様式 C-19

# 科学研究費補助金研究成果報告書

平成２２年５月３１日現在

研究種目：若手研究(B)
研究期間：2008～2009
課題番号：20700101
研究課題名（和文）複数の検索システムを動的に組み合わせた統合型情報検索システムの構築

研究課題名（英文）Integrated Information Retrieval using Multiple Retrieval Systems

研究代表者
鈴木　優（SUZUKI YU）
京都大学・大学院情報学研究科・特定研究員
研究者番号：40388111

研究成果の概要（和文）：複数の検索システムを動的に組み合わせることによって，検索システムの精度を向上させるための手法の提案を行った．本研究では，検索システムが出力する関連度を複数の検索システムの間で相互に比較可能な値へ変換する方法，および検索対象そのものの質を測定する方法などを提案している．評価実験によって，確かに提案手法を利用することによって精度が 10%程度向上することが確かめられた．

研究成果の概要（英文）：In this study, we propose an integrated information retrieval system using multiple information retrieval systems, for improving the accuracy of retrieval. This system includes a method to convert scores, which are calculated by multiple information retrieval systems to be equivalent, and a method to measure the quality of retrieval targets. In our experiment, we confirmed that our proposed methods improve the accuracy about 10%.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| ２００８年度 | 1,200,000 | 360,000 | 1,560,000 |
| ２００９年度 | 1,500,000 | 450,000 | 1,950,000 |
| 年度 | | | |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 2,700,000 | 810,000 | 3,510,000 |

研究分野：　総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：ディレクトリ・情報検索，アルゴリズム，コンテンツ・アーカイブ，情報基礎，情報図書館学

## １．研究開始当初の背景

現在，計算機システム上では様々な検索システムが構築されている．これらの検索システムは様々な目的が想定されており，例えば Yahoo!や Google などの検索エンジンは不特定多数の用途に用いられていることに対し，特許検索システムや論文検索システムなど特定の用途に利用するための検索エンジンがある．また，検索エンジンごとに特性があることも多く，ある種類の問い合わせに対する精度が高いが，他の種類の問い合わせに対

して精度が低いことも考えられる．これらの検索エンジンはインターネット上に大量に存在している．ところが，利用者はその検索エンジンすべてを把握することは困難である．そのため，利用者がある意図をもって検索を行うとき，その意図に対して適切な検索エンジンを選択することは困難である．

## ２．研究の目的

本研究の目的は,ある問合せに対して適している検索システムを複数の検索システムから自動的に選び出し,組み合わせることによって,高精度な情報検索システムを実現するための方式を確立することである.
本研究で提案する統合型検索とは,どのような検索目的を持った人であっても,特別な知識を必要とせずに,必要な検索結果へアクセスできることを意味する.現在,インターネットの普及と共に,司書 などの検索の専門家から,検索システムについて理解していない初心者までが,インターネット上の検 索システムを利用している.一方,従来の検索システムは以前の主な利用者であった検索の専門家が 利用していたものと本質的に同一である.そのため,初心者が検索を行った場合に必要な検索結果を得ることができないという問題がある.
特定分野向け検索システムは,検索要求の種類と検索システムの特性双方が合致した場合に,検索精 度を最大限に発揮することができる.しかし,これらの特定分野向け検索システムの種類を増やさなければ,利用者の多様な検索要求に応えることができない.一方,検索システムの種類を数多く用意 できた場合は,利用者の要求に応じることが可能な検索システムを利用者自身が探し出す必要がある．さらに,利用者の検索要求を利用者自身が把握していない場合も多いため,利用者にとって最適な検索 システムを選択すること自体が困難である.そのため,利用者の検索意図に応じた検索システム選択を 自動的に行う必要があると考えられる.
特定分野向け検索システムを効果的に利用するために,本研究では複数の検索システムを利用者の 検索要求に応じて動的に統合する方式を確立することを目指している.例えば,利用者が “データベースに関する論文”について調査したいと考えているとする.汎用の検索システムに “データベース”や “論文”などのキーワードを入力することによって,ある程度の情報を得ることができるが,十分な検 索結果を得ることは難しい.このとき,情報科学に関する論文を専門に検索することが可能な検索システムを利用した場合に,より必要な検索結果を得ることができると考えられる.ところが,WWW 上 には様々な種類のデータベースがあるため,どの検索システムが検索目的に適しているのかが分からないことが多い.したがって,本研究ではある問合せに対して適している検索システムを複数の検索システムから自動的に選び出すことによって,統合型情報検索システムを実現するための方式を確立することを目指す.
本研究の特徴は,検索精度だけを重点的に追求し,検索速度を重視しない点である.現在,多くの検 索システムでは,大量の検索対象に対して検索を行うために,検索速度を重視した手法であることが多 い.ところが,記憶装置の容量が増加していることや,計算機の速度が飛躍的に増加していることなど から,検索速度は今後大きな問題とはならないと思われる.一方,検索精度は計算機の速度や記憶装置 の容量と比例するわけではないため,精度の向上は最も重視すべき課題であると考えた.

## ３．研究の方法

まず,複数の検索システムを統合する際に必要と考えられる,スコアの正規化手法について,スコア統合のための関数についての課題に取り組んだ.まず,検索対象としてはテキストを考え，NTCIR や TREC など既に開催された情報検索に関するコンペティションに投稿された,多くの参加チー ムによる検索結果を統合することを考える.結果として,これら複数の検索結果を統合することによって,どのチームの検索結果よりも高い検索結果を統合システムによって得ることを目標としている.
予期される問題点として,性能面での問題である応答速度の低下,統合することによる検索精度の低下が考えられる.一般的な情報検索システムにおける処理時間に加えて,スコアの統合にかかる時間を考えなければならない.そこで,スコアの統合にかかる時間をできるだけ小さくするための解決策として,検索結果が利用者の検索意図に適合している上位 k 件の検索結果だけを利用した統合を行う手法を 考える.また,複数の検索結果を統合することによって必ずしも検索精度が向上しないという問題点を 解決するために,統合前の検索結果相互の重複度を計算する.これら二つの手法を提案することによって,予期される問題点を解決することが可能になると考えられる.
次に,多様な統合方法のうち,最も適した統合方法を，利用者の検索意図から推測する手法についての課題に取り組んだ.従来の検索システムと比較して検索精度,速度が共に良い検索システムを構築することを目標として いる.

予期される問題点として,テキスト検索システムにおける異種の検索システムから出力される検索結果を統合する際に,どのような統合の手法を用いたら良いかという問題点が考えられる.この問題点を解決するために,検索結果そのものから得られる特徴,検索結果の精度との相関関係 についての調査を行う.検索結果から得られる特徴として,スコアの分布や高スコアの含有率などが考えられる.
また,利用者の問合せをどのように検索システムに入力するかという点も問題点として予想された．テキスト検索システムへの入力はキーワードが主であったが,キーワードで画像を検索することは困難 であり,また必ずしも最適な入力方法であるとは限らない.そこで,これらの問題を解決するために,複数のサンプル文書を利用者に提示する手法を考えた.

４．研究成果

本研究では，複数の検索システムとして画像検索エンジンの統合を行った．それぞれ異なる特徴を利用して画像を検索することができるシステムを構築し，それらのシステムを統合することによって，異なる複数の特徴を統合したシステムを構築した.
ここで，それぞれのシステムを統合する際に.複数のシステムが検索対象に対して適合しているかどうかを表す数値であるスコアを統合することを考えた．ここで，数値を統合する方法として，最大値，最小値，p ノルムなどを含む 28 種類の関数を用意した.
スコアを統合する際に，スコアを統合可能となるように正規化を行わなければならない.そこで，一般的な正規化方法により 0，1 の間に正規化する方法を用いるとともに，より精度が高い正規化手法である，シャノンの情報量を利用した手法を開発した.
三つの検索システムおよび 28 種類の関数，および二つの正規化手法を比較することにより評価実験を行った．その結果，提案したスコア正規化手法によって確かに精度が向上することが分かった．このとき，精度は提案手法を用いたとき，従来の正規化手法と比較して 10% の精度向上が見られた.
スコアを統合するための関数においては，p ノルムを利用した方法が最も精度が高いことが分かった．このとき，他の手法と比較して最大で 15%の精度向上を確認することができた．特に，再現率が低いときの精度向上を確認することができた.
また，三つの検索システムすべてを利用した場合が，一つもしくは二つの検索システムだけを利用する場合と比較して精度が向上することが分かった.
以上の評価実験の結果，異なる分布をもつ複数の数値群を統合するための手法が確立した．一般に，複数の値を統合する際には最大値を 1 に，最小値を 0 にする正規化手法や，加算，乗算などの単純なスコア統合関数が利用されることが多いが，これらの正規化手法や統合関数について各場合において検討されることは少ない．本研究では，この部分について明確な指針を与えることが可能となったと考えている.
本研究では，検索エンジンの精度の観点から手法を確立したが，今後は検索の質も重要な観点となりつつあると考えられる．本研究で確立した正規化手法，およびスコア統合手法を利用することによって，検索された文書の質も測定することが可能となるような方法を考案する必要があると考えている.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

①Seint Seint Aye, Yu Suzuki and Kyoji Kawagoe: “Supporting Appropriate Communication Media Selection Thr ough Context-Awareness” International Journal of Hybrid Information Technology (IJHIT), ISSN: 1738 -9968, Vol.4, No.1,pp.79-88, Oct.,2008

② 大野 和久，鈴木 優，川越 恭二：「楽曲全体における特徴量の傾向に基づいた類似検索手法」，日本データベース学会論文誌，Vol. 7, No. 1, pp. 233-238, 2008

③ Kazuhisa Ono, Yu Suz uki, and Kyoji Kawagoe: “A Music Retrieval Method based on Di stribution of Feature Segments”, Proceedings of the Fourth IEEE International Workshop on Multimedia Information Processing and Retrieval (MIPR 2008), IEEE CS Press, pp. 613－618, Berkeley, CA, USA, December 2008

④ Yu Suzuki , Masahiro Mitsukawa, and Kyoji Kawagoe: “A Content-based Image Retrieval Method based o n TFIDF-based Weighting Scheme”, Proceedings of the 2nd International Workshop on Multimedia Data Mining (MDMM 2008) In conjunction with the 19th International Conference on Database and Expert Systems Applications (DEXA 2008), IEEE CS Press, pp. 112－116, Turin, Italy, September 2008

⑤ Yu Suzuki, Jun Ishizuka, and Kyoji Kawagoe: “ A Similarity Search of Trajectory Data using Textual Information Retrieval Techniques” , Proceedings of the 13th In ternational Conference on Database Systems for Advanced Applications (DASFAA 2008), Lecture Notes in Computer Science, pp. 627-634, New Delhi, India, March 2008.

〔学会発表〕(計 2 件)

鈴木 優, 吉川 正俊:「Wikipediaにおけるキーパーソン抽出による信頼度算出精度および速度の改善」, 第21回セマンティックウェブとオントロジー研究会 (第2回 Wikipedia ワークショップ) 研究報告, SIG-SWO-A901-01, 2009 年 11 月.

金本 径卓, 鈴木 優, 川越 恭二:「編集履歴に基づく Wikipediaにおける記事の信頼度算出手法」, 情報処理学会第 144 回データベースシステム研究会, pp. 31 - 38, 2008 年 1 月.

６．研究組織
(1)研究代表者
鈴木 優 (SUZUKI YU)
京都大学・大学院情報学研究科・特定研究員
研究者番号：40388111