Acquisition of Typical Occupancy Schedules for Commercial Buildings from Social Networks

Xing Lu
Department of Mechanical Engineering
The University of Alabama
Tuscaloosa, AL, USA
xlu25@crimson.ua.edu

Fan Feng
Department of Mechanical Engineering
The University of Alabama
Tuscaloosa, AL, USA
ffeng2@crimson.ua.edu

Zheng O'Neill
Department of Mechanical Engineering
The University of Alabama
Tuscaloosa, AL, USA
zoneill@eng.ua.edu

ABSTRACT

Occupancy behavior in buildings has been a hotspot of research as building systems become more sophisticated. Traditional sensing technologies suffer from high costs, inevitable sensor errors, and scalability issues and thus are not widely implemented in buildings.

In this notes paper, two different approaches for extracting the typical occupancy schedules for the input to the building energy simulation are explored based on the data from social networks.

The first approach is to use text classification algorithms to identify whether people are present in space where they are making posts on social networks. To achieve this, word embedding and machine learning algorithms for the classification are used. On top of that, we could extract the typical occupancy schedules by assuming certain people counting rules. The second approach is to utilize the processed GPS location tracking data provided by social network giants such as Facebook and Google Map. Web scraping techniques are used to obtain the data and extract the building typical occupancy schedules.

Two preliminary case studies demonstrate these two approaches as a proof of concept using a museum building, the Art Institute of Chicago. The results show that the extracted building occupancy schedules from different social network data sources (Twitter, Facebook, and Google Map) share a similar trend but slightly distinct with each other, which requires more explorations and further validations and corrections. However, the methodology and promising results from this preliminary study will lay the foundation of the occupancy sensing through the social network data mining, which aims to provide another data source for occupancy sensing in buildings at the building level. This will provide another alternative to estimate the occupancy at the community- and urban-scale.

UrbSys '19, November 13–14, 2019, New York, NY, USA

ACM ISBN 978-1-4503-7014-1/19/11…$15.00
https://doi.org/10.1145/3363459.3363537https://doi.org/10.1145/1234567890

CCS CONCEPTS

• Applied computing → Computers in other domains
• Applied computing → Operations research

KEYWORDS

Building Occupancy; Data Acquisition; Machine Learning; Machine Learning;

1 Introduction

Occupancy behavior in buildings plays a vital role in managing building energy performance and has become a hotspot of research as building systems are more sophisticated [1]. The occupancy behavior largely influences modeling of building energy performance, design of the future building system, and the operations of intelligent building systems [2]. In the ASHRAE HVAC Application Handbook [3], the occupancy schedule in commercial buildings is in the form of a static schedule for the HVAC design and sizing. However, it is well known that the static occupancy schedule may cause the discrepancy and different buildings of the same type may have distinct occupancy pattern. To accurately acquire the occupancy information, various types of occupancy sensors (such as image-based, threshold and mechanical, motion sensing, and radio-based sensing) are commonly used in

buildings [4]. However, those sensing technologies suffer from high cost, inevitable sensor errors, and scalability issues. Thus, they are not widely implemented in buildings. With the vast development of information technologies in the era of the internet-of-things (IoT), occupant sensing and data acquisition are not limited to traditional approaches.

The prevalence of social networks could provide myriad of publically available social media data that might contain occupancy information in space and in time. Only a few studies targeted at the occupancy estimation for the building level. The Population Density Tables (PDT) project by Oak Ridge National Laboratory estimated the ranges for average day and night population density for over 50 building types using Bayesian learning model with different open source data [5]. Stewart et al. [6] proposed a social network unit occupancy model to extract the social media-based occupancy curve for a museum during its operating hours. The uncertainty of the model is also quantified. Sims et al. [7] applied social media data to high-resolution mapping of a special event population. Twitter posts and Facebook check-ins were calculated for the Game Day at the University of Tennessee Knoxville. Population distributions for game hours and nongame hours of the game day were modeled using social media data. It is noted that it used a linear relationship to describe the event population with the social media activity. Bentz et al. [8] designed a thermostat in which the setpoint could be adjusted based on by the expected occupancy and the social media activity. These studies indicate the feasibility of extracting the building occupancy information from social networks, but they do not target at the building or urban energy modeling.

In this notes paper, two different approaches for extracting the typical occupancy schedules of commercial buildings are explored based on the data from social networks. The first approach is to use text classification algorithms to identify whether people are present in space where they are making posts on social media. On top of that, we could extract the typical occupancy schedules by assuming certain people counting rules. The second approach is to utilize the processed GPS location tracking data provided by social network makers such as Facebook and Google Map. Web scraping techniques are used to obtain the data and extract the building typical occupancy schedules. In the following sections, we will use a museum building to show this concept and methodology.

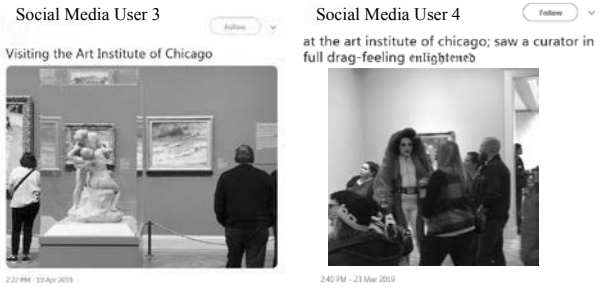
2 Case Studies

The Art Institute of Chicago is selected as the case study building. Founded in 1879 and located in Chicago's Grant Park, it is one of the oldest and largest art museums in the United States. It opens daily from 10:30 a.m. to 5:00 p.m. except on Thursdays until 8:00 p.m. In the subsections, two different methods are illustrated to extract typical occupancy schedules for this building.

2.1 Case 1: Extract Typical Occupancy Schedule from Twitter

In this approach, the collected raw data are the posts from the social media of Twitter, as illustrated in Figure 1. Apparently, we could deduct the people presence from semantics of the text. Therefore, the idea behind this approach is to use text classification algorithms to identify whether people are present in space where they are making posts (Tweets in this case) on social network (Twitter). To achieve this, word embedding and machine learning algorithms for classification will be used. On top of that, we could extract the typical occupancy schedules by assuming certain people counting rules.

a) Geo-tagged posts b) Check-in posts



c) Implicit geotagged posts example 1 d) Implicit geotagged posts example 2

Figure 1: Sample of the collected data in Case 1.

The methodology for this approach involves four key elements: data collection and pre-processing, feature generation, classifier formulation, and result evaluation, as illustrated in Figure 2.

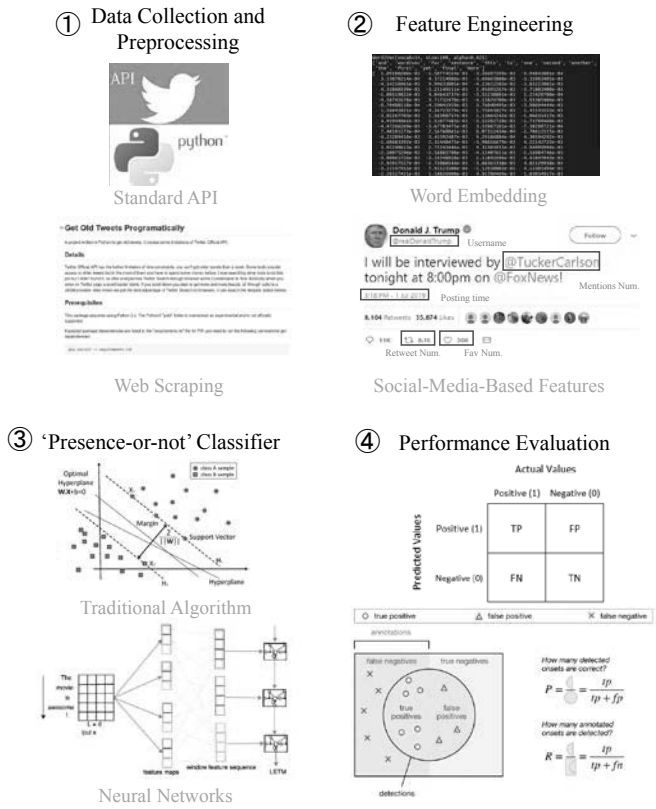


Figure 2: Framework of Case Study 1.

In the step of data collection and preprocessing, the history posts could be collected either through social network official APIs [9] or web-scraping tools such as GetOldTweets [10]. In this case, we collected all the available ~30,000 history posts with search key words like 'art institute chicago' from December 2016 to June 2019 using the GetOldTweets approach. We manually labeled the latest 3,000 history posts which indicated whether the user was present or not. Then we lowered the case of the Tweets, conducted the tokenization, and removed the stop words. For the feature selection, we generated the word embeddings using pre-trained Word2Vec [11], where each word is presented by a high dimension vector. The dimension of the vector space is 300. For each Tweet, the aggregated vector is weighted by the value of the TF-IDF. In addition, we also considered social media content-based features, such as whether the posted time is within opening hours, whether the username of users contain the keywords like 'art', whether the domain name in the Url contains check-in apps, hashtags counts, mentions counts, favorite/like counts, retweet counts, etc. For the classifier formulation, we tested the performance of different categories of classifiers. We selected SVM (a traditional classifier), random forest (an ensemble classifier), and shallow neural network (that contains three types of layers). The training/testing data ratio is 8:2. The performance metrics of different classifiers are listed in Table 1 in terms of accuracy, precision, recall, and F1-score. It can be seen that the accuracy of the different classifiers is in a similar range with the random forest and neural network slightly being higher.

Table 1. Summary of the performance metrics of different classifiers

Performance Metric	SVM	Random Forest	Shallow Neural Network
Accuracy	0.8485	0.9091	0.9091
Precision	0.6000	0.8333	0.7500
Recall	0.8571	0.7143	0.8571
F1-score	0.7059	0.7692	0.8000

Using the formulated classifier, we labeled all the historical ~30,000 Tweets. We added up all the positive labeled data in the same time slot (an hour) on the same day type using these historical Tweets. We assumed that people would stay for three hours before and after the posting time based on the visit duration from Google Map App. In this way, we obtained the weekly typical occupancy schedules to be used as inputs for the building energy models, as shown in Figure 3, which shows two different types of typical occupancy daily schedules. It is noted that this is only a preliminary result, which needs further and more comprehensive investigations to alleviate the uncertainties associated with assumptions.

2.2 Case 2: Extract Typical Occupancy Schedule from Facebook and Google Map

The second approach is to utilize the processed GPS location tracking data provided by social network makers such as Facebook and Google Map. Web scraping techniques are used to obtain the data and extract the building typical occupancy schedules.

Figure 4 shows the sample of "Popular Times" by Google Map and "Popular Hours" by Facebook. The principle behind these types of data lies in that these social network

giants use aggregated and anonymized data from users who have opted in to share their real-time location. These companies also have Points-of-Interest (POI) building footprints (polygons) which determine the location, shape, and size of a place. Based on these data, machine learning algorithms are used to join the GPS data against the building footprints to derive the occupancy information.

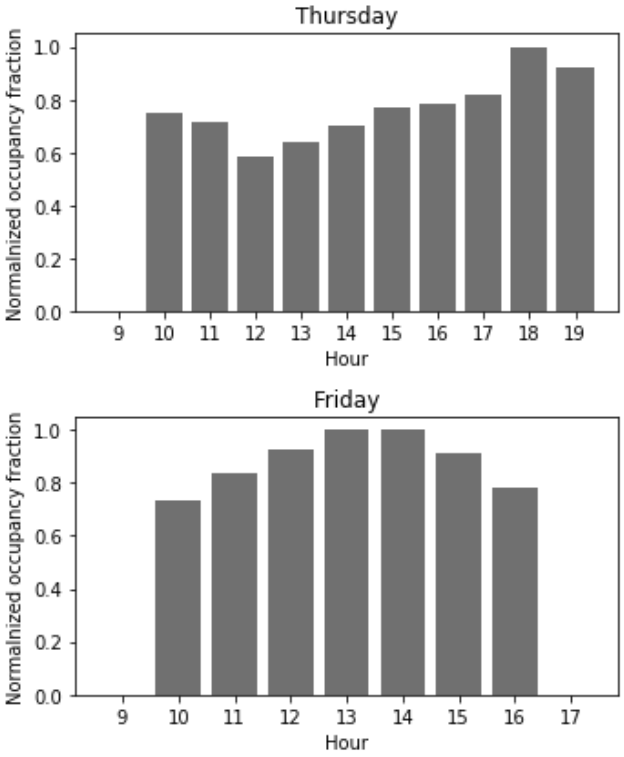


Figure 3: Preliminary typical occupancy schedules for two days extracted from social media data.

a) Facebook Popular Hour b) Google Map Popular Times

Figure 4: The sample of "Popular Times" by Google Map and "Popular Hour" by Facebook.

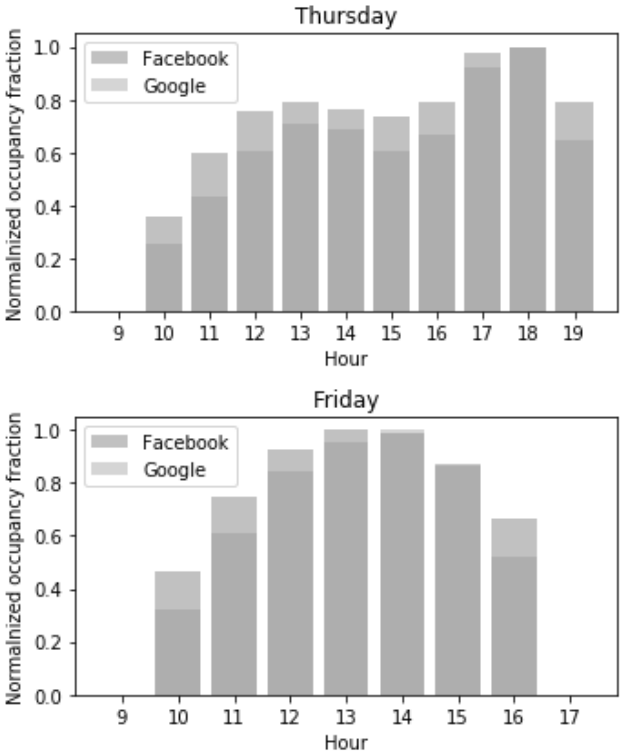


Figure 5: Comparisons of typical occupancy schedules for two days extracted from Facebook and Google Map.

We scraped and normalized the data from their websites. Figure 5 depicts the bar chart of the extracted typical building occupancy schedules from Facebook and Google Map. These two bars have a similar trend but there still exist deviations for the two occupancy schedules. We still need to assess the fidelity of the extracted data as there might be several users who would be visiting who do not have Google Maps or location history enabled. Compared with the extracted schedule in Case 1 for the same museum, these two schedules in Case 2 resemble each other more. However, we can see that all three schedules share a similar trend. The schedule from Facebook Popular Hour has a more similar trend with one in Case 1 than the schedule from Google Map.

3 Conclusions and Future Work

In this paper, we explored two approaches to extract the typical occupancy schedules for the input to the building energy simulation based on the data from social networks. As a preliminary study, the Art Institute of Chicago is selected as the case study building. The first approach uses text classification algorithms to identify whether people are

present in the space where they are making posts on social media. On top of that, we could extract the typical occupancy schedules with assumed people counting rules. The second approach utilizes the processed GPS location tracking data provided by social network makers such as Facebook and Google Map. Web scraping techniques are used to obtain the data and extract the building typical occupancy schedules. It can be seen from the preliminary results that the extracted building occupancy schedules from different data sources (Twitter, Facebook, and Google Map) share a similar trend but slightly distinct with each other which requires further validation and corrections.

The future work includes improving, validating, and correcting the occupancy schedule estimation from two proposed approaches for building energy performance modeling. We will investigate a lot of uncertainties arising from the data from the social network. For example, the ratio of users to non-user of social media should be observed due to the age bracket of the social network users. There also exists cyber associated risks that attackers might create fake accounts and publicly tweet their presence or absence and the approach is not robust against such adversarial scenarios. As another example, the posts may occur within the normal hours of operation but not occur at the time of occupancy, which results in inaccurate timestamp data. These uncertainty will all lead to inaccurate extracted occupancy schedules.

Another direction we would like to study is to evaluate the value of the extracted building occupancy schedules. A data-driven building energy model for a university museum is being constructed to see whether additional feature regarding the occupancy at the building level will facilitate the improvement of the accuracy and the fidelity of the building energy model. Furthermore, such occupancy information at the building level will be incorporated with the urban-scale community and city energy modeling.

REFERENCES

[1] T. Hong, S. Taylor-Lange, S. D'Oca, D. Yan, and S. Corgnati, "Advances in research and applications of energy-related occupant behavior in buildings," Energy and Buildings, 116, pp. 694-702, 2016.

[2] D. Yan, W. O'Brien, T. Hong, X. Feng, H. B. Gunay, F. Tahmasebi, and A. Mahdavi, "Occupant behavior modeling for building performance simulation: Current state and future challenges," Energy and Buildings, 107, pp. 264-278, 2015.

[3] M. Owen, "ASHRAE Handbook: HVAC Applications," Atlanta: American Society of Heating, 2016.

[4] B. Dong, V. Prakash, F. Feng, and Z. O'Neill, "A review of smart building sensing system for better indoor environment control," Energy and Buildings, 199, pp. 29-46, 2019.

[5] R. Stewart, M. Urban, S. Duchscherer, J. Kaufman, A. Morton, G. Thakur, J. Piburn, J. Moehl, "A Bayesian machine learning model for estimating building occupancy from open source data," Natural Hazards, 81(3), pp. 1929-1956, 2016.

[6] R. Stewart, J. Piburn, E. Weber, M. Urban, A. Morton, G. Thakur, B. Bhaduri, "Can Social Media Play a Role in the Development of Building Occupancy Curves?" In Advances in Geo-computation, pp. 59-66, 2017.

[7] K. Sims, E. Weber, B. Bhaduri, G. Thakur, D. Resseguie, "Application of social media data to high-resolution mapping of a special event population," In Advances in Geo-computation , pp. 67-74, 2017.

[8] J. Bentz , B. Rigg, T. Gillette, T. Tasker, T. McCune, J. Uerkvitz, S. Atchison , D. Middleton, A. Noor, "Thermostat with occupancy detection based on social media event data," United States patent application US, 15, pp. 260-295, 2017.

[9] TwitterDeveloper, "Standard search API," Available online: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html, 2019.

[10] J. Henrique, "A project written in Python to get old tweets which could bypass some limitations of Twitter official API," Available online: https://github.com/Jefferson Henrique/GetOldTweets-python.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality," Paper presented at the Advances in neural information processing systems, 2013.