音楽ランキングにおけるクラスタリング手法の研究

2004MM033 金森弘晃
指導教員: 松田眞一

1　はじめに

私はWeb上のTop Hits Online[1]（以下THO）という音楽ランキングを高校生の時に発見し、よく見て楽しんでいた。このランキングはＣＤの売り上げなどから作成されているチャートとは違い、登録されている世界中のランキングの全てを合計し毎週ランキングを発表しているため、人々が今どの曲を好んで聴いているかが良く分かるのが利点である。
本研究では音楽ランキングにおける曲の変動パターンをつかむクラスタリング手法を提案する。特に視覚化したとき動きが似ていると感じたものの距離が近くなるようにしたい。データは順位をそのまま使用するのではなく、共に発表されている各週ごとのポイントの変遷・推移から解析を行う。

2　関連研究

この研究は井手[2],G.Das et al.[4]の応用である。関連研究では部分時系列クラスタリングという時系列データからのパターン抽出方法を提案している。この方法は1本の長い時系列データを多数の部分系列に分け、それらをそれぞれ独立なベクトルとしてクラスタリングしようというものである。このクラスタリングをし、できたクラスタ平均を代表パターンとして取り出すことが目的である。また、この長い時系列ベクトルから多数の部分ベクトルを作る際に用いるのがsliding window techniqueという方法である。この方法は一つの長い時系列データを固定窓幅ωで区切り、少しずつずらしていって部分時系列ベクトルを作ることである。

3　解析方法

方針としては一般のクラスタリング手法を用いる。まず、それぞれの曲に対し2曲間の距離を求め、その距離にもとづいてクラスタリングを行う。問題は距離の算出方法であるが曲によって登場週数、つまりデータ数にばらつきがある。例えばある曲は18週しかチャートインしなかったため、18週分のデータしかないが、ある曲は26週分のデータがあるためデータ数に差があるものを比較する新たな方法が必要となる。

3.1　週数幅ω

長い一つの曲をどのくらいの幅で分けていくかという週数幅の決定に関しては井出ら[3]に記述されている。しかし今回のようにランクイン週数の違う2曲を扱う場合にどのような週数幅を使えばよいのかは不明である。そのためいくつかの曲で週数幅を変えて実験したところ、ある曲の上昇期に対し別曲の下降期との距離が小さくなってしまうことがあった。この問題を解決するためある程度安定する15週に週数幅を設定した。

3.2　補正

2曲間距離を計算する前に、それぞれの曲についてポイントの補正が必要となる。もしそのままクラスタリングを行うと「大ヒットした群」「中くらいのヒットした群」「ヒットしなかった曲」の3群に分かれてしまうのは明らかで、変動パターンをつかむ今回の目的とは一致しない。補正には平均ポイントや最高ポイントなどそれぞれの曲に関する何らかのポイントを計算し、そのポイントで各週のポイントを割ることによって全ての曲のポイントをそろえる必要がある。今回曲のポイント補正方法として3種類考えた。以下にそれぞれの特徴をまとめる。

1. 最大値で割った場合は、ヒット規模で群分けされにくく最高ポイントをある程度無視したクラスタリングができる。しかしポイント変動タイプが加味されにくく、ヒットタイミングで差ができにくい。また外れ最高値がある場合その結果に左右されてしまう。
2. 平均値で割った場合は、ポイント変動タイプによって平均値が変わるため、ヒットタイミングを考慮したクラスタリングができる。しかしヒット規模で群分けされやすく、大ヒット群と小ヒット群に分かれてしまう。また平均値なので下位が長く続くとその結果に左右されてしまう。
3. 上位値の平均で割る方法は最高値と平均値の間をとったような方法である。外れ値があっても、下位が長く続いても左右されにくく、安定している。

3.3　2曲間距離の算出方法

2曲間距離の算出にはまず、井出[2]のsliding window techniqueを使って１曲をいくつかのパートに分ける。井出[2]では対象となる時系列が1つのみだったため一方のみをスライドさせる方法であったが、ここでは2曲間比較であるため両方のデータをスライドさせる点で違いがある。本研究ではこの音楽ランキングの場合を例に出す。固定窓幅ω（週数幅）を15週とすると1つ目$A_{1r}(r = 1, 2, ..., \omega)$は1週目から15週目まで、2つ目$A_{2r}(r = 1, 2, ..., \omega)$は2週目から16週目まで、3つ目$A_{3r}(r = 1, 2, ..., \omega)$は3週目から17週目までという具合に部分時系列を作っていく。最終的には1つの曲から登場週数nに対し$n - \omega + 1$個の部分時系列が生成されることになる。 できた1曲目の部分時系列Aと2曲目の部分時系列Bを

$$\begin{cases} A_{ir}(i = 1, 2, ..., n - \omega + 1) \\ B_{jr}(j = 1, 2, ..., m - \omega + 1) \end{cases} \quad (1)$$

とおくことにする。nは1曲目の登場週数、mは2曲目の登場週数である。

次に実際にこの2曲間の距離を計算するのであるが、それにはまず1曲目と2曲目の部分時系列同士のユークリッド距離を計算する必要がある。$d(x, y)$をxとyベクトル同

士のユークリッド距離として、部分時系列間の距離行列$\{X_{ij}\}$をA_{ir}、B_{jr}を使い以下のように定義する。

$$X_{ij} = d(A_{ir}, B_{jr}) \tag{2}$$

3.4 距離行列の意味

距離行列$\{X_{ij}\}$は曲の全ての一部分どうしの距離をあらわしたものである。なぜこのような操作をしたのかというと曲によってヒットしはじめるタイミングが違うからである。ただ単に人気が急上昇した時期が少しずれていただけの2曲間距離は近くなるのが理想とされる。そこでこの距離行列は全ての一部分どうしの距離を計算しているため、2曲間で最も距離の近くなった週どうしを選んでその2曲の総合的な距離として適用することができる。

3.5 2曲間距離

次に2曲間の距離行列を一つの数値にまとめる。2曲間距離Dは以下の式で定義する。

$$D = \frac{1}{2}\left(\frac{\sum_{k=1}^{n-\omega+1} \min X_{kj}}{n-\omega+1} + \frac{\sum_{k=1}^{m-\omega+1} \min X_{ik}}{m-\omega+1}\right) \tag{3}$$

まずそれぞれの行方向について最小値をとり、その最小値の平均をとる。同様にそれぞれの列方向についても最小値をとり、その平均をとる。最後に行方向の平均値と列方向の平均値の平均を2曲間距離とする。

4 解析結果

以上の方法で解析を行うこととする。今回の解析では週数幅ωを15週、またポイント補正方法は上位5値の平均で割る方法をとった。今回使用するデータは2003年の100曲である。解析で得られた結果を左から大きく5群に分けることにする。

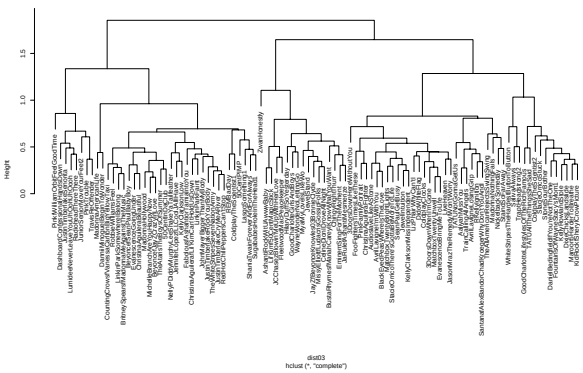

図 1: 2003年デンドログラム

- 第1群
  全ての曲が21週以下とランクイン期間が短い曲がこの群を占める。
- 第2群
  この群の登場週数を見てみると18週から32週まで幅広いが、第1群に近いスピードで上昇しヒットタイミングが早いのが特徴である。
- 第3群
  全ての曲が22週以下とランクイン期間が短い。しかしランクイン期間が短いがヒットタイミングは平均的で、その分ピーク後の下降スピードが速い。
- 第4群
  この群は20週から最高は46週まで幅広い曲が属すロングヒット群である。ロングヒットであるがどの曲も上昇スピードは速く、わりとしっかりしたヒットの山を持つ曲群である。
- 第5群
  この群の大きな特徴は序盤のポイント上昇が他のどの群よりも緩やかなことである。曲のピークが20週前後の曲が多いが、その反面ピーク後は一気に下降しランク外になる曲が多い。

4.1 まとめ

第1群と第2群と第3群は良く似ていて違いが分かりづらいが、単にランクイン期間が短い第1群の「短期間ヒット群」に対し第2群は下降スピードが平均的なため「ヒットタイミング早期群」。また第3群はランクイン期間が短い割にヒットタイミングが少し遅めで、下降が急激なため「ヒットタイミング末期群」とそれぞれ現行の手法でとても細かい特徴まで読み取れていることが分かった。他にも第4群と第5群のようにどちらも30週前後のロングヒット曲が多い中でヒットタイミングによる群分けができていた。第4群はヒットがしっかりしていた「ロングヒット群」。また第5群は上昇スピードは緩やかであるが一気に下降する「急下降群」であった。

5 おわりに

本研究の目的は音楽ランキングにおける曲の変動パターンをつかむクラスタリング手法の提案であった。変動パターンという点に関しては解析結果にも反映され良い手法ができていた。ランクイン週数やヒットタイミングが違っても似ている変動をしたものについてはまとめることができ、「視覚化した場合」という大きな指標は満たしていた。

参考文献

[1] Top Hits Online:http://www.tophitsonline.com/.

[2] 井手剛：部分時系列クラスタリングの理論的基礎, 第20回人工知能学会全国大会予稿集, 2A1-2, 2006 http://spinglass.hp.infoseek.co.jp/.

[3] 井手剛、井上恵介:非線形変換を利用した時系列データからの知識発見, 第4回データマイニングワークショップ, 日本ソフトウェア科学会データマイニング研究会, 研究会資料シリーズ ISSN 1341-870X, No.29, 2004, pp.1-8. http://spinglass.hp.infoseek.co.jp/.

[4] G.Das, K.-I. Lin, H.Mannila, G. Renganathan, and P.Smyth. Rule discovery from time series. In Proc. the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1998.