WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 1, Number 1, Year 2023
DOI: https://doi.org/10.62051/ijcsit.v1n1.05
Journal homepage: https://wepub.org/index.php/IJCSIT/index

The implementation of an AI-driven advertising push system based on a NLP algorithm

Qi Xin [1], Yuhang He [2], Yiming Pan [3], Yong Wang [4], Shuqian Du [5]

[1] Management Information Systems, University of Pittsburgh, Pittsburgh, PA, USA
[2] Computer Science and Technology, Tianjin University of Technology, Tianjin, China
[3] Computer Science, Individual Contributor, Austin, TX, USA
[4] Information Technology, University of Aberdeen, Aberdeen, United Kingdom
[5] Information Studies, Trine University, Phoenix, Arizona, AZ, USA

ABSTRACT

The advertising industry is developing very rapidly, especially outdoor advertising, which has attracted people's attention. All kinds of commercial advertisements can be seen everywhere in outdoor public places, but the advertising delivery system on the market is relatively simple in function, and the evaluation of advertising effect lacks effective automatic analysis means, mainly carried out by manual observation, which is low in efficiency and difficult to conduct quantitative evaluation, which directly leads to the lack of targeted advertising. Artificial intelligence advertising refers to the use of artificial intelligence technology (such as voice recognition, face recognition, deep learning, machine learning, etc.), investigation, production, publishing advertising and other fields, the combination of artificial intelligence and advertising brings great value and convenience to people's lives. Among them, the most common algorithm that can achieve accurate and intelligent advertising is NLP, and the artificial intelligence advertising push system based on natural language processing (NLP) algorithm can provide a variety of useful applications in the advertising field. These applications can help advertisers better understand user needs, improve the accuracy and effectiveness of ads, and provide a better user experience. In order to fully tap the advertising information value contained in unstructured data, this paper introduces the text mining technology based on natural language processing, explores from principle to practice, and analyzes the push process and application of intelligent advertisements in daily life by analyzing the implementation steps of NLP algorithm model.

KEYWORDS

Intelligent recommendation; NLP algorithm; Data mining; Artificial intelligence.

1. INTRODUCTION

In recent years, with the large-scale popularization of Internet and the improvement of enterprise informatization, more and more information is accumulated in enterprises. According to statistics, corporate data is growing at a rate of 200% every year, of which 80% of the data is stored in unstructured data such as files, emails, and pictures in the computer system of the enterprise. In the face of the above massive data, if there is no powerful tool to make the storage of a large number of raw data to be fully utilized and transformed into "knowledge" to guide production, then the data collected in many databases will become a "data grave", and form a "data of the ocean knowledge desert such a strange phenomenon[1]." As a result, important decisions are often based not on the rich data in the database, but on the intuition of decision makers, who lack the tools to extract valuable

knowledge from the vast amount of data. In order to screen and push web advertisements intelligently, a design of web advertisement intelligent push system based on the principle of visual information transmission is proposed. In order to better drive the vigorous rise of the Internet finance industry, the online advertising industry, especially the reasonable push of outdoor advertising, has gradually received more attention[2]. At present, the content of push advertisements on web pages is uneven, and deceptive advertisements are still widespread, so it is necessary to further check and screen the advertising content before pushing web ads. However, in the current search of web advertising content based on visual information transmission, because the current online advertising push mode is generally simple, most of them adopt the analysis and evaluation mode of manual background, which fails to form an efficient independent analysis structure, the overall efficiency of advertising screening and push is not high, and it is difficult to conduct quantitative analysis of advertising content. It is easy to lead to the lack of pertinence of the advertising market model.

Therefore, combining artificial intelligence and deep learning algorithms to realize the push system of Internet advertising web pages has become the best publicity tool. Through relevant advertising push algorithms, more convenient and accurate advertising transmission can be achieved, so as to better improve the self-delivery and push mode of online advertising and achieve reasonable push of web ads[3].

2. RELATED WORK

Text mining technology includes Natural Language Processing (NLP), information extraction, data mining and other technologies. Unstructured data can be processed to extract potential and important information that customers are interested in, which is a process of transforming unstructured data into structured data

Generally speaking, the classification application of text mining mainly involves the following processes: First, the sample is classified and labeled according to the label system set by manual recognition, and the model training set is constructed; second, text classification tools are used for text segmentation and preprocessing. Extract text features and transform text data into structured data that can describe text content. Thirdly, based on naive Bayes algorithm, the feature vector and classification contribution (TF-IDF value) are automatically calculated, and the classification rule table is output to build a model; Finally, data mining techniques such as classification, clustering and association analysis are used to discover new concepts and obtain corresponding relationships according to this structure[4-6]. Text Grocery, Open NLP, Weka, GATE, etc. Text Grocery is a short text classification tool based on Lib Linear and Jieba segmentation. It is efficient and easy to use, and supports both Chinese and English corpus[7]. Text Grocery is used for text model training in Python environment. The complaint content of work order is intelligently classified based on well-trained and accurate model.

2.1. Natural language processing

NLP technology involves a variety of processing algorithms and thus derived a variety of models, Hidden Markov Model (HMM) is one of them. Suppose that the hidden state sequence generated by the hidden Markov chain is Q=q1,q2... qr, its set of all possible states, S = {s1, s2,... sN}; The observation sequence is O = {o1, o2,... ,oT}, the observations at any one time are from a finite set of observations, denoted V = {v1, v2,... , vM}. This particular type of "memory lessness" is called the Markov property. The recurrence formula can be obtained:

$$P(o_t|o_1, o_2, \cdots, o_{t-1}, s_1, s_2, \cdots, s_t) = P(o_t|s_t) \quad (1)$$

The hidden Markov model is determined by the initial probability distribution, the state transition probability distribution and the observed probability distribution. The form of the hidden Markov model is defined as follows:

Let Q be the set of all possible states and V be the set of all possible observations.

$$Q = \{q_1, q_2, \cdots q_N\},\ V = \{v_1, v_2, \cdots, v_M\} \tag{2}$$

Where S is the state sequence of length T, and O is the corresponding observation sequence.

$$S = (s_1, s_2, \cdots, s_T)\ ,\ O = (o_1, o_2, \cdots, o_T) \tag{3}$$

A is the state transition probability matrix :

$$A = [a_{ij}]_{N \times N} \tag{4}$$

$$a_{ij} = P(s_{t+1} = q_j | s_t = q_i),\quad i = 1, 2, \cdots, N; j = 1, 2, \cdots, N \tag{5}$$

Based on the above (Viterbi) algorithm based on dynamic programming, the prediction problem of HMM can be solved, and the result of text word segmentation can be obtained. Therefore, the advantage of Hidden Markov Model (HMM) in advertising intelligent push is that it can model the hidden interest and behavior pattern of users, and identify the possible interest transformation and behavior change of users by analyzing the historical behavior data of users. This more accurately predict the needs of users and provide personalized advertising content, improve the effect of advertising push and user satisfaction.

2.2. TF-IDF algorithm

After text segmentation, in order to obtain the most characteristic information and generate a classification rule table to train the model, it is also necessary to use tools based on algorithms such as naive Bayes to calculate the feature vector of word segmentation, including word frequency Term Frequency)(TF) and Inverse Document Frequency (IDF). Assessing the importance of a word in a document can be described using the category Contribution TE-IDE value TF-IDF can be expressed as:

$$\mathrm{TFIDF}_i(d) = tf_i(d) \times \ln\left(\frac{N}{n_i}\right) \tag{6}$$

Where, the TF value of feature word i in text d is t(d), n is the number of texts containing feature word i, and V is the number of texts. This shows that the importance of words increases with the number of occurrences in the document and decreases with the number of occurrences in all documents.

Therefore, before the implementation of NLP algorithm push, it is necessary to label the complaint work order reply text that typically belongs to the network side of the training text classification model to form the training set of the model, which is labeled as weak coverage, interference, and fault alarm high load. The Grocery tool was used for modeling and training based on training sets. By processing the data set and tuning the model parameters, the accuracy of the output model is more than 80%[8]. The trained model is used to classify the work order text, and the classification results of network problems are obtained. Then, through the physiochemical problem analysis, it is easy to see that a large number of customer complaints are actually only clustered in a few areas - such as a residential area, which is the focus of attention and treatment. After the residential community with a large number of complaints and belonging to the problem of weak network coverage is given priority to open new base stations, combined with MR Data analysis, it is found that the final weak network coverage of relevant areas has been improved. The customer network end experience has been further improved.

3. METHODOLOGY

Personalized recommendation is an indispensable technology in advertising, which plays an important role in the fields of e-commerce, information distribution, computational advertising, Internet finance and so on. Specifically, personalized recommendation plays a core role in efficient utilization of traffic, efficient distribution of information, improvement of user experience, and mining of long-tail items[9]. In the recommendation system, it is often necessary to deal with various text data, such as product description, news information, user message and so on. Specifically, the NLP algorithm model can be implemented through the following steps:

3.1. Text data integration

In the process of realizing intelligent advertising push, it is necessary to realize the collection and combing of intelligent data, which is also the first step of the recommendation process to generate the collection of items to be recommended. The core operation of this part is to obtain the corresponding item set according to various recommendation algorithms[10-12]. Text data is a very important type of recall algorithm, which has the advantages of independent user behavior and rich diversity, and plays a very important role in the case of rich text information or lack of user information. In the daily advertising data collection, text data is a kind of large, complex and rich data, which plays an important role in the recommendation system. Therefore, the model analysis of common text processing methods in the recommendation system will be carried out based on the abovementioned aspects.

The first is the Bag of Words model (BOW model for short) is the simplest text processing method, and its core assumption is very simple, that is, a document is composed of multiple sets of words in the document (multiple sets and ordinary sets are different in considering the number of occurrences of elements in the set). This is the simplest assumption that does not take into account other important factors in the document such as grammar, word order, etc., and only takes into account the number of occurrences of words. Such a simple assumption obviously throws away a lot of information, but the benefit is that it is simpler to use and calculate, and also has greater flexibility.

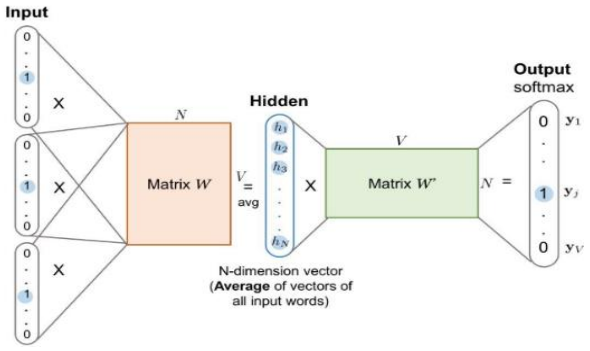

Figure 1. Bag of words model flow

For data collection, the bag of words model will first carry out word segmentation. After word segmentation, we can get the word-based features of the text by counting the number of occurrences of each word in the text. If these words of each text sample are put together with the corresponding word frequency, it is often called vectorization. After vectorization is completed, TF-IDF is generally used to correct the weight of the features, and then the features are standardized. After some other feature engineering, the data can be brought into the machine learning algorithm for classification and clustering.

There are obviously a lot of problems with this simple approach:

First of all, after the text is divided into words, not every word can be used for recall and sorting, for example, "stop words" such as "you and me" should be removed, in addition, some words with high

or low frequency also need to be special treatment, otherwise it will lead to low relevance of recall results or recall results are too few and other problems.

Second, using word frequency to measure importance is not reasonable enough. Using the above "down jacket" recall as an example, if you use the frequency of the word "down jacket" in the description to measure the relevance of the product in the down jacket category, it will result in all down jackets having similar relevance because everyone will use a similar amount of the word in the description. So we need a more scientific way to measure the correlation between texts.

In addition to the above usage, we can also add each word in the bag to the ranking model as a one-dimensional feature. For example, in a CTR ranking model modeled on LR, if the weight of this one-dimensional feature is w, it can be interpreted as "samples that include this word have higher log odds of click-through rates than samples that do not include this word." In order to enhance the distinguishing ability of features, we often use N-gram bag model, an upgraded version of simple bag model, when using word features in ranking model.

3.2. Weight calculation and vector space model

From the above, we see that the simple bag of words model can be used to recall candidate items in the recommendation system after proper preprocessing. However, when calculating the correlation between items and keywords, as well as the correlation between items, it is obviously unreasonable to only use simple word frequency as a ranking factor. In order to solve this problem, we can introduce a more expressive weight calculation method based on TF-IDF. In the TF-IDF method, the weight of a word t in document d is calculated as:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t = tf_{t,d} \times log\frac{N}{df_t} \tag{7}$$

Where tft,d represents the frequency of occurrence of t in d, while dft refers to the number of documents containing t, and N represents the number of total documents.

The TF-IDF and its various modifications and variants (For a detailed description of the variants and variants of the TF-IDF, see Chapter 6 of the Introduction to Information Retrieval.) The core improvement over the simple TF method lies in the importance measurement of a word, such as:

a. The original TF-IDF adds the consideration of IDF on the basis of TF, thus reducing the importance of words with high frequency resulting in no distinguishing ability, typical stop-words.

b. Because the importance and frequency of occurrence of a word in a document are not completely linearly correlated, non-linear TF scaling logs TF scaling to reduce the weight of words with particularly high frequency of occurrence.

c. The frequency of words appearing in documents is not only related to their importance, but also to the length of documents. In order to eliminate this difference, the maximum TF can be used to normalize all TFS.

The purpose of these methods is to make the measurement of the importance of words in documents more reasonable. On this basis, we can improve the method based on word frequency. For example, we can improve the method of sorting items by word frequency to TF-IDF score.

3.3. Application of LDA

In this part, we introduce some points that need attention when LDA is used as similarity calculation and ranking features, and then introduce more applications of text topics represented by LDA in recommendation systems from different perspectives.

(1) Similarity calculation

As mentioned above, LSA can be directly applied to VSM for similarity calculation, and similar calculation can also be done in LDA. The specific method is to vectorize the topic distribution value of the document and calculate it with the cosine formula. However, replacing cosine similarity with KL divergence or Jensen-Shannon divergence has a better effect, because the subject distribution given by LDA is probability value with clear meaning, and it is more reasonable to measure the similarity between probabilities.

(2) Sorting features

Having an item's LDA theme as a feature of the sorting model is a natural way to use it, but not all themes are useful. There are generally two types of theme distribution on items:

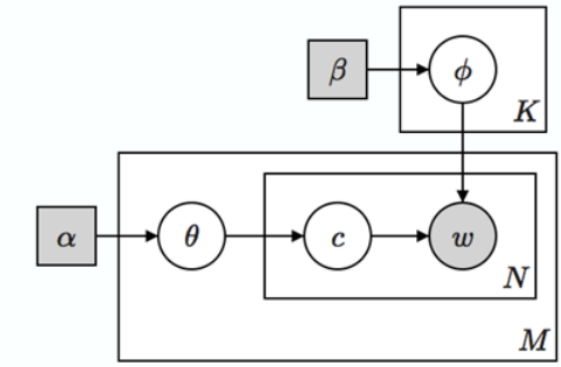

Figure 2. LDA generation process

a. There are a few topics (three or fewer) that occupy a relatively large probability, and the remaining topics add up to a relatively small probability.

b. All topics have similar probability values and are relatively small. In the first case, only the first few topics with high probability are useful, while in the second case, basically all the topics are useless. So how do you identify these two situations? In the first method, you can make a simple K-means clustering on the topic according to the probability value of the topic, K is chosen as 2, if the first case, then the number of topics in the two classes will be very different - one class contains a few useful topics, the other class contains other useless topics; In the second case, the number of topics is not very different, and you can use this method to identify the importance of the topic. The second method can calculate the information entropy of the topic distribution. The information entropy corresponding to the first case will be relatively small, while the second case will be relatively large. Selecting the appropriate threshold can also distinguish the two cases.

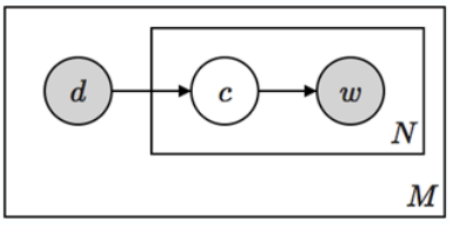

Figure 3: pLSA generation process

(3) item labeling & user labeling

After calculating the corresponding theme for the item and the corresponding word distribution under the theme, we can select several topics with the greatest probability, and then select several words with the greatest probability under these topics as the label of the item. Based on this, these labels can be spread to the user if the user acts on the item.

The labels printed in this method have a very intuitive explanation, and can act as a reason for recommending explanations in appropriate scenarios. For example, when we do mobile personalized push, the space available for display copy is very small, we can first label the item in the above way, and then spread the label to the user according to the user, and use these label words as the recall source and recommendation reasons when pushing, so that the user can understand why he made such a recommendation[13].Such an assumption makes LDA lose some important information, and in recent years, the neural probabilistic language model represented by word2vec, which has received more and more attention, just forms a certain degree of complementary relationship with LDA in this respect, so as to capture information that LDA cannot capture.

4. CONCLUSION

After the text topic model was proposed, the advertising intelligent autonomous push system was designed. NLP algorithm has been widely used in various industries of the Internet because of its good probabilistic properties and meaningful clustering abstraction ability for text data[14]. The search giant Google makes extensive use of the text theme model in all aspects of its system and has developed the massive text theme system Rephil for this purpose. For example, in the process of generating advertisements for users' searches, text themes are used to calculate the match between web content and advertisements, which is one of the important factors for the success of its advertising products. In addition, text themes can be used to improve matching recall and accuracy when matching relationships between user search terms and web pages[15]. Yahoo! Also makes extensive use of LDA theme features in its search ranking model, and has also open-source the famous Yahoo! LDA tool.

Therefore, it shows that text mining technology based on natural language processing can fully explore the implementation and application of intelligent advertising recommendation[16-17]. Under the background of continuous development and development of artificial intelligence and big data, the information contained in unstructured data in online public opinion can be better understood by classification and exploration using this information. NLP algorithm combines the multi-path push principle for intelligent push path selection, and removes interference factors in the network environment, so as to better achieve, and web advertising push guides the relevant network perception improvement work, which can better help improve customer satisfaction.

ACKNOWLEDGEMENT

This paper expresses deep gratitude and respect to Tian Miao, Zepeng Shen and other authors for their research work. Their article in The Academic Journal of Science and Technology, "The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier provides us with valuable knowledge and inspiration. This paper discusses in detail the application of artificial intelligence in the field of medical diagnosis and its cutting-edge nature, which greatly enriches our research vision and guides us a new research direction.

Link: Tian, Miao, et al. "The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier." Academic Journal of Science and Technology 8.2 (2023): 57-61.

REFERENCES

[1] Zhang Le, Tang Liang. Opportunities for linguists in the age of artificial intelligence[J]. Computer Knowledge and Technology, 2020 (24) :195-197.

[2] LIU Hualiang, Du Kun, Qin Chunxiu. Chinese language based on semantic similarity of Know net Research on text classification [J]. Modern Library and Information Technology, 2015 (2) :39-45.

[3] Liu, B., Yu, L., Che, C., Lin, Q., Hu, H., & Zhao, X. (2023). Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms. arXiv preprint arXiv:2312.12872.

[4] Jiang Z ,Gao S . An intelligent recommendation approach for online advertising based on hybrid deep neural network and parallel computing [J]. Cluster Computing, 2019, 23 (3): 1-14.

[5] Che, C., Liu, B., Li, S., Huang, J., & Hu, H. (2023). Deep learning for precise robot position prediction in logistics. Journal of Theory and Practice of Engineering Science, 3(10), 36-41.

[6] Hao Hu, Shulin Li, Jiaxin Huang, Bo Liu, and Change Che. Casting product image data for quality inspection with xception and data augmentation. Journal of Theory and Practice of Engineering Science, 3(10):42–46, 2023. https://doi.org/10.53469/jtpes.2023.03(10).06

[7] Chang Che, Qunwei Lin, Xinyu Zhao, Jiaxin Huang, and Liqiang Yu. 2023. Enhancing Multimodal Understanding with CLIP-Based Image-to-Text Transformation. In Proceedings of the 2023 6th International Conference on Big Data Technologies (ICBDT '23). Association for Computing Machinery, New York, NY, USA, 414–418. https://doi.org/10.1145/3627377.3627442

[8] Lin, Q., Che, C., Hu, H., Zhao, X., & Li, S. (2023). A Comprehensive Study on Early Alzheimer's Disease Detection through Advanced Machine Learning Techniques on MRI Data. Academic Journal of Science and Technology, 8(1), 281–285.DOI: 10.1111/jgs.18617

[9] Che, C., Hu, H., Zhao, X., Li, S., & Lin, Q. (2023). Advancing Cancer Document Classification with R andom Forest. Academic Journal of Science and Technology, 8(1), 278–280. https://doi.org/10.54097/ajst.v8i1.14333

[10] Fu Xiaofeng, Wu Jun, NIU Li. Spontaneous expression classification in deep migration network of small Data samples [J]. Chinese Journal of Image and Graphics, 2019,24 (5) :753-761.

[11] Research on network marketing strategy based on big Data precision marketing [J]. Business Economics Research, 2017 (11) :46-47.

[12] Yang Xian shun, Molly. Empirical Research on "Availability emergence" of personalized advertising based on Algorithm recommendation in Intelligent Marketing Communication [J]. University press, 2022, (11) : 1-15 + 116.

[13] Tian, Miao, et al. "The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier." Academic Journal of Science and Technology 8.2 (2023): 57-61.

[14] Wan, Weixiang, et al. "Development and Evaluation of Intelligent Medical Decision Support Systems." Academic Journal of Science and Technology 8.2 (2023): 22-25.

[15] Pan, Linying, et al. "Research Progress of Diabetic Disease Prediction Model in Deep Learning." Journal of Theory and Practice of Engineering Science 3.12 (2023): 15-21.

[16] Shen, Zepeng, et al. "The Application of Artificial Intelligence to The Bayesian Model Algorithm for Combining Genome Data." Academic Journal of Science and Technology 8.3 (2023): 132-135.

[17] Zong, Yanqi, et al. "Improvements and Challenges in StarCraft II Macro-Management A Study on the MSC Dataset." Journal of Theory and Practice of Engineering Science 3.12 (2023): 29-35.