Improving Accident Statistics: The Role of Social Media

Napong Wanichayapong
Department of Computer Science, Thammasat University
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
napong.wanichayapong@nectec.or.th

Monsak Socharoentum
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
monsak.socharoentum@nectec.or.th,

Wasan Pattara-atikom
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
wasan@nectec.or.th

Abstract

Road safety is a critical issue for Thailand. In 2013, we were ranked the third in terms of high road accident fatality rate according to the World Health Organization [9] and ranked the sixth from World Life Expectancy website [10]. ThaiRSC, Thai Road Accident Data Center for Road Safety Culture, reported that there are 27,627 accidents occurred in Bangkok Area in 2014 [5]. To examine the completeness of an official road accident report from ThaiRSC, we collected road accident reports from an independent and voluntary source, i.e. Twitter, via our natural language processing system, Traffy, and then compared with ThaiRSC source. The result shows that only 1.5% of the reports appear on both ThaiRSC and Twitter. This indicates that data from both sources could significantly complement each other and researchers using ThaiRSC source should be aware that single source of information may not be accurate in terms of completeness.

Author Keywords

Intelligent Transport System; Social Networks; Twitter; Road Safety; Road Accident.

ACM Classification Keywords

H.2.2. Information Search and Retrieval

Introduction

Twitter, despite its original purpose as micro blog social network, can be used to share many kinds of information for many purposes. Among others, people use twitter to report incident or crisis such as earthquake, typhoon [3] and influenza epidemics [1][4]. In 2008, we started Traffy, the project for the social-based road traffic information acquisition, classification and dissemination. Traffy system collects traffic related tweets from independent commuters and three main traffic radio stations with the total of 1.4 million followers out of 4.5 million Twitter users in Thailand.

Thai Road Accident Data Center for Road Safety Culture (ThaiRSC) is maintained by Road Accident Victims Protection (RVP) which is a company established (in 1998) according to the Protection for Motor Vehicle Accident Victims Act. The objective of establishing RVP is to provide a convenient point of contact to victims of road accidents through RVP service centers established nationwide. In Thailand, it is required that road accidents involving a motor vehicle must be reported to RVP through its local service centers and coordinators. In consequence, all accident reports are collected in a database from which ThaiRSC retrieves and publishes. Despite RVP's official authority, a road accident may not be reported when the vehicle involved does not have insurance or accidents are minor such that victims and vehicle owners do not concerns.

Based on the two sources of road accident information (ThaiRSC and Twitter), we raised a research question that “How accurate is the number of accident reports from official data source?”

Tweet Acquisition

We used three Twitter APIs to obtain tweets. Twitter’s Search API for taking query terms and retrieve search results. Twitter’s user_timeline API to obtain tweets from well-known traffic reporters, e.g., a traffic radio station’s Twitter account. Twitter’s mention_timeline API to retrieve tweet that mention @traffy, Traffy Twitter account.

Preprocessing

Unlike English, Thai writing does not use space to separate individual words. Due to this writing rule, we use LEXTO [3] to tokenize individual words from a sequence and then perform word filtering to remove stop words and other irrelevant words. Finally, we converted synonyms and abbreviations into common and official words.

Processing

We identify word attributes, e.g. road name, point of interest (POI) name. Then, performed POI inference to find and tag POI names that are relevant to the tweets using ontology inference rules. Finally, the tweets are classified into one of two roles (question or report). If the tweet is a question then a proper answer is generated and returned. If the tweet is a report, the tweet is converted into ontology instances and recorded in a database which will be used to generate proper answer for an incoming tweet question.

System Accuracy

The system uses incident classification method that we proposed in Social-based Traffic Information Extraction and Classification [8], question interpretation and information acquisition and summarization that we proposed in An Ontology Design for Traffic Incident

Q&A [6] and Road Traffic Question Answering System Using Ontology [7]. The overall system accuracy is as follows.

- 76.85% for POI-based incident classification
- 93.23% for road segment-based incident classification
- 85% for question interpretation
- 76.92% for information acquisition and summarization

Comparison between Twitter and ThaiRSC

We compared the number of accidents between those of Twitter and ThaiRSC. The data are 24 hour basis collected for the entire year in 2014. The results are as follows.

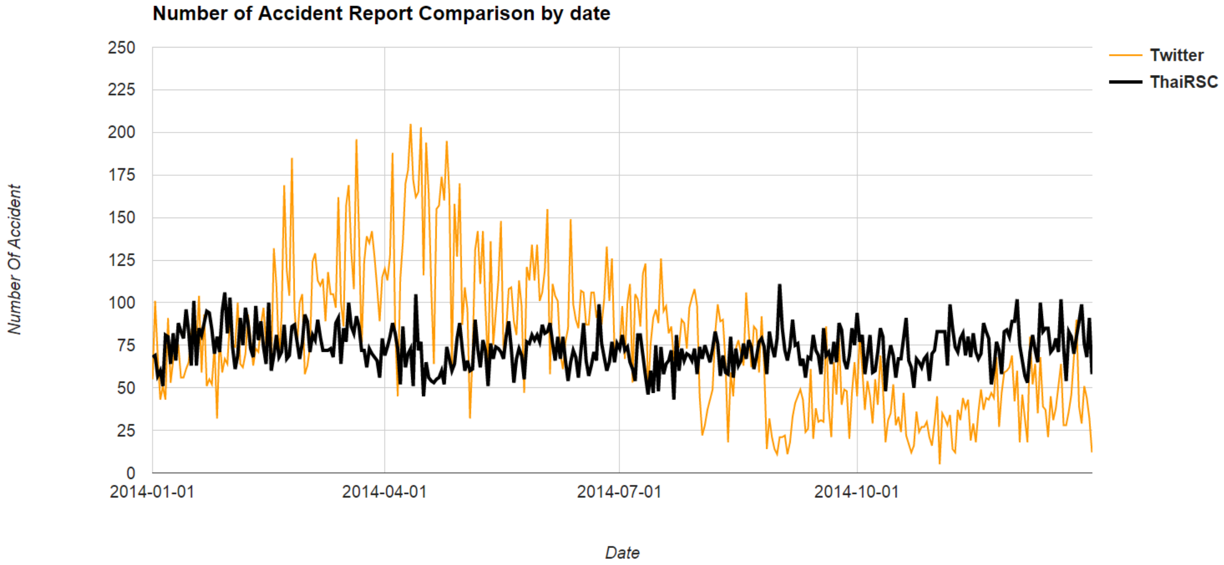

Figure 1: Twitter and ThaiRSC Comparisons: Number of accident by date

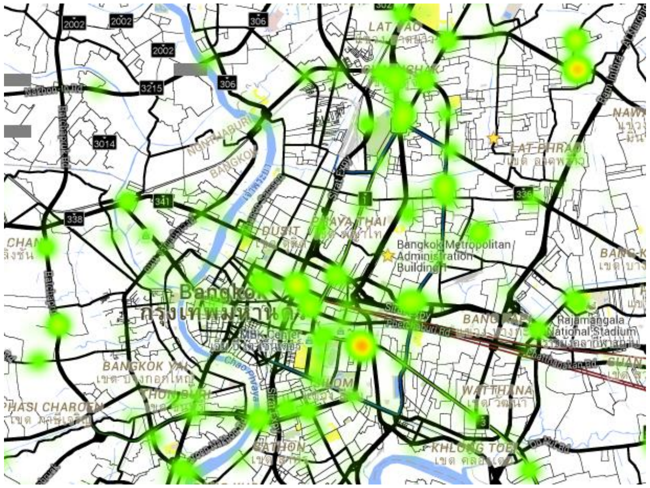

Figure 2: Heat map of accidents from Twitter

As shown in figure 1, ThaiRSC has much narrower range of number of accidents than Twitter (ThaiRSC: 50-100, Twitter: 5-205). The wide range of Twitter is in line with the assumptions that accident report is a voluntarily task such that people may or may not tweet an accident.

In figure 2, accidents in Twitter reports usually appear at junction as plotted on the heat map. The detailed location estimation method is available in Social-based Traffic Information Extraction and Classification [8].

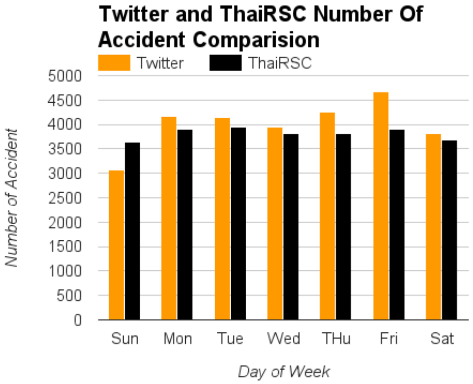

Figure 3: Twitter and ThaiRSC Comparisons: Number of accident by day of week

In general, as shown in figure 3, Twitter reports the higher number of accidents than ThaiRSC from Monday through Saturday. Sunday is the only day that the number of accidents reported from Twitter is lower. The lower accidents reported is reasonable. The number of Twitter users on the road is usually lower; therefore, the number of tweets reduces.

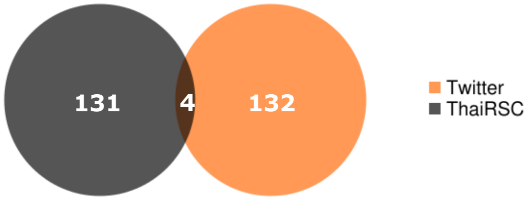

Figure 4: Venn diagram of accident report related to Sathon road by source

To be identified as a common accident, the three criteria must be met simultaneously. The paper is also revised to improve its clarification.

1. Tweet and ThaiRSC reports are created on the same date.
2. Tweet and ThaiRSC reports refer to the same POI.
3. Tweet and ThaiRSC reports refer to the same road and direction.

To verify the completeness of the official accidents reported by ThaiRSC, Sathon road, a major road in Bangkok, is selected. The total of 267 accidents associated with Sathon road was reported. Of the 267 accidents, only 4 accidents are reported from both sources. ThaiRSC contains 131 accidents (49.1%) that do not appear in Twitter while Twitter contains 132 accidents (49.4%) that do not appear in ThaiRSC, as shown in the Venn diagram in figure 4.

Discussion & Conclusion

In this paper we compare accident report from our social-based road traffic report system with ThaiRSC which is an official source of accident report. The result shows that only 1.5% of the reports appear on both ThaiRSC and Twitter. The social-based traffic report is better than ThaiRSC in four aspects: a) it is real-time, b) it has more precise timestamp, c) it contains more detail description, and d) it provides the capability to attach pictures and videos. On the other hand, ThaiRSC is better than our system in two aspects: a) it contain death and injured toll, b) it has accurate geolocation. Due to large portion of non-overlapping reports/tweets, each source can be considered as a complement to each other. Our findings reveal that Twitter has strong potential to improve data completeness of the official source in order to achieve better accuracy of accident reports.

References

1. Aramaki, E., Maskawa, S., and Morita, M. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. 2011. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 1568–1576.
2. NECTEC. Thai Lexeme Tokenizer. http://www.sansarn.com/lexto/.
3. Sakaki, T., Okazaki, M., and Matsuo, Y. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *Proceedings of the 19th International Conference on World Wide Web*, ACM , 851–860.
4. Signorini, A., Segre, A.M., and Polgreen, P.M. 2011. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE 6*, 5, e19467.
5. ThaiRSC. Thailand road accident location Map. http://www.thairsc.com/ReportAccidentGMapV3.aspx.
6. Wanichayapong, N., Pattara-Atikom, W., and Peachavanish, R. 2013. An Ontology Design for Traffic Incident Q&A System.
7. Wanichayapong, N., Pattara-Atikom, W., and Peachavanish, R. 2014. Road Traffic Question Answering System Using Ontology. In T. Supnithi, T. Yamaguchi, J.Z. Pan, V. Wuwongse and M. Buranarach, eds., *Semantic Technology*. Springer International Publishing, 422–427.

8. Wanichayapong, N., Pruthipunyaskul, W., Pattara-atikom, W., and Chaovalit, P. 2011. Social-based traffic information extraction and classification. *2011 11th International Conference on ITS Telecommunications (ITST)*, 107–112.

9. World Health Organization. Road Traffic Deaths and Proportion of deaths by Road User, by country/area. .

10. World Life Expectancy. World Life Expectancy. http://www.worldlifeexpectancy.com/.