A Spatiotemporal Data Analysis System for Microblogs

Chung Wai Yeung, Ma Man Chung and Mak Yee Ting

Advised by Prof. Ke Yi

Introduction

Overview

Microblogs, like Twitter and Instagram, are now widely used for people to update their social statuses regularly. With GPS equipped in mobile devices, a post includes text, photos, timestamp and also location, as known as spatiotemporal data, which provides useful information for researchers to extract data and conduct analyses on the distributions of their interested phenomena in Hong Kong.

Our project is to develop a spatiotemporal data analysis system for microblogs. With the use of our analysis system, researchers can easily achieve the result with different visualized presentations.

Objectives

- Extract data from microblogs
- Build a database for storing collected data
- Develop a web-based graphical user interface (GUI)
- Analyze the collected data
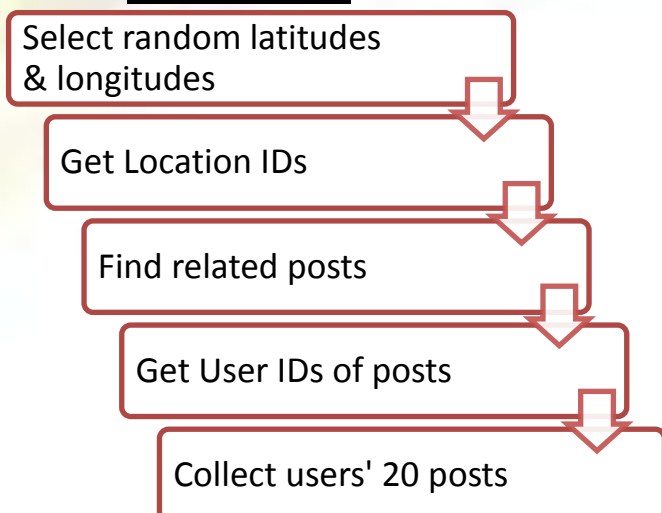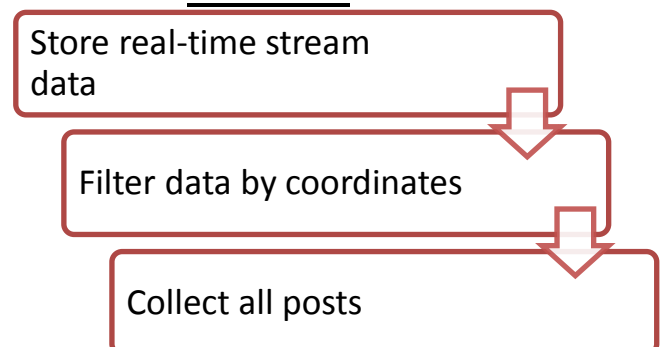- Design visualized presentation methods

Methodology

Data Collection

Twitter and Instagram are chosen as the platforms for collecting data with two different approaches.
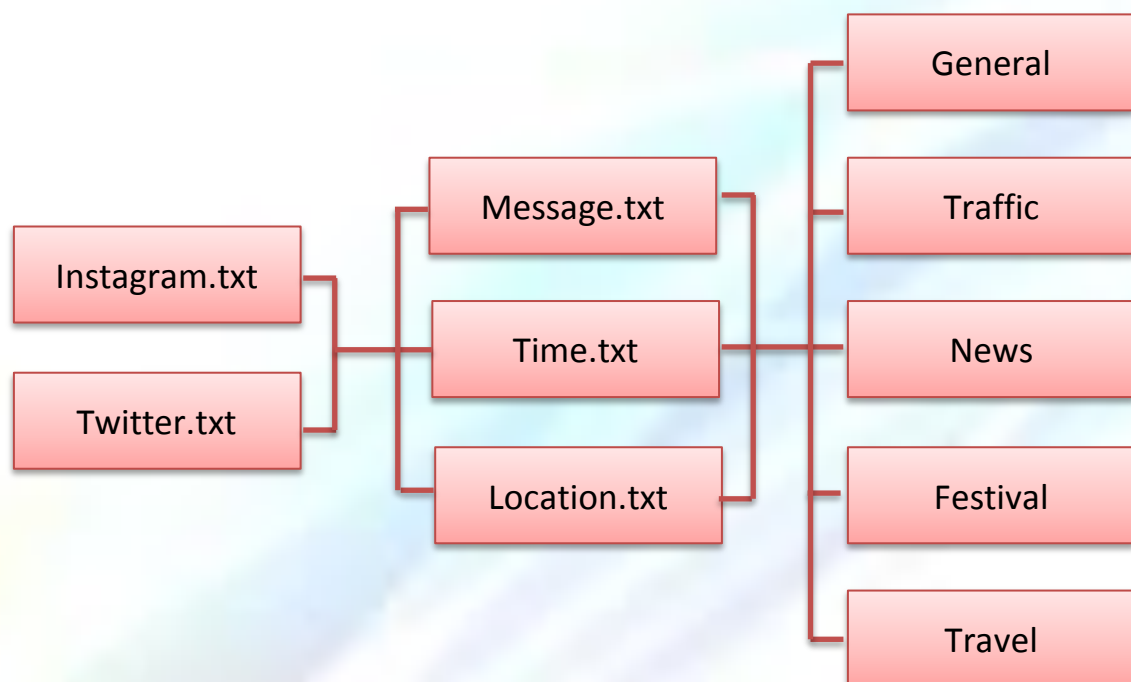
Instagram

Select random latitudes & longitudes

↓

Get Location IDs

↓

Find related posts

↓

Get User IDs of posts

↓

Collect users' 20 posts

Twitter

Store real-time stream data

↓

Filter data by coordinates

↓

Collect all posts

Data Handling and Indexing

After collecting data from Instagram and Twitter, data in both text files (i.e. Instagram.txt and Twitter.txt) are merged into three text files, with each file storing only message, time and location (i.e. message.txt, time.txt and location.txt).

Then, the three text files are converted into five data files in TREC format according to the defined categories (i.e. General, Traffic, News, Festival and Travel). Lastly, these data files perform indexing by a powerful data analysis tool, Lemur.

```
Instagram.txt ─┐        ┌─ Message.txt ─┐       ┌─ General
               ├────────┼─ Time.txt ────┼───────┼─ Traffic
Twitter.txt ───┘        └─ Location.txt ┘       ├─ News
                                                ├─ Festival
                                                └─ Travel
```

Data Analysis

After doing indexing, Indri Query Language in Lemur is used for information retrieval. For different types of search, we can use different queries.

- Keyword search:
  A or B or C
  →#combine(A B C)

- Location/Time search:
  coordinates/timeslots inside the boundary A B
  →#between(Field A B)

Lemur

Graphical User Interface
Responsive Web Design

A responsive website is designed with Foundation Framework, which allows different end-users to access and use our application.
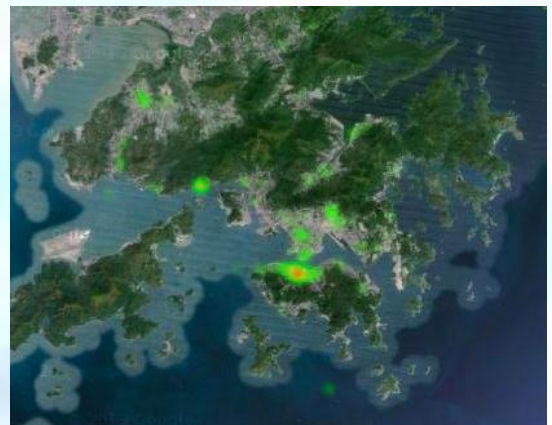

Various Presentation Methods

Different visualized presentations are used to show the relationship between the spatiotemporal data (location and time) and message.

Map Presentation

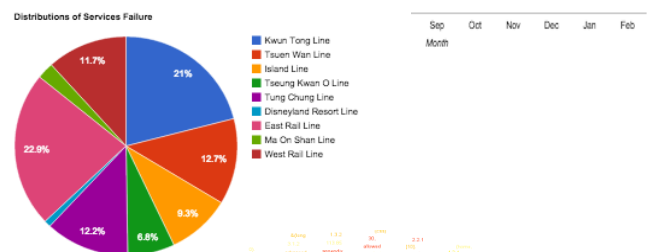We use maps to show the spatial distribution of microblogs posted in different time periods.
- Point Map
- Heat Map

Chart Presentation

We use charts to show the statistics of data in a visualized way.
- Bar Chart
- Pie Chart
- Line Chart

Word Cloud Presentation

We use word clouds to show the most frequent words within an area and a time period.