2003年日本オペレーションズ・リサーチ学会
秋季研究発表会

マーケットバスケット分析のための アルゴリズムの比較と提案

01601360　慶應義塾大学　森雅夫　MORI Masao
02005620　慶應義塾大学　*阿部島誉幸　ABESHIMA Takayuki

0　はじめに

今日、センサーやコンピュータの発達によって、企業の持つデータベースの情報量が巨大化している。その巨大化したデータベースなどの大量のデータを用いて、分析を行うデータマイニングという手法が発達してきた。この研究では、その中の“マーケットバスケット分析”の手法について取り扱っている。

1　マーケットバスケット分析

“マーケットバスケット分析”は顧客が何と何を同時に購入するか(購入する可能性が高いか)を分析する手法である。その結果を用いて販売計画などに役立てることが、“マーケットバスケット分析”の目的である。

この問題では、おっ客が購入するもの(分析の対象)を「アイテム」と呼び、その集まりを「アイテムセット」、顧客それぞれのバスケットを「トランザクション」と呼ぶ。アイテムセットが購入されている割合を「サポート」と呼び、研究者はこの「サポート」の最低値(ミニマムサポート)を設定する。この「ミニマムサポート」の値をあるアイテムセットの「サポート」が超えていたら、「大きい」アイテムセットであると判断し、相関ルールを調べる際の分析対象にする。

2　アルゴリズムの解説

「大きい」アイテムセットを判別しなくてはならない。この中でもAprioriのアルゴリズム(1994)が著名ある。このAprioriのアルゴリズムを改良したDynamic Itemset Counting (DIC)(1997)も知られている。また違うアプローチを用いた方法として、Vertical Search(2001)がある。この研究では、これら3つのアルゴリズムの良い点を組み合わせることによって vDIC という新しいアルゴリズムを提案する。

2.1　Apriori

トランザクションの数が100万、10億といった膨大な量のデータベースの中から相関ルールを見つけるためにアイテムのすべての組み合わせについて調べていたら、時間がかかり過ぎる。そこで1994年にIBMのアルマデン研究所のR. AgrawalらはAprioriというアルゴリズムを提唱した。このアルゴリズムの高速化の原理は、例えばA, B, Cという3つのアイテムがあるとき、もし{A, B}(AとBを同時に購入する)が「小さい」のとき、{A, B, C}は調べるまでもなく「小さい」ということである。このことによって調べる必要のあるアイテムセットの数

を減らし、計算時間を短縮させるアルゴリズムである。

2.2 Dynamic Itemset Counting

DICのアルゴリズムは1997年にS. Brinらによって開発されたアルゴリズムである。このアルゴリズムは、Apriori の原理に加えて、データに偏りがないときには1部を取り出して調べてもあるアイテムセットのサポートの値には大差ないということを利用し、推測、並列処理(、修正)をすることによって Apriori よりも速いアルゴリズムとしている。

しかし、DIC においては並列処理をする際の分割数というものも問題になってくる。分割数を大きくすると、当然誤判別が増加し、それを修正するための時間が多くかかってしまうことになる。

2.3 Vertical Seearch

このアルゴリズムは2001年にZ.Huらによって開発されたアルゴリズムで、Apriori や DIC と違い、データとしてトランザクションごとに並べたものではなく、アイテムごとに並べたものを使う。またこのとき、アイテムを購入されている回数が多い順に並び替え、そのことによって「小さい」アイテムがでてきたときにそれ以降を調べなくてもすむようにもしている。このようにすることによってトランザクション全体を調べる回数を減らし、計算時間を短縮しているのである。

2.4 vDIC

これは上記の2つのアルゴリズム、DIC と Vertical Search をあわせたアルゴリズムで、Vertical Search のようにアイテムごとにアイテムを並び替え、DIC と同じように推測、並列処理、(修正)を用いて上記の2つのアルゴリズムより更に計算時間を短縮するアルゴリズムである。

3 シミュレーション

2章で示した4つのアルゴリズムを、作成したデータに適用し、その計算時間を比較した。またこのときのデータは、トランザクション数が50,000、アイテム数が20、各々のアイテムを購入する確率を30%としたデータである。実験に際し、ミニマムサポートの値によって計算時間が異なってくるので、ミニマムサポートの値を5%、10%、20%、30%、40%と5段階について実験した。その結果は表1のようになる。表1を見ても明らかであるが、vDICがもっとも計算時間が短いということがわかる。

表1：4つのアルゴリズムのパフォーマンス (秒)

	5%	10%	20%	30%	40%
Apriori	113.31	26.91	27.10	21.62	14.81
DIC	104.89	23.30	14.89	25.60	2.69
Vertical Search	41.13	6.48	6.54	2.69	2.10
vDIC	27.89	3.00	3.00	1.39	0.59

4　実際データへの適用

次にそれぞれのアルゴリズムを日用品に関する実際のデータに適用した。このときアイテムはカテゴリー別に分類し、アイテム数を 22、トランザクション数を 2,015,634 とした。この結果は表 2 のようになる。また表 2 を見れば明らかであるが、vDIC がもっとも計算時間が短い。またデータをトランザクションごとからアイテムごとに並び替えるためにかかる時間は 58.31 秒である。

表 2：4 つのアルゴリズムのパフォーマンス　(秒)

	5%	10%	20%	30%	40%
Apriori	646.17	580.46	553.09	553.29	537.07
DIC	218.57	73.36	47.06	47.27	42.09
Vertical Search	34.63	21.95	15.83	16.07	15.69
vDIC	31.10	17.06	10.41	10.41	9.37

5　まとめと課題

提案したアルゴリズム、つまり vDIC を使用することによって、マーケットバスケット分析を行う際の「大きい」アイテムセットを判別するためにかかる時間を Apriori のものに比べて 20 倍以上速くすることができた(表 2 参照)。

6　参考文献

[1] R.Agrawal and R.Srikant "Fast Algorithms for Mining Association Rules" 1994

[2] S.Brin, R.Motwani, J.D.Ullman and S.Tsur "Dynamic Itemset Counting and Implication Rules for Market Basket Data" 1997

[3] Z.Hu, W.Chin and M.Tkakeichi "Calculating a New Data Mining Algorithm for Market Basket Analysis" 2001

[4] M. J.A. Berry(原著) Gordon Linoff(原著) 江原淳(訳) 佐藤栄作(訳) SAS インスティテュートジャパン(訳)『データマイニング手法－営業、マーケティング、カスタマーサポートのための顧客分析』 1999