The breakthrough of the Big Data Process Model under the Fourth Paradigm

Bingru Yang*
School of Computer & Communication Engineering
University of Science and Technology Beijing
Beijing, China
bryang_kd@126.com

***Abstract*—at the critical point of data accumulation from quantitative change process to qualitative change process (i.e. data intelligence), profound philosophical speculation reveals that: BDM work is most critical and valuable at this critical point [33] [34]. BDM which is in perplexed state shall not and unable to continue massive data mining mainstream development track----efficient and scalable algorithm and its application research; because process model for algorithm operation and more important creative research of internal mechanism exist. With this fact, and based on massive data mining research for nearly more than 20 years [22][32], this thesis is intended to achieve breakthrough of BDM mechanism and model which has an essential distinction with massive data mining against the fourth paradigm, namely providing relatively systematic and complete overall structure of process model and laying a foundation for subsequent explorative and guiding methodology research on certain specific links.**

Keywords-big data mining; overall process;creative idea;Fourth Paradigm;Process Model

Main text: due to the inherent characteristics of big data, actual development of the original massive data mining and limitation in human cognition, BDM is logically bound to encounter theoretical bottleneck and application dilemma that are difficult to break through. However, current basic situation of BDM development may reappear with the conclusion made by scientists on Washington Conference (KD & DM) on August 27, 2003: "in a long run of scientific development, the greatest obstacle may be the lack in basic theories and explicit clarification of the problems and challenges we are facing". "The application of Web and hype from manufacturers will influence the development of this field in short term, the fundamental research of KDD must target at solving real fundamental problems of KDD while eradicating those disturbances." [22][32]

Based on above background and on the premise of expanding BDM definition basically agreed and strictly defining distinction between BDM and big data analysis; under scientific research fourth paradigm which is complementary to experimental science, theoretical deduction and analog simulation----- meaning of data-intensive scientific BDM discovery, discuss as follows.

I. Overall Framework of BDM Process Model

Overall framework design of process model mainly depends on the following five points:

(1) Internal mechanism of BDM (as mentioned above).

(2) Features of big data itself: features of big data: large scale (marking A), great variety (marking B), fast speed (marking C) and low value density (marking D).

*[Introduction to Author] Bingru Yang, male, lifetime professor and doctorial supervisor of School of Computer & Communication Engineering, University of Science and Technology Beijing, China; director of Knowledge Engineering Institute.

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.11

(3) Contradiction analysis of features: data comes in the way of one or multiple streams, in case of not processing or storing timely, data is lost forever, existence is primary. So, (C) Extremely quickly speed, it is impossible to store whole in movable memory or interact in selected time, and (D) ----"Abandon wastes, relieve burden"; (B) distinguish type theory; (A) ---conduct "higher unit transformed into lower unit" by type. So processing order is C、D→B→A. Such analysis will be presented in later "overall structure diagram of BDM process model".

(4) Technical framework [33] infrastructure [35], fundamental resource [19], basic technique [33], ubiquitous technique, etc. (be omitted)

(5) Modern research reveals that multilayer hierarchical structure is the most effective processing method for reducing system complexity, while sequential granularity space theory is one of most effective method for establishing multilayer hierarchical structure of complex system.

In conclusion, the brand-new overall framework of BDM process model differing from massive data mining essentially is proposed (as shown in Figure 1).

II..Interpretation

1)Virtual collaborative filter (sieve) layer:

(1) Stream filtration [4] [6] [8] 12][13]:

a) Bloom filter technology: Remove most of tuples which can not meet the standard by hash operation.

b) Web page filtration: Iteratively compute the authority and navigation degree of each page to make choice.

(2) Sieve of data field [19]:

a) In data field, the relation of data set scale and radiation coefficient is shown in Figure 2: [19] [27][33]

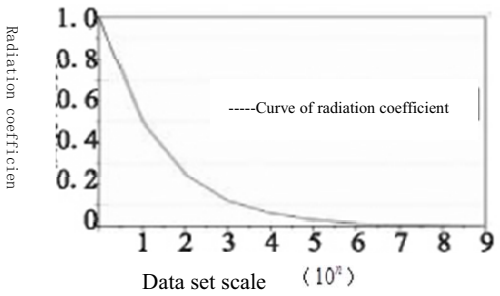

Figure 2 Relationship Diagram of Radiation Coefficient and Data Sets Scale

b) According to the Euclidian distance between σ and node, any value of nodal potential function can be calculated:

$$\varphi(v_i)=\sum_{j=1}^{n} e^{\frac{\|d(i,j)^2\|}{2\sigma^2}}$$

For higher potential function value at a certain node position in data stream, approval structure can be formed; for lower value, removal structure can be formed.

c) Allow all key values to pass the streaming element in S, and block most of key values passing the streaming element in S. (namely, if the corresponding bytes are all 1, the streaming element will be allowed to pass; otherwise, it will be refused).

(3) Sieve of information entropy: Thermodynamic entropy is applicable in studying the distribution law of massive particles. After transforming entropy[32] (the parameter for measuring unorganized degree) into knowledge information entropy, describe the process of data in data mining transforming to knowledge; screen out when information content required for any element classification in sample space less than Iv (see information quantity theorem in 1.2).

(4) Sieve of chain of causality [22][21]: Define a partial ordering relation between various causal metamorphoses to form the

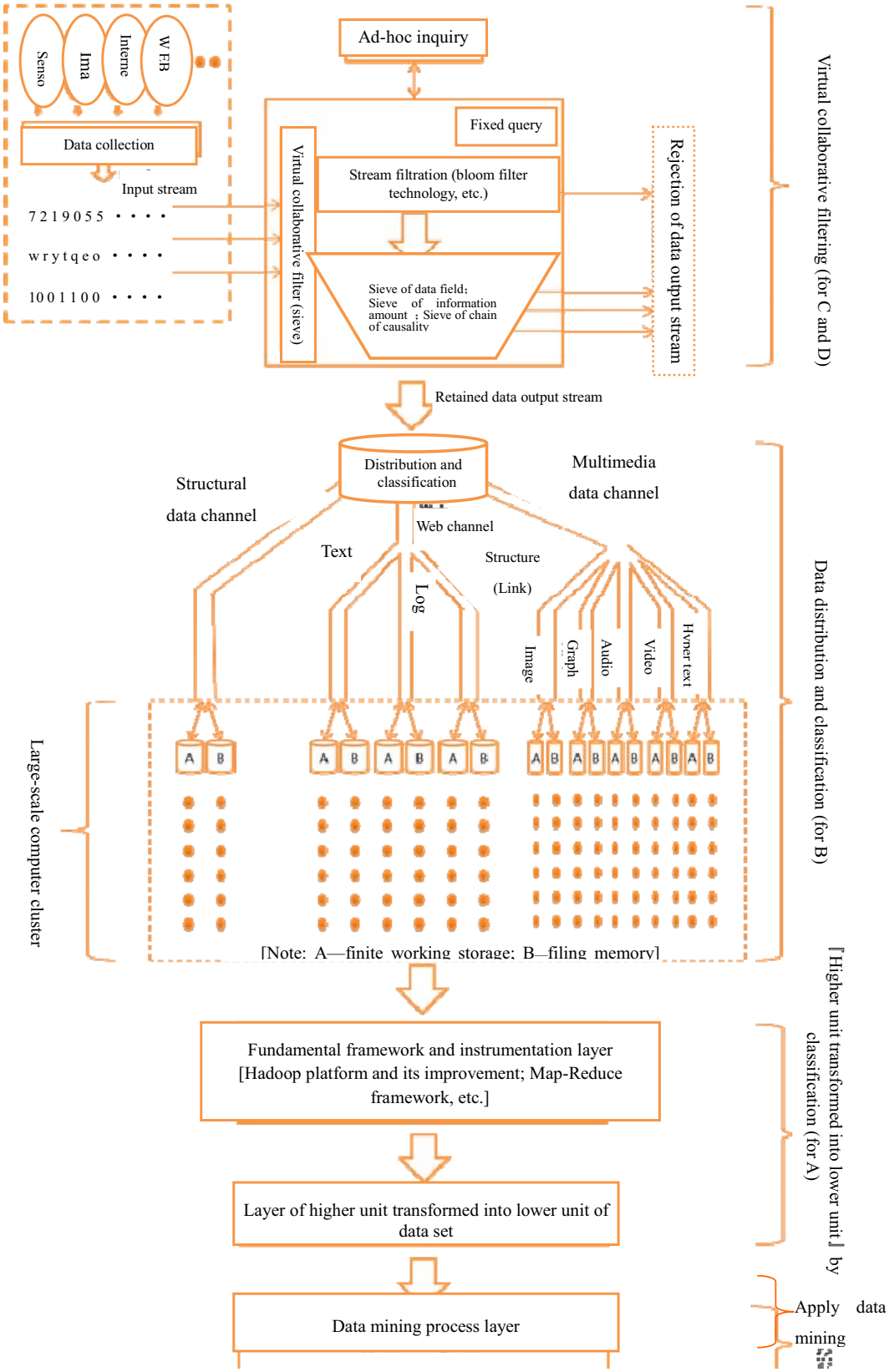

Figure 1 The general architecture graph of BDM process model

"causal metamorphosis chain table". The data element is deemed as "cause" and the mining target is deemed as "effect"; angles of view shall be obtained to determine the respective sequence of major cause by result reason method—it shall be abandoned while the data series which have major impact on mining target shall be retained.

2) Distribution and classification layer: on the basis of ideology of distinguishing type theory. (Classifying by structurization, Web, multimedia data and refinement)

3) Fundamental framework and instrumentation layer: after distribution, various data shall enter Hadoop server and respective memory of huge cluster of computers according to the new concept of form progressive. [19] [33] [35]

4) Layer of "higher unit transformed into lower unit" of data sets:

(1) "Method of magnetic effect"[29] :

a) If the mining task T and precision δ are given, the "minimum" data subset K ≤D (D is a real data set) will exist. Its potential is Ω (Ω<|D|), which makes T carrying out mining task in K have precision of δ at least, and Ω is estimable, so K is the "core set" of D. K is essentially composed of heavy quality data in data field, Ω is probably obtained by optimizing two objective functions (quadratic programming), namely the min J (under language field) and max Iv (under information entropy).

b) Construction algorithm of K: According to the estimated value Ω, make finite extensions (carry out "magnetic attraction" by "semantic measure") for some initial data samples (known as "nucleus attractor") until its potential has reached to Ω.

(2) "Intersection method" [16] [22]: According to the "Double-base cooperating mechanism", under the description of linguistic values of given mining task: i) the layer (value) of sub-catalog structure corresponding to knowledge node of relevant linguistic value is intersected; ii) according to the record in relational data base corresponding to intersection (value), create new mining data set.

(3) "Focusing method"[32] : When the user interest (or OLAP, etc.) and "knowledge deficiency" (acquired by incidence matrix of directed hyper-graph---large-scale sparse matrix) exist at the same time and are identical, the record set in corresponding relational data base consists of the mining data subset.

(4) "information entropy method ": achieve the goal of forecasting mining effect and finding minimum data size for mining with relationship of knowledge entropy and rule intensity; seek for cluster center with "big data set quick spectral clustering method based on CCMEB" and so forth and then form proper subset for mining with the method of "nuclear attractor".

(5) Several available and effective technologies such as data compression, dimension reduction, attribute reduction and record reduction [4] [11].

5) Data mining process layer [4] [12] [14]:

Mining task (scene imagination) → preprocessing (similar to massive data mining) → multi-element focusing (interestingness + OLAP + knowledge deficiency +…) → various process models (mainly including two categories: ① For structural data mining --KDD* Process [32][16]; ②For complex data mining--DFSSM Process Model[28][32]) → various algorithms (mainly including two categories: ① "higher unit transformed into lower unit" post-processing, several algorithms of massive data mining can be used continuously [4][9][10][11][26][18][30][31][36][37] ; ② primary exploration of several creative algorithms under big data background

[17][21][13][6][2] [19][12][23]) → post-processing[20][1][5][15].

References

[1]Kim Y G., Suh J H, Park S C. Visualization of patent analysis for emerging technology[J]. Expert Systems with Applications, 2008, 34(3): 1804-1812

[2]Kleinberg J M., Authoritative sources in a hyperlinked environment [J]. Journal of the ACM (JACM), 1999, 46(5): 604-632.

[3]Yoon B, Park Y., A text-mining-based patent network: Analytical tool for high-technology trend [J]. The Journal of High Technology Management Research, 2004, 15(1): 37-50.

[4] J.W. Han, Data Mining: Concepts and Techniques [M]. Second Edition, (China Machine Press, 2006).

[5] N. Siddiqi, Credit risk scorecards: developing and implementing intelligent credit scoring [M] (John Wiley & Sons, 2005).

[6] Rajaraman A., Big Data: Mining and Distributing Process of Massive Datasets [J] (2012).

[7] M. S. Victor, C. Kenneth, Big Data: A Revolution that will transform How We Live, Work and Think. Zhejiang People's publishing House, 2013.

[8] Linden G., Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. Internet Computing, IEEE, 2003, 7(1): 76-80.

[9] [Portuguese] Liuis Torgo, Data Mining with R: Learning with Case Studies [M] (China Machine Press, 2013).

[10] Atkeson C G., Moore A.,and Schaal S., Locally weighted learning[J]. Artificial Intelligence Review, 1999.

[11]Hornik K., Buchta C., and Zeileis A., Open-source machine learning: R meets Weka[J]. Computational Statistics, 2009, 24(2): 225-232.

[12]Hong S., Use of contextual information for feature ranking and discretization [J]. Knowledge and Data Engineering, IEEE Transactions on, 1997, 9(5): 718-730.

[13] Klinkenberg R., Learning drifting concepts: Example selection vs. example weighting [J]. Intelligent Data Analysis, 2004, 8(3): 281-300.

[14] Weiss G., Provost F., Learning when training data are costly: The effect of class distribution on tree induction [J]. J. Artif. Intell. Res. (JAIR), 2003, 19: 315-354.

[15] [American] Douglas W. Hubbard, Datamation Decision-Big Data Time [M].1. (World Publishing Corporation, 2013).

[16] B. R. Yang, H.H. Sun, The Mining Association Rules Algorithm Based on Double-base Cooperating Mechanism Maradbcm[J]. Comput. Res. Devel., 39(11): 1447-1455 (2002).

[17] S. F. Lu, Z. D. Lu, Fast Mining Maximum Frequent Itemsets, [J]. Journal of Software, 12(2): 293-297 (2001).

[18] Agrawal R., Srikant R., Mining sequential patterns[C]//Data Engineering, 1995. Proceedings of the Eleventh International Conference on. IEEE, 1995: 3-14.

[19] Gang-yi W S D., Ming Z., On Spatial Data Mining under Big Data[J]. Journal of China Academy of Electronics and Information Technology, 2013, 1: 003.

[20] Hernández M A., Stolfo S J., Real-world data is dirty: Data cleansing and the merge/purge problem [J]. Data mining and knowledge discovery, 1998, 2(1): 9-37.

[21] Reshef D N., et al., Detecting novel associations in large data sets [J]. science, 2011, 334(6062): 1518-1524.

[22] B. R. Yang, Knowledge Discovery Theory Based on Inner Mechanism [M]. (National Defense Industry Press, 2009).

[23] Benaroch M., Toward the notion of a knowledge repository for financial risk management [J]. Knowledge and Data Engineering, IEEE Transactions on, 1997, 9(1): 161-167.

[24] Bingru Y., Fanlun X., KD (D&K) and Double-Bases Cooperating Mechanism [J]. Systems Engineering and Electronics, Journal of, 1999, 10(2): 48-54.

[25] Jicheng W., Jinggui P., Fuyan Z., Research on Web Text Mining[J].Journal of Computer Research & Development, 2000, 37(5):513-520.

[26] Roussinov D., Zhao J L., Automatic discovery of similarity relationships through web mining [J]. Decision Support Systems, 2003, 35(1): 149-166.

[27] Bing-ru Y., FIA and CASE based on fuzzy language field [J]. Fuzzy sets and systems, 1998, 95(1): 83-89.

[28] B. R. Yang, J. Tang. The Research of Discovery Feature Sub-space Model (DFSSM) Based on Complex Type Data [J] Eng. Sci., 5(1): 56-61 (2003).

[29] B. R. Yang, Two Conjectures of Data Mining. Comm. Comput. Chinese and English version, 4(4): 1-4 (2007).

[30] Yang B R., Shen J, Song W., KDK based double-basis fusion mechanism and its process model [J]. International Journal on Artificial Intelligence Tools, 2005, 14(03): 399-423.

[31] Yang B R., Song W, Xu Z Y. ,New construction for expert system based on innovative knowledge discovery technology[J]. Science in China Series F: Information Sciences, 2007, 50(1): 29-40.

[32] Yang B R., Knowledge Discovery Based on Theory of Inner Cognition Mechanism and Application [J]—Construction , Realization and Application. Elliott & Fitzpatrick, Inc.(USA) ,2004.

[33] Y. Z. Wang, X. L. Jin, Network Big Data: Present and Future [J] Chinese J. Comput., 36(6): 1125-1138, (2013).

[34] Hot Issues of Big Data and Analysis on 2013 Development Tendency, CCCF, (12) 8, (2012).

[35] Nimmagadda S.L., Dreher H.V., Big-data integration methodologies for effective management and data mining of petroleum digital ecosystems, IEEE DEST(7th),2013:148-53.

[36]Lin Yu, Maximum frequent itemsets mining algorithms in large data environment, the power industry informatization excellent essays, 2013.

[37]Li Yue, Liu Ran, Analytic hierarchy process (ahp) construct the mining model in the treatment of the efficiency in the process of multidimensional data mining research, Proceedings of 2013 the Fourth International Conference on Information ,Communication and Education Application (ICEA 2013),Vol.31,2013.

[38] B. R. Yang, , J.X.Wang, Study on the double base cooperating mechanism in KDD (Ⅰ) [J] Eng. Sci., 4(4) , 2002:41-51, 57.