Jurnal
Teknologi

Full Paper

TWEET DATA EXTRACTOR FOR CREATING A TWITTER TRAFFIC MAP MASHUP

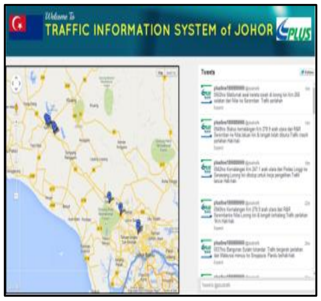
Amirul Afif Jasmi[a], Mohamad Hafis Izran Ishak[b], Nurul Hawani Idris[a]*

[a]Geoinformation Department, Faculty of Geoinformation and Real Estate, Universiti Teknologi Malaysia, Malaysia
[b]Faculty Electrical Engineering, Universiti Teknologi Malaysia, Malaysia

Article history
Received
6 April 2015
Received in revised form
12 August 2015
Accepted
23 August 2015

*Corresponding author
hawani@utm.my

Graphical abstract



Abstract

Over recent years, there has been a growth of interest in the use of social media including Facebook and Twitter by the authorities to share and updates current information to the general public. The technology has been used for a variety of purposes including traffic control and transportation planning. There is a concern that the use of new technologies, including social media will lead to data abundance that requires effective operational resources to interpret the big data. This paper proposes a tweet data extractor to extract the traffic tweet by the authority and visualise the reports and mash up on top of online map, namely Twitter map. Visualisation of traffic tweet on a map could assist a user to effectively interpret the text based Twitter report by a location based map viewer. Hence, it could ease the process of planning itinerary by the road users.

Keywords: Twitter, map mashup, transportation planning, web data extractor, big data

Abstrak

Sejak kebelakangan ini, media sosial telah banyak digunakan oleh pihak berwajib untuk berkongsi maklumat terkini dengan pengguna awam. Antaranya di dalam operasi pengawalan dan perancangan perjalanan trafik. Walaubagaimanapun, kecanggihan teknologi baru dalam pengurusan trafik ini memberi kesan kepada penghasilan data-data yang terlampau banyak sehingga memerlukan pengurusan operasi yang efektif untuk menguruskan dan menterjemahkan data kepada informasi yang lebih berguna untuk pengguna akhir. Artikel ini mencadangkan pembangunan pengekstrak data Twitter, iaitu di dalam konteks penyebaran maklumat trafik, untuk mengekstrak tweet berkaitan laporan trafik untuk divisualisasikan di atas peta Twitter. Visualisasi 'tweet berkaitan laporan trafik di atas peta boleh memudahkan pengguna untuk menterjemahkan laporan berasaskan teks yang di kongsi menggunakan Twitter di atas pemapar berasaskan lokasi. Seterusnya memudahkan pengguna membuat perancangan perjalanan.

Kata kunci: Twitter, peta mashup, perancangan pengangkutan, ekstrak web data, data besar

1.0 INTRODUCTION

Traffic congestion has become a common issue, particularly at the urban areas and during festive seasons. Traffic congestion are varies from time to time. It negatively affect the duration of journey which could be longer in time, decelerate and increase the number of vehicle [1]. An effective road traffic planning and managing is crucial to help avoiding frequent traffic problems on the road.

Nowadays social media has been used as a communication technology to share traffic information to the general public. Web 2.0 applications such as Twitter and Facebook have been used to share traffic updates by the authority such as the Projek Lebuhraya Usahasama Berhad (PLUS) and Lembaga Lebuh Raya Malaysia (LLM).

The phenomenal growth of Web 2.0, cloud computing and cyber-infrastructure and the advancement of low cost mobile computing with built-in geo-positioning technologies has led to the emergence of 'neogeography'. The term is coined by Turner [2] as a 'new geography' and consists of a set of techniques and tools that fall outside the realm of traditional GIS. Conventionally, a professional geo-literate users might use ESRI based products, talk of local coordinates projections, and to assess suitable area that high risk to landslides events. In contrast, a neogeographers, would uses free mapping APIs such as Google Maps, talks about KML format, and geotags their photos to share with friends using social media such as Twitter, Instagram and Facebook. In a simple word, neogeography platform allows users more flexibility in creating and sharing the location based information. This platform has been used to collect traffic updates from the road users such as through Waze mobile application.

The availability of free mapping APIs such as Google Map and OpenStreetMap provides users low cost solutions in developing map applications. Twitter map could use to share information and to gather data and information from various sources. For example, in earthquakes tragedy occurred in Haiti in 2010, many volunteers have been recruited by a non-profit organisation namely, Usahidi.org to share the messages of the occurrence of the tragedy and to set up operation as the respond to the tragedy [3].

There are still little applications that integrate 'tweets' submitted via Twitter and viewed it on online map. But Twitter has been used widely to share traffic updates such as Live Traffic Sydney (https://twitter.com/LiveTrafficSyd), KL Traffic Updates (https://twitter.com/ kltrafficupdate), Honolulu Traffic Updates (http://livewire.kitv. com/ Event/ LIVE_Oahu_Traffic_Updates).

The abundance of data related to road traffic that mostly available in timely version has led to difficulty in managing and visualising the big data. Big data demands big machine including the storage and speed [4]. Nevertheless, big data requires effective methods to manage, interpret, visualise and disseminate the information to the end users. Big data leads to data abundance that requires operational resources that able to interpret the data into meaningful information. Otherwise, the data could become wasted and expensive resources [5]. Several studies have demonstrated ways to extract the unstructured data on the web such as [6].

The traffic updates reported via Twitter platform are presented on textual based information; hence unable to assist users to generate instant view of the reported location. The end users are not able to interpret the location reported in a tweet, thinking spatially and making decision through a medium of space. According to [7] spatializing non spatial data or using space to view conceptualise problems are one effective cognitive strategy to make decision; spatializing non spatial data and expressed graphically could aid in analysis and comprehension; graphic presentation is necessary when the public users' access to the data increasing.

This paper proposed a tweet data extractor to produce Twitter map that able to extract the tweet timeline that reporting traffic events shared by expressway operator. The rest of the paper is organised as follows. Section 2 provides an overview of the proposed Twitter traffic map with a discussion on the components in the application framework. Section 3 describes the workflow of the proposed tweet data extractor. Section 4 presents an overview of the implementation of prototype application where PLUS Twitter traffic is used as a case study. Section 5 discusses the benefits and limitations of the proposed workflow and concludes remarks.

2.0 A PROPOSED ARCHITECTURE OF TWITTER TRAFFIC MAP

In this section, the architecture of Twitter Traffic Map is proposed. Figure 1 depicts the application framework in general. There are five components in this architectures – 'traffic data' from Twitter account, reference table of kilometre expressway section, tweet data extractor, base map from Google Map, and Twitter Map online application that mashed up tweet traffic data on top of base map from Google Map.

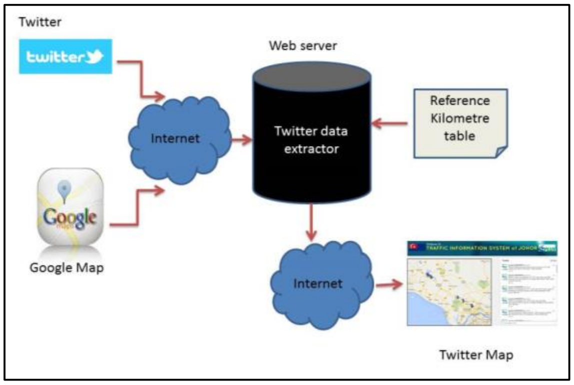
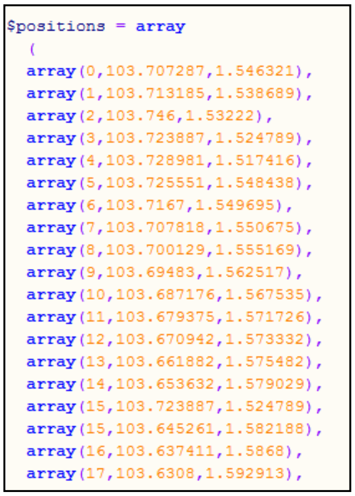


Figure 1 The proposed architecture for twitter traffic map

2.1 A Reference Table

The purpose to create a reference table of expressway section (KM) data was to match the location of KM expressway section (in points) with data and locations extracted from the traffic Twitter account, for example the PLUS and LLM. In this study, the reference data was in ESRI Shapefile format and contained attributes of each location point supplied by the PLUS. The attribute data include kilometre, name, region, point_x (longitude), point_y (latitude) and section. However, the created reference table contains only a few columns including kilometre (KM) id, point_x (longitude) and point_y (latitude). The kilometre data were integer data type. While the point_x and point_y is a float data type that represents the latitude and longitude of a point. For an example, the 8th kilometre of Southern region of PLUS expressway is refer as 8 km with latitude and longitude of 103.700356 and 1.555392.

The reference table was then stored as arrays in *.php file. The purpose was to match these data with data extracted from the Twitter account. Figure 2 shows a snapshot of arrays that stored reference data.

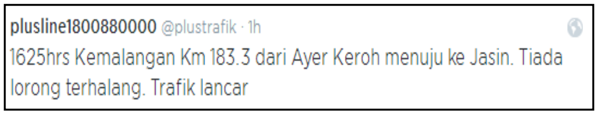


```
$positions = array
  (
  array(0,103.707287,1.546321),
  array(1,103.713185,1.538689),
  array(2,103.746,1.53222),
  array(3,103.723887,1.524789),
  array(4,103.728981,1.517416),
  array(5,103.725551,1.548438),
  array(6,103.7167,1.549695),
  array(7,103.707818,1.550675),
  array(8,103.700129,1.555169),
  array(9,103.69483,1.562517),
  array(10,103.687176,1.567535),
  array(11,103.679375,1.571726),
  array(12,103.670942,1.573332),
  array(13,103.661882,1.575482),
  array(14,103.653632,1.579029),
  array(15,103.723887,1.524789),
  array(15,103.645261,1.582188),
  array(16,103.637411,1.5868),
  array(17,103.6308,1.592913),
```

Figure 2 A snapshot of reference data that stored as arrays

2.2 Tweet traffic data from Twitter™ account

The tweet was real-time data which extracted from the Twitter account. In this study, the account used was the PLUS trafik (@*plustrafik*). These data were in textual form known as tweet and dynamically changed. To match the reference data (stored in arrays) with extracted traffic data (from Twitter™), the common structures that used by the expressway crews to report traffic events have to be identified. From this study, the organisation does not have a common format in composing a tweet. There were several structures of tweet reported on the Twitter timeline. Generally, the tweet structure of traffic report starts with time (24-hours format), type of report, kilometre's expressway section id, and name of place and report description. This study chose to use this tweet structure to be extracted. Figure 3 below shows a sample of tweet structure.



plusline1800880000 @plustrafik · 1h

1625hrs Kemalangan Km 183.3 dari Ayer Keroh menuju ke Jasin. Tiada lorong terhalang. Trafik lancar

(Source: https://Twitter.com/plustrafik)

Figure 3 Sample of tweet data by @*plustrafik* account

2.3 Tweet Data Extractor

Tweet data extractor was developed as a tool to extract tweet that successfully composed and shared in the Twitter timeline by the authority. The purpose of tweet data extractor was to extract the raw tweet data from the traffic Twitter account and transform it into useful information.

This tool extracts the raw tweet data, stores the data as arrays, and filters the data according to the required 'tweet timeline' patterns, currency and locations (area). This tool then matches the filtered tweet with the reference table before mash up the data on top of online map application.

2.4 Twitter Account

Twitter is medium of online social media which allow their user to read and send their message or so-called 'tweets' of 140 characters or less [8]. Twitter users are able to share information on Twitter via web interface, short message service (SMS) and via tablet or smartphone. This study used Twitter account from PLUS trafik as a source of traffic data. The Twitter account, however is not limited to this organisation, but could use other Twitter accounts that disseminate traffic information. However, the tweet patterns of traffic data may vary and depends on the users who write the tweets. This study used Twitter API version 1.1 to extract the tweet traffic data.

2.5 Google Map APIs

Google Map is a free commercial map released by Google for general public. The end users able to mash up their data on top Google Map via free mapping Google Map APIs since 2005 [9]. This study used Map APIs version 3.0 to mash up the extracted tweet data on top of Google Map.

2.6 Twitter Traffic Map

A Twitter map is a web based application that integrates the traffic information shared via traffic Twitter account with the online map from the Google Map. This application enable end user to visualise the tweet data shared in a series of textual based information on a map based format. This map view could enhance end user to make decision by spatially think of the specific location reported via Twitter.

3.0 THE WORKFLOW OF TWEET DATA EXTRACTOR

This section describes the workflow of the proposed tweet data extractor to create Twitter map. Figure 4 shows the diagram of the workflow from the stage of accessing all the tweet of the traffic data until the stage of visualising the tweet on top of online map application.

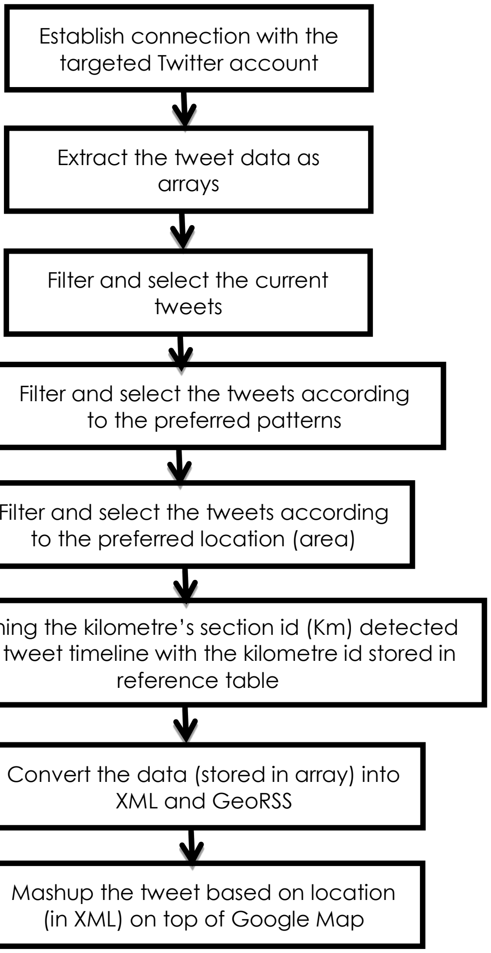
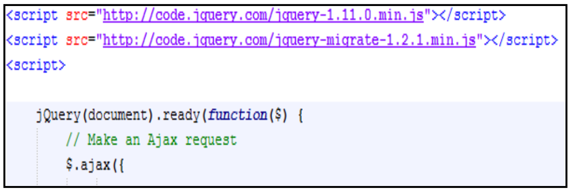

Establish connection with the targeted Twitter account
Extract the tweet data as arrays
Filter and select the current tweets
Filter and select the tweets according to the preferred patterns
Filter and select the tweets according to the preferred location (area)
Matching the kilometre's section id (Km) detected in the tweet timeline with the kilometre id stored in reference table
Convert the data (stored in array) into XML and GeoRSS
Mashup the tweet based on location (in XML) on top of Google Map

Figure 4 The traffic twitter map application workflow

3.1 Establish Connection to Access Tweet Data

JavaScript for Twitter API version 1.1 was used to extract the tweet data. The current Twitter™ APIs requires a user to make an 'OAuth connection' which is to establish connection before able to extract 'timeline' from a Twitter™ account. Figure 5 shows a snapshot of jQuery function in timeline_request.html file to request the 'tweet timeline' from the traffic Twitter account.



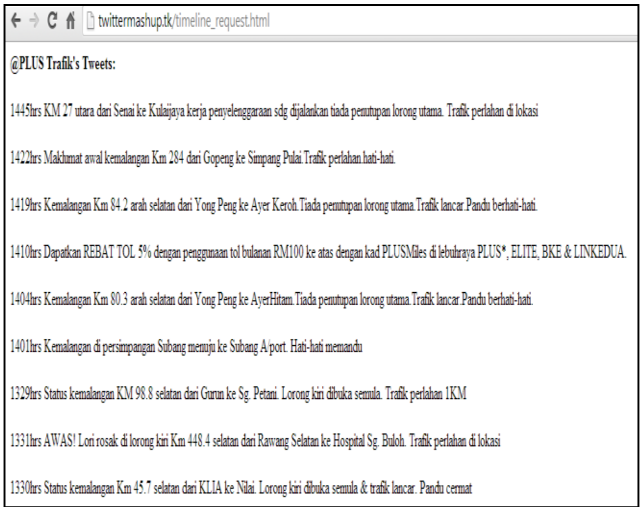
```
<script src="http://code.jquery.com/jquery-1.11.0.min.js"></script>
<script src="http://code.jquery.com/jquery-migrate-1.2.1.min.js"></script>
<script>

    jQuery(document).ready(function($) {
        // Make an Ajax request
        $.ajax({
```

Figure 5 The jQuery function

3.2 Extract Tweet Data as Arrays

After the 'tweet timelines (reports)' are obtained from the traffic Twitter™ account, it automatically extracts the 'tweets' and assembled the data into a single HTML string and displays the results in textual form. Figure 6 below presents a series of tweet that has been filtered via tweet data extractor. The tweet data is based on real-time data extracted from the tweet published via Twitter account where the data dynamically changes.



@PLUS Trafik's Tweets:

1445hrs KM 27 utara dari Senai ke Kulaijaya kerja penyelenggaraan sdg dijalankan tiada penutupan lorong utama. Trafik perlahan di lokasi

1422hrs Makhumat awal kemalangan Km 284 dari Gopeng ke Simpang Pulai.Trafik perlahan.hati-hati.

1419hrs Kemalangan Km 84.2 arah selatan dari Yong Peng ke Ayer Keroh.Tiada penutupan lorong utama.Trafik lancar.Pandu berhati-hati.

1410hrs Dapatkan REBAT TOL 5% dengan penggunaan tol bulanan RM100 ke atas dengan kad PLUSMiles di lebuhraya PLUS*, ELITE, BKE & LINKEDUA.

1404hrs Kemalangan Km 80.3 arah selatan dari Yong Peng ke AyerHitam.Tiada penutupan lorong utama.Trafik lancar.Pandu berhati-hati.

1401hrs Kemalangan di persimpangan Subang menuju ke Subang A/port. Hati-hati memandu

1329hrs Status kemalangan KM 98.8 selatan dari Gurun ke Sg. Petani. Lorong kiri dibuka semula. Trafik perlahan 1KM

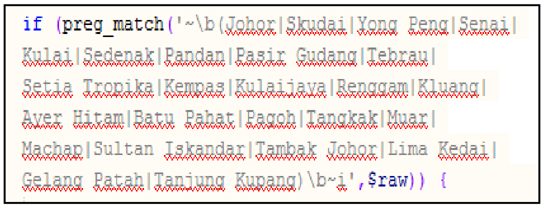
1331hrs AWAS! Lori rosak di lorong kiri Km 448.4 selatan dari Rawang Selatan ke Hospital Sg. Buloh. Trafik perlahan di lokasi

1330hrs Status kemalangan Km 45.7 selatan dari KLIA ke Nilai. Lorong kiri dibuka semula & trafik lancar. Pandu cermat

Figure 6 A series of tweet data that has been extracted via tweet data extractor

3.3 Data Filtration

The next workflow was data filtration. Filtration of data was conducted to allow the selection of specific tweet records was based on pre-defined criteria. First, filtration was conducted on the raw tweet data to select only the current 100 tweets. The common structure of 'tweet time lines' (report) was also examined. The big challenge in this stage was the variety of 'tweet time lines' structures used by the expressway crews to share the traffic information. Some of the 'tweets' were composed to share latest promotion and public awareness campaign. Therefore, the first level of filtration was conducted to eliminate the 'tweets' that related with unnecessary information. The second level of filtration was filtering the location (area) composed on the 'tweets'. The scope of this study was to visualise the traffic tweet data that covering the expressway sections in Johor, therefore the locations which were stated outside of Johor were eliminated. The common place names along the expressway were identified to produce one reference list of locations along the north-south expressway of Johor such as Skudai, Senai and Machap. The locations were stored in the *process.php* file to use a reference to filter the tweet by location. Figure 7 shows the keywords of location name in the source code.



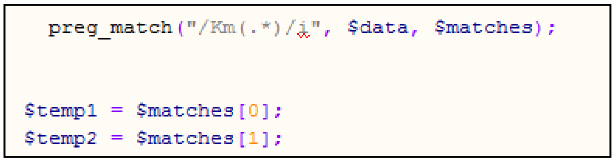
```
if (preg_match('~\b(Johor|Skudai|Yong Peng|Senai|
Kulai|Sedenak|Pandan|Pasir Gudang|Tebrau|
Setia Tropika|Kempas|Kulaijaya|Renggam|Kluang|
Ayer Hitam|Batu Pahat|Pagoh|Tangkak|Muar|
Machap|Sultan Iskandar|Tambak Johor|Lima Kedai|
Gelang Patah|Tanjung Kupang)\b~i',$raw)) {
```

Figure 7 The list of locations in *process.php* file

3.4 Data Matching

After filtering process, the tweet data were matched with the reference kilometre section table. Automated data matching was conducted to match the location of tweet data, according to the stated kilometre section id to the coordinate of the section id stored in the reference table. Then the matched data will be map as a point feature. Due to the inaccurate location used by the default geotag function in Twitter application, this study extracted kilometre data from composed tweet instead of extracting the location of geotag data embedded with tweet.

First, the tool will scan the keyword that stated kilometre in the tweet report such as 'Km' and 'KM'. Figure 8 shows the line of source code to detect and stored the kilometre id stated in a tweet.



```
  preg_match("/Km(.*)/i", $data, $matches);


$temp1 = $matches[0];
$temp2 = $matches[1];
```

Figure 8 The code to detect and store the kilometre id

Second, the stored kilometre's section id data from the tweet will be matched with the reference table. As the reference table was stored as array, the tool will read the kilometre data line by line until the value is matched. Once the kilometre id from both sources matched, it will automatically read the latitude and longitude data of that particular location that stored in the reference table. The matched data were stored in Extensible Markup Language (XML) and GeoRSS data format.

3.5 Mashup Traffic Tweet Data on Google Map

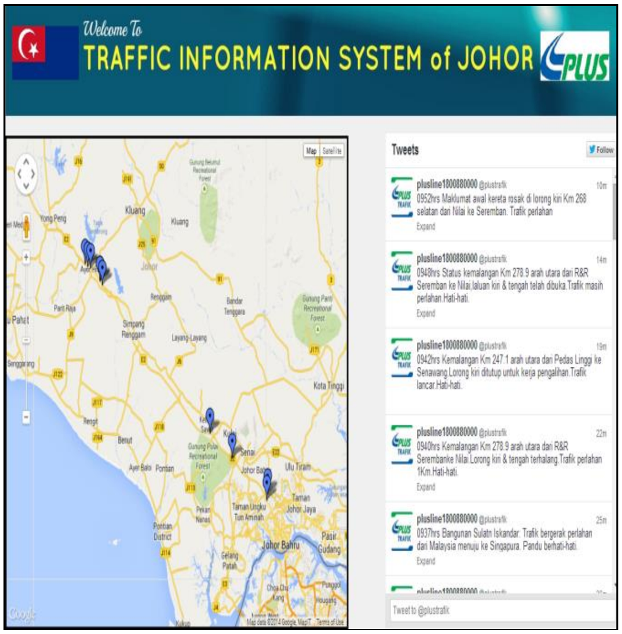
The last process was mapping the matched tweet reports on top of a map. Google Map was chose as base map to map the tweet traffic events according to the coordinate values of kilometre section id. Google Map APIs version 3 were used integrate the tweet data.

4.0 THE PROTOTYPE IMPLEMENTATION

According to the architecture and workflow proposed in the previous sections, a prototype Twitter map application known as the Traffic Information System of Johor was developed. The application integrates traffic reported via PLUS Twitter account (@*plustrafik*) with online map. The application able to visualise the near-real time traffic reported via Twitter on top of online map.

4.1 The Main Interface

Figure 9 below shows the user interface for the prototype application. The interface is divided into two partitions which is the right and left side. The right side of the interface is to embed the list of current 'tweet' from @*plustrafik* account. The list of tweet timelines will be updated if there is a new tweet composed via the Twitter account. At the bottom of the list, there is a form for other Twitter users to comment or giving feed back to the @*plustrafik* account regarding the traffic report's update.



plusline1800880000 @plustrafik 10m
0952hrs Maklumat awal kereta rosak di lorong kiri Km 268 selatan dari Nilai ke Seremban. Trafik perlahan

plusline1800880000 @plustrafik 14m
0948hrs Status kemalangan Km 278.9 arah utara dari R&R Seremban ke Nilai,laluan kiri & tengah telah dibuka.Trafik masih perlahan.Hati-hati.

plusline1800880000 @plustrafik 19m
0942hrs Kemalangan Km 247.1 arah utara dari Pedas Linggi ke Senawang.Lorong kiri ditutup untuk kerja pengalihan.Trafik lancar.Hati-hati.

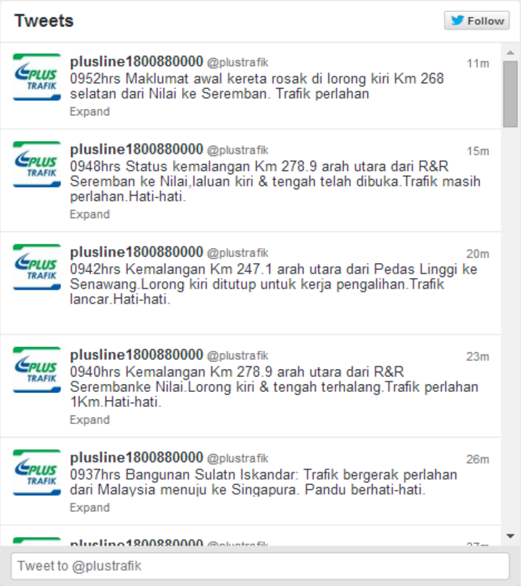
plusline1800880000 @plustrafik 22m
0940hrs Kemalangan Km 278.9 arah utara dari R&R Serembanke Nilai.Lorong kiri & tengah terhalang.Trafik perlahan 1Km.Hati-hati.

plusline1800880000 @plustrafik 25m
0937hrs Bangunan Sulatn Iskandar: Trafik bergerak perlahan dari Malaysia menuju ke Singapura. Pandu berhati-hati.

Tweet to @plustrafik

Figure 9 The Twitter Map of Traffic Information System of Johor

The right-side of the web contains the main element which is a map that displays the tweets on top of the medium. Figure 10 shows the Twitter timelines that contain a series of tweets.



Tweets

plusline1800880000 @plustrafik 11m
0952hrs Maklumat awal kereta rosak di lorong kiri Km 268 selatan dari Nilai ke Seremban. Trafik perlahan
Expand

plusline1800880000 @plustrafik 15m
0948hrs Status kemalangan Km 278.9 arah utara dari R&R Seremban ke Nilai,laluan kiri & tengah telah dibuka.Trafik masih perlahan.Hati-hati.
Expand

plusline1800880000 @plustrafik 20m
0942hrs Kemalangan Km 247.1 arah utara dari Pedas Linggi ke Senawang.Lorong kiri ditutup untuk kerja pengalihan.Trafik lancar.Hati-hati.

plusline1800880000 @plustrafik 23m
0940hrs Kemalangan Km 278.9 arah utara dari R&R Serembanke Nilai.Lorong kiri & tengah terhalang.Trafik perlahan 1Km.Hati-hati.
Expand

plusline1800880000 @plustrafik 26m
0937hrs Bangunan Sulatn Iskandar: Trafik bergerak perlahan dari Malaysia menuju ke Singapura. Pandu berhati-hati.
Expand

plusline1800880000 @plustrafik 27m

Tweet to @plustrafik

Figure 10 Twitter timelines contain a series of tweets

4.2 Twitter Map

Google Map was used as a base map to display the traffic information from Twitter @*plustrafik* account. The matched tweet data stored as GeoRSS that contains latitude and longitude of a tweet location will be mapped. A point which represents a tweet will be displayed along the highway using blue-coloured marker. Figure 11 depicts the real time tweet extracted from @*plustrafik* Twitter account.

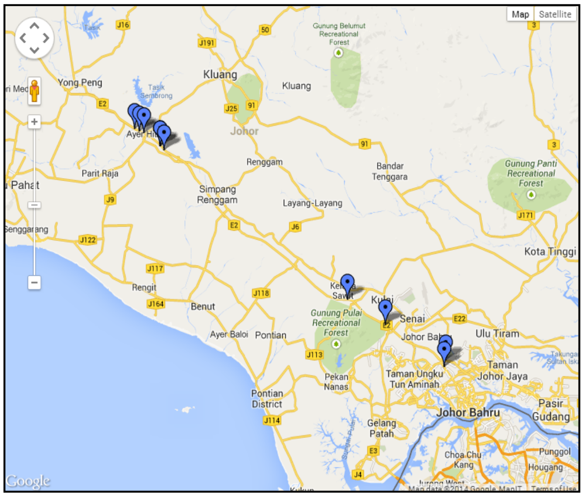
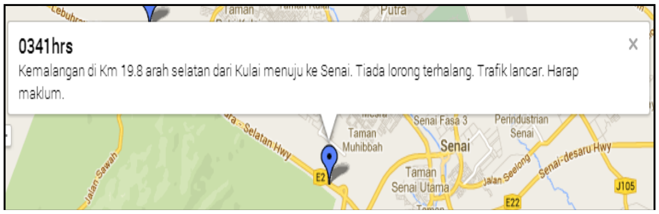


Figure 11 The snapshot of prototype Twitter Traffic map

End users could click on a blue-coloured marker on the map to get the details of tweet report including the type of information, kilometre's section id, location and event's description. The information window will be popped-out as in Figure 12 below.



0341hrs
Kemalangan di Km 19.8 arah selatan dari Kulai menuju ke Senai. Tiada lorong terhalang. Trafik lancar. Harap maklum.

Figure 12 A snapshot of information window that represent the tweet timeline on Twitter map

5.0 CONCLUSION

This study proposed the use of tweet data extractor to create Twitter Traffic Map. There are five components involves in the framework to create Twitter Traffic Map– Twitter account and APIs, Google Map APIs, tweet data extractor, a reference table (i.e. kilometre's section id), and a Twitter map application. The proposed tweet data extractor able to learn the common composes patterns of 'tweet time lines' and match with the reference table to map the traffic location. The limitation on this method is that it requires the keywords and text structures to be pre-defined before the extractor tool able to detect the matching patterns. Twitter traffic map allows end users to spatialise the traffic tweet supplied by the authorities, hence assist them to think spatially the traffic location that has been reported. Further research could examine other 'tweet timelines' patterns used by the traffic authority to compose traffic reports, enhance the Twitter map viewers that able support layers management, allow performing basic queries etc. The proposed data extractor could use by other authorities to enhance traffic reports dissemination through Twitter platform by presenting on Twitter Map.

Acknowledgement

The authors would like to acknowledge the University Grant (GUP) vot 08J94 (2013) from the Ministry of Education Malaysia and the Universiti Teknologi Malaysia. We were grateful to the PLUS Malaysia Berhad for supplying the relevant data to support this study.

References

[1] Sofia, A. S. S. D., Nithyaa, R. and Arulraj, P. G. 2013. Minimizing the Traffic Congestion Using GIS. *International Journal of Research in Engineering & Advanced Technology*. 1(1): 1-6.
[2] Turner, A. 2006. *Introduction to Neogeography*. O'Reilly Media
[3] Goolsby, R. 2010. Social Media as Crisis Platform: The Future of Community Maps/Crisis Maps. *ACM Transactions on Intelligent Systems and Technology*. 1(1): 1-11.
[4] Sui, D. Goodchild, M., Elwood, S. 2013. Chapter 1: Volunteered Geographic Information, the Exaflood, and the Growing Digital Divide. In Sui, D. *et al*. (eds). *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Springer. 1-12.

[5] Hamilton, A., Waterson, B., Cherret, T., Robinson, A. and Snell, I. 2012. The Evolution of Urban Traffic Control: changing policy and technology. *Transportation Planning and Technology*. 36(1).

[6] Idris, N. H, Jackson, M. J, Said, M. N, Ishak, M. I. I, Hashim, G. H, Ismail, Z. 2014. *Semi-Automated Metadata Detection for Assessing The Credibility of Map Mashups*. The XXV FIG Congress. 16-21 June 2014, Kuala Lumpur.

[7] Bednarz, R. S. and Bednarz, S.W. 2008. Chapter 16: The Importance of Spatial Thinking in an Uncertain World. In Sui, D. Z. (ed). *Geospatial Technologies and Homeland Security*. Springer: 315-330.

[8] Deller, R. 2011. Twittering On: Audience Research And Participation Using Twitter. *Journal of Audience and Reception Studies*. 8: 1.

[9] Idris, N. H. 2014. Credibility Assessment and Labelling on Map Mashup. PhD Thesis. The University of Nottingham.