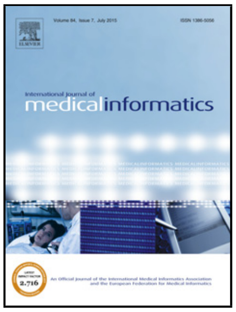


Title: Search Engines, News Wires and Digital Epidemiology: Presumptions and Facts

Authors: Fatemeh Kaveh-Yazdy, Ali-Mohammad Zareh-Bidoki

PII: S1386-5056(18)30271-5
DOI: https://doi.org/10.1016/j.ijmedinf.2018.03.017
Reference: IJB 3683

To appear in: *International Journal of Medical Informatics*

Received date: 14-12-2017
Revised date: 30-3-2018
Accepted date: 31-3-2018

Search Engines, News Wires and Digital Epidemiology: Presumptions and Facts

Fatemeh Kaveh-Yazdy[1] Ali-Mohammad Zareh-Bidoki[2]*

[1]Research Assistant at Computer Engineering Department, Yazd University, Yazd, Iran,
Head of Natural Language Processing Dept. at Parsijoo Persian Search Engine

[2]Associate Prof. at Computer Engineering Department, Yazd University, Yazd, Iran,
Founder and CEO of Parsijoo Persian Search Engine

*Corresponding Author:
Ali-Mohammad Zareh-Bidoki
Associate Professor at Computer Engineering Department, Yazd University, Yazd, Iran
University Blvd, Safaieh, Yazd, Iran, Tel: +98-35-3123-2358

Emails:

Fatemeh Kaveh-Yazdy:	fkavehy@stu.yazd.ac.ir	fkavehy@parsijoo.ir
Ali-Mohammad Zareh-Bidoki:	alizareh@yazd.ac.ir	alizareh@parsijoo.ir

Search Engines, News Wires and Digital Epidemiology: Presumptions and Facts

Highlights

Points

- Users' information-seeking behaviors varied by time of the day.
- Trend belonging to "Immunization and Vaccination" is seasonal we well as the trends of diseases such as, Mumps, Flu, Chicken Pox, and Meningitis.
- Trends belong to diseases which received minor social attention are vulnerable to be misclassified as seasonal trends.
- News and search trends are weakly correlated, while they still be co-integrate and move towards with a stable distance from each other.
- Analyzing search queries without taking into account the impact of news and external events can be deceiving.

Abstract

Background

Digital epidemiology tries to identify diseases dynamics and spread behaviors using digital traces collected via search engines logs and social media posts. However, the impacts of news on information-seeking behaviors have been remained unknown.

Methods

Data employed in this research provided from two sources, (1) Parsijoo search engine query logs of 48 months, and (2) a set of documents of 28 months of Parsijoo's news service. Two classes of topics, i.e. macro-topics and micro-topics were selected to be tracked in query logs and news. Keywords of the macro-topics were automatically generated using web provided resources and exceeded 10k. Keyword set of micro-topics were limited to a numerable list including terms related to diseases and health-related activities. The tests are established in the form of three studies. Study A includes temporal analyses of 7 macro-topics in query logs. Study B considers analyzing seasonality of searching patterns of 9 micro-topics, and Study C assesses the impact of news media coverage on users' health-related information-seeking behaviors.

Results

Study A showed that the hourly distribution of various macro-topics followed the changes in social activity level. Conversely, the interestingness of macro-topics did not follow the regulation of topic distributions. Among macro-topics, "Pharmacotherapy" has highest interestingness level and wider time-window of popularity. In Study B, seasonality of a limited number of diseases and health-related activities were analyzed. Trends of infectious diseases, such as flu, mumps and chicken pox were seasonal. Due to seasonality of most of diseases covered in national vaccination plans, the trend belonging to "Immunization and Vaccination" was seasonal, as well. Cancer awareness events caused peaks in search trends of "Cancer" and "Screening" micro-topics in specific days of each year that mimic repeated patterns which may mistakenly be identified as seasonality. In study C, we assessed the co-integration and correlation between news and query trends. Our results demonstrated that micro-topics sparsely covered in news media had lowest level of impressiveness and, subsequently, the lowest impact on users' intents.

Conclusion

Our results can reveal public reaction to social events, diseases and prevention procedures. Furthermore, we found that news trends are co-integrated with search queries and are able to reveal health-related events; however, they cannot be used interchangeably. It is recommended that the user-generated contents and news documents are analyzed mutually and interactively.

Keywords

Big Data, Digital Epidemiology, Healthcare Informatics, News, Search Query Log, Text Mining;

1 Introduction

Up to now, required information for disease surveillance and epidemic outbreaks have been collected via formal reporting forms of health organizations and institutes. Gathering data, aggregating them and inferencing epidemiological facts out of them is time-consuming; while hours are critical in decision making, information inference takes weeks to complete, e.g., CDC Influenza-like illnesses reports have a delay of 1-3 weeks [1]. These reports solely include results of observations of health care personnel in fixed time and location in the past; although demographic information, living location, time between appearing symptoms to calling for health services, movement patterns and social interactions are critical parameters in epidemiological studies ignored in formal reports. As a consequence, health control agencies make their decisions based on incomplete and inaccurate data while ignored parameters are able to alter the decisions.

Nowadays, internet based media are frequently used by people. Social networks collect various types of data, such as spatio-temporal logs, social interactions, movement information and textual descriptions of daily life. Interestingly, online data can cover the information insufficiency of formal disease reports. According to a study [2], the benefits that data of internet media bring to epidemiological analysis can be listed as: (1) early detection of disease incident, (2) continuous monitoring, (3) assessing sentiments and behaviors around the disease control procedures, and (4) providing data to analyze the time period between the first incident and the outbreak announcement.

Utilizing digitally-generated data in epidemiological analyses leads to initiation of a research field called Digital Epidemiology. Digital epidemiology uses online information, e.g., social media data and search engine logs, to analyze distribution of disease incidence and dynamics of spread. Google Flu Trends (GFT) [3,4] was the first operational software framework developed to track flu-related searches and predict flu incident cases month by month. GFT faced with two main problems, i.e. big data hubris and algorithm dynamics, which have led to over-estimated incidents of the flu [5]. GFT finally failed but it has opened doors to new horizons and disseminated applications of digital media data in epidemiological analyses. Besides epidemiological analyses, online data can be employed in other health-related studies.

Based on the types of health-related issues analyzed using online data, studies are categorized into three major groups. The first group of studies try to identify the relation between incidents of diseases and user-generated online contents. Infectious diseases, and particularly influenza-like illnesses [6,7], have been mainly considered in different studies. Early detection of outbreaks of diseases such as Zika virus disease (ZVD) [8], Dengue Fever [9] and Ebola [10] encouraged researchers to extend the domain of digital epidemiology to study distribution of non-infectious diseases, e.g., brain aneurysm[11] and kidney stone[12].

In addition to incident cases, people's reaction to cancer screening [13], effectivity of vaccination [14–16] and public reaction to life-threaten events [17] can be monitored through the content of social media and search logs.

The second group of studies considers assessment of the online health-related and medical data. According to different surveys, more than half of the internet users search for health-related issues [18,19]. Online information seeking can increase people's level of awareness; however irrelevant and inaccurate information may direct people to self-treatment and self-diagnosis [20] and it should be noted that recommendations of online resources are most often inconsistent with technical guides [21–23].

The third group analyzes patient-physicians relation by taking into account the impacts of online information-seeking [24]. In most of the studies conducted in this group, patients were targeted in surveys. Results of analyses showed that patients suppose that physicians have central role in medical consultation [25]. Online information can bridge the knowledge gap between physicians and patients [26] and improve patients' communications with doctors [20]. Fortunately, several surveys indicated that not only does online searching negatively affect patient-physician relationship but also promotes mutual understanding of symptoms and diagnosis [19].

In this study, we analyzed information-seeking behaviors of Persian-native users in the Parsijoo Persian search engine which is the leading Persian search engine and the second frequently-used search engine in Iran after Google. In addition, the impact of the news on searching behaviors of users were assessed. We employed advanced big data analysis methods [27] as well as machine learning procedures to extract health-related queries and news.

2 Data & Methods

The established research involves three studies. Study A included temporal analyses of seven major topics (aka macro-topics) intended in most of health-related search queries. Study B considered analyzing seasonality of searching patterns of 9 minor topics (aka micro-topics), and Study C assessed the impact of the news media coverage on the users' behaviors.

2.1 Data

We collected our data from two sources: (1) search query logs of the Parsijoo search engine from April 2013 to April 2017, and (2) logs of the Parsijoo news service from January 2015 to April 2017. Removing the spam queries, logs of 48 months of the Parsijoo web search included 208,925,978 queries and logs of the news service contained 8,520,472 news documents collected from 62 Persian news agencies mostly visited by Persian natives.

2.1.1 Macro-topic Keywords

At the first step, we selected health-related topics that were mostly covered by similar studies and then we collected keywords related to each topic. Textual contents of various sources of health-related repositories (listed in Table 1.) were collected, preprocessed and expanded to prepare keywords.

Table 1. Sources of keyword set of macro-topics.

#	Topic	Source
1	Medical Services	List of medical centers provided by the Ministry of Health and Medical Education (MHME)
2	Signs and Symptoms	ICD-10 (2017) Diagnosis Codes R00-R09
3	Diagnosis	ICD-10-CM Codes, Britannica's list of medical tests and diagnostic procedures[1], index of Lab Tests Online[2]
4	Diseases	ICD-10-CM list of disease and injuries
5	Pharmacotherapy	List of approved by Food and Drug Administration of Iran
6	Public Health Threats	Top causes of mortality in Iran, Iran's notifiable infectious diseases
7	Weight Issues and Nutrition	Automatically extracted from Wikipedia "Nutrition" and "Diets" categories

[1] https://www.britannica.com/topic/list-of-medical-tests-and-diagnostic-procedures-2074273

[2] labtestsonline.org

Raw ICD-10 codes and descriptions are not applicable to related query retrieval. Furthermore, technical phrases and expressions are in English which are unlikely to be used by Persian native users. Therefore, we decided to preprocess the collected data. The preprocessing step included windowing, translation, transliteration and expansion (details of preprocessing steps are covered in Appendix A).

2.1.2 Micro-topic Keywords

We selected a list of diseases and health-related activities which were used in similar researches [1,13–15,28,29]. The size of the keyword set of these studies (in both English and Persian languages) was very small and the set was limited to a numerable list of terms (listed in Table 2).

Table 2. Selected diseases/procedures to be employed as micro-topics and their keywords.

#	Topic	Keywords
1	Allergy	Allergy, Seasonal Allergy, Respiratory Allergy
2	HIV	Aids, HIV, HIV+
3	Flu	Flu, Influenza
4	Mumps	Mumps, Parotitis
5	Chicken Pox	Chicken Pox, Varicella Zoster Virus (VZV), Varicella
6	Meningitis	Meningitis, Meningococcal disease
7	Cancer	Cancer
8	Screening	Screening
9	Immunization & Vaccination	Immunization, Vaccine, Vaccination

2.2 Query and News Selection

In this study, we adopted more than 10K keywords to retrieve relevant queries from a repository of size 208,925,978 queries. In such a big-data-scale problem, simple word-matching techniques were not applicable; consequently a topic-based document retrieval method [30] was applied to the query log (details of the retrieval framework and formulation are denoted in Appendix B). Retrieved queries associated with each topic were counted in one-minute periods to form macro-topic time-series (Study A). In studies B and C, micro-topic keywords (in Table 2) were utilized and micro-topics' time-series were extracted in time period from January 2015 to April 2017.

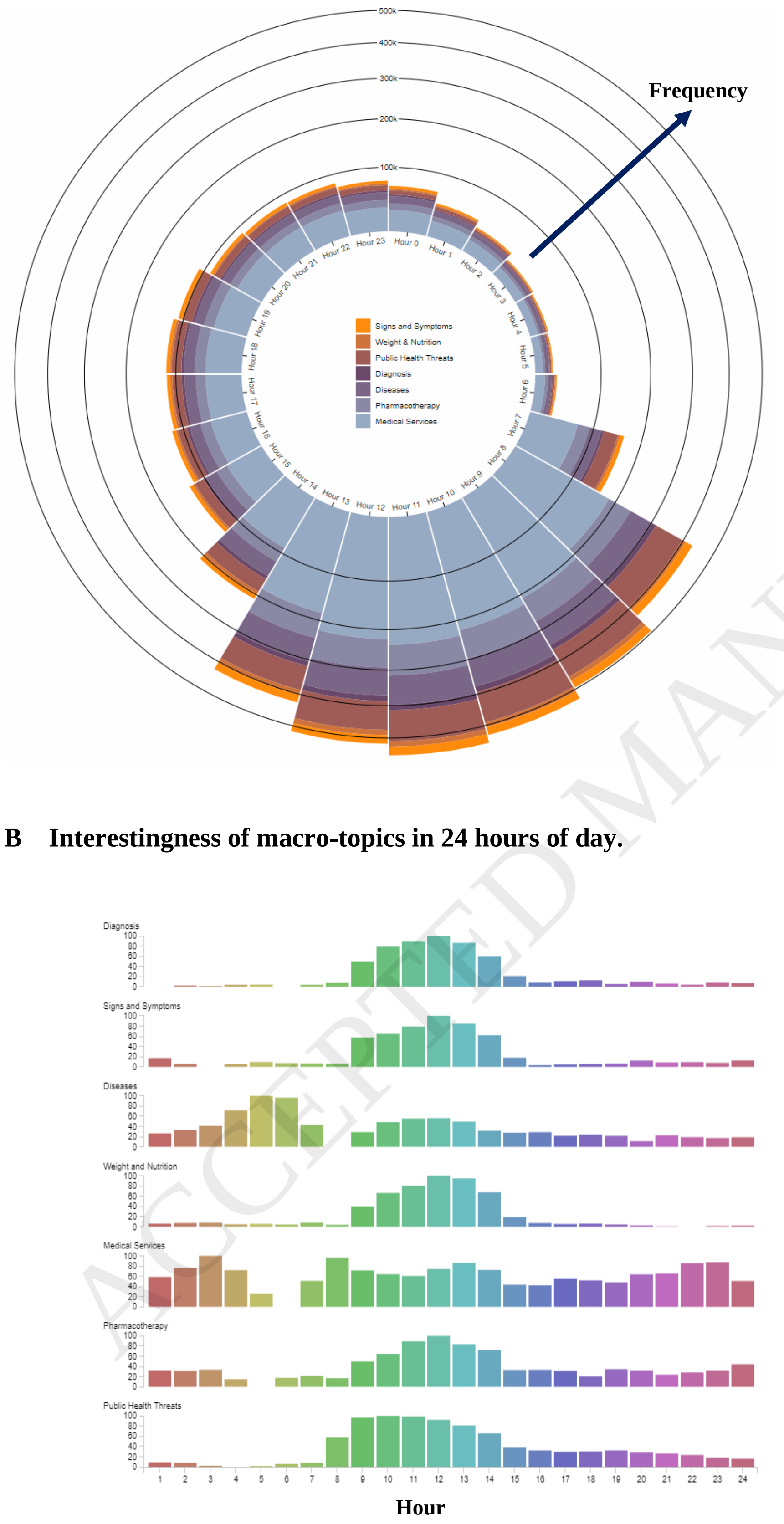
2.2.1 Study A: Temporal Analyses of Macro-topics

Temporal analyses of queries were divided into four major groups: (1) hourly distribution of topics, (2) interestingness of topics (3) intra-topic jumps, and (4) topic drop. Queries in one-hour time periods were counted to extract hourly distribution over each topic. (Figure 1-A).

Various measures were introduced in data mining [31] to present interestingness of association patterns. The improved Gini index which has been widely employed in text categorization application [32,33] was selected to estimate the interestingness of topics in one-hour periods (Figure 1-B).

Users tend to change their queries across the search sessions. We tracked the users in 20-minute sessions to extract frequent patterns of intra-topic jumps in health-related searches. In analyzing the intra-topic jumps, we added the label "Others" to present the jumps from a macro-topic to another topic in one session (Figure 2-A).

Ratio of the users who followed their intended topic is greater than that of those who switched between two topics in one session, thus we removed sessions without topic-switching to expose the intra-topic jumps. Jumps between each pair of classes were counted and the probability of dropping a topic during a search session was estimated (Figure 2-B). Furthermore, similarity of probability distribution of all pairs of macro-topic time-series are estimated using Jensen–Shannon divergence (JSD). The 7×7 matrix of JSD is employed in an Agglomerative hierarchical categorization framework [34] to disclose similarity between the users' behaviors in searching for topics (Figure 2-C).



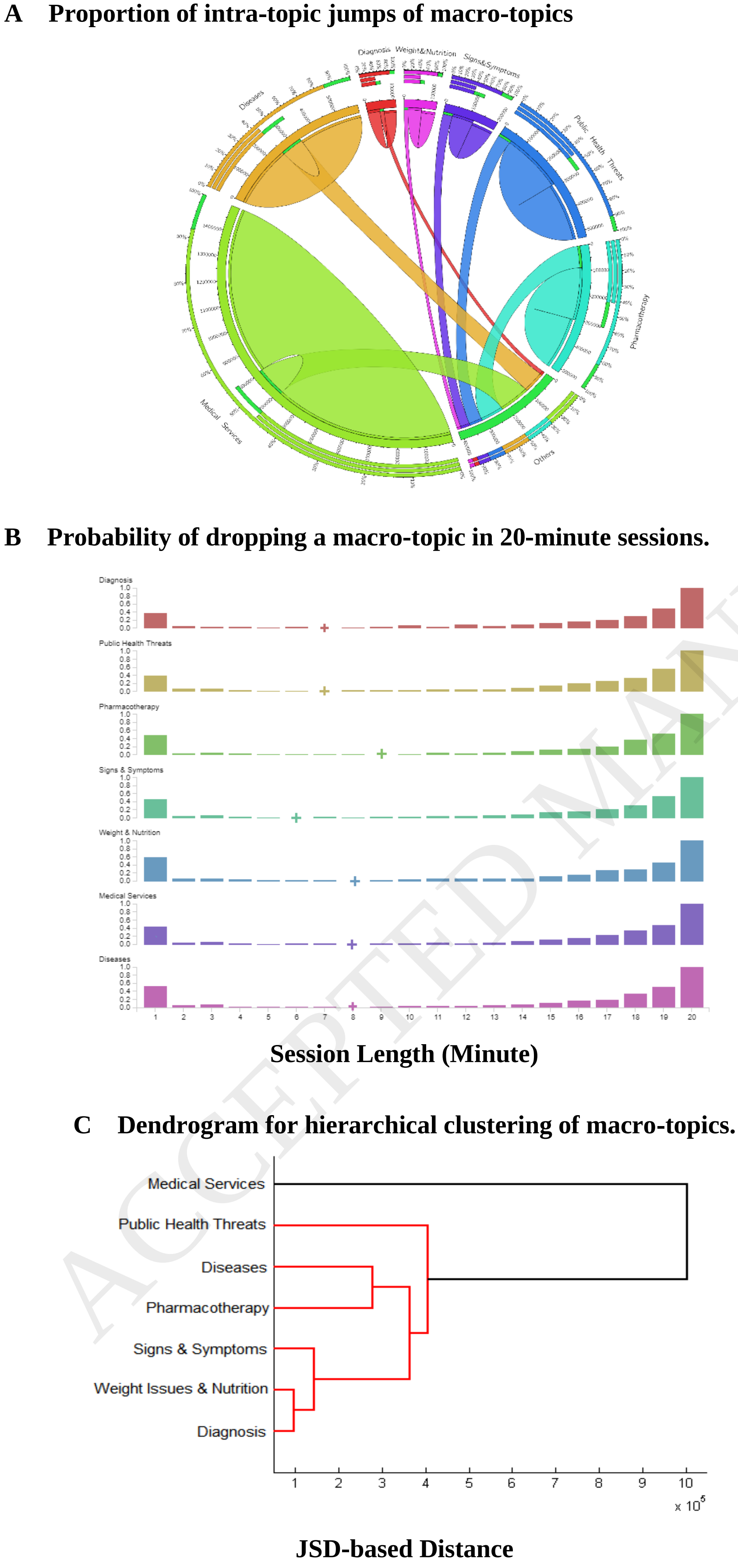
A Hourly distribution of macro-topics in search query repository.

B Interestingness of macro-topics in 24 hours of day.

Figure 1. Distribution and interestingness of macro-topics in 24 hours of day.

Panel A shows distribution of various macro-topics in a day. Hours of a day are organized in a circle in the center and the frequencies of topics are demonstrated using heights of stacks. Each tier of the diagram shows 100k queries and macro-topics are distinguished by colors.

Panel B shows interestingness levels of mentioned macro-topics in a period of 24 hours. The raw values of interestingness are scaled to the range of [0, 100] on y-axis and hours are organized on x-axis. Spectrum-derivate colors of the bars facilitate the discrimination between adjacent pairs of bars.



A Proportion of intra-topic jumps of macro-topics

B Probability of dropping a macro-topic in 20-minute sessions.

C Dendrogram for hierarchical clustering of macro-topics.

Figure 2. Diagrams of intra-topic jumps and probability of dropping macro-topics and dendrogram of HAC of topics.

Panel A shows the proportion of intra-topic jumps. The user may continue searching for a topic or switch to any other health-related topic. Each sector of the plot is assigned to a topic and size of the sector is determined proportionally to the normalized size of the topic. Width of edges between the "Others" and other sectors are estimated with respect to the ratio of within-sessions jumps to total count of searches in sessions.

Panel B demonstrates seven plots, each one shows the probability of dropping corresponding macro-topic in 20-minute sessions. The first minute in which the dropping probability reaches the zero is marked by + sign. The delay before zero probability is in the range of [6, 9] minutes.

Panel C shows dendrogram for hierarchical clustering of probability distributions. Time-series of probabilities were extracted and the Jensen–Shannon divergence of each pair was computed to form 7×7 matrix. At last, pairs of topics are clustered together one-by-one regarding weighted average linkage measure (see Supplementary Appendix for more comments on agglomerative hierarchical clustering).

2.2.2 Study B: Seasonality analyses of Micro-topics

In this study, the entire search logs were employed to extract temporal patterns of information-seeking using EPIPOI toolbox [35]. EPIPOI received raw time-series and smoothed them using digital filter proportioned the values per average number of searches in each year[3]. This tool visualizes the heat grid of the time-series as well (Figure 3).

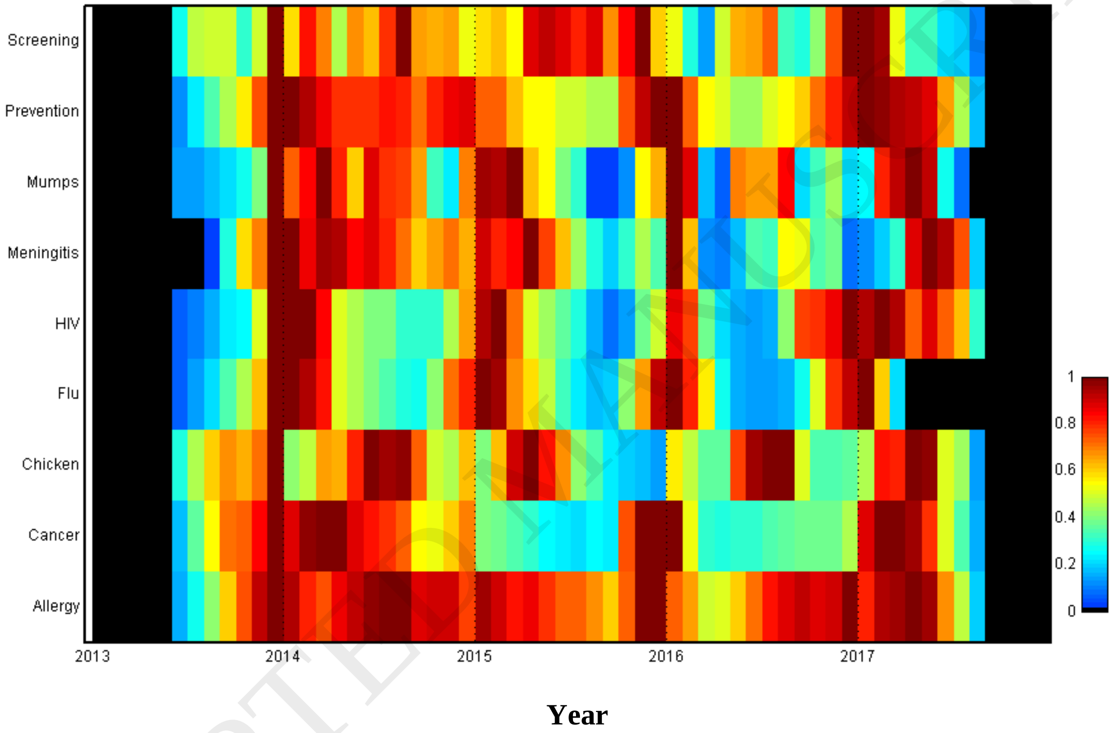


Figure 3. Diagrams of heat-grid of micro-topics.

This heat-grid diagram shows normalized frequency of micro-topics in entire query repository. As demonstrated above, the time-series of topics "Allergy", "Chicken Pox", "Flu", "HIV", "Meningitis", "Mumps" and "Immunization & Vaccination" were seasonal.

2.2.3 Study C: Correlative Analyses of Macro-topics

Disease spread prediction by employing searching trends is applicable to limited number of diseases which receive minor news coverage or that of rare with higher audience [36]. Thus, we conducted correlative analysis to study the impact of news coverage on users' information-seeking behaviors. The time-series of queries and news related to miscellaneous keywords are extracted and analyzed using Spearman's

[3] This option is provided by the EPIPOI toolbox, as one of built-in options.

correlation and Engle-Granger co-integration test. Our results (in Table 3) show that the Engle-Granger test rejects the null hypothesis in favor of alternative co-integration relation between pairs of news and query trends. Before computing Spearman's correlation, we perform lag analysis to determine delayed impact of news. The Spearman' Rho values range from 0.01 to 0.42 and show weak correlation between news and query trends.

3 Results

3.1 Temporal Analyses

Intuitively, the hourly distribution of queries was correlated with social activity level; thus, count of queries was minimum in midnight and maximum in midday (Figure 1A) and "Medical Services", "Public Health Treats", "Disease" and "Pharmacotherapy" were top interesting macro-topics, respectively.

In Figure 1-B, the interestingness levels of topics in various hours of the day did not follow the overall distribution of searches, e.g., interestingness levels of "Medical Service", "Diseases" and "Pharmacotherapy" topics were higher in early hours of the day (i.e. 0-8 am) that seems to be meaningful. Because in hours with lower level of medical services, users try to find emergency departments and medical centers serving at nights or try to self-medication by searching for diseases and their associated pharmacotherapy information. This plot shows that users prefer to search for their diseases and pharmacotherapy information instead. Interestingly, the peaks of interestingness of two topics "Medical Services" and "Diseases" are located respectively, which means people in the period between 1 am to 4 am prefer using medical services, while in the period between 4 am to 7 am they prefer self-diagnosis and self-medication.

Intra-topic jumps show the dependency and inter-relation of topics (Figure 2-A). We tracked users in 20-minute sessions and labeled their queries to enumerate topic-switching states. If a session involves *k* topics, then the session includes *k-1* topic-switching states. In this study, the minimum and maximum numbers of searched topics in sessions were 2 and 4, respectively. Therefore, sessions included at least 1 switching state and at most 3.

The sequence of macro-topic labels were used to draw topic's alluvial diagram (Figure 4). The alluvial diagram were designed to represent changes in structure of large networks and streams over time [37]. Alluvial diagram of topics, illustrated in Figure 4, shows the flow of searches routed from one topic to another. Each column of the diagram shows a switching state and each row represents a flow of searches. The width of the row (i.e. flow) is proportional to the frequency of users who select the topic in their sequence of searches. This diagram demonstrates the rational dependency of health-related topics and reveals the users' patterns of searching and deducing.

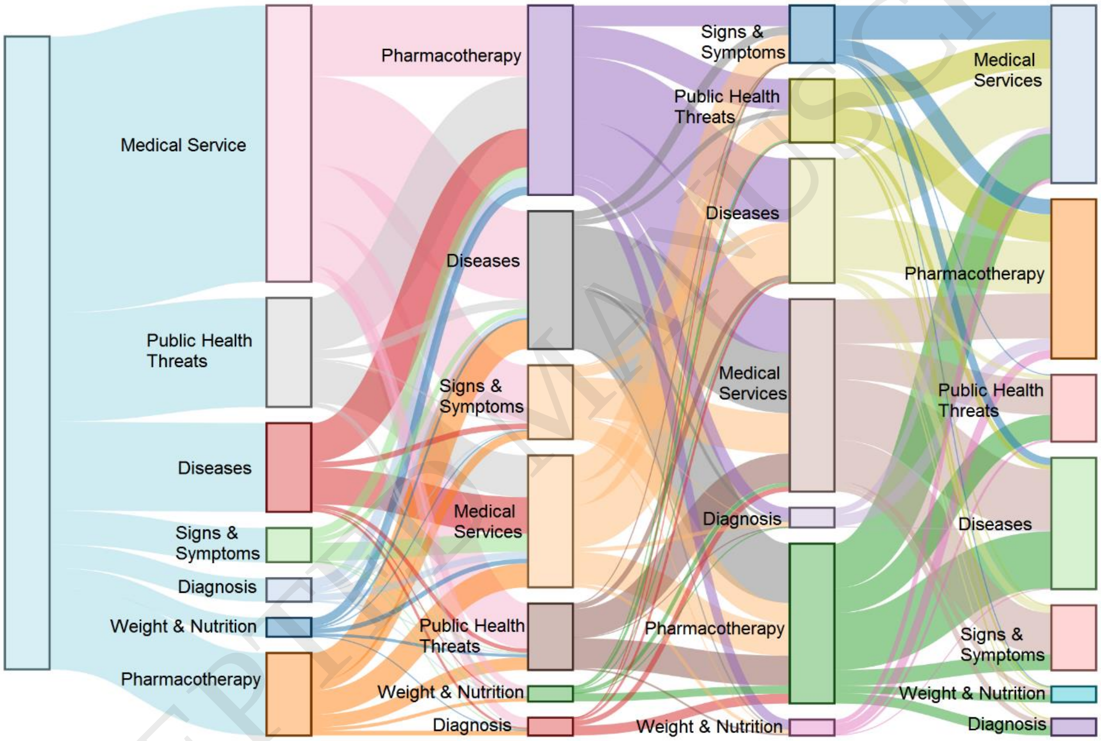


Figure 4. Alluvial diagram of topic switching in 20-minute search sessions.
Users may change their intended topic or switch between them in search sessions. Each column of this diagram shows a switching state and each row illustrates flow of searches routed through the corresponding sequence of topics. Maximum number of topics were covered in sessions was 4; therefore, the corresponding alluvial diagram has four columns. As shown in the figure, the major proportion of the flow belongs to topics, "Medical Services", "Pharmacotherapy", "Diseases" and "Public Health Threats".

To highlight the most important paths through the diagram, we selected two most likely branches succeeding each topic and draw the corresponding topic tree in Figure 5. This tree shows that in the first switching state (which is often the most important step), users preferred to seek information about "Medical Services" and "Pharmacotherapy". The topics "Diseases" and "Pharmacotherapy" were the mostly attractive topics in the second state. Pruning makes the tree sparse and subsequently clear to understand. The pruned tree reveals "Pharmacotherapy" and "Medical Services" are two frequent topics in the second searches of users.

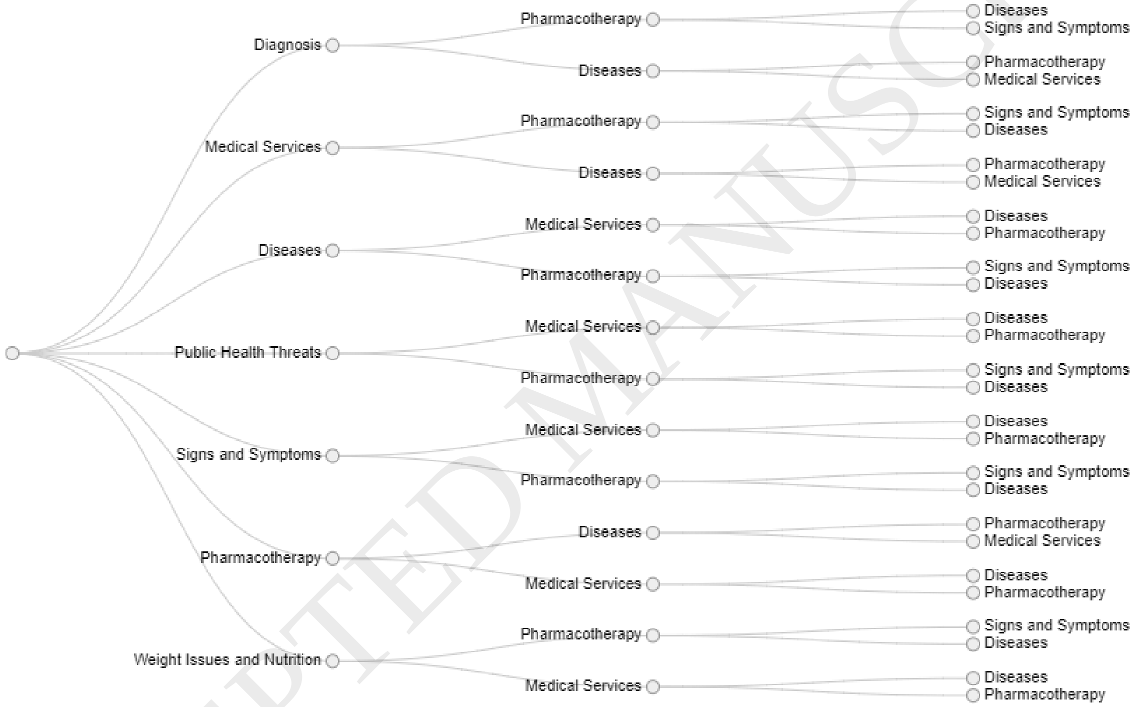


Figure 5. Tree diagram of major topics in 20-minute search sessions.
Each column of the alluvial diagram consists seven spots for topics. Thus, the associated tree must have same number of nodes in each layer; however, we decided to prune branches that were not likely to be selected by the users. From the second layer, we sorted the nodes according to the proportion of flow passing through. Then we selected the two nodes with highest flow to form the succeeding branches.

Similar to the second layer, in the third one, the "Pharmacotherapy" is the major choice. The topic "Diseases" is the second frequent choice in the third layer. The rationale is that, users search for "Signs and Symptoms" before searching for "Diseases" and then seeking for the "Pharmacotherapy" solutions.

However, results imply that users guess their diseases and subsequently search for information about “Diseases” and “Pharmacotherapy” solutions.

The probability of dropping a topic reflects the importance of the topic. The smaller the dropping probability is, the more important the topic is. Our results, demonstrated in Figure 2-B, showed that the dropping probability decreases to zero in first 6-9 minutes and then increases from 0 to 1 in the remained 12-13 minutes of the session. Among macro-topics, the “Pharmacotherapy” had the longest delay before reaching zero probability and the “Signs and Symptoms” had the shortest delay period.

3.2 Seasonality Analyses

Seasonality of searching trends are considered as evidences for seasonality of occurrences of diseases [11,36]. Our results show that time-series of topics “Allergy”, “Chicken Pox”, “Flu”, “HIV”, “Meningitis”, “Mumps” and “Immunization & Vaccination” were seasonal (Figure 3). Seasonality of mentioned diseases were studied in similar researches; however, seasonality of searches for immunization and vaccination have not been considered specifically. Most of infectious diseases supported by Iran’s government-demanded and optional vaccination programs are seasonal and subsequently searching for information about the immunization and vaccination were seasonal. “Cancer” and “Screening” micro-topics were not seasonal; however, they had peaks on the awareness days dedicated to different cancers and the week of awareness about cancer screening that imitates seasonality.

Among infectious diseases involved in seasonality analyses, HIV is less-studied disease. To the best of our knowledge, the seasonality of HIV infection has not been considered in similar studies [29,38]. Therefore, we turned our attention toward inspecting the seasonality of HIV in search trends. A study on the number of admissions of undiagnosed HIV cases in two urban hospitals [39] showed higher seroprevalence of undiagnosed HIV in the fall-winter admissions.

Panels A and B of Figure 6 show the average of monthly frequency of searched and seasonality diagram of HIV from April 2013 to April 2017, respectively. According to our collected data, users paid particular attention to HIV between months November and February (Figure 6-A). Seasonality diagram of the HIV (drawn by EPIPOI toolbox[35]) illustrates the repeated pattern of searching for HIV-related contents (Figure 6-B). In the winter of the year 2017, HIV attracted more interest due to awareness programs established in the last weeks of the year 2016 and continued to the second week of the year 2017. In addition to the nation-wide awareness program, an awareness campaign started by news promoted the wave of interest by the September of 2016.

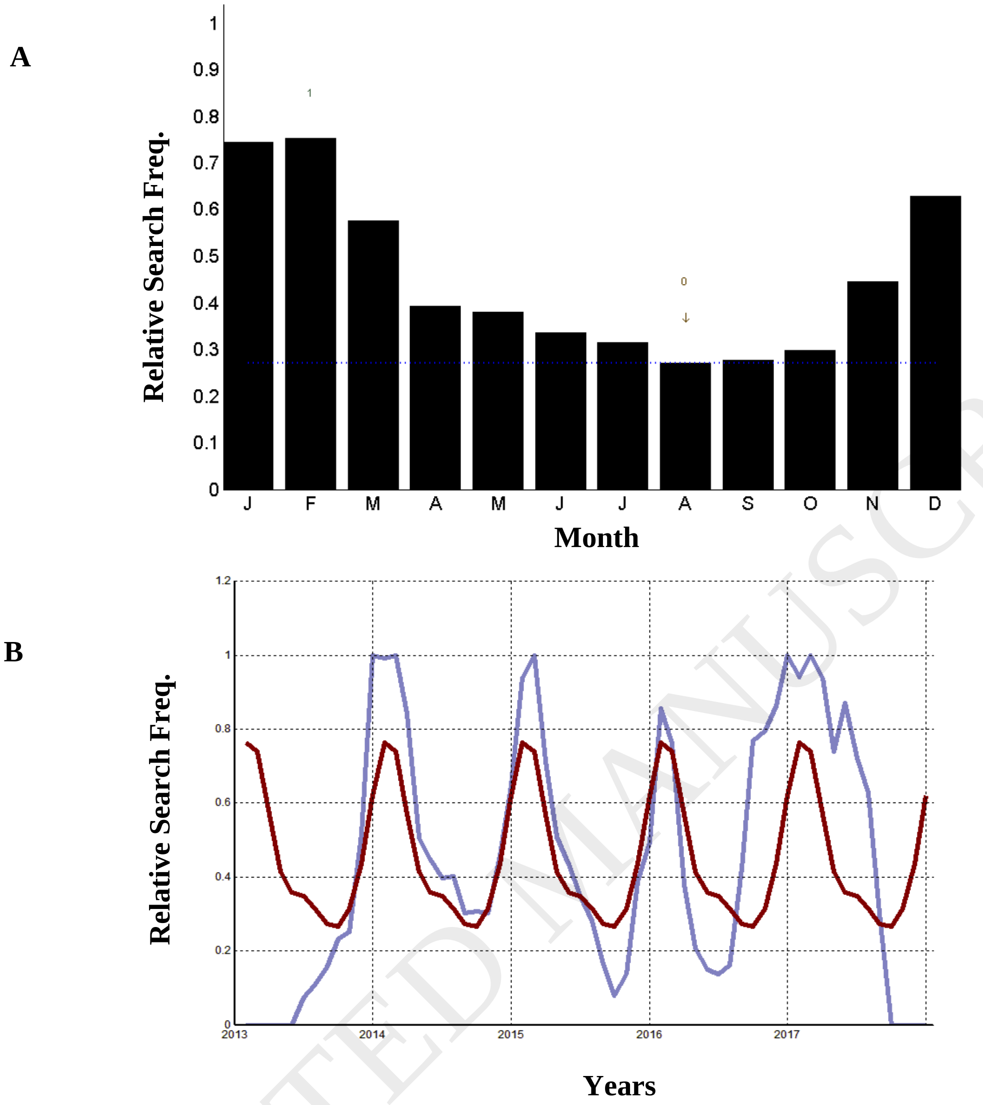


Figure 6. Seasonal HIV search trend and monthly frequency bar-chart used in seasonality study. Frequency of HIV-related searches were processed using EPIPOI's digital filter and normalized based on averages searches in each year. Panel A shows the normalized average of searches in months. In this plot, month February have the highest values and month August has the minimum value. Our findings support the results of a study which showed that higher seroprevalence of undiagnosed HIV in the fall-winter admissions [38]. Panel B illustrates the trends of HIV-related searches (blue line) and the seasonal pattern (red line).

3.3 Correlative Analyses

We decided to investigate the impact of the news on query trends using Engle-Granger co-integration test and Spearmen's correlation. According to the summarized results shown in Table 3, micro-topics can be divided into two groups. The first group includes "Meningitis", "Chicken Pox", "Mumps" and "HIV", and the second one contains "Allergy", "Flu", "Cancer", "Screening" and "Immunization & Vaccination". The first group of topics had lower Spearman's Rho values (i.e. Rho<0.2) than the second group. Results showed that most of topics with lower Rho values were sparsely covered in news media (i.e. News-count<2000) and the lags between news and query trends of the first group were higher than the second group. Delayed impact of news on searches demonstrated the lower level of impressiveness of these news.

Table 3. Results of Co-integration and Correlation Analyses of Micro-topics.

#	Topic	News Count	Co-integration		Spearman's Correlation		
			Label	p-value	Rho	p-value	Time Lag
1	Allergy	4886	True	<0.001	0.28	<0.001	0
2	HIV	1746	True	<0.001	0.14	0.007	8
3	Flu	1482	True	<0.001	0.42	<0.001	0
4	Mumps	108	True	<0.001	0.05	0.33	5
5	Chicken Pox	290	True	<0.001	0.10	0.051	7
6	Meningitis	384	True	<0.001	0.01	0.72	6
7	Cancer	13480	True	<0.001	0.36	<0.001	0
8	Screening	500	True	<0.001	0.20	<0.001	0
9	Immunization & Vaccination	6006	True	<0.001	0.41	<0.001	0

For each micro-topic studied in the correlative analysis, a time-lag value has been estimated. The time-lag determines the time period that is required to maximize the effectivity of the news published about the topic [29]. Hence, designers of awareness campaigns can make use of this time-lags to timely publicize their programs. It means that it is better for designers to make their awareness programs ready and wait until passing the time-lag after a news event to publicize their campaigns. By considering the time-lag in publishing the event, designers can ensure that their campaigns would attract particular attention timely and conveniently.

Co-integration test showed that trends of all micro-topics were co-integrated with their corresponding news pair. Time-series with positive correlation are synchronous in increasing/decreasing movements, but co-integrate time-series do not necessarily increase/decrease in the same manner. Co-integrate time-series move towards with a stable distance from each other.

Time-series of micro-topics were analyzed employing an agglomerative hierarchical categorization based on Jensen–Shannon divergence (JSD). Figure 7 illustrates the hierarchical clustering of the topics. "Mumps", "Meningitis", "Chicken Pox" and "Flu" were clustered together, respectively. Interestingly, "Screening" was very similar to the group of "Mumps" and "Meningitis" which were clustered before. Time-series of topics "Immunization & Vaccination" and "Allergy" were similar; however center of their cluster was farther away from the center belongs to the cluster of diseases. Time-series of the "Cancer" is the last one merged into the cluster containing other time-series.

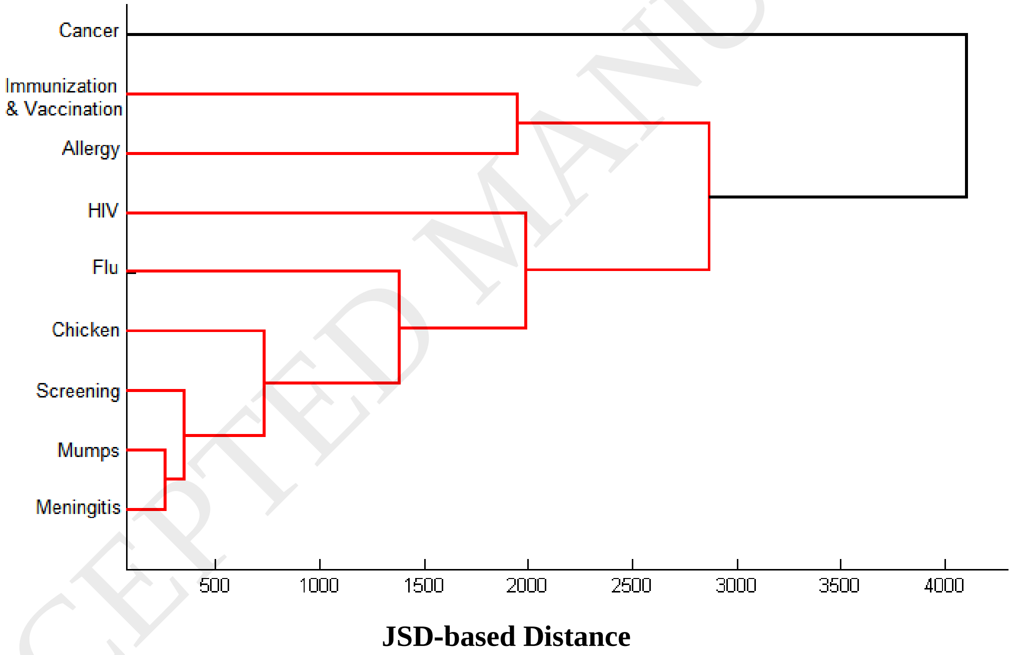


Figure 7. Dendrogram for hierarchical clustering of probability distribution of micro-topics. Micro-topics' time-series of probabilities were extracted and the Jensen–Shannon divergence of each pair was computed to form distance matrix. At last, pairs of topics are clustered together one-by-one regarding weighted average linkage measure.

4 Discussions

In this research we aimed to analyze users' information-seeking behaviors in health domain. An important point may be considered here is whether or not the results collected from a local search engine can be generalized to search engines used by users all over the world. Results of our previous analyses showed that search patterns and interests of Parsijoo's Persian native users resembles the patterns revealed by employing logs of world most popular search engines, e.g., Google and AOL, as well as local search engines, such as Baidu and Yandex. [40]

We selected two kinds of topics called macro/micro-topics. Analyses on macro-topics trends extracted from 48 months of search logs revealed that in cases where the topic was related to the treatment of diseases and where the availability level of health services were low, the interestingness level of the topic was higher than that of other. Agglomerative clustering of topics using Jensen-Shannon divergence showed that people searched for diseases and pharmacotherapy information similarly. In addition, the patterns of seeking information about diagnostic procedures and "Signs and Symptoms" were similar. Interestingly, searching patterns of "Weight issues and Nutrition" were analogous to diagnosis-related topics that suggested users interested in identifying cause of weight-related issues. The next fact can be extracted from the analyses is that the diagnosis-related topics have higher probability than treatment-related ones to be dropped by the users. By considering the intra-topic jumps, we concluded that requests related to medical services, pharmacotherapy and disease-specific information held majority proportion of intra-topic jumps.

Recently conducted studies supported a presumption in favor of search logs' ability to reveal the seasonality of diseases. Evidences in searches showed that seasonal diseases have seasonal search trends [6,11,14,41,42].

But, what if the search trend of a diseases is seasonal? During our investigation, we found out that search trend of the disease "Epidermolysis Bullosa" was seasonal. We drilled down to the deepest level of the logs and compared the time-series of search queries with corresponding news series. Our findings showed that monthly average of query related "Epidermolysis Bullosa (EB)" was less than 10. The EB-related trend had four peaks in 4 years of log which were happened in world Rare Disease Days. Since Trends of searches in different years were very similar and peaks were happened at equal time intervals, the trend of this disease imitated seasonal behavior. The trend of "Cancer" included more numbers of searches than the trends of "EB"; however, they showed comparable repeated pattern imitated the seasonality, too. Interestingly, similar phenomenon reported for trends of mushroom poisoning [36]. Although, we cannot ignore invaluable knowledge extracted from search trends analyses. Hence, it may seem rational to rectify the patterns extracted from search trends with a supplementary data resources, such as news documents. Finally, we should note that search queries can reveal aggregate behaviors of society facing health-related issues, but reliance on search trends alone may be deceiving.

Conclusion and Future Works

Search trend analyses provide benefits such as, (1) early outbreak detection, (2) studying effectivity of vaccination plans, (3) revealing seasonality patterns, and (4) demonstrating geospatial patterns of disease spread. Changes in time series of diseases reveal outbreaks earlier than aggregated formal reports, which makes control diseases departments (CDDs) able to prepare for the outbreaks and supply sufficient volumes of vaccine doses timely. Furthermore, CDDs are benefitted from the advantage of temporal similarity of disease outbreaks and search trends during previous seasons to predict outbreak intension. We should note that seasonality of a search trends does not imply seasonality of disease. That is, why we suggested rectifying search trends with external data sources. There exist various studies [12,14,43] which showed that seasonal disease activities follow latitudinal trends. Health authorities can take advantage of this fact to prioritize immunization programs in different regions. We plan to expand our research to geospatial dimension of digital epidemiology and explore the impact of location on time lags between outbreaks in various regions.

Authorship contributions

- Study conception and design: Fatemeh Kaveh-Yazdy, Ali-Mohammad Zareh-Bidoki
- Manuscript writing and preparation: Fatemeh Kaveh-Yazdy, Ali-Mohammad Zareh-Bidoki
- Contributed Data and Analysis Tools: Fatemeh Kaveh-Yazdy, Ali-Mohammad Zareh-Bidoki
- Date Collection: Fatemeh Kaveh-Yazdy

Funding

None.

Disclosure of conflicts of interest

None.

Acknowledgement

Authors would like to thank to Parsijoo's CTO, Dr. Sajjad Zarifzadeh and members of distributed computation team, M. Fallah, A. Raeiszadeh and M. Bagheresmaeili for their helpful suggestions and technical supports. We thank Dr. Mohammad-Ali Kaveh-Yazdy, MD and Dr. Samane Jahanabadi, PhD of Pharmacology, for their helpful comments.

Appendix A: Data Preparation

A1: Macro-Topics' Keyword Preparation

Major preprocesses used in this paper were:

- Sliding window on text
- Translating phrases
- Transliterating phrases
- Keyword Expansion

Long descriptions of ICD-10 codes are unlikely to be used by the users, thus we slided a window of length 3 on descriptions to break them into shorter phrases (i.e. trigrams). Trigrams extracted from the ICD-10 descriptions were in English. We used three services to translate English trigrams into Persian. In the first step, English phrases were send to the API provided by the local publisher of the English-Persian Dorland's Medical dictionary. Remained phrases were sent to the Google translate and the Parsijoo translation services simultaneously and the most probable translations were chosen to be used as Persian equivalents.

Furthermore, trigrams did not have probable Persian translation were removed from the set to avoid noise of ambiguous and meaningless phrases.

In addition to list of the medical centers provided by the MHME[4], general words such as “office”, “physician”, “pharmacy”, “drug store” and etc. in both English and Persian languages were added to the keyword set of the topic “Medical Services”.

Watzlaf et al. [44] studied public health threats using a list containing CDC’s notifiable infectious diseases, top 10 leading causes of mortality in US and ICD-9-CM classification for morbidity/mortality related to terrorism. We localized the proposed list by replacing the CDC’s list with the list of Iran’s infectious diseases [45,46] and the top causes of mortality in US with the identical list for Iran [47].

Weight issues and nutrition-related keywords were extracted from phrases of an automatically generated category-graph [48] using pages categorized under the English Wikipedia “Nutrition” and “Diets” categories. Similarly, same category-graph was extracted from identical categories of the Persian Wikipedia. The keywords of Pharmacotherapy topic included Iran’s FDA approved drug list. Given that the Persian native users use Persian transliteration of drug name more likely; we used the Parsijoo transliteration service to generate Persian transliterations of drug names.

The collected set of Persian and English macro-topics’ keywords included 10209 unique keywords that were employed in retrieving and labeling queries.

A2: Micro-Topics’ Keyword Preparation

Keyword set of micro-topics was limited to the numerable list shown in Table 2. This list of keywords was not expanded or broken into shorter phrases. However, the Persian equivalent phrases were added. These phrases and words were translated using an API provided by the local publisher of the Dorland’s medical dictionary.

[4] The Ministry of Health and Medical Education (MHME)

Appendix B: Time-series Generation

B1: Query & News Retrieval

The topic-related algorithm utilized in this research involves two parts called Topic Detection using AGF (TDA) and Topic Clustering and Tweet Retrieval (TCTR) [30]. The TDA method detects topics using frequent patterns that was done in this research by selecting keywords and topics manually. Thus, the process of retrieving related queries/news was limited to the TCTR process. In the TCTR step, proximity between each query/news and keywords of each topic were calculated and the queries/news which were most similar to a topic were labeled.

B2: Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering (AHC) is a bottom-up method which starts from a set of objects that are labeled by their own cluster. It means in a dataset containing N unique objects, AHC is initialized with N unique clusters. This algorithm iteratively merges clusters with their nearest clusters until reaching one single cluster. In each iteration, AHC estimates the similarity of cluster pairs using a proximity measure called linkage measure. There are several linkage measures, such as minimum, maximum and average distances. In this study, we used “Weighted Pair Group Method with Arithmetic Mean (WPGMA)” algorithm [49] that is a variation of AHC based on weighted average linkage measure. In WPGMA algorithm, the distance between two merging clusters is computed recursively using sub-clusters of them. Suppose that X is a cluster made up of two clusters, i.e. $X.a$ and $X.b$, and Y is the second cluster. The distance between X and Y, denoted by $d(X,Y)$ is calculated as,

$$d(X,Y) = \frac{d(X.a\,,Y) + d(X.b\,,Y)}{2} \tag{B1}$$

Whereas the equation (B1) denoted the linkage measure, the distance measure has not been described yet. We employed the Jensen–Shannon divergence (JSD) as distance measure. JSD is defined based on two Kullback-Leibler divergence. Let P and Q be discrete probability distributions.

The Kullback-Leibler divergence of P and Q is defined as,

$$D_{KL}\,(P \parallel Q) \sum_{i} P(i)\ \log \frac{P(i)}{Q(i)} \tag{B2}$$

Accordingly, the Jensen-Shannon divergence is computed as,

$$JSD\,(P\ \parallel Q) = \frac{1}{2}(D_{KL}\,(P \parallel M)) \tag{B3}$$

where M is the mean of P and Q, i.e. $M = \frac{1}{2}(P + Q)$. The JSD measure is a symmetric measure in the range of $[0.0, +\infty)$. We normalized the values of time-series by dividing the number of searches in days into total number of searches in the corresponding year. The normalized values of time-series are fall between 0.0 and 1.0. Then, the matrix of distances between these distributions were computed to be employed by the WPGMA algorithm to form a single cluster. Results of clustering are illustrated in the form of dendrogram plots (Figures 2-C and 7).

Summary Points

- Users' information-seeking behaviors varied by time of the day.
- Trend belonging to "Immunization and Vaccination" is seasonal we well as the trends of diseases such as, Mumps, Flu, Chicken Pox, and Meningitis.
- Trends belong to diseases which received minor social attention are vulnerable to be misclassified as seasonal trends.
- News and search trends are weakly correlated, while they still be co-integrate and move towards with a stable distance from each other.
- Analyzing search queries without taking into account the impact of news and external events can be deceiving.

References

[1] S. Yang, M. Santillana, S.C. Kou, Accurate estimation of influenza epidemics using Google search data via ARGO, Proc. Natl. Acad. Sci. 112 (2015) 14473–14478. doi:10.1073/pnas.1515373112.

[2] S. Marcel, F.C. C., M.S. R., T.A. F., B.J. S., Influenza A (H7N9) and the Importance of Digital Epidemiology, N. Engl. J. Med. 369 (2013) 401–404. doi:10.1056/NEJMp1307752.

[3] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature. 457 (2009) 1012–1014. http://dx.doi.org/10.1038/nature07634.

[4] S. Cook, C. Conrad, A.L. Fowlkes, M.H. Mohebbi, Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic, PLoS One. 6 (2011) 1–8. doi:10.1371/journal.pone.0023610.

[5] D. Lazer, R. Kennedy, G. King, A. Vespignani, The Parable of Google Flu: Traps in Big Data Analysis, Science (80-.). 343 (2014) 1203–1205.

[6] P. Guo, J. Zhang, L. Wang, S. Yang, G. Luo, C. Deng, Y. Wen, Q. Zhang, Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model, Sci. Rep. 7 (2017) 46469. doi:10.1038/srep46469.

[7] H. Woo, Y. Cho, E. Shim, J.-K. Lee, C.-G. Lee, S.H. Kim, Estimating Influenza Outbreaks Using Both Search Engine Query Data and Social Media Data in South Korea., J. Med. Internet Res. 18 (2016) e177. doi:10.2196/jmir.4955.

[8] Y. Teng, D. Bi, G. Xie, Y. Jin, Y. Huang, B. Lin, X. An, D. Feng, Y. Tong, Dynamic Forecasting of Zika Epidemics Using Google Trends, PLoS One. 12 (2017) 1–10. doi:10.1371/journal.pone.0165085.

[9] K. Liu, T. Wang, Z. Yang, X. Huang, G.J. Milinovich, Y. Lu, Q. Jing, Y. Xia, Z. Zhao, Y. Yang, S. Tong, W. Hu, J. Lu, Using Baidu Search Index to Predict Dengue Outbreak in China, Sci. Rep. 6 (2016) 38040. doi:10.1038/srep38040.

[10] C. Alicino, N.L. Bragazzi, V. Faccio, D. Amicizia, D. Panatto, R. Gasparini, G. Icardi, A. Orsi, Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes, Infect. Dis. Poverty. 4 (2015). doi:10.1186/s40249-015-0090-9.

[11] J.C. Ku, N.M. Alotaibi, J. Wang, G.M. Ibrahim, T.A. Schweizer, R.L. Macdonald, Internet search volumes in brain aneurysms and subarachnoid hemorrhage: Is there evidence of seasonality?, Clin. Neurol. Neurosurg. 158 (2017) 1–4. doi:10.1016/j.clineuro.2017.04.008.

[12] S.D. Willard, M.M. Nguyen, Internet Search Trends Analysis Tools Can Provide Real-time Data on Kidney Stone Disease in the United States, Urology. 81 (2013) 37–42. doi:10.1016/j.urology.2011.04.024.

[13] M. Schootman, A. Toor, P. Cavazos-Rehg, D.B. Jeffe, A. McQueen, J. Eberth, N.O. Davidson, The utility of Google Trends data to examine interest in cancer screening, BMJ Open. 5 (2015). http://bmjopen.bmj.com/content/5/6/e006678.abstract.

[14] K.M. Bakker, M.E. Martinez-Bakker, B. Helm, T.J. Stevenson, Digital epidemiology reveals global childhood disease seasonality and the effects of immunization, Proc. Natl. Acad. Sci. 113 (2016) 6689–6694. doi:10.1073/pnas.1523941113.

[15] E. Yom-Tov, L. Fernandez-Luque, Information is in the eye of the beholder: Seeking information on the MMR vaccine through an Internet search engine, AMIA Annu. Symp. Proc. 2014 (2014) 1238–1247.

[16] L.Y. Fu, K. Zook, Z. Spoehr-Labutta, P. Hu, J.G. Joseph, Search Engine Ranking, Quality, and Content of Web Pages That Are Critical Versus Noncritical of Human Papillomavirus Vaccine., J. Adolesc. Heal. 58 (2016) 33–39. doi:10.1016/j.jadohealth.2015.09.016.

[17] Y. Cha, C.A. Stow, Mining web-based data to assess public response to environmental events, Environ. Pollut. 198 (2015) 97–99. doi:10.1016/j.envpol.2014.12.027.

[18] S. Fox, Health Topics: 80% of Internet Users Look for Health Information Online, Pew Internet & American Life Project, Washington, DC, USA, 2011. https://books.google.com/books?id=y7NNnQAACAAJ.

[19] N. Van Riel, K. Auwerx, P. Debbaut, S. Van Hees, B. Schoenmakers, The effect of Dr Google on doctor–patient encounters in primary care: a quantitative, observational, cross-sectional study, BJGP Open. (2017). http://bjgpopen.org/content/early/2017/05/12/bjgpopen17X100833.abstract.

[20] S.S.-L. Tan, N. Goonawardene, Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review, J. Med. Internet Res. 19 (2017). doi:10.2196/jmir.5729.

[21] M. Chung, R.P. Oden, B.L. Joyner, A. Sims, R.Y. Moon, Safe Infant Sleep Recommendations on the Internet: Let's Google It, J. Pediatr. 161 (2017) 1080–1084.e1. doi:10.1016/j.jpeds.2012.06.004.

[22] H. Rayess, G.F. Zuliani, A. Gupta, P.F. Svider, A.J. Folbe, J.A. Eloy, M.A. Carron, Critical Analysis of the Quality, Readability, and Technical Aspects of Online Information Provided for Neck-Lifts, JAMA Facial Plast. Surg. 19 (2017) 115–120. doi:10.1001/jamafacial.2016.1219.

[23] T. Roughead, D. Sewell, C.J. Ryerson, J.H. Fisher, A.M. Flexman, Internet-Based Resources Frequently Provide Inaccurate and Out-of-Date Recommendations on Preoperative Fasting: A Systematic Review, Anesth. Analg. 123 (2016) 1463–1468. doi:10.1213/ANE.0000000000001590.

[24] H. Pamela, G. Jerome, Untangling the Web — Patients, Doctors, and the Internet, N. Engl. J. Med. 362 (2010) 1063–1066. doi:10.1056/NEJMp0911938.

[25] B. Schrank, I. Sibitz, A. Unger, M. Amering, How Patients With Schizophrenia Use the Internet: Qualitative Study, J. Med. Internet Res. 12 (2010). doi:10.2196/jmir.1550.

[26] J. Laugesen, K. Hassanein, Y. Yuan, The Impact of Internet Health Information on Patient Compliance: A Research Model and an Empirical Study, J. Med. Internet Res. 17 (2015). doi:10.2196/jmir.4333.

[27] Z. Obermeyer, E.J. Emanuel, Predicting the Future — Big Data, Machine Learning, and Clinical Medicine, N. Engl. J. Med. 375 (2016) 1216–1219. doi:10.1056/NEJMp1606181.

[28] G. Murray, C. O'Rourke, J. Hogan, J.E. Fenton, Detecting internet search activity for mouth cancer in Ireland, Br. J. Oral Maxillofac. Surg. 54 (2016) 163–165. doi:10.1016/j.bjoms.2015.12.005.

[29] C. APY, L. Q, H. D., News trends and web search query of HIV/AIDS in Hong Kong, PLoS One. 12 (2017).

[30] A. Benny, M. Philip, Keyword based tweet extraction and detection of related topics, in: Procedia Comput. Sci., Elsevier, 2015: pp. 364–371. doi:10.1016/j.procs.2015.02.032.

[31] L. Geng, H.J. Hamilton, Interestingness Measures for Data Mining: A Survey, ACM Comput. Surv. 38 (2006).

[32] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, Z. Wang, A novel feature selection algorithm for text categorization, Expert Syst. Appl. 33 (2007) 1–5. doi:10.1016/j.eswa.2006.04.001.

[33] J. Yang, Z. Qu, Z. Liu, Improved feature-selection method considering the imbalance problem in text categorization., ScientificWorldJournal. 2014 (2014) 625342. doi:10.1155/2014/625342.

[34] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

[35] W.J. Alonso, B.J.J. McCormick, EPIPOI: A user-friendly analytical tool for the extraction and visualization of temporal parameters from epidemiological time series, BMC Public Health. 12 (2012) 982. doi:10.1186/1471-2458-12-982.

[36] G. Cervellin, I. Comelli, G. Lippi, Is Google Trends a reliable tool for digital epidemiology? Insights

from different clinical settings., J. Epidemiol. Glob. Health. 7 (2017) 185–189. doi:10.1016/j.jegh.2017.06.001.

[37] M. Rosvall, C.T. Bergstrom, Mapping Change in Large Networks, PLoS One. 5 (2010) 1–7. doi:10.1371/journal.pone.0008694.

[38] A.B. Jena, P. Karaca-Mandic, L. Weaver, S.A. Seabury, Predicting new diagnoses of HIV infection using internet search engine data., Clin. Infect. Dis. 56 (2013) 1352–1353. doi:10.1093/cid/cit022.

[39] K.A. Brady, S. Berry, R. Gupta, M. Weiner, B.J. Turner, Seasonal variation in undiagnosed HIV infection on the general medicine and trauma services of two urban hospitals., J. Gen. Intern. Med. 20 (2005) 324–330. doi:10.1111/j.1525-1497.2005.40300.x.

[40] F. Kaveh-Yazdy, A.-M. Zareh-Bidoki, M.-R. Pajoohan, Linguistic Analysis of Interaction Patterns and Users' Query Reformulation Strategies in Persian Search Engine, Lang. Linguist. Article in (2018).

[41] H. Yu, W.J. Alonso, L. Feng, Y. Tan, Y. Shu, W. Yang, C. Viboud, Characterization of Regional Influenza Seasonality Patterns in China and Implications for Vaccination Strategies: Spatio-Temporal Modeling of Surveillance Data, PLOS Med. 10 (2013) 1–16. doi:10.1371/journal.pmed.1001552.

[42] D.R. Olson, K.J. Konty, M. Paladini, C. Viboud, L. Simonsen, Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales, PLoS Comput Biol. 9 (2013) e1003256. doi:10.1371/journal.pcbi.1003256.

[43] J. Paireau, A. Chen, H. Broutin, B. Grenfell, N.E. Basta, Seasonal dynamics of bacterial meningitis: a time-series analysis, Lancet. Glob. Heal. 4 (2016) e370-7. doi:10.1016/S2214-109X(16)30064-X.

[44] V.J.M. Watzlaf, J.H. Garvin, S. Moeini, P. Anania-Firouzan, The Effectiveness of ICD-10-CM in Capturing Public Health Diseases, Perspect. Heal. Inf. Manag. 4 (2007) 6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2047296/.

[45] M. Askarian, R. Mansour Ghanaie, A. Karimi, F. Habibzadeh, Infectious diseases in Iran: A bird's eye view, Clin. Microbiol. Infect. 18 (2012) 1081–1088. doi:10.1111/1469-0691.12021.

[46] A. Khoshdel, M. Noori Fard, R. Pezeshkan, A. Salahi-Moghaddam, AWT_TAG, Mapping the Important Communicable Diseases of Iran, J. Heal. Dev. 1 (2012) 31–46. http://jhad.kmu.ac.ir/article-1-124-en.html.

[47] S. Saadat, M. Yousefifard, H. Asady, A. Moghadas Jafari, M. Fayaz, M. Hosseini, The Most Important Causes of Death in Iranian Population; a Retrospective Cohort Study, Emergency. 3 (2015) 16–21.

[48] M. Alemzadeh, R. Khoury, F. Karray, Exploring Wikipedia's Category Graph for Query Classification, in: M. Kamel, F. Karray, W. Gueaieb, A. Khamis (Eds.), Auton. Intell. Syst. Second Int. Conf. AIS 2011, Burn. BC, Canada, June 22-24, 2011. Proc., Springer Berlin Heidelberg, Berlin, Heidelberg, 2011: pp. 222–230. doi:10.1007/978-3-642-21538-4_22.

[49] R.R. Sokal, C.D. Michener, A statistical method for evaluating systematic relationships, Univ. Kansas Sci. Bull. 38 (1958) 1409–1438.