Information Fusion 46 (2019) 33–43

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Characterizing online health and wellness information consumption: A study

Aastha Nigam[a], Reid A. Johnson[b], Dong Wang[a], Nitesh V. Chawla[*,a,c]

[a] Interdisciplinary Center for Network Science and Applications (iCeNSA), Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA

[b] Concur Technologies, Bellevue, WA, USA

[c] Department of Computational Intelligence, Wrocław University of Science and Technology, Wrocław, 50-370, Poland

ARTICLE INFO

Keywords:
Online health seeker
User behavioral analysis
Factorization models
Content consumption

ABSTRACT

To seek answers to health queries, we often find ourselves on a quest to assimilate information from varied online sources. This information search and fusion from different sources elicits user preferences, which can be driven by demographics, context, and socio-economic factors. To that end, we study these factors as part of health-information seeking behavior of users on a large health and wellness-based knowledge sharing online platform. We begin by identifying the topical interests of users from different content consumption sources. Using these topical preferences, we explore information consumption and health-seeking behavior across three contextual dimensions: user-based demographic attributes, time-related features, and community-based socio-economic factors. We then study how these context signals can be used to explain specific user health topic preferences. Our findings suggest that linking demographic features to user profiles is more effective in explaining health preferences than other features. Our work demonstrates the value of using contextual factors to characterize and understand the content consumption of users seeking health and wellness information online.

1. Introduction

The increasing volume of health and wellness information now available online has transformed the pursuit of self-diagnoses and assessment on a multitude of disease conditions or wellness goals [1–9]. The query patterns and demographics of online consumers of health and wellness information hold the potential to provide insight into health concerns and choices at a population level, albeit segmented by region and time.

While many confounding factors may impact an individual's health, the main influences can be thought of as a series of social determinant layers [10], beginning with general socio-economic, cultural and environmental conditions that include housing, education and employment, followed by social and local community networks, and culminating with individual lifestyle factors that include personal habits (e.g. food choices and smoking) and attributes (e.g. age, gender, and race) [11–14]. Together, these factors influence the health of an individual and subsequently inform their information consumption patterns and online health seeking behavior. While characterizing health needs and interests is deemed as a challenging problem, increased dependency on specialized health websites [15] such as WebMD Health,[1] Everyday Health[2] and Healthline[3] provides a novel opportunity to empirically understand health interests and information-seeking behavior at scale.

These massive user (and/or population) centered datasets not only allow us monitor individual health concerns, but provide an exciting opportunity to glean insights about the factors guiding information search and consumption on health issues, especially through the lenses of demographics, socio-economic factors, and regions (derived from public data sources such as the census data). Fusing the online health consumption patterns with community data, such as socio-economic factors and demographics, can allow us to begin to understand a broader story about social-determinants of health and wellness, as well as factors that largely drive the online search. This can not only help us in personalizing a consumer's experience, but at a community level, it can also be used to develop a deeper understanding of societies and population segments to build better-informed health systems [16].

Most previous research has focused on using single data source such as surveys, interviews, questionnaires [14,17–24] to understand the factors governing search and consumption of healthcare information.

* Corresponding author at: Interdisciplinary Center for Network Science and Applications (iCeNSA), Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA.
E-mail address: nchawla@nd.edu (N.V. Chawla).

[1] http://www.webmd.com/ .

[2] https://www.everydayhealth.com/ .

[3] http://www.healthline.com/ .

https://doi.org/10.1016/j.inffus.2018.04.005
Received 1 November 2017; Received in revised form 2 April 2018; Accepted 23 April 2018
Available online 24 April 2018

Table 1
Diversity of the counties in terms of area, population, unemployment rate, median age and median income based on 2010 U.S. Census. ID is arbitrarily assigned.

ID	County	State	County seat	Area (in mi^2)	Population (in Millions)	Median age	Unemployment rate	Median income (in USD)
0	St. Joseph County	IN	South Bend	461.38	0.266	36.4	4.9%	45,012
1	Cook County	IL	Chicago	1635	5.194	35.7	6.2%	54,828
2	King County	WA	Seattle	2307	1.931	37.2	3.3%	73,035
3	New York County	NY	–	33.6	1.585	36.6	5.2%	71,656
4	Santa Clara County	CA	San Jose	1304	1.781	36.6	3.8%	93,854
5	San Mateo County	CA	Redwood City	744	0.718	39.4	3.2%	91,421
6	San Francisco County	CA	–	231.89	0.805	38.6	3.4%	78,378
7	Orange County	CA	Santa Ana	948	3.010	36.7	4.1%	75,998
8	Marion County	IN	Indianapolis	403.01	0.903	34.0	4.9%	42,378
9	Bronx County	NY	–	57	1.385	33.1	8.3%	34,384
10	Queens County	NY	–	178	2.230	37.5	5.4%	57,210
11	Richmond County	NY	–	102.5	0.468	39.0	6.1%	74,043
12	Kings County	NY	–	97	2.504	34.3	3.3%	46,958
13	Multnomah County	OR	Portland	466	0.735	36.3	4.6%	52,845
14	Providence County	RI	Providence	436	0.626	37.0	6.0%	49,139
15	Denver County	CO	Denver	155	0.600	34.0	4.2%	51,800
16	Suffolk County	MA	Boston	120	0.722	31.9	3.8%	54,169
17	Bexar County	TX	San Antonio	1256	1.714	33.1	3.4%	50,867
18	Travis County	TX	Austin	1023	1.024	32.7	2.9%	59,620
19	Dallas County	TX	Dallas	909	2.638	32.9	3.9%	49,925
20	Jackson County	MO	Kansas City	616	0.674	36.4	6.7%	46,917

More recently, researchers have also leveraged search queries and logs retrieved from search engines to identify health goals and to study evolution of illnesses among users [25–27]. However, these studies have been limited in their fusion of information or data across multiple sources.

We present a novel study that utilizes data from one of the largest online health content providers in the United States of America (USA). The provider manages health and wellness content through several focused knowledge sharing websites allowing users to consume articles on a variety of health topics. Our data, comprising of deidentified 1.2 million users, includes user logs of interactions through newsgroups, forums, and site-specific searches, and records data associated with various user actions. We posit that even if the website presents only a segment of possible avenues of procuring information, it still empowers us to better understand individual and population-level health interests, as the individuals are subscribing to or searching for health related information on this website that has a wide reach across the country. Further, we fuse this online data with census data to obtain aggregates for population segments. We focus on personal factors that are either self-reported by the user, such as demographics, or that can be inferred from their activity, such as browsing time, in combination with community driven factors that can be inferred from their self-reported location, such as socioeconomic status.

Such large scale analysis can help discern emerging health needs, improve health outcomes, and aid in public health surveillance systems, targeted care management and healthcare resource allocation [16,28–32]. For example, arming health-care providers and resource allocators with the knowledge that at a given time middle-aged women in *county A* are more interested in articles about *Calcium Deficiency* while middle-aged women in *county B* are more interested in content related to *Gonorrhea* can help them design care that is better targeted to the health needs of these population segments.

Contributions. We address two primary research tasks: 1) characterizing the large-scale consumption of online health and wellness information across the user-based contextual dimensions of demographics, time and socio-economic attributes, and 2) explaining individual health-related topical preferences and studying the influence of user-based contextual signals in their determination. Our findings suggest that information consumption is heavily impacted by gender, age and location. We also find that a factorization model based on users' demographic attributes statistically significantly outperforms all other models studied ($p \ll 0.01$), producing the lowest error (RMSE = 0.3217) in modeling user interest. Finally, we provide novel insights about differences in health concerns across regions, irrespective of the age and gender, suggesting that we can leverage this targeted information fusion to address population health management within resource-constrained environments such as community health programs where it can be used to provide insight and resource planning into targeted care for the population served.

2. Data characteristics

Our work uses data from a large U.S.-based digital health content provider. The company manages several websites that serve as health and wellness knowledge-sharing communities. On each website, users can navigate through hyperlinked content or search for specific health and wellness content via built-in search options. Users can select to register on these websites, permitting the company to interact with them through forums and published newsletters sent to user mailing lists.[4]

During user registration, the company requests information on the user's age, gender and location. To populate an initial set of preferences, users are also encouraged to select health topics of interest from a predefined topic list. The available topics cover a wide-range of health-related subject matter (e.g., *cancer, pregnancy, senior health*, etc.). Each article available for consumption is manually annotated by company domain experts with an appropriate health topic based on its content. For instance, an article describing life practices to prevent diabetes would be tagged with *diabetes* by the company. Note, we use topic and content consumption interchangeably.

The health-based knowledge sharing social platform has found widespread use across the continental United States. We explore data collected from 1,263,426 users residing within 21 regionally diverse U.S. counties. Details on these counties are provided in Table 1 (which includes a manually annotated ID for convenient reference throughout the paper). Based on the 2010 U.S. Census,[5] the counties vary substantially according to measures such as population (listed in Table 1), area, unemployment rate, median age and median income.

For this research, we infer user's interest from three sources: 1) **Topic Selection (S1)**: Health topic(s) selected by the user from a list of

[4] Company-provided anonymized data. Can be shared upon request.
[5] http://factfinder.census.gov/ .

predefined topics (some at the time of registration, however a user can select topics at any point); 2) **Newsletter (S2)**: Topics corresponding to the articles user clicked on from the company newsletters via the user mailing lists and newsgroups; 3) **Website Search (S3)**: Topics corresponding to the articles returned to a user from their search on the company's content. The website search data only records the articles that a user clicks on and the date of the click. The search query or any external web search engine information is not recorded by the company.

As discussed earlier, one of our research goals is to study the effect of demographics on a user's reading profile. To study this effect, we investigate user demographics across three self-reported attributes: age, gender and county. In total, we provide analysis on 1,263,426 users, of whom 1,085,298 (85.9%) are female and 178,128 (14.1%) are male. That females represent a disproportionately large number of the registered users is partially attributable to one of the company's websites providing only maternity-based content. Previous work has also shown that women tend to seek health information more actively than men [18,22,33–35]. Further, the provided data categorizes users into 6 distinct age groups (defined in years): 18–24 (13.2%), 25–34 (33.3%), 35–44 (21.6%), 45–54 (14.4%), 55–64 (11.6%) and 65–80 (5.6%).

The data was collected between 2006 and 2015 and covers information across 263 unique health topics, with each topic identified by a numeric ID annotated by the company. A user's topic profile can be described as a list of topics they are interested in and their corresponding activity for those topics. From among the users who provided age, gender and county information, we were able to build topic profiles for 75%, resulting in 952,078 users with topic profiles; the remaining users did not perform any of the actions used to build a topic profile, and are thus omitted from the remainder of this work.

3. Modeling and analysis methods

Our work leverages data from active users of a large-scale specialized health website to characterize content consumption and infer interests, without data from surveys and questionnaires. User interests are derived from heterogeneous sources including reading activity and explicitly reported preferences, and do not rely on search logs. In our study, the user interests range from disease-specific topics to general well-being issues. We characterize consumption and interest across the three contextual dimensions of demographics, time and socioeconomics, and study how it varies between different segments of the population as discussed in Section 3.1. Additionally, in Section 3.2, we present a supervised approach to explain individual preference towards a given topic using these contextual dimensions.

3.1. Health seeking behavioral analysis

We characterize health seekers and study variability in information consumption across three contextual dimensions: demographics, time and socioeconomics. For each dimension, we investigate multiple factors to draw insights into the health seeking behavior of individuals.

As previously discussed, demographics of an individual can influence the choices of a health seeker. To examine, the role of demographics, we consider self reported user attributes such as age, gender and location for the analysis. Further, for the time dimension, we derive various factors leveraging user actions based on sources (described in the previous section): topic selection (S1), newsletter (S2) and website search (S3). Since each action is associated with a timestamp representing the date, month and year, we compute the following factors for the time dimension:

- **User activity**: Defined as the total number of actions performed by the user in each category (S1, S2 and S3).
- **User involvement**: Defined as the average time a user spends consuming information via each action type. To determine the average consumption time, we identify the first and last timestamps for a user seen on a given action type (S1, S2 and S3) and compute the total time the user spent consuming information through the given medium.
- **User engagement**: User engagement is defined based on the average period of time elapsed between two actions of the same kind performed by a user. To compute this metric, for each category, we calculate the mean difference in time between two consecutive actions related to that category. In general, a lower average time difference for a category suggests higher user engagement.
- **User inclination**: User inclination can be measured as the time difference between the user's initial enrollment date and the timestamp recorded for the first topic action performed by the user in each category.

While, the time factors allow us to examine consumption based on user activity, to fully explore information consumption, we also consider a third dimension retrieved from census data: socioeconomic factors derived from the region (i.e., county) in which a user resides. More specifically, we investigate four factors: population, poverty, unemployment and education. While these factors might not have a direct or immediate association with a user's choices, they may nonetheless affect their health-seeking behavior.

We note that, with this analysis, our goal is not to generate recommendations, but to explore the underlying structure of the available factors and identify the insights they may provide.

3.2. Explaining topical preference

With the previous analysis, our aim is to study aggregate consumption of health information according to demographic, time, and socioeconomic factors. In this analysis, we present a framework to explore if these factors can be used to explain each health consumer's preference for a given topic. For example, if a user is mostly interested in articles on *Asthma*, can we leverage the previously presented contextual factors to explain this preference? Does the age and gender of the individual affect their interest in *Asthma* or is it mostly influenced by the community they live in or the interplay of all three dimensions cause the interest?

Our framework begins by defining a sparse topic preference vector that captures each user's preference towards a given set of topics. We then use factorization machines, a class of models that works well under high sparsity, to identify the factors that contribute most to user preferences.

While, many factors, such as medical history, can influence the preference of a health seeker—we leverage the demographic, time, community-driven factors explored in the previous analysis to explain the topical preferences. We note that our goal is not to predict user preference or to provide recommendations, but instead to understand factors that best explain individual preferences. By better understanding these preferences, health and wellness management services/systems can deliver, engage and empower each consumer in a more targeted way.

Given the problem statement, we next explain the method used to capture preference representation of an individual followed by a detailed description about the models used for preference understanding.

3.2.1. Preference representation

A nearly infinitely diverse array of health content is available on the Internet. However , user interests are typically restricted to a limited set of topic choices, with varying levels of preference towards them. Obtaining explicit feedback or ratings from users on their topic interests can be challenging. Nonetheless, implicit feedback can be estimated from user's online reading habits—specifically from the web pages they visit and the topic associated with those web pages. Their preference towards a given topic can also be approximated by the number of

articles they have consumed on the topic. In this work, we infer user topical interests from the previously defined sources - S1, S2 and S3.

To capture user preferences, we computed a topic preference vector for each user [36]. Formally, we represent this vector as $TP_{uj} = [tp(1), tp(2), ..., tp(j)]$, where j is a subset of topics consumed by user u and $tp(j)$ is the preference score given to topic j by user u. The score is calculated as the proportion of articles consumed by user u on topic j from the total number of consumed articles (normalized to a 0–1 scale). For example, if user u has read 3 articles on *breast cancer* and 1 article on *diabetes*, then TP_{uj} would be given as [Breast Cancer: 0.75, Diabetes: 0.25].

3.2.2. Factorization models

Factorization machines (FM) [37] are general predictors that leverage factorized interactions between real-valued feature inputs. FM perform well under high sparsity and are able to mimic various factorization models by varying the feature vector, providing a flexible means of feature engineering. In this work, we use different sets of engineered features as input to the same FM algorithm (implemented via libFM) to evaluate the applicability of a suite of factorization models to our problem. We note that a deep mathematical relationship between FM and other factorization models has already been established [38].

We define the data matrix as $X \in \mathrm{IR}^{n \times m}$, where n is the number of instances and each $x_i \in \mathrm{IR}^m$ capture m-real-valued variables (feature vector). The target vector for each x_i is given by y_i, as shown in Fig. 1. A FM can model up to order d-way variable interactions. A FM model with order degree = 2 is defined as:

$$\hat{y}(x) := w_0 + \sum_{i=1}^{m} w_i x_i + \sum_{i=1}^{m} \sum_{j=i+1}^{m} x_i x_j \sum_{f=1}^{k} v_{if} v_{jf}, \tag{1}$$

where the model parameters are $w_0 \in \mathrm{IR}$, $w \in \mathrm{IR}^n$, $V \in \mathrm{IR}^{m \times k}$. w_0 is the global bias, w is a vector of weights for individual interactions and V encapsulates the factored parameters for two-way interactions for k factors.

As shown in Fig. 1, each feature vector x_i is composed of **U** binary user bits to indicate an active user and **I** binary topic bits to denote the topic selection. Other related features may also be appended to this feature vector, such as **D** demographic, **T** time and **S** socioeconomic features. Using FM framework, we are able to infer the effect of each set of features on user preferences while including the pairwise interactions between these factors.

Using the framework explained above, we employ 6 different models, 4 of which are FM variants. For each model, the target y is obtained using the topic preference vector for the user-topic pair under consideration. We treat this as a regression problem where, given the features (x_i), we learn a model to estimate the target variable (y). The models are:

1. **FM-All**: All contextual information—namely, user demographics, time features about user actions and county-level socioeconomic data—is encoded into the feature, along with the current user and topic pair. The feature vector x_i is as shown in Fig. 1.
2. **FM-Demographic**: We aim to understand the effect of a user's demographics on their preference for given a topic. Our demographic features include the user's age, gender and location (county user resides in) information. In this model, we capture interaction between pairs of attributes, such as a user's age and gender. The feature vector x_i is given by Eq. (2). This resembles the attribute aware model [39], using user information and user-item pairs, as proved in [38].

$$x_i = \underbrace{0, 0, 1, ..., 0, 0}_{\mathrm{U}}, \underbrace{0, 0, 0, ..., 1, 0}_{\mathrm{I}}, \underbrace{d_1, d_2, ..., d_D}_{\mathrm{D}} \tag{2}$$

3. **FM-Time**: To study the contribution of time features to a user's reading habits, we use the time features as described in Section 3.1. The feature vector x_i is given by Eq. (3), where t_t are the time feature values:

$$x_i = \underbrace{0, 0, 1, ..., 0, 0}_{\mathrm{U}}, \underbrace{0, 0, 0, ..., 1, 0}_{\mathrm{I}}, \underbrace{t_1, t_2, ..., t_T}_{\mathrm{T}} \tag{3}$$

4. **FM-Socioeconomic**: The feature vector x_i, given by Eq. (4), can be used to analyze individual and pair-wise interactions of county-level poverty, population, education and unemployment data on reading choices, where s_s are the socioeconomic features.

$$x_i = \underbrace{0, 0, 1, ..., 0, 0}_{\mathrm{U}}, \underbrace{0, 0, 0, ..., 1, 0}_{\mathrm{I}}, \underbrace{s_1, s_2, ..., s_S}_{\mathrm{S}} \tag{4}$$

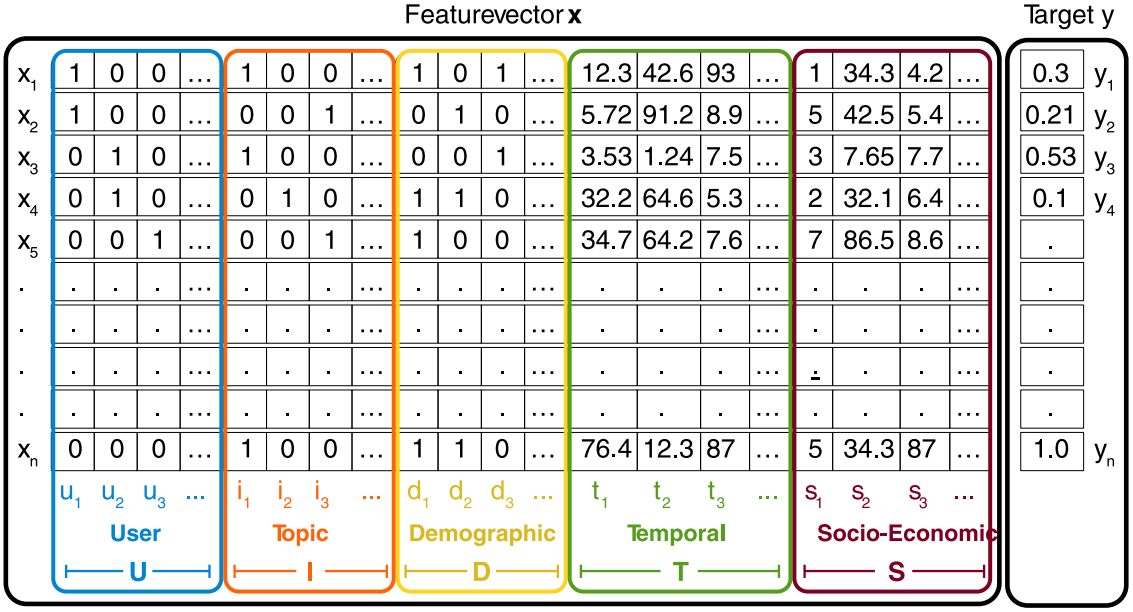


Fig. 1. Example of a factorization machine (based on [38]) with random mock-up data. A feature vector x_i contains a representation of U users, I topics, D demographic, T time and S socioeconomic features. Each x_i is associated with a target y_i which is the topic preference score for the given user and topic.

5. **Matrix Factorization**: Using the feature vector x_i as described in Eq. (5) corresponds to a biased matrix factorization model [38,40,41]. Latent factors based on the user-item interactions are used to predict user preference.

$$x_i = \underbrace{0, 0, 1, ..., 0, 0}_{\text{U}}, \underbrace{0, 0, 0, ..., 1, 0}_{\text{I}} \tag{5}$$

6. **User Attribute Aware**: The feature vector x_i given by Eq. (6) first encodes the active user using U bits followed by user attributes—age, gender and county—without encoding topic information. This model is similar to that proposed in [42], as shown by Rendle [38]. In this model, we learn the mapping of attributes to latent factors.

$$x_i = \underbrace{0, 0, 1, ..., 0, 0}_{\text{U}}, \underbrace{\underbrace{1}_{\text{gender}} \underbrace{0, 1, 0, 0, 0, 0}_{\text{age}} \underbrace{1, 0, ..., 0}_{\text{county}}}_{\text{D}} \tag{6}$$

4. Results and discussion

In the following sections, we present our results and discussions for two analyses as previously described: 1) Health seeking behavioral analysis, and 2) Explaining topical preference.

4.1. Health seeking behavioral analysis

In this section, we present how user activity and consumption patterns vary across three key dimensions: demographic, time and socioeconomics. Demographic features are self reported whereas the time features are derived from the user activities. Further, socioeconomic features are obtained using census data.

4.1.1. Demographic factors

Demographics can play a vital role in determining the kind of health articles a user will consume and in governing their interests. For example, an 18-year-old woman's health interests are likely to be very different from those of a 60-year-old man. Similarly, a user's location can play a critical role in influencing interests. For example, a user in an urban city may have very different health interests than a user in a rural area.

As shown in Fig. 2, we study three demographic factors: age, gender and location (i.e., the county in which a user resides). We observe that the users are not uniformly distributed across the counties, with most users residing in counties 0–5. We also observe that the ratio of female users to male users is high and constant across different age groups, though there is a higher population of male readers in the older age segment compared to other ages. Fig. 3 displays the number of active users—that is, users who consume content and from whom topic interests can be inferred—for each county relative to the base population. We observe that the participation is independent of the base population, and that counties with fewer users also show a high level of participation.

To investigate individual reading habits, we analyzed the unique number of topics in which each user expressed interest. As shown in Fig. 4a, we observe a heavy-tailed distribution, with few users expressing interest in many topics. On average, most users consume around two distinct topics, suggesting that their interests are focused and specific.

So far, we have analyzed basic characteristics of the data spread. Next, we present results for the three demographic features: gender, location and age.

User gender. The topics most frequently consumed by both males and females were those relating to general health conditions such as *Diet & Nutrition, Sexual Health, Weight Management, Pregnancy* and *Diabetes*. However, while these topics found support across genders, we did find several topics that demonstrated a distinct stratification according to gender. For example, articles on *Testicular Mass (Varicocele, Hydrocele, Spermatocele)* and *Enlarged Prostate/BPH* were read exclusively by men, whereas content on topics such as *Calcium Deficiency* and *Diabetes & Hormones* were read exclusively by female users.

User location. To study how topical interest varies according to region or location, we analyzed the variation in consumption/interest of 263 health topics across the 21 counties. Fig. 5a reports the distribution (standardized on topics) of users interested in topic t that reside in county c. The gradation in intensities support our assertion that degree of interest in a particular topic may vary according to region. For example, we observe a high rate of consumption for topic 215 (*Hearing Disorders and Deafness*) in county 13 (Multnomah County). As conditions relating to *Hearing Disorders and Deafness* are more commonly observed in older age groups, this disproportionate interest may be accounted for by the relatively large aging population in Multnomah County (refer to Fig. 2).

User age. We further analyzed how the activity of a user changes with age. As shown in Fig. 4b, the activity of the users increases with age, and female users seek information more actively than men till the age of 45–54. However, men in the ages of 55–85 are more active than female users. In addition, we observe an increase in the activity for older age groups, despite fewer users in that age segment. Our findings suggest that while older individuals may be more hesitant to consume health information online, those enrolled with health content websites tend to be very active.

To investigate the connection between users' age and health interests, we analyzed the 10 most popular topics for each age group. Table 2 lists the results for each group. Interestingly, we find that there are specific health choices that appear to set users of different ages apart. For example, articles on topics such as *menopause, heart health* and *emotional health* are more actively consumed by the older population groups. By contrast, we observe users in younger age groups tend to read articles on *pregnancy, parenting* and *sexual health*, providing insight into their preferences. We also find that as user age increases, there is a gradual shift in user interests from *weight management* and *pregnancy* to greater emphasis on *diet and nutrition*. We note that users beyond the age of 55 have highly overlapping and consistent interests.

4.1.2. Time factors

Thus far, we have studied how information consumption changes based on various demographic features. In this section, we present our analysis driven by time based features.

As discussed previously, the dataset captures user actions from 2006 to 2015, where each action (from S1, S2 and S3) is associated with a timestamp denoting the date, month and year. This time-based information raises several salient questions about when and how users consume health topics.

We first investigate whether the time of year impacts topic consumption. Using actions from 2014, which captures the vast majority of the recorded actions (and is the year in which the company went public), we study the spread and consumption pattern of 263 topics across the year. Fig. 5b illustrates the results (standardized on topics), with higher intensities denoting that a topic is consumed relatively more heavily. In general, we observe little activity during November and December, which can be explained by users consuming (or searching for) less health information during the holiday season. We also observe an increase in activity for many topics during January. This may be the result of individuals carrying out their New Year's resolutions, many of which tend to relate to increased health and wellness activities.

We also investigate whether there is a connection between topic consumption and user behavior across time. Using the timestamp

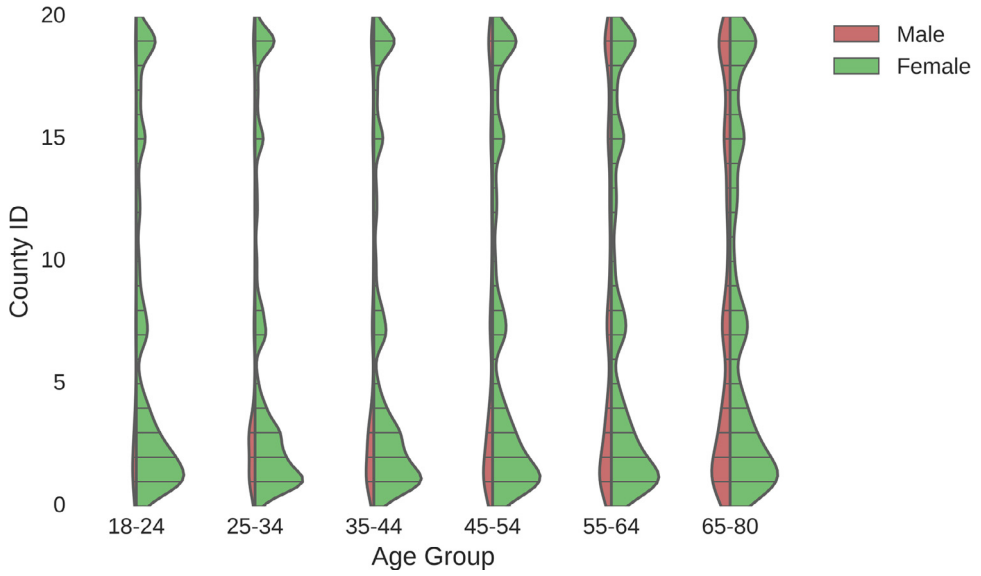


Fig. 2. Distribution of users across the three demographic attributes of age, gender and county. Color denotes gender; *x*-axis denotes age; *y*-axis denotes county. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

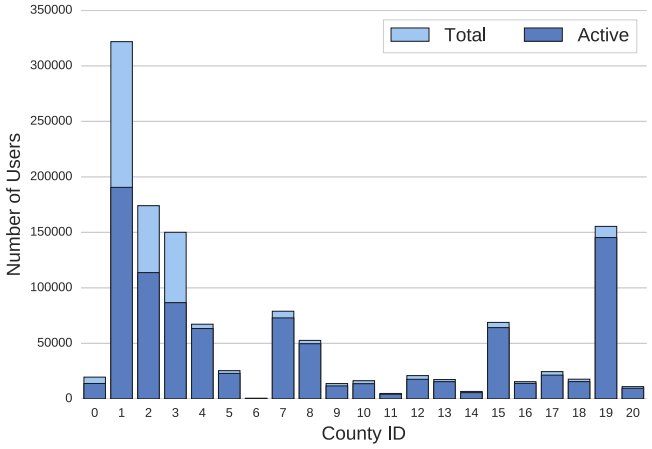
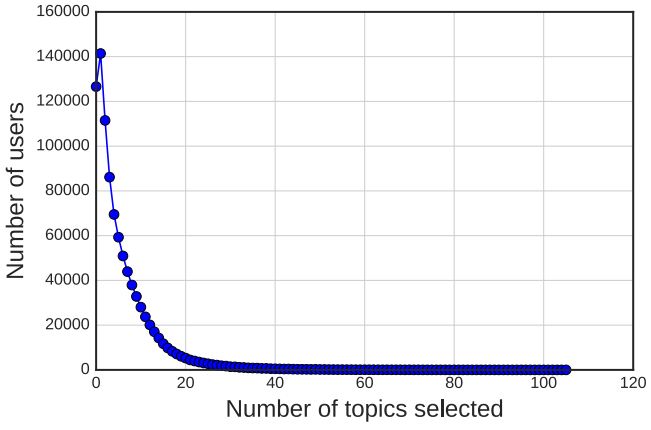
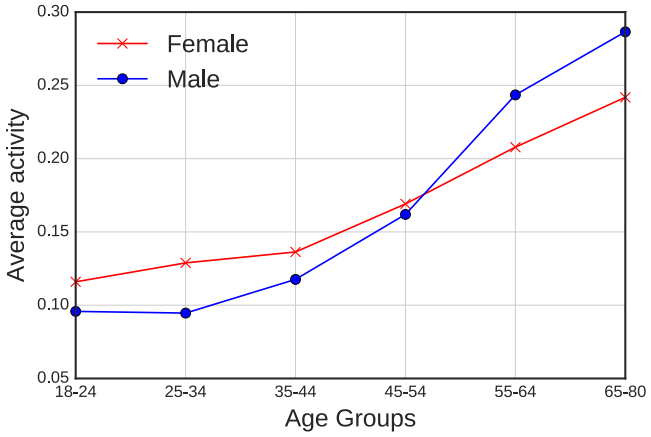


Fig. 3. Number of Users per County: Comparison between the active and total (base) users across counties as seen from the data. The data is not normalized to highlight difference between and within each county.



(a) Topic Selection Histogram

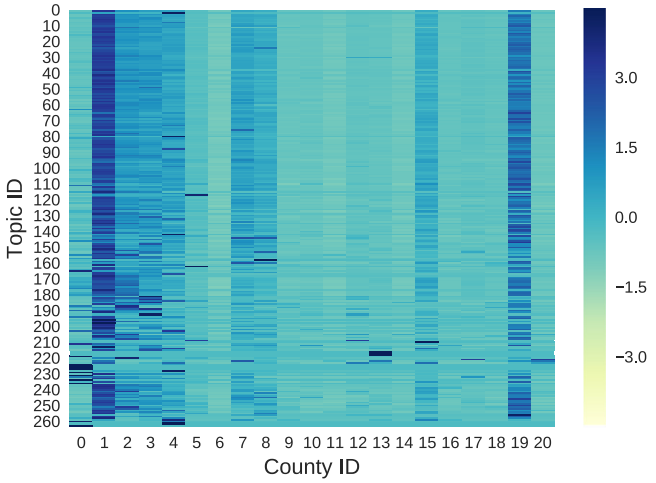


(b) Average Activity by Gender and Age

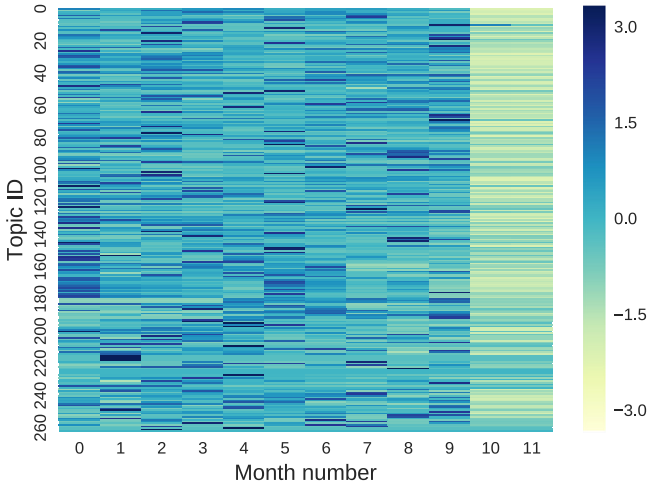
Fig. 4. (a) Distribution of users based on number of unique topics in which they are interested. (b) Distribution of average activity (total actions of users divided by user count) by age and gender.

associated with each user action, we study four user-based measures: user activity, user involvement, user engagement, and user inclination. Each measure captures a different time-based association between user behavior and consumption. For this investigation, we elect to study only those users with at least 10 or more actions in the 2014 calendar year; this filtering simplifies our analyses while retaining the vast majority of the recorded actions.

User activity:. As defined in Section 3.1, user activity is captured for a topic category as the total number of actions performed by the user in that category. To identify the dominant topic medium for user activity, we analyze the differences in the average user activity for each county across the three sources of topic consumption: S1 (topic selection), S2 (newsletter) and S3 (website search). In Fig. 6a, we observe that topic selection is the least contributing factor, indicating that not many users list their preferences explicitly. This encourages us to study consumption and infer topical preferences from alternate mediums such as S2 and S3. Moreover, we observe users prefer consuming information through website search rather than by clicking on links sent in a newsletter, demonstrating preference for more active modes of reading.



(a) Counties



(b) Months (0–11) for 2014.

Fig. 5. Heatmap of topic consumption across (a) counties and (b) time. For both subfigures, the *y*-axis denotes the topics. Values are standardized across topics. Darker colors indicate greater consumption. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

User involvement. We define user involvement (as discussed in Section 3.1) as the average time a user spends consuming information via each action type. To determine the average consumption time, we identify the first and last timestamps for a user seen on a given action type and compute the total time (in days) the user spent consuming information through the given medium. Fig. 6b provides a histogram of the average time duration for users across the three action types. We observe that users are more involved year round with consuming content through newsletters and website search. They prefer a more active mode of consumption than by simply selecting topic preferences. This highlights the importance of using implicit user feedback (such as S2 and S3) to understand a health seeker better.

User engagement. We define user engagement, in Section 3.1, based on the average period of time elapsed between two actions of the same kind performed by a user. To compute this metric, for each category, we calculate the mean difference in time between two consecutive actions related to that category. In general, a lower average time difference for a category suggests higher user engagement. Fig. 6c presents a histogram of number of users corresponding to the average time spent between actions. We observe a heavy-tailed distribution for all three action types, indicating that most users are engaged and check content frequently. They seem most engaged with content delivered through newsletter which can be attributed to its periodic nature. In addition, website search is engaging but used for specific questions.

User inclination. As previously discussed in Section 3.1, we define user inclination as the time difference between the user's initial enrollment date and the timestamp recorded for the first topic action performed by the user. Fig. 6d shows the distribution of user inclination across all users. We observe very similar patterns for all three sources, indicating that most users start consuming and listing preferences immediately after enrollment.

Based on these analyses, we build time profiles for each user to understand their preferred mode of consuming health information consisting of the following features: 1) user activity, capturing the frequency of each action type; 2) user involvement, capturing the total time spent for each action type; 3) user engagement, capturing the average gap between actions of the same type; and 4) user inclination, capturing the average time between enrollment and first action.

4.1.3. Socioeconomic factors

We previously studied factors that can be directly derived from a user and his or her actions. In this section, we investigate the degree to which indirect factors, such as the community a user lives in and its economic status, may influence topical interests and health-seeking behavior.

Social and economic conditions [43–48] such as poverty, education, unemployment and population can have substantial influence on individual's health outcomes. Positive associations between education and health have been observed in the literature, with higher levels of education associated with a longer and healthier life [49,50]. Lack of available resources due to poverty, unemployment or overpopulation can also affect individual health and, consequently, health-information consumption. While obtaining social and economic indicators for each user is difficult, county-level aggregations of these factors can serve as

Table 2
Top 10 topics consumed by users across age groups. Distinct pregnancy topics marked by the company as (1),(2).

18–24 years	25–34 years	35–44 years	45–54 years	55–64 years	65–80 years
Weight Management	Pregnancy (1)	Diet and Nutrition	Diet and Nutrition	Diet and Nutrition	Diet and Nutrition
Pregnancy (1)	Pregnancy (2)	Pregnancy (1)	Weight Management	Weight Management	Diabetes
Diet and Nutrition	Diet and Nutrition	Weight Management	Sexual Health	Diabetes	Weight Management
Sexual Health	Weight Management	Sexual Health	Diabetes	Sexual Health	Sexual Health
Pregnancy (2)	Sexual Health	Pregnancy (2)	Beauty	Heart Health	Heart Health
Diabetes	Parenting	Parenting	Menopause	Digestive Health	Digestive Health
Beauty	Dental Health	Diabetes	Digestive Health	Beauty	Beauty
Digestive Health	ADD/ADHD	Beauty	Heart Health	ADD/ADHD	ADD/ADHD
ADD/ADHD	Beauty	ADD/ADHD	ADD/ADHD	Emotional Health	Emotional Health
Parenting	Children's Health	Dental Health	Pregnancy (1)	Menopause	Menopause

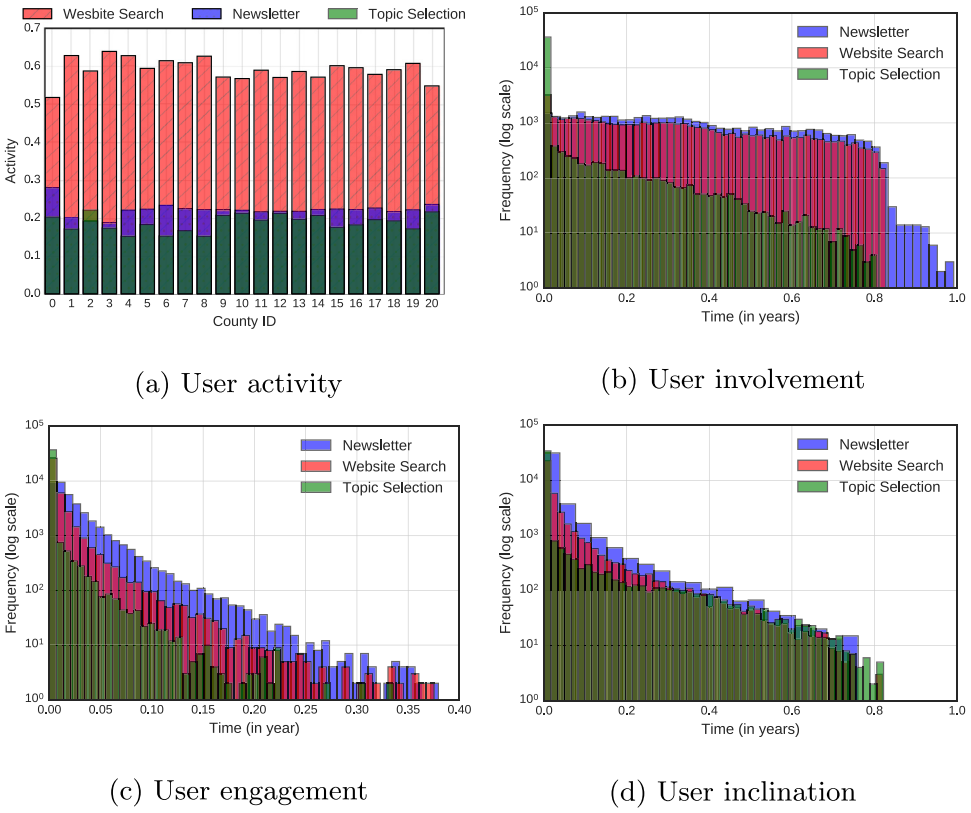


Fig. 6. (a) County-based distribution of activity. (b), (c) and (d) Time-based distributions for topic consumption by (b) topic selection, (c) newsletter and (d) website search. For (b), (c) and (d), the *y*-axis is in log scale and the *x*-axis is for the year 2014.

proxy indicators for individual socioeconomic conditions. These proxy indicators can provide additional insight into user preferences.

Our investigation of social and economic factors focuses specifically on data from 2014, representing the preponderance of the data collected by our partner company, which IPOed in the same year. We obtained county-based socioeconomic data for 2014 from the USDA Economic Research Service,[6] extracting the following factors for each county:

1. **Population**: Data reported between 2010 and 2014 that includes attributes such as the census population, yearly population change, number of births, and international and domestic migration.
2. **Poverty**: Data reported for 2014 that represents attributes such as estimates of individuals, families and children in poverty, and median household income.
3. **Unemployment**: Data reported between 2006 and 2014 that includes attributes such as civilian labor force, number of people employed and unemployed, and median household income.
4. **Education**: An aggregation of data between 1970 and 2014 at 10-year intervals (i.e., 1970, 1980, 1990, 2000, and 2010–14) that includes attributes based on education level, such as percentage of individuals with high school diploma or college degrees.

4.2. Explaining topical preference

Thus far, we have studied the variation in consumption of health information based on demographic, time, and socioeconomic factors. We now present results by applying factorization machines (as discussed in Section 3.2) to explore if these factors can be used to explain each health consumer's preference for a given topic.

[6] http://www.ers.usda.gov/data-products/county-level-data-sets.aspx .

Due to computational limitations, we perform our experiments on a representative subset of the 21 available counties. We use county information from the 2010 U.S. Census Report to ensure that the selected counties represent a diverse range of median income, area, population, median age and unemployment rate. Based on these factors, the following 8 counties were selected for inclusion: St. Joseph (0), King (2), San Mateo (5), Bronx (9), Richmond (11), Denver (15), Suffolk (16) and Bexar (17). These counties account for 346,553 users, 265,839 of whom have topic profiles. We further limit our study to users with at least 3 topic actions, thereby restricting our experiments to 183,665 users with data across 254 topics.

We divide the data used in our experiments into training and testing sets based on each user's distinct topic choices. For each user, 75% of the topics are randomly selected as training topics, with the user-topic interactions recorded for these topics used as the training data. The remaining 25% of the user's topics are selected as testing topics, with the corresponding user-topic interactions used as the testing data. In total, the training and testing data consist of 1,173,144 and 474,473 interactions, respectively. From these interactions, topic preference is computed for each user-topic pair as described in Section 3.2, with user topic preferences calculated independently for the training and testing data sets. The models are evaluated using root mean square error (RMSE), defined as $\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i - \hat{y_i})}$, where y_i and $\hat{y_i}$ are the actual and predicted rating, respectively.

4.2.1. Experiment 1: context-based modeling

Our first experiment investigates the effectiveness of context-based factors on estimating user topical preference. The data matrix X is computed for both the training and testing data sets, where each feature vector contained a combination of user (U = 183,665), topic (I = 254), demographic (D = 15), time (T = 12) and socioeconomic (S = 176) bits, resulting in m= 184,122. The data matrix containing all features was extremely sparse, with an average of only 50 ($\ll m$) non-zero

Table 3
Root mean square error (RMSE) of context-based topic preference models on testing data. Each model is statistically significant ($p \ll 0.01$) compared to all models listed below it.

Method	RMSE
User Attribute Aware	0.3217
FM-Demographic	0.3421
Matrix Factorization	0.3725
FM-Socioeconomic	0.4643
FM-Time	0.5515
FM-All	0.5636

elements per feature vector. Given this sparsity, which makes higher-order interactions more difficult to compute [38,51], we restricted our experiments to second-order FMs (implemented via libFM [38]). Markov chain Monte Carlo (MCMC) inference [37] was used as the learning method, which requires an initial σ parameter. Grid search was performed to select parameters values for σ ([0.1, 0.2, 0.5, 1.0]), the level of latent factors ([8, 16, 32, 64]) and the number of iterations ([100, 200, 300, 400, 500]) over a validation set (330,104 instances) for each model.

Results. The performance of each model on the testing data is shown in Table 3. To establish statistical significance, we use the Wilcoxon signed-rank test to compare pairs of models with a significance level of 0.01. Given the high degrees of freedom, each improvement in the performance metric (RMSE) is statistically significant ($p \ll 0.01$). We observe that the User Attribute Aware model statistically significantly outperforms the other models ($p \ll 0.01$). The User Attribute Aware model utilizes attribute-to-factor mappings and user-attribute interactions, which leads to its improved ability for estimating user preference. The FM-Demographic model leverages pairwise interactions between demographic attributes in addition to the latent factors, resulting in it outperforming Matrix Factorization ($p \ll 0.01$). The User Attribute Aware and FM-Demographic models both learn from pairwise interactions—such as interactions between *age* and *gender*—and are able to better explain a preference score for a user-topic pair.

We find that incorporating demographic information is able to model individual preference better than simply using other contextual features, such as socioeconomic and time attributes. We also find that using the full set of contextual signals actually negatively impacts the performance. This emphasizes the vitality of each context category and demonstrates that they are not equally explanatory of a user's preference. The effective performance of the User Attribute Aware and FM-Demographic models can be explained by their use of demographic features (e.g., gender, age and location), which are directly associated with a user, rather than the sole use of indirect features (e.g., socio-economic attributes), which are derived from the county in which a user resides.

Based on our findings that demographics is an important factor in inferring user topical preference, we conduct a further experiment to investigate the role of different demographic data in explaining topical preference, as described below.

4.2.2. Experiment 2: demographic segmentation

Our second experiment investigates the performance of each model according to various demographic segments. For this experiment, we segment the user population based on age, gender and county, and build different matrix factorization (MF) models for each segment. By segmenting the population, we place more emphasis on the user's demographics and preclude potential biases that may arise due to collective trends in the data. For example, our data is representative of far more female users than male users, which may skew results toward female preferences. To avert this potential bias, we segment users based on gender to build models that leverage user-topic interactions for women and men separately.

Inspired by the extensive success of ensemble methods, we also investigate whether these segmented models can be used to collectively estimate for a test user-topic pair. For example, to estimate the preferences of a 19-year-old woman residing in County 0 for the topic t, we can leverage model estimations from three segmented models: Female MF, 18–24 MF and County 0 MF. To combine these models, we examine two ensembles methods: average and weighted average. For the average ensemble, we average the scores obtained from the three models based on a test user's demographic (age, gender and county) values. For the weighted average ensemble, the individual model predictions are weighted according to the number of users in each category before calculating the average. All matrix factorization models are computed using the following parameters: $\sigma = 1.0$, factors $= 8$ and iterations $= 500$ (based on the grid search performed for the FM-Demographic model).

Results. The results obtained from segmenting the population based on demographics and the ensemble models are provided in Table 4. Based on RMSE, we observe improved scores for some segments, such as 65–80 MF and County 0 MF. Overall, however, we do not observe any improvement over the User Attribute Aware and FM-Demographic models. Comparing the User Attribute model and ensemble methods, we find that the User Attribute Aware model performs statistically significantly better ($p \ll 0.01$) than the ensembles. This indicates that pairwise interactions between demographic attributes (demonstrated by the User Attribute Aware and FM-Demographic models) capture user preferences better than simply using them as a data segmenting criteria. We note that as each segmented model is evaluated within its own population, we cannot perform significance tests to compare the performance of the segmented models with each other.

Table 4
Topic preference models learned on users segmented based on their demographics and the corresponding root mean square error (RMSE) on testing data. For each model, the number of users, topics, and training and testing instances are listed.

Demographic	Method	Topic count	User count	Train instances	Test instances	RMSE
Gender	Female MF	254	158,920	1,016,328	410,723	0.3606
	Male MF	254	24,745	156,816	63,750	0.3667
Age	18–24 MF	254	24,182	126,364	52,641	0.3974
	25–34 MF	254	54,629	272,484	114,744	0.4176
	35–44 MF	254	37,278	224,348	91,773	0.3776
	45–54 MF	254	28,642	211,471	83,896	0.3363
	55–64 MF	254	25,448	211,213	82,445	0.3119
	65–80 MF	254	13,486	127,264	48,974	0.2917
County	County 0 MF	254	10,592	88,757	34,427	0.2794
	County 2 MF	254	77,986	484,939	196,837	0.3808
	County 5 MF	254	15,718	102,983	41,550	0.3621
	County 9 MF	254	7766	48,829	19,824	0.3627
	County 11 MF	254	2705	16,707	6825	0.3697
	County 15 MF	254	45,268	283,373	115,072	0.3767
	County 16 MF	254	9553	60,786	24,599	0.3699
	County 17 MF	254	14,077	86,770	35,339	0.3703
Ensemble	Average	254	183,665	1,173,144	474,473	0.4972
	Weighted average	254	183,665	1,173,144	474,473	0.5216

4.2.3. Discussion

We observe that contextual features such as demographics, time and socioeconomic status can provide rich information that is integral to understanding how health information is consumed, but may not contribute equally to inferences made regarding users' topical preferences. Based on our experiments, without prior knowledge of medical background or search logs, demographic attributes (age, gender and location) can be used to effectively model an online health seeker's topical preferences, as demonstrated by the success of our User Attribute Aware model.

We note, however, that despite these encouraging results, user interests can also be influenced by various confounding factors such as an underlying medical condition. Or, as suggested by previous research [23], a user's behavior could be performed on behalf of a loved one: a wife might show interest in topics relevant to her husband or a father might read articles for his children. Unfortunately, it is difficult to incorporate or compensate for such factors.

5. Conclusion

Vast adoption of the Internet has revolutionized the healthcare industry with large amounts of information available through specialized health websites. While many factors can affect a person's health, an individual's online activity can be indicative of their health preferences and interests. In this work, we harvested individuals' health interests by fusing information from varied sources of user activity. Using unique, real-world data from a large digital health content provider, we presented a large-scale analysis of health seekers and characterized health content consumption across three contextual dimensions (demographics, time and socioeconomics) motivated by the social determinants of health.

We also investigated several models that leverage different kinds of contextual features along with implicit feedback to learn factors that can explain individual user preferences. We found that some factors, such as demographics, play a pivotal role in content consumption and the explanation of user topical preferences, while other factors, such as socioeconomics and time, are not able to model individual user preference that closely. Our results help to distinguish between factors that influence content consumption and those that explain individual's topical interest.

This work directly contributes to the understanding of the health information needs of an individual and community and of the factors governing those needs. At an individual level, this work can help in inferring medical conditions and designing more personalized online experiences. At a community level, this work characterizes the interests of different population segments, which can prove helpful in health surveillance systems, providing targeted care and resource allocation, and subsequently improve health outcomes for the general public by making them more informed through preferred mediums.

The next steps would be to include other factors in our analysis, such as race and individual income, to further study how consumption of health information and interests vary. Additionally, as our study focuses on individuals who use Internet for consuming health content, many factors such as availability of Internet access in different regions and the penetration of Internet usage among various age groups may affect user activity and could be addressed by future work.

Acknowledgments

This work is supported in part by the National Science Foundation (NSF) Grant IIS-1447795 and in part by the National Science Centre, Poland under the research project no. 2016/23/B/ST6/01735.**Author disclosure statement**

The authors declare no competing financial interests.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.inffus.2018.04.005.

References

[1] S. Fox, M. Duggan, Health online 2013, Pew Internet & American Life Project, Washington, D.C, (2013).
[2] J. Morahan-Martin, How internet users find, evaluate, and use online health information: a cross-cultural review. Cyberpsy. Behav. Soc. Netw. 7 (5) (2004) 497–510.
[3] A.J. Morgan, Identity and the health information consumer, Health Syst. 5 (1) (2016) 1–5.
[4] R.J. Cline, K.M. Haynes, Consumer health information seeking on the internet: the state of the art, Health Educ. Res. 16 (6) (2001) 671–692.
[5] T. Lang, Advancing global health research through digital technology and sharing data, Science 331 (6018) (2011) 714–717.
[6] G. Eysenbach, Consumer health informatics, Br. Med. J. 320 (7251) (2000) 1713.
[7] D. Lewis, B.L. Chang, C.P. Friedman, Consumer health informatics, Consumer Health Informatics, Springer, 2005, pp. 1–7.
[8] G. Eysenbach, C. Köhler, How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews, BMJ 324 (7337) (2002) 573–577.
[9] M. De Choudhury, M.R. Morris, R.W. White, Seeking and sharing health information online: comparing search engines and social media, CHI '14, (2014), pp. 1365–1376.
[10] G. Dahlgren, M. Whitehead, Policies and Strategies to Promote Social Equity in Health, Institute for future studies, Stockholm, 1991.
[11] W. Liang, M.C. Shediac-Rizkallah, D.D. Celentano, C. Rohde, A population-based study of age and gender differences in patterns of health-related behaviors, Am. J. Prev. Med. 17 (1) (1999) 8–17.
[12] Y. Benyamini, E.A. Leventhal, H. Leventhal, Gender differences in processing information for making self-assessments of health, Psychosom. Med. 62 (3) (2000) 354–364.
[13] S. Birch, S. Chambers, To each according to need: a community-based approach to allocating health care resources. Can. Med. Assoc. J. 149 (5) (1993) 607.
[14] D. Musoke, P. Boynton, C. Butler, M.B. Musoke, Health seeking behaviour and challenges in utilising health facilities in Wakiso district, Uganda, Afr. Health Sci. 14 (4) (2014) 1046–1055.
[15] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D.P. Lorence, S. Ozmutlu, H.C. Ozmutlu, A study of medical and health queries to web search engines, Health Info. Libr. J. 21 (1) (2004) 44–51.
[16] M. Salathe, L. Bengtsson, T.J. Bodnar, D.D. Brewer, J.S. Brownstein, C. Buckee, E.M. Campbell, C. Cattuto, S. Khandelwal, P.L. Mabry, et al., Digital epidemiology, PLoS Comput. Biol. 8 (7) (2012) e1002616.
[17] B.P. Kennedy, I. Kawachi, R. Glass, D. Prothrow-Stith, Income distribution, socioeconomic status, and self rated health in the united states: multilevel analysis, BMJ 317 (7163) (1998) 917–921.
[18] J. Powell, N. Inglis, J. Ronnie, S. Large, The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study, JMIR 13 (1) (2011) e20, http://dx.doi.org/10.2196/jmir.1600.
[19] G. Peterson, P. Aslani, K.A. Williams, How do consumers search for and appraise information on medicines on the internet? a qualitative study using focus groups, JMIR 5 (4) (2003) e33.
[20] Y. Zhang, Searching for specific health-related information in medlineplus: behavioral patterns and user experience, J. Assoc. Inf. Sci. Technol. 65 (1) (2014) 53–68.
[21] R.W. White, E. Horvitz, Experiences with web search on medical concerns and self diagnosis. AMIA '09, (2009).
[22] M.L. Ybarra, M. Suman, Help seeking behavior and the internet: a national survey, Int. J. Med. Inform. 75 (1) (2006) 29–41.
[23] S.R. Cotten, S.S. Gupta, Characteristics of online and offline health information seekers and factors that discriminate between them, Social Sci. Med. 59 (9) (2004) 1795–1806.
[24] C. Escoffery, K.R. Miner, D.D. Adame, S. Butler, L. McCormick, E. Mendell, Internet use for health information among college students, J. Am. Coll. Health 53 (4) (2005) 183–188.
[25] M.-A. Cartright, R.W. White, E. Horvitz, Intentions and attention in exploratory health search, SIGIR '11, (2011), pp. 65–74.
[26] S.L. Ayers, J.J. Kronenfeld, Chronic illness and health-seeking information on the internet, Health: 11 (3) (2007) 327–347.
[27] M.J. Paul, R.W. White, E. Horvitz, Diagnoses, decisions, and outcomes: web search as decision support for cancer, WWW '15, ACM, 2015, pp. 831–841.
[28] J.S. Brownstein, C.C. Freifeld, B.Y. Reis, K.D. Mandl, Surveillance sans frontieres: internet-based emerging infectious disease intelligence and the healthmap project, PLoS Med. 5 (7) (2008) e151.
[29] D.M. Berwick, T.W. Nolan, J. Whittington, The triple aim: care, health, and cost, Health Aff. 27 (3) (2008) 759–769.
[30] E.H. Wagner, B.T. Austin, C. Davis, M. Hindmarsh, J. Schaefer, A. Bonomi, Improving chronic illness care: translating evidence into action, Health Aff. 20 (6) (2001) 64–78.
[31] J.S. Brownstein, C.C. Freifeld, L.C. Madoff, Digital disease detectionharnessing the web for public health surveillance, N. Engl. J. Med. 360 (21) (2009) 2153–2157.
[32] D. Revere, A.M. Turner, A. Madhavan, N. Rambo, P.F. Bugni, A. Kimball, S.S. Fuller,

Understanding the information needs of public health practitioners: a literature review to inform design of an interactive digital knowledge management system, J. Biomed. Inform. 40 (4) (2007) 410–421.
[33] S. Fox, D. Fallows, Internet health resources, Pew Internet & American Life Project: Online Report, TPRC, 2003.
[34] J.F.S. Patricia A. Looker, Getting to know the women's health care segment. Mark. Health Serv. (2001).
[35] L.M. Miller, R.A. Bell, Online health information seeking: the influence of age, information trustworthiness, and search challenges. J. Aging Health 24 (3) (2012) 525–541, http://dx.doi.org/10.1177/0898264311428167.
[36] F. Qiu, J. Cho, Automatic identification of user interest for personalized search, WWW '06, ACM, 2006, pp. 727–736, http://dx.doi.org/10.1145/1135777.1135883.
[37] S. Rendle, Factorization machines, ICDM '10, IEEE Computer Society, Washington, DC, USA, 2010, pp. 995–1000, http://dx.doi.org/10.1109/ICDM.2010.127.
[38] S. Rendle, Factorization machines with libfm, ACM Trans. Intell. Syst. Technol. 3 (3) (2012) 57:1–57:22, http://dx.doi.org/10.1145/2168752.2168771.
[39] D. Agarwal, B.-C. Chen, Regression-based latent factor models, KDD '09, ACM, New York, NY, USA, 2009, pp. 19–28, http://dx.doi.org/10.1145/1557019.1557029.
[40] A. Paterek, Improving regularized singular value decomposition for collaborative filtering, Proceedings of KDD Cup and Workshop, vol. 2007, (2007), pp. 5–8.
[41] J.D.M. Rennie, N. Srebro, Fast maximum margin matrix factorization for collaborative prediction, ICML '05, ACM, New York, NY, USA, 2005, pp. 713–719, http://dx.doi.org/10.1145/1102351.1102441.
[42] Z. Gantner, L. Drumond, C. Freudenthaler, S. Rendle, L. Schmidt-Thieme, Learning attribute-to-feature mappings for cold-start recommendations, ICDM '10, IEEE, 2010, pp. 176–185, http://dx.doi.org/10.1109/ICDM.2010.129.
[43] N.E. Adler, J.M. Ostrove, Socioeconomic status and health: what we know and what we don't, Ann. N. Y. Acad. Sci. 896 (1) (1999) 3–15, http://dx.doi.org/10.1111/j.1749-6632.1999.tb08101.x.
[44] C. Friestad, K.-I. Klepp, Socioeconomic status and health behaviour patterns through adolescence: results from a prospective cohort study in Norway, Eur. J. Public Health 16 (1) (2006) 41–47, http://dx.doi.org/10.1093/eurpub/cki051.
[45] A.-M. Taloş, Influence of population lifestyle on local health profile case study: Ialomita county, Procedia Environ. Sci. 32 (2016) 311–317 https://doi.org/10.1016/j.proenv.2016.03.036.
[46] M.A. Winkleby, D.E. Jatulis, E. Frank, S.P. Fortmann, Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. Am. J. Public Health 82 (1992).
[47] K. Fiscella, P. Franks, M.R. Gold, C.M. Clancy, Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care, JAMIA (2000).
[48] I.S. Kickbusch, Health literacy: addressing the health and education divide, Health Promot. Int. 16 (3) (2001) 289–297.
[49] E.R. Eide, M.H. Showalter, Estimating the relation between health and education: what do we know and what do we need to know? Econ. Educ. Rev. 30 (5) (2011) 778–791.
[50] C.E. Ross, J. Mirowsky, Refining the association between education and health: the effects of quantity, credential, and selectivity, Demography 36 (4) (1999) 445–460, http://dx.doi.org/10.2307/2648083.
[51] S. Rendle, L. Schmidt-Thieme, Pairwise interaction tensor factorization for personalized tag recommendation, WSDM '10, ACM, New York, NY, USA, 2010, pp. 81–90, http://dx.doi.org/10.1145/1718487.1718498.