# News Recommendation Systems in the Era of Information Overload

Shuaishuai Feng[1], Junyan Meng[1,*] and Jiaxing Zhang[2,3]

[1]*School of Sociology, Wuhan University, Wuhan, Hubei, China*
[2]*The Institute of Social Development Studies, Wuhan University, China*
[3]*Shenzhen Qianhai Siwei Innovation Technology Ltd. Co., Shenzhen, China*
*E-mail: JYMeng@whu.edu.cn*
[*]*Corresponding Author*

Received 30 November 2020; Accepted 07 December 2020;
Publication 09 March 2021

## Abstract

The internet has reconstructed information boundaries in the modern world, and along with mobile internet has become the most important source of information for the public. Simultaneously, the internet has brought humanity into an era of information overload. In response to this information overload, recommendation systems backed by big data and smart algorithms have become highly popular on information platforms on the internet. There have already been many studies that attempted to improve and upgrade recommendation algorithms from a technical perspective, but the field lacks a comprehensive reflection on news recommendation systems. In our study, we summarize the principles and characteristics of current news recommendation systems and discuss "unexpected consequences" that might arise from these algorithms. In particular, technical bottlenecks include cold starts and data sparsity, and moral bottlenecks are presented in the form of information imbalance and manipulation. These problems may cause new recommendation systems to become a "warped mirror".

**Keywords:** Information overload, internet news, recommendation systems, User-CF, Item-CF, reflection on technology.

*Journal of Web Engineering, Vol. 20_2,* 459–470.
doi: 10.13052/jwe1540-9589.20210

## 1 Introduction

As a new and powerful carrier of information, the internet opened up the age of new media. Today, new media has deeply integrated itself to people's lives. All kinds of news and advertisements automatically pop up in windows as one turns on a computer, and all sorts for recommendations for various merchandise immediately appears as one logs in to an online shopping platform. More surprisingly, you might suddenly discover that the news and shopping information that pops up is not "irritating" but about things you actually care about. There seems to be an "invisible person" behind the computer screen that continues to observe each user, pondering what kind of news they like to read, and what things they want to buy. In reality, this "invisible person" exists – this is what a recommendation system is. Computer scientists first invented this system to solve the problem of information overload in the age of big data. Since the amount of data provided by the internet far exceeds the amounts of data a user can accept and process, the rate at which a user can effectively process data has decreased dramatically [1]. In the field of recommendation algorithms, news recommendation systems is a hot topic [2, 3]. Before the internet, popular methods of recommendation utilized newspapers, radio, and television. During the starting stage of the development of the internet, the popular way to recommend something was to put the information on a website with many users like Yahoo. Nowadays, personalized recommendation algorithms and information visualization is mainstream and is used heavily by big news companies all over the word, such as News Republic, Flipboard, Google News, Zaker, and TOPBuzz, a newly-emerged popular news company in China.

## 2 Technical Principles of News Recommendation Systems

A recommendation system is developed by experts on the basis of computing technology and statistics, combining data, algorithms, and computers to create mechanisms that relate users with personalized resources that can reflect the user's consumer behaviors and information intake. To date, recommendation systems have developed a relatively complete methodology. The most common algorithm categories are collaborative filtering recommendations, content-based recommendations, association rule-based recommendations, utility-based recommendations, and knowledge-based recommendations.

First, we talk about collaborative filtering recommendations. These algorithms can be categorized into User-Based Collaborative Filtering and Item-Based Collaborative Filtering. User-Based CF finds other users that has a high similarity with the target user, i.e., similar users, and uses the preferences of those similar users to predict the preferences of the target user. Item-Based CF uses the ratings that users give an item or a message, analyzes the similarity between different items and messages, and recommends the items that has a high similarity with the items that the target user likes. The basics of these algorithms include association algorithms, categorization algorithms, clustering algorithms, regression algorithms, matrix decomposition, graph models, word connotation models, and neural networks. Next, there are content-based recommendations. These algorithms do not need to utilize the item ratings of users, but instead use the browsing history of a user to predict what the user might have never seen but might like. Third, there are association-based recommendations. These algorithms smartly utilize the ability to find association in big data. By using the association rule and digging through datasets, the relationship between different items during their sales and usage can be discovered and be used to predict the users' needs. Fourth, there are utility-based recommendations. The core of this algorithm is to create an utility function for each user that has the item characteristics and customer satisfaction as input. After calculating the different values of utility given the different inputs, the item with the highest utility is recommended to the user. Fifth, there is knowledge-based recommendations. This algorithm uses the knowledge structure in the user data that can support inference (such as regularized user search history), creating a knowledge base of how an item can satisfy a particular user, and using that knowledge to make a prediction [4]. In these common categories of recommendation algorithms, collaborative filtering is the main algorithm basis of news recommendation system. We talk about the principles and processes of collaborative filtering below.

## 2.1 Item-Based Collaborative Filtering

Item based collaborative filtering is one of the main algorithms used in news recommendation systems. Its fundamental principle is finding pieces of news similar to the news that the user prefers based on the user's history. From a computational angle, it takes all the preferences a user has towards a piece of news and converts that into a vector to determine the similarity between the two pieces of news. Once the similarity has been calculated, based on past preferences, the pieces of news which the user has not yet expressed preferences for are predicted, finally getting a sorted list of news recommended to the user. For instance, for a piece of military news A, based on the preferences of all the users in the past, users who like news A like political news C. Therefore, military news A and political news C are similar, so if user U like military news A, it can be predicted that user U might also like political news C.

The implementation process is as follows:

(1) Calculate the similarity between two pieces of news. The formula is as follows:

$$w_{ij} = \frac{|N(i) \cap N(j)|}{\sqrt{|N(i)||N(j)|}} \tag{1}$$

In formula (1), $N(i)$ is the collection of users who like reading news $i$, and $N(j)$ is the collection of users who like reading news $j$. The calculated $w_{ij}$ is the number of people who like news $i$ who also like news $j$, and is the similarity between news $i$ and news $j$.

(2) Constructing the users – reversed news list

We can express all the historical data of news in a matrix $D_{n*m}$.

$$D_{n*m} = \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} \tag{2}$$

In formula (2), $n$ is the total number of users that has seen the piece of news; $m$ is number of news the $n$ users have ever seen, $d_{ij}$ represents whether user $i$ has seen user $j$, 1 means they have read it and 0 means the opposite. We construct a reverse user list based on $D_{i*j}$, as shown in Table 1.

In particular, 1, 2, and $3 \ldots, m$, and $n$ are the pieces of news seen by the users. $x_{ij}$ is the number of users who have read news $i$ and news $j$.

(3) Calculate how interested a user is in a piece of news

We use the similarity formula given above to recommend to a user the type of news they like or similar to what they like. The interest a user $u$ has

**Table 1** The inverted table to access of news-users

| | 1 | 2 | … | $i$ | … | $m$ |
|---|---|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{12}$ | … | $x_{1i}$ | … | $x_{1m}$ |
| … | … | … | … | … | … | … |
| $j$ | $x_{j1}$ | $x_{j2}$ | … | $x_{ij}$ | … | $x_{jm}$ |
| … | … | … | … | … | … | … |
| $m$ | $x_{m1}$ | $x_{m2}$ | … | $x_{mi}$ | … | $x_{mm}$ |

in a piece of news $j$ is calculated as follows:

$$p_{uj} = \sum_{i \in N(u) \cap S(j,k)} w_{ji} r_{ui} \tag{3}$$

In formula (3), $S(j,k)$ are the $k$ most similar pieces of news to news $j$. $N(u)$ is the collection of news that user $u$ likes. Since we need to predict how much interest user $u$ has in news $j$, we should naturally pick from the top $k$ pieces of news the ones that are related with $j$. We take the intersection. $w_{ji}$ is the similarity between news $j$ and news $i$. $r_{ui}$ is the interest user $u$ has to news $i$.

(4) Top-N analysis

We reverse sort the $p_{uj}$ values calculated, and choose the first $N$ items to recommend to user $u$.

## 2.2 User-Based Collaborative Filtering

In user-based collaborative filtering, we can recommend based on the similarities of how two different users rate the same piece of news. In simpler terms, we group people into categories and recommend to users news that other users in that category like. For instance, if a user U1 reads news of type A and type C, and user U2 also reads those types of news, then if U2 also reads news of type B, there is a high probability that U1 also likes reading news of type B.

The implementation process is as follows:

(1) Calculate the similarity between two users. Similar to formula (1), the formula to calculate user similarity is as follows:

$$w_{uv} = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)||N(v)|}} \tag{4}$$

In formula (4), $N(u)$ is the collection of news liked by user $u$, and $N(v)$ is the collection of news liked by user $v$. The calculated $w_{uv}$ is the similarity between user $u$ and user $v$.

(2) Constructing the users – reversed news list

This step is the same as item-based CF, and is therefore omitted here.

(3) Calculate how interested a user is in a piece of news

The interest a user $u$ has in a piece of news $i$ is calculated as follows:

$$p_{(u,i)} = \sum_{v \in S(u,k) \cap N(i)} w_{uv} r_{vi} \quad (5)$$

In formula (5), $S(u,k)$ are the $k$ most similar users to user $u$. $N(i)$ is the collection of users that has interacted with news $i$. Since we need to predict how much interest user $u$ has in news $i$, we should naturally pick from the top $k$ users the ones that is related with $i$. We take the intersection. $w_{uv}$ is the similarity between user $u$ and user $v$. $r_{vi}$ is the interest user $v$ has to news $i$.

(4) Top-N analysis

This step is the same as item-based CF, and is therefore omitted here.

### 2.3 A Comparison Between Item-based CF and User-based CF

From a concrete application perspective, the two recommendation algorithms each have their own advantages. In particular, item-based CF has the advantages of better reflecting the interests of an individual user and the recommendations can be updated real-time, but it requires more computational power because it is more requires more computation. In contrast, user-based CF is great at reflecting what is popular in a given time frame, and since it does not need to calculate how similar two pieces of news are, it requires less computational power. The pros and cons of each algorithm can be found below in Table 2.

**Table 2** Comparison between Item-based CF & User-based CF

| Character | Item-based CF | User-based CF |
|---|---|---|
| Advantages | – Reflects the interest of an individual user<br>– Real time changes of recommendation results<br>– Higher level of personalization in the recommendations | – Better reflects what is popular in a given time period<br>– Reflects the relations between a user's interests with the group of users<br>– No need to calculate similarity between news |
| Disadvantages | – Higher complexity for item similarity calculations<br>– Higher requirements for system response speed<br>– Hard to construct a model of the object | – Higher complexity for user similarity calculations<br>– New user behavior cannot be reflected in real time<br>– Not enough reasons for a recommendation |

## 3 Problems with News Recommendation Systems

The success of recommendation algorithms in commercial news have excited experts and businesspeople, incentivizing them to continue to evolve and optimize the algorithms in order to achieve more precise recommendations and increase customer retention. Anderson went as far as to declare, “Better data and better tools for analysis can win over the world” [5]. Yet, with the ubiquitous applications of news recommendation systems, a few bottlenecks have become more apparent, and we can discuss them from a technical perspective and a moral perspective.

### 3.1 Technical Bottlenecks

First, there is the problem of data sparsity. As the scale of recommendation systems increases and there are tens of millions of users and news content, the possibility of some pieces of chosen news overlapping between two users becomes very small. If we look at data sparsity using a ratio of the existing relations between a user and some pieces of news to all possible existing relations, we can observe that as the number of news content grows exponentially, data becomes more and more sparse, which will significantly increases the computation complexity of the system. The root of this problem cannot be entirely solved, but there are some ways to reach a middle ground. If it is possible to use a kind of expanding algorithm, changing the original order-one relations (how similar are two users or the news content that read) to relations of order-two or higher (given that relations and similarity itself is expandable) [6], then some default parameters can be used [7] and increase the resolution of similarities. The bigger the scale of data, the sparser it usually is. There are some sparse data algorithms that are believed to be very helpful in the future (e.g. expanding [6], iterative optimization [8], similarity transference [9] etc.)

Second, there is the problem of cold starts. For a new user that has just joined a system, there is no effective behavioral data that can be used as reference and it is therefore hard for the system to give the user a precise recommendation. On the flip side, some news content has only been seen by users a few times and it is therefore difficult to recommend this piece of news to users. One solution is to use the content as an assistive recommendation, and another is to gather some data such as age, city, education level, gender, and occupation [10, 11] upon signup or via a questionnaire. Recently, tagging systems that have been used in many places are also a possible solution [12], because tags can been thought of as the essence of content, and they simultaneously reflect the personal preferences of users. Two users can watch the same news but be interested in different parts of the same piece. Of course, the usage of tags can only improve recommendations given to users with minimal behavioral data but does not help users who are from a completely cold start, because no tags are associated with these users.

### 3.2 Moral Bottlenecks

First, there is the problem of information imbalance. Prasier, the founder of a subscription-based news website Upworthy, stated that media given to users by algorithms can create a bubble-like environment that wraps around the users. He named this phenomenon as filter bubbles [13]. Personalized news algorithms use recommendation systems to create a "personalized daily digest" for the users, giving the users what they want to see. Filtering based on user preference causes internet media to provide to users their own echoes. If the user is unaware of this fact and believes the content given to him by the media is reflective of the real world, then the user becomes even more subjective. This subjectivity will in return affect the algorithm, causing a positive feedback cycle that enhances itself over and over. A "personalized daily digest" will soon become "Information Cocoons". The most dangerous part of being inside a information cocoon is the inability to correct cognitive bias, hence seeing bias or misinformation as the truth.

Second, there is the problem of information manipulation. Recommendation systems are in part driven by commercial and economical incentives. Some users with malicious intent can increase or decrease the probability of a piece of news being recommended [14]. Therefore, an important characteristic of a recommendation algorithm is its ability to to a certain extent remain neutral even under malicious attacks. Take a simple association-rule based algorithm as an example, the Apriori algorithm is much better at remaining neutral that the k-nearest-neighbors algorithm [15]. Some technology have been developed to increase an algorithm's capability to remain neutral even under malicious attacks, such as analyzing the behavioral differences of a real user and a malicious user, judging ahead of time which behaviors are malicious, and stopping it from entering the system or giving it less impact when it enters the system [16–18]. Besides commercial manipulation, research has shown that it is possible for news recommendation systems to serve political interests. As early as 2015, a research that focused on Facebook showed that broadcasting different pieces of news can affect the participation of American voters [19]. Epstein also stated that a biased search engine can change voter intent, manipulating the outcome of politics [20].

## 4 Conclusions

There are five commonly seen recommendation algorithms: Collaborative Filtering Recommendation, Content-Based Recommendation, Association Rule-Based Recommendation, Utility-Based Recommendation and Knowledge-Based Recommendation. In particular, collaborative filtering is the one used most often in news recommendation systems. This paper summarizes the principles and processes of the two collaborative filtering algorithms: Item-based CF and User-based CF. We also reflected on the technical and moral bottlenecks of these news recommendation systems. The prior includes cold start and data sparsity, and the latter materializes in forms such as information imbalance and information manipulation. We believe that in an age of information overload, the government, companies, and the public should work together. Concretely, the government needs to come up with new regulations to guide new media, encouraging companies to create algorithms that put more weight on good content. Companies need to be self-disciplined and take up the associated responsibility of an information carrier and medium. They cannot distribute vulgar content only because it is profitable. Users also need to keep up with the times and increase their information literacy by strengthening their abilities to acquire, read, analyze, and rate all kinds of information.

## Acknowledgements

This paper is Supported by the National Social Science Fund of China (20BSH002), the Independent Research (Humanities and Social Sciences) at Wuhan University (413000057), and the Philosophy and Social Science Fund of Hunan (19YBA201).

## References

[1] Maes P. Agents that reduce work and information overload [J]. Communications of the ACM, 1995, 377:30–40.

[2] Jing Qiu, Lejian Liao, Peng Li. News Recommender System Based on Topic Detection and Tracking [M]. Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2009.

[3] Chen W, Zhang LJ, Chen C, Bu JJ. A Hybrid Phonic Web News Recommender System for Pervasive Access [C]. Wri International Conference on Communications & Mobile Computing. IEEE, 2009, pp. 122–126.

[4] Liu Cundi, Xu Wei. Can algorithms define society? News algorithm recommendation system from the perspective of media sociology [J]. Academic Forum, 2018, 41(04):28–37.
[5] Anderson C. The end of theory [J]. Wired Magazine, 2008, 16(7):16–17.
[6] Huang Z, Chen H, Zeng D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering [J]. ACM Transactions on Information Systems, 2004, (22):116–142.
[7] Breese JS, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering [J]. Uncertainty in Artificial Intelligence, 2013, 98(7):43–52.
[8] Ren J, Zhou T, Zhang YC. Information Filtering via Self-Consistent Refinement [J]. Epl, 2008, 82(5):1–4.
[9] Sun D, Zhou T, Liu JG, et al. Information filtering based on transferring similarity [J]. Physical Review E, 2009, 80(1):17101–17101.
[10] Schein AI, Popescul A, Ungar LH, Pennock DM. Methods and metrics for cold-start recommendations. in: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, 2002, pp. 253–260.
[11] X. N. Lam, T. Vu, T. D. Le, A. D. Duong, Addressing cold-start problem in recommendation systems [C]. in: Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, 2008, pp. 208–211.
[12] Zhang ZK, Liu C, Zhang YC, et al. Solving the cold-start problem in recommender systems with social tags [J]. EPL (Europhysics Letters), 2010, 92(2):28002p1–p6.
[13] Chen Chang. Today's headlines and Zhang Yiming's eyes on future media: it's time to give the power of filtering information to social relationships and algorithms [EB/OL]. http://www.cyzone.cn/a/20160115/288570.html,2016-01-15.
[14] Mobasher B, Burke RD, Bhaumik R, et al. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness [J]. ACM Transactions on Internet Technology (TOIT), 2007, 7(2):1–41.
[15] Sandvig JJ, Mobasher B, Burke R. Robustness of collaborative recommendation based on association rule mining [C]. in: Proceedings of the 2007 ACM Conference on Recommender Systems, ACM Press, 2007, pp. 105–112.
[16] Shyong K, Frankowski D, Riedl J. Do You Trust Your Recommendations? An Exploration of Security and Privacy Issues in Recommender Systems [C]. Springer, Heidelberg, Germany, 2006, pp. 14–29.

[17] Resnick P, Sami R. The influence limiter: provably manipulation-resistant recommender systems [C]. in: Proceedings of the 2007 ACM Conference on Recommender Systems, ACM Press, 2007, pp. 25–32.
[18] Shi C, Kaminsky M, Gibbons PB, Xiao F. DSybil: Optimal Sybil-Resistance for Recommendation Systems [C], IEEE Press, 2009, pp. 283–298.
[19] Tufekci, Z. Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency [J]. Colorado Technology Law Journal, 2015(2):396–412.
[20] Epstein R, Robertson RE. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections [J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(33):4512–21.

## Biographies

**Shuaishuai Feng** is a PhD candidate in sociology at Wuhan University. Now he is a researcher member of the Institute of Social Development Studies, Wuhan University, China. Shuaishuai Feng received his bachelor's degree and master's degree in sociology from Northwest A&F University and Wuhan University respectively. His current focus is on computational social science research.

**Junyan Meng** is a doctoral candidate in the School of Social Sciences, Wuhan University, graduated with a bachelor's and master's degree in journalism and communication from the School of Public Administration, Hohai University, and his main research direction is communication sociology.

**Jiaxing Zhang** attended the Wuhan University where she received her B.Sc. in Software Engineering in 2009. She then went on to pursuit a M.Sc. in software Engineering from Wuhan University, China in 2011. After that, she got a M.Sc. in Digital Media from Wuhan University, China in 2013. Jiaxing Zhang has held solution and software engineering senior positions at Shenzhen since 2014. And she got some awards from some other research institutes in her research areas. Her Ph.D. work centers on Block Chain Technology and Social Governance.