組み合わせ制約の大域的特徴分析による 複合商品の効率的な検索支援方法

志賀 隆之† 岩井原 瑞穂‡ 上林 彌彦‡

†京都大学工学部情報学科 〒606-8501 京都市左京区吉田本町
‡京都大学大学院情報学研究科 〒606-8501 京都市左京区吉田本町
E-mail: †tshiga@db.soc.i.kyoto-u.ac.jp, ‡{iwaihara, yahiko }@i.kyoto-u.ac.jp

あらまし 従来の電子商取引システムにおいて単一商品を価格などの属性値で検索する場合は既存の情報検索システムを用いて検索することが可能である．しかし，パーソナルコンピュータやパック旅行など複数の商品を組み合わせて商品を購入する，いわゆる複合商品を扱う場合に至っては従来の検索方法では不十分で，組み合わせの制約表現による検索が必要である．本論文では，この組み合わせの制約から，各部品間の関連度をカイ２乗検定やデータマイニングの手法を用いて求め，利用者に効率のよい選択順序を提示する検索方法について述べる．さらに，検索方法の有用性を実験結果とともに示す．本研究により，数多くのデータから，与えられた制約を用いて，求める情報を効率よく抽出することが可能になる．

キーワード E-Commerce，情報検索，データマイニング，動的制約代数，複合商品

Efficient retrieval of configurable goods by analysis of combination constraints

Takayuki SHIGA† Mizuho IWAIHARA‡ and Yahiko KAMBAYASHI‡

†School of Infomatics, Faculty of Engineering, Kyoto University

‡Department of Social Infomatics Graduate School of Informatics, Kyoto University

Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 JAPAN
E-mail: †tshiga@db.soc.i.kyoto-u.ac.jp, ‡{iwaihara, yahiko }@i.kyoto-u.ac.jp

Abstract This paper discusses electronic commerce dealing with configurable goods. A typical example of configurable goods is personal computers. Such configurable goods need a special search strategy in consideration of association rules between items because existing search strategies are inadequate. In this paper, we introduce a search strategy to show users an efficient selection order based on the chi-squared tests and the correlation of relations. Furthermore we show advantages of the strategy with various experiments.

Keyword E-Commerce, Information Retrieval, Data Mining , Dynamic Constraint Algebra, Configurable goods

1. はじめに

電子商取引において，単体で販売されている商品を，価格やカテゴリなどの属性値で検索する場合は従来の検索システムを用いた方法で検索することが可能である．しかし，複数の商品を組み合わせる，いわゆる複合商品を扱う場合の検索方法では，組み合わせの制約や売買条件，ルールといったものが存在するため，これらを考慮した検索が必要になる．典型的な複合商品である PC を例にすると，CPU として A を選んだとき，B，C のマザーボードを選ぶことはできるが，D のマザーボードは選べないといったような制約が組み合わせの制約である．いわゆる BTO [*1] の電子商取引は PC を典型例として，自動車や住宅，パック旅行など様々な商品が出現し，また物流管理システムと連動した巨大BTOサーバも検討されている．

複合商品の制約を記述するデータベースモデルとしての動的制約[6]を用いることができる．動的制約データベース(DCDB : Dynamic Constraint Database)はこうした制約

[*1] BTO (Build To Order) … 受注生産で，顧客の注文を受けてから，パソコンを生産すること．BTO パソコンとはカスタマイズ（改造）可能なパソコンのことである．

を記述するデータベースで, 動的制約代数(DCA : Dynamic Constraint Algebra)はその質問言語である.

典型的な PC の例でも可能な組み合わせは 10 億通りを超えるものが珍しくなく，その中から利用者の欲しい組み合わせを見つけるのは困難であり，現状では，いくつかの典型的な構成例から利用者がカスタマイズしていく方法が取られている．今後，多様な複合商品が出現すると予想され，人間の知識や経験で，適切な選択メニューを構成することは困難になり，システム化する必要が生じると思われる．莫大な組み合わせから目標の組み合わせにたどり着く上で，必要なデータの取捨選択という行為が行われる．その上で利用者に適切な部品の選択順序を提示すること，あるいは部品間の関連を集約して大域的な特徴を抽出して利用者に提示し，利用者に見通しの良い選択手段を提供することが重要である．図 1 は PC の例における部品間に存在する制約の例を図式化したものである.

データマイニングでは支持度(support)と確信度(confidence)というパラメータを用いて関連の強さを表すことが一般に行われているが，相関ルールの価値の評価基準としてサポートとコンフィデンスによる評価のみでは必ずしも適切とはいえない. Brin らは通常のサポート・コンフィデンスの枠組みに代わり，χ^2 値（カイ 2 乗値）により相関の強さを評価し，マイニングを行う方法を提案している[3]．我々は複合商品の組み合わせのルールがブール関数で表現され，その変数間の相関の強さの評価にも χ^2 値が使えることに着目し, 複合商品の適切な商品選択順序を求める問題に，χ^2 値や χ^2 値の問題点を克服したクラメールの連関係数，2 変数がどれだけ独立しているかを表す結合率を用いることによって関連度の評価を行うことを検討する.

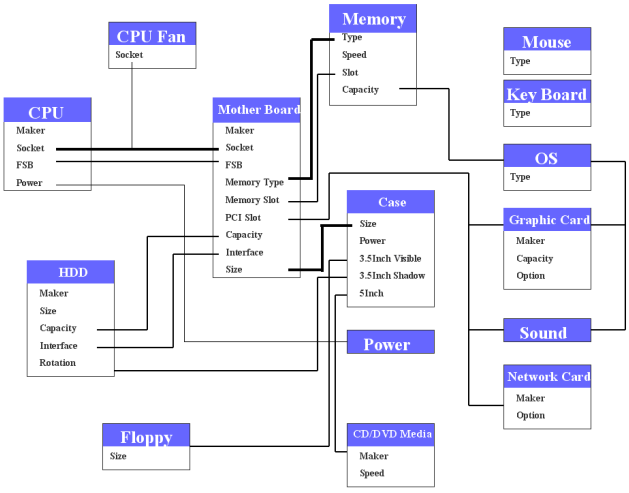

図 1 ：PC における関連図

本論文では，第 2 章で理論的な背景として，複合商品の特徴やデータマイニングで用いる相関ルール，カイ 2 乗検定の有用性など関連度の評価方法について述べ，第 3 章で代表的な複合商品の例である PC を用いて，その具体的な例や部品間の関連度についての実験と考察をする．続いて，第 4 章で，制約を考慮したユーザインターフェースについての説明と関連度を用いた効率的な検索方法を提案し，評価する．最後に，第 5 章で本論文の結論と今後の課題について述べる.

2. 複合商品の関連度の評価方法

2.1. 複合商品の組み合わせ制約の特徴

複合商品を取り扱う上で，BTO のユーザインターフェースの生成が最終的な目標である．そこで，各部品の関連を表す連関係数を求めることにする．これは，各部品が独立しているか，あるいは依存しているかを示す指標であり，複合商品のもつ様々な組み合わせ制約から統計的性質を見つけ出すことにする．複合商品のもつ組み合わせ制約には下記のような例がある.

- CPU が Pentium4 かつ Socket が 478 であるものを選べば，マザーボードが Socket478 かつ FSB533 のものであれば選ぶことができる.
- グラフィックカードに Matrox Parhelia 128MB を選ぶと，OS に Windows95 を選ぶことはできない.
- マザーボードにサウンド機能がついていれば，サウンドカードはなくても良い.
- メモリの合計の容量がマザーボードの限界容量を越えてはならない.

このような，複合商品の組み合わせ制約の持つ特徴として，以下のものがある.

- 複雑なブール関数が制約として存在する.
- 負の含意の制約など複雑な制約が存在し，これを扱わなければならない:「$x = a$ ならば $y \neq b$ 」.
- 制約はいくつかの部品クラスにまたがるが，そのパターンは部品インスタンスごとに異なることが多い.

こうした，複合商品特有の特徴をいかに処理して関連度を抽出し, 部品間の関連度を比較できるかが課題である.

2.2. 相関ルール(Association Rule)

あるアイテム集合を $I = \{i_1, i_2, ..., i_k\}$ とし，アイテムの部分集合からなるトランザクションデータ $b_i \subset I$ の集合 $B = \{b_1, b_2, ..., b_n\}$ をバスケットデータとして定義する．相関ルールとは B から求まる共起事象を表すルールである．アイテム集合 X を含むバスケットデータはアイテム集合 Y もまた含むという相関ルールは

$$X \Rightarrow Y$$

と表される[2]．この相関ルールを，サポートとコンフィデンスに基づいて評価する．サポートとコンフィデンスは，相関ルールの重要度を評価する指標として用いられ，これらの値が大きい相関ルールほどそれだけ有用であると考えられている．ここで，相関ルールは，強い共起関係を持つアイテム集合の対を求めるということができる.

2.3. サポート･コンフィデンスの問題点

サポートとコンフィデンスは幅広く用いられているが，問題がいくつか指摘されている．一つは条件部と結論部の相関関係を正しく評価できないということで，これはS.Brin らによって示されている[3]．次のような相関ルールを考える．

(CPU.Maker=”intel”)⇒(Mother.Maker=”gigabyte”)…(※)

この相関ルールは「Intel の CPU と GIGABYTE のマザーボードに関連がある」ということを意味している．ここで，このルールを説明するために，表１の例を用いて説明する．

	$intel$	$\overline{intel}$	Σ_{row}
$gigabyte$	20	5	25
$\overline{gigabyte}$	70	5	75
Σ_{col}	90	10	100

表１：CPU とマザーボードを買う顧客の相関を表す分割表

この分割表から，上記の相関ルールのサポートが 20 / 100 = 0.2 (20%)，コンフィデンス 20 / 25 = 0.8 (80%) と求まる．両方とも高い値を示しているが，これで(※)のルールが有益かというと，「Intel の CPU と GIGABYTE でないマザーボードに関連がある」という相関ルールについてもコンフィデンスが 70 / 75 ≅ 0.93 (93%) であり，Intel の CPU と GIGABYTE のマザーボードに関連があるということはできない．サポートとコンフィデンスによる評価が非常に有用な場合として，データベースの膨大なデータの中から最小支持度と最小確信度を設定し，有用な $X \Rightarrow Y$ 型の相関ルールを抽出するのには適しているが，相関の高いアイテムの対がアイテム集合間の全体の相関にどの程度寄与しているかを評価したり負の含意の制約(「$x = a$ ならば $y \neq b$」)を扱ったりすることができないといったことが既に多くの研究者によって指摘されている．

2.4. カイ2乗検定

χ^2 値（カイ２乗値）は直感的には，観測値と期待値のズレ具合を表す指標である．ここでカイ２乗検定を行う手順を示す．

2.4.1. 前提

・帰無仮説 H_0：２変数は独立である(関連がない)．
・対立仮説 H_1：２変数は独立ではない(関連がある)．
・有意水準 α で両側検定を行う．

2.4.2. カイ２乗検定の手順

【１】２個の変数 A，B がそれぞれ k 個，m 個のカテゴリを持ち，$k \times m$ 個の枡目を持つ分割表を考える．

【２】$k \times m$ 分割表で，変数 A の第 i カテゴリ，変数 B の第 j カテゴリの観察値を O_{ij} とする．また，$n_{i\,.}$ を第 i 行の合計，$n_{.\,j}$ を第 j 列の合計とする．

【３】帰無仮説のもとでは，変数 A の第 i カテゴリ，変数 B の第 j カテゴリの期待値は次式で表される．

$$E_{ij}=n_{i\,.}\,n_{.\,j}/n$$

【４】全ての枡目について $(O_{ij}-E_{ij})^2/E_{ij}$ の合計をとったものを χ_0^2 とする．χ_0^2 は自由度が $(k-1)\times(m-1)$ の χ^2 分布に従う．

$$\chi_0^2=\sum_{i=1}^{k}\sum_{j=1}^{m}(O_{ij}-E_{ij})^2/E_{ij}$$

【５】有意確率を $P=\Pr\{\chi^2 \geq \chi_0^2\}$ とし，帰無仮説の採否を決める．$P > \alpha$ のとき，帰無仮説を採択する．このとき２変数は独立でないとはいえない（関連があるとはいえない）．$P \leq \alpha$ のとき，帰無仮説を棄却する．このとき２変数は独立ではない（関連がある）．

2.4.3. カイ２乗値を利用する有用性

複合商品は可能な組み合わせの数が 10 億通りを超えることが珍しくないが，そのような組み合わせをデータベース問い合わせで求めるのは現実的ではないため，適切な商品の選択順を提示することが重要である．複合商品の組み合わせは少数の例外を除いて，任意の組み合わせが可能といった例が多く見られるが，これは２つの部品の関連がほぼ独立である．つまり，依存関係が少ないと捉えることができる．これは，関連の χ^2 値を求めた場合，χ^2 値が 0 に近いという性質となって現れる．このように，部品インスタンスの個々の接続ルールの集合の大域的特徴を表す統計量として，χ^2 値を利用することが考えられる．ほぼ独立な部品間は利用者に自由に選ばせても良いが，逆に χ^2 値が大きい部品では，注意して選ばないと禁止されている組み合わせを選ぶ可能性が大きくなる．そのため，依存関係の強い部品間から利用者に選ばせていくと探索失敗となる確率を小さくできると考えられる．

2.5. クラメールの連関係数

χ^2 値を複合商品の関連度の指標として用いるのには問題がある．それは，χ^2 値は分割表の度数の大小に影響を受けるため，異なる部品間の関連度を直接比較できないことにある．度数による影響を取り除くために，次のような式で関連度を求めたものが，クラメール(Cramer)の連関係数と呼ばれる．ここで，$\min(k,m)$ は $k \times m$ 分割表の k と m の小さい方を表す．

$$V=\sqrt{\frac{\chi^2}{N\cdot\min(k,m)}}$$

クラメールの連関係数は 0 から 1 の値をとり，行と列の変数が独立のとき 0 を示す．そこから関連が強くなるほど大きくなる傾向がある．例として表１の例で考えると，$\chi^2 = 3.704$ で，$k = m = 2$ のためクラメールの連関係数は $V = 0.1924$ となり，これよりあまり関連があるとはいえないことがわかる．

2.6. 結合率

部品クラス間の関連を求めるもう一つの方法として，

結合率θを用いることが考えられる．これは，関係データベースモデルにおいて，２つの関係を結合する際に，１つのタプルが他方の関係のタプルとどれくらいの割合で結合するかを表す指標である．複合商品において，２つの部品クラス間のアイテムの結合の度合いを表現することが考えられる．集合X,Yが存在し，その間に関係R_1が成立しているときに

$$\theta_{XY} = \frac{|R_1|}{|\pi_X(R_1)| \cdot |\pi_Y(R_1)|}$$

によって表される．ここでπ_Xは属性Xによる射影を表し，$|R_1|$は関係R_1の組数を表す．θは0～1の値を示し，$\theta \cong 1$のときほぼ直積，つまり，独立であるといえる．逆に$0 < \theta << 1$のとき，依存関係が強いといえる．

	coffee sugar	coffee cream	Σ_{row}
tea sugar	8	1	9
tea cream	1	7	8
Σ_{col}	9	8	17

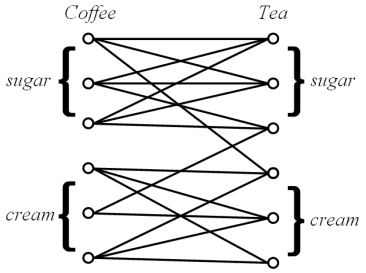

図２：砂糖かクリームを入れる人に関連があるかを調べる分割表＆グラフ（例１）

	coffee sugar	coffee cream	Σ_{row}
tea sugar	4	4	8
tea cream	4	5	9
Σ_{col}	8	9	17

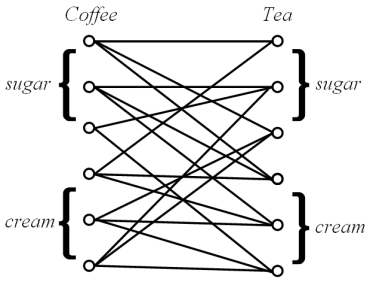

図３：砂糖かクリームを入れる人に関連があるかを調べる分割表＆グラフ（例２）

しかしながら，結合率は結合する相手の割合を表す指標であって，それぞれが相関しているかどうかを示すものではない．

例を用いて説明する．６人の人物に対して，何回かコーヒーと紅茶を与え，砂糖かクリームのどちらかを入れてもらうようにした．砂糖もクリームも３種類用意し，その結果，図２，図３のような関連が示された．ここで，コーヒーに砂糖を入れる人と，紅茶に砂糖を入れる人に関連があるかどうかを確かめると，図２でも図３でも結合率は$\theta = 0.472$と同じ値なのに対し，クラメールの連関係数は図２では$V = 0.764$，図３では$V = 0.055$と全く異なった結果となった．このように，結合率では相関関係を識別できないが，クラメールの連関係数では違いを識別できている．次節で，複合商品における相関関係を比較する．

2.7. 関連度を求める手法についてのまとめ

本章では，商品間の関連度を求める手法として，サポート・コンフィデンスによる手法，カイ２乗検定を用いたχ^2値による評価と，さらにそれを改良したクラメールの連関係数による手法，そして，データベースのテーブル間の結合率による手法を紹介した．

サポート・コンフィデンスは共起関係の大きなアイテム集合を求めることができるが，２つのアイテム集合間の相関がどのくらいかを直接的に表現することはできない．

一方，χ^2値やクラメールの連関係数による評価では，$X \Rightarrow Y$という形式の相関ルールについて，サポートやコンフィデンスを一つ一つ求めるのではなく，２つの集合間の関連度を数値化するため，比較も行いやすい． Brinらの行った2×2分割表のχ^2値によるマイニングとは異なり，我々は複合商品を構成する各部品クラス間の連関係数を評価する必要がある．そのためアイテムの数に対応した$k \times m$分割表に適用させる．

Num	Item-1	Item-2	Element-1	Element-2	χ^2	V	sup^*	θ
1	CPU	Mother	Socket	Socket	2346	1.000	0.222	0.338
2			FSB	FSB	1497.26	0.647	0.078	0.338
3			Socket	FSB	1253.40	0.727	0.104	0.338
4			FSB	Socket	1191.92	0.711	0.125	0.338
5	Memory	Mother	Memory Type	Memory Type	328	1.000	0.194	0.348
6			Memory Type	FSB	36.24	0.332	0.000	0.348
7			Capacity	Memory Type	25.73	0.280	0.056	0.348
8	Case	Mother	Size (Type)	Size (Type)	1014	1.000	0.194	0.525
9			Size (Type)	Memory Type	10.91	0.104	0.056	0.525
10	CPU	Memory	FSB	Memory Type	77.83	0.172	0.167	0.259

***sup 値は support＞20%,confidence＞80%を満たす support の割合**

表２：部品間の関連度

3. 関連度の評価実験

3.1. 実験の設定

複合商品を購入するに当たり，顧客が自らの望む商品の構成を容易にするために，各商品をどのような順でどのような選び方をさせ，どのようにナビゲートすれば効率が良いのかを求め，それに従ったインターフェースを設計する必要がある．そこで，データベースに格納されている各部品(部品の持つ要素)間の制約を用いて，その関連の強さを求めることにした．関連の強さは χ^2 値，クラメールの連関係数，サポート・コンフィデンス，結合率を求めることにより比較する．サポート・コンフィデンスについては，相関の高いアイテムの対が含まれる割合により，部品クラス間の相関度を表現することを試みる．これはサポート・コンフィデンスに一定の閾値を設定し，これらの値以上の，アイテムの対の割合を求めるということである．実験にあたって必要となるデータは，マザーボードのメーカーであるギガバイト社[*1]，日本のパソコンのパーツを販売しているパソコン工房[*2]，DOS/V パラダイス[*3] というサイトから収集した．

3.2. 概要

いくつかの部品について，部品間の各関連度を求めた．その結果が表２である．例として，CPU とマザーボードとの関連についてみてみる. Item-1,2 は比べた 2 つの部品クラスで，Element-1,2 はそれぞれの部品クラスの持つ属性である．実験番号１～４の例について言えば, CPU を属性値 Socket の値で分けたものと，マザーボードを Socket の値で分けたものを比べたということである．

3.3. 関連度の評価

結果を見ると CPU とマザーボードの相関については, CPU の Socket の値とマザーボードの Socket の値に他よりも強い何らかの相関があることを示している. 現在 CPU やマザーボードは Socket の規格により分類されており，かなりの数の組み合わせが，その分類に従って決められている．各々を違う要素で分類したり，FSB で分けたりするよりも Socket で分類したときの値が，関連度が大きくなることを裏付けたという今回の結果は納得のいく結果である．同様のことが，実験番号５～９を見ても観測できる．実験番号 10 は CPU ⇒ Mother board と Mother board ⇒ Memory という２つの関係を結合(join)して生まれた, CPU とメモリの推移的な関連度を前と同様の評価方法を用いることにより求めた．実際に，４つの評価方法を用いて比較を行ったが，χ^2 値に関しては，比較対象の度数の影響もあり各々の χ^2 値による評価の比較には適さず, またサポート･コンフィデンスによる評価については，サポートの最小閾値を設定するのが困難で，結果を見てもわかるように値が非常に小さくなってしまう. さらに, 例えば，実験番号 1 と 5 を比較したときに CPU とマザーボードのほうが，メモリとマザーボードより強い相関があるということは考えられず，相関関係を正確に表しているとは言い難い．結合率に関しても，実験 1～4 や実験 5～7 で値が同一となり，属性値の分類により生じる関連度の違いを表現できていない. 以上よりクラメールの連関係数が部品間の相関関係を評価するのに最も適しているということができる．

また，2.4 で述べた，カイ２乗検定を用いたことにおける有用性である仮説の採用・棄却(相関ルールを取捨選択)であるが，ほとんどの実験について χ^2 値が大きい値になって $\chi^2 > \chi^2(0.01)$ となり有意水準１％で有意差が存在する状況となった. これにより, 統計的には独立ではないという結論になる．しかし，χ^2 値の間に大きな差がついている．これは関連の強さをよく反映して折り，選択の効率に関連が深いと考えられる．例えば，表２の実験番号５と６を比較してみると，双方とも，$\chi^2 > \chi^2$ (0.01)となったが，圧倒的に実験番号５の χ^2 値もクラメールの連関係数も大きい．実際，実験番号６のようにメモリをその形状で分けたものとマザーボードを FSB 値で分けたものに，意味のある関係があるとは考えられず，関連度の大きさはそれを反映していると考えられる．このように，χ^2 値やクラメールの連関係数の差によって，関連の大域的特徴を表す一つの指標になっていると考えられる．

4. ユーザインターフェース構造の生成

前節では複合商品の部品クラス間の関連度の評価手法について考察した．本節ではその応用として，複合商品のためのユーザインターフェースについて検討する．インターフェースの要件としては以下のものがある．

- 部品クラスや部品に対する利用者の嗜好の反映
- 利用者の選択した組み合わせが制約違反になることを少なくする
- 部品クラス間の関連度の表示
- その他，商品探索機能，クエリ送信機能，価格表示機能

利用者が対話的に選択する複合商品の Web インターフェースとしてよく利用されている「根付き順序木型」および「画面遷移グラフ型」がよく利用されている．両者は，あらかじめ用意しておいた選択メニューに従って利用者が商品を構成していくものであり，以下では選択メニューの構造をユーザインターフェース構造と呼ぶことにする．以下，部品クラス間の関連度および利用者の嗜好等の情報を用いて，適切なユーザインターフェース構造を生成する手法について検討する．

4.1. 根付き順序木型メニューによるインターフェース

複合商品を生成するユーザインターフェースのひとつとして，根付き順序木(rooted ordered tree)が挙げられる．根付き順序木によるユーザインターフェースの例を図４に示す.

根付き順序木型は，根を持ち，かつ共通の親を持つ兄弟のノード同士に全順序を持つ．１つのノードを１つの部品クラスの選択メニューに対応させる．兄弟には番号が付与されて

[*1] Gigabyte　http://www.gigabyte.com/
[*2] パソコン工房　http://www.pc-koubou.jp/
[*3] DOS/V パラダイス　http://www.dospara.co.jp/

いて，これに従って画面に配置を行う．番号の小さい順を画面の目立つ位置に対応させる．利用者が注目しているノードから，その祖先および子孫を近い順に表示可能な範囲にまで表示し，残りはリンクを設ける．表示量は利用者の環境に応じて調節すればよい．以下，部品クラス間の関連度や利用者の嗜好を用いた根付き順序木型インターフェースの構造の生成方法について考察する．

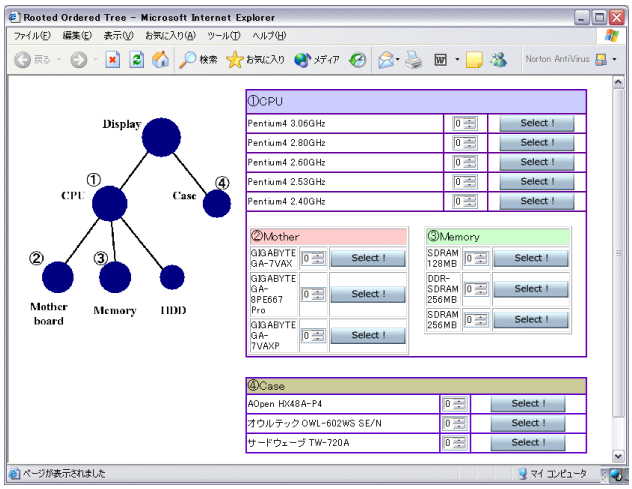

図４：根付き順序木型メニューによるユーザインターフェース

顧客の嗜好には，多様なものがあるが，ここでは次の４つを考える．

ⅰ．部品クラスに対する嗜好
ⅱ．部品に対する嗜好
ⅲ．属性に対する嗜好
ⅳ．属性値に対する嗜好

上記４つのうち，利用者による指定が容易な部品クラスに対する嗜好を以下で用いる．利用者がメニューをよく調べて選択したい部品クラスと，どれでも良いとする部品クラスの２種類に，利用者が指定するフラグで分類するものとする． 根付き順序木型メニューの生成アルゴリズムを以下に示す．

根付き順序木型メニューの生成アルゴリズム

入力

部品クラス，クラメールの連関係数による部品クラス間の関連度，部品の度数，顧客の嗜好を反映した部品クラスの集合

出力

根付き順序木型ユーザインターフェース構造

手順

【１】すべての部品クラスについて，各部品クラス間の関連度をクラメールの連関係数により求める． ２つの部品クラス間で複数の関連度がある場合は，その関連度の大きい方を採用する．ノードを部品クラス，枝を部品クラス間の関連に対応させる．任意のノードを始点に選び木を作っていく．始点から始め，既に存在する木に連結する枝の中から，関連度が最も大きい枝を選び，それを木の一部として追加する．すべてのノードが木に含まれるようになるまで繰り返す．

【２】作成した木の中からノード数が一番大きい木を選ぶ．

【３】選ばれた集合の中で，次のように根を決定する．顧客の嗜好によりフラグの立っているノードが存在する場合は，フラグの立てられたノードの中から任意に一つを選び根とし，フラグの立てられたノードがない場合は，関連度の最も大きい枝に連結したノードの片方を任意に選び，根とする．

【４】根に直接連結するノードを１階層下の子供として，再帰的に子孫を作っていき木を形成する．兄弟の順番は顧客の嗜好によるフラグが立っているノードがある場合は，そのノードを優先させ，フラグがない場合は関連度の高い順に決定する．関連度が同じ場合は，できるだけ早く選択可能な組み合わせを絞れるように，部品の度数を小さいものを優先する．

【５】【２】で選ばれた以外の木に関しても【２】～【４】を繰り返す．

4.2. 画面遷移グラフによるインターフェース

画面遷移グラフはいくつかのWeb画面を使用して利用者が作業を行っていく際の，画面間の状態遷移を有向グラフで表現したものである．画面遷移グラフの例とそれに対応する画面の例を図５に示す．

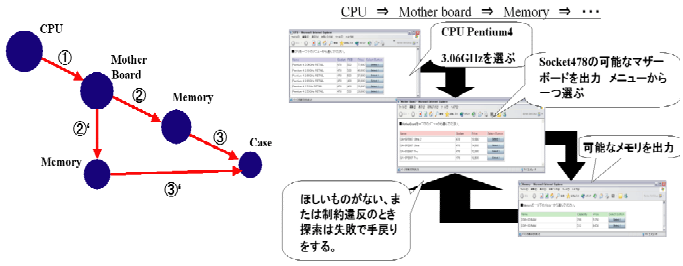

図５：画面遷移グラフによるユーザインターフェース

画面遷移グラフは開始点と終了点を持ち，また手戻り可能な部分グラフは強連結になる性質を持つ．画面遷移グラフによるユーザインターフェースでは，一画面でまとまった作業を行い，１ノードでのデータの表示量や利用者の作業量は根付き順序木型よりも多いと考えられる．複合商品の選択に画面遷移グラフ型を用いる場合は，１つの画面に１つの部品クラスを対応させることや，あるいは，まとまった部品クラスの集合を１つの画面とし，同一画面内では根付き順序木型ユーザインターフェースを用いることが考えられる．複数の根付き順序木を開始点から終了点まで結ぶ有向グラフを構成する，画面遷移グラフの生成アルゴリズムを以下に示す．

画面遷移グラフの生成アルゴリズム

入力

部品クラス，クラメールの連関係数による部品クラス間の関連度，部品の度数，顧客の嗜好を反映した部品クラスの集合，関連度の閾値

出力

画面遷移ユーザインターフェースの構造

手順

【１】ノードを部品クラス，枝を部品クラス間の関連に対応させる． ２つの部品クラス間で複数の関連度がある場合は，

その関連度の大きい方を採用する．ここで，関連度の閾値より低い枝を枝刈りする．

【２】存在する木の集合の中からノード数が一番大きい木を選ぶ．

【３】選ばれた木の中で，次のように根を決定する．顧客の嗜好によりフラグの立てられたノードが存在する場合は，フラグの立てられたノードの中から任意に一つを選び根とし，フラグの立てられたノードがない場合は，関連度の最も大きい枝に連結したノードの片方を任意に選び，根とする．

【４】根に直接連結するノードを１階層下の子供として，再帰的に子孫を作っていき木を形成する．兄弟の順番は顧客の嗜好によるフラグが立っているノードがある場合は，そのノードを優先させ，フラグがない場合は関連度の高い順に決定する．関連度が同じ場合は，できるだけ早く選択可能な組み合わせを絞れるように，部品の度数を小さいものを優先する．

【５】最初の順序木を有向グラフの開始点とし，他の木に関しても【２】～【４】を繰り返す．各木を開始点から結んでいき，最後の順序木を終了点とする有向グラフを生成する．

4.3. 部品クラス間の関連度と選択順序の関係

部品クラスの関連度と選択順序の関係について考える．一般に，商品選択時に部品クラス間の独立性が高いということはクラメールの連関係数が小さい値で関連度が低いということであり，任意に２つの部品を選ぶと制約を満たす確率が高いということである．逆に依存関係の強いものを選ぶと，制約非充足となる確率が高いということができる．このような性質を用いて，最善の順序を選択する方法として次のものが考えられる．

- 探索コストの少ない順序に従う
- 関連度の大きなものから選んでいく

ただし，仮定として，利用者は与えられたメニューから１つ部品を選択するが，希望のものがないときに，手戻りする，または失敗として探索を終了することにする．また，成功の場合にどの部品を選ぶかは等確率とする(一様確率)．

4.4. 関連度と選択順序の比較実験

本節では，前節の関連度と選択順序の関係が存在するかについて，実際のデータに基づいて比較実験を行う．クラメールの連関係数および他の関連度を求め，それにより求めた部品クラスの順序が，実際の探索コストと一致するかを確認する．

ここで，例として図６を用いる．これは，上記 PC のルールのそれぞれいくつかを CPU，Mother board，Memory の３商品から取り出し，その組み合わせ可能性を表したものである．簡単のため，CPU を A，Mother board を B，Memory を C と表すことにする．すると，その選択順序は次の６通り存在する．

① A ⇒ B ⇒ C　　② A ⇒ C ⇒ B
③ B ⇒ A ⇒ C　　④ B ⇒ C ⇒ A
⑤ C ⇒ A ⇒ B　　⑥ C ⇒ B ⇒ A

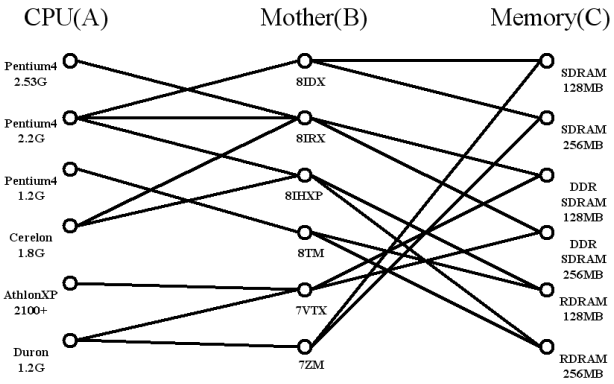

図６ ： ３商品間の組み合わせ制約

次に３部品クラス間のそれぞれの関連度を求める．ここで A-B 間，B-C 間，A-C 間のそれぞれにおいて，前節で述べたような分割表を作成する．ただし，A-C の関連については A-B と B-C の関連を結合(join)した，推移的な関連として考える．これらについて χ^2 値，クラメールの連関係数 V，結合率 θ を求めると図７のようになった．

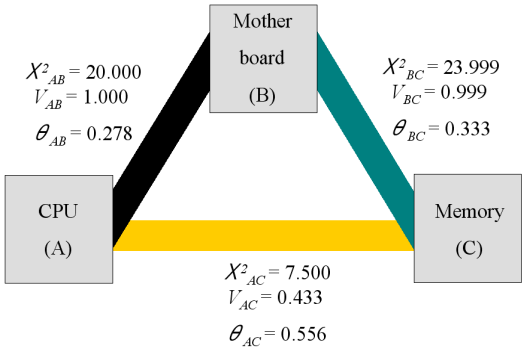

図７ ： ３商品間の関連

クラメールの連関係数と結合率による関連度は実験結果とほぼ同様，A-B ≅ B-C ＞ A-C とすることができる．これらの関連度を用いて 4.2 で述べた，関連度の大きい順で選択していく手法をとる方が，他の選択順を採用するよりもコストが小さくなるかを検証する．コスト計算は選択可能性のある部品の個数をもとにして計算する．

表３と表４は図６の CPU, Mother board, Memory すべての組み合わせについて，それが成功探索および失敗探索の場合，どれだけのコストで成功か失敗と判断できるかを記述したものである．ただし，A,B,C それぞれの部品クラスに存在する数字はそれぞれの部品の表示してある順番を表すことにする．例えば Memory の列の２(C2)は SDRAM 256MB を指す．この表を，例を用いて説明する．A2-B2-C3 (Pentium4 2.2G - 8IRX - DDR SDRAM 128MB という組み合わせ)を探索するとして設定する．このとき，図６において制約を満たす関係は存在するので，結果は成功となる．これを順序①で行うと，A1 を選ぶのに６個の A の中から１つ選ばなければならないので，コストは６となる．次に，A2 から B へ向けて出ている枝は３

本で，そのうちの１本が B2 と接続しているため，コストは３である．最後に，同様にして，B2 から C へ向けて出ている枝は２本で,そのうち１本が C3 へとつながっているので，コストは２である．すべての和をとると，この順序を選択したときのコストは 6+3+2=11 となる．ここで，もし C3 が C1 であるならば試行は失敗で，コストは B2 を選んだ時点で次に C1 へ向けた枝が存在しないため，ここで探索は終了となり，失敗と判明するまでにかかったコストは 6+3=9 となる．このようにして表３と表４を完成させる．求めた Total は，それぞれの順序をとったときの成功または失敗と判明するまでのコストの合計で，Average は一回あたりの平均コストである．

組み合わせの総数	216
成功探索	20
失敗探索	196

表３：３部品クラス間の組み合わせ内訳

順序番号	成功探索	失敗探索	Total	Average
①	198	1256	1454	6.73
②	260	2352	2612	12.09
③	200	1256	1456	6.74
④	200	1280	1480	6.85
⑤	260	2352	2612	12.09
⑥	200	1280	1480	6.85

表４：３部品クラス間の探索コストの詳細

表３および表４より，順序①または③を採用した場合が，コストの少ない選択を行えるということが判明した．続いて順序④，⑥，そして最後に②，⑤となる．これはクラメールの連関係数で表現される関連度の大きなものから選択を行う順序と一致し，コストが２分の１程度に少なくなるということがわかる．結合率も同様にコスト最小の順序と一致した．これは，関連度が強い，つまり依存関係の大きいものから選ぶ方が充足可能な組み合わせが多くなり，その結果，失敗確率が小さくなったからであると考えられる．これに対し χ^2 値はコスト最小の順序とは一致しなかった．これは，前述したように，部品数の母数の影響を受け，正しい相関関係が得られなかったのが原因と考えられる．以上より，これらが関連度を評価するのに有用な指標であることも確かめられた．

5. まとめ

本稿では電子商取引における複合商品のサイト構築において，各部品クラス間に存在する組み合わせの制約をもとに，その関連度をカイ２乗検定やデータマイニングの手法を用いて求め，利用者に効率のよい選択順序を提示する検索方法を提示した．相関関係の評価にクラメールの連関係数を用いることは，他の統計的指標よりも有用である事が分かった．さらに，この関連度を用いたユーザインターフェースの設計手法も示した．また，その後の選択順序という面においても，クラメールの連関係数が有効に利用できることが分かった．今後の課題としては，以下が挙げられる．

・組み合わせの制約を格納したデータベースとの連携
・複雑なルールをどのように処理するか

また課題のひとつに，既存システムとの連携を考えた上でのユーザインターフェースの実装・システムの高度化や相関ルールの可視化ということも挙げられる．

参　考　文　献

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami: Mining Association Rules between Sets of Items in Large Databases, *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD'93)*, pp. 207--216, 1993.

[2] Rakesh Agrawal and Ramakrishnan Srikant : Fast Algorithms for Mining Association Rules, *Proc. of the 20th VLDB Conference*, pp.487--499, 1994.

[3] Sergey Brin, Rajeev Motwani, and Craig Silverstein: Beyond Market Baskets: Generalizing Association Rules to Correlations, *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '97)*, pp. 265--276, 1997.

[4] M.Iwaihara : Supporting Dynamic Constraints for Commerce Negotiations. *2nd Int. Workshop in Advanced Issues of E-Commerce and Web-Information Systems* (WECWIS), IEEE Press, 12--20, June 2000.

[5] M. Iwaihara : Matching and Deriving Dynamic Constraints for E-Commerce Negotiations, *Workshop on Technologies for E-Services*, (informal proceedings), Cairo, Sep. 2000.

[6] M.Kozawa, M.Iwaihara, Y.Kambayashi : Constraint Search for Comparing Multiple-Incentive Merchandises, *Proc. 3rd International Conference on Electronic Commerce and Web Technologies (EC-Web)*, Aix en Provence, pp. 152-161, Sep. 2002.

[7] Shinichi Morishita, and Akihiro Nakaya: Parallel branch-and-bound graph search for correlated association rules. *Proc. of ACM SIGKDD Workshop on Large-Scale Parallel KDD Systems*, San Diego, August 1999. (Revised version: Lecture Notes in Artificial Intelligence, Springer, Vol.1759, pages 127-144, 2000.)

[8] 小山聡, 石田亨 : 情報ナビゲーションへの連想ルールの適用，電子情報通信学会論文誌，Vol.J84-D-I, No.8, pp.1266-1274, 2001.

[9] 福田剛志, 森下真一 : 相関ルールの可視化について, 電子情報通信学会技術研究報告, 95-81, pp.41-48, 1995.

[10] R.T.Ng, L.V.S.Lakshmanan, J.Han, and A.Pang : Exploratory mining and pruning optimizations of constrained association rules. *Proc. of ACM SIGMOD*, pp13--24, 1998.

[11] G.Kuper, L.Libkin and J.Paredaens(Eds) : Constraint Database, Springer-Verlog, 2000.

[12] J.Han and Y.Fu : Discovery of Multiple-Level Association Rules from Large Databases, *Proc. of 1995 Int'l Conf. on Very Large Data Bases (VLDB'95)*, Zürich, Switzerland, pp.420-431, September 1995.

[13] J.Han and M.Kamber: Data Mining : Concepts and Techneques, MORGAN KAUFMANN, San Francisco, 2001.