用途別統合検索における図書情報検索

4Y－4

嶺岸 則宏　　田中 聡
三菱電機(株) 情報技術総合研究所

1.はじめに

将来の電子図書館システムは、ネットワーク化されることが予想され、老若男女あらゆる人が利用者として想定される。このため、図書情報の検索は、情報検索の専門家でない一般の利用者でも、容易に行なえることが重要である。

中核を成す大規模な総合的図書館から、その他、地方図書館、大学や研究機関の図書館、また海外の図書館など様々な図書館が電子化・ネットワーク化されていくことにより、利用者の地域／専門／時間を問わない各種の問い合わせやシステムへのアクセスが容易になる。

このような総合電子図書館に対して、各電子図書館のある一分野の図書情報を統合し、用途別インデクスを備えた、仮想的な専門電子図書館を提案する。関連する世界中の電子図書館を連携することによって、世界規模の分野・メディア別電子図書館を作り出すことができる。ある専門分野の図書情報を検索する場合には、複数の図書館サーバにアクセスしなくても、この仮想的な専門電子図書館にだけアクセスすれば、統合された結果を得ることができる。これにより、利用者は世界中の図書資産を目的や用途に合わせて自由に効率良く利用することが可能になり、また、ネットワークの負荷分散も図れる。

このような仮想的な専門電子図書館は、バーチャルインデクスの集まりであるので、通常の電子図書館の１つのサービスとして同じサーバ上に構築することも可能である。

現在の図書情報検索システムは、システム毎に異なる図書分類体系で整理されているため、検索者は、システム毎の図書分類体系を十分に理解しなければ、目的の図書情報を効率的に検索することができない。ここで提案する用途別統合検索とは、上記の仮想専門電子図書館システムを構築し、その上で個別に構築され異なる観点で分類された様々な情報を、検索者が理解し易い分類で整理統合することによって、検索者のいろいろな検索目的に柔軟に対応できるものである。例えば、医学分野を例にとると医学生や医師向けには、米国医学図書館分類法のような極めて専門的な分類基準によって検索できるようにし、一般人向けには、家庭向け医学全集といった医学書の目次を基にした分類基準からも検索できるようになる。

しかし、人手による分類の場合、分類体系を理解した上で臨まなければならず、次々と多様な分類体系を持ち込まれた場合に対応できなくなる可能性がある。また、従来から研究されている自動分類の方法では、分類基準に最も適切な図書サンプル、あるいは典型的な図書のサンプルを設定し、それと合致、似ているものを分類する等、分類者の意図が反映される問題がある。

本研究では、それら問題を解決するため、用途別統合検索に必要な用途別インデクス作成について、分類基準の単語を基に分類対象の図書群全体を統計的に処理し、自動的に作成する手法を検討した。

2.用途別インデクスの作成

複数の図書館サーバから収集してきた図書情報に対し、検索者の利用しやすい分類体系に従って分類し対応付けをするのが用途別インデクスである。

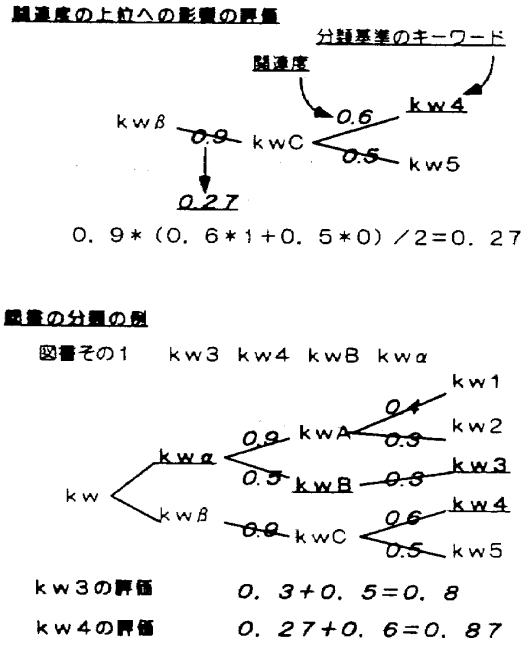

図1　用途別インデクス作成の例

Special Purpose Retrieval for Digital Library Information
Norihiro MINEGISHI, Satoshi TANAKA
Information Technology R&D Center
Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa, 247-8501, Japan

分類は2段階の手順により行われる。まず与えられた分類体系の各概念間の階層関係に関連度を設定した関連度マップを作成する。そして、分類対象となる図書に含まれている単語をその分類体系関連度マップにマッピング・評価を行い、最も適切な場所へ分類を行う。

関連度マップの作成には、分類体系で使用されている概念を表す単語が分類対象図書全体に出現する頻度を用いて算出する。ある図書中に2つの単語が出現した場合、何らかの関連があると考え共出現頻度を算出し、各単語の総出現頻度に対する比率で関連度を定義する。これを全ての分類基準の階層関係に対して算出し単語関連度マップを作成する。

次に分類対象となる図書1冊毎に関連度マップをもとに評価を行う。図1に示すように、これから分類される対象となる図書に含まれる単語を関連度マップ上にマッピングし、ルートから各末端ノードへたどるパス上の単語間の関連度を累計して最も高いところに分類を行う。

しかし、ある1つのパス上の単語がすべて1冊の図書に含まれるというのは希であり、実際は関連度マップ上に散在する場合が多い。この場合、図書に含まれず直接マッピングされなかった単語についても、その周辺（特に下位の単語）が図書に含まれていれば影響を及ぼすとし、仮の関連度を評価する。例えば図1に示した場合のように、キーワードkwCについて、その下にあるキーワードkw4が分類対象となる図書に含まれているため、その影響を考慮した比率を元の関連度マップ上の関連度に掛けた仮の関連度を算出している。

3.試作システム

以上の考えに基づき、ワークステーション上でインデクス作成を行い、ネットワーク経由でパソコンから図書検索するシステムを試作した。

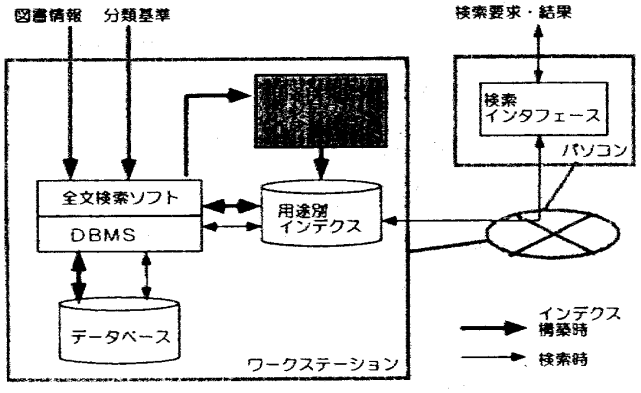

図2　システム構成図

試作システムに試験データを投入し実験、評価を行う予定である。実験は、医学分野を対象とし、図書の代わりに技術文献を用い、複数の分類基準に対して実験を行なう。

・使用文献データ:JICST科学技術文献速報CD－ROMライフサイエンス編の文献（約9万件）
・使用分類基準:日本10進分類（NDC）、米国医学図書館分類法（NLM）、看護大全（一般向医学書）目次
・頻度評価対象:各文献の200文字程度の抄録部分

評価項目を明らかにするため、この方式における問題点を机上評価した。今後これら項目について実験を行い評価する予定である。

(1)計算量

関連度マップを作成する際の計算量は、分類基準に含まれている単語を基に各文献毎に走査、相互の関連度を計算するため1文献あたりの単語含有数の階乗と、対象文献数および分類基準に出現する総単語数の2乗に比例する。また、実際に各文献を分類対応づけする際の計算量は、各文献に含まれる単語数と文献数に比例する。関連度マップ作成に関して、対象となるデータの規模により計算量が爆発する可能性がある。関連度の定義や、計算アルゴリズムの見直しの必要がある。

また、分類基準や対象文献に変更が生じた場合の関連度マップの見直しについては、始めから再計算をするのではなく、差分情報による修正方法の検討が必要である。

(2)分類基準の性質

分類基準の階層数が少ない場合（平たく詳細に分類されるようなタイプ）、本方式による効果は現れず、単純に分類基準の出現するキーワードにより直感的に分類する場合と大差のない結果となる。

これは、今回の評価式は分類基準の階層の深さに意味をもたせるような考え方をとっている為である。このような場合にも効果を発揮させるためには別の評価式を検討する必要がある。

4.おわりに

我々は、用途別検索のためにインデクス構築の手法を検討し、試作を行った。机上評価では、この方式では、計算量の問題、一部タイプの分類基準については極端に性能が落ちること等、明らかになっている。

今後、実験、評価を行い、計算量の削減、用途別インデクスを生かした新しい検索手法の検討などを課題として進めていく。

謝辞

本研究開発は、通商産業省「次世代電子図書館システム研究開発事業」に参画し、実施したものです。本事業は，国からの委託を受けた情報処理振興事業協会（IPA）より，（財）日本情報処理開発協会（JIPDEC）が再委託され実施しているものです。