Disaster Consequence Analysis of Thailand's Severe Flood

Phat Jotikabukkana[*], Virach Sornlertlamvanich[*], Yukari Shirota[†], and Takako Hashimoto[‡]
[*] School of ICT, Sirindhorn International Institute of Technology, Thammasat University, Thailand
[†] Faculty of Economics, Gakushuin University, Japan
[‡] Faculty of Commerce and Economics, Chiba University of Commerce, Japan

***Abstract*— Severe Flooding in Thailand during year 2011 has destroyed many Industrial Estates infrastructure. The estimated amount of economic damages at 1,425 billion baht is reported by the World Bank, the world's fourth costliest disaster. There are a lot of related information, conclusions, and suggestions after this critical event. We have focused on defining an exactly demand and important information of any state during the flood period. We utilized related online news article and Wikipedia to verify a transition state of this big flood. The four main states, Before Flood (BF) period, Heavy Flood period (HF), Ebb Away period (EA), and Help&Recovery period (H&R), are verified by time series. Term Frequency-Inverse Document Frequency (TF-IDF) technique is used to extract the essence from bag-of-words. While a text normalization by utilizing the fuzzy set of words is used to improve model efficiency. Finally, we found the significant keywords which be a representative of demanding and solution for each state. This result could be a good clue or guideline for government to deal with upcoming disaster.**

Keyword: Serve Flood, transition state, Term Frequency-Inverse Document Frequency (TF-IDF), bag-of-words, text normalization, fuzzy set of words

I. INTRODUCTION

In 2011, Thailand encountered a critical disaster of deluge and inundation, the server flood. This is one of the most serious disaster which has been occurred in Thailand. Most of industrial estate infrastructure are destroyed. The manufacturing industry and manufacturing supply chains are interrupted. This is a main reason of a global hard disk drive shortage problem and affected regional automobile production.[1] However, bad event usually generates a good experience, we found many related information as a valuable chronicle. There are a lot of online news articles, Wikipedia, and video from YouTube which have been recorded in time series format. Most of online news websites have the event's news articles as archive files. While Wikipedia, and some government's websites already summarized the event as time series annals which contain all important information, cause of the problems, related news, solutions and suggestions. All data can be separated into four main transition states: Before Flood period (BF), Heavy Flood period (HF), Ebb Away period (EA), and Help and Recovery period (H&R). These data source are the precious data source. Analyzing and summarizing all of these important data could generate a worthy data for executive level and all.

We have an idea to analyze this precious data by using bag-of-words model, Term Frequency-Inverse Document Frequency (TF-IDF) [2], and text normalization with fuzzy set of words technique to find the essential keywords which become a good guideline for organization and government to keep up with the next disaster.

In section II, main techniques and preprocessing are explained. Our approach is described in section III. Then, experiment result is shown in section IV. Finally, section V, a discussion in conclusions and suggestions is illustrated.

II. MAIN TECHNIQUES AND PREPROCESSING

There are three main parts in this section, Main Transition State Identification, Text Normalization with Fuzzy Set of Words, and Term Frequency-Inverse Document Frequency (TF-IDF), which described as follows:

A. Main Transition State Identification

Refer to our considered data is a time series format one, and the essential keywords from each state of event would be gleaned, we need to identify the exactly date of state transition: BF, HF, EA, and H&R, to limit the scope of the data (online news articles, and Wikipedia) to generate the bag-of-words model. We used the electricity consumption value of a company in one of affected industrial estate area [3] as shown in Fig. 1. to identify the specific date.

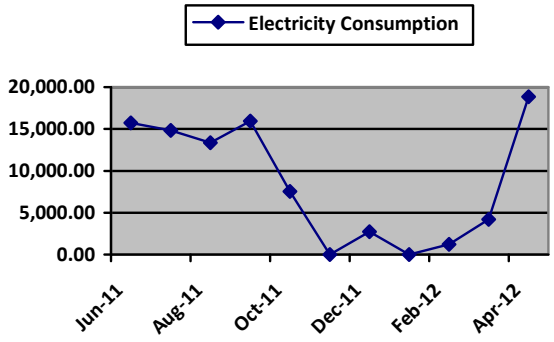


Fig. 1 An Electricity Consumption data of a company in the affected Industrial Estate area.

We reaffirmed the date of state transition by using the record of events data in the Wikipedia, and human judging after reading all related articles. Finally, we can identify the date of the four main transition states: BF, HF, EA, and H&R as shown in Table I.

TABLE I
DETAIL OF THE MAIN TRANSITION STATES

Transition States	Date
Before Flood period (BF)	June 1 – Oct 5, 2011
Heavy Flood period (HF)	Oct 6 – Nov 15, 2011
Ebb Away period (EA)	Nov 16 – Dec 30, 2011
Help&Recovery period (H&R)	Oct 10, 2011 – Jan 30, 2012

For H&R period, we consider all help and recovery solutions along the HF, EA and after EA period to see all essential data of aid and supports.

B. Text Normalization with Fuzzy Set of Words

A nature of Thai's online news article, we mostly found the abbreviation of person name, long organization name, and long topic name when they are mentioned as the next mention in the article. For example, 'Prime Minister Yingluck Shinawatra' is a full name format as the first mention in the news article, while 'Prime Minister Yingluck' or 'Miss Yingluck' are the abbreviation format of the same person as the next mention in the same article. This semantic [4] problem can lead a bag-of-words model and TF-IDF to generate incorrect significant value of words. We utilized the fuzzy set of words as preprocessing (text normalization) technique to change all words with same meaning into one word. Some sample of the fuzzy set are shown in Table II.

TABLE II
A SAMPLE OF THE FUZZY SET OF WORDS

Original Word	Fuzzy Set
'นายกรัฐมนตรี ยิ่งลักษณ์ ชินวัตร' 'Prime Minister Yingluck Shinawatra'	'นายก ยิ่งลักษณ์' (Prime Minister Yingluck), 'นางสาว ยิ่งลักษณ์' (Miss Yingluck), 'นายกปู' (Prime Minister Puu) *** Puu is the nickname of the Prime Minister
'นิคมอุตสาหกรรมนวนคร' 'Nawanakorn Industrial Estate'	'นิคมนวนคร' or 'นิคมฯนวนคร' (Nawanakron I.E.), 'นวนคร' (Nawanakorn)
'บิ๊กแบ๊ก' 'Big Bag'	'บิ๊กแบ๋ก' or 'บิ๊กแบ๊ค' or 'บิ๊กแบ๋ค' (Big Bag) ** Different vowel, alphabet
'พายุโซนร้อนไหหม่า' 'Haima Tropical Storm'	'พายุไหหม่า' (Haima storm), 'ไหหม่า' (Haima)

After normalization, our data is ready to verify the word boundary by using a word segmentation tool. Then we can generate a precise bag-of-words model which can generate the most efficient significant value of the essential keywords.

C. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a very useful technique to extract keywords from the articles. It is consisted of two main part, Term Frequency (TF) and Inverse Document Frequency (IDF). TF is the value of the frequency of word or term in the article, while IDF is important weight of the words or terms appearance in the considered articles. The words or terms which appear in one article frequently, and rarely appear in other articles should be the significant words or significant terms with high TF-IDF score. While the words/terms which appear frequently in all articles should be a conjunction words or common words that get a very low TF-IDF score. The TF-IDF formulas are shown as (1), (2), (3), and (4).

$$TF(t,d) = \sum_{x \in d} fr(x,t) \tag{1}$$

The $TF(t,d)$ is actually the total number of the term t that appears in the document d, and the $fr(x,t)$ is a simple function defined as (2):

$$fr(x,t) = \begin{cases} 1, & if\ x = t \\ 0, & otherwise \end{cases} \tag{2}$$

The IDF is defined as the logarithm of the number of all documents in a collection divided by the number of documents which the observed term appears (3).

$$IDF(t) - log \frac{|D|}{1+|\{d:t \in d\}|} \tag{3}$$

The 1+ $|\{d: t \in d\}|$ is the number of documents where the term t appears, when the term-frequency function satisfies $TF(t,d) \neq 0$, we apply "1 +" to avoid divide by zero case. Then, the TF-IDF formula is defined as (4):

$$TF - IDF(t) = TF(t,d) \times IDF(t) \tag{4}$$

We also normalized the TF-IDF value in our model to generate the most precise words significant value as shown in (5), and (6).

$$||\vec{X}||_2 = \sqrt{X_1{}^2 + X_2{}^2 + X_3{}^2 + \cdots + X_n{}^2} \tag{5}$$

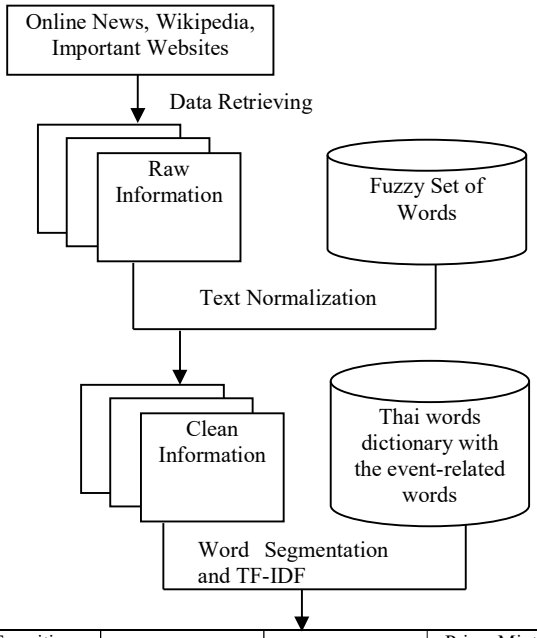
The $||\vec{X}||_2$ is the Euclidean norm factor, L2-normalization. The $X_1, X_2, \ldots, X_n$ are the TF-IDF value of terms (1) to terms (n) in the corpus.

$$TF - IDF_{term(i)} = \frac{X_i}{||\vec{X}||_2} \tag{6}$$

The $TF - IDF_{term(i)}$ is the normalized TF-IDF value of the term (i). While X_i is the TF-IDF value of term (i) and $||\vec{X}||_2$ is the L2-normalization factor of the corpus.

III. Our Approach

There are four main steps in our approach model as shown in Fig. 2.



Transition State \ Words	Haima Tropical Storm	Big Bag	Prime Minter Yingluck Shinawatra
BF	0.5655	0	0.2892
HF	0.0956	0.6599	0.5636
EA	0	0.1954	0.5545
H&R	0.1569	0.3655	0.5199

Bag-Of-Words model with normalized TF-IDF values

Fig. 2 An Approach Model

First, we retrieved all related online news articles and Wikipedia data from the online news websites, Wikipedia websites, and related government's organization websites. The number of articles is shown in Table III.

Table III
A Numbers of Articles

Transition State	Number of Articles
BF	154
HF	789
EA	282
H&R	799

Second, we created the fuzzy set of words and perform the text normalization process.

Then, we have updated our dictionary file for the Thai word segmentation to with the event-related words such as 'บิ๊กแบ๊ก' (Big Bag), 'นายกรัฐมนตรี ยิ่งลักษณ์ ชินวัตร' (Prime Minister Yingluck Shinawatra), 'ศูนย์ปฏิบัติการช่วยเหลือผู้ประสบภัยน้ำท่วม (ศปภ.)' (Flood Relief Operations Center (FROC)), etc., for more precise of Thai word boundary identification.

Finally, we perform the word segmentation process and created the bag-of-words model [5] with the effective normalized TF-IDF score for the four main transition states: BF, HF, EA, and H&R.

IV. Experiment Result

After the bag-of-words model is created, we found the number of all words/terms for each transition states, and unduplicated words/terms for the model as shown in Table IV.

Table IV
A Numbers of Words/Terms in the Bag-of-Words Model

Transition State	Number of all Words/ Terms
BF	1,206
HF	4,014
EA	1,765
H&R	4,523
Total	11,508
Number of unduplicated words/terms in the model	6,632

In the Before Flood period (BF), we found a set of significant keyword terms as shown in Table V.

Table V
A Significant Keyword Terms with Normalized TF-IDF Value in the BF State

Significant Keyword Terms	Normalized TF-IDF value
'ฝนตกหนัก' (Heavy Raining)	0.1828
'น้ำป่าไหลหลาก' (Flash Flood rushing down)	0.0924
'พายุไต้ฝุ่น' (Typhoon)	0.0647
'พายุโซนร้อน' (Tropical Storm)	0.0557
'ศูนย์เตือนภัยพิบัติแห่งชาติ' (National Disaster Warning Center)	0.0281
'ผลกระทบจากการระบายน้ำ' (The effect from water draining)	0.0117

The terms 'Heavy Raining', 'Flash Flood rushing down', 'Typhoon', 'Tropical Storm', 'National Disaster Warning Center' and 'The effect from water draining' are the significant keyword terms which can roughly explained that there is concerned issue about the typhoon and the tropical storm in the BF state. This topic lead to found the effect from water draining of the dam, and the warning from the National Disaster Warning Center. These are the sign of the beginning of the severe flood.

Then, in the Heavy Flood period (HF), we found a set of significant keyword terms as shown in Table VI.

TABLE VI
A SIGNIFICANT KEYWORD TERMS WITH NORMALIZED TF-IDF VALUE IN THE HF STATE

Significant Keyword Terms	Normalized TF-IDF value
'บิ๊กแบ็ก' (Big Bag)	0.0868
'นิคมอุตสาหกรรม' (Industrial Estate)	0.0482
'เครื่องสูบน้ำ' (Feed Pump)	0.0232
'แนวกั้นน้ำ' (Water Barrier)	0.0107
'คันกั้นน้ำพัง' (Dike collapse)	0.0055
'ศูนย์ปฏิบัติการช่วยเหลือผู้ประสบอุทกภัย' (Flood Relief Operations Center)	0.0024

The terms 'Big Bag', 'Industrial Estate', 'Feed Pump', 'Water Barrier', 'Dike collapse', and 'Flood Relief Operations Center' are the significant terms which can illustrated some major event of the industrial estate infrastructure destruction. They also need some help from government related to the stronger Big Bag, or the Water Barrier, and Dike which can be protect their area than the usual one that already collapsed.

Then, in the Ebb Away period (EA), we found a set of significant keyword terms as shown in Table VII.

TABLE VII
A SIGNIFICANT KEYWORD TERMS WITH NORMALIZED TF-IDF VALUE IN THE EA STATE

Significant Keyword Terms	Normalized TF-IDF value
'เปิดประตูระบายน้ำ' (Floodgate opening)	0.0616
'จระเข้' (Crocodile)	0.0252
'การฟื้นฟู' (Restoration)	0.0218
'การกู้นิคมฯ' (Recover the Industrial Estate)	0.0145
'สินเชื่อ' (Credit)	0.0097
'เคลียร์รถบนโทลเวย์' (Remove all parked car on the Toll-way)	0.0092

The terms 'Floodgate opening' can explained the main action when we want to drain the water from the flood area, the main action in EA period. When the terms 'Crocodile' is an awareness to all people that after the flood we can found the dangerous reptiles. And the terms 'Restoration', 'Recover the Industrial Estate' , and 'Credit' can showed us the most important things to do in the EA period, recover the most effective economic area of the country, helping entrepreneur and manufacturer by giving them the credit to recover their company. Finally, the terms 'remove all parked car on the Toll-way' could illustrated the way to protect people owns asset like cars. They need some secure area to move their car to locate and keep away from the flood. Government must consider about this minor issue also.

Finally, in the Help&Recovery period (H&R), we found a set of significant keyword terms as shown in Table VIII.

TABLE VIII
A SIGNIFICANT KEYWORD TERMS WITH NORMALIZED TF-IDF VALUE IN THE H&R STATE

Significant Keyword Terms	Normalized TF-IDF value
'บริจาค' (Donate)	0.2420
'กระทรวงสาธารณสุข' (Ministry of Public Health)	0.0716
'งบประมาณ' (Budget)	0.0569
'เยียวยา' (Remedy)	0.0471
'สุขภาพจิต' (Mental health)	0.0469
'กระทรวงแรงงาน' (Ministry of Labor and Social Welfare)	0.0443

The terms 'Donate' is the common word for H&R period. We need more donation, money, and aid packages, along the flood period and after that also (HF, EA, H&R). The terms 'Ministry of Public Health', 'Mental Health' are the related keyword terms for Government to think about the mental health treatment for people. The severe food could generate more stress to all people. Finally, the terms 'Budget', 'Remedy', and 'Ministry of Labor and Social Welfare' are also the important keywords for the Government to take care all labors with enough budget to prevent the 'Unemployed state' which could generate a worst impact for the country's economic.

V. CONCLUSIONS AND SUGGESTIONS

The bag-of-words model with TF-IDF and text normalization with fuzzy set of words techniques can produce some promising result of the essential keywords extraction from the various data sources related to Thailand severe flood event in year 2011. The model efficiency and accuracy depend on the correctness of the words/terms in the model. Dictionary updating with the event-related words and using fuzzy set of words to clean all same semantic words become the interesting point. More words in both of text database can expand the coverage area of model enhancement. However, this task also be a manual task which consume more man hour to do it. Deep learning and Name Entity Recognition should be the next spot of interesting area to improve the model for NLP, and Decision Support System.

ACKNOWLEDGMENT

We would like to show our gratitude to the Provincial Electricity Authority (PEA), Mr. Akanit Kwangkaew, and Ms.Nungnut Rodsri, for assistance with data source, data mining techniques and comments that greatly improved the manuscript during the course of this research.

REFERENCES

[1] Wikipedia, "2011 Thailand Flood", *https://en.wikipedia.org/wiki/2011_Thailand_floods*, May 2016.

[2] Zhang et al, "An improved TF-IDF approach for text classification", *Journal of Zhejiang University SCIENCE*, pp.49-55.

[3] Provincial Electricity Authority (PEA), "The electricity consumption value of a company in the affected Industrial Estate Area," Jan 2011 – Dec 2012.

[4] L. Leydesdorff , and K. Welbers, "The semantic mapping of words and co-words in contexts", *Journal of Informetrics,* vol.5, pp.469-475.

[5] Y. Zhang, R. Jin, and Z.H. Zhou, "Understanding bag-of-words model: a statistical framework", *International Journal of Machine Learning and Cybernetics*, vol. 1, pp.43-52.