Hu and Xu *EURASIP Journal on Wireless Communications and Networking* (2019) 2019:139
https://doi.org/10.1186/s13638-019-1441-1

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Research on real-time network data mining technology for big data

Jing Hu[1,2*] and Xianbin Xu[2]

Abstract

The data distribution in big data environment is very different, and it is difficult to mine the data because of the strong interference of redundant data and frequent items. The traditional data mining algorithm uses closed frequent item feature extraction algorithm. Due to the uneven distribution of web data in big data environment, the mining accuracy of closed frequent item feature extraction is not high. A real-time web data mining model is proposed based on high order spectral feature fuzzy neural network learning in big data environment. The transmission channel model and statistical time series model of web data under big data environment are constructed, and the redundant information flow is removed and reprocessed, and the web data after redundant filtering is analyzed by fusion clustering. The feature of high order spectrum is extracted, and the optimal mining of web data is realized by using fuzzy neural network learning classification method. The simulation results show that this web data mining method has good timeliness, high mining precision, and superior performance.

Keywords: Big data web data, Mining, Feature extraction

1 Introduction

With the development of computer network information technology, all kinds of large websites are constantly established and updated in real time. In the network society, large-scale websites are information browsing and data publishing through the way of establishing web. Web has become an important platform for people to interact and interact with each other. It is also an important module for storing and transmitting massive data information. In the big data environment, a large number of images, sounds, data, and text information are stored and displayed in the client in order to provide users with information sharing and use. In the big data environment, the distribution of the data is more complex and varied, so it is difficult to mine the web data. Moreover, in the network big data environment, the web data adopts the form of packet-switched network data communication, which usually results in the interference of redundant data. It leads to data offset and error in web data mining and access, and reduces the accurate mining and access probability of data [1]. It is necessary to study the optimization mining model of web data based on massive web model to improve the ability of accessing and managing web data. People pay great attention to the related algorithms.

Traditionally, web data mining models in large data environment mainly adopt high order cumulant feature extraction, time-frequency analysis and feature extraction, wavelet analysis, support vector machine classification mining algorithm, and data mining algorithm based on rough set classification in large data environment; there are many drawbacks in the web data model, such as the inaccuracy of data, the lack of effectiveness [2], the error resulting in the rule pattern of the system coding, and the other most important thing is that in the process of mining the data [3], it is not possible to determine whether the system is safe or not, if the data is excavated into an unsafe system, not only the data are data, but the data are not the same in this case; in reference [4], a feature data mining algorithm is proposed based on the distributed feature partition extraction of mass web access time, and improved the web data by multi-layer autoregressive vector analysis. Classification mining ability, but the computation cost of the algorithm is large, and the time delay error occurs in web information retrieval. In reference [5], a data mining and text retrieval method is proposed based on the large data

* Correspondence: caiduhuan3@163.com
[1]Wuhan Qingchuan University, Wuhan 430204, China
[2]Wuhan University, Wuhan 430072, China

environment based on the decision time classification search engine. By constructing the web search engine, the text feature extraction is realized and the number of semantic matching is improved by strict semantic matching. According to the convergence ability of the mining, the problem of the algorithm is that the accuracy of data mining is limited when the efficiency of the data attribute classification is not obvious or the interference data of the approximate web is large, but the algorithm has poor convergence and complex computation [6].

In order to solve the above problems, this paper presents a real-time web data mining model based on higher-order spectral feature fuzzy neural network learning in big data environment. In this paper, the transmission channel model and statistical time series model of web data under big data environment are constructed, and the redundant information flow is removed and reprocessed, and the web data after redundant filtering is analyzed by fusion clustering. On this basis, the feature of high order spectrum is extracted, and the optimal mining of web data is realized by using fuzzy neural network learning classification method. Finally, the performance test is carried out through the simulation experiment, which shows the superior performance of this method in improving the ability of data mining.

2 Data distribution model based on big data and anti-interference preprocessing of web data information flow

2.1 Data distribution model in big data environment

In order to realize the web data mining under the big data environment, it is necessary to analyze the web data in the big data environment first. In the big data environment, a large amount of data is stored in the deep web database for data cloud storage. Information browse and result display are realized on the web through web server [7]. The mining of web data under big data environment is realized by establishing database fusion and data clustering model, and data feature extraction and fusion clustering to realize real-time data mining. In big data environment, a large amount of data information is sorted and matched by similarity degree. Search engine and deep web database search engine and deep web database through data links and query results to carry out intelligent retrieval [8]. The overall model of web data mining under big data environment is shown in Fig. 1.

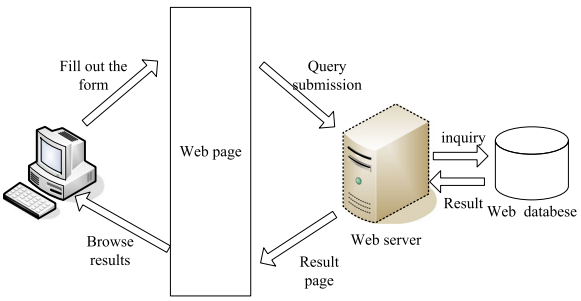

Fig. 1 Overall structure model of web data mining in big data environment

According to the above design of the overall structure model of web data mining in big data environment, the distribution of web data in big data environment is analyzed. It is assumed that the web data to be mined is distributed in the web database through the topic crawler method. Firstly, the attributes of the continuous data set of the information flow sequence under the big data environment are discretized, $A = \{a_1, a_2, \ldots, a_n\}$ is the initial vector of the information flow under the big data environment. The associated attribute set of web data in big data environment is $B = \{b_1, b_2, \ldots, b_m\}$. The mathematical model of the whole network database system is expressed as follows:

$$\dot{x} = f(x, u) \tag{1}$$

where u is the bit rate of data access, x is the initial access scalar time series, and the first-order vector group $x_1, x_2, \cdots x_n \in C^m$ (m-dimensional complex space) of the higher order cumulative vector of the data information is given, where:

$$u = [u_1, u_2, \cdots, u_N] \in R^{mN} \tag{2}$$

According to the global search ability of chaotic differential evolution algorithm, the optimal value of clustering center is found [9]. If there is no $G_3 = (M^{\alpha}{}_3, M^{\beta}{}_3, Y_3)$, the inherent mode function of big data state space is:

$$y(t) = \frac{1}{\pi} P \int \frac{x(\tau)}{t-\tau} d\tau = x(t) * \frac{1}{\pi t} \tag{3}$$

In order to reflect the diversity of the web data groups, the adaptive beamforming estimation of the web database information flow is carried out under the big data environment. The global convergence retrieval beam of the output is obtained as follows:

$$l(t) = \left(\sum_{m=1}^{M} u_m \right) \cos(2\pi f_0 t) - \left(\sum_{m-1}^{M} v_m \right) \sin(2\pi f_0 t) \tag{4}$$

By the above processing, the network crawler method is used to focus the data stream information in the high dimensional feature space of web to form a cluster center of web data access, which improves the ability of data feature mining and web data mining.

2.2 Web statistical time series model construction and anti-interference pretreatment

In order to improve the ability of accessing and mining web data in big data environment and combining modern data information processing algorithm to construct the information flow of massive data, because the web data is interfered by adjacent web data, it is necessary to carry out anti-interference suppression [10]. After processing, the empirical mode decomposition and Hilbert spectrum analysis of the web data information model under the big data environment are carried out, and the state transfer equation of the data information flow distribution in the process of web data access is obtained as follows:

$$x(n) = s(n) + v(n) = \omega_{k-1}^{(i)} \frac{p\left(y_k | X_k^{(i)}, Y_{k-1}\right) p\left(x_k^{(i)} | X_{k-1}^{(i)}, Y_{k-1}\right)}{q\left(x_k^{(i)} | .\right)} \tag{5}$$

In the above formula, $s(n)$ is the distributed time sampling sequence of web data under big data environment, $v(n)$ denotes interference component, and considering the phase difference of network differential characteristic behavior in the information source i of web node, the web number under big data environment is needed [11]. According to the information flow access process, the error square of delay estimation is:

$$\begin{aligned} \varepsilon^2(k) = {} & d^2(k) - 2d(k)X^T(k)W \\ & + W^T X(k) X^T(k) W \end{aligned} \tag{6}$$

Thus, the web statistical time series model under the large data environment is constructed. The interference suppression filtering of web data is designed, and the interference suppression of web data is suppressed by the multipath adaptive cascade filtering method [12]. It is assumed that the time-varying multipath correlation dimension of the web data in the large data environment is expressed as follows:

$$x(t) = \lambda \ \mathrm{Re}\left\{ a_n(t) e^{-j2\pi f_c \tau_n(t)} s_l(t - \tau_n(t)) e^{-j2\pi f_c t} \right\} \tag{7}$$

An adaptive cascade tracking filter is designed to suppress interference [13]. The system transmission function of the filter is obtained as follows:

$$H(z) = Am \cdot \frac{1 + 2z^{-1} + z^{-2}}{(1 - \rho e^{j\phi} z^{-1})(1 - \rho e^{-j\phi} z^{-1})} \tag{8}$$

By the above filtering process, assuming that the symbol width of web data in web is T_a, $T_a = 1/R_a$, the output amplitude of web data mining after interference suppression is:

$$a(t) = \sum_{n=0}^{\infty} a_n g_a(t - nT_a) \tag{9}$$

The big data is sampled by pseudorandom sequence by using binary phase shift keying (BPSK) modulation, and the sampling value is $c(t)$. Thus, the mining accuracy of web data under big data environment can be improved effectively.

3 Methods

3.1 Learning algorithm of higher order spectrum feature fuzzy neural network for web data in data environment

In this paper, a real-time data mining model based on higher-order spectral feature fuzzy neural network learning in big data environment is proposed. High-order spectral features are extracted from the web data in big data environment, and the clustering fusion analysis is carried out by using the adaptive learning method. The higher-order spectral features of the big data are extracted by the Agung Rai Museum of Art (ARMA) model [14]. Assuming the spectral width of the Doppler time slice is T_c, $T_c = 1/R_c$, then:

$$c(t) = \sum_{n=0}^{N-1} c_n g_c(t - nT_c) \tag{10}$$

Based on higher-order spectral feature extraction method, the time-delay and scale estimation of p-dimensional vector in web data mining model is carried out, and the following binary hypothesis testing problems are obtained:

$$\begin{cases} H_0 : r(t) = n(t) \\ H_1 : r(t) = g(t) + n(t) \end{cases} \quad t \in [0, T] \tag{11}$$

In the formula, $r(t)$ is the unilateral exponential distribution of fusion clustering, $g(t)$ is the center vector of data clustering, σ^2 is the color noise with zero mean value, and ϕ_{mi} is the variance. The analytical expression of phase deviation ϕ_{mi} of web data mining in web is obtained by taking mathematical expectation on both sides of the above formula:

$$\phi_{mi} = \frac{2\pi r_i}{\lambda} \left(\sqrt{1 + \frac{m^2 d^2}{r_i^2} - \frac{2md \sin\theta_i}{r_i}} - 1 \right) \tag{12}$$

Table 1 Training data sets

Training set	Size
Web mode 1	4345
Web mode 2	2435
Web mode 3	1344
Web mode 4	3532

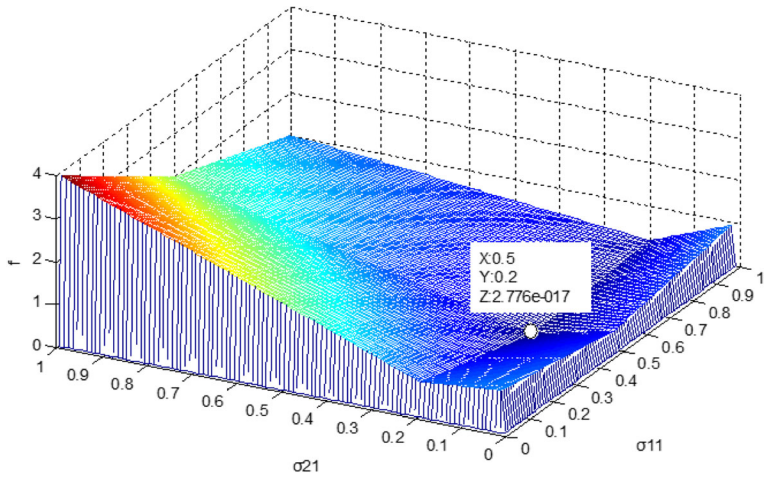

Fig. 2 Data mining output. Figure shows that this method can accurately mine the web data under the big data environment, and the feature expression ability of the data mining output is strong

Based on the fuzzy neural network learning algorithm, the error tracking and fitting of web data mining is carried out, and the fitting state function is:

$$p(Q_s) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left[-\frac{(Q_s - \langle Q_s \rangle)^2}{2{\sigma_s}^2}\right] \tag{13}$$

$$\int_{-\infty}^{\infty} p(Q_s) dQ_s = 1 \tag{14}$$

Where $\{\eta_i\}$ is an independent and uniformly distributed fuzzy neural network learning tracker with a mean value of 0 and a variance of σ^2. The error of data mining is reduced and the mining accuracy is improved by the learning of the fuzzy neural network [15].

3.2 Implementation of data real-time mining model

On the basis of the data mining fusion clustering analysis and fuzzy neural network learning, the improved design of data mining model is carried out. The sampling interval of data mining is assumed to be $n \in [n_1, n_2]$, and the data mining is studied by fuzzy neural network [16]. The phase characteristics of the web data are described as follows [17]:

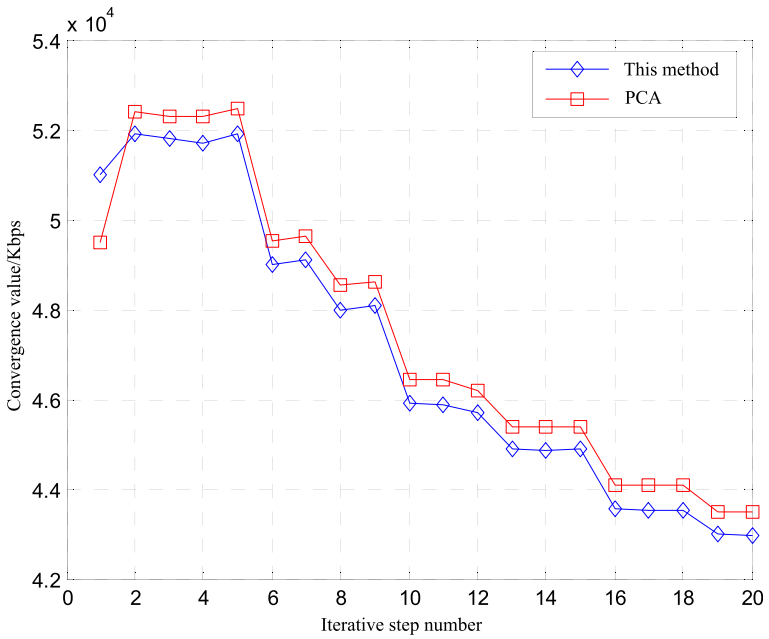

Fig. 3 Comparison of convergence curves of data mining. Figure shows the convergence of different methods for web data mining is tested

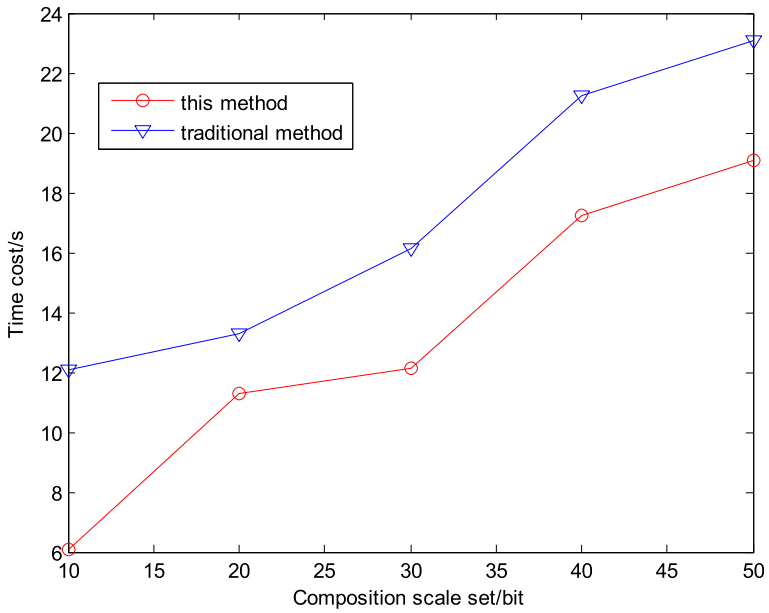

Fig. 4 Comparison of time cost for data mining

$$s(v) = \int_0^v \sin\left(\frac{\pi}{2}x^2\right)dx, \quad c(v) = \int_0^v \cos\left(\frac{\pi}{2}x^2\right)dx \tag{15}$$

In big data environment, the probability of any web data mining can be expressed as:

$$x(t) = R\left(a(t)e^{i\theta(t)}\right) = a(t)\ \cos\theta(t) \tag{16}$$

Based on decision tree classification, the semantic similarity information attributes of web data are extracted, and the feature classification functions of data mining under big data environment are obtained:

$$v_1 = \sqrt{BT}\frac{1+2(f-f_0)/B}{\sqrt{2}} \tag{17}$$

$$v_2 = \sqrt{BT}\frac{1-2(f-f_0)/B}{\sqrt{2}} \tag{18}$$

Based on the above processing, the high order spectrum feature of web data is studied by fuzzy neural network under the environment of big data, which improves the confidence and accuracy of data mining, reduces the false alarm probability, and realizes the real time mining and accurate mining of web data.

4 Experience

In order to test the application performance of this algorithm in the implementation of web data mining in big data environment, the simulation experiment is carried out. Based on Matlab platform, the simulation experiment is carried out. The computer simulation experiment platform is configured as Intel: core i5 processor, the main frequency is 2.8 GHz/4G memory and Windows 10 professional edition 32 Bit SP2 operating system. The test data is a deep web database under the big data environment of a large website. CWT200G data combination mode is used to start visa resource manager for data loading. More than 200,000 big data information in big data environment are obtained. The data collected are 16-bit vertical accuracy. The massive data were divided into training set and test set, assuming that the interference intensity in data mining was − 15 dB Gaussian color noise. The simulated dataset consists of two partitions of 25.2 MB in size, and the size distribution of the dataset for training and testing is shown in Table 1.

According to the above simulation environment and parameter setting, the data mining under the big data environment is carried out, and the output of the data mining is shown in Fig. 2.

Table 2 Test data set

Test data	Size大小
Web mode 1	2452
Web mode 2	6433
Web mode 3	3532
Web mode 4	1344

Figure 2 shows that this method can accurately mine the web data under the big data environment, and the feature expression ability of the data mining output is strong. The convergence of different methods for web data mining is tested, and the comparison results are shown in Fig. 3. The execution time comparison of data mining is shown in Fig. 4. The analysis shows that the proposed method has better convergence, shorter execution time, and better real-time performance (Table 2).

5 Results and discussion

In this paper, a real-time web data mining model is proposed based on high order spectral feature fuzzy neural network learning in big data environment. The transmission channel model and statistical time series model of web data under big data environment are constructed, and the redundant information flow is removed and reprocessed, and the web data after redundant filtering is analyzed by fusion clustering. The feature of high order spectrum is extracted, and the optimal mining of Web data is realized by using fuzzy neural network learning classification method. The simulation results show that this web data mining method has good timeliness, high mining precision, and superior performance. This method has good application value in real-time data mining and feature extraction of web data.

Abbreviations

ARMA: Agung Rai Museum of Art; BPSK: Binary phase shift keying

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

JH prepared the first draft of the full text and adjusted the structure of the full text. XX has written the model formula part of the article and simulated the experimental part of the article. Both authors read and approved the final manuscript.

Authors' information

Jing Hu, Master of computer application technology, Associate Professor. Graduated from the Wuhan University in 2009. Worked in Wuhan Qingchuan University. Doctoral candidate of computer system structure in Wuhan University. Her research interests include image processing, pattern recognition, embedded system, and high performance computing. Xianbin Xu, Doctor of computer software and theory, Professor. Graduate as an undergraduate in e Huazhong University of Science and Technology in 1977, and got PhD degree at Wuhan University majoring in computer software and theory. Worked in Wuhan University. His research interests include high performance computing and massive information storage.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 6 January 2019 Accepted: 18 April 2019
Published online: 28 May 2019

References

1. G. Kirchner, F. Koidl, F. Friederich, et al., Laser measurements to space debris from Graz SLR station[J]. Adv. Space Res. **51**(1), 21–24 (2013)
2. S.J. Xue, W.L. Shi, X.L. Xu, A heuristic scheduling algorithm based on PSO in the cloud computing environment [J]. Int. J. u-and e-Serv., Sci. Technol. **9**(1), 349–362 (2016)
3. W. Pao, W. Lou, Y. Chen, et al., Resource allocation for multiple input multiple output-orthogonal frequency division multiplexing-based space division multiple access systems [J]. IET. Commun. **8**(18), 3424–3434 (2014)
4. B. ORM, S.M. Senouci, M. Feham, A novel secure aggregation scheme for wireless sensor networks using stateful public key cryptography [J]. Ad Hoc Netw. **32**(C), 98–113 (2015)
5. S. Chen, G. Wang, W. Jia, Cluster-group based trusted computing for mobile social networks using implicit social behavioral graph [J]. Futur. Gener. Comput. Syst. **55**, 391–400 (2016)
6. Y. Zhang Q, C. Wang R, C. Sha, et al., Node correlation clustering algorithm for wireless multimedia sensor networks based on overlapped FoVs [J]. J. China Univ. Posts and Telecommun. **20**(5), 37–44 (2013)
7. C. Yong-jun, Z. Yong-hua, Linux system dual threshold scheduling algorithm based on characteristic scale equilibrium[J]. Comput. Sci. **42**(6), 181–184 (2015)
8. W. Zhi-jun, P. Bao-song, The detection of LDoS attack based on the model of small signal. Chin. J. Electron. **39**(6), 1456–1460 (2011)
9. W. Jie, L. Jianzhu, Z. Xiaofei, Data aggregation scheme for wireless sensor network to timely determine compromised nodes[J]. J. Comput. Appl. **36**(9), 2432–2437 (2016)
10. J. Gubbi, R. Buyya, S. Marusic, et al., Internet of things (IoT): a vision, architectural elements, and future directions [J]. Futur. Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
11. M. Hayman, J.P. Thayer, General description of polarization in lidar using stokes vectors and polar decomposition of Mueller matrices[J]. JOSA A **29**(4), 400–409 (2012)
12. Z. Liu, Y. Yuan, X. Guan, et al., An approach of distributed joint optimization for cluster-based wireless sensor networks [J]. IEEE/CAA Journal of Automatica Sinica **2**(3), 267–273 (2015)
13. Y.-l. Cao, X.-m. Wang, Z.-b. He, Optimal security strategy for malware propagation in mobile wireless sensor networks[J]. Acta Electron. Sin. **44**(8), 1851–1857 (2016)
14. Y. Tan Q, H. Leung, Y. Song, et al., Multipath ghost suppression for through-the-wall-radar[J]. IEEE Trans. Aerosp. Electron. Syst. **50**(3), 2284–2292 (2014)
15. Y. Xu, S. Tong, Y. Li, Prescribed performance fuzzy adaptive fault-tolerant control of non-linear systems with actuator faults[J]. IET Control Theory Appl. **8**(6), 420–431 (2014)
16. G. Gennarelli, F. Soldovieri, Multipath ghosts in radar imaging: physical insight and mitigation strategies[J]. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **8**(3), 1078–1086 (2014)
17. X. Huang, Z. Wang, Y. Li, et al., Design of fuzzy state feedback controller for robust stabilization of uncertain fractional-order chaotic systems[J]. J. Franklin Inst. **351**(12), 5480–5493 (2015)