Predicting the Evolution of Service Value Features from User Reviews for Continuous Service Improvement

Xu Chi[1], Haifang Wang[1], Zhongjie Wang[1](✉), Shiping Chen[2], and Xiaofei Xu[1]

[1] Harbin Institute of Technology, Harbin 150001, Heilongjiang, China
{chixujohnny,wanghaifang,rainy,xiaofei}@hit.edu.cn
[2] CSIRO DATA61, Eveleigh, NSW 2109, Australia
shiping.chen@data61.csiro.au

Abstract. Facing with a highly competitive service market where customers have more choices on services to fulfill their demands, service providers have to improve their services continuously to make them adapt to constantly-changing value expectations of customers. An enormous quantity of reviews published by customers who have experienced services is an essential basis for service providers to understand which fine-grained features are cared more by customers and what others are less. In this paper, we present a method (VFAMine) for extracting Service Value Features (VF) from review texts by text mining and measuring customers' attention degrees on VFs by sentiment analysis. As a result, a Time-series Service Value Feature Distribution model (TSVFD) is constructed to delineate the evolution history of attention degrees on various VFs. To help providers identify VFs which are to be extensively concerned by customers and improve them in advance, we give a convolutional sliding window and random forest based algorithm (CSRF) for predicting the future trend of the attention degree on one VF, either for a single service or for services belonging to the same region/domain. In terms of Maximum Information Coefficient (MIC) based correlation analysis, we find that there are latent correlations between the evolution history of different VFs, and such correlation would help service providers improve multiple correlated VFs together. Experiments are conducted on a Yelp dataset and the results demonstrate the effectiveness of our approach.

Keywords: Service Value Feature (VF) · Service improvement · User reviews · Attention degree · Evolution trend · Prediction

1 Introduction

More and more services have been deployed on Internet and thus offer a wider range of choices to customers for fulfilling their demands [18]. On the condition that there are abundant mutually-substitutable competitive services for customers to choose, services providers are faced with a great pressure on improving their services to cater to the common value expectations of a larger scale

M. Maximilien et al. (Eds.): ICSOC 2017, LNCS 10601, pp. 142–157, 2017.
https://doi.org/10.1007/978-3-319-69035-3_10

of customers. Besides, customers' demands and preferences evolve as time goes by [16], which requires service providers to continuously improve their services accordingly. User reviews, which are direct feedback submitted by users after they use a service, contain valuable information that have been considered as an important basis for service improvement.

Take a service named `Liholiho Yacht Club` in San Francisco as an example. It joined Yelp in February 2015. Because it is very popular in the local region, in most cases customers have to wait in line in order to enjoy the service; thus it is as expected that most user reviews on Yelp are mainly focused on "wait time". In July 2017, this club improved its service by adding an *online reservation* feature; from then on, its ratings on Yelp increased significantly.

However, as user reviews are numerous and the amount increases drastically along with time, service providers cannot go through all the reviews piece by piece [4]. It is needful to extract high-level valuable information from reviews and offer them to service providers for references. A customer tends to focus on one or several specific features in each of his review. For example, a customer has a review on a restaurant: *The staff was super friendly and most importantly the food was tasty and fresh*, and we see that he cares about `staff` and `food`. We call them Service Value Features (VF). Compared with the numeric ratings (e.g., scores between 1 and 5), VFs delineate more specific facets that customers care about and are considered as "focus of attentions" or "preferences" of customers [15], hence they are useful for service providers to have a deeper understanding on their customers and could be used as important evidences for further improvement on their services. There are diversified VFs in each service, but common VFs across different services in the same domain do widely exist.

Due to constant changes of the service market, changes of user identities and social positions, etc., the scope of VFs that customers are concerned and their attention degrees on each specific VF at different times change frequently, too. In order for service providers to improve their services in advance as far as possible, it is useful to predict the future changing trend of massive customers' attention degrees on VFs and find out those VFs that would have growth spurts in the recent future. In RQ1 of this paper, we focus on such prediction for one service provider with the objective of providing service improvement suggestions to it. In RQ2, we focus on the prediction for a group of services that belong to the same "region" (such as the Bay Area in California) or the same "service domain" (such as `food` and `nightlife`), with the objective of analyzing and predicting the holistic evolution trend of one region or service domain, so that new service providers who would like to enter this region or service domain may have a deeper acquaintance on how to set up their new services to better cater to the future's customer expectations.

Besides, as there are many VFs hidden in user reviews, we wonder if there are latent correlations between the attention degrees of different VFs. For instance, for three VFs $\{f_1, f_2, f_3\}$, when the attention degree on f_1 increases along with time, the one of f_2 increases synchronously, while the one on f_3 decreases synchronously. If such correlations really exist, VFs are to be grouped and VFs in each group should be simultaneously considered when service providers improve their services. If they do like this, the difficulty of service improvement would be significantly decreased. RQ3 of this paper is to present a method of validating the existence of the correlation among the changing trends of attention degrees on different VFs and measure the correlation degrees.

To mine VFs from massive user reviews, we propose a mining algorithm `VFAMine` based on text mining and sentiment analysis. It looks for VFs by analyzing grammatical structures of review texts by applying a set of heuristic rules and represents a VF as one or several keywords, and sentiment analysis is employed to measure the attention degree by emotional factors such as dissatisfaction, criticism, complains or praise exposed in review texts.

For RQ1, we use a machine learning approach and propose a convolutional sliding window method (`CSRF`) to build a model that depict the underlying characteristics of the evolution of the attention degree on one VF in a certain period, then use this model to predict the attention degree on the same VF in the recent future. RQ2 adopts a quite similar approach but the prediction is for a group of services in the same region or service domain. In both RQ1 and RQ2, a time-series prediction accuracy index (*loss*) is used to assess the prediction accuracy. For RQ3, we adopt Maximum Information Coefficient (MIC) to measure the correlation between the evolution history of attention degrees on multiple VFs. Experiments are conducted on a dataset released by Yelp Dataset Challenge[1] (including 800,000 services, 680,000 users, and 2.68 million user reviews published between Jan. 2010 and Jun. 2016), and the results validate the effectiveness of the proposed methods.

In summary, this paper makes the following contributions:

- We define a Time-series Service Value Distribution model (TSVFD) to quantitatively delineate the evolution history of users' attention degrees on VFs. It is a useful tool for the prediction and correlation analysis of VFs.
- Based on the experiments conducted on Yelp dataset, the text mining and sentiment analysis based VF mining method `VFAMine` is effective for identifying VFs from review texts, with the accuracy being more than 86%.
- The prediction model for the future evolution of attention degrees of VFs can reach good performance for both one service provider and multiple providers in one region/domain, and the average loss value is limited within 0.15.
- Certain correlations do really exist between the evolution of different VFs' attention degrees, but the density of highly correlated VF pairs is rather low. The adopted correlation measurement (i.e., MIC) can reach at an accuracy rate above 0.85.

The remainder is organized as follows. Section 2 presents the VF mining algorithm `VFAMine`. Section 3 presents the TSVFD model and gives the method `CSRF` for predicting the evolution trend of one VF. Section 4 gives the correlation

[1] https://www.yelp.com/dataset_challenge

analysis method for the co-evolution of multiple VFs' attention degrees. Section 5 is related work, and Sect. 6 is conclusions and future work.

2 Service Value Feature and the Mining Algorithm

2.1 Service Value Feature (VF)

User reviews contain latent information on how a user looks upon a service, i.e., what features he prefers more when he chooses candidate services to fulfill his demand [16]. If he does not mention a feature in his review at all, there are two possibilities: (1) the performance that the service exhibits on this feature is equal to or beyond his expectation on this feature; (2) he does not care about the performance on this feature. To sum up, user reviews reveal what a user minds and indirectly, what he does not mind; or to say, his value preferences. We define it by "Service Value Features (VF)".

Definition 1. A review is denoted by $r = (s, u, d, text)$, representing that a user u publishes a review r with $text$ in natural language on a service s on the date d.

Definition 2. Service Value Feature (VF). A VF is a noun or a noun phrase describing a specific feature that a service could deliver to its customers, and there is a numeric value associated with the VF to quantitatively measure the degree with which it is concerned by one or a group of customers (namely, attention degree).

Since service are significantly "personalized", different users have quite diversified value propositions and value expectations, thus the VFs hidden in the reviews of different users might be quite diversified. Still, there are some common VFs that multiple users together care about.

2.2 VFAMine: Mining Service Value Features from User Reviews

It is difficult to get value propositions/expectations directly from users, i.e., most users cannot express their preferences explicitly before he uses a service. Only after he uses a service and has got rich experiences on it, does he find out what he cares and what does not. As text mining has been proved to be an effective way of extracting structural information from texts, here we use a heuristic text mining approach to identify VFs from user reviews.

The heuristic rules are straightforward:

- Rule 1: If a noun is modified by one or multiple qualifiers (e.g., adjectives) which appear in a limited range before or after this noun, then there is a significant probability that it represent a VF;
- Rule 2: If two candidate VFs identified by Rule 1 are neighbors in review texts or they are connected by conjunctions such as `of` and `for`, and they are modified by the same qualifiers, then they are combined into one VF.

Although these heuristic rules cannot cover all possible circumstances (especially on the condition that users seldom follow strict grammar rules when they write review texts), our approach tries to reach at a tradeoff between the mining precision and the computation time by avoiding complicated semantics analysis. The mining process includes three steps:

- Review texts are separated into words and part-of-speech (POS) tagging is conducted to give each word a tag such as NN (Noun), NNS (Noun, plural), J (Adjective), JJR (Adjective, comparative), and JJS (Adjective, superlative). This is implemented based on NLTK APIs[2];
- Irrelevant words such as articles (`a`, `an`, `the`) and verbs are removed;
- For each sentence in review texts, above-mentioned heuristic rules are applied and a set of candidate VFs are identified.

Here is an example review on a restaurant service: *The environment of the restaurant is nice, but size of food is too big.* There are four nouns: `environment`, `restaurant`, `size`, and `food`. The nouns `environment` and `restaurant` are both modified by the qualifier `nice` and they are connected by `of`, so they are combined into one VF: `environment of restaurant`. Similarly there is another VF `size of food` modified by the adjective `big`. Here the range where qualifiers are detected is 3–5 words before and after a noun.

After VFs are identified, the next step is to measure the attention degree on each VF. In natural languages, different adverbs or adjectives exhibit different degrees of emotions (e.g., `excellent food` and `good food`), or called emotional factors. For each VF which has been represented by a noun phase, we first extract all the emotional words that appear in a specific range before and after the noun phase, and then look up the emotional factors (ranging from 1 to 5) of these words from the publicly-available emotional term dictionary[3]. The aggregated value of these emotional factors are used to measure the user's attention degree on this VF. If there are no emotional words found around the VF, the value is set to 2.5 (indicating it is a VF with neutral attention degree).

The mining process is conducted on each review r_i correspondingly, and one or several VFs are identified. After all reviews are dealt with (no matter which services they belong to), synonym dictionary based similarity analysis is conducted to merge similar candidate VFs. The final identified VFs are $VF = \{f_1, f_2, ..., f_n\}$. $\forall f_j \in VF$ and a review r_i, v_{ij} is the emotional factor of f_j in r_i, i.e., the attention degree of the user $u(r_i)$ on f_j in r_i. If r_i does not cover f_j, then $v_{ij} = 0$, indicating that $u(r_i)$ does not care about f_j in this review.

To sum up, we identify VFs from review texts and merge similar VFs, and approximately measure the attention degrees of each VF in each review. Due to limited space, we do not show the detailed algorithm for this process.

[2] http://www.nltk.org/api/nltk.tag.html.
[3] http://www.keenage.com/download/sentiment.rar.

We conduct an experiment on Yelp dataset to validate the accuracy of `VFAMine`. We select top-5 hottest service domains in terms of the amount of reviews: `Food`, `Nightlife`, `Shopping Medical`, and `Home-Service`. From each domain we use a random sampling approach to choose 230 reviews, respectively, and manually annotate VFs covered by these reviews. The annotation results are used as the test set. Then, `VFAMine` is applied to automatically identify VFs from the same reviews, and the results are compared with manually-annotated results.

Experiment result: for the five domains, the precisions are 91.3%, 92.1%, 88.6%, 85.4%, and 89.7%, respectively. All precisions keep above 85%, indicating that the mining accuracy is relatively high. The *nightlife* domain receives the highest precision, while the *medical* domain has the lowest precision. As for the reasons, we guess that customers of *nightlife* services are mostly young guys who would like to carefully write reviews on the services they have experienced, while customers of *medical* services tend to be older and they are apt to write shorter reviews which are more unlikely to follow strict grammar rules, and consequently, the accuracy of VF mining is deteriorated.

2.3 Time-Series Service Value Feature Distribution (TSVFD)

To facilitate predicting the changing trend of VFs in the future, for each service s, we construct a matrix $M(s)$ to represent the distribution of aggregated attention degrees by all users on each VF over a long period. It is called Time-Series Service Value Feature Distribution matrix (TSVFD). Table 1 shows the visual form of TSVFD.

Table 1. Time-series service value feature distribution (TSVFD) matrix

	t_1	t_2	...	t_m
f_1	$v^{\Diamond}_{1,1}$	$v^{\Diamond}_{1,2}$	...	$v^{\Diamond}_{1,m}$
f_2	$v^{\Diamond}_{2,1}$	$v^{\Diamond}_{2,2}$	...	$v^{\Diamond}_{2,m}$
...	...	...	...	...
f_n	$v^{\Diamond}_{n,1}$	$v^{\Diamond}_{n,2}$	...	$v^{\Diamond}_{n,m}$

Rows of the matrix are n VFs (i.e., $f_1, f_2, ..., f_n$) identified from reviews of all services, and columns are m consecutive but non-overlapping time intervals with equal lengths (i.e., $t_1, t_2, ..., t_m$), e.g., each t_j is a calendar month. $v^{\Diamond}_{k,j}$ is the aggregated attention degree on f_k in the reviews that are published for the service s during the period t_j and is calculated by:

$$v^{\Diamond}_{k,j} = \sum_{\forall r_i, s(r_i)=s, d(r_i)\in t_j} v_{ik} \tag{1}$$

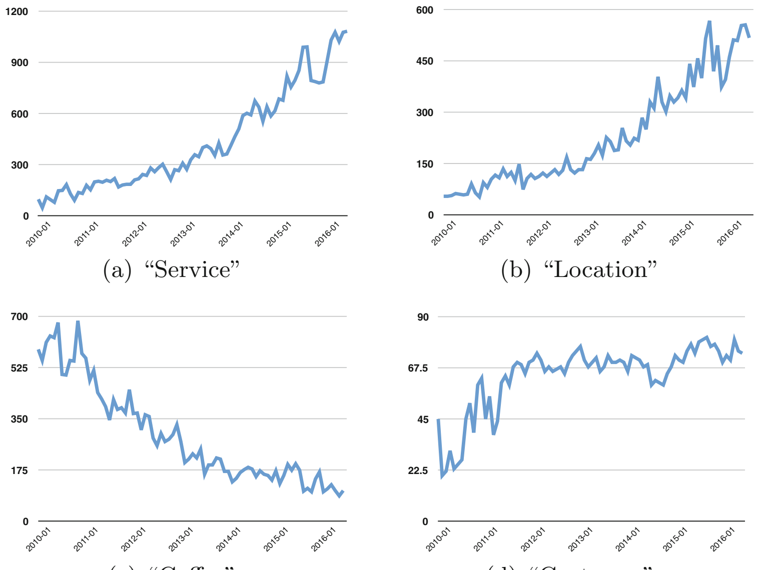
where v_{ik} is the attention degree of f_k in the review r_i.

Similarly, we can construct the TSVFD for a group of services belonging to the same region R or domain D, denoted by $M(R)$ and $M(D)$. The only difference is that when calculating $v^{\diamondsuit}_{k,j}$, we replace the condition $s(r_i) = s$ by $s(r_i) \in R$ or $s(r_i) \in D$, i.e., to group reviews by regions or domains instead of just for one service.

$$v^{\diamondsuit}_{k,j} = \sum_{\forall r_i, s(r_i) \in R, d(r_i) \in t_j} v_{ik} \tag{2}$$

TSVFD of a service/region/domain demonstrates the global view of how massive users care about various VFs at different times, i.e., the changing history of "user concerns". It is used for VF's evolution analysis and prediction in subsequent sections.

Figure 1 shows the evolution of four VFs (abbr. `service`, `location`, `coffee` and `customer`) in the TSVFD of the *nightlife* service domain. The history of how attention degree changes from Jan. 2010 to Jun. 2016 is visualized by line charts. We can see that different VFs shows quite diversified changing trends: the first and the second VFs show an increasing trend, the third VF shows a decreasing trend, while the fourth VF fluctuates with a smaller scale compared with the other three.



(a) "Service" (b) "Location"

(c) "Coffee" (d) "Customer"

Fig. 1. Evolution history of attention degrees of 4 VFs in *Nightlife* domain

3 Evolution Analysis and Trend Prediction of an VF

In this section we focus mainly on one VF and study how to predict the future changing trend of users' attention degree on this VF. In Sect. 4 the focus is switched to the correlation between the evolution history of multiple VFs.

3.1 CSRF: A ML-based Model for VF Evolution and Prediction

In the matrix of TSVFD, the changing history of the attention degree of a VF f_k in m consecutive time intervals is represented by the row vector $V(f_k) = (v^{\diamondsuit}_{k,1}, ..., v^{\diamondsuit}_{k,m})$. Our goal is to predict the values of $(v^{\diamondsuit}_{k,m+1}, v^{\diamondsuit}_{k,m+2}, ..., v^{\diamondsuit}_{k,m+\rho})$, where ρ is the prediction horizon, i.e., the number of future time intervals in which the attention degrees on f_k are to be predicted.

Since $V(f_k)$ is time-series data, this is a typical time series prediction problem. There are many time series prediction models such as ARIMA (Auto Regressive Integrated Moving Average) which has been proved to have good prediction accuracy. Nevertheless, before prediction by ARIMA, manual interventions have to be conducted for stationary handling in case that original data is non-stationary. This is a time-consuming task, especially when there are hundreds of VFs each of which requires prediction. To deal with this issue, we propose a machine learning based method (`CSRF`) which learns the fluctuation characteristics of the time series data, so that the pre-processing and prediction process becomes more efficient.

Specifically, `CSRF` has two phases: a convolutional sliding window model (`CS`) is firstly used to split the time-series data into multiple time-series samples in terms of specific size and step value of sliding windows, and then a random forest regression model (`RF`) is applied on these samples to learn the latent fluctuation patterns and to predict the changing trend of $V(f_k)$ in the recent future. Detailed steps are shown in Algorithm 1.

`CSRF` algorithm has four inputs: V is the row vector for a specific VF whose changing trend is to be predicted; ws_{min} and ws_{max} are the minimal and maximal sizes of sliding windows, respectively; ρ is the prediction horizon (the number of time intervals during which the values of attention degree on the VF is to be predicted); and δ is the step value when V is split into samples by the convolutional sliding window approach.

The outer loop (Steps 3–19) is to predict the attention degree of the VF in the next period, i.e., $v^{\diamondsuit}_{k,m+1}$ where m is the length of current V; in the next loop for predicting $v^{\diamondsuit}_{k,m+2}$, the predicted value in the first loop is added into V (Step 18) and thus its length becomes $m + 1$; the loop continues until all the expected values within the time intervals in the prediction horizon ρ are obtained and recorded in V_{pred} as the output.

The inner loop (Steps 4–17) is to look for a best size of sliding windows that could result in minimal *loss* (measuring the error between the real value and the predicted value) and get the best prediction value *bestPrediction* by looking for the minimal *loss* (see Steps 14–15). In terms of the selected size

Algorithm 1. The CSRF Algorithm

```
Require: $V, ws_{min}, ws_{max}, \rho, \delta$
Ensure: $V_{pred}$
 1: $V_{pred} \leftarrow \emptyset, bestLoss \leftarrow 1, bestPrediction \leftarrow 0$
 2: $train \leftarrow V[: -\rho]$, $test \leftarrow V[-\rho :]$
 3: for $\forall round \in [1, \rho]$ do
 4:     for $\forall w \in [ws_{min}, ws_{max}]$ do
 5:         $Samples \leftarrow \emptyset, y \leftarrow \emptyset, i \leftarrow 0$
 6:         while $i \leq length(train - w - 1)$ do
 7:             $Samples.add(train[i, i + w])$
 8:             $y.add(train[i + w + 1])$
 9:             $i \leftarrow i + \delta$
10:         end while
11:         $Regression\_model \leftarrow sklearn.$RandomForestRegressor$(Samples, y)$
12:         $prediction \leftarrow$ Predict$(Regression\_model, train[-w, ])$
13:         $loss \leftarrow$Loss$(test, prediction)$
14:         if $loss < bestLoss$ then
15:             $bestLoss \leftarrow L, bestPrediction \leftarrow prediction$
16:         end if
17:     end for
18:     $train.add(bestPrediction), V_{pred}.add(bestPrediction)$
19: end for
20: return $V_{pred}$
```

of sliding windows (w in Step 4), Steps 6–10 are to use convolutional sliding window modeling to split V into a set of samples. Each sample is composed of w time-series features (denoted by w columns in Table 2, i.e., $t'_1, t'_2, ..., t'_w$) and a target value (i.e., the last column y in Table 2). All obtained samples (denoted by *Samples* in the algorithm) are used as the train set for training the regression model between the first w attention degrees and the $(w + 1)$-th one. We will discuss the regression process later.

Here we take f_i as an example to demonstrate the process of constructing samples from $V(f_i)$. The first sample starts from the first time interval and ends with the w-th time interval, and the attention degrees are $<v^{\diamondsuit}_{i,1}, v^{\diamondsuit}_{i,2}, ..., v^{\diamondsuit}_{i,w}>$ (see the second row of Table 2), and the value $v^{\diamondsuit}_{i,w+1}$ in the $(w+1)$-th time interval is used as the target value y. Hence, the first sample has been constructed. For the second sample, in terms of the step value δ, it should start from the $(\delta + 1)$-th time interval and ends with $(\delta + w)$-th time interval with the corresponding attention degrees, namely $<v^{\diamondsuit}_{i,\delta+1}, v^{\diamondsuit}_{i,\delta+2}, ..., v^{\diamondsuit}_{i,\delta+w}>$, and $v^{\diamondsuit}_{i,\delta+w+1}$ is used as the target value y. Repeatedly, total $k = m - w - 1$ samples are to be constructed (where m is the length of vectors in the train set of the current loop) and they are shown in Table 2.

Based on the constructed training set (*Samples*), Step 11 is to use random forest as the regression model for training. Here we use `RandomForestRegressor`

Table 2. Constructing samples by convolutional sliding window approach

	t'_1	t'_2	...	t'_w	y
$Sample_1$	$v^{\diamondsuit}_{i,1}$	$v^{\diamondsuit}_{i,2}$	...	$v^{\diamondsuit}_{i,w}$	$v^{\diamondsuit}_{i,w+1}$
$Sample_2$	$v^{\diamondsuit}_{i,\delta+1}$	$v^{\diamondsuit}_{i,\delta+2}$	...	$v^{\diamondsuit}_{i,\delta+w}$	$v^{\diamondsuit}_{i,\delta+w+1}$
...	...	...	...	...	...
$Sample_k$	$v^{\diamondsuit}_{i,(k-1)\times\delta+1}$	$v^{\diamondsuit}_{i,(k-1)\times\delta+2}$	...	$v^{\diamondsuit}_{i,(k-1)\times\delta+w}$	$v^{\diamondsuit}_{i,(k-1)\times\delta+w+1}$

provided by *sklean* ML library[4]) to fulfill this task. Afterwards, Step 12 makes the prediction, and the *loss* is measured by comparing the prediction value and the test set in Step 13.

3.2 Predicting a VF's Future Trend for One Service and for One Region or Service Domain

We first apply the CSRF algorithm on the reviews of one single service and predict the changing trend of a VF's attention degree in the future ρ times intervals. The result would be a valuable reference for the service provider to know which perspectives should be improved with higher priority in the future.

The prediction horizon (the parameter ρ in Algorithm 1) could be of any length, but along with the increasing ρ, the prediction accuracy would decrease drastically. In our experiments, we set $\rho = 6$, i.e., we predict the attention degrees of a VF in the subsequent $1^{st}, 2^{nd}$, ..., and 6^{th} months, respectively.

The CSRF algorithm could also be applied on the reviews of services belonging to the same region or the same service domain to predict the changing trend of a VF's attention degree, so that new service providers who would like to join this region or domain may have a clear acquaintance on how to set up their new services to better cater to user expectations. Compared with the prediction for a service provider, this prediction involves a broader range of services.

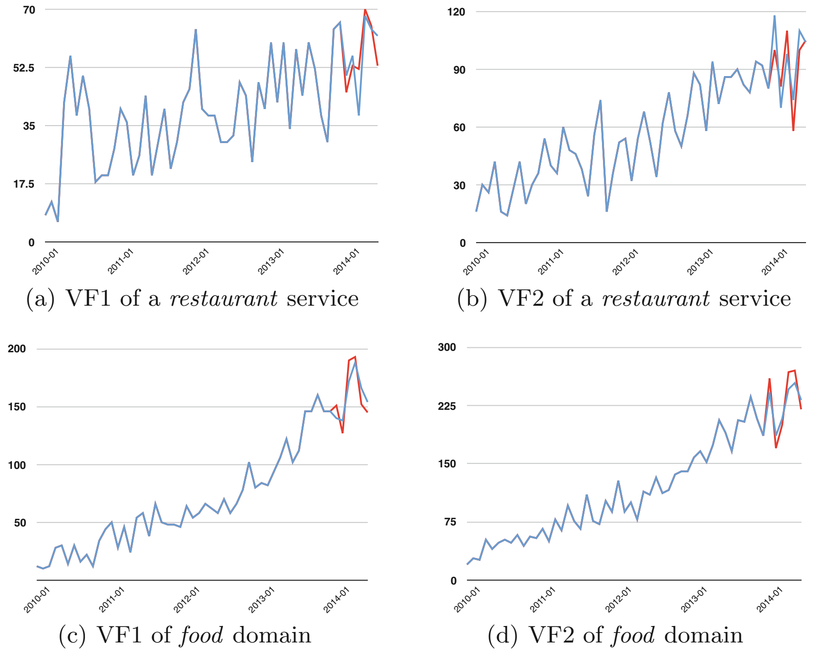
Figure 2(a) and (b) shows the prediction results of two VFs of a restaurant service, and Fig. 2(c) and (d) is the results of two VFs from the food domain. Blue lines are the evolution history of VFs' attention degrees, and red lines are the prediction results.

To evaluate the prediction accuracy, we use the *loss* metrics which measures the aggregated errors between the prediction and the real values of all the features in one service or in all services belonging to the same region/domain:

$$loss = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{t=1}^{m} \left| \frac{v^{\diamondsuit}_{i,t} - v^{P}_{i,t}}{v^{\diamondsuit}_{i,t} + v^{P}_{i,t}} \right| \tag{3}$$

where $v^{\diamondsuit}_{i,t}$ is the real value of the attention degree of the i-th VF in the t-th month, and $v^{P}_{i,t}$ is the corresponding prediction value.

[4] http://scikit-learn.org.



(a) VF1 of a *restaurant* service

(b) VF2 of a *restaurant* service

(c) VF1 of *food* domain

(d) VF2 of *food* domain

Fig. 2. Comparison between the actual evolution history and the prediction

Figure 3 shows the distribution of *loss* values for multiple services in each service domain. In the five domains, the values are mostly distributed in the range $[0.08, 0.15]$ with the medium in the range $[0.11, 0.13]$. This indicates that the prediction results are accurate and acceptable. For comparison, the `Food` domain has more amount of reviews and thus has higher prediction accuracy than other domains, and we do find that `CSRF` performs better on the services that have more reviews than on the services having fewer reviews.

4 Correlation Analysis for the Evolution of Multiple Service Value Features

4.1 MIC-based Correlation Analysis on Multiple VFs

We conjecture that the evolution of multiple VFs might not be absolutely independent but sometimes they are correlated, i.e., there is a phenomenon called "co-evolution of VFs". In this section we validate whether this hypothesis is true. If it is valid, it is possible to group highly-correlated VFs together so that service providers would improve them holistically and consequently, more efficiently.

The first step is to determine which correlation measure is suitable for this goal. Pearson correlation coefficient[5], Kenall's rank coefficient[6], and Spearman's

[5] https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.

[6] https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient.

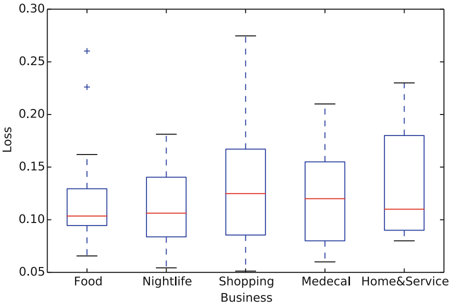


Fig. 3. Distribution of *loss* values for VFs in 5 service domains

rank correlation coefficient[7] are all widely-adopted correlation measures. However, because a VF's attention degree usually evolves periodically, there would be nonlinear correlations between the evolution history of different VFs, and unfortunately, above three correlation measures are all weak in handling such nonlinear correlation. If they are applied in this scenario, some closely-related VFs might be considered as weakly- or non-correlated ones.

Here we use MIC (Maximal Information Coefficient)[8] to measure the correlation between non-linearly correlated VFs. In statistics, MIC is a measure of the strength of the linear or non-linear association between two random continuous variables X and Y. It uses binning as a means to apply mutual information on X and Y, i.e., $I[X;Y] = \int_Y \int_X p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$, and the rationale is that the bins for both variables should be chosen in such a way that the mutual information between the variables be maximal.

MIC coefficient falls in the range $[-1, 1]$, and the sign of MIC (i.e., <0 or >0) indicates whether it is negative or positive correlation. If the correlation between the evolution history of attention degrees of two VFs falls in the range $[0.8, 1]$ or $[-1, -0.8]$, the two VFs are closely correlated.

In terms of one service or one region/domain, for arbitrary two VFs f_i and f_j and their corresponding attention degrees' time-series evolution vectors $V(f_i)$ and $V(f_j)$, we calculate their MIC correlation coefficient $MIC(f_i, f_j)$ using the MIC API provided by `minepy` library[9]. In order to evaluate the effectiveness, similar as the approach in Algorithm 1, we split each $V(f_i)$ into train set and test set, measure $MIC(f_i, f_j)$ on the two sets separately, and then compare their results to measure *precision*, *recall* and F1-score, respectively.

4.2 Experiments

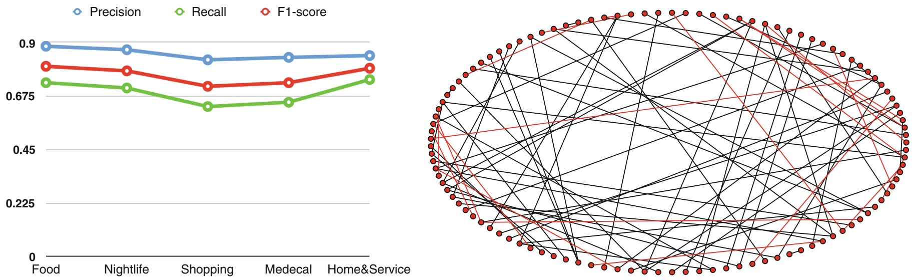
We select top-100 popular services from five domains and conduct MIC-based correlation analysis on their VFs. We manually identify and label some corre-

[7] https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.

[8] https://en.wikipedia.org/wiki/Maximal_information_coefficient.

[9] https://pypi.python.org/pypi/minepy.

lations, then compare them with the analysis results of the proposed approach to measure the performance. It is shown in Fig. 4(a). The *precision* is above 85% in average, the *recall* is above 65% in average, and F1-Score is above 0.75, indicating that MIC has good performance for correlation analysis.



(a) Precision/Recall/F1-score (b) A Partial Correlation Network

Fig. 4. Experiment results of VF correlation analysis

Experiment result shows that there are 73.86% VF pairs that have no correlation or are weakly correlated ($|MIC| \in [0, 0.2)$, 23.59% VF pairs being slightly correlated ($|MIC| \in [0.2, 0.4)$), 1.62% VF pairs being moderately correlated ($|MIC| \in [0.4, 0.6)$), 0.72% VF pairs being relatively correlated ($|MIC| \in [0.6, 0.8)$), and only 0.21% VF pairs being closely correlated ($|MIC| \geq 0.8$). This shows that such time-series correlation between the attention degrees' evolution of different VFs does really exist but the density of highly or moderately correlated VF pairs is fairly low, and most of VF pairs keep relatively independent. Because of this, it is of great significance to identify those highly correlated VF pairs from a mass of VFs and recommend them to service providers, so that these VF pairs are to be considered simultaneously when services are improved.

Specifically, Table 3 demonstrates detailed distribution of various correlation levels in five popular domains in Yelp. Chi-square test shows there is no significant difference among the MIC distribution in different domains, indicating that different domains exhibit similar correlation characteristics among their VFs.

Table 3. Distribution of MIC correlation coefficient between VFs in 5 domains

$\lvert MIC\rvert$	Food	Nightlife	Shopping	Medical	Home & Service
[0.8, 1]	0.21%	0.20%	0.12%	0.18%	0.15%
[0.6, 0.8)	0.72%	0.83%	0.64%	0.92%	0.79%
[0.4, 0.6)	1.62%	2.06%	1.98%	2.59%	2.28%
[0.2, 0.4)	23.59%	26.09%	24.32%	30.33%	20.01%
[0, 0.2)	73.86%	70.55%	72.94%	64.98%	76.77%

Here are two examples of highly related VF pairs: a VF `flavor` is positively correlated with another VF `size of food`, while `size of food` is negatively correlated with the VF `price`, and the correlation degree between `flavor` and `ize of food` (0.893) is higher than the one between `size` and `price` (−0.831).

Another interesting phenomenon is that, in terms of those closely correlated VF pairs, there are about 77.9% correlations being positive ones, and only 22.1% correlations being negative ones; for those relatively correlated VF pairs, the two numbers are 71% and 29%, respectively. This can be observed from Fig. 4(b) which is partial correlation network among VFs in the `food` domain. Closely correlated VF pairs with $|MIC| \geq 0.8$ are connected by lines, red lines are for negative correlated VFs, and black lines are for positive correlated VFs. The ratio of positive ones is much higher than the ratio of negative ones.

5 Related Work

Values are the ultimate goal that providers and customers expect to get from delivering and using a service. Provider value is often exhibited by the earning from the economics perspective, while customer value concerns mainly with experiences and satisfactions, i.e., whether and to what degree a service could meet a customer's demand. Zeithaml [19] defined customer value as "customers overall assessments on products". Gale [7] defined customer value as the relative price in market and product quality adjustment. Patricio et al. [13] proposed a multilevel service design method which takes customer value into full consideration. On the other hand, research on how to improve provider value seems inadequate. Wang et al. [17] suggested that dynamic service selection and composition should consider costs and earning of service providers. Wancheng et al. [14] put forward a competitive mechanism in commodity market to maintain balance in service selection by pricing based on member services. Chawathe et al. [3] proposed a method of distributing combined service earnings.

Mining user reviews has attracted wide attentions in previous research. Using available training corpus from open websites where each review has been appointed a class (e.g., thumbs-up and thumbs-downs, or some other quantitative or binary ratings), Hu et al. [9] designed and experimented a number of methods for building sentiment classifiers of reviews. Eirinaki et al. [6] presented a method to mine users' opinions from blogs and social network. Zhang et al. [20] gave a method to extract entity from users' opinions. To analyze the sentimental opinion expressed in a review, sentiment analysis techniques are typically conducted at two levels: (1) in the document level: to distinguish positive reviews from negative ones [2]; (2) in the sentiment level or phrase level: to perform tasks such as multi-perspective question answering and summarization, and opinion-oriented information extraction [11]. However, these methods are of limited usefulness for deriving useful information to represent the value features of services that are cared by customers.

Time-series correlation analysis is an important issue in data mining [1,5] and is applied in various domains, e.g., Kumar et al. [10] adopted the ARIMA algorithm to forecast the ambient air pollutants and achieves good performance; Gao et al. [8] used the random forest regression model to predict the volume of railway freight. CNN for extracting and modeling samples is also used in time-series data mining to create convolution sliding window modeling method [12].

6 Conclusions and Future Work

Numerous user reviews on third-party service platforms such as Yelp are a great treasure for service providers to collect valuable feedbacks from customers so as to improve their services. We propose two methods (`VFAMine` and `CSRF`) to help service providers extract Service Value Features (VFs) from review texts, quantify the evolution history of the attention degrees on these VFs (e.g., TSVFD), and predict the future trends of their attention degrees. They are not only useful for a single service provider to improve his service in advance in terms of user concerns (e.g., the VFs with increasing attention degrees in the future), but also for providers who plan to enter a new region or a new service domain to be full aware of the trend of massive users' attention degrees on specific VFs. Based on MIC-based correlation analysis, we also find that the evolution of different VFs are sometimes closely correlated.

Future work include: (1) Deep semantics analysis techniques are required to further improve the precision and recall of `VFAMine` (currently only some simplified heuristic rules are used); (2) After the prediction of a VF's attention degree in the future time intervals is obtained, how are service providers to be given more specific suggestions for improving the VF? (3) A method for grouping VFs in terms of the MIC correlation degree between them is required, and operational suggestions on how to take highly-correlated VFs into consideration at the same time during service improvement is of significance to service providers, too.

Acknowledgments. This work is supported by Natural Science Foundation of China (No. 61772155, 61472106)

References

1. Bankó, Z., Abonyi, J.: Correlation based dynamic time warping of multivariate time series. Expert Syst. Appl. **39**(17), 12814–12823 (2012)
2. Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. **28**(2), 15–21 (2013)
3. Chawathe, S.S.: Strategic web-service agreements. In: International Conference on Web Services, pp. 119–126. IEEE (2006)
4. Chen, N., Lin, J., Hoi, S.C., Xiao, X., Zhang, B.: AR-Miner: mining informative reviews for developers from mobile app marketplace. In: International Conference on Software Engineering, pp. 767–778. ACM (2014)
5. Dorr, D.H., Denton, A.M.: Establishing relationships among patterns in stock market data. Data Knowl. Eng. **68**(3), 318–337 (2009)
6. Eirinaki, M., Pisal, S., Singh, J.: Feature-based opinion mining and ranking. J. Comput. Syst. Sci. **78**(4), 1175–1184 (2012)

7. Gale, B., Wood, R.C.: Managing Customer Value: Creating Quality and Service that Customers can See. Simon and Schuster, New York (1994)
8. Gao, J., Lu, X.: Forecast of china railway freight volume by random forest regression model. In: International Conference on Logistics, Informatics and Service Sciences, pp. 1–6. IEEE (2015)
9. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
10. Kumar, U., Jain, V.: ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). Stoch. Env. Res. Risk Assess. **24**(5), 751–760 (2010)
11. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: International World Wide Web Conference, pp. 342–351. ACM (2005)
12. Papandreou, G., Kokkinos, I., Savalle, P.A.: Modeling local and global deformations in deep learning: epitomic convolution, multiple instance learning, and sliding window detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 390–399 (2015)
13. Patricio, L., Fisk, R.P., Cunha, J.F., Constantine, L.: Multilevel service design: from customer value constellation to service experience blueprinting. J. Serv. Res. **14**(2), 180–200 (2011)
14. Wancheng, N., Lingjuan, H., Lianchen, L., Cheng, W.: Commodity-market based services selection in dynamic web service composition. In: IEEE Asia-Pacific Service Computing Conference, pp. 218–223. IEEE (2007)
15. Wang, H., Chi, X., Wang, Z., Xu, X., Chen, S.: Extracting fine-grained service value features and distributions for accurate service recommendation. In: International Conference on Web Services. IEEE (2017)
16. Wang, H., Wang, Z., Xu, X.: Time-aware customer preference sensing and satisfaction prediction in a dynamic service market. In: Sheng, Q.Z., Stroulia, E., Tata, S., Bhiri, S. (eds.) ICSOC 2016. LNCS, vol. 9936, pp. 236–251. Springer, Cham (2016). doi:10.1007/978-3-319-46295-0_15
17. Wang, X.Z., Xu, X.F., Wang, Z.J.: A profit optimization oriented service selection method for dynamic service composition. Chin. J. Comput. **33**(11), 2104–2115 (2010)
18. Yao, L., Sheng, Q.Z., Segev, A., Yu, J.: Recommending web services via combining collaborative filtering with content-based features. In: IEEE International Conference on Web Services, pp. 42–49. IEEE (2013)
19. Zeithaml, V.A.: Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. J. Mark. **52**, 2–22 (1988)
20. Zhang, L., Liu, B.: Aspect and entity extraction for opinion mining. In: Chu, W. (ed.) Data Mining and Knowledge Discovery for Big Data. Studies in Big Data, vol. 1, pp. 1–40. Springer, Heidelberg (2014). doi:10.1007/978-3-642-40837-3_1