METHOD AND DEVICE FOR TRAINING A TOPIC CLASSIFIER, AND COMPUTER-READABLE STORAGE MEDIUM

[0001] **FIELD**

[0002] The present disclosure relates to the field of information processing, and more particularly to a method and device for training a topic classifier, and computer-readable storage medium.

[0003] **BACKGROUND**

[0004] In recent years, with the rapid development of the Internet, information resources are growing exponentially. Abundant Internet information resources have brought great convenience to people's lives. People can obtain various types of information resources such as audio and video media, news reports, and technical literatures only by connecting a computer to the Internet.

[0005] However, in the era of big data, the classification efficiency and accuracy of existing classification techniques are relatively low, as a result, it difficult for users to obtain relevant topic information accurately and quickly in front of massive information resources. Therefore, how to improve the efficiency and accuracy of topic classification is a technical problem to be solved by those skilled in the art.

[0006] **SUMMARY**

[0007] The present disclosure is to provide a method and device for training a topic classifier, and computer-readable storage medium, which aims to improve the efficiency and accuracy of topic classification, so that users can effectively obtain relevant topic information from massive information.

[0008] In order to achieve the above aim, the present disclosure provides a method for training a topic classifier which includes:

[0009] obtaining a training sample and a test sample, wherein the training sample

is obtained by manually labeling after a corresponding topic model having been trained based on text data;

[0010] extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and

[0011] drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier.

[0012] Furthermore, in order to achieve the above aim, the present disclosure provides a device for training a topic classifier which includes: a memory, a processor, and a topic classifier training program stored in the memory and executable on the processor, the topic classifier training program when executed by the processor performing the above operations of the method for training the topic classifier.

[0013] Furthermore, in order to achieve the above aim, the present disclosure provides a computer-readable storage medium, wherein a topic classifier training program is stored in the computer-readable storage medium, the topic classifier training program when executed by the processor performing the above operations of the method for training the topic classifier.

[0014] Furthermore, in order to achieve the above aim, the present disclosure provides a device for training a topic classifier which includes:

[0015] a first obtaining module, configured for obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data;

[0016] a first training module, configured for extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm

based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and

[0017] a second training module, configured for drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier.

[0018] In the present disclosure, the training sample and the test sample are obtained, wherein the training sample is obtained by manually labeling after the corresponding topic model having been trained based on text data; features of the training sample and the test sample are extracted respectively using the preset algorithm, and optimal model parameters of the logistic regression model are computed by the iterative algorithm based on the features of the training sample, the logistic regression model containing the optimal model parameters is trained and got; the ROC curve of receiver operating characteristic is drawn based on the features of the test sample and the logistic regression model containing the optimal model parameters, and the logistic regression model containing the optimal model parameters is evaluated based on the area AUC under the ROC curve, the first topic classifier is trained and got. Through the above method, the present disclosure performs feature extracting to the training sample and the test sample using the preset algorithm which shortens the time of feature extracting and model training and improves the classification efficiency. The present disclosure selects the training sample by manually labeling which could improve the accuracy of the training sample, so as to improve the classification accuracy of the topic classifier, meanwhile, performing evaluating the logistic regression model containing the optimal model parameters through the area AUC under the ROC curve to train the topic classifier, so as to perform classification to the text data, which could further improve the accuracy of topic classification.

[0019] **BRIEF DESCRIPTION OF THE DRAWINGS**

[0020] Fig. 1 is a structural diagram illustrating a topic classifier device of an

embodiment according to the present disclosure.

[0021] Fig. 2 is a flowchart illustrating a first embodiment of the method for training a topic classifier according to the present disclosure.

[0022] Fig. 3 is a detailed flowchart illustrating obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data of an embodiment according to the present disclosure.

[0023] Fig. 4 is a detailed flowchart illustrating drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier of an embodiment according to the present disclosure.

[0024] Fig. 5 is a flowchart illustrating a second embodiment of the method for training a topic classifier according to the present disclosure.

[0025] Fig. 6 is a detailed flowchart illustrating collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set of an embodiment according to the present disclosure.

[0026] Various implementations, functional features, and advantages of the present disclosure will now be described in further detail with reference to the accompanying drawings and some illustrative embodiments.

[0027] **DETAILED DESCRIPTION OF THE EMBODIMENTS**

[0028] It is to be understood that, the specific embodiments described herein portrays merely some illustrative embodiments of the present disclosure, and are not intended to limit the patentable scope of the present disclosure.

[0029] Due to the low classification efficiency and accuracy of the existing classification technology, it is difficult for a user to obtain the relevant topic information required by the user accurately and quickly in front of massive information resources.

[0030] In order to achieve the above aim, the present disclosure provides a

method for training a topic classifier: obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data; extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier. Through the above method, the present disclosure performs feature extracting to the training sample and the test sample using the preset algorithm which shortens the time of feature extracting and model training and improves the classification efficiency. The present disclosure selects the training sample by manually labeling which could improve the accuracy of the training sample, so as to improve the classification accuracy of the topic classifier, meanwhile, performing evaluating the logistic regression model containing the optimal model parameters through the area AUC under the ROC curve to train the topic classifier, so as to perform classification to the text data, which could further improve the accuracy of topic classification.

[0031] Referring to Fig. 1, Fig. 1 is a structural diagram illustrating a topic classifier device of an embodiment according to the present disclosure.

[0032] The device in the embodiment of the present invention may be a PC, or may be a terminal device with a display function, such as a smart phone, a tablet computer, or a portable computer.

[0033] As shown in FIG. 1, the device may include a processor 1001, such as a CPU, a network interface 1004, a user interface 1003, a memory 1005, and a communication bus 1002. Among them, the communication bus 1002 is configured to facilitate the connection communication between these components. The user interface 1003 may include a display, an input unit such as a keyboard, and the user

interface 1003 may optionally also include a standard wired interface, a wireless interface. The network interface 1004 may optionally include a standard wired interface, a wireless interface (such as a WI-FI interface). The memory 1005 may be a high speed RAM memory or a non-volatile memory such as a magnetic disk memory. The memory 1005 may optionally be a storage device that is separate from the aforementioned processor 1001.

[0034] Optionally, the device may further include a camera, an RF (Radio Frequency) circuit, a sensor, an audio circuit, a WiFi module, and the like. Among them, the sensor such as a light sensor, a motion sensor, and other sensors. Specifically, the light sensor may include an ambient light sensor and a proximity sensor, wherein the ambient light sensor may adjust the brightness of the display based on the brightness of the ambient light, and the proximity sensor may turn off the display and/or the backlight when the device moves near the ear. As a kind of the motion sensor, a gravity acceleration sensor can detect the magnitude of acceleration in each direction (usually three axes). When at rest it can detect the magnitude and direction of the gravity, and it can be used in applications for identifying the attitude of a mobile terminal (e.g., switching between landscape and portrait screen modes, related games, magnetometer attitude calibration), vibration identification related functions (e.g., pedometer, tapping), and so on. The mobile terminal can of course also be equipped with other sensors such as a gyroscope, a barometer, a hygrometer, a thermometer, an infrared sensor, etc.; however it will not be detailed herein.

[0035] Persons skilled in the art can understand that the device structure illustrated in FIG. 1 is not meant to limit the device, the device may include more or fewer components than illustrated, or some components may be combined, or different component arrangements may be implemented.

[0036] As illustrated in FIG. 1, the memory 1005 as a computer storage medium may include an operating system, a network communication module, a user interface module, and a topic classifier training program.

[0037] In the device illustrated in FIG. 1, the network interface 1004 is mainly used to connect to a backend server and perform data communication with the

backend server. The user interface 1003 is mainly used to connect to a client and perform data communication with the client. The processor 1001 can be used to invoke the topic classifier training program stored in the memory 1005 and perform the following operations:

[0038] obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model is trained based on text data;

[0039] extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and

[0040] drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier.

[0041] Further, the processor 1001 can invoke the topic classifier training program stored in the memory 1005 and perform the following operations:

[0042] collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set;

[0043] computing a distribution of the text data on a preset number of topics using a preset topic 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 get the corresponding topic models of the text data; and

[0044] selecting from among the text data the training samples that correspond to a target topic 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 sample.

[0045] Further, the processor 1001 can invoke the topic classifier training program

stored in the memory 1005 and perform the following operations:

[0046] extracting the features of the training sample and of the test sample respectively using a preset algorithm, and correspondingly establishing a first hash table and a second hash table;

[0047] substituting the first hash table into the logistic regression model, and calculating the optimal model parameters of the logistic regression model using the iterative algorithm, to train and get the logistic regression model containing the optimal model parameters.

[0048] Further, the processor 1001 can invoke the topic classifier training program stored in the memory 1005 and perform the following operations:

[0049] substituting the second hash table into the logistic regression model containing the optimal model parameters to obtain true positive TP, true negative TN, false negative FN, and false positive FP;

[0050] drawing the ROC curve based on TP, TN, FN and FP;

[0051] calculating the area AUC under the ROC curve, and evaluating the logistic regression model containing the optimal model parameters based on the AUC value;

[0052] when the AUC value is less than or equal to a preset AUC threshold, determining that the logistic regression model containing the optimal model parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get the logistic regression model containing the optimal model parameters;

[0053] 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 requirement, and trains to get the first topic classifier.

[0054] Further, the processor 1001 can invoke the topic classifier training program stored in the memory 1005 and perform the following operations:

[0055] calculating a false positive rate FPR and a true positive rate TPR based on TP, TN, FN, and FP, wherein their respective calculation formulas are FPR = FP / (FP + TN), TPR = TP / (TP + FN); and

[0056] drawing the ROC curve taking the FPR as the abscissa and the TPR as the ordinate.

[0057] Further, the processor 1001 can invoke the topic classifier training program stored in the memory 1005 and perform the following operations:

[0058] substituting the second hash table into the first topic classifier to obtain a probability that the test sample belongs to a corresponding topic;

[0059] adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on TP, FP, and FN;

[0060] 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 get the second topic classifier.

[0061] Further, the processor 1001 can invoke the topic classifier training program stored in the memory 1005 and perform the following operations:

[0062] classifying the text data using the second topic classifier.

[0063] Further, the processor 1001 can invoke the topic classifier training program stored in the memory 1005 and perform the following operations:

[0064] collecting the text data, and segmenting the text data;

[0065] removing stop words in the text data after the segmentation based on a preset stop word list, to obtain a second keyword set;

[0066] calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.

[0067] Further, the processor 1001 can invoke the topic classifier training program stored in the memory 1005 and perform the following operations:

[0068] calculating the term frequency TF and the inverse document frequency IDF of each keyword in the second keyword set;

[0069] calculating the term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose

TF-IDF value is lower than the preset threshold of TF-IDF, to obtain the corresponding first keyword set.

[0070] Referring to Fig. 2, Fig. 2 is a flowchart illustrating a first embodiment of the method for training a topic classifier according to the present disclosure.

[0071] In the present disclosure, the method for training the topic classifier includes:

[0072] S100, obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data;

[0073] S200, extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters;

[0074] In this embodiment, the training sample and the test sample required by training topic classifier are obtained, the training sample is obtained by manually labeling after the corresponding topic model is trained based on the text data, configured for optimizing the parameters of the model, while the test sample is the text data other than the training sample, configured for evaluating the performance of the established model. In a specific embodiment, the acquisition of the training sample and the test sample can also be sampled directly from the microblogs found in the Internet by a program, such as the Svmtrain function in the mathematical software Matlab.

[0075] Further, the features of the training sample and the test sample are respectively extracted using the preset algorithm. In this embodiment, the features of the training sample and the test sample are respectively extracted by using a Byte 4-gram algorithm of a binary hash table. Each training sample or test sample is correspondingly represented as a feature vector consisting of a set of features. The method extracts all consecutive 4 bytes in each training sample or test sample data as a key, converts the string into a byte array corresponding to UTF-8 encoding of the

string, whose value is 32 bit integer. Further, a hash function is established through the remainder method, and a first hash table and a second hash table are respectively correspondingly established. Among them, it should be noted that the hash function formula for the hash table with a length m is: f(key) = key mod p, (p≤m). Wherein the mod represents the remainder. In a specific implementation method, in order to reduce the occurrence of conflicts, and to avoid the hash table distribution to be too sparse, usually p is the largest prime number smaller than the hash table length.

[0076] Further, substitute the first hash table into the logistic regression model, and iteratively compute the optimal model parameters by an optimization method, and the logistic regression model is trained, wherein the logistic regression model is configured to estimate the possibility of a certain thing, or to determine the probability that a sample belongs to a certain category. The logistic regression model is:

[0077] $$h_\theta(x^{(i)}) = g(\theta^T x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}}, \quad \theta^T x^{(i)} = \sum_{j=0}^{n} \theta_i x_j^{(i)},$$

[0078] Wherein, xj represents the eigenvector of the jth training sample, x(i) represents the ith sampling, and θ represents the model parameters.

[0079] In addition, it should be noted that the iterative algorithm includes gradient descent, conjugate gradient method and quasi-Newton method. In a specific embodiment, the optimal model parameters of the logistic regression model can be computed by any of the above iterative algorithms, and the logistic regression model containing optimal model parameters is trained. Certainly, in a specific embodiment, other methods may be used to respectively extract features of the training sample and of the test sample, such as a vector space model VSM, an information gain method, and a expected cross entropy. ?

[0080] S300, drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier.

[0081] In this embodiment, the second hash table established based on the test sample is substituted into the logistic regression model containing the optimal model parameters, thereby obtaining true positive TP, true negative TN, false negative FN and false positive FP, wherein TP is the number of positive class after using the logistic regression model to judge the positive class in the training sample, TN is the number of negative class after using the logistic regression model to judge the negative classes in the training sample, FN is the number of positive class after using the logistic regression model to judge the negative classes in the training sample, FP is the number of negative class after using the logistic regression model to judge the positive classes in the training sample. The positive class and the negative class refer to two categories labeled manually to the training sample. That is, if a sample is manually labeled as belonging to a specific class, the sample belongs to the positive class, and the sample that does not belong to that particular class belongs to the negative class. Based on TP, TN, FN and FP, a false positive rate FPR and a true positive rate TPR are calculated. The ROC curve is drawn with FPR as the abscissa and TPR as the ordinate. The ROC curve is a characteristic curve of each indicator obtained, configured to demonstrate a relationship between the indicators. Further, the area AUC under the ROC curve is calculated, AUC is the area under the ROC curve, the greater the AUC, the better it is, meaning that the diagnostic value of the test is higher. Perform evaluating to the logistic regression model containing the optimal model parameters, when the AUC value is less than or equal to a preset AUC threshold, it is determined that the logistic regression model containing the optimal model parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get the logistic regression model containing the optimal model parameters, until the AUC value is greater than the preset AUC threshold, it is determined that the logistic regression model containing the optimal model parameters meets the requirement, and the first subject classifier has been trained.

[0082] In the present disclosure, the training sample and the test sample are

obtained, wherein the training sample is obtained by manually labeling after the corresponding topic model having been trained based on text data; features of the training sample and the test sample are extracted respectively using the preset algorithm, and optimal model parameters of the logistic regression model are computed by the iterative algorithm based on the features of the training sample, the logistic regression model containing the optimal model parameters is trained and got; the ROC curve of receiver operating characteristic is drawn based on the features of the test sample and the logistic regression model containing the optimal model parameters, and the logistic regression model containing the optimal model parameters is evaluated based on the area AUC under the ROC curve, the first topic classifier is trained and got. Through the above method, the present disclosure performs feature extracting to the training sample and the test sample using the preset algorithm which shortens the time of feature extracting and model training and improves the classification efficiency. The present disclosure selects the training sample by manually labeling which could improve the accuracy of the training sample, so as to improve the classification accuracy of the topic classifier, meanwhile, performing evaluating the logistic regression model containing the optimal model parameters through the area AUC under the ROC curve to train the topic classifier, so as to perform classification to the text data, which could further improve the accuracy of topic classification.

[0083] Based on the first embodiment illustrated in Fig.2, referring to Fig.3, Fig. 3 is a detailed flowchart illustrating obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data of an embodiment according to the present disclosure, S100 includes:

[0084] S110, collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set;

[0085] In the embodiment, the text data can be obtained from all major social networking platforms, such as Weibo, QQ Space, Zhihu, Baidu Tieba, etc., and can also be obtained from all major information resource databases, such as Tencent Video,

CNKI, and EPaper, etc. In this embodiment, Weibo text is taken as an example, specifically, Weibo text data can be collected through Sina API (Application Programming Interface), and the text data includes main body and comment.

[0086] In the embodiment, the process of preprocessing the text data includes segmenting the text data, performing part-of-speech tagging, and then removing stop words in the text data after the segmentation based on a preset stop word list to obtain a second keyword set. Further, calculate the term frequency TF, the inverse document frequency IDF, and the term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and remove the keyword whose TF-IDF value is lower than the preset threshold of TF-IDF, to obtain the corresponding first keyword set.

[0087] S120, computing a distribution of the text data on a preset number of topics using a preset topic 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 get the corresponding topic models of the text data;

[0088] In the embodiment, the preset topic model is an LDA topic model, which is an unsupervised machine learning technology, configured to identify underlying topic information in large-scale document sets or corpuses, using probability distribution of underlying topics to represent each document in the document set, using probability distribution of lexical items to represent each underlying topic. Specifically, in the embodiment, when the terminal receives the input first keyword set and the set number of topics, the LDA topic model computes the distribution of the text data on a preset number of topics using a preset topic model based on the first keyword set and the preset number of topics. Further, clustering is performed based on the distribution of the text data on the topics, and the topic model corresponding to the text data is trained.

[0089] S130, selecting from among the text data the training samples that correspond to a target topic 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 sample.

[0090] In this embodiment, since the LDA model is a topic generation model, the type of the obtained topic cannot be controlled. Therefore, the obtained topic needs to be manually labeled to filter out the text data corresponding to the target topic as the training sample of the topic classifier, which facilitates to improve the classification accuracy of the topic classifier. In addition, text data other than the training sample is used as the test sample for evaluating the trained logistic regression model.

[0091] Based on the first embodiment illustrated in Fig.2, referring to Fig.4, Fig. 4 is a detailed flowchart illustrating drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier. S300 includes:

[0092] S310, substituting the second hash table into the logistic regression model containing the optimal model parameters to obtain true positive TP, true negative TN, false negative FN, and false positive FP;

[0093] S320, drawing the ROC curve based on TP, TN, FN and FP;

[0094] S330, calculating the area AUC under the ROC curve, and evaluating the logistic regression model containing the optimal model parameters based on the AUC value;

[0095] S340, when the AUC value is less than or equal to a preset AUC threshold, determining that the logistic regression model containing the optimal model parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get the logistic regression model containing the optimal model parameters;

[0096] S350, when the AUC value is greater than the preset AUC threshold, determining that the logistic regression model containing the optimal model parameters meets the requirement, and trains to get the first topic classifier.

[0097] In this embodiment, the second hash table is substituted into the logistic regression model containing the optimal model parameters to analyze the test sample,

there exist the following four situations: if the text data belongs to a certain topic, meanwhile it is predicted to belong to the topic by the logistic regression model containing the optimal model parameters, then is true TP; if the text data does not belong to a certain topic meanwhile it is predicted not to belong to the topic, it is true negative TN; if the text data belongs to a certain topic but is predicted not to belong to the topic, it is false negative FN; if the text data does not belong to a certain topic but is predicted to belong to the topic, it is false positive FP.

[0098] Further, the ROC curve is drawn based on TP, TN, FN and FP. Specifically, the ROC curve takes the false positive rate FPR as the abscissa and the true positive rate TPR as the ordinate. The specific calculation formula is as follows:

[0099] FPR = FP / (FP + TN), TPR= TP / (TP + FN).

[0100] Further, calculate the area AUC under the ROC curve, the calculation formula is as follows:

$$\mathrm{AUC} = \int_0^1 TPR\, dFPR = \frac{1}{(TP+FN)(TN+FP)} \int_0^1 TP\, dFP$$

[0101] In this embodiment, the larger the AUC value, the better performance of the logistic regression model containing the optimal model parameters. When the calculated AUC value is less than or equal to the preset AUC threshold, it is determined that the logistic regression model with the optimal model parameters does not meet the requirement, and returns to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get the logistic regression model containing the optimal model parameters. Until the AUC value is greater than the preset AUC threshold, it is determined that the logistic regression model containing the optimal model parameters meets the requirement, the first subject classifier is trained.

[0102] Based on the first embodiment illustrated in Fig.2, referring to Fig.5, Fig. 5 is a flowchart illustrating a second embodiment of the method for training a topic classifier according to the present disclosure. The method further includes:

[0103] S400, substituting the second hash table into the first topic classifier to obtain a probability that the test sample belongs to a corresponding topic;

[0104] S500, adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on TP, FP, and FN;

[0105] S600, 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 get the second topic classifier;

[0106] S700, classifying the text data using the second topic classifier.

[0107] It should be noted that, with respect to the first embodiment shown in FIG.2, the difference between the second embodiment shown in FIG. 4 is that, in actual use, due to excessive text data, the labor force for manual labeling sample is too large, may not be cover all possible text data, resulting in poor performance. In addition, when using the area AUC under the ROC curve to evaluate the logistic regression model containing the optimal model parameters, 0.5 is used as the preset AUC threshold by default, and if it is greater than 0.5, the predicted result of the logistic regression model is 1, indicating that it belongs to the topic; if less than or equal to 0.5, the prediction result of the logistic regression model is 0, which means that it does not belong to the topic. Therefore, in the second embodiment, by adjusting the preset AUC threshold, the classification accuracy of the second topic classifier is further improved while ensuring the precision rate p and the recall rate r.

[0108] In this embodiment, the second hash table is substituted into the first topic classifier to obtain the probability of the test sample belonging to the corresponding topic. Further, the preset AUC threshold is adjusted, and the precision rate p and the recall rate r are calculated based on TP, FP, and FN, the calculation formula is as follows:

[0109] $$p = \frac{TP}{TP + FP}, \quad r = \frac{TP}{TP + FN},$$

[0110] When p is less than or equal to a preset threshold of the precision rate, or r is less than or equal to a preset threshold of the recall rate, returning to the following operation: adjusting the preset AUC threshold until p is greater than the preset

threshold of the precision rate, and r is greater than the preset threshold of the recall rate, the second subject classifier is trained, classifying the text data using the second topic classifier.

[0111] Based on the first embodiment illustrated in Fig.3, referring to Fig.6, Fig. 6 is a detailed flowchart illustrating collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set of an embodiment according to the present disclosure. S110 includes:

[0112] S111, collecting the text data, and segmenting the text data;

[0113] S112, removing stop words in the text data after the segmentation based on a preset stop word list, to obtain a second keyword set;

[0114] S113, calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.

[0115] In the embodiment, the text data can be obtained from all major social networking platforms, such as Weibo, QQ Space, Zhihu, Baidu Tieba, etc., and can also be obtained from all major information resource databases, such as Tencent Video, CNKI, and EPaper, etc. In this embodiment, Weibo text is taken as an example, specifically, Weibo text data can be collected through Sina API (Application Programming Interface), and the text data includes main body and comment.

[0116] Further, perform preprocessing to the text data which includes segmenting the text data and performing part-of-speech tagging. It should be noted that the word segmentation process can be carried out through a word segmentation tool, such as Chinese Lexical Analysis System ICTCLAS, Tsinghua University Lexical Analyzer for Chinese THULAC, Language Technology Platform LTP, and the like. The word segmentation mainly divides each Chinese text in the sample data into one by one word based on the characteristics of the Chinese language, and performs part-of-speech tagging.

[0117] Further, the pre-processing process further includes removing stop words in the text data after the segmentation based on the preset stop word list to obtain the

second keyword set. The removal of the stop words is beneficial to increase the density of the keywords, thereby facilitating the determination of the topic to which the text data belongs. It should be noted that the stop words mainly include two categories: the first category is some words which are used too frequently, such as "I", "just", etc., such words appear in almost every document; the second category is some words which appear frequently in the text but have no real meaning, such words only have a certain meaning when they are put into a complete sentence, including modal auxiliary words, adverbs, prepositions, conjunctions, etc., such as "of" , "in", "then" and so on.

[0118] Further, the preprocessing process includes calculating the term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set. Specifically, first the term frequency TF and the inverse document frequency IDF are calculated, wherein TF represents the frequency of a certain keyword appearing in the current document, and IDF represents the distribution of the keyword in all of the documents of the text data, which is a measure of general importance of a word. The formula for calculating TF and IDF is as follows:

[0119] $$\mathrm{TF} = \frac{n_i}{n}, \quad \mathrm{IDF} = \log\frac{N}{1+N_i}$$

[0120] Wherein, ni represents the number of times the keyword appears in the current document, n represents the total number of keywords in the current document, N represents the total number of documents in the data set, and Ni represents the number of documents in which the keyword appears in the text data set.

[0121] Further, calculate the TF-IDF value based on the formula TF-IDF=TF×IDF, remove the keyword whose TF-IDF value is lower than the preset threshold of TF-IDF, to obtain the corresponding first keyword set.

[0122] In addition, the present disclosure further provides a computer-readable storage medium, a topic classifier training program is stored on the computer-readable

storage medium, the above operations of the method of training a topic classifier are performed when the topic classifier training program is executed by the processor.

[0123] The operations performed when the topic classifier training program is executed by the processor refer to various embodiments of the method of training the topic classifier of the present disclosure, details are not described herein.

[0124] In addition, the present disclosure further provides a device for training a topic classifier, which includes:

[0125] a first obtaining module, configured for obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data;

[0126] a first training module, configured for extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and

[0127] a second training module, configured for drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier.

[0128] Further, the first obtaining module includes:

[0129] a collecting unit, configured for collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set;

[0130] a first training unit, configured for computing a distribution of the text data on a preset number of topics using a preset topic 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 get the corresponding topic models of the text data; and

[0131] a classifying unit, configured for selecting from among the text data the training samples that correspond to a target topic 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 sample.

[0132] Further, the first training unit includes:

[0133] an establishing unit, configured for extracting the features of the training sample and of the test sample respectively using a preset algorithm, and correspondingly establishing a first hash table and a second hash table;

[0134] a second training unit, configured for substituting the first hash table into the logistic regression model, and calculating the optimal model parameters of the logistic regression model using the iterative algorithm, to train and get the logistic regression model containing the optimal model parameters.

[0135] Further, the second training module includes:

[0136] an obtaining unit, configured for substituting the second hash table into the logistic regression model containing the optimal model parameters to obtain true positive TP, true negative TN, false negative FN, and false positive FP;

[0137] a drawing unit, configured for drawing the ROC curve based on TP, TN, FN and FP;

[0138] an evaluating unit, configured for calculating the area AUC under the ROC curve, and evaluating the logistic regression model containing the optimal model parameters based on the AUC value;

[0139] a determining unit, configured for when the AUC value is less than or equal to a preset AUC threshold, determining that the logistic regression model containing the optimal model parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get the logistic regression model containing the optimal model parameters;

[0140] a third training unit, configured for when the AUC value is greater than the preset AUC threshold, determining that the logistic regression model containing the optimal model parameters meets the requirement, and trains to get the first topic classifier.

[0141] Further, the drawing unit includes:

[0142] a calculating sub-unit, configured for calculating a false positive rate FPR and a true positive rate TPR based on TP, TN, FN, and FP, wherein their respective calculation formulas are FPR = FP / (FP + TN), TPR = TP / (TP + FN); and

[0143] a drawing sub-unit, configured for drawing the ROC curve taking the FPR as the abscissa and the TPR as the ordinate.

[0144] Further, the device for training a topic classifier further includes:

[0145] a second obtaining module, configured for substituting the second hash table into the first topic classifier to obtain a probability that the test sample belongs to a corresponding topic;

[0146] a first adjusting module, configured for adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on TP, FP, and FN;

[0147] a second adjusting module, configured for 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 get the second topic classifier;

[0148] a classifying module, configured for classifying the text data using the second topic classifier.

[0149] Further, the collecting unit includes:

[0150] a collecting sub-unit, configured for collecting the text data, and segmenting the text data;

[0151] a removing sub-unit, configured for removing stop words in the text data after the segmentation based on a preset stop word list, to obtain a second keyword set;

[0152] a calculating sub-unit, configured for calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.

[0153] Further, the calculating sub-unit includes:

[0154] a first calculating sub-unit, configured for calculating the term frequency

TF and the inverse document frequency IDF of each keyword in the second keyword set;

[0155] a second calculating sub-unit, configured for calculating the term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than the preset threshold of TF-IDF, to obtain the corresponding first keyword set.

[0156] The operations performed when executing each module refer to various embodiments of the method of training the topic classifier of the present disclosure, details are not described herein.

[0157] It should be noted that, throughout this disclosure, the terms "include", "comprise" or any other variations thereof are intended to encompass non-exclusive inclusions, so that a process, method, article, or system that includes a series of elements would include not only those elements, but it may further include other elements that are not explicitly listed or elements that are inherent to such processes, methods, articles, or systems. In the absence of extra limitations, an element defined by the phrase "includes a..." does not exclude the presence of additional identical elements in this process, method, article, or system that includes the element.

[0158] Sequence numbers of the embodiments disclosed herein are meant for the sole purpose of illustrative and do not represent the advantages and disadvantages of these embodiments.

[0159] Through the above description of the foregoing embodiments, those skilled in the art can clearly understand that the above methods of the embodiments can be implemented by means of software plus a necessary general hardware platform; they certainly can also be implemented by means of hardware, but in many cases, the former is a better implementation. Based on this understanding, the essential part of the technical solution according to the present disclosure or the part that contributes to the prior art can be embodied in the form of a software product. Computer software products can be stored in a storage medium as described above (e.g., ROM/RAM, a magnetic disk, an optical disc) which includes instructions to cause a terminal device (e.g., a mobile phone, a computer, a server, an air conditioner, or a network device,

etc.) to perform the methods described in the various embodiments of the present disclosure.

[0160] The foregoing description portrays merely some illustrative embodiments of the present disclosure, and are not intended to limit the patentable scope of the present disclosure. Any equivalent structural or flow transformations based on the specification and the drawing of the present disclosure, or any direct or indirect applications of the present disclosure in other related technical fields, shall all fall within the protection scope of the present disclosure.

CLAIMS

WHAT IS CLAIMED IS:

1. A method for training a topic classifier, comprising:

obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data;

extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and

drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier;

wherein the step of extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters comprises:

extracting the features of the training sample and of the test sample respectively using a preset algorithm, and correspondingly establishing a first hash table and a second hash table;

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

2. The method of claim 1, wherein the step of obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having 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 topic 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 get the corresponding topic models of the text data; and

selecting from among the text data the training samples that correspond to a target topic 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 sample.

3. The method of claim 1, wherein the step of drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier comprises:

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

drawing the ROC curve based on 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 parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get 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 requirement, and trains to get the first topic classifier.

4. The method of claim 3, further comprising:

substituting the second hash table into the first topic classifier to obtain a probability that the test sample belongs to a corresponding topic;

adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on 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 get the second topic classifier;

classifying the text data using the second topic classifier.

5. The method of claim 2, wherein the step of collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set comprises:

collecting the text data, and segmenting the text data;

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

calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.

6. A device for training a topic classifier, comprising:

a memory,

a processor, and

a topic classifier training program stored in the memory and executable on the processor, the topic classifier training program when executed by the processor

performing the following operations:

obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data;

extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and

drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier;

wherein following operations are further performed when the topic classifier training program executed by the processor:

extracting the features of the training sample and of the test sample respectively using a preset algorithm, and correspondingly establishing a first hash table and a second hash table;

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

7. The device of claim 6, wherein following operations are further performed when the topic classifier training program executed by the processor:

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 topic 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 get the corresponding topic models of the text data; and

selecting from among the text data the training samples that correspond to a target topic 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 sample.

8. The device of claim 6, wherein following operations are further performed when the topic classifier training program executed by the processor:

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

drawing the ROC curve based on 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 parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get 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 requirement, and trains to get the first topic classifier.

9. The device of claim 8, wherein following operations are further performed when the topic classifier training program executed by the processor:

substituting the second hash table into the first topic classifier to obtain a probability that the test sample belongs to a corresponding topic;

adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on 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 get the second topic classifier;

classifying the text data using the second topic classifier.

10. The device of claim 7, wherein following operations are further performed when the topic classifier training program executed by the processor:

collecting the text data, and segmenting the text data;

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

calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.

11. The device of claim 10, wherein following operations are further performed when the topic classifier training program executed by the processor:

calculating the term frequency TF and the inverse document frequency IDF of each keyword in the second keyword set;

calculating the term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than the preset threshold of TF-IDF, to obtain the corresponding first keyword set.

12. A computer-readable storage medium, wherein a topic classifier training program is stored in the computer-readable storage medium, the topic classifier training program when executed by the processor performing the following operations:

obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained

based on text data;

extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and

drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier;

wherein following operations are further performed when the topic classifier training program executed by the processor:

extracting the features of the training sample and of the test sample respectively using a preset algorithm, and correspondingly establishing a first hash table and a second hash table;

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

13. The computer-readable storage medium of claim 12, wherein following operations are further performed when the topic classifier training program executed by the processor:

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 topic 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 get the corresponding topic models of the text data; and

selecting from among the text data the training samples that correspond to a target

topic 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 sample.

14. The computer-readable storage medium of claim 12, wherein following operations are further performed when the topic classifier training program executed by the processor:

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

drawing the ROC curve based on 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 parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get 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 requirement, and trains to get the first topic classifier.

15. The computer-readable storage medium of claim 14, wherein following operations are further performed when the topic classifier training program executed by the processor:

substituting the second hash table into the first topic classifier to obtain a probability that the test sample belongs to a corresponding topic;

adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on 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 get the second topic classifier;

classifying the text data using the second topic classifier.

16. The computer-readable storage medium of claim 12, wherein following operations are further performed when the topic classifier training program executed by the processor:

collecting the text data, and segmenting the text data;

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

calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.

17. The computer-readable storage medium of claim 16, wherein following operations are further performed when the topic classifier training program executed by the processor:

calculating the term frequency TF and the inverse document frequency IDF of each keyword in the second keyword set;

calculating the term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than the preset threshold of TF-IDF, to obtain the corresponding first keyword set.