スパイダリングを用いた楽曲検索システムの開発

06H091 藤田康孝

1 はじめに

今日インターネットで様々な Web サイトが見ることが出来る。しかしその中で自分の欲しい情報を入手するのは難しい。自分の欲しい情報が複数の検索サイトに渡ってあった場合、それらをひとつひとつ確認していき、データの集約を手作業でするのは非常に手間であり、時間の無駄である。今回はその中でも楽曲データに着目してみた。楽曲データを取り扱っている Web サイトは数多く存在するが、必ずしも全てのデータが揃っているとは限らない。よって多くの Web サイトを見るのは時間がかかるので無駄である。そこで様々な Web サイトから情報を集約し、加工する技術であるスパイダリングを用いる。スパイダリングを用いて楽曲データを集約し、ユーザが求める情報を表示することにより時間の短縮が可能である。そこで本研究では楽曲データを集約し、表示できるシステムを開発した。

2 目的

様々な Web サイトにある楽曲データを集約し、表示することの出来るシステムの開発を目的とする。また、1 度の検索で多くの情報を入手することにより、検索回数や Web ページを表示する時間を短縮することを目的とする。

3 本システムについて

本システムでは楽曲データを取り扱っている Web サイトから楽曲データを抽出し、結果を表示する。まず各サイトにユーザが指定した検索ワードを渡す。そこでまず検索結果から楽曲データのある URL を検索結果の HTML から抽出する。そこで楽曲データの抽出には正規表現を用いた。正規表現はパターンマッチであり、各サイトにより楽曲データの掲載されている URL の表記に違いがあるのでそれぞれ用意しておく。

次に楽曲データが掲載されている URL を開きその HTML から楽曲データを正規表現で抽出する。抽出したデータはデータベースに入れ、検索結果の全ての URL で同じことを行う。データベースに入れたデータは重複したものがあるので、重複したものは 2 回表示することのないように処理する。データベース内で処理したものを表示をする。表示はアーティスト名、曲名、その曲を収録しているアルバム名、曲番号をそれぞれ HTML 形式で 1 つにまとめて表示する。

4 まとめ

4.1 利点と問題点

今回の研究でスパイダリングを用いた楽曲検索システムを開発した。これにより複数の Web サイトの楽曲データをまとめることによって、実際にページを見る手間と時間の短縮につながった。また PHP での作成なので、インターネット環境があれば使えるという利点が挙げられる。問題点として重複処理の時に空白文字の有無で同じ内容の結果が表示されてしまうことがあった。これは重複処理を完全に一致していないと重複とみなさないことが問題である。

4.2 今後の課題

今後の課題として用いた Web サイトでは曲名検索やアルバムのタイトルなどでも検索できる。今回のシステムでもそれらの検索方法を追加する。また、表示の形式がユーザで指定できないので、表示するものを選択できるような改良を行う。

次に、問題点でもある空白文字の有無で同じ結果が表示されることがあるが、それらにも重複処理ができるようにする。本システムでは指定した Web サイトから情報を入手するので、プログラムに検索対象を追加することができる。しかし追加をすればそのサイトに合った正規表現も追加しないとならない。そこで正規表現も自動で行う改良をする。