Reclassification of Traffic Crashes Using Traffic Crash Report Data and Keyword Analysis

Harim Jeong[1], Sangmin Park[2], Yongjoo Jun[3], Jungwook Choi[4], Byong-Ho Choe[5], Ki-Hwan Park[6], Ilsoo Yun[7]

[1] Affiliation Information: 206, Wordl cup-ro, Suwon, Gyeonggi-do, 16499, and gkfla0731@ajou.ac.kr
[2] Affiliation Information: 206, Wordl cup-ro, Suwon, Gyeonggi-do, 16499, and stylecap@ajou.ac.kr
[3] Affiliation Information: 169-30, Boorim-ro, Anyang, Gyeonggi-do, 14051, and richard@ilmile.com
[4] Affiliation Information: 135, Jungdae-ro, Seoul, 05717 and jwookchoi@neighbor21.co.kr
[5] Affiliation Information: 17, Hyeoksin 6-ro, Gimcheon, Gyeongsangbook-do, 39660, and byongho.choe@ts2020.kr
[6] Affiliation Information: 17, Hyeoksin 6-ro, Gimcheon, Gyeongsangbook-do, 39660, and foxpark@ts2020.kr
[7] Affiliation Information: 206, Wordl cup-ro, Suwon, Gyeonggi-do, 16499, and ilsooyun@ajou.ac.kr

In Korea, a total of 223,552 traffic crashes happen in 2014. In the traffic crashes, 4,762 people were killed and 337,497 people were wounded. The number of traffic crashes per 100,000 people that was 315 in 1980 increased to 443 in 2014. However, the number of traffic crashes per 10,000 vehicles showed a big drop from 1,615 in 1980 to 94 in 2014. It is very encouraging that the numbers of traffic crashes and fatalities have been decreasing recently. In this study, the types of the violation of the traffic regulations of total 6,023,731 traffic crashes since 1990 in Korea were examined. As a result, the violation of the duty of safe driving occupies 59.1%. The violation of traffic signal (8.4%), the violation of safety distance (7.4%), crossing the center line (6.4%) followed the violation of the duty of safe driving. The violation of the duty of safe driving involves diverse violation activities such as careless driving, aggressive driving, the use of cellular phone, etc. In other words, the scope of the violation of the duty of safe driving is too wide and ambiguous. Thus, it is very hard to establish appropriate alternatives and policies to overcome the violation activities associated with the violation of the duty of safe driving. Therefore, there is a strong need for an in-depth study on reclassifying the actual types of the violation activities which are categorized as the violation of the duty of safe driving. This study was initiated to examine the more detailed and actual types of violation activities categorized as the violation of the duty of safe driving using traffic crash description data recorded in the traffic crash report and keyword analysis. Conclusively, a few types of violations such as drunken driving or the use of cellular phone, etc. are found to be even categorized as the violation of the duty of safe driving even though the violations have a dedicated category. In order to avoid such a confusion, the data written in an unstructured form may be utilized for the reclassification of the types of traffic crashes.

I. INTRODUCTION

Since the introduction of cars, traffic accidents have been one of most important issues in our modern life. Recently 44 persons have been killed or wounded in a traffic crash happened in Young-dong expressway in Korea due to drowsy driving. In order to control such critical traffic crashes, it is very important to establish proper alternative which should be developed based on the right reason of the traffic crash under study.

In Korea 223,552 traffic crashes reported to the national police agency in 2014. Only the traffic crashes are recorded in the statistics of traffic crashes published by the national police agency. The statistics summarizes the traffic crashes by month, hour, day, type of cars involved in crashes, type of roads, and type of violation, etc.

In depth, traffic crashes by the type of violation are categorized into twenty types, including overwork, speeding, passing method violation, center line violation, traffic signal violation, safe distance violation, suspending violation, unfair turn, failure of priority compromise, failure of direction compromise, safe driving violation, etc.

However, the violation type of 56.5% of total 223,552 traffic crashes are designated as "safe driving violation." Here, when the police officer in charge of a traffic crash cannot find a specific and major violation involved in the traffic crash, the police officer may categorize the violation type of the traffic crash as "safe driving violation." In other words, the violation types of many traffic crashes are designated as very ambiguous type, which is not helpful for finding the alternative to avoid such a traffic crash. Therefore, there has been a strong desire to subdivide the traffic crashes categorized as the safe driving violation into a specific violation types.

To this end, this study was initiated to identify important information which is able to dismantle the traffic crashes categorized as the safe driving violation using traffic description data and text mining technique. For the analysis, the traffic crashes data happed in 2014 within Seoul managed by the national police agency are used. The traffic crash data consists of structured data and unstructured data. The structured data includes date, time, day, location, personal information, car information, type of violation, etc. In addition, the unstructured data means the description which was written by a police officer in charge of the traffic crash in order to explain the conditions and situation of the crash in detail.

In this study, the text mining technique was utilized to extract meaningful words out of the description recorded in the traffic crash data. This study was conducted according to the following process.

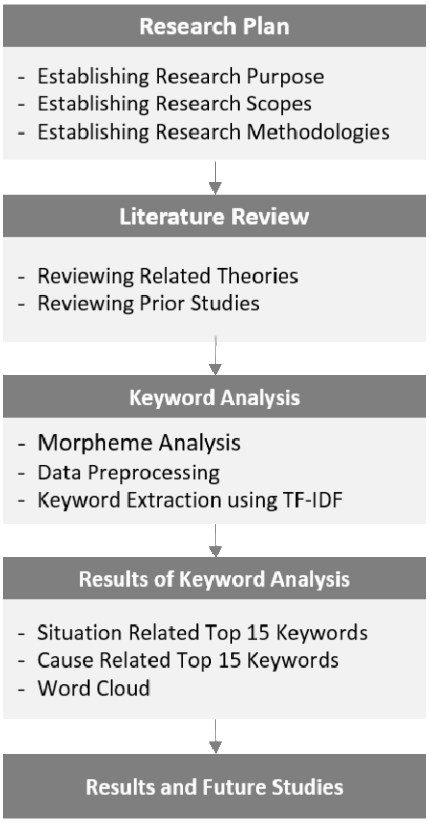
Research Plan
- Establishing Research Purpose
- Establishing Research Scopes
- Establishing Research Methodologies
Literature Review
- Reviewing Related Theories
- Reviewing Prior Studies
Keyword Analysis
- Morpheme Analysis
- Data Preprocessing
- Keyword Extraction using TF-IDF
Results of Keyword Analysis
- Situation Related Top 15 Keywords
- Cause Related Top 15 Keywords
- Word Cloud
Results and Future Studies

Fig. 1 Study Process

II. TEXT MINING

A program developed for this study was utilized to conduct the text mining for the unstructured description data explaining the details of each traffic crash in the traffic crash data. Total 22,644 traffic crash data are used in this study. The traffic crash happened in Seoul between January 1 and December 31, 2014.

The traffic crash data includes date, time, day, local government code, area code, HIT based area code, location (address if possible), description of the crash, type of violation, etc. The description says the details of the traffic crash based on the five W's.

The information recorded by the police in the traffic crash report consists of structured and unstructured data. The structured data means the information written in the fixed format in the report, including age, sex, weather, location, vehicle type, etc. Whereas, the unstructured data is written in the report form in order to explain the additional and detailed description on each traffic crash which may be hard to be recorded in a structured form. The unstructured data is technically hard to be analyzed. However, due to the recent development in big-data analysis, the unstructured data is able to be analyzed using the keyword analysis technique. The keyword analysis focused on identifying important words within the sentence recorded in an unstructured-form in the report. For the keyword analysis, keyword identification methodologies were first reviewed and then important keywords were extracted. With the extracted keywords, the types and frequencies of the keywords were analyzed in order to determine the content features in the description of traffic crashes.

For the text mining, the description including the five W's in an unstructured format should be dismantled into morphemes, which mean the smallest meaningful grammatical unit in a language. And then, potential keywords are selected among the entire morphemes where the morphemes meets the criteria for the keywords. The selected potential keywords are enrolled in a temporary dictionary for the test mining. Through this process, a description in the traffic crash data can be abbreviated with a few morphemes. However, the entire morpheme in the temporary diction cannot be the final influential keywords as a result of text mining. Therefore, the final influential keywords out of the entire morpheme in the temporary dictionary are selected using the TF-IDF model, which is able to provide a score explaining the level of influence of each keyword using the frequency of appearance and the number of descriptions including the keyword.

III. RESULT OF TEXT MINING

3.1 Top Fifteen Influential Keywords Related with Crash Situation

In this study, the final influential keywords were identified through the preprocess, morpheme analysis and keyword analysis. As a result, the keywords related with crash situation and crash reason were identified.

First, the top fifteen keywords related with cash situation based on the score from the TF-IDF model are shown in Table 1.

Table 1. Top Fifteen Keywords related with Crash Situation

Ranks	Keywords
1	Right-side
2	Left-side
3	Bumper
4	Lane 2
5	Lane 1 (the most inside lane in Korea)
6	Crahs
7	Impact

8	Lane 3
9	Crossing
10	Rear-end
11	Pedestrian
12	Lane 4
13	Contact
14	Fault
15	Right turn

The word cloud based on the keywords related with crash situation is presented in Fig. 2

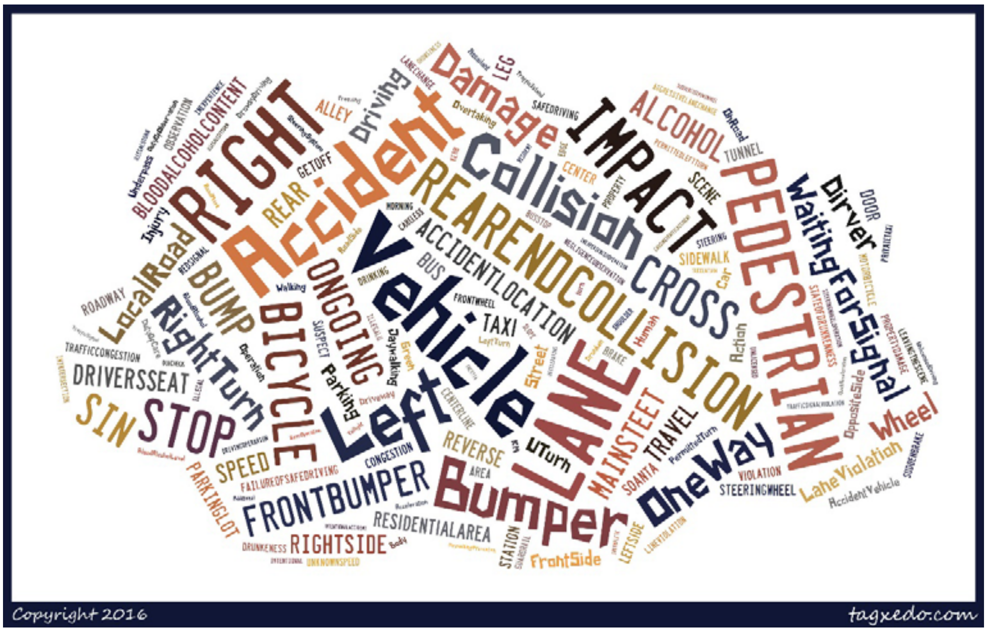

Fig. 2 Word Cloud for Keywords related with Crash Situation

3.2 Top Fifteen Influential Keywords Related with Violation Type

As mentioned earlier, the majority of traffic crashes are categorized as “safe driving violation” in terms of the type of traffic regulation violation. The safe driving violation includes all kinds of ambiguous violation types except for typical important violation types, such as speeding, violation of traffic signal, drunk driving, etc.

Under such a circumstance, this study was initiated to conduct in-depth examination of traffic crashes categorized as the safe driving violation in terms of crash reason using text mining.

The results of text mining was surprising. First, the keywords related with alcohol, including “alcohol”, “blood-alcohol level”, “drunk driving” were found within the top fifteen keywords related with violation types. The other keywords are related with driving maneuver, including “lane change”, “U-turn”, “keeping eyes forward”, “sudden braking”, etc. Also “jaywalking” was found in the keywords. The fifteen keywords are listed in Table 2.

Table 2. Top Fifteen Keywords related with Violation Types

Ranks	Keywords
1	Alcohol
2	Lane change
3	Blood-alcohol level
4	U-turn
5	Keeping eyes forward
6	Jaywalking
7	Sudden braking
8	Drunk driving
9	Negligence
10	Baking
11	Passing
12	Maneuver
13	Blood-alcohol
14	Driving
15	Steering

The word cloud based on the keywords related with violation type is presented in Fig. 3

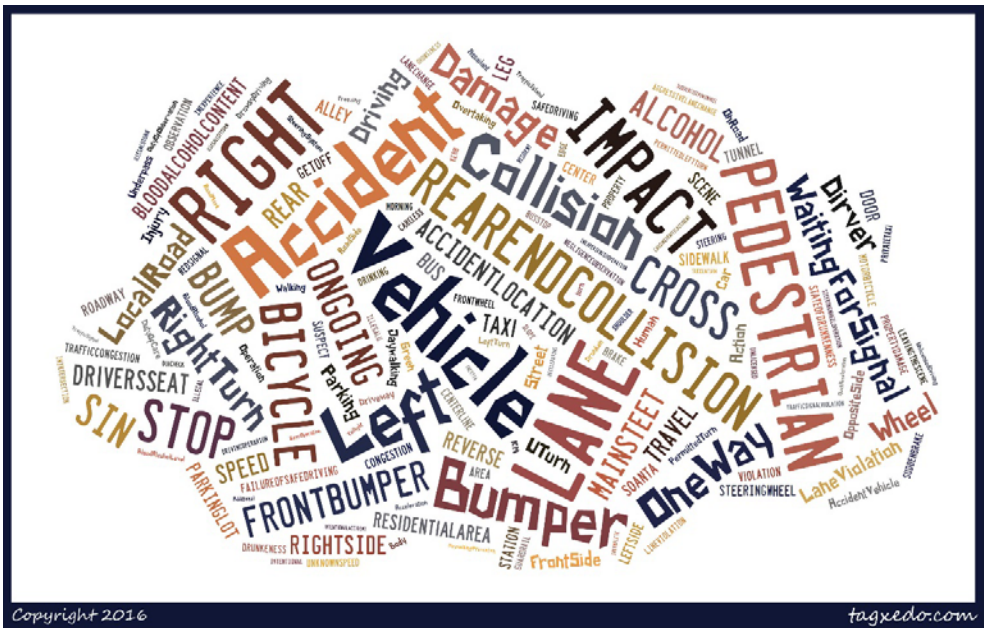

Fig. 3 Word Cloud for Keywords related with Violation Type

II. Conclusions

This study was initiated to emphasize the necessity of re-categorization of traffic crash conducted by the national police agency. This is because the majority of traffic crashes are categorized as “safe driving violation” even though the crash is happened due to a specific and major violation which cannot be ignored as just safe driving violation.

In order to certify the assumption, in this study the description recorded in traffic crash data for the entire traffic crashes reported to the national police agency were analyzed using the text mining technique. The text mining is well-known big-data analysis technique, especially justified to unstructured data such as the description for individual traffic crash.

As a result, many keywords related with alcohol, including “alcohol”, “blood-alcohol level”, “drunk driving” were found within the top fifteen keywords related with violation types. This means that many traffic crashes categorized as safe driving violation are involved in drunk driving.

REFERENCES

1. Pollak, S., Coesemans, R., Daelemans, W. and Lavrac, N (2011), Detecting Contrast Patterns in Newspaper Articles by Combining Discourse Analysis and Text Mining, International Pragmatics Association, Vol. 21, No. 4, pp. 647-683
2. Zhao, W., X, Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H. and Li, X. (2011), Proceedings of the 33rd European conference on Advances in information retrieval, pp. 338-349