Application of Collaborative Filtering Personalized Recommendation Algorithms to Website Navigation

Hao Wu*
School of Computer Science, JLUZH, Zhuhai, China
*Corresponding author email: haowu_mouse@hotmail.com

Abstract. This article builds a personalized management and push design for the functional and performance requirements of website personalized recommendations. The system uses PHPstorm as the development platform and Hadoop distributed file system combined with DBMS design algorithm to achieve data level research, and uses a combination of customer-based collaborative filtering and project-based collaborative filtering to categorize clients and web pages in personalized recommendation, which improves the calculation and personalization of website navigation to a certain extent.

1. Introduction
Nowadays, more and more information is available on the Internet, and navigation websites provide the same website navigation and search functions to different users. Faced with the huge number of portals, it is difficult for users to quickly get the information they need from the dizzying number of pages. How to classify such a huge amount of information, which is increasing in real time, and assign it to different users has become the most important issue nowadays.

In this paper, the collaborative filtering personalized recommendation algorithm is applied to the website navigation system. According to the user's browsing history and information needs, the data mining is carried out on the user's reading and search records, and the interest features are classified to predict the user's possible future reading behavior, so as to provide the user with personalized website navigation.

2. Co-filtering Recommended Algorithm
Collaborative filtering is the most commonly used personalized recommendation technology and it consists of both online collaboration and offline filtering. Online collaboration refers to the analysis of online data to identify items that users may like, while offline filtering refers to the analysis of the results of the filtering, so as to remove some meaningless recommendations.

Collaborative filtering algorithms can be divided into two categories: the first category is user-based collaborative filtering; the second category is content-based collaborative filtering. User-based collaborative filtering algorithms are computed through the similarity between users, which is based on the principle that the user's favorite products will be analyzed and will be compared to similar products that the same computed user has browsed or purchased. Both algorithms are essentially based on the following steps: (1) collecting and analyzing the user's browsing history or operation data; (2) finding users or items that are similar to the analysis results; (3) filtering the data from the previous steps to get the most compatible information for recommendation in the client.
3. System Architecture Design

The personalized website navigation recommendation system needs to solve the following problems: (1) the storage capacity of massive web page information and user browsing information requires the system to be very strong and capable of good horizontal expansion; (2) the storage of massive web page information and user browsing information; (3) as the navigation system must take into account the growth in the number of web pages in the future development process, especially in the project's The demand for the computational power of the system may gradually intensify in terms of matrix construction and computation of users and projects, thus requiring performance requirements for personalized navigation systems; (4) the system should have a highly scalable and fault-tolerant computational framework to handle large amounts of data and to analyze historical user behavior data. The design chooses Hadoop platform as the background framework of the system. The system flow analysis is as figure1.

The system first stores the user behavior logs and website contents into HDFS, then configures the recommendation model and submits it to Hadoop for calculation, and finally Hadoop filters the calculation results to get the user recommendation results.

3.1. User Management Module Design

The user management module is mainly responsible for user tag management, user information management and user behavior diary management.

- Management of user's tags: each user can have multiple tags, which in the navigation system are mainly topics of interest to the user or the type of that topic. Users can manage their own tags to manage the type of recommendations for their navigation.
- User information management: the management of user account passwords and basic personal information, mainly for recommendations and website information cloud login.
- User's behaviour diary: this management is mainly used to record analysis to derive recommended websites that users are matched to for management.

In addition, judgment is required when the user logs in to the account, and the judgment flow when the user logs in to the system is shown in Figure 2.

Figure 2 shows that after a user logs in, the system first determines whether the user is a new user or not based on his or her information, and if the user is a new user, then the system can only recommend various types of hot topics in real time.

3.2. Personalized Recommendation Module Design

Personalized recommendation is the core function of the whole system, which mainly generates user logs according to the user's usual usage, and then takes the data and compares it with other data through Pearson's correlation coefficient calculation to obtain the similarity of the user's interest in a certain kind of things. This result is stored in the database. Every time the user opens the navigation, the system will give priority to recommend computer information to the user, and will push the gold mining, segmentfault computer-related URLs to the user to choose.

So, personalized recommendation system, not just to record user behavior, the next time you open the user's most commonly used website, but more importantly, recommendations, it will be based on your short-term needs, will be a variety of similar, may be helpful to you recommend the website to you,
greatly increasing the width of the user's vision. The personalized recommendation module, closely linked between the various modules, collaborative operation, the process is shown in Figure 3.

Figure 3. Implementation flow of the personalized recommendation block.

From the analysis of Figure 3 above, the website navigation system is divided into three modules: model modeling module, similarity calculation module and personalized recommendation module.

- The model modeling module is responsible for analyzing user operation logs and new web page information, and providing feedback to HDFS. this process first gets the user operation log data or web page information from the database, and then analyzes the data through modeling, which can lead to the type of web page or a user tag.

- The similarity calculation module is mainly responsible for clustering users and web pages, which is based on the types of web pages and users calculated by the previous model. After the model is obtained from HDFS, the information obtained is calculated and the pages with higher similarity are returned to the DBMS, while the users with higher similarity are returned to HDFS.

- Personalized recommendation module is the last part of the whole system, based on the previous two modules, it matches and recommends the clustered web pages with clients by using the client-based collaborative filtering algorithm. First, it gets the clustered users with high similarity to other users from HDFS, then it gets the similarity ranking of web pages from DBMS, and then it compares and matches the information from the two tables and filters it, and finally recommends it to the client.

Setting a separate module to cluster users and web pages is an effective measure to shorten computation time, clustering users and web pages first, shortening the search space time, and helping the system to make larger recommendations, as well as meeting the requirements of the navigation on the system, which can effectively improve the personalized service of Clients.

A flowchart of the personalized service is shown in Figure 4.

Figure 4. Personalized service implementation process.

As can be seen in Figure 4, this process is a pre-processing of the data by the theme module, with the main objective of modeling to obtain the web page and the model to which the client belongs, in preparation for the subsequent steps. It is shown in Figure 5.

Figure 5. Cluster analysis of users and web pages. Figure 6. Matching and filtering the list.

From Figure 5, it can be seen that this paper mainly adopts a collaborative filtering technique to cluster users and web pages separately, in order to recommend web services for subsequent users, and this simultaneous clustering analysis of items and users. Improve the accuracy and calculation time when large systems make recommendations to users, making the website navigation system much more efficient. It is shown in Figure 6.
From Figure 6, we can see that the above filter is mainly to filter some Clients have visited the web page or some interested, Clients unfollow or poor evaluation of the page, the multi-layer matching filter is to give Clients a more accurate way to recommend, improve Client usage and use satisfaction.

4. Database Design

For personalized navigation database design, it includes web page label information, Client labels and personal information, the data within the module in this system is stored in HDFS, while the DBMS stores the module output information, mainly designed four tables, the first one is the Client information table, as shown in the following table 1. The second one is the personalized tag classification table, as shown in Table 2 below. The third table is a web page type table, as shown in Table 3 below. The fourth table is the web page information table, as shown in Table 4 below.

| Lists | Data Type | Length | Null or void? | Property Name      | Descriptions       |
|-------|-----------|--------|---------------|-------------------|--------------------|
| Sbu_ID| int       | 20     | No            | Client ID         | Primary key        |
| Sbu_No| varchar   | 42     | No            | Client Account    | /                  |
| Sbu_Password| varchar | 42     | No            | Client Password   | /                  |
| Sbu_Name| varchar | 42     | No            | Client Name       | /                  |
| Sbu_Cookie| varchar | 50     | Yes           | Client Cookie     | /                  |
| Sbu_RFID| text     |        | Yes           | Client Personalized tags | /               |
| Sbu_Behaviour| text  |        | Yes           | Client Behavior   | /                  |

Where Sbu_Behaviour is used to record the behavior of Client operations, mainly for logging events.
- Web pages viewed by the Client.
- Indicates Client comments on the web page.

The main purpose of a personalized testimonial list is to store information about the page and the Client, mainly the Client id, the collection of page ids and the time of creation.

| Lists | Data Type | Length | Null or void? | Property Name        | Descriptions       |
|-------|-----------|--------|---------------|----------------------|--------------------|
| Sbr_ID| int       | 20     | No            | Labels ID            | Primary key        |
| Sbr_name| varchar | 42     | No            | Labels Name          | /                  |

From Table 2, the personalized label classification table has two attributes, where Sbr_ID represents the id of the label and Sbr_name represents the name of the label.

| Lists | Data Type | Length | Null or void? | Property Name                  | Descriptions       |
|-------|-----------|--------|---------------|--------------------------------|--------------------|
| Sbmeta_ID| int      | 20     | No            | Web Type ID                    | Primary key        |
| Sbmeta_NAME| varchar | 32     | No            | Web Type Name                  | /                  |
| Sbmeta_URL| varchar | 54     | No            | Jumping Address                | /                  |
| Sbmeta_RFID| text    |        | Yes           | Suitable Personalized Labels   | /                  |

From Table 3, the web page type table is divided into four attribute attributes, where the primary key is Sbmeta_ID for the type id of the web page, Sbmeta_NAME for the type name of the web page, Sbmeta_URL for the jump address, and Sbmeta_RFID for the personality tag of the individual.

| Lists | Data Type | Length | Null or void? | Property Name                  | Descriptions       |
|-------|-----------|--------|---------------|--------------------------------|--------------------|
| Sbmi_ID| int       | 20     | No            | Web Information ID             | Primary key        |
| Sbmi_Title| varchar | 32     | No            | Webpage Title                  | /                  |
| Sbmi_URL| varchar | 84     | No            | Web Bounce Information         | /                  |
| Sbmi_Type| int      | 10     | No            | Type of Information            | /                  |

From Table 4, it can be seen that the web page information table is divided into four attribute attributes, where the primary key is Sbmi_ID which represents the type id of the web page, Sbmi_Title which
represents the title of the web page, Sbmi_URL which represents the jump information, and Sbmt_Type which represents the type to which the web page information belongs.

5. System Implementation

5.1. Client Management Module Implementation
Client management module is mainly used for the management of Client tags, Client account information, the management of Client behavior logs, Client tag management implementation is the addition or deletion of Client tags, Clients can add or delete their own tags. Client account information includes Client accounts, collection pages, etc., while the Client's action log is the page information generated by the Client when browsing or searching the page.

5.2. Implementation of the Sexualized Recommendation Sub-module
The personalized recommendation module uses a combination of customer-based co-filtering and project-based co-filtering, and this paper uses the combination of these two algorithms to categorize web pages and customer tags, and to categorize customers and web pages respectively before personalized recommendation, which will improve the calculation of website navigation, a system that accommodates relatively large amounts of data, as well as the degree of personalization of the system.

5.3. Similarity Calculations
The similarity calculation of the more important part of the collaborative filtering algorithm has a role in determining whether a recommendation system is good enough. Currently, the main methods of calculation are Euclidean distance, Pearson's correlation coefficient, Cosine similarity, and Tanimoto coefficient.

The formula for the Euclidean distance is shown in Eq. (1).

\[ d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1) \]

In equation (1), x and y are the parameters of the two targets, and the similarity calculation refers to the distance between two points in two-dimensional space, or the distance from the point to the origin of the vector, and in two- and three-dimensional space to the actual distance between the two points.

The formula for the Cosine similarity is shown in Eq. (2).

\[ T(x, y) = \frac{x \cdot y}{\|x\| \times \|y\|} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}} \quad (2) \]

Where x, y is the component of the two items, T(x, y) is the cosine similarity of the two items, which can be analyzed by Eq. (2), the closer the cosine similarity is to 1 i.e. the higher the similarity of the two items.

Cosine similarity is a measure of the size of the difference between individuals, and is obtained by calculating the cosine of the angle between the vectors in vector space. The closer the cosine value is to 1, the greater the similarity between the two vectors. Cosine similarity and the Euclidean distance algorithm child are more likely to calculate differences in the direction of the objects, whereas the results from the European distance calculation are biased towards the numerical differences between the two items.

The formula for Pearson's correlation coefficient is shown in Eq. (3).

\[ p(x, y) = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1)S_x S_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (3) \]

Where x, y are the two item variables and n is the number of samples. The analysis of Eq. (3) shows that P(x, y) varies in the range [-1, 1], as the absolute value of Pearson's coefficient is closer to 1, i.e., the more similar the two variables are.
The correlation coefficient is obtained by dividing the covariance by the standard deviation of the two variables. Although it is possible to indicate the degree of correlation of random variables by whether the covariance is greater than zero, the structure of the calculation in this way is more dependent on the data and it is difficult to obtain more accurate results.

The Tanimoto coefficient (generalized Jaccard similarity coefficient), also known as the generalized jaccard similarity coefficient, uses the data to find the degree of concentration, or similarity, of the sample through the similarity coefficient. Generalized Jaccard coefficients can be used on document data and are normalized to Jaccard coefficients in the binary property case.

The formula for the Tanimoto coefficient is shown in Eq. (4).

\[ p(x, y) = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1)s_x s_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \] (4)

Where \( s_x \) and \( s_y \) is the sample standard deviation of \( x \), \( y \), and \( p(x, y) \) ranges from \([-1, 1]\), and the correlation is stronger when the absolute value of \( p(x, y) \) is greater. Tanimoto coefficient reflects a statistic of the degree of linear correlation between two variables and is used to compare the similarity in the sample set. Jaccard coefficient can accurately find the degree of similarity, i.e., the intersection of the samples and Contrast ratios are formed with the sample set.

6. Conclusion

This paper is based on the collaborative filtering algorithm for website recommended navigation design, according to the user's browsing history information needs, the user's reading and search records for data mining, the interest characteristics and other categories, to predict the user's possible future reading behavior, to provide users with personalized website navigation, personalized navigation recommendation system to solve the problem of information overload when users go online.

The system adopts Hadoop to analyze and process customer's operation diary offline, and with the advent of Internet information age, the number of web pages updated every day is countless, which makes it difficult for the system to make recommendation in real time and may even cause delay problems. In terms of web pages, because some keywords are ambiguous or have lexical meanings, the model has some errors when classifying web information.

References

[1] Huang Ren, Meng Tingting. An overview of personalized recommendation algorithm [J]. SME Management and Technology (Midterm), 2015(03):27 1
[2] Jingwei Wu. Research on collaborative filtering algorithm [J]. Computer Knowledge and Technology, 2019, 15(03).
[3] Wang Lin, PHP advantages and specific application analysis in dynamic website development [J]. Information Records Material, 2019, 20(03):30-31.
[4] Li Jing. Principles and practice of computer database software design [J]. Computer Products and Circulation, 2019(03):16.
[5] Zhao Tao. Analysis of website navigation system design [J]. Computer Knowledge and Technology, 2017, 13(36):224.
[6] Wang Ning, He Zhen, Huang Ze, Zhou Yipeng, Wu Xinliang. Improved collaborative filtering algorithm for apparel personalized recommendation[J]. Journal of Hunan Academy of Engineering (Natural Science Edition), 2019, 29(01):33.
[7] Zhu Manzhou. Design and implementation of a personalized news recommendation system based on collaborative filtering [D]. Nanjing University of Technology, 2019.
[8] Wang Ping. Design and implementation of a news personalized push system based on collaborative filtering algorithm [D]. Hunan University, 2016.
[9] Tang Xiaobo, "Personalized Recommendation Based on Ontology and Tags", Information Studies: Theory & Application, vol. 39, no. 12, pp. 114-119, 2016.
[10] Dong Yue-hua and Liang Xue-lei, "Collaborative Filtering Recommendation Algorithm Based on Tag Importance", Science Technology and Engineering, vol. 18, no. 14, pp. 172-178, 2018.