A Proposal for Construction of a Floor Noise Map Using Web Scraping

Yang, Jong Hyeon*, Kim, Jung Ok**, Yu, Kiyun*

* Dept. of Civil & Environmental Engineering, Seoul National Univ.
** Institute of Construction and Environmental Engineering, Seoul National Univ.

Extended Abstract

In Korea, the proportion of multi-unit dwellings was 60% in 2014 and this proportion has been increasing annually. As a result, the number of civil complaints regarding floor noise have increased steadily. According to the Center for Floor Noise (CFN) operated by the Korea Environment Corporation, the number of civil complaints filled annually exceeds 15,000, with 15,455 in 2013, 16,370 in 2014, and 15,419 in 2015. The CFN provides services to tackle floor noise problems. However, as this center has only been operational on a national scale since June 3rd, 2014, it is hard to identify the national status of floor noise. This research aims to develop a floor noise map by city and province and analyze the spatial distribution of noise using data collected from March 2012 to June 2016 using web scraping.

The flow chart for the construction of a floor noise map is shown in *Fig. 1* and we used R 3.2.5 for our experiment. First, we searched for posts on Naver Cafe, the most popular portal site in Korea, with the key words “Floor Noise” to develop the floor noise map. The search period was limited to May, 2012 to June 2016 and the total number of search results was 116,465. We collected the title, contents of the post, and name of the cafe by using web scraping on the search results and excluded data with duplicate contents. Second, we conducted natural language processing on the 7996 data sets with no duplicate contents. We removed the URL, English, special characters, consonants and vowels in Korean. Then, we conducted morphological analysis with SimplePos22 in the KoNLP package. Third, we extracted the spatial position from the name of the cafe. We used the road name address basic map (20160809 version) downloaded from the National Spatial Information Clearinghouse as a reference. We investigated the name of the cafe to determine whether the administrative district, Gun or Gu, are included in the data. For duplicated Gun or Gu, we investigated whether the name of the province or city was included, and if it was not, we concluded that the data had no spatial location. We then extracted the data with the Gun or Gu included in the name of the cafe. The rest of the data were excluded from the research. Third, we examined the 2,565 posts with spatial location to check whether they were about floor noise. We divided the posts into 2 classes – a positive class and a negative class - using a Support Vector Machines(SVMs). We applied linear SVMs and used Term Frequency (TF) as term weights. We randomly extracted 826 samples and divided them into 550 training sets and 276 test sets. Finally, we constructed the floor noise map with 511 data sets classified as the positive class.

SVMs have been proven to outperform several other learning algorithms for text categorization. The procedure for text categorization with SVMs is explained in *Fig.2*. We constructed the data with results for NLP as in *Table 1*. With their ability to generalize well

LBS 2016
13th Conference on Location-Based Services
Vienna, 14–16 November 2016

Published in “Proceedings of the 13th International Conference on Location-Based Services”, edited by Georg Gartner and Haosheng Huang, LBS 2016, 14-16 November 2016, Vienna, Austria.

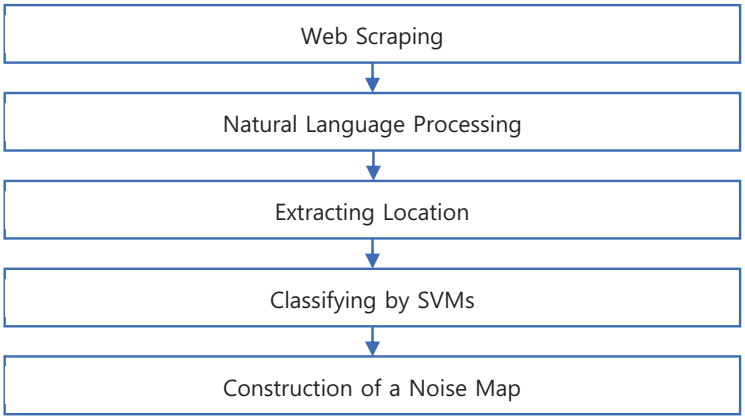

Figure 1. Flow chart

in high dimensional feature spaces, SVMs eliminate the need for feature selection (Joachims 1998). Therefore, we eliminated the feature with Document Frequency (DF) 1, to avoid being influenced by information for specific documents. We used the Linear SVMs because they provide good classification accuracy, are easy to learn and quick to classify new instances (Dumais et al. 1998). Moreover, we used TF for term weighting because it performs well (Lim 2000). We trained linear SVMs with a cost of 100 and γ 1.0. Finally, we tested SVMs with 276 test sets and evaluated performance with the $F_1 score$ in *Eq.(1)*.

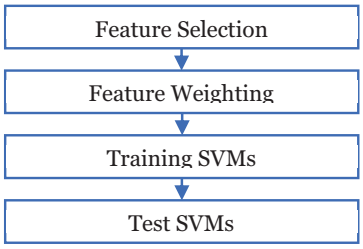

Figure 2. Procedure of SVMs

Table 1. Data type for SVMs

Type	Feature
Data1	NC (common noun)
Data 2	NC, PA (adjective)
Data 3	NC, PV (verb)
Data 4	NC, PV, PA
Data 5	Whole parts of speech in SimplePOS22

Table 2. A contingency table

	YES is correct	No is correct
Assigned YES	a	b
Assigned No	c	d

$$F_1 \ score = 2\frac{p*r}{p+r} \quad (1)$$

$$\text{where } p = \frac{a}{a+b} \ and \ r = \frac{a}{a+c}.$$

The classification results are in *Table. 3*. Data 4 showed the best results and therefore we used the data set in our experiment. Before using data 4 in our experiment, we tested whether the value of cost and γ affects the results. We changed γ from 0.5 to 1.0 in increments of 0.1 and changed cost from 100 to 1,000 in increments of 100. The result was a value of γ 0.8 and cost 100. However, this value showed the same results as γ 1.0 and cost 100. By using data 4,

we obtained 511 data points classified as positive. We analyzed the contents of the posts and constructed the noise map. By analyzing the contents of the posts, we could determine that not only does the victim post, so does the inflictor and the people who are suspected to be the inflictor. In addition, we analyzed the floor noise map in *Fig. 3* and determined that many people complain of floor noise in Sejong-si, Buchon-si, Ilsan-gu, Osan-si, Suji-gu, Paju-si, Gumi-si, and Gimhae-si.

Table 3. SVMs results

	Data 1	Data 2	Data 3	Data 4	Data 5
p	0.742	0.779	0.736	0.810	0.779
r	0.712	0.760	0.764	0.763	0.747
$F_1\ score$	0.727	0.770	0.750	0.786	0.763

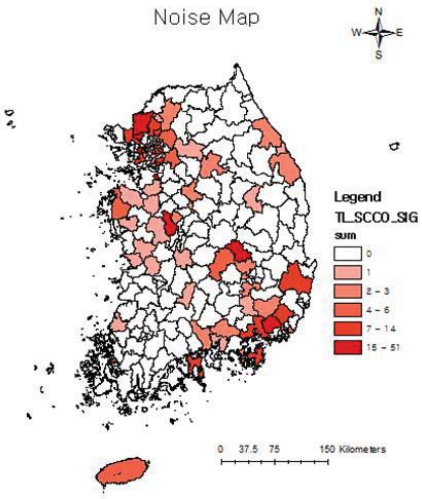

Figure 3. Noise Map

In this research, we developed a nationwide floor noise map using web scraping. We found that the combination of NC, PV, and PA showed the best results and the corresponding $F_1\ score$ was 0.786, which is quite high. Moreover, we found that not only victims posted civil complaints. Inflictors and people who were suspected to be inflictors also posted in the cafe. Finally, we identified the district with many complaints of floor noise. There is a need for further research examining the relationship with socioeconomic values and the frequency of the post.

Acknowledgement

This research, "Geospatial Big Data Management, Analysis and Service Platform Technology Development," was supported by the MOLIT (The Ministry of Land, Infrastructure and Transport), Korea, under the national spatial information research program supervised by the KAIA (Korea Agency for Infrastructure Technology Advancement) (16NSIP-B081011-03).

References

Dumais S, Platt J, Heckerman D, Sahami M(1998) Inductive Learning Algorithms and Representations for Text Categorization. In Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management (Bethesda, MD, 1998), p.148–155

Joachims T (1998) Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. Proceedings of the 10th European Conference on Machine Learning, April 21-23 p.137-142

Lim H Y (2000) An Experimental Study on Text Categorization using an SVM Classifier, Master's thesis, Yonsei University, Seoul, Korea, 81p