# WWWベース横断検索システムにおける知識作成支援環境

3T－10

柳本　豪一　谷　幹也　市山　俊治
NEC　ヒューマンメディア研究所

## 1. はじめに

インターネット上でデータベースの検索サービスを提供する検索サーバの一例として、図書情報を扱うオンライン図書目録(OPAC:Online Public Access Catalog)の検索サービスがある。このOPACサーバを対象として、統一した検索インターフェイスを提供するWWWベース横断検索システムを開発[1][2]している。このシステムでは、各OPACサーバごとに送信する検索式のテンプレートや検索結果を整形する方法を知識として持つことにより、ユーザに統一したインターフェイスを提供する。新たにOPACサーバを追加するには、登録作業者がテンプレートや整形知識を生成する必要がある。このため、知識生成を自動的に行うツールを提供することで、登録作業者の負担を軽減することができる。

本稿では、WWWにより検索サービスが提供されているOPACサーバを対象に、検索手順のテンプレートであるスクリプトと検索結果の整形を行う知識であるフィルタとを実際に検索を行った検索ログから自動的に生成するスクリプト作成ツールと、フィルタ作成ツールについて説明する。

## 2. 提案方式

スクリプトを作成するためには、ユーザがブラウザから入力する検索語がOPACサーバに送られる文字列へどのように変換されるかを決定する必要がある。この決定は、検索画面のHTMLファイルを解析して送信文字列を生成する方法と検索画面と検索ログを解析する方法がある。HTMLファイルを解析する場合、1）HTML上のテキストを自然言語処理で解析し変数の同定を行う必要があるので、解析のために大規模な知識が必要である。2）ユーザへの説明画面において画像が利用されている場合があり、このときには変数の推定ができない。3）HTMLのヴァージョンのちがいに対応がとりにくいなどの問題がある。そこで、入力画面から入力された文字列とユーザからOPACサーバへ送信される検索ログを利用することで、検索画面のHTMLファイルの解析を行わず、スクリプトを作成する方式を採用した。

また、フィルタを作成するためには、検索結果から文献情報が記述されている部分を見つけ、必要な情報のみを抽出する必要がある。検索結果を解析する手法としては、自然言語処理を利用して、必要情報のみを抽出する方法があるが、検索結果には様々な単語が含まれるため文献情報に含まれた単語がタイトルか著者か出版者であるかを判断するためには大規模な辞書を用意する必要がある。現在公開されている国内約100ヶ所のOPACサーバを調査したところ、タグや改行を用いた文献情報の整形や、区切り文字でタイトル・著者・出版者などを分割する記述形式を、すべてのOPACサーバで採用していることが分かった。このため、区切り文字やタグなどのフォーマット情報を用いることで、自然言語処理を利用せずに精度よく検索結果の抽出が行えると考え、自然言語処理ではなくフォーマット情報を用いた抽出方法を採用した。

スクリプト作成ツールは、検索ログを収集する際に利用する単語を限定することでサーバへ送信される検索式からテンプレートを自動的に生成することを特徴としている。フィルタ作成ツールは、フォーマット情報に着目し文献情報だけを切り出すことと、区切り文字に使用条件を付け、区切り文字の適用を制御することで必要な項目のみを抽出することを特徴としている。

## 3. 環境構成

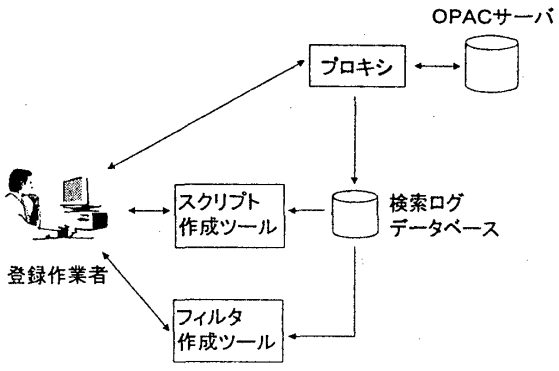


図 1 システム構成

知識作成支援環境は、図 1のような構成をしている。まず登録作業者がプロキシを介して登録するOPACサーバの検索を行う。このとき、プロキシでは、OP

Support system for making knowledge of WWW library system
Hidekazu YANAGIMOTO, Mikiya TANI, Shunji ICHIYAMA
Human Media Research Laboratories, NEC Corporation

ＡＣサーバへ送信される文字列や登録作業者へ送信される文字列を検索ログデータベースにすべて保存する。登録作業者は一連の検索操作を実行し、検索が正常に動作したことを確認した後、スクリプト作成ツールやフィルタ作成ツールを起動する。起動されたツールは、検索ログデータベースに登録されているデータを解析し、スクリプトやフィルタを自動生成する。

## 4. スクリプト作成ツール

スクリプト作成ツールは、登録作業者からＯＰＡＣサーバに送信した検索要求の文字列を解析し検索式作成用テンプレートを生成する。

検索を行う際、スクリプト作成ツールが提示した単語を用いて登録作業者が検索を行うことで、ＯＰＡＣサーバへ送信する文字列と検索画面から入力された単語との対応をとることが可能となる。このため、サーバへ送信する検索式のみを用いるだけでテンプレートを作成することが可能となる。例えば、「流通」をキーワードとして設定した場合、検索ログ中の送信文字列で「流通」が含まれている場所がキーワードを入力すべき場所であると推定できる。タイトル・著者などに対しても同様に単語を決め、その単語を用いた検索ログを利用して、スクリプトを作成することにより、異なった属性を組み合わせた検索が行えるスクリプトを作成できる。

## 5. フィルタ作成ツール

フィルタ作成ツールでは、ＯＰＡＣサーバから送られてきた検索結果を解析しフィルタの作成を行う。検索結果には、特定の単語がタイトルに含まれるよう検索が行われており、文献情報の推定や項目の推定に利用する。フィルタ作成ツールでは以下の２つの処理を行う。

1)文献情報抽出処理

ＯＰＡＣサーバから送られてくる一件ごとの文献情報は同じフォーマットを持っている。このため、複数の文献情報が含まれている時には、整形用のタグや改行記号などのフォーマット情報が繰り返し現れる。例えば、WWWにより検索サービスを提供しているＯＰＡＣサーバでは、<LI>、<OL>、<TR>などのタグが文献を一つずつ分離するために用いられている。

フィルタ作成ツールはこのようなタグが繰り返し出現する部分を見つけ、そこを文献情報が並んでいる部分であると推定する。その後、文献情報以外の情報を排除するため、ヘッダー部分の最後の文字列とフッターの先頭の文字列を取り出し、文献情報抽出知識を作成する。

また、上記タグが出現するごとに文献情報を分割することで、得られた検索結果を一件ごとの文献情報に分ける。ここで、文献情報の先頭のタグを抽出し、検索結果を一件の文献情報ごとに分割する知識を作成する。

2)項目抽出処理

検索結果

<LI><A HREF="http://www.aitech.ac.jp/lib/cgi-bin/?sdid=2343>
建築設計資料集成 /日本建築学会編</A> <BR>丸善 , 1985<P>

分割 ⇩ 利用した区切り文字
「<A HREF>」「/」「</A>」「<BR>」 「,」「<P>」

| 区切り文字により分割されたブロック | ブロックごとの正規表現 |
|---|---|
| <LI><A HREF="http://www.aitech.ac.jp/lib/cgi-bin/?sdid=2343> | <LI>([^>]+)> |
| 建築設計資料集成/ | ([^/]*)/ |
| 日本建築学会編</A> | ([^<]*)</A> |
| <BR> | <BR> |
| 丸善 , | ([^,]*), |
| 1985<P> | *$ |

⇩

抽出知識:<LI>([^>]*)>([^/]*)/ ([^<]*)</A> <BR>([^,]*), *$

**図 2 抽出知識の作成例**

文献情報は区切り文字を用いて整形されている。このため、フィルタ作成ツールでは、区切り文字を用いて文献情報を複数のブロックに分割する。区切り文字には使用条件が付けられており、分割の際にはその使用条件に従って区切り文字が適用される。例えば、出版者を含むブロックのみに利用するような使用条件を区切り文字に付けることで、他の項目を含むブロックを不必要に細かく分割しすぎることを防ぎ、必要な情報を抽出できるようになる。

分割が終わった後、すべてのブロックを正規表現で表して、抽出知識を作成する。図 2に抽出知識作成の一例を示す。この抽出知識により、主要な書誌事項の項目であるタイトル・著者・出版者を切り出すことが可能である。

## 6. おわりに

インターネット上の図書目録を統一的な検索インターフェイスで検索できるWWWベース横断検索システムにおいて、ＯＰＡＣサーバを登録する作業を支援するため、検索ログより検索式作成用テンプレートであるスクリプトと、検索結果の整形知識であるフィルタを生成する知識作成支援環境を構築した。これにより、新たにＯＰＡＣサーバを登録する際、知識作成の作業時間をおよそ 10 分の１にすることができた。

**参考文献**

[1]柳本　他,「WWWベース図書館情報横断検索システム」,情処第５４回全国大会,1997

[2]柳本　他,「検索先の自動選択を行う WWW ベース横断検索システム」,情処第５６回全国大会,1998