Discovering Seasonal Patterns of Smoking Behavior Using Online Search Information

Zhu Zhang[1], Xiaolong Zheng[1,2], Daniel Dajun Zeng[1,3], Kainan Cui[4,1], Chuan Luo[1], Saike He[1], Scott Leischow[5]
[1]The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]Dongguan Research Institute of CASIA,
Cloud Computing Center, Chinese Academy of Sciences, Dongguan, China
[3]Department of Management Information Systems,
The University of Arizona, Tucson, USA
[4]The School of Electronic and Information Engineering,
Xi'an Jiaotong University, China
[5]Mayo Clinic in Arizona, Phoenix, USA
{zhu.zhang, xiaolong.zheng, dajun.zeng, chuan.luo, saike.he}@ia.ac.cn, kainan.cui@live.cn, Leischow.Scott@mayo.edu

***Abstract*—Discovering temporal patterns and changes in tobacco use has important practical implications in tobacco control. This paper presents one of the first comprehensive international studies of seasonal smoking patterns based on online searches performed. Using periodogram and cross-correlation, we find that smoking-related search behavior shows strong seasonality effect across countries. In addition, there are significant pairwise associations between such seasonality in different countries.**

Keywords—behavior informatics; search behavior; smoking; seasonality

I. INTRODUCTION

Tobacco is one of the main behavioral risk factors leading to a substantially large number of potentially preventable deaths worldwide [1]. About the deaths of 5 million people are caused by direct tobacco use across the world each year, and many of these deaths occur prematurely [2]. Given its importance, tobacco control is a global public health priority.

One important aspect of smoking behavior analysis is concerned with discovering related temporal patterns. Findings on seasonality of smoking may have major implications for the timing of tobacco control measures and interventions [3]. There have been some studies about seasonality of smoking behavior [3-7]. These studies show that seasonality exists in cigarette sales [3], onset of youth smoking [7], the initiation of smoking among adolescents [5], and sales of nicotine replacement therapies [4], using U.S. survey data. Such data are costly to obtain, often lack timeliness, and might not be representative of the ground truth in the populations [8, 9]. In addition, to the best of our knowledge, no comprehensive studies have been conducted at the international scale.

In this paper, we make use of Web search queries to fill in this important gap in tobacco control research. Web search queries have been used to monitor infectious and noninfectious diseases such as influenza epidemics [10-12] and other diseases [13, 14]. Tobacco control researchers have used Internet search queries to explore issues such as tracking the popularity of electronic cigarettes and monitoring tax avoidance and smoking cession after cigarette tax increases [15, 16].

Our study is based on publically-available data provided through Google Trends [17]. We collected search volume data for query term "smoking" from five English-speaking countries and for query term "吸烟" (i.e., the Chinese translation of smoking) from China. Searching for the term "smoking" and "吸烟" means that the user has the need for smoking information. Based on the previous studies [10-15], we assume that the trend of smoking behavior in real life can be reflected by the trend of the corresponding searching behavior on the Internet. Then we applied the periodogram method to detect the seasonality, and calculated the pairwise cross-correlations among these countries. Empirical findings indicate that the seasonality effect of smoking-related search behavior is quite strong in these six countries and there are significant pairwise associations between the seasonality of smoking-related search behavior in different countries.

II. DATA SET

The search query trend data are downloaded from Google Trends, a public service provided by Google Inc., which allows Internet users to examine trends of certain search terms by time, geographic location, and category. We used "smoking" as the search query term to collect the search query trend data about smoking behavior in five main English-speaking countries (i.e., USA, UK, Canada, Australia, and New Zealand), and used "吸烟" as the query term for the trend data in China. The query term "smoking" is chosen as it is general to the area of smoking behavior. We compared the search volumes of "smoking", "cigarette smoking", "tobacco smoking", "cigarette" and "tobacco", then we found that "smoking" was the most commonly searched term. We chose the query term "吸烟" in the same way. The category of the above search volumes is limited to health, because such category can exclude search queries irrelevant with the act of smoking tobacco or other substances. The time interval covered by our research is from January 4, 2004 to December 29, 2012. Note that the query index on Google Trends is not the raw search volume, but rather the total search volume for a search query term in a

geographic region divided by the total search volume of that region at a given time.

III. METHODOLOGY

The time series plots of the raw trend data are shown in Fig. 1. The horizontal axis is the time line with one week as the unit time. The periodogram, ideal pass filter and cross-correlation in the time series tools of MATLAB [18] were applied to analyze the trend data.

Periodogram (also known as spectral plot) is applied to determine whether the seasonality of smoking exists in the Google Trends data. In this paper, the periodogram is computed as the scaled absolute value of the square of fast Fourier transform of the search trend data, which is the unbiased estimate of the power spectral density of a set of search trend data. The corresponding frequency vector is calculated in cycles per week and has the same length as the power vector. The periodogram is scaled in order that the mean equals the variance of the periodogram. A peak in the periodogram indicates there is a seasonal component near the value corresponding to the peak. To extract a seasonal component from the raw trends data, ideal pass filter is applied for the periodogram to filter the raw trend data, with the frequency interval covering the peak shape in the periodogram.

To examine the associations between the trends in different countries, we calculated the pairwise cross-correlations of the raw search trend data in the mentioned six countries. Cross-correlation is used to measure the degree of the linear relationship between two search trends with limited time lags (fifty weeks in this study). A specific lag corresponding to a high correlation of two trends data may indicate that these two trends have certain same influencing factors and a time delay.

IV. RESULTS

Fig. 2 shows the periodograms of the search trend data. The salient peak near 0.0192 on the horizontal line in the periodograms indicates that there is an important seasonal component in each set of the six search trend data and the period is about one year (i.e., 52 weeks ≈ 1/0.0192 cycles/week). Another salient peak near 2.137×10^{-3} on the horizontal line indicates the seasonal component whose period is the whole time interval (i.e., 468 weeks ≈ $1/2.137\times10^{-3}$ cycles/week) in the six search trend data. In addition, the peak near 0.04 indicates the seasonal component with half year as the period.

As illustrated in Fig. 3, the seasonal components with one year as their period are extracted from the raw smoking-related trend data by using ideal pass filter, when we select the frequency range from 0.015 to 0.025 in the periodograms. This frequency range is selected to only cover each peak shape near 0.0192 but not cover other peak shapes in Fig. 2. For Fig. 1 and 3, we can observe that the relative search volume of the countries on the northern hemisphere (i.e., the U.S., Canada, UK, and China) has an approximate sinusoidal pattern with the peak nearby January and the trough nearby August. The relative search volume of the countries on the southern hemisphere (i.e., Australia and New Zealand) also has an approximate sinusoidal pattern with the peak nearby June and the trough nearby December.

In Table I, the values in each cell contain two parts: a correlation coefficient and a time lag in the parenthesis (week as unit time). These lags are selected to provide the highest correlation for each pair. From Table I, we can observe that the lengths of the time lags corresponding to the correlation

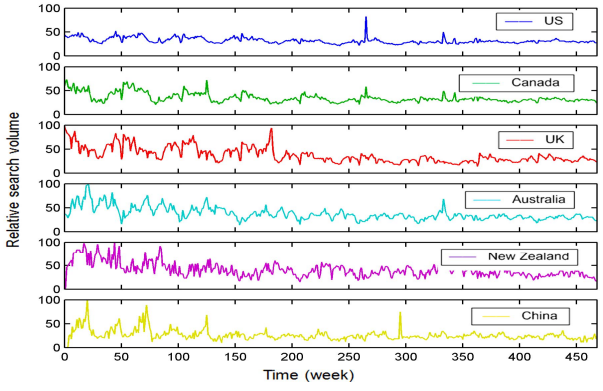

Fig. 1. Search trends on smoking in USA, UK, Canada, Australia, New Zealand and China from Google Trends, from January 4, 2004 to December 29, 2012.

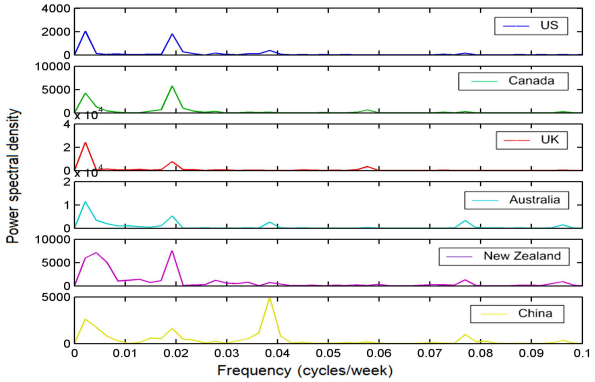

Fig. 2. The periodograms of the search trends data of USA, UK, Canada, Australia, New Zealand and China.

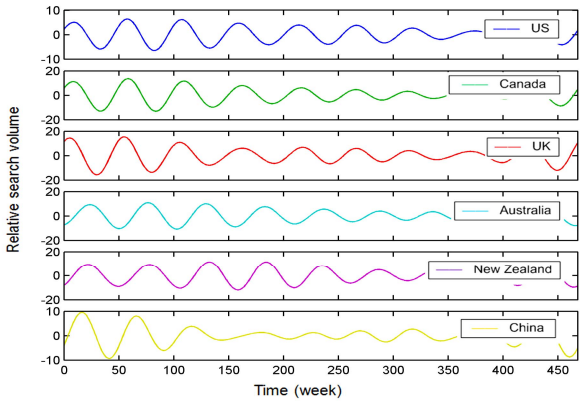

Fig. 3. The seasonal components of the search trends data of USA, UK, Canada, Australia, New Zealand and China.

TABLE I. CROSS-CORRELATIONS OF THE SEASONAL COMPONENTS OF THE SMOKING-RELATED SEARCH TRENDS

Countries	Countries					
	US	***UK***	***Canada***	***Australia***	***New Zealand***	***China***
US	1.000	0.920(2)	0.956(-1)	0.844(-19)	0.783(-21)	0.780(-8)
UK	--	1.000	0.956(2)	0.889(-20)	0.796(-20)	0.859(-8)
Canada	--	--	1.000	0.908(-18)	0.828(-19)	0.895(-10)
Australia	--	--	--	1.000	0.960(0)	0.769(-8)
New Zealand	--	--	--	--	1.000	0.670(-8)
China	--	--	--	--	--	1.000

coefficients on the same hemisphere are larger than on the different hemisphere except for China. The reason for the observation may be that the similar weather condition (especially temperature) on the same hemisphere leads to the similar smoking-related search behavior on the same hemisphere, because some studies show that the weather condition (especially temperature) may be a factor contributing to the seasonal variation of smoking behavior [3, 6]. As shown in Table I, except for the correlation between New Zealand and China, all the correlations of the seasonal components of the smoking-related search trend data are above 0.769, then we can infer that seasonal components in the smoking-related search trend data of these English-speaking countries and China are similar each other and the seasonality of smoking behavior in these countries are also similar each other.

V. DISCUSSION AND CONCLUSION

The findings of this paper have the following limitations. First, the study focuses on English and Chinese search terms only and covers six countries only. Second, the factors that attribute to the seasonality of the search trends about smoking are not investigated in detail. Our ongoing research is addressing these limitations, and we plan to explore the collective intelligence in the collective search behavior using social computing [19-22] approaches and new data analysis methods [23-24].

The findings indicate that the seasonality of smoking-related search behavior widely exists in many countries. The findings can help tobacco community with the timing of tobacco control effort and provide a novel Internet-based evidence for the seasonality of smoking behavior.

ACKNOWLEDGMENT

Reported research is partially funded through NNSFC Grants #71025001, #71103180, #91124001, #70890084 and #91024030, MOH Grants #2013ZX10004218 and #2012ZX10004801.

REFERENCES

[1] WHO, WHO global report: mortality attributable to tobacco. Geneva: World Health Organization, 2012.
[2] C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," PLoS Med, vol. 3, pp. e442, 2006.
[3] S. Chandra and F. J. Chaloupka, "Seasonality in cigarette sales: patterns and implications for tobacco control," Tob Control, vol. 12, pp. 105-107, 2003.
[4] S. Chandra, J. G. Gitchell, and S. Shiffman, "Seasonality in sales of nicotine replacement therapies: patterns and implications for tobacco control," Nicotine Tob Res, vol. 13, pp. 395-398, 2011.
[5] B. Colwell, N. Ramirez, L. Koehly, S. Stevens, D. W. Smith, and S. Creekmur, "Seasonal variations in the initiation of smoking among adolescents," Nicotine Tob Res, vol. 8, pp. 239-243, 2006.
[6] D. Momperousse, C. D. Delnevo, and M. J. Lewis, "Exploring the seasonality of cigarette-smoking behaviour," Tob Control, vol. 16, pp. 69-70, 2007
[7] R. J. Wellman and J. R. DiFranza, Seasonality in onset of youth smoking parallels seasonality in cigarette sales: Tob Control. 2003 Sep;12(3):339.
[8] J. W. Ayers, B. M. Althouse, J.-P. Allem, D. E. Ford, K. M. Ribisl, and J. E. Cohen, "A Novel Evaluation of World No Tobacco Day in Latin America," Journal of Medical Internet Research, vol. 14, 2012.
[9] M. Boland, M. Sweeney, E. Scallan, M. Harrington, and A. Staines, "Emerging advantages and drawbacks of telephone surveying in public health research in Ireland and the U.K," BMC Public Health, vol. 6, pp. 208, 2006.
[10] G. Eysenbach, "Infodemiology: tracking flu-related searches on the web for syndromic surveillance," AMIA Annu Symp Proc, pp. 244-8, 2006.
[11] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," Nature, vol. 457, pp. 1012-1014, 2009.
[12] Y. Luo, D. Zeng, Z. Cao, X. Zheng, Y. Wang, Q. Wang, and H. Zhao, "Using multi-source web data for epidemic surveillance: A case study of the 2009 Influenza A (H1N1) pandemic in Beijing," presented at Service Operations and Logistics and Informatics (SOLI), 2010 IEEE International Conference on, 2010.
[13] A. C. Yang, N. E. Huang, C.-K. Peng, and S.-J. Tsai, "Do Seasons Have an Influence on the Incidence of Depression? The Use of an Internet Search Engine Query Data as a Proxy of Human Affect," PloS one, vol. 5, pp. e13728, 2010.
[14] K. Cui, Z. Cao, X. Zheng, D. Zeng, K. Zeng, and M. Zheng, "A Geospatial Analysis on the Potential Value of News Comments in Infectious Disease Surveillance," in PAISI'11, 2011, pp. 85-93.
[15] J. W. Ayers, K. Ribisl, and J. S. Brownstein, "Using Search Query Surveillance to Monitor Tax Avoidance and Smoking Cessation following the United States' 2009 "SCHIP" Cigarette Tax Increase," PloS one, vol. 6, 2011.
[16] J. W. Ayers, K. M. Ribisl, and J. S. Brownstein, "Tracking the rise in popularity of electronic nicotine delivery systems (electronic cigarettes) using search query surveillance," American journal of preventive medicine, vol. 40, pp. 448-453, 2011.
[17] Google, "Google Trends", http://www.google.com/trends/, retrived March 4, 2013.
[18] MathWorks, "MATLAB 7.11.0 (R2010b)", Natick, Massachusetts.
[19] X. Li, W. Mao, D. Zeng, and F.-Y. Wang, "Agent-based social simulation and modeling in social computing," in Intelligence and Security Informatics: Springer Berlin Heidelberg, 2008, pp. 401-412.
[20] D. Zeng, H. Chen, R. Lusch, and S.-H. Li, "Social media analytics and intelligence," Intelligent Systems, IEEE, vol. 25, pp. 13-16, 2010.
[21] D. Zeng, F.-Y. Wang, and K. M. Carley, "Guest Editors' Introduction: Social Computing," Intelligent Systems, IEEE, vol. 22, pp. 20-22, 2007.
[22] W. Fei-Yue, K. M. Carley, Z. Daniel, and M. Wenji, "Social Computing: From Social Informatics to Social Intelligence," Intelligent Systems, IEEE, vol. 22, pp. 79-83, 2007.
[23] W. Chang, D. Zeng, and H. Chen, "Prospective spatio-temporal data analysis for security informatics," presented at Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE, 2005.
[24] H.-M. Lu, D. Zeng, and H. Chen, "Prospective infectious disease outbreak detection using Markov switching models," Knowledge and Data Engineering, IEEE Transactions on, vol. 22, pp. 565-577, 2010.