Survey of Analytical Methods for Big Data of Quality Inspection

Yingcheng Xu[1], Wei Jiang[2*], Xiuli Ning[1], Bisong Liu[1] and Ya Li[1]

[1]Quality Management Branch, National Institute of Standardization, Beijing 100191, China
[2]School of Computer and Information, Anqing Normal University, Anqing 246011, China
Email: wjiangaqtc@163.com

Abstract. Big data of quality inspection is consist of structured data, semi-structured data and unstructured data. It brings great challenge for data analysis as the features of huge volume, various types, high timeliness and low value density. This paper introduced the research status of semantic analysis technology based on depth learning, migration learning algorithm based on depth neural network, reinforcement learning algorithm based on depth neural network and visualization analysis from the characteristics of quality inspection data. The existence problems are analyzed and the research direction is looking forward of the on data analysis in this paper.

1. Introduction

Some serious quality events have consecutively happened in China during recent years, which severely affects people's lives and assets, e.g. 2012 "Anxin toxic floor board" event, 2013 "Cancerigenic toxic school uniform", 2014 "Excessive lead level in the tap" and "Shaft break of Volkswagen Sagitar", 2015 "Mobile power explosion" and 2016 "Toxic runway" event. The quality security events are caused due to plentiful reasons, but how to eliminate structural barriers among different sources and data types, effectively analyze isolated and fragmented data, mine contained information, knowledge and intelligence from data according to actual requirements, make a correct decision as soon as possible, and prevent against risk has become a key issue for the quality inspection authority to solve. The State Council issued "Action outline for promoting big data development" in Aug, 2015, which proposes to "promote big data application in fields such as quality security", therefore, it is very necessary and significant to analyze and study big data in the quality inspection field.

The quality inspection data are from diversified sources, which not only include traditional state product quality supervision and sampling, 12315/12365 consumer complaints, WTO/TBT recall notification, product quality arbitration, lab product detection, product injury and accident, but also includes emerging social media data with extensive user interaction such as Blog, WiKi, Microblog, forum, social network and content communities. The generalized quality inspection data covers data such as measurement, standard, inspection, certification and special equipment detection data, different real-time data generated by new products in IoT and cloud computing and external data such as macro economy and environmental protection. These data is the integrated part of the big data of quality inspection and are diversified in information source, information type, descriptive structure, text feature, expression mode and transmission channels. These data includes text information, image information, audio information and video information, are classified into structural information, semi-

structural information and non-structural information, feature high magnitude, diversified types, high timeliness and low value density, so it brings huge challenges to data mining and analysis.

2. Current Conditions on Big Data Analysis and Research

2.1. Semantic Analysis Technology Based on Deep Learning

Unlike the traditional statistical machine learning, to use the deep learning model in the natural language processing, first the features are expressed from discrete one hot embedding to the continuous dense embedding, namely distributed expression. One strength of the distributed expression is "distance" among features, which is very helpful for plentiful natural language processing tasks. Bengio et al. (2003) proposed a n-gram language model composed of one L3 neutral network in the early application of the deep learning. The first layercan map single word to the low-dimension real number embedding space via the one-hot expression of words and dot product of the word embedding matrix. The word embedding based on this mode is called as distributed representation of word or word embedding. The word embedding replaces the traditional one-hot for vocabulary expression, which solves dimension disaster due to one-hot [1]. Mnih et al. (2009) proposed the hierarchical neutral network language model to speed up model training and deduction process [2]. Hinton(2007) proposed the Log-Bilinear language model based on RBM, which features a novel utilization mode based on the history information. The history vocabulary embedding is linearly transformed and the sum is calculated to get the hidden layer expression h (h is the summary expression of the history information semantics). Next the inner product of h and candidate word embedding is calculated as the output of the output layer, so it can play the role of the word embedding to most extent [3]. Collobert et al. (2008) processed other natural langue tasks by using the word embedding. The word embedding obtained from the unsupervised training language model is used in training as the initial value to improve performance of related supervision tasks, get better word embedding and make word embedding get task-related information [4]. Based on word embedding, plentiful researchers have tried to get feature expression of sentences and documents. The feature expression learning of sentences and documents is also one hot spot of current deep learning in NLP field. On the whole, the machine learning is divided into supervised learning and unsupervised learning. The supervised learning orients to different classification tasks, e.g. text relation classification task, sentiment analysis task, etc. Huang et al.(2013) proposed the deep semantic similarity match model (DSSM) to learn the sentence expression. The model will design the target function according to the semantic similarity between sentences and guide model training by using the user clicking feedback data in the webpage search. Kai et al. (2015) proposed to construct a tree LSTM network by using the Tree LSTM model and learn the sentence expression via the sentiment analysis and sematic similarity determination [7]. Zhu et al. (2015) proposed the tree LSTM. These models construct a tree by using the prior knowledges such as dependence tree and parsing tree [8]. Li et al. (2015) proposed the hierarchical LSTM model on the 2015 ACL conference to process word, sentence and paragraph input by using different LSTM, guide parameter learning via re-establishment of input texts, and detect the information saving and compression capabilities of LSTM by using the reestablishment effect of the autocoder. The feature expressions of the word, sentence and document is from LSTM output at different levels [9].

2.2. Transfer Learning Based on Deep Neutral Network

The generalized transfer learning involves multiple learning frameworks such as multi-task learning, field adaptation, variance deviation, sample selection offset, concept drift and robust learning. The narrow transfer learning includes the data set deviation, field adaptation and multi-task learning. The transfer learning is mainly used for cross-language text classification and retrieval. The training data and test data are from different language types. Bickel S et al. (2007) corrected the edge distribution and conditional distribution differences via the instance weighting method and proposed the concept of the data set offset [10]. Zhong E et al. (2009) reduced the edge distribution and conditional distribution differences via the feature expression method and studied the isomorphic learning [11]. Wei et al. (2010) first translated the auxiliary field language to the target field language via automatic

translation, next processed the probability distribution mismatch problem of the uniform language, and transformed the heterogeneous transfer learning to the isomorphic transfer learning [12]. Platt et al. (2010) learnt the language-related feature project from the training documents of different languages, mapped different feature spaces to one "language-independent" abstract space, and got association among different languages by using the canonical correlation analysis[13]. Dai et al. (2008) proposed the translation learning method, established the supervision model by using the training data in one feature space, and predicted the test data in another feature space [14]. Shi et al. (2010) got the conditional probability of the given label time word translation via the collaborative learning based on expectation maximization algorithm of the double-language dictionary, auxiliary field data and target field data [15]. Zhu et al. (2011) proposed the knowledge transfer method between text and images [16]. Li et al. (2014) proposed the generic isomeric transfer learning method based on feature alignment, expansion and support vector machine, which features good effect in cross-language and cross-media task. The type and space inconsistency of the auxiliary field and target field attract extensive attention in the text mining and image understanding [17].

2.3. Intensive Learning Algorithm Based on Deep Neutral Network

Generally the issues faced in deep intensive learning feature strong time dependence. The recurrent neural network is suitable for processing issues related to time sequences. Combination of the intensive learning and the recurrent neural network is also one key form for deep intensive learning. Narasimhan et al. (2015) proposed the deep network architecture based on combination of a long and short time memory network and intensive learning to process the text game. Such method can map the text information to the vector expression space and get the sematic information of the game status [18]. The deep Q network processes the time sequence information by adding the experience playback mechanism, but the memory capabilities of the experience playback is limited. Each decision point shall get the whole input image for perception and memory. Sorokin et al. (2015) combined the long and short time memory network with the deep Q network and proposed the deep recurrent Q network (DRQN). This network shows higher robustness in partially observable Markov decision process (POMDP) and gets the better experimental results when some frame images are missing [19]. With success of the vision attention mechanism in the target tracking and machine translation field, inspired by this mechanism, Sorokinet al proposed deep attention recurrent Q network (DARQN). It can selectively pay key attention to the related information area to reduce parameter number and computing consumption of the deep neutral network [20].

2.4. Visual Analysis

The visual analysis of the big data aims to utilize the automatic analysis capabilities of computers, fully mine recognition capability strength of the human being in the visual information, seamlessly fuse strengths of the human being and the machines, and assist the human being to intuitively and efficiently perceive the information, knowledge and intelligence behind big data by using the human-machine interactive analysis method and interactive technology [21]. The domestic and foreign scholars deeply study technologies such as text visualization, network (graph) visualization, spatio-temporal data visualization and multi-dimension data visualization in big data mainstream applications and have achieved a series of achievements. Keim et al. (1996) induced the basic methods of multi-dimension visualization, including geometric graph, icon, pixel, hierarchical structure, graph structure and mixing method [22]. Plaisant et al. (2002) proposed a concept of the tree browser and dynamically adjusted the size of the twigs to best adapt the displaying area [23]. Herman et al. (2000) summarized the basic methods and technologies for graph visualization and studied applications [24]. Gou et al. (2011) summarized the tree visualization technology of the radiogram and space filling method and proposed the TreeNetViz algorithm, but the algorithm is difficult to support visualization of the large-scale graphs [25]. Slingsby et al. (2011) proposed the graph visualization technologies based on the framework, calculated the framework according to the edge distribution law, and bound the edges based on the framework[27]. Tominski et al. (2012) introduced the stack graph and extended the multi-dimension attribute display space in the space-time cube[28].

3. Issues and Research Direction in Big Data of Quality Inspection Analysis

3.1. Issues in Analysis and Research of Current Big Data of Quality Inspection

From the references related to the big data analysis, now the big data analysis and mining mainly focus on the sematic analysis based on deep learning and transfer learning and intensive learning based on the deep neutral network. Although these method solve some problems in big data analysis, some weaknesses exist. The research on big data of quality inspection analysis, mining and application are limited in current references. The generic methods and models are lack of industry pertinence, so they cannot satisfy the realistic requirements for deep analysis on the quality inspection data.

(1) Knowledge acquisition: The big data of quality inspection features continuous growth and continuous upgrade. With continuous growth of the problem solving scale, the demand of the traditional massive knowledge acquisition method for space and time also quickly grows under the continuously changing dynamic data, so the knowledge learning technologies cannot keep abreast with the data update rate. Therefore, the timeliness is a big challenge for quality inspection data acquisition. Low data quality, data loss or data incompleteness result from missing data attributes, damage of storage medium, failure of collection device and jamming of transfer matched body in actual applications of the big data of quality inspection, which is a challenge faced in the quality inspection data analysis.

(2) Characteristic adaptation: with continuous accumulation and update of quality-related data, the "quality big data" era is upcoming. The large data result in severe missing of data labels and statistical isomerism. How to transfer knowledge and adapt the model among the cross-organization and cross-industry data is a problem faced in the big data of quality inspection. The sparsity of the labeled data will result in severe over-fitting of the traditional supervised learning. Although the traditional half-supervised learning, proactive learning and collaborative training can solve the data sparsity, some labeled data are required for the target field. The labeled data are very sparse and the acquisition cost is higher, a proper method is required to solve the labeling sparsity of the quality inspection data.

(3) Evolution and prediction: The core of the big data of quality inspection includes evolution and prediction. The mathematical algorithms are used in the massive data to predict possibility of quality security events in evolution and prediction. Now the statistical analysis of the big data of quality inspection is limited to analysis on the existing data and after-event analysis on the happened quality security events, but the events are not predicted and analyzed prior to occurrence. How to predict and analyze quality security events, discover possible quality risks in time, strengthen proactive, early and pertinent statistical analysis, and predict and alarm the product quality security are the urgent issues to solve in the quality inspection industry.

(4) Visual analysis: The visual analysis of big data depends on data. The big data of quality inspection are from diversified resources and heterogeneous. Therefore, it is difficult to guarantee data integrity, accuracy and consistency. The data quality issues will directly affect science and accuracy of the visual analysis. Most references are based on the statistical analysis level and do not aim at features of diversified sources and complicated isomerism of big data, and cannot deeply study the big data visualization technology from the integrity, consistency and accuracy view.

3.2. Research Directions in Big Data Analysis of Quality Inspection

Data analysis is the process to get hidden information and knowledge from massive, complicated, irregular, random and fuzzy data, which are not perceived in advance and have potential values[29-32]. The big data of quality inspection is from different organizations such as General Administration of quality Supervision and Quarantine, General Administration of Industry and Commerce and National Development and Reform Commission. Generally these data involve different fields and orient to different industries. Cross-organization and cross-field big data is subject to different probability distribution, so it generates barriers for model generalization among different fields and labels are sparse in case of execution of machine learning. Therefore, it is urgent to design effective algorithms in cross-organization and cross-industry big data of quality inspection analysis in order to reduce distribution differences and manual labeling.

(1) Research on dynamic knowledge acquisition method based on incomplete data

The big data of quality inspection features incompleteness and dynamic update. How to mine and analyze massive, incomplete and dynamically updated data and acquire knowledge is a big challenge for big data of quality inspection analysis. The research on knowledge acquisition of the big data of quality inspection focuses on the following aspects: establish the dynamic probability rough set model based on data object update for the dynamic insertion and deletion of data objects in the incomplete information system, discuss incremental update method of similar set based on the rough set theory according to influences of the data objects to add or delete on different updates of decision classification in the information system for the dynamic insertion and deletion of data objects in the set value information system, establish the matrix method for uniform maintenance of similar set in the dominant rough set model for the dynamic addition and removal of attribute features in the set value information system, and establish the dynamic dominant rough set model based on data value change for the random update and change of the data value in the set value decision system.

(2) Research on characteristic adaptation method of big data of quality inspection based on transfer learning

For features of the cross-organization and cross-industry industry data of the big data of quality inspection, the transfer labeling data of the auxiliary field is established to improve the learning effect of the target field. Based on it, the training data and test based on the supervised learning is subject to the basic assumption of the independent identical distribution. The characteristic representation based on the constant transfer learning field is studied by combining the transfer learning theory for inconsistent feature space. The depth features are learnt for constant fields and probability distribution differences are corrected based on the deep transfer learning framework. The generic non-linear characteristic learning is constructed, the autocoder model for invariant denoising is improved, and the non-linear distribution differences of cross-organization data are minimized. Finally the transfer crossed validation strategy is proposed to select the unsupervised transfer learning model for the non-labeled data in the target fields.

(3) Research on evolution and prediction method of big data of quality inspection based on intensive learning

By combining the deep neutral network and intensive learning model, the deep neutral network is used to place the environmental statuses and decision actions into the continuous sematic space from the view of environmental status, decision action and status transformation rules in order to study Q learning of the continuous semantic space and construct the deep intensive association network. Based on it, for the features of the big data of quality inspection, the evolution and prediction model of the big data of quality inspection is established, the status and actions are described in a text string, and some decisions and predictions are made via the text string to generate corresponding decision support actions under the new environmental statuses.

(4) Research on visualization technology of big data of quality inspection based on theme model

The large-scale and high-dimension data visualization technology is studied based on deep learning mining and analysis of the big data of quality inspection and time, space and object alignment and fusion theory of multi-source heterogeneous data by combining immersed loop screen visualization technology and human-machine interaction technology and the high-quality fine fusion and reduction of the heterogeneous data is implemented. The multi-variant, multi-level and multi-dimension integrated and interactive display method for the large-scale fusion data is studied. The pattern discovery and real-time decision method based on large-scale visual data is studied, the feature data dimensions are reduced or are directly displayed in the collaborative analysis of single view and multiple views and information in the limited space is displayed as much as possible. By combining the multi-body sensor interactive recognition device, the visual analysis and mining technology based on dynamic feedback is studied, the multi-level visual display method based on cross-screen and cross-area interactive operations is constructed to provide reliable tools for pattern discovery and real-time decision method.

4. Conclusions

When the big data of quality inspection is analyzed, on the one hand, the multi-source heterogeneous data with huge differences shall be faced. On the other hand, the real-time stream data will be faced

and the data shall be effectively expressed, explained and learnt. Effective analysis on the quality inspection data can assist to grasp generation and development law of the product quality security events and further provide technical support for correct governmental decisions. This paper analyzes the current research conditions of the big data analysis in the sematic analysis technology based on deep learning, transfer learning algorithm based on deep neutral network, intensive learning algorithm based on deep neutral network and visual analysis and the issues and research direction of the big data of quality inspection analysis are discussed.

5. Acknowledgments

We would like to acknowledge that this research are supported and funded by the National Science Foundation of China under Grant No. 91646122 and 91746202, the Basic Scientific Research Business Projects 552018Y-5927, the National Key R&D Program of China under Grant No.2016YFF0202600, No.2016YFF0202604 and 2017YFF0209604, and 2016 Natural Science Research Key Project of Department of Education for Anhui Province under Grant No. KJ2016A866.

6. References

[1] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. The Journal of Machine Learning Research, 2003, 3: 1137-1155.
[2] Mnih A, Hinton G E. A scalable hierarchical distributed language model[C] // Advances in neural informa-tion processing systems. 2009: 1081-1088.
[3] Hinton, G. E. To recognize shapes, first learn to generate images[J]. Progress in brain re-search, 2007, 165(6):535-547.
[4] Collobert R., Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C] // Proceedings of the 25th international conference on Machine learning. ACM, 2008:160-167.
[5] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data[C] // Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013: 2333-2338.
[6] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modellingsentences[J]. Advances in Neural Information Processing Systems, 2015, 3:2042-2050.
[7] Kai Sheng Tai, Richard Socher& Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks[J]. Computer science, 2015, 23(3):1212-1222.
[8] Zhu B., Cheng J, et al. Error Rate Bounds for Equal-Gain Combining Over Arbitrarily Correlated Rician Channels[C] // 2015 IEEE Global Communications Conference (GLOBECOM). IEEE, 2015:203-211.
[9] Li J, Jurafsky D, Hovy E. When Are Tree Structures Necessary for Deep Learning of Representations[J]. Computer Science, 2015, 23(6):1324-1340.
[10] Bickel S, Bruckner M, Scheffer T. Discriminative learning for differing training and test distributions[C] // Proceedings of the 24th international conference on Machine learning, 2007.
[11] Zhong E, Fan W, Peng J, et al. Cross Domain Distribution Adaptation via Kernel Mapping[C] // Pro-ceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.
[12] Wei B, Pal C. Cross lingual adaptation: an experiment on sentiment classifications[C] // Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics, 2010: 258-262.
[13] Platt J, Toutanova K, Yih W T. Translingual document representations from discriminative projections[C] // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010:251-261.
[14] Ling X, Xue G R, Dai W, et al. Can Chinese web pages be classified with English data source[C] // Pro-ceedings of the 17th international conference on World Wide Web, 2008:969-978.

[15] Shi L, Mihalcea R, Tian M. Cross language text classification by model translation and Semi-Supervised learning[C] // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010:1057-1067.
[16] Zhu Y., Chen Y., Lu Z., et al. Heterogeneous Transfer Learning for Image Classification[C] // Proceedings of the 25th AAAI Conference on Artificial Intelligence, 2011:56-64.
[17] Li W, Duan L, Xu D, et al. Learning with Augmented Features for Supervised and Semisupervised He-terogeneous Domain Adaptation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(6):1134-1148.
[18] Narasimhan K, Kulkarni T, Barzilay R. Language understanding for text-based games using deep rein-forcement learning[J]. Computer Science, 2015, 40(4):1-5.
[19] I Sorokin, A Seleznev, M Pavlov, A Fedorov, A Ignateva , Deep Attention Recurrent Q-Network[J]. Computer Science , 2015, 22(5):103-109.
[20] Sorokin A.A.KorolevS.P.UrmanovI.P.VerkhoturovA.L.MakogonovS.V.ShestakovN.V.,Software Platform for Observation Networks Instrumental Data[C] // 2015 International Conference on Computer Science and Environmental Engineering, 2015:943-949.
[21] Ren Lei, Du Yi, Ma Shuai et al. Review of big data visual analysis[J]. Journal of Software, 2014, 25 (9):1909-1936.
[22] Keim DA, Kriegel HP. Visualization techniques for mining large databases: A comparison[J]. IEEE Trans. on Knowledge and Data Engineering, 1996, 8(6):923-938.
[23] Plaisant C, Grosjean J, Bederson BB. Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation [C] // In: Proc. of the IEEE Symp. on Information Visualization (InfoVis 2002). Washington: IEEE Computer Society, 2002, 57-64.
[24] Herman I, Melancon G, Marshall MS. Graph visualization and navigation in information visualization: A survey[J]. IEEE Trans. on Visualization and Computer Graphics, 2000, 6(1):24-43.
[25] Gou L, Zhang X. Treenetviz: Revealing patterns of networks over tree structures[J]. IEEE Trans. on Visualization and Computer Graphics, 2011, 17(12):2449-2458.
[26] Slingsby A, Dykes J, Wood J. Exploring uncertainty in geodemographics with interactive graphics [J]. IEEE Trans. on Visualization and Computer Graphics, 2011, 17(12): 2545-2554.
[27] Ersoy O, Hurter C, Paulovich FV, Cantareiro G, Telea A. Skeleton-Based edge bundling for graph visua-lization[J]. IEEE Trans. on Visualization and Computer Graphics, 2011,17(12):2364-2373.
[28] Tominski C, Schumann H, Andrienko G, Andrienko N. Stacking-Based visualization of trajectory attribute data [J]. IEEE Trans. on Visualization and Computer Graphics, 2012, 18(12):2565-2574.
[29] Liu Zhihui, Zhang Quanling. Review of Big Data Technology Research[J]. Journal of Zhejiang University (Engineering Science), 2014, 6: 957-972.
[30] Meng Xiaofeng, Ci Xiang. Big Data Management: Concepts, Technologies and Challenges [J]. Journal of Computer Research and Development, 2013, 1: 146-169.
[31] Wu Xindong, Ji Shengkai. Comparative Study on Map Reduce and Spark for Big Data Analytics[J]. Journal of Software, 2018, 29(6):1770-1791
[32] Weijing Song, Lizhe Wang, Yang Xiang, Albert Y. Zomaya. Geographic spatiotemporal big data correlation analysis via the Hilbert–Huang transformation[J]. Journal of Computer and System Sciences, 2017, 89-94.