Research on algorithm recommendation mechanism and characteristics of personalized news app—Taking "Toutiao" as an example

Yuan Yuan
Beijing Normal University, Beijing, China

Abstract. With the development of mobile Internet, new technologies such as algorithm distribution and personalized recommendation came into being. The media appeared intelligent trend and the era of intelligent media came. This article uses China's most popular algorithm to push news app "Toutiao" as an example. The app's recommendation system is analyzed from the basic principles of algorithm recommendation, text recommendation and personalized tag. For news app, algorithm push is an inevitable trend, and the principle of algorithm recommendation mechanism of Toutiao is studied and beneficial. The platform regulates its own behavior and explores the solution for further development.

1. Introduction
With the development of the mobile Internet, new technologies such as algorithm distribution and personalized recommendation emerged as the times require, and the media appeared to be intelligent, and the era of intellectual media came. According to the 42nd China Internet Network Development Report released by China Internet Network Information Center (CNNIC), as of June 30, 2018, the number of Chinese Internet users has reached 802 million, and the Internet penetration rate is 57.7%. The scale reached 788 million, and the proportion of Internet users accessing the Internet through mobile phones was as high as 98.3%. According to the 2018 New Media Trends Report released by Penguin Think Tank in October 2018, the average daily consumption of information for Internet information consumers in China reached 76.8 minutes per person, and 73.7% of users exceeded 30 minutes per day. App is the first way for netizens to get information.

Toutiao are the fastest growing new media companies in China's Internet industry. From a valuation of 500 million in 2014 to a value of 75 billion in 2018, it has increased by 150 times in just 4 years. The interest recommendation algorithm is one of the key factors in the success of this news app. The algorithm breaks through the limitation of the number of manual editing recommendations, and realizes the personalized customization of the content, thereby improving the user's usage time and opening frequency, making the application quickly become a first-line platform. However, since 2018, the controversy over algorithmic checks has also kept it in a regulatory crisis and negative public opinion. For news APP, algorithm push is an inevitable trend. The principle of the algorithm recommendation mechanism of Toutiao and its existing problems and reasons are beneficial to the platform to regulate its own behavior and explore solutions to further develop.
2. The basic principle of Toutiao recommendation system

The recommendation system, if used to describe in a formal way is actually a function that fits a user's satisfaction with the content, this function needs to input variables of three dimensions. This function is described as:

\[ y = F(X_i, X_u, X_c) \]

First, the content is the first dimension. "Toutiao" is now a comprehensive content platform, including graphics, video, short videos, Q&A, and micro-headlines. Within this context, each content contains its own characteristics, and you need to consider how to extract feature optimization recommendations for different content types. The second dimension is the user feature. Including various interest tags, occupations, ages, genders, etc., as well as a number of models portrayed implicit user interests. The third dimension is the environmental characteristics. This is a recommended feature in the era of mobile Internet. Users can move anywhere, anytime, in different situations such as work, commute, travel, etc., and information preferences are offset. Combined with the three dimensions, the model will give an estimate of whether the recommended content is appropriate for this user in this scenario.

In a recommendation model, click-through rate, reading time, likes, comments, and forwards are all quantifiable goals. You can use the model to directly fit the estimate, and see the online promotion to understand the recommended effect. However, for a large-scale recommendation system, there are many service users, and it is not completely evaluated by indicators. It is also very important to introduce elements other than data indicators, such as advertising and special content frequency control. A question and answer card is a special form of content. The recommended goal is not to let users browse, but also to attract users to answer content for the community. How to mix and sort these content and common content and how to control the frequency need to be considered. In addition, the platform is based on the consideration of content ecology and social responsibility, such as the review of vulgar content, the review of title party, low-quality content, the topping, weighting, and insertion of important news, and the reduction of rights of low-level account content are all impossible by the algorithm itself. Further intervention in the content is needed.

In fact, the formula \( y = F(X_i, X_u, X_c) \) is a very classic supervised learning problem. There are many achievable methods, such as the traditional collaborative filtering model, the supervised learning algorithm Logistic Regression model, the deep learning based model, the Factorization Machine and the GBDT. An excellent industrial-level recommendation system requires a very flexible algorithmic experimental platform that can support multiple algorithm combinations, including model structure adjustments. Because it is difficult to have a common model architecture for all recommended scenarios. It is now very popular to combine LR and DNN. In the past few years, Facebook has also combined LR and GBDT algorithms. Toutiao products are using the same powerful algorithm recommendation system, but the model architecture will be adjusted according to different business scenarios.

Data processing is the process of transforming data into user item scoring matrix. The user's rating of the project is generally divided into two types: one is the score system, that is, the user gives 1-10 points to the project, the higher the score is, the more satisfied he is with the project, and the common five levels of liking (very like, like, general, dislike, very dislike) also belong to the score system. The second is the 0-1 scoring system, that is, the user's interaction with the project is recorded as 1, otherwise it is recorded as 0. After data processing, a \( m \times n \) user item rating matrix \( R \) can be obtained.

\[
R = \begin{bmatrix}
    r_{u_1, i_1} & r_{u_1, i_2} & \cdots & r_{u_1, i_n} \\
    r_{u_2, i_1} & r_{u_2, i_2} & \cdots & r_{u_2, i_n} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{u_m, i_1} & r_{u_m, i_2} & \cdots & r_{u_m, i_n}
\end{bmatrix}
\]

Formula 1. User-item rating matrix
There are four types of features that play a more important role in recommending features. The first category is the relevance feature, which is to evaluate the attributes of the content and whether it matches the user. Explicit matching includes keyword matching, classification matching, source matching, topic matching, and the like. There are also some implicit matches in the FM model, which can be derived from the distance between the user vector and the content vector. The second category is environmental characteristics, including geographic location and time. These are both bias features and can be used to build some matching features. The third category is the heat feature. Including global heat, classification heat, theme heat, and keyword heat. Content popularity information is very effective in large recommendation systems, especially when the user is cold-starting. The fourth category is the synergistic feature, which can help to solve the problem that the so-called algorithm is narrower and narrower. The collaborative feature does not take into account the user's history. Rather, user behavior is used to analyze the similarity between different users, such as click similarity, similar interest classification, similar theme, similar interest words, and even vector similarity, thus expanding the exploration ability of the model.

In the training of the model, "Toutiao" use real-time training. Real-time training saves resources and feedback is fast, which is very important for information flow products. The user needs behavior information that can be quickly captured by the model and fed back to the recommended effect of the next brush. The online data processing is based on the storm cluster in real time, including action types such as click, display, collection, and sharing. The model parameter server is a high-performance system developed internally. Because the data scale grows too fast, the stability and performance of similar open source systems cannot be met, and the bottom layer of their self-developed system has made many targeted optimizations and provided perfection. The operation and maintenance tools are more suitable for existing business scenarios. Currently, the "Toutiao" recommendation algorithm model contains tens of billions of original features and billions of vector features. The overall training process is: online server records real-time features, imported into the Kafka file queue, and then further imported into the Storm cluster to consume Kafka data, the client returns the recommended label to construct the training samples, and then updates the model parameters according to the latest sample online training. The final online model is updated. The main delay in this process is the user's motion feedback delay. Because the user does not necessarily look at it immediately after the article is recommended, the entire system is almost real-time if this part of the time is not considered.

3. Text Analysis: An Important Component of User Data Modeling
A very important role of text analysis in recommendation systems is user interest modeling. There is no content or text label to get the user's interest tag. On the other hand, the label of the text content can directly help recommend the feature. If the recommended narrowing phenomenon occurs because the recommended main channel is not ideal for a certain period of time, the user can read it in other channels, and the recommendation effect will be more obvious. Because the entire model is open, the sub-channel exploration space is smaller and it is easier to meet user needs. It is more difficult to improve the recommendation accuracy only by single channel feedback, so the subchannel is very important, and this is based on content analysis. For information products, most users consume the same day's content, no text features, new content is the basis, cold start is very difficult, and the collaborative feature can't solve the cold start problem of the article.

The main text features extracted by the headline recommendation system today include semantic tag class features, text similar features, and text features. The first is the semantic tag class feature, which explicitly tags the article with semantic tags. This part of the label is a feature defined by the person, each label has a clear meaning, and the label system is predefined. In addition, there are implicit semantic features, mainly topic features and keyword features. The topic features are descriptions of word probability distributions, and there is no clear meaning; while keyword features are based on some unified feature descriptions, and there is no clear set.

In addition, the online classification of the recommendation system uses a typical hierarchical text classification algorithm. Compared with the separate classifier, the hierarchical text classification
algorithm can solve the problem of data skew more effectively. There are some exceptions. If you want to improve the recall, you can connect the flying line. This architecture is universal, but depending on the difficulty of the problem, each meta-classifier can be heterogeneous. Some SVMs are very effective, some are combined with CNN, and some are combined with RNN. For example, based on word segmentation results and part-of-speech tagging, candidates may need to be spliced according to the knowledge base. Some entities are combinations of several words. It is necessary to determine which words are combined to map the description of the entity. If the result maps multiple entities, the differences are removed by word vector, topic distribution, and even word frequency itself, and finally a correlation model is calculated.

4. User tags: an important part of information recommendation

Content analysis and user tags are the two cornerstones of the recommendation system. The user tags commonly used in Toutiao include categories and topics of interest to users, keywords, sources, interest-based user clustering, and various vertical interest features. There are also gender, age, location, and other information. The gender information is obtained by logging in with the user's third-party social account. Age information is usually predicted by the model and estimated by model, reading time distribution, and so on. The resident location is from the user authorized access location information, and the resident point is obtained by the traditional clustering method based on the location information. The resident point combined with other information can be used to guess the user's work location, business location, and travel location. These user tags are very helpful for recommendations.

The simplest user tag is the content tag that was viewed. But here are some data processing strategies involved. Mainly includes: The first is to filter noise. Filter the title party by clicking on the short stay time. The second is hot spot punishment, which is to reduce the user's actions on some popular articles. The reason for this is that, theoretically speaking, the content of a large spread will reduce the confidence. The third is that as time decays, user interest shifts, so the strategy is more biased towards new user behavior. Therefore, as the user's motion increases, the old feature weight will decay with time, and the feature weight of the new action contribution will be greater. Fourth, the punishment shows. If an article recommended to the user is not clicked, the relevant feature (category, keyword, source) weight will be penalized. Of course, at the same time, we must also consider the global background, whether it is more relevant to push, and related off and dislike signals.

User tag mining is generally simple. The first version of the headline user label is a batch calculation framework. The process is relatively simple. The action data of the previous day's daily active users in the past two months is extracted every day, and the results are calculated in batches on the Hadoop cluster. But the problem is that with the rapid growth of users, the types of interest models and other batch processing tasks are increasing, and the amount of calculation involved is too large. Tight cluster computing resources can easily affect other work, and the pressure to focus on distributed storage systems is beginning to increase, and user interest tag update delays are getting higher and higher. In 2014, “Toutiao” went online with the user tag Storm cluster streaming computing system. After changing to the streaming mode, as long as the user action updates the tag, the CPU cost is relatively small, which can save 80% of the CPU time and greatly reduce the computing resource overhead. At the same time, only a few dozen computers can support the interest model update of tens of millions of users every day, and the feature update speed is very fast, and the basic real-time can be achieved.

5. Algorithm recommendation mechanism based on graph structure and social media

In addition, Toutiao also use the algorithm recommendation mode based on graph structure and the algorithm recommendation mechanism based on social network.

Graph structure is a special structure in network. It only contains two kinds of graph nodes and the connection between them. The set of nodes and lines constitutes the graph. Suppose there is a undirected graph network, which is represented by $G = \{V, e\}$. If the fixed point set $V$ of a graph can be represented by two disjoint subsets $u$ and $I$, that is, $u \in V, I \in V, u \cup I = \phi$. The graph $G$ is a
bipartite graph if the two fixed points \( u \) and \( I \) connected by any edge have \( u \in u \) and \( I \in I \). The simple bipartite diagram is shown in Fig. 1.

Fig 1. A typical bipartite graph structure

Recommendation based on social network mainly relies on some indicators of social relations to quantify the similarity or trust relationship between users. The similarity is used to judge the neighbor users of the target users, and the items interested by the neighbor users are formed into the recommended candidate set. Then, according to the similarity between users, the score of items is weighted, and the items of interest are recommended to the target users. In the recommendation of social network, the method to calculate the similarity between users is to know the familiarity and interest among friends. The familiarity of two users can be measured by the number of common friends. Generally speaking, the more common friends, the closer the familiarity. Its calculation is shown in formula 2.

\[
\text{familiarity}(u,v) = \frac{|\text{out}(u) \cap \text{out}(v)|}{|\text{out}(u) \cup \text{out}(v)|}
\]

\[
\text{similarity}(u,v) = \frac{|I(u) \cap I(v)|}{|I(u) \cup I(v)|}
\]

Formula 2. Calculation method of user’s interest similarity and familiarity

In the personalized recommendation of Toutiao, we can analyze the interest and similarity between users and friends to recommend the items that friends like to target users by weighted sorting. In "Toutiao", the user's comprehensive similarity calculation formula is shown in Formula 3:

\[
P_{ui} = \sum_{v \in \text{out}(u)} w_{uv} r_{vi}
\]

Formula 3. User’s comprehensive similarity calculation method

In this formula, \( w \) consists of two parts, one is the familiarity of user \( u \) and user \( v \), and the other part is the similarity of interest between user \( u \) and user \( v \).

6. The shortcomings and prospects of "Toutiao"

Toutiao as a platform should be responsible for the quality of the content products received by users. Therefore, the platform should purify the environment and improve the content quality of the platform. Signed a contract with a well-known original author to strengthen cooperation with relevant media organizations to resolve copyright issues. At the same time, some manual editing can be added to the content supervision to control the quality of the content. As a supplement to the algorithm recommendation technology, the editor can provide users with content that they still need to understand beyond their interests, such as news events with significant influence, real valuable news, and so on.

Most of today's algorithms are still in the initial stage. As a platform that is known for its algorithms, we should focus on perfecting the algorithm, and strive to achieve digital and intelligent, so that the algorithm can make the image and the whole picture of people's information needs as complete and accurate as possible, and better understand the user. The improved algorithm should not only push the user's favorite content, but also infiltrate some mainstream social value content to help users break the information squatting, so that users have a more comprehensive understanding of the society, so as not to be out of touch with the society.
References
[1] Zhao Chenxi, Liu Wei, Du Haihong. Research on video recommendation mode of short-jumping video platform under the view of algorithm [J/OL]. Publishing wide angle, 2019(18): 76-78[2019-10-28]. Https://doi.org/10.16491/j.cnki.cn45-1216/g2. 2019. 18. 023.
[2] Zhang Kaihui, Zhou Zhiping, Zhao Weidong. Collaborative filtering recommendation algorithm combined with CFDP and time factor [J/OL]. Computer Engineering and Applications: 1-6[2019-10-28]. http://kns.cnki.net/kcms/detail/11. 2127. TP. 20191018. 1142. 008. html.
[3] Cao Yin, Li Wei. Collaborative Filtering Recommendation Algorithm Based on Trust and Graph Clustering [j]. Information Technology, 2019(10): 77-82.
[4] Wen Kai, Zhu Chuanliang, He Shaoyuan. Research on recommendation algorithm combined with user community and scoring matrix joint community [j]. Small microcomputer system, 2019, 40(10): 2119-2124.
[5] Yang Peng, Shao Wei, Huo Xing, Zhang Yangyang, Jing Yongjun. Social recommendation algorithm based on user implicit preference [j]. Small microcomputer system, 2019, 40(10): 2039-2045.
[6] Zhou Yi. On the defects of personalized recommendation algorithm in news push [j]. Science and Technology Economic Guide, 2019, 27 (28): 152.
[7] Liu Chunling, Zhang Li. Recommended algorithm for improving asymmetric similarity and associated regularization [J/OL]. Computer Engineering and Applications: 1-6[2019-10-28]. http://kns.cnki.net/kcms/detail/11. 2127. TP. 20190925. 1357. 018. html.
[8] Li Xiaofeng, Li Dong. High-dimensional sparse data combination recommendation algorithm based on deep learning [J/OL]. Computer Technology and Development, 2020(01): 1-6[2019-10-28]. http://Kns.cnki.net/kcms/detail/61. 1450. tp. 20190925. 1525. 078. html.
[9] Cao Tienan, Wang Yingjie. Application of Recommendation Algorithm in Power Equipment Defect Repair Scene [j]. Power Big Data, 2019, 22(09): 8-13.
[10] Li Xiaoyu. Research on collaborative filtering recommendation algorithm [j]. Computer and Digital Engineering, 2019, 47 (09): 2118-2122+2136.