2005－FI－81（4）
2005／11／2

プレーリストからの曲目やアーティストの相互関連抽出

廣川佐千男 †1　伊東栄典 †1　下司義寛 †2
Yufeng Dou †3　池田大輔 †4

iPod に代表される携帯メディアプレーヤーの出現により音楽業界の状況が大きく変わりつつある。iTMS に登録された個人的プレーリスト群に現れる曲目やアーティストの出現頻度の解析により、相互関連を表す概念グラフを構築する方式を提案する。

Relation Extraction of Songs and Artists from iPod Playlists

Sachio Hirokawa †1 ,Eisuke Itoh †1 ,Yoshihiro Shimoji †2
,Yufeng Dou †3 and Daisuke Ikeda †4

Portable music players are chaning the music entertainment environment. They can download the music through network. This paper proposes a method for extracting relation and popularity of songs and artists which appear in playlists of iTMS(iTunes Music Store).

1. はじめに

Web の発達にともない、誰でも自由に様々なデータを提供できるようになってきた。個人でも企業でも、単独のデータだけでなく、多数の同種類のデータをまとめて提供するサイトが増えている。例えば、商品、本、音楽など、個人の好みのものをリストとして他の人が見られるようにするページが増えている。従来の文字情報だけでなく、iPod などの携帯音楽プレーヤーとネットワークによる音楽配信サービスという新たなネットワークビジネスが起こっている。例えば、iTMS☆ Music Store には自分が好きな曲を選んでプレーリストを作り登録できる。これらのプレーリストは専用のソフト iTunes を使うことで、他の人に伝えることができ、さらに検索したり曲ごとにダウンロード（購入）できる (図 1)。一つのプレーリストには、20 曲程度の曲が並んでいる。各曲は、曲名、アーティスト名、ジャンルなどの情報がそれぞれ固有の ID とともに表されている。

図 1　iTunes

このように多数の人の推薦情報を用いて、本や映画を推薦するシステムは協調フィルタリングとして知られている [8]。また、近年、音楽配信サービスの普及にともない、研究の段階から具体的なビジネスのための特許として出願されるようになってきた。Creative Technology 社は、メタデータを使った曲の自動階層分類についての特許 [5] を取得している。Microsoft 社は曲の中で認知される音量、テンポ、楽器などと信号

†1 九州大学情報基盤センター
Kyushu Univeristy, Computing and Communications Center
†2 九州大学理学部物理学科
Kyushu Univeristy, Department of Physics
†3 九州大学大学院システム情報科学府
Kyushu University, Graduate School of Information Science and Electrical Engineering
†4 九州大学附属図書館
Kyushu Univeristy, Library
☆ http://www.apple.com/itunes/store/

処理技術を合わせることにより曲を分類し、類似曲の検索を可能にする特許[9]を取得している。

本論文では、このようなプレーリストから、関連する曲を抽出しさらにそれらの曲の相互関係を可視化した概念グラフを構築する手法を提案する。このような概念グラフの構成は、曲だけに限らず、アーティストや曲のタイトルに現れるキーワード群についても同様に行なうことができる。

2. プレーリスト群に現れる曲とアーティストの頻度解析

Apple の音楽配信サイト iMix には 2005 年 8 月 26 日の時点で 332,560 個のプレーリストが登録されていた。我々は、その中から約 1/20 にあたる 13,480 個のプレーリストをランダムに選択しダウンロードした。これらについて予備的分析[2]を行ない、約 60%に当たる 7,919 個のプレーリストでは 5〜20 曲程度しか含まれていないことを確認した。全体の平均としては、一つのプレーリストに 24 曲が含まれている。また、一つのプレーリストに含まれる曲数の分布については、Zip の法則が成り立つことを確認した。

より具体的な分析として、出現頻度 20 位までの曲やアーティストの一覧を表 1 と表 2 に示す。このようにランキングにより人気の度合を表すことは広く用いられている。しかし、単純に並べただけでは、単独の曲やアーティストが人気で何位になっているかという局所的な解析しかできない。曲の推薦を考えた場合、このようなランキングでは単に現在人気のある曲を、ユーザーの嗜好に関係なく示すことしかできない。

ある曲、あるいはあるアーティストを好きなユーザーにとっては、単に人気のある曲やアーティストよりも、その曲やそのアーティストが好きな人が推薦するものに興味があると考えられる。そこで、プレーリストに出現する曲やアーティストのうち 100 回個以上のプレーリストに出現するものを対象として、どの曲とどの曲が共起する度合が高いか、どのアーティストとどのアーティストが共起する度合が高いか調べた。表 4、表 3 は曲のペア、およびアーティストのペアで上位 20 位までの一覧を表す。二つ並んだもののうち、左側のものの方が単独順位は上位となっている。

しかし、このように共起頻度するものについて頻度を求めても、2 つ間に強い関係があることが分かるだけで、全体の中でその二つの関係がどのような位置を占めるかという大局的な分析はできない。曲のペアについての人気度を示す表 4 において、第一キーを曲、第二キーを共起頻度としてソートすると、一つの曲と共起する曲を共起の度合いが高い順番で見ることができる。しかし、これはその曲についての解析だけしかできない。その曲と共起する曲がさらに他のどのような曲と共起するか調べるには、2 番目の曲についてソートし直さなければならない。本論文で提案する方法は、このような共起する様子を大局的に眺めることができる。

表 1 上位 20 曲

順位	頻度	曲目
1	583	This Love
2	513	Ocean Avenue
3	510	The Reason
4	493	Boulevard of Broken Dreams
5	480	She Will Be Loved
6	413	One, Two Step
7	383	Float On
8	376	Let's Get It Started (Spike Mix) [Bonus Track]
9	356	American Idiot
10	336	Soul
11	320	Mr. Brightside
12	320	Holiday / Boulevard of Broken Dreams
13	306	Hey Mama
14	298	American Idiot
15	286	Are You Gonna Be My Girl
16	281	Somebody Told Me
17	279	I Fought the Law
18	269	Harder to Breathe
19	268	In da Club
20	264	Hey Ya!

表 2 上位 20 アーティスト

順位	頻度	アーティスト
1	1792	Green Day
2	1262	Maroon 5
3	1195	U2
4	1157	Eminem
5	970	OutKast
6	913	50 Cent
7	883	Coldplay
8	866	Jet
9	850	Black Eyed Peas
10	840	Blink-182
11	780	Yellowcard
12	761	Hoobastank
13	745	Nirvana
14	726	Modest Mouse
15	709	The Killers
16	646	Britney Spears
17	630	Counting Crows
18	619	The White Stripes
19	609	Ludacris
20	602	Guns N' Roses

表 3　上位 20 組（アーティスト）

順位	頻度	アーティストのペア	
1	440	Eminem	50 Cent
2	403	Green Day	Blink-182
3	390	Green Day	Maroon 5
4	362	Green Day	Eminem
5	357	Green Day	The Killers
6	342	Green Day	Yellowcard
7	338	Green Day	Jet
8	320	Green Day	U2
9	312	50 Cent	Ludacris
10	307	Green Day	Simple Plan
11	303	Green Day	Good Charlotte
12	298	Maroon 5	Hoobastank
13	298	Green Day	Sum 41
14	297	Eminem	Ludacris
15	294	Green Day	Bowling for Soup
16	287	Eminem	OutKast
17	284	Maroon 5	Jet
18	283	Green Day	Hoobastank
19	282	OutKast	Black Eyed Peas
20	271	Yellowcard	Hoobastank

3. ランキングやクラスタリングによる検索結果表示の問題

ランキング

プレーリストを文書と考え曲をキーワードと捉えると、複数の文書群における特徴的キーワードの抽出と抽出されたキーワードの関連を分析する問題と捉えることができる。特に、特定の曲、あるいはアーティストを含むプレーリストを対象文書群とすると、その曲やアーティストに関連する曲やアーティストを求めることになる。

多数のデータを大局的にみる方法として様々な可視化技術の研究がある。検索結果のファイル群やそれらに現れるキーワードの分析としては、何らか一つの尺度によるランキングをつけて、一次元に表示にすることが広く用いられている。しかし、ランク付けしたとしても、読むためには時間がかかり実質的に数十個しか読まない。内容的に異なるものが隣接して表示されることもある。近い内容的のものが遠く離れて表示されることがある。複数の観点で眺めることができるデータであれば、どのように工夫しても、一次元表示を選べば、関連の強い項目であっても離れて表示されることになる。

クラスタリング

クラスタリングは項目間の関係を平面的あるいは立体的に配置することにより可視化する。ばらばらだったものをまとめるクラスタリングしたり、ひとつにまとまっていたものについてより詳細な構造を調べるために分割したりする。また、可視化された単語群の意味をつかむために、その単語群に名前をつける、あるいは、単語群から代表元を求める。各クラスタを理解するために適当な名前をつけなければならない。クラスタの配置では「似たものを近く」配置することができるが、上下左右の一関係に特別な意味はない。

4. 頻度を用いた概念グラフ

本稿では、プレーリスト群に現れる曲の出現頻度を用いて、二つの曲の上位下位関係について新しい定式化を与える。二つの曲がランキングにおいて上位下位関係にあったとしても、聞く人達がそれぞれ異なっていればその二つの曲は関係ない。しかし、下位の曲を聞いている人の大半が上位の曲を聞いているような状況では、上位の曲はその下位の曲しか知らない人にとって興味のある曲と考えられる。また、上位の曲の曲しか知らない人にとっても、下位の曲の曲はそれほどメジャーではなくても、ちょっと気になる曲と考えられる。

このような観点から、文書集合における二つの単語の上位下位の関係を定義する [6]。D を文書集合、w を単語とする。w が現れる D 中の文書の個数（文書頻度）、すなわち、$\sharp\{d \in D \mid w \text{ が } d \text{ 中に現れる}\}$ を $df(w, D)$ で表す。二つの単語 u, v の両方が現れる文書数を $df(u * v, D)$ で表す。単語 u と v について、$df(u*v, D)/df(v, D) > 0.5$ かつ $df(u, D) > df(v, D)$ となっているとき、「文書頻度の観点から u は v の上位である」ということにする。

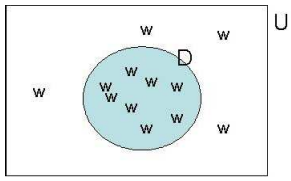

図 2　上位下位関係

5. システム概要

ダウンロードした 13,480 個のプレーリストについての検索システムを作った。検索エンジンの主要部分には、GETA[4] を用いた。ユーザーは、曲あるいはアーティストを探すためのキーワード w を入力する。システムはまず、そのキーワードを曲名あるいはアーティスト名の中に含むプレーリストの一覧 D を求める。w と関連あるキーワード u は、$df(u, D)/df(U) > 0.5$

となる単語として定義する。ただし、U は 13,480 個のプレーリスト全体の文書群とする。曲 w に関連のある曲 u とは、その曲 u を含むプレーリストの過半数のプレーリストにおいて、w が現れていることを意味する（図 2）。概念グラフのノードとして現れる u としては、曲名、アーティスト名、単語の 3 種類のグラフを作ることができる。

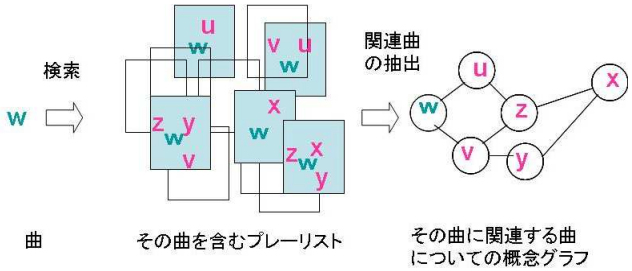

図 3 システム概要

6. 概念グラフの例

アーティストのペア（表 3）の一位には、“Eminem” と “50 Cent” のペアが現れている。単独のアーティストの順位では、Eminem が 4 位、50 Cent が 6 位となっている。単独順位では Eminem よりも下位のアーティストで、Eminem のペアとして上位 20 位に入っているものとして他に、Ludacris、OutKast があるので、Einem のファンにこの二つのグループを推薦することができる。さらに、OutKast より下位のもので OutKast と共起してペア 20 位に入っているもの探すと、”Black Eyed Peas” があること分かる。このような繋がりをたどることにより、Eminem と共起の度合いは高くないもので推薦候補を探すことができる。しかし、このように表を順番にたどるのは人間にとって容易なことではないし、このように推薦されても直感的には理解しずらい。

図 4 は、キーワードとして Eminem として与え、関連するアーティストで共起頻度が 20 回以上のものについて概念グラフを描いたものである。Eminem の一つ下に現れるアーティストとして、“Jay-Z & Linkin Park”、“50 Cent”、“Eminem & Dido”、“Snoop Dogg & Pharrell Williams” の 4 つのアーティストがあることが分かる。“Snoop Dogg & Pharrell Williams (22)” のようにアーティストの横にある数値は、Eminem とこのグループが共起しているプレーリストが 22 件あることを示す。左側にあるものの方が共起の度合いが高い。例えば、“Eminem(1156)” と “Snoop Dogg & PharrellWilliams (22)” を結ぶ枝は、Snoop Dogg & PharrellWilliams を聞いている人の過半数が、Eminem を聞いていることを表す。なお、線で直接繋がれているものは、上に定義した上位下位の関係で、1 段階の関係になっているものを表す。すなわち、二つの点 P と Q が枝で直接繋がっているときには、上位下位の関係でその間にくるような点はないことを意味する。

このように、図 4 は単独のランキング（表 2）やペアのランキング（表 3）からは理解することができないアーティスト間の関連を可視化している。さらに、グラフの配置については、左側にあるものの方が人気が高いという意味を持っている。このように、検索結果の表示方法としてランキングやクラスタリングで問題となっていた事柄が解決できている。

7. まとめと今後の課題

プレーリスト群に現れる曲やアーティストの共起頻度を用いて上位下位関係の定式化を与え、曲をノードとする概念グラフやアーティストをノードとする概念グラフを構成する方法を提案した。iTMS からダウンロードした一万件以上のプレーリストについて、検索と分析を行なうシステムを実装し、単純なランキングや共起頻度のランキングでは得られない関連情報を発見できることを示した。

単語や概念の関連を分析する理論としては、概念束 [1),3)] やグラフィカル・モデル [7),10)] が知られている。本稿で提案する概念グラフの構成法とこれらの関連を明らかにすることは今後の課題である。

本研究の一部は、平成 17 年度科学研究費 16650030、16016267 による。検索エンジン GETA の利用を認めていただいた高野先生を始めとする GETA 開発グループの方々に感謝します。

参 考 文 献

1) Claudio Carpineto, Giovanni Romano, Concept Data Analysis: Theory and Applications John Wiley & Sons, 2004
2) An Approach to Analyzing Correlation between Songs/Artists Using iTMS Playlists, Yufeng Dou, Eisuke Itoh, Sachio Hirokawa, Daisuke Ikeda, Proceeding of the International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC'2005), to appear.
3) Bernhard Ganter, Rudolf Wille, C. Franzke, Formal Concept Analysis : Mathematical Foundations, Springer-Verlag, 1999

4) 汎用連想計算エンジン (GETA), http://geta.ex.nii.ac.jp
5) Ron Goodman, Howard N. Egan, Automatic hierarchical categorization of music by metadata 米国特許 6928433, 2005
6) 廣川佐千男, 下司義寛, 和多太樹, 文書群からの概念グラフの構成, 情報処理学会研究会報告 2005-NL-169, pp.79–84, 2005
7) 宮川雅己, グラフィカル・モデリング, 朝倉書店, 1997
8) P.Resnick, N.Iacovou, M.Suchak, P.Bergstrom, J.Riedl, GroupLens: Open Architecture for Collaborative Filtering of Netnews, In Conference on Computer Supported Cooperative Work, pp. 175–186, 1994
9) Geoffrey R. Stanfield, Eric Bassman, System and methods for training a trainee to classify fundamental properties of media entities 米国特許 6913466,2005
10) Rohini Srihari, Miguel E. Ruiz, Munirathnam Srikanth, Concept Chain Graphs: A Hybrid IR Framework for Biomedical Text Mining, Proceedings of the SIGIR 2003 Workshop on Text Analysis and Search for Bioinformatics, 2003
11) 下司義寛, 和多太樹, 安元裕司, 関隆宏, 廣川佐千男, 文書群の局所性と大域性の差を利用したキーワード粒度評価, 情報処理学会研究会報告 2005-NL-168, pp.7–12, 2005

表 4 上位 20 組（曲）

順位	頻度	曲のペア	
1	181	This Love	She Will Be Loved
2	165	Ocean Avenue	The Reason
3	157	This Love	Harder to Breathe
4	156	This Love	The Reason
5	147	American Idiot	Holiday / Boulevard of Broken Dreams
6	126	This Love	Ocean Avenue
7	121	Boulevard of Broken Dreams	American Idiot
8	115	Let's Get It Started (Spike Mix) [Bonus Track]	Hey Mama
9	113	Boulevard of Broken Dreams	One, Two Step
10	106	One, Two Step	Since U Been Gone
11	106	She Will Be Loved	Harder to Breathe
12	101	One, Two Step	Soul
13	101	The Reason	She Will Be Loved
14	101	One, Two Step	My Boo (Bonus Track)
15	100	Holiday / Boulevard of Broken Dreams	Give Me Novacaine / She's a Rebel
16	100	Mr. Brightside	Somebody Told Me
17	96	Like Toy Soldiers	Mockingbird
18	95	Ocean Avenue	She Will Be Loved
19	94	Holiday / Boulevard of Broken Dreams	Wake Me Up When September Ends
20	92	Ocean Avenue	Way Away

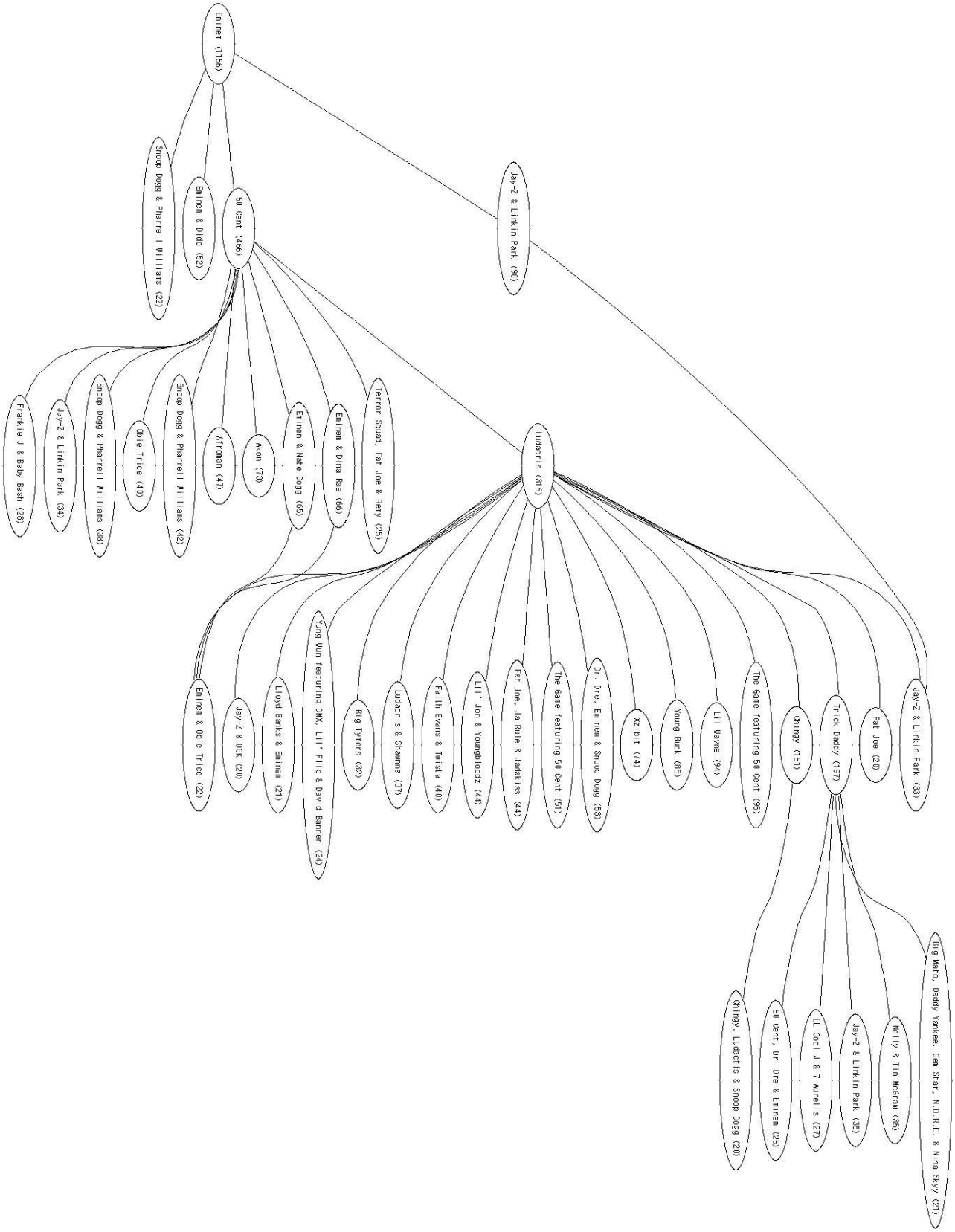

図 4 “Eminem” 関連アーティスト 概念グラフ