Proceedings of the Second International Conference on Smart Electronics and Communication (ICOSEC).
IEEE Xplore Part Number: CFP21V90-ART; ISBN: 978-1-6654-3368-6

2021 2nd International Conference on Smart Electronics and Communication (ICOSEC) | 978-1-6654-3368-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOSEC51865.2021.9591853

Application of Web Data Mining in Large Scale E-Commerce Platform

Qi Liu
Guangdong University of Science&Technology, Dongguan Guangdong, 523083, China
liuqidongguan@gmail.com

***Abstract*— With the rapid development of mobile Internet technology and social network, social media has gradually integrated into people's daily life, and its function has gradually evolved from social entertainment to Internet based application. The scale and frequency of social online shopping users are increasing, and the application of new technologies and models has driven the diversification of e-commerce formats, which makes the rapid development of social e-commerce. In this paper, a k-means algorithm based on density standard deviation is proposed_ Means improved algorithm. In this method, the average and standard deviation of the samples are calculated firstly, then the density distribution function of each data point is calculated, and finally the average density and standard deviation of the samples are calculated.**

Keywords*— *Web Data Mining, K-Means Algorithm, PageRank Algorithm, Python, Pycharm

I. INTRODUCTION

In the United States, nearly 80% of adults use social networks, 23% of those who often surf the Internet spend their time on social networks and blogs, and 70% of them have online shopping behavior. In China, according to the report data of CNNIC, the scale of social e-commerce market in 2014 was 96 billion yuan, and the scale of merchants was 9.16 million. It is estimated that the scale of social e-commerce users in China will reach 24 million in 2020, and the market scale will exceed one trillion yuan. At the end of 2016, the Ministry of Commerce, the central network information office and the national development and Reform Commission jointly issued the 13th five year development plan for e-commerce, which clearly proposes to encourage the development of social e-commerce. The plan clearly proposes to encourage social networks to give full play to the advantages of content, creativity and user relations. Establish the operation mode of linked e-commerce, support the healthy and standardized development mode of micro commerce, provide personalized e-commerce services for consumers, and stimulate the continuous growth of online consumption [1-4]. Encourage and standardize social network marketing innovation. E-commerce enterprises are encouraged to rely on the emerging video, streaming media, live broadcast and other diversified marketing methods to carry out fan interaction, truthfully transmit commodity information, and establish a healthy and harmonious social network marketing mode. According to the statistics of the Ministry of Commerce, in 2016, the scale of China's online retail market was 4.7 trillion yuan, and the scale of micro business market was 360 billion yuan, accounting for 7.7% of the total scale of online retail transactions. In 2020, China's online retail market is expected to reach 9.6 trillion yuan, and the social e-commerce market is expected to reach 3 trillion yuan, accounting for 30% of the online retail transaction scale, with huge market space [5-8].

Social e-commerce has the advantages of both traditional e-commerce and social networking sites. It is a new business model formed by the integration of e-commerce and Web2.0. Under the traditional e-commerce mode, users obtain the relevant information and path of purchasing products through active search. Under the mode of social e-commerce, social e-commerce takes advantage of people's habit of trusting acquaintances' shopping evaluation in social life to accurately locate the user group. The use of social group word-of-mouth, a high degree of recognition and loyalty of users, and then get a high conversion rate and high repurchase rate. In this situation mode, the display and transaction promotion of goods is based on the social information dissemination and interaction of the user's social relationship chain, and the information released by each person has different degrees of guidance to other users in the social circle. Social e-commerce combines user interaction with business transactions, which makes shopping behavior penetrate into fragmented social scenes. Search for user information and get more recommendations from friends in social circle or the same interest groups. Through user social interaction, user generated content and other means to assist the purchase and sale of goods, most of the content and behavior of the platform are generated and dominated by end users. Users are the core of the social e-commerce platform and the key subject to determine the development of the platform and the profit of the enterprise. In the context of social e-commerce, users' behavior of publishing, obtaining and using platform information has also changed greatly [9-12].

The world wide web has many characteristics, which makes mining useful information and data more challenging.

(1) The data on the world wide web is huge and increasing. The data subject is very wide, the content of data is rich and colorful, and users can find almost any information needed on the world wide web.

(2) There are various types of data on the world wide web. These data include structured tables and semi-structured web pages, as well as unstructured text and multimedia files, which refer to pictures, audio and video.

(3) The information on the world wide web is heterogeneous. Different web designers may design the same page with different text and format, which makes it a new challenge to integrate the information and data of multiple pages.

(4) The vast majority of information on the world wide web is connected through hyperlinks, and the web pages inside and between websites are also linked through hyperlinks.

(5) The world wide web contains noise information. There are two reasons for noise generation. One is that only part of the information contained in the web page is useful to specific users, and most of the useless information is noise; Second, anyone can make any comments on the world wide web, which leads to a lot of low-quality and useless information on the web page.

(6) The World Wide Web will also provide business services, such as goods purchase, bill payment, etc.

(7) The world wide web is highly dynamic. The information on the world wide web is constantly updated and changing.

(8) The world wide web is also a virtual society. It not only provides data services and information services, but also provides interaction between people and organizations and automation systems.

II. The Proposed Methodology

A. Analysis of Web Data Mining Algorithm

Web data objects include content data, structure data and log data. Content data is the data based on page, which mainly includes structured and unstructured multimedia data and text data. Multimedia data refers to pictures, audio and video. Structural data is the data that records the internal structure of a page and the organizational structure between pages, mainly including hyperlinks, HTML, XML, etc. Log data is the interactive data generated by users accessing the page, mainly including web server log, browser side log, application server log and user access mode.

Web structure mining is a process of mining useful information from hyperlinks which represent web structure. The hyperlink information of web structure includes: XML tree structure inside web page, hyperlink structure between web pages and directory path structure in web address. It includes not only the relevance of the page content, the quality and structure of the page, but also the reference relationship, inclusion relationship and subordination relationship of the page. Web structure mining finds potential knowledge by analyzing the organization, reference and link relationship between web pages. Web structure mining plays an important role in website optimization. It can not only analyze the quality of web pages, optimize the link design of web pages, but also optimize the search engine.

Because of the interconnection of web pages, the world wide web can provide structural information in addition to the content of the page itself. After obtaining this information, you can sort the pages to find important pages. When page a has a hyperlink to page B, it means that page B is very important to page a. When a large number of hyperlinks point to the same web page, it indicates that the web page is particularly important and is usually considered as an authoritative page, as shown in Figure 1.

$$X = \{x_1, x_2, ..., x_n\} \tag{1}$$

$$Dist(x, y) = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2} \tag{2}$$

$$SSE = \sum_{k=1}^{k} Dist(x, m)^2 \tag{3}$$

$$P_D(i) = \frac{d_I(i)}{n-1} \tag{4}$$

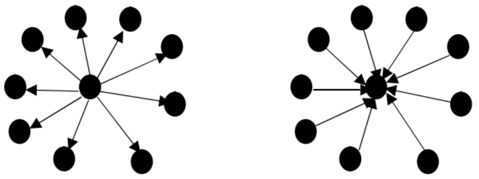


Fig. 1. Authority page and center page

B. Connotation of User Information Adoption in Social E-Commerce

The evaluation and recommendation information of acquaintances in social e-commerce has high credibility, so it can better meet the needs of users who want to eliminate the information asymmetry in the shopping process. Combined with the characteristics of social e-commerce, this study considers that information adoption of social e-commerce users is a complex dynamic process connecting information needs and search, information selection and evaluation, information absorption and utilization. Information adoption is the premise to change the attitude and behavior of social e-commerce users, which is related to the success or failure of business transactions.

According to Martin kiduff, there are three sources of ideas in the field of social science. The first aspect is that German researchers, such as Kurt Lewin, are inspired by the physical field theory and apply the concept of network to study social interaction. Moreno, an American psychologist, and Lu Yan, a social psychologist, believe that field theory can be used to explore the interdependence between group and environment in relational system. The second aspect is to use mathematical methods to study social interaction. In the United States, many graph theorists, such as Cartwright, use mathematical methods to study social interaction and push social network from descriptive analysis to analytical research. The third aspect comes from anthropological methods. In the 1920s, Harvard Business School took Western Electric Company as the research object and carried out the famous Hawthorne experiment. This is the first time that social network graph has been used for research.

Larson's research shows that the loyalty of social network members can be maintained by strong relationship, and strong relationship has a strong effect on the promotion of inter subject trust. If social network members want to get more resources and maintain long-term contact with social users, they must establish strong relationships. Strong relationship is an intimate and special relationship, which is not only a relationship of voluntary investment, but also a relationship of mutual understanding and support. They can establish an emotional connection, enhance the sense of belonging and highlight personal value. Strong relationship is established through long-term cooperation among network members. Strong relationship is a very important and directly accessible source of information and an important driving force of social integration.

C. *Analysis of Information Flow in Social E-commerce*

The information flow of social e-commerce is a process of information transmission from one information subject to another based on the relationship of social network, business transaction and subject cooperation. The information flow of social e-commerce mainly includes four types: information flow between platform and user, information flow between business and user, information flow between business and platform and information flow between users. From the micro perspective, the information flow of social e-commerce is the process of information exchange and transfer and experience sharing among users. From the macro perspective, the information flow of social e-commerce is the interaction process between users and the social e-commerce platform. Users will their own understanding, views, evaluation and doubts and other content published on the social e-commerce platform. The platform uses various functional modules to obtain and integrate information, and forms the information resource base of the social e-commerce platform. Users with information requirements retrieve information they need from the information resource library. The information flow of social e-commerce is the forerunner and foundation of commercial flow, logistics and capital flow, and is an important guarantee to promote positive feedback of supply information and negative feedback of demand information and two-way flow, so as to balance demand and supply, reduce the space-time distance and promote the realization of transactions.

Social e-commerce information dissemination depends on the relationship network formed by users' mutual concern. Information spreads in the social e-commerce network, and information users have different influences on the dissemination and diffusion of information. The social relationship of social e-commerce users is complex, their friends may come from different industries, have different occupations and different identities, but users tend to establish friend relationship with users who have similar interests to some extent. In the social network of social e-commerce users, the higher the interest similarity, the easier it is for users to establish relationships. The integration of information needs between friends makes them easier to become the object of some types of information exchange. Therefore, when information is transmitted, forwarded and shared, it will spread rapidly and widely to more people in the social network. Among them, some people will ignore the information, produce immunity to information, and prevent the viral transmission and spread of information.

The media of social e-commerce information flow are not only traditional devices, but also more widely used mobile communication devices, which facilitate the release and dissemination of user information. The form of information dissemination is usually one to many or many to many divergent, which is convenient for synchronous communication, real-time interaction and effective information acquisition. The information flow mode of social e-commerce real-time interaction can accelerate the speed of information flow and meet the users' real-time demand for information.

III. Experiment

The traditional K-means algorithm and the improved k-means algorithm based on density standard deviation are compared. In this paper, iris data set, wine data set and simulation experiment data set in UCI database are selected as test data set.

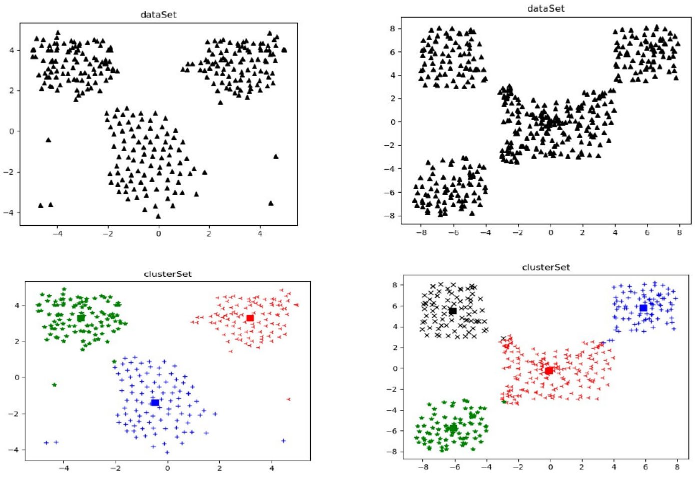


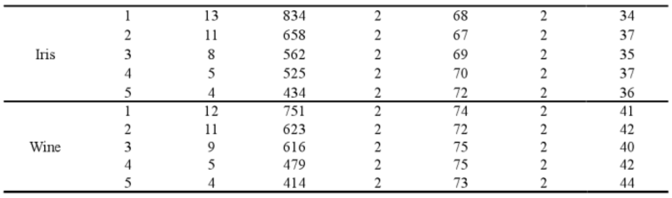
Fig. 2. Data analysis results

The test results of traditional K-means algorithm running 10 times are shown in Figure 3.

	1	2	3	4	5	6	7	8	9	10
Iris	88.6	88.6	89.3	88.6	88.6	64.0	88.6	89.3	89.3	88.6
Wine	70.1	53.4	70.1	53.4	70.1	53.4	53.4	70.1	70.1	53.4
Iris	9	12	8	10	11	8	7	7	5	13
Wine	5	7	6	13	10	8	12	4	6	12
Iris	47	57	39	50	55	41	36	38	26	63
Wine	49	64	55	114	89	74	105	40	54	108

Fig. 3. The traditional K-means algorithm runs for 10 times

The test results of the three algorithms on iris and wine datasets are as follows.



Iris	1	13	834	2	68	2	34
	2	11	658	2	67	2	37
	3	8	562	2	69	2	35
	4	5	525	2	70	2	37
	5	4	434	2	72	2	36
Wine	1	12	751	2	74	2	41
	2	11	623	2	72	2	42
	3	9	616	2	75	2	40
	4	5	479	2	75	2	42
	5	4	414	2	73	2	44

Fig. 4. Test results of three algorithms on iris and wine datasets

IV. Conclusion

Social e-commerce is the integration of e-commerce and social media, which focuses on information exchange and interaction between users. Based on the perspective of information ecology and information dissemination, this paper constructs the model of information flow of social e-commerce, further enriches and improves the theoretical research of information flow of social e-commerce, and lays the foundation for the follow-up analysis. Secondly, based on the perspective of information ecology and information dissemination, this paper constructs a model of information flow of social e-commerce from the elements, motivation and process of information flow of social e-commerce.

REFERENCES

[1] Cai Z, Lin J, Business S O. Trust-aware collaborative filtering recommendation method for social E-commerce[J].Journal of Computer Applications, 2015,35(1):167-171

[2] Yan X X, Hu Z Q, Xu J, et al. Research on the Social E-commerce Marketing Model Based on SICAS Model in China[J]. International Journal of Marketing Studies, 2017,9(3):113

[3] Iten L，Arnold K，Pistilli M． Mining Real－time Data to Improve Sdudent Success in a Gateway Course[EB/OL]．West Lafayette: Purdue University[2016-11-1] . http://www.purdue.edu/apps/facultyfocus/Newsletter/Details/17

[4] Bruno Zuga, Atis Kapenieks, Aleksandrs Gorbunovs, Merija Jirgensons, Janis Kapenieks, Janis Kapenieks, Ieva Vitolina, Guna Jakobsone-Snepste, Ieva Kudina, Kristaps Kapenieks, Zanis Timsans, Rudolfs Gulbis. Concept of Learner Behaviour Data Based Learning Support[J]. Procedia Computer Science,2015,43

[5] Wanli Xing, Rui Guo, Eva Petakovic, Sean Goggins. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory[J]. Computers in Human Behavior,2015,47

[6] P. J. Guo, J. Kim, and R. Rubin. How video production affects student engagement. In Proceedings of the first ACM conference on Learning@ scale conference, pages 41–50, New York, New York, USA, 2014. ACM Press

[7] Kim J, Li S. W, Cai C.J, Gajos K.Z.& Miller R. C.. Leveraging video interaction data and content analysis to improve video learning[C].[2016-03-06]//CHI2014:Learning Innovation at Scale Work-shop https://dash.harvard.edu/bitstream/handle/1/22719144/kim141everaging.pdf? sequence=1

[8] Crossley S, Paquette L, Dascalu M, et al. Combining click-stream data with NLP tools to better understand MOOC completion[C]// International Conference. 2016

[9] Larry, et al. ViaCa [DB/OL]. http://www. visualcatch. org/visca/web/home.jsp,2016-11-01

[10] Shi C, Fu S, Chen Q, et al. VisMOOC: Visualizing video clickstream data from Massive Open Online Courses[C]// Visual Analytics Science & Technology. 2014

[11] Lin J W , Huang H H , Chuang Y S . The impacts of network centrality and self-regulation on an e-learning environment with the support of social network awareness[J]. British Journal of Educational Technology, 2015, 46(1):32-44

[12] Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, et al. Modeling and Predicting Learning Behavior in MOOCs[C]// the Ninth ACM International Conference. ACM,2016

[13] Wu Guixian, Gao Xiangmin. Research on the application of Web Data Mining Technology in e-commerce [J]. Journal of Chengdu aviation vocational and technical college, 2018, V.34; No.114(01):41-42+50.

[14] Research on the application of Web Data Mining in e-commerce transactions [J]. Knowledge guide, 2019, 000 (022): 21-22

[15] Hou Limin, Yan Jianmin. On the application of Web Data Mining Technology in e-commerce [J]. Digital design.cg world, 2018, 007 (024): p.27-27

[16] Meng Qiang. Research on Web data mining model and application for e-commerce user behavior [D]. Heilongjiang University, 2017

[17] Gao Yu, Yan JUANJUAN, Sun Jian. Exploration on the combination of e-commerce management and web data mining technology [J]. Industry and Technology Forum, 2019, 018 (011): 57-58

[18] Zhu Yanqing. Research and implementation of e-commerce recommendation system based on data mining [D]. Hunan University, 2019

[19] Yang Junwei. Design and development of e-commerce system based on cloud computing technology [D]. North China Electric Power University, 2016

[20] Jiang Ning, Niu Yongjie. Application of Web Data Mining in E-commerce -- Taking Taobao as an example [J]. Computer age, 2016, 000 (007): 49-52

[21] Ji Shanshan. Research on the application of Web Data Mining Technology in Dongguan e-commerce [J]. Modern information technology, 2018, 002 (004): p.21-23

[22] Liu Jiaqi. Application of Web Data Mining in e-commerce commodity recommendation [J]. Fujian quality management, 2019, 000 (024): 104-106

[23] Zhang Chunzheng. Computer web data and its application in e-commerce [J]. Modern information technology, 2018, 002 (011): 133-134

[24] Wang Li. Computer web page data and its application in e-commerce [J]. Science public, 2018, 000 (010): p.12-12