. In this effort, we were gathering annotations continuously in order to find a large collection of stories, and we utilized each annotation to refine the query. As users annotated documents, and then searched again for new documents to annotate, the search was performed with an updated query based on all annotations made by the user. This results in a strategy of query refinement, where over time the query migrates toward the center of the space of relevant documents.

In the traditional vector-space model of information retrieval, all of the terms that appear in the original query and annotated documents are dimensions of the query vector. To ensure that search results could be computed at interactive speeds (consistently executed in less than 60 seconds using the full index of 17.4 millions stories), we truncated this vector to 50 terms. These 50 terms were selected as the terms with the greatest weight after the relevance feedback calculation. We discuss an alternative term selection approach in the results section of this paper.

IV. Evaluation: Stories of Strokes

In November and December of 2011, we evaluated our search technologies in collaboration with a team of faculty and students in the psychology and cognitive science departments of Pomona College and Claremont McKenna College in Claremont, California. The aim of this team of users was to identify stories that describe the experience of having a stroke to be subsequently analyzed to investigate gender differences in how these experiences are narrated. We present the motivation for collecting stories on stroke and the methodology used during this formative evaluation.

A. *Motivation*

Receiving treatment in a timely manner is critical to surviving a stroke. The FAST (Face Arms Speech Time) warning signs of stroke published by the National Stroke Association note the importance of seeking treatment quickly. Past medical studies have concluded that women who

experienced a stroke received a particular beneficial stroke treatment (tissue plasminogen activator) less than men and also were more delayed in receiving treatment, with both in and out of hospital delays (Barr, 2006; Lisabeth, 2009). One study has investigated the cause of this delay disparity concluding that women experience different symptoms; women are more likely to report nontraditional symptoms including "altered mental status" (a.k.a. confusion) (Lisabeth, 2009).

To further understand stroke diagnosis and treatment delay disparities between men and women, we look to narrative accounts of stroke experiences in blogs. If we are able to analyze gender differences in descriptions of stroke symptoms, perhaps we can better understand how to recognize these differences in emergency situations, educating triage nurses and emergency room doctors.

B. Methodology

Among our collaborators, a Cognitive Psychology faculty member and two undergraduate pre-medical students (all with prior interest in stroke) utilized our search tools over a two-month period. The three began their search together in order to establish their process and agree upon what they were looking for. Using the improved retrieval system, they first issued a "boring story" query of a general stroke experience, shown in Figure 1(a). Given a set of results, a relevance ordered list of stories (links to blog posts containing the stories), their task was then to assess the fit of each story to their needs, first and third person narratives of stroke experiences. The three discussed and decided as a group whether each story was relevant or not. This label was then passed to the Rocchio relevance feedback system to update the search query and deliver a new set of results.

In many situations, it was best to not label a story as relevant or not relevant. Using the new interface, the searchers were able to 'skip' a story. They were instructed to skip results that were relevant but were out of genre – perhaps a news article about stroke. Labeling these out of genre posts as 'not relevant' would have unfortunate side effects in the Rocchio relevance feedback system, signaling that words used in these posts should be deemed less important in the search context, when in fact, the words used in the post were relevant, just not occurring in the appropriate genre (personal stories). The searchers were also instructed to skip links that lead to a 'page not found' as the blog author had removed the post after it was indexed by our system.

The group labeled first and third person narratives of a stroke experience as 'relevant.' Each must include a diagnosis of stroke, or experiences of stroke symptoms that are undiagnosed. We suggested descriptions of stroke symptoms that were diagnosed as heart attack or Bell's Palsy be 'skipped,' so as not to adversely affect the search query as stroke symptoms were discussed in the story.

After the initial group search session, the three used the search interface independently. Results were added to a single story collection, and the relevance feedback altered a single query for the group. Each story relevance assessment resulted in more relevant result sets for the entire group. They found 210 relevant narratives (28.6%); 244 posts were labeled as not relevant (33.2%), while 281 posts were skipped (38.2%).

V. Results

This formative evaluation of our search technologies was very positive from the perspective of our users, in that they identified a suitably large corpus of stroke stories for subsequent analysis. The three users read and labeled 776 weblog posts in different sessions over the course of seven weeks. To better understand the performance of the system during this period, we conducted a series of post-hoc analyses on their annotation efforts.

A. Relevant, non-relevant, and skipped items

First, we conducted analyses to better understand the contents of each of the three categories of annotated posts. Reviewing the posts in each category, we sorted them based on what we judged to be the dominant high-level feature of that category. For the 'relevant' stories, this was the division between the first-person and third-person perspective. 77 (36.7%) of the 'relevant' stories were written by the person who had the stroke. The remaining 'relevant' stories were written by a third party, with 38 (18.1% of all 'relevant' stories) written by witnesses of the stroke, 10 (4.8%) by medical professionals, 50 (23.8%) by acquaintances of the victim who did not witness the stroke, and 36 (17.1%) by people previously unacquainted with the patient, such as reporters interviewing the patient about his or her stroke. One story was written jointly by the person who experienced the stroke and his spouse, with each author providing a personal account of the event.

Stories in the 'not relevant' category reveal an interesting side effect of our search strategy. 'Not relevant' stories were often about other medical events. Only 67 (27.5%) of the 'not relevant' stories were not about a medical condition. 65 (26.6%) of the 'not relevant' stories were about a neurological condition—such as Bell's Palsy or complex migraines—where often a stroke was initially suspected but later ruled out; 85 (34.8%) of the stories in this category were about various non-neurological medical conditions. The remaining 27 stories exhibited no obvious common features, and included some stories that were no longer accessible on the web.

The 'skip' category proved to be difficult to analyze, with no dominant feature among its stories. We expected that this category would be dominated by stories that were not accessible on the web at the time of the search, e.g. the posts were removed by their authors. However, only 82 (29.2%) of the 'skipped' posts met these criteria. Instead, this category seemed to overlap significantly with the 'not relevant' category, suggesting that our users had difficulty deciding between these two categories when judging the relevance of retrieval results.

B. Evaluation of relevance feedback

We evaluated whether our approach to relevance feedback reduced the amount of time our users took to collect a sufficiently large corpus for narrative analysis. Specifically, we asked whether relevance feedback after identifying 100

a) Original "boring story" query

I was watching TV and I slumped over on my side. I just didn't feel right. I felt weak and I didn't feel normal on one side. I had a hard time speaking and I could not find the words I wanted to use. My speech did not make sense and was slurred. I felt confused and could not pick up my arm or leg on one side. I felt anxious and upset. I felt scared. My head felt strange. My friend called 911. The paramedics arrived. I was loaded into the ambulance and taken to the emergency room. There, at the triage unit I was diagnosed with a stroke.

b) Relevance feedback query after 100 relevant items

felt:3.739 side:2.873 stroke:2.86 not:1.868 one:1.832 on:1.446 feel:1.379 speech:1.324 could:1.288 didnt:1.27 slurred:1.066 speaking:1.059 arrived:1.054 unit:1.05 called:1.04 taken:1.02 use:1.01 right:1.007 slumped:1.007 arm:1.006 triage:0.989 emergency:0.986 ambulance:0.985 sense:0.98 tv:0.978 loaded:0.976 confused:0.971 find:0.965 anxious:0.957 911:0.955 leg:0.953 pick:0.94 words:0.938 scared:0.936 upset:0.921 diagnosed:0.92 watching:0.916 or:0.906 normal:0.9 strange:0.894 weak:0.892 paramedics:0.879 friend:0.875 hard:0.854 make:0.805 room:0.72 did:0.696 i:0.643 wanted:0.64 my:0.58

c) Relevance feedback query after 100 relevant items, terms selected using TF×IDF weights

stroke:2.86 triage:0.989 slurred:1.066 felt:3.739 speech:1.324 side:2.873 paramedics:0.879 slumped:1.007 911:0.955 ambulance:0.985 unit:1.05 diagnosed:0.92 anxious:0.957 loaded:0.976 emergency:0.986 didnt:1.27 weak:0.892 speaking:1.059 confused:0.971 leg:0.953 arm:1.006 upset:0.921 sense:0.98 tv:0.978 scared:0.936 arrived:1.054 strange:0.894 normal:0.9 taken:1.02 words:0.938 use:1.01 pick:0.94 feel:1.379 watching:0.916 called:1.04 hard:0.854 find:0.965 brain:0.412 friend:0.875 clot:0.195 could:1.288 one:1.832 right:1.007 room:0.72 make:0.805 yves:0.126 not:1.868 wanted:0.64 hospital:0.272 suffered:0.173

d) Optimal query for stroke stories after 100 relevant items (α=1.2, β=1, γ=1)

stroke:4.884 triage:1.179 slurred:1.313 speech:1.806 side:3.309 slumped:1.195 ambulance:1.206 unit:1.301 paramedics:0.952 diagnosed:1.029 felt:3.338 911:0.971 loaded:1.164 anxious:1.116 emergency:1.14 speaking:1.318 weak:0.986 confused:1.147 leg:1.136 arm:1.208 sense:1.157 tv:1.16 upset:1.046 arrived:1.289 scared:1.068 strange:0.981 normal:1.023 taken:1.187 brain:0.848 use:1.219 words:1.058 pick:1.085 clot:0.387 didnt:0.836 watching:1.023 marty:0.34 called:1.187 yves:0.25 find:1.141 feel:1.209 hospital:0.532 hard:0.888 englert:0.13 suffered:0.316 right:1.223 friend:0.908 pedro:0.21 patients:0.244 one:1.996 therapy:0.266

Figure 1. Queries for stroke stories

relevant items yielded a query that was better able to find the remaining relevant items than the original "boring story" query, using the metric of average precision. Here average precision is computed as the mean of precision values at each point in the ranked results where the recall changes. This metric is useful in judging the relative effectiveness of different queries when the relevance of retrieved items are known. The average precision of the original query (Figure 1a) was 0.073, compared to 0.416 for the relevance feedback query (Figure 1b). Ranks of relevant items were significantly improved in the relevance feedback query ($p<0.01$, one-tail sign test).

C. *Evaluation of term selection methodology*

We noted that many common, seemingly insignificant terms were included in the relevance feedback queries. We evaluated whether our approach to the selection of 50 query terms could be improved by factoring in the information value of each term, as indicated by its inverse document frequency in our larger corpus of stories. Specifically, we asked whether it would have been better to select the 50 terms with the highest TF×IDF weights for use as relevance feedback query after identifying 100 relevant items, using the average precision of the remaining relevant items as the metric. The average precision of the TF×IDF query (Figure 1c) was improved to 0.471. Ranks of relevant items were also significantly improved ($p<0.01$). This result indicates that it would have been better if we had incorporated inverse document frequency in our selection of query terms.

D. *Evaluation of optimal Rocchio parameters*

When developing our relevance feedback approach, we selected parameters for our Rocchio implementation based on intuition alone ($\alpha=2$, $\beta=1$, $\gamma=1$). In a post-hoc analysis, we sought to identify the optimal Rocchio parameters based on methods used in previous research (Buckley & Salton, 1995). We performed a grid-search for optimal parameters given incrementally more annotations, from 10 to 100 relevant items in 10 item increments. At each increment, we computed the average precision based on all remaining relevant items, using queries generated by fixing the weight of relevant items ($\beta=1$) and varying α and γ values between 0 and 2 at increments of 0.2. Figure 2 plots the optimal values for α and γ when $\beta=1$. Results indicate that optimal parameters do not vary greatly given different amounts of relevant feedback, with the best results obtained when the original query is given slightly more weight than both relevant and non-relevant items. This result indicates that our intuitions led us to a good guess, but it would

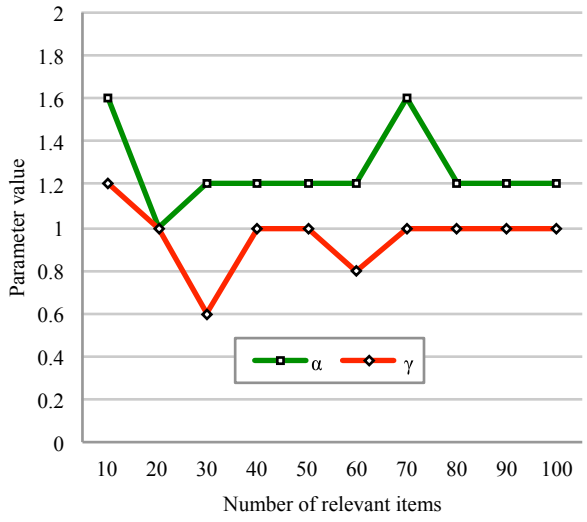

Figure 2. Optimal Rocchio parameters, $\beta=1$

have been better if we had reduced the weight given to the original query during our evaluation. The highest average precision (0.524) was achieved using the query shown in Figure 1d, which is significantly better than the one produced using our original Rocchio parameters ($p<0.05$).

VI. Discussion

The evaluation of our search technologies was formative, yielding several recommendations for how system parameters and our user interface should be tweaked to more effectively identify stories related to specific health issues. However, much of the value of our approach is difficult to assess outside of the context of the subsequent analyses of these narratives, where they are used as evidence in the investigation of specific health-related questions. As these analyses are undertaken in the future, we believe that an important concern will be the sample biases that are inherent to this genre of social media and those introduced by our technical approach.

All health-related investigations have sample biases. They are introduced in the recruitment of experimental subjects, the recording of medical data, and the acquisition of records from institutions. Health narratives from weblogs are not random samples of the experiences of people. Their authors are literate, technologically savvy people that survived the health experience that they have chosen to share publicly. The use of "boring story" initial queries and relevance feedback creates additional sample biases, favoring the inclusion of health narratives that include some terms over others. When the appearance of these terms is strongly correlated with the one possible answer to a research question, then statistics gleaned from these health narratives will be invalid. For example, the sample of narratives collected in our evaluation may not be helpful in investigating the frequency of slurred speech in strokes, as the terms "slurred" and "speech" figured prominently in the evolving queries.

Understanding these additional sample biases will be important when investigators formulate the research questions that they intend to explore using weblog stories, and again when they are analyzing the results of their collection efforts. To better surmise how the sample differs from other populations that have been previously studied, investigators should include in their analyses the annotation of demographic or other information that can be directly compared to published statistics obtained using traditional sampling methods. Where samples appear similar along many dimensions, the unique insights afforded by health narratives from weblogs can be better appreciated with greater confidence in their validity.

Acknowledgment

The authors thank Deborah Burke, Sukjin Koh, and Stephanie Morley for their collaboration in the evaluation of this research. The projects or efforts depicted were or are sponsored by the U. S. Army. The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] Adams, S. (2010) Blog-based applications and health information: Two case studies that illustrate important questions for Consumer Health Informatics (CHI) research. International Journal of Medical Informatics 79: e89 - e96.

[2] Atkinson, P. (1997) Narrative Turn or Blind Alley? Qualitative Health Research 7(3):325-344.

[3] Barr, J., McKinley, S., O'Brien, E., and Herkes, G. (2006) Patient Recognition of and Response to Symptoms of TIA or Stroke. Neuroepidemiology. 26: 168 – 175.

[4] Buckley, C. and Salton, G. (1995) Optimization of Relevance Feedback Weights. Proceedings of ACM SIGIR 1995, Seattle, WA.

[5] Charikar, M. (2002) Similarity Estimation Techniques from Rounding Algorithms, Proceedings of the thirty-fourth annual ACM symposium on Theory of computing, ACM Press, 2002

[6] Charon, R. (2006) Narrative Medicine: Honoring the Stories of Illness. Oxford & New York: Oxford University Press.

[7] Gordon, A. and Swanson, R. (2009) Identifying Personal Stories in Millions of Weblog Entries. ICWSM Data Challenge Workshop, San Jose, CA, May 20, 2009.

[8] Gordon, A. and Swanson. R. (2008) StoryUpgrade: Finding Stories in Internet Weblogs. International Conference on Weblogs and Social Media, March 31-April 2, 2008, Seattle, WA.

[9] Hollenbaugh, E. (2011) Motives for maintaining personal journal blogs. Cyberpsychology, Behavior, and Social Networking. 14(1-2): 13-20

[10] Lisabeth, L., Brown, D., Hughes, R., Majersik, J., and Morgenstern L. (2009) Acute stroke symptoms: comparing women and men. Stroke. 40(6): 2031 – 2036.

[11] Manku, G., Jain, A., and Sarma, A. (2007) Detecting Near-Duplicates for Web Crawling. 16th international conference on World Wide Web, ACM Press, 2007.

[12] Miller, E. and Pole, A. (2010) Diagnosis Blog: Checking Up on Health Blogs in the Blogosphere. American Journal of Public Health 100.8:1514 – 1519.

[13] Munson, S. and Resnick, P. (2011) The Prevalence of Political Discourse in Non-Political Blogs. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, July 17 – 21, 2011, Barcelona, Spain.

[14] Ounis, I., Lioma, C., Macdonald, C., and Plachouras, V. (2007) Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web. Upgrade 7(1):49-56.

[15] Pratt, W., Unruh, K., Civan, A., and Skeels, M. (2006). Personal Health Infomration Management. *Communications of the ACM* 49.1:51-55.

[16] Rocchio, J. (1971) Relevance Feedback in Information Retrieval. In Salton: The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, 313-323.

[17] Thomas, C. (2010) Negotiating the contested terrain of narrative methods in illness contexts. Sociology of Health & Illness 32(4):647-660.