様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １７ 日現在

機関番号：２５４０３
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３３０２６３

研究課題名（和文）新情報の追加と書き換え技術を用いたサーベイ論文の自動作成

研究課題名（英文）Automatic Generation of Survey Articles Based on Update Summarization and Text Revision Techniques

研究代表者

難波　英嗣（NANBA, HIDETSUGU）

広島市立大学・情報科学研究科・准教授

研究者番号：５０３４５３７８

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：引用論文データベースから，(1)研究者が過去に執筆したサーベイ論文を自動検出し，(2)そのサーベイ論文に含まれていない新しい論文を検索し，(3)それらの新論文の情報を，サーベイ論文の文脈に応じて書き換えて追加することにより，最新の研究動向を含んだサーベイ論文を作成する．引用論文データベースCiteSeerを用いた実験を行い，提案手法の有効性を確認した．

研究成果の概要（英文）：We propose a method of generating a survey article. Our method consists of the three steps: (1) detecting survey articles in a research paper database, (2) retrieving research papers that should be contained in the survey articles, which were detected in Step (1), and (3) updating the survey articles based on the research papers retrieved in Step (2). We conducted several experiments using a citation index "CiteSeer", and confirmed the effectiveness of our method.

研究分野：自然言語処理

キーワード：テキスト要約　学術論文　サーベイ論文　SNS

１．研究開始当初の背景

研究者数の増加，学問分野の専門分化と共に学術情報量が爆発的に増加している今日，研究者が入手できる論文の量も増える一方で，人間の処理能力の限界から，入手した論文全てに目を通し利用することが益々困難になっている．

このような状況にあって，特定の研究分野に関連したサーベイ論文の必要性は高まる一方である．例えば医学分野ではコクラン共同計画のもと，サーベイ論文の作成と更新が定期的に行われている．しかし，それ以外の研究領域では同様の仕組みが整備されていないため，研究者が知りたい分野のサーベイ論文を見つけても，その論文が何年も前に執筆されていて，最新の研究動向を含んでいなかったり，そもそも当該分野のサーベイ論文自体が存在しなかったりすることもある．

このような状況を改善するため，これまでに，研究者が行うサーベイ論文の執筆を支援するシステムを開発する研究や[Nanba 2005, 2010]，サーベイ論文そのものを複数の論文から自動生成する研究[Mohammad 2009]が行われてきている．本研究では，論文間の引用関係に着目し，引用論文データベースからサーベイ論文を自動的に検出する手法を提案している[Nanba 2005]．この手法により，もし研究者が必要とする分野のサーベイ論文が検出されれば，その分野の効率的なサーベイが可能となるが，上述のとおり，検出されたサーベイが最新の研究動向を含んでいるとは限らない．

本研究では，この他に，国立情報学研究所が主催する評価ワークショップ NTCIR において論文と特許から技術動向マップを自動作成するプロジェクトを企画，実施している．このプロジェクトの最終目標は，技術分野ごとに，論文と特許を，「要素技術」と「その効果」という観点で分類することで技術動向マップを作成するというものであり，国内外の企業および大学などの学術研究機関総計 15 団体が参加し，この分野の裾野を広げるなど一定の成果を上げている．しかしながら，サーベイ論文は「要素技術」と「その効果」だけでなく，この他にも様々な観点から研究動向をまとめられることがあるため，このプロジェクトの成果のみで十分なサーベイ支援が行えるわけではない．

Mohammad らは，論文間の引用関係などを利用して，サーベイ論文の自動作成を試みている[Mohammad 2009]．この手法は，要約対象となる複数のテキストから重要文を抽出し，抽出された文間の結束性を考慮して並べることにより要約作成を行う，一般的な複数テキスト要約の手法に基づくものである．しかしながら，現状では数文程度の非常に短い分野の概要を生成する程度にとどまっている．

参考文献

[Mohammad 2009] Mohammad, S. et al. (2009) “Generating Surveys of Scientific Paradigms”. In Proceedings of HLT-NAACL 2009.

[Nanba 2010] Nanba, H. et al. (2010) “Overview of the Patent Mining Task at the NTCIR-8 Workshop”. In Proceedings of the 8th NTCIR Workshop Meeting, pp.293-302.

[Nanba 2005] Nanba, H. et al. (2005) “Automatic Detection of Survey Articles”. In Proceedings of ECDL 2005, pp.391-401.

２．研究の目的

本研究では，過去の研究成果[Nanba 2005]を用いて，引用論文データベースから研究者が過去に執筆したサーベイ論文を自動検出し，そのサーベイ論文に含まれていない研究を追加することにより，最新の研究動向を含んだサーベイ論文の作成を目指す．また，実験により，その有効性を検証する．

図１にサーベイ論文自動作成の手順を示す．図中の各ステップの処理内容について，以下に説明する．

- **入力**

「テキスト要約」や「統計的機械翻訳」など研究分野の名称をサーベイ論文自動作成システムの入力とする．

- **手順(1)サーベイ論文検出**

過去の研究成果[Nanba 2005]を用い，論文データベースから入力されたキーワードに関するサーベイ論文を検出する．

- **手順(2)追加すべき最新論文の検索**

手順(1)で検出されたサーベイ論文に追加すべき論文を，論文データベースから検索する．このステップでは，まず，手順(1)で検出されたサーベイ論文を章ごとに分割し，次に，各章の文脈を解析し，文脈に適合しており，かつその文脈では言及されていない新しい論文を論文データベースから検索する．ここで，サーベイ論文に追加すべきかどうかの判定は，単に新しいだけでなく，当該分野内でのその論文の重要性も考慮した上で行う．論文の重要性の評価には，申請者の過去の研究成果[Nanba 2010]を用いた手法と Twitter を用いた手法を提案する．また，一般的な情報推薦手法との比較により提案手法の有効性を確認する．

- **手順(3)既存サーベイ論文への最新論文の追加**

手順(2)で検索された新しい論文から追加すべき文を抽出し，それらをサーベイ論文の文脈に応じて書き換えて既存のサーベイ論文に追加することにより，最新の動向情報を含んだサーベイ論文を作成する．また，同一分野で異なる時期に執筆されたサーベイ論文に本提案手法を適用し，どの程度過

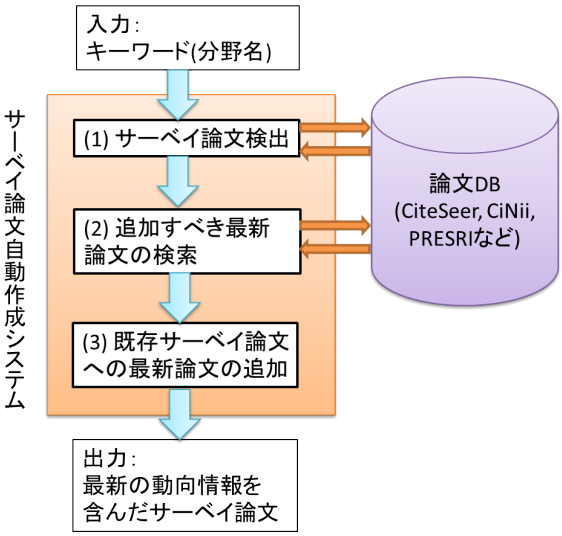

図1 サーベイ論文自動作成手順

去に執筆されたサーベイ論文まで適用可能であるか明らかにする．

● **出力**

最新の動向情報を含んだサーベイ論文を出力する．

3．研究の方法

本研究では，次の 4 つの手順でサーベイ論文作成システムを構築する．

[課題 1] サーベイ論文検出
[課題 2] 追加すべき最新論文の検索
　(2-1) サーベイ論文の文脈に適合する最新論文の検索
　(2-2) 検索された最新論文の重要度の評価
[課題 3] 既存サーベイ論文への最新論文の追加

以下に，各課題について述べる．

[課題 1] サーベイ論文検出

論文データベースから入力されたキーワードに関するサーベイ論文を検出する．このステップでは，過去の研究成果[Nanba 2005]を用いる．

[課題 2-1] 追加すべき最新論文の検索（サーベイ論文の文脈に適合する最新論文の検索）

課題 1 で検出されたサーベイ論文に追加すべき論文を，論文データベースから検索する．例えば，サーベイ論文に以下の文が記載されているとする．

> (例 1) 隠れマルコフモデル(HMM)[1]やサポートベクトルマシン(SVM)[2]などの機械学習を用いた形態素解析手法が提案されている．

この文脈に追加すべき論文は，(1)「形態素解析」という研究課題が一致しており，なおかつ(2)要素技術として機械学習を使っている，という 2 点を満たしていることが条件となる．このうち(1)に関しては，同一分野の論文を探せばよいので，従来の tf*idf などを用いた 2 文書間の内容の類似度を測る尺度が利用できる．一方，(2)に関しては，従来手法をそのまま適用することはできない．なぜならば，この文脈で追加したい論文は，HMM や SVM に関するものではなく，例えば条件付き確率場(CRF)などの新しい機械学習手法を使ったものだからである．このためには，この文脈には「機械学習」と「形態素解析」という 2 つのトピックが含まれていることを自動認識し，それに応じた論文を検索する必要がある．そこで，既存のサーベイ論文で言及されている論文集合と，追加候補論文が引用する論文の関係に着目し，引用関係と論文中の内容語を素性としたランキング学習に基づいた論文検索手法を検討する．

[課題 2-2] 追加すべき最新論文の検索（検索された最新論文の重要度の評価）

課題 2-1 で検索された論文の重要度を評価し，サーベイ論文に追加すべき論文を決定する．発表されてある程度年月が経過した論文については，被引用数等を用いて論文の重要度を評価することが可能であるが，発表されて間もない論文については，被引用数に基づく手法が利用できない．そこで，Twitter を利用した新たな論文の重要度評価手法について検討する．近年，学会等でそのイベントごとにハッシュタグを取り決め，研究発表に対するコメントを Twitter に投稿する研究者も少なくない．また，興味深い発表については，数多くの聴衆がツイートないしリツイートする傾向にある．そこで，Twitter のハッシュタグに着目して，特定の会議のツイートを収集，論文と対応付けを行い，ツイート数の多い論文を重要と考える手法についても同時に検討する．

[課題 3] 既存サーベイ論文への最新論文の追加

課題 3 では，課題 2 で検出された論文を用い，既存サーベイ論文に新情報を追加することで新たなサーベイ論文を生成する手法について検討する．

4．研究成果

[課題 2-1] 追加すべき最新論文の検索（サーベイ論文の文脈に適合する最新論文の検索）

サーベイ論文または専門書籍の特定のトピックに関する一部（節、章）s と，そこで言及されている論文集合 D をシステムの入力とし，s に追加すべき論文 p を検出する，という課題について考える．一般に，論文

集合 D 中の論文 d との共引用数の多い論文は，s との関連性が高く、なおかつ重要性も高いと考えられる．また，ある論文に関する引用個所は，他の研究者がその論文に見出した関連性や新規性などの注目すべき点を示しており，引用箇所間の類似性が高ければ，対応する論文対の類似性，関連性が高いと考えられる．

以上の仮定に基づき，引用論文データベース CiteSeer 約 200 万論文を用いて提案手法の有効性を確認するための実験を行った．また，比較手法として，s との類似度の高い論文をサーベイ論文に追加すべき候補の論文として出力する手法を含む，いくつかの手法でも実験した．実験の結果，提案手法は比較手法よりも高い検出精度が得られることが分かった．

[課題 2-2] 追加すべき最新論文の検索（検索された最新論文の重要度の評価)

本研究は 3 つのステップ, (1) 有益なツイートの自動分類，(2) ツイートと論文との自動対応付け，(3) 有益なツイートに基づく論文の重要度の評価から構成される．提案手法の有効性を確認するため実験を行った．実験の結果，ツイートの自動分類では，再現率 0.591，精度 0.588 を，自動対応付けでは, 再現率 0.483, 精度 0.525 を，論文の重要度の評価では，ε 値 0.236 を，それぞれ得た．

[課題 3] 既存サーベイ論文への最新論文の追加

情報分野の専門書籍 4 冊の(1)旧版一章と(2)対応する新版の章で追加された論文をシステムの入力とし，新版の章で新たに記載された個所を正解要約と考え，提案システムで正解要約をどの程度再現できるかにより評価を行った．提案システムでは，「(2)対応する新版の章で追加された論文」の概要を使った場合，概要と引用個所を使った場合，概要と引用箇所と本文を使った場合の 3 種類で実験を行った．実験では 36 個の正解要約を対象に ROUGE を用いて評価を行った．実験の結果，概要と引用個所と本文を使った場合に，もっとも高い ROUGE 値が得られた．

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 1 件）

Iinuma, S., Fukuda, S., Nanba, H., and Takezawa, T.. (2015) “Evaluation of the Industrial and Social Impacts of Academic Research Using Patents and News Articles” The International Association for Computers & Information Science (ACIS) International Journal of Computer & Information Science, Vol.16, No.1, 12-21.

〔学会発表〕（計 6 件）

Fukuda, S., Nakahashi, H., Nanba, H., and Takezawa, T. (2015) “Quick Evaluation of Research Impacts at Conferences using SNS”. In Proceedings of the 12th International Workshop on Text-based Information Retrieval, in conjunction with DEXA 2015 .

中橋光, 難波英嗣, 竹澤寿幸. (2015) “SNS を用いた迅速な論文の重要度の評価” 第 7 回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2015).

飯沼俊平, 難波英嗣, 竹澤寿幸. (2015) “サーベイ論文作成支援のための引用論文推薦” 第 7 回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2015).

飯沼俊平, 福田悟志, 難波英嗣, 竹澤寿幸. (2014) “ニュース記事と特許を利用した科学技術の重要性の評価” 人工知能学会全国大会(第 28 回).

飯沼俊平, 難波英嗣, 竹澤寿幸. (2014) “新情報の追加によるサーベイ論文の作成支援” 言語処理学会 第 20 回年次大会.

中橋光, 難波英嗣, 竹澤寿幸, 高須淳宏. (2013) “Twitter と論文との自動対応付け” 日本データベース学会 第 4 回ソーシャルコンピューティングシンポジウム講演論文集 (SoC2013)

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
難波 英嗣 （NANBA HIDETSUGU）
広島市立大学・大学院情報科学研究科・准教授
研究者番号：50345378

(2)研究分担者
(3)連携研究者