Detecting Epidemic Tendency by Mining Search Logs[1]

Weize Kong
School of Software
Beihang University
Beijing, China, 100086
kongweize@gmail.com

Yiqun Liu, Shaoping Ma, Liyun Ru
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
DCST, Tsinghua University, Beijing, China, 100084
yiqunliu@tsinghua.edu.cn

ABSTRACT
We consider the problem of detecting epidemic tendency by mining search logs. We propose an algorithm based on click-through information to select epidemic related queries/terms. We adopt linear regression to model epidemic occurrences and frequencies of epidemic related terms (ERTs) in search logs. The results show our algorithm is effective in finding ERTs which obtain a high correlation value with epidemic occurrences. We also find the proposed method performs better when combining different ERTs than using single ERT.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Miscellaneous

General Terms: Algorithms, Measurement, Experimentation

Keywords: Epidemic detection, search log mining, query

1. INTRODUCTION
A large and increasing number of people are using Web search engine to seek information today. As a consequence, search log collects massive amount of data which records people's search behavior. This provides us an alternative way to discover patterns of group behaviors or topics of people's interests. An interesting application is Google Trends, with which people can see how often one topic have been searched on Google over time. In another wok, Ginsberg et al. [1] used Google search queries to track influenza-like illness (ILI). Their method of selecting ILI-related query is based on testing each of 50 million most common search queries to see how their model performs when using the single query. We consider the same question that whether user search behavior is effective in detecting epidemic tendency. In particular, we examine the correlation between frequencies of ERTs and epidemic occurrences in a specific location. However, differently from [1], we utilize click-though information to find epidemic related queries (ERQs). The results show the ERQs selected by our algorithm can effectively model epidemic occurrences.

2. FINDING EPIDEMIC RELATED QUERIES
Our method of selecting ERQs is based on click-through information. We assume that queries which share a common clicked URL may be related to a same topic. Therefore, we can select more ERQs by finding other queries which share same clicked URLs with the original ERQ.

WWW 2010, April 20–24, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

To represent the query and URL relationship, we construct Click-through Graph, which was described as directed bipartite graph in [2]. We define Click-through Graph as G = (*V*, *E*); the set of vertices *V* is composed of queries and Web pages which appear in search log. The set of edges *E* is composed of edges connecting a query and a Web page, with which the query leads to the selection of the Web page. An example of a Click-through Graph is showed in Figure 1.

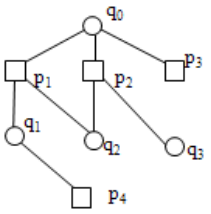

Figure 1: An example of Click-through Graph.

In Figure 1, query q_0 and q_2 are both connected to page p_2. It means that users click on p_2 while searching query q_0 and q_2. Here we assume that q_0 and q_2 may share a same topic because users click a same page p_2 while questioning them. Particularly, if q_0 or q_2 is related to one specific epidemic, it is likely that the other query is also related to the epidemic.

Based on Click-through Graph, our iterative algorithm for finding ERQ is:

1. ERQ set Q= {initial ERQs}, epidemic related page set $P = \emptyset$
2. For any query q in Q, add new pages connected to q into P
3. For any page p in P, add new queries connected to p into Q
4. Go to step 2 until exit condition is satisfied.

The resulting ERQs are in set Q. In our experiment, we pick the epidemic name as initial epidemic related query. The exit condition can be set as that the size of Q or the iterative times reach given thresholds. In our task, we end the iteration at the second round, when we harvest 771 unique ERQs.

3. DATE SET
To obtain epidemic occurrence data, we extract occurrence information for four kinds of epidemics in Beijing from the official website of Beijing Centers for Diseases Control and Prevention. The result is showed in Figure 2. Some data is missing due to absence of reports from the official website. Since the epidemic occurrence data is relatively sufficient for varicella

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

and hand-foot-mouth disease (HFMD) from Feb. 2009 to May 2009, we consider these two kinds of epidemics and analyze search query log from a widely-used commercial Chinese search engine. Specifically, we used query log data from Feb. 23, 2009 to May 31, 2009 (98 days) in Beijing. Only query and user clicks, no other information related to user privacy, were used in our work.

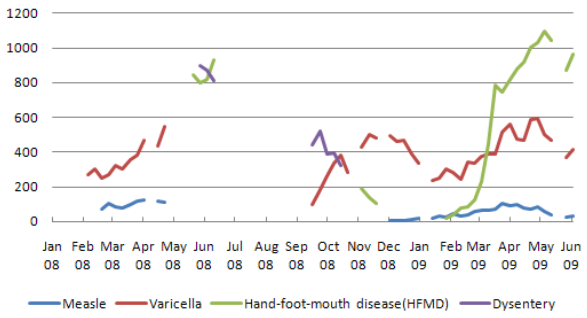

Figure 2: Four kinds of epidemic occurrences in Beijing.

4. EXPERIMENT AND RESULTS

We performed our Click-through Graph based algorithm described in section 2, and harvested 771 unique ERQs consisted by 1301 unique terms. Terms with high frequencies in these ERQs are supposed to be highly related to the epidemic. So we picked these terms as ERTs (We also filtered some terms are intuitively not related to epidemic, such as stop words). For example, in varicella-related queries, the frequently occurring terms are *varicella* with 2831 occurrences, *herpes zoster* with 287 occurrences and *infant* with 230 occurrences.

In our data set, we computed a time series of weekly frequencies for ERTs and epidemic occurrences. Both the frequencies for ERT and epidemic occurrences were normalized by dividing their maximum counts, respectively. By performing linear regression between frequencies of epidemic occurrence and single ERT, we found the selected ERTs showed a high correlation with epidemic occurrences. Tabel 1 gives us four varicella related terms with top correlation values for varicella occurrence counts. We can see a strong correlation between varicella related terms and varicella occurrences. The term *varicella*, the epidemic name, obtained the highest correlation value 0.77. The term which describes one symptom of vermicelli, *fever*, also has a close correlation value 0.73.

Table 1. Four varicella related terms with top correlation values for varicella occurrence counts

Term	Correlation Value
Varicella	0.7720
Fever	0.7343
Herpes zoster	0.5935
Scab	0.5087

Figure 3 shows the fitted linear regression line when only using frequencies of term *varicella* as independent varible. Comparing the fitted line and the line of varicella occurrences, we can see the trendence of varicella can be mirrored by frequencies of ERT *varicella*, the epedimic name.

To combine different ERT frequency data, we conducted multiple linear regression [3], using ERTs *varicella*, *fever*, *herpes zoster*, *scab* as independent variables. The result in Figure 4 shows that the fitted line of multiple linear regression fits varicella tendency more accurately than the fitted line in Figure 3. The correlation value improves from 0.77 (the highest correlation value when using single ERT counts) to 0.91. This can be explained by that the ERTs describe different aspects of the epidemic. *Varicella*, *herpes zoster* and *scab* are different symptoms of varicella. It's likely that users might adpot different ERTs, considering their own situation, when searching information about the epidemic.

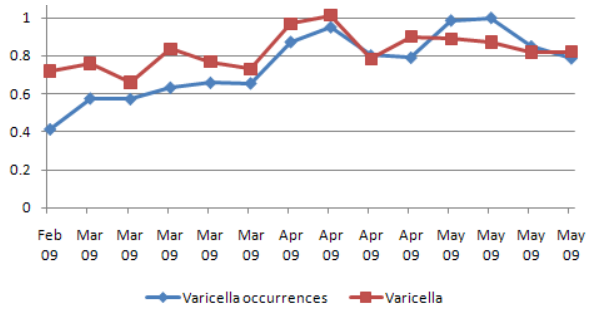

Figure 3: Line of varicella occurrences and fitted linear regression line, with frequencies of ERT *varicella* as independent variable.

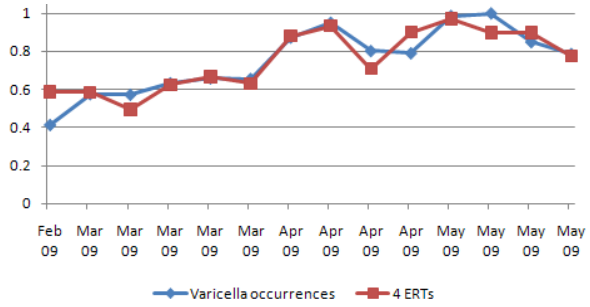

Figure 4: Line of varicella occurrences and fitted linear regression line, with frequencies of 4 ERTs *varicella*, *fever*, *herpes zoster*, *scab* as independent variables.

5. CONCLUSIONS

In this paper, we use linear regression to model frequencies of ERTs and epidemic occurrences. We proposed a Click-through Graph based algorithm to select ERQs/ERTs. The results show that the selected ERTs are effective in detecting epidemic tendency and the performance is improved when combining different ERT information in multiple linear regression model.

6. REFERENCES

[1] Ginsberg J, Mohebbi M. H, Patel R. S, Brammer L, Smolinski M. S, et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457: 1012–1014.

[2] Yi, J. and Maghoul, F. 2009. Query clustering using click-through graph. In Proceedings of the 18th international Conference on World Wide Web (Madrid, Spain, April 20 - 24, 2009). WWW '09. ACM, New York, NY, 1055-1056.

[3] Weisberg, S., Applied Linear Regression, 3rd Ed. published by Wiley/Interscience in 2005 (ISBN 0-471-66379-4).