Vol.2012-DBS-155 No.17
2012/11/19

確率推論処理を有するストリーム処理実行方式の検討

川島 英之[†1]

ベイジアンネットワークにおいて厳密解を求める確率伝播法としてメッセージパッシング法がある．この技法はノード数・生起事象数の増加に伴い計算時間が増大する．ストリームデータ処理では高速性が求められるため，同技法の利用は好ましくない．この問題を解決するために，確率伝播の事前計算化法を示す．第一の技法は確率値を事前に計算しておき，問合せ評価時に表探索を行う．第二の技法はλ値とπ値を事前計算しておき，問合せ評価時に表探索と確率値を求める．第二の技法はノード数・生起事象数の増加に対して性能劣化度が優れる性質を示す．本発表で手法の評価実験結果，効率的な事前計算アルゴリズム，スケジューリングアルゴリズム，ならびにストリーム処理システムへの実装に関して述べる．

Consideration of Stream Data Processing with Probabilistic Reasoning

Hideyuki Kawashima[†1]

Message passing technique is a belief propagation method which provides strict answers on Bayesian networks. Using the technique, computation time increases according to the number of nodes and the number of occurring events. Since stream data processing should provide high performance, adopting the technique is not preferred. To solve the problem, we describe prior computation techniques of belief propagation. First method computes probabilities in prior, and it executes table lookup on query evaluation. Second method computes λ value and π value in prior, and it executes table lookup and probability computation on query evaluation. In addition, we describe experimental results, an efficient prior computation algorithm, operator scheduling algorithm, and implementation on a real stream data processing system.

1. Introduction

To accurately obtain occurrence probability in a Bayesian network, message passing algorithm should be used. Unfortunately it is known that message passing requires long time since (1) it should propagate a message to most of nodes in a DAG, and (2) the procedure should be conducted for all of occurrence nodes. When a Bayesian network is applied for stream data processing, the occurrence probability should be obtained with low latency. The number of nodes in a Bayesian network can be large because of the progress with LDPC or recommendation by association analysis. Therefore the acceleration of probability computation has been a severe problem. To deal with the problem, this paper proposes two acceleration techniques based on prior computation. The first proposal computes probabilities in prior, and it executes a simple table lookup operation on query evaluation. Since the address of occurrence patterns is obtained offset computation, the lookup requires just a single disk access. Though it provides excellent performance, it is not scalable. It is because for k nodes, the table should have $O(v^k)$ entries for v variables.

Our second proposal solves the scalability issue. Instead of storing probabilities, second proposal stores λ value and π value. On a message passing algorithm, with the occurrence of an event, lambda messages and pi messages are sent to upper and lower directions respectively. Then λ value and pi value are computed in each node. Finally, occurrence probability is computed by using lambda values and pi values in a node. This strategy requires just $O(k^2)$ entries for the lookup table with a little degree of degradation on time complexity. To the best of our knowledge, this is the first proposal which accelerates to compute accurate probability values by prior computation. Most of acceleration techniques adopt approximation strategies. Our previous study [1] adopts omission of message passing with a restriction by considering retrieval targets. It should be noted that the idea of prior computation is novel, and it does not require any restrictions.

2. Proposal

In our presentation, we will show experimental results on our prior computation method, an efficient prior computation algorithm, operator scheduling algorithm, and implementation on a real stream data processing system.

Especially, it should be noted that a Bayesian network generates may generate massively amount of tuples. This nature is different from usual relational operators. Even if cartesian product, it generates at most O(mn) tuples. However a Bayesian network may generate more than n tuples for an input. This phenomenon can be observed by other data mining techniques such as classification techniques or clustering techniques.

[†1] 筑波大学
University of Tsukuba.

[†1] 筑波大学
University of Tsukuba.

To cope with this problem, operator scheduling should be refined. In the field of stream data processing, Chain [2] was proposed for relational data stream processing. However it does not cover our motivating operators. For our problem, we think selection combination to Bayesian network operator and intra operator scheduling is effective to realize pipelined data processing which is preferred in stream data processing rather than materialized strategy.

Acknowledgement

The work is partially supported by "New generation network R&D program for innovative network virtualization platform and its application(s)", the Commissioned Research of National Institute of Information and Communications Technology (NICT), and KAKENHI (#24240015).

References

1) Ryo Sato, Hideyuki Kawashima, Hiroyuki Kitagawa, "The Integration of Data Streams with Probabilities and a Relational Database using Bayesian Networks", Proc. SeNTIE'08.
2) Brian Babcock, Shivnath Babu, Rajeev Motwani, and Mayur Datar. 2003. Chain: operator scheduling for memory minimization in data stream systems. SIGMOD'03.