Extracting knowledge patterns in a data lake for management effectiveness

Ziyi Cheng[1,a], Haitong Wang[2,b], Hongyan Li[2,c]

[1]International Business School, Shaanxi Normal University, Xi'an, China
[2]International Business School, Shaanxi Normal University, Xi'an, China

Abstract—With the correlation collision between different types of data becomes more and more intense, a meaningful and far-reaching data revolution has arrived. Enterprises urgently require a hybrid data platform that can effectively break data silos, and unify data aggregation and sharing. Once the data lake was born, it has been a promising method for enterprises to profoundly improve their Business Intelligence. In this paper, we combine principle component analysis (PCA) with a network-based approach to extract a visual knowledge pattern from data sources in data lake, so as to improve management effectiveness.

1 Introduction

Big data, once it appeared, has the power to show amazing methods instead of traditional ways of doing business, as well as the management of organizational knowledge [1]. Due to the advent of big data, managers today understand their enterprise, their surroundings, their customers and their competitors better. Therefore, they tend to be able to achieve a faster decision making, accurately develop new products according to customer needs, increase customer loyalty, explore a series of new markets, improve operating income, and ultimately reinforce management effectiveness. [2]. In other words, big data analytics can be a great investment for companies to improve organization performance and stay competitive with their business competitors. [3].

Nevertheless, traditional data warehouses extract data and integrate data after converting it to a common static pattern. Thus, they are called centralized data stores [4]. Now that this way of data integration ultimately leads to severe information silos, companies require finding an effective method to link and structure the various flows from the information silos that help address specific issues flexibility [5].

Data Lake is the most promising and relevant solution for this emergency [6], and is used as a multifunctional data repository to store largescale of raw data, offering extraction, exploration and monitoring capabilities [7]. Although data lake breaks the data silos, if there is no proper method of data integration in a heterogeneous environment, the data lake will fall into an awkward situation and eventually become a useless swamp.

Network analysis is always regarded as a flexible and intuitive analysis method. In the paper, we propose an approach based on the method of network analysis to extract an inner relationship involving constructs coming from heterogeneous source of a data lake, trying to help companies build a pattern that makes it more intuitive to understand the relationship between various kinds of performance and capabilities.

It is universally acknowledged that, big data has 4 main characteristics (4V): volume, velocity, variety, and value. In a big data scenario, it is significant to reduce dimensions of the data in order to increase the velocity, meanwhile, cut back data volume and type. Principle component analysis (PCA) allows considering all the elements and finally achieving dimensional reduction. Hence, we combine PCA with the network-based approach to extract a visual knowledge pattern to improve management effectiveness. The main contribution of this article includes:

- This paper uses a network-based model to extract visual knowledge patterns.
- Introduce the PCA algorithm and achieve dimension reduction by this method.

We used the following structure in this paper: In section 2, we review related works. In section 3, we narrate and propose the network model based on PCA. Section 4 shows steps to extract visual knowledge patterns. In section 5, we summarize the paper, and draw conclusions.

2 Related Literature

2.1 Data lake

First proposed by James Dixon, data lake is described as a tool that works across a variety of data sources [8]. Although there is no generally accepted concept or

[a]czy@snnu.edu.cn, [b]1219769143@qq.com, [c]2585941175@qq.com

definition here, the relevant definitions of different scholars or institutions are highly correlated.

For instance, Tyagi and Demirkan claim that in the data lake, any type of data can be used and analyzed directly without being predefined [9]. Analogously, Alserafi A., Abelló A., Romero O., and Calders T. introduce data lake as "Data lakes save the crude data format without conversion or pre-order and can be accessible using read mode." [10]. Miloslavskaya and Tolstoy define data lake as "A data lake holds structured, unstructured and semi-structured data in its native format and can capture data without damaging the data structure" and "This can be considered as a huge data set that has access to almost all historical and new information in real time. In this case, once the data is queried, the schema will be defined" [11]. Endris K.M., Rohde P.D., Vidal ME., and Auer S. say that "To solve the problem of data integration, especially the heterogeneous data, a number of data lake systems are appeared, focusing more on data input and on metadata management." [4]. Llave [12] defines a data lake from a business perspective as "an ability to improve business effectiveness where you can get raw, unaltered data from various source systems." or "the place where you can find all the data in your company."

Obviously, there is a strong agreement in the concept of data lake. However, there are various management approaches, paradigms and frameworks to solve the problems due to limitations caused by the characteristics of data lake.

2.2 Approaches and techniques in the big data context

Compared with traditional data warehouses, data lakes have many clear advantages, but they do have some limitations as well [13]. While data lake aims at getting rid of human effort before using the data, the dilemma is just postponed because there were still difficulties in preparing and cleaning the database. Introducing Kayak, a framework, Maccioni and Torlone [14] tackle this problem and help scientists to define and optimize data pipelines. Using Kayak, users can tailor their needs according to the precise requirements and create an output that meets those requirements. To integrate data silos and machine learning technologies into a data lake, Wibowo M., Sulaiman S., and Shamsuddin S.M. present Rough Set as a predicting method, because the significant features can be well considered by using reduce computation and non-critical features in silos will be eliminated. [15]. The combination of different data sources provides the need for metadata management. In [16], M. Farid, A. Roatis, I. F. Ilyas, H.-F. Hoffmann, and X. Chu propose CLAMS system, which focuses on integrity constraints of data. Jian Liu, X.X. Zhang, and Lei Zhang [17] propose a novel framework based on tree-pattern used to handle obscure XML queries.

In the literature, we saw a large number of methods based on network analysis are proposed [6][18][19][20]. Paolo Lo Giudice, Lorenzo Musarella, Giuseppe Sofo, and Domenico Ursino [6] proposed a new network-based model to create a structured representation method of using keywords which generally representing unstructured data sources. In [18], Ma and Yuan suggest that the use of neural networks is rapidly growing with the trend of neural networks combining traditional algorithms and transformation learning technologies. In [20], Xin Li and Rob Law take advantage of a comprehensive network method to analyze and find out the current situation of big data research in the tourism industry by examining multi-disciplinary contributions relevant to big data.

As the network level continues to improve, the dimensions of the data are also increasing, so that subsequent modeling requires a dimensionality reduction algorithm. Simultaneously, the relevance can be easily analyzed using the multivariate data modeling methods [21]. For instance, the principal component analysis (PCA) is a good one.

PCA, a multivariate analysis technique, is generally used to transform the coordinates of a sample into another coordinate system that is ideal for analyzing data. It allows the identification and the representation of patterns in a data series and it can identify identities and differences, reducing the dimensionality without losing too much information [22]. Based on the above advantages, PCA has been widely used for the purpose of reducing dimensions. In [18], Ma and Yuan use the principal component analysis algorithm to extract information from image features, aiming at dimension reduction.

3 A Network Model Based on PCA

3.1 Technical details of PCA

Principal component analysis (PCA), a statistical analysis method, typically aims to convert multiple indicators into a few comprehensive indicators, using the concept of dimension reduction. By means of this method, we not only cut back the number of variables, but also reflects most of the original variables with fewer principal components.

In this model, in order to find factors related to the company's ability or performance level (z), let each factor be x_i (i=1,2,3, ...,p).By means of coordinate transformation, the original related variables x_i are normalized and then linearly combined to transform into another set of unrelated variables y_i. Here:

$$\begin{cases} y_1 = \mu_{11}x_1 + \mu_{12}x_2 + \mu_{13}x_3 + \ldots + \mu_{1p}x_p \\ y_2 = \mu_{21}x_1 + \mu_{22}x_2 + \mu_{23}x_3 + \ldots + \mu_{2p}x_p \\ y_3 = \mu_{31}x_1 + \mu_{32}x_2 + \mu_{33}x_3 + \ldots + \mu_{3p}x_p \\ \ldots\ \ldots \\ y_p = \mu_{p1}x_1 + \mu_{p2}x_2 + \mu_{p3}x_3 + \ldots + \mu_{pp}x_p \end{cases}$$

In the above formula, $\mu^2_{i1} + \mu^2_{i2} + \mu^2_{i3} + \ldots + \mu^2_{ip} = 1$ (i=1,2,3 3..., p).

Since there are countless kinds of the above transformations, for the purpose of obtaining the best calculation results, μ_{ij} in the above-mentioned formula is solved according to the following principles:

a) y_i and y_j ($i \neq j$, $j = 1, 2, 3..., p$) should be independent of each other;

b) y_1 refers to be one of the linear combinations of x_1, x_2, x_3, ..., x_p with the largest variance; y_2(not related to y1) is one of the linear combinations of x_1, x_2, x_3, ..., x_p with the second largest variance. Analogously, yp(not related to y_1, y_2, y_3, ..., y_p) has the smallest variance among all linear combinations of x_1, x_2, x_3, ..., y_p .

The variables y_i defined according to the above principles are sequentially called the first, second, third, ... principal components of the original variables x_i. Among them, y_1 holds the largest proportion in the total variance, and it has the strongest ability to integrate the original variables x_1, x_2, x_3, ... x_p, and the remaining principal components y_2, y_3, ..., y_p gradually decrease in proportion to the total variance. That is, their ability to integrate the original variables x_1, x_2, x_3, ..., x_p weakens in turn.

In the practical application of principal component analysis, generally only the previous few principal components with larger variances are selected. We determine the number of principal components based on the cumulative variance contribution rate and the eigenvalues of principal components, combined with actual needs. Thus, we implement the reduction of dimension by keeping the components with higher variance and a larger amount of information, and removing components with little differences and inadequate information [18].

3.2 Build a network

Consider a situation where z can be explained by three principal components, and each principal component can be explained by two or three factors from different data sources, as shown in Fig. 1.

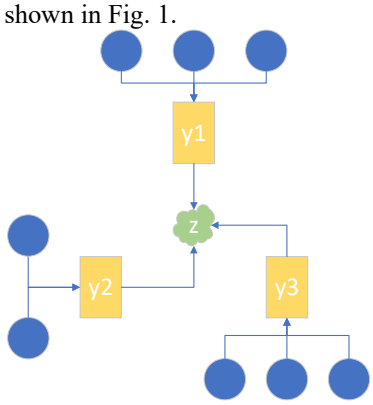

Figure 1. Graphical representation of single construct analysis

As we can see in Fig. 1, green dots represent the concept of company performance or capabilities; the selected principal components are colored in yellow; each blue node represents a factor related to the company's ability or performance level.

Going one step further, we connect the results of the principal component analysis of three certain constructs together as an example. The network is represented in Fig. 2.

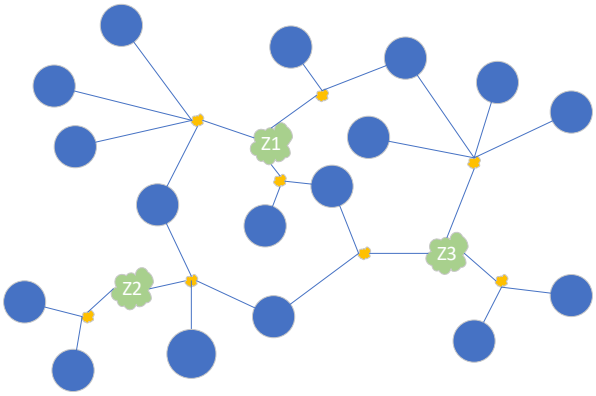

Figure 2. The network of three constructs analysis

4 Extraction of Visual Knowledge Patterns

In this section, we show how to extract the visual knowledge patterns, using the previously stated network model. Because the factors from different data sources and the target constructs are in the same network, extracting the pattern simply needs to calculate the shortest path between the two of the target constructs. Since in this model, the principal component y_i is the concept that we build, rather than the real data we extract from a data lake, it is ignored when calculating the shortest path. We discussed paths from z1 to z2, z2 to z3, and z3 to z1, in the network model proposed in the previous section, as is shown in Fig. 3, Fig. 4, and Fig. 5.

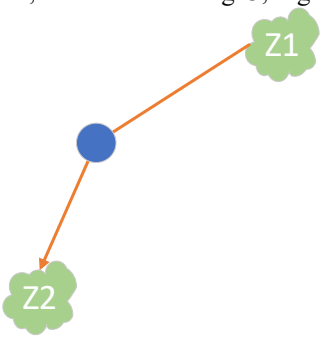

Figure 3. The konwledge pattern from z1 to z2

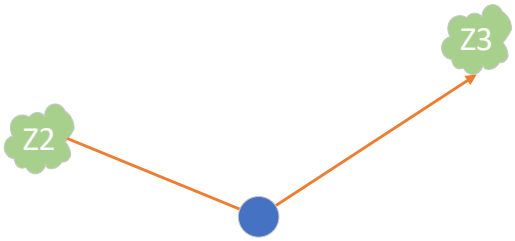

Figure 4. The konwledge pattern from z2 to z3

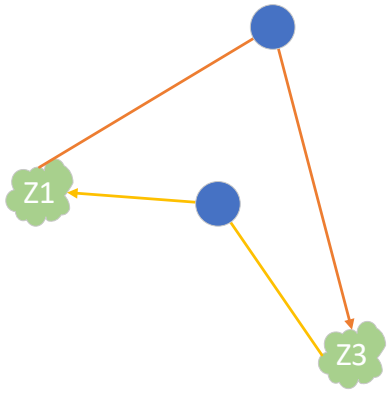

Figure 5. The konwledge pattern from z3 to z1

As we can see in the three visual knowledge patterns, it can be found that from one construct, through related factors, to another construct, at least one path can be drawn (there are two paths connecting z1 with z3). In real-world situations, it might be more complicated and not limited to studying the relationship among only a few constructs. More parameters will be considered in order to better and more accurately model the real situation.

As we mentioned in the introduction, the patterns extracted from the network model flexibly and intuitively show the relevance of the constructs. In the business context, it reflects the internal relationship between the various capabilities, development potentials and performances that companies are concerned about. Enterprise managers can use this method to effectively use the data in their databases plus external data and extract patterns to help them gain an in-depth understanding of their companies, surroundings and then make decisions effectively. Also, the approach can be broadened to more common areas to extract patterns from a certain network or complex networks [23], including social network, electrical transmission network, railway network, mobile signal network, and many other kinds of networks.

5 Conclusion

Nowadays, data lakes have been applied more and more widely, and some companies have started building their own data lakes. Compared with traditional data warehouses, data lakes have many clear advantages, but they do have some limitations as well. Helping organizations to enhance their management skills and effectiveness, in this paper, we put forward a network-based approach combined with PCA algorithm to extract visual knowledge patterns in a data lake scenario. As a result, the approach makes it more intuitive to appreciate the relationship between various kinds of performance and capabilities from heterogeneous business data. Thus, this method can enable enterprises to improve their competitiveness comprehensively.

The method we used in this paper can easily deal with traditional structured and semi-structured data, while it is difficult to handle the various types of unstructured data at the same time. Therefore, future work will continue to attach importance to novel and flexible approaches that can deal with all three types of data simultaneously from heterogeneous sources in data lakes.

Acknowledgment

The authors would like to acknowledge the excellent comments and suggestions provided by the anonymous reviewers.

References

1. Riccardo Rialti, Lamberto Zollo, Alberto Ferraris, Ilan Alon, Big data analytics capabilities and performance: Evidence from a moderated multi-mediation model, Technological Forecasting and Social Change 149 (2019) 119781.
2. Jiwat Ram, Changyu. Zhang, Andy Koronios, The Implications of Big Data Analytics on Business Intelligence: A Qualitative Study in China, Procedia Computer Science 87 (2019) 221-226.
3. Nadine Côrte-Real, Tiago Oliveira, Pedro Ruivo, Assessing business value of Big Data Analytics in European firms, Journal of Business Research 70 (2017) 379-390.
4. Endris K.M., Rohde P.D., Vidal ME., Auer S. (2019) Ontario: Federated Query Processing Against a Semantic Data Lake. In: Hartmann S., Küng J., Chakravarthy S., Anderst-Kotsis G., Tjoa A., Khalil I. (eds) Database and Expert Systems Applications. DEXA 2019. Lecture Notes in Computer Science, vol 11706. Springer, Cham
5. Yuanzhu Zhan, Kim Hua Tan, An analytic infrastructure for harvesting big data to enhance supply chain performance, European Journal of Operational Research, 281 (2020) 559-574.
6. Paolo Lo Giudice, Lorenzo Musarella, Giuseppe Sofo, Domenico Ursino, An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake, Information Sciences, 478 (2019) 606-626.
7. H. Fang, "Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem," in 2015 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, IEEE-CYBER 2015, 2015, pp. 820–824.
8. Dixon,J.:Pentaho,Hadoop,andDataLakes. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
9. Tyagi, P., Demirkan, H.: Data lakes: the biggest big data challenges. Analytics 9(6), 56–63 (2016)
10. Alserafi A., Abelló A., Romero O., Calders T. (2019) Keeping the Data Lake in Form: DS-kNN Datasets Categorization Using Proximity Mining. In: Schewe KD., Singh N. (eds) Model and Data Engineering.

MEDI 2019. Lecture Notes in Computer Science, vol 11815. Springer, Cham

11. Natalia Miloslavskaya, Alexander Tolstoy, Big Data, Fast Data and Data Lake Concepts, Procedia Computer Science. Sci. 88 (2016) 300-305.
12. Marilex Rea Llave, Data lakes in business intelligence: reporting from the trenches, Procedia Computer Science, 138 (2018) 516-524.
13. Mehmood, Hassan (University of Oulu, Finland); Gilman, Ekaterina; Cortes, Marta; Kostakos, Panos; Byrne, Andrew; Valta, Katerina; Tekes, Stavros; Riekki, Jukka Source: Proceedings - 2019 IEEE 35th International Conference on Data Engineering Workshops, ICDEW 2019, p 37-44, April 2019
14. Maccioni A., Torlone R. (2018) KAYAK: A Framework for Just-in-Time Data Preparation in a Data Lake. In: Krogstie J., Reijers H. (eds) Advanced Information Systems Engineering. CAiSE 2018. Lecture Notes in Computer Science, vol 10816. Springer, Cham
15. Wibowo M., Sulaiman S., Shamsuddin S.M. (2017) Machine Learning in Data Lake for Combining Data Silos. In: Tan Y., Takagi H., Shi Y. (eds) Data Mining and Big Data. DMBD 2017. Lecture Notes in Computer Science, vol 10387. Springer, Cham
16. M. Farid, A. Roatis, I. F. Ilyas, H.-F. Hoffmann, and X. Chu, "CLAMS: Bringing Quality to Data Lakes," Proceedings of the 2016 International Conference on Management of Data SIGMOD 16, 2016.
17. Jian Liu, X.X. Zhang, Lei Zhang, Tree pattern matching in heterogeneous fuzzy XML databases, Knowledge-Based Systems, 122 (2017) 119-130.
18. Ji Ma, Yuyu Yuan, Dimension reduction of image deep feature using PCA, Journal of Visual Communication and Image Representation, 63 (2019)
19. Y.Yuan, G.Wang, L.Chen, B.Ning, Efficient pattern matching on bigun certain graphs, Inf.Sci.339(2016)369–394.
20. Xin Li, Rob Law, Network analysis of big data research in tourism, Tourism Management Perspectives, 33 (2020)
21. Qiang Liu, Dezhi Kong, S. Joe Qin, Quan Xu, Map-Reduce Decentralized PCA for Big Data Monitoring and Diagnosis of Faults in High-Speed Train Bearings IFAC-Papers OnLine, 51 (2018) 144-149.
22. Cheilane T. de Souza, Sarah A.R. Soares, Antonio F.S. Queiroz, Ana M.P. dos Santos, Sergio L.C. Ferreira, Determination and evaluation of the mineral composition of breadfruit (Artocarpus altilis) using multivariate analysis technique, Microchemical Journal, 128(2016) 84-88,
23. Nasrin Kalanat, Eynollah Khanjari, Extracting actionable knowledge from social networks with node attributes, Expert Systems with Applications: X, 3(2019) 100013.