Research on fused sorting based on logical regression in news recommendation system

Xing Lijie, Feng Xiwei, Chen Haiming, Wang Ying and Zhang Yue
Liaoning Shihua University, Fushun City, Liaoning Province, China
Corresponding author’s e-mail address: jiemailbox@163.com

Abstract. A fusion sorting model based on logical regression is used to solve the problem of information overload in the news recommendation system and meet the personalized needs of users. First, the fusion sorting model is trained based on the article data and user data, and the original candidate news set is obtained by multi-channel recall method. Then the original candidate set is filtered by the fusion sorting model, and the recommendation list is generated. Finally, the experiment is carried out based on the Wuli News Recommendation System. Under the evaluation indexes of precision, recall and ILS, the fusion sorting model adopted significantly improves the recommendation ability of the system, and achieves the goal of making multiple single recommendation strategies make up for each other's shortcomings.

1. Instruction
There are some problems with the current personalized recommendation algorithm. For example, collaborative filtering recommendation algorithm [1] has typical cold start and sparse matrix problems; news pushed by content-based recommendation algorithm [2] may lack novelty for users and make it difficult for users to find new types of interest [3]; hot spot-based recommendation strategy [4] focuses only on shallow statistical analysis of relevant data. When using a variety of recommendation algorithms in personalized news recommendation system, how to make more effective use of each algorithm to provide users with a better reading experience, improve news click-through rate and recall rate becomes a direction worth studying when improving the recommendation system.

By fusing the recall results of multiple recommendation algorithms, the fused ranking model studied in this paper is designed to make multiple recommendation algorithms make up for each other, improve the news click-through rate and recall rate, and improve the user experience. To achieve this goal, this paper first builds article features and user features based on article data and user data, and trains a fused sorting model based on logical regression using feature data. In candidate news sets recalled based on a variety of recommendation algorithms, the sorting model filters out the eligible news as the recommendation news. In the online test of the Wuli News Recommendation System, the fused sorting model based on logistic regression implemented in this paper significantly improves the click-through and recall rate of the recommendation system.

2. Related work
2.1. Building article features
Sample features and sample tags need to be used as model inputs to train and use the fusion sorting model. Sample features are constructed based on article features and user features. Before building the article features, we need to build the key words of the article and its weight. After the article is...
segmented by the word breaker, TF-IDF [5] and TextRank [6] are used to calculate the word weight. Take the product of TF-IDF value and TextRank value of each word as the weight of the word, and the processing results are shown in Table 1. The top 20 words were selected as keywords. The 30 words with the highest TextRank value and the 30 words with the highest TF-IDF value are the theme words.

Table 1. Word weight example.

| Article Id | Word | TF-IDF value | TextRank value | Weight |
|------------|------|--------------|----------------|--------|
| 12983      | progress | 5.1971     | 1.5098         | 7.8466 |
| 45695      | Progress  | 12.6935    | 2.6482         | 33.6149|
| 95263      | progress  | 2.5986     | 0.1977         | 0.5134 |
| 12983      | AI        | 74.0591    | 2.1647         | 160.3157|
| 45695      | AI        | 12.9928    | 2.6482         | 34.4075|
| 95263      | AI        | 9.095      | 5.3624         | 48.7710|

To construct article features, we need the word vector of the word. Because the meaning of the same word is slightly different in different channels, this paper trains word vector model for each channel. First, all the historical articles in the selected channel are segmented, and the Word2Vec model [7] is trained with the segmentation results as the training sets to get the word vectors of all the words in the channel. The article key words and word vector results are correlated to get the word vector of each key word. Then calculate the eigenvector $V'_j$ of the article based on the following formula:

$$V'_j = \frac{\sum_{j=1}^{K} W'_j * v'}{K}$$

Where, $W'_j$ is the weight of the j-th key word arranged in reverse order of weight in article $D_i$, and $v'$ is the word vector corresponding to this key word.

2.2. Building user features

User features are represented by one-dimensional array consisting of user basic information and user tag weight list. Because user's preferences vary greatly in different channels, the features of users in different channels should be constructed. User base information is converted from non-numeric to numeric using different conversion methods, some of which are treated as shown in Table 2.

Table 2. User basic information processing example.

| Feature name          | Feature sample | Processing mode           | Result         |
|-----------------------|----------------|---------------------------|----------------|
| Gender                | male, female  | one-hot                   | 01, 10         |
| Registration platform | android        | one-hot                   | 00000100       |
| Registration time     | 1520486537     | intercept the first four digits | 1520        |
| Province              | Beijing        | one-hot                   | 00001000       |
| Weekly browsing volume| 238, 25        | equal frequency division  | 12, 2          |
| Registration type     | other          | one-hot                   | 00010000       |

Through the article ID, the user's historical behavior data is associated with the article data, and the subject words of the article are used as the user tags. The tag weight is calculated according to the user's behavior and timeliness of the article. The calculation formula of user label weight is:

$$w = \frac{1}{log(t) + 1} \sum_{a} Q_a$$

(2)

Where, $t$ is the number of days from the time when the behavior occurred to the current time. $A$ is the set of all behaviors of the user to the article, and $Q_a$ is the corresponding weight of behavior $a$.

2.3. Multiple recalls
Multiple recalls run multiple recommendation strategies in parallel, and takes the recall results as the original candidate news set. The recommendation strategies used in this paper are ALS (Alternative Lead Squares) recommendation model [8], content-based recommendation model [9] and hotspot based recommendation model [10].

3. Construction of Fusion Sorting Model

3.1. Logistic Regression

Logistic regression is a generalized linear regression [11] analysis model, which is based on linear regression and adds a logical function to predict the probability that users will click on articles in this paper. The hypothetical function of logistic regression is defined as:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$  \hspace{1cm} (3)

The function value is the probability that the user will click, $\theta$ is a set of model parameters, and $X$ is the sample eigenvector. The key to learning Logistic Regression Model is to find the optimal solution of parameter $\theta$ to minimize the loss function, which is defined as:

$$L(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2$$  \hspace{1cm} (4)

Where $m$ is the number of samples. The second half of the function is a regularization item, which is designed to prevent the model from overfitting, enhance the generalization ability of the model, and $\lambda$ is the regularization coefficient. When clicking occurs, the $y$-value of the sample label is 1, otherwise $y$ is 0. The objective function derives $\theta$:

$$\frac{\partial}{\partial \theta_j} L(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left( h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)} - \lambda \theta_j$$  \hspace{1cm} (5)

In this paper, a gradient descent method is used to find the optimal value of parameter $\theta$. First, the parameters $\theta$ are initialized randomly, and then the following formula is used to iterate until the value of the objective function meets the requirements or the iteration steps reach the upper limit.

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} L(\theta)$$  \hspace{1cm} (6)

3.2. Construction of fusion sorting model

This paper uses a fused sorting model based on logistic regression to recommend the article. First, the fused sorting model is trained with historical data, and the original recommendation list is derived from multiple recalls. Then the fusion ranking model is used to predict the click rate of candidate news in the list, and the top 10 news of click rate is retained as the final recommendation list. The algorithm flow is shown in Figure 1.
Figure 1. Algorithm flow.

**Stage 1:** fusion sorting model training  
*Input:* user's historical behavior record, user features, article features  
*Output:* fusion sort model  
*Algorithm steps:*  
   a) Use user features and article features to build sample features, read user historical behavior data, mark the sample with click behavior as 1, otherwise mark as 0.  
   b) Balance the sample set so that the number ratio of the two types of samples is 1:1.  
   c) Use the logistic regression algorithm to train the model and adjust the parameters, and finally get the model.

**Stage 2:** generating recommendation list  
*Input:* fusion sorting model, user features, candidate set  
*Output:* recommendation list  
*Algorithm steps:*  
   a) Use multiple recalls to get the original candidate article set.  
   b) Combine candidate article set and user feature to build sample feature.  
   c) Use the ranking model to predict the click through rate of candidate articles, and the 10 news with the highest click through rate is retained as the final recommendation list.

4. Results and analysis of experiment

This paper uses the Precision [12], Recall [13] and Intra-List Similarity (ILS) [14] to objectively evaluate the validity of the fused sorting model, and uses the Wuli News Recommendation System to test the fusion sorting model online. The specific test methods are as follows: by the ABTest method, the user traffic is randomly divided into four parts, and the results are taken as the control group by using ALS model, content-based model and hotspot based model respectively, and the results are taken as the experimental group by using the fusion sorting model, so as to compare and analyze the recommendation effect of the fusion ranking model.

According to the above design, online experiments are carried out, and the experimental results are statistically analyzed. The average precision and average recall of different recommended strategies are calculated respectively, and the experimental results are shown in Figure 2. The average value of ILS of the recommendation results of the four recommendation models for a single user is calculated, and the results are shown in Figure 3. The users are divided into different categories according to the accumulated active days in the past week, and the average precision of each recommendation strategy is calculated for different users. The analysis results are shown in Figure 4.
It can be seen from Figure 2 that under the two indicators of precision and recall, the fusion sorting model is superior to the other three independently used recall models, which proves that the fusion sorting model based on logical regression adopted in this paper can effectively improve the recommendation ability of the news system. As shown in Figure 3, compared with hotspot based model and ALS model, the ILS value of recommendation results of content-based model is significantly higher, that is, the diversity of recommendation results of content-based model is poor. Compared with the other three recommendation models, the ILS value of the fusion sorting model is significantly lower, that is, the fusion sorting model can improve the diversity of the three recommendation models, especially for the content-based recommendation model.

The analysis results in Figure 4 show that under the precision index, the ALS model is significantly lower than the hotspot based model and content-based model for the recommendation ability of users with lower activity. For users with high activity, ALS model is better than hotspot based model and content-based model. This shows that there is a typical cold start problem in ALS model. No matter what kind of users, the recommendation ability of the fusion sorting model is higher than the other
three, that is, fusion sorting model can significantly improve the cold start problem of ALS model, effectively use the other two recommendation models to recommend low-active users, and effectively use the ALS model to recommend high-active users. These results show that the fusion sorting model based on logistic regression can effectively improve the recommendation ability of news system.

5. Conclusion and prospect
Aiming at the defect of single recommendation model in personalized news recommendation system, this paper adopts a fusion sorting model based on logical regression. Firstly, based on the article data, user data and user historical behavior data, the sample features and sample tags are constructed, and then the fusion sorting model is trained. Second, we use multiple recalls to get the original candidate article set, and by using the fusion sorting model to predict the click rate of candidate articles, retain the 10 articles with the highest click rate as the final recommendation list. Finally, the experiment is carried out in the Wuli News Recommendation System. The experimental results show that the fusion sorting model adopted in this paper effectively improves the recommendation effect of the news system, and achieves the expected goal. In the future research work, we can further consider using other recall strategies in multi-channel recalls, and build more accurate user and article models to better improve the recommendation ability of the system.

Acknowledgments
This research was financially supported by the Natural Science Foundation of Liaoning Province of China(20180550130).

References
[1] Lei Man, Long Qing and Wang Jichao 2019 Collaborative filtering recommendation algorithm based on label weight Computer application 39(3) p 634
[2] Yan Dongdong and Li Hongqiang 2016 An improved content-based personalized recommendation model Software guide 15 (4) p 11
[3] Wang Guoxia and Liu Heping 2012 Overview of personalized recommendation system Computer engineering and application 48 (7) p 66
[4] Wang Wei and Xu Xin 2009 Discovery and analysis of hot issues in Internet public opinion based on clustering Information analysis and research 3 p 74
[5] Li J and Zhang K 2007 Keyword extraction based on tf/ï Ifd for Chinese news document Wuhan University Journal of Natural Sciences 12 (5) p 917
[6] Mihalcea R and Tarau P 2004 TextRank: Bringing Order into Texts Emnlp 1 p 404
[7] Mikolov T, Sutskever I and Chen K 2013 Distributed Representations of Words and Phrases and their Compositionality Advances in Neural Information Processing Systems 26 p 3111
[8] Deshpande,Mukund,and George Karypis 2004 Item-based top-n recommendation algorithms ACM Transactions on Information Systems(TOIS) 22.1 p 143
[9] Pazzani M J and Billsus D 2007 Content-based recommendation systems (The adaptive web. Springer Berlin Heidelberg) pp 325-341
[10] Liu Xing, he Tingting and Gong Haijun 2008 Design of network hot event discovery system Chinese Journal of information 22 (6) pp 80-85
[11] Gao Huixuan 2005 Statistical calculation (Beijing: Peking University Press) p 44
[12] Hu Y, Koren Y and Volinsky C 2008 Collaborative filtering for implicit feedback datasets (Pisa: Eighth IEEE International Conference on Data Mining) pp 263-272
[13] Jiang M, Cui P and Liu R 2012 Social contextual recommendation (Maui, HI, USA: Proceedings of the 21st ACM international conference on Information and knowledge management) pp 45-54
[14] Wu Kuo-Lung and Yang Miin-Shen 2002 Alternative C-means Clustering Algorithms Pattern Recognition 35(10) pp 2267-2278.