web 上からの段階的なイベント情報特定による バンドメンバーの経歴自動収集

澤田真吾† 吉谷幹人‡ 浜中雅俊††‡

筑波大学理工学群工学システム学類† 筑波大学大学院システム情報工学研究科‡

科学技術振興機構さきがけ††

1. はじめに

本研究では，ミュージシャンの生存期間，在籍バンドとその在籍期間といった，ミュージシャンの活動の経歴情報を Web から自動抽出する手法について述べる．あるミュージシャンについての情報が欲しい時，Web 検索エンジンを用いて情報検索をしても，ミュージシャン単体についての情報が主体となったページが存在することは少ないため，ユーザは多くのページから少しずつ情報を収集しなければならない．本手法により，複数のページから自動的に情報を集めることで，ユーザは自分の好きなミュージシャンの経歴をくまなくたどり，ゲストで参加したバンドなど，今まで発見することが難しかった情報を得ることができる．

Web を利用して，人物の経歴情報を抽出する手法には，いくつか先行研究が存在する．文献[1]では人物名を含む Web ページから，日付表現と人物の行動表現を抽出し，人物に関する年表の自動生成を行った．文献[2]では，職業別人名リストから人物プロフィールを抽出するシステムを提案した． これらの研究では，人物と職業の関係や，人物と出来事との関係を求めていたのに対し，我々の研究では，在籍していた「期間」を求めるという問題を扱う．

本研究では，Web 上の情報から目的のミュージシャンの活動経歴を大局的な期間から，詳細な期間へ絞り込みながら特定することでくわしい活動経歴を抽出する手法を提案する．これにより，ミュージシャンについて書かれている Web ページ中の様々な日付表現から，適切な経歴の情報を抽出できる．本稿では，本研究が最終的に目指すアルバムのリリース時期や，ライブ日時などの詳細な経歴の情報取得の基本となる，ミュージシャンの生存期間とバンドの在籍期間の抽出方法を示す．

Automatic Collection of Band Member's Career by Stepwise Specifying event from World Wide Web

† ShingoSAWADA

‡ MikitoYOSHIYA, Masatoshi HAMANAKA

† College of Engineering Systems,University of Tsukuba

‡ Graduate School of Systems and Information Engineering,University of Tsukuba

†† PREST JST

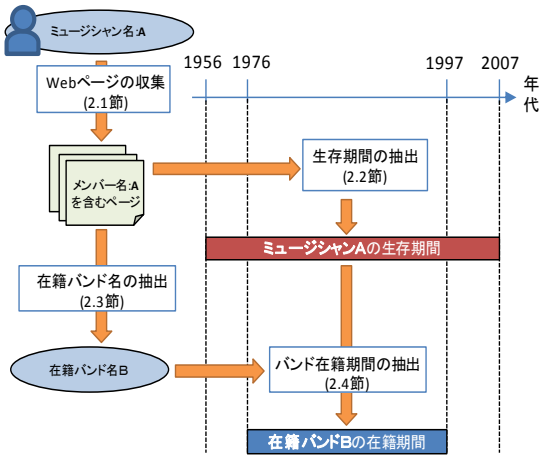

図 1. 処理の流れ

2. 経歴情報抽出範囲の段階的特定

経歴情報を抽出する処理の流れを**図 1** に示す．人物の経歴情報は人物の生存期間にしか存在しないため，はじめにミュージシャンの生存期間を特定する(2.2 節)．次に，ミュージシャンが在籍していたバンドの情報を抽出するため，在籍していたバンド名を特定し(2.3 節)，最後にバンドの在籍期間を特定する(2.4 節)．

2.1. Web ページの収集

本研究では抽出する情報源を Web ページの HTML 文中とする．Web ページの収集には，Google 検索 API を利用し，「“ミュージシャン名” + band」というクエリで検索し，ミュージシャン 1 人につき 50 ページを取得する．

2.2. 生存期間の抽出

メンバーの生存期間の抽出には，正規表現を用いて出現パターンを記述することで抽出をおこなう．調査の結果，頻度が多かった誕生年・死亡年の**表 1** のパターンを用いる．表中で「…」で表わした部分は任意の単語列である．

表 1. 誕生年・死亡年の出現パターン

抽出対象	出現パターン
誕生年	…名前…born…誕生年…
死亡年	…名前…died…死亡年… …名前…passed away…死亡年…

2.3. 在籍バンド名の抽出

バンドの在籍期間を求めるためのキーワードとして，在籍していたバンド名を求める．バンド名の抽出には N-gram モデルを用いた抽出をおこなう．

取得した Web ページから既知のバンド名の直前に来る 2 単語を全て抽出する．その後，2 単語が文章中でバンド名の直前に出現する頻度 C(2 単語 バンド名)と，2 単語が文章中に出現する頻度 C(2 単語)をそれぞれ求め，次式により 2 単語の直後にバンド名が出現する条件付き確率 P(バンド名|2 単語)を求める．

$$P(\text{バンド名}|2\text{ 単語}) = \frac{C(2\text{ 単語} + \text{バンド名})}{C(2\text{ 単語})}$$

求めた条件付き確率 P の例を**表 2** に示す．Web 上からメンバー名を含む一文を抜き出す．その中に条件付き確立が一定以上の 2 単語を含んだ場合，2 単語の次の単語をバンド名として抽出する．

表 2. 条件付き確率の例

2単語	P(バンド名|2単語)
rejoined in	0.896
met members	0.833
left back	0.800

2.4. バンド在籍期間の抽出

在籍期間の抽出には加入・脱退を表現する動詞とのパターンマッチングによりおこなう．

取得した Web ページから加入年か脱退年が存在する文中の動詞 V を全て抽出する．それらの動詞 V が文章中で加入年・脱退年と同時に出現する頻度 C(V+加入年) ，C(V+脱退年)と，動詞 V が年号表現と共に出現する頻度を求め，次式より動詞 V と共に出現する年号表現が加入年・脱退年である条件付き確率を求める．

$$P(\text{加入年}|\text{V}) = \frac{C(\text{V} + \text{加入年})}{C(\text{V} + \text{年号})}$$

$$P(\text{脱退年}|\text{V}) = \frac{C(\text{V} + \text{脱退年})}{C(\text{V} + \text{年号})}$$

求めた動詞の出現確率の例を**表 3** に示す．Web 上からメンバー名，在籍バンド名，年号表現が同時に存在する一文を抜き出す．その中に条件付き確率が一定以上の動詞を含んだ場合，文中の年号表現を加入年・脱退年であるとして抽出する．抽出した加入年と脱退年の間の期間をメンバーのバンドの在籍期間とする．

表 3. 動詞の出現確率の例

動詞 V	P(加入年|V)	動詞 V	P(脱退年|V)
replacing	0.667	fired	0.433
joined	0.630	leaving	0.429

3. 評価実験

Web 上の情報を元に，バンド数 24 組，メンバー数 233 人の生存期間，在籍バンド名，バンド加入年とバンド脱退年をそれぞれ手作業で収集し，正解データとした．2 節で示した抽出方法で実験を行い，求めた再現率・適合率を**表 4** に示す．

表 4. 経歴情報の収集精度

抽出対象	再現率	適合率
生存期間	0.930	0.927
バンド名	0.781	0.759
バンド加入年	0.674	0.624
バンド脱退年	0.459	0.512

生存期間の抽出は再現率，適合率共に非常に高い値を示した．しかし，バンド脱退年の精度は低いことがわかった．この原因は，バンド脱退年が存在する文中に，加入年を抽出するための動詞 V が存在することが多いため，脱退年が加入年として抽出されることが多かったためと思われる．これに関しては，動詞の係り受けなどを考慮することで，改善が見込めると考えられる．

4. まとめ

本稿では Web 上からバンドメンバーの情報を抽出することにより，バンドメンバーの生存期間，在籍バンド名，バンドの在籍期間を特定する手法を示した．

今後は取得した期間内での CD のリリースやライブなどのより詳しい経歴情報の抽出を試みる．

参考文献

[1] 木村塁，小山聡，田中克己，“web からの人物事典生成のための経歴情報の自動収集，” 日本データベース学会 Letters Vol. 5, No. 2, pp. 29-32, 2006

[2] 山本あゆみ，佐藤理史，“ワールドワイドウェブからの人物情報の自動収集，” 情報処理学会研究報告， No. 2000-ICS-119， pp. 165-172， 2000