Predicting Epidemic Tendency through Search Behavior Analysis*

Danqing Xu, Yiqun Liu, Min Zhang, Shaoping Ma, Anqi Cui, Liyun Ru
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
xudanqing06@gmail.com

Abstract

The possibility that influenza activity can be generally detected through search log analysis has been explored in recent years. However, previous studies have mainly focused on influenza, and little attention has been paid to other epidemics. With an analysis of web user behavior data, we consider the problem of predicting the tendency of hand-foot -and-mouth disease[1] (HFMD), whose outbreak in 2010 resulted in a great panic in China. In addition to search queries, we consider users' interactions with search engines. Given the collected search logs, we cluster HFMD-related search queries, medical pages and news reports into the following sets: epidemic-related queries (ERQs), epidemic-related pages (ERPs) and epidemic-related news (ERNs). Furthermore, we count their own frequencies as different features, and we conduct a regression analysis with current HFMD occurrences. The experimental results show that these features exhibit good performances on both accuracy and timeliness.

1 Introduction

Seasonal epidemics have posed a tremendous threat to public health. The panic caused by both Influenza A (H1N1) and Severe Acute Respiratory Syndrome (SARS) flu has had a terrible impact on both economic and social development worldwide. With the rapid development of the Internet, search engines have become an important gateway to obtaining information. Consequently, many social events, including the spread of epidemics, can be traced from users' search logs. The Web provides abundant medical resources for users. Approximately 80% of consumers turn first to the Internet when confronted with health problems [Fox, 2006]. The idea that epidemic tendency can be generally detected through Web information has been explored in recent years [Ginsberg *et al*., 2009; Heather *et al.*, 2004; Hulth *et al.*, 2009; Pelat *et al.*, 2009; Philip *et al*., 2008]. These studies show that the frequency of online search queries is strongly correlated with epidemic events. This correlation makes it possible to detect epidemic outbreaks in areas with a large population of web users.

However, most studies focus on influenza, and their results are not applied to other diseases. In recent years, HFMD has become a serious disease for Chinese infants from one to three years old. A large outbreak of HFMD in 2010 resulted in a great panic in China. At the peak of the epidemic, tens of thousands of children were infected every week. The cycle of this infectious disease is brief, and it can be observed. No suitable HFMD vaccine has been developed so far, and infants may die without timely treatment. Thus, studies of HMFD have become indispensable for epidemic research [Dong and Sun, 2008; Wang *et al.*, 2009]. In an attempt to provide early detection, [Wang *et al.*, 2009] establish a seasonal HFMD trend model to predict future HFMD trends. This approach may have some viability, but a sudden outbreak cannot be predicted in a timely manner through this model. In this paper, we will track HFMD activity by a method of analyzing huge volumes of search logs.

Generally speaking, most online health information seekers first take actions such as self-diagnosis and self-medication when they show slight symptoms [Fox, 2006]. If the symptoms get worse, they may again search online for specific symptoms or diseases and decide whether or not to see a doctor. For example, when a child with HFMD begins to have a slight cough, his/her parents may not know the specific disease and will merely search for this symptom online. With more severe symptoms, the specific disease can be recognized, and more detailed information will have to be obtained from search engines again. Finally, the child is diagnosed as an HFMD case. Through this whole process, we find that there is a considerable lag between searching and reporting but that users' search logs may contain related information much earlier. Therefore, it is possible to detect or even predict an epidemic tendency on the basis of information-seeking behaviors.

Previous studies have discovered that the frequencies of search queries can not only estimate current influenza activity [Ginsberg *et al.*, 2009] but also predict an increase of

* Supported by Natural Science Foundation (60736044, 60903107) and Research Fund for the Doctoral Program of Higher Education of China (20090002120005)

[1] http://en.wikipedia.org/wiki/HFMD

influenza in advance [Philip *et al.*, 2008]. These search queries come from patients and non-patients, and different user interactions may represent different user needs. If a user clicks related medical pages, we think that he/she has a higher probability of being infected than those who do not click any medical pages. In addition, some news websites provide alerts of important infectious events and outbreaks during the HFMD season. These news reports may produce a positive effect on web users and make some users submit related queries to search engines. [Fox, 2006] reveals that approximately 7% of respondents without signs of the disease follow news stories about epidemics every day. However, little attention is paid to these situations. If the frequency of ERQs alone is considered, some situations may be misidentified or not covered. In addition to the ERQs proposed in previous studies [Ginsberg *et al.*, 2009; Heather *et al.*, 2004; Hulth *et al.*, 2009; Pelat *et al.*, 2009; Philip *et al.*, 2008], this paper introduces two new features: epidemic related pages (ERPs) and news (ERNs). The "ERPs" feature records the frequency of medical pages being visited, while the "ERNs" represents the popularity of epidemic-related news stories. On the basis of these features, this paper provides a solution to predict the number of future HFMD cases one week in advance of when the outbreak occurs.

The main contributions of this paper are as follows: 1.In addition to search queries, we combine user interaction into analysis. Besides ERQs, this paper introduces two new features: ERPs and ERNs, and other search situations will be covered; 2.This is the first time that online search surveillance has been developed for HFMD. Previous works mainly focus on influenza. Given search logs, our results show that web surveillance is also suitable for HFMD.

The rest of this paper is organized as follows: After a description of the related work in the next section, we introduce the information of user behavior data sets and details of the three proposed features in Section 3. In Section 4, different features are combined on the basis of linear models, and the performance of each model is evaluated. We discuss and conclude this work in Sections 5 and 6.

2 Related Work

With the widespread use of the Internet, many patients tend to obtain information online when they meet with any medical symptoms. Fox [Fox, 2006] shows that approximately 80% of American adults search online for medical information about specific diseases or symptoms every year. This survey also shows that 7% of users search for health information or follow epidemic news stories on a typical day. On average, 66% seekers begin their online search from a search engine, while 27% begin at a health-related website. The Web can help users better understand health status within a short amount of time. This fact makes it possible to utilize search logs to detect or predict epidemic activity in a timely manner.

Ginsberg *et al*. use Google search logs in the United States to estimate influenza activity in each state [Ginsberg *et al*., 2009]. They proposed a method of analyzing a large amount of Google search queries to track the trend of influenza-like illness (ILI). They select influenza-like illness related queries (IRQs or ERQs) and draw a comparison between the influenza activity and the total number of the influenza related queries submitted in some areas during the outbreak period. Their results suggest a high correlation between ILI-related queries and official released ILI occurrences. Philip *et al.* also examine the relationship between influenza-related searches and actual influenza cases on the basis of Yahoo! search logs [Philip *et al*., 2008]. They find that the frequency of influenza-like symptom searches can not only detect influenza activity but also predict an increase in the mortality of influenza in advance. All of their results show that the selected ERQs strongly correlate with the current level of influenza.

In addition, [White and Horvitz, 2009] survey 515 individuals' online health-related searches, and they perform a log-based study of how people search for related medical information online. Their result shows that there may be some escalation of medical concern on user search logs, where queries about serious illness commonly follow behind initial queries about common symptoms. This result suggests that the Web has the potential to increase search engine users' anxieties, leading to "cyberchondria". Thus, if the frequency of ERQs alone is counted, errors may be introduced in the estimation. If user interaction is taken into consideration, this error may be avoided. In addition to online queries, Heather *et al.* study the relationship between actual influenza occurrences and the number of health-related website visits [Heather *et al*., 2004]. They discover that the frequency of influenza-related article accesses is strongly positively correlated with the CDC's traditional surveillance data.

3 Dataset and Features

3.1 Dataset

HFMD typically occurs in small epidemics in kindergartens during the spring and summer months. During this period, the Chinese Center for Disease Control and Prevention (CDC) reports the total number of HFMD-infected cases every week to monitor HFMD occurrences. These cases are collected weekly from clinical laboratories and hospitals in 31 provinces. Taking data integrity and continuity into consideration, HFMD data from February 2009 to September 2010 (two HFMD seasons) were selected as our experimental data. These data are available at http://www.chinacdc.cn/.

In addition, with the help of a widely-used Chinese commercial search engine, anonymous search logs were gathered for the same period. Approximately 70 million search entries with 12 million users are collected every day. In consideration of user privacy, only queries submitted and URLs clicked by users were extracted in our experiment.

3.2 Features

The Internet provides an abundance of resources, and it determines how people search for information. Typically, users submit queries to search engines and click the corresponding items according to their needs. Both queries and click interaction are treated as the representation of users' intents. If a user only submits a query but does not click any result, we assume that this kind of behavior is not as meaningful as querying with result being clicked. In addition, an increase in search queries can be caused by important news reports. Previous studies [Ginsberg *et al.*, 2009; Heather *et al.,* 2004; Hulth *et al.*, 2009; Pelat *et al.*, 2009; Philip *et al.*, 2008] have mainly adopted the frequency of epidemic related search queries to track the current epidemic activity. Thus, the above situations cannot be well distinguished from each other. In additional to search queries, both click interactions and the influence caused by public news are introduced. In our experiment, the number of HFMD-related medical articles (pages) being clicked is an important basis to judge the current level of HFMD activity. At the same time, we select the number of related web news reports to represent the HFMD popularity of public media.

Epidemic related queries (ERQs). A total of 66% of health seekers submit a health inquiry to a search engine for online information [Fox, 2006]. Hence, queries can be treated as an important resource for web mining and can play key roles in representing users' search intent. On the basis of this idea, Ginsberg *et al.* find that the frequencies of certain influenza-related queries are highly correlated with the number of influenza cases [Ginsberg *et al.*, 2009]. In addition, Philip *et al.* discover that the total counts of some queries of influenza-like symptoms can predict an increase of influenza activity [Philip et al., 2008]. According to previous work, we select ERQs as the first feature, and we quantify them by counting the total frequency during a certain period. In our experiment, the ERQs are the set of HFMD related queries, and they are obtained by the method of query clustering [Baeza-Yates and Tiberi, 2007; Beeferman and Berger, 2000; Chan *et al.*, 2004; Wen *et al.,* 2001].

Epidemic related medical pages (ERPs). As we all know, the Internet provides abundant information, and users click different results according to their needs. [Fox, 2006] shows that 27% of health information seekers obtain online health information through websites. Medical websites are an important source of information for health seekers. If the number of visiting HFMD medial websites or pages increases, an outbreak of HFMD may occur during this period. In this paper, we collect HFMD-related medical articles from popular medical websites, and we gather them into the set ERPs. The frequencies of these pages being clicked are summed to quantify this feature.

Epidemic related web news (ERNs). In addition to medical websites, news websites can provide health information. [Fox, 2006] reports that the most recent health news can affect 53% of health seekers' behaviors, and 7% people follow health-related news every day. Thus, many searchers may be guided by public media. An increase of public concern may bring about a sharp rise in the frequency of ERQs and ERPs, but this increase may not mean that the number of infected patients increases. Similarly with ERPs, HFMD related news stories are extracted from certain popular news websites, and the total count of these reports is the quantification of ERNs.

Given collected search logs, queries are clustered into different categories. To some extent, queries can be regarded as a kind of description for those clicked documents. These pages are also representations of the corresponding queries. We can assume that queries connected a common page have a similar topic, and the contents of these web pages can be an expression of this topic. A click-through graph is constructed from the collected search logs. A simple click-through graph is shown in Figure 1.

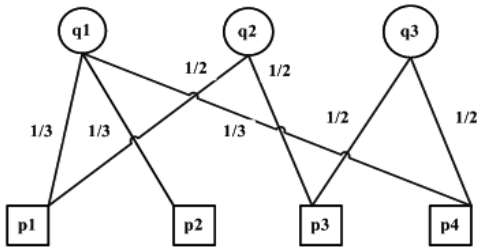

Figure 1: An example of a click-through graph

The query-clustering steps are listed as follows:

Step 1: Define the initial sets Q and P; Q stands for the ERQ set, and P represents the set of connected pages. Q= {initial ERQ}, P= Φ .

Step 2: For any query in Q, calculate the weight between this query and its close web pages. If the weight is no less than the fixed threshold (here the threshold is set to 1.0), add this page to P.

Step 3: For any page in P, calculate the weight between this page and every query connected to it. If the weight is no less than the fixed threshold, add this query to Q.

Step 4: Go to Step 2 until the count of the Q set remains unchanged or the iteration time reaches a given value.

Typically, queries and web pages of high frequency have a broad topic and link different topics together. Therefore, to avoid a bad effect on the weight calculation, we filter the high-frequency terms at the very start of our experiment. Here, we select the medical name of the HFMD disease as the initial ERQ. Thus, after the above iterative process, we obtain numerous HFMD-related queries. Some search queries such as "hay fever" may coincide with HFMD seasons but have no relation with HFMD. To make our study more accurate, these queries in the ERQs are manually filtered. Finally, the size of the ERQ set is 66 (the size is 45 in [Ginsberg *et al.*, 2009]). In addition, HFMD medical articles and news stories were gathered from medical sites and news sites, respectively. These two kinds of pages constitute the ERP and ERN sets. These features are quantified by counting their own fre-

quencies of different sets. Examples of ERQs, ERPs and ERNs are illustrated in Table 1.

Table 1: Examples of ERQs, ERPs and ERNs

Item	Type
Sore throat	ERQ
Fever	ERQ
Headache	ERQ
HFMD	ERQ
http://www.qqbaobao.com/s/shouzukoubing	ERP
http://jbk.39.net/keshi/erke/6adc9.html	ERP
http://news.hnjkw.net/hyzx/jbyw/2011/031632815.html	ERN

In our experiment, we adopt the correlation coefficient as the evaluation indicator, which is usually adopted in previous studies. The correlation coefficient between two vectors（A and B）is calculated as follows:

$$\rho_{A,B} = \frac{\mathrm{Cov}(A,B)}{\sigma_A . \sigma_B} = \frac{\sum_i (a_i - a')*(b_i - b')}{\sqrt{\sum_i (a_i - a')^2}\sqrt{\sum_i (b_i - b')^2}} \quad (1)$$

, where a' and b' represent the average values of vectors A and B, respectively. All three features and the actual HFMD occurrences are normalized by dividing their own maximum values. A comparison between the normalized ratios of these features is illustrated in Table 2.

Table 2: Coefficient-analysis over different time lags

Lag(weeks)	ERQs	ERPs	ERNs
0	0.725	0.707	0.415
1	0.763	0.765	0.509
2	0.667	0.667	0.561
3	0.658	0.650	0.606
4	0.491	0.473	0.595
5	0.251	0.243	0.583

In Table 2, the average correlation between actual HFMD occurrences and the total frequency of ERQs is 0.725. This correlation value shows that the search frequency of HFMD-related queries can generally reflect current HFMD activity. However, this value is smaller than the result (0.91) in [Ginsberg *et al.*, 2009], which is mostly ascribed to the following two aspects: 1.The methods of the ERQ set's selection are different. We automatically select our ERQs on the basis of click-through data, while they use linear regression to select the top 45 queries; 2. Our study target is HFMD, while they focus on influenza. These two diseases may differ from each other. In addition to ERQs, we also study the relationship between the number of HFMD cases and ERPs and ERNs. As described in [Heather *et al.*, 2004], the total number of people visiting HFMD-related medical pages is strongly correlated with HFMD's morbidity, and this correlation value reaches to 0.707.

The total frequencies of ERQs and ERPs have a similar tendency over time, and their own notable peaks precede the peak of actual HFMD activity. Different from ERQs and ERPs, the number of reported news stories is not completely positively correlated with occurrences. The peak of ERNs is much earlier than the peak of HFMD activity. The ERQ and ERP features obtain a best fit with a one-week preceding lag, while the coefficient between the ERNs and actual HFMD occurrences reaches the maximum value with a three-week lag. These lag values will provide an important basis for the following predictive models.

From the medical point of view, the incubation period of HFMD is commonly a week, that is, the transitional period from showing early slight symptoms to getting sick is a week. This fact may provide a reasonable explanation for the lags of ERQs and ERPs. If a user shows slight symptoms, that user may submit one or more ERQs, will visit medical treatment pages (ERPs) and will be confirmed as a case after a week. For ERQs, ERPs and ERNs, we think their frequencies are approximately linear with the actual occurrences. With the time lags taken into consideration, a simple log-odd linear model (adopted in [Ginsberg *et al.*, 2009]) is established as follows:

$$\log it(occ_t) = \partial_1 * \log it(ERQs_{t-1}) + \partial_2 * \log it(ERPs_{t-1}) + \partial_3 * \log it(ERQs_{t-3}) + \varepsilon \quad (2)$$

, where $\partial_1, \partial_2, \partial_3$ are multiplicative coefficients, ε is the error term, and logit(x) is the natural log of x/(1-x).

4 Experiments and Analysis

Let us consider the two following circumstances when an epidemic breaks: 1.A user shows slight symptoms and submits ERQs for medical treatments; 2.Another user without symptoms also submits ERQs in response to the recent epidemic news. ERQs cannot distinguish these two situations from each other. The number of ERPs will increase in the first situation, while the second situation will merely lead to increasing ERNs. We think these problems will be better solved by introducing both ERPs and ERNs.

To study their roles in HFMD prediction, these features are introduced one at a time in our experiment. A simple multivariable logistical linear regression [Christensen, 1997] is selected. We collect search logs for two HFMD seasons, and we establish different linear models for the data of 2009 (the training set). The final model is validated on untrained data of 2010. To make full use of the existing data, we adopt the following learning algorithm: on the basis of the training set, we learn an initial model for the first iteration of test set. During each iteration on the test set, previous weeks are also added into training set and relearn a model. Different models are illustrated in Table 3.

In Equation 2, we first fit the relationship between the

Table 3: Results of different models on the training and test sets

Different models	Training set(2009)	Test set (2010)
1: $\text{logit}(\text{occ}_t) = 0.9297 * \text{logit}(\text{ERQs}_{t-1}) - 0.9090$	0.763	0.735
2: $\text{logit}(\text{occ}_t) = 0.9536 * \text{logit}(\text{ERQs}_{t-1}) - 0.1198$	0.746	0.729
3: $\text{logit}(\text{occ}_t) = 0.5137 * \text{logit}(\text{ERNs}_{t-3}) - 0.3408$	0.612	0.595
4: $\text{logit}(\text{occ}_t) = 0.3157 * \text{logit}(\text{ERQs}_{t-1}) + 0.6582 * \text{logit}(\text{ERPs}_{t-1}) - 0.3624$	0.784	0.813
5: $\text{logit}(\text{occ}_t) = 0.8478 * \text{logit}(\text{ERQs}_{t-1}) + 0.1616 * \text{logit}(\text{ERNs}_{t-3}) - 0.7152$	0.802	0.824
6: $\text{logit}(\text{occ}_t) = 0.2042 * \text{logit}(\text{ERQs}_{t-1}) + 0.6830 * \text{logit}(\text{ERPs}_{t-1}) - 0.0739 * \text{logit}(\text{ERNs}_{t-3}) - 0.1326$	0.836	0.891

actual HFMD occurrences and ERQs, ERPs and ERNs using the unary linear regression method. The correlation value of Model 1 is the largest for all single features, and the ERNs feature has the worst coefficient (0.595). As a single feature, the "ERPs" or "ERQs" feature can not only detect epidemic occurrences but also predict future tendency. However, the "ERNs" feature is not a good single feature. Hence, we need to combine these features using binary linear regression to modify these single models.

At the peak of HFMD outbreaks, many healthy people are likely to follow HFMD related reports to search for HFMD-related queries. These users are misidentified as patients in Model 1. In addition, some patients with the disease begin their search at a website, and these users are also not identified as patients in the single model. According to these situations, we add other selected features, ERPs and ERNs, into the single models. ERPs_{t-1} ERPs_{t-1} represents the click frequencies of epidemic-related medical pages during the time t-1, and ERNs_{t-3} is the number of news items during this period. The correlation values between empirical occurrences and these fitted results are calculated. The comparison between the baseline and the binary feature models are illustrated in Table 3. For Model 4, the correlation between the actual occurrences and the model combining both the "ERQs" and "ERPs" features reaches 0.813, and its performance is better than the models using only the "ERQs" or "ERPs" feature. This fact indicates that the "ERPs" feature may have a good assistant function of predicting future improvements over the method of only using the "ERQs" feature. Next, we place emphasis on why the "ERPs" feature promotes predictive performance. According to [Fox, 2006], approximately 27% people seek health-related information from a health-related website. Some healthy people submit HFMD related queries, but do not click any HFMD-related medical articles. They may have other needs. The "ERPs" feature only records the frequency of medical articles being clicked. If adding the ERP features into the predicted model, this situation can be removed and the prediction will become closer to actual HFMD occurrences, which is why the ERPs are introduced. Next, we will use the ternary linear regression method to continue modifying the models.

To some extent, high ERQs and ERPs do not mean large numbers of patients; they may also be caused by high ERNs. The frequencies of the "ERQs" come from both patients and non-patients. When an HFMD epidemic breaks out, epidemic-related news stories also increase, and many healthy people may follow these stories. In Model 6, the coefficient of ERNs is negative, meaning that it can partly separate news concern queries from latent patients. Model 6 is the best in all models, and we select it as our final predictive model. Figure 2 shows our predictive result at four time points through the 2010 HFMD season. A notable increase is predicted during April, and the peak is reached on May 24. Next, the occurrences begin to decline during the following period. All of these results are later validated by official CDC data.

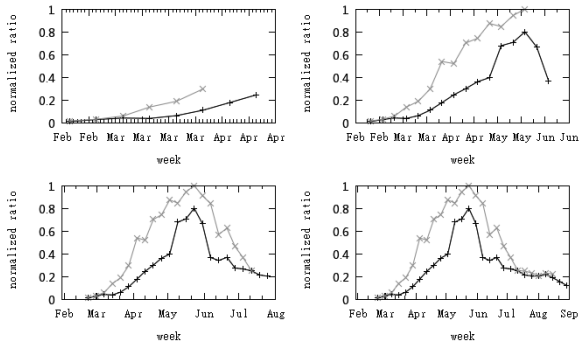

Figure 2: Predictive effect of our final model (black: predicted values, gray: actual values)

To monitor the predictive effect in a timely manner, we develop a simple HFMD monitoring system based on our predictive model. Admittedly, our system still has some drawbacks. This system is not designed to be a replacement for traditional monitoring networks. We hope our system will be useful for future epidemic researches and will enable public health officials to be well-prepare for epidemic emergencies. In the future, we will continue to improve our method and system.

5 Discussion

Some healthy people follow epidemic news by submitting epidemic related queries (ERQs) while some patients search for information from medical websites. Consequently, ERQs cannot draw a complete picture of epidemic patients. Therefore, we propose two novel features to separate the healthy users from the patients. Different from previous studies, this paper makes an attempt to predict the future numbers of epidemic cases. From the comparison between different models, we can see that the ternary-feature model has the best predictive effect, indicating that these new features play important roles in the prediction of an epidemic. However, our method has some limitations: 1.Our predictive model can only estimate the tendency of future epidemic cases (rise or drop), and concrete occurrence rates are not accurately calculated; 2.User reliability is not taken into consideration in our experiment. Hence, some noisy data may have a negative effect on the final result.

This paper develops web resource surveillance for HFMD for the first time, and it obtains a better effect on both accuracy and timeliness. This method of web mining is not limited to influenza and HFMD, and it can also be used to predict other infectious diseases. We expect that our features and methods may provide helpful information to health officials for seasonal epidemics.

6 Conclusion

Previous studies show that influenza activity can be traced from web query logs. On the basis of these earlier works, this paper develops a similar approach to HFMD prediction. In this paper, two novel features, ERPs and ERNs, are introduced to separate the healthy from the latent patients. Furthermore, we conduct a log-based study, and we obtain HFMD-related queries, pages and news. To validate the effect of these features, we conduct a systematic comparison between these features, and the experimental result indicates that both the ERPs and ERNs play key roles in predicting epidemic activities. Finally, we develop an online prediction system, and the monitored results show that our model is fairly effective at predicting future epidemic tendency. We hope that our predictive method will be helpful for future epidemic research and will help provide people with earlier alerts to give them enough time to prepare for an imminent epidemic outbreak by taking certain measures, such as obtaining vaccines.

References

[Baeza-Yates and Tiberi, 2007] Ricardo Baeza-Yates, and Alessandro Tiberi. Extracting semantic relations from query logs. *In proceeding of the 13th international conference on Knowledge discovery and data mining*, pages 76-85, San Jose, California, August 12-15, 2007.

[Beeferman and Berger, 2000] Doug Beeferman, and Adam Berger. Agglomerative clustering of a search engine query log. *In proceeding of sixth international conference on Knowledge discovery and data mining*, pages 407-416, Boston, MA, USA, August 20-23, 2000.

[Chan *et al*., 2004] Wing Shun Chan, Wai Ting Leung, and Dik Lun Lee. Clustering search engine query log containing noisy clickthroughs. *In proceeding of 2004 International Symposium on Applications and the Internet*, pages 305-308, Tokyo, Japan, January 26-30, 2004.

[Christensen, 1997] Ronald Christensen. *Log-Linear Models and Logistic Regression*, Second Edition, Spring-Velag, 1997.

[Dong and Sun, 2008] Zhaohua Dong, and Ping Sun. Clinical analysis of 325 children with hand-foot-mouth disease in 2005 and 2007. *Journal of Clinical Pediatrics*, pages. 470-472, 2008.

[Fox, 2006] Susannah Fox. Online Health Search 2006. *Pew Internet and American Life Project,* 2006.

[Ginsberg *et al.*, 2009] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S.Smolinski and Larry Brilliant. Detecting influenza epidemic using search engine query data. *Nature,* 457: 1012-1014, 2009.

[Heather *et al.*, 2004] Heather A. Johnson, Michael M. Wagner, William R. Hogan, Wendy Chapman, Robert T Olszewski, John Dowling and Gary Barnas. Analysis of web access logs for surveillance of influenza. *Stu Health Technol Inform*, 107:1202-1207, 2004.

[Hulth *et al.*, 2009] Anette Hulth, Gustaf Rydevik, and Annika Linde. Web queries as a source for syndromic surveillance. *pLos ONE 4(2):* e4378, 2009.

[Pelat *et al.*, 2009] Camille Palat, Clement Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Vallernon. More diseases tracked by using Google trends. *Emerging infectious diseases,* 15(8): 1327-1328, 2009.

[Philip *et al.*, 2008] Philip M. Polgreen, Yiling Chen, David M. Pennock, Forrest D. Nelson, and Robert A. Weinstein. Using internet searches for influenza surveillance. *Clinical Infectious Diseases 47(11)*: 1443-1448, 2008.

[Wang *et al.,* 2009] Ruiping Wang, Yali Chun, and Yiling Wu. Predicting hand-foot-mouth disease condition in Songjiang District of Shanghai City based on seasonal trend model. *China Preventive Medicine*, 10(11): 1025-1028, 2009.

[Wen *et al.*,2001] Ji-Rong Wen, Jian-Yun Nie, and-Hong-Jiang Zhang. Clustering user queries of a search engine. *In proceeding of the 10th international World Wide Web conference*, pages 162-168, Hongkong, May 1-5, 2001.

[White and Horvitz, 2009] Ryen White, and Eric Horvitz. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems*, 23(4): 770-812, November, 2009.