E-14

歌詞を用いた音楽の自動分類
Automatic Classification of Music Information Using Lyrics

杉浦 康友 †
Yasutomo Sugiura

木本 晴夫 ‡
Haruo Kimoto

1. はじめに

現在、インターネットが普及し、日本のインターネット人口は3,263万6千人（2001年）いるといわれている。また、普及に伴ない誰でも簡単にホームページを作成・公開できる環境が整備されてきておりホームページ数も爆発的な増加をしている。そのためユーザがリンクをたどって目的の情報を得ることは非常に難しくなっている。

ユーザが欲しい情報をインターネットで検索しようとするときに誰もが検索エンジンを利用する。検索エンジンにはgoo等のロボット型検索エンジンとYahoo! Japan等のディレクトリ型検索エンジンの2種類がある。しかし、この2つの検索エンジンでは音楽情報などの曲の雰囲気や人の感性で聞くものは知っているアーティスト名や曲名では検索できるが、曲の雰囲気などでは検索できない。その原因として一般的な音楽のカテゴリ体系は「アーティスト名」、「ジャンル」で作成されていることが挙げられる。また近年、あるジャンルに絞って選曲したオムニバスアルバム、コンピレーションアルバムなどが流行しておりあるテーマにそった選曲は社会でも注目を浴びている。そこで、本研究ではWWW音楽情報に的を絞り音楽の歌詞を解析することにより「夏に聞きたい曲」、「卒業で聞きたい曲」などの音楽の雰囲気に着目した独自のカテゴリに音楽を自動分類し、ユーザに今までは知っているアーティストでしか検索できなかったものが自分の好きな曲の雰囲気で検索することができる新しい検索スタイルを提供するとともに試作システムを作成し評価を行い検証する。

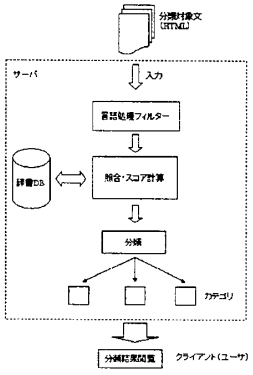

図 1: システム概要図

†東京電機大学大学院
‡名古屋市立大学大学院

2. システム構成

2.1 概要

図1にシステム概要図を示す。まず、インターネットロボットなどで収集した音楽情報のホームページ(歌詞)をシステム入力する。入力されたホームページを言語処理フィルターに通してタグを取り除き、形態素解析を行う。次に辞書データベースと照合して各カテゴリのスコアをそれぞれ計算して、一番得点の高いカテゴリのスコアを持つものを分類するカテゴリとみなしてカテゴリを付与する。以下の節にそれぞれの機能の詳細を説明する。

2.2 言語処理フィルター

言語処理フィルターでは後で文書のスコアを計算する際に不必要なものを取り除く機能や形態素解析を行なう。図2に言語処理フィルターの処理フロー図を示す。

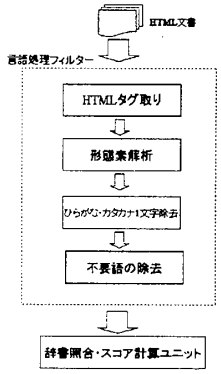

図 2: 言語処理フィルターフロー図

ホームページはHTML（Hyper Text Markup Language）で書かれているため、ブラウザを通さないで読むと「タグ」が入ってしまう為、自然言語処理には適さない。そこでこの「タグ」を除去する。次に形態素解析を行い単語を切り分ける。ひらがな・カタカナの1文字除去とは形態素解析を行い各品詞に分解された時に出る「お」、「ク」などの意味を持たない1文字のひらがな・カタカナを除去する機能である。この1文字を除去することにより後で行なうスコア計算の際に無駄で意味のない単語を計算せずに済む。その結果、分類精度が向上すると考えられる。同様に不要語とは「こと」などのそのものでは意味を持たない語を表し、文章の特徴を表す語とはなりえない。それを除去することにより分類対象文書の特徴を抽出するのに大変効果であり分類精度が向上すると考えることができる。

2.3 カテゴリ体系

その際に使用するカテゴリ体系を図3に示す。このカテゴリは「夏の音楽」→「海」というのは「海で聞きた

い曲」というように表現したものである。

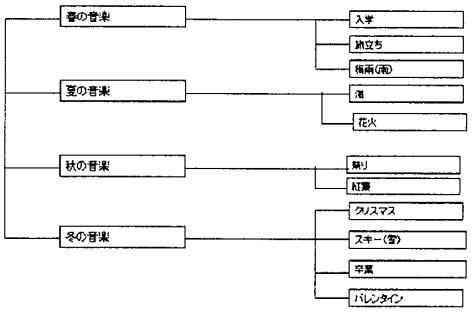

図 3: 分類カテゴリ体系図

辞書は分類を行なうためのスコア計算を行なう際に使用する。作成した辞書のデータを表 1 に示す。

表 1　辞書のデータ

フォーマット	CSV
要素	単語、出現頻度、カテゴリ
登録単語数	5922 語

2.4　辞書データベース照合・得点計算

言語処理フィルターで処理された分類対象文書は辞書データベースとの照合を行う。

はじめに言語処理フィルターで処理した分類対象文書上の単語を抽出しそれを $w_k(k = 1, 2, \cdots, n)$ とする。n は分類対象文書上の単語数を表す。w_k からそれぞれ各単語の出現頻度数を算出し、その出現頻度数をそれぞれ $a_k(k = 1, 2, \cdots, n)$ とする。次に、分類対象文書の単語 w_k を辞書データベースとの照合を行い辞書上の単語の出現頻度数 $b_k(k = 1, 2, \cdots, n)$ を得る。以上から得た a_k、b_k をもとに分類対象文書を各カテゴリに対してそれぞれスコア S の計算行う。

$$S = \sum_{k=1}^{n} b_k \frac{a_k}{N} \tag{1}$$

ここで N はカテゴリ別の辞書に登録されている全単語数を表す。

スコア S が一番高い値になったカテゴリを該当するカテゴリとみなし分類する。

3.　評価実験

本システムを評価する際に人手によって分類されたテキストを正解データとして利用する。今回の実験は「海で聞きたい曲」、「入学で聞きたい曲」、「スキー（雪）で聞きたい曲」「クリスマスで聞きたい曲」、「卒業で聞きたい曲」、他 10 カテゴリのカテゴリを対象としてそれぞれのカテゴリに対して 15 曲の正解データ、合計 146 曲を作成し、それを分類対象文書として入力し、分類を行う。正解データの作成方法は大学生のアンケートにより歌をランキング付けているカレッジチャート [1] を参考として曲を選定し、20 代 5 人の方に協力していただいて作成した。

その結果を情報検索システムの評価によく利用される再現率 *(Recall)*、適合率 *(Precision)* の計算を行い評価を行なう。

再現率は以下の (2) 式で定義する。

$$\text{分類の再現率} = \frac{\text{正しく分類された文書数}}{\text{カテゴリの全正解分類文書数}} \tag{2}$$

この (2) 式は分類結果ににどれだけ「漏れ」がないかを表す。

適合率は以下の (3) 式で定義する。

$$\text{分類の適合率} = \frac{\text{正しく分類された文書数}}{\text{カテゴリに分類された文書数}} \tag{3}$$

この (3) 式は分類結果にどれだけ「ゴミ」がないかを表す。

一般的に再現率と適合率が大きい値ほど性能がよいことになる。

正解データをシステムに入力し再現率・適合率を計算した結果を表 2 に示す。

表 2　再現率・適合率結果

カテゴリ	再現率	適合率
海で聴きたい曲	0.8	0.8
スキー（雪）で聴きたい曲	0.9	0.34
入学で聴きたい曲	0.7	0.7
卒業で聴きたい曲	0.2	0.2
クリスマスで聴きたい曲	0.5	0.35

表 2 からカテゴリによって分散してしまっていてよく分類できるものとできないものが分かれてしまった。原因として「海で聞きたい曲」などは海を表す特徴語（季語）を含んでいるが卒業を表す特徴語はほとんどないため再現率、適合率ともに悪かったと考えられる。

4.　まとめ

好きな音楽を雰囲気で検索する手法を提案し、独自のカテゴリに自動分類を行うシステムを試作し実験を行った。今後は実験で使用した正解データの数を増やし、すべてのカテゴリについて実験を行い検証する予定である。また、システム改善点として以下のものがあげられる。

- そのカテゴリを表す特徴語の抽出方法
- スコアの算出アルゴリズムの検討
- 辞書データベースの強化
- 不要語の再検討
- 曲のテンポなどの利用

参考文献

[1] カレッジチャート http://www.collegechart.co.jp/

[2] 情報検索と言語処理　徳永健伸 著
東京大学出版会