CAM法を用いた個人嗜好モデルに基づく 商品推薦システム

Product Recommendation System Based on Personal Preference Model Using CAM

村上 知子

Tomoko Murakami

(株) 東芝 研究開発センター 知識メディアラボラトリ^{*1}

Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba
tomoko.murakami@toshiba.co.jp

吉岡 信和

Nobukazu Yoshioka

国立情報学研究所

National Institute of Informatics
nobukazu@nii.ac.jp

折原 良平

Ryohei Orihara

(株) 東芝 研究開発センター 知識メディアラボラトリ

Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba
ryohei.orihara@toshiba.co.jp

古川 康一

Koichi Furukawa

慶應義塾大学大学院 政策・メディア研究科

Graduate of Media and Governance, Keio University
furukawa@sfc.keio.ac.jp

keywords: recommendation system, feature construction, kansei, text mining

Summary

Product recommendation system is realized by applying business rules acquired by data mining techniques. Business rules such as demographical patterns of purchase, are able to cover the groups of users that have a tendency to purchase products, but it is difficult to recommend products adaptive to various personal preferences only by utilizing them. In addition to that, it is very costly to gather the large volume of high quality survey data, which is necessary for good recommendation based on personal preference model. A method collecting kansei information automatically without questionnaire survey is required. The constructing personal preference model from less favor data is also necessary, since it is costly for the user to input favor data.

In this paper, we propose product recommendation system based on kansei information extracted by text mining and user's preference model constructed by Category-guided Adaptive Modeling, CAM for short. CAM is a feature construction method that can generate new features constructing the space where same labeled examples are close and different labeled examples are far away from some labeled examples.

It is possible to construct personal preference model by CAM despite less information of likes and dislikes categories. In the system, retrieval agent gathers the products' specification and user agent manages preference model, user's likes and dislikes. Kansei information of the products is gained by applying text mining technique to the reputation documents about the products on the web site. We carry out some experimental studies to make sure that preference model obtained by our method performs effectively.

1. はじめに

近年の World Wide Web(WWW)の普及により、多くのショッピングサイトや商品情報サイトが出現し、それらのサイト上の商品情報にアクセスすることによる WWW を通じた商品の購入が日常的に行われるようになった。個人がアクセスできる商品情報の量が飛躍的に上昇したのに伴い、利用者が欲しい商品を推薦する商品推薦システムの開発が注目を集めている [喜連川 01, Riedl 01]。

最も基本的な商品推薦システムでは、提供する商品情

報に関するドメインの知識や専門家の知見 [Nelson 70, Bettman 79] を定義したビジネスルールに基づいて商品を推薦する。しかし、記述すべきビジネスルールが複雑であったり多量である場合には、記述に要する時間やコストが問題になっている。そこで、利用者のプロフィール情報や購買履歴情報に対してデータマイニング技術を適用して購買の傾向に関するパターンを獲得し、それに基づいて商品を推薦するという、推薦規則の自動獲得の研究が進められている [Pazzani 99, Berry 04]。しかし初期の研究では、プロフィールや購買履歴に応じた利用者層を対象に購買パターンを発見していたため、利用者

^{*1} 現在、慶應義塾大学大学院政策・メディア研究科にも所属

ごとに異なる嗜好や主観を反映した推薦ができないという問題点があった。

そこで、利用者自らが検索語を入力することにより、希望する商品のリストを獲得するための商品推薦システムの研究も進められている。利用者が商品価格や商品仕様を検索語として入力すると検索語にマッチした商品のリストを返すシステム [Etzioni 97, Doorenbos 97] が一般的であるが、価格や仕様以外の属性が重要となる商品のためには、商品のイメージや使用感などに関する検索語を扱うことが可能な商品推薦システムが重要となる [加藤 01, 矢野 03]。

商品イメージや使用感などの感性情報は、商品仕様などと同じく、同一商品であっても利用者の嗜好や主観を反映するため利用者ごとに評価が異なる。商品推薦システムでは感性情報を表す検索語は感性語と呼ばれるが [木本 99]、例えば「かわいい」などの感性語で表される商品イメージは利用者の嗜好や主観を伴うので、利用者に関する情報を用いずに検索語のみから利用者の商品イメージにマッチする商品に対応付けるのは困難である。そこで、アンケートにより利用者の感性情報を収集し、その情報から個人感性モデルを構築して利用者の嗜好にかなう商品を検索するシステムが研究されている [矢野 03]。

しかし、上記のシステムでは個人感性モデルを構築するために、各商品についてのアンケートに利用者が回答しなければならない。例えば、ある商品に対して「明るい」、「都会的」などの具体的な感性語を回答したり、あらかじめ用意された感性語に段階的な評価値を入力しなければならない。このような多量のアンケートを必要とするアプローチは、利用者負担が重いという問題点がある。また、商品提供側が各商品に適切な感性語を割り当てる負担も問題となる。そこで、利用者ごとに異なる嗜好や主観を反映した商品推薦を実現するために、利用者の感性的な情報を収集するアンケートを必要とせずに個人感性モデルをデータから自動獲得し、各商品に適切な感性語を割り当てることが可能な商品推薦システムが必要とされている。

本論文では、属性として感性語を割り当てる必要がある商品を、利用者の嗜好や主観を反映して推薦するために、知識辞書を用いたテキストマイニングと個人感性モデルのひとつである嗜好モデルを構築する CAM 法 [Murakami 01] を組み合わせた商品推薦システムを提案する。

CAM(Category-guided Adaptive Modeling)法は、複数の属性変数とクラス変数から成るデータに対して、特徴量空間において、同じ事例は近くに異なる事例は遠くに配置されるような新しい属性を、クラス変数値が設定されている事例から生成することが可能な属性生成手法である。CAM法は商品の選好情報から嗜好モデル(個人感性モデル)を構築することが可能であり、生成する属性には任意の非線形関数を用いることが可能である。CAM法を用いて嗜好モデルを構築することにより、利用者層

ではなく個人の嗜好や主観を反映した商品推薦システムを、比較的少ない利用者負担で実現する事が可能になる。

本論文の実験では、コスメ商品を対象とした商品推薦システムを構築し、利用者によるアンケート評価を用いて提案するシステムの有効性を確認した。なお、商品価格や商品仕様などの情報、商品の感性情報、利用者の選好情報は商品口コミサイトなどからエージェント技術を用いて収集した。

本論文は以下のように構成されている。2章で商品推薦システム、嗜好や主観を扱ったシステムに関する過去の研究を紹介し本論文との関連について説明する。3章ではエージェントに基づいたシステム構成と嗜好モデル構築のための商品の感性情報の収集、CAM法による嗜好モデルの構築に関して述べる。4章では嗜好モデルの評価実験と実験結果について述べる。最後に5章で今後の課題を述べる。

2. 関 連 研 究

市場が成熟して利用者の価値観が多様化したため、サービスや商品の提供者が利用者のニーズに的確に対応するための one-to-one マーケティングが重要になっている [神場 00]。one-to-one マーケティングを実現するためには利用者に対する様々な働きかけをパーソナライズする必要がある。利用者によっては「ワンストップ」でショッピングをしたいと思っている傾向にあると言われており、購買単価を上げるために様々なパーソナライズを行う商品推薦システムが導入されている [Pazzani 99]。商品推薦システムのパーソナリゼーションには、利用者層の購買の傾向を利用する方法と利用者の嗜好をモデル化する方法がある。以下にそれぞれの方法の特徴を述べる。

2.1 購買パターンの利用

ビジネスルールによるパーソナリゼーションは、購買パターンに関する専門家の知識に基づいた商品推薦規則をビジネスルールとしてルールベースに蓄積し、蓄積されたビジネスルールに基づいて商品を選択し利用者に推薦する手法である。例えば、利用者のデモグラフィック情報から推薦商品を決定的なビジネスルールが用いられる。このような手法は、質の高いビジネスルールベースが維持できれば有効なマーケティング手法になるが、ビジネスルールベースの構築にコストがかかることに加え、一旦蓄積したビジネスルールが永遠に有効であるという保証はないので定期的にチェックして更新する必要がある点が欠点である。

ビジネスルールを用いた手法の弱点を克服するために、現在世の中で最も多く実システムで採用されている手法が協調フィルタリングである。

協調フィルタリングは利用者間あるいは商品間の相関関係を利用したパーソナリゼーション手法であり、利用者

と購買履歴の近い(似た商品を購入している)利用者集合を特定し、利用者が未購入の商品の中で利用者集合における購入頻度の高い商品を推薦する [Resnick 94]。また、購買履歴の類似判定に商品相関を考慮した手法 [Pazzani 99] や、オンライン学習を導入した手法 [Nakamura 98] が提案されている。

具体的に述べると、例えば利用者 A が商品 α を買うとする。すると商品情報システムは過去に α を買った利用者を洗い出し、その利用者が α 以外に買った商品をピックアップして A に紹介する。またあるときには商品 β を買おうとしている利用者に β と同じカテゴリに入る商品も一緒に提示して比較検討してもらう。

協調フィルタリングのように蓄えられた購買データベースから購買パターンを発見する手法は、ビジネスルールベースを構築する必要がなく、メンテナンスを省力化できるのが利点である。しかし、例えば α を買った利用者のように、類似した嗜好を持つと推測される利用者層を対象にしているため、利用者個人のためのきめ細かな推薦ができないという欠点がある。

2.2 利用者の嗜好のモデル化

利用者個人の嗜好に応じた情報を提供するための技術は、ユーザプロファイリング技術と呼ばれる。ユーザプロファイリング技術は、嗜好に関する情報を利用者が直接入力する明示的手法とシステムが観測した利用者の振舞いに基づいて嗜好を推定する暗黙的手法に分類できる [土方 04]。

代表的な明示的ユーザプロファイリング手法である Syskill&Webert [Pazzani 97] では、推薦対象を複数の属性によって表し、利用者の嗜好をクラス属性としてクラスがうまく分類されるような属性を予測情報利得によって選択し、利用者の嗜好をベクトル空間として表現する。例えば推薦対象が文章の場合、文章中に表れる各単語が属性となる。そして、ベクトル空間上のユークリッド距離に基づき推薦を行う。

明示的手法のひとつである FoxTrot [Middleton 03] では、推薦対象をオントロジを用いて抽象化した複数の属性によって表し、利用者の興味の度合いをクラス属性として、クラスがうまく予測されるような各属性の重み係数を決定し、利用者の嗜好を重みベクトルとして表現する。例えば推薦対象が文章の場合、単語によって表される意味的な概念が属性となる。そして、k-nearest neighbor 法の類似アルゴリズムによって推薦を行う。

これらの手法により利用者個人の嗜好に基づいた情報の推薦を行うことが可能になる。また、嗜好モデルを用いることにより、協調フィルタリングなどで生じる新商品などの新しいアイテムに対する推薦精度の低下 (cold-start 問題) が緩和されるという利点もある。特に、オントロジ (知識辞書) を用いた FoxTrot は、Syskill&Webert と比較して属性数をはるかに少なくすることが可能なので、

より少ないデータで cold-start 問題を解決できる。

本論文で提案する嗜好モデル構築方法は明示的な利用者の嗜好情報を用いる。そして、FoxTrot と同様に知識辞書を用いて属性数の削減を行うが、非線形関数を用いることが可能な点と、Syskill&Webert と同様に利用者の嗜好をクラス属性とする点が特徴である。また、推薦対象の属性として個人差の大きい感性属性を用いる点が本論文で提案する商品推薦システムの特徴である。

2.3 感性情報の利用

値を決定するのが難しく、利用者によって商品を認知する際のバイアスに差がある感性情報を扱う必要がある商品の推薦システムを実現するため、利用者の感性的な情報を収集し、利用者の嗜好をモデル化する研究が進められている。

加藤らの研究グループ [Kato 96, 加藤 01, 矢野 03, 吉田 01] は、正準相関分析法やベイジアンネットワークを用いて対象の感性情報を利用した個人感性モデルを構築する手法を提案している。[矢野 03] では、商品の使用感に基づいた「やわらかい」などの感性語と商品イメージに基づいた「かわいい」などの感性語との関係をイメージ語ネットとして表現し、これを嗜好モデルとみなして利用者の嗜好にかなう商品を推薦している。一方 [吉田 01] では、色などの画像の物理的特徴と「明るい」などの感性語との関係をベイジアンネットワーク型の個人感性モデルとして構築し、利用者の嗜好にかなう画像の検索に応用している。

これらの研究では、アンケートにより利用者の感性的な情報を収集し、その情報をもとに個人感性モデルを構築して利用者の嗜好にかなう商品や画像を検索するシステムを実現している。しかし、上記のシステムでは個人感性モデルを構築するために、各商品についてのアンケートに利用者が回答しなければならない。例えば、ある商品に対して「明るい」、「都会的」などの具体的な感性語を回答したり、あらかじめ用意された感性語に段階的な評価値を入力しなければならない。このような多量のアンケートを必要とするアプローチは、利用者負荷が重いという問題点がある。また、商品推薦システムにおいては、商品提供側が各商品に適切な感性語を割り当てる負荷も問題となる。本論文で提案する商品推薦システムでは、知識辞書を用いたテキストマイニングにより商品の感性情報を推定する。

3. 商品推薦システム

3.1 システムの概要

本論文で提案する商品推薦システムは、商品価格や商品仕様を収集したり商品へ感性語を割り当てるデータ収集システムと、利用者の嗜好情報を含む利用者情報と嗜好モデルを管理する利用者嗜好管理システムから構成され

図 1 システム構成

ている。これらのシステムはエージェント技術 [Beegent 00, Yoshioka 00] によってそれぞれ検索エージェントとユーザエージェントとして実装されている。図 1 にシステム構成図を示す。以下ではユーザエージェントと検索エージェントについて説明する。

(1) ユーザエージェント

ユーザエージェントはクライアント側（利用者の端末など）に常駐して利用者の選好情報などの利用者に関する様々な情報と嗜好モデルを管理する。利用者の選好情報とは利用者が好きな商品やブランドもしくは嫌いな商品やブランドのことを指す。本論文の実験ではあらかじめ利用者に選好情報を入力してもらう方法を採用したが、利用者への負荷を考慮すれば、クリック履歴などの利用者行動情報から推測する方法も可能と思われる。

ユーザエージェントは検索エージェントと連動しており、利用者がある商品ジャンルを指定すると、検索エージェントに対して情報収集を指示し、検索エージェントからもたらされたデータとあらかじめ管理していた利用者の選好情報を用いてその商品ジャンルの嗜好モデルを CAM 法により構築する。

さらに、ユーザエージェントは構築された嗜好モデルを用いて、新たに収集された商品の中から推薦すべき商品を決定する。

(2) 検索エージェント

検索エージェントは、ユーザエージェントからの検索要求を受けると、あらかじめ登録された商品情報サイトから、指定されたジャンルの商品に関する商品価格と商品仕様を自動的に収集する。

そして、あらかじめ登録された口コミサイトから収集された商品に関する評判情報を文書データとして自動的に収集し、後述するテキストマイニングを適用することにより商品に対して感性語を割り当てる。

さらに、検索エージェントは商品情報サイトに登場する新商品情報を常に監視する機能も備えている。この機能により新商品に関する情報を収集することが可能になり、利用者の嗜好が変化しない場合にはユーザエージェントが新製品を利用者に推薦すべきかどうかの決定を行うことができる。

3.2 商品の感性情報

商品価格や商品仕様の他に商品の選好に影響を与える要因として、商品のイメージや使用感などの商品の感性情報がある。商品の感性情報を収集するためにはアンケート調査や専門家の意見を用いる方法があるが、アンケート調査に要するコストや適切な専門家が存在しないなどの問題点がある。商品情報サイトには一般的な利用者からの商品の評判情報がテキストデータとして蓄積されることが多く、テキストマイニングを利用した商品の感性情報の抽出が期待できる。すなわち、商品へ感性語を割り当てることが可能となる。

テキストマイニング手法の代表例として、テキストデータからの概念抽出手法がいくつか開発されている [小林 03, 那須川 01]。テキストデータからの概念抽出には、同じ概念が様々な表現によって記述されることによる困難さがある。そこで、本研究では表現の多様性を吸収するため知識辞書を用いた概念抽出手法 [市村 03] を利用して商品に感性語を割り当てる。

表 1 は、知識辞書の一例である。知識辞書は上位概念から「概念クラス」「キー概念」「表層表現」の順に構成されている。概念クラスはテキストの主題を表し、キー概念はその主題を述べるためのひとつの抽象的な表現を表し、表層表現はあるキー概念を述べるために文書中に現れる記述を表す。このような知識辞書により、文書中に存在する様々な表現を吸収することが可能になる。

表 1 知識辞書の登録例

概念クラス	キー概念	表層表現
うるおい	みずみずしい	みずみずしい
うるおい	みずみずしい	水っぽい
うるおい	ウェット	ウェット
うるおい	ウェット	ウエット
うるおい	ウォーター	ウォーターブルー
うるおい	ウォーター	ウォーターブルー
うるおい	ウォーター	ウォーターブルータイプ
低刺激	やさしい	やさしい
低刺激	やさしい	優しい
低刺激	よわい	弱い

本論文の実験で用いた知識辞書は、文書の語尾（例えば「～だ」「～だった」）のバリエーションと一般に日本語の文書で用いられる表層表現を集めた知識辞書テンプレート [市村 03] をベースに、対象文書を参照しながら、推薦商品ジャンルにふさわしい概念クラスやキー概念を手入力した。許容できる感性語の抽出率となるために、約 300 件の文書を参照した。

ここで、知識辞書構築の際にキー概念が不足したとしても、感性語の抽出率が若干低くなる分、対応した感性属性値が全体的に低くなるだけなので、推薦精度への影響は限定的と考えられる。しかし、知識辞書構築の際に概念クラスが不足してしまうと、重要な感性属性の抽出に失敗することになるので、推薦精度へ重大な影響があると考えられる。

本システムでは、上記の方法によって得られた知識辞書を用いて、ある商品の評判情報のテキストデータに表層表現がマッチした場合に、その表層表現に対応する概念クラスを感性語としてその商品に割り当てる。ただし、商品の感性情報を定量的に表現するため、商品の感性語に重み付けをする。

ある商品の感性語の重みを算出する場合、その商品に関する評判情報の全文書のうちどの程度の文書に、その感性語に対応する概念クラスがマッチし、かつ、その概念クラスがマッチする商品は全商品のうちどの程度存在するかという指標が参考になると考えられる。この指標は tf-idf 法 [Salton 97] として知られ、全文書中の概念の分布を考慮した代表的な索引語重み付け手法である。本研究では、tf-idf 法を利用して商品の感性語の重みを算出する。

例えばある商品に関して収集した「…使用後はみずみずしい感じでした」という文書に対してテキストマイニングを実行すると、表 1 に示す知識辞書中の「みずみずしい」という表層表現とマッチし、知識辞書中の上位概念クラス「うるおい」が商品の感性語として付与される。

収集した全文書に対してこのようなテキストマイニングを実行すると、知識辞書で定義された概念クラスのうちいくつかの商品の感性情報として抽出される。ある商品 p の i 番目の感性語の重み $W(p, i)$ は、テキストマイニングの結果商品 p に付与された i 番目の概念クラスの出現頻度 $tf(p, i)$ と i 番目の概念クラスが抽出された商品数 $df(i)$ を利用して以下のような式で求める。

$$W(p, i) = \frac{tf(p, i) * idf(i)}{\sqrt{\sum_i (tf(p, i) * idf(i))^2}} \quad (1)$$

$$idf(i) = \log \frac{N}{df(i)}$$

ここで N は全商品の数とする。なお、テキストマイニングの結果マッチする文書がひとつもなかった感性語の重みは 0 になる。

上記の方法により、利用者や商品の提供者に負担なく、商品価格や商品仕様、商品の感性情報を併せた属性をひ

とつの商品属性ベクトルとして表現することが可能になる。その結果、ある商品ジャンルの収集された全商品がベクトル空間上にマッピングされる。

3.3 CAM 法

利用者に適応した商品推薦を行うための嗜好モデルのひとつとして、商品属性ベクトルによって生成されるベクトル空間を利用者の選好情報を用いて適応させるという方法がある。このとき、利用者の選好情報が全商品について得られるわけではないことに注意しなければならない。CAM 法は、全商品に対する商品属性ベクトルと一部の商品に付与された利用者の選好情報を入力データとして、利用者の嗜好を反映した全商品がマップされたベクトル空間を生成することが可能な属性生成手法である。CAM 法を用いて少ない利用者の選好情報から嗜好モデルを構築することにより、利用者層ではなく個人の嗜好を反映した商品推薦システムを、比較的少ない利用者負担で実現する事が可能になる。

i 番目の事例が n 個の特徴量ベクトル $r_i = \{r_{i1}, \dots, r_{in}\}$ ($r_i \in R^n$) と、離散的クラス情報が割り当てられたクラスデータ $l_i \in \{L_1, \dots, L_M, unknown\}$ から成る事例数 N の学習データ $\{\{r_1, l_1\}, \dots, \{r_N, l_N\}\}$ が与えられたとする。ここで、クラス情報が与えられていない事例に対して、クラスデータ $unknown$ が付与されているものとする。そのとき、CAM 法は新たな属性となる R^n の部分空間上の (変換) 関数 $F_j (j = 1, \dots, k)$ を生成する。

$F_j (j = 1, \dots, k)$ には様々な関数を与えることができるが、任意の変換関数を自動生成することを許してしまうと非常に複雑で意味を理解しがたい関数や意味のない関数を多く生成する結果にいたることが報告されているので [Muggleton 87]、CAM 法では先見知識に基づいてパラメータ化された関数をあらかじめ与える。例えば、 F_j に R^n の部分空間に対するパラメータ化されたシグモイド関数を設定することができる。

$$y = \frac{1}{1 + \exp(-p(x - x_0))} \quad (2)$$

ここで x_0 は x の平均値を表し p は $x = x_0$ における関数の傾きを表しており、 p と x_0 をパラメータとみなす。CAM 法では関数 F_j にどの属性を対応させるべきか、どのような関数を用いるべきか、および、いくつの関数を用いるべきかに関して、先見知識に基づいて専門家によって決定されることを想定している。

空間中の事例の散布は F_j のパラメータにより決定されるが、CAM 法では空間中に事例を配置した時、クラス値の異なる事例は離れて配置され、クラス値の同じ事例は近づいて配置されるように、以下の 2 つの空間の評価尺度に従いパラメータを最適化する。

- (i) クラス値の異なる事例は離れて存在するものとして、クラス値の異なる事例間の距離が大きいほど評価が高い。

- (ii) クラス値の同じ事例は近傍に存在するものとして、クラス値の同じ事例間の距離が小さいほど評価が高い。

具体的には、まず、クラス情報が既知の事例に与えられるクラスデータ $l_i \in \{L_1, \dots, L_M\}$ に対して、各クラス $L_m (m = 1, \dots, M)$ の中心座標の第 j 成分 $C_j(L_m) (j = 1, \dots, k)$ を

$$C_j(L_m) \stackrel{\text{def}}{=} \frac{\sum_{\{r_i | l_i = L_m\}} F_j(r_i)}{|\{r_i | l_i = L_m\}|}.$$

で定義し各クラスの代表ベクトルとする。そして、評価尺度 (i) に従い F_j におけるクラス値の異なる事例間の距離 D_j を

$$D_j \stackrel{\text{def}}{=} \sum_{m, m' \in \{1, \dots, M\} \wedge m \neq m'} |C_j(L_m) - C_j(L_{m'})|,$$

と定義する。さらに、評価尺度 (ii) に関してクラス値が同じ L_m である事例間の距離 $V_{j,m}$ を

$$V_{j,m} \stackrel{\text{def}}{=} \frac{\sum_{\{r_i | l_i = L_m\}} (F_j(r_i) - C_j(L_m))^2}{|\{r_i | l_i = L_m\}|}.$$

と定義して、 V_j を

$$V_j \stackrel{\text{def}}{=} \max(V_{j,1}, \dots, V_{j,M}),$$

によって定義する。すなわち、各クラスの事例間距離 $V_{j,m}$ の最大値とする。

最後に、 D_j, V_j を用いて、評価尺度 (i), (ii) を満たし、かつ、評価尺度 (i) を優先して評価する空間の評価関数 $E(D_j, V_j)$ を以下のように定義する。

$$E(D_j, V_j) = \frac{\alpha \cdot D_j}{M(M-1)} - V_j. \quad (3)$$

ここで D_j と V_j のスケールを同一にするために $M(M-1)$ を利用し、評価尺度 (i) と (ii) のどちらかをどれだけ優先して評価するかを決定する設計パラメータとして係数 α を導入した。

式 (3) を評価関数として、評価値 $E(D_j, V_j)$ が最大になるように式 (2) におけるパラメータ (例えば p と x_0) を最適化アルゴリズムを用いて最適化することにより F_j を決定し、空間中に事例を配置する。本論文の実験では、評価関数の複雑さや非線形性、局所最適解に陥る危険性を考慮して遺伝的アルゴリズムを最適化手法とした。

遺伝的アルゴリズムの世代交代や交叉に関して様々なアルゴリズムが提案されているが、初期収束を回避し、多種多様な個体の生存を意図して世代間の個体分布の差異を最小化した世代交代モデル Minimal Generation Gap(MGG)[佐藤 97] を採用した。MGG では適応度を無視して個体集団をランダムに生成し、家族の中から最良 1 個体およびルーレット選択により選ばれた 1 個体を次世代に残す戦略を取る。また、両親の遺伝子が表す実数値

ベクトルの区間を両方向に a だけ拡張した区間から、一様乱数に従ってランダムに子を生成する blx- a [Eshelman 97] を交叉オペレータとして採用した。なお、突然変異は行わないものとした。

嗜好モデル構築では、事例 $r_1 \dots r_N$ が N 個の商品、事例の特徴量 $r_i = \{r_{i1}, \dots, r_{in}\}$ が商品の n 個の特徴量、クラスデータが利用者の選好情報 ($\{L_1: \text{好き}, L_2: \text{嫌い}\}$) に対応する。本研究では利用者が商品を知覚する際に嗜好に基づいたバイアスがかかると考え、このバイアスを変換関数 F_j のパラメータに対応付け、利用者の嗜好を反映した全商品がマップされた空間を F_j のつくる空間として表現し、バイアスに基づいた商品の選好予測を空間上への商品のマッピングとして対応付けた。

F_j は構成的帰納法 [Liu 98] における新述語、あるいは 3 層ニューラルネットワークにおける隠れ層ノードに相当する。本研究では F_j としてシグモイド関数を導入し、関数の形を決定する中心座標値と中心点における傾きのふたつをパラメータとした。すなわち、空間を構成する新しい属性はシグモイド関数 F_j のパラメータにより生成され、 F_j によって作られる空間中に事例を配置した時、 L_1 の事例と L_2 の事例が離れて配置され、 L_1, L_2 それぞれの事例は近づいて配置されるようにパラメータは最適化される。これは、商品の選好はある評価尺度に従い決定されていて好きな商品や嫌いな商品にはそれぞれ何らかの共通の要素があるため、嗜好を反映した空間ではクラス値の同じ事例は広域に点在するよりもコンパクトに散布しクラス値の異なる事例は離散して散布する、という仮定に基づいている。このような仮定に基づいてベクトル空間を生成する CAM 法を用いることにより、事例の分布の自由度が限定されるため、少ない利用者の選好情報から効率的に嗜好モデルが構築されることが期待できる。

ニューラルネットワークなどと同様に CAM 法でも過学習の危険性があるので、データ数に対してなるべく少ない数のパラメータのみを用いることが重要になる。本研究では、専門家の知識辞書を用いてテキストマイニングを行うことにより属性数を減少させ、パラメータ数を少なくしている。また、専門家の先見知識によって関数 F_j のパラメータ数を少なくすることも可能であるため、過学習の問題を緩和できると考えられる。

利用者の嗜好を獲得する際には非線形性の扱いが重要であり、CAM 法による利用者個人のベクトル空間の作成は非線形変換を行うことに等しく、これはサポートベクターマシーンにおけるカーネル変換に代表される様々な非線形変換のひとつと捉えることができる。

また、情報検索の分野で用いられる LSI(Latent Semantic Indexing)[Deerwester 90] は、高次元空間にある文書ベクトルを低次元空間へと射影することにより、独立に扱われる単語ベースではなく、意味的に相関関係を持つ概念に基づいて検索を行い、検索精度の向上を図る

図 2 推薦のイメージ図: $\circ$ が正例, $\times$ が負例. 未知の商品 ($\star$) に対して正例や負例からの距離を考慮して推薦すべきか判断する. (a) は商品属性をそのまま用いた場合, (b) は CAM 法を用いて作成された認知空間, (c) は LSI によって次元を圧縮された縮退空間. CAM 法は教示信号を利用するが, LSI は属性分布のみを利用し教示信号は利用しない.

属性生成手法である. しかし, 図 2 に示すように, LSI は線形変換であり, かつ利用者個人の嗜好を表すベクトル空間を作成するものではなく, 多次元空間を圧縮することによる万人共通のベクトル空間を作成する手法であるため, CAM 法とは利用の目的が異なる.

3.4 商品推薦

CAM 法を用いて F_j を決定することにより個人の嗜好に基づいたバイアスを反映させた利用者の認知空間を作成する. この認知空間では, 利用者が好き/嫌いとした商品を含めた全商品が, 好き/嫌いな商品同士は近く, そうでない製品は遠くなるようにマッピングされている. そこで, 認知空間において, 利用者が好きと判断した商品集合の中心座標から近い順に推薦対象の商品としてランク付けすることにする. 利用者が好きと判断した商品に与えられるクラスを L_{m_0} とした際, L_{m_0} に属さない商品に対して以下の距離計算に準じた式

$$D_{Rank} \stackrel{\text{def}}{=} \sum_{j=1}^k (F_j(r_i) - C_j(L_{m_0}))^2, \quad (4)$$

(このとき r_i は $l_i \neq L_{m_0}$ を満たす)

に基づき D_{Rank} の値が小さい順にランク付けし, 閾値 β を決めて $D_{Rank} \leq \beta$ を満たす商品 r_i を推薦する.

4. 実験

提案した商品推薦システムによって妥当な嗜好モデルが獲得され, 効果的な商品推薦が可能かを検証するため, 実在する商品の推薦実験を行いアンケート結果により推薦内容を評価する.

4.1 実験設定

対象とする商品ジャンルは, 嗜好が購買に影響を及ぼすと言われていること, 購入頻度が比較的高いこと, 評判情報が入手可能であることなどを考慮してコスメ商品に決定した. まず, 検索エージェントが, あらかじめ登録

されたコスメ商品に関する商品情報サイトから商品名や商品価格, 商品仕様などの一般的な属性を収集した. そして, あらかじめ登録された口コミサイトから該当商品に関する評判情報を収集し, テキストマイニングによりそれらの商品へ感性語を割り当てた. その結果得られたデータは:

- 商品数: 374 件 (化粧品)
- 商品の一般的な属性: 価格, 容量, 人気度 (10 段階評価)
- 商品の感性語: うるおい, アンチエイジング, 低刺激, 毛穴角質ケア, 引き締め, てかり防止, にきびケア, 美白, 日焼け止め

となった.

このとき, テキストマイニングで用いた知識辞書の内容は以下の通りである.

- テキストマイニング用知識辞書の構成: 概念クラス 9 個, キー概念 90 個, 表層表現 207 個
- テキストマイニング対象文書: ひとつの商品に対して 30 件, 計 11,220 件

実験に用いた辞書に登録された表層表現を用いると, 得られた 100 件分の文書データから人手で抽出した感性語の約 86% が自動的に抽出されていた. 上記の抽出率が得られていれば, 口コミサイトの評判情報を十分反映した感性語が抽出されていると考えられる. 感性語が抽出された文書データを用いて, 先に述べた式 (1) によって各感性語の重みを各商品に割り当てた. 商品属性ベクトルは, 商品価格や商品仕様, 商品の感性情報を併せた 12 個の属性で構成される.

それぞれの属性に対して, パラメータ p_j , x_{0j} からなるシグモイド関数 F_j を導入し, CAM 法の評価値である式 (3) の値が最大になるように, 24 個のパラメータを遺伝的アルゴリズムを用いて最適化した. 遺伝的アルゴリズムのパラメータは, 予備実験の結果から以下のように設定した.

- 個体集団数: 50 個
- 交叉パラメータ (a): 0.5
- 世代数: 10000 世代

表 2 空間におけるデータの分布

被験者 i	認知空間			商品属性空間		
	L_1 の分散	L_2 の分散	クラス間距離	L_1 の分散	L_2 の分散	クラス間距離
$i = 1$	0.38	0.30	3.16	0.33	0.54	2.42
$i = 2$	0.31	1.07	2.25	0.33	0.98	1.27
$i = 3$	0.40	0.79	2.37	0.49	0.74	1.14
$i = 4$	0.24	0.34	2.43	0.39	0.49	1.48
$i = 5$	0.27	0.39	2.44	0.42	0.42	1.95
$i = 6$	0.29	0.54	2.49	0.40	0.69	2.08
$i = 7$	0.40	0.66	2.48	0.50	0.70	2.06
$i = 8$	0.44	0.70	3.12	0.47	0.77	2.44
$i = 9$	0.37	0.32	3.02	0.44	0.43	2.50
$i = 10$	0.30	0.29	2.44	0.61	0.54	3.07
平均	0.34	0.54	2.62	0.438	0.63	2.041

また、同様に予備実験から、CAM 法の評価関数式 (3) の係数 α は評価尺度 (i) を比較的優先して評価する $\alpha = 4$ とし、式 (4) のランク付けされた商品に対して推薦する商品リストを決定するための閾値は $\beta = 0.15$ とした。

4.2 商品推薦の評価

本来、推薦システムの有効性を評価するためには、推薦を行わなかった場合と推薦を行った場合の売上げの比較が必要とされる [Lawrence 01]。しかし、被験者に実際に商品を購入してもらう実験は多くの場合困難であるため、本研究では推薦した商品に対するアンケート結果をもとに推薦システムの評価を行う。本実験では、女性 10 人を被験者として 2 段階のアンケートに対して回答を依頼した。まず、第 1 段階として、被験者に好き/嫌いなコスメ商品名（もしくはブランド名）を選好情報としてアンケート調査した。調査方法は、自由記述では書きづらい、思い出せないという場合を考慮して選択式の設問と自由記述の両方を用意した。収集した選好情報をもとに CAM 法によりそれぞれの被験者の嗜好モデルを獲得し、被験者への推薦商品リストを作成した。

次に第 2 段階として、同じ被験者に対して推薦商品リストを提示し、リストの各商品に対して (1. すでに購入・利用している, 2. 購入してみたい, 3. 購入したくない) の 3 種類の回答のいずれかを求めるアンケート調査を行った。その際、知らない商品、利用体験のない商品については、エージェントが収集した商品に関する WWW 上のページを閲覧して商品に関する知識を得た上で回答できるようにした。

また、比較のため、被験者に提示する推薦商品リストには、提案システムが嗜好モデルに基づいて推薦した商品に加え、認知空間の適応を行わずに決定した推薦商品を追加した。具体的には、全商品を商品属性空間にプロットした場合に、被験者の選好情報に含まれる商品集合の中心座標からユークリッド距離が近い順に、推薦商品を

決定した。

第 2 段階のアンケート回答を集計し、個人の嗜好モデルに基いた推薦が商品属性空間上のユークリッド距離に基づく推薦と比較して利用者の嗜好にかなうかを検証する。

4.3 実験結果

まず、CAM 法によって作成された認知空間と商品属性空間における全商品データの分布の統計量を表 2 に示す。表 2 は、第 1 段階のアンケート回答で各被験者 i ($i = 1 \cdots n, n = 10$) が好き/嫌いと判断した商品集合 $\{L_1: \text{好き}, L_2: \text{嫌い}\}$ のそれぞれの集合の分散と集合の中心間の距離を計算した結果を示している。表 2 から CAM 法によって作成された認知空間では商品属性空間よりも、 L_1 の商品と L_2 の商品が離れて配置され L_1, L_2 それぞれの集合の商品は近づいて配置されていた。この結果から、CAM 法によって作成された認知空間ではクラス値の同じ事例はコンパクトに散布しクラス値の異なる事例は離散して散布しており、CAM 法の評価関数に従った認知

表 3 アンケートによる推薦内容に対する評価

被験者 i	嗜好モデル				ユークリッド距離			
	r_i	a_i	b_i	c_i	r'_i	a'_i	b'_i	c'_i
$i=1$	18	8	7	3	18	8	2	8
$i=2$	8	3	4	1	8	2	0	6
$i=3$	11	5	3	3	13	3	5	5
$i=4$	23	7	12	4	20	5	7	8
$i=5$	23	15	3	5	18	10	1	7
$i=6$	25	12	5	8	23	5	8	10
$i=7$	19	3	13	3	18	3	7	8
$i=8$	15	5	8	2	18	2	10	6
$i=9$	17	10	5	2	15	1	2	12
$i=10$	20	8	9	3	18	3	7	8
平均	17.9	7.6	6.9	3.4	16.9	4.2	4.9	7.8

表 4 母平均の差の検定

	嗜好モデル	ユークリッド距離
標本数 (n)	10	10
標本平均	$\bar{x} = \frac{1}{n} \sum_{i=1}^n \frac{a_i + b_i}{r_i} = 0.8165$	$\bar{y} = \frac{1}{n} \sum_{i=1}^n \frac{a'_i + b'_i}{r'_i} = 0.5175$
標本分散	$s_x^2 = 0.0231$	$s_y^2 = 0.0252$
s^2	0.0241	
検定統計量 T	4.304	
$t_{18}(0.01)$	2.552	

空間が形成されていることが確認された。

次に、個人の嗜好モデルに基づく推薦と商品属性空間上のユークリッド距離に基づく推薦に対するアンケート結果を比較する。各被験者 i に対して、個人の嗜好モデルに基づいて推薦される商品の数を r_i とする。この r_i 個の推薦商品のうち、被験者 i が「すでに購入・利用している」と回答した商品数を a_i 、「購入してみたい」と回答した商品数を b_i 、「購入したくない」もしくは「分からない」と回答した商品数を c_i とする。ここで、 $r_i = a_i + b_i + c_i$ である。同様に、商品属性空間上のユークリッド距離に基づいて推薦される商品の数を r'_i とする。この r'_i 個の推薦商品のうち、被験者 i が「すでに購入・利用している」と回答した商品数を a'_i 、「購入してみたい」と回答した商品数を b'_i 、「購入したくない」もしくは「分からない」と回答した商品数を c'_i とする。

各被験者 i への推薦商品に対する各回答の件数を表 3 に示す。表 3 から、個人の嗜好モデルに基づく推薦は商品属性空間上のユークリッド距離に基づく推薦に比べて、「購入してみたい」もしくは「すでに購入・利用している」という回答の比率 ($\frac{a_i + b_i}{r_i}$) が平均して約 27.2% 高いことが確認された。

本実験では、被験者の数が 10 人と比較的小数のため、統計的検定を行い結果を検証した。2 種類の方法による推薦内容に対する全被験者の評価結果を 2 つの母集団と考え、ともに母分散が等しい正規母集団と仮定して母平均の差の検定を行った。すなわち、個人の嗜好モデルに基づく推薦内容に対して被験者が「購入してみたい」もしくは「すでに購入・利用している」と回答した商品の占める割合の母平均を μ とする。同様に商品属性空間上のユークリッド距離に基づく推薦内容に対して同様の回答をした商品の占める割合の母平均を μ' とする。そして、

$$\text{帰無仮説 } H_0 : \mu = \mu'$$

$$\text{対立仮説 } H_1 : \mu > \mu'$$

として 2 種類の方法による推薦内容に対する被験者の評価に差があるかどうかを優位水準 0.01 で右側検定を行った。その結果表 4 に示すように $T > t_{18}(0.01)$ となり、仮説 H_0 は棄却され仮説 H_1 が採択された。この結果から、提案した商品推薦システムは被験者の嗜好にかなう商品推薦を行っていると考えられるため、提案手法の有効性

が確認されたと結論づけられる。

また、表 3 に示した実験結果から、商品属性空間上のユークリッド距離に基づく推薦内容に対して、「購入してみたい」もしくは「すでに購入・利用している」という回答の比率が約 53.8% になることが確認された。この結果はテキストマイニングによる商品の感性語を利用した効果であると考えられるため、商品属性として感性語を扱うことには意味があると推測できる。

5. お わ り に

本論文では、個人感性モデルのひとつである嗜好モデルを商品の選好情報から CAM 法を用いて構築し、テキストマイニングによって商品に割り当てた感性語と得られたモデルに基づいて利用者の嗜好に適した商品推薦システムを提案した。

コスメ商品を対象とした商品推薦システムを構築し、利用者によるアンケート評価を用いて提案するシステムの有効性を確認した。

本論文の実験では嗜好モデルの効果が示されたので、今後は大規模な実験で評価する必要があると考えている。また、商品属性空間上のユークリッド距離に基づく推薦内容に対する被験者の評価が高かったことから、テキストマイニングによる商品の感性語の利用に関して、商品価格や商品仕様などの情報のみを対象とした推薦との比較実験が必要であると考えている。

さらに、利用者の好き/嫌いな商品を利用者の選好情報として CAM 法により嗜好モデルを構築したが、利用者の好きな商品に対する好感度などの多値や連続値などのさまざまな選好情報が与えられた場合の処理方法に関しても検討していきたい。

謝 辞

貴重なコメントをくださった査読者の方々にこの場を借りて御礼申し上げます。

◇ 参 考 文 献 ◇

- [Beegent 00] Bee-gent: Muti-agent framework, <http://www.toshiba.co.jp/beegent/>, (2000)

- [Berry 04] M. J. Berry and G. S. Linoff: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, John Wiley & Sons Inc., (2004)
- [Bettman 79] J. R. Bettman: An Information Processing Theory of Consumer Choice, Addison-Wesley, (1979)
- [Deerwester 90] C. S. Deerwester et al.: Indexing by latent semantic analysis, Journal of the American Society for Information Science, Vol.41, No.6, pp.391-407, (1990)
- [Doorenbos 97] R. B. Doorenbos et al.: A Scalable Comprison-Shopping Agent for the World-Wide Web, In proc. of Int. Conf. on Autonomous Agents, pp.39-48, (1997)
- [Eshelman 97] J. Eshelman, E. Mathias, D. Schaffer: Crossover Operator Biases: Exploiting the Population Distribution, In proc. of Int. Conf. on Genetic Algorithms, pp.354-361, (1997)
- [Etzioni 97] O. Etzioni: Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web, AI magazine, Vol.18, No.2, pp.11-18, (1997)
- [土方 04] 土方 嘉徳: 情報推薦・情報フィルタリングのためのユーザプロファイリング技術, 人工知能学会誌, Vol.19, No.3, pp.365-372, (2004)
- [市村 03] 市村 由美ほか: 日報分析システムと分析用知識記述支援ツールの開発, 電子情報通信学会誌, Vol.J86-D-II, No.2, pp.310-323, (2003)
- [神場 00] 神場 知成, 古閑 義幸: インターネットマーケティングの技術と応用, 人工知能学会論文誌, Vol.15, No.3, pp.412-418, (2000).
- [加藤 01] T. Kato: 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～感性工房～, 情報処理学会研究報告「情報学基礎」, Vol.062, No.5, pp.33-40, (2001)
- [Kato 96] T. Kato: Cognitive user interface to cyber space database: human media technology for global information infrastructure. In proc. of Int. Sympo. on Cooperative Database Systems for Advanced Applications, pp.184-190, (1996)
- [木本 99] 木本 晴夫: 感性語による画像検索とその精度評価 情報処理学会論文誌, Vol.40, No.3, pp.886-898, (1999)
- [喜連川 01] ウェブパーソナリゼーション技術の動向, ACM SIGMOD 日本支部主催第 17 回大会 配布資料, (2001)
- [小林 03] 小林 のぞみほか: テキストマイニングによる評価表現の収集, 情報処理学会研究報告「自然言語処理」, No.154-12, (2003)
- [Lawrence 01] R. D. Lawrence et al.: Personalization of supermarket product recommendations, Journal of Data Mining and Knowledge Discovery, Vol.5, pp.11-32, (2001)
- [Liu 98] H. Liu and H. Motoda: In H.Liu and H.Motoda, editors, Feature Extraction, Construction, and Selection, Kluwer Academic Publishers, (1998)
- [Middleton 03] S. Middleton et al.: Capturing Interest Through Inference and Visualization: Ontological User Profiling in Recommender Systems, In proc. of Int. Conf. on Knowledge Capture, pp.62-69, (2003)
- [Mugleton 87] S. Mugleton: DUCe, an Oracle based Approach to Constructive Induction, In proc. of Int. Joint Conf. on Artificial Intelligence. pp.287-292, (1987)
- [Murakami 01] T. Murakami et al.: Specification of kansei patterns in an adaptive perceptual space, In proc. of Australian Joint Conf. on Artificial Intelligence, (2001).
- [Nakamura 98] A. Nakamura and N. Abe: Collaborative Filtering Using Weighted Majority Prediction Algorithms, In proc. of Int. Conf. on Machine Learning, pp.395-403, (1998)
- [那須川 01] 那須川 哲哉ほか: テキストマイニング基盤技術, 特集「テキストマイニング」, 人工知能学会誌, Vol.16, No.2, (2001)
- [Nelson 70] P. Nelson: Information and Consumer Behavior, Journal of Political Economy, Vol.78, pp.311-329, (1970)
- [Pazzani 97] M. J. Pazzani: Learning and Revising User Profiles: the Identification of Interesting Web Sites, Journal of Machine Learning, Vol.27, No.3, pp.313-331, (1997)
- [Pazzani 99] M. J. Pazzani: A Framework for Collaborative, Content-Based and Demographic Filtering, Journal of

Artificial Intelligence Review, Vol.13, No.5-6, pp.393-408, (1999)

[Resnick 94] P. Resnick et al.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, In proc. of Int. Conf. on Computer Supported Cooperative Work, pp.175-186, (1994)

[Riedl 01] J. Riedl: Recommender Systems in Commerce and Community, In proc. of Int. Conf. on Knowledge Discovery and Data Mining, pp.15, (2001)

[Salton 97] G. Salton et al.: Term Weighting Approaches in Automatic Text Retrieval, Readings in Information Retrieval, pp.323-328, (1997)

[佐藤 97] 佐藤浩ほか: 遺伝的アルゴリズムにおける世代交代モデルの提案と評価, 人工知能学会誌, Vol.12, No.5, pp.734-744, (1997)

[矢野 03] 矢野 絵美ほか: 消費者の感性モデルを利用したレコメンデーションシステムの構築, 情報処理学会論文誌, Vol.44, No.SIG8(TOD18), pp.46-54, (2003)

[吉田 01] 吉田 香, 本村 陽一: 感性データベースと連携したパーソナルモデル構築ツール IPA Technology Expo(ITX) 2001, (2001)

[Yoshioka 00] N. Yoshioka et al.: Evaluation of a Multi Agent Framework for Open Distributed Systems, IEICE TRANS. FUNDAMENTALS, Vol.E85-A, IEICE, No.11, pp.2396-2406, (2002)

〔担当委員: 庄司 裕子〕

2005 年 2 月 25 日 受理

著者紹介

村上 知子(正会員)

1996 年青山学院大学理工学部経営工学科卒業。1998 年慶應義塾大学政策・メディア研究科修士課程終了。同年(株)東芝入社。現在、同社研究開発センター知識メディアラボラトリーに勤務。機械学習、データ・テキストマイニングの研究に従事。2003 年より、慶應義塾大学大学院政策・メディア研究科後期博士課程に在籍。情報処理学会、行動計量学会、感性工学会各会員。

吉岡 信和

1993 年富山大学工学部電子情報工学科卒業。1998 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了。博士(情報科学)。同年(株)東芝入社。2002 年より国立情報学研究所に勤務。2004 年より同研究所特任助教授、エージェント技術の研究、ソフトウェア工学の研究に従事。現在に至る。情報処理学会、日本ソフトウェア科学会、電気通信情報学会各会員。

折原 良平(正会員)

1986 年筑波大学第 3 学群情報学類卒業。1988 年同大学院工学研究科電子・情報工学専攻博士前期課程修了。同年、(株)東芝入社。現在、同社研究開発センター知識メディアラボラトリーに勤務。1993-95 年にかけて University of Toronto 客員研究員。2005 年より東京工業大学連携助教授。博士(工学)。発想支援技術、類推、機械学習、データ・テキストマイニングの研究に従事。情報処理学会、日本ソフトウェア科学会各会員。

古川 康一(正会員)

1965 年東京大学工学部計数工学科卒業。1967 年同大学院修士課程(計数工学)修了。同年、電気試験所入所。1982 年(財)新世代コンピュータ技術開発機構へ出向。1992 年慶應義塾大学環境情報学部教授。1994 年より同大学大学院政策・メディア研究科教授。工学博士。この間、人工知能言語、論理プログラミング、帰納論理プログラミング、暗黙知の言語化などの研究に従事。情報処理学会、日本ソフトウェア科学会、日本認知科学会、バイオメカニズム学会、

各会員。