マーケット・バスケット

アソシエーション分析（association analysis）は、連関分析ともいわれ、データマイニングの分析手法の中核を成すものです。もう一つよく使われる「マーケットバスケット分析」は、アソシエーション分析のひとつです。スーパーやコンビニなどの買い物かごに何が一緒に入っているか、何と何が一緒に買われるかを分析する手法で、POSシステムの出現によって、一躍有名になりました。今では都市伝説といわれることの多い、「おむつとビールの法則」も、マーケットバスケット分析で発見されたといわれています。

アソシエーション分析もマーケットバスケット分析も、データマイニングの分析方法の1つですが、多変量解析の手法の一種ではありません。従って、同じマーケットバスケット分析でも、用いる手法は様々であり、共起性に基づくAprioriという手法もあれば、決定木やニューラルネットワークを用いたマーケットバスケット分析もあり、方法と手法の区別を明確にしておく必要があります。

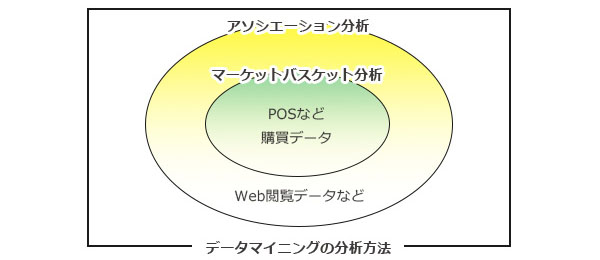


図14.アソシエーション分析とマーケットバスケット分析の関係

**②マーケットバスケット分析で用いるデータ**

実際に用いるデータは、買い物かごのデータ＝レシートのデータが基本です。レシートのデータは図15のように表すことができます。購入者全員、販売品目全体のデータは表11のようになり、実際には買った個数のデータを使うこともできます。

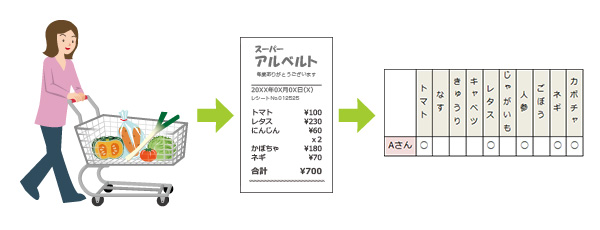
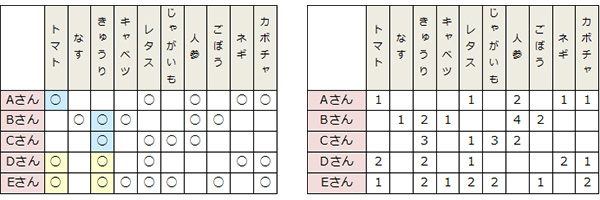


図15.マーケットバスケット分析のデータイメージ

表11.マーケットバスケット分析で用いるデータ



**③共起性の指標**

アソシエーション分析の具体的な説明の前に、共起性を表す尺度について説明しておきましょう。式や集合の記号が出てくると一気に難しく感じることがあるので、ここではベン図で概念を示してみました。最も基本となるのは、図16の共起頻度です。表11の例でトマトがX、きゅうりがYだとすると、トマトもきゅうりも購入している人はDさんとEさん2人なので、共起頻度は2となります。

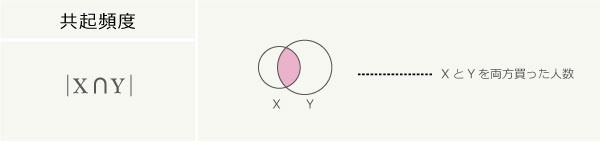


図16.共起性の指標(1)

アソシエーション分析で用いられることのある、そのほかの共起性の指標を参考まで以下に示しました。これらに共通していえる重要なことは、XとYを逆にしても意味が変わらないということです。

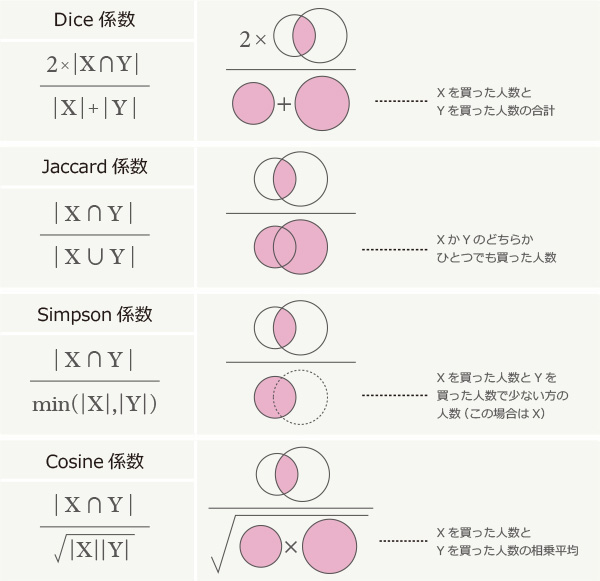


図17.共起性の指標(2)

**④レコメンデーションに用いられる3つの指標**

**■Confidence（信頼度）**

次にレコメンデーションで最も基本となる指標が、Confidence（信頼度）です。表11の例では、トマトを買った人のうち、どれくらいの人がきゅうりも買ったかという確率となります。トマトを買った人は3人で、このうち2人がきゅうりも買っているので、信頼度は66.7％ということになります。

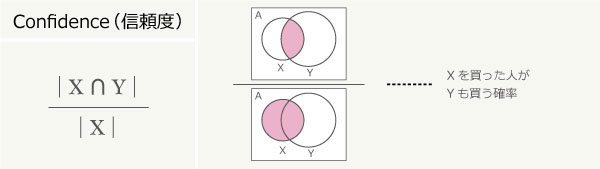


図18.Confidence（信頼度）のイメージ

共起性の指標との大きな違いは、XとYに方向性があることです。共起性の指標は、XとYを入れ替えることができますが、レコメンデーションで用いられる指標は、XとYに方向性があります。おむつとビールが一緒に買われた話は、おむつを買った後にビールを買ったのか、おむつを買う前にビールを買ったのかには言及していませんが、レコメンデーションの場合は、今まさにおむつを買おうとしている人に、ビールをすすめる、おむつを買った人に、その後ビールをすすめるという順番が問題になることがしばしばあります

Vリンターを購入した人に、用紙をおすすめすることは問題ありませんが、用紙を購入した人にプリンターをおすすめするのは、あまりよいレコメンデーションとはいえないでしょう。書籍の上巻、下巻なども同様なことがいえます。レコメンデーションにおけるアソシエーション分析には、方向性があることを考慮しなくてはなりません。

**■Support（支持度）**

次によく用いられるのが、Support（支持度）です。これは、そもそもXとY、表11の例ではトマトときゅうりがどのくらい一緒に売れているのかという指標です。\*

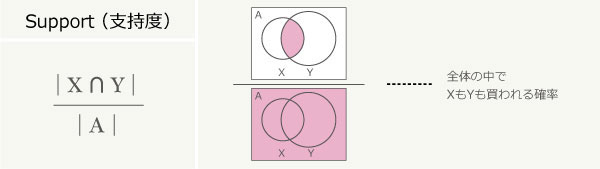


図20.Support（支持度）のイメージ

なぜSupport（支持度）が必要かというと、例えば、たまたま1人のお客様がXという商品を購入しており、Yという商品も購入していたとすると、 Confidence（信頼度）は100％となり、次にXを買われたお客様に最もおすすめすべき商品はＹであることになってしまうからです。あまり売れていない商品に何らかのおすすめをする場合には注意を要するということと、もう一つは、膨大な商品がある中で、ほとんど売れていない商品のアソシエーション分析を行なうということは効率的ではなく、あらかじめルール計算から除くというときに用いる場合もあります。

\*Supportは、Xの生起確率とする定義もある。（Borgelt and Kruse 2002）

**■Lift（リフト）**

最後にご紹介するのは、Lift（リフト）という指標です。これは、Support（支持度）が、あまり売れていない商品をどう扱うかのための指標であったのと逆に、売れ筋商品をどう扱うかという指標です。

リフトの概念（ルールの改善）については、[こちら](https://www.albert2005.co.jp/knowledge/marketing/crm_basics/lift)でも詳しく説明しているのでご覧いただきたいのですが、いくらConfidence（信頼度）が高くても、推薦される商品が誰もが知っていて、誰もが買っている商品だとしたら、あまりよいレコメンデーションではないでしょう。最近のスーパーでは、レジ袋を有料で販売しているところもあり、レジ前で自分でかごに入れることがあります。

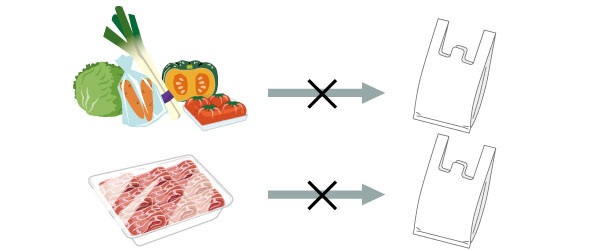


図21.売れ筋商品のレコメンデーション

これは極端な例ですが、 Confidence（信頼度）だけに頼ってレコメンドルールを生成していると、非常に多くの人のレシートの中に含まれているようなレジ袋のような商品は、その店最大の売れ筋商品であり、野菜を買っても肉を買っても、レジ袋の必要のない人にまでおすすめされてしまうことになります。

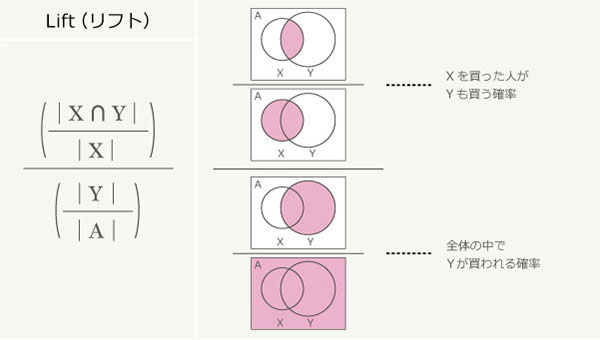


図22.Lift（リフト）のイメージ

図22がLift（リフト）のイメージですが、分子はConfidence（信頼度）で、分母はYの人気度を表しています。つまり、何もしなくてもどれくらいの人がYを買っているかという確率に対して、Xを買った人でYも買う人がどれくらいいるかの確率を比較しています。そもそもXを買っていなくてもYを買う人が50％いたとして、Xを買った人の中の50％の人もYを買うとすると、このおすすめはしてもしなくても同じだということになり、よいレコメンデーションとはいえません。通常、何もしない時よりも購入確率が上がる場合に、おすすめルールとして採用しますので、Lift（リフト）が1より小さい場合は、おすすめしないようにします。

一般のレコメンドエンジンは、 例えば「Confidence（信頼度） 0.7以上、Support（支持度）0.5以上、Lift（リフト）1以上」をルールとして採用するという方法を採用しますが、ALBERTのレコメンドロジックは、これらの指標にさらにランキング情報や情報エントロピーの考え方を独自に組み合わせ、商品特性などに合わせて適用しています。「Logreco」のロジックについては[こちら](http://www.albert2005.co.jp/logreco/recommend.html#logic)をご参照ください。

**⑤レコメンドルール生成事例**

先ほどの野菜の購買データを用いて、具体的なレコメンドルール生成事例をご紹介します。まず今回は、最低Support（支持度）50％に設定します。つまり、5人中3人が買っているアイテムだけについて計算をすればよいことになるので、なす、キャベツ、じゃがいも、ごぼう、ネギは脱落します。

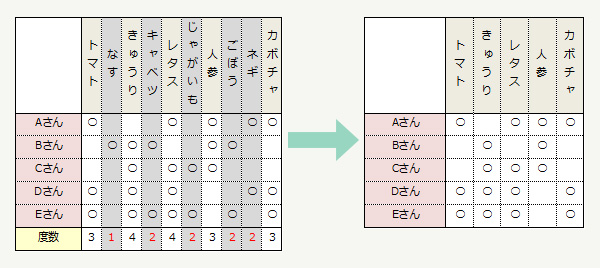


図23.最低Support（支持度）による絞り込み

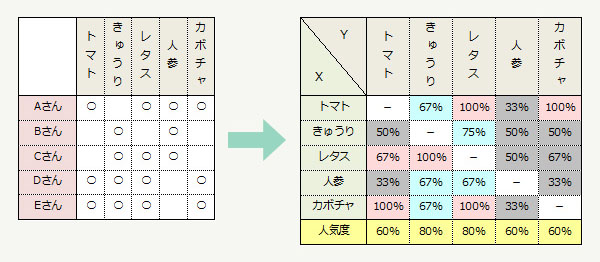
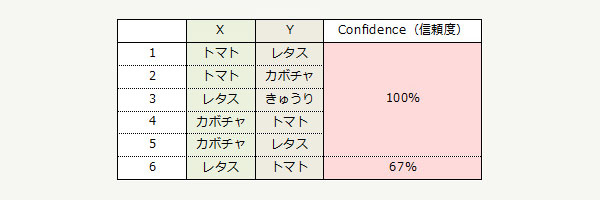


図24.Confidence（信頼度）、Lift（リフト）による絞り込み

残ったトマト、きゅうり、レタス、人参、カボチャの購買データから、それぞれのConfidence（信頼度）と人気度を出してみます。トマトを買った3人のうち、きゅうりを買ったのはDさんとEさんの2人ですので、Confidence（信頼度）は2/3＝67％となります。同様にトマトを買った3人は全員レタスも買っているので、Confidence（信頼度）＝3/3＝100％となります。人気度は全体の何人が買っているかなので、4人が買っているきゅうりとレタスは80％、その他は60％となります。

Confidence（信頼度）60％以上かつLift（リフト）1以上を採用すると、Confidence（信頼度）でグレーの部分は脱落し、人気度よりConfidence（信頼度）の小さいブルーの部分も脱落し、残るルールはピンクの部分6つだけになりました。きゅうりと人参は、ここでルールがないことになりますので脱落し、考慮すべきは、トマトとレタスとカボチャの3アイテムに絞り込まれ、90種類あるルールは以下の6種類に絞りこまれました。

表12.絞り込まれたルール



このようにして、最低Support（支持度）や最低Confidence（信頼度）を調整しながら、最終的に何に対して何を推薦するかを決めていきます。

**⑥Aprioriアルゴリズム**

ここまでは、トマトを買った人に何をおすすめすべきかといった、1対1のレコメンデーションについて言及してきましたが、実際のアソシエーション分析では、トマトときゅうりを買った人が何を買うか、じゃがいもと人参と玉ねぎを買った人は何を買うか等の、複数の組み合わせを考える必要があります。しかし、10アイテムあると、1アイテム同士の組み合わせだけでも90種類になり、すべての組み合わせを計算しようとすると、約5万7000回の共起計算が必要となります。総合通販会社などでは、数千から数万のアイテム数になることもあり、すべての計算をリアルタイムに行うのは、事実上不可能です。

この問題を解決するために、R. Agrawal氏らが1994年に発表したのが、高速で相関ルールを検出するAprioriアルゴリズムです。Aprioriアルゴリズムは、簡単にいえばSupport（支持度）が一定以下のアイテムを含む組み合わせは、最初からConfidence（信頼度）を計算しないというものです。

先ほどと同様に、Support（支持度）50％以上のみを用いるとすると、なす、キャベツ、じゃがいも、ごぼう、ネギは脱落しますので、表12のように残ったアイテムの共起頻度を算出するわけですが、ここからはトマトときゅうりが一緒に買われている事象をXと考え、そのときに一緒に買われるYが何かを考えます。つまり、トマトときゅうりが一緒に買われたときに、何が一緒に買われるかを調べることになります。

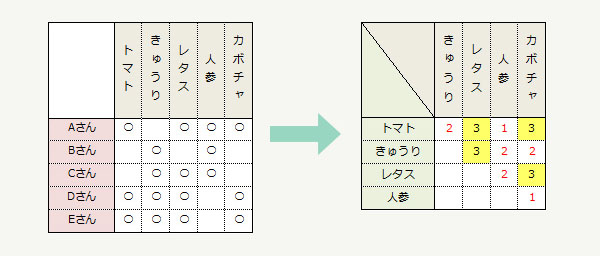


図25.Aprioriアルゴリズム(1)

図25の通り、トマトときゅうりが一緒に買われている事象は2回であり、この時のSupport（支持度）は2/5＝40％となります。したがって、この事象は50％を超えていないので不採用となります。一方、トマトとレタスが一緒に買われている事象は3回ですので、Support（支持度）＝3/5＝60％となり採用となりました。このように計算していくと、図25の黄色の部分の事象が最低Support（支持度）を超えたことになりました。さらに、この条件にもう1カテゴリー加わった場合を考えてみます。

表13.Aprioriアルゴリズム(2)



結果は、表13のように、トマト、レタス、かぼちゃが一緒に買われている場合のみSupport（支持度）50％を超えているので、結果として得られたXの条件は以下の5つとなり、Confidence（信頼度）は以下の通りになりました。

表14.Aprioriアルゴリズムによるレコメンドルール

