REAL-TIME PUBLIC SENTIMENTS ANALYSIS AND INFORMATION INTEGRATION PLATFORM FOR DISASTER PREVENTION AND VICTIMS OF DISASTER RESCUE BASED ON SOCIAL NETWORKS

Nai-Wei Lo[1], Kuo-Hui Yeh[2],Raylin Tso[3],Kuo-Yu Tsai[4], Bor-Shiun Lin[5], Tzu-Yin Chang[6], Chih-Hao Liu[7]

[1]Department of Information Management, National Taiwan University of Science and Technology, nwlo@cs.ntust.edu.tw

[2]Department of Information Management, National Dong Hwa University, khyeh@mail.ndhu.edu.tw

[3]Department of Computer Science, National Chengchi University,raylin@cs.nccu.edu.tw

[4]Department of Applied Mathematics, Chinese Culture University, cgy13@ulive.pccu.edu.tw

[5]Sinotech Engineering Consultants, INC., bosch.lin@sinotech.org.tw

[67]National Science and Technology Center for Disaster Reduction (NCDR), [6]geoct@ncdr.nat.gov.tw; [7]liuchihhao@ncdr.nat.gov.tw

Abstract

With the rapid advancement of social networking services, people tend to exchange and share information online. Massive amounts of global information are aggregated promptly and circulated quickly via social networks, such as Facebook, LINE, PTT and Dcard. How to extract useful information from social networks to support decision making on public events is one of the most important research issues for the Taiwan government.In this paper, we focus on the information integration of disaster events and the analysis of public sentiment on social networks. Three subjects are thoroughly investigated: (1) predictive analysis on disaster information in social networks via natural language processing and semantic model analysis; (2) information extraction and sentiment trend analysis for disaster events on social networks and (3) a crowdsourcing based correctness verification approach toward information on social networks.

Keywords: Crowdsourcing, Disaster Prevention, Refugee Rescue, Public Sentiments Analysis, Social Networks, Semantic Model Analysis

1. INTRODUCTION

Due to the rapid development and advancement of Internet services and search engines, people tend to exchange and share information online. Massive global data is aggregated promptly and circulated quickly via multi-modal network technologies. Government, industry and academia are aware of the significant benefits of knowledge potentially existing inside the aggregated data (and data set). Hence, how to extract useful information and valuable knowledge to support decision making on public events has become a hot research topic in recent years. Among the numerous data sources, social networking services has become one of the most popular and important platforms for information dissemination and sharing among a multitude of Internet users. Social networks provide a set of perspectives for analyzing the whole structures of involved entities as well as for explaining phenomena and the patterns observed in these structures. Predictive analytics, including data capture, data analysis, data storage, visualization and data curation, on social service platforms, such as Facebook, LINE, PTT and Dcard, have been highly focused on by industrial and academic communities. For example, in the 2014 Kaohsiung gas explosions, two downtown districts, i.e. Cianjhen and Lingya, were ripped apart by a series of explosions, which were triggered by underground gas leaks, killing 32 people and injuring 321. During the disaster rescue organized by the Taiwan government, a social networking platform, called gov-zero, was established by society to support disaster information integration and rescue resource allocation. The temporarily-built social networking platform played a critical role in the rescue of the victims. It is obvious that information management and rescue resource allocation in public disaster events may be significantly affected by the knowledge and intangible assistance gained from social networks. Therefore, in this study we would like to design an information integration and management platform for social networks. The platform is utilized to support disaster prevention and refugee rescue and real-time public sentiments analysis as well. Three subjects are thoroughly investigated. First of all, we develop an automatic data collecting mechanism to retrieve disaster relevant information from social networks and implement semantic model analysis techniques to extract precise information related to disaster events. Secondly, by combining the Jieba system, i.e. a Chinese word segmentation utility with the semantic orientation pointwise mutual information (SO-PMI) algorithm, we implement a sentiment analysis tool to determine public sentiments on social networks via semantic orientation on disaster events. Thirdly, as data correctness plays a crucial role from the viewpoint of data quality, it is important to provide a judgement mechanism for evaluating the correctness of information retrieved from social networks. We thus propose a crowdsourcing based approach which is based on feedback from many people to verify the correctness of extracted information on social networks.

2. RELATED WORKS

Latent Semantic Analysis (LSA) was proposed by Deerwester *et al.* (1990) as an information retrieval technique, where two concepts, i.e. dimensionality reduction and singular value decomposition, are exploited to conduct the potential correlation between words and documents. Various applications based on LSA have been proposed. For example, Kuo *et al.* (2013) introduced a multiple-type LSA model for analyzing the potential connection between the background music and the video shot. Lintean *et al*. (2010) and Klein *et al*. (2011) focused on the assessment of text responses in an automatic way. Both the two researches are based on LSA.Ozsoy *et al*. (2011) which utilized LSA as the data analysis technique to process documents and integrate all the information, while Luh*et al*. (2012) combined LSA and a genetic algorithm to predict the rank of the search engine. On the other hand, Hofmann *et al*. (1999) introduced Probabilistic Latent Semantic Analysis (PLSA) which provides better perfomance on identifying latent semantic and utilizes an expectation maximization algorithm to improve the similarity. Famous applications based on PLSA are as follows. Zhang *et al*. (2010) used structured PLSA to analyze the image and obtain identification efficiency. Later, Shen *et al*. (2011) adopted PLSA as a cluster technique for the classification of documents and further retrieve the abstracts of the documents automatically.

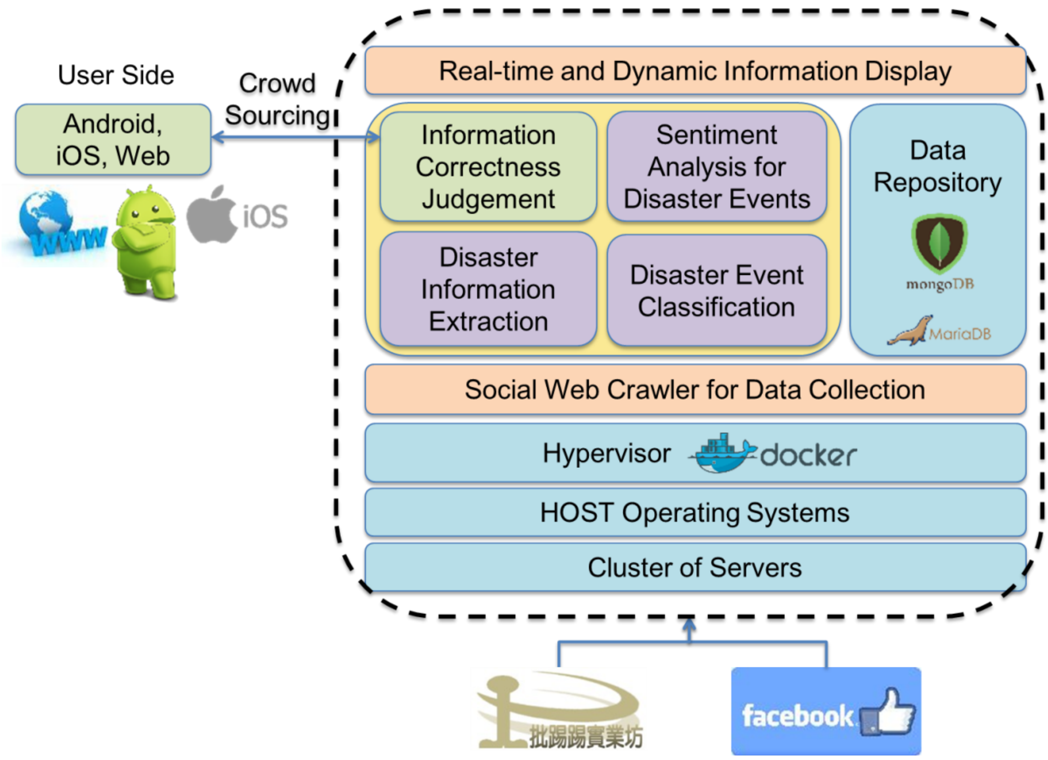
Sentiment analysis is the procedure for the natural language processing which deals with the extraction of objective and subjective statements. (Pang and Lillian, 2008) By using natural language processing (NLP), sentiment analysis collects opinion or sentiment words and determines whether opinion or sentiment words are positive, negative or neutral. That is, sentiment analysis is used to recognize the sentiments of some people about some topic or event. Hai *et al*. (2014) pointed out that it is an attractive research area to determine subjective attitudes in big social data. Turney, P. (2002) presents a two-word-phrase learning algorithm for classifying reviews as recommended or not recommended, in which the classification of reviews is predicted according to the average semantic orientation of the phrases in the reviews that contain adjectives or adverbs. Esuli and Sebastiani (2006) described a publicly available lexical resource for opinion mining (also called SENTIWORDNET), in which each WORDNET synset is associated with three types of numerical scores: Obj(s), Pos(s) and Neg(s). The algorithm is based on the quantitative analysis of the glosses associated with synsets. Subrahmanian and Reforgiato (2008) proposed an adjective verb adverb framework (AVA for short) to determine opinions about some topic or event. By using supervised machine learning (Support Vector Machines), Kessler and Nicolov (2009) presented a ranking algorithm for potential target mentions of a sentiment expression, in which the main feature is the syntactic configuration that connects the sentiment expression and the mention. For extracting sentiments from the text, Taboada *et al*. (2011) presented a lexicon-based algorithm which uses dictionaries of words annotated with the semantic orientation (polarity and strength), and incorporates intensification and negation. Romanyshyn (2013) proposed a rule-based algorithm for clause-level sentiment analysis of reviews in Ukrainian, in which the main emphasis is made on the design of rules for computing sentiments. Bandyopadhyay and Mallick (2014) presented a shortest path based hybrid measure algorithm for ontological similarity between two terms, combining both the structure of the GO graph and the information content of the terms. Recently, Agarwal and Mittal (2016) adopted adjectives and single word phrases to determine whether sentiments of sentences are positive, negative or neutral. Agarwal and Mittal's proposed algorithm is also a lexicon based algorithm for sentiment analysis.

The term crowdsourcing first appeared in Howe (2006). It is the act of taking a job traditionally performed by a designated employee and outsourcing it to an undefined generally large group of people in the form of an open call. The features of crowdsourcing are as follows: (1) Enterprise can take response a few times and gather the outcome; (2) Not outsourced to a specific organization or entity but to the public through the Internet; (3) People participate when they have free time and (4) Participants do not primarily work for money, but to prove themselves. Since then, the crowdsourcing concept has been widely used in many applications. For example, Santani *et al.* (2015) mentioned the main functionality of crowdsourcing and gave an example to show how it can be done. Nairobi is one of the major business and technology powerhouses in Africa. Since Nairobi lacks road monitoring equipment, it is hard to obtain real-time traffic information. In order to overcome this problem, the use of mobile crowdsourcing is investigated to collect and record Nairobi's road quality information. The main idea was to establish a crowdsourcing application, testing it using thirty participants over two weeks. Users can upload road information through this application (including photos, description and user's location) if they find problems such as potholes, speed bumps and so on. Therefore, Nairobi can gather the road quality information that the city lacks. To examine the correctness, they proposed to use online crowdsourcing using Amazon's Mechanical Turk (MTurk) to verify whether uploaded information indeed described the road situation. At the end of the test, they found that 92% of users submitted reports that were the same as the MTurkers judgements.

3. THE PROPOSED PLATFORM

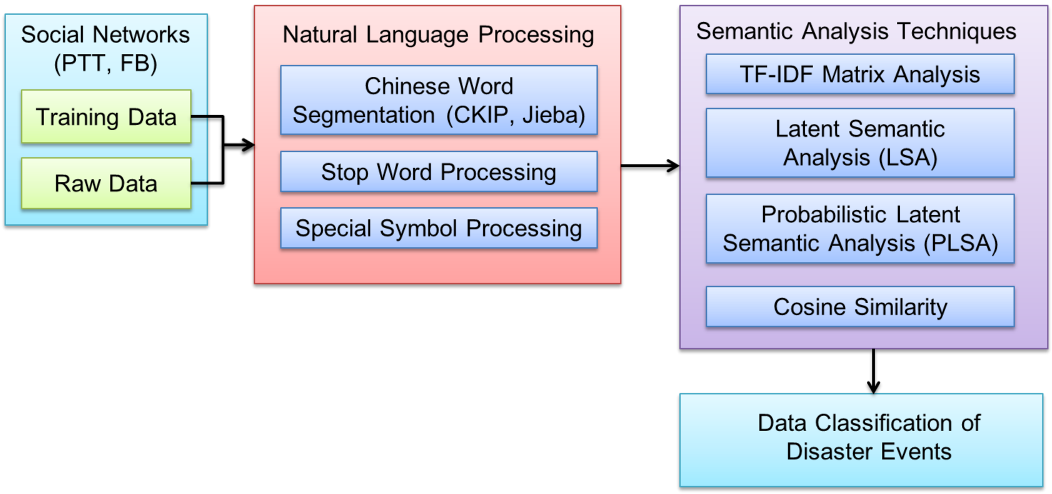
In this study, we focus on disaster information circulation and public sentiment analysis on social networks. A service platform for integrating disaster information and analyzing public sentiment is introduced to effectively and correctly integrate the disaster information of social networks.The proposed system platform integrates data from various social networks via self-designed and automatic data collecting mechanisms (i.e. web data crawlers). Disaster prevention information is then analyzed and filtered.That is, notifications of disaster events will be conducted and sent from the platform and then forwarded to the government and the local population as the newest disaster prevention information. The system architecture of the proposed information integration platform is shown in Figure 1. The data source is based on PTT and Facebook which are the most popular social networks in Taiwan. Cloud based data management tools, such as Hadoop and Docker, are adopted as major data processing techniques. Under this cloud cluster architecture, three topics are investigated. Those are (1) Predictive analysis on disaster information in social networks via natural language processing and semantic model analysis (i.e. Figure 2), (2) Information extraction and sentiment trend analysis for disaster events on social networks (i.e. Figure 3) and (3)A crowdsourcing based correctness verification approach of information on social networks (i.e. Figure 4). Finally, using a real-time and dynamic information display board, our platform is able to provide timely and correct information with a user-friendly interface. In the following, we describe the detailed procedures and designs for each topic.

Figure 1. The proposed Information Integration Platform for Disaster Prevention and Refugee Rescue based on Social Networks



In the first topic (Figure 2), we collect an array of data from social networks (i.e. PTT or Facebook) as the training data set. Using natural language processing, such as chinese word segmentation tools (CKIP) and stop word and special symbol elimination, the training data are pre-processed to remove all inadequate data when the data is analyzed. Semantic analysis techniques, such as TF-IDF matrix analysis, latent semantic analysis and probabilitistic latent analysis, are then performed to classify the data into categories which possess the same characteristics (or belong to the same disaster event).

Figure 2. Predictive Analysis on Disaster Information in Social Networks via Natural Lanaguage Processing and Semantic Model Analysis



In the second topic (Figure 3), based on the classification of disaster events, we then perform user content analysis to filter and extract the user's sentiment. The Jieba system is used to perform the chinese word segmentation and a so-called semantic orientation-pointwise mutual information algorithm is adopted to extract the user's sentiment. In addition, with the self-designed user comment filter and the disaster vocabulary base, we can correctly and actually identify the abstract information of disaster events discussed on the social networks.

In the third topic (Figure 4), due to the functionality that crowdsourcing can gather the reliable data that we are lacking, we adopt this concept to our study to filter out the incorrect information of disaster events and keep the reliable information. The process is as follows:

(1) Users can upload information of disaster events, such as photos, comments, disaster location and user's location on to our database.
(2) The database manager receives the information and makes a simple judgement; if no error happens, then the manager pushes the hazard formation to users.

(3) Users receive push notifications and send back their comments (or evidence). The database manager decides whether to save or remove the information according to the user's comments.

Figure 3.Information Extraction and Sentiment Trend Analysis for Disaster Events on Social Networks

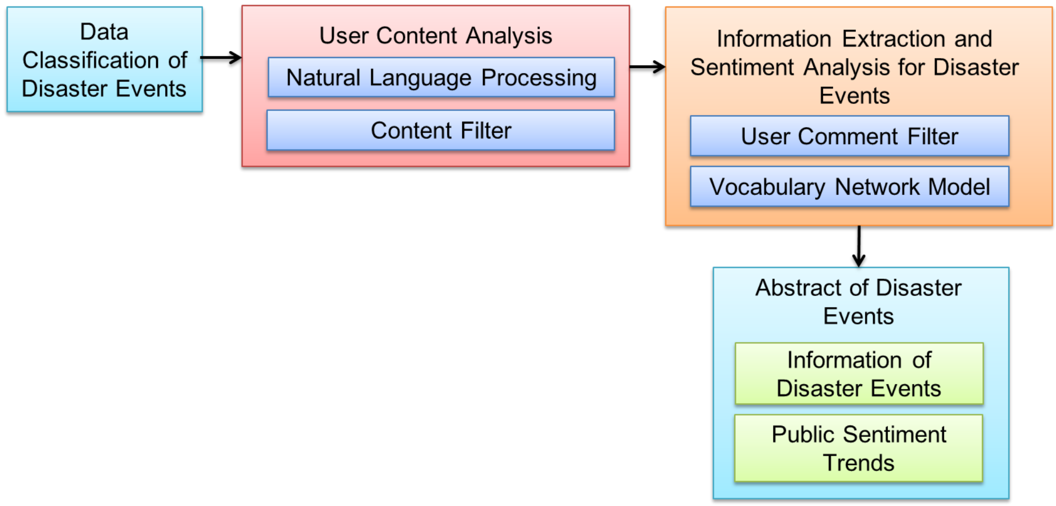
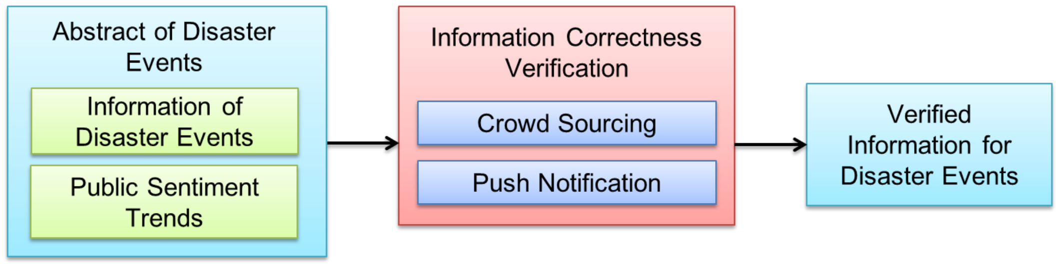


Figure 4. ACrowdsourcing based Correctness Verification Approach of Information on Social Networks



4. IMPLEMENTATION

Our system implementation is written with Node js and operated as follows. First, we present the disaster relevant informaiton extracted from the social networks via TGOS, i.e. Taiwan Geospatial One-Stop (2013), made by the Taiwan government.

Each point is represented as each pieces of news is retrieved from social networks (i.e. PTT or Facebook), and all of them are able to be dynamically and real-time displayed on the map constructed by the TGOS.We define three colors for the status of each point: green is news that is true and confirmed, yellow is news that is not true and rejected and red is news that is unable to be confirmed. In addition, the blue item is the location information of this news.Note that all the points on the TGOS based information display board can be represented as a list based format (i.e. Figure 6).For extracting useful information for disaster events, we employ the web data crawler, natural language processing and semantic analysis techniques in the background without the user's notices. All of the detailed procedures of data processing are presented as those in section 3.

Figure 5. TGOS based Information Display Board

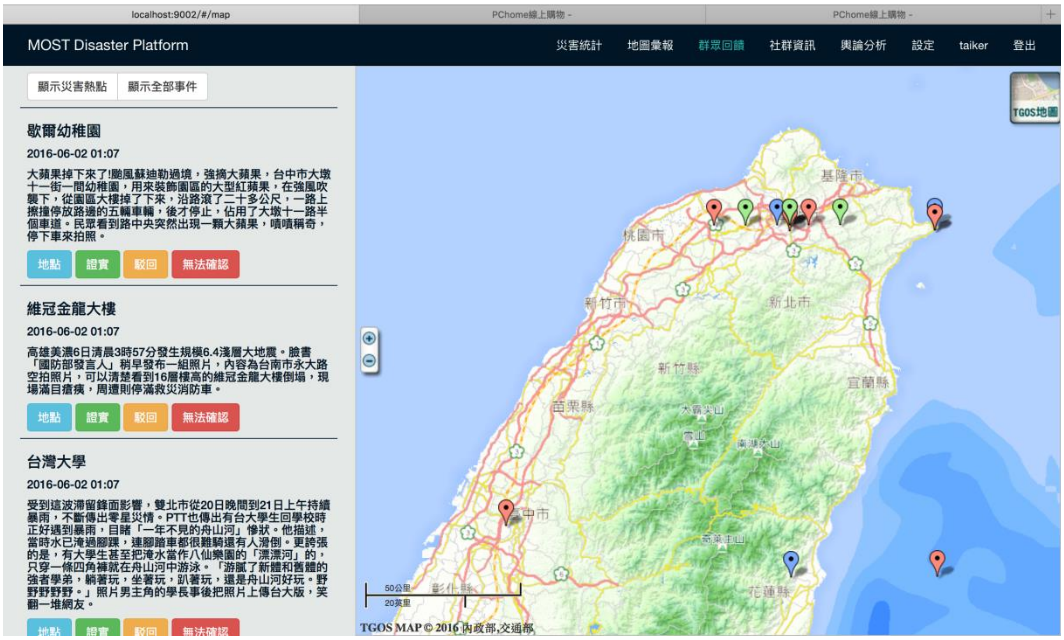
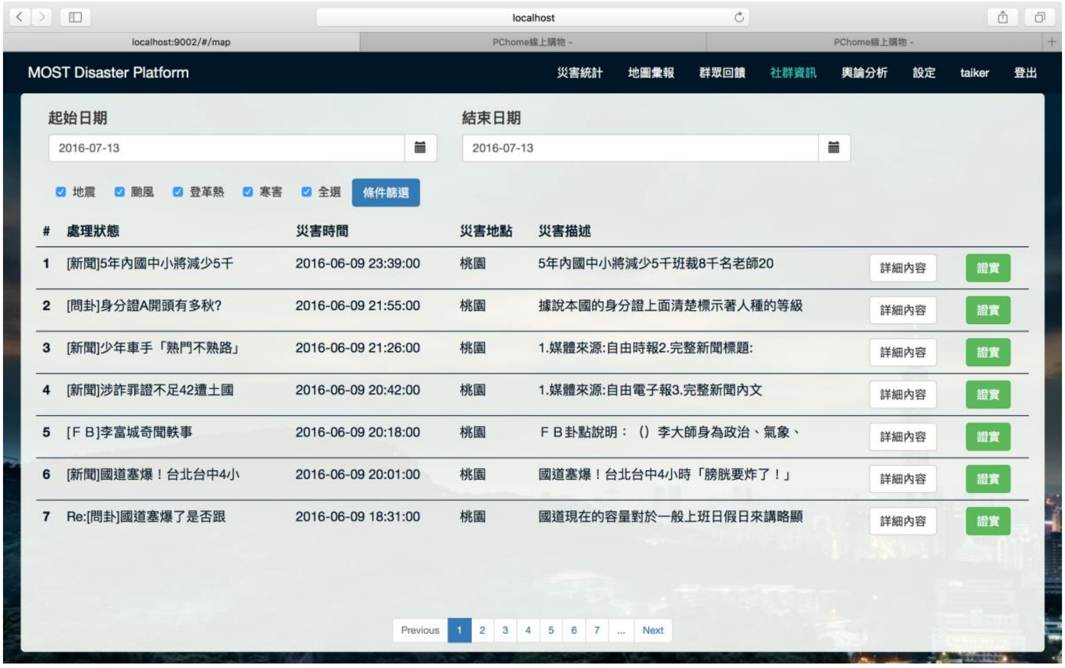


Figure 6. Listed Information Retrieved from Social Networks



4.1. A Sentiment Analysis Tool for Disaster Events on Social Networks

This sentiment analysis tool includes the Jieba analysis (a Chinese word segmentation utility) and the semantic orientation-pointwise mutual information (SO-PMI) algorithm. The process is as shown in the following figure 7. Firstly, the Text Reader opens and reads the opinion file. According to the segmentation dictionary, the Jieba analyzer then divides the text from the opinion file into meaningful units, such as words, sentences, or topics. The segmentation, consistency and granularity of Chinese words is important for the following sentiment analysis. In addition, some Chinese words about natural disasters can be added into the segmentation dictionary file, such as "水災", "淹水", "警戒水位," etc. Finally, the sentiment analyzer determines the attitude of citizens in Taiwan with respect to some natural disaster by referring to the text analysis result and the sentiment dictionary. Different Chinese sentiment words also can be inserted into the sentiment dictionary file, such as "棒", "不錯", "糟糕", etc.

The sentiment analysis tool focuses on earthquakes, typhoons, dengue fever, and winter injuries (including low temperature injury, frost injury, desiccation injury, snow and ice breakage, etc.). Our implemented user interface is shown as Figure 8. The user can set up the beginning date (起始日期) and the ending date (結束日

期) for some disaster events, and choose the earthquake event (地震), the typhoon event (颱風), the dengue fever event (登革熱), the winter injury event (寒害) or all (全部) by clicking on the checking box. Two or more events chosen at once is allowed in our implement interface.

Figure 7. The process of the sentiment analysis tool

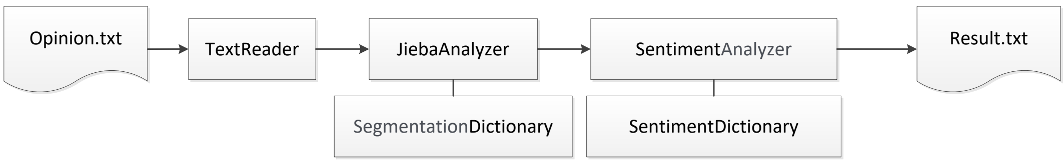
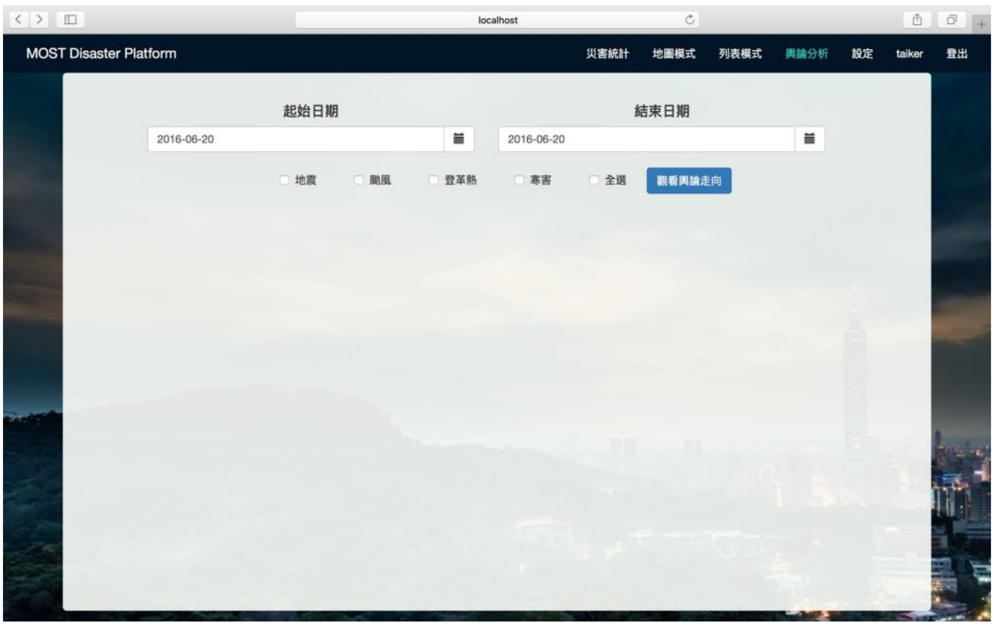


Figure 8. The interface for sentiment analysis



After pushing the “觀看輿論走向” button, the sentiment analysis results will be stored in the database, as shown in Figure 9. There is a graphic display showing the sentiment analysis results in our implemented user interface. The x-coordinate and the y-coordinate represent the date that the user chose and the number of the sentiments in the broken line graph, respectively. The green broken-line stands for the positive sentiments, while the red broken-line stands for the negative sentiments. In Figure 10, the green bar and the red bar represent the percentages of the positive sentiments and negative sentiments (正負面評價)，respectively.

Figure 9. The sentiment analysis results stored in the database

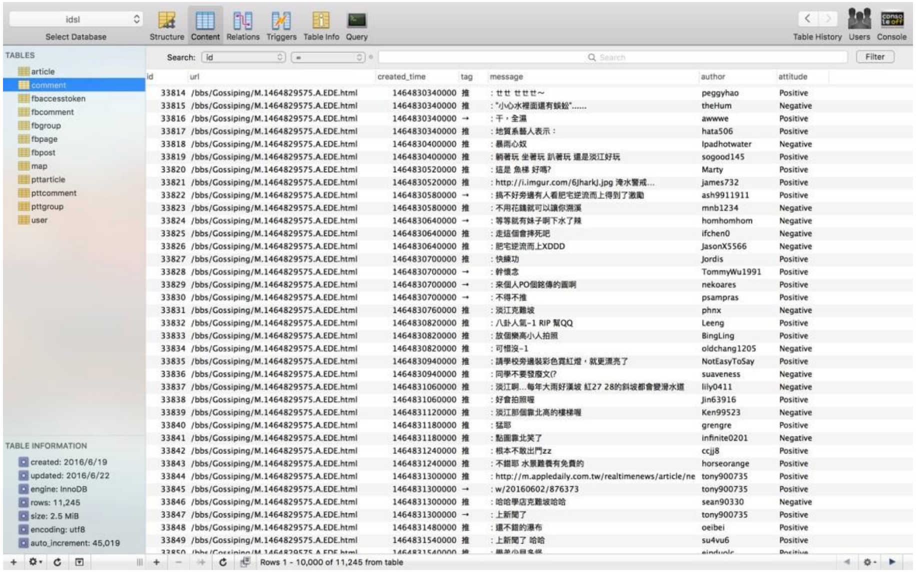


Figure 10. The interface for sentiment analysis

4.2. A Crowdsourcing Based Correctness Verification Approach to Information on Social Networks

In our correctness verification approach, we develop an app which is a system of providing real-time disaster information. Upon using the app, users can take real-time photos through the camera of its mobile device along with Google Map's location features, then upload the information to the social network. This feature is useful and available for the public to understand the actual situation about a disaster in real time. Moreover, users can see a response or a feedback about the situation from other users.Our app is developed by PhoneGap which is an open-source mobile development framework. PhoneGap allows developers to use programming language such as HTML, Javascript, CSS Web APIs to develop a cross-platform mobile app. In detail, the features of the app can be divided into the following functions:

- Real time photo taking and reporting: Users can take pictures in real time with their mobile device or choose the pictures which are already in their mobile devices. After selecting a photo, it will be displayed on the upload page, and uploaded after pushing the confirmation button. Photographs taken through the site allow users to understand the actual situation and also get the related information. The main purpose of this function is to let NCDR (National Science and Technology Center for Disaster Reduction) to understand the real situation of a disaster in real time, and to make the right decision in response and to send a rescue team if necessary.

- Locating the accurate place of disasters by GPS: A user with a GPS-enabled mobile device can mark his current location by the app. Also, he can mark the disaster area where the user wants to report and send the information back to the server for further confirmation.

With this function, we can find out the location of the user through the location-based system in our mobile device, and mark the location on the map with a flag (see figure 11 for example). The location of the user will be transformed into coordinates and the street through the location system. If the user does not open the location system, the app will not be able to find the current location. In this case, our app will show the initial location which is defined as the location of the National Chengchi University of Taiwan (see figure 12).

Users can report their location and/or the location of disasters when the GPS system is open. To do this, a user can just open the app, touch the map shown up on the screen of his/her mobile device. The map will display the location which the user wants to mark (see figure 13 and figure 14). Then, the position, coordinates, distance and street of the place will be displayed on the map as the information of the location.

Figure11. Location Marked

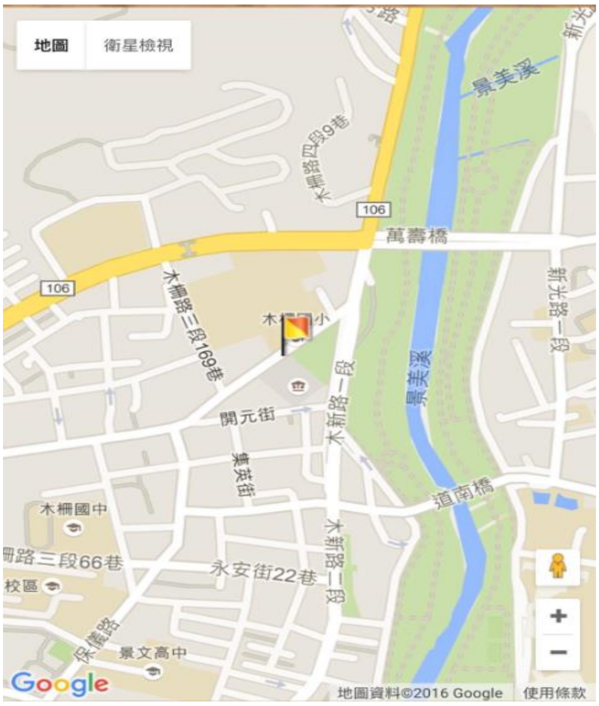


Figure12. Location Display on a Device when GPS is Unavailable

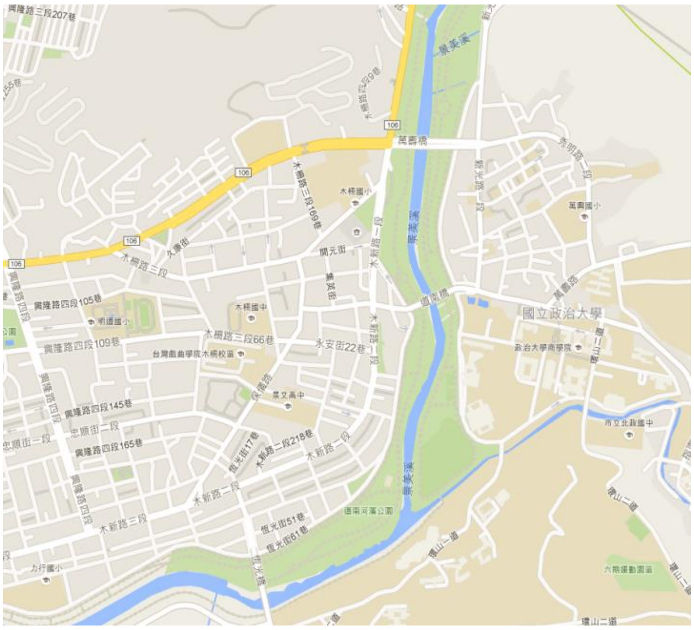


Figure13. Disaster Location Marked by a User

Figure14. Details of a Location Marked

After marking the position, users can choose the types of disasters happening at the marked location. There are three types of disasters we have defined currently in this app. They are typhoons, earthquakes and floods. The types of disasters is scalable and can be added in the future if necessary.Users can upload the corresponding photos after marking the location and defining the type of a disaster. The uploading page provides some confirmation information for the user. He can decide whether to upload the image and/or other comments about the disaster (see figure 15).

Figure15. UploadingPage

To ensure the integrity of the uploaded information, if the user does not open the GPS or choose the position, he will not be able to upload the photos and comments. When photos and/or comments are uploaded successfully, the database will update the information including checking the user ID to make sure who is reporting this disaster (see figure 16).

Figure 16. Information Stored on the Database

		mark_lng	user_dis	user_addr	damage	user_id	comment
6	726682686	121.57163858413696	469.34	116台灣台北市文山區木新路二段79巷12巷24號 台灣台北	Flood	1	水災
7	9216167168	121.53865814208984	242.71	106台灣台北市大安區基隆路高架道路 台灣台北	Earthquake	1	
8	9216167168	121.53865814208984	242.71	106台灣台北市大安區基隆路高架道路 台灣台北	Earthquake	1	
9	0725148348	121.53634071350098	502.51	116台灣台北市文山區汀州路四段33號 台灣台北	Fire	1	
.0	498331363	121.56794786453247	379.75	116台灣台北市文山區木柵路三段102巷19號 台灣台北	Earthquake	1	地震發生！

Any user of this app can check a disaster including the location and disaster type as well as the corresponding photos and comments by clicking on the map (see figure 17, figure 18). Users can confirm the correctness of the uploaded information.With the help of this app, the rescue team can narrow down the scope and the area of an unknown disaster. This will help them to accelerate the rescue operation.

Figure 17. Disaster Location Shown on the Map

Figure 18. Detailed Information Shown on the Map

5. CONCLUSIONS

In this study, we focus on information circulation and sentiment trend analysis on public disaster events in social networks. An information integrated system is implemented with various data analysis techniques, such as natural language processing, Chinese word segmentation, semantic analysis model, semantic orientation pointwise mutual information processing, to effectively and correctly retrieve and integrate disaster information from social networks. The extracted information can truly be reflected the sentiment trend of the masses and is useful for refugee rescue and resource allocation. In addition, to evaluate the correctness of the information retrieved from social networks, we implement a correctness judgement mechanism which is based on feedback from the public as evidence of event judgement. We develop a mobile application and a corresponding web platform to collect meaningful information from the public and are able to identify the correct information for a specific disaster event.

6. REFERENCES

Agarwal, B. and Mittal, N. (2016). Prominent Feature Extraction for Review Analysis: An Empirical Study, *Journal of Experimental and theoretical Artificial Intelligence*, 28(3): 485-498.

Bandyopadhyay, S. and Mallick, K. (2014). A New Path Based Hybrid Measure for Gene Ontology Similarity, *IEEE/ACM Transactions on Computational Biology andBioinformatics*, 11(1): 116-127.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). "Indexing by Latent Semantic Analysis," American Society for Information Science (41:6), pp 391-407.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining, *Proceedings 5th International Conference on LanguageResources and Evaluation (LREC)*, pp, 417–422.

Hai, Z., Chang, K., Kim, J.J., Yang, C.C. (2014). Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance, *IEEE Transaction on Knowledge and Data Engineering*, 26(3): 623-634.

Hofmann, T. (1999). "Probabilistic latent semantic analysis," Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., pp. 289-296.

Howe,J. (2006) The Rise of Crowdsourcing. *Wired Magazine*.

Kessler, J. S. and Nicolov, N. (2009) Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations, *Proceedings 3rd International AAAI Conference on Weblogs and Social Media*.

Klein, R., Kyrilov, A., and Tokman, M. (2011). "Automated Assessment of Short Free-Text Responses in Computer Science using Latent Semantic Analysis," Innovation and Technology in Computer Science Education, pp. 158-162.

Kuo, F.-F., Shan, M.-K., and Lee, S.-Y.(2013). "Background Music Recommendation for Video Based on Multimodal Latent Semantic Analysis," International Conference on Multimedia and Expo, pp. 1-6.

Lintean, M. C., Moldovan, C., Rus, V., and McNamara, D. S. (2010)."The Role of Local and Global Weighting in Assessing the Semantic Similarity of Texts Using Latent Semantic Analysis," Florida Artificial Intelligence Research Society, pp. 235-240.

Luh, C.-J., Yang, S.-A., and Huang, D. T.-L.(2012). "Estimating Search Engine Ranking Function with Latent Semantic Analysis and a Genetic Algorithm," International Conference on E-Business and E-Government, pp. 439-442.

Ozsoy, M. G., Alpaslan, F. N., and Cicekli, I. (2011)."Text summarization using Latent Semantic Analysis," Journal of Information Science (37:4), pp 405-417.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis, *Foundations and Trends in Information Retrieval*, 2(1–2): 1-135.

Romanyshyn, M. (2013). Rule-Based Sentiment Analysis of Ukrainian Reviews, *International Journal of Artificial Intelligence & Applications (IJAIA)*, 4(4): 103-111.

Santani, D., Njuguna, J., Bills, T., Bryant, A. W., Bryant, R., Ledgard, J., and Gatica-Perez, D. (2015). CommuniSense: Crowdsourcing Road Hazards in Nairobi. *Proceedings of MobileHCI'15*, 445-456.

Shen, C., Li, T., and Ding, C. H. (2011). "Integrating Clustering and Multi-Document Summarization by Bi-Mixture Probabilistic Latent Semantic Analysis (PLSA) with Sentence Bases," Association for the Advancement of Artificial Intelligence, pp. 914-920.

Su, E. C., Chang, J.-M., Cheng, C.-W., Sung, T.-Y., and Hsu, W.-L.(2012). "Prediction of nuclear proteins using nuclear translocation signals proposed by probabilistic latent semantic indexing," Bio Med Central Bioinformatics (13:S17-S13), pp 1-10.

Subrahmanian, V. S. and Reforgiato, D. (2008). Ava: Adjective-Verb-Adverb Combinations for Sentiment Analysis, *IEEE Intelligent Systems*, 23(4): 43–50.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis, *Computational Linguistics*, 37(2): 267-307.

Taiwan Geospatial One-Stop (2013),http://tgos.nat.gov.tw/tgos/EngWeb/TGOS_ENG_Home.aspx

Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *Proceedings 40th Meeting of the Association for Computational Linguistics*, pp. 417–424, Philadelphia, PA.

Zhang, J., and Gong, S. (2010). "Action Categorization by Structural Probabilistic Latent Semantic Analysis," Computer Vision and Image Understanding (114:8), pp 857-864.