APPLICATION OF MACHINE LEARNING IN INTELLIGENT KNOWLEDGE SEARCH ALGORITHM

H.M. LI†, Q.P. ZHANG† and Z.L. ZHENG‡

† *School of economics and management, Harbin Institute of Technology，Harbin, China*
‡ *School of business，Suzhou University of Science and Technology, Suzhou, China*
morrisonsOU@yahoo.com

***Abstract*— With the rapid development of Internet technology, the scale of knowledge and data production continues to expand. Such huge data has brought convenience to our lives and has brought challenges to our research. In order to obtain Internet data quickly and accurately, this paper uses machine learning algorithms to build a multidimensional knowledge search model. The model adds a real-time learning mechanism to acquire various knowledge, so that the system has intelligent features such as self-learning and self-adaptation. Through the comparative analysis of traditional knowledge search algorithms, the results show that the intelligent knowledge search algorithm based on machine learning has obvious advantages in efficiency and accuracy, which greatly shortens the search time of users.**

***Keywords*— machine learning, intelligent retrieval, knowledge acquisition, Unsupervised Learning.**

I. INTRODUCTION

The rapid development of Internet technology is accompanied by the production. According to statistics, in the past three decades, the total amount of human-generated data has exceeded the sum of all previous data in history, and the trend of such data blowouts continues to occur. In this way, a problem of knowledge search has arisen (Huang *et al.*, 2014) How to quickly and efficiently retrieve the information resources that people really need from the massive data on the Internet has become an urgent problem in the information society (Zhao *et al.*, 2016). The data acquisition and intelligent sensing network plays an important role in a robot teleoperation system. It is composed of sensing, controlling, computing, information processing, network communication, artificial intelligence and many other technologies (Lei *et al.*, 2016). Compared with the traditional method, great changes have taken place in the design and production of intelligent sensors and actuators. Its future development direction is networking with intelligent interfaces (Narudin *et al.*, 2016). The article introduces the theory and application of field bus-based data acquisition and intelligent sensing technology in robot teleoperation system (Narudin *et al.*, 2016). It also discusses several key problems and corresponding solutions in this issue. LambdaMART is the boosted tree version of LambdaRank, which is based on RankNet (Allix e*t al.*, 2016). RankNet, LambdaRank, and LambdaMART have proven to be very successful algorithms for solving real world ranking problems: for example, an ensemble of LambdaMART rankers won Track 1 of the 2010 Yahoo! Learning To Rank Challenge (Kulkarni *et al.*, 2016). The details of these algorithms are spread across several papers and re-ports, and so here we give a self-contained, detailed and complete description of them (Bougoudis *et al.*, 2016).

II. STATE OF THE ART

Knowledge and data on the Internet can be divided into the following categories according to their types and uses: (1) Text-based data: various online newspapers, electronic editions of electronic periodicals or printed journals exclusively distributed on the Internet, academic conference papers directly published on the Internet, and online colleges of various major colleges and universities (Gao *et al.*, 2017). The full texts of publications, standards, etc. of various government agencies. (2) Factual information: introduction of regions, cities or scenic spots, project facts and records, list of enterprises and institutions, guidelines, dictionary, dictionary, encyclopedia, yearbook, handbook, product samples and other reference tools, program previews, and weather forecasts. (3) Numerical information: specifications and prices of products or commodities, various statistical data, and so on (Taherkhani *et al.*, 2016). (4) Database information: It is often a secondary document that rearranges the bibliographical documents and is a networked information of various traditional databases. Database retrieval systems have a wide range of information (Soueidan and Nikolski, 2016). For example, the DIALOG system, which has long been networked, contains more than 500 document databases covering almost everything in business, technology, humanities, and social sciences (Zhou *et al*, 2016). The domestic professional service organization CNKI Data Network Center provides a document retrieval system on the Internet that also allows users to quickly and easily use dozens of valuable Chinese database information. A regional CD-ROM database network retrieval system (such as a university library) also enables users to obtain a variety of optical disc data information from a local area network (Haddoud *et al.*, 2016). (5) Real-time activity type information: various types of investment market and analysis, entertainment, chat, online news discussion group, mail discussion group, BBS, online shopping, etc. (Fatima *et al.*, 2017).

Other types of information: various media such as graphics, music, movies, and advertisements (Navarro *et al.*, 2016).

III. METHODOLOGY

A. PSO algorithm model

Due to the fast growth of the Web and the difficulties in finding desired information, efficient and effective information retrieval systems have become more important than ever, and the search engine has become an essential tool for many people (Suominen *et al.*, 2016). The ranker, a central component in every search engine, is responsible for the matching between processed queries and indexed documents. Tie-Yan Liu is a lead researcher at Microsoft Research Asia. He leads a team working on learning to rank for information retrieval, and graph-based machine learning. The information world is changing rapidly, the pattern of network information is changing, the content and form of information are changing, and users' information needs are changing (Wu *et al.*, 2018). If the network retrieval system cannot adapt to these changes, it will fall into a very passive situation or even be eliminated. The best way to solve the problem is to add real time learning mechanism in the retrieval system to obtain all kinds of knowledge, make the system have intelligent characteristics such as self-learning and self-adaptive, and realize intelligent and knowledge-based retrieval (Gao, 2018; Zhu, 2017). For this reason, we propose a machine learning based intelligent and efficient knowledge search model (Du and Liu, 2014).

In this system, the first two search results obtained from each retrieval service agent are considered as associated. Those classified as related (or related) results will be placed on the list of search results, and the unrelated results will be placed below the list. The results of the same correlation are simply sorted in their natural order.

The CC4 algorithm is a neural network based on the intersection classification training algorithm. Determining the relevancy of web pages to a query term is basic to the working of any search engine. In this paper, we present a neural network based algorithm to classify the relevancy of search results on a metasearch engine. The fast learning neural network technology used by us enables the metasearch engine to handle a query term in a reasonably short time and return the search results with high accuracy. It is a more effective algorithm in all "Comer Classification" neural network algorithms. Because their mechanism is to divide the intersection angles (vertex angles) of an n-dimensional data cube through a hyperplane, and these angles are usually represented as training vectors. The important feature of the CC4 algorithm is that the training paradigm submitted to the network is only trained once at a time, which is called indicative learning. Compared with BP network, the intersection classification method has been proved to have a good generalization ability in pattern recognition and problem prediction, and the speed is much faster. The algorithm has been successfully applied in many applications, such as lossless text compression using CC4.In this paper, we will use this algorithm to determine the relevance of the results of distributed retrieval system, filtering useless information, in order to achieve more accurate results.

The CC4 network maps between an input binary vector X and an output vector Y. The architecture of the network is shown in Fig. 1.

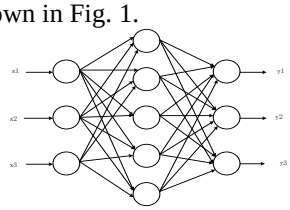

Figure 1. Architecture of CC4 network.

The input and output layers are fully connected. Neurons are binary neurons, and the activation functions are as follows:

$$y = f(\sum x_i) = \begin{cases} 1, \sum x_i \succ 0 \\ 0, \sum x_i \le 0 \end{cases} \quad (1)$$

Here's x_i =1, or 0. The number of neurons input is equal to the length of the input vector plus 1. The additional neuron is a constant 1 input neuron. The number of hidden layer neurons (hidden layer) is equal to the number of training samples, and each hidden neuron corresponds to a training sample.

B. Algorithm training

The RBM training process is divided into the following steps: Initialization weights and offsets. The bias value is generally initialized to 0.The weight matrix W is usually initialized with a Gaussian random function; the visible layer nodes are used to estimate the state of the hidden layer nodes. $(v \to h)$。

One by one calculates the probability value of each hidden layer node taking a value of 1.Assuming that the number of nodes in the visible layer and hidden layer is D and F respectively, we estimate the *j* hidden layer node and h_j.

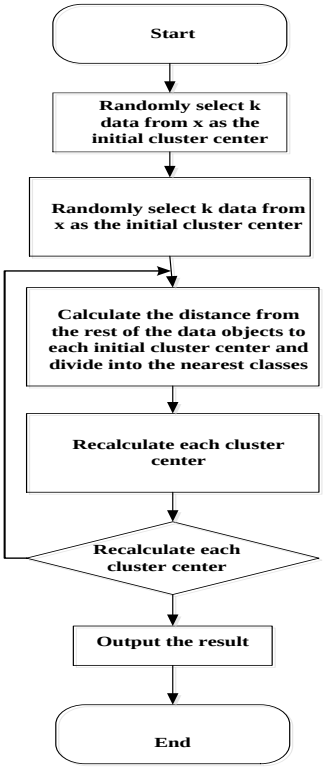

Figure 2. The flow chart of PSO algorithm.

The weights and offsets used are w_{ji} and b_i, respectively. The value of each node's value v_i is multiplied by the weight of the corresponding connection weight w_{ji}, and the sum is added to the bias value b_i of h_j. Then a sigmoid function is applied: $f(x)=1/\exp(-x)$. The value of $p(h_j=1|v)$ represents the activity of h_j under the effect of each visible layer node. Its calculation formula is as follows:

$$p(h_1=1|v)=sigmoid\left(\sum_{i=1}^{D}v_i w_{1i}+b_1\right)=\frac{1}{1+\exp\left(-\sum_{i=1}^{D}v_i w_{1i}-b_1\right)} \quad (2)$$

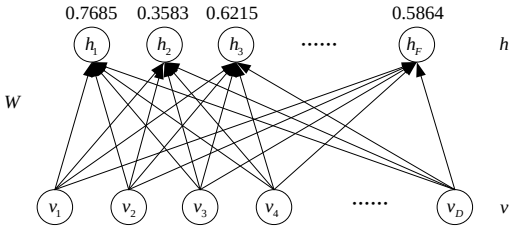

Figure 3. Computing p(hj=1/v).

In Fig. 3, the probability that each hidden layer node has a value of 1 is:

$$p(h_1=1|v)=0.7585,$$

$$p(h_2=1|v)=0.3583,$$

$$p(h_3=1|v)=0.6215,$$

$$p(h_F=1|v)=0.5864$$

Record such probabilities in the ph array. After getting the ph, a random sampling is performed to obtain the state phstate of each hidden layer. First generate a set of random numbers with dimension F. In the jth position, compare ph(j)with rand(j). If ph(j)>rand(j), then phstate(j)is 1, otherwise, phstate(j)0 (Fig. 4).

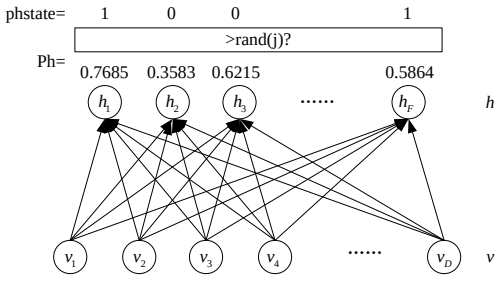

Figure 4. Computing p (phstate).

The significance of random sampling is that by sampling, highlighting the role of the nodes that play a major role in the hidden layer and ignoring the role of other nodes (node states are assigned a value of 1 and 0, respectively), such a prominently-ignored goal is to

make the data It is easier to represent in another feature space while exploring features that may be contained in the data, making prediction or classification more accurate.

Using Hidden Layer Nodes to Estimate the State of Visible Layer Nodes After finding the phstate of each hidden layer, using the same weight matrix *W*, reconstruct the original data v (using the image as an example) to obtain the original image. An image (negdata, or confabulated images), denoted as v.

The process of using h -to- v' estimation is similar to using estimated $(h \to v')$. Using formula (1) to calculate the probability $p(v_j = 1|h)$ of the node in the virtual image of the i-th visible layer node (Fig. 5). Compared with Eq. 2, in Eq. 3, the weight matrix is once transposed, and the offset value used is the offset value a_i of the i-th visible layer node.

$$p(h_j = 1|v) = sigmoid\left(\sum_{i=1}^{D} v_i w_{ji} + b_1\right) = \frac{1}{1+\exp\left(-\sum_{i=1}^{D} v_i w_{ji} - b_j\right)} \quad (3)$$

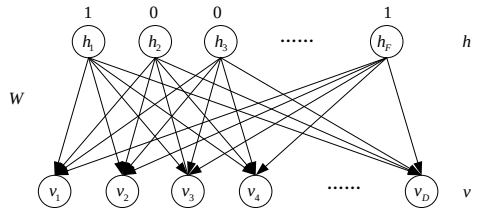

Figure 5. Computing p(vj=1lh).

Once again, the visible layer node is used to estimate the state of the hidden layer node $(v' \to h')$.The difference is that starting from the virtual image v, the virtual layer image is used to estimate the state of the hidden layer node again. In the first few steps of the RBM training process, the weight and offset values are updated.

In essence, the estimation from the visible layer v to the hidden layer h is performed twice. The difference between the last time and the previous one is that the latter time is from the virtual image. The v starts to get h, and v' is obtained from the previous h estimate, as shown in Fig.6.

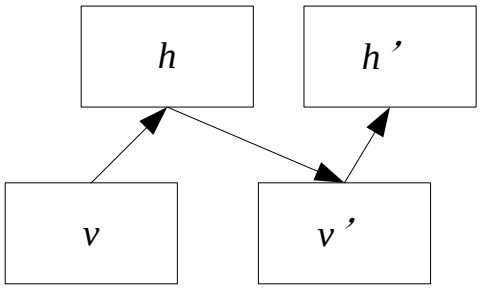

Figure 6. A schematic diagram of the state estimation of the visible and hidden layer nodes.

Table1. Iris data sampling attributes and their value ranges.

Property name		Value range
Sepal	Length	3~9
	Width	0~4
Petal	Length	0~9
	Width	1~5

Table 2. Attributes and value ranges of wine data samples.

Property name	Value range
Alcohol	11~14
Malic acid	4~8
Alcalinity of ash	1~3
Magesium	10~30
Flavanoids	70~162
Hue	4~7
Proline	278~688

IV. RESULT ANALYSIS AND DISCUSSION

In order to verify the performance of the proposed algorithm, the test data set used is the famous Iris dataset and Wine dataset in the UCI database. Then the corresponding comparison algorithm is used to simulate the two datasets.

The dataset is introduced: The Iris dataset Iris is an iris plant. The dataset contains three flowers named Setosa, Versicolor, and Virginica. Each flower has 50 samples, so the dataset contains a total of 150 data samples. Each sample contains four properties: Spetal length, Spetal width, Petal length, and Petal width. Therefore, each data sample can be represented as a 4-dimensional vector of four characteristics of the plant. The following table shows the attributes of the Iris dataset and their range of values: Wine dataset: The Wine dataset is a collection of three wine analysis results for the same region of Italy. The three wines contain different chemical components. In fact, this dataset contains a total of 178 data samples, and is divided into 3 categories. These three categories respectively include 59, 71, and 48 samples. Each sample contains 13 attributes such as Alcohol, Malic acid, and Ash. Therefore, each sample can be represented as a 13-dimensional vector. The following table shows the various attributes of the Wine dataset and their range of values:

This article focuses on the concept of the PSO algorithm, as well as the value of several related parameters in the velocity and position formulas. Some improvements are made to the standard particle swarm optimization algorithm.

A non-linear degressive time factor is added to the position formula to make the variable's dimension on both sides of the position formula consistent, which ensures that the particles have a better global search in the initial stage of the algorithm. Ability to further improve the accuracy of the algorithm. In order to prevent particles from "premature" during flight, chaotic search techniques are introduced.

The randomness and ergodicity of chaos technology are used to maintain the diversity of particle swarms. At the same time, in order to prevent particles from leaving the prescribed motion space, boundaries are introduced. The buffer wall technology can dynamically adjust the flight position of particles and the accuracy of clustering to some extent.

V. CONCLUSIONS

In order to solve the problem of rough search results in traditional knowledge search system, the adaptive ability of user knowledge personalized selection is not strong enough. An efficient intelligent knowledge search system based on machine learning is proposed in this paper. By using the method of machine learning in the traditional intelligent knowledge search system, the intelligence of knowledge search is greatly improved. However, we should know that network knowledge search emphasizes a kind of system function, that is, it is not only to search knowledge. Simple combination of cable technology and machine learning method can achieve the desired function. It is necessary to combine the two organically, that is to use appropriate learning methods according to the specific type of knowledge, in order to make our search system more intelligent, so that users can get more accurate search results more quickly.

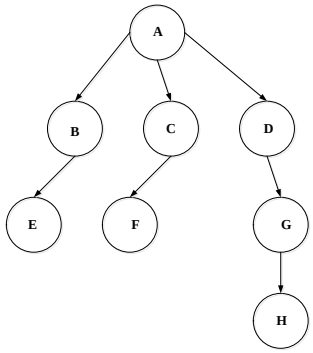

Figure 7. The figure of crawling strategy.

REFERENCES

Allix, K., T.F. Bissyandé, Q. Jérome, J. Klein, and Y. Le Traon, "Empirical assessment of machine learning-based malware detectors for Android", *Empirical Software Engineering*, **21(1)**, 183-211 (2016).

Bougoudis, I., K. Demertzis and L. Iliadis, "HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens", *Neural Computing and Applications*, **27(5)**, 1191-1206 (2016).

Du, C.H. and J.R. Liu, "A new utilization approach of natural soda ash: to manufacture sodium percarbonate", *Latin American Applied Research*, **44(2)**, 179-183 (2014).

Fatima, M. and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic", *Journal of Intelligent Learning Systems and Applications*, **1(1)**, 1-16 (2017).

Gao, X. and T. Zhang, "Unsupervised learning to detect loops using deep neural networks for visual SLAM system", *Autonomous Robots*, **41(1)**, 1-18 (2017).

Gao, Y., M. Reza Farahani, and W. Gao, "Ontology optimization tactics via distance calculating", *Applied Mathematics and Nonlinear Sciences*, **1(1)**, 159-174 (2018).

Haddoud, M., A. Mokhtari, T. Lecroq, and S. Abdeddaim, "Combining supervised term-weighting metrics for SVM text classification with extended term representation", *Knowledge and Information Systems*, **49(3)**, 909-931 (2016).

Huang, G., S. Song, J.N. Gupta, and C. Wu, "Semi-supervised and unsupervised extreme learning

machines", *IEEE Trans. Cybern.*, **44(12)**, 2405-2417 (2014).

Khalique, C.M. and I.E. Mhlanga. "Travelling waves and conservation laws of a (2+1)-dimensional coupling system with Korteweg-de Vries equation", *Applied Mathematics and Nonlinear Sciences*, **3(1)**, 241-254 (2018).

Kulkarni, A.J. and H. Shabir, "Solving 0–1 knapsack problem using cohort intelligence algorithm", *International Journal of Machine Learning and Cybernetics*, **7(3)**, 427-441 (2016).

Lei, Y., F. Jia, J. Lin, S. Xing and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data", *IEEE Trans. Ind. Electron.*, **63(5)**, 3137-3147 (2016).

Narudin, F.A., A. Feizollah, N.B. Anuar and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection", *Soft Computing*, **20(1)**, 343-357 (2016).

Navarro, P.J., F. Pérez, J. Weiss and M. Egea-Cortines, "Machine Learning and Computer Vision System for Phenotype Data Acquisition and Analysis in plants", *Sensors*, **16(5)**, 641 (2016).

Soueidan, H. and M. Nikolski, "Machine learning for metagenomics, methods and tools", *Metagenomics*, **1(1)**, 1-19 (2016).

Suominen, A. and H. Toivanen, "Map of science with topic modeling, comparison of unsupervised learning and human-assigned subject classification", *Journal of Association for Information Science and Technology*, **67(10)**, 2464-2476 (2016).

Taherkhani, N and S. Pierre, "Centralized and Localized Data Congestion Control Strategy for Vehicular Ad Hoc Networks Using a Machine Learning Clustering Algorithm", *IEEE Transactions on Intelligent Transportation Systems*, **17(11)**, 3275-3285 (2016).

Wu, Yan and Fu Jing-Li "Noether's theorems of variable mass systems on time scales," *Applied. Mathematics and Nonlinear Sciences*, **3(1)**, 229-240 (2018).

Zhao, J., M. Mathieu, R. Goroshin, and Y. Lecun, "Stacked What-Where Auto-Encoders," *arXiv preprint arXiv:* 1506.02351, (2015).

Zhou, W., M. Yang, X. Wang, H. Li, Y. Li and Q. Tian, "Scalable feature matching by dual cascaded scalar quantization for image retrieval", *IEEE Transactions on Pattern Analysis and machine intelligence*, **38(1)**, 159-171 (2016).

Zhu, L., Y. Pan and J. Wang, "Affine Transformation Based Ontology Sparse Vector Learning Algorithm", *Applied Mathematics and Nonlinear Sciences*, **2(1)**, 111-122 (2017).

Received: December 15th 2017
Accepted: June 30th 2018
Recommended by Guest Editor
Juan Luis García Guirao