http://dx.doi.org/10.14257/ijdta.2016.9.2.16

Classification Model for Intent Mining in Personal Website Based on Support Vector Machine

Shuang Zhang, Nianbin Wang

School of Computer Science and Technology, Harbin Engineering University, Harbin, 150001, PR China
15204699538@yeah.net, wangnianbin@hrbeu.edu.cn

Abstract

With the rapid growth of personal website influence, the advertisement placing has become an important investment in personal websites. But in order to accurate the advertisement placing, the specific quest for the specific users with their specific interesting need to be concerned. Acquiring, preprocessing and classifying consumption intention of the released information that published in the personal websites is the main task of this essay. We regard consumption intention mining as a binary classification problem, and extract multi-dimensional features from the raw corpus. Finally, we propose models based on SVM, Naïve Bayes and deep learning to solve the consumption intention classification problem. The experimental result shows that the deep learning based method achieves the highest F-measure.

Keywords*: User Intent, Intent Mining, Query Log, Consumption Intention*

1. Introduction

With the rapid growth of personal website influence, the advertisement placing has become an important investment in personal websites. But in order to accurate the advertisement placing, we must focus on the specific quest for the specific users with their specific interesting. Just like Traditional advertisement it has "scatter gun shoot" shortcomings, which means that the placing of advertisement does not has the exact target and the users click rate of the advertisement was quite low. Therefore, for manufacturers, a kind of new way of network which is media advertisement has greater risk, compared with traditional media advertising, the earning of this new way does not appear enough stability. On the other hand, the network provides a vast platform for personal website such as personal blog. On the personal blog, people always put some of their interest on their home page or in the information they released. In those information people always reveal their demand for some specific product. This is the excellent opportunity for media to perception the trend of the demand of the consumers.

Until recent years, personal blog and other social media is gradually raised and has more and more influence, therefore the consumption intention recognition of personal blog is a relatively new research direction. The wrong target is a quite obvious disadvantage of the traditional way of advertisements, but as far as personal blog to be concerned, cause it can easily collect the behavior of any user, as result the behavior information can be used to recognize the intention of consumers, or to determine whether the user is interested in some particular category in order to make more targeted advertisement.

This paper discussed the way to obtain the initial corpus of consumption intention, and get the initial corpus of consumer intentions from some specific website and pre-processing the corpus. In this paper, the consumption intentions tread like a binary classification problem, extracted multiple dimensional characteristics of the obtained data of the consumption intention. In this paper, we decided a method based on SVM in order to classify the intention of consumption. First of all, the personal websites were divided into two parts which are with consumption intention and without consumer intention. And if we give a classification for the personal website with consumption intention, then the consumption intention could be divided into explicit consumption intention and implicitly consumption intention. Among them, the explicit consumption intention defined as in the released information of someone's personal website, they have explicit intention to buy a certain product or service. On the other hand, the definition of implicit intention is that the released information of someone's personal website not directly and not clearly expressed the intention to purchase a product or service. But we could predicate the intention of buying some certain product or service through some of the reasonable processes. Finally, this paper proposed a model of consumption intent classification based on the SVM, Nave Bayes and deep learning. Among them algorithm of classification method based on consumption intention, the value of F (F-measure) of deep learning was the highest.

2. Intent Analysis and SVM Classification Theory

In economics, consumption intention refers to the demand for commodities for the specific consumer groups in different periods of time. It depends on the level of purchasing ability, the types of goods and some other factors. In this information era, with the popularized of the Internet, more and more users prefer to consume on the Internet, a variety of online shopping websites have gathered a considerable number of users. In addition, many users also choose to appeal to others in the network, in order to gain more information at the time of consumption to make better decisions. For manufacturers, if manufacturers adopt the consumption intention of some products, manufacturers could keep up with these potential consumers. At the same time, if manufacturers keep dynamic tracking with these users, the feedback of their services and advices of products of these potential users can also give manufactures directly help. This is undoubtedly very valuable information.

This paper proposed a method based on SVM classifier for classification of personal website consumption intent. The classification of this paper is two dimensions, which means for a given classifier the output will be with consumption intent or without consumption intent.

The designing goal of classifier is that the classifier could automatically send data to the known categories after learning. The essence of the classifiers is mathematical model. According to the different models, there are many branches, such as: Bayes classifier, BP neural network classifier, decision tree algorithms, SVM (support vector machine) classifier. The SVM classifier is proposed in this paper to give a classification for consumption intention, the classification principle as shown in Figure 1.

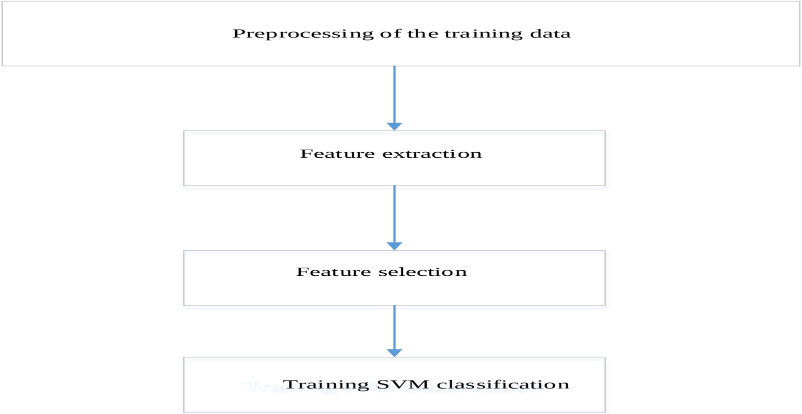
Preprocessing of the training data
Feature extraction
Feature selection
Training SVM classification

Figure 1. The Classification Principle

Support vector machine (Support Vector Machine) is first proposed by Cortes and Vapnik in 1995, it shows many unique advantages in solving the problems of small sample, nonlinear and high dimensional pattern recognition problems, and can be promoted and applied to the function fitting to other machine learning. The theoretical basis of SVM is the algorithm of the VC dimension theory and structural risk minimization principle of the statistical learning theory, which is according to few amount of the sample information, find the best compromise of the complexity of model (i.e., the learning accuracy of training samples, Accuracy) and the ability learning (i.e., the ability of identifying arbitrarily samples without error), in order to obtain best ability of promotion (or generalization ability).

SVM classifier could solve the problem of linear classification, as well as the problem of nonlinear classification. The SVM uses the method of kernel function to solve the problem of nonlinear classification. In this way the problem of nonlinear in low dimensional space could be mapped as the problem in high dimensional space, and this solution without increasing the computational complexity. In some certain degree, this solution avoided the "dimension disaster" problem which happened with other machine learning algorithm. Some common Kernel functions of SVM classifier include:

(1) The linear kernel functions, as shown in the formula (1); $K(x, y) = x.y$

(2) Polynomial kernel function, as shown in the formula (2); $K(x, y) = (x.y + 1)^d$

(3) The radial basis functions, as shown in formula (3); $K(x, y) = e^{\frac{-|x-y|^2}{2\delta^2}}$

(4)The two layer neural network By kernel function, as shown in the formula (4) $K(x, y) = \tanh(a(x, y) + b)$

SVM classifier has many advantages, at the same time it also has some shortcomings, for example:

1) Difficulty for achieving the processing of large-scale data based on SVM algorithm: SVM gets the solution of support vector of the samples through the second planning methods The second times planning design related to the calculation of the matrix of order m. As result, a large amount of computing time and storage space will be needed during the processing of a large number of samples.

2) Indirectly provide solutions for multi class problem base on SVM: the classic SVM algorithm suitable binary classification problem. The SVM algorithm not fit for the problem involved with multi classification problem.

Although the SVM classifier has some shortcomings, otherwise a few high quality training data would be chose in order to reduce the computation of SVM classifier. In addition, the combination of multiple SVM would solute the problem of multivariate.

3. Classification model for Intent Mining based on SVM

In this section, we proposed a model of consumption intention mining based on SVM, which is used to classified consumption intention.

3.1 The Fusion of Multiple Attribute based on SVM

During the processes of using the SVM classifier, the training data could not directly used for classification. First of all, the feature would be extracted from the training data, and then provided the features to classifier for learning, so that the model from learning of the classifier emerged. As result the model will be used to complete the classification of unknown types of data. On the other hand, if a large number of training data is used, then a lot of features will be extracted. Therefore, all the features are used in the SVM learning processes seems not reasonable. To solve this problem, before the classification of data, the features need to be chosen first. The features are most suitable for classification, and using these features to training the classification. The most commonly used method for choosing features is shown as follow.

Table 1. Computational Formula for Choosing Feature

Chose Features	Computational Formula
Mutual Information	$MI(t_k, c_i) = \log \dfrac{p(t_k, c_i)}{p(t_k).p(c_i)}$
Odds Ratio	$OR(t_k, c_i) = \log \dfrac{p(t_k \mid c_i).[1 - p(t_k \mid \overline{c_i})]}{[1 - p(t_k, c_i)].p(t_k \mid \overline{c_i})}$
Information Gain	$IG(t_k, c_i) = \sum_{c \in \{c_i, \overline{c_i}\}} \sum_{t \in \{t_k, \overline{t_k}\}} p(t, c).\log \dfrac{p(t, c)}{p(t).p(c)}$
Chi-Square	$\chi^2(t_k, c_i) = \dfrac{N.[p(t_k, c_i).p(\overline{t_k}, \overline{c_i}) - p(t_k, \overline{c_i}).(\overline{t_k}, c_i)]^2}{p(t_k).p(\overline{t_k}).p(c_i).p(\overline{c_i})}$

3.2 Classification Model of Intent Mining based on Fusion Multiple Attributes based on SVM

There are a lot of various type of data in the information that the personal websites released, these information are quite useful for user to express their opinion. However, during the processed of this information, the various type of data leads to a great difficulty.

During the analysis of quite lots of personal websites, some common types of the information the user released are extracted. These types are uses as characteristics to classified the consumption intention The characteristics of the SVM model are as follow:

1) Feature of word: characteristic value is determined by the appearance of the word in some certain sentence among a few of picked up words.

2) Emoticons: Emoticon is the specific characteristic of personal website. In the information uses released, they always would like to add in some emoticons. In that way, the users probably add the products name with emoticons. In a great extent, this may means the consumption intention for some products. As result, if the personal website appears emoticons, the EXIST_SYMBOL features were set as true.

3) URL: the user of personal website sometimes would like to use some other way to increase the information they released such as adding a URL which lead to some other websites. As result, if the personal website appears URL, the EXIST_URL features were set as true.

4) Length of text: From the observation of corpus of consumption intention of personal websites, it shows that the length of the released text with consumption intention is relatively short. This is because the released information of personal websites are relatively simple. On the other side, the length of advertisement, events, news is usually longer. Based on the above reasons, a text length threshold was set to compare with the length of corpus of with consumption intent and without consumption intent. According to the experimental results, if the length of a text length is less than threshold,

the LONG_SENTENCE feature was set as true.

In this section a method based on data filtering of consumption intention of personal websites is proposed. The specific algorithm is shown as follow.

$$\phi(h_{Amazon}(d_i)) = \left\{ \begin{matrix} 1, h_{Amazon}(d_i)=1 \\ 0,(othercase) \end{matrix} \right\}$$

$$\phi(h_{Amazon}(d_i)) = \left\{ \begin{matrix} 1, h_{Amazon}(d_i)=1 \\ 0,(othercase) \end{matrix} \right\}$$

First of all, a method of preprocessing is used to preprocess all the data appears in the information the users released, then extract features of text contains for the classifier, next, the extracted features would be chose from all the features. The method for feature extraction used the method of information gain shown as follow:

$$CC(t_k, c_i) = \frac{\sqrt{N} \; [p(t_k, c_i).p(\overline{t_k, c_i}) - p(t_k, \overline{c_i}).p(\overline{t_k}, c_i)]}{\sqrt{p(t_k).p(\overline{t_k}).p(c_i).p(\overline{c_i})}}$$

Then, the extract features will be submitted to the SVM, and then training as a classification model. In this study, the libsvm1 toolkit of SVM is used. Libsvm is software toolkit of an integrated support vector classification(C-SVC, nu-SVC), Regression (epsilon-SVR, nu-SVR) and estimation of distribution (one-class SVM). This toolkit supports not only the SVM classification of some of the basic operations, but also provides a lot of useful tools, such as functions of call parameters, file format conversion and so on.

4. Experiment

In this section, we evaluate the performance of the classifier proposed in this paper using the corresponding testing set. The experiment is based on SVM to classify the consumption intention according to the information the user released in the personal websites. All the data used in this chapter are from Amazon.com.

In the websites of Amazon there is 10 categories and more than 600000 purchasing data, the data of amazon.com used in this experiment shown in table 2. From all the purchasing data, 2000 records were picked up randomly which are selected from the digital equipment category. We removed the short those information, the last are remaining 1898. After labeling, we obtained 990 with consumption intent information and 908 do not have the intention of information consumption.

Table 2. Training Data

No.	Positive Data	Negative Data	Data Quality
1.	1020	980	100%
2.	6000	5000	69.38%
3.	10000	10000	65.76%
4.	16000	16000	59.23%
5.	20000	20000	56.98%
6.	25000	25000	57.36%
7.	30000	30000	60.15%
8.	35000	35000	59.35%
9.	40000	40000	57.78%
10.	50000	50000	59.34%

The methods to obtain the test data manual annotation, we randomly sampled records of released information from personal websites, and the data from the data in training data by personal websites. And then manually annotated these records whether with consumption intention or not. The scale of the data is shown in Table 3. It is not difficult to find, the distribution of positive and negative examples in the manual annotation data is extremely unbalanced, which will seriously affect the performance of classifier, in order to avoid this problem, we do not have random consumption intention personal websites in the same number of cases of personal websites.

Table 3. The Scale of Test Data

Classification	No. of personal websites records
With Consumption Intention	410
Without Consumption Intention	15460

Table 4 gives the performance of the classifier using personal websites data as training data, corresponding to the number of the data in the data in Table 4number, and the number 0 of the data, we manually labeled data.

Table 4. The Performance of the Classifier based on Multi Feature Fusion

No.	Precision (%)	Recall (%)	F-measure (%)
1.	20.89	50.74	100%
2.	23.08	52.25	69.38%
3.	22.21	53.86	65.76%
4.	23.4	16000	59.23%
5.	23.42	20000	56.98%
6.	24.38	25000	57.36%
7.	21.86	30000	60.15%
8.	18.73	35000	59.35%
9.	19.53	40000	57.78%
10.	18.34	50000	59.34%

This result is based on transfer learning and multi feature fusion classification method to mining the data. Through comparative analysis obtained an acceptable result. The text is short, the expression of micro-blog randomly and other characteristics, it is a big challenge for our study; at the same time, some features do not have common text has micro-blog text, such as emoticons and so on, which gave us a lot of opportunities.

5.Conclusion

At present, personal website has become the largest social network platform. There are more than 300000000 users and more than 100000000 records of information released it per day. As result, in the massive information released in personal website, there is a large number of users expressed consumption intention of some products. Accurate mining these consumption intentions of some specific products, it will greatly help the manufactures delivery their advertisements to the users with purposes. In addition, the research of consumption intention of personal websites could also promote the development of the other research work on the social network. This paper studies the following three aspects according to the subject of the consumption intention mining:

(1) Obtain of the consumption intention

During the research of the consumption intention, the collection of corpus database is an important problem. For some traditional research direction, the corpus database of emotion analysis had been researched for quite a long time. As result, researchers do not need to put too much energy in the corpus, because there has been a large number of existing corpus used for testing and evaluating in this field. The only things that need to do is put the focus on the result of experiment. But the consumption intention mining is a relatively new research field, it does not have a mature corpus database for the study. So in the paper at first we chose some corpus from existed corpus database for classified.

(2) Preprocessing

During the process of tagging corpus of personal websites, we noticed that some times the released information of some specific product in the personal websites is not from buyers, but from sellers or advertisers. Therefore, a determination of whether the released information is from the buyer or not is needed. A preprocessing of this information is proposed in this paper. The SVM classifier based on RBF kernel function is used to determine whether the released information is from users for buyers. SVM classifier is one of the most commonly used two dimensions of classifier, and applied in text classification field widely.

(3) Classification

In this paper, the consumption intention is regarded as a classification problem from the consumption of two dimensions. A classification of consumption intention based on the SVM, and the depth of learning (Deep Learning) is proposed. The classification method of depth study of the consumption intention based on the F value (F-measure) is highest, reaching more than 0.8. The paper also constructed a mining model of the data of amazon.com transfer learning based on consumption intention, results show that the classification results of the method is better than the only use personal websites data obtained.

Acknowledgement

The work was funded by the National Natural Science Foundation of China (No. 6127218 5), Natural Science Foundation of Heilongjiang Province of China (F201238, F020510) and Fundamental Research Funds for the Central Universities (No. HEUCFZ1219, HEUCF1006 08 and HEUCF100613)

Reference

[1] S. S. P. Salin, “Improving pattern quality in web usage mining by using semantic information”, Knowledge Information System, vol. 30, no. 3, **(2012)**, pp. 527–541.

[2] C. A. A. Clarke, “Impact of query intent and search context on click through behavior in sponsored search”, Knowledge Information System, vol. 34, no. 2, **(2012)**, pp. 425–452.

[3] M. Wan, A. Jönsson, C. Wang, L. Li and Y. Yang, “Web user clustering and Web prefetching using random indexing with weight functions”, Knowledge Information System, vol. 33, no. 1, **(2011)**, pp. 89–115.

[4] D. Beeferman and A. Berger, “Agglomerative clustering of a search engine query log”, In: Proceedings of the 6th ACM SIGKDD international conference on knowledge discovery and data mining, **(2000)**, pp. 407–416.

[5] S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder and D. Grossman, “Temporal analysis of a very large topically categorized Web query log”, Journal Am Social Information Science Technology, vol. 58, no. 2, **(2007)**, pp. 166–178.

[6] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen and H. Li, “Context-aware query suggestion by mining click-through and session data”, In: Proceeding of the 14th ACM SIGKDD international conference on

knowledge discovery and data mining, **(2008)**, pp. 875–883.

[7] A. Z. Broder, "A taxonomy of web search", SIGIR Forum, vol. 36, no. 2, **(2002)**, pp. 3–10.

[8] P. Bille, "A survey on tree edit distance and related problems", Theory Computer Science, vol. 337, no. 1–3, **(2005)**, pp. 217–239.

[9] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, H. Li, "Context-aware query suggestion by mining click-through and session data", In: Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, **(2008)**, pp. 875–883.

[10] N. Craswell, O. Zoeter, M. Taylor and B. Ramsey, "An experimental comparison of click position-bias models", In: Proceedings of the international conference on Web search and Web data mining, **(2008)**, pp. 87–94.

[11] N. Craswell, R. Jones, G. Dupret and E. Viegas, "Proceedings of the 2009 workshop on Web search click data", **(2009)**, pp. 95.

[12] O. Chapelle, D. Metlzer, Y. Zhang and P. Grinspan, "Expected reciprocal rank for graded relevance", In: Proceeding of the 18th ACM conference on information and knowledge management, **(2009)**, pp. 621–630.

[13] K. E. Arini and C. Guestrin, "Beyond keyword search: discovering relevant scientific literature", In: Proceedings of the 17thACMSIGKDDinternational conference on knowledge discovery and data mining, **(2011)**, pp. 439–447.

[14] S. Gu, J. Yan, L. Ji, S. Yan, J. Huang, N. Liu, Y. Chen and Z. Chen, "Cross domain random walk for query intent pattern mining from search engine log", In: Proceedings of the 2011 IEEE 11th international conference on data mining, **(2011)**, pp. 221–230.

[15] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. M. Wang and C. Faloutsos, "Click chain model in web search", In: Proceedings of the 18th international conference on World Wide Web, **(2009)**, pp. 11–20.

[16] F. Guo, C. Liu and Y. M. Wang, "Efficient multiple-click models in web search", In: Proceedings of the 2nd ACM international conference on Web search and data mining, **(2009)**, pp. 124–131.

[17] B. J. Jansen, A. Spink, C. Blakely and S. Koshman, "Defining a session on Web search engines: research articles", Journal Am Social Information Science Technology, vol. 58, no. 6, **(2007)**, pp. 862–871.

[18] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski and G. Gay, "Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search", ACM Trans. Information System, vol. 25, no. 2, **(2007)**, pp. 1–27.

[19] R. Landis and G. Koch, "The measurement of observer agreement for categorical data", Biometrics, vol. 33, no. 1, **(1977)**, pp. 159–174.

[20] X. Li, Y. Y. Wang and A. Acero, "Learning query intent from regularized click graphs", In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, **(2008)**, pp. 339–346.