Comparison Between Collaborative Filtering and Content-Based Filtering

Xinyi Wu

Department of Statistical Science, University College London, London, British

zczlx23@ucl.ac.uk

Abstract. With the rapid development of Internet technology nowadays, how to quickly obtain the effective information needed by users has become the key point of the scientific and technological academia. Therefore, various kinds of recommendation algorithms have been invented. Based on the previous research, this paper introduces the most famous and widely used recommendation algorithms among many recommendation systems, which are collaborative filtering and content-based filtering. In this paper, the core ideas and operation principles of the two algorithms are introduced in detail. In addition, by describing the steps of these two algorithms gradually and analyzing their processes step by step, we can accurately analyze and summarize their advantages and disadvantages respectively. And on this basis, the respective areas which they are good at are mentioned. Moreover, this paper points out the shortcomings and limitations that still exist at present, and the direction for further improvement in the future. Finally, at the end of the paper, there are some overall comparison and summation about the two algorithms. And the hot research points of them in the future are discussed.

Keywords: Recommendation algorithms, collaborative filtering, content-based filtering, information filtering.

1. Research Background

With the continuous expansion of the Internet in the world and the large-scale popularization of mobile terminals, people's access to and exchange of information has become more convenient and diverse. A great deal of information is added to the database of the Internet every day. The increase in the amount of information data does enrich people's lives and accelerate the communication and cognition within the communities. However, at the same time, the overload of information also brings a burden to people. Too much information also makes people unable to find the information meeting their needs, which forms a phenomenon of "information overload" [1]. In this case, the utilization rate of information would decrease instead. Therefore, how to eliminate invalid information from a large amount of disordered data and find out effective information for people quickly and effectively has become the primary issue of all major Internet platforms.
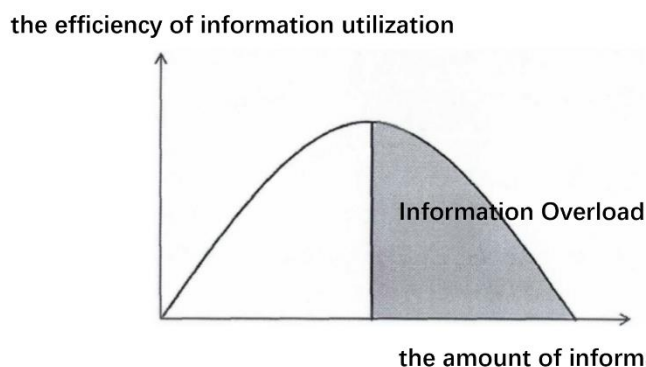

Figure 1. Relationship between users' ability to process information and the amount of information

In order to deal with it, the recommendation system was invented as a tool for filtering information. The recommendation system exists in the backstage of the software platform. It filters out the invalid content through the specified calculation and recommends the information according to the weighted

level. The recommendation system effectively solves the problem of information overload and meets the needs of website customers' personalized services [2].

According to the difference of the recommendation systems' algorithm principles, the recommendation algorithms are generally divided into two major types, namely collaborative filtering (CF) and content-based filtering (CBF). Because among all classical recommendation algorithms, collaborative filtering algorithm and content-based filtering algorithm are the most widely used nowadays [3].
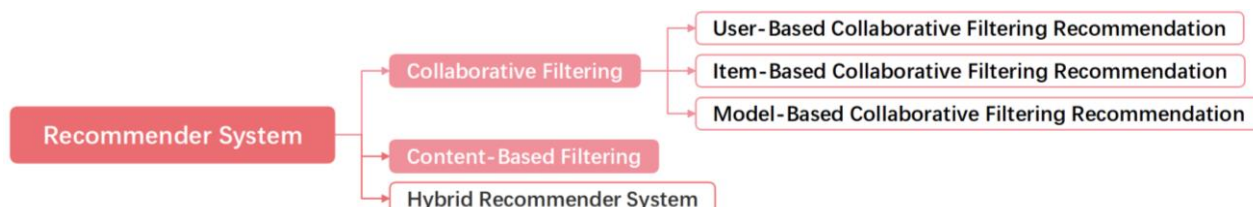
Figure 2. Classification of the recommender system

2. Collaborative Filtering

With the development of the Internet, a large number of recommendation algorithms have been invented and applied to various website platforms. Among them, collaborative filtering (CF) stands out among many kinds of recommendation systems and is widely used by many social platforms because its advanced concept and giving the full play to its unique property advantages of personalized recommendation.

The core idea of collaborative filtering recommendation algorithm is that "customers with the same preferences may have other similar interests". Therefore, software developers carry out collaborative filtering algorithms by comparing the similarities between objects. Collaborative filtering algorithm uses multiple methods to collect various information left by users when they surf the Internet. Following, by comparing a large amount of different user information and analyzing the similarity (i.e., "collaboration") between these information. Through this way, it could be easy to predict platform users' potential preferences and potential consumption products [4].

2.1. Principle of Collaborative Filtering

2.1.1 Process of collaborative filterinng

1) Collect users' preference data and item information data

Collaborative filtering algorithm has the characteristics that it has diversity of channels for collecting various user information. The comprehensive information collection can effectively help this algorithm to improve the information needed by the database. Therefore, it is conducive to the website platform to do further development on personalized recommendations for users. As long as the user are browsing through the website platform, the user's behaviors would be recorded in the back stage's database, such as clicking, saving and leaving a mark. In addition, the collaborative filtering algorithm would also collect some kind of the records that even the users might not notice, such as user's click order, personal Internet protocol address and specific page stay time. Moreover, the collaborative filtering algorithm might take the response of the target user to certain event or item as the screening criteria, such as scoring and voting [5].

It is worth noting that the recorded responses are not limited to the objects that the users are interested in or that are given positive comments. The advantage of collaborative filtering algorithm is, for its recommendation system, the objects that users are not interested in, given negative comments or even ignored are all recorded, which is in order to achieve the goal of filtering information.

2) Uniformly clean and normalize the collected data

As to make the recommended results more accurate, it is an important link for data to do page denoise. When the collaborative filtering algorithm are collecting data, there is a certain probability

for both the non-core content part and the non-content part both to be captured from the page. In addition, the user may make some mis operation during the process of using the application. These are the potential interference items to the recommended algorithm. In order to analyze the accurate data, the background programmers would usually choose to carry some further classical and convenient data mining algorithms to filter out the irrelevant information in the captured data.

Expect for the noise reduction, the program also needs to normalize the available data. Normalization is to map the data of each behavior to a specified range and then process it, so that the overall preference obtained by weighted summation would become more accurate. Normalization can process data more conveniently and quickly, moreover, it also enables objects of different units or magnitudes to be compared and weighted [6]. Normalization not only simplifies the complex calculation, but also furtherly excavates the available data.

3) Quantitatively compare the similarity between user or item

After completing the above steps, the algorithm would gather the collected information into the feature sets of different individuals, and then conduct specific quantitative comparison.

In collaborative filtering algorithm, there are many methods to calculate the similarity. Among these methods, comparing the similarity based on distance is a widely used method. Its formula can be seen below [7]. The distance d can be obtained by Euclidean distance and Manhattan distance.

$$similarty = \frac{1}{1+d} \qquad (1)$$

Moreover, comparing the similarity of set content is also a common method. For example, the Jaccard similarity shows the coincidence degree of the same elements in individual A and B through Venn diagram as figure 3. And its calculation formula is shown below.
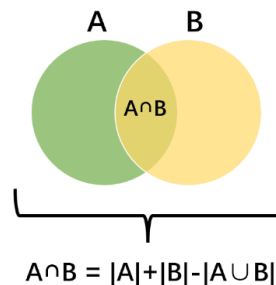

Figure 3. The Jaccard similarity shown by Venn diagram

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \qquad (2)$$

In general, the wisdom of collaborative filtering algorithm is to use the collective intelligence in the cluster. As the program begins to run, the database of user object preferences would be built first, and then the similarity between user characters would be calculated. This is to match users with similar interests and hobbies, which would be helpful for completing the whole recommendation.

2.1.2 Specific classification of collaborative filtering

Because of its wide application field, collaborative filtering algorithms are derived and refined into different categories. Although these different kinds of algorithms might seem complex and diverse, in fact, their basic theories and construction principles are actually highly similar. The difference is that what the similarity needed to be calculated is based on. Therefore, collaborative filtering algorithm can be broadly divided into three subcategories downward, which can be specifically subdivided into:
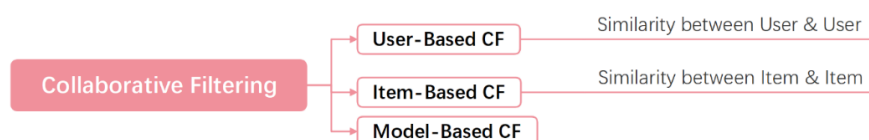
Figure 4. Classification of collaborative filtering

1) User-based collaborative filtering

User-based collaborative filtering algorithm is appropriate for the platforms with relatively less user volume and individual differences, and it is famous for its good diversity and timeliness [8].

In recent years, this algorithm has been very popular among dating websites for minority group with unique hobbies. By comparing neighborhood users' common interests and predilections, it picks out the users with many similarities from the vast crowd to make mutual recommendations between each other. Moreover, when there is something new happens among friends, it can also be updated in time.
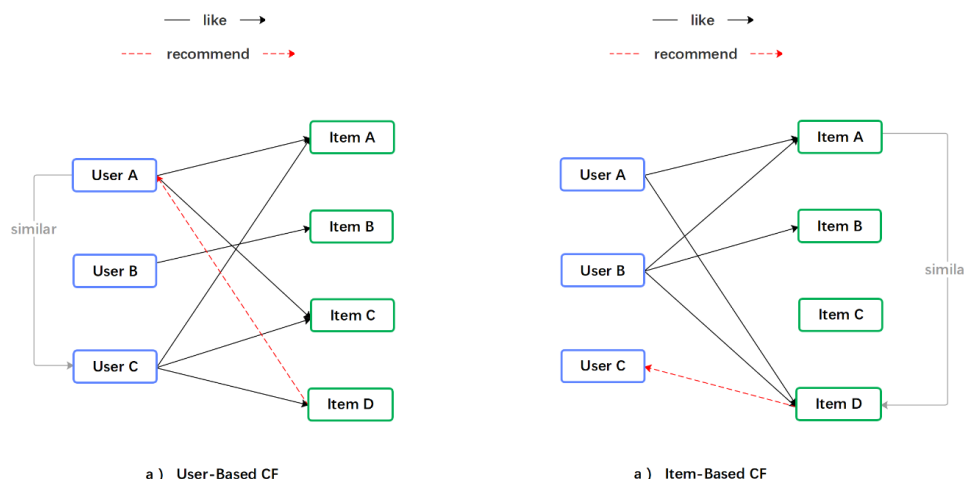
2) Item-based collaborative filtering

Item-based collaborative filtering algorithm is suitable for the situation where the number of items is less than the number of users. In addition, it is also applicable to fields with strong personalized demands, like movie recommendations and book reviews. Moreover, it has also been widely used in online e-commerce shopping platforms that have sprung up in recent years. When a customer wants to find something similar to the target product, the system would make recommendations in the background database.

3) Model-based collaborative filtering

Model-based collaborative filtering is a recommendation system based on users' behavior and preference information. It makes recommendations through similar models.

Compared with user-based collaborative filtering and item-based collaborative filtering, model-based collaborative filtering is relatively less used in the related fields.

Figure 5. Comparison between User-based CF and Item-based CF

2.2. Characteristics and Superiorities of Collaborative Filtering Algorithm

1) Easy to extract feature information

The information of collaborative filtering algorithm is based on the behavior of individual users, such as the purchase and evaluation of a series of items. Therefore, there is no need for the software background programmers to manually extract and summarize the information. Therefore, compared with other recommendation algorithms, it is very convenient to extract item features from objects in collaborative filtering algorithm.

Besides, collaborative filtering algorithm breaks through the defects in the automatic analysis and filtering of previous computer machine. It can analyze and classify things that are vague and difficult to define accurately, such as art and music.

2) Novelty of recommendation results

Collaborative filtering can broaden the field of information received by users, and it is also one of the important reasons for the popularity of this algorithm. Collaborative filtering algorithm is a recommendation method based on the similarity of behavior patterns, which is equivalent to introducing the group behavior of other similar Internet users to users. Hence, the results of

recommendation increases and changes following others' behavior, which could be helpful for users to break through the restrictions of the Internet filter bubbles. Moreover, it also effectively strengthens the communication between individual users and their communities.

3) Intellectualization and personalization

Compared with other similar algorithms, collaborative filtering algorithm has faster personalized learning speed. When using the platform for the first time, users do not need to provide and select their own interests. On the contrary, the platform would automatically judge and supplement users' hobby types according to their online history footprints. Collaborative filtering algorithm can make use of other users' responses to similar information to automatically improve the recommended content. Thereby, it can achieve an excellent degree of automation with its high level of intellectualization and personalization.

2.3. Disadvantages of Collaborative Filtering Algorithm

1) Problem about old start

The cold start problem has always been the most fatal problem for collaborative filtering algorithms. This is because that collaborative filtering algorithms are based on feedback from users [9]. While as a user has just registered for the software to become a member, it is impossible to predict this user's interests because the platform lacks the historical data about his behavior. Therefore, when users are new to the platform, it could always be a common phenomenon that they would be easily pushed a lot of irrelevant content. The same problem often occurs with some new commodities for sale as well. Due to the lack of customer reviews, online shopping platforms are also unable to accurately determine the potential purchasing groups. In simple terms, when there is a new user with little recorded historical data, it is less possible for the platform to provide high-quality recommend results for the user.

2) Problem about data sparsity

In reality, the recommendation systems of some well-known platforms usually contain billions of users and their data tracking records. However, it is like a drop in the ocean for individual users' data recorded in the background system compared with all the user data in the entire system. Therefore, for users, there is still a large amount of unfilled essential detailed data in the database. In other words, the larger size of the database, the more sparsity would appear [10]. And this is detrimental for collaborative filtering algorithms which is recommend relayed on user similarity. Because if users just make a few positive or negative evaluations on a small number of objects, it would be hard for the recommend system to make further inference on user's preference areas only from their reviews. The failure in building similarity relationships leads to the failure of the recommended algorithm.

It is important to note that the data sparsity problem does not mean the recorded information accumulated in the database is all invalid, instead there is no effective means have been found to dig out the valid information. So far, although scholars and researchers from various professional fields have been constantly attempting to improve and update the solving methods, the issue about data sparsity is still the most important problem faced by collaborative filtering algorithm.

3) Problem about maintainability and extensibility

As the number of users increases and the fields of interest expand, the number of items in the database increases as well. Thereby, the scale of information in the database also continues to expand, which would lead to the amount of computing required increase exponentially. Finally, the pressure on the platform's background data processor would keep continuing to aggrandize. As a result, when the information recorded in the database is accumulated to a certain scale, the large amount of data comparison would inevitably slow down the platform's operation speed.

3. Content-based Filtering

"Information retrieval technology" is the most critical part of content-based filtering (CBF). The basic idea of this algorithm is to quantify the similarity of objects' text information, and then build a

specific model on this basis. Further on, it would dig deep into users' related data, and then accurately predict users' possible behavior, so as to accurately recommend the information that users might be interested in to them.

The principle of content-based filtering algorithm is almost the same as that of collaborative filtering algorithm. The difference is that collaborative filtering algorithms usually improve its recommendation system through users' responses to the projects provided by the platform, while content-based filtering algorithm requires the system to accurately extract items' precise information and users' operation behaviors. This is to construct users' preference portraits of items, so as to find the target objects which are matching their traits.

3.1. Process of Content-based Filtering

1) Extract information from people or events

Content-based filtering algorithm is to compare the similarity degree by extracting the text information of things. It can extract the text information of users or items in the background of the software platform, such as the basic personal information of the user and the marketing description text of the product. These are the text feature attributes required by the content-based filtering algorithm. The algorithm would organize them into feature sets based on the individuals [11].

It should be noted that the content-based filtering algorithm is not only to summarize its users, but also to record users' operation behaviors on the target objects. Such as clicking, commenting, buying, collecting, playing, etc. Collecting the behavior data of these users would help content-based filtering to concretely improve the feature sets of users.

2) Transform feature sets into feature vectors

After completing the steps above, the content-based filtering algorithm would uniformly digitize the collected feature sets, that is, convert them into feature vectors. In the content-based filtering algorithm, it is the most common choice to select the vector space model as the carrier form of the feature vectors. This algorithm would express all the essential information in the form of three-dimensional space vectors for further comparison [12].

3) Import into the same vector space and then make comparison

After the feature vectors are converted, content-based filtering algorithm would uniformly import those vectors which are needed to be compared into the same vector space. Here, cosine similarity is always used to quantify and calculate the similarity between vectors.

Cosine similarity is the most commonly used measuring method in content-based filtering algorithm. In addition, sine similarity and weighted cosine similarity are often used as auxiliary tools as well.
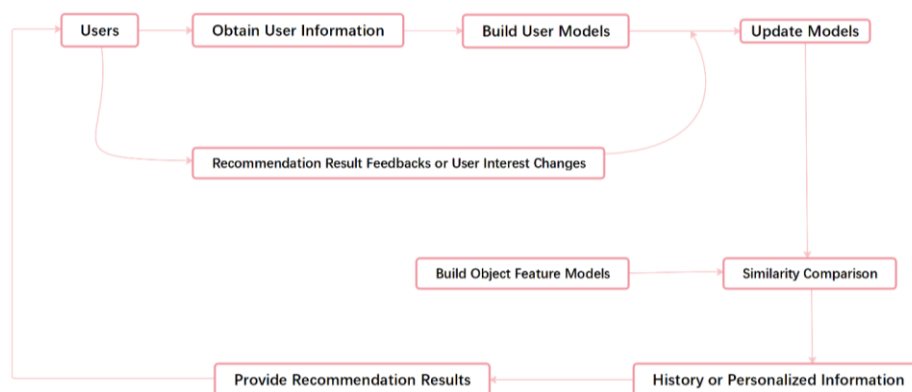
Figure 6. Basic flow diagram of content-based filtering recommendation algorithm

3.2. Advantages of Content-based Filtering

1) Not depend on other users' data, and have less impact on the problem of cold start and data sparsity

Content-based filtering algorithm is based on user behavior information to find possible items of interest. Thereby, even if there is only one user operation behavior on the platform, content-based filtering algorithm can also recommend relevant content for users based on the content of the operation. Because of this, the problem of start problem is avoided availably and the problem of data sparsity is solved easier [13].

2) Suitable for minority groups

For the less popular things, the platform can only collect a relatively small number of user behavior responses. In this case, it might be difficult to expand collaborative filtering algorithm, but relatively, it has a much smaller impact on the content-based filtering algorithm. Therefore, the content-based filtering algorithm is very suited for minority groups who might have a specific single interest but a high degree of consistency.

3) Explainable for the recommended results

Since the content-based filtering algorithm is completely based on user's existing behavior history to carry out the push. Therefore, the objects recommended to the user by the algorithm are supposed to be something the user already knows about, rather than something completely new that may cause the user's aversion. Compared with the collaborative filtering algorithm, the platform has sufficient reasons to explain why the product is recommended to users, and users also have a higher degree of acceptance and recognition for the recommended items as well.

3.3. Disadvantages of Content-based Filtering

1) Difficult in extracting feature information

In order to simplify the process of comparing various kinds of information, before repeatedly comparing various data of user communities, content-based filtering algorithm first needs to convert users' basic information into feature vectors with clear meanings. This requires that the user's information should be clearly expressed and described, rather than some ambiguous and conceptual features. However, the individuals in the society have the diversity of their culture and background, which would lead that the information of various pictures, music and videos cannot be accurately extracted with the existing Internet technology [14]. This would inevitably cause errors in the process of finding relevant information.

Compared with collaborative filtering algorithm, the information extracted by content-based filtering algorithm has a greater probability of containing noise, which increases the processing difficulty. In addition, extracting information according to content-based filtering is more dependent on professional knowledge of information processing than collaborative filtering algorithm. Meanwhile, the comprehensiveness, completeness and accuracy of the content understanding of the processor also have a great impact on the effect of recommendation.

2) Unable to get the core content of objects

The quality of text features extracted from text information often determines the recommendation effect of content-based filtering algorithm. However, content-based filtering algorithm sometimes fails to extract the essential information and core content of the text, resulting in its content concentration of objects often stays on the superficial analysis. And this has become an obvious drawback of content-based filtering algorithm. One of the typical examples is the processing of homophonic jokes by this algorithm. Homophonic jokes usually consist of several words with similar sounds but diametrically opposite meanings. Most homophonic jokes rely on inexplicable combinations of unrelated things and phenomena to achieve the humorous or ironic purpose, which is exactly what the algorithm cannot understand.

Moreover, content-based filtering algorithm is not available to make hierarchical recommendation for the things that their qualities are judged by users' subjective feelings. For example, some art appreciation websites are supposed to take visitors' subjective feelings as the first priority, while the quality of these web pages is determined by their beauty and aesthetics. However, content-based filtering algorithm that relies on keyword extraction can only make recommendations according to relevance. Therefore, it often cannot judge the actual quality of these web pages exactly.

3) Low novelty and narrow recommended range

Content-based filtering algorithm takes the personal behavior of individual users as the main measurement standard. Therefore, before the user takes the initiative pay attention to other types of interest things, in another word, there is no new user information is added to the database, it would be difficult for the algorithm to recommend diversified new things for the user [15]. As a result, the user would repeatedly view a large number of recommended contents which have high similarity but low overall value. To sum up, collaborative filtering algorithm also has the shortcomings that its recommendation items are single and lack of novelty. Moreover, it is unable deeply explore users' potential interests and hobbies.

This problem often occurs in newsfeed software using content-based filtering algorithm. For news with similar content, users only need to read a few of them to meet their needs in fact. However, given the operation mechanism of content-based filtering, many outdated old news and news with the same topic are easily pushed to users over and over again, while such recommendation is meaningless actually.

Table 1. Summary of the advantages and disadvantages of the two algorithms

Algorithm	ADV/DV	Point
Collaborative Filtering	Advantage	▪ easy to extract objects' feature information ▪ able to process complex unstructured objects (e.g. music and video) ▪ good novelty of recommendation results ▪ easy to find community's new interesting points ▪ high intellectualization, personalization and automation ▪ performance improving as the increase of user number ▪ no need for professional knowledge in specific fields
	Disadvantage	▪ problem about old start ▪ problem about data sparsity ▪ problem about maintainability and extensibility for the system ▪ the quality of system limit by the history dataset
Content-Based Filtering	Advantage	▪ less impact on the problem of cold start and data sparsity ▪ not depend on other users' data, ▪ no require for users' history data ▪ suited for minority groups ▪ intuitive and explainable for the recommended results ▪ supported by the mature technology of classification learning
	Disadvantage	▪ limited by its method in feature extraction ▪ difficult to catch the core content of objects ▪ unable to deal with ambiguous and metaphysical things ▪ low novelty of recommendation results ▪ narrow recommended range

4. Comparsion between Collaborative Filtering and Conntent-baed Filtering

4.1. Operation Principles

1) Collaborative filtering algorithm:

Collaborative filtering algorithm has two starting points. For one thing, by using group wisdom, users with similar interests are possible to like the same things. And for another is from the user himself, that is, the user himself might be interested in the goods that are similar to the items what he has purchased before. This algorithm collect data on community's behavior trajectories and make comparative deductions on this basis [16]. In general, the operation mechanism of the collaborative filtering algorithm ensures the highly personalization of the recommended results.

2) Content-based filtering algorithm:

Content-based filtering algorithm relies on the similarity based on the content itself. This algorithm only considers the nature of the object itself. Content-based filtering algorithm extracts the text feature information of the object and converts them into labels to form a set, so as to find the results with high similarity to the feature set in the dataset for recommendation.

4.2. Page Numbers

1) Collaborative filtering algorithm:

Collaborative filtering algorithms can fully promote fresh knowledge by tracking and analyzing the information about communities. The outstanding features of this algorithm are because its high convenience, intelligence and personalization, and its ability to accurately deal with ambiguous things. But at the same time, as to make its customer have a better user experience, collaborative filtering technology is supposed to solve its problems about cold start and data sparseness. In addition, the ability to appropriately increase the novelty of the recommendation results is also one of the important factors that why programmers prefer to apply collaborative filtering to accomplish the application of recommendation systems. Platforms should guide customers to choose their own interests and hobbies at an appropriate frequency, so as to better match customers with similar background information and reduce the pushes that customers are not interested in.

2) Content-based filtering algorithm:

As a recommendation algorithm which has achieved great success before collaborative filtering algorithm, content-based filtering algorithm is popular because of its simplicity in building recommendation system. In many minority group applications with small investment, content-based filtering algorithms are well preferred by developers. This is not only because it is easy to build a recommendation system based on this algorithm, but also because it does not depend on the individual data of customers. But at the same time, content-based filtering algorithm has its fatal weakness. Once there are too many recommendations of the same content, it would inevitably cause users' disgust. Therefore, how to introduce the novel information properly is a good reform point for content-based filtering algorithm.

4.3. Suitable Population

1) Collaborative filtering algorithm:

For the past few decades, collaborative filtering algorithm has been rapidly distinguished among many recommended algorithms because of its ability to precisely define the ambiguous objects according to the perceptions of the human community. One of the highlights is its rapid development in music software, such as Spotify. The superiority of collaborative filtering solves the problem that it takes a lot of effort for the software platforms before to manually classify the art which is difficult to define. Now the platform could categorize users with similar interests and preferences so that they can make personalized recommendations based on the trends of the groups.

2) Content-based filtering algorithm:

Content-based filtering algorithm has the characteristic of accurate classification according to the text feature information of objects. Thereby, it has a wide range of applications in areas that rely on key information for searching and recommendation. One typical example of this is its application in Quora. Content-based filtering algorithm can help users accurately locate the information they want by typing keywords in the search bar on Quora. Besides, content-based filtering algorithms have been extensively used in the applications for accurate book recommendation.

5. Summary

With the rapid development of Internet technology nowadays, the research of recommendation algorithms has become a key point in the scientific and technological academia. Based on the previous studies, this paper concretely introduces two main computer algorithms which have been widely used in recommendation system: collaborative filtering and content-based filtering.

By analyzing the specific processes of these two algorithms step by step, the advantages and disadvantages of them can be accurately analyzed and summarizes respectively. Through the research, it can be found that collaborative filtering algorithm achieves a high level of intelligence, personalization and automation by using the power of community, which enables it to keep the novelty of its recommendation results; however, on the other hand, the problems about cold start and data sparsity are also its Achilles heel which still need to be solved. While for content-based algorithm, the simplicity to build a recommendation system and the convenience of its easy operation have always been its advantages, but similarly, how to improve the accuracy of its extraction for important information by this algorithm is also a worth thinking problem.

References

[1] Zhou ZhiMin. Research and application of friend recommendation algorithm in social network. Kunming: Kunming University of Science and Technology, 2016.

[2] Li Jing. Research on recommendation strategy of academic friends based on paper cooperation network. Dalian: University of Technology, 2014.

[3] Yu Yan. Research on friend recommendation algorithm in social network. Nanjing: Nanjing University of Aeronautics and Astronautics, 2014.

[4] Fei Zhongyou. Research and application of random walk technology. Chengdu: Southwest Jiaotong University, 2014.

[5] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992, 35(12), 61-70.

[6] Kim, S. C., Sung, K. J., Park, C. S., & Kim, S. K. Improvement of collaborative filtering using rating normalization. Multimedia tools and applications, 2016, 75(9): 4957-4968.

[7] Wu Tao. Research and application of recommendation algorithm in recommendation system. Beijing: Beijing Jiaotong University, 2019.

[8] Zhao Fengyue. Research on collaborative filtering and content-based hybrid recommendation algorithm. Tianjin: Tianjin University of Finance and Economics, 2016.

[9] Reng Lei. Research on Key technologies of recommender system. Shanghai: East China Normal University, 2012.

[10] Liu Shuai. Research on correlation measurement of high dimensional sparse data. Beijing: Capital University of Economics and Business, 2014.

[11] Lops P, Gemmis M, Semeraro G. Content-based recommender systems: State of the art and trends. Recommender systems handbook, 2011: 73-105.

[12] Barbu, T. Content-based image retrieval using gabor filtering. IEEE In 2009 20th International Workshop on Database and Expert Systems Application, 2009.

[13] Feng Shan. Research on Friend recommendation algorithms in social networks. Beijing University of Posts and Telecommunications, 2015.

[14] Tkalčič, M., Burnik, U., & Košir, A. Using affective parameters in a content-based recommender system for images. User Modeling and User-Adapted Interaction, 2010, 20(4): 279-311.

[15] Nie Shuaihua. Design & Implementation of intelligent dating website based on content recommendation or collaborative filtering recommendation algorithm. Wuhan: Central China Normal University, 2016.

[16] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. Grouplens: An open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work, 1994.