平成 17 年度情報処理軽井沢セミナーレポート

筑波大学附属図書館
名波　一明

１．取り組んだ課題

arXiv プレプリント簡易検索システム

２．概要

プレプリント・サーバ e-Print archive (http://arxiv.org/) から OAI-PMH を用いてハーベストしたメタデータをもとに，簡易検索システムを構築した。検索システムは二種類作成した。一つは，a) 検索結果のタイトルから専門用語（キーワード）を抽出し，その専門用語（キーワード）を使って再検索できる検索システム，もう一つは，b) GETA による連想計算を実装した検索システム，である。

３．演習とその成果

(1) 演習第１日

OAI-PMH を使って約 20,000 件のメタデータを e-Print archive からハーベストした。ハーベスティングしたメタデータのうち，検索システムでは，"著者名"，"タイトル"，"発表日"，"サブジェクト"，"抄録"，"識別子" を使用し，XML 形式のファイルからこれらの項目を抽出するスクリプトを作成した。抽出した項目を１レコード１行の形式に整形したものをソースファイルとした。

次に，a) の検索インターフェイスを作成した（図１）。この検索システムは，従来からよくみられる検索システムように，入力した検索語が一致しているレコードを返すというものである。なお，先に作成したソースファイルを検索対象としている。

(2) 演習第２日

a) の検索システムで，検索結果のタイトルから専門用語（キーワード）を抽出して，再検索をかける機能を追加した（図２）。専門用語を抽出する部分は，Perl モジュール TermExtract を用いた。なお，TermExtract は「茶筌」などの日本語形態素解析システムと組み合わせて使用すれば，日本語の文章からも専門用語を抽出することができる。

また，取り組んだ課題を離れて，OAI-PMH 以外のメタデータの取得方法について調査した。主に ZING SRW/SRU プロトコルによる横断検索のドキュメント（http://www.loc.gov/z3950/agency/zing/zing-home.html 以下のページ）を眺めた。

(3) 演習第３日

b）の汎用連想計算エンジン GETA による連想検索を実装した（図３）。GETA のインストールから実装までは，導入マニュアルにしたがって作業した。演習第１日で作成したソースファイルから，単語出現頻度ファイル（freq ファイル）を作成した。次に，freq ファイルから GETA の機能を使って，連想検索用インデックスを作成した。検索インターフェイスは，配付に含まれる検索用 CGI をカスタマイズした。

４．研修で学んだ技術及び知識

・汎用連想計算エンジン(GETA)

http://geta.ex.nii.ac.jp/

GETA の導入および操作マニュアルに関する情報。

・専門用語抽出(TermExtract)

http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

テキストデータから専門用語（キーワード）を取り出すための Perl モジュール。

５．事前準備として必要と思われるもの

Perl 等のスクリプト言語によるテキスト処理の知識。XML の処理に関する基礎的な知識。プレゼンテーションの技術。日本語の処理を行うなら，日本語形態素解析システム（例えば「茶筌」）を準備しておくと何かと応用が利く。

６．今後の課題

メタデータを取得できるサイトの調査。OAI-PMH に対応するデータ・プロバイダだけでなく，REST あるいは SOAP の API を公開しているデータ・プロバイダ（例えば，Amazon Web Service）や ZING SRW/SRU に対応する検索サイトからのデータの取得方法も調査したい。

７．今後の計画

農学分野の論文情報が数千件あるので，そのデータをもとに GETA を使った連想検索システムの構築に取り組む予定である。

８．演習の感想

連想による検索という考えは新鮮だった。GETA については名称だけは聞いたことがあったが，研修で講義を聴かなかったら，関心を持つことはなかったと思う。カリキュラムに盛り込まれていたのは幸いだった。

演習は自分で課題を設定して学術ポータルの作成に取り組むという形式だったが，他の参加者の課題から刺激を受けるところが多かった。また，学術ポータルを作成していくうえで，講師の方に直接，質問できたことも貴重な経験だった。実際に作業をすると，コンピュータおよびプログラミングの知識不足がよくわかった。

一連の講義を聴いて，インターネット環境下での情報処理は，XML が標準になる方向にあるように思われる。今後は，XML および XML 関連技術の知識の必要性を感じている。それに関連して，異種データベース，異種システム間の連携，さらには情報サービスのオートメーション化といった話題に注目していきたい。また，利用者が学術ポータルに対して求める機能は何か，といった基本的な問題も踏まえ，要求と技術のバランスのとれたシステムを立案できることを目標としたい。

９．備考，その他

このような機会を与えてくださった NII に感謝致します。また，親切にサポートしてくださった NII 研修担当，講師の皆様に御礼申し上げます。

図１．キーワード検索の実行結果

図２．タイトル関連検索をクリックしたところ

論文タイトルから専門用語が自動抽出されて，検索フォームが生成される

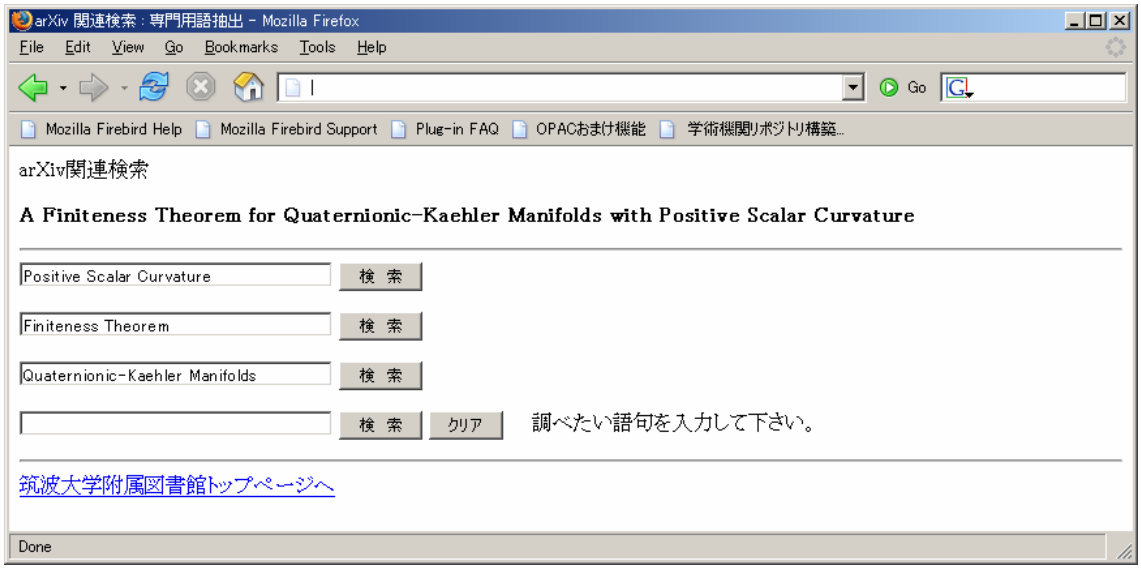

図３．連想計算を利用した検索の検索結果