Wikipediaと図書館情報資源による調べ方自動提示システム

坂井 哲†　　増田 英孝†　　清田 陽司‡　　中川 裕志‡
†東京電機大学未来科学部　　‡東京大学情報基盤センター

1 はじめに

情報を探す際に利用者の要求が曖昧なことがある．このような場合，情報探しのテーマを利用者に推薦することが求められる．情報探しのテーマを推薦するためにはサービス提供側は「カバレッジ」「分類体系の存在」「信頼性」の3つの条件を満たしていることが必要である．

日常の情報探しにおいて最も広く用いられているWebサーチエンジンは，上記の3つの条件のうち「カバレッジ」は満たしているが，「分類体系の存在」「信頼性」においては不十分である．

一方，図書館の代表的なWeb情報サービスであるOPACでは，「分類体系の存在」「信頼性」の条件は満たしているが，「カバレッジ」については不十分である．

オンライン百科事典Wikipediaは，この3つの条件を考慮したとき興味深い位置に存在する．誰でも編集が可能であることから「信頼性」については課題があるが，「カバレッジ」「分類体系の存在」の条件は満たしていることから，Webと図書館システムの間のギャップを埋める架け橋として利用できる可能性がある．

本稿では，Wikipediaと図書館情報資源を統合的に活用することでこれらの条件を満たすことを示し，Wikipediaを手がかりとして図書館情報資源へ誘導することによる調べ方自動提示システム（リサーチ・ナビ検索システム）について述べる．

2 Wikipediaと件名標目表を統合的に活用した分類自動導出

Wikipediaは「テーマ推薦の要件」と「カテゴリの構造」の観点からみたときに，きわめてユニークな特徴をもっている．この特徴をうまく用いて，情報探索の出発点としてWikipediaを利用し，そこから概念を一般化することによって図書館の分類体系に導いていくという方法を我々は既に提案している[1]．

2.1 アルゴリズムの概要

図1に導出アルゴリズムの概要を示す．まず，Wikipediaカテゴリの構造について説明する．Wikipediaの記事「阪神・淡路大震災」には，カテゴリとして「日本の経済史」「地震の歴史」が付与されている．さらに，カテゴリ「日本の経済史」には上位カテゴリとして「経済史」が，カテゴリ「地震の歴史」には上位カテゴリとして「災害と防災の歴史」「地震」が付与されている．このように，Wikipediaの記事を一つとりあげてみると，関連するカテゴリ群をツリー構造として取り出せることがわかる．

次に，Wikipediaカテゴリと図書館の分類体系の対応付けについて説明する．Wikipediaカテゴリと図書館の件名の間には，カテゴリ名が一致するものが存在する．図1では，「経済史」「災害」「地震」が一致している．

よって，「阪神・淡路大震災」につながるカテゴリが構成する有向グラフの構造を再帰アルゴリズムによってたどることで，「阪神・淡路大震災」に関連する分類を自動的に導出することができる．リサーチ・ナビ検索システムでは，グラフのエッジに対する重みスコアをノード間の文字列類似度によって定義し，ビームサーチによって重みスコアが相対的に大きい件名を絞り込むアルゴリズムを採用している．

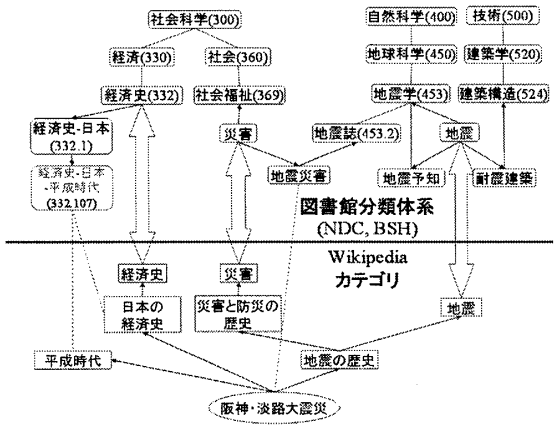

図1: 図書館分類体系とWikipediaのカテゴリの対応付け

An Automated Pathfinder System Using Wikipedia and Library Infomation Resources
†Satoshi Sakai　†Hidetaka Masuda
‡Yoji Kiyota　‡Hiroshi Nakagawa
†School of Science and Technology for Future Life, Tokyo Denki University
‡Information Technology Center, The University of Tokyo

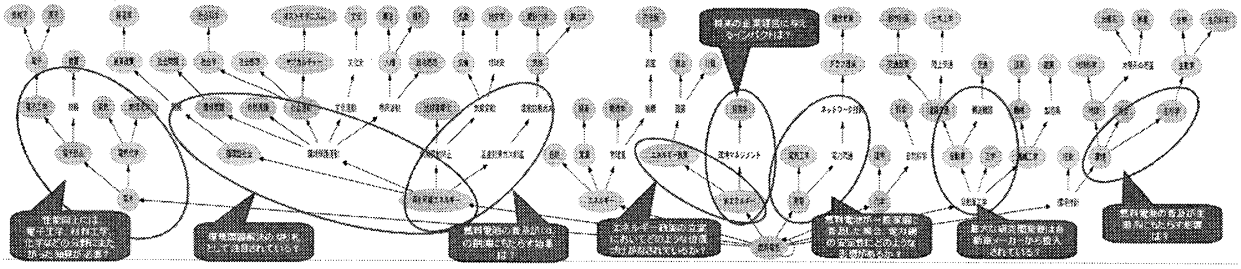

図 2:「燃料電池」のテーマグラフ

2.2 テーマグラフによる分類の俯瞰

上記に示したように，Wikipedia カテゴリ構造と図書館件名標目表を活用することで，任意の Wikipedia の項目に対して，関連する件名を自動的に導出することができる．リサーチ・ナビ検索システムでは，この導出の過程をグラフ構造として自動的に描画する機能を実装している．このグラフ構造は，与えられた Wikipedia の項目がどのようなテーマを有しているのかを示していることから，テーマグラフと名付けている [2].

図 2 に,「燃料電池」という検索キーワードから導き出されたテーマグラフの例を示す．テーマグラフの内容を考察していくことで,「燃料電池」という概念がどのようなテーマと関連を持っているのかを知ることができる．例えば,「再生可能エネルギー」「循環型社会」「環境問題」などの件名と「燃料電池」との関連性を考察すると,「燃料電池は環境問題解決の切り札として注目されている」という背景を見いだすことが可能である．

3 調べ方自動提示システムの開発

上記の考え方にもとづき，国立国会図書館リサーチ・ナビ [3] の検索システムの開発を行った．国立国会図書館の Web サイトにて 2009 年 5 月から一般公開されている．図 3 に，リサーチナビ検索システムの検索結果例を示す.

システム評価として，日常的に検索サービスを利用している 50 人（理系大学生:3 割 IT 系社会人:5 割 その他:2 割）に対しアンケート調査を行った．アンケート項目の 1 つにある「テーマグラフは調べ物にとってどのようなヒントが得られたか」という質問に対しては，約 80%の被験者が何らかのヒントを得られたと回答した．回答例には，別の観点で調べ物を進めることができた，自分が気づかなかった・思いつかなった関連語が表示された，調べ物がどういう分野のものかが一目で分かった，などがあった．このことから，テーマグラフには調べ物のヒントを得るための様々な効果があるといえる.

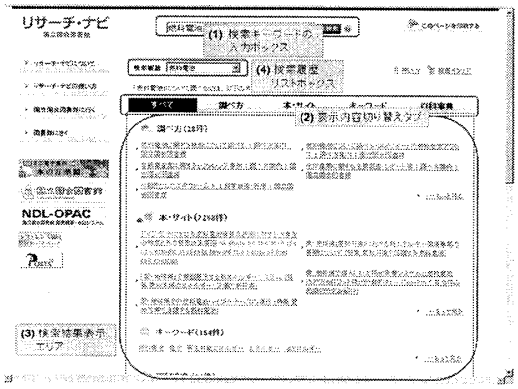

図 3: リサーチ・ナビ検索システムの検索結果例

4 まとめ

本稿では Wikipedia と図書館情報資源を統合的に活用し，さまざまなキーワードから関連する件名を自動的に導出する機能を実装することで，調べ方のためのヒントを利用者に与えるようにした.

謝辞

本研究は，科研費若手研究（B）及び特定領域研究（情報爆発）の支援を受けた.

参考文献

[1] 坂井哲, 増田英孝, 清田陽司, 中川裕志. 国立国会図書館リサーチ・ナビにおけるテーマグラフの生成. 情報処理学会 第 96 回情報学基礎 第 37 回デジタル図書館ワークショップ合同研究会 FI, 2009.

[2] 坂井哲, 清田陽司, 増田英孝, 中川裕志. 図書館と web の分類体系を統合的に活用したテーマグラフ可視化インタフェース. 情報処理学会 第 70 回全国大会, 2008. 4ZK-9.

[3] リサーチ・ナビ. http://rnavi.ndl.go.jp/.