IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

A Cross-Distribution User Intention Identification Based on Topic Transfer

To cite this article: Shan-shan Wang 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **490** 062059

View the article online for updates and enhancements.

This content was downloaded from IP address 110.16.80.106 on 15/10/2019 at 00:23

A Cross-Distribution User Intention Identification Based on Topic Transfer

Shan-shan Wang*

Anhui Institute of International Business, Hefei, 230011, China

*Corresponding author e-mail: 258693224@qq.com

Abstract. At present, in the field of financial science, so as to make intelligent recommendation,the potential financial intention is usually analyzed by the behavior patterns of users. However, due to the diversity of user' behavior and distribution on different platforms, the feature of users' financial behavior are also composed of weak features. In this paper, a cross-distributed user' behavior recognition method based on topic transfer model is proposed. To training a transfer model with APP data in source domain, the model can weaken the influence of different distributions and recognize the intention of the target user. Finally, based on the user's APP online data, we prove that the model based on topic transfer has better performance on cross-distributed data.

1. Introduction

In the traditional financial field, obtaining the financial intention of users mainly depends on the strong features and the experience of financial practitioners. With the development of artificial intelligence, the weaker features are lower in cost than the strong features, and the financial intentions of a large number of users are hidden behind the weak features in Internet. However, due to the diversification of the platform and the distribution of user feature, the majority of machine learning algorithms are based on the assumption of independent and identical distribution [1]. Therefore, how to accurately find the financial intention hidden in the weak features has become a very important project in the financial technology field.

Transfer learning can deal with these scenarios by leveraging the already existing labeled data of some related task or domain. To try to store this knowledge gained in solving the source task in the source domain and apply it to our problem of interest as can be seen in Figure1.

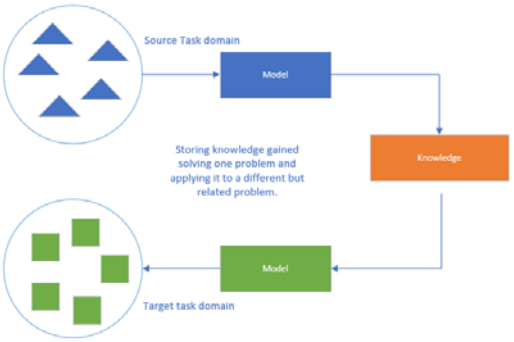

Figure 1. Transfer learning

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Given a source domain $D_S = \{X_s, f_s(X)\}$ and learning task T_s, a target domain $D_t = \{X_t, f_t(X)\}$ and learning task T_t, transfer learning aims to help improve the learning of the target predictive function $f_t(\cdot)$ in D_t using the knowledge in D_s and T_s Where $D_s \neq D_t$ or $T_s \neq T_t$.

According to the idea of Transfer learning [2], We propose a model based on topic Transfer, which is used to transfer the knowledge of user behavior and to recognize user intention in different distribution of users. The main idea is to train a topic model using the user's APP behavior data of the source domain, using the topic to represent the higher level semantics of user behavior. Through this topic models to transfer the knowledge, the target domain can be represented through the same semantics. At this time, the influence of different distributions will be weakened or even eliminated.

2. LDA topic model

We select LDA[3] (Latent Dirichlet allocation) to build the topic model. In 2003, the LDA was first proposed by Blei[4] and used widely for document topic modeling. In LDA, each document is represented as a mixed distribution of *K* implicit topics, and each topic is a multinomial distribution over *W* words. The probability graph is shown in Figure 2:

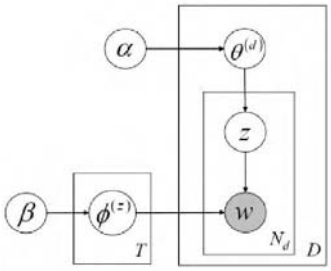

Figure 2. The probability graph of the hidden topic model LDA

The main idea of the LDA model is to view each article as a mixed probability distribution of all topics, and each topic is regarded as a probability distribution on a word. The process of the LDA probability topic model is as follows:

(1) For the topic *z*, the polynomial distribution vector ϕ on the topic of a word is obtained based on the Dirichlet distribution $Dir(\beta)$;
(2) The number of the words *N* obtained from the Poisson distribution *P*;
(3) According to the Dirichlet distribution $Dir(\alpha)$, a topic distribution probability vector θ of the document is obtained;
(4) For each of the *N* words in the document, W_n:
   (a) Randomly select a topic z from the polynomial distribution Multinomial (θ) of θ;
   (b) Select a word as W_n from the polynomial conditional probability distribution Multinomial(φ) of the subject *z*.

In this paper, the user-installed APP is considered as a word, and the user's APP daily is considered as a document. The APP data installed by the user during the time period T is treated as a document; APP can be processed with reference to text, extract key-APPS under *z* topics, and generate user topic vectors.

3. Recognizing intention based on topic transfer model

Transfer learning is to mine the interconnections between different distributions from a higher level and to transfer of knowledge from cross-domain by discovering potential shared subspaces. First, We propose a high-level semantic representation of APP data of user. Then, explain how to build cross-distribution topic features. Finally, the topic features are transferred to recognize the user intention.

3.1. Data representation

In the Internet finance, the data of user behavior is huge and different between platforms. We use APP data as a user behavior feature, which has no direct connection with the user's financial intention. Use the improved bag of word model feature to describe the user intentional features in cross-distribution.

The traditional bag of word model feature uses the dictionary length |V| as the dimension of one-hot vector, adding the time's dimension T, the user's APP features are a matrix which has the dimension |T|*|D|, where each row represents the t day, the user-installed APP list corresponds to the APP category n, and is set to 1 if the user installs such an APP. Otherwise set to 0; each column shows the number of days used by the nth APP user, as shown in Figure 3 as a matrix of the bag of word feature. The source and target domain users' APP features are converted into a matrix composed of APP types and time. Because the underlying features are less robust and vulnerable on cross-distribution interference, the topic transfer model can solve cross-distribution problems very well.

	APP_1	APP_2	APP_3	APP_4	APP_5	……	APP_{n-2}	APP_{n-1}	APP_n
T_0	1	1	0	1	1	……	0	0	0
T_1	0	0	0	1	0	……	1	0	1
T_2	1	0	0	0	0	……	0	0	1
……	0	1	0	1	0	……	1	1	1
T_{n-2}	0	0	0	1	0	……	0	1	1
T_{n-1}	0	1	0	1	0	……	1	0	1
T_n	1	0	0	1	1	……	1	1	1

Figure 3. Character matrix of user's bag of words

3.2. Topic Transfer Model

This model is mainly used to establish transitive high-level semantic features. As shown in Figure 4, the establishment of topic transfer model mainly has the following three steps:

Step1: According to the above, the APP data in the source and target domains are featured by an improved bag of word model. The user's APP data is represented as a time span of *T* and the APP type is the bag of word form of *D*. The source domain user APP data is described as $X_s = \{\chi_{s1}, \chi_{s2,\ldots}\chi_{sd}\}$.The target domain user APP data is described as $X_t = \{\chi_{t1}, \chi_{t2,\ldots}\chi_{td}\}$.

Step2: In the source domain, the topic model is trained using the APP bag feature and it can obtain the probability distribution of each user under the number of topics which are defined *N*. Specifically, LDA is used to generate the topic of the user's APP and discover potential interests of the user; Z topic generated regarded as high-level semantics which can be delivered in different distribution. User behavior is abstracted into a corresponding topic, and it can be described as $Z_s = \{Z_{s1}, Z_{s2,\ldots}, Z_{sn}\}$.

Step3: Although the source domain and the target domain belong to the different distribution, but by improving the bag of word model, APP user data has become a consistent feature of the uniform bag of word. Therefore, through the LDA topic model, the bag of word features in the target domain can be directly translated into the high-level semantics of the source domain, so that the target domain feature data Z_t and the source domain feature data Z_S obey the uniform distribution in high-level semantics.

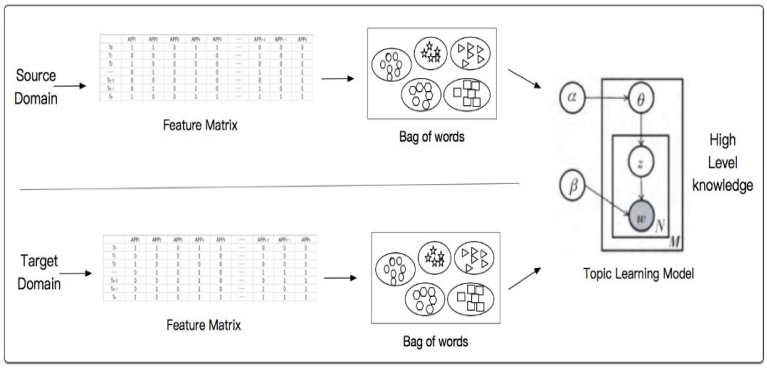

Figure 4. The topic transfer learning framework

3.3. Recognize the Intention

Recognize the Intention refers to training a classifier using the high-level semantic features Z_S converted by the topic model, thereby recognizing and classifying user intentions in the target domain feature Z_t . In this paper, Boosted Tree [5] is used as a classifier to train on the source domain topic

feature Z_S and label collection that was defined L. Using the trained classifier to identify the Z_t which is the topic feature of the target domain, and to recognize the intentions of different users.
The topic-based transfer model:

(1) Extract the original feature of the source and target domain, and generate the word of bag model feature based on the improved bag of word model;
(2) Train the LDA topic model in the source domain, obtain the topic of the source domain data, and calculate the topic feature of source domain;
(3) Use the bag of word feature of target to calculate target domain topic feature;
(4) Training a classifier in source domain and testing the user intention in target domain.

4. Experiments

This paper selects the APP list in real scene as the test data, in which the source domain 40,000 users have a positive and negative ratio of 1:1 and the time span is one month; the target domain has 30,000 users, the positive and negative ratio is 1:1, and the time span is one month. Select recall rate, precision rate, accuracy, and AUC (Area Under Curve) as performance measures.

The source domain and target domain are distributed from user APP data collected by different platforms. First, experiments verify the impact of cross-distribution on the classification in non-subject transfer models. The experimental results are shown in Table 1.

Table 1. Bag of word's feature on cross-distribution

	Recall Rate	Precision	Accuracy	AUC
Source domain	0.71	0.73	0.71	0.70
Target domain	0.60	0.61	0.63	0.60

Secondly, experiments verify the impact of topic transfer model with different topic numbers for the performance of classification. The experimental results are shown in Table 2. It can be seen that after 300 topics, the growth of the performance is no longer obvious, so this paper selects 300 as the dimension of topic transfer feature.

Table 2. The topic feature of different dimension

	Recall Rate	Correct rate	Accuracy	AUC
100 dimension	0.65	0.63	0.61	0.62
300 dimension	0.82	0.82	0.80	0.81
500 dimension	0.83	0.81	0.81	0.81
1000 dimension	0.83	0.82	0.82	0.82

Finally, the experiment compares the recognition performance of the 300-dimensional topic transfer feature and the non-topic transfer feature on the different data's distribution. The experimental results are shown in Table 3. It can be seen that the topic transfer feature has better robustness across the distributed data and there is almost no loss of recognition.

Table 3. The topic feature of 300 dimension on cross-distribution

	Recall Rate	Correct rate	Accuracy	AUC
Source domain	0.82	0.82	0.80	0.81
Target domain	0.80	0.81	0.79	0.80

5. Conclusion

We propose a topic transfer model to solve the cross-distribution to recognize the user's intention, when the distribution of source and target domains is different, the model is established in the source domain, and recognizing the intention on the target domain's data. First, an improved the bag of word model is proposed to express the original features. Secondly, the topic model is used to convert the bag of word features into topic-based representations in the source domain. The topic is transferred to the target domain, and the feature transfer under different distributions is achieved through transferring learning. Finally, xgboost [6]is used to recognize the classification of intention. This paper validates the proposed method using on-line user data. It has a higher recognition rate for cross-distributed users' data and achieves better results.

References

[1] Vapnik V. Statistical Learning Theory. John Wiley, 1998.
[2] Pan S J,Yang P Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010; 22(10):1345-1359
[3] Momtazi, Felix Naumann. Topic modeling for expert finding using latent Dirichlet allocation. Saeedeh Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2013(5).
[4] Blei D M, Ng A Y, Jordan M I.Latent Dirichlet Allocation.Journal of Machine Learning Research, 2003, 3:993-1022.
[5] Choi, Myung Jin,Torralba, Antonio,Willsky, Alan S. A tree-based context model for object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012.
[6] Tianqi Chen,Carlos Guestrin.XGBoost:A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD 16 . 2016.