東京大学 OPAC Plus "言選 Web"
-関連学術用語による日本語文献情報への簡易ナビゲーションシステム-

前田　朗[†]　　中川　裕志[‡]
東京大学社会科学研究所[†]　　東京大学情報基盤センター[‡]

１．はじめに

東京大学情報基盤センターでは「東京大学 OPAC Plus "言選 Web"」[1]を 2008 年 7 月から試行サービスしている。これは入力したフレーズから日本語関連学術用語（キーワード候補）一覧を提示し、東京大学附属図書館の蔵書目録（東京大学 OPAC）及び雑誌記事索引（国会図書館 PORTA）にナビゲートするサービスである。

学術文献データベースにおいて、キーワード候補提示機能を備えたシステムは珍しくない。「東京大学 OPAC Plus "言選 Web"」もそのひとつに位置づけられ、１）システム構築が簡易、２）カスタマイズやアレンジの余地が大きい、という特徴を持つ。

本論ではこの「東京大学 OPAC Plus "言選 Web"」を具体例とし、Web サービス化されたデータベース（Web 検索エンジンを含む）と「言選 Web」による関連用語抽出をコアに、一般に入手可能なシソーラス、学術文献データベースの外部インターフェイスを複数種組み合わせ、関連学術用語を提示し文献へナビゲートするシステムを、簡易にかつカスタマイズの余地が大きい形で構築する手法を示す。

２．「東京大学 OPAC Plus “言選 Web”」

「東京大学 OPAC Plus “言選 Web”」は入力したフレーズから関連用語一覧を提示し、各用語から東京大学 OPAC 及び雑誌記事索引へとナビゲートする。各用語をクリックすると、その用語の関連学術用語一覧が別ウインドウで開き、用語の関連を順次たどれるようになっている（図１）。

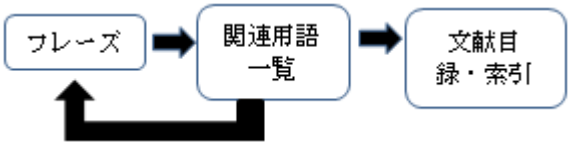

図１　利用モデル

The University of Tokyo OPAC Plus “Gensenweb” : The simple navigation system that based on related academic terms.

†Akira Maeda.
Institute of Social Science, the University of Tokyo
‡Hiroshi Nakagawa
Information Technology Center, the University of Tokyo

吉川ら(2007)[2]によると、キーワード候補を提示するツールは、「ブラブラとながめながら探す」（散策）という検索行動に向いている。当システムもその用途をメインに考えている。ただし場合によっては、キーワードを思い出す用途に使えるかもしれない。例えば、姉川の戦いで織田信長と戦った「浅井長政」を思い出せなくとも、「織田信長 敵対」といった複数フレーズ指定などの工夫で見つけられる(2010.1.9 時点)。

関連用語の提示は、２次元でグラフィカルに用語の関連を示すものではなく、用語の一覧である。ただし、重要度でランキングをしている。重要度の高い用語ほど一覧の上位に、かつ文字のフォントサイズを大きくし、重要度が直感的に分かるようにしている。

関連用語には日本語学術用語をできるだけ提示することを目指している。そのため、関連用語の情報源は、国内学術サイト(ドメイン ac.jp の Web サイト)の Yahoo！検索結果をデフォルトに、CiNii(NII 論文情報ナビゲータ)など学術情報をメインに用意している。また、提示する関連用語が実際に日本の学術論文で利用されているかどうかを、CiNii での文献ヒットの有無を元に判断するオプション機能もある。

関連用語一覧では、上部にシソーラスすなわち人手で関連づけた用語を、下部には上述のように Yahoo!などの Web 上の情報源をキーワード検索した結果を「言選 Web」を適用して機械的に抽出した関連用語を提示している。シソーラスとしては国会図書館件名標目表と日本語 WordNet を組み合わせて利用している。

各関連用語の左には、「OPAC」ボタンと「雑誌記事」ボタンがある。それぞれ、「東京大学 OPAC」と「雑誌記事索引（国会図書館 PORTA)で該当キーワードを検索した結果を別ウインドウで表示する。文献がヒットしないことをできるだけ避けるため、１）キーワードのフレーズ検索、２）キーワードを形態素に分割し AND 検索、の両方で検索する仕組みとした。

３．関連用語の提示手法

関連用語の提示は、入力したフレーズで学術

文献サイトもしくは学術文献データベースを検索し、そのタイトルや要約（抜粋）を専門用語抽出システム「言選 Web」[3]にかけることで行なっている。

Web 検索エンジンの検索結果から複合語を取り出し頻度順一覧を出力することで、関連用語を提示するサイトに「Web 関連語抽出」[4]があるが、「東京大学 OPAC Plus"言選 Web"」と違い、用語の抽出とランキングに「言選 Web」を使っていない。「言選 Web」は専門用語抽出システムであり、文章からの用語の切り出しは学術向けに調整してある。加えて頻度のみによるランキングと異なり、用語を構成する単名詞の連接も考慮するため Web ページ抜粋のような少ないテキスト量でも、より細かい重要度ランキングを得られる。

他に近い仕組みに、佐藤ら(2003)[5]の手法がある。大きな相違点は佐藤らが Web 検索エンジンの出力する Web ページ抜粋ではなく、Web ページそのものを対象としていること、入力語と関連用語候補の検索ヒット件数から共起度を得て、共起度が小さい結果を排除したことである。「東京大学 OPAC Plus"言選 Web"」はリアルタイム性を確保するため、精度が落ちるが簡便な処理で済ませている。

関連用語の抽出には文書構造を利用した手法もあるが、「言選 Web」を用いた方式であれば、情報源の文書構造に関わらず関連用語を抽出できる。

４．システム構成にみる手法のメリット

既存の学術文献データベースにキーワード候補提示機能を付与するには、検索システムそのものを改修できればよいが大がかりになる。簡易に実現するには外付けの方式が有力である。

「CiNii with 関連検索ワード」[6]のように、Web ブラウザのアドオンで既存のデータベースに機能を付与する方法がある。別の方法として大阪市立大学の"Subject World"や「東京大学 OPAC Plus "言選 Web"」のように、関連用語を提示するサイトを中継して学術文献データベースにリンクを張る手法が考えられる。

特に「東京大学 OPAC Plus "言選 Web"」では一般に利用可能な学術情報資源のみモジュールとして組み合わせることで、実現が容易かつ、モジュールの組み換えによるカスタマイズの余地が大きい（図２）。例えば、現在は実現していないが、ある学術分野で Web サービス化されたデータベースがあれば、その特定分野に絞った用語提示も可能である。

また、サーバシステムの負荷もさして大きくはない。情報源の検索や文献へのリンクは外部サーバを活用する。サーバシステムで行なうのは、検索結果を「言選 Web」にかける処理と、内部のシソーラスから関連用語を取り出す処理のみで済んでいる。

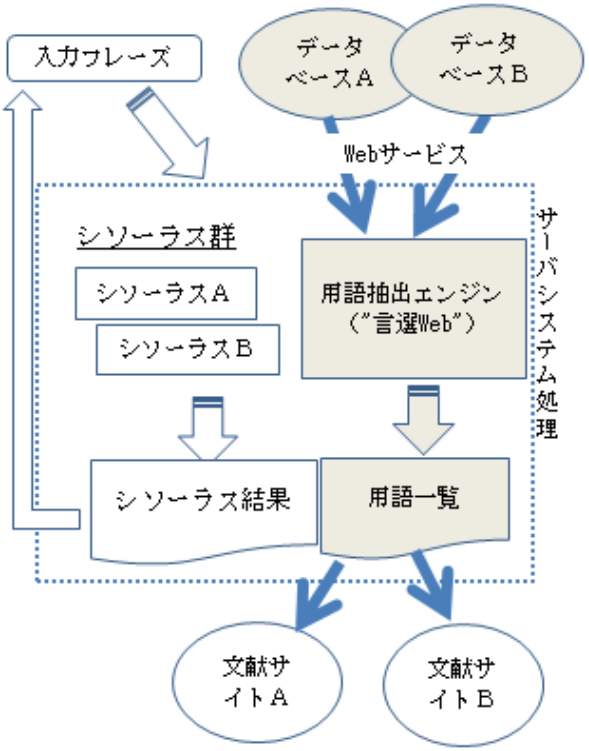

図２　システム構成モデル

５．おわりに

本手法は、関連用語の提示という観点でも、使いやすいインターフェイスという観点でも、最良を求めたものではない。しかし、関連用語の情報源に融通がきき、実現が容易な手法としては、意義があるかと考えている。

参考文献

[1] 東京大学 OPAC Plus “言選 Web” https://mbc.dl.itc.u-tokyo.ac.jp/UT_OPAC_Plus_gensenweb/ [accessed 2010-1-9].
[2] 吉川日出行編. サーチアーキテクチャ：「さがす」の情報科学. 2007, 271p.
[3] 言選 Web http://gensen.dl.itc.u-tokyo.ac.jp/ [accessed 2009-12-30].
[4] Web 関連用抽出 http://yapi.ta2o.net/kanrenp/ [accessed 2009-12-30].
[5] 佐藤理史, 佐々木靖弘. ウェブを利用した関連用語の自動収集. 情報処理学会研究報告 自然言語処理研究会報告. 2003, vol.2003, no.4, p.57-64.
[6] CiNii with 関連検索ワード https://mbc.dl.itc.u-tokyo.ac.jp/related_term/cinii_relatedterm.html [accessed 2009-12-30].