Paper Reports

Siyu Zhou
100995161

1 Report on "An Overview of Business Intelligence Technology" [1]

Abstract.

This paper introduces ten technologies used in BI. Four of them are mature but still having challenge problems while another six are new technologies with open research challenges.

Summary of [1].

Data Storage includes two aspects: access structure and data compression. Index structure is through the index scan to get the value of rows while column-oriented storage just scans columns. Index structure can reduce the operation to the base tables if index have enough information. Column-oriented storage has greater data compression ability than row-oriented store.Materialized views answer to the query directly while partitioning increasing manageability by making tables and indexes smaller. Based on automated physical tools, users choose the most suitable access structure from above four methods, according to their work requirements. Data compression reduces the amount of scanned data and required storage while sometimes uses compressed data to answer the query directly. The author also introduced compression techniques: null suppression, dictionary compression and run-length encoding. Query Processing includes OLAP servers, Relation Servers, Distributed Systems using MapReduce Paradigm and Near Real-Time BI.OLAP servers includes MOLAP servers,ROLAP servers,HOLAP servers and in-memory BI engines. MOLAP servers are helpful for storage parse data while ROLAP servers using snowflake schema to express data in multidimensional and query it by SQL. HOLAP stores data liking ROLAP servers and precomputing in MOLAP. In-memory BI engines are not suitable to update data but suitable to read data-only which can be used for OLAP engines to response queries interactive. Relation Servers provides query optimization and parallel processing and appliances working for complex SQL queries.Near Real-Time BI is used for reduce the time gap between acquired operational data and analysis result. Enterprise Search is keyword searching from various data sources and formats.Data Profiling and de-duplication are used for improving data quality which is used in ETL tools to load large volumes data.ETL includes data quality,accurately extracting structure,deduplication and data load and refresh. Data mining is always used for constructing prediction mode while text analysis can extract data from the text on the Internet. Cloud data services take advantages of virtual machine to improve query efficiency.

My Personal View of [1]

After reading this paper I am interested in MBI, even though it is just be mentioned in one sentence. All the ten technologies discussed in this paper also important to MBI.I think the most useful one is cloud data service. Since one of the aims of MBI is real time (Rouse, 2010), cloud can provide parallel computing for queries while no constraints in hardware. MBI also need to consider security. Encryption and authorization play an important role in vendors no matter how what delivery methods used (Rouse, 2010). I think we also should consider security from mobile itself. Since mobile is more easily to be lost, how should we protect the query answers in the MBI apps. Finally, I think we also should consider the usability of MBI. When users want to make decisions from query answers, they need the supports from graphics and tables. How to make analytic results show clearly in mobile screen is also need to explore.

2 Report on "A Few Useful Things to Know About Machine Learning" [2]

Abstract.

This paper introduces twelve lessons from folk knowledge in machining learning which summarized by the author. It includes the relation between data amount, learning algorithm and learning process. The focus of this paper is classification.

Summary of [2].

This paper tell us twelve lessons about classification in machine learning

The first lesson told us when choosing learning algorithm, learner should consider three aspects: representation, evaluation and optimization. After evaluation, learner can find the weakness of the algorithm, and then optimize it. The second rule is its generalization that counts, which reflects that the classification should be general and avoid overfitting. The third lesson told us when to do classification, only having data is not enough, learner should have data related background. The fourth lesson states that overfitting may in many forms and difficult to find it immediately. However, cross-validation and a regularization term are helpful to alleviate overfitting. The fifth lesson is Intuition fails in high dimension. It is easy to construct classifier in two or three dimensions, but it is difficult to do it in high dimension and understand what happens there. The sixth lesson tells us theoretical is helpful for learners to understand the design of algorithm but not work as the reason for the practical of machine learning. The seventh lesson is about feature selection. Some features look irrelevant in one by one doesnt mean the combination of them is also irrelevant. The eighth lesson tells us dumb algorithm with more data is better than clever one with fewer data. But in reality, learner may not get enough data or have constraints to manage large amount data. The ninth lesson is better to learn from more models by the methods of bagging, boosting and stacking. The tenth lesson is simplicity does not imply accuracy. There is no relation between the number of parameters and accuracy. The eleventh lesson told us representable just means classifier can be handled by computer but not means it can be learned. The last lesson exposes that correlation is just the symbol of relation not for the causation.

My Personal View of [2]

After reading this paper, I have some thinking about evaluation and learning algorithms. I think evaluation should consider the constructed time of model(Han et al.,2012). For example, the construction time of neural network is long which is not suitable for large amount training data. I also think evaluation should consider interpretability(Han et al.,2012). If learner gets some classifications by one algorithm, when put these classifications into practical use, it should be explained clearly how to get these classification and why it is correct. Decision tree is interpretable based on its clearly graph. Conversely, neural network is difficult to explain. The other part of my opinion is about choosing learning algorithm. I think it is confused to choose which one is more suitable. Every method has its strengths and drawbacks. According to what I said about neural network, it seems like not a good algorithm. But it has a wide application in deep learning. So, I it is worthwhile to explore how to choose the suitable algorithm for the specific situation.

References

[1] Surajit Chaudhuri, Umeshwar Dayal, Vivek R. Narasayya. An Overview of Business Intelligence Technology. *Communications of the ACM*, 2011, 54(8):88-98.

[2] Pedro Domingos. A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, 2012, 55(10):78-87.

[3] H.V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi. Big Data and Its Technical Challenges. *Communications of the ACM*, 2014, 57(7):86-94.

[4] Rouse, M.(2010, September).Mobile business intelligence. Retrieved December 10,2016 from http://searchbusinessanalytics.techtarget.com/definition/mobile-business-intelligence

[5] Han, J., Kamber, M.,Pei, J.(2012). Data mining concepts and techniques(3rd edition). Retrieved December 11,2016 from http://www.cs.uiuc.edu/ hanj/bk3/