Vol.2013-MUS-98 No.9
Vol.2013-EC-27 No.9
2013/3/15

Web 中の文章とリスト構造体を用いたバンドメンバー自動収集手法に関する研究

澤田真吾[†1]　浜中雅俊[†1†2]

本論文ではWeb中にある文章とリスト構造体から音楽のバンドやグループに関連するメンバーを自動的に収集する手法を述べる．音楽のバンドやグループには，ボーカルやギタリスト，ベーシストなど，多くのミュージシャンがメンバーとして参加している．これらの情報を収集する手段として，従来提案されていた手法では特定のパターンとWebテキストとのパターンマッチングによりバンドのメンバーを抽出していた．しかし特定のパターンを用いた抽出では汎用性に乏しく精度が不十分であった．本研究ではN-gramモデルを用いたメンバー抽出とWeb上に存在する「メンバーリスト」を利用した手法を提案し，精度の向上を目指す．我々はまず，Web上でメンバー情報が主に記載されている文章とリスト構造体を取得する．次に，N-gramモデルを用いて文章からメンバー抽出をおこなう．さらにその結果を用いてメンバーリストを判定し，リスト中のメンバーを抽出する．文章からの抽出とリストからの抽出で生じる誤情報のパターンには違いがあるため，それらの共通データのみを利用することで，正しいメンバーだけを収集することができる．我々は提案手法を用いてメンバー収集の正確性と汎用性を評価した．その結果提案手法は従来手法より精度の高い収集が可能であることを示した．

Extraction of Band Members from Sentence and List Structure on the Web

SHINGO SAWADA[†1]　MASATOSHI HAMANAKA[†1†2]

In this paper, we propose a method of automatic collecting members that belongs to music band or group. Music band or group contains various kinds of members, vocalist, guitarist and bassist. As a means to gather this information, conventional method has been extracted the band members by the pattern matching between specific patterns and Web texts. But conventional method was poor in versatility and accuracy. We propose the method of extraction band members using N-gram model and member list on web, to improve the accuracy. First, we obtain sentences and list structure from web texts. Next, we extract band members from sentences using the N-gram model. Determining the member lists using the result, we extracted the members in the lists. We can extract the correct members by using the intersection of the data obtained from sentences and lists, because their incorrect information is difference each other. We evaluated the accuracy and versatility of the collection member by using the proposed method. As a result, he proposed method was able to collect accurate than the conventional methods.

1. はじめに

本論文では，音楽のバンドやグループに関連するメンバーを収集するため，Web中にある文章とリスト構造から自動的に情報抽出をする手法を述べる．音楽のバンドやグループには,「メンバー」として多くのミュージシャンが関わっている．ボーカリストやギタリスト，ベーシスト，ドラマーなど，様々なミュージシャンによってバンドが構成されている．そして，どんなボーカルが歌っているか，どんなギタリストが演奏しているかという情報は，エンドユーザが自分の好みに合ったバンドを選択する際の重要な要素である．

しかし，バンドの現メンバーや過去のメンバーなどのすべてのメンバーを集約したデータが記載されたデータベースやWebサイトは存在しないため，バンドのメンバー情報を得るのは容易ではない．従来，Wikipediaから関連人物を抽出した手法もあったが[1]，Wikipediaなどの Webサービスでは，有名なバンドやミュージシャンに関する情報は詳細に記載されているというメリットがある反面，知名度の低いバンドに関しては，極端に情報量が少ない場合や，ページそのものが存在しないことが多いため，情報を収集できなかった．検索エンジンにより取得した複数のページを利用した手法[2]では，メンバー名と共に出現しやすい単語列と Web テキストとのパターンマッチングによりメンバー収集を行うことで，知名度の低いバンドのメンバー収集を可能にした．しかし，収集後の閾値処理によって多くの正解データが失われてしまっていたため，精度が不十分であった．閾値処理の問題に対応した手法[3]では，パターンマッチングにより得られたメンバーの中で特に信頼性の高いデータを選択し，そのデータと同じパターンで出現する

†1 筑波大学大学院システム情報工学研究科
University of Tsukuba, Graduate School of System and Information Engineering
†2 JST さきがけ
PRESTO JST

Vol.2013-MUS-98 No.9
Vol.2013-EC-27 No.9
2013/3/15

メンバーを再取得することで，閾値処理によって削除された正解データの再取得を可能にした．これにより，バンドに一時的に参加したメンバーや，演奏支援を行ったミュージシャンなどの，バンドと関連性の低いメンバーの抽出を実現した．しかし，特定のパターンを用いたマッチングによる抽出では得られる情報量が少なく，さらに Web 上でパターンの共起が起こる頻度が低かったため，大幅な精度向上には至らなかった．

そこで我々は，メンバー抽出に適切なパターンを発見するために N-gram を用いる．そうして得られたパターンとのマッチングにより Web テキスト上からメンバーを抽出し，さらに閾値処理によって削除されたメンバーを，Web 中のリストを用いて再取得する手法を提案する．本手法でメンバーを抽出するパターンを N-gram モデルを用いて生成することで，Web 上のメンバー情報の様々な出現パターンに対応した抽出を可能とする．また閾値処理により削除されたデータを，メンバーリストとのマッチングにより再取得することで，削除されたデータの中から，正しいメンバーのデータのみを抽出することが可能となる．

我々は提案手法が正しいメンバーの収集が可能であるかという点と，すべてのメンバーを収集可能であるかという点を調べるため，メンバーの収集実験を行った．その結果，提案手法はメンバーリストとのマッチングを行わなかった場合と比べて，高い精度でのメンバー収集が可能であることを示した．

2. メンバーの自動収集

Web 上に存在するメンバー情報は，主に「文章」と「リスト構造体」に存在する．その出現例を図 1 に示す．

文章中に出現するメンバー名は，その周囲に単語列が存在するため，確率的手法による抽出が可能である．しかし，その出現パターンの類似性から，誤った情報としてバンドのアルバム名やライブツアー名を少量抽出してしまい，抽出回数の少なかったメンバーと紛れてしまう．

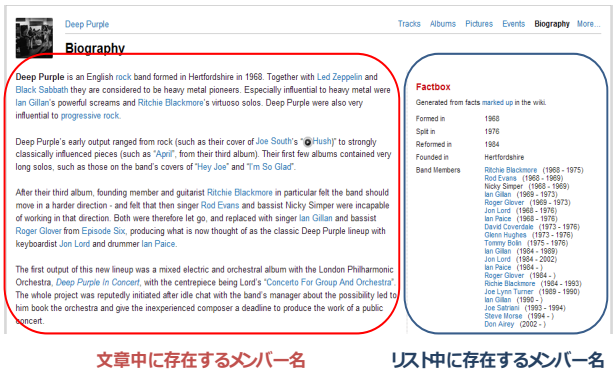

図 1. Web 上に存在する文章とリストの例(文献 11 より引用)

リストを用いた抽出では，Web 上に存在するリスト構造体から，バンドのメンバーリストである可能性の高いものを抽出し，リスト中のデータをすべて取得することで多くのメンバーの収集が可能である．しかし，Web 上に大量に存在するリスト構造体からバンドのメンバーリストのみを抽出することは非常に困難であり，類似バンド・ミュージシャンリスト(図 2)などのメンバーリストでないリストから，バンド名や他のバンドのメンバーなどの誤った情報を多く収集してしまう．

我々は文章とリストからの抽出で得られる誤情報のパターンの違いに着目し，二つの結果の共通項のみを利用することで正解データのみの抽出が出来ると考えた．その手法の全体図を図 3 に示す．

はじめに Web ページの収集をし，情報源を取得する．出来るだけ対象とするバンドの情報のみが記載されていて，さらにメンバーの情報を含んでいる可能性の高いページを収集する(2.1 節，図 3. a).

次に，得られたページから文章とリスト構造体を抽出する．それぞれの HTML 文中での出現の特性を利用し，正規表現により抽出する(2.2 節，図 3. b).

HTML 文より抽出した文章から，N-gram モデルを用いて作成したパターンとのマッチングによりメンバーを抽出する．このとき抽出数が多いデータは信頼性が高く，抽出数の少ないデータは誤ったデータが多く含まれている．そこで閾値 N を設け，抽出数が N 以上のデータを正解のメンバー，N に満たないものをリザーブメンバーとして一時保存する(2.4 節，図 3. c).

HTML 文より抽出したリスト構造体からメンバーを抽出する．Web ページ上にはリスト構造体が大量に存在するため，どのリストがメンバー情報を表したリストであるかを選定しなければならない．そこで先ほど文章中から抽出した正解のメンバーが含まれているリストをメンバーリストであると判断し，そのリスト中の人名をすべて抽出する

図 2. 誤って抽出しやすいリスト例(バンドに類似するバンドやミュージシャンのリスト)

(2.5 節，図 3. d).

最後にリストから得たメンバーと，文章から抽出したメンバーの中でリザーブメンバーとなったデータとのマッチングをおこなう．この作業によりお互いのデータの正しい情報のみを利用することが可能となる(図 3. e).

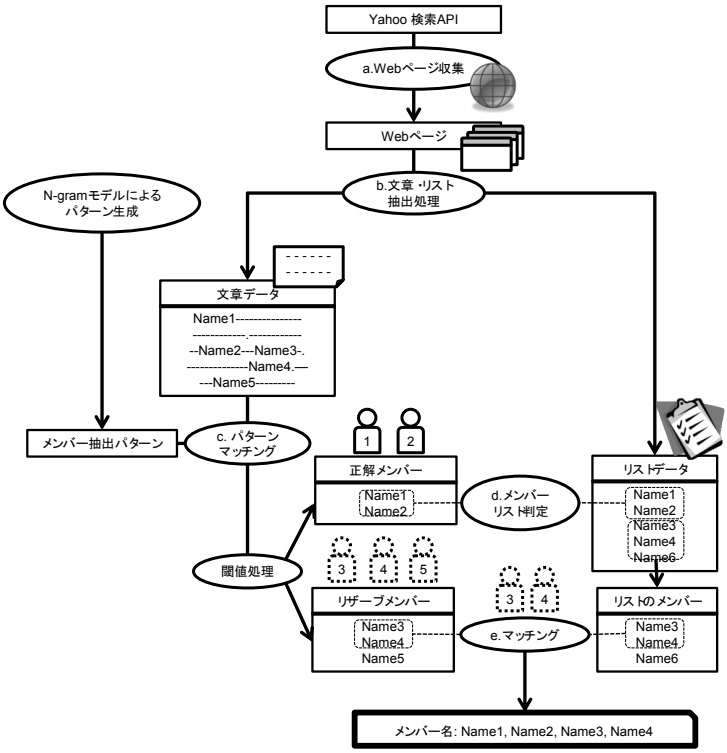

図 3. メンバー収集の全体の流れ

2.1 Web ページの収集

Web ページの収集には Yahoo の検索 API を用いる. 検索 API に適切なクエリを渡すことで，よりバンドのメンバー情報を含む確率の高い Web ページを取得できる. 従来手法[3]では, バンド名に以下の 4 種類の検索ワードを加えたクエリで検索を行なった．実験の結果,”バンド名” band members のクエリで検索した結果が最も精度が高かったと記されている.

“バンド名” music
“バンド名” music members
“バンド名” band members
“バンド名” band lineup

本研究でもこの結果に従い，クエリは ”バンド名” band members を採用する.

また，検索 API を用いて得た結果には，クエリとの関係性の強さに応じて，ランキングが付けられている．つまりランキングの高いページは目的のバンドのメンバー情報が記載されている可能性が高いが，ランキングの低いページは，メンバー情報が載っていない場合や，目的のバンドとは違うバンドの情報が書かれているということもある．本手法では，誤った情報の抽出を避けるためランキング上位 50 までの Web ページを利用する.

2.2 文章とリスト構造体の抽出

文章とリスト構造体は, それぞれの HTML 文中での出現の特徴から，正規表現を用いて抽出する.

文章は基本的にアルファベットの大文字から始まり，ピリオド ”.” で終わる．また，必ず二つ以上の単語が連続して存在していなければ文章は成立しない．これらの仮定から正規表現を生成し，Web ページ中の文章を抽出する.

リスト構造体の抽出は HTML タグを利用する. Web ページ上のリストは, HTML 文中では項目ごとに, リストタグ ”<li>” によって囲まれ，それらが連続して表記されることで構成されている(図 3). この性質を利用し，リストタグが連続して記載されている部分を正規表現により抽出する.

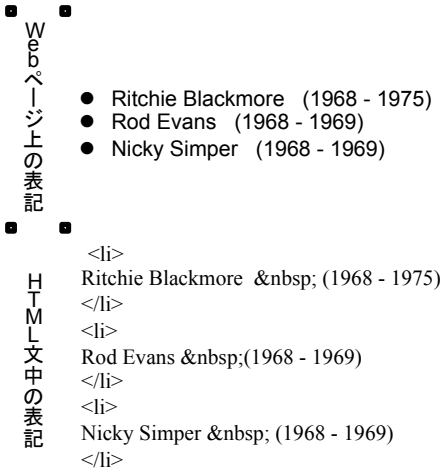

図 4. HTML 文中のリスト構造体の記述例

2.3 メンバー抽出パターン生成

N-gram を用いて, バンドのメンバー名の周囲に出現しやすい単語を求め，抽出モデルを作成する．本来，N-gram は目的の語句の直前に出現する単語列を求めるのが一般的だが，バンドのメンバー名は文頭に来ることが多く，直前に単語列が存在しない場合も多いため，本研究では，メンバー名の直前と直後に出現する確率の高い単語列をそれぞれ求め，抽出モデルを作成する.

モデル作成に用いるバンドのメンバーは，音楽情報データベース Musicbrainz[13]に存在するバンド 50 組と, それらに演奏参加したことのあるメンバーとする．はじめにメンバー情報の記載されている Web ページを取得するため「"バンド名” band members」のクエリで Yahoo API で検索をかけ，上位 50 ページを取得する．次に，取得したページからピリオドで終わる文章をすべて抽出する．文章中にメンバー名が存在する場合, メンバー名(MemberName)と直前の

Vol.2013-MUS-98 No.9
Vol.2013-EC-27 No.9
2013/3/15

2 単語($w_0\ w_1$)との Trigram と，メンバー名と直後の 2 単語($w_2\ w_3$)との Trigram をすべて抽出し，それぞれの抽出数

$$C(w_0\ w_1\ MemberName)$$
$$C(MemberName\ w_2\ w_3)$$

を取得する．次に抽出した単語列がすべての文章中で何回出現するのかをカウントし，それぞれの抽出数

$$C(w_0\ w_1)$$
$$C(w_2\ w_3)$$

を取得する．最後にそれらの単語列がメンバー名の直前，直後に出現する条件付き確立を式 1 , 式 2 により計算する．

$$P_{\mathrm{pre}}(MemberName|w_0\ w_1) = \frac{C(w_0\ w_1\ MemberName)}{C(w_0\ w_1)} \quad (1)$$

$$P_{\mathrm{suf}}(MemberName|w_2\ w_3) = \frac{C(MemberName\ w_2\ w_3)}{C(w_2\ w_3)} \quad (2)$$

2.4 文章からのメンバー抽出

作成したモデルを用いて，HTML 文中の文章とのパターンマッチングによりメンバーを抽出する．このとき，正しいメンバーである可能性の高いデータの抽出を目指すため，2.3 で取得した単語列のうち，メンバー名の前に出現する条件付き確立が 70%以上の単語列を用いる．また，抽出する文字列は人名であることを考慮して，以下の条件を付加する

- 先頭文字は大文字
- 2 つ以上 5 つ以下の連続した単語列

次に抽出したメンバーをソートする．例えば，正解のメンバーが「Freddie Mercury」であるとき，誤って「Freddie Mercury Live」など，名前に余計な単語が付加されたものを抽出してしまうことがある．この場合，抽出した単語列同士を比較し，ある単語列が，別の単語列に含まれている場合，2 つの単語列の抽出数を比べ，抽出数の低いものを削除し，その抽出数を残った単語列の抽出数に加算する．

最後に抽出したメンバーを，抽出数によりフィルタリングし，正解メンバーとリザーブメンバーに分類する．正解とする抽出数 N の適切な値は 4 章の実験で求める．

2.5 リストからのメンバー抽出

リストからメンバーを抽出するためには，はじめに Web 上に大量に存在するリスト構造体からメンバー情報を表すリストを選び抜かなければならない．図 5 は Web 上に存在するメンバーリストの例である．

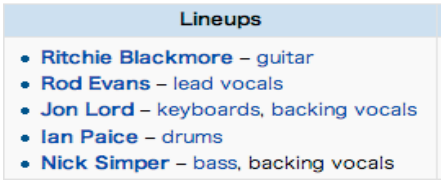

図 5. Web 上に存在するメンバーリストの例
(文献[12]より引用)

このようにメンバーリストはいくつかのメンバーが箇条書きのような形で表記されている．つまり，一人でも正しいメンバーが分かっていれば，そのメンバーを含むリストはメンバーリストである可能性が高い．そこで文章中から抽出したメンバーの中で，信頼性が高いと判断された正解のメンバーがリスト中に含まれていれば，そのリストをメンバーリストと判断し，リスト中に記載されている人名をすべて抽出する．このとき抽出する人名は 3.4 の場合と同じく，先頭文字は大文字，2 つ以上 5 つ以下の連続した単語列であることをを条件とする．

3. BandNavi HD

提案手法により収集したバンド・グループのメンバー情報を用いてミュージシャンを探索するアプリケーション”BandNavi HD”を構築した．BandNavi HD はミュージシャンのつながりを利用して, 新たなバンド, ミュージシャン, 楽曲を発見できることが出来る iPhone アプリ「BandNavi」[5]を iPad 上で実装し, さらにミュージシャンの関係性を可視化する機能を加えたアプリケーションである．バンドに在籍するミュージシャンの中には，複数のバンドを掛け持ちしていたり，サポートメンバーとして他のバンドのアルバムに演奏参加していたり，作曲家としての楽曲提供やライブでのゲスト出演などの活動により，一人の人物がたくさんのバンドと関係を持っていることがある．このような

図 6. BandNavi HD 画面図

関係性を持つバンドとミュージシャンを繋いでいくと，ネットワークが形成される．BandNavi HD はこのバンドとミュージシャンのネットワークを次々と辿っていくことで，お気に入りのボーカルやギタリストが参加している他のバンドの楽曲を発見することができる新しい楽曲探索インタフェースである．

4. 評価実験

提案した手法が，バンドの正確なメンバーを収集できるかを調べるため，メンバーの収集実験を行った． 実験に用いたバンドは，大型音楽情報データベース MusicBrainz に存在するバンド 50 組を無作為に選択しそれらのバンドのメンバーの収集を試みた(4.1)．提案手法により収集した結果と従来手法との結果を比較した(4.3)．またリザーブメンバーとのマッチングを行った提案手法と，マッチングを行わず，リストのメンバーをすべて正解のメンバーとした結果，リストを用いず N-gram モデルによる抽出のみの結果とを比較し，提案手法の有効性を評価した(4.4)．

4.1 正解データ

正解データであるバンドのメンバーはバンドの情報が記載されている Web サイトから手動で収集した．メンバーは，現在所属しているメンバーや過去に所属していたメンバーだけでなく，ライブやツアーなどに参加したメンバーや，リリースしたアルバムでの演奏にゲストとして参加したメンバーも正解のメンバーとした．

4.2 文章からの抽出における閾値の設定

正しいメンバーリストを抽出するためには，閾値処理によって正解メンバーとなったデータが十分に信頼性のあるデータでなければならない．表 1 は抽出数 N の閾値を変化させたときのそれぞれの結果である．

本手法では 9 割以上の適合率を示した N=6 以上の閾値処理を採用する．また，最大抽出数が 6 に満たなかった場合は，最も大きい抽出数を 1 とした場合のそれぞれの抽出数の割合により閾値を設ける．こちらも，9 割以上の正解率を示した，閾値 0.8 を採用する．

表 1. N-gram での抽出結果

抽出数Nの閾値	適合率	再現率	F値
N=1以上を正解	0.15	0.74	0.25
N=4以上を正解	0.83	0.55	0.66
N=5以上を正解	0.89	0.50	0.64
N=6以上を正解	0.91	0.47	0.62

4.3 従来手法との比較

提案手法を用いてメンバー収集をした結果と，パターンマッチングとパターンの共起によりメンバー収集を行った従来手法[3]による結果を表 2 に示す．表より，提案手法により収集されたメンバーは，従来手法より適合率，再現率ともに高いという結果を得た．これは，提案手法は従来手法より，メンバーの収集精度が高ことを示す．適合率が向上した要因は，従来手法が特定のパターンのマッチングによる抽出であったのに対し，提案手法は N-gram モデルを用いたため，メンバーである確率の高いデータのみを抽出できたためである．再現率が向上した要因は，従来は閾値削除されてしまった正解データを，信頼性の高いデータとのパターンの共起を用いて再取得していたのに対し，提案手法ではリスト上での共起により再取得を行ったためと考えられる．これは，メンバー同士のパターンが共起する頻度より，リスト内での共起の頻度の方が多いため，正解データの再取得に適していたためである．

表 2. 提案手法と従来手法の収集結果

メンバー収集手法	適合率	再現率	F値
提案手法	0.87	0.67	0.75
従来手法	0.68	0.62	0.65

4.4 リストの利用方法の比較

表 3 はザーブメンバーとのマッチングを行った提案手法による結果と，リストのメンバーをすべて正解メンバーとした結果，リストを利用しなかった場合の結果を示す．表より，提案手法は N-gram のみでの収集に比べ，適合率を大きく下げることなく，再現率を大幅に上げることができるという結果を得た．これはリストを用いることで，正確なメンバーのみをリザーブメンバーから抜き出すことができたためである．リザーブメンバーとのマッチングをせずに，リストのメンバーをそのまま正解とした場合は，最も多くの正解のメンバーを抽出することができるが，適合率が著しく低くなるという結果となった．これは対象のバンドのメンバーリストだけでなく，他のバンドのメンバーリストやその他のリスト構造体が混在してしまったため，誤ったデータを多く抽出してしまったことが原因である．これより，リスト上の正しいメンバーのみを抽出するために，リザーブメンバーを利用することは有効であった．

表 3. リストの利用方法の結果比較

メンバー収集手法	適合率	再現率	F値
リザーブメンバーとのマッチングあり（提案手法）	0.85	0.67	0.75
リザーブメンバーとのマッチングなし	0.66	0.69	0.68
N-gramのみ	0.91	0.47	0.62

5. まとめ

本論文では，Web 上の文章とリスト構造体を用いることでバンドのメンバーを収集する手法を述べた．

本論文の意義は次の二つである．第一の意義は，Web 上の文章中に存在するメンバー情報は，N-gram モデルを用いたパターンマッチングにより抽出可能であることを明らかにした点である．従来の手法は特定のパターンのみを利用したメンバー抽出であったが，提案手法は N-gram モデルを用いてパターンを生成した．そのため Web 中での様々なメンバー名の出現パターンに対応し，多くのメンバーの収集が可能となった．

第二は，閾値処理によって失われた正解データは，Web 上のリストを用いることで正しいデータの再抽出が可能であることを明らかにした点である．閾値によって削除された誤った情報は，バンドのアルバム名やライブ名であった．リスト中の誤った情報は，他のバンド名やそれに在籍するメンバー名であった．これら二つの誤情報のパターンの違いを利用し，互いの共通部分のみを利用することで正解データの再抽出が可能であることを確認した．

今後，英文で記載された情報抽出のみでなく，日本語で書かれた情報抽出を行うことで，J-POP や J-ROCK などのバンドにどのようなミュージシャンが関連しているかという情報も明らかにしていきたい．

参考文献

1) Yulan Yan, Yutaka Matsuo and Mitsuru Ishizuka: Unsupervised Relation Extraction by Mining Wikipedia Texts with Support from Web Corpus, Proc. the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing 2009 (2009).
2) M.Schedl and G.Widmer, ”Automatically Detecting Members and Instrumentation of Music Bands via Web Content Mining,” Proceedings of the 5th Workshop on Adaptive Multimedia Retrieval (2007).
3) 吉谷幹人，宇佐美敦志，浜中雅俊: “メンバー情報に基づくバンドネットワークの構築と利用”，情報処理学会 音楽情報科学研究会 研究報告 2009-MUS-82-5 (2009).
4) Xiaoxin Yin, Wenzhao Tan, Xiao Li, Yi-Chin Tu “Automatic extraction of clickable structured web contents for name entity queries”, WWW 2010, 991-1000 (2010).
5) 吉谷幹人，宇佐美敦志，浜中雅俊: “BandNavi: バンドメンバーの変遷情報を辿るアーティスト発見システム”，情報処理学会 音楽情報科学研究会 研究報告 MUS-86-16 (2010).
6) Grosche, P., Mu¨ller, M. and Serr`a, J.: Au- dio Content-Based Music Retrieval, Multimodal Music Processing (Mu¨ller, M., Goto, M. and Schedl, M., eds.), Dagstuhl Follow-Ups, Vol. 3, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, pp. 157–174 (online), DOI: http://dx.doi.org/10.4230/DFU.Vol3.11041.157 (2012).
7) 吉井和佳,後藤真孝:音楽推薦システム,情報処理 (情報 処理学会誌), Vol. 50, No. 8, pp. 751‑755 (2009).
8) Celma, O.: Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space, Springer (2010).
9) 松尾豊，友部博教，橋田浩一，中島秀之，石塚満. Web 上の情報からの人間関係ネットワークの抽出．人工知能学会論文誌, Vol. 20, No. 1E, pp. 46‑56 (2005).
10) 濱崎 雅弘，後藤 真孝: Songrium: 多様な関係性に基づく音楽視聴支援サービス，Vol.2012-MUS-96 No.1 (2012).
11) Last.fm Deep Purple’s Biography. http://www.last.fm/music/Deep+Purple/+wiki?setlang=en (2013).
12) Wikipedia Deep Purple’s Biography.http://en.wikipedia.org/wiki/List_of_Deep_Purple_band_members (2013).
13) MusicBrainz.http://musicbrainz.org/ (2013).