Personalized information recommendation simulation system based on compound recommendation algorithm——A research tool to study the push effect of algorithm

Yuqiao Yan, Weiping Zhang, Fangdan Luo, Hongli Zhang
Beijing Normal University, Beijing, China

Abstract. The benefits and effects of algorithmic push are affecting all walks of life. It has also aroused the interest of media research and the discussion on the technical and moral issues involved in algorithmic push. In recent years, the discussion about "information cocoons" has become more and more intense. How to make effective quantitative calculation for algorithm recommendation is the key to empirical research on this problem. Therefore, this study designed a tool to effectively operate this variable around this problem. Based on the analysis of the main recommendation mechanism of algorithmic information distribution, this paper points out a method of how to measure the accuracy of algorithm recommendation in practical research, and designs a simulation personalized information recommendation system based on composite recommendation algorithm combining two mainstream recommendation mechanisms, collaborative filtering and text analysis.

1. Introduction
With the advent of the intellectual media era, algorithmic information distribution technology has penetrated into various media. The benefits and effects of algorithmic push are affecting all walks of life, and have also attracted interest in media research. The deep integration of algorithms and the media industry continues to shape the practice of the media industry, bringing new mechanisms and rules to the process of information collection, production, distribution, etc., and also has a considerable impact on communication ethics. Based on the analysis of the main recommendation mechanism of algorithmic information distribution, this paper points out a method for measuring the accuracy of algorithm recommendation in actual research, and combines two mainstream recommendation mechanisms of collaborative filtering and text analysis. A simulated personalized information recommendation system for a composite recommendation algorithm. The software can quantify the algorithm recommendation level by controlling the accuracy of the recommendation information, and can observe the user's behavior data, cognitive data and user attitude changes as independent variables. At the same time, the system can interface with most psychology and user experience scales, enabling researchers to measure the impact of user-effects caused by the accuracy of algorithm recommendations in the laboratory environment, thus solving the "technical mortuary" that has been debated for many years on technical ethics. This proposition provides a useful research tool.

2. Summary of recommendation mechanism for algorithm information distribution
The task of the recommendation system is to contact users and information. On the one hand, it helps users to find information that is valuable to them, on the other hand, it allows information to be displayed in front of users who are interested in it, thereby achieving a win-win situation between information consumption and information producers. Currently, the recommendation system mainly has three
recommendation modes: recommendation based on collaborative filtering, recommendation based on text content, and recommendation based on association rules.

For collaborative filtering recommendations, memory-based collaborative filtering can be divided into user-based collaborative filtering and project-based collaborative filtering according to different objects. The basic idea is to calculate the similarity between users, sort according to the similarity level, set the similarity threshold or set the nearest neighbor user threshold, select several users, let the products evaluated by these users form a candidate set, and These items are weighted to calculate scores and sorted, recommending the highest rated items to the user. The calculation of similarity is the core part of the collaborative filtering algorithm. The mainstream similarity calculation methods have cosine similarity and Pearson correlation coefficient method. Compared with the calculation method of distance similarity, the cosine similarity pays more attention to the difference of the directions of two vectors, rather than the distance or length, which is calculated as shown in Formula 1.

\[
sim(u,v) = \frac{\vec{R}_u \times \vec{R}_v}{\| \vec{R}_u \|_2 \times \| \vec{R}_v \|_2} = \frac{\sum_{i \in I_u} R_{ui} \times R_{vi}}{\sqrt{\sum_{i \in I_u} R_{ui}^2} \times \sqrt{\sum_{i \in I_v} R_{vi}^2}}
\]

Formula 1. Cosine similarity calculation method

However, memory-based collaborative filtering recommendations rely on the entire user history database within the system as a raw material for its recommendation system. When the data is seriously scarce, there will be problems such as poor cold start and reduced recommendation accuracy.

Content-based recommendation is to extract and filter text information features based on user history items, generate models, and recommend information similar to historical item content to users. One of its advantages is to solve the problem of sparse data and cold boot in collaborative filtering. However, if the information is only recommended based on user historical data for a long time, it will cause excessive personalization and information cocoons. In addition, the algorithm is more adept at the extraction and analysis of text information features, and has defects in the analysis ability of unstructured data such as audio and video. Therefore, it is mostly used for the recommendation of text information such as web pages and text news; The recommendation is to mine the relevant associations behind user data based on user historical data to analyze the potential needs of users and recommend information that may be of interest to users. In addition, there are social network-based recommendations that rely mainly on some indicators in social relationships to quantify the similarity or trust relationship between users and users. The neighboring users of the target users are judged by the similarity, and the items of interest of the neighboring users are formed into a recommended candidate set. Then, the score of the item is weighted according to the similarity between the users, and the item of interest is recommended to the target user. In the recommendation of social networks, the method of calculating the similarity between users is to understand the familiarity and hobbies between friends.

In summary, recommendations based on collaborative filtering, content-based recommendations, and recommendations based on association rules have considerable limitations, and each has its own problems that are difficult to resolve.

3. information cocoons: Ethical issues and communication problems of algorithms

As early as 1995, Bill Gates predicted that in the future world, tailor-made information will naturally increase, and everyone can arrange a "daily newspaper" that is completely in line with their own interests. In the book "Digital Survival" published the following year, scholar Negroponte also predicted that The Daily Me is about to appear. Sunstein strongly agrees with "My Daily" and further proposes the concept of "information cocoons" on this basis. He pointed out that in the Internet era, with the development of network technology and the rapid increase of information volume, everyone can choose the topic of interest at will, and can create a personal daily report according to their own preferences, but this information selection behavior will lead to The formation of "information cocoons". This "information cocoons" refers to the consequences of the closure of information caused by the personalization of the
When an individual only pays attention to self-selected or pleasing content, and reduces contact with other information, over time, it will gradually become a self-woven "mortuary" like a silkworm. According to Sunstein, this will lead to narrow vision and closed or even polarized thoughts, which will strengthen prejudice and create irrational extremism. However, the concept of information cocoons has gradually been questioned in recent years. When Sunstein proposed the concept of "information cocoons", the law was still a vague concept. When the algorithm technology gradually matured and became the common rule of content distribution, people were surprised to find that the "information cocoons" seems to be more apt to describe the impact of the algorithm. In the traditional media era, media organizations produce news for the masses and rely on manual editing to distribute them. There is no clear sense of audience segmentation. Based on the algorithm-recommended content distribution, the user's personal characteristics are used as the standard for information screening, and the user is recommended to personalized information that is highly matched with his interests and values, thereby forming a "thousands of thousands of face" content consumption patterns.

Based on the above problems, in recent years, relevant experts and scholars have become more and more fierce about the discussion of "information cocoons", and even have reached the point where they cannot be opened. The advocates of traditional media ethics and media ethics believe that the narrowing of user acceptance information caused by algorithmic push is the chief culprit of group polarization in recent years, and the increasing public opinion extremes and public opinion reversal are their evidence. However, the technical optimist with Yu Guoming as the main representative insists that the phenomenon of partial eclipse of information comes from the beginning of the phenomenon of communication, not the problem brought by the algorithm push. Therefore, the necessity of algorithm push is adhered to by the theory of technological progress.

At present, from the empirical level, the views of these two factions lack the support of practical arguments. First of all, the technical optimist of the algorithm harmless does not deny that the phenomenon of partial eclipse that humans insist on has been amplified by algorithm push in recent years, and to what extent is it harmful, and whether it is not harmful to society. The extent of democracy still lacks strong evidence. Similarly, the evidence that the defenders of media morality insist on can only be a correlation hypothesis of "causal cause". Obviously, there are many reasons for the phenomenon of group polarization, network Balkanization and reversal of public opinion in recent years. It is only because of the popularity of algorithm push in recent years that the main reason is that the algorithm is also suspected of being speculative. To take a step back, although some scholars have already explained or implied that the individual's visual narrowing is related to algorithmic push through questionnaires, the correlation does not represent causality, and it is impossible to explain what kind of changes the algorithm pushes to the bottom.

Therefore, if it is necessary to solve the current debate on "information cocoons" from the research side, it is necessary to study the effect of the "advancement of algorithm recommendation" on the user. For the proof of causality, the most effective method is solved by experiments with laboratory control variables. Judging from the current research on the correlation research and the lack of causality, this problem still lacks a practical and effective research tool. Since the Second World War, the effects research has been the subject of mainstream social science research such as psychology and communication. Therefore, there is no shortage of the dependent variable research tool for this problem. However, people can't quantify the "algorithm recommendation" variable. From this point of view, how to optimize the algorithm to make effective quantitative calculation is the key to study this problem. Therefore, this study has designed a tool that can effectively manipulate this variable around this problem.

4. Recommended level evaluation: the theoretical basis and the design

How to measure the level of information recommendation, evaluation recommendation algorithm is very difficult in nature. Different algorithms perform differently on different data sets. For example, user-based collaborative filtering algorithms perform well on systems where the number of users is much larger than the number of products; conversely, such algorithms are not very practical. Relevant
influencing factors include scoring sparsity, scoring scales, and other characteristics of the data set are evaluated for different purposes. However, most recommendation systems use accuracy to evaluate the quality of the recommendation algorithm. Suppose the user can view the information of all products, and can sort the products according to their own preference for the products, then the accuracy can be defined as the closeness of the predicted ranking of the recommendation algorithm and the actual ranking of the user. In the computer field, the most widely used is the classification accuracy indicator, which includes accuracy and recall. Billsus first introduced the accuracy and recall rate into the recommendation system to evaluate the recommendation system. Assuming that the total number of products in the system is \( N \), the total number of recommended products is \( N_s \), where \( N_s = N_{rs} + N_{is} \), \( N_{rs} \) and \( N_{is} \) are the number of products that the user likes and dislikes in the recommended products, respectively. Accordingly, \( N_{rn} \) and \( N_{in} \) are respectively the number of products that the user does not like and dislike in the unrecommended product.

Therefore, the accuracy calculation should be: \( p = \frac{n_{(the number of products the user likes)}}{n_{(the number of products presented to the user)}} \).

That is to say, for an algorithm push platform, the personalization level, that is, the accuracy rate, can be decomposed into: (accuracy rate = number of pieces of information preferred by the user / total number of pieces of information seen by the user)

Therefore, the independent and dependent variable decomposition of this study is:

Independent variable: the accuracy of personalized push
Dependent variables: likes, number of forwards, forwarding (quantity, relationship, etc.)

5. The design logic and user interface of the system

Therefore, the program is developed using the Java environment, and a user text library with adjustable accuracy is set. In general, the programming idea is: input the keyword language of the user's interest, perform text matching and collaborative filtering to generate the preferred keyword, search on the microblog according to the keyword and crawl the text or picture information and make the information stream. The end user browses the information stream text and uses the UI virtual button to count the user's behavior.

The overall logic is:
1. with experimental UI interface, easy to operate
2. there are virtual likes, forward buttons, and statistical user behavior stored in the database
3. Match and reproduce according to the user's favorite content, search and crawl the news content on the Internet, and use it to build the information base (text matching)
4. Use collaborative filtering to predict user interest tags, search and crawl news content online, and build information base (collaborative filtering)

The overall programming is divided into three steps. The first step is a program that stores user basic information and search information to store user data and habits. The second step is a preference mining based on user data and habits of the first step. And in the whole network according to each user's preferences for data capture and storage, and ultimately each number of users to generate a (preference information / see the total information) ratio of the user database. The third step is to store each database to each participant to see the storage statistic that can be stored in the different numbering library.

The above three steps will be described in detail below.

The first step is to use python as the development language. In the Google Crome browser, Weibo (headline, Baidu?) search nested a storage mechanism for user header information to achieve a "cold start" of user information. "The specific steps and part of the key code logic are shown in Figure 1. The user interface is in the box, and the red word is working in the background. This step requires networking.
The second step is a key step in the overall programming, that is, the mining based on the user's personal preference data and the generation of the main experimental material. This step requires network operation. The main process is: full mining and personalized feature collection based on interest tags, search history, links, etc., and then crawl or download relevant information on the entire network text. (If it is difficult to achieve, you can download the relevant information directly according to the interest tag.) After that, randomly crawl the irrelevant information on the Internet, and then mix the two proportionally to generate operable experiments according to personal interests at different accuracy rates. The database is shown in Figure 2.
The third step is the overall experimental interface design. The users who measured in the first step went to the lab again, this time they asked them to view the database that has been compiled according to their previous user habits, and then not counted, only the statistics of the participants in different proportions of the database. And propagation bias, specific logic and database storage information as shown in Figure 3.

Finally, the displayed interface can view videos and pictures, and can perform keyword association in real time. Finally, it will be integrated into an online test system and presented to the participants, as shown in Figure 4.
Combined with the dependent variable research tools such as the Emotional Tendency Scale and the Critical Thinking Scale, the system has been verified to perform smooth and accurate recording of behavioral data.

Appendix: Text Online Simulation Recommendation System
URL: http://47.104.167.57/wrcmd/index2.jsp

References
[1] Yu Guoming, Qu Hui. The Misunderstanding of "information cocoons" and the Necessity of Algorithm Pushing--On the Solution of Social Ethical Dilemma in Content Distribution [J/OL]. Journal of Xinjiang Normal University (Philosophy and Social Sciences) Edition, 2020 (01): 1-7 [2019-10-30].https://doi.org/10.14100/j.cnki.65-1039/g4.20190826.001.
[2] Liu Chunling, Zhang Li. Recommended algorithm for improving asymmetric similarity and associated regularization [J/OL]. Computer Engineering and Applications: 1-6[2019-10-30].http://kns.cnki.net/kcms/detail/11.2127.TP.20190925.1357.018.html.
[3] Li Xiaolong, Li Dong. High-dimensional sparse data combination recommendation algorithm based on deep learning [J/OL]. Computer Technology and Development, 2020(01): 1-6[2019-10-30].http://Kns.cnki.net/kcms/detail/61.1450.tp.20190925.1525.078.html.
[4] Shi Fan, Chen Jie. The cause, harm and standardization strategy of algorithm recommendation mode [jj]. Science and Technology Communication, 2019, 11 (17): 126-127.
[5] Wang Shiyong. Technical innovation and ethical dilemma of algorithm recommendation news: a review [jj]. Chongqing Social Sciences, 2019 (09): 123-132.
[6] Fan Qingli. Rethinking about the easy formation of "information squatting" by algorithm recommendation [jj]. Editing and writing, 2019 (04): 33-35.
[7] Guo Qiuhua. The future of wisdom: the ability of journalism communication talents in the intelligence media era [jj]. China Editor, 2019 (09): 45-49.
[8] Li Xiaoyu. Research on collaborative filtering recommendation algorithm [jj]. Computer and Digital Engineering, 2019, 47 (09): 2118-2122+2136.
[9] Wang Hui. Talking about the formation of "information squatting" from personalized news push [jj]. Reporter's cradle, 2019 (08): 119-120.
[10] Ji Nan, Li Ping. Protection of User Privacy in the Age of Algorithms [jj]. Young Reporter, 2019 (26): 78-79.
[11] Du Juan. Going to human-computer collaboration: news ethics in the era of algorithmic news [jj]. News enthusiasts, 2019 (09): 21-25.
[12] Mo Yazhen. Algorithms recommend conflict and balance between news and personal information protection [jj]. Editorial Journal, 2019 (05): 40-45.
[13] Microscopy; Recent Findings from New York University Has Provided New Information about Microscopy [Mechanics of cocoon secretion in a segmented worm (Annelida: Hirudinidae)][J]. Science Letter,2016.
[14] Fang Lingling. Beware of the "information cocoons" effect of community learning in the Internet age [jj]. China Education Journal, 2019 (10): 79-83.
[15] Chen Wei. Opportunities and Challenges of News Publicity Communication under Artificial Intelligence Technology [jj]. Media Forum, 2019, 2(16): 120-121.