Relevance prediction between domestic dust and
by the search data in China (Baidu)

2015.05.31

휴먼ICT융합학과
정희석
지도교수 : 조준동 교수님
Contents

I. Introduction

II. Prior Research, Research Questions, and Hypothesis

III. Research Design

IV. Results

V. Conclusion
I. Introduction

Background (1)

- An increase of yellow sand/dust (PM10) and accrual dates
  - Increase the number of the average in the 1980s 3.9 days in Seoul, In the 1990s, 7.7 days, 12.4 days since 2000 (source: Korea Meteorological Administration)

- Features of yellow sand: Domestic intrusion on usually 2 to 3 days, as early as within 12 hours from China
  - There are various prediction algorithm but occurrence is a difference based on the research variables
    - Lack of relevance of days of occurrence considering regional

Moving path and PM10 epicenter
I. Introduction (Cont.)

Background (2)

- Epidemic prediction of Google of the US regional influenza
  - Read the search pattern, related search on cold, confirm the overall location information
  - Predict the fast-forward fashion 1-2 weeks than Center for Disease Control

- Since 2001, Baidu maintain the top position in China portal site
  - In the case of the NAVER in Korea
  - Account for more than 70% of market share in China’s online search market
  - Facilitate the collection of search data
    - Amount of data buzz
Prior Research

- **Lee Geunghui** (Korean Statistical Society, 2014)
  - Economic prediction using the search data
  - Internet users find important information by entering a keyword in the search portal, call the search data that organize an overview of the search query language
  - It is possible to grasp the change of the search comparatively
  - Keywords centered, so have a small error than atypical text analysis in SNS

- **Eun-Hee Lee** (Asian Dust Research Laboratory, National Institute of Meteorological Research, 2011)
  - Performance Analysis of Simulation of Asian Dust Observed in 2010 by the all-Season Dust Forecasting Model, UM-ADAM2

- **Ki-Kwang Lee** (dankook Univ, 2012)
  - The Fluctuation Patterns of Conjunctivitis Cases Caused by Asian Dust Storm (ADS): Focused on the ADS Density and the Accuracy of ADS Forecast

- **Jae-Hwuen Jung** (SKTelecom Information Technology Institute, 2009)
  - A Study on the Factors Affecting Knowledge Contribution and Knowledge Utilization in an Online Knowledge Network

- **Choe Byeongheon** (chinastudy, 2010)
  - BlueOcean strategic research of Baidu
Research Question

- Display the changes on search volume, traffic such as fine dust, yellow sand in Baidu
  - Show the warning symptoms

- So, is there relationships that time until PM10 to reach the Korea & division of China epicenter & Baidu search volume

The purpose of this study

- The purpose of this study is to verify the proper association time & source classification. Time is that it takes PM10 arrive in Korea from each region in China

Hypothesis

Main H0: There are not relationships between predicting PM10 by the search data and existing in KMA data(relationships: time from China to Korea, regional division in China)

Sub H0

H0: There are not relationships between search volume in Baidu and PM10 increase date in Korea
H0: We will not be able to grasp the relationship between China birthplace on PM10
Ⅲ. Research Design

Data and Sample Selection

- The sample obtained from 百度指数(Baidu Index)
  - Classification: Words is related PM10 (沙尘暴, 雾霾)
    - Year, month, daily statistics
  - Sample Period: 2012 ~ 2014 (3 years)

- The sample obtained from KMA
  - Classification: PM10 observation data
    - By the beta-ray PM10 observation device (domestic 28 places)

Test Design and Measurement

- Test Using statistic Tools
  - Time Series & ARIMA model statistical method by index classification
    - Comparison of the Index date and domestic PM10 numerical slope
    - Calculate the difference between the number of days
  - Check the originated PM10 regions in China by related domestic PM10
    - Implementation of the correlation analysis
IV. Results

Data Collection

- Data Collection - Baidu
  - China area is divided into 34 regions
  - Specifically divided into 340 single regions
  - In this study, I analyze the search data of 34 pieces of regional data

Results of sample data

- It represents a change of Baidu search data and domestic PM10 by a day
- Region of Gansu Province as follows:

example) 2014년 11월 4일
### Data Collection

#### Results of sample data

Creating a sample result table as follows (reference domestic PM10 200 or more) – Scheduled to add partial

<table>
<thead>
<tr>
<th>China areas (34 places)</th>
<th>Correlation coefficient</th>
<th>Date difference(about)</th>
<th>The difference between the existing measurements</th>
<th>The difference between the existing source</th>
<th>Comparison of the accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>北京(베이징)</td>
<td>0.58</td>
<td>2일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>甘肃(간쑤)</td>
<td>0.65</td>
<td>1일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>河北(허베이)</td>
<td>0.53</td>
<td>2일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>湖北(후베이)</td>
<td>0.56</td>
<td>2일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>吉林(길림)</td>
<td>0.71</td>
<td>0일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>辽宁(요녕성)</td>
<td>0.70</td>
<td>0일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>青海(칭하이)</td>
<td>0.60</td>
<td>1일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>台湾(대만)</td>
<td>0.62</td>
<td>0일</td>
<td>-</td>
<td>Bias doubt</td>
<td>-</td>
</tr>
<tr>
<td>西藏(티베트)</td>
<td>0.59</td>
<td>2일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>香港(홍콩)</td>
<td>0.69</td>
<td>1일</td>
<td>-</td>
<td>Demand F.R</td>
<td>-</td>
</tr>
<tr>
<td>新疆(신장)</td>
<td>0.72</td>
<td>3일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>云南(운남)</td>
<td>0.52</td>
<td>2일</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
IV. Conclusion

Pros & Cons

✓ Pros
  ✓ May want to understand the relevance of the new China region of the main sources
  ✓ It is able to reveal the fact that search frequency (the amount of buzz) or traffic of one country is likely to affect other countries (Korea)
  ✓ It is able to quickly predict than KMA through the major regions
  ✓ May confer validity that PM10 increasing phenomenon is spread in China inland

✓ Cons
  ✓ There are various environment variables such as wind direction changes in the Chinese, it is necessary to study comprehensively the several variables in the future

Future Work

✓ Implementation of the continuous analysis of whether the result is not satisfied
✓ Since improve the validity of predictive analysis like pattern. So share the ideas for the information to develop
Thank you