E-003

Webサービスを利用した総合電子文書検索システム

The comprehensive electronic document search engine using Web service

小池 勇治† 矢島匡人† 高野 明彦‡ 絹川 博之†
Yuji Koike Tadahito Yajima Akihiko Takano Hiroshi Kinukawa

1. はじめに

近年多くの検索システムが登場している。各検索システムはそれぞれ独立しており、利用者はそれぞれの検索システムに対して検索を行う必要があった。また、単語を入力する「キーワード検索」では、利用者がキーワードを考える必要があり、検索漏れや、求めていない結果が返ってくるといった問題が生じ、うまくヒットするような単語を選ぶのは困難であった。ユーザの検索の支援を行いながら複数の検索システム同士が連携を取り検索することができれば、このような問題は解決することができる。

そこで本研究では複数の検索システムが存在する場合に、お互いの検索システムの結果を利用して検索することのできる総合電子文書検索システムを提案する。

2. 総合電子文書検索システム概要

総合電子文書検索システムは、ある検索システムからの検索結果を利用し、複数の異なる検索システムに対して検索実行を指示することができるシステムである。検索結果の文書群そのものや、検索結果の文書群から抽出した特徴語を利用して、複数の異なる検索システムの横断的な検索を可能とする。

複数の検索システム同士が連携しながら検索することのできる統合文書検索システムは次の機能を備えている必要がある。

（1） 検索機能

(a) 一斉検索

キーワードやテキスト文を検索条件としてシステムに与え、利用者が指定する複数の検索システムに実行を支持できること。

(b) 検索結果の特徴語の表示

検索結果の文書群を要約するような特徴語を表示する。この特徴語により、利用者に思いつかなかった検索キーワードを知るきっかけを与え、うまくヒットするような単語キーワードを見つける支援となる。

(c) 検索結果の文書群からの検索

検索結果として得られた文書群そのものを検索条件としてさらに検索を行う。当初の検索結果を得た検索システムとは異なる検索システムに対して検索実行を支持可能である。

（2） 分散処理

ネットワーク上の異なる検索システムを利用するには、異なるマシンで処理を実行する必要がある。また、検索システムを自由に組み合わせて利用するシステムであるため、各システムに対して柔軟に連携できなければならない。さらに新たな検索システムを登録したい場合も容易に追加することのできる仕組みが必要である。

† 東京電機大学大学院 工学研究科

‡ 国立情報学研究所

3. 総合電子文書検索システムの実装

総合電子文書検索システムを実現するために検索機能として連想検索エンジン、分散処理の仕組みとして Web サービスを用いる。

3.1 連想検索エンジン

連想検索エンジンとして、汎用連想計算エンジン GETA[1]を使用し、単語群から文書群、文書群から文書群、文書群から単語群の検索など、2 章で述べた検索機能を実現する。GETA は各検索システムに組み込まれている。

GETA は、大規模な文書を高速に連想計算することができるソフトウェアである。

3.2 Web サービス

Web サービスは分散した複数のサービスを組み合わせて連携させ利用することを可能とするものである。

Web サービスの仕組みでは異なる検索システムを自由に組み合わせ連携させることが可能となり、新たな検索システムの加入にも対応できる。

3.3 システム構成

総合電子文書検索システムではサービスプロバイダとして各検索システム、サービスリクエスタをポータルサイトとして構成し、ポータルサイトでは Web アプリケーションとしてエンドユーザからの問い合わせを受ける。このとき、エンドユーザは利用したい検索システムを自由に組み合わせて選ぶことができる。今回は下記の検索システムを対象としたシステム構成を図 1 に示す。

- XML 電子辞典システム[2]
- 電子新書システム
- 新書マップ[3]
- 新聞記事検索システム

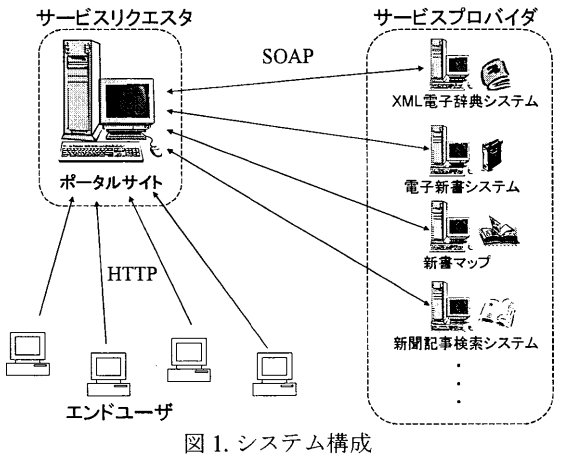


図 1. システム構成

3.3.1 Web サービスが提供するサービス

サービスプロバイダである各検索システムは連想検索機能のサービスを提供する。各機能を実現するために必要となるサービスは以下の 2 つである。

（1） 単語群から文書群の類似度計算
（2） 文書群から単語群の類似度計算

サービスリクエスタはこの 2 つのサービスを組み合わせて、2 章で述べた検索機能を実現する。

これらのサービスを WSDL に記述することによって、サービスの利用者はサービスを利用できるプログラムの作成が可能になり、また、新たに加わる検索システムも WSDL を用いてサービスを提供することが可能となる。

3.3.2 ポータルサイト

サービスリクエスタであるポータルサイトについて述べる。エンドユーザによって利用され、連想検索の機能を持つ複数のサービスプロバイダを自由に組み合わせて検索することの出来るポータルサイトである。

- 一斉検索
- 検索結果の文書群からの検索
- 検索結果の特徴語の表示

ポータルサイトの機能として以上の機能を備えている。これらの機能は 3.3.1 で述べたサービスを持つサービスプロバイダを組み合わせて利用することで実現している。利用者は存在する検索システムから利用したいサービスを自由に選び、組み合わせることができる。それぞれの機能についての手順を述べていく。

（1）一斉検索の手順
(a) エンドユーザがキーワードやテキスト文を検索条件として入力
(b) ポータルサイトが検索文に形態素解析を行い形態素単位に分割
(c) 各検索システムの「単語群から文書群の類似度計算」のサービスに問い合わせ、検索結果の文書群を取得
(d) 取得した結果の文書群をサービスごとに分けて表示

（2） 検索結果の特徴語の表示の手順
(a) 得られた検索結果の文書群を各サービスの「文書群から単語群の類似度計算」のサービスに問い合わせ、単語群を取得
(b) 取得した単語群を特徴語として表示

（3） 検索結果の文書群からの検索の手順
(a) エンドユーザが検索結果の中からさらに検索したい文書群を指定
(b) ポータルサイトは指定された文書群をその文書を持つ検索システムに「文書群から単語群の類似度計算」のサービスに問い合わせ、単語群を取得
(c) 2 で得られた単語群を各検索システムに「単語群から文書群の類似度計算」のサービスに問い合わせ、検索結果の文書群を取得
(d)取得した結果の文書群をサービスごとに分けて表示

（4） 統合電子文書検索システムにおける横断検索例として XML 電子辞典システムの検索結果から新書マップ検索を実行した例を図 2 に示す。

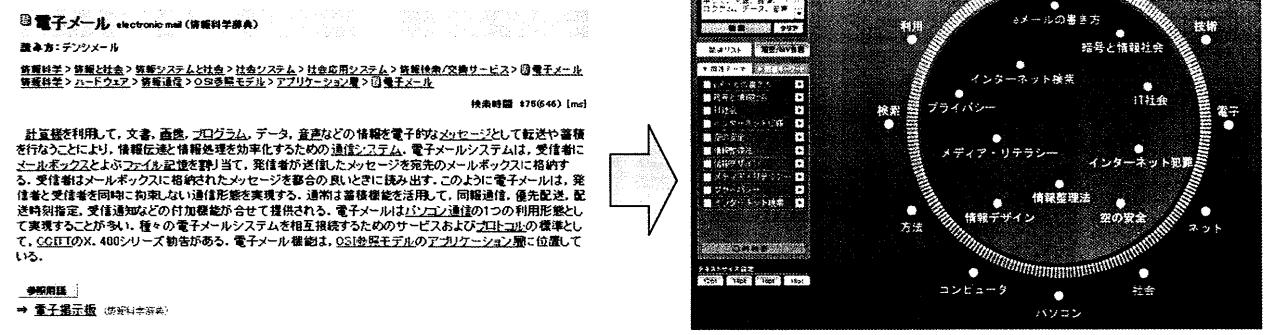
4 おわりに

本研究では連想計算機能を持った Web サービスのサーバを利用した統合検索システムについて述べた。既存の検索システムやこれから開発される検索システムに今回の連想検索機能のインタフェースを持つ Web サービスの窓口を持たせるだけで、統合検索システムとして利用することができるようになる。今回は総合電子文書検索システムでは Web サービスのサービスリクエスタとしてポータルサイトとしたが、検索システム同士がお互いに Web サービスを利用するといったことも可能である。今後は利用方法についても検討していく。

Web サービスには UDDI によるディレクトリサービスがある。本システムも UDDI と組み合わせることで、サービスの登録、発見などを行うことができるので、今後 UDDI を利用できるシステムにしていく予定である。

参考文献

[1] 高野明彦, 他:汎用連想計算エンジンの開発と大規模文書分析への応用, http://geta.ex.nii.ac.jp/ (2002)
[2] 小池勇治, 高野明彦, 絹川博之: 複数辞典の鳥瞰が可能な XML 電子辞典システム, FIT2003 (2003)
[3] 新書マップ, http://shinshomap.info/
[4] 奈良先端科学技術大学院大学自然言語処理学講座: 形態素解析システム茶筌, http://chasen.aist-nara.ac.jp/



XML 電子辞典システム　　新書マップ

図 2. 統合電子文書検索システムにおける横断検索例