様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　６月　１日現在

機関番号：６２６１５
研究種目：基盤研究（B）
研究期間：2009～2012
課題番号：２１３０００５８

研究課題名（和文）　データベースとウェブの連携による情報の獲得と利用に関する研究

研究課題名（英文）　A Study on Database and Web Linkage for Information Acquisition and Utilization

研究代表者
相澤　彰子（AIZAWA AKIKO）
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：90222447

研究成果の概要（和文）：本研究は、大規模なデータベースとウェブの連携による知識獲得のフレームワークの提案と実証を目的とする。研究期間内では、特に学術論文の著者同定問題に焦点を絞り、(1)ネットワーク構造を用いた情報同定手法の開発、(2)情報同定結果の要約と推薦システムへの適用手法の提案と実証、の２点について研究を進めた。そして、提案する情報同定手法を大規模な実問題に適用するとともに、得られた同定結果を利用した論文推薦システムを実装して、有用性を示した。

研究成果の概要（英文）：In this research, we investigated a new framework of information acquisition based on database and Web linkage. We specifically focused on the identification of authors of academic papers and developed (1) a network-based identification system, and also (2) a recommender system that utilizes the aggregated information. In our evaluation, we applied the proposed identification method to a large-scale identification problem and demonstrated the usefulness of the aggregated information through our implementation and a practice of a scientific paper recommender system.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	5,300,000	1,590,000	6,890,000
２０１０年度	4,000,000	1,200,000	5,200,000
２０１１年度	4,400,000	1,320,000	5,720,000
年度			
年度			
総　計	13,700,000	4,110,000	17,810,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：情報リンケージ、学術コンテンツ、データベース、著者同定

１．研究開始当初の背景

データベース分野においては、従前より、データベースの整合性チェックや異種データベース統合の目的で、重複レコードの検出に関する研究が行われてきた。その中心的な検討課題として、(1)大規模なレコード集合を対象とするための高速な同定処理の実現、(2)人手判定のコストを削減するための照合関数の高精度化などがあり、後者については、機械学習手法の適用が多く行われてきた。

従来の研究がデータベース・レコードどうしの同定を問題として想定しているのに対

して、申請者はこれまで、フォーマットを持たないテキストをデータベース・レコードに対応づけるための情報同定基盤について研究を進めてきた。そして、挑戦的萌芽研究「データベース照合に基づくテキスト・エンティティの同定に関する研究」（代表）をはじめとする研究の成果として、(a)テキスト解析処理、(b)高速あいまい検索、(c)同一性判定処理の各モジュールを構成要素とする「リンケージエンジン」を提案した。適用例として、国立情報学研究所が事業サービスする数千万件の書誌・図書データベースを用い、ウェブ文書等を対象としたオンラインリンケージエンジンの試作版を公開するとともに、数百万件～１千万件規模のテキスト入力に対する同定処理を実際に行い、事業サービス用データへのリンク埋め込み等により有効性を実証した。

ここで、これまでの研究は書誌や図書など一貫性が保証されたデータベース・レコードに焦点をあてていた。すなわち、同定対象となる書誌や図書には、一意に定まる識別子が付与され、共通の属性を持ち欠損値も少ないことが想定されていた。このため同定においては、識別子の対応がとれればよく、得られる情報を統合し提供するといった機能は必ずしも必要ではなかった。これに対して本研究では、必ずしもこれらの前提が成り立たない一般的な情報同定問題に焦点をあて、論文著者の同定を中心に、大規模な同定と情報要約のための手法を検討する。

２．研究の目的

本研究の目的は、大規模なデータベースとウェブの連携による知識獲得のフレームワークの提案と実証である。異種データベース間で共通するレコードや、テキスト中で特定のレコードを指示する記述を同定し、その結果を手がかりとして、共起する属性値どうしをさらに同定したり、周辺に出現する文脈を単語ベクトルとしてあいまい性解消に利用したりする手法を検討する。研究期間内では、以下の 2 つの目標を設定して研究を進めた。

(1)ネットワーク構造を用いた情報同定手法の開発

注目するレコードやテキスト中の「エンティティ」（＝特定のレコードを指示する記述、同定対象となる情報）をノードとして、同じであると判定されたノードどうしをリンクでつないだ大規模なネットワークの構築手法を検討する。

(2)情報同定結果の要約と推薦システムへの適用手法の提案と実証

情報同定の結果として得られるネットワーク表現を利用した情報統合・提示法を検討するとともに、情報推薦システムへの適用を研究・実証する。

３．研究の方法

本研究では、上記で述べたように、同一の事物を参照すると判定された異なるデータベース上のレコードやテキスト中の記述どうしを同一指示リンクで結んだものを「情報同定ネットワーク」と呼び、その構成法および利用法を検討した。

(1)ネットワーク構造を用いた情報同定手法の開発

本研究で提案した情報同定手法の流れを図１に示す。提案手法で用いるネットワーク表現は単純だが大規模データベース上の数十～数千万規模のノードを扱うことが可能で、以下の特徴を備えている。

◆入力の多様性：リンケージエンジンによって、ウェブや一般の文書、他のデータベース等に存在する多様な情報を入力として利用し、中核とするデータベースの ID 集合に対応付けることができる。

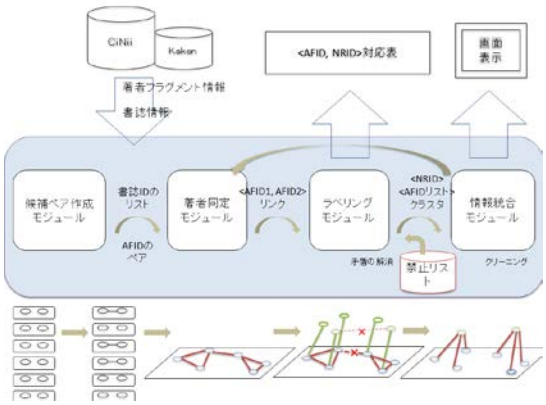

図 1　提案手法における処理の流れ

◆学習機能：同定結果をリンクの形で追加保存することで、表記揺れや別表記を知識としてネットワーク内に蓄積し、逐次的に同定性能を高めることができる。

◆適応的な計算コスト配分：外部からの入力をトリガーとして処理を行うことで、利用頻度が高い部分に優先的に同一性判定処理のための計算コストを割り当てられる。

(2)情報同定結果の要約と推薦システムへの適用手法の提案と実証

著者情報の同定結果に基づく共著関係の解析、著者情報をプロファイルとする論文推薦、

それをサポートするための論文間の関係解析やマイニング技術などの要素技術について研究を進め、プロトタイプの論文推薦システムを構築・評価した。

４．研究成果

(1)ネットワーク構造を用いた情報同定手法の開発

本研究では､論文著者の同定問題を対象として手法の開発および評価を進めた。情報同定の処理を、(a)同定ペア候補抽出、(b)機械学習手法による候補ペアの同一性判定、(c)獲得されるネットワーク構造上でのクラスタリングおよび統合によるクリーニング、の 3 つのステップにまとめ、各々について手法を検討した。

①同定ペア候補抽出

情報同定ネットワークでは、同定のための計算コストの配分のために、外部から類似した書誌のリストをトリガーとして受け取り、同一リンクの有無を判定する。トリガーとしては多様な入力を想定することが可能であり、具体的には、科研費報告書の成果リスト、編集距離に基づきタイトルが互いに類似した論文どうし、３名以上の共著者が共通している論文リスト、引用・被引用関係にある論文ペア、氏名や所属がほとんど一致する著者、ウェブ上の発表文献をトリガーとして実装した。

特にウェブ上の情報について、書誌・図書リンケージエンジンを利用すると、明示的な構造を持たないウェブのテキスト文書を入力として、指示先となっているデータベースレコードの集合を得ることができる。そこで、研究者のウェブページを検索して同定用の資源として活用する仕組みを実現するため、(i) URL で指定されるテキスト文書全体をクエリ入力とするデータベースの検索、および (ii) データベース・レコードから自動生成したクエリに基づくウェブ検索の手法を実装した。

②機械学習手法による候補ペアの同一性判定

現実の同定処理では、著者 A と著者 B、著者 B と著者 C が同じであるときに著者 A と著者 C は同じであるか(推移律)の判断が大きな問題となるが、同定候補ペアに機械学習を適用して得られるネットワーク表現上でクラスタリングを適用することで、あいまい性の解消を行う手法を提案して有効性を検証した。また、作業の効率から正例に偏りがちな人手判定正解データに対して疑似的に負例を追加することで、性能が大幅に改善することを示した。

③獲得されるネットワーク構造上でのクラスタリングおよび統合によるクリーニング

機械学習手法による判定では､誤りが一定の確率で存在することは避けられない。そこで、同定ネットワークに対してグラフアルゴリズムを適用することにより、大局的にみたリンクの確からしさを判定する手法を提案し、正規化相互情報量（NMI）で測定したクラスタリング性能が向上することを示した。さらに、同定結果に基づき得られるクラスタのばらつき具合をチェックすることで、過剰な統合を防ぐ手法を実装した。

上記により､所属の変遷や表記揺れなどからデータベース中でばらばらに存在していた論文著者の情報を、比較的高い精度で同定することが可能になった。実証のため、大規模な論文データベースの論文の著者を対象として、著者が同一人物であるかの判定を行うエンジンの構成を設計し、プロトタイプを試作してのべ数約８千万人の著者の同定を行った。

また、これらの実装を通して、信頼性の高い情報源からの情報と低い情報源からの情報を区別する重要性が判明した。図２に示すような異なる可能性を検討した結果、識別子の間に優先順位を持たせる手法を採用し、さらに人手によるデータ修正結果を柔軟に取り込めるよう実装を工夫した。

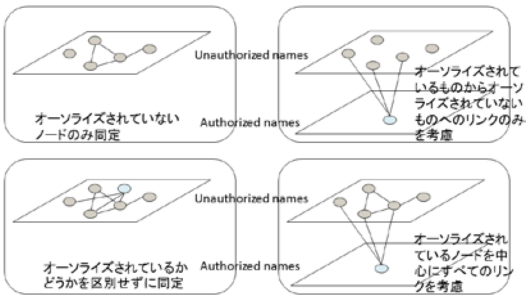

図 2　識別子の階層化

(2)情報同定結果の要約と推薦システムへの適用手法の提案と実証

まず､機械学習で判断が困難である候補ペアについて、関連文書の内容(抄録・本文)解析による同一性判定を行うため、情報距離に基づく類似度計算法を新たに提案した。具体的には Ziv-Merhav crossparsing に競合的な N グラム選択の仕組みを導入して、高速に著者同定を行う手法を提案し、論文の抄録を用いて有用性を示した。

また、著者同定の結果を用いて、研究者ごとに共著関係、引用・被引用文献、投稿学会、

専門用語等の情報を集約するとともに、これを利用した情報推薦の手法について検討した。特に Content-based Filtering（内容に基づく推薦手法）と呼ばれる手法に焦点をあてて検討を進めるとともに、評価のための実証基盤の開発に取り組んだ。ユーザの多様な検索要求に対応するため、多視点情報推薦システムを設計・構築し（図３）、有効性を実証的に確認した。

さらに、意味的な関連性に基づく論文の推薦を可能にするため、自然言語処理技術を用いて論文の本文を解析し、引用文脈から情報を自動獲得する手法の研究に取り組み成果発表を行った。

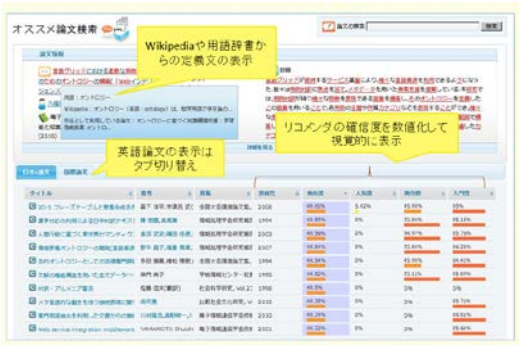

図 3　多視点論文推薦システム

本研究で試作した著者同定のプロトタイプシステムは、国立情報学研究所の学術情報サービス（http://ci.nii.ac.jp/）の著者検索機能として改めて実装され、実サービスとして運用されている。

最後に、本研究で開発した論文および著者の同定技術は、研究者の個人ページ、研究プロジェクトのページ、研究ポータル、機関レポジトリが発信する論文や個人業績、ＯＣＲで読み込まれた書籍など、様々なウェブ上の資源に適用可能で、学術的なコンテンツに関する情報統合の基盤技術となることが期待される。

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 3 件）

① 乾伸雄，相澤彰子：“SAT ソルバーを用いた最小無矛盾 DFA の生成” 人工知能学会論文誌 27(3)，151-162 (2012)査読有

② 相澤彰子：“多クラス文書分類問題における Ziv-Merhav Crossparsing の適用と評価”情報処理学会論文誌，52(10)，2953-2964(2011)査読有

③ 金城敬太，相澤彰子，尾崎知伸：“調査データに基づく社会構造変化の抽出” 人工知能学会論文誌 25(3)，452-463 (2010) 査読有

〔学会発表〕（計 24 件）

① Akihiro Kameda, Kiyoko Uchiyama, Hideaki Takeda, Akiko Aizawa: “Extraction of Semantic Relationships from Academic Papers using Syntactic Patterns” The Fifth International Conference on Information, Process, and Knowledge Management (eKNOW 2013) -Best Paper Award (20130224). Nice, France

② Yuan Li, Akihiro Kameda, Kiyoko Uchiyama and Akiko Aizawa: “A Method for Corresponding Paragraphs with Sentences in Academic Paper’s Abstract” 第11回情報科学技術フォーラム(FIT2012) (20120904) 法政大学小金井キャンパス（小金井）

③ Pannawit Samatthiyadikun, Atsuhiro Takasu, Saranya Maneeroj: “Multicriteria Collaborative Filtering by Bayesian Model-based User Profiling” 13th International Conference on Information Reuse and Integration (IRI 2012) (20120808) Las Vegas, USA

④ Panot Chaimongkol, Pontus Stenetorp, Akiko Aizawa: “Utilising Bilingual Lexical Resources for Technical Term Extraction” 第 26 回人工知能学会全国大会 (JSAI 2012) International Organized Session (20120614)山口県教育会館（山口）

⑤ 亀田尭宙，内山清子，武田英明，相澤彰子：“構文パターンを用いた論文の引用文脈からの関係情報抽出” 第 26 回人工知能学会全国大会（JSAI 2012）(20120615) 山口県教育会館（山口）

⑥ 亀田尭宙，内山清子，宮尾祐介，武田英明，相澤彰子：“論文中の引用文における構文パターンを用いた論文・概念間の関係抽出” 人工知能学会知識ベースシステム研究会（20111215）慶応義塾大学日吉キャンパス（横浜）

⑦ Takafumi Suzuki, Ryota Tomisaka, Kiyoko Uchiyama, Akiko Aizawa: “Analyzing the Characteristics of Academic Paper Categories by Using an Index of Representativeness” Pacific Asia Conference on Language, Information and Computation (20111216) Singapore

⑧ Atsuhiro Takasu, Saranya Maneeroj: “A Recommendation Algorithm Using Positive and Negative Latent Models” IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2011) (20111211) Vancouver, Canada

⑨ Takafumi Suzuki, Shin Hasegawa, Takayuki Hamamoto, Akiko Aizawa: “Document Recommendation Using Data Compression”, Pacific Association For Computational Linguistics (PACLING 2011) (20110719) Kuala Lumpur, Malaysia

⑩ 亀田尭宙, 武田英明, 相澤彰子: “関連研究に関する記述の分析による論文間の意味的関係の抽出” 2011 年度 人工知能学会全国大会 (第 25 回) (20110601) アイーナ岩手 (盛岡)

⑪ 内山清子, 高須淳宏, 相澤彰子, 難波英嗣, 宮尾祐介: “オススメ論文検索システム:OSUSUME” 2011 年度 人工知能学会全国大会 (第 25 回) (20110601) アイーナ岩手 (盛岡)

⑫ Atsuhiro Takasu, Saranya Maneeroj: “A Recommendation Algorithm Using Positive and Negative latent Models” IEEE Symposium on Computational Intelligence and Data Mining (20110411) Paris, France

⑬ Pakapon Tangphoklang, Saranya Maneeroj, Atsuhiro Takasu: “A Novel Weighting Scheme for a Multi-Criteria Rating Recommender System” IADIS International Conference on Information Systems (IS2011) -Best Applied Research (20110311) Avila, Spain

⑭ 冨坂亮太, 鈴木崇文, 相澤彰子: “話題推薦システムのためのモデル構築手法” 情報処理学会第 73 回全国大会 (20110302) 東京工業大学(東京)

⑮ 冨坂亮太, 鈴木崇文, 相澤彰子: “発話を意識した文推薦システムの構築と評価” 情報処理学会研究報告.情報学基礎研究会報告 (20110121)NHK 技研(東京)

⑯ Kiyoko Uchiyama, Akiko Aizawa, Hidetsugu Nanba, Takeshi Sagara: “OSUSUME : cross-lingual recommender system for research papers” Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation (CaRR 2011) (20110213) Palo Alto, USA

⑰ 相澤彰子: “情報検索における圧縮距離の適用に関する考察” 情報処理学会研究報告.自然言語処理研究会報告 (20101118) 広島市立大(広島)

⑱ Atsuhiro Takasu: “Cross-lingual keyword recommendation using latent topics” International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2010) (20100926) Barcelona, Spain

⑲ 金城敬太, 相澤彰子, 市瀬龍太郎, 小暮厚之: “稀な事象同士の関連性指標～異常値間の関連性抽出のための時系列データマイニング” 2010 年度人工知能学会全国大会(第 24 回) (20100611) 長崎ブリックホール (長崎)

⑳ 長谷川新, 相澤彰子, 浜本隆之: “パーソナライゼーションにおけるトピックを意識しない類似度測定” 人工知能学会 2009 年全国大会(第 23 回) (20090619) サンポートホール高松(高松)

㉑ 相澤彰子, 宮田淳平: “参照記述の表記揺れ同定問題に対するアプローチ” 第 25 回ファジィシステムシンポジウム (FSS2009) (20090715) 筑波大学(つくば)

㉒ 中村智洋, 相澤彰子, 馬場康維: “論文データベースに見る統計分野の研究動向” 2009 年度統計関連学会連合大会 (20090909) 同志社大学(京田辺)

㉓ 港真人, 相澤彰子: “名前同定のための SVM 特徴素の抽出と適用” 情報処理学会創立50周年記念全国大会 (20100308) 東京大学(東京)

㉔ 長谷川新, 相澤彰子, 浜本隆之: “視線情報を用いたユーザプロファイル獲得と文書推薦における有用性” 電子情報通信学会 2010 年総合大会 (20100318) 東北大学(仙台)

６．研究組織

(1)研究代表者

相澤　彰子 (AIZAWA AKIKO)

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：90222447

(2)連携研究者

高須　淳宏 (TAKASU ATSUHIRO)

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：90216648