Combining Online News Articles and Web Search to Predict the Fluctuation of Real Estate Market in Big Data Context

Daoyuan Sun
Department of Information Systems
National University of Singapore
Singapore
daoyuan@comp.nus.edu.sg

Yudie Du
School of Information
Renmin University of China
Beijing, China
duyudie.605@163.com

Wei Xu*
School of Information
Renmin University of China
Beijing, China
weixu@ruc.edu.cn

Meiyun Zuo
School of Information
Renmin University of China
Beijing, China
zuomy@ruc.edu.cn

Ce Zhang
School of Information
Renmin University of China
Beijing, China
fanczc@foxmail.com

Junjie Zhou
Henan University of Economics and Law
Zhengzhou, Henan, China
jjzhou@huel.edu.cn

Abstract

The real estate price is of paramount importance in both economic and social fields. It is a key indicator of the operation of real estate market and its prediction is essential in the decision-making process of both average people and official governments. Past researchers on this topic have already proposed several prediction methods including linear regression models, nonlinear regression models and machine learning models. Nevertheless, those models have generally

neglected the impact of human behavior, which we believe is a significant factor of the real estate price prediction. What's more, past studies have shown that news sentiments could improve the prediction performance of real estate price. Search engine query data were studied to reflect web users' behavior by analyzing the frequency of words searched by online users. Researchers have already used the news sentiments and query data for prediction, respectively. But none have combined them together as an integrated model. In this paper, we propose an integrated method that throws new light on the prediction of real estate price in China by integrating these two factors into the forecasting model. In our method, we extract sentiment series from both news data and search engine query data by adding weights to original sentiment series that are produced by news data alone. Then both the weighted series and original ones are used as inputs of several well-acknowledged data mining models, including SVR, RBFNN and BPNN, to produce prediction results.

To validate the integrated model, we apply it to four representative cities in China respectively, and compare the results produced by the integrated model using weighted inputs with non-integrated ones using original inputs. The results show that for every one of the four cities, the integrated model generally leads to lower prediction errors than the non-integrated ones. This not only validates the model's accuracy and universality, but also proves the hypothesis that human searching behavior as a strong impact in typical Chinese cities' real estate market and can enhance the prediction accuracy of real estate prices.

Keywords: Real Estate Market, Prediction, Web Mining, Online News Articles, Search Engine Queries

https://aisel.aisnet.org/pajais/vol6/iss4/2
DOI: 10.17705/1pais.06403

Introduction

As an important industry in China, real estate has consistently exerted huge influence on social and economic fields, contributing to a large proportion of the nation's GDP. Because of its tight connection to almost every possible social sector, research on real estate market prediction can significantly facilitate the decision-making process of individuals, enterprises and government. The global real estate market has suffered and become more volatile and unpredictable since the sub-prime mortgage financial crisis in 2008 (Profile of Home Buyers and Sellers 2011).This, however, has led to more intensive research on real estate market the world over.

A profile by National Association of Realtors (NAR) indicates that 90% of the real estate buyers use the Internet to find useful information and 92% sellers use the Internet to publicize trading information (Profile of Home Buyers and Sellers 2011). Apart from that, according to an unofficial survey made by SouFun.com in China, more than 60% of Chinese web users obtain real estate information through the Internet and approximately more than 80
% Chinese web users access the website of real estate per month. These survey results suggest that online users' behavior is almost a mirror of real world human behavior in the real estate market. It can reflect the situation of real estate market to a large extent. Thus it is reasonable and feasible to take online users' behavior into consideration when forecasting the real estate market.

Findings in behavioral economics tell us that emotions can affect individual behavior profoundly, and emotions or sentiment of words in online opinions or articles could serve as an effective indicator of real world and therefore could be used for prediction. Extant research already shows that sentiment expressed in web articles or other forms of web data has great potential for the market prediction. Multiple types of web data have been used to make all kinds of predictions by extracting their market sentiments and news analysis. Das and Chen (2007) used word play on certain message boards to discover the relationship between web sentiment and stock returns. Moreover, Bollen and Mao (2011) used twitter mood as the input of fuzzy neutral network to prove that public mood and the Dow Jones Industrial Average close value are closely related. Similarly, online opinion ensemble was also used to predict the financial market (Xu et al. 2012). These examples remind us that sentiment in web data can be integrated into the model of real estate market prediction.

Although news sentiment can reflect part of human behaviors in the real estate market, the sentiment alone is not enough to make the prediction accurate. The content of a news article with few viewers cannot indicate the actual fluctuation the real estate market, however strong its sentiment is. To solve this problem, researches have now considered incorporating data on online users' behavior and search engine query data into research. The idea behind search engine query data is that it is a time series that accurately reflects online users' searching behavior. Since people search what they focus on, these data are a direct reflection of their interests and specific behavior such as purchasing desires, investing possibilities, bargaining or selling motivations during a certain period of time (Wu and Brynjolfsson, 2009). Researchers have already used search engine query data for prediction. Wu and Brynjolfsson (2009) used Google search engine query data to predict the real estate price index and quantity volume which gained profound outputs. Furthermore, Sun et al. (2014) used search engine query data to predict the real estate prices and achieve better results.

Among various forms of web data, online news articles from reliable sources contain abundant, timely, convincing and valuable information about real estate market in reality. Therefore, the sentiment used in

these articles could be adopted to make predictions and is introduced into the proposed prediction model. Meanwhile, as online users' behavior is critically important in market prediction, search engine query data is also integrated into our model in order to make more precise prediction. Although previous studies have used news sentiment and search engine query data respectively for prediction, no one has ever combined these two indicators in a single integrated model. In this paper, a new method of real estate price index prediction is proposed by introducing the human behavioral factor into the forecasting model. By combining online daily news' sentiment and search engine query data, we constructed an integrated data mining model that has satisfactory forecasting performance. In our model, web information and the lags of real estate price index time series data are integrated into the model, and the state-of-the-art data mining tools are used to model the relationship of real estate market price and web information to achieve a better forecasting performance.

Literature review

To predict real estate market prices more precisely, a web mining-based model is proposed for real estate market price prediction. Novel web information-based indicators and state-of-the-art data mining models are introduced for prediction purposes. To find the research gap and address our contributions, this section reviews related literatures in both real estate market analysis and prediction, and web mining-based market prediction.

Real estate market prediction

In the past years, a number of models and methods have been put forward for real estate market prediction. The sensitive indicator of real estate market, usually the real estate price index, has been studied by many researchers who hope to predict the future trend of real estate market. Factors affecting real estate market have been well studied. For example, Grebler (1979) pointed out that some indicators (e.g., income, the Consumer Price Index (CPI), seasonal factors, vacancy rate, and previous real estate price) can be used to predict the real estate price. Nellis and Longbottom (1981) found that the determinants of real estate price were based on real disposable income, loan interest rate, and total loan. Case and Shiller (1990) forecasted the real estate price and excess return in the real estate market using the percentage change in real per capital income, real construction cost, adult population, marginal tax rate and housing starts. Clapp and Giaccotto (1994) studied the influence of economic variables on local real estate price dynamics and found that some variables such as population, employment and income had a considerable forecasting ability for housing price. To identify the factors influencing yearly urban real estate price, Potepan (1998) used a number of indicators including the privately owned dwelling price index, monthly rent based on the hedonic model, land price, medium income, population, quality of public services, crime rate, air pollution, non-dwelling consumable price, mortgage rate, construction cost, farm land price, land restriction et cetera. Baffoe-Bonnie (1998) used a nonstructural estimation technique to analyze the dynamic effects of four key macroeconomic variables on real estate price and the stock of houses sold. The impulse response functions derived from vector auto-regression (VAR) model suggested that the real estate market was very sensitive to shocks in the employment growth and mortgage rate at both national and regional levels. Malpezzi (1999) used a time-series cross-section regression to find that real estate price did not change according to a random walk, and at least it could be partially forecasted. Seko (2003) thought that there was a strong correlation between Japanese dwelling price and economic fundamentals in some areas, and used a time series model to forecast housing price with some indicators, including average sales price of private ownership dwelling,

https://aisel.aisnet.org/pajais/vol6/iss4/2
DOI: 10.17705/1pais.06403

annual household income, population, new-started dwelling, the CPI and the vacancy rate.

By exploiting these factors, several forecasting models are offered for real estate market price prediction. Mei and Liu (1994) employed a multi-factor latent variable model for real estate price prediction. Fingleton (2008) applied a spatial model with moving average errors to predict real estate price, with a generalized method of moment's estimator. Pace et al. (2000) used a spatial–temporal forecasting model to predict real estate price. Anglin (2006) set up a VAR model to forecast Toronto real estate price, with the predictors' three lags for the average real estate price growth rate, CPI, mortgage rates and unemployment rate. Wilson et al. (2002) used neural networks to predict real estate price. Similarly, Khalafallah (2008) applied a neutral network model to the real estate price prediction with only slight forecasting error. Liu et al. (2006) applied a fuzzy neural network for real estate prediction. Li et al. (2009) suggested an SVR-based forecasting approach for real estate market price prediction. Wang et al. (2008) used a RS-based SVM model in real estate price prediction. Yan et al. (2007) pointed out the limitations and shortcomings in previous research and proposed an integrated method for real estate price forecast based on TEI@I methodology.

As the world is now moving into an era where the Internet and other Information technologies become commonplace, more and more people are beginning to use the Internet to obtain or create information of real estate market. Thus, it is an urgent challenge for researchers to propose new methods for real estate market predictions. For example, Wu and Brynjolfsson (2009) used Google search engine query data to predict the real estate price index and quantity volume, which gained profound outputs.

Web information-based market prediction

Web information is regarded as a treasure for market intelligence, and has been used for market analysis and prediction. Web content has been suggested for market prediction. For example, Schumaker and Chen (2009) proposed a quantitative stock prediction system based on financial web news. Meanwhile, sentiment analysis has been used in market prediction. For example, Das and Chen (2007) made the study using sentiment of words on certain message boards to discover the relationship between web sentiment and stock returns. Moreover, Bollen and Mao (2011) used twitter mood as the input of fuzzy neutral network to prove that public mood and the Dow Jones Industrial Average close value are closely related. Similarly, Xu et al. (2012) used an online opinion ensemble method to predict the financial market. These examples remind us that sentiment in web data can also be integrated into real estate market prediction model.

Web usage, especially users' search data, also has been proposed to make market prediction better. Ginsberg, et al. (2009) applied the search engine query data to detect the seasonal influenza epidemics. Xu et al. (2012) used search engine query data, together with machine learning models, to predict the unemployment rate and received significant results. Furthermore, Xu et al. (2014) built an unemployment ontology first and then applied it to unemployment rate prediction with search engine query data. Finally, Wu and Brynjolfsson (2009) used Google search engine query data to predict the real estate price index and quantity volume which gained profound outputs.

To combine the advantage of social media and web usage information, this paper proposes an integrated forecasting model to predict real estate prices using both the sentiment of news articles and the search volumes of online users.

The rest of this paper is organized as follows. Section 3 gives the theoretical background of the proposed data mining methods. Section 4 presents the research

framework and the modeling process in detail. Then the empirical analysis is carried out to verify the effectiveness and feasibility of the proposed method and to compare the proposed integrated method with other non-integrated models. Finally the last section gives conclusions and future work.

Technical foundation

In this section we present the background of several time series analysis and data mining tools as the theoretical foundation of prediction.

Artificial neutral networks

Artificial Neural Network (ANN) is a computational model imitating the operation of animals' central nervous systems, particularly the brain. It has been well acknowledged and applied in a variety of issues, especially in detecting the non-linear relationship between the input and output data. ANN consists of several layers of interconnected neurons, namely an input layer, hidden layer(s) and an output layer. Connections between neurons are given values called 'weight'. The specific structure of ANN is determined during the training procedure. A typical ANN is demonstrated in Figure 1.

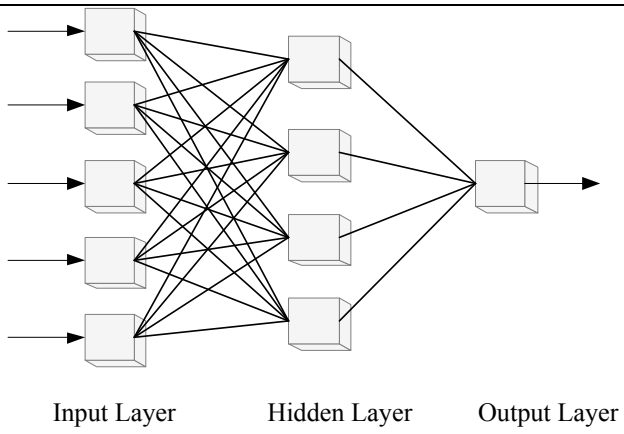

Figure 1 – Structure of Artificial Neutral Networks

Among all kinds of ANN models, Back Propagation Neural Network (BPNN) and Radio Basis Functions Network (RBFNN) are two widely employed networks well suited for prediction. For BPNN, the crucial process is iteratively feeding the training data to the network, and modifying the weights of connections based on the deviation between the predicted and the actual output until the error is in an acceptable range. As for RBFNN, input is mapped onto each radio basis function in the 'hidden' layer. The function chosen is often a Gaussian. When applied in regression problems, the output layer is a linear combination of hidden layer values representing mean predicted output.

Support vector regressions

Support vector machines (SVM) are supervised machine learning models with associated learning algorithms that can analyze data and recognize patterns in them. It is initially used for classification. Given a set of training data, which is marked as belonging to one of two categories, a SVM training algorithm would build a model that recognizes the pattern in the training dataset. Therefore, the model is able to assign new examples into one category or the other, making it a feasible way of prediction. The process can be shown in Figure 2.

Specifically, a SVM model is a representation of the instances as points mapped in a space so that the examples of the separate categories are divided by a

https://aisel.aisnet.org/pajais/vol6/iss4/2
DOI: 10.17705/1pais.06403

clear gap. New examples are then mapped into the same space and predicted to belong to a category based on which side of the gap they fall on. SVMs can efficiently perform non-linear classifications using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Support vector regression (SVR), which is an extension of SVM, was proposed by Vapnik in 1996. There are two types of SVR models, each of which has four types of kernel functions performing the 'kernel trick' mentioned before. Both have gained lots of achievement on issues of non-linear regression. Table 1 presents different types of SVR models.

Table 1 - Machine Learning Methods and Kernel Functions

Kernel Functions	Kernel Formula	SVR Types
Linear	$x^T x_i$	$\varepsilon - SVR$
Polynomial	$(\gamma x^T x_i + 1)^q$	
Radial basis function	$\exp(-\| x_i - x \|^2/\sigma^2)$	$\nu - SVR$
Sigmoid	$\tanh[\gamma {x_i}^T x + c]$	

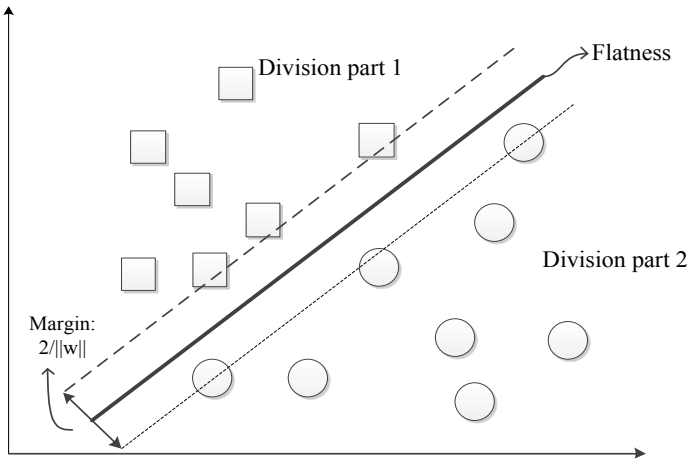

Figure 2 - Core Content of SVM

Research methodology

In this section，we propose a novel method which combines online daily news sentiments with search engine query data. We assume the latter is able to reflect human searching behavior by recording the frequency of words searched by online users. Therefore, the prediction model we have put forward is an integrated one that addresses both real estate news and human searching behavior. We hope to validate the integrated model by detecting whether it can improve the accuracy of real estate market prediction compared to non-integrated models.

In this study, we choose the house price index as an indicator of Chinese real estate price and also as the predicted variable in regression models. In the prediction phase, we apply several well acknowledged data mining tools including different types of artificial neural networks (ANN) and support

vector regression (SVR) to obtain the prediction results, respectively. After obtaining experimental results, we make a comparison between models with search engine queries (integrated models) and models without query data (non-integrated models) in order to justify the integrated model we propose. The framework of the proposed method is shown in Figure 3, and more details will be discussed in the following subsections.

What is noteworthy is that the house price index data can only be obtained at the city level; hence, we could only examine one specific city at a time. Therefore, to examine the universality of our model for typical Chinese cities, we conduct experiments using data of four Chinese cities in this paper: Beijing, Shanghai, Chengdu and Hangzhou, the locations of which are indicated in Figure 4. These four cities serve as representatives of different tiers of Chinese cities in terms of different sectors such as geographical feature, economy, policy, tourism and culture, which are all important factors affecting real estate price *i.* We believe by examining these four cities, we can conclude the model with both satisfactory performance and fine universality; therefore practically it can be extended to predict other cities besides the four cities addressed in this paper.

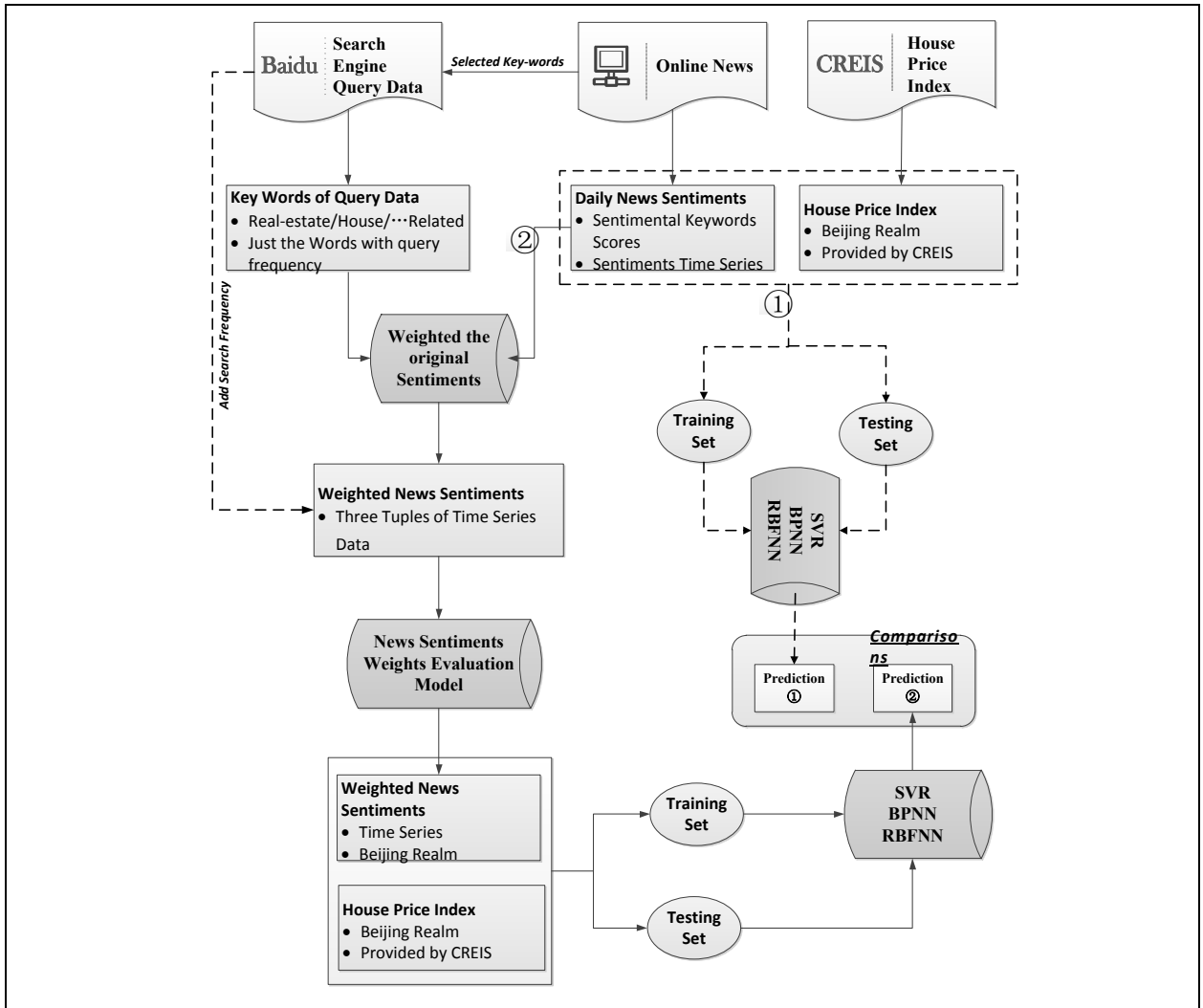

Figure 3 - The Research Framework of Proposed Methodology

https://aisel.aisnet.org/pajais/vol6/iss4/2
DOI: 10.17705/1pais.06403

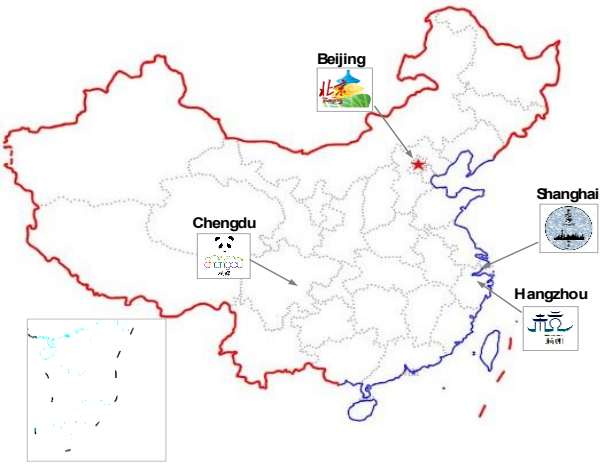

Figure 4 - The Four Empirical Cities in China

Data acquisition and data processing

The acquisition and processing of data follow the steps below. All these steps are applied to data of the four selected cities in question, respectively.

Step 1: Acquisition of online news articles & real estate price index

In order to reflect the actual operation and situation of four cities' real estate market, the news articles in our experiment ought to come from reliable sources whose information is true to reality. Therefore, it is imperative that we crawl the news from authoritative and convincing websites at regular intervals. As the biggest and leading web station in China, *www.sina.com* has enjoyed a large user volume for decades and its impacts on society cannot be underestimated. More importantly, because of its huge influence in many social aspects, the reliability of its content is closely watched by institutions, agencies, and the public so as to prevent rumors and malicious information from circulating around the country. These qualities make it a feasible and valid source of news articles in our experiment. Furthermore, we choose the House Price Index (HPI) as the predicted variable in this paper, as it is an authoritative index reflecting the real estate market situations poignantly. These two data sets can be gathered by using a self-programmed web crawler.

Step 2: Construction of query data set

After all the news articles have been collated, we are able to gather some most frequent words in the news articles as a whole. In this paper, we record the first 138 most frequent words in the crawled news articles because they are proved to be adequate enough to represent the overall sentiments of corresponding articles (Sun et al. 2014).

Next, the search frequency by online users of these words is also recorded from a most influential search engine in China, *www.baidu.com*. Ultimately, the recorded words, together with their search frequency, make up the key words query data set which is a crucial part of our experiment.

Step3: Generation of original sentiment series

The crawled news articles often contain words which could reveal public moods and opinions. These words are also called sentiment words because they contain certain types of human sentiment in their meanings. To extract the sentiment in those words, and process it into quantitative data

which can be easily computed and calculated, we use the open-source package ICTCLAS50 and the sentiment dictionary, from the Department of Chinese, Tsinghua University, to calculate the sentiment score of each article, and generate four types of sentiment scores according to the different attributes of human moods: positive, negative, the addition and subtraction of positive and negative.

By adding up the four types of sentiment scores of all the articles published at the same time point t respectively, we can obtain four sentiment series: $\mathrm{P_t}$ (positive sentiment score), $\mathrm{N_t}$ (negative sentiment score), $\mathrm{U_t}$ ($\mathrm{P_t} + \mathrm{N_t}$), and $\mathrm{S_t}$ ($\mathrm{P_t} - \mathrm{N_t}$).The specific equation is as follows.

$$P_t = \sum_i P_{ti},$$

$$N_t = \sum_i N_{ti},$$

$$U_t = \sum_i P_{ti} + \sum_i N_{ti} = \sum_i U_{ti},$$

$$S_t = \sum_i P_{ti} - \sum_i N_{ti} = \sum_i S_{ti}.$$

Where P_{ti}, N_{ti}, U_{ti}, S_{ti} represent the four types of sentiment score of article i at time point t.

The four sentiment series in this step are called original sentiment series, since they only contain information provided by the online news sentiment and have not included the human searching behavior factor. However, we can still predict the operation of real-estate market with the original sentiment series.

Step 4: Generation of weighted sentiment series

This step can be divided into two parts. The first part is to label the crawled news articles with the help of query data set constructed in step 2. In other words, we attach a key word as a label to every single piece of news. The key word of an article is one belonging to both the article and the query data set that appears most frequent in this article compared with other words in query data set. There may be cases where an article happens to contain no words in the query data set. However, these cases are so rare relative to the total number of crawled articles that when we come across them, we can simply ignore these 'strange' articles. In this way, each news article is associated with a key word, which leads to the second part of this step.

In the second part, the search volume (query data) of the key word for each article is added as weight to the original sentiment score to produce a new sentiment score called weighted sentiment score. The same as in step 3, the weighted sentiment score also has four types, which can be calculated with equation (1), where *W_sentiments*, *O_sentiments* represent the weighted, original sentiment scores respectively, and subscript it means the corresponding variable for article no. i at time point t.

$$W\_sentiments_{it} = O\_sentiments_{it} \cdot \frac{Searching\ Volume_{it}}{\sum Searching\ Volume_{it}} \ldots\ldots\ldots\ldots\ldots (1)$$

Hence, we can acquire four weighted sentiment series just as in step 3. These are series that contain information about both news sentiments and human searching behaviors.

When we finish data processing, we can obtain 9 crucial time series for each city for later experiment: 4 original sentiment series, 4 weighted sentiment series and the HPI time series. With these time series, we are capable of constructing specific data mining

https://aisel.aisnet.org/pajais/vol6/iss4/2
DOI: 10.17705/1pais.06403

tools for HPI forecasting. We will discuss more details in the following sections.

Construction of prediction models

In this section, we use the obtained time series data to construct specific prediction models with three different data mining tools: SVR, BPNN, and RBFNN. All three tools are very successful data mining models that are widely adopted by researchers in predicting time series data. For the purpose of enhancing the accuracy of prediction, we apply all the three models: SVR model, BPNN model and RBFNN model to the data of four cities, respectively. In each run of a specific model, we use different combinations of the four types of sentiment series, the HPI time series and their lags as input data. The combinations of the four sentiment series can be seen in Table 2. As is indicated, there are altogether 15 combinations of sentiment series, each of which has two types: original sentiment series and weighted sentiment series. Both

are produced in the former section.

As we have mentioned in former sections, SVR model can be further divided into 8 concrete models according to four different kernel functions ($K(\mathcal{X}_i, \mathcal{X})$) and two types of SVR models. The detail of the 8 SVR models can be seen in aforementioned section.

Table 2 – Combinations of Sentiment Series

ID	Series	ID	Series	ID	Series	ID	Series	ID	Series
1	P_i	2	N_i	3	U_i	4	S_i	5	P_i, N_i
6	P_i, U_i	7	P_i, S_i	8	N_i, U_i	9	N_i, S_i	10	U_i, S_i
11	P_i, N_i, U_i	12	P_i, N_i, S_i	13	P_i, U_i, S_i	14	N_i, S_i, U_i	15	P_i, N_i, S_i, U_i

Evaluation criteria for models

In order to evaluate the performance of each model as mentioned above, and find out the best prediction model for each city, we propose three measures, namely Root Mean Square Error ($RMSE$), Mean Absolute Error (MAE) and Relative Absolute Error (RAE) as standards for model evaluation. Given n pairs of actual value of price index (RI_i) and predicted value of price index (PI_i), the three measures are defined as follows. The smaller their values are, the better the model performs.

$$RMSE = \sqrt{\sum_{i=1}^{n}(\mathrm{PI}_i - \mathrm{RI}_i)/_n}$$

$$MAE = \sum_{i=1}^{n}|\mathrm{PI}_i - \mathrm{RI}_i|/_n$$

$$RAE = \frac{\sum_{i=1}^{n}|\mathrm{PI}_i - \mathrm{RI}_i|}{\sum_{i=1}^{n}|RI_i - \overline{\mathrm{RI}}|}$$

These three criteria are employed to evaluate the accuracies of models and data inputs.

Empirical analysis

Data collection

The crawled news articles cover the time span from December, 2012 to July, 2014 from "Sina Real Estate News" (*http://house.sina.com.cn/#*). And the key words search frequency from search engine can be recorded from "*http://index.baidu.com*".

In addition, the monthly House Price Index (*HPI*) is obtained from China Index Academy (*CIA*), a professional real estate

research and statistical organization in China.

Experimental results

In this part, we apply the proposed method to the data of four cities respectively. Table 3 and Figure 5 show the results of our experiments, including values of *RMSE, MAE* and *RAE* of different models and the corresponding input data (combinations of time series) as indicated by indexes in brackets in the table. We point out that, for SVR models, we only present results produced by linear kernel of $\varepsilon - SVR$ since it outperforms all the other types of SVR models. With these results, we are able to compare the performance of integrated models with weighted series and that of non-integrated models with original ones. Furthermore, we can also identify the best model for each city, and check out whether it is an integrated model or a non-integrated one; in this way we are capable of testing whether the introduction of human searching behavior in real estate price prediction would improve the prediction.

Table 3 – Errors & Inputs of Prediction Models for Four Cities

Cities/Series \ Models			SVR	RBFNN	BPNN
Beijing	RMSE	Original	452.2323 (1)	1723.3880 (3)	363.7115 (13)
		Weighted	395.4115 (8)	1532.8960 (2)	312.3568 (12)
	MAE	Original	548.0784 (1)	1379.2960 (3)	278.3885 (13)
		Weighted	486.1089 (8)	1199.9760 (2)	267.6558 (12)
	RAE (%)	Original	16.7408 (1)	51.1332 (3)	10.3204 (13)
		Weighted	14.3775 (8)	44.4855 (2)	9.9225(12)
Shanghai	RMSE	Original	273.4290 (2)	908.1365 (4/5)	179.1604 (8)
		Weighted	269.2164 (3)	692.9583 (1)	167.9230 (2)
	MAE	Original	273.4290 (2)	711.9938 (4/5)	142.8453 (8)
		Weighted	271.7455 (3)	591.9469 (1)	133.7654 (2)
	RAE (%)	Original	15.5422 (2)	40.1037 (4/5)	8.0459 (8)
		Weighted	15.1202 (3)	32.1489 (1)	7.3230 (2)
Hangzhou	RMSE	Original	193.0334 (2)	340.5245 (2)	180.6028 (13)
		Weighted	166.5122 (2)	333.3563 (1)	178.9342 (3)
	MAE	Original	231.3926 (2)	295.8521 (2)	134.3566 (13)
		Weighted	196.9983 (2)	259.2744 (1)	119.8374 (3)
	RAE (%)	Original	36.4606 (2)	55.4513 (2)	25.1823 (13)
		Weighted	31.4512 (2)	47.1287 (1)	21.8356 (3)
Chengdu	RMSE	Original	44.5431 (3)	103.5110 (4)	91.8181 (3)
		Weighted	54.4161 (4)	94.5965 (3)	87.1117 (2)
	MAE	Original	51.2437 (3)	86.6173 (4)	74.0267 (3)
		Weighted	66.3225 (4)	75.9824 (3)	70.7321 (2)
	RAE (%)	Original	30.3690 (3)	58.5835 (4)	50.0678 (3)
		Weighted	38.5105 (4)	51.9446 (3)	47.8395 (2)

Notes: The numbers in the brackets are the index of combinations of input sentiments series, which means the same as in Table 2; the colored ones are results of best model for each city.

https://aisel.aisnet.org/pajais/vol6/iss4/2
DOI: 10.17705/1pais.06403

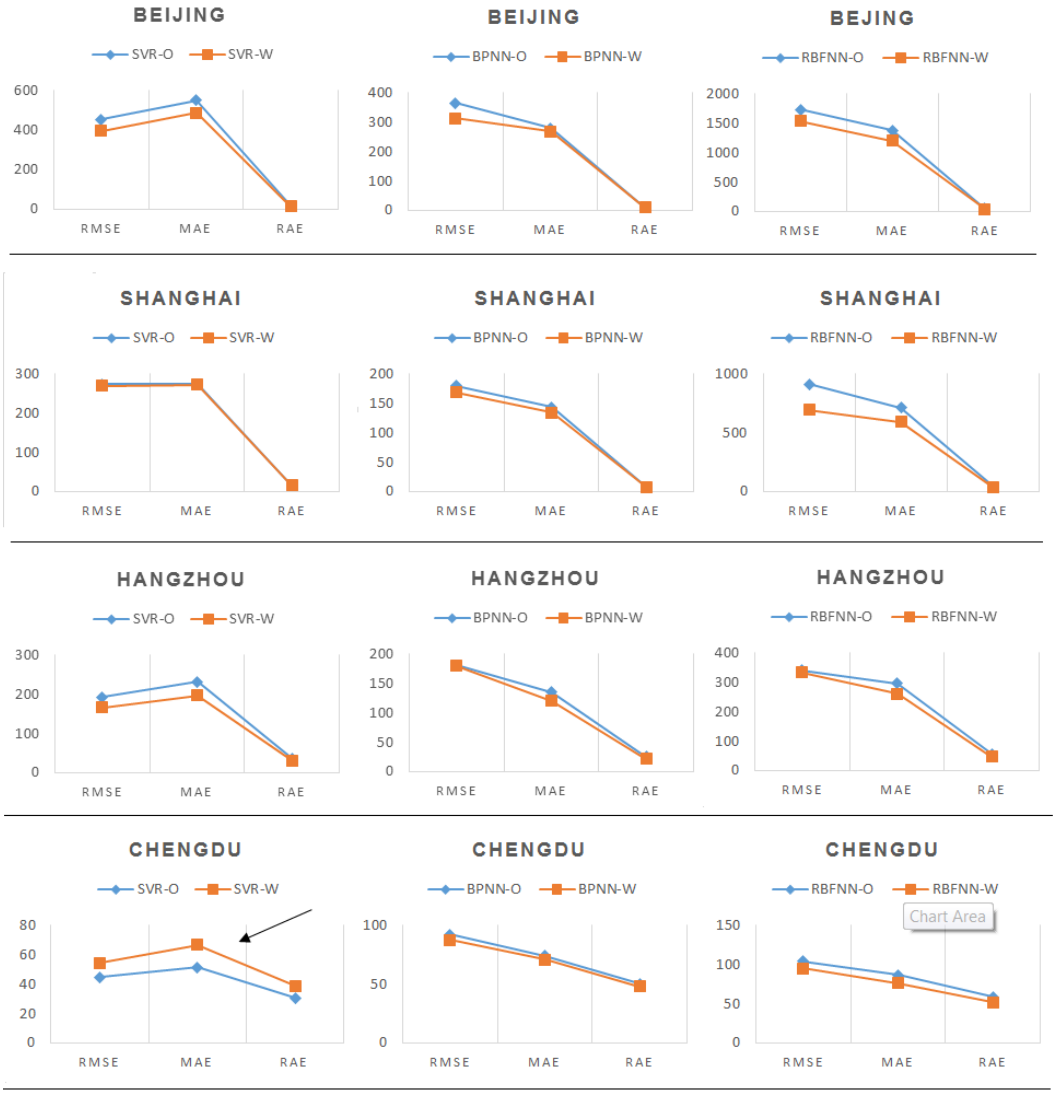

Figure 5 – Comparisons of Models of Weighted and Original Inputs

As can be seen in the table and figure above, for Beijing and Hangzhou, all the three criteria (RMSE, MAE, RAE) for models with weighted series inputs show noticeably lower values than the same models with original series inputs. This indicates that search engine query data helps to enhance the accuracy of prediction in real estate price of both Beijing and Hangzhou, which further suggests the significance of human searching behavior in the prediction of real estate market of these two cities. What's more, the model with best performance for Beijing comes from BPNN model with input of weighted series combination No.12 (See Table 2). Similarly, the best model for Hangzhou also comes from BPNN model with weighted input No. 3 (See Table 2).

As for Shanghai, it bears almost the same feature. Integrated models generally outperform non-integrated ones in terms of three criteria except for the MAE value in SVR model, which records 273.4290 for original series input but a higher 321.7455

for weighted input. However, on the whole, weighted inputs still have better prediction accuracy than original ones. Again, the best model comes from BPNN model with input of weighted series No. 2.

For Chengdu, the only one landlocked city in our sample, the picture is quite different. Although for all the RBFNN and BPNN models, weighted inputs still present higher accuracy of prediction than original ones. But for SVR models, they show a converse trend, in which the non-integrated models with original inputs lead to lower errors for all the three criteria. Furthermore, the best model for Chengdu actually comes from SVR model with original series input No. 3. To sum up, for most cases, models of any types with weighted inputs produce better results than those with original ones. Specifically, BPNN models with weighted inputs are proved to be best models for all the three cities: Beijing, Shanghai and Hangzhou. As for Chengdu, however, the best model is SVR model with original inputs.

Comparisons

Since the former research only considered the prediction effects of search engine queries or online news alone, our integrated model is more comprehensive in using the prediction effects from both online news and human searching behaviors. We have proved in the previous section that models using search queries weighted sentiments have better performances than the original sentiments. In this part, we compare models using search queries weighted sentiments with the models using solo search queries to find whether our new method has better performances. The empirical results demonstrate that our integrated model surpasses the model of solo search queries, which means our new method is better than models using only search queries.

Table 4 – Errors & Inputs of Prediction Models for Four Cities

Cities/Series \ Models			SVR	RBFNN	BPNN
Beijing	RMSE	Solo	3196.2627	2985.4865	3240.8107
		Integrated	395.4115 (8)	1532.8960 (2)	312.3568 (12)
	MAE	Solo	2639.6579	2500.1441	2865.6879
		Integrated	486.1089 (8)	1199.9760 (2)	267.6558 (12)
	RAE (%)	Solo	97.8573	92.5507	97.3139
		Integrated	14.3775 (8)	44.4855 (2)	9.9225(12)
Shanghai	RMSE	Solo	1998.7399	1898.5442	2004.1762
		Integrated	269.2164 (3)	692.9583 (1)	167.9230 (2)
	MAE	Solo	1941.2369	1738.2724	1770.6208
		Integrated	271.7455 (3)	591.9469 (1)	133.7654 (2)
	RAE (%)	Solo	99.9616	97.628	99.7319
		Integrated	15.1202 (3)	32.1489 (1)	7.3230 (2)
Hangzhou	RMSE	Solo	588.4369	754.7163	707.954
		Integrated	166.5122 (2)	333.3563 (1)	178.9342 (3)
	MAE	Solo	504.2105	662.5704	579.8735
		Integrated	196.9983 (2)	259.2744 (1)	119.8374 (3)
	RAE (%)	Solo	92.0009	97.6818	108.6852
		Integrated	31.4512 (2)	47.1287 (1)	21.8356 (3)
Chengdu	RMSE	Solo	202.2317	166.581	112.8352
		Integrated	54.4161 (4)	94.5965 (3)	87.1117 (2)
	MAE	Solo	139.501	146.0755	94.3931
		Integrated	66.3225 (4)	75.9824 (3)	70.7321 (2)
	RAE (%)	Solo	94.3513	99.0429	64.6472
		Integrated	38.5105 (4)	51.9446 (3)	47.8395 (2)

https://aisel.aisnet.org/pajais/vol6/iss4/2
DOI: 10.17705/1pais.06403

As can be seen in Table 4, integrated models have much better performances than solo models. For each city and each tool (SVR, BPNN, RBFNN), the corresponding integrated model could yield better prediction than the solo model. Hence we could conclude that our new model based on both online news sentiments and search queries excels in the prediction of real estate price. One possible explanation of the better performance of integrated models could be that, as our model combines two sources of information, the explanation and prediction ability is improved than models with only one source of information. Besides, the weighting factor that we used in the integrated model helps us amplify the useful sentiments and eliminate redundancy. Hence the empirical study based on the samples of four cities shows that the integrated model has much better performance in real estate price prediction.

According to the two empirical findings above, we find that the integration of human searching factor into the three popular types of data mining models (SVR, RBFNN, and BPNN) proves to be successful and meaningful for all the four cities except Chengdu. Since Chengdu is the only one landlocked city in our sample, the real estate price fluctuation is not as volatile as the other two coastal cities and the capital. That is a possible reason for the converse results in the aforementioned analysis that may need further investigation. On the whole, the integrated model we propose can enhance the accuracy of prediction of real estate price. Additionally, the hypothesis that human searching behavior has impact on Chinese real estate market is also tested true by our experiments.

Conclusions and future work

In this paper, we present a novel integrated model for real estate price prediction, and conduct experiments to detect the significance of human searching behavior on real estate price prediction in China.

A fact in our study is that, the house price index data, which are supposed to be the predicted variable, can only be obtained at the city level monthly. In other words, we are not able to examine the real estate market of China as a whole. Instead, we can only examine one city at a time. Therefore, to validate the universality of our model, or in other words, to prove that the integration of human searching behaviors can enhance the accuracy of prediction for different cities, we have conducted experiments on four cities instead of just one particular city. These four cities are chosen as qualified examples.

In the proposed method, we primarily crawl online news articles concerning real estate market in China and generate four types of original sentiment series reflecting the human moods included in the news articles. Next, search engine query data, as an indicator of online users' searching behavior, is integrated into the prediction model by adding weights to the original sentiment series to generate the weighted sentiment series. After that, different types of data mining tools including the SVR model, the RBFNN model and the BPNN model are employed to implement the prediction. The models with inputs of original series are called non-integrated models while those with weighted series are referred to as integrated models. In order to validate the integrated model and test our hypothesis of significance of human behavior, we make a comparison between the non-integrated models with inputs of original series and the integrated ones with weighted inputs. Additionally, a comparison between the integrated model and the solo model (only use search queries) are also made to find out whether the integrated model is better than the solo one. The empirical analysis reveals that the integrated models generally outperform non-integrated ones, which supports the hypothesis that online users' behavior has great value in real-estate price prediction research, and that the integrated model we propose is more effective than the non-integrated one. Also the integrated

model outperforms the solo model with only search queries. In this way we prove that the integrated model with human searching factor and the online sentiments is a better prediction model than model with only news sentiments or with only search queries.

However, there is still room for improvement and further study. Questions such as 'How to improve the quality of data and eliminate the bias of chosen data?', 'Can the four cities really represent the overall situation in China?', 'Can the model be applied to other regions and countries besides China?' and so on should be answered. Most importantly, the process of selecting key words for news articles and weighting the sentiment series should undergo more consideration, as they are crucial steps in our method.

All in all, we believe our model has thrown a new light on the field of prediction. We are eager to apply it to other research areas such as stock market prediction, sales prediction in e-commerce and so many other social science sectors, and we also hope more research could be done to further validate our model presented in this paper.

Acknowledgments

This research work was partly supported by 973 Project (Grant No. 2012CB316205), National Natural Science Foundation of China (Grant No. 71001103, 91224008, 71273265, 91324015), National Social Science Foundation of China Major Program (Grant No. 13&ZD184), Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJA630075), Beijing Social Science Fund (No. 13JGB035), Beijing Natural Science Foundation (No. 9122013), Beijing Nova Program (No.Z131101000413058), and Program for Excellent Talents in Beijing.

References

Anglin, P. (2006). "Local dynamics and contagion in real estate markets," *The International Conference on Real Estates and Macro Economy*, Beijing.

Baffoe-Bonnie, J. (1998). "The dynamic impact of macroeconomic aggregates on housing prices and stock of houses: a national and regional analysis," *The Journal of Real Estate Finance and Economics, 17(2)*, 179-197.

Bardhan, A., Edelstein, R., and Tsang, D. (2008). "Global financial integration and real estate security returns," *Real Estate Economics, 36(2)*, 285-311.

Basak, D., Pal, S., and Patranabis, D. C. (2007). "Support vector regression," *Neural Information Processing-Letters and Reviews*, 11(10), 203-224.

Bollen, J., and Mao, H. (2011). "Twitter mood as a stock market predictor," *IEEE Computer, 44(10)*, 91-94.

Case, K. E., and Shiller, R. J. (1990). "Forecasting prices and excess returns in the housing market," *AREUEA Journal, 18(3)*, 253-273.

Cashin, P., McDermott, C. J., and Scott, A. (2002). "Booms and slumps in world commodity prices," *Journal of Development Economics, 69(1)*, 277-296.

Chang, C. C., and Lin, C. J. (2002). "Training v-support vector regression: theory and algorithms," *Neural Computation, 14(8)*, 1959-1977.

Chang, C. C., and Lin, C. J. (2011). "LIBSVM: a library for support vector machines," A*CM Transactions on Intelligent Systems and Technology*, 2(3), 1-27.

Clapp, J. M., and Giaccotto, C. (1994). "The influence of economic variables on local housing price dynamics", *Journal of Urban Economics, 36(2)*, 161-183.

Das, S. R., and Chen, M. Y. (2007). "Yahoo! for Amazon: Sentiment extraction from small talk on the web," *Management Science, 53(9)*, 1375-1388.

Fingleton, B. (2008). "A generalized method of moments estimator for a spatial model with moving average errors, with application to real estate prices," *Empirical Economics, 34*, 35–57.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2008). "Detecting influenza epidemics using search engine query data," *Nature, 457 (7232)*, 1012-1014.

Grebler, L. (1979). "The Inflation of Housing Price, Its Extent, Cause and Consequences," Lexington Books.

Guyon, I., and Elisseeff, A. (2003). "An introduction to variable and feature selection," *The Journal of Machine Learning Research, 3*, 1157-1182.

Kummerow, M., and Lun, J. C. (2005). "Information and communication technology in the real estate industry: productivity, industry structure and market efficiency," *Telecommunications Policy, 29(2)*, 173-190.

Li, D., Xu, W., Zhao, H., and Chen, R. (2009). "A SVR based forecasting approach for real estate price prediction," *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*, 970-974.

Li, Z., Xu, W., Zhang, L., and Lau, R. Y. K. (2014). "An ontology-based Web mining method for unemployment rate prediction," *Decision Support Systems, 66*, 114-122.

Liu, J., Zhang, X., and Wu, W. (2006). "Application of fuzzy neural network for real estate prediction," *LNCS, 3973,* 1187-1191.

Lu, C. J., Lee, T. S., and Chiu, C. C. (2009). "Financial time series forecasting using independent component analysis and support vector regression," *Decision Support Systems, 47(2)*, 115-125.

Malpezzi, S. (1999). "A simple error correction model of housing prices," *Journal of Housing Economics, 8(1)*, 27-62.

Mei, J. and Liu, C. H. (1994). "The predictability of real estate returns and market timing," *Journal of Real Estate Finance and Economics, 8*, 115-135.

Melville, P., Gryc, W., and Lawrence, R. D. (2009). "Sentiment analysis of blogs by combining lexical knowledge with text classification," *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1275-1284.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., and Tiao, G. C. (1998). "Forecasting the US unemployment rate," *Journal of the American Statistical Association, 93 (442)*, 478-493.

Nellis, J. and Longbottom, J. (1981). "An empirical analysis of the determination of house prices in the United Kingdom", *Urban Studies, 18(1)*, 9-21.

Pace, R. K., Barry, R., Gilley, O. W., Sirmans, C.F. (2000). "A method for spatial-temporal forecasting with an application to real estate prices," *International Journal of Forecasting, 16*, 229–246.

Potepan, M. J. (1998). "Explaining intermetropolitan variation in housing price, rent and land price," *Real Estate Economic, 24(2)*, 219-245.

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). "New support vector algorithms," *Neural Computation,12(5)*, 1207-1245.

Schumaker, R. P. and Chen, H. (2009). "A quantitative stock prediction system based on financial news," *Information Processing and Management, 45*, 571–583.

Seko, M. (2003). "Housing prices and economic cycles," *The International Conference on Housing Market and the Macro Economy*, Hong Kong.

Smola, A. J., and Schölkopf, B. (2004). "A tutorial on support vector regression," *Statistics and Computing,14(3)*, 199-222.

Sun, D., Zhang, C., Xu, W.,Zuo, M., and Zhou, J. (2014) "Does Web news media have opinions: evidence from real estate market prediction," *Proceedings of the 18th Pacific Asia Conference on Information Systems*, Paper 374, 1-12.

Tokgoz, S., Wailes, E., and Chavez, E. (2011). "A quantitative analysis of trade policy responses to higher world agricultural commodity prices," *Food Policy, 36(5)*, 545-561.

Tong, S., and Koller, D. (2002). "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research, 2*, 45-66.

Vapnik, V. N. (1999). "An overview of statistical learning theory," *IEEE Transactions on Neural Networks, 10(5)*, 988-999.

Wang, T., Li, Y., and Zhao, S.(2008). "Application of SVM based on rough set in real estate prices prediction," *IEEE Conference*, 1-4.

Wilson, I. D., Paris,S. D., Ware, J. A., and Jenkins, D. H. (2002). "Residential property price time series forecasting with neural networks," *Knowledge-Based Systems, 15*, 335-341.

Wu, C. H., Tzeng, G. H., and Lin, R. H. (2009). "A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression," *Expert Systems with Applications, 36(3)*, 4725-4735.

Wu, L., and Brynjolfsson, E. (2009). "The future of prediction: how Google searches foreshadow housing prices and quantities," *Proceedings of the 30th International Conference on Information Systems,* Paper 147, 1-14.

Xie, H., Yu, Z., and Wu, J. (2011). "Research on the sustainability of China's real estate market," *Procedia Engineering, 21*, 243-251.

Xie, P. (2008). "Development ranking in real estate strategy management in China based on TOPSIS method," *International Conference on Computing, Communication, Control, and Management, 3*, 254-258.

Xu, W., Li, Z., Cheng, C., and Zheng, T. (2013). Data mining for unemployment rate prediction using search engine query data. *Service Oriented Computing and Applications, 7(1)*, 33-42.

Xu, W., Sun, D., Meng, Z., and Zuo, M. (2012). "Financial market prediction based on online opinion ensemble," *Proceedings of the 16th Pacific AisaConference on Information Systems*, Paper 137, pp. 1-15.

Yan, Y., Xu, W., Bu, H., Song, Y., Zhang, W., Yuan, H. and Wang, S. (2007). "Method for housing price forecasting based on TEI@I methodology," *Systems Engineering - Theory & Practice, 27(7)*, 1-9.

Zhou, W. X., and Sornette, D. (2006). "Is there a real-estate bubble in the US?" *Physica A: Statistical Mechanics and its Applications, 361(1)*, 297-308.

About the Authors

Daoyuan Sun is a Ph.D. student with Department of Information Systems (Econometric Track) at National University of Singapore. He obtained his bachelor degree in Management Information Systems from School of Information, Renmin University of China. His current research interests include crowdfunding, market prediction, big data analytics, and business intelligence.

Yudie Du is an undergraduate student majoring in Applied Mathematics at School of Information, Renmin University of China. Her current research interests focus on econometrics and statistics.

Wei Xu is an associate professor at School of Information, Renmin University of China. His research interests include web mining, business intelligence and decision support systems. He has published over 60 research papers in international journals and conferences, such as Decision Support

https://aisel.aisnet.org/pajais/vol6/iss4/2
DOI: 10.17705/1pais.06403

Systems, European Journal of Operational Research, Fuzzy Sets and Systems, IEEE Trans. Systems, Man and Cybernetics, and International Journal of Production Economics.

Ce Zhang received his bachelor degree in Management Information Systems from School of Information, Renmin University of China. His research interests include business intelligence and decision support systems.

Meiyun Zuo is a professor at School of Information, Renmin University of China. His research interests include knowledge management and HCI.

Junjie Zhuo is an assistant professor at Henan University of Economics and Law. His research interests include knowledge management and E-commence.