ISSN: 1813-4890

Performance improvement of content-based recommendation systems in big data environments

Song Sun

Ukraine Odesa I. I. Mechnikov National University, Ukraine

1010492983@qq.com

Abstract

likely items of interest to users among all items (products, content, etc.) and recommend them to them. Like online retail websites, they will spend a lot of effort building recommendation systems to achieve more accurate recommendations and increase item sales. From a data source perspective, it is generally obtained through user feedback, such as ratings after watching movies, search content, purchase history, etc., or obtaining some likes and dislikes information about users and products from other sources. From the user's perspective, users can feel understood by the system and know what they want, leading to more behaviors in the system. These behaviors can be fed back to the recommendation system to optimize their understanding of the user, forming positive feedback and increasing user stickiness.

Keywords

Big data, Recommendation system, data mining.

1. Introduction

A classic recommendation algorithm that generally only relies on the content and behavioral attributes of the user and the item itself, without involving the behavior of other users[1-4]. It can still make recommendations in cold start situations (i.e. new users or items).

For all customer-oriented product sales and service system backgrounds, there is an increasing amount of data in enterprises. How to quickly transform business data into understanding of the market and operational conditions, thereby assisting enterprise decision-making, continuously optimizing decision-making management processes, and enhancing responsiveness to market changes has become an urgent problem that sales departments need to solve, A content guided product service recommendation system is an important tool for improving sales performance and is always worth continuous research and improvement[5].

2. Previous technical references and current major issues

Content based recommendation systems often erroneously recommend incorrect or biased guidance information to users or objects in need of information, while also having shortcomings in efficiency and the scope and speed of information content collection[6]. Feature extraction is difficult in the recommendation calculation process, making it difficult to discover other potential interests of users and lacking diversity.

The purpose of a recommendation system is to search for the most likely items of interest to users among all items (products, content, etc.) and recommend them to them. Like online retail websites, they will spend a lot of effort building recommendation systems to achieve more accurate recommendations and increase item sales[7]. From a data source perspective, it is generally obtained through user feedback, such as ratings after watching movies, search

ISSN: 1813-4890

content, purchase history, etc., or information about users and products' likes and dislikes obtained from other sources

3. Recommendations for countermeasures

Improve the practicality of guessing in different ways by combining big data and other materials with limited information to provide customers or information seekers with the fastest and most accurate internal recommendations, and more accurately locate the demand segmentation in similar content, achieving more accurate and fast implementation of recommendation functions.

3.1. Basic content-based recommendation algorithms

Content based recommendation algorithms are prone to the phenomenon of "homogenization", where recommendation results are too similar and lack diversity[8].

3.2. Nearest Neighbor Classification Algorithm

The nearest neighbor based classification algorithm is easy to fall into the "overfitting" state, that is, it performs well on the training set, but poorly on the test set. This is because content-based recommendation algorithms only consider the characteristics of the item, without considering user behavior[9].

3.3. Algorithm Based on Correlation Feedback

Rocchio algorithm is a well-known algorithm in the field of information retrieval, mainly used to solve relevance feedback (RF) problems[10]. When using the Rocchio algorithm to construct a user profile vector, it is usually assumed that the vector has the highest correlation with the features of the item that the user likes and the lowest correlation with the features of the item that the user does not like. Once the algorithm is given an error at a key selection point, it will lead to a completely opposite result of the customer's wishes appearing foolish and comical[11].

3.4. Decision Tree Based Recommendation

Content based recommendation algorithms need to extract and update the features of items, so it is necessary to continuously maintain and update the attribute information of items. For some large-scale recommendation systems, the cost may be higher

3.5. Naive Bayesian classification

3.6. Content Recommendation Algorithm Based on Linear Classification

Generally, gradient descent or least squares methods can be used to find the optimal parameters of a linear model. For candidate movies, determine whether the features of the movie meet the conditions, and then sort and recommend based on the classification results. The disadvantage is that in reality, the actual data belongs to the logical inertia of true data rather than arrays[12].

3.7. Recommended Text Representation Based on Unstructured Content and Recommended Non Text Representation of Unstructured Content, such as Images and Videos[13].

These data are difficult to represent using table structures in the database. Compared with structured data, unstructured data is irregular and fuzzy, which makes it difficult for computers to understand[14]. Although unstructured data has the disadvantages of complex structure, non-standard and high processing threshold, the high data stock and rich connotation information determine that unstructured data is a treasure to be explored by the recommended system[15]. All kinds of unstructured data have their own unique representation methods, but the processing ideas are interlinked[16].

ISSN: 1813-4890

3.8. The Best Creative Method Familiarity Coefficient Method

The data acquisition process in this method includes two aspects: firstly, the user actively sends information recommendation requests, and secondly, the recommendation system actively recommends information to the user[17]. In the process of users actively sending information recommendations, users send HTTP requests to obtain corresponding responses to match the information they need to obtain and send it to the recommendation system. The recommendation system obtains the corresponding matching content through queries and pushes it to the user[18]. In this mode, the data acquisition process is called active data request. In this process, the analysis of the data acquisition The process of processing and filtering is the same as the data analysis, processing, and filtering process carried out by the recommendation system when actively recommending information to users, which will be explained in detail in the following text. In the process of data acquisition, multiple methods can be used. In this article, web crawler technology is adopted, which downloads behavioral attribute data that users have browsed in search engines such as web pages[19]. The behavior here is not specific to the user's actions, but rather the user's browsing behavior left behind when browsing web pages, apps, social circles, or Weibo. The system framework of web crawlers includes three parts: control module, analysis module, and database. The control module is responsible for assigning work to various crawler threads of multiple threads. It is believed that the analyzer downloads web pages and processes them, such as processing JS script tags, CSS code content, space characters, HTML tags, etc., and then stores the downloaded web resources through the database. The data acquisition device in this article includes the system of the web crawler to obtain user behavior attribute data for one or more clients. After obtaining behavioral attribute data, it is necessary to analyze and process it. As mentioned earlier, user behavioral attribute data includes user basic information set, user interest information set, and user feedback information set[20].

In this article, based on this reason, the user behavior attribute data is divided. When users browse the internet and obtain relevant network information[21], there are several possible situations: (1) there is no purpose, only a randomness to obtain data. In the process of randomly obtaining data, they often suddenly obtain content of interest, which may also have a certain purpose. However, the purpose is relatively vague, which is called approximate search, In this case, we call it the user basic information set, which has randomness and the extraction of keywords is difficult and scattered; (2) Information acquisition with strong purpose is relatively easy to extract keywords or calculate similarity, and the concentration of information that users pay attention to is relatively high, which is referred to as the user interest information set in this article. (3) In existing recommendation systems, recommendation systems have information push, but do not process the feedback content. That is to say, although it is recommended, is it appropriate, Whether it is useful or not does not result in interaction with users. In this article, a set of user feedback information has been set up. The above is the three types of information sets mentioned in this article. When receiving one or more of these information sets, corresponding labels need to be set accordingly to process these information in a targeted manner. The data analysis device receives the behavioral attribute data, analyzes the attribute data, adds corresponding content labels, and sends the behavioral attribute data corresponding to the content label to the knowledge base content storage device. The process is as follows: (1) After receiving the behavioral attribute data, it determines which type the behavioral attribute data belongs to, If it is a user basic information set or a user interest information set, proceed to step (12). If it is a user feedback information set, add a feedback content label to the user feedback information set and proceed to step (14); (12) After obtaining the user basic information set or user interest information set, extract the user attribute information from the user basic information set or user interest information set; (13) The data acquisition device analyzes the user attribute information and retrieves personal

credit evaluation data from the knowledge base content repository to determine whether the personal credit data of one or more users obtained is within a reasonable range. If it is within a reasonable range, the personal credit warning content label is not added, otherwise the personal credit data label of the user is added to the user behavior attribute data; (14) Send behavioral attribute data consisting of one or more user basic information sets, user interest information sets, or user feedback information sets with corresponding content labels to the knowledge base storage device[22]. The knowledge base content storage device includes multiple knowledge base content storage databases identified by three corresponding content labels, including a user basic content storage database corresponding to the user basic information set content label, a user interest content storage database corresponding to the user interest information set content label, and a user feedback information content storage database corresponding to the user feedback information set content label[23]. The user basic content storage database corresponding to the content label of the user basic information set is used to store the user's initial content. The user basic information is a randomly generated basic information set when the user browses a webpage, app client, or friend circle. By randomly selecting keywords, the corresponding content is obtained from the knowledge pool, and recommended to the customer through the server recommendation device.

4. Conclusion

The present invention proposes a knowledge base recommendation system based on content tags, which recommends recommended content to users[24]. The recommendation system includes a server; And a client device that establishes a communication link with the server through a network; The server includes a data acquisition device to obtain behavioral attribute data of one or more users, wherein the behavioral attribute data includes a user basic information set, a user interest information set, and a user feedback information set[25]; A data analysis device that receives the behavioral attribute data, analyzes the attribute data, adds corresponding content labels, and sends the behavioral attribute data corresponding to the content labels to the knowledge base content storage device; A knowledge base content storage device, including multiple knowledge base content storage databases and knowledge pools, wherein the knowledge base content storage device receives the behavioral attribute data and extracts or stores corresponding behavioral attribute data in the knowledge pool based on the content label; The server recommendation device pushes the content that needs to be pushed to the user. The recommendation system of the present invention improves the effectiveness and accuracy of system data push.

Acknowledgements

Thank you to the school and the country for their nurturing efforts. Thank you to the professors of Odessa National University in Ukraine, Professor Eugene V. Malakhov and Professor Oleksandr Antonenko.

References

[1] CASTRO M, LISKOV B. Practical Byzantine fault tolerance and proactive recovery[J]. ACM Transactions on Computer Systems, 2002, 20(4): 398-461.

[2] TAO F, ZHANG L, VENKATESH V C, et al. Cloud manufacturing: acomputing and service-oriented manufacturing model[J]. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engi-neering Manufacture, 2011, 225(10): 1969-1976.

[3] TAO F, LAILI Y J, XU L D, et al. FC-PACO-RM: a parallel method for service composition optimal-selection in cloud manufacturing system[J]. IEEE Transactions on Industrial Informatics, 2013, 9(4): 2023-2033.

[4] ZHANG G, ZHANG Y F, XU X, et al. An augmented Lagrangian coordination method for optimal allocation of cloud manufacturing services[J]. Journal of Manufacturing Systems, 2018, 48: 122-133.

[5] LIU Y K, WANG L H, WANG X V, et al. Scheduling in cloud manu-facturing: state-of-the-art and research challenges[J]. International Journal of Production Research, 2019, 57(15/16): 4854-4879.

[6] MOURAD M H, NASSEHI A, SCHAEFER D, et al. Assessment of interoperability in cloud manufacturing[J]. Robotics and Computer-Integrated Manufacturing, 2020, 61: 101832.

[7] SUKHWANI H, MARTÍNEZ J M, CHANG X L, et al. Performance modeling of PBFT consensus process for permissioned blockchain network (hyperledger fabric)[C]//Proceedings of 2017 IEEE 36th Symposium on Reliable Distributed Systems. Piscataway: IEEE Press, 2017: 253-255.

[8] Nakamoto, S.Bitcoin.A peer-to-peer electronic cash system[J], Consulted,2008(1):2645-2652.

[9] L. Lamport, R. Shostak, M. Pease. The Byzantine generals problem[M]. New York: ACM book, 2019, 203-226.

[10] LAMPORT L.Paxos made simple[J].ACM SIGACT News,2001,21(4):51-58.

[11] ONGARO D, OUSTERHOUTJ K.In search of an understandable consensus a lgorithm [C]//USENIX Annua l Technica l Conference.2014:305-319.

[12] Han R, Foutris N, Kotselidis C. Demystifying CryptoMining: Analysis and Optimizations of memory-hard PoW Algorithms [C]. 2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Boston, 2019: 22-33.

[13] Bentov I , Lee C , Mizrahi A , et al. Proof of Activity: Extending Bitcoin's Proof of Work via Proof of Stake[J]. 2014.42(3):34-37.

[14] Miguel Castro, Barbara Liskov. Practical Byzantine fault tolerance[P]. Operating systems design and implementation,1999.

[15] Ren Zhongwen. Blockchain: A Reader for Leaders [M]. Beijing: People's Daily Publishing House, 2018:3-5.

[16] Wang Chongyu. Blockchain Technology and Its Value Outlook [J]. Economic Dynamics and Reviews, 2018, (2): 149-182.

[17] Tang Wenjian. How Blockchain Will Redefine the World [M]. Beijing: Mechanical Industry Press, 2016:81.

[18] Han Feng. Blockchain from Digital Currency to Credit Society [M]. Beijing: CITIC Publishing House, 2018:260.

[19] Xiong Jiankun. The Rise of Blockchain Technology and a New Revolution in Governance [J]. Journal of Harbin Institute of Technology: Social Sciences Edition, 2018, (5): 15-18.

[20] Gao Hongye. Microeconomics [M]. Beijing: Renmin University of China Press, 2018.

[21] [English] Adam Smith. National Wealth Theory [M]. Chongqing: Chongqing Publishing House, 2015:290.

[22] Lin Yifu. Interpreting the Chinese Economy [M]. Beijing: Peking University Press, 2016.

[23] Jia Kai. Research on Blockchain Governance: Technology, Mechanism, and Policy [J]. Administrative Forum, 2019, (2): 80-85.

[24] Liu Yibo. Under the Shadow of Artificial Intelligence: Ethical Dilemmas in Government Big Data Governance [J]. Administrative Forum, 2018, (3): 97-103. [18]

[25] Qi Xuexiang. Application of Blockchain Technology in Government Data Governance: Advantages, Challenges, and Countermeasures [J]. Journal of Beijing University of Technology: Social Sciences Edition, 2018, (5): 105-111