A Method for Crawling and Analyzing Web-based Real Estate Listings

Huy Nguyen
University of Science
Ho Chi Minh, Vietnam
nhphuy@apcs.vn

An Vo
University of Science
Ho Chi Minh, Vietnam
vtan@apcs.vn

Vu Nguyen
University of Science
Ho Chi Minh, Vietnam
nvu@fit.hcmus.edu.vn

ABSTRACT

There are hundreds of thousands of real estate listings generated daily on the Internet, and they are created by brokers or property owners on websites specialized in housing trading. These housing listings are not fully utilized and they are usually overlooked because they are used only for advertising properties. Instead, they cannot reflect the market activities or the current state of Vietnam real estate market. To solve the this problem, the authors propose a method to use these data sources and offer a novel insight into the housing market across Vietnam. We collected housing items from various real estate websites, then we cleaned, analyzed, and visualized the data by a website. In this study, we apply this method to the real estate market in Vietnam, and we analyzed the web-based real estate listings of Ho Chi Minh city to help people have a deep insight into the housing market of this city.

CCS CONCEPTS

• **AI and Big Data Analytics → Data mining**; *Data crawling*; *Data analyzing*;

KEYWORDS

Data crawling, data analyzing, data visualization, real estate listings, web scraping.

INTRODUCTION

Today, most of the real estate activities usually take place on websites specialized in real estate field. However, these websites cannot offer the customers a view about the current state of the real estate market. Instead, they are just commercial websites used for advertising properties, and their purpose is mainly to allow people to start real estate transactions such as posting, searching, buying or renting housing items. Regardless of the above drawback, these websites still have many promising applications because they have a large volume of real estate listings. There are thousands of housing items generated on these housing websites everyday, and they are usually overlooked because of their inability to provide reliable estimates or information about the current housing market.

To make use of these real estate listings in an efficient way, we built a system that automatically collected, cleaned, and analyzed hundreds of thousands of real estate items from various resources. Then, a website was developed to visualize the processed data by many different graphs. Each of these graphs can reveal an unique insight into the housing markets such as the price trend over the year, the distribution of a property, the popularity of a house type, or the average price in a specific area.

There have been some efforts to solve the problem, and they are presented in different ways. Take the research of Geoff Boeing and Paul Waddell in 2016 as an example. By using the Craigslist rental listings, they provided the American a deep intuition into the rental housing market across the United States. His paper successfully presented many trends as well as the spatial patterns of cities through the nation. In addition, it also introduced a strong methodology to deal with big data (data collecting and cleaning)[2]. However, the study still has some drawbacks. Their way of removing extreme value only by removing a percentage of largest and smallest values is not robust and there is no guarantee that it would not remove actually sensible data. Furthermore, instead of covering all transaction types (sale and rent), the research only focuses on rental and this is definitely a shortcoming. Still, this paper inspired us to apply a similar process to extract insightful views into Vietnam's real estate market.

There are two objectives for this study. The first is to share with housing scholars and practitioners an original methodology for collecting and analyzing housing data. The method can collect user-generated housing items from different real estate websites. Afterward, they are analyzed by different steps to remove duplicate and unreasonable items before. The second is to present different insights into the housing market based on these web-based real estae listings and their implications for the housing market. These available dataset is very comprehensive to examine the real estate market in Ho Chi Minh city, and they can provide much richer detail about the housing market in the city than other publicly available data sources .

We begin by providing a brief background for our work and other related work to solve the problem. Next, we explain our methodology for collecting this unique data set, cleaning it, validating it, and analyzing it. Then, we present our findings and discuss the practical implications of these housing insights-and urban big data generally-for planners. We conclude with a discussion of the generalizability of our methodology, and the prospects and challenges of our system for real estate business in Vietnam.

RELATED WORK

New Approaches In Discovering US Rental Market: Craigslist Data Analysis

Even though local residents usually have a longstanding tradition of ownership-occupancy, United States has experienced a dramatic rise in the significance of housing rental market recent years, with rental households increasing from 31% (2004) to 35% (2012), and presenting forty-three million households in 2013 (Joint Center for Housing Studies 2013). However, most of these rental activities take place mainly in privacy and the estate data is often inconsistent. Meanwhile, data from commercial sources (newspaper, websites, and television) tend to cover large apartments, and the data is not reliable in some respects. With the development of Information Technology, most of rental activities has moved online by using rentals websites. In that situation, Craigslist plays an important role in the rental housing market because of its considerably enormous data source, where hundreds of thousands of rentals are generated each day, in Figure 1.

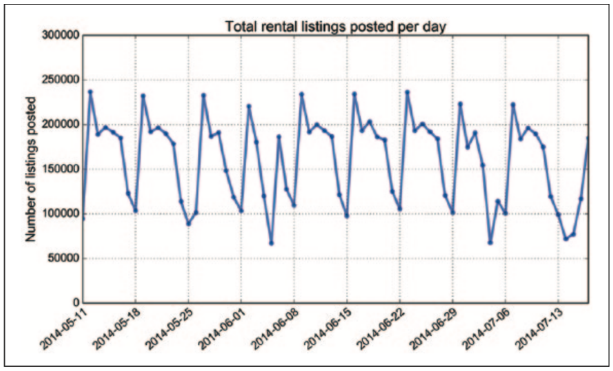


Figure 1: Count of rental listings posted per day from Craigslist filtered data set[2]

Unfortunately, the potential of Craigslist rental housing lists is still not explored and it would be a waste of valuable data. To solve the problem, Geoff Boeing and Paul Waddell from University of California conducted a research into the rental housing market in US based on eleven million rental listings of Craigslist. The research provided a new insight about rental housing market in mega cities (e.g. New York, San Francisco) and some detailed information which could not be found publicly. During this research, the authors had two targets. Firstly, due to the fact that Craigslist rental property data has an important place in the US rental estate market, the authors wanted to find the trend in the data and provide users an overview of the US rental market, in Figure 2. Secondly, there is no doubt that dealing with big data is not an easy task, and the authors wished to introduce the methodology of processing the data. In terms of technical aspect, the authors collected data from Craigslist (eleven million rental listings over the United States), afterwards, the data cleaning process consisted of web scraping, duplicate record detection, and outlier removal. In the data analysis section, the author analysed and presented data characteristics such as house type distribution trend, price distribution, and distribution per square foot. Affordability is another utilitarian information that users can find in the research. Despite being an invaluable data source, Craigslist rental listings still have some drawbacks. Most importantly, price in Craigslist rental listings is not the final price of the property because buyers can negotiate with property owners to cut the price down. Moreover, most of mega cities like New York and San Francisco have the rental housing market nominated by brokers. All the above reasons can affect the analysis result[2].

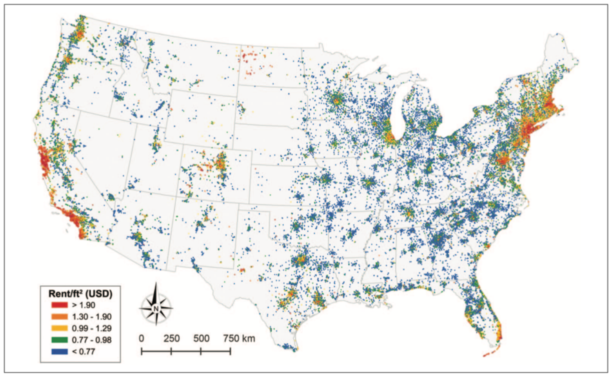


Figure 2: Rental listings map in the United States from Craigslist data set [2]

APPROACH

To get a closer insight about the real estate market in Ho Chi Minh city, we crawled nearly 500,000 listings from six real estate websites in Vietnam (*nhadat24h.net, batdongsan.com.vn, 123nhadat.vn, diaoconline.vn, muabannhadat.vn, alonhadat.com.vn*). Then, we extracted the useful data, developed tools to clean them, and visualize them to analyze the real estate market. It is notable that the price we gathered from the internet were only the advertised price, which mean it could, and most likely would, deviate slightly in the process of negotiation. Nevertheless, these listings still have its use as a source of reference for people interested in real estate.

Data crawling

There is a large amount of data transferred through the Internet every second, and it will be easier for data to navigate by using a formal and structured format that can be understood by computer. However, Internet contains a large portion of data that is unstructured, or semi-structured, human-readable but not easily machine-readable. Even though these data are abundant, they are hard to utilize because they do not follow any general format. In this project, the authors used web scraping to attempt to put this source of information to use[5]. More specifically, when accessing a web page, a web scraper will find certain elements of this page based on user requirements, extract these elements, transform them to a machine-readable format, and save the transformed data as a structured dataset on computer. The web scraper can apply a general format to many websites which have similar displayed style, or it can customize the format to use for a specific website. In essence, web scraping automates the process of collecting unstructured, messy data from a large number of web pages.

To build our dataset, we use Python and Scrapy to create spiders for each of our targeted website[3]. Each spider is tasked with the job of scraping any publicly available listings on the websites. Starting from a specified first page, the spiders travel through the structure of each website with the content defined by HTML and accessed an individual page for each listing[7]. Then, for each of those pages, XPath will extract the HTML elements that contain useful data[4]. Finally, the listings are saved on the hard drives for further processing. We schedule our crawler to run on each website starting from 1:00 AM daily, where they will gather all the listings posted yesterday. During the course of our crawling, we have collected roughly 600,000 listings. The acquired data is then saved in the following format:

- Post ID: the post's ID. Each website has their own format for this field
- Website: the website that this listing was crawled from.
- Title: title of the house listing
- Location: location of the house for sale/rent, include the following information:
  - Province
  - County
  - Ward (optional)
  - Road (optional)
  - Location detail: the whole address in the house listing, which can include further information like road number
- Author: name of the property owner
- Post time: the time that this house listing was posted
- Project (optional): the project that a house listing belongs to (only available for apartment)
- Bed count (optional): number of bedrooms of a house listing
- Area: area is extract from the raw area text by removing the unit (m^2)
- Price: price is extract by multiply the value with its unit. There are various types of unit: thousands, millions, billions thousands/m^2, millions/m^2, billion/m^2, thousands/m2/month, millions/m^2/month, billion/m^2/month. If the price is not saved in the per square metre format, it would be converted to it.
- Transaction type: either for rent or for sale
- House type: the real estate type of this house listing. The range of this variable change from website to website.
- Description: description provided by author of this house listing.

Data analyzing

Data collected from the Internet are usually messy and with a lot of error, either in typing, or in formatting. Our data is similar in that manner. Because of that, in order to have a more reliable material to conduct our analysis, we need to do some cleaning on the raw data set. Overall, there are three things that we need to do: first, we divide the listings into groups with the same location and house type; second, we remove the duplicate listings, which are listings that advertise for a single property; third, we remove the listings with unrealistic values in area or price. The removal of these two types of listing is because, if left alone, they would heavily skew the representation of the real estate market.

Data grouping. Before we do any cleaning, we need to divide the listings into smaller groups, where listings in the same group have the same location and house type. The motive behind this is to reduce the number of comparisons we need to make to remove listings. For duplication removal, listings with different value in location (province/ county) or house type have a very low chance to be duplicate and would only happen for house in the border of two different location entities, or when there is a mistake in inputting. Still, those occurrences are very rare and tolerable. Similarly, we can reduce the number of comparisons to remove the unrealistic values by grouping data since listings of the same region and of the same house type would have a lower deviation in price and area, which make detecting the anomaly values easier and more accurate.

From the initial dataset, we first divide them into listings for rent and listing for sale. Then, we split them into groups with the same province and county. Finally, for each of those group, we divide the listings further by their house type. At this point, we have our dataset divided into groups with the same transaction type, location, and house type.

Duplication removal. In our dataset, there are a lot of listings that advertise for the same property. These listings is posted by brokers repeatedly in a website or in multiple websites to increase the chance that it would catch the interest of some potential buyers. Because of that, we have a large number of duplicate listings that need to be removed since they would make the dataset biased if left alone. To detect the duplicate listings, we use a method call rule-based record linkage, which is a sequence of conditions that two listings is going to be compared on. If two listings satisfy all the conditions, they are regarded as duplicate and one of them is removed[1].

The sequence of condition is as follow: first, we compare the descriptions of two listings to see of they are similar; then, if two listings have similar description, we compare their price and area; finally, if the two listings also have the same values in price and area, we compare the posted day, and if they are more than 15 days apart, they are considered as duplicate. In order to determine if two descriptions are similar, we used term frequency - inverse document frequency, which is a way to describe a document based on its terms. Term frequency - inverse document frequency is a measurement of how important a term is in representing a document. Term frequency is how frequent a word appears in a document, while inverse document frequency is how common a word is in a collection of documents[9]. The formula to calculate term frequency is:

$$tf(t,d) = \begin{cases} 1 + \log(f_{t,d}) & (f_{t,d} > 0) \\ 0 & otherwise \end{cases} \tag{1}$$

, with **tf(t,d)** as the term frequency of term **t** in document **d** and $f_{t,d}$ is the number of time **t** appears in **d**. Inverse document frequency is calculated as follow:

$$idf(t) = \log(\frac{N}{n_t}) \tag{2}$$

, where **idf(t)** is the inverse document frequency of term **t**, N is the total number of documents, and n_t is the number of documents term **t** appears in.

$$tfidf = tf(t,d) \times idf(t) \tag{3}$$

We calculated term frequency - inverse document frequency **(tfidf)** by multiplying term frequency with inverse document frequency. If a term appears a lot in a single document while it is rare in other document, it has a high value in representing that document since both its term frequency and inverse document frequency is high. On the other hand, when a term has a high frequency in one document while also appears in a lot other documents, it is not as valuable since it is not unique enough since while term frequency is high, its inverse document frequency is low. A document can be represented by a vector of the term frequency - inverse document frequency of the terms in the collection of documents and by comparing these vectors using cosine similarity, we can determine if two descriptions is similar or not. To compute similarity between two vectors, we have:

$$similarity = \cos(\theta) = \frac{A.B}{\|A\|_2 \|B\|_2} = \frac{\sum_{i=1}^{n} A_\mathrm{i} B_\mathrm{i}}{\sqrt{\sum_{i=1}^{n} {A_\mathrm{i}}^2} \sqrt{\sum_{i=1}^{n} {B_\mathrm{i}}^2}} \tag{4}$$

In our case, if two vectors have their cosine similarity larger than 70%, they are considered as similar.

Unreasonable Value Removal. In all the websites that we gather data from, no website has any kind of restriction for the input of house price and area. Because of that, there exist some listings that have abnormal values compared with others in our dataset, either too high (thousands billion VND) or too low (zero VND). Outlier detection is an important part to maintain an insight that is as close to reality as possible. In this study, we tried many methods for removing the unrealistic properties.

First, we used the 3SD method which has formula:

$$\bar{x} \pm 3SD \tag{5}$$

,with $\bar{x}$ is mean, and SD stands for standard deviation. Any values stay outside these intervals (from mean to 3SD) are considered as outliers (unreasonable values)[8]. Nonetheless, the method of using standard deviation to detect outliers has some disadvantages. To detect outliers of a given data, the method supposes that the data follows a normal distribution including the outliers. Moreover, the mean and standard deviation can be greatly influenced by the outliers. Consequently, we use an alternative method called Interquartile Range (IQR) to detect outliers.

IQR is computed by subtracting the first quartile from the third quartile:

$$IQR = Q3 - Q1 \tag{6}$$

We multiply both ends of the box with three, and any real estate items having price outside the range from Q1 - 3xIQR to Q3 + 3xIQR are considered as unreasonable item (contain extreme price) and will be removed[6]. Despite being superior to 3 SD method and showing potential in removing outliers, Interquartile Range still has a serious drawback. In real life, there are many cases that Q1 is equal Q3, and IQR will be 0. In these cases, Interquartile Range cannot detect extreme values because the with of the expected range is 0, which show the limitation of Interquartile Range in outlier detection. Finally, we proceed to removing the listings with unrealistic price by modified z - score method on each of the groups that were divided in the first step.

We chose modified z - score because it is a robust way to calculate the deviation and detect anomalies since it is not affected by the extreme value like z - score, which depend on mean, whereas modified z - score depends on median. To compute modified z-score of a document, we first compute the median absolute deviation, which is the median of the deviation of each value in the groups compare to the median value. Modified - z score is calculated using the following equation:

$$M_i = \frac{0.6745(A_i - \tilde{A})}{MAD} \tag{7}$$

, where A_i is the i-th values in the array, $\tilde{A}$ is the median, and MAD stands for Median Absolute Deviation. If Mi, which is the modified z - score, is larger than **3.5**, the value is considered as extreme and will be removed[8].

RESULTS

Listing Distribution Pattern

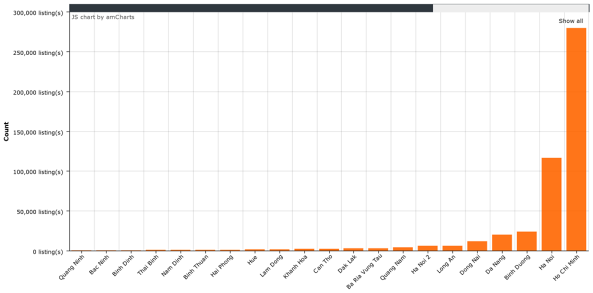


Figure 3: National listing distribution for sale

Most of the listings we scrape mainly belong to Ho Chi Minh and Hanoi for both sale and rent. Figure 3 shows that Ho Chi Minh and Hanoi stand out from other cities in the number of housing listings for sale across Vietnam, with Ho Chi Minh having 279,726 listings and Hanoi 116,533 listing. Because of that, we chose Ho Chi Minh as a case for validation of our proposed method as this method can be applied to other places in Vietnam and elsewhere.

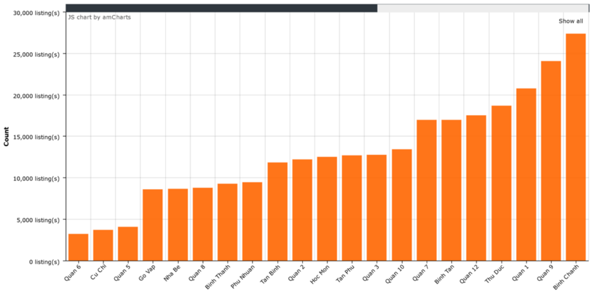


Figure 4: Distribution of total house for sale in districts of Ho Chi Minh

Across the entire dataset for Ho Chi Minh city, there are a total of 279,726 listings for sales and 44,569 listings for rent. Figure 4 show the distribution of listings in districts of Ho Chi Minh city. Among those districts, Binh Chanh has the biggest portion of real estates for sale (9.8% of Ho Chi Minh's listings), with District 9 following right behind (8.6%). These two districts are currently in the beginning of their development phase, since both of them either just have a new system of transportation infrastructure established

(1A Highway and Provincial Highway 10 in Binh Chanh) or have a new industrial zone with the gathering of many big company (Tan Thuan Export Processing Zone in District 9). District 1, being one of the most developed districts of Ho Chi Minh, boasts a large number of listings, which takes up to 7.4% of the whole city. District 12 and Thu Duc also have a large number of real estate listings, albeit not as large as the three highest districts, rest at around 3/4 of their total number of listings, which is 6.3% and 6.7% listings respectively. The construction of the new metro system really boosts the popularity of District 12, making its real estate market developed. Metro construction together with other existing transport infrastructures (Binh Trieu bridge, Binh Loi bridge, Tan Son Nhat - Binh Loi rim) have also made Thu Duc into a bustling market. The remaining districts do not have such a step decrease as the first four district. The real estate market in districts like District 2, 10, 3, Tan Phu, Binh Tan, etc., which had its infrastructure in place for a while but still have several projects in development phase had a moderate number of listings. These market is not as big as the new developing district but is still quite active. We did not get any listings for Thu Dau and Gia Lam, which maybe because brokers did not use the pages we selected to post their info. Can Gio also has a significantly low number of listings.

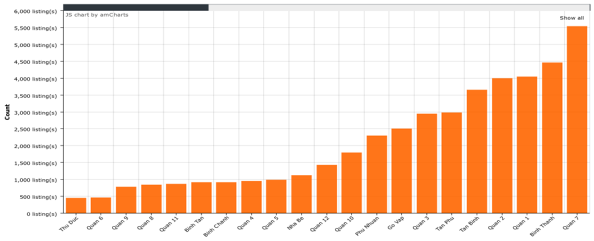


Figure 5: Distribution of listings for rent in Ho Chi Minh

Figure 5 shows the distribution of listings for rent in Ho Chi Minh city. The listings distribution of real estate for rent is a bit different from real estate for sale. That the total listing is much less than that of real estate for sale (about 1/3) indicates that the market for rental real estate is not as big as its counterpart. Many real estate listings gather in the central part of the city and the east side, with the most note-worthy is District 7, possesses the most listings (12.1%), with many new apartment projects and many people joining in the housing rental service. Other central counties like District 1, Tan Binh, Binh Thanh, etc. also have relatively large number of listings. The suburban counties like Cu Chi, Can Gio, Hoc Mon have really small numbers of listings because of smaller need for rental house (either because not many workers gather there or because the house price is lower) and many of their transaction happened offline, without advertising on the net.

Property Type

From Figure 6, which showed the real estate listings for sale grouped by type, we can see that the real estate market in Ho Chi Minh has 7 significant types: Townhouses (27.50%), apartments (21.81%), residential lands (19.68%), detached houses (8.79%), project lands (8.57%), alley houses (6.89%), and villa (2.91%). There are other remaining types, though each of them did not even make up 1% of the whole market. While there are many types of real estate, not all districts had a market for them, or at least, not significant. Most urban districts of Ho Chi Minh had town houses or/and apartment projects take a large portion in their market: District 1 (77.90% for town houses), District 10 (55.85% for town houses, 15.44% for apartments), Binh Thanh (23.49% for town houses, 37.34% for apartments), etc. On the other hand, most suburban districts have residential lands as the most dominant real estate type: Cu Chi (58.75%), Hoc Mon (34.50%), Thu Duc (34.51%).

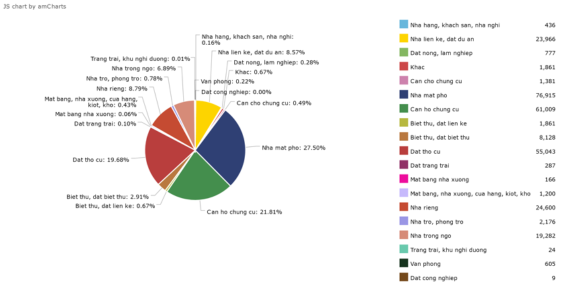


Figure 6: Ho Chi Minh real estate listings for sale by type

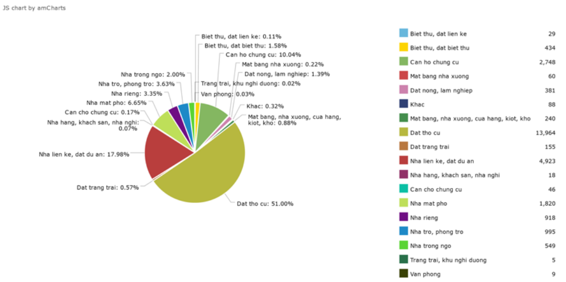
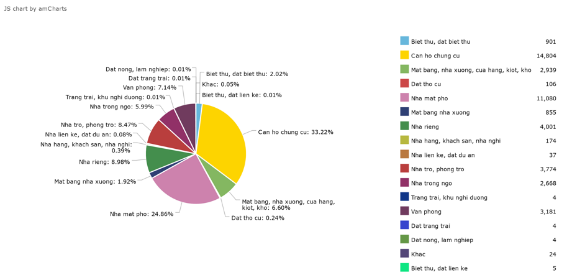


Figure 7: Binh Chanh's real estate listings for sale by type

One of the main districts of Ho Chi Minh that we focus in our scope is Binh Chanh, which has the highest number of listings for sale in Ho Chi Minh. From Figure 7 which represents the real estate market of Binh Chanh by types, we can see that there is a wide variety of real estate type, from hotel, farm lands, garages to residential lands, villa, and town houses. Overall, residential lands dominated the market and take up to 51%. This conforms to reality since Binh Chanh is a large suburban district with lots of free lands. With the recent establishment of many transportation infrastructure like: A1 route and Provincial Highway 10 connecting Binh Chanh with Duc Hoa Industrial Zone, Nguyen Van Linh road connecting A1 route to District 7, Vo Van Kiet road connecting District 2 and Dong Nai, etc., the real estate market in Binh Chanh has become much more vibrant. Other than residential lands, there is also some new project lands (An Lac Resident, New Life Resident) and apartment projects (HQC Plaza, South Saigon Complex), which take up 17.98% and 10.04% respectively.



Biet thu, dat biet thu	901
Can ho chung cu	14,804
Mat bang, nha xuong, cua hang, kiot, kho	2,939
Dat tho cu	106
Nha mat pho	11,080
Mat bang nha xuong	855
Nha rieng	4,001
Nha hang, khach san, nha nghi	174
Nha lien ke, dat du an	37
Nha tro, phong tro	3,774
Nha trong ngo	2,668
Trang trai, khu nghi duong	4
Van phong	3,181
Dat trang trai	4
Dat nong, lam nghiep	4
Khac	24
Biet thu, dat lien ke	5

Figure 8: Ho Chi Minh's real estate listings for rent by type

In Ho Chi Minh, apartment is the dominant type of rental housing (33.76%), with a lot of existing apartment complexes or in construction in Figure 8. These apartments litter throughout the districts of Ho Chi Minh but in all of them, District 7 stands out as the district with the most listings for apartments (18.7% of total listings in Ho Chi Minh). Another dominant type is town houses, which mostly belonged to urban districts like District 1 (16.7% of total town house listings), District 3 (11.5%), District 10 (7.79%), etc. To our surprise, even though rooms for rent is a common need for both university students and workers alike, it only took up a portion of about 8.47% of the whole market. These listings mostly belong to Binh Thanh, District 3, Tan Binh, District 7, District 10, and Phu Nhuan (12.5%, 12.24%, 10.47%, 8.92%, 6.22%, and 5.3% respectively in 3,774 total listings).

Price Pattern

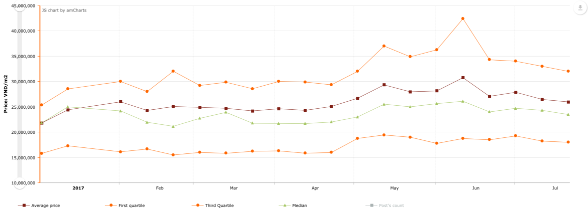


Figure 9: Price trend of apartments for sale in Ho Chi Minh

From the beginning of 2107, the price of apartment for sale in Ho Chi Minh city has increased steadily in spite of some fluctuations. At the end of July 2017, the average price reached 26 million VND per square metre, with median price being 23.4 million VND. The corresponding values of the first and third quartile are 18 million and 32 million VND per square metre (Figure 9).

Across the entire data set of apartment for sale in Ho Chi Minh from Figure 10, the median square meters is 28,412,166 VND and the mean square meters is 28,722,612 VND. Furthermore, while district 1 has the highest value of mean and median price (approximately 67 million VND per square metre), Cu Chi and Hoc Mon are two districts which have the least mean and median price of apartment for sale, at 5.5 million VND.

In district 3, there is a significant gap between mean and median price (Figure 10), with mean being 51 million VND and median 39.2 million VND per square metre. Because the median value is smaller than the mean value, we can conclude that there is a number of apartments having price smaller than the average price of the whole district. There are many causes for this problem such as location, area, number of bedrooms, and age of the property. To be specific, from Figure 11, we can see that although the price of apartment for sale in district 3 ranges mainly from 25 million to 75 million, there are a majority of units having a price of 45 million VND per square metre which is smaller than the average price of 51 million VND in district 3 (Figure 12).

Residential land is another property that we should consider because this property takes up a large portion of total real estate listings of the whole city (Figure 6). Ho Chi Minh has experienced a significant rise in the price of residential land for sale from the first month of 2017. From Figure 15, the average price of residential land was 17.8 million VND and the median price was 13.13 million VND per square metre at the end of July 2017. However, while there was little change in the first quartile (at 4 million VND), the third quartile increased considerably and reached 25 million VND per square metre.

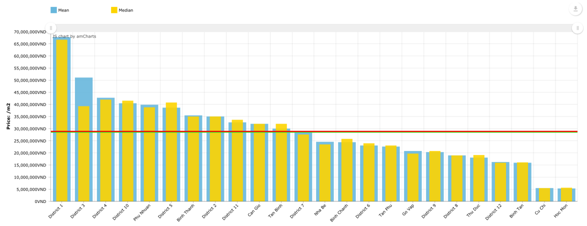


Figure 10: Mean and median price of apartments for sale in Ho Chi Minh

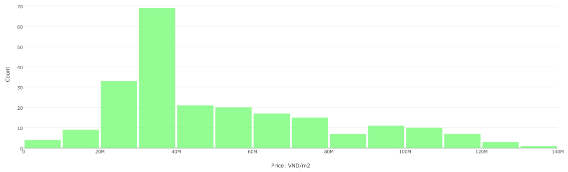


Figure 11: Price distribution of apartment for sale in District 3

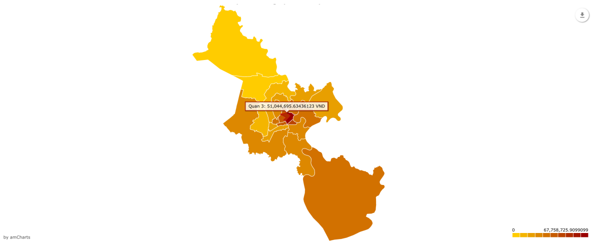


Figure 12: Average price distribution of apartment for sale in Ho Chi Minh

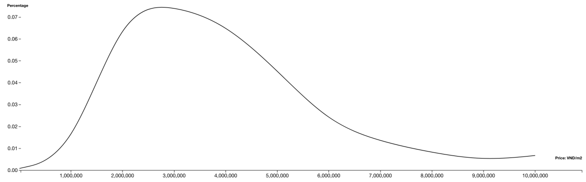


Figure 13: Price density probability of residental land in Binh Chanh

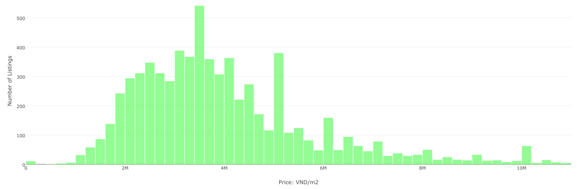


Figure 14: Price distribution of residental land for sale in Binh Chanh

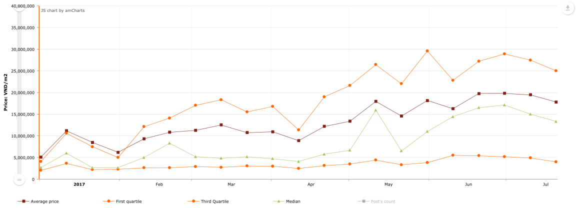


Figure 15: Price trend of residental land for sale in Ho Chi Minh

In addition to the general price pattern of residential land in Ho Chi Minh, we also examine the price pattern of residential land in Binh Chanh because this district has the greatest number of real estate listings for sale(Figure 4). As we can see from Figure 13, the area between 2 million and 4 million has a very large area under the curve, which means there is a large chance that a house in this area would have the price between 2 million and 4 million VND per square metre. This has been proved by Figure 14. From the figure, the greatest number of residential lands for sale in Binh Chanh has the price of roughly 3.5 million VND per square metre.

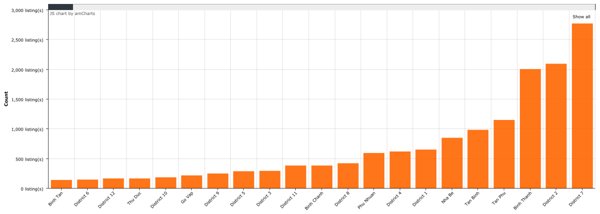


Figure 16: Distribution of apartment for rent in Ho Chi Minh

On account of rental housing market in the city, Figure 16 and Figure 5 depicts that most of the rental housing activities take place on district 7, and this district also accounts for the largest number of apartments for rent of the whole city, at 2,771 units. According to Figure 17, despite some fluctuations, the price of apartment for rent in district 7 has increased slightly from July 2017. At the end of July 2017, the mean and median price were approximately 178,000 VND per square metre. First quartile and third quartile also experienced a steady rise from January 2017 and reached 136,00 VND and 227,272 VND, respectively.

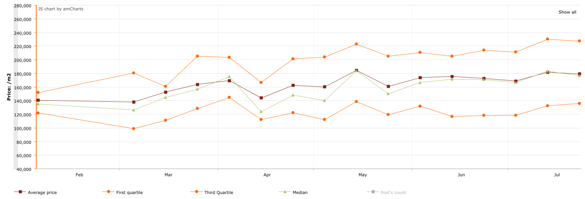


Figure 17: Price trend of apartment for rent in District 7

CONCLUSION

There are hundreds of thousands of real estate listings generated everyday on the housing websites. However, most of these real estate listings are usually overlooked due to their informal characteristics and commercial purpose. In this study, we propose a method for crawling and analyzing these web-based real estate listings.

By applying our method to the real estate market of Ho Chi Minh city and visualize these little-understood dataset, We provide different insights into the housing market of this city by our results. We assessed the distribution of real estate listings across Vietnam and found that Ha Noi and Ho Chi Minh makes up most of the housing items. Considering Ho Chi Minh city, we also found that while Binh Chanh constitutes the majority of housing items, it has little of apartments. By contrast, district 7 has the largest number of apartments for both sale and rent. Moreover, in the property type section, we tried to figure out the most populous real estate items for sale and rent in Ho Chi Minh. When examining Binh Chanh, we also found that residential land is the main real estae items for sale in this district. Our price pattern section discuss about the price trend, price distribution and price density of the whole city and also some specific districts.

To help people have an up-to-date view about the current housing market, we built a system that automatically crawl data from different resources at 1:00AM daily. For data visualization, we also developed a website so that people can view the current state of the market. The website has 9 different graphs by several graphs, and each of them has their own unique way to visualize the data.

REFERENCES

[1] Adelaide Ariel, Bart Bakker, Mark de Groot, Gerard van Grootheest, Jan van der Laan, Jan Smit, and Bep Verkerk. 2014. Record Linkage in Health Data: a simulation study. *The Hague/Heerlen: Statistics Netherlands* (2014). https://www.biolink-nl.eu/public/2014%20Record%20linkage%20simulation.pdf

[2] Geoff Boeing and Paul Waddell. 2016. New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. *Journal of Planning Education and Research 1-20* (2016), 20. https://doi.org/10.1177/0739456X16664789

[3] Scrapy Community. 2015. Scrapy: A Fast and Powerful Scraping Web Crawling Framework. (2015). Retrieved June 19, 2017 from https://scrapy.org/doc

[4] Michael Kay. 2008. *XSLT 2.0 and XPath 2.0 Programmer's Reference* (4th ed.). Indianapolis, IN : Wiley, 2004.

[5] Ryan Mitchell. 2015. *Web Scraping with Python: Collecting Data from the Modern Web* (1st ed.). Sebastopol, CA: O'Reilly Media.

[6] Purplemath. 2010. Interquartile Ranges & Outliers. (2010). Retrieved August 3, 2017 from http://www.purplemath.com/modules/boxwhisk3.htm

[7] Jonathan Reid. 2015. *HTML5 Programmer's Reference* (1st ed.). New York: Apress.

[8] Songwon Seo. 2006. *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. Master's thesis. Kyunghee University.

[9] Tfidf. 2017. Tf-idf :: A Single-Page Tutorial - Information Retrieval and Text Mining. (2017). Retrieved August 2, 2017 from http://www.tfidf.com/