音楽サーチエンジンのシソーラスの設計

大田勝覚 、高橋和史(大同工業大学) 、 小川　清 (名古屋市工業研究所)

はじめに

インタネットには、膨大な情報が存在する。マルチメディアとよばれ、音声、音楽、アニメーション、そして、動画など様々な種類の情報が存在する。

その膨大な情報を、効率よく取り出すためのサービスが、検索エンジンである。

検索エンジンに登録されている情報も、膨大な数があり、検索を行っても、膨大な情報が表示される。そのため、必要な情報を、その中から探すのが困難な場合がある。

例えば、集合演算が可能な検索エンジンにおいて、５つのキーワードの共通部分（ＡＮＤ）を取っても、１万件以上の情報が表示される場合がある。

つまり、一般的な検索エンジンで、すべての人に効率的な情報を提供することはできない。そこで、太田・高橋がインタネットを通じて音楽活動を行っていること(http://www.sun-inet.or.jp/~mb/)から、音楽に絞って、検索エンジンで効果的にデータを引き出すためのシステムを検討した。

1　検査エンジン

インタネットの電話帳

この数年間で、インタネットは爆発的な成長を遂げ、現在でもその成長は続いている。インタネット上には様々な種類の情報が発信されている。インタネットに置かれた情報は、誰にも知られることが無い場合もある。その情報のある URL を知る者のみがその情報を閲覧することが出来る場合もある。

そこで検索エンジンは、WWW（World Wide Web）の「リンク」を利用し、電話帳で電話番号を探すように、知りたい情報のある URL へのリンクをリストアップする。

インタネットは巨大なデータベース

インタネットは巨大なデータベースである。検索エンジンで任意の単語を入力すれば、それに関連する情報が得られる。

また、インタネットで提供されているデータの多くは無償で公開されていることもあり、これからのデータベース検索技術の研究は、インタネットの検索エンジンをベースに進められる。

検索エンジンの仕組み

一般的に検索エンジンは、ユーザーによって登録されるか、または、検索ロボットが集めてきたＷｅｂサイトの情報をキーワードなどによって分類し、クライアントの要求した条件から、目的の情報のあるＵＲＬをリストアップするサービスのことを指す。

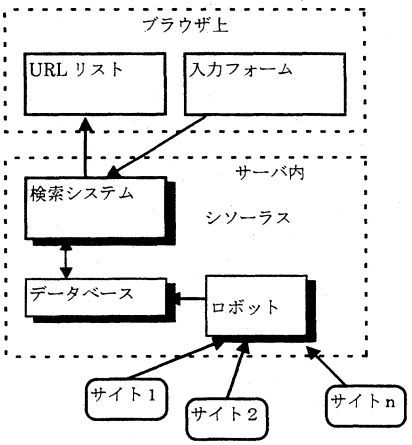

Fig. 3.1 検索エンジンの構成

現在主に使用されている検索エンジンのサービスは、クライアントが入力したキーワードは、サーバにある検索システムに渡される。検索システムはデータベースの中からキーワードを探し、そのキーワードに関連づけられている URL を HTML で出力する。

検索ロボット

検索ロボットは、Web 上のハイパーテキスト構造を自動的にたどるプログラムである。ある一つの html ファイルを取得し、 そのページの中から参照されているすべてのページを見に行き、それを再帰的に繰り返す。

シソーラス

シソーラスは、用語を体系的に整理したもので、用語に上位、下位の概念のある物を言う。音楽のように、科学的に厳密な定義のある分野ではないところでは、概念の間には、複雑な関係が存在している。

たとえば、同じような意味の言葉がたくさんあり、なおかつ、少しずつニュアンスが違

いがある。この違いをより鮮明にするため、複数の観点から、それぞれの意味を見直す必要がある。それぞれの観点から見ると、同一として扱っていい場合と、かなり異なる概念であることがわかる場合がある。そのため、より単純な関係として記述しなおすことができれば、シソーラスは簡単になる。

2　音楽検索エンジン

音楽検索エンジンの設計

汎用的な検索エンジンでは、シソーラスの作成、必要なデータの種類などに関する細かい指定を行っていない。そのため、音楽活動の傍ら、音楽用検索エンジンの設計を行うこととした。

インタネット上の音楽データ

ブラウザを通して音楽データを取り扱う場合、音楽データは、文字やグラフィックデータと同じように扱うことが出来る。ただし、音楽データの場合、「すべての音楽データを誰もが聞けるわけではない。」場合がまだある。ただし、最近は PCM 音源さえ搭載していれば、どんなデータでも再生できるようになってきている。

音楽データ検索エンジン

音楽データを扱うためには、音楽用のシソーラスと音楽データを集める為のロボットが必要になる。

ロボットは HTML 内のキーワードだけでなく、リンクされている音楽データ自体のヘッダを読み、タイトルやアーティスト名や著作権情報を取得する。これにより、音楽データの検索は、HTML 上のキーワードからでは無くデータ自体が持っている情報で検索することが可能になる。
また、音楽データのフォーマットによってはヘッダに著作権情報を書き込むことが出来ない場合もある。このような場合は、データと共に著作権情報ファイルを用意することを提案する。

また、音楽データの中でも、.mid（スタンダードミディ）データは、著作権以外に、使用しているトラック数や楽器の番号（プログラムチェンジ番号）などを取得することが出来る。従って検索エンジンは、使用している楽器の名前などからも検索が可能になる。

音楽のシソーラス

音楽のシソーラスは、正確な定義がないため、上下関係が非常に曖昧になっている。そこで、検索される情報と、その情報の分類方法を定めて、情報検索を行えるような、シソーラスを作成する。
第一層には、検索される情報の種類を示す。
第二層には、情報の分類方法を示す。
第三層以降には、情報の分類を示す。
また、関連のある語に対しては、関連語を付加する。

従来のシソーラスには、一つの用語がシソーラスの中で、１箇所にしか登場しないものもあった。ここでは一つの概念を複数の観点に登録するため、１つのキーワードがシソーラスの複数箇所に現れるものになる。

人名とジャンルの関係

しかし、この方法だけでは十分ではないため、人名とジャンル名とをペアで持ち、人名からジャンル名を一覧し、そのジャンル名に存在する人名から、またそのジャンル名一覧を表示することにより、音楽の情報をたどる経路を提供することとした。

人名を取得するためには、著作権を利用することとした。

実際に整理したシソーラス。

- アーティスト
  - 性別と人数による分類
    - 男性一人
      - 人名リスト(ジャンル名リスト)
    - 女性一人
    - 男性グループ
    - 女性グループ
    - 男女グループ
  - 所属プロダクションによる分類
    - アマチュア
  - ＜ここには、実際にはプロダクション名を並べる＞
- ジャンル
  - 時代による分類
    - ポップス(人名リスト)
    - ロック...
      - ハードロック...
      - プログレッシブロック
    - フォーク
    - クラシック
  - 地域による分類
    - ジャズ
    - フュージョン
    - カントリー
    - レゲエ
    - ソウル
    - 演歌
    - 民謡
    - ブルース
- 状況
  - 主目的
    - ゲームミュージック
    - アニメソング
    - ダンスミュージック
      - （ディスコミュージック）

```
        (クラブミュージック)
    映画音楽
楽器情報
    楽器構造
        弦楽器
            擦弦楽器
            撥弦楽器
            打弦楽器
        管楽器
            木管楽器(シングルリード)
            木管楽器(ダブルリード)
            木管楽器(エアリード)
            金管楽器
        打楽器
            太鼓類
            木製のもの
            金属製のもの
            その他のもの
        電気楽器
            電気ピアノ
            エレキギター
        電子楽器
            シンセサイザー
            電子ピアノ
            サンプラー
        制御装置
            人間
            コンピュータ
```

4　音楽用検索ロボット

検索ロボットの動作

検索ロボットの動作を大きく分けると、次のようになる。

1．HTMLファイルを読み出す
2．ハイパーテキスト（HTML）の解析
3．リンクを再起的にたどる

この様にして、検索ロボットはWWW上にある情報を収集していく。

ＨＴＭＬファイルを直接読む

検索ロボットがＷｅｂサイトにアクセスする方法は、「Netscape Navigator」や「Microsoft Internet Exproler」がHTMLにアクセスする方法と基本的に同じで、サーバのport 80 （httpd の設定により変わることがある）に接続することで行う。

Ex.4.1 telnet コマンドで port 80 に接続する。

```
% telnet 192.168.0.2 80<CR>
Trying 192.168.0.2 ...
Connected to 192.168.0.2.
Escape character is '^]'.
GET / HTTP/1.0<CR>
<CR>
HTTP/1.0 200 OK
Date: Wed, 29 Jan 1997 06:32:58 +0000
Content-type: text/html
Server: RoxenChallenger/1.0
Last-Modified: Wed, 13 Nov 1996 23:39:5
4 +0000
Expires: Wed, 29 Jan 1997 06:31:18 +00
00
Content-length: 3615

<HTML>
<HEAD>
   <TITLE>owai's home page</TITLE>
</HEAD>
<BODY TEXT="#FFCCFF" BGCOLOR="#
005500" LINK="#FFFF33" VLINK="#FF0
099">
<FONT SIZE=+3></FONT>
<CENTER><TABLE BORDER=6 CELLP
ADDING=0 >
<TR>
    .
    .
</BODY>
</HTML>
Connection closed by foreign host.
```

上のリストは名古屋市工業研究所のあるサーバに接続した例を示す。この方法で、ブラウザを使用せずにＷｅｂサイトの HTML をのぞくことが可能である。
ここで、Connect した後に送っている文字列

```
GET / HTTP/1.0<CR>
<CR>                    *<CR>は改行コード
```

は、httpd に HTML を送るように要求するコマンド（メソッド）で、この他にも「HEAD」，「POST」，「PUT」，「DELETE」，「LINK」などがある。

ヘッダ

http で情報をやりとりする場合には、ヘッダが最初に送られる。

Ex.4.2 http://owari.nmiri.city.nagoya.jp のヘッダ

```
HTTP/1.0 200 OK
Date: Wed, 29 Jan 1997 06:32:58 +0000
Content-type: text/html
Server: RoxenChallenger/1.0
Last-Modified: Wed, 13 Nov 1996 23:39:5
4 +0000
Expires: Wed, 29 Jan 1997 06:31:18 +00
00
Content-length: 3615
```

ここで重要なのは「Content-type: text/ht

ml」で、これは、この送られてきたデータが何なのかを現している。

サーバが嘘をつく

例えば、「www.sun-inet.or.jp/~mb/mus/tly.ra」のヘッダを見ると、「Content-type: text/plane」となっている。「tly.ra」というのはリアルオーディオのファイルです。これは、サーバ側の設定に拡張子「.ra」が設定されていないために起こる。これを回避するには、拡張子の判別を行う。

Ex.4.3 サーバが嘘をつく例

```
% telnet www.sun-inet.or.jp 80
Trying 202.231.73.4...
Connected to mail.sun-inet.or.jp.
Escape character is '^]'.
HEAD /~mb/mus/tly.ra HTTP/1.0

HTTP/1.0 200 OK
Server: Netscape-Communications/1.1
Date: Monday, 03-Feb-97 04:18:20 GMT
Last-modified: Sunday, 12-Jan-97 09:37:0
6 GMT
Content-length: 57588
Content-type: text/plain

Connection closed by foreign host.
```

Content-typeだけでなく、ファイルの拡張子もみる必要がある。

リンクを再起的にたどる

リンク先が以前に調べたことのあるページかどうかをチェックしないと、無駄な情報がどんどん増えてしまう。特に、今調査中のページにリンクが張ってある場合は注意が必要である。

これらの問題を解決するためには、現在調査中のページのファイル名を知っていなくてはならないが、URLからファイル名を特定できない場合がある。次の3種類の場合、サーバは同じファイルを返す。

http://www.sun-inet.or.jp/~mb
http://www.sun-inet.or.jp/~mb/
http://www.sun-inet.or.jp/~mb/index.html

また、同一サーバで違うドメイン名のばあいや、IPアドレスを複数持ったサーバも注意が必要である。

著作権情報ファイル

本来なら、データ自体に著作権情報を書き込むことがベストなのですが、そのようにできないフォーマットも存在します。そのような場合に、同じファイルネームで拡張子が「.(c)」の著作権情報ファイルを置くことで著作権を主張する事が出来るようにしようとする規格がこの著作権情報ファイルである。

著作権情報ファイルの記述例

```
Title: Song of Seibutubu
Genre: pops
Artist: NEX
Compose: Masayoshi Ohta
Arrenge: Masayoshi Ohta
Copyright: Masayoshi Ohta
Date: 1990/11
```

MIME-Type

audio/basic	au snd
audio/x-aiff	aif aiff aifc
audio/x-wav	wav
audio/realaudio	ra ram
text/html	html htm

著作権情報とジャンルの提案

人名とジャンルとをペアにするシソーラスを成するために、検索ロボットにより情報収集する中で、必要であると考えられるのは、著作者の情報と、曲名、ジャンル名、楽器などの名称が、標準フォーマットで記録されることであり、共通フォーマットを提唱する。

A．3 参考文献

1 書籍

(1) ＨＴＭＬ＆ＣＧＩ入門, エーアイ出版, 笹木 望・藤崎真美・太田昌宏
(2)JAVA使いへの道, ソフトバンク,武田圭史
(3) Ｊａｖａ入門,翔泳社,有我成城.江藤敏寿.佐藤治.白神一久.西村利浩.村上列
(4)Harvest User's Manual Version1.4 patchlevel2, Darren R.Hardry. Michael F.Schwartz; Duane Wessels
(5)Java&JavaScripﾌﾟﾛｸﾞﾗﾐﾝｸﾞ,ｿﾌﾄﾊﾞﾝｸ,田中

2インタネット・リソース

(7) ＪａｖａＦＡＱ日本語版, http: //www.webcity.co.jp/info/andoh/java/javafaq.html
(8)WWW/HTML Memos(in Japanese), http://w3.lab.kdd.co.jp/technotes/WWW/
(9)JASRAC(日本音楽著作権協会),http: //www.jasrac.or.jp/
(10) WWW Robot Q&A (Japanese Version), http://fml.ec.tmit.ac.jp/robofaq-j.html
(11)Java API Documentation 1.0 日本語版, http://www.sun.co.jp/java.jp/docs/japi/
(12) Java 入門 日本語版, http://www.sun.co.jp/java.jp/docs/jtutorial/index.html
(13)勝手にＰｅｒｌリファレンス, http://www.threeweb.ad.jp/~takagaki/perl/
(14)Search Engines in Japan, http://www.ingrid.org/w3conf-bof/search.html
(15) A Standard for Robot Exclusion, http://info.webcrawler.com/mak/projects/robots/