

多変量解析を用いた感性データベース

坂井 伸明* 大塚 真吾* 宮崎 収兄**

*千葉工業大学大学院 工学研究科 情報工学専攻 **千葉工業大学 情報学部 情報工学科

千葉県習志野市津田沼 2-17-1

Tel 047-478-0541

{novuaki,otsuka,miyazaki}@mz.cs.it-chiba.ac.jp

あらまし 現在、感性データベースは画像や音響、デザインなどの分野で多く研究されている。しかし、情報化が進む現在、多くのデータベースに対して感性検索を行えることはデジタルデバイドを減少させる意味でも有効であると言える。そこで本研究ではあまり研究されていない料理に関して感性 DB を作成する。また感性 DB ではデータのモデル化の手法も問題である。それについては最も正確であると考えられるのがアンケートなどによりデータに対して感じる各印象語の度合いを単純に重み付けすることであると考えられる。しかし、データが多くなった場合などの検索効率が悪いという問題がある。それに対し本研究では料理の感性 DB を対象とし印象語とデータとの関係に因子分析とクラスター分析を用いた分析を行い、それぞれの評価を行う。これにより料理の感性 DB に因子分析、クラスター分析を用いたときの性能や傾向などについて明らかにする。

感性データベース、因子分析、クラスター分析、料理データベース

A Kansei Database by Multivariate Analysis

Nobuaki Sakai and Shingo Otsuka and Nobuyoshi Miyazaki

Department of Computer Science,

Chiba Institute of Technology

2-17-1 Tsudanuma, Narashino, Chiba, Japan

Tel +81-47-478-0541

{novuaki,otsuka,miyazaki}@mz.cs.it-chiba.ac.jp

Kansei databases are studied in various fields such as pictures, sounds and design. One of the most common ways to construct Kansei database is attaching impressional words to objects. Weighting factors between these words and objects are used to search most adequate objects. Because, it is not very efficient to directly use words and objects, factor analysis and cluster analysis are applied in order to reduce search space. We apply these methods to cooking databases, and evaluate their advantages and disadvantages.

Kansei database, factor analysis, cluster analysis, cooking database

1 はじめに

昨今のマルチメディア情報の増加により、データベースにおいて従来のキーワード・索引語による検索方法に加え、利用者の主観に基づいた感性検索が可能な感性データベースが研究されている。

感性検索の一つとして、「印象語」をキーワードとして検索を行うものがある。検索するデータに対しての知識が少ない場合などに有効であると考えられている。検索の対象とするメディアデータにも様々なものが考えられ、静止画、動画、音響などが多く研究されている[2]、[3]、[5]。しかし、情報化が進む現在では様々なデータベースが存在している。

また感性検索ではデータのモデル化の手法も問題である[6]。モデル化においてはデータの特徴と印象語との複雑な関係をいかに正確にモデル化できるかということが重要である。その方法で最も正確であると考えられるのがアンケートなどによりデータに対して感じる各印象語の度合いを単純に重み付けすることであると考えられる。しかし、データが多くなった場合などの検索効率が悪いという問題がある。そのような背景からモデル化には多変量解析を用いる方法が一般的であり、ニューラルネットワークを用いた方法なども研究されている[1]。しかし、その場合、検索結果の適、不適は個人の主観に依存するという理由から、作成者の主観的な性能の評価に止まっている。また感性DBにおけるそれぞれのモデル化の手法にどのような傾向があるかなどは不鮮明である。

本研究ではあまり研究されていない料理に関しその感性検索の方法について検討を行う。料理は画像などに比べ、単純に料理名のみをデータとして扱う場合、人による味の解釈など複雑な問題がある。本研究では多変量解析のうち

因子分析とクラスター分析を用いそれぞれで感性情報をモデル化し印象語検索を可能とした感性DBを作成する。そして、作成したDBの評価を行い、性能やデータ数を増やした場合の傾向などについても検討を行う。

2 料理の感性DB

2-1 対象データ

分析するデータとしてファミリーレストランなどのメニューから基本的な料理21種類を選出した。選出した料理を表1に示す。また検索に使用する印象語には事前調査によりそれらの料理に適切であると思われる41語を選出した。選出した印象語を表2に示す。

それらを元にSD法のアンケートを作成し被験者21名に対して実施した。表3に一部を示す。またSD法においては対象の意味を持つ形容詞対を用いる方法が一般的である。しかし形容詞の意味の解釈は多様であり「辛い」の反対を「甘い」とするのは必ずしも適切であるとは言えない。そこで本研究では「辛い」の反対を「辛くない」というようにしたものを用いた。

表1 選出した料理

トンカツ	ピザ	うどん
ハンバーグ	シチュー	牛丼
肉じゃが	カレー	かつ丼
ロールキャベツ	から揚げ	焼きそば
野菜サラダ	ステーキ	醤油ラーメン
フライドポテト	てんぷら	チャーハン
オムライス	そば	ギョーザ

表2 選出した印象語

辛い	ジューシー	春らしい
甘い	ナチュラル	秋らしい
しょっぱい	男っぽい	上品な
酸っぱい	女っぽい	グロテス
渋い	大人向け	さわやか
苦い	子供向け	涼しげな
冷たい	田舎的な	おしゃれ
熱い	都会的な	色合いが良い
あっさり	お袋の味	ポリュー感がある
こってり	豪勢な	重い
さっぱり	素朴な	軽い
まったり	オーソド	体に良さそう
ふっくら	夏らしい	美容に良い
さくっと	冬らしい	

表3 SD 法記入例

	非常に感じる	やや感じる	普通	あまり感じない	特に感じない	トンカツ	ハンバーグ
	5	4	3	2	1		
辛い						2	1
ジューシー						3	4

アンケートの集計結果に対しそれぞれ因子分析、クラスター分析を用い、感性DBを作成した。

2-2 因子分析を用いた感性DB

因子分析とはある領域での一見複雑そうに見える種々の現象もきわめて少数の潜在的因子によって説明できるという考えに基づく分析手法である。一般的に P 個の要因を変数 X

とし、それらが潜在的な m 個の因子で表すことができるすると変数 X を表す式は以下のようになる。

$$X_1 = b_{11}f_1 + b_{12}f_2 + \dots + b_{1m}f_m + e_1$$

$$X_2 = b_{21}f_1 + b_{22}f_2 + \dots + b_{2m}f_m + e_2$$

.....

$$X_p = b_{p1}f_1 + b_{p2}f_2 + \dots + b_{pm}f_m + e_p$$

ここで f は共通因子という。この変数 f は上記のような要因で決定される事柄の対象となる事象個々でそれぞれ具体的な値を持ち、それぞれの値は因子得点と呼ばれる。b は因子負荷量とよばれ、それぞれの事象によらない定数である。また共通因子と因子負荷量だけでは説明しきれない部分を表すものとして、e を独自因子と呼ぶ。

本研究では SD 法により重み付けを行ったデータに対し STATISTICA を用い因子分析を行った。因子数は因子負荷量の分散共分散行列の固有値が 1 以上のものを用いることにより 8 つとした。全ての因子によって全体の何%が表されるかを示す累積寄与率は約 80%である。

作成した DB では因子負荷量と因子得点を用い、入力された印象語を 8 つの因子を媒体として検索する。具体的な検索過程を以下に記す。

- ① 利用者が印象語を入力する。(3 語までとした。)
- ② 因子負荷量空間において入力された印象語の因子負荷量が最も大きい因子を選択する。(入力された印象語の数に対応し 3 つまで選択可能)
- ③ 因子得点空間において②で選択された因子を軸とする空間ですべて正の因子得点をもつ料理を選択する。
- ④ 選択された料理の中で因子得点ベクトルが最も大きいものから順に 3 つを検索結果として出力する。

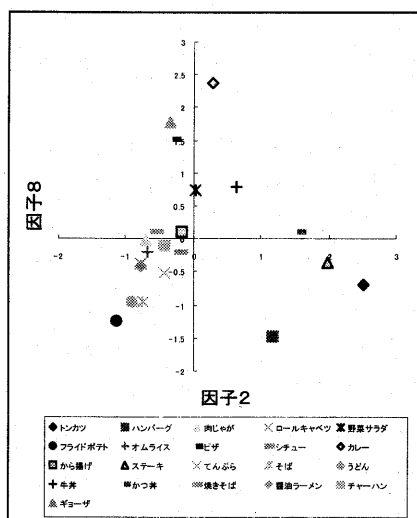


図2 因子2・8を軸とした因子得点空間

2-3 クラスタ分析を用いた感性DB

本研究ではK-means法によるクラスタ分析を用いた。K-means法での分類は次のように行われる。

データAの持つ変数を X_{A1} から X_{Ap} までとすると別のデータBの持つそれと対応する変数とでのユークリッド距離は次のようになる。

$$d^2(A, B) = \sum_{i=1}^p (X_{Ai} - X_{Bi})^2$$

対象とするデータが q 個のクラスターに分類できると仮定し、 p 次元空間上での q 個のクラスターの中心をあらかじめ決める。そしてその中心に近いデータを割り当て重心を計算する。次にデータを別のクラスターに移動し、そのクラスター内での各データとクラスターの重心間の距離の和が小さくなるようならば、移動したデータが含まれるようにクラスターの境界を移動させる。以上を繰り返し、データをどう移動してもクラスター内でのデータとそのクラスターの重心間との距離の和が小さくならないような q 個のクラスターになるまで繰り返す手法である。

本研究では因子分析と同様にSD法による重み付けデータに対しSTATISTICAを用いK-means法によるクラスタ分析を行った。クラスタ数は因子分析における因子数をそのまま適用し8つとした。

検索方法としては分析の過程で使用する各次元の各クラスターに対する平均と各クラスター内でのデータのクラスター中心からの距離を用いる。具体的な検索過程を以下に記す。

- ① 利用者が印象語を入力する。(因子分析と同様3語までとした。)
- ② 各クラスターの中から入力された印象語の平均値が最も大きいクラスターを選択する。(複数入力された場合はそれに対応する印象語の平均の合計が最も大きいクラスターを選択する。)
- ③ 選択されたクラスター内でクラスターの中心からの距離が近い料理を順に3つ検索結果として出力する。(選択されたクラスター内の料理が3つ以内の場合は存在する料理のみ出力する。)

3 比較・評価

3-1 評価方法

本研究では作成したDBについて適合率、検索速度を測定し評価を行った。適合率は情報検索で性能評価の一手法として使用され、

$$\text{適合率} = \frac{\text{適合データ}}{\text{検索されたデータ}}$$

と定義されている。また、

$$\text{再現率} = \frac{\text{検索された適合データ}}{\text{検索されたデータ}}$$

として再現率も定義される。

比較の基準としは被験者 21 名による SD 法アンケートの合計を単純に重みとした感性 DB を用いた。また利用者の主観に依存する感性検索においては、一般的に適切ではなくてもそれに近い値であれば利用者の主観に当てはまる確率も高いといえる。そこで、それについても考慮する為、適合率に加え適合率の基準を緩めた準適合率も測定した。さらに、データ数を増加させた場合の感性 DB を新たに作成し、検索速度の変化を測定した。新たに作成した DB には表 1 の料理とは別に新たに 100 種類の料理を選出し、印象語には表 2 のものを使用した。それを元に著者による SD 法のデータを作成した。そしてそのデータの中からデータ数を 50 にした場合と 100 にした場合でそれぞれ感性 DB を作成し結果を測定した。また初めに被験者 21 名のアンケートによって作成した感性 DB と整合性も考慮しデータ数を 20 にした感性 DB も作成し、実験した。被験者 21 名の SD 法アンケートを用いた感性 DB を「基本料理データ」とし、データ数を増加させた場合での感性 DB を「拡張料理データ」とする。

3-2 基本料理データにおける適合率

基本料理データでの検索結果の適合率を測定した。入力した印象語は 1 語入力した場合で「辛い」、「さっぱり」、「田舎的」など 5 種類、2 語入力した場合で「辛い」・「さっぱり」など 5 種類、3 語入力した場合で「辛い」・「さっぱり」・「田舎的」など 5 種類をそれぞれ測定した。また適合率の適合データとしては、重みづけを用いた感性 DB の検索結果上位 3 つの料理を用い、さらに準適合率として上位 5 つの料理を用いるとした。結果は表 4 のようになった。

表 4 適合率、準適合率

	分析法	感性語1つ	感性語2つ	感性語3つ	平均
適合率	因子	0.733	0.800	0.566	0.700
(重み検索3)	クラスター	0.800	0.866	0.866	0.844
準適合率	因子	0.933	0.800	0.800	0.844
(重み検索5)	クラスター	0.933	0.866	0.866	0.888

それぞれの分析での個々の印象語での検索結果については極端に不適なものもなく概ね正解といえるものであった。適合率を見ると因子分析においては感性語を組み合わせたときに多少誤差が生じる場合もあるが、準適合率の値は良いことから、検索結果においては適合していなくとも近い料理を検索することが可能であるといえる。クラスター分析を見ると因子分析よりも良い結果が出ている。しかし、クラスター分析の感性 DB は検索アルゴリズムの関係上、選択されたクラスター内にある料理しか検索結果として出力しない為、再現率で評価すると必ずしも良い結果にはならない。

3-3 基本料理データにおける検索速度

次に基本料理データでの検索速度を測定した。測定の仕方は適合率と同様に印象語を 1 語、2 語、3 語入力した場合でそれぞれ 5 種類ずつ行った平均値を検索速度とした。結果は表 5 に示す。

表 5 平均検索時間

	感性語1つ	感性語2つ	感性語3つ	平均
因子	0.09 秒	0.11 秒	0.11 秒	0.10 秒
クラスター	0.08 秒	0.07 秒	0.07 秒	0.07 秒
重み	0.18 秒	0.18 秒	0.18 秒	0.18 秒

検索速度においては、データ数が少ない為あまり大きな差は見受けられないが、因子やクラスターを通して検索する場合、重みづけしたデ

ータを直接検索する方法よりも 2 倍近く速くなるという結果となった。

3-4 拡張料理データでの比較

次に拡張料理データでの検索速度、適合率、準適合率を測定した。データを増やした場合での因子数は因子負荷量の分散共分散行列の固有値が 0.5 以上のもので全体の 70%以上表すことができるように設定した。その結果、データ数 50 の時に因子数を 15、データ数 100 の時に 20 とした。またクラスターの数も同様とした。それぞれの適合率、準適合率の値を表 6、図 2 に示す。検索速度の変化を図 3 に示す。尚、データ数が増加した場合、適切な検索結果に近い検索結果の数も増加すると考えられるため、準適合率において重み検索の結果を 5 つにしたものに加え 10 にしたものについても測定した。前者を準適合率 A、後者を準適合率 B とする。表 6 の適合率を見てみるとデータ数を増やした場合の因子分析の適合率・準適合率は決して良い値とはいえない。これは因子分析により因子によって表せない部分を省いた誤差が印象語を組み合わせた場合などに大きく影響しているからだと考えられる。摘出する因子数を増やせば誤差は小さくなるがデータ数が多くなる場合は、因子負荷量の分散共分散行列での固有値一つ一つの寄与率も小さい。データ数が大きくなるにつれ精度を高めるには寄与率の低い因子も多く摘出する必要がある。それは重み付け検索に近づくことを意味し、大幅な検索速度の向上にはつながらない。一方クラスター分析は適合率は低が準適合率はデータ数が大きくなっても高い値を保っている。データ数が大きくなるにつれ、的確なデータが検索される確率は低い但那に近いデータをコンスタントに検索することができるといえる。図 3 を見るとデータ数を 50 とすると検索速度は、因

子分析は重みづけ検索の約 2 倍、クラスター分析は約 4 倍の速度となり、データ数を 100 にした場合はそれぞれ、約 4 倍、約 8 倍という結果となった。

表 6 拡張料理データでの適合率

データ数20

	分析法	感性語1つ	感性語2つ	感性語3つ	平均
適合率	因子	0.600	0.444	0.222	0.422
(重み検索3)	クラスタ	0.667	0.733	0.767	0.722
準適合率 A	因子	0.733	0.556	0.222	0.504
(重み検索5)	クラスタ	0.667	0.900	0.833	0.800
準適合率 B	因子	0.733	0.667	0.889	0.763
(重み検索10)	クラスタ	0.733	0.900	0.900	0.844

データ数50

	分析法	感性語1つ	感性語2つ	感性語3つ	平均
適合率	因子	0.467	0.267	0.367	0.367
(重み検索3)	クラスタ	0.767	0.467	0.367	0.533
準適合率 A	因子	0.667	0.400	0.567	0.544
(重み検索5)	クラスタ	0.833	0.567	0.467	0.622
準適合率 B	因子	0.733	0.600	0.667	0.667
(重み検索10)	クラスタ	0.900	0.900	1.000	0.933

データ数100

	分析法	感性語1つ	感性語2つ	感性語3つ	平均
適合率	因子	0.533	0.533	0.267	0.444
(重み検索3)	クラスタ	0.200	0.200	0.400	0.267
準適合率 A	因子	0.733	0.600	0.467	0.600
(重み検索5)	クラスタ	0.267	0.200	0.400	0.289
準適合率 B	因子	0.800	0.800	0.533	0.711
(重み検索10)	クラスタ	0.800	0.800	0.800	0.800

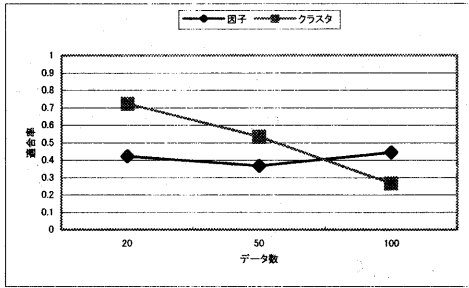


図 2 - 1 データ数の変化と適合率

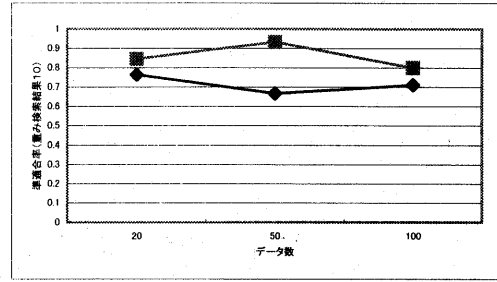


図 2 - 3 標準適合率 B

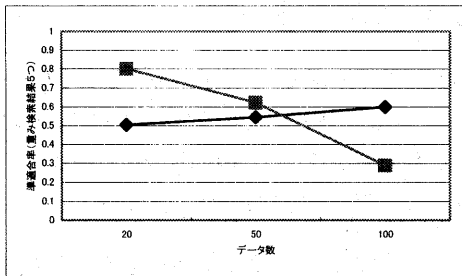


図 2 - 2 標準適合率 A

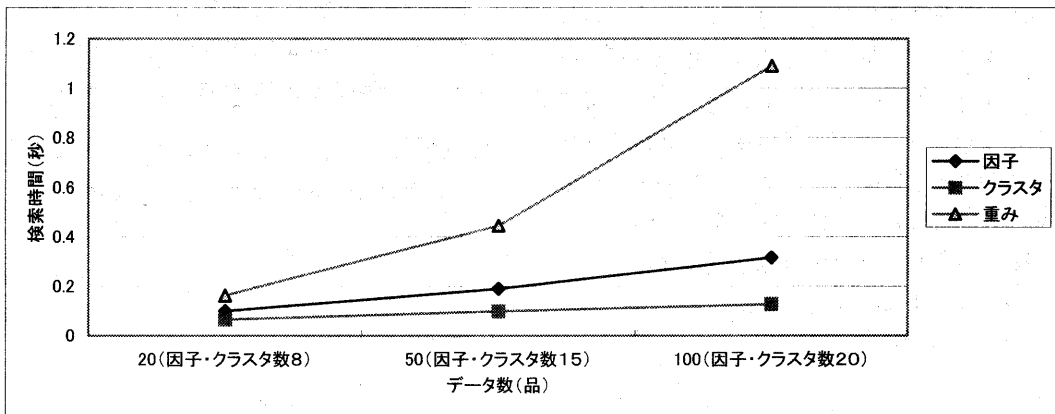


図 3 検索速度の変化による変化

3-4 それぞれの分析での傾向

以上より、それぞれの分析での大まかな特徴や傾向などについて、正確性、検索速度、拡張性の3点からまとめたものを表7に示す。

表7 それぞれの分析での評価

	正確性	検索速度	拡張性
因子分析	△	△	○
クラスター分析	△	○	△
重み付け	○	×	×

3-5 その他の問題

その他の問題点として感性DBにおいては人により違う感性にどう対応するかという点がある。それについて考えると因子分析は印象語間の関係やデータ間の関係を動的に変化させることにより、比較的容易に検索結果を調整することができる。すなわち因子負荷量と因子得点の値を調整することによりそれが可能である。

さらに、印象語や料理を後から追加する場合、重みづけ検索では新たに追加したデータの重みを集計する必要がある。しかし因子分析を用いれば、因子負荷量や因子得点をうまく設定することにより追加データは適切な位置に容易に追加可能である。クラスター分析の場合も印象語の追加は印象語間の関係を別に測定する必要があるが、データの追加は各印象語の平均の値を設定すれば可能である。

4 まとめ

本研究では印象語により料理を検索する感性DBについて、因子分析とクラスター分析を用いることにより、ある程度の精度を保って印象語検索を高速化することができた。またそれ

ぞれの分析での特徴も把握できた。

今後はデータ数が増加した場合の適合率を高くする方法など検討していきたい。また前述した人により違う感性についての対応も考えていきたい。

5 参考文献

- [1] 辻 三郎：感性の科学、サイエンス社（1997）
- [2] 栗田 多喜夫、加藤 俊一、福田 郁美、坂倉 あゆみ：印象語による絵画データベースの検索、情報処理学会論文誌、Vol33, No11, Nov1992
- [3] 宝珍 輝尚、山田 恵一、都司 達夫：感性の主因子に基づく画像・自然語アクセスに関する一検討、情報処理学会研究報告、Vol2000, No44, データベースシステム 121-12、情報学基礎 58-12（2000. 5.26）
- [5] 三石 大、多田 和彦、佐々木 淳、船生豊：媒介変数を用いた分散分析による感性情報検索の提案、Vol2000, No.44、データベースシステム 124-1、情報学基礎 62-1（2000. 5. 21）
- [6] 加藤 俊一：感性の工学的モデル化とヒューマンメディアデータベース、Vol2000, No.44、データベースシステム 124-5、情報学基礎 62-5（2000. 5. 21）
- [7] 河野 浩之、川原 稔：Web 検索におけるテキストマイニング、人工知能学会誌、16 巻2 号（2001 年3 月）