Research on Recommendation Technology Based on User Portrait

Min Huang

School of Business, Information System, Stevens Institute of Technology, U.S.

Keywords: Big Data, Recommendation System, User Portrait, Recommendation System Algorithm, User Portrait Algorithm

Abstract: Under the background of the rapid development of Internet applications, various data amounts in the big data platform have shown explosive growth. Although both enterprises and the government have a large number of original data of the customers, how to properly use them requires a comprehensive system to support. The recommendation system in a wide variety of systems has been sought after by countless people as an effective and accurate marketing tool. In the recommendation system under the background of big data, the application of the user portrait technology has gradually become a research hotspot for countless people. Through introducing the user portrait, this paper has a deep exploration of the algorithm used in the recommendation system and the meaning and construction framework of the user portrait. In the end, it expects the challenges and solutions confronted by the user portrait technology recommendation system.

1. Introduction

1.1 Background

In the recent years of rapid science and technology development, customers or users are more likely to purchase or have other similar behaviors through online and network. Specifically, in the context of the straight-line development of the e-commerce industry, a lot of consumption data has been generated. Through comprehensive analysis of massive data and multi-angle and multi-dimensional data, enterprises can implement marketing methods to consumers in a more direct way, so as to realize precise marketing.

1.2 User Portrait Profile

User portrait is just an effective tool derived from such a large background -describing the target user and understanding user demand acts as the guidance direction of the final product design, and user portrait is also one of the most effective tools of helping enterprises accurately identify and analyze users. The consumption data and user information that were previously floating in the network are classified into various simple labels under the background of big data, and then these labels are put on the user to objectivise the user's image and preferences, so that the enterprises can provide users with accurate products and services that meet their needs, thus achieving the purpose of precision marketing.

Online social networks such as WeChat and MicroBlog in China, as well as Facebook and Twitter in foreign countries have become the most important networks in the world, so that they play an important role in the daily life of all the people. Therefore, applying the user portrait technology to these online social networks becomes very effective. First of all, the user's behavioral data can be used to evaluate and design the recommended content launch strategy in products, to deliver on demand and preferences, and to effectively assist product designers in product design. Secondly, an accurate user behavior model can enable marketing salespeople to disseminate product content or arrange promotions faster and more extensively.

Figure 1 Presents the Connotation of User Portrait in a Simpler Way:

  DOI: 10.25236/scmc.2019.094

Data		Business		Modeling
Social attribute; living habit; consuming behavior	+	Meet business requirements; Specific user	+	Data Mining; Data Visualization

Figure 1 Basic Meaning of User Profile [1]

2. Basic System Framework of User Portrait Recommendation System

The explosive growth of the Internet scale and coverage have brought about the problem of information overload. Such a large amount of information is presented to the user at the same time, and it will be difficult for the user itself to extract the useful part. As an important means of information filtering, the recommendation system plays an important role in solving the problem of information overload. The user portrait of the recommendation system at the present stage is to vector the user to show. The reason is also because the user portrait is a secondary product generated by the recommendation system in a single process in the process of perfecting its own system, it requires a computer instead of a person to understand.

The basic architecture of the recommendation system based on the user portrait is given below:

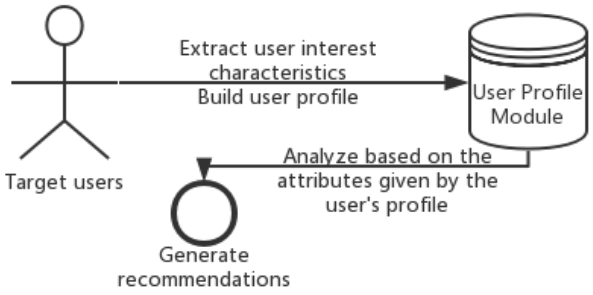

Figure 2 Recommendation System - User Profile Architecture

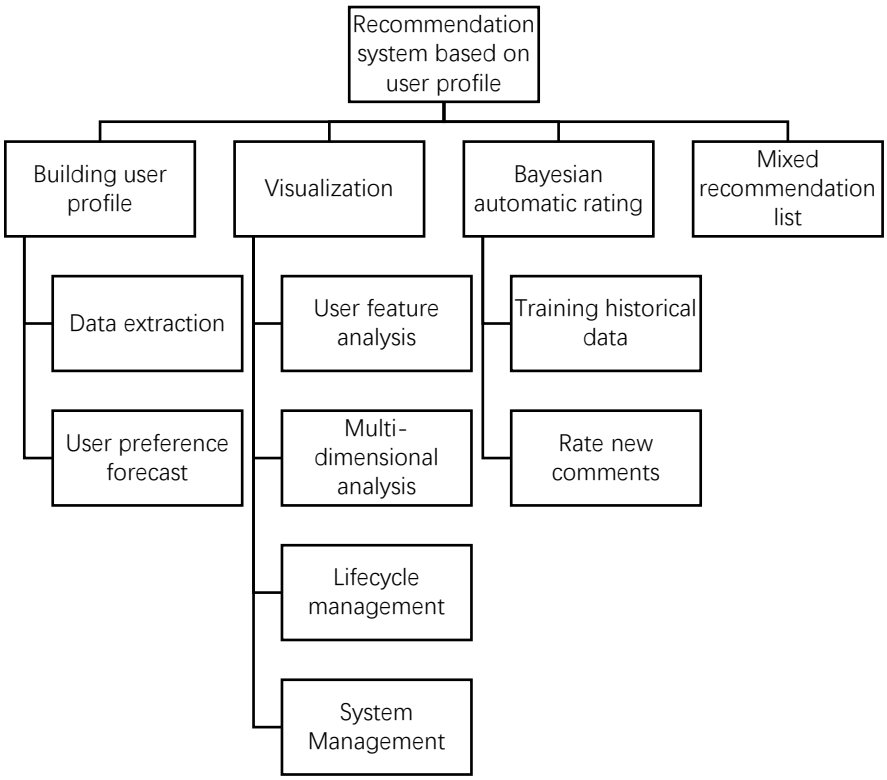

Figure 3 Overall Structure of Recommendation System Based on User Profile

2.1 Build the User Portrait

The multi-dimensional user portrait data extraction method is a relatively popular user image data collection method. Multi-dimensional data extraction is to take out the data of the user footprint from the data warehouse Hive, then classify it step by step, collect relevant data according to the classified center point, iteratively judge according to the source of the URL in classifying, eliminate the unmatched ones through some filtering conditions, obtain the “clean” data and store the data in the data warehouse, and then carry out excavated data analysis and predication, and finally obtain the predication and analysis results based on the user portrait. From the perspective of the technical level, most of the current user portrait techniques are used jointly with data mining techniques. Since the user's data is used to store information such as user interest characteristics on a user-by-user basis, although it is easy to obtain the visualized user data for the present enterprises, it is unable to find the best solution in such a massive data collection methods, instead the enterprises can only analyze according to practical conditions, so as to find a most appropriate method.

The general algorithm for user preference analysis is the linear regression algorithm. The following is an analysis and prediction flow chart for constructing the user portrait:

1) Start
2) User information and browsing footprint
3) Add the analysis dimension, collect user information and store it in the data warehouse.
4) Data mining algorithm analysis - linear regression equations and prediction of user preferences
5) Build user portraits to show users' future behavioral preferences
6) Place the target predication result in the warehouse according to different weights.
7) End

2.2 Visualization

The visualization process is divided into four modules:

1) User feature analysis is to make the presentation of user information more intuitive and accurate. The information data also includes some information attributes such as user behavior information and purchase information;

2) Multi-dimensional analysis is to mine the results of user feature analysis, and then present the user’s multi-dimensional results with other predication results being combined;

3) Lifecycle management, just as its name implies, is to manage the lifecycle of data, focusing on data source updates and user data lifecycle detection, reducing useless data and accuracy caused by lifecycles;

4) System management is the management of the above entire system, including the jurisdiction and page, as well as system modification and other operations caused by external reasons.

2.3 Bayes Automatic Grading

Bayes classifier [2], also known as simple Bayes classification or Bayes filter, its classification principle is to use the Bayes formula to calculate the probability of the prior probability of the object in a certain object category, and choose the highest probability in the probability as the classification scope of the object.

The reason of selecting the Bayes classifier for automatic grading among various classification algorithms is that it is suitable for enterprises in most cases, because the existing recommendation strategy set of the enterprise has been perfected in most conditions, it is very risky to propose a kind of new recommendation strategy to replace the original strategy. The Bayes classifier can be used to add and mix a new clue on the basis of the participating methods of the original strategy, so as to increase the recommendation effect and accuracy. The main understanding of Bayes automatic grading is to automatically score items based on historical reviews of the items, and then operate the list after obtaining these grading results: ranking (in the order of customer satisfaction from high to low). In this way, the time for the user in checking the reviews, making comparisons in purchasing

is reduced at the same time of recommending, which is beneficial for the sales.

2.4 Mixed Recommendation List

After completing the three stages of user portrait construction, visualization, and Bayesian automatic scoring, the processed data will be presented for being used according to the recommendation results of the mixed recommendation list, so as to complete the construction of the recommendation system based on the user portrait technology.

3. Detailed Content of User Portrait

To sum up, the user portrait is a very useful system under the platform of big data analysis. It is suitable for being used in various systems and processes, and can help realize the accurate and personalized effect of applying to each individual.

3.1 Basic Ideas and Principles of the User Portrait

The basic idea of the user portrait is to label users' behavioral characteristics. There are labels with single content and complicated content. The labels of complicated content are derived from the original and single labels according to a certain logic. Through this series of labeling behaviors, the user's behavior and characteristics can be highly summarized and understood, and it is also convenient for artificial intelligence for further treatment.

Building user portrait is promoted because the data in a certain scope has insufficient dimension and satisfactory demand on the basis of scenarios and objectives.

3.2 Labels of Different Dimensions and Label Processing

Let’s go back to the label for constructing the user's portrait. The user portrait, as its name suggests, is to model various types of information for a real user, but it can also be classified into a person's basic information, a person's behavior, a person's social attributes, and the like for the user. Different insights into the constituent dimensions have also emerged during the construction process. A few typical descriptions are listed below:

1) Retrieving user characteristic information mainly includes two aspects: stable factors related to the user (such as the user's personal basic information, behavioral information and long-term habit information) and variable information (such as the detection environment, search target and other possible changing factors). [3]

2) User portrait can be divided into attributes of different types, including the natural attribute, relationship attribute, interest attribute, ability attribute, consumer behavior attribute, and credit attribute. [4]

3) Taking the Internet communication industry as an example, adding user behavior preference attribute (access preference, latest attention, search information, business usage, application usage ranking, social media analysis, traffic consumption, and terminal) on the basis of the user's basic information attributes. User portrait with the preference attribute will better show the demographic characteristics, as well as the differences of the user in habit, attitude and behavior track and so on[5].

The labeling system makes the analysis of user portrait into a process of classifying users with common characteristics into one group through the clustering means, and then discovering more core groups through group analysis. The analysis process is generally divided into five steps:

1) Qualitative analysis of the user portrait. That is to analyze the user's needs according to the current market environment, and lay the theoretical and data foundation for the subsequent steps.

2) Obtain the user image data source. The data collecting method is also introduced in the previous chapter.

3) User portrait similarity calculation.

4) User portrait clustering.

5) Production of group user portraits.

3.3 Data Storage Method and Key Word Extracting Method of User Portrait

As a model structure describing the target user, the user portrait is the relevant system design and user requirements. The effective methods sought have been widely used in many fields. [6] The user portrait usually has the following manifestations:

1) The keyword method, that is, the feature attribute of the user is represented by several feature words.

2) The grading matrix method is represented by a two-dimensional matrix, the row represents the user, the column represents the item, and the intersection represents the user grading or interest tendency of the item.

3) Vector Space Model (VSM) will give the given key signing weight value, ie the tag weight (Tab-Value).

4) Ontology representation, stores user relationships and attributes through an ontology model.

The storage methods of user portrait labeling data include: relational database, NoSQL database, data warehouse, and so on.

The labels using these existing key words must use the key word extraction method. The commonly used methods are TF-IDF and TextRank. Both methods have their own strengths, but one thing in common is that there is no need to label data. They both belong to the non-supervision means.

3.3.1 Tf-Idf

TF is called the word frequency, and IDF is the inverse document frequency. The thought of the TF-IDF method is quite simple: words that appear repeatedly in a text are more important, and words that appear in all texts are less important. These two points are quantified into two indicators: TF and IDF:

1) TF, the number of occurrences in the text. Since the word frequency is usually 1 in a short text, TF is more useful in long text.

2) IDF is counted in advance. In all existing documents, the number of texts with the appearance of each word is counted(recorded as n), which is the document frequency, and the number of texts is counted as well(recorded as N).

IDF is calculated like this: $$IDF = log\frac{N}{n-1}$$

After calculating the TF and IDF values, the following methods can be used to extract keywords:

• Keep labels with the highest TopN weight.

• Set a threshold and retain the label on the threshold.

• Calculate the weighted mean and retain the label on the mean.

In addition, in some scenarios, some other filtering measures will be added, such as extracting only verbs and nouns as keywords.

As to this method, there is only one precondition, which is to calculate the IDF value of the dictionary in advance. In particularly, for short text, TF does not take effect, the IDF value ranking is almost the sole dependence.

3.3.2 Textrank[7]

The idea of the TextRank algorithm is similar to that of PageRank and it can be summarized as:

1) In the text, a window width is set, such as the word of K, which is used to count the co-occurrence relationship between words and words in the window, and regard it as an undirected graph.

2) The importance of all word initialization is 1.

3) Each node distributes its weight equally to other nodes that connect to it.

4) Each node regards the sum of the weights assigned to it as its new weight.

5) Repeat steps 3 and 4 in this way until the converge of all node weights.

Through weighing the words calculated by TextRank, this feature is presented: those who have a co-occurrence relationship will support each other to become keywords.

4. Specific Contents of the Recommendation System

4.1 Present Recommended Algorithm

The recommended algorithm is the lifeblood of the huge recommendation system. The type and performance of the whole system are largely determined by its performance. At present, the mainstream recommendation algorithms are the following [8]: content-based recommendation, collaborative filtering recommendation, knowledge-based recommendation, and combined recommendation.

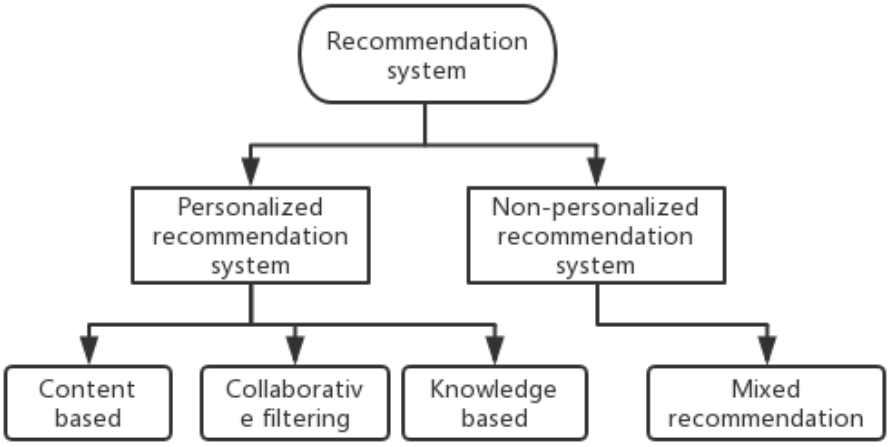

Figure 4 Basic Algorithms Commonly Used in Existing Recommendation Systems

4.1.1 Content-Based Recommendation

Content-based Recommendation means to recommend objectives with similar attributes through analyzing the objectives selected by the user in the past. The original information retrieval method is the source of such algorithm[9]. The features of the objective content extracted through the feature extraction method is used to represent the attribute of the objective, and then the system studies the user interest through the features of the user evaluated objectives, so as to comprehensively investigate the complete user materials and the matching degree with the objectives to be predicated.

The extraction of most target object content features (content(s)) is based on the text description of the target object. In addition, the content based profile of the user can always be used no matter which machine learning method, data mining algorithm is used, such as: the decision tree, the above-mentioned Bayesian classification algorithm, and the neural network and so on. In summary, the utility function can be defined as [10]:

u(c,s)=score(ContentBasedProfile(c), Content(s)).

There can be many different types of Score calculation methods depending on the algorithm used. The final value of u is the basis for sorting, and the item with the highest value and the foremost place is recommended.

Cold boot does not cause shortcomings because of the heat limitation, and it is not related to the browsing history. However, the content-based recommendation algorithm will always recommend the closely related items to the users due to excessive specialization, thus losing the diversified recommendations.

4.1.2 Collaborative Filtering Technology

collaborative filtering technology is one of the maturest technologies in the recommendation system, and one of the most successful technologies in the recommendation system. Its algorithm can be classified into the user-based collaborative filtering and item-based system filtering.

4.1.2.1 User-Based Collaborative Filtering

Briefly, user-based collaborative filtering means selecting any user that is most similar to the target user for recommendation. [11] The specific steps are as follows:

1) Collect and analyze the evaluation of several users to the item;
2) Calculate the similarity value among all users based on these evaluations;
3) Select several users that are most similar to the target user;
4) Screen the items with the highest evaluation of the user and the items that the target user has not purchased or browsed, and select the high-score items to recommend to the target users.

4.1.2.2 Item-Based Collaborative Filtering

Item-based collaborative filtering doesn't need to collect the browsing records of several customers to the items, instead, it is only required to analyze one single browsing record. [12] The specific steps are as follows:
(1) Collect and analyze the browsing record of several users to one item;
(2) Obtain the similarity among all items according to the analysis;
(3) Find several items with the similar attributes to the items with the highest evaluation of the target users;
(4) Recommend these items to the target user;
Although collaborative filtering is simple and practical, it still has some problems:
1) Require excessively a too accurate user evaluation;
2) Hot items are more likely to be recommended to users;
3) In case of a cold boot, there might be condition that when the new user joins or new product is put on shelf, it is unable to recommend;
4) If the item has a short life cycle and a quick updating speed, there might be recommendation missing caused by the sparse grading matrix.

4.1.3 Knowledge-Based Recommendation Algorithm

Knowledge-based recommendation algorithm [13] uses a ready-made expert knowledge base to semantically extend the classification features of items, and extract semantically similar content for recommendation. The knowledge base will also identify specific nouns, entities, subjects, new terms and so on to improve the accuracy of the recommendation. The shortcoming of this kind of knowledge-based recommendation is that it requires a high-cost expert knowledge base and a short-term and timely update.

4.1.4 Mixed Recommendation Algorithm

In most of the practical applications, the above single algorithm is seldom used to form a complete recommendation system, instead, the mixed recommendation algorithm is always used. The principle is to avoid and remedy the weaknesses and defects of each recommendation technology. The results of different algorithms can be given a weight suitable for the project, the results are integrated, or different algorithms can be used in different calculations, and then the algorithm is selected according to the fact that the situation is closer to the direction of the business. The specific steps are shown in the table [14]:

Table 1 Mixed Recommendation Algorithm

Combination Method	Description
Weighting	Weighted combination of multiple recommendation algorithms
Transformation	According to the background of the problem and the actual situation or requirements, the decision to use different recommendation algorithms
Mixing	At the same time, a variety of recommendation algorithms are used to give a variety of recommendation results to the user
Cascading	First use a recommendation algorithm to generate rough recommendation results, and then use another recommendation algorithm to make more accurate recommendations on this basis
Feature expansion	An algorithm generates additional information features embedded in the feature input of another recommendation algorithm
Meta level	A model generated by a recommendation algorithm is used as input to another recommendation algorithm

5. Challenge and Expectation

At present, the user portrait under the big data platform is to extract feature making labels from the user data to represent users on the basis of machine learning technology, and use these attribute labels as the labeling data to be trained in the user portrait predication model, so as to predicate more users without the attribute label. The seemingly complete system steps still have a certain challenges:

1) Most of the existing methods use discrete features of manual extraction. These features do not have a context-related information for the entire user data, thus having certain limitations on the user's expressive force;

2) The algorithms used by user portraits are basically simple classification models or linear regressions, thus being unable to automatically learn some deep abstract features from a large amount of user data, and it is also unable to model the relationship among features, thus making the entire model simple and not rich;

3) User portrait has not yet considered the timeliness of attribute label, and it is difficult to describe the dynamic changes of a user.

The above three challenges are very likely to be realized in the future. It is because, after the deep learning technology becomes matured, the deep neural network can be used to deeply and abstract the user's original data. These features can then help to effectively enhance the accuracy of the user portrait. Secondly, the deep neural network user representation model is promising to effectively overcome the shortcomings of the existing linear equations and classification models. In the end, the speed of network development is unimaginable, and the information of the mass users has been advanced from the basic information from the very beginning to data of different modalities, so it is possible to form a very large and diverse user portrait through the multi-source data in the future.

References

[1] Wang Xianming, User Portrait Construction Based on Video Big Data[J], 2017-01-01

[2] Mooney RJ, Bennett PN, Roy L. Book recommending using textctegorization with extracted information. In: Proc. Of the AAAI’98/ICML’98 Workshop on Learning for Text Categorization. Madison: AAAI Press, 1998.49-54.

[3] Lafouge T, Lardy J P, Abdallah N B. Improving information retrieval by combining user profile and document segmentation [J]. Information Processing Management an International Journal, 1996(3): 305-315.

[4] Li Yingkun, User Portrait Statistics Method Practice Research under Big Data Background [D], Beijing: Capital University of Economics and Business, 2016

[5] Ma ANHUA, Accurate Marketing System Design and Realization Based on User Behavior Analysis [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2013.

[6] Xu Luyao, Jiang Zengqi, Huang Tingting et al. Profile of User Portrait System Based on Big Data[J], Electronic World, [6]2018(2):64-65.

[7] Resnick P, Varian HR. Recommender systems. Communication of the ACM, 1997,40(3):56-58.

[8] Arekar, T., M.R. Sonar, and N.Uke, A Survey on Recommendation System. 2014.

[9] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. New York: Addison-Wesley Publishing Co., 1999.

[10] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the art and possible extensions. IEEE Trans. On Knowledge and Data Engineering, 2005, 17(6):734-749.

[11] Delgado J, Ishii N. Memory-Based weighted-majority prediction for recommender systems.

In:proc. Of the ACM SIGIR’99 Workshop Recommender System: Algorithms and Evaluation. New York: ACM Press, 1999.

[12] Marlin B. Modeling user rating profiles for collaborative filtering. In: Proc. Of the 17th Annual Conf. on Neural Information Processing System. Cambridge: MIT Press, 2003.627-634.

[13] Burke R. Knowledge-Based recommender systems. Encyclopedia of Library and Information Systems, 2000,69(32): 180-200.

[14] Jannach, D. et al. Recommender Systems: an introduction. 2010: Cambridge University Press.