2017 13th International Conference on Semantics, Knowledge and Grids

Dynamic Sale Prediction of the Time-Sensitive Products Based on Conformity

Qingwei Liu, Ming Han, Ling Feng

Dept. of Computer Science and Technology, Tsinghua University, Beijing 100084, China
{liuqw10, hanm13}@mails.tsinghua.edu.cn; fengling@tsinghua.edu.cn

***Abstract*— In these years, many online shopping and investment becomes more time-sensitive such as crowdfunding, online auction, time-limit sale and group buying. Most of them have deadlines. Consequently, people need to make faster decisions than other products leading to more conformity in this area. Though there is much work about the conformity and social influence in social networks, there lacks work of conformity in the online shopping and investment, because in these scenarios there is usually no such "strong" connection. However, we think the conformity information is also important for the sale prediction of time-sensitive products, since conformity impacts people's purchase decisions. In this paper, we propose a model which captures the dynamic conformity and apply it to the similarity computation and sale prediction of time-sensitive products. It can be dynamically adjusted to fit the target product and the prediction moment, which means it is well self-adaptive. Since it also gives the consideration to other static and latent features in a collaborative way, it alleviates the "cold start" problem nicely and is free from the presetting hypotheses. Experiments on the real large-scale dataset shows that our method exceeds other comparison methods by 33.2% in the sale prediction.**

I. Introduction

Nowadays, more and more online products and investments have deadlines which means their values are much more time-sensitive. Some products have the deadlines due to their commodity instinct nature such as crowdfunding, group buying and online auction, while others have deadlines because it is the sellers' promotion means to stimulate the market and improve their products sale, such as the time-limit sale.

The fact of some online shopping becoming more time-sensitive also changes people online shopping behaviors and leads more conformity in these areas. We argue that when people shop online, their purchase decisions not only come from their preference but also from their estimation of the product value at the moment of buying.This estimation consists of rational and irrational parts. The inclination of conformity makes people more irrational when they estimate the value of time-sensitive product. Because people need to make the decision faster, so they feel more pressure of time urgency and more likely refer to others [1].

Though there is much work about the conformity or social influence in social networks, there lacks work of conformity in the field of online shopping and investment. Compared with social networks, in online shopping, there lacks the "strong" connection such as "following" in twitter. When shopping online, people are usually influenced by the former buyers through the comments or the rates, and they are usually strangers to each other. Besides, the conformity effect is more implicit than the influence in social networks: people even don't know how venerable they are to others' actions and they usually effected by the conformity unconsciously. So it is hard to capture the conformity information in online shopping. Besides, we argue the conformity stimuli come from various aspects, and people's conformity response also reflects on various actions. As there is no public dataset with enough information for the conformity computation, no work about applying conformity to sales prediction has been proposed.

However we think it is important to bring the conformity into online shopping, especially for the time-sensitive products. Because we think it is a predictive factor for their sales. In the personal level, people's conformity effects their final purchase decision. In the market level, the conformity distribution of the customer group determines the group's response for the following consumption stimuli.

Besides, the conformity information and sale prediction also benefit all the trade participants. For sellers, if they foreknow which kind of buyer group their products are more prone to attract, they can take pertinent actions or bring some referral incentives [2] to improve their sales. For customers, since they can not get in touch with the products when shopping online, they usually need to make investment decisions based on limited static information. So the dynamic prediction based on the conformity can help them to make their decisions more wisely [3]. Besides, knowing their and others' conformity levels can also remind them to make less irrational decisions. For the platform administrators, by knowing the conformity information of customers can help them to coordinate their resources in time to optimize their utilization.

In this work, we propose a novel model for time-sensitive products, which can predict the sales at any time of its duration by adopting the conformity related features. Firstly, it measures individual conformity levels and combines them to get the group distribution. Then along with other static and latent features, it finds the most similar neighbors of the target one. Finally, it combines the *known* neighbors' sale trends and their similarities to predict the sale trend of target product.

Our model adopts a collaborative way to predict, it unifies both dynamic and static, explicit and implicit features under one framework to compute the similarity between two products. Consequently, it is easy to expand the model to incorporate new features. Besides, the model we propose follows a data-driven schema, which means it needs no presetting

0-7695-6304-X/17/31.00 ©2017 IEEE
DOI 10.1109/SKG.2017.00023

assumptions and can be adaptive to any other dataset with the similar setting. Since our prediction model is based on the item-item similarity computation, for those newcome products, though we know little about its temporal information, we can also infer its future sale trend based on its neighboring products' history trends. So it alleviates the "cold start" and data sparsity problem to some extent.

The structure of this paper is following: In Section 2, we present the related work. Section elaborates our model step by step. Next in Section 4, we describe our dataset including its main features and the preprocessing technique. Then in Section 5, we demonstrate our method on the real-world dataset. Finally, we conclude this study in Section 6.

II. Related Work

We present the related work from two respective: the conformity related research in different fields and sale prediction.

Conformity. Conformity was first studied in psychology and economics. Kelman et al. identified and categorize the conformity in [4], then Bernheim et al. presented a study of modeling the conformity process in [5]. After that, considerable qualitative work about the social influence has been conducted. Until 2009, Tang et al. firstly proposed an approach to quantify the peer influence in large network [6], and in 2013, they furthered their work by formally defining the problem of conformity influence analysis and propose a method to address this problem. Besides, Goyal et al. used the number of correlative social actions to learn the influence probabilities [7]. Tan et al. studied the effects of influence, correlation and users' action dependency [8]. Zhang et al. presented a method to learn the social influence locality and study how people' behaviors are influenced by their friends in their ego networks[9]. Li et al. studied the correlation between the users' conformity and the influence[10]. However, these work mainly focus on the social networks which has the explicit "strong" connection relationship. In online shopping and investment, most of buyers do not know each other and their influence happens via a "weak" connection like a message after the transition or clicking the like button to express the approval.

Sale Prediction. The main prediction methods in economics are based on the causal forecasting which takes account of past relationships between variables and sale: if one variable has been approximately linearly related to sale for a long period of time, it may be appropriate to extrapolate such a relationship into the future. And the most popular method used in this area is regression analysis.As the machine learning booming there years, some researchers try to apply machine learning models to solve the sale prediction problem. For example, Y Yi et al. adopted SVR model to predict the tobacco sale [11], and Thiesing et al. used neural networks to predict the sale of supermarket [12]. However, most of these methods are domain-constricted and do not adapt to the online shopping scenario. As far as we know, there is no work on the sale prediction of the online time-sensitive products. Due to the variance of the online shopping market situation, we need a novel way to fit the sales of these time-sensitive products.

III. Model Framework

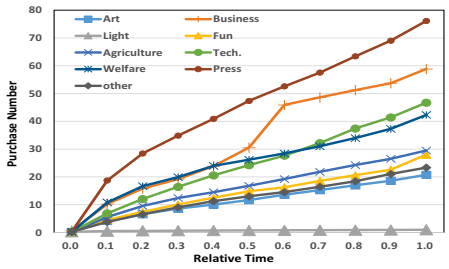


Fig. 1. Purchase Trend in Different Categories.

Since the crowdfunding is the most representative field of time-sensitive online investment, in this study we mainly use the application in crowdfunding to elaborate our model. Fig. 1 is the sales growth trend for each category in $zhongchou.com$. We can see that different categories have distinct trend patterns and the products from the same category have similar patterns. Inspired by this observation, we assume that *the more similar any two products are, the more possibility they have similar sale growth trends*. So we can use the information of similar *known* products to infer the *unknown*.

Our algorithm framework includes three parts: conformity quantification, similarity computation and sale prediction (Fig.2). Firstly we measure the conformity level of each customer and for each product we combine its customers conformity levels at each moment to get the dynamic conformity distribution. Then for the product to be predicted at the moment t, we first find the most similar neighbors by comparing their historical conformity distribution, conformity stimuli dynamics as well as other static and latent attributes. Then in a collaborative manner, we use the similar products trends after timestamp t to predict the follow-up change of the target product by timing the similarity.

A. Conformity Distribution

According to Donelson Forsyth [13], after submitting to group pressures, individuals may find themselves facing one of several responses to conformity, and these types of responses to conformity vary in their degree of public agreement versus private agreement which can be seen to vary along a continuum from conversion [1] to anticonformity. Our goal is to predict the target product sale trend by utilizing the purchase group response to conformity . So we first measure each customer conformity level by his/her historical purchase log. Then for each moment t, we combine the individual conformity levels to get the real-time conformity distribution for the target product.

We argue people's response to conformity influence their purchase decisions, so we use the temporal information of products and the market as well as interactive history to infer customers' conformity responses in the reverse direction. The

[1]Change that occurs when group members personally accept the influencer's position [13].

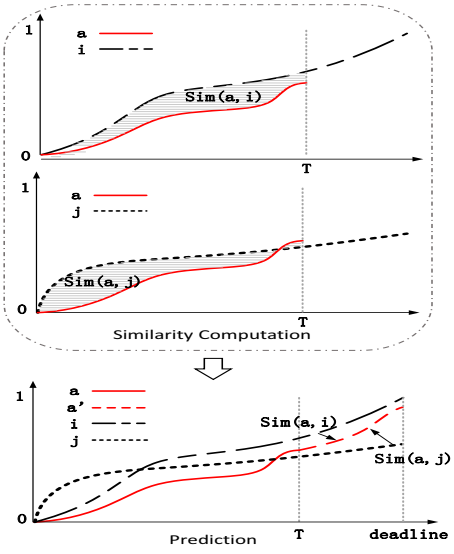


Fig. 2. Product Similarity Computation and Sale Prediction. Line *a* stands for the known sale trend of the target product, Line *a'* stands for the sale to be predicted, Line *i* and Line *j* stand for the known neighbors sale trends.

features which can reflect customers conformity are following (*u* represents the customer and *i* represents *i*th product):

- **Dose the product have a deadline?** Though we focus on the products with deadline, we can infer the customer social response from all the products no matter whether it has a deadline or not. If the customer buys more products with deadline, as discussed above, we think there is more chance that his/her purchase decision comes from irrational evaluation. We use a boolean value E_i to represent this information which $E_i = 1$ if the product has a deadline, otherwise $E_i = 0$.
- **When did he buy in the product duration?** We think the temporal factor is a very important indicator for the customer conformity response. Intuitively, the later a customer buy the product, the more chance he will be prone to wait and see the market situation and influenced by the crowd attention. So we think the individual social response has positive correlation to the purchase relative time of the duration which is represented as $X_i(t) = \frac{t - T^i_{start}}{T^i_{end} - T^i_{start}}$, where T^i_{start} and T^i_{end} are the start time and end time of the i's duration.
- **How much are stimuli when he buys?** If there are more stimuli when the customer buy some product, he is more prone to rely his decision on these stimuli. So we think the amount of stimuli $Y_i(t)$ is also relative to the individual social response. To measure it, we weightedly sum the various stimuli including the number of i's purchase, share, comment and like at time t: $Y_i(t) = a * N^p_i(t) + b * N^s_i(t) + c * N^c_i(t) + d * N^l_i(t)$.
- **Is it hotter compared with other products in the same category?** Besides the influence from the product *inner* information, the *outside* information also can indicate the degree of social response. For example, if a customer buys some unpopular product when there are many other hotter ones in the same category, s/he may have a higher degree of independence. Thus we think the social response is correlative to the product popularity which can be represented as $Z_i(t) = \frac{1}{Rank_i(t)}$ where $Rank_i(t)$ is the popular rank of product i measured by $Y_i(t)$ in the same category.

vspace-3.5mm

Individual Conformity Response. The individual conformity response is defined as the average degree of his/her purchase conformity propensity, which reflected by the combination of features above and the money s/he spends on the product. Specifically, after quantifying the features above, for each record of each customer, we apply z-score standardization to scale them: $X_i(t)' = \frac{X_i(t) - \mu_x}{\sigma_x}$, $Y_i(t)' = \frac{Y_i(t) - \mu_y}{\sigma_y}$, $Z_i(t)' = \frac{Z_i(t) - \mu_z}{\sigma_z}$; then we time their linear combination by the money(also scaled by z-score standardization $M'_u(i) = \frac{M_u(i) - \mu_m}{\sigma_m}$) s/he spends which represents the degree s/he is influence by the conformity. More precisely, we get the following formula to measure individual conformity response:

$$P_u = \sum_{i=1}^{n} \sum_{t \in T_i} (\theta * E_i + \alpha * X_i(t)' + \beta * Y_i(t)' + \gamma * Z_i(t)') \cdot \frac{M'_u(i)}{n \cdot |T_i|} \quad (1)$$

where n is the number of customer's purchase record and T_i is the set of purchase time for product i.

Inspired by the work of [14], [15], we argue that when the number of customers increases to enough extent, the conformity response of purchase crowd presents the normal distribution. So to estimate the parameters involved above, we adopt the normal distribution to fit the conformity response. In doing so, we get the value of parameters: $a = 0.7$, $b = 0.1$, $c = 0.1$, $d = 0.1$, $\alpha = 0.6$, $\beta = 0.2$, $\gamma = 0.2$, $\theta = 0.1$. After obtaining social response of each customer, we rank and bin them into five levels of conformity defined below.

Individual Conformity Level. Inspired by the work of [13], [4], we classify the responses to conformity of online customer into five levels according with the conceptions of *independence, convergence, compliance, identification* and *internalization* (in this study we use L1...L5 to present them). The *independence* (L1) stands for the lowest conformity response bin, while *internalization* (L5) stands for the highest one. For example, if the conformity responses of a group is [0.15, 0.18, 0.45, 0.51, 0.61, 0.7, 0.79, 0.97] then the according conformity levels are [L1, L1, L3, L3, L4, L4, L4, L5].

Conformity Distribution. For each product, its conformity distribution is represented as a five-tuple $\vec{D} = [N_{independence}, N_{convergence}, N_{compliance}, N_{identification}, N_{internalization}]$, which combines the number of five-level social responses. Firstly, we initial it to $\vec{0}$. Then if a new customer has some interactive behavior with the product at time t (such as purchase or comment), the corresponding part of $\vec{D}$ which accords with his/her social response level increases by 1. Loop this procedure we can get the conformity distribution for

product i at time t. For instance, the conformity distribution of above example is [2, 0, 2, 3, 1].

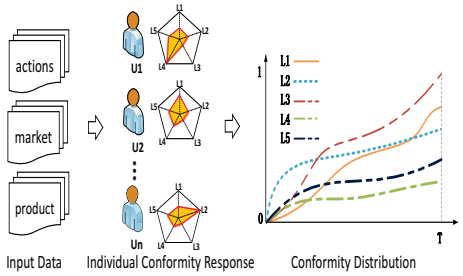


Fig. 3. Conformity Distribution Computation. L1 to L5 refer to the numbers of costumer whose social response levels are *independence, convergence, compliance, identification* and *internalization* respectively.

B. Conformity Stimuli

An online product may have many predictive conformity stimuli, and they have more powerful indication when treated in a dynamic respective. We classify these features into three categories: spread dynamics on social network, customer dynamics and product dynamics.

Spread Dynamics on Social Networks. We can treat an online product as a set of "activities" which comes from the participation of crowd in various roles: sellers, customers and platform administrators, and more people awareness means more opportunity to sale success. So the spread dynamics on social networks which directly impact the crowd awareness has important influence on the prediction and similarity computation [16], [17] accordingly.

Customer Dynamics. The attitude from the purchase crowd also impacts the product sale [18],and the public opinion and sentiment are reflected on different aspect: explicitly on the *pledge* or *like* [16], [17], implicitly on the comments [16]. Intuitively, a faster increase on these features stands for more probability of sale success and the more similar increase trends means more similarity on this feature.

Product Dynamics. *Updates* is also an important tool of managing a campaign or product sale. The form of a product update is like a blog post which is used to "keep backers (funders) informed of a product's progress" [19]. As Anbang Xu and Jisun An proposed in [16], [20] respectively, a high frequency of *updates* reflects the seller effort and the intimate interaction between sellers and customers which means that the specific utilization of *updates* has a strong associations with campaign success and the similarity computation.

To sum up, we add the content variance of share, pledge, like, comment and update to $\overrightarrow{D}$(the detail quantification process is presented in Section 5).

Besides, like what Yehuda Koren did in [21], we not only consider the explicit content of the stimuli , but also take the numbers of them as the implicit supplement. For example, the dataset not only tell us the pledge values, but also *which* products the customer buy, regardless of *how* they support these products. So we add the number variance of share, pledge, like, comment and update to $\overrightarrow{D}$.

C. Conformity Similarity

Inspired by the cross-correlations between stocks [22], We combine these explicit and implicit conformity attributes to give the similarity definition of conformity features:

Definition 1: Assume D^k is the kth conformity feature, the similarity on D^k between product i and j at time t is defined as:

$$Sim_{D^k}(i,j,t) = \frac{\langle R_i R_j\rangle - \langle R_i\rangle\langle R_j\rangle}{\sqrt{|\langle R_i^2 - \langle R_i\rangle^2\rangle\langle R_j^2 - \langle R_j\rangle^2\rangle|}} \quad (2)$$

where $R_i(t') \equiv \ln D_i^k(t') - \ln D_i^k(t'-1)$ and $t' \in (1,t]$. $R_i(t')$ is the logarithm change of the D^k value for product i. The angular brackets indicate a time average over the period from the beginning to time t. By this definition, the $Sim_{D^k}(i,j,t)$ range from -1 to 1, where 0 indicates no correlated changes and 1/-1 indicate completely correlated/anti-correlated changes on D^k between product i and j.

Intuitively, the conformity attribute similarity measures "*how close two conformity-relative feature trend lines are*" (Fig.2). For example, if some dynamic conformity attribute of the target product (such as comment number) increases as the one of its counterpart, we think they are similar in this attribute until the prediction moment.

D. Other Features

Because the conformity distribution and stimuli are dynamic features, when few or even no *known* data is fed into the model, the similarity accuracy suffers a lot. To overcome this shortage and improve effectiveness, our model also includes some static and latent features.

Static Features. As shown in [23], many static features are predictive for the sale success too, so we combine the static features listed in Table IV as the supplement to the similarity computation. Intuitively, the static attribute similarity measure "*How close the essences of two product are*". From the practice functionality, the static attributes can be classified into two category: description and sale features.

(a) Description Features. The description features present the inherent features presenting "*what the product is and how it looks like*". They have three main functions:*classification description* and *design*. As observed in our two-year dataset, products from different categories have different trend in the pledge (Fig. 1), so it is natural to consider the product with same category or labels are more similar. As shown in [24], it is important to describe a product persuasively in order to make crowd want to invest, so we use the title, abstract and text as the static feature for description. Besides, the design features (such as picture and video number) also impact the customer behavior [16], [25] and influence both community and business developments. Besides, it also impact the diversity of sellers' thoughts. So we take them into consideration when computing the static similarity.

(b) Sale Features. The sale attributes reflect the product fixed information which depict the sale. Firstly, the economic

factor is an indispensable sale feature, it is natural to think that products with comparable goals and reward are more likely similar to some extent. Besides, the economic factors also impact investors behaviors [16]. Thus, we include them into sale features. As shown in [16], the geography of crowdfunding plays an important role during the campaign, for example,"*A local product is likely to be supported by occasional investors*[2]". Obviously, products from the same or similar initiators are more likely have similar features and as shown in [26] the experience of the initiators accounts for the success of product to an extent, so we argue the location and initiator information are important sale features. Besides, the timing factor affects the dynamics of product investors [27], which indirectly influence the product success, so we include the *launch time* and *deadline*.

Static Similarity. Like what people did in the recommender system field, we use the cosine correlation to compute the similarity between each two static features and synthesize them to infer the final static similarity.

Definition 2: Assume S^l is the lth static feature, the similarity on S^l between product i and j is defined as:

$$Sim_{S^l}(i,j) = \frac{S^l(i) \cdot S^l(j)}{\|S^l(i)\| \|S^l(j)\|} \tag{3}$$

Latent Features. We argue that though not reflected on the explicit content, some latent features are also important for the similarity calculation. For example, there may be some products that have strong attraction for teenagers while no explicit content with this information is available on the homepage. In this study, we extract the latent feature from the action sets of *pledge*, *like* and *comment*. So if two products are invested, liked or commented by many investors from the same group, it means that they have more common attraction to that group and there is more latent similarity on that feature between them correspondingly.

Latent Similarity. This kind of feature is hard to represent in a readable way, however, we can use the common selection of the same crowd to mine these latent features of product and compute the similarity based on them [28].

Definition 3: Assume $U_i(t)$ is the investor set of product i at time t and L^q is the qth latent feature, the similarity between product i and j on L^q is defined as:

$$Sim_{L^q}(i,j,t) = \frac{|U_i(t) \cap U_j(t)|}{|U_i(t) \cup U_j(t)|} \tag{4}$$

E. Model Learning

Coherence Parameter. Inspired by the idea of collaborative filtering in recommender system [21], we linearly combine similarities above to compute the final similarities between the target product and its neighbors. We assume that *the pledges of similar products increase in a similar way*, so we use the pledge money trends of similar products weighted by the combination of similarities to estimate the target product money trend:

[2]Occasional investors are those who funded less than four products.

TABLE I
MAIN PARAMETERS LIST

Notation	Meaning
a	product to be predicted
$Sim_{D^k}(i,j,t)$	kth conformity similarity of i and j at time t
$Sim_{S^l}(i,j)$	the lth static similarity of i and j
$Sim_{L^q}(i,j,t)$	the qth latent similarity of i and j at time t
$\hat{M}_a(t)$	sale estimation for a at time t
$M_a(t)$	real sale for a at time t
$G(i)$	sale goal of i
T	end time of training set
$H^v(a)$	top k similar neighbors of a
γ	the decay parameter of data importance

$$\begin{aligned} Sim(a,i,t) = &\sum_{l=1}^{F} w_l \cdot Sim_{S^l}(a,i) + \sum_{k=1}^{V} w_k \cdot Sim_{D^k}(a,i,t) \\ &+ \sum_{q=1}^{Q} w_q \cdot Sim_L(a,i,t) \end{aligned} \tag{5}$$

$$\hat{M}_a(t) = \frac{\sum_{i \in H(a)} Sim(a,i,t) \cdot M_i(t) \cdot \frac{G(a)}{G(i)}}{\sum_{i \in H(a)} Sim(a,i,t)} \tag{6}$$

Where $H^v(a)$ is the set of v neighbors most similar to a determined by the similarity measure. F, V and Q are the number of static, conformity and latent features respectively, and t is the timestamp between start time 0 and the end of training time T. From the Eq. 6 we can see that for the product to be predicted, the more similar one neighboring product is, the more "contribution" it offer to the estimation, and we use $H^v(a)$ instead of the whole product set to reduce the complexity of the model.

Cost Function. To find the fittest coherence parameters for the target product a at time T, cost function is as follows:

$$\min_{W_a} \sum_{t=1}^{T-1} \left(\frac{M_a(t) - \hat{M}_a(t)}{M_a(t)}\right)^2 \cdot e^{\gamma(t-T)} + \lambda \|W_a\|^2 \tag{7}$$

Here we use exponential decay formed by the function $e^{\gamma(t-T)}$ [29] to penalize the old data so that the recent data plays a more important part in the parameter training process. Thus the parameter W_a is specific to the training time T and the product a, and to overcome the overfitting problem, we add the two norm of W to the cost function.

Gradient Descent Solver. The least square solvers from standard linear algebra packages is optional for this convex optimal problem. However, the following gradient descent solver works faster. We loop through all known records from time 0 to T. For a given training case $M_a(t)$, we modify the parameters by moving in the opposite direction of the gradient, yielding:

$$\frac{\mathrm{d}\hat{M}_a(t)}{\mathrm{d}w} = \frac{\sum_{i \in H(a)} Sim_x \cdot (M_i(t) \cdot \frac{G(a)}{G(i)} - \hat{M}_a(t))}{\sum_{i \in H(a)} Sim(a,i,t)} \tag{8}$$

$$w \leftarrow w - \alpha \cdot \left(\sum_{t=1}^{T-1} \frac{\mathrm{d}\hat{M}_a(t)}{\mathrm{d}w} \cdot e^{\gamma(t-T))} \cdot \left(\frac{\hat{M}_a(t) - M_a(t)}{M_a(t)^2}\right) + \lambda \cdot w\right) \tag{9}$$

Where Sim_x refer to $Sim_{S^l}(a,i)$, $Sim_{D^k}(a,i,t)$ and $Sim_L^q(a,i,t)$ when w represents w_l, w_k and w_q respectively, and the meta-parameters α (step size), λ and γ are determined by cross-validation. Since the intermediate results of similarity computation are used repeatedly, to accelerate the computation, when implementing the solver we store the similarity values in the first place.

Prediction. With the coherence parameters, we can predict the pledge of target product after time T by utilizing the trends of similar products. Like the estimation of formula Eq.6, the prediction function is following:

$$\hat{M}_a(T) = \frac{\sum_{i \in H(a)} Sim(a,i,T-1) \cdot M_i(T) \cdot \frac{G(a)}{G(i)}}{\sum_{i \in H(a)} Sim(a,i,T-1)} \quad (10)$$

We use the same static feature similarity computed on the training set. For the latent and dynamic feature similarity, we use the data right before the prediction moment T.

IV. DATASET

A crawler is built to collect the real-world dataset in $zhongchou.com$, in which we gathered all the information of projects and customers who are explicitly involved with those project, spanning from February 2013 to March 2015 and the information obtained includes funding goal, description, location, rewards, time-stamping updates, launch time and deadline. What's more, we also collected all the information of interactions such as launch, comment, like and pledge. Finally, we have gathered about 6,477 projects which funded by 117,698 investors with a total number of 184,514 pledges. The projects are classified into 9 categories, and the most popular ones are light crowdfunding[3], entertainment, technology and public benefit.It is worthy to mention that our dataset has sufficient temporal information, more concretely, each interaction record has a timestamp accurate to second.

	Successful	Failed	Ongoing	Total
Projects	1,883	4,091	503	6,477
Proportion	29.1%	63.1%	7.8%	100%
Investors	-	-	-	117,698
Initiators	-	-	-	4605
Pledges	146,566	23,620	14,328	184,514
Pledged(CNY)	65,710,159	3,346,806	2,005,515	71,062,481
Likes	369,902	183,202	21,855	574,959
Shares	26,562	14,317	9,282	50,161
Updates	6,288	2,387	628	9,303
Time Span	-	-	-	736 days

TABLE II
STATISTICS FOR THE ZHONGCHOU DATASET.

Besides, we also collected the sharing information on 122 main SNS platforms in China such as Weibo, Weixin, Qzone and Facebook. By doing so, we get a total of 50,161 shares.

Table II and Table III report the general statistics of our dataset and specific project information respectively. The numbers in Table III are the average of all projects. Among the 6,477 projects, 1,883 projects (29.1%) were successfully funded, 4,091 (63.1%) failed to achieve their pledging goals and the rest was still ongoing by the end time of crawling. For these projects, 71.1M yuan were pledged at last.

[3]Projects launched from mobile and spreading on social networks

	Successful	Failed	Ongoing	Total
Goal(CNY)	32,699	26,122	90,023	32,996
Duration(days)	33.69	35.43	57.18	36.61
# of investors	60.57	5.03	24.80	22.71
Pledge per user(CNY)	576.14	162.64	160.77	483.11
Final amount	106.72%	3.13%	4.42%	33.25%

TABLE III
STATISTICS FOR THE ZHONGCHOU PROJECTS

.

V. EXPERIMENT

In our model, the conformity features are used to predict the sale trend and compare with other regression models.

A. Data preprocessing

Feature Quantification. Since the Zhongchou dataset have various data types (Fig. IV), we cannot directly put them into our model or comparison methods. So apart from the ordinary data cleaning (such as removing noisy or redundancy data, correcting error and filling the missing data), we need to quantify the data into structured forms.

Short and Long Text. The Zhongchou dataset has many string-type features and we classify them into *short text* and *long text* according to the number of words. The *short text* have several words including title, abstract and the descriptions of updates, comments and rewards. We use the TextRank model to extract keyword vector from them. Then we remove the keywords that appear only once, since they have no discrimination in similarity computation and classification. In doing so, every feature can be mapped into a vector where each tuple represents the number of the responding word in that position. Finally, we apply the PCA algorithm to these vectors to further reduce the dimensionality. For the *long text* such as the project description, we use the LDA model to reduce the themes and map each description into a theme vector, and the rest of work is the same with *short text*.

Categoric, Temporal and Boolean type. For the categoric features such as location and category, we map them into numeric code ranging from 1 to the size of feature range. Besides, we map the boolean into 0/1 and map the temporal feature to wall-clock time with the unit of second.

Composite Type. As shown in Fig. IV, there are four composite features consisting of several sub-features. We unfold them into a plain form, then use the methods above to quantify each of them.

Sentiment Relative Type . For the comment update, we not only consider their frequency but also use their sentiment as complement. For each tuple of comment or update, we analyze it positive possibility[4]. Then we put them along the time line to get the sentiment trend for each project.

[4]The SnowNLP is applied for sentiment analysis.
[5]Including level number, description, pledge amount, sending date, postage.
[6]It includes investor id, pledge money and timestamp.
[7]It includes critic id, comment content and timestamp.
[8]It includes update content and timestamp.

TABLE IV
SCRAPED AND CALCULATED FEATURES.

Description		Sale		Conformity	
Attribute	**Type**	**Attribute**	**Type**	**Attribute**	**Type**
title	String	*goal*	Double	*pledge*	Composite[5]
category	String	*launch time*	Date	*#pledge*	Integer
abstract	String	*deadline*	Date	*#like*	Integer
labels	String	*reward*	Composite[6]	*comment*	Composite[7]
project text	String	*initiator id*	Integer	*#comment*	Integer
# picture	Integer	*location*	String	*#share*	Integer
has_video	Boolean			*update*	Composite[8]

B. Sale Prediction

In this part, we conduct two experiments to demonstrate the conformity related features' effectiveness in our model. First we compare our model with other regression models from different aspects and then analyze the factor contribution to find how much important every component in our model is.

Experiment Setup. Since we want to predict the sale of each project individually, different from the data partition in the classification, we divide every project history data into independent training and test set. We split every project history into four equal length segments, and we pick the last one as the test dataset. We use the first one, first two and first three segments as the training set respectively so as to find the effect variation from different amount of training set. The metric we use in the prediction problem is the Root Mean Square Error (RMSE), and we evaluate it on every bin in the test segment and use the average value as the final measurement.

Regression Analysis: We regard the time as independent variable and the sale as dependent variable, then we use the regression analysis to estimating the relationships between them and predict the future sale.

SVR: The Support Vector Regression (SVR) is a version of SVM for regression. Y Yi et al. used it to predict the tobacco sale, so we try it on our crowdfunding dataset.

Gaussian Process: Since many features in our dataset present the power-law distribution which can be transformed into Gaussian distribution. So we also adopt the Gaussian processes to our dataset as the comparison method.

CART: CART constructs binary trees with the features and thresholds yielding the largest information gain at each node. We use CART as the representative of decision tree regression models for its simplicity to interpret and little data preparation.

Nearest Neighbors Regression: Since our model is similar with Neighbors-based regression: they both pick up nearest neighbors to predict the target value in a collaborative way, we also take the Nearest Neighbors Regression (NNR) as our comparison method. Because our model selects the k nearest products of the target, where k is an integer specified in advance, so we employ another type of Nearest Neighbors Regression—the Radius Neighbors Regressor, using neighbors with a fixed radius, as our comparison.

KRR: Kernel Ridge Regression (KRR) combines Ridge Regression with the kernel trick. It learns a linear function in the space induced by the respective kernel and the input data. Due to its popularity in regression problem, we also include it as our counterpart method.

Performance Analysis. From the Fig. 4, we can see that our method is 33.2% better than the best comparison method in general. For different training/test ratios, we find that our method is 36.5%, 28.3% and 34.8% better than the best of comparison methods. In terms of different aggregation granularity of the timestamp, we can see that our method is about 29.4%, 43.2% and 27.0% better than the best of the rest. So the experiment result presents that our method is good at predicting the sale of next week, next day and next hour no matter how much training data are fed into.

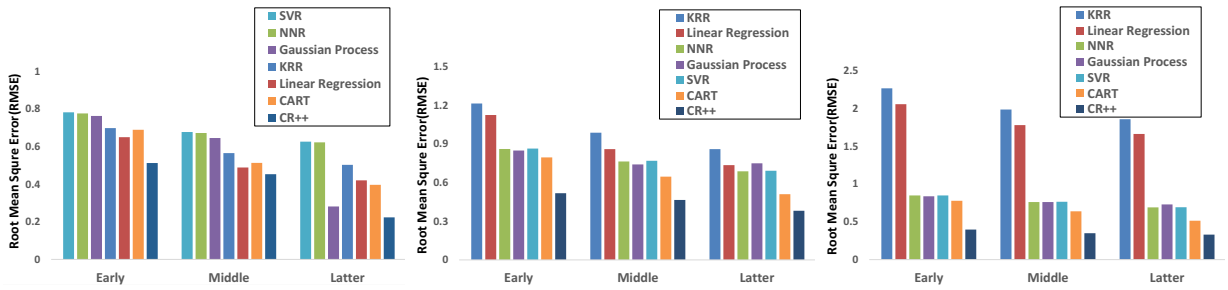
Factor Contribution. Our model is based on the similarity computation between two projects, and the similarity come from three parts: (1) conformity-related features, (2) static features and (3) latent features. Besides, the conformity-related features can be further divided into the conformity stimuli and the conformity distribution. In order to find how much contribution of each component has, we also conduct four experiments with different factors. All the experiments are conducted on the day-level dataset with the same amount of training dataset as the first two segments. As the Fig. 5 shows, with more factors involved in the similarity computation, the prediction performance of the model get better: the RMSE is reduced by 0.02, 0.11 and 0.16 respectively. Besides, our model with only the stimuli features can still get the result comparable with others: about 6.91% higher than the decision tree and better than the rest.

Discussion. The prediction result (Fig. 4) shows that with the complexity of prediction problem increasing (finer time granularity), it is more difficult to predict the sale trend, so the hour level aggregation experiment has the highest prediction error. Our method improves the prediction performance in all three time aggregation situations, which means our methods is less vulnerable to the problem complexity. Besides, it also demonstrates our assumption: people are more prone to conformed after knowing the deadline so the conformity information is relative to the prediction of the sale trend.

From the experiment of factor contribution analysis, we find that each factor has positive effect for the prediction and the combination of them can synthesize their contribution. It is also interesting to find that compared with static features adding the customers, conformity distribution and latent features bring significant improvement. From this we can learn that the interactive information has more indicative effect for the sale prediction than the static information.

VI. CONCLUSION

When faced with time-sensitive products, people need to make the purchase decision quickly. So they may feel more pressure and more easily conformed by others' opinions. In this study, we extract the conformity stimuli and compute the customer group conformity distribution. Then we utilize these conformity related information as well as other static and latent features to predict the sale. The core of our work is the computation of conformity distribution: we first compute individual conformity level based on the customer and market



(a) week (b) day (c) hour

Fig. 4. Prediction Result. *Early Middle and Latter* stand for different prediction start point in the project duration. *CR++* stand for our model.

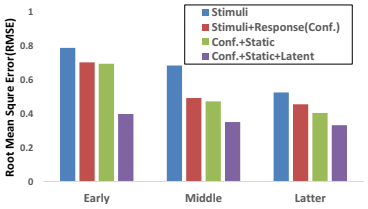


Fig. 5. Factor Contribution. *Conf* stands for the features related to conformity including the conformity stimuli and the conformity distribution.

history information, then we synthesize them to get the product conformity distribution. Our model adopts a collaborative framework which finds the nearest neighbors to infer the sale trend, weighted by the similarity. We construct a real-world dataset which collected information from *zhongchou.com*. Then we predict the sale trend and decompose our model to analyze the factor contribution. The result shows that our method exceeds over other comparison methods by 33.2%. So we can conclude that the conformity is positive for the improvement of sales prediction and our model based on the conformity has good effectiveness.

REFERENCES

[1] G. Vaughan and M. A. Hogg, *Introduction to social psychology*. Pearson Education Australia, 2005.

[2] V. Naroditskiy, S. Stein, M. Tonin, L. Tran-Thanh, M. Vlassopoulos, and N. R. Jennings, "Referral incentives in crowdfunding," 2014.

[3] F. Thies and M. Wessel, "The circular effects of popularity information and electronic word-of-mouth on consumer decision-making: Evidence from a crowdfunding platform," 2014.

[4] H. C. Kelman, "Compliance, identification, and internalization: Three processes of attitude change," *Journal of Conflict Resolution*, vol. 2, no. 1, pp. 51–60, 1958.

[5] B. D. Bernheim, "A theory of conformity." *Journal of Political Economy*, vol. 102, no. 5, pp. 841–77, 1994.

[6] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," *Proceedings of Acm Sigkdd Conference on Knowledge Discovery and Data Mining*, 2009.

[7] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 241–250.

[8] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang, "Social action tracking via noise tolerant time-varying factor graphs," in *In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD?10*, 2010.

[9] J. T. J.Zhang, B.Liu and J.Li, "Social influence locality for modeling retweeting behaviors," in *IJCAI 13*, 2013.

[10] S. B. H.Li and A.Sun, "Casino: towards conformity-aware social influence analysis in online social networks," in *CIKM 11*, 2011.

[11] Y. Yi, F. Rong, H. Chang, and Z. Xiao, "Svr mathematical model and methods for sale prediction," *Journal of Systems Engineering & Electronics*, vol. 18, no. 4, pp. 769–773, 2007.

[12] F. M. Thiesing and O. Vornberger, "Sales forecasting using neural networks," in *Neural Networks, 1997., International Conference on*, vol. 4. IEEE, 1997, pp. 2125–2128.

[13] D. R. Forsyth, "Methodological advances in the study of group dynamics." *Group Dynamics Theory Research Practice*, vol. 2(4), no. 4, pp. 211–212, 1998.

[14] G. Casella and R. L. Berger, "Statistical inference," *Oxford University Press New York*, vol. 25, no. 98, pp. xii,328, 1995.

[15] S. B. Gall, B. Beins, and A. Feldman, "The gale encyclopedia of psychology," *Gale Encyclopedia of Psychology*, 1996.

[16] J. An, D. Quercia, and J. Crowcroft, "Recommending investors for crowdfunding projects," in *Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee, 2014, pp. 261–270.

[17] E. Mollick, "The dynamics of crowdfunding: An exploratory study," *Journal of Business Venturing*, vol. 29, no. 1, pp. 1–16, 2014.

[18] V. Kuppuswamy and B. L. Bayus, "Crowdfunding creative ideas: The dynamics of project backers in kickstarter," *UNC Kenan-Flagler Research Paper*, no. 2013-15, 2014.

[19] J. Hui, E. Gerber, and M. Greenberg, "Easy money? the demands of crowdfunding work," *Proceedings of the Segal Technical Report: 12*, vol. 4, p. 2012, 2012.

[20] A. Xu, X. Yang, H. Rao, W.-T. Fu, S.-W. Huang, and B. P. Bailey, "Show me the money!: An analysis of project updates during crowdfunding campaigns," pp. 591–600, 2014.

[21] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," pp. 426–434, 2008.

[22] N. Rosario, H. Mantegna, and E. Stanley, "An introduction to econophysics: Correlations and complexity in finance," 1999.

[23] M. D. Greenberg, B. Pardo, K. Hariharan, and E. Gerber, "Crowdfunding support tools: predicting success & failure," pp. 1815–1820, 2013.

[24] Y.-F. Kuo and L.-T. Liu, "The effects of framing and cause-related marketing on crowdfunding sponsors' intentions: A model development," in *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia*. ACM, 2014, pp. 439–443.

[25] P.-Y. Kuo and E. Gerber, "Design principles: crowdfunding as a creativity support tool," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2012, pp. 1601–1606.

[26] E. Harburg, J. Hui, M. Greenberg, and E. M. Gerber, "Understanding the effects of crowdfunding on entrepreneurial self-efficacy," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 3–16.

[27] J. Solomon, W. Ma, and R. Wash, "Dont wait! how timing affects coordination of crowdfunding donations."

[28] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 285–295.

[29] Y. Koren, "Collaborative filtering with temporal dynamics," *Communications of the ACM*, vol. 53, no. 4, pp. 89–97, 2010.