External Information Sharing on Health Forums: An Exploration

Dana M. Nguyen, 1 Alexandra Olteanu, 2 Emre Kıcıman 2
1 University of California, Santa Barbara
2 Microsoft Research
dananguyen@ucsb.edu, {alexandra.olteanu, emrek}@microsoft.com

Abstract

Online health forums are an important avenue for receiving support and learning about fellow patients’ experiences with similar diagnoses. We seek to characterize the external information shared (via web links) on health forums as a proxy to participants’ information needs. For this purpose, using a dataset of web links shared publicly on a lung cancer forum over a period of 16 years, we perform a comparative analysis with three different website typologies, uncovering a diverse ecosystem of websites. We also examine typological variations as this forum gains and then loses popularity over time.

Introduction

Over the past decades, patients and caregivers have turned to a diversity of online resources for support and information about their or their loved one’s health issues. Such resources range from health and well-being news and information websites (like WebMD) to a rich ecosystem of health communities formed on dedicated forums or social media platforms (like csn.cancer.org hosted by the American Cancer Society or r/Health/ on Reddit). A 2013 Pew Research study reported that about 7 in 10 US internet users searched for health topics online (Fox and Duggan 2013)–up from 38% in a ’98 Harris Poll (The Harris Poll 2011). Their searches included diverse topics such as medical diseases and diagnoses, treatments and procedures, prescription drugs, and health insurance. Patients have also taken on a more active role by striving to produce and maintain reliable online resources that include “the best online links, the best medical centres, the best treatments, and the latest research” and cover “topics that clinicians may consider secondary but are very important to the patient—their quality of life, the impact of their disease on their friends and family, and the psychological aspects of their illness” (Ferguson 1996).

Within the ecosystem of online health communities, health forums remain a critical avenue to directly seek out and exchange support, information, and stories about health conditions. Patients often seek health information online as a key complement to information received from their doctors (Zhao and Zhang 2017; Ofran et al. 2012), using social platforms to satisfy different kinds of information needs than they do via web search (Morris, Teevan, and Panovich 2010). While the information shared on health forums is centered around personal experiences and word-of-mouth information from friends and family (Bond and Ahmed 2016; Zhao and Zhang 2017), users also bring in information from external sources via web links to external websites.

The process of sharing and discussing information on health forums is akin to collective sensemaking, wherein participants—by posting, questioning, iterating—construct a shared knowledge structure (Mamykina, Nakikj, and Elhadad 2015). Under this framing, suggesting external resources is a key step in the process of collective sensemaking. Our descriptive study examines what external resources are being suggested in a health forum over time, and how that is changing. This provides context for research on the information needs of health patients, data mining of health forums and social media activities, and other work that contextually interprets the activities of health communities.

Figure 1: Normalized temporal patterns in information sharing on a health forum over a 16 years period. See in color.
Table 1: Dataset figures, including for the first vs. last 8 years. P – average prevalence of URLs (resp. posts/threads with URLs) per 100 posts (resp. threads). * indicates estimated values due to e.g., expired shortened URLs.

<table>
<thead>
<tr>
<th></th>
<th>Total Posts</th>
<th>Posts &lt;2011</th>
<th>Posts &gt;2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posts</td>
<td>369,284</td>
<td>244,257</td>
<td>25,927</td>
</tr>
<tr>
<td>Threads</td>
<td>38,209</td>
<td>34,451</td>
<td>3,758</td>
</tr>
<tr>
<td>URLs</td>
<td>28,302</td>
<td>7,548</td>
<td>34,451</td>
</tr>
<tr>
<td>Posts with URLs</td>
<td>11,112</td>
<td>3,841</td>
<td>7,267 (P = 8.8)</td>
</tr>
<tr>
<td>Threads with URLs</td>
<td>9,490</td>
<td>24.8</td>
<td>21.9 (P = 8.8)</td>
</tr>
<tr>
<td>Unique URLs</td>
<td>11,655</td>
<td>9,141</td>
<td>2,314</td>
</tr>
<tr>
<td>Unique Domains</td>
<td>4,645*</td>
<td>3,841</td>
<td>1,135</td>
</tr>
</tbody>
</table>

To characterize the websites being shared, we started with a qualitative review of small samples of 25-50 URLs, containing either full URLs or their domains. We see patterns emerging from the open coding of the full URLs: a majority of webpages are health related, with a sizable fraction linking back to forum posts or to other links on the hosting NGO website. Among these are self-help websites (health information), research articles (medical research) and coverage of such articles (health news), drug manufacturing, hospital or other health service providers’ websites (health care services), as well as links to other health forums. When coding at domain level, other types of sites also emerge including personal blogs, diaries, or memorial sites, as well as fundraising websites. Links to NGOs, mainstream news, and governmental websites were also present in these samples. Drawing from these qualitative insights and our focus on the types of information shared via web links, we center our analysis on the URL domains.

**Websites Coding**

To examine the information shared via web links, we juxtapose multiple website typologies, including a domain typology drawn from prior studies of online health communities and our qualitative exploration and two existing open-domain and general purpose typologies of online content. This also ensures that our observations are not an artifact of a certain typology. Throughout the paper, for comparison purpose (when applicable) we group similar categories.

**Webpage version.** Given that many websites in our dataset were shared several years back, we used the Internet Archive’s API (https://archive.org) to recover their archived versions. When available, we based the typological assessments on the version of the website archived closest to the time the website was initially posted onto the forum.

**General purpose typologies.** We classified web domains using two web typologies: 1) the Open Directory Project (ODP) classification (AOL, Inc. 2016), and 2) the Interactive Advertising Bureau (IAB) classification (IAB Tech Lab 2017). To categorize domains according to the IAB typology we used the WebShrinker API (DNSFilter, Inc. 2020), while for ODP we used an existing classifier (Bennett, Svore, and Dumais 2010). We could not categorize 12.3% of URLs with ODP, and 7.5% with IAB (in part due to outdated links) and removed them from the analysis.

**Domain specific (DS) typology.** In addition to these typologies—designed to be open domain and for general purpose—we also developed a typology grounded in our application domain (see Table 2). To identify the most likely category for each URL domain, we employed crowd judges to manually annotate them by selecting the most applicable category from our typology. A brief definition and examples for each category were also provided, and the judges had to pass a qualification task before being allowed to annotate. We randomly sampled 1200 domains, and for each domain we gathered 3 annotations and kept the majority label.

**Exploratory Analysis & Discussion**

We begin with an overview of general patterns in web links sharing. To address our guiding RQs, we then use the typology...
Table 2: The domain typology of websites (with examples) and the relationship to aspects mentioned in prior work.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Mentions in related work</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainstream News Media</td>
<td>news media sites that cover a diversity of topics</td>
<td>celebrity news (Ofran et al. 2012; Quinn et al. 2013); news sources (Madden et al. 2012)</td>
<td>cm.com, bbc.com, news-wise.com</td>
</tr>
<tr>
<td>Health News &amp; Info.</td>
<td>medical &amp; health related news articles, blogs, information or education sites</td>
<td>technical medical information (Ferguson 1997); specialized websites (Fox and Duggan 2013); health information (Devin et al. 2016); cancer-related blogs, educational sites (Quinn et al. 2013) encyclopedic medical sites (Madden et al. 2012)</td>
<td>cancerindex.org, emedicine.com, webmd.com</td>
</tr>
<tr>
<td>Medical/Health Research</td>
<td>health and medical research articles, journals or proceedings, or research institutions</td>
<td>research on conditions or treatments (Ferguson 1996; De Choudhury et al. 2014; Ofran et al. 2012); .edu websites (Nath et al. 2016); academic, technical &amp; scientific resources (Gliwniak 1995; Madden et al. 2012); peer-reviewed journals (Quinn et al. 2013; Reynolds et al. 1995)</td>
<td>health.ucdavis.edu, nichd.nih.gov, pubmed.org, rider.edu</td>
</tr>
<tr>
<td>Health Care Services</td>
<td>doctors &amp; other health care providers, pharmaceutical &amp; other health services, products &amp; tech</td>
<td>medical centers, treatments (Ferguson 1996; Gliwniak 1995; Madden et al. 2012; Barney, Griffiths, and Banfield 2011); commercial, health facility (Quinn et al. 2013); health products &amp; services (Devin et al. 2016)</td>
<td>spectrumhealth.org, ppax.org, novax.com, eduma.com</td>
</tr>
<tr>
<td>Health Forums</td>
<td>forums and online messaging boards focused on health related issues</td>
<td>support groups &amp; self-help communities (Ferguson 1997); online health or patient communities and forums (Nath et al. 2016; Elhadad et al. 2014; Kanthawala et al. 2016; Huh, Patel, and Pratt 2012; Mao et al. 2013); health discussion forums (Quinn et al. 2013; Bond and Ahmed 2016); medical message boards (Benton et al. 2011, Barney et al. 2011)</td>
<td>cancergrace.com, healthboards.com, forums.cancerhealth.com</td>
</tr>
<tr>
<td>Fundraising &amp; NGOs</td>
<td>non-profit organizations, fundraising &amp; awareness campaigns (by both for- &amp; non-profit entities)</td>
<td>for-profit organizations (Nath et al. 2016; Madden et al. 2012); medical crowdfunding (Kim et al. 2017); non-profit (Madden et al. 2012; Nath et al. 2016); charity (Quinn et al. 2013); support organizations, awareness, and novelty items (Ofran et al. 2012)</td>
<td>speciallove.org, jimmyfund.org</td>
</tr>
<tr>
<td>Governmental Sources</td>
<td>governmental websites &amp; information, not necessarily on health</td>
<td>.gov websites (Nath et al. 2016); government-initiated webpage (Kanthawala et al. 2016); governmental agencies (Madden et al. 2012; Quinn et al. 2013)</td>
<td>ssa.gov, statistics.gov.uk</td>
</tr>
<tr>
<td>Social Media</td>
<td>any type of social media sites</td>
<td>social media (Mao et al. 2013; Griffiths et al. 2014; Ofran et al. 2012); social network (Fox and Duggan 2013; Quinn et al. 2013); discussion boards, social bookmarking (Koskan et al. 2014)</td>
<td>twitter.com, facebook.com, reddit.com</td>
</tr>
<tr>
<td>Personal Sites</td>
<td>personal sites like blogs, personal diaries, memorial websites</td>
<td>personal websites (Kanthawala et al. 2016; Nath et al. 2016); individually run websites (Quinn et al. 2013); personal patient blogs (Guaitieri and Akhtar 2013)</td>
<td>bonobike.blogspot.com, griefhealing.com</td>
</tr>
</tbody>
</table>

Figure 3: Variations in cross topical entropy over time.
When measuring the dispersion of shared domains \(^4\) — i.e., are many domains shared, or are mostly a small number of domains shared — using two different metrics, entropy and Gini coefficient of inequality (Fig. 5), we see that there is a significant decrease in entropy and increase in Gini over time. This indicates that, indeed, a smaller set of external resources are collecting a greater degree of the attention.

**Implications & prior work.** Changes over time can reflect shifts in both the needs of individuals and of their communities. As users get diagnosed or as the communities establish, users may be more likely to seek information about disease and treatment related information (Ofran et al. 2012). As their treatments evolve or as the shared knowledge structure matures and saturates, their focus may also shift.

**Conclusions & Limitations**

As social platforms became popular, so did website sharing as a way to discover information via word-of-mouth (Rodrigues et al. 2011). The history of self-help websites and health communities, and of the efforts to understand these communities and their experiences goes over two decades back (Rice and Katz 2000; Ferguson 1997; Reynolds, Sharma, and Jack 1995; Ferguson 1996). Though empirical and descriptive in nature, our longitudinal study extends this literature by analyzing patterns in external resource sharing over the course of 16 years on a cancer forum, showing that such online communities appear to maintain a focus on health information, even as the forum loses popularity and as we see a drop in the diversity of content being shared.

**Limitations & future work.** While the scope of our study is limited by ethical considerations regarding re-use of sensitive, publicly shared content for research purposes, our findings motivate future investigations of broader scope. We observe news is shared more than government resources (Fig. 4), and the degree of attention to domains becomes unequal over time (Fig. 5): Does this behavior also appear in other websites or the web in general? Could the inequality be a side effect of search algorithms promoting sites that are already popular? Our analysis also does not capture variations of shared content (and its’ heterogeneity over time) across distinct health forums and offers limited cues about factors that cause temporal variations. Future studies may

\(^4\)To avoid a skew in the Gini coefficient from infrequently shared domains, we include only web domains posted in \(\geq 4\) years.

---

**Figure 4:** Topical trends over time across the three typologies. The inset plots show the normalized distribution. See in color.

**Figure 5:** Yearly entropy & Gini coef. for the number of posts per web domains. \(p < 0.01 (***)\). See in color.

**Typological variations over time (R2).** By analyzing the entropy of the typological distribution (Fig. 3), we observe that as the forum loses popularity, the entropy also decreases — consistent with prior work showing a convergence of norms and conventions, e.g. (Danescu-Niculescu-Mizil et al. 2013), particularly as a community shrinks. Inspecting the distribution of topical interests as the popularity of the forum declines over time, we see that the core focus of the remaining participants is on health (IAB) topics, while ancillary topics become less prominent among shared web links. We hypothesize that the decrease in cross topical entropy could, in part, be a result of a convergence toward a smaller set of external resources that users turn to.
juxtapose patterns of web link sharing across multiple forums via deeper content, contextual and behavioral analyses.

References

Eschler, J.; Dehlawi, Z.; and Pratt, W. 2015. Self-characterized illness phase and information needs of participants in an online cancer forum. In *ICWSM ’15*.