

METHOD AND DEVICE FOR IDENTIFYING A USER INTEREST, AND COMPUTER-READABLE STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application is the National Stage of International Application No. PCT/CN2017/103944, filed September 28, 2017, which claims the benefit of China Patent Application No. 201710749088.0, filed August 25, 2017 with the State Intellectual Property Office and entitled “Method and Device for Identifying a User Interest, and Computer-readable Storage Medium”, the entirety of which is hereby incorporated herein by reference.

TECHNICAL FIELD

[0002] The present disclosure relates to the field of data processing, and more particularly relates to a method and a device for identifying a user interest, as well as a computer-readable storage medium.

BACKGROUND

[0003] In recent years, with the rapid development of the Internet, especially the sudden rise of social media, people are increasingly aware of the social media's various influences on themselves and on the information dissemination environment. In the past, people have been passively obtaining information from the Internet all along. Nowadays, however, more and more people are actively participating in the generation and dissemination of information on social media, resulting in a huge amount of user information and social relationship information.

[0004] But currently the internal data of many enterprises are usually mostly transaction records. The included customer information is not comprehensive enough to accurately determine potential customers and understand user needs. Therefore, how to identify a user's interest and fully understand the user through

CLAIMS

WHAT IS CLAIMED IS:

1. A method for identifying a user interest, comprising:

obtaining training samples and test samples, wherein the training samples are obtained by manually labeling after the corresponding topic models have been trained based on text data;

extracting characteristics of the training samples and of the test samples respectively using a first preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the characteristics of the training samples, to train and obtain a logistic regression model containing the optimal model parameters;

plotting a receiver operating characteristic (ROC) curve based on the characteristics of the test samples and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on an area AUC under the ROC curve, to train and obtain a first theme classifier;

classifying the text data using the first theme classifier to determine a theme to which the text data belongs, and computing a score of the theme to which the text data belongs based on the logistic regression model containing the optimal model parameters; and

computing a confidence score of the user who writes the text data being interested in the theme based on the score and a second preset algorithm, and identifying an interest of the user based on the confidence score.

2. The method of claim 1, wherein the block of “computing a confidence score of the user who writes the text data being interested in the theme based on the score and a second preset algorithm, and identifying an interest of the user based on the confidence score” comprises:

computing an average score of the theme to which the text data belongs based on the score and a third preset algorithm; and

computing the confidence score of the user who writes the text data being interested in the theme based on the average score and the second preset algorithm, and identifying the interest of the user based on the confidence score;

wherein the computation formula of the third preset algorithm is as follows:

$$Avg(u_i, topic_j) = \frac{\sum_{m=1}^n s(u_i, tweet_m, topic_j)}{n},$$

where $Avg(u_i, topic_j)$ represents the average score of user u_i on theme $topic_j$, $s(u_i, tweet_m, topic_j)$ represents a score of the user u_i 's text data $tweet_m$ belonging to the theme $topic_j$ after having been classified, and n represents the total number of text data messages related to the theme $topic_j$ in the user u_i 's text data messages $tweet_m$.

3. The method of claim 1, wherein the computation formula of the second preset algorithm is as follows:

$$k_j = \frac{10}{\max(TN_j) - \min(TN_j)}, \quad x_{j0} = \text{median}(TN_j),$$

$$s(u_i, topic_j, TN_{ij}) = \frac{TN_{ij} * Avg(u_i, topic_j)}{1 + e^{-k_j * (TN_{ij} - x_{j0})}},$$

where TN_j represents the number of texts of all the users where the users are interested in theme $topic_j$, x_{j0} denotes a median of TN_j , TN_{ij} represents the number of microblogs related to the theme $topic_j$ that are published by user u_i , and $s(u_i, topic_j, TN_{ij})$ denotes the confidence score of the user u_i being interested in the theme $topic_j$.

4. The method of claim 2, wherein the computation formula of the second preset algorithm is as follows:

$$k_j = \frac{10}{\max(TN_j) - \min(TN_j)}, \quad x_{j0} = \text{median}(TN_j),$$

$$s(u_i, topic_j, TN_{ij}) = \frac{TN_{ij} * Avg(u_i, topic_j)}{1 + e^{-k_j * (TN_{ij} - x_{j0})}},$$

where TN_j represents the number of texts of all the users where the users are interested in theme $topic_j$, x_{j0} denotes a median of TN_j , TN_{ij} represents the number of microblogs related to the theme $topic_j$ that are published by user u_i , and $s(u_i, topic_j, TN_{ij})$ denotes the confidence score of the user u_i being interested in the theme $topic_j$.

5. The method of claim 1, wherein the block of “obtaining training samples and test samples, wherein the training samples are obtained by manually labeling after the corresponding topic models have been trained based on text data ”comprises:

collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set;

computing a distribution of the text data on a preset number of topics using a preset theme model based on the first keyword set and the preset number of topics, and clustering the text data based on the distribution of the text data on the topics, to train and obtain the corresponding topic models of the text data; and

selecting from among the text data the training samples that correspond to a target theme classifier based on the manual labeling results on the text data based on the topic models, and using the text data other than the training samples as the test samples.

6. The method of claim 1, wherein the block of “extracting characteristics of the training samples and of the test samples respectively using a first preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the characteristics of the training samples, to train and obtain the logistic regression model containing the optimal model parameters ”comprises:

extracting the characteristics of the training samples and of the test samples respectively using the first preset algorithm, and establishing a first hash table and a second hash table, respectively; and

substituting the first hash table into the logistic regression model, and computing the optimal model parameters of the logistic regression model using the iterative algorithm, to train and obtain the logistic regression model containing the optimal model parameters.

7. The method of claim 6, wherein the block of “plotting an ROC curve based on the characteristics of the test samples and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on an area AUC under the ROC curve, to train and obtain a first theme classifier”comprises:

substituting the second hash table into the logistic regression model containing the optimal model parameters to obtain a true positive TP, a true negative TN, a false

negative FN, and a false positive FP;

plotting the ROC curve based on the TP, TN, FN, and FP;

calculating the area AUC under the ROC curve, and evaluating the logistic regression model containing the optimal model parameters based on the AUC value;

when the AUC value is less than or equal to a preset AUC threshold, determining that the logistic regression model containing the optimal model parameter does not meet the requirements, and returning to the following operation: computing the optimal model parameters of the logistic regression model using the iterative algorithm so as to train and obtain the logistic regression model containing the optimal model parameters;

otherwise when the AUC value is greater than the preset AUC threshold, determining that the logistic regression model containing the optimal model parameters meets the requirements, and training to obtain the first theme classifier; wherein plotting the ROC curve based on the TP, TN, FN, and FP comprises:

calculating a false positive rate FPR and a true positive rate TPR based on the TP, TN, FN, and FP, wherein their respective calculation formulas are $FPR = FP / (FP + TN)$, $TPR = TP / (TP + FN)$; and

plotting the ROC curve taking the FPR as the abscissa and the TPR as the ordinate.

8. The method of claim 7, wherein subsequent to the block of “plotting an ROC curve based on the characteristics of the test samples and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on an area AUC under the ROC curve, to train and obtain the first theme classifier”, the method further comprising the following operations:

substituting the second hash table into the first theme classifier to obtain a probability that the test samples belong to a corresponding topic;

adjusting the preset AUC threshold, and calculating an accuracy rate p and a recall rate r based on the TP, FP, and FN;

when the p is less than or equal to a preset p threshold, or the r is less than or equal to a preset r threshold, returning to the following operation: adjusting the preset AUC threshold until the p is greater than the preset p threshold and the r is greater than the preset r threshold, and training to obtain a second theme classifier;

wherein classifying the text data using the first theme classifier to determine the theme to which the text data belongs comprises the following operation:

classifying the text data using the second theme classifier to determine the theme to which the text data belongs.

9. The method of claim 5, wherein the block of “collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set ”comprises:

collecting the text data, and performing word segmentation on the text data;

deleting stop words in the text data after the word segmentation based on a preset stop word table, to obtain a second keyword set;

computing a term frequency (TF) and an inverse document frequency (IDF) of each keyword in the second keyword set; and

computing a term frequency-inverse document frequency (TF-IDF) value of each keyword in the second keyword set based on the TF and the IDF, and removing the keywords whose TF-IDF values are lower than a preset TF-IDF threshold, to obtain the corresponding first keyword set.

10. A device for identifying a user interest, the device comprising a memory, a processor, and a user interest identification program that is stored on the memory and operable on the processor, the user interest identification program when executed by the processor performing the following operations:

obtaining training samples and test samples, wherein the training samples are obtained by manually labeling after the corresponding topic models have been trained based on text data;

extracting characteristics of the training samples and of the test samples respectively using a first preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the characteristics of the training samples, to train and obtain a logistic regression model containing the optimal model parameters;

plotting a receiver operating characteristic (ROC) curve based on the characteristics of the test samples and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on an area AUC under the ROC curve, to train and obtain a first theme classifier;

classifying the text data using the first theme classifier to determine a theme to which the text data belongs, and computing a score of the theme to which the text data belongs based on the logistic regression model containing the optimal model parameters; and

computing a confidence score of the user who writes the text data being interested in the theme based on the score and a second preset algorithm, and identifying an interest of the user based on the confidence score.

11. A device for identifying a user interest, comprising:

an acquisition module, configured to obtain training samples and test samples, where the training samples are obtained by manually labeling after the corresponding topic models have been trained based on text data;

a first computation module, configured to extract characteristics of the training samples and of the test samples respectively using a first preset algorithm, compute optimal model parameters of a logistic regression model by an iterative algorithm based on the characteristics of the training samples, to train and obtain a logistic regression model containing the optimal model parameters;

a first training module, configured to plot a receiver operating characteristic (ROC) curve based on the characteristics of the test samples and the logistic regression model containing the optimal model parameters, and evaluate the logistic regression model containing the optimal model parameters based on an area AUC under the ROC curve, to train and obtain a first theme classifier;

a second computation module, configured to classify the text data using the first theme classifier to determine a theme to which the text data belongs, and compute a score of the theme to which the text data belongs based on the logistic regression model containing the optimal model parameters; and

an identification module, configured to compute a confidence score of the user who writes the text data being interested in the theme based on the score and a second preset algorithm, and identify an interest of the user based on the confidence score.

12. The device of claim 11, wherein the identification module comprises:

a first computation unit, configured to compute an average score of the theme to which the text data belongs based on the score and a third preset algorithm; and

a second computation unit, configured to compute the confidence score of the

user who writes the text data being interested in the theme based on the average score and the second preset algorithm, and identify the interest of the user based on the confidence score;

wherein the computation formula of the third preset algorithm is as follows:

$$Avg(u_i, topic_j) = \frac{\sum_{m=1}^n s(u_i, tweet_m, topic_j)}{n},$$

where $Avg(u_i, topic_j)$ represents the average score of user u_i on theme $topic_j$, $s(u_i, tweet_m, topic_j)$ represents a score of the user u_i 's text data $tweet_m$ belonging to the theme $topic_j$ after having been classified, and n represents the total number of text data messages related to the theme $topic_j$ in the user u_i 's text data messages $tweet_m$.

13. The device of claim 11, wherein the computation formula of the second preset algorithm is as follows:

$$k_j = \frac{10}{\max(TN_j) - \min(TN_j)}, \quad x_{j0} = \text{median}(TN_j),$$

$$s(u_i, topic_j, TN_{ij}) = \frac{TN_{ij} * Avg(u_i, topic_j)}{1 + e^{-k_j * (TN_{ij} - x_{j0})}},$$

where TN_j represents the number of texts of all the users where the users are interested in theme $topic_j$, x_{j0} denotes a median of TN_j , TN_{ij} represents the number of microblogs related to the theme $topic_j$ that are published by user u_i , and $s(u_i, topic_j, TN_{ij})$ denotes the confidence score of the user u_i being interested in the theme $topic_j$.

14. The device of claim 12, wherein the computation formula of the second preset algorithm is as follows:

$$k_j = \frac{10}{\max(TN_j) - \min(TN_j)}, \quad x_{j0} = \text{median}(TN_j),$$

$$s(u_i, topic_j, TN_{ij}) = \frac{TN_{ij} * Avg(u_i, topic_j)}{1 + e^{-k_j * (TN_{ij} - x_{j0})}},$$

where TN_j represents the number of texts of all the users where the users are interested in theme $topic_j$, x_{j0} denotes a median of TN_j , TN_{ij} represents the number of microblogs related to the theme $topic_j$ that are published by user u_i , and

$s(u_i, topic_j, TN_{ij})$ denotes the confidence score of the user u_i being interested in the theme $topic_j$.

15. The device of claim 11, wherein the acquisition module comprises:

a collection unit, configured to collect the text data and preprocess the text data to obtain a corresponding first keyword set;

a first training unit, configured to compute a distribution of the text data on a preset number of topics using a preset theme model based on the first keyword set and the preset number of topics, and cluster the text data based on the distribution of the text data on the topics, to train and obtain the corresponding topic models of the text data; and a selection unit, configured to select from among the text data the training samples that correspond to a target theme classifier based on the manual labeling results on the text data based on the topic models, and use the text data other than the training samples as the test samples.

16. The device of claim 11, wherein the first training module comprises:

an establishing unit, configured to extract the characteristics of the training samples and of the test samples respectively using the first preset algorithm, and establish a first hash table and a second hash table, respectively; and

a second training unit, configured to substitute the first hash table into the logistic regression model, and compute the optimal model parameters of the logistic regression model using the iterative algorithm, to train and obtain the logistic regression model containing the optimal model parameters.

17. The device of claim 16, wherein the first training module further comprises:

a third computation unit, configured to substitute the second hash table into the logistic regression model containing the optimal model parameters to obtain a true positive TP, a true negative TN, a false negative FN, and a false positive FP;

a plotting unit, configured to plot the ROC curve based on the TP, TN, FN, and FP;

an evaluation unit, configured to calculate the area AUC under the ROC curve, and evaluate the logistic regression model containing the optimal model parameters based on the AUC value;

a determination unit, configured to determine that the logistic regression model containing the optimal model parameter does not meet the requirements when the AUC value is less than or equal to a preset AUC threshold, and return to the following operation: computing the optimal model parameters of the logistic regression model using the iterative algorithm so as to train and obtain the logistic regression model containing the optimal model parameters;

a third training unit, configured to determine that the logistic regression model containing the optimal model parameters meets the requirements when the AUC value is greater than the preset AUC threshold, and train to obtain the first theme classifier; wherein plotting the ROC curve based on the TP, TN, FN, and FP comprises:

calculating a false positive rate FPR and a true positive rate TPR based on the TP, TN, FN, and FP, wherein their respective calculation formulas are $FPR = FP / (FP + TN)$, $TPR = TP / (TP + FN)$; and

plotting the ROC curve taking the FPR as the abscissa and the TPR as the ordinate.

18. The device of claim 17, further comprising:

a third computation module, configured to substitute the second hash table into the first theme classifier to obtain a probability that the test samples belong to a corresponding topic;

a fourth computation module, configured to adjust the preset AUC threshold, and calculate an accuracy rate p and a recall rate r based on the TP, FP, and FN;

a second training module, configured to return to the following operation when the p is less than or equal to a preset p threshold, or the r is less than or equal to a preset r threshold: adjusting the preset AUC threshold, until the p is greater than the preset p threshold and the r is greater than the preset r threshold, and train to obtain a second theme classifier;

where the identification module is further configured to classify the text data using the second theme classifier to determine the theme to which the text data belongs.

19. The device of claim 15, wherein the collection unit comprises:

a collection sub-unit, configured to collect the text data and perform word segmentation on the text data;

an acquisition sub-unit, configured to delete stop words in the text data after the word segmentation based on a preset stop word table, to obtain a second keyword set;

a first computation sub-unit, configured to compute a term frequency (TF) and an inverse document frequency (IDF) of each keyword in the second keyword set; and

a second computation sub-unit, configured to compute a term frequency-inverse document frequency (TF-IDF) value of each keyword in the second keyword set based on the TF and the IDF, and remove the keywords whose TF-IDF values are lower than a preset TF-IDF threshold, to obtain the corresponding first keyword set.

20. A computer-readable storage medium having stored therein a user interest identification program, which when executed by a processor performs the following operations:

obtaining training samples and test samples, wherein the training samples are obtained by manually labeling after the corresponding topic models have been trained based on text data;

extracting characteristics of the training samples and of the test samples respectively using a first preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the characteristics of the training samples, to train and obtain a logistic regression model containing the optimal model parameters;

plotting a receiver operating characteristic (ROC) curve based on the characteristics of the test samples and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on an area AUC under the ROC curve, to train and obtain a first theme classifier;

classifying the text data using the first theme classifier to determine a theme to which the text data belongs, and computing a score of the theme to which the text data belongs based on the logistic regression model containing the optimal model parameters; and

computing a confidence score of the user who writes the text data being interested in the theme based on the score and a second preset algorithm, and identifying an interest of the user based on the confidence score.