研究初心者のための論文サーベイ支援システムの試作

加藤 健太 †　辻野 友孝 †　清水 堅 †　高崎 隼 †
白松 俊 †　大囿 忠親 †　新谷 虎松 †
名古屋工業大学大学院工学研究科情報工学専攻 †

1 はじめに

研究者が研究をする際，既存の研究について論文サーベイを行う必要がある．論文サーベイを行うことで自分の研究に深く関係する分野について調べることができ，どのような手法を用いれば自分の研究の目的を達せられるか，自分の研究が当該分野においてどのような位置に位置づけられるのか，また，他の研究との差異を知ることで自分の研究の強みを知ることができる．以上のことから，研究者が研究をする上で，既存の研究について論文をサーベイすることは重要である．

しかし，研究活動を始めたばかりの大学生のような研究初心者は，専門的な知識が少ないため論文を検索するためのキーワードが分からず，興味ある論文を見つけることが困難である．また，論文中に気になる用語がある場合，同じ用語を使用している別の論文を閲覧することにより，その用語に対してより多くの情報を得ることができ，理解を深めることが期待できるが，研究初心者は論文中のどの用語が専門的な用語なのか分からない．よって，研究初心者には，論文サーベイ時に，論文中の専門的な用語の特定や，それらの用語を使用してる他の論文を閲覧できるような支援が必要である．

本稿では，ユーザが論文を検索する際の支援に主な焦点を当て，研究初心者に対して論文サーベイ支援を行う．具体的には，ユーザが閲覧する論文中の文献を引用している箇所の情報を解析し，同じ手法，既存のシステムなどの要素技術[1]を扱っている他の論文へのリンクを可能にする．これにより，閲覧する論文中の要素技術を次々と調べることが可能となる．つまり，ユーザが気になる専門的な手法やシステムなどを芋づる式に調べることが可能となるため，研究初心者の論文サーベイの効率向上が期待できる．引用情報をもとに，同じ要素技術を用いている論文を提示するためには多量の論文データが必要になる．そのため，本稿では，過去の情報処理学会全国大会論文集を用いる．

2 関連研究

既存の論文サーベイ支援システムでは，適切な論文を検索するため，シソーラスを構築しクエリ拡張を行う研究や，収集した論文を分類する研究など様々な支援システムが存在する．本稿でも論文検索の支援に焦点を当てるが，文献の引用情報の利用，専門的な用語の特定を行う点でそれらの研究とは異なる．論文の引用文献を利用する研究では，難波ら[1]が多くの研究成果を残している．難波らは，どのような理由で文献を引用しているかという引用情報の理由を考慮し，論文間の類似度を計算，関連用語の自動収集，研究動向情報の抽出などを行っている．文献の引用情報を利用する点では類似しているが，本稿では，引用箇所に出現する要素技術を特定し，更に同じ要素技術を扱う論文へのリンクを生成し，論文検索の支援を行う点で難波らの研究とは異なっている．また，研究支援として本研究室では，鈴木ら[2]がユーザモデルを持たない研究初心者に対して，対話的ユーザモデルを構築し，興味ある論文の推薦を行っている．本稿で試作する検索支援機能と，鈴木らの論文推薦機能を組み合わせることで，より研究初心者に特化した研究支援が期待できる．

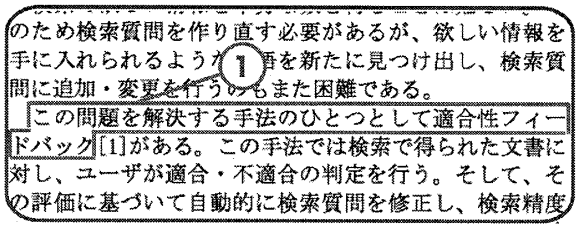
のため検索質問を作り直す必要があるが、欲しい情報を手に入れられるようか ① 語を新たに見つけ出し、検索質問に追加・変更を行うのもまた困難である。
この問題を解決する手法のひとつとして適合性フィードバック[1]がある。この手法では検索で得られた文書に対し、ユーザが適合・不適合の判定を行う。そして、その評価に基づいて自動的に検索質問を修正し、検索精度

図1: 本稿で扱う引用情報の例

3 論文サーベイ支援システム

3.1 論文データの詳細

本稿では，情報処理学会第71回全国大会論文集を解析対象の論文データとして用いる．論文集には約1,200の論文がPDF形式で収録されている．1つ1つの論文には，タイトル，著者名とその所属，本文，参考文献が含まれており，キーワードとアブストラクトは含まれていない．本稿では，本文と参考文献を主に利用する．文献には，引用文献と参考文献が存在する．引用文献とは，自分の著作と非常に関連があり，その文献中の文言を引用した文献である．参考文献とは、自分の著作を書く上で漠然と参考になった文献である．本稿で扱う論文はA4サイズ2ページで構成されており，引用文献が主であると仮定する．実際，多くの論文が，関連性が低い文献は引用していないのが一般的であった．よって，本稿では，「参考文献」と記載されいる場合でも，関連性が高い引用文献とみなしシステムを試作する．また，本システムでは，PDF形式の論文からテキストを抽出する必要がある．PDF形式からのテキスト抽出のため，本稿では，Xpdfパッケージのツール pdftotext[1]を用いた．

3.2 手法

以下に，本システムの手法を示す．まず，抽出したテキスト形式の全論文データを対象として，同じ文献を引用している論文同士を1つのグループとする．これは，同じ文献を引用している論文中の引用情報には，同じ要素技術が出現する可能性が高いという考えに基づき，同じ要素技術を特定し，その要素技術に引用文献へのリンクの貼り付けを行っていくためである．本稿では，引用箇所の前の文章の句点から引用のマークまでの文章を引用情報として扱う．取得する引用情報の例を図1に示す．図1の例では枠①の中の文章，つまり，「この問題」～「フィードバック」が扱う引用情報となる．本稿では，ある文献 d の引用情報に出現する用語 w は，$w \in citext(d)$ のように表す．被引用文献 $D=\{d_1,d_2,d_3,...\}$．引用文献をもとに各論文をグループにまとめる際，1つの論文が複数のグループに属しても構わない．各グループのそれぞ

Prototyping survey support system for research beginner

Kenta KATO, Tomotaka TSUJINO, and Ken SHIMIZU, and Jun TAKASAKI, and Syun SHIRAMATSU, and Tadachika OZONO, and Toramatsu SHINTANI

Dept. of Computer Science and Engineering, Graduate School of Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, 466-8555 JAPAN

[1] http://www.foolabs.com/xpdf/download.html

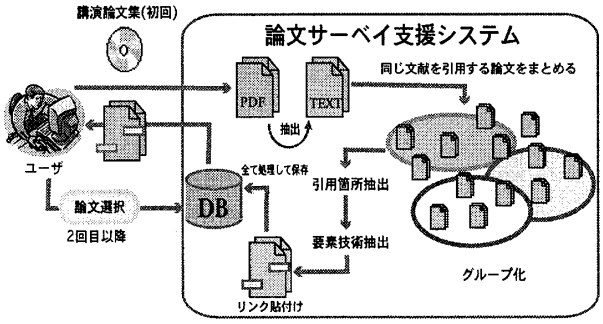

図 2: システムの概要

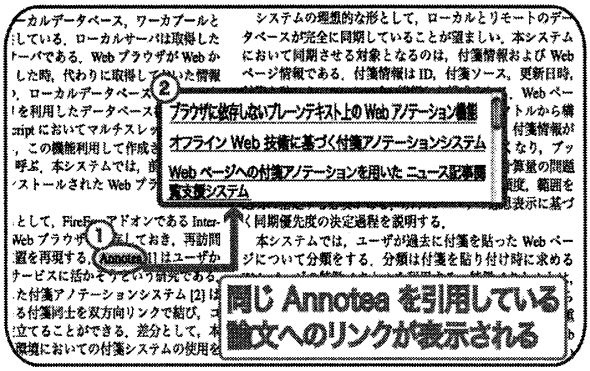

図 3: 本システムの実行例

れの論文から引用情報を取得し，それぞれのグループの引用情報の集合 $\{citext(d)\}$ 内で出現する用語 w ごとの頻度 r を計算する．用語 W は，形態素の N-$gram$ である．すなわち，用語 $W=\{w_1,w_2,w_3,...\}$ のように形態素解析と N-$gram$ を組み合わせた用語となる．尚，形態素解析には，MeCab[2]を用いる．r が閾値 θ_r 以上の場合，用語 w を求める要素技術の候補とする．要素技術の候補が存在した場合，$p(w|citext(d))$ を求める．$p(w|citext(d))$ は式 (1) のように表す．

$$p(w|citext(d)) = \frac{\{citext(d)\} \text{ 中の } w \text{ の出現回数}}{\{citext(d)\} \text{ 中の } N-gram \text{ の総数}} \quad (1)$$

次に対象とする全論文を形態素解析し，$p(w)$ を求める．$p(w)$ は式 (2) のように表す．

$$p(w) = \frac{\text{全論文中の } w \text{ の出現回数}}{\text{全論文中の全ての } N-gram \text{ の総数}} \quad (2)$$

最後に，式 (1)，(2) を用いて以下の値を計算する．

$$\frac{p(w|citext(d))}{p(w)} \geq \theta \quad (3)$$

条件式 (3) を満たす場合，用語 w を要素技術とする．閾値 θ より低い場合，もう一度適切な用語 w の候補を求めに戻る．これにより，要素技術に適さない一般的な用語が取り除かれる．そしてその要素技術に，同じ要素技術を扱う論文，すなわち，同じグループ内の全ての論文へのリンクを貼る．

3.3 システム概要

図 2 に本システムの概要を示す．初めに，ユーザがシステムに対し講演論文集を入力として与える．システムは，与えられた PDF 形式の全論文データからテキスト形式の論文データへと変換する．テキスト化された論文データの解析を行い，同じ文献を引用している論文を発見しそれらをグループ化する．次に，各グループにおけるそれぞれの論文から，該当する引用情報の抽出を行う．抽出した引用情報の集合から要素技術の候補を，閾値を用い決定する．図 1 の場合，「適合性フィードバック」という要素技術を取得することが望ましい．そして，論文中の求めた要素技術の箇所に，同じグループに含まれる各論文へのリンクを貼る．同じグループの論文全てに対して同様の処理を行う．最後に新たにリンクを生成した上記の論文をデータベースへ保存する．以降，データベースに保存されたリンク付きの論文データをユーザへの出力とする．その際，講演論文集に含まれていた各論文のインデックスページも共に出力する．インデックスページには各論文のタイトルがカテゴリ別に記載されている．本稿では，このインデックスページと生成したリンク付きの論文の両方を出力として用いる．具体的には，初めに，ユーザにインデックスページから，論文を選択してもらう．選択された論文には要素技術へのリンクが既に生成されている．ユーザは選択した論文を閲覧するだけでも良いし，気になる要素技術をリンクを辿り芋づる式に調べることも可能である．インデックスページは常にユーザへの出力へ含めるため，ユーザは論文閲覧時において，常にインデックスページへ戻り再び別の分野の論文を閲覧することが可能である．

図 3 に本システムの実行例を示す．

図 3 の例は，論文中の要素技術に対しリンクが貼られており，その要素技術が扱われている他の論文を提示している様子を表している．この場合，論文中の要素技術は ① の「Annotea」であり，赤い枠 ② の中に「Annotea」を引用している他の論文のタイトルが表示されている．本システムを利用する際，要素技術 ① は黄色でハイライトされており，要素技術をクリックすると赤い枠 ② のリンク集が表示される．そして，赤い枠 ② 内の論文タイトルをクリックすると同じ要素技術を扱ってるその論文が新たに表示される．新たに表示された論文中に，異なる要素技術に対して既にリンクが生成されている場合がある．その場合，別の要素技術を新たに調べることも可能となる．

4 おわりに

本稿では，ユーザが論文を検索する際の支援を主な目的とし，論文中の引用情報をもとに，同じ要素技術を利用している論文をユーザへ提示する論文サーベイ支援システムを試作した．本システムを用いることにより，ユーザは閲覧する論文中に引用されている要素技術と同じ要素技術を利用している他の論文へリンクを辿ることが可能である．また，本システムでは，1 つの論文に複数の種類の要素技術へのリンクが作成されている場合があるため，論文間の簡易なネットワークを構築したと考えることができる．ユーザは気になる要素技術を発見した場合，講演論文集内の同じ要素技術を扱っている全ての論文を容易に調べることが可能である．本システムを用いることにより，ユーザが専門的な用語を知らない場合でも，興味ある論文を次々と検索できるため，研究初心者に対する論文サーベイを十分支援できると考える．

参考文献

[1] 難波英嗣，谷口裕子，“学術論文データベースからの研究動向情報の抽出と可視化” 言語処理学会 第 12 回年次大会 2006.

[2] 鈴木亮詞，工藤聖広，辻野友孝，清水堅，白松俊，大囿忠親，新谷虎松，“論文リポジトリに基づく研究支援のための対話的ユーザモデル構築手法の提案” 第 72 回情報処理学会全国大会論文集，Mar，2010(掲載予定).

[2]http://mecab.sourceforge.net/