

決定木を用いた音楽情報フィルタリングシステムと その有効性の検証

土方 嘉徳[†] 岩濱 数宏[†] 西田 正吾[†]

インターネット上で流通する情報の中から、ユーザの好みに適合するものを推薦するサービスとして情報フィルタリングがある。近年は、テキストで記述された情報だけでなく音楽などのマルチメディアデータに対しても、その必要性が高まりつつある。情報フィルタリングの一方式として、ユーザの興味に関する情報を表すユーザプロフィールと対象情報(コンテンツ)の内容をモデル化したコンテンツモデルを比較する内容に基づくフィルタリングがある。音楽データを対象とした内容に基づくフィルタリングに関する既存の研究として、コンテンツモデルとユーザプロフィールの両方を音楽の特徴量のベクトルで表し、ベクトル間距離を用いて推薦する手法に関する研究や、それらの特徴量に任意の評価関数を用いるフレームワークに関する研究などが存在する。本研究はこれらの方法とは性質が異なるアルゴリズムである決定木を用いた内容に基づく音楽情報フィルタリングシステムを提案する。決定木は、ユーザによって重視する音楽の特徴が違って、うまくそれを表現できると思われる。実際の音楽データとユーザを用いて実験を行い、他の手法と比較しての本システムの有効性を検証する。

Content-based Music Filtering System using Decision Tree and its Evaluation

YOSHINORI HIJIKATA,[†] KAZUHIRO IWAHAMA[†] and SHOGO NISHIDA[†]

Information filtering systems, which recommend appropriate information to users from enormous amount of information on the Internet, are becoming popular. One method of information filtering is content-based filtering that compares a user profile with a content model. Many systems using the content-based filtering deal with text data, and few systems deal with music data. We propose a content-based filtering system for music data by using decision tree. And we conduct an experiment by using real music data and users, and validate the effectiveness of our system compared with other filtering methods.

1. はじめに

従来から、ユーザの情報獲得を支援する目的で、情報フィルタリングシステム(または情報推薦システム)の研究が行われてきた¹⁾²⁾。情報フィルタリングシステムとは、膨大な情報の中からユーザの好みや興味に合致する情報をユーザに提供するシステムである。情報フィルタリングシステムを実現する方式には、内容に基づくフィルタリングと協調フィルタリングの二種類がある³⁾⁴⁾。両手法とも始めはテキストデータを対象として研究が行われてきたが、特に前者の内容に基づくフィルタリングについては、フィルタリングに用いる特徴量をマルチメディアデータから抽出する困難さもあり、依然としてテキストデータを対象とした研

究が多く、マルチメディアデータを対象とした研究は少ない。

音楽を対象とした内容に基づくフィルタリングでは、音楽データから曲の特徴を表す特徴量を抽出し、それらを用いて推薦対象の音楽コンテンツのモデル(以降、コンテンツモデル)を作成する。また、ユーザの嗜好を表すユーザプロフィールも、コンテンツモデルと同様の特徴量を用いて表現する。そして、コンテンツモデルとユーザプロフィールを比較することで推薦する音楽を決定する。コンテンツモデルとユーザプロフィールをどのようにモデル化するかの、それらの比較方法にはいくつかの方法が考えられる。我々が調査したところ、コンテンツモデルとユーザプロフィールの両方を特徴量のベクトルで表し、ベクトル間距離を用いて推薦する手法に関する研究や、それらの特徴量に任意の評価関数を用いるフレームワークに関する研究などが存在する。ベクトル間距離を用いる方法は、すべ

[†] 大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University

ての特徴量を均等に扱うことに等しくなる。しかし、ユーザの嗜好は人によって重視する特徴が違っていると考えられ、必ずしも上記の比較手法が有効とは限らない。なぜなら、興味の分類に関係のない軸では、その方向に1つのクラスタが広がってしまい、距離の計算に大きな影響を与えてしまうからである。評価関数を用いた場合は、個々の特徴量の重みをユーザの好みとの相関係数などにより変化させることができるため、ユーザの音楽の好みに影響を与えない特徴量による外乱を回避することができる。しかし、評価関数を用いた場合は1つの超平面でしか分類できないため、嗜好がいくつかの塊(クラスタ)に分散するような複雑な分類を行うことができない。

また、一般的に情報フィルタリングにおいては、ユーザプロファイルのカスタマイズの容易性が重要視されることがある⁵⁾。これは、機械学習したユーザプロファイルには誤って学習された部分が避けられず、それらを微修正する必要があるためである。そこで本研究は、決定木⁶⁾を用いた内容に基づく音楽情報フィルタリングシステムを提案する。ユーザごとにユーザプロファイルとなる決定木を構築することにより、そのユーザにとって重要な特徴量のみを用いて分類することが可能になる。さらに、決定木の分類ルールは人にとって可読であるため、誤って学習された箇所を修正することが可能である。実際の音楽データとユーザを用いて実験することで、距離に基づく方法と比較しての有効性を検証する。

本論文では、2章で本研究のアプローチについて述べる。3章で関連研究について述べ、本研究との違いを明確にする。4章で本研究で使用する特徴量について述べ、フィルタリング方式を提案する。また、5章でシステムの実装と動作例を示す。そして、6章で評価実験を行い、本研究で提案する推薦手法の有効性を検証する。最後に、7章でまとめを述べる。

2. 研究のアプローチ

本研究では、音楽データの形式として MIDI を用いる。また、対象とする音楽のジャンルとしてポップス(歌謡曲)を選択する。推薦する音楽のジャンルを限定するのは、音楽のジャンルによりパートや曲の構成が大きく異なるため、推薦に使う特徴量も大きく異なってしまうためである。また、ポップスを選択したのは、最も流通量が多いというビジネス上の理由と、多くのユーザが聴いているため後の被験者実験が行いやすいという理由からである。

本研究におけるフィルタリングの概要を図1に示す。

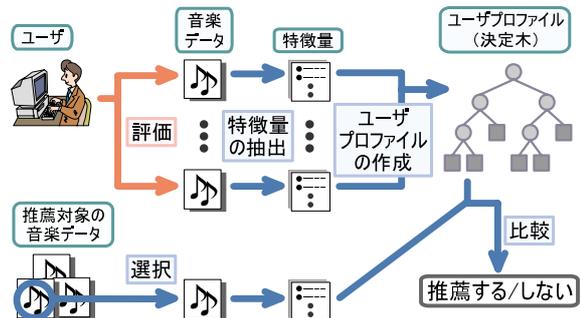


図1 本研究におけるフィルタリングの概要
Fig. 1 Overview of our filtering system.

ユーザはいくつかの音楽データを自分の好みに合うか合わないかの観点から評価付けを行う。評価付けされた音楽データからシステムは曲の特徴量(属性とその値から成る)を抽出し、ユーザの好みを表すユーザプロファイル(実際には、決定木)を作成する。次に、システムは推薦対象の音楽データから特徴量を抽出し、それとユーザプロファイルを比較する。比較した結果から、システムはユーザが気に入りそうな音楽データを選択する。そして、選択した音楽データをユーザに提示する。

また、本研究では以下の手順で研究を進めていく。

- Step1: ポップスを対象とした MIDI データの調査を行い、フィルタリングに用いる特徴量を決定する。
- Step2: Step1 で決定した特徴量を用いたフィルタリング方式を考案する。
- Step3: 蒸気方式に基づくフィルタリングシステムを実装する。
- Step4: 実際の MIDI データとユーザを用いた実験を行い、システムの有効性を検証する。

3. 関連研究

コンテンツに基づくフィルタリングは、これまでテキストデータを対象に多くの研究が行われてきた¹⁾²⁾。多くの研究で用いられている手法は、ベクトル空間モデル⁷⁾に基づき、閲覧文書を用いて適合性フィードバック⁸⁾を行うものである⁹⁾。ここでのユーザプロファイルのモデル化は、単語の生起確率に基づいており、長期間ユーザが文書閲覧していると、もともと興味のある単語だけが高頻度なベクトル要素として残ることを利用している。音楽を対象としたコンテンツに基づくフィルタリングでは、テキストにおける単語のように、曲によって有無がはっきりと出る特徴量がなく、音楽には音楽に適したフィルタリング方法を考え

る必要がある。

音楽情報処理の分野では、音楽を対象とした情報検索は数多く研究が行われている¹⁰⁾。これまで、音楽から多くの特徴量を取得することに成功しているため、これらをすぐにでもフィルタリングに応用することは可能であるように見えるが、そのような研究はまだ少ないと言える。音楽データを対象とした情報推薦に関する研究としては、まず Uitdenbogerd ら¹¹⁾の研究がある。この研究で、Uitdenbogerd らは年齢、性別及び職業などの要素が音楽の好みに影響を与えると述べている。また、音楽のジャンル分類や音楽モード重要な要素として、テンポ、調性及びリズムを挙げている。しかし、実際のフィルタリング方式を考案し、推薦システムを構築・評価するには至っていない。

また、実際に音楽推薦システムを構築した例として、Chen ら¹²⁾と黒瀬ら¹³⁾の研究がある。これらの研究は、内容に基づくフィルタリングを用いて曲の推薦を行っている。Chen らのシステムでは、音楽データから曲の特徴量を抽出し、音楽データを特徴量のベクトルで表現する。そのベクトル間距離に基づき音楽データをカテゴリに分類する。黒瀬らのシステムも同様に、音楽データから曲の特徴量を抽出する。そして、任意の評価関数を用いて推薦する音楽データを決定する。Chen らの方法では、すべての特徴量を均等に扱うことに等しくなる。そのため、ユーザごとに異なる特徴量を用いて推薦曲を決定することはできない。黒瀬らの方法では、重みを変化させることでユーザごとに重要な特徴量を選択できるが、1つの超平面で2分割することになるので、複雑な分類平面は構築できない。本研究では、ユーザごとにユーザプロファイルとなる決定木を構築し、そのユーザにとって重要な特徴量のみを用いて推薦曲を決定することにより、細かな識別能力をもつ推薦方式を提案する。

4. 特徴量とフィルタリング方式

文献¹⁴⁾において、MIDI形式の音楽データの特徴量に関する調査を行うことによって、フィルタリングにどのような特徴量を用いるかを決定した。具体的には、既存の音楽情報推薦に関する研究及び音楽情報検索に関する研究で一般的に用いられている特徴量と、この他に音楽の分類に利用できそうな特徴量を可能な限り列挙して、それらの特徴量を抽出するプログラムを実装し、実際の音楽データにおけるそれらの特徴量の分布を調査した。そして、分布のばらつきに基づきフィルタリングに用いる特徴量を決定した(表1)。具体的な特徴量の抽出方法については、文献¹⁴⁾におい

表1 フィルタリングに用いる特徴量

Table 1 Feature parameters used in our filtering method.

拍子 (曲全体)
調性 (曲全体)
平均テンポ (曲全体)
リズム (曲全体)
メジャーコードの割合 (曲全体)
マイナーコードの割合 (曲全体)
sus4コードの割合 (曲全体)
キー (曲全体)
音色 (メロディCH)
平均音高差 (メロディCH, ベースCH, コードCH)
平均音長 (メロディCH, ベースCH, コードCH, ドラムCH)
平均音長差 (メロディCH, ベースCH, コードCH, ドラムCH)

て発表済みであるので、ここでは割愛する。

次いで、決定した特徴量を用いたフィルタリング方式を考案する。分類アルゴリズムには、特徴量をベクトルで表現し、そのベクトル間距離に基づき分類する方法(K-means法¹⁵⁾、凝集法¹⁵⁾)と、特徴量の1つ1つに注目して閾値による条件分岐を行うことで分類する方法(決定木⁶⁾)の2種類が考えられる。本研究では、後者の手法、すなわち決定木を用いることとする。その理由は1章で述べたとおりであるが、ユーザの嗜好は人によって重視する特徴が違っていると思われるためと、ユーザプロファイルのカスタマイズを可能にするためである。本方式では、ユーザごとにユーザプロファイルとなる決定木を構築する。決定木のアルゴリズムとしては、パラメータとして連続値とカテゴリ変数を扱うことのできるC4.5⁶⁾を採用することとする。

ユーザは、システムが提示する音楽データに対して、“好き”、“嫌い”または“どちらでもない”の3段階で評価付けを行う。ユーザの評価をもとにして、学習を行い、決定木を作成する。ここで、決定木のノードは曲の特徴量についての条件を持ち、葉ノードは“好き”、“嫌い”及び“どちらでもない”というクラスを持つ。

本研究では、コンテンツモデルを音楽データごとに作成する。システムが音楽データから自動的に抽出した特徴量がコンテンツモデルになる。フィルタリングを行う時は、各音楽データのコンテンツモデルを使って決定木の根ノードから探索する。順に探索を行い、葉ノードに達した時にその音楽データが“好き”というクラスになれば、システムはその音楽データをユーザに推薦する。“嫌い”または“どちらでもない”というクラスになれば推薦を行わない。また、ユーザはシステムが推薦した音楽データに対して、“好き”、“嫌い”または“どちらでもない”で評価を行う。そして、このユー

ザの答えをもとに決定木を再学習させることでユーザプロファイルの更新を行う。

5. プロトタイプシステム

上述のフィルタリング方式を持つプロトタイプシステムを Java と Java サブレットで実装した。このシステムを”C-base MR (Content-based Music Recommender)”と呼んでいる¹⁶⁾。ここでは、5.1 節でプロトタイプシステムの構成について説明する。そして、5.2 節でシステムの動作例を示す。

5.1 プロトタイプシステムの処理フロー

プロトタイプシステムの構成を図 2 に示す。このシステムは、ユーザインタフェース層 (Web ブラウザ)、サブレット層及びデータベース層の 3 層から構成される。

システムの処理の流れを以下に示す。まず、メニューページでユーザがユーザ名を入力し、メニュー (アンケートを行うか、推薦を行うか) を選択する。そしてサブレットのリクエストマネージャがユーザ名とユーザが選択したメニューを受け取る。ユーザが”アンケート”を選択した場合は、アンケートページ作成モジュールがアンケートページ (図 3(a)) 用の HTML ファイルを作成し、Web ブラウザに送信する。そして、ユーザが音楽データに評価付けを行った後、システムはユーザの評価値をレイティングデータベースに保存する。ここで、レイティングデータベースとは音楽データに対するユーザ評価値を保存するもので、決定木を構築するために使用される。

ユーザが”推薦”を選択した場合は、レイティングデータベースのユーザ評価値と特徴量データベースの対象音楽データの特徴量を用いて、決定木作成モジュールがユーザプロファイルとなる決定木を構築する。次に、比較モジュールが推薦対象の音楽データを用いて決定木を探索し、推薦すべきか否かを決定する。推薦ページ作成モジュールが、推薦する音楽データを表示する推薦ページ (図 3(b)) 用の HTML ファイルを作成し、Web ブラウザに送信する。また、特徴量の抽出は特徴量抽出モジュールが MIDI データベース中の MIDI データを用いて、オフラインで行う。

5.2 システムの動作例

図 3 に、プロトタイプシステムによる推薦例を示す。ユーザは、アンケートページで音楽データに評価付けを行う (図 3(a))。ここでは、このユーザはテンポが速い音楽データに対して、”好き”という評価をつける傾向がある。次に、このユーザの評価を反映したユーザプロファイルが構築される (図 3(c))。根ノードにお

けるルールは、”テンポ (tempo) が速い”というこのユーザの評価の最も顕著な特徴を反映している。そして、システムは構築したユーザプロファイルを用いて推薦する音楽データを選択し、ユーザに音楽データを推薦する (図 3(b))。この推薦結果では、ユーザの評価を反映して、テンポが速い音楽データが推薦されている。

また、図 3(c) の決定木は、図 3(a),(b) のメニュー中のユーザプロファイルをクリックして表示されたものである。決定木中のノードとリンクには編集ページへのアンカーが埋め込まれており、これをクリックすることで、該当箇所のユーザプロファイル変更用の画面 (図 3(d)) が表れる。この画面では、中間ノードであれば特徴量の変更、特徴値の変更、葉ノード化が行える。葉ノードであれば、クラス変更、中間ノード化が行える。リンクであれば特徴値の変更が行える。また、選択している特徴量が連続値の場合は、その特徴量の最大値、最小値、平均と標準偏差を表示している。ここでは、ドラム CH の平均音長のノードで特徴量の値を変更している。

6. システムの評価

本章では、プロトタイプシステムの評価を行い、決定木を用いた推薦方法が有効であるかどうかを定量的に検証する。具体的には、実際の 200 曲の MIDI データと 10 人のユーザを用いて実験を行い、精度、再現率及びそれらの改善率を用いて他の推薦手法との比較を行う。6.1 節で実験方法について説明する。6.2 節でランダムに推薦する方法との比較を行う。これは、全く推薦に妥当性のない方法に比べてどの程度推薦の程度が良くなるのかを見るためと、学習用データとして最低どの程度の数が必要かを確かめるためである。6.3 節と 6.4 節でベクトル距離に基づき分類する方法の代表的なアルゴリズムである K-means 法と凝集法を用いて推薦した場合との比較を行う。

6.1 実験用データ

システムの評価にあたり、MIDI データ 200 曲を使用した。使用した曲の内訳は、RWC 研究用音楽データベース¹⁷⁾と、インターネット上で公開しているオリジナルの音楽データで作者に実験での利用許可をとったもの、JASRAC¹⁸⁾が著作権を持つ音楽を第 3 者が MIDI データ化したもので JASRAC と MIDI データの作者双方に利用許可をとったものを利用した。

ここで、これらのデータはユーザが知らない曲とした。これは知っている曲の場合、曲の内容ではなくそれ以外の要素 (アーティスト名など) がユーザの評価

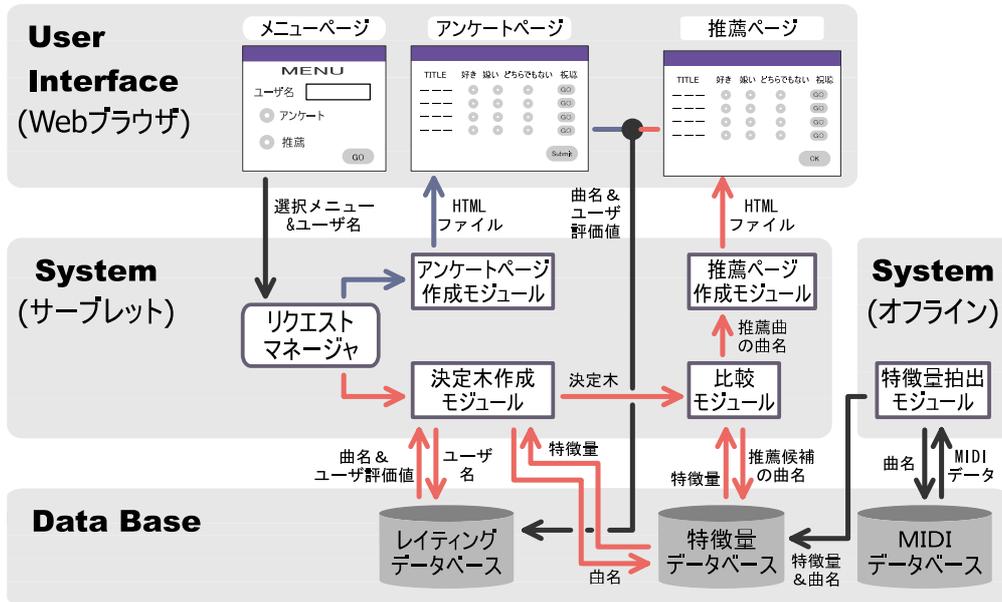


図 2 プロトタイプシステムの構成
Fig. 2 System architecture of prototype system.

表 2 ユーザごとの評価の分布
Table 2 Users' ratings.

ユーザ	好き	嫌い	どちらでもない
ユーザ A	87	53	60
ユーザ B	54	51	95
ユーザ C	71	43	86
ユーザ D	69	91	40
ユーザ E	91	29	80
ユーザ F	109	45	46
ユーザ G	72	74	54
ユーザ H	79	12	109
ユーザ I	109	43	48
ユーザ J	76	15	109

に影響を与える可能性があるからである。次に、ユーザ 10 人にこれらの 200 曲を試聴してもらい、すべての曲に対して好き、嫌いまたはどちらでもないの 3 段階で評価付けを行ってもらった。ユーザごとの評価値の分布を表 2 に示す。そして、MIDI データ 200 曲を評価値のばらつきが均等になるように 2 分割し、100 曲を決定木を構築するための学習用データ、残り 100 曲を決定木の評価用データとした。

6.2 ランダムに推薦する方法との比較

ここでは、決定木を構築するための学習用データを、25、50、75 及び 100 と変化させ、各ケースにおいて提案手法とランダムに推薦する方法の比較を行う。そして、提案手法が有効に働くために必要な学習用データ数を調べる。図 4 に、学習用データ数を変化させた場合のランダムに推薦する方法に対する精度について

ランダム推薦との比較

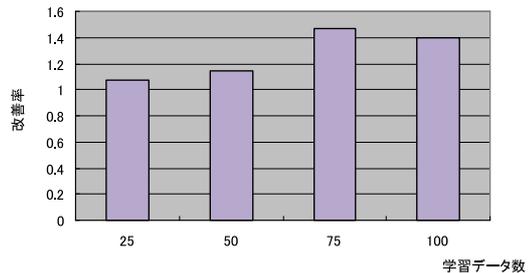


図 4 ランダム推薦に対する精度の改善率
Fig. 4 Improvement rate of precision to random recommendation.

での改善率の平均を示す。再現率の改善率は精度の改善率に等しくなるため、ここでは精度の改善率のみを示す。

図 4 より、学習データ数が少ないときは、改善率が 1.1 程度であり提案手法とランダムに推薦する方法の精度の間にあまり差が見られなかった。これは、学習データ数が少ないときは決定木を構築するために十分な学習を行うことができなかったからである。また、図 4 より学習データ数が 75 曲以上の場合は改善率が 1.4 を超えており、学習データ数が 50 曲以下の場合と比較して精度が大きく向上していることが分かる。75 曲(及び 100 曲)の学習データを用いた場合に提案手法とランダムに推薦する方法の精度に有意差があるかどうかを調べるために t-検定(片側)を行った。その

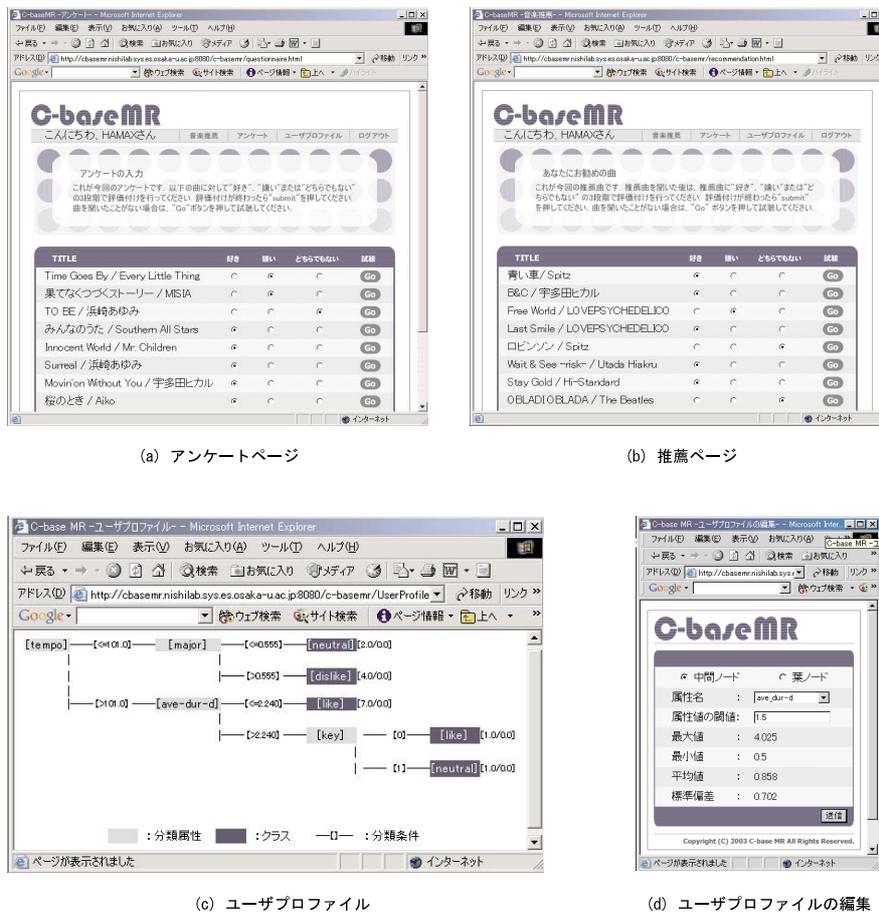


図 3 プロトタイプシステムの画面例
Fig. 3 Screenshot of prototype system.

結果，有意水準 $\alpha = 0.05$ で提案手法の精度とランダム推薦の精度で有意差が見られた．このことから，学習データ数が 75 曲以上の場合は提案手法の方がランダムに推薦する方法よりも有効であるといえる．すなわち，本研究で提案する推薦手法が有効に働くには最低 75 曲程度の学習データが必要になると考えられる．

6.3 K-means 法を用いた推薦方法との比較

学習用データとして 75 曲を使用し，提案手法と K-means 法による推薦方法の比較を行う．各手法の精度と再現率を図 5 に示す．図 5 より，再現率は同等で，精度は提案手法を用いた場合の方が K-means 法を用いた場合よりも精度が高くなったことが分かる．これは，提案手法を用いて推薦したデータ数が K-means 法を用いて推薦した場合よりも少なかったからである (図 6 参照)．提案手法を用いた場合と K-means 法を用いた場合で，精度に有意差があるかを調べるために t-検定 (片側) を行った．その結果，有意水準 $\alpha = 0.05$ で有意差が見られた．このことから，K-means 法を用

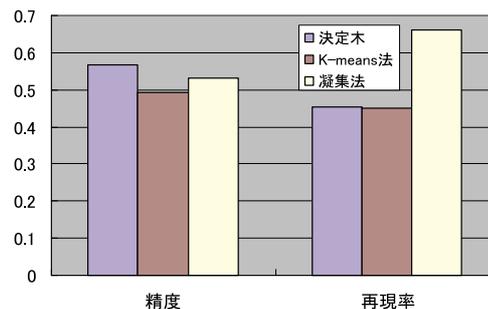


図 5 提案手法 (決定木) とベクトル距離に基づく推薦方法 (K-means 法と凝集法) の比較
Fig. 5 Comparison between our method (decision tree) and distance-based method (K-means method and tree clustering).

いるより提案手法を用いて生成した分類ルールの方が，好きな曲をより絞り込んで推薦できたことが分かる．次に，なぜ提案手法を用いて生成した分類ルールの方が精度が高くなったのかを調べるために，実験に参

ユーザ名	A	B	C	D	E	F	G	H	I	J
提案手法	27	7	7	23	43	81	61	5	87	11
K-means法	35	3	3	52	46	85	47	11	100	13
凝集法	96	0	5	0	95	85	3	0	99	0

図 6 各手法を用いた場合の推薦データ数
Fig. 6 The number of recommended music data in each method.

表 3 アンケート結果の例

ユーザ名	アンケート結果の例
ユーザ A	<ul style="list-style-type: none"> テンポが遅く、メロディラインがしっかりしている曲に好きという評価を付けた。 テンポが速いと感じた曲やロック（低音が多い）やパンク色の強い曲には嫌いという評価を付けた。 どちらにも該当しないものに、どちらでもないという評価を付けた。
ユーザ B	<ul style="list-style-type: none"> ドラムが一定速度かつ速い曲に好きという評価を付けた。 同じベース音が繰り返している曲に好きという評価を付けた。 ミディアムテンポで普通な感じの曲に嫌いという評価を付けた。 どちらにも該当しないものに、どちらでもないという評価を付けた。
ユーザ C	<ul style="list-style-type: none"> テンポが速い曲に好きという評価を付けた。 テンポが普通かつ明るい感じの曲に好きという評価を付けた。 テンポが遅くかつ明るい感じの曲に好きという評価を付けた。 テンポが普通かつ暗い感じの曲にどちらでもないという評価を付けた。 テンポが遅くかつ暗い感じの曲に嫌いという評価を付けた。

加したユーザ 10 人にどのような基準で曲に評価値をつけたかを尋ねるアンケートを行った。アンケート結果の例を表 3 に示す。表 3 から、今回実験に参加したユーザは曲の全体的な雰囲気ではなく、数種類の特徴に着目して評価を付けている傾向があることが分かる（「テンポが速い曲が好き」「テンポの遅い曲が嫌い」など）。ユーザ A の決定木を図 7 に示す。図 7 より、1 つ目の分類属性としてテンポ、2 番目の分類属性として平均音長差（メロディCH）が選択されていることが分かる。アンケート結果からテンポとメロディCHに関する特徴がユーザ A の音楽の嗜好に強い影響を与える特徴量になっていることが確認でき、このことが決定木に強く反映されていることが分かる。それに対して K-means 法を用いて分類ルールを生成した場合は、このように数種類の特徴量が音楽の好みに強い影響を与えていると、そうでない特徴量はベクトル間距離を計算するとき外乱になる。

さらに表 3 から、テンポなど多くのユーザの音楽の嗜好に強い影響を与える特徴もあるが、メロディに関

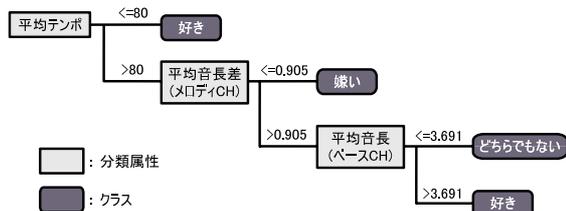


図 7 ユーザ A の決定木 (学習用データ 75 曲)
Fig. 7 Decision tree of User A (75 data for learning).

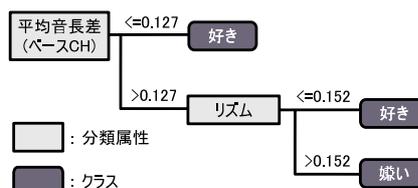


図 8 ユーザ B の決定木 (学習用データ 75 曲)
Fig. 8 Decision tree of User B (75 data for learning).

する特徴やベースに関する特徴などは、それに注目するかどうかはユーザによって異なることが分かる。具体的には、ユーザ A はテンポやメロディに関する特徴を重視しているが、ユーザ B はドラムやベースの音を重視している (表 3 参照)。ユーザ B の決定木を図 8 に示す。図 7 と図 8 より、各決定木が使用している分類属性はユーザによって異なり、各ユーザの好みを反映した分類属性を用いていることが確認できる。K-means 法では全てのユーザにおいて重要な属性が決まっていれば、その属性だけを用いてより良い精度で推薦を行うことも可能である。しかし、そのような仮定は現実には有り得ず、各ユーザにとって重要でない特徴量が入ることはやむを得ないと言える。これらのことから、ユーザの嗜好に特定の特徴が強く影響を与え、なおかつユーザによって重要な特徴が異なる音楽には、K-means 法よりも提案手法の方がユーザの好みをよく表すとと言える。

6.4 凝集法を用いた推薦方法との比較

学習用データとして 75 曲を使用して、提案手法と凝集法による推薦方法の比較を行う。凝集法は他の手法と違って、10 人中 4 人のユーザにおいて、一曲も推薦されない結果となった (図 6 参照)。これらのユーザは他の手法においても有効なクラスタができていなかったグループではあるが、1 曲も推薦されなかったのは凝集法のクラスタリング方法に原因があると考えられる。凝集法は、ボトムアップにクラスタリングを行うため、上位ノードにおいては、多次元空間において離れた空間に位置する 1 つまたは少数のデータからなるクラスタを分岐していきがちである (図 9 に、ユーザ

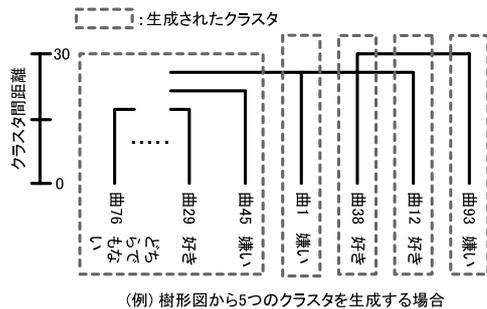


図 9 ユーザ B の凝集法により生成されたクラスター
Fig.9 Clustering result of User B by tree clustering.

B の凝集法のクラスタリング結果を示す。図の下部の”曲 n ”は、各曲に対応しその番号が記述してある。決定木の葉ノードの数になるように上位ノードで区切るため、ユーザの評価における最も多い評価値が”好き”でない場合、極端なケースでは少数のデータからなる $n - 1$ 個のクラスターと、多数のデータからなる 1 個の一般的なクラスター (評価値の分布から”好き”以外のクラスターとなる) ができることがある。この場合は、1 つもデータが推薦されないことも有り得る。データが推薦されたユーザは、再現率が高くなっているが (図 5)、これはユーザの評価における最も多い評価値が”好き”の場合、上記で説明した 1 個の一般的なクラスターの最頻出のクラスが”好き”であったためである。図 6 から分かるように、推薦されているユーザには非常に多くの曲が推薦されている。このように凝集法は、学習データのクラス分布に依存してしまうため、適用が困難であると言える。

7. むすび

本稿では、音楽情報を対象とした決定木を用いた内容に基づくフィルタリング方式を提案し、Java と Java サブレットを用いてプロトタイプシステムを構築した。従来手法である、ベクトル距離に基づく方法と比較したところ、決定木の方がユーザの重視する音楽の特徴量をうまく表現でき、ユーザの好む曲をより絞り込んで推薦することが分かった。今後ユーザプロフィール編集に関する評価を行う予定である。

参考文献

- 1) Resnick, P. and Varian, H.R.: Recommender Systems, *Comm. of the ACM*, Vol.40, No.3, pp.56-89 (1997).
- 2) Loeb, S. and Terry, D.: Information Filtering, *Comm. of the ACM*, Vol.35, No.12, pp.26-81 (1992).
- 3) Ramakrishnan, N.: PIPE: Web Personaliza-

- tion by Partial Evaluation, *IEEE Internet Computing*, Vol.4, No.6, pp.21-31 (2000).
- 4) Riecken, D.: Personalized Views of Personalization, *Comm. of the ACM*, Vol.43, No.8, pp.26-158 (2000).
- 5) Vassileva, J.: A Practical Architecture for User Modeling in a Hypermedia-Based Information Systems, *In Proceedings of 4th International Conference on User Modeling*, pp.115-120 (1994).
- 6) Quinlan, J.R.: C4.5 Programs for Machine Learning, Morgan Kaufmann (1993).
- 7) Salton, G. and McGill, M.J.: Introduction to Modern Information Retrieval, McGraw-Hill (1983).
- 8) Meadow, C.: Text Information Retrieval Systems, Academic Press (1992).
- 9) 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, *人工知能学会学会誌*, Vol.19, No.3 (2004).
- 10) ISMIR: The International Conferences on Music Information Retrieval, <http://www.ismir.net/>
- 11) Uitdenbogerd, A. and Schyndel, R.V.: A Review of Factors Affecting Music Recommender Success, *In Proceedings of International Symposium on Music Information Retrieval (ISMIR 2002)*, (2002).
- 12) Chen, H. and Chen, A.L.P.: A music recommendation system based on music data grouping and user interests, *In Proceedings of the tenth international conference on Information and knowledge management*, (2001).
- 13) 黒瀬崇弘, 梶川嘉延, 野村康: 感性情報を用いた楽曲推薦システム, 第 14 回データ工学ワークショップ (DEWS2002), 8-P-6 (2003).
- 14) 岩濱数宏, 土方嘉徳, 西田正吾: 内容に基づく音楽情報フィルタリングシステム, データベースと Web 情報システムに関するシンポジウム (DB-Web2003), pp.69-76 (2003).
- 15) Berry, M.J.A. and Linoff, G.: Data Mining Techniques, For Marketing, Sales, and Customer Support, Wiley Computing Publishing (1997).
- 16) インターメディアフォーラム 2003, C-base MR: 内容に基づく音楽情報フィルタリングシステム, 大阪大学西田研究室, 2003 年 10 月, マイドーム大阪, <http://www.jma.or.jp/imf/>
- 17) RWC 研究用音楽データベース, <http://staff.aist.go.jp/m.goto/RWC-MDB/index-j.html>
- 18) 社団法人日本音楽著作権協会 JASRAC, <http://www.jasrac.or.jp/>