EPIDEMIC OUTBREAK AND SPREAD DETECTION SYSTEM (EOSDS) BASED ON TWITTER DATA

Xiang Ji, James Geller
New Jersey Institute of Technology, Newark, NJ, USA

Soon Ae Chun
CUNY College of Staten Island, Staten Island, NY, USA
OVERVIEW

Introduction
Research challenges and approaches
Overview of EOSDS System
Methods of EOSDS System
  Data Collection and Preprocessing
  Static Map
  Distribution Map and Two-Step Coding
  Filter Map
Results and Discussion
Related Work
Conclusions and Future Work
Introduction

Research challenges and approaches

Overview of EOSDS System

Methods of EOSDS System

- Data Collection and Preprocessing
- Static Map
- Distribution Map and Two-Step Coding
- Filter Map

Results and Discussion

Related Work

Conclusions and Future Work
INTRODUCTION

Epidemics are a major threat for humanity

1918 Flu Pandemic (Spanish Influenza)
(Killed 50-100 million. 1918-1920)

SARS
(Killed 962. 2003)

Swine Flu
(Killed 18,400. 2009)

Listeria
(Killed 30. 2011)
Traditional way of monitoring epidemics

CDC manually collected data from sentinel medical practices.

Often 2 weeks delay between a patient is diagnosed and a report is published

Search Engine was used to address the problem

[Ginsberg et al. (2009)]

Limitation 1: Search engine query terms not available to outsiders

Limitation 2: Users’ locations are not explicitly recorded.
INTRODUCTION - 3

Twitter, a micro-blogging Service, shows the potential to overcome the above limitations.

Advantage 1: Up-to-date event streams

Advantage 2: Most of the tweets are public.

Advantage 3: Nearly all related information, including users’ locations can be retrieved through APIs.
Introduction

Research challenges and approaches

Overview of EOSDS System

Methods of EOSDS System

  Data Collection and Preprocessing
  Static Map
  Distribution Map and Two-Step Coding
  Filter Map

Results and Discussion

Related Work

Conclusions and Future Work
Research Challenges

Challenge 1: Tweets are unstructured. Tweets and related information need to be extracted and preprocessed.

Challenge 2: Location information, which is embedded in user profile, (1) is noisy (40% are meaningless) and (2) Spatial granularity is inconsistent.

Challenge 3: In order to monitor the outbreak and spread of epidemics, spatial and temporal information must be used by different visualization methods.
Our Approaches

Approach 1: Keyword-Based Extraction
  • Five disease keywords are specified
  • Tweets are stored into database

Approach 2: Location Information Processing
  • A Text Mining method is used to identify and clean spam locations.
  • Granularity of locations is judged by searching the US National Gazetteer

Approach 3: Three visualization methods:
  • Static Map (spatial)
  • Distribution Map (spatial)
  • Filter Map (spatial and temporal)
Introduction

Research challenges and approaches

Overview of EOSDS System

Methods of EOSDS System

Data Collection and Preprocessing

Static Map

Distribution Map and Two-Step Coding

Filter Map

Results and Discussion

Related Work

Conclusions and Future Work
Internet

Twitter Streaming API

Data Collector

Geonames_Gazetteer

Location Granularity and Timestamp

Location Latitude and Longitude

Standardized Location

Reverse geocode

Dataset

EOSDS-Backend

EOSDS-User Interface

Unstructured Twitter Data

Twitter Streaming API

Store relational database tables

Cleaning

geocode

Timestamp

Location Granularity and Timestamp

Geonames_Gazetteer

Static Map

Distribution Map

Timeline Filter Map

EOSDS-Backend

EOSDS-User Interface
Introduction
Research challenges and approaches
Overview of EOSDS System
Methods of EOSDS System

Data Collection and Preprocessing
Static Map
Distribution Map and Two-Step Coding
Filter Map

Results and Discussion
Related Work
Conclusions and Future Work
Five epidemics are monitored since Sep. of 2011
Tuberculosis, listeria, influenza, swine flu, measles.

• Whenever a tweet contains one of the above disease names, it is automatically processed
• Tweet is recorded by EOSDS in the form of a relational database.
The core PHP code for collecting epidemics-related real-time tweets

```php
class Consumer extends Phirehose {
    public function enqueueStatus($status) {
        $tweet_object = json_decode($status);
        // ...
    }
}

$stream = new Consumer (STREAM_ACCOUNT, STREAM_PASSWORD, Phirehose::METHOD_FILTER);
$stream->db_connect();
$stream->setTrack(array('health'));
$stream->consume();
```
### tweets

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet_id</td>
<td>bigint</td>
</tr>
<tr>
<td>tweet_text</td>
<td>varchar</td>
</tr>
<tr>
<td>entities</td>
<td>text</td>
</tr>
<tr>
<td>created_at</td>
<td>datetime</td>
</tr>
<tr>
<td>geo_lat</td>
<td>decimal</td>
</tr>
<tr>
<td>geo_long</td>
<td>decimal</td>
</tr>
<tr>
<td>user_id</td>
<td>int</td>
</tr>
<tr>
<td>screen_name</td>
<td>char</td>
</tr>
<tr>
<td>name</td>
<td>varchar</td>
</tr>
<tr>
<td>profile_image_url</td>
<td>varchar</td>
</tr>
<tr>
<td>location</td>
<td>varchar</td>
</tr>
<tr>
<td>url</td>
<td>varchar</td>
</tr>
<tr>
<td>description</td>
<td>varchar</td>
</tr>
<tr>
<td>created_at</td>
<td>datetime</td>
</tr>
<tr>
<td>followers_count</td>
<td>int</td>
</tr>
<tr>
<td>friends_count</td>
<td>int</td>
</tr>
<tr>
<td>statuses_count</td>
<td>int</td>
</tr>
<tr>
<td>time_zone</td>
<td>varchar</td>
</tr>
<tr>
<td>last_update</td>
<td>timestamp</td>
</tr>
</tbody>
</table>

### users

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>user_id</td>
<td>bigint</td>
</tr>
<tr>
<td>screen_name</td>
<td>char</td>
</tr>
<tr>
<td>name</td>
<td>varchar</td>
</tr>
<tr>
<td>profile_image_url</td>
<td>varchar</td>
</tr>
<tr>
<td>location</td>
<td>varchar</td>
</tr>
<tr>
<td>url</td>
<td>varchar</td>
</tr>
<tr>
<td>description</td>
<td>varchar</td>
</tr>
<tr>
<td>created_at</td>
<td>datetime</td>
</tr>
<tr>
<td>followers_count</td>
<td>int</td>
</tr>
<tr>
<td>friends_count</td>
<td>int</td>
</tr>
<tr>
<td>statuses_count</td>
<td>int</td>
</tr>
</tbody>
</table>

### tweet_mentions

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet_id</td>
<td>bigint</td>
</tr>
<tr>
<td>source_user_id</td>
<td>bigint</td>
</tr>
<tr>
<td>target_user_id</td>
<td>bigint</td>
</tr>
</tbody>
</table>

### tweet_tags

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet_id</td>
<td>bigint</td>
</tr>
<tr>
<td>tag</td>
<td>varchar</td>
</tr>
</tbody>
</table>

### tweet_urls

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet_id</td>
<td>bigint</td>
</tr>
<tr>
<td>url</td>
<td>varchar</td>
</tr>
</tbody>
</table>
SQL query for integrating tweet and user information:

```sql
```
A Text Mining method is used to detect and delete Meaningless locations (e.g. in the universe, heaven)

Top five single-grams in meaningless locations

<table>
<thead>
<tr>
<th>Concept</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Gram Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>19</td>
<td>1.00</td>
<td>single</td>
</tr>
<tr>
<td>In</td>
<td>17</td>
<td>0.89</td>
<td>single</td>
</tr>
<tr>
<td>in</td>
<td>13</td>
<td>0.68</td>
<td>single</td>
</tr>
<tr>
<td>The</td>
<td>8</td>
<td>0.42</td>
<td>single</td>
</tr>
<tr>
<td>you</td>
<td>8</td>
<td>0.42</td>
<td>single</td>
</tr>
</tbody>
</table>

Stop-word detection method achieved a precision of 97.1% and a recall of 95.8%

<table>
<thead>
<tr>
<th></th>
<th>S− (not spam)</th>
<th>S+ (spam)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detect+</td>
<td>7</td>
<td>347</td>
<td>354</td>
</tr>
<tr>
<td>Detect−</td>
<td>631</td>
<td>15</td>
<td>646</td>
</tr>
<tr>
<td>Total</td>
<td>638</td>
<td>362</td>
<td>1000</td>
</tr>
</tbody>
</table>
Introduction

Research challenges and approaches

Overview of EOSDS System

Methods of EOSDS System
   Data Collection and Preprocessing
      Static Map
      Distribution Map and Two-Step Coding
      Filter Map

Results and Discussion

Related Work

Conclusions and Future Work
STATIC MAP

Provides a direct way to display locations of all tweet instances.

EOSDS geocodes the geographic information into (latitude, longitude) coordinates.

Geocoding is done by Google Map API.

Rochester -> (43 North, 77 West)
Epidemic: Listeria  Period: 09-26-2011 to 09-28-2011
Visually recognize wherever there is an unusual cluster of “markers”

Limitation 1: Not always easy to judge whether a particular area is unusual

Limitation 2: In some states, such as California, there are more tweets because there are more people.

Is there always an epidemic in these areas?
Introduction

Research challenges and approaches

Overview of EOSDS System

Methods of EOSDS System

Data Collection and Preprocessing

Static Map

Distribution Map and Two-Step Coding

Filter Map

Results and Discussion

Related Work

Conclusions and Future Work
Granularity:
The different granularities of Twitter locations make it difficult to identify what state or city a tweet comes from.

Different levels of granularity:

<table>
<thead>
<tr>
<th>Granularity</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place</td>
<td>Newark, New Jersey</td>
</tr>
<tr>
<td>State</td>
<td>Colorado</td>
</tr>
<tr>
<td>Country</td>
<td>Netherlands</td>
</tr>
<tr>
<td>World</td>
<td>Somewhere</td>
</tr>
</tbody>
</table>

How to know “Rochester”, “Rochester, USA”, “NYC” are in the State of New York?
Algorithm:

1. Geocode all the locations into latitudes and longitudes.
2. We reverse-geocode the obtained latitudes and longitudes.
3. Use the resulting standardized locations (e.g. county, state, country or state, country) to know how many tweets belong to each state.
EOSDS knows that all 5 locations belong to the state of New York.
Introduction

Research challenges and approaches

Overview of EOSDS System

Methods of EOSDS System

Data Collection and Preprocessing

Static Map

Distribution Map and Two-Step Coding

Filter Map

Results and Discussion

Related Work

Conclusions and Future Work
FILTER MAP

Provides users with a dynamic interface to monitor and analyze dynamic trends derivable from health-related tweets.

Three filters are incorporated into the filter map:

Granularity Filter: US National Places Gazetteer contains 29,000 US geographic locations, + 50 US states and 245 countries

Influence Filter: A range of follower counts may be set by an EOSDS user. Only tweeters in the range are displayed. Useful to estimate where the “seed tweet” originated.

Timeline Filter: Every tweet has a timestamp. Insights into the temporal distribution and development of an epidemic.
FILTER MAP

(a) Timeline: 3:00 am; (b) Timeline: 4:00 am; (c) Timeline: 5:00 am
Introduction
Research challenges and approaches
Overview of EOSDS System
Methods of EOSDS System
  Data Collection and Preprocessing
  Static Map
  Distribution Map and Two-Step Coding
  Filter Map
Results and Discussion
Related Work
Conclusions and Future Work
RESULTS

In September of 2011, there was a sudden outbreak of Listeria in US.

Big clusters in California, Washington, Colorado, Texas, Oklahoma, Florida, Maryland, and New York.
RESULTS -2

Distribution Map

09-26-2011 absolute

09-26-2011 relative

09-27-2011 absolute

09-27-2011 relative
RESULTS -3

CDC’s report, as of 11am EDT on September 29, 2011


- 84 persons were infected with listeria as reported by CDC.
- The states with the largest numbers of infected persons were: Colorado (17), Texas (14), New Mexico (13), Oklahoma (11), Nebraska (6), Kansas (5).
DISCUSSION

In six worst states indicated by CDC report (blue line), EOSDS result correlated well with CDC report in four states (red cycled).

There are two states (blue cycled) showing conflict between EOSDS results and CDC report, what happened?
DISCUSSION -2

- Wyoming’s frequency is relatively low in the official CDC report, but its relative frequency is the highest in the EOSDS distribution map.

- The reason is that on September 26th, a death was confirmed by the health department of Wyoming [1] but that death was not in the CDC report until October 6th [2].


• We beat the 9.2 billion-budgeted CDC by 10 days !!
Introduction
Research challenges and approaches
Overview of EOSDS System
Methods of EOSDS System
  Data Collection and Preprocessing
  Static Map
  Distribution Map and Two-Step Coding
  Filter Map
Results and Discussion
Related Work
Conclusions and Future Work
Related Work

Brownstein: [Brownstein et al. (2008)]
- Used online news-based data
- Collects reports from online news aggregators (Google News and ProMED-mail)
- Categorize the news and filter documents into “breaking news,” “warnings,” and “old news”
- Trigger alerts based on the “breaking news”

Cheng: [Cheng et al. (2010)]
- Determine users’ positions when location information is absent
- Place 51% of the Twitter users within 100 miles of their actual locations
- Rely on detecting “local” words, with a high local specificity and a fast dispersion (e.g. howdy)
Comparison with Related Work

The EOSDS system is different from the above research in terms of data sources and processing of locations.

Data sources:

- Microblogging service is used for extracting data
- More information about user is accessible in Twitter

Location Processing:

- Provide levels of granularity, because location is arbitrarily specified by the users
- Increase the location estimation by identifying higher-level locations
Introduction

Research challenges and approaches

Overview of EOSDS System

Methods of EOSDS System

Data Collection and Preprocessing

Static Map

Distribution Map and Two-Step Coding

Filter Map

Results and Discussion

Related Work

Conclusions and Future Work
CONCLUSIONS

1. EOSDS was faster than CDC in one case. The distribution map made it possible to discover an unusual listeria outbreak situation in Wyoming, which was not reported by the CDC until 10 days later.

2. Modules to clean noisy geographic locations, and to identify the levels of granularity for different location specifications
CONCLUSIONS -2

4. Visualize the Twitter data from three different perspectives.

5. All three maps displayed listeria patterns that correlated well with the CDC reports on the same topic.
FUTURE WORK

1. Extend the EOSDS to be able to detect epidemic trends happening worldwide.

2. Analyze sentiments to quantify the “degree of panic” caused by each epidemic.

3. Introduce other parameters besides the number of followers, to better judge the influence of tweeters.
THANKS!
QUESTIONS?