Vol.2014-HCI-159 No.11
2014/8/5

他者の検索履歴を利用した研究者のための論文検索システムの提案

杉山　暢彦[†1]　澤本　潤[†2]　杉野　栄二[†2]　瀬川　典久[†2]

論文サーベイは研究活動において、動向調査や新規性の確認をするための非常に重要な過程の一つである。近年では、Google Scholar や Cinii などといった論文検索サービスによって簡単に論文を探して取得できるようになった。しかし、論文検索は目的に沿った論文を探すために、より深くページを辿り、多くのキーワードを使う必要がある。そのため論文検索は、少数の優良なページを探したら活動を終了できる通常の検索活動とは大きく異なる特徴を持つ。そこで、本研究では他者の検索結果を利用する事で論文検索活動を容易にするシステムを提案・実装した。本システムは論文検索機能に加えて、キーワードに関する論文グループを提示し、グループ内で利用されるキーワードを提示する機能を持つ。これにより、論文検索における手間の軽減を狙う。

Proposal of paper retrieval system for researchers using the search history of others

NOBUHIKO SUGIYAMA[†1]　JUN SAWAMOTO[†2]
EIJI SUGINO[†2]　NORIHISA SEGAWA[†2]

In research activities, paper survey is one of the process very important for checking the novelty trends and research. In recent years, it has become possible to find and fetch the paper simply by article search services such as Cinii and Google Scholar. However, in order to find the paper in line with the objective, it is necessary to follow a deeper page, use many keywords paper search. Search paper is characterized by very different from the normal course of search that can terminate the activity if you looking for a good few pages for that. Therefore, we propose and implement a system that facilitates the search paper activities by making use of the results of others in this study. In addition to the article search function, the system has the ability to present a paper on the keyword group presents a keyword that is used within the group. As a result, I aim to reduce the time and effort in the paper search.

1. はじめに

論文サーベイは研究活動において、研究動向の調査や新規性の確認をするための非常に重要な過程の一つである。論文サーベイする方法として、学会のポータルサイトや、Google Scholar[(2]、Cinii[(1] などといった論文検索サービスを利用する事などが挙げられる。また、グループ内で協調作業する際の支援を行うシステム[(5]等の研究も行われている。これらのシステムにより、効率良く関連研究を探せるようになっている。

論文サーベイのための論文検索活動は、通常の検索活動と異なる性質を持つ。通常の検索では、目的に対して少数の優良なページを発見し、問題や疑問が解決できれば検索活動を終了できる。論文検索は関連論文が少数であるとは限らないため、多くのキーワードを使い、膨大な検索結果を深く辿らなければならないため、論文検索には多大な時間と労力がかかってしまうという問題がある。

本研究では上記の問題を解決するために、キーワードを基にユーザの検索履歴を蓄積し、他のユーザに提示する事で論文検索活動を容易にするシステムを提案する。本システムでは、一度誰かが検索に利用したキーワードと実際にダウンロードした論文のリストを蓄積し、再び他のユーザが同一キーワードで検索行動をした際に蓄積してある行動履歴を提示する。これにより、提示された行動履歴を参照する事で、膨大な検索結果を辿らず一度に関連論文リストを得られ、論文検索における手間を軽減できると考える。また、提示した行動履歴の内キーワードを提示する事で、検索活動を支援できると考える。

以下､本論文では､２章で既存の論文検索について述べ、３章で提案手法とプロトタイプシステムについて述べ、４章でプロトタイプシステムの実装について述べ、５章で本研究のまとめについて述べる。

2. 関連研究

論文探索に関する研究やサービスは非常に多くの種類があるため、様々な関連研究について述べる。

2.1 既存の検索システムのアルゴリズム

通常の検索では、良質なページを上位に表示させるための仕組みとして Google のページランクが存在する。ページランクとは、多くのページから被リンクされているページは優良なページであり、優良なページからリンクされてい

†1 岩手県立大学大学院
Iwate Prefectural University Graduate School
†2 岩手県立大学
Iwate Prefectural University

るサイトもまた優良だという考え方に基づいたアルゴリズムである。同様に、論文検索ではより多くの論文から引用されている論文は重要であるという考え方に基づいたシステムが存在する。

また、論文検索環境として、Google scholar や Cinii といった論文検索サービスが近年の論文検索でよく利用される。これらのサービスは様々な論文検索サイトと対応している事や、様々な検索条件を指定できる事から、あらゆる視点からの検索を可能としている。さらに、検索キーワードと関連度の高い順に論文をソートできるため、前半のページに関連度の高い論文が集まり、後半のページは関連度の低いほとんど関係のないような論文の並びとする事ができる。

しかし、Google scholar は類義語の検索に対応していない事や、通常の検索と違ってキーワードのサジェストが行われない事から、関連キーワードの提示についての問題が解決されていない。

2.2 グループ検索の関連研究

武田らは、グループ内では興味・関心が似る事から、グループ内のメンバーの検索履歴を蓄積し、非同期でグループ内の検索が上手な人の力を利用できるようにするインタフェースを提案している。検索履歴には、ブックマークしたページ、クエリ入力間隔、特定ページへの訪問、ウィンドウの開閉といった動作に着目し、クエリとページが結びつくかを判別する事で返す決定木を生成している。

本システムでもユーザの検索履歴を蓄積して非同期な作業を実現しているが、対象は不特定多数のユーザであり、蓄積した検索履歴をグループ化して提示する事から手法が異なっている。

2.3 関連論文検索の関連研究

林らは、ユーザが検索キーワードについて考慮する必要の無い、英語論文の一部をクエリとする事で論文を検索できる意味概念に基づいた関連論文検索システムを提案・実装している〔6〕。ユーザの検索意図に適合させるために、単語間の相関性や単語の持つ意味といった概念を考慮する事で、意味的な関わりの強いクエリの生成に成功している。

2.4 関連論文の組織化の関連研究

学術論文には参照・被参照関係にある事から、２つの論文間で同じ論文を参照していれば、その論文は同じテーマを扱っていると考えられる書誌結合や、２つの論文が他の論文によって共に引用されている場合は２つの論文が同じテーマを扱っているとする共引用分析がある。

難波らは、学術論文間の参照・被参照関係の組織化について、３つのタイプの参照の理由を考慮した組織化を提案している〔7〕。タイプ B の論説根拠型では、他の研究者が提唱する理論や手法を用いて新しい理論を提唱する場合に利用する。タイプ C の問題点指摘型では、新しく提案した理論やシステムの新規性を証明・比較するために述べる場合等に利用する。タイプ O はその他のタイプを表す。これらのタイプを、被参照論文について記述している箇所を自動で抽出して解析、分類する。特に書誌結合について参照のタイプを適用させる事で、ノイズとなる結合を削減できるというものである。

この手法では論文から文書を読み取って機械的に組織化しているのに対して、本システムでは、ユーザの行動とクラスタリングによって論文同士をグループ化しているという点で異なる。

2.5 研究動向の可視化

吉田らは、電子的に入手可能な論文が増えた事で、全ての情報を入手し利用する事は困難であるとして、共引用分析とアソシエーションルールを用いる事で重み付き有効グラフを形成し、研究動向のマクロな流れの抽出に成功している〔8〕。流れの抽出には、論文をノード、アプリオリアルゴリズムの結果で得られたルールをコンフィデンス(信頼)値を重みとした有効枝とすることで重み付き有向グラフを作成している。このグラフに対してクラスタリングを行い、クラスタリングは、閾値を与えて、その閾値よりも大きい重みを持っていればクラスタ化する事でマクロな流れの抽出を実現している。

この研究では、参照構造を利用した研究動向の抽出と可視化に成功しているが、アルゴリズムの関係上まだあまり引用されていないような新しい論文にはうまく適用できていない。本研究では、ユーザの行動によるデータの蓄積を行うため、新しい論文であっても取得できる。

3. 提案手法

本章では、本研究における検索キーワードの定義、ユーザの行動履歴の解く責、提案するアプローチ、また、要素技術ついて述べる。

3.1 検索キーワードの定義

キーワードはスペース区切り等を考慮しない１つの文字列として考える。例えば、検索キーワードを「ユーザ　距離」とした場合、「ユーザ　距離」を一つの文字列としてみなす。そのため、「ユーザ」と「距離」がキーワードとして抽出される訳ではない。

3.2 ユーザの行動履歴の蓄積

本研究では、ユーザが必要とする論文をダウンロードする事でシステムにデータを蓄積する。蓄積するデータは、利用したキーワードとダウンロードした論文の URL の組とする。

3.3 本研究のアプローチ

論文検索において、同一のキーワードを利用したときに必要な論文はユーザの目的によってそれぞれ異なる。しかし、似た目的のユーザ同士が DL する論文は似ていると考えられる。例えば、検索キーワードを「データベース　仮想」とした場合、データベースを仮想的に複数のデータベースとして扱いたいという目的と、仮想環境内でのデータベースの利用という目的が考えられる。このように同一のキーワードであっても目的が異なるため、それぞれの目的に応じて必要な論文は異なる事が考えられる。

既存の論文検索サービスでは、著者名やキーワードとの関連度などといった様々なソート方法が提供されているが、目的に応じたソートが提供されていない事や、論文検索に関してクエリ拡張機能がない。そのため、大量の検索結果の中から欲しい論文を探さなければならない事や、様々なキーワードを使って繰り返し検索しなければならないという問題が発生する。

そこで、既に蓄積されているキーワードと同一のキーワードで検索した際に、蓄積されたデータはある目的に沿ったデータセットとなる。そのため、このデータセットを提示する事で、目的が一致した場合は検索結果を深く辿る事無く必要な論文が得られる。また、目的と合致しない論文が提示された場合には、違った目的の論文がシステムに蓄積されるため、これを繰り返す事によってユーザの目的を網羅できるようなデータを蓄積していく事が可能となる。

さらに、あるキーワードで検索した事のあるユーザ間において類似度を求め、ユーザ間でクラスタリングしグループを抽出する事で、類似ユーザがダウンロードした論文集合を得られる。また、類似ユーザによってグループ化したため､利用されるキーワードの目的が似ると考えられ得る。このデータを検索時に候補として提示する事によって検索の支援が可能となる。概念図を以下に示す。（図１）

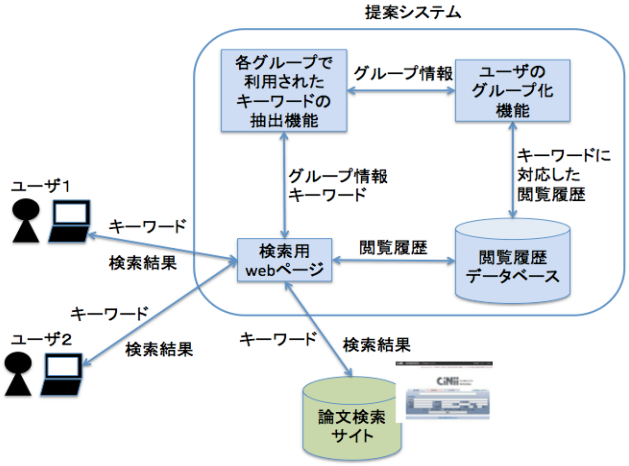

図 1　システム概念図

Figure 1　System conceptual diagram

3.4 ユーザ間の距離

本システムでは、同一のキーワードを利用したユーザ同士の類似度をユーザ間の距離としてクラスタリングで利用する。

ユーザ間の類似度の抽出のために､蓄積したデータの内、同一のキーワードを利用したユーザと、そのキーワードを利用した時にダウンロードした論文を表形式で表現した閲覧履歴データとする（表１）。閲覧履歴データは、論文をダウンロードした場合は１、論文をダウンロードしていない場合は０といった２値で表現する。この表形式のデータを以下のような行列に変換する。

$$U = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

変換した行列に対して､Jaccard 係数を用いて距離を算出する。Jaccard 係数とは、２つの集合間の類似度を表す係数で、集合間の類似度が高いほど１に近づく性質を持つ。Jaccard 係数は以下の式で表現される。

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

本システムでは、ユーザ間で同一の論文を持っているか持っていないかを基に類似度を求めたいと考えたため、jaccard 係数を利用する。また、類似度が高いほど０に近づけたいため､1 から Jaccard 係数により求められた距離を引く事でユーザ間の距離を抽出する。

表 1　ユーザ間の距離抽出に利用するサンプルデータ

Table 1　Sample data for extracting the distance between users.

	論文 A	論文 B	論文 C	論文 D	論文 E
ユーザ 0	DL	DL	DL	x	x
ユーザ 1	DL	DL	DL	DL	X
ユーザ 2	x	x	DL	DL	DL
ユーザ 3	x	x	x	DL	DL

3.5 検索履歴のクラスタリング

本システムでは、上記で抽出したユーザ間の距離を用いてクラスタリングを行い、類似したユーザが属するグループを抽出する。

クラスタリングには、主に階層的クラスタリングと非階層的クラスタリングの２種類があるが、抽出したいグループの個数は蓄積したデータによって異なるため、本システムではグループ数が不定でも実行可能な階層的クラスタリングを利用する。

階層的クラスタリングとは、類似性の高いデータ同士をグループ化し、さらにそのグループ同士をまとめて親グループ化するという作業を繰り返す事で、グループの階層構

造を作り出す手法である。クラスタを生成する際の手法はいくつかあるが、本システムではユーザ間の距離が近い順にグルーピングしたいため最短距離法を用いる。（図 2）

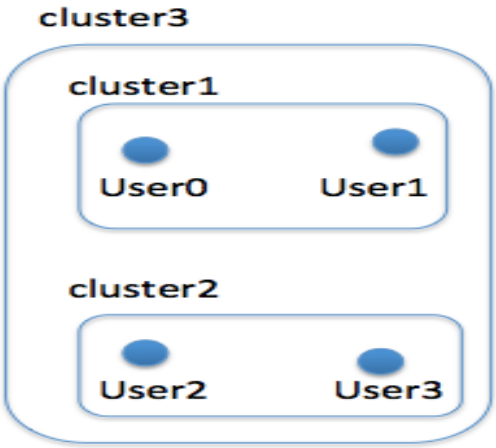

図 2　階層的クラスタリングについて
Figure 2　Figure of the hierarchical clustering

階層的クラスタリングのアルゴリズムを以下に示す。

1. データ集合の中から、互いの距離が最も近くなるデータ項目の対を探す。
2. その項目対を、1 つの項目対を一つのクラスタに統合する。
3. そのクラスタと残りのデータ集合の中から、互いの距離が最も近くなるよう訴追を探して統合する。
4. 上記の処理を、データ全体が一つのクラスタに統合されるまで繰り返す。

以上の手順により、ツリー状のクラスタ構造が形成される。このツリー上のクラスタはデンドログラムと呼ばれる。（図３）このクラスタ構造は、クラスタをノードとし、１つのノードはそれぞれ２つクラスタと距離を持つ。この距離を閾値として設定する事で指定した閾値以下のクラスタを抽出する。本システムでは、クラスタ同士の距離は 0〜1 として設定し、0 に近いほど似ているユーザとした。閾値は試験的に 0.6 に設定してクラスタを抽出している。

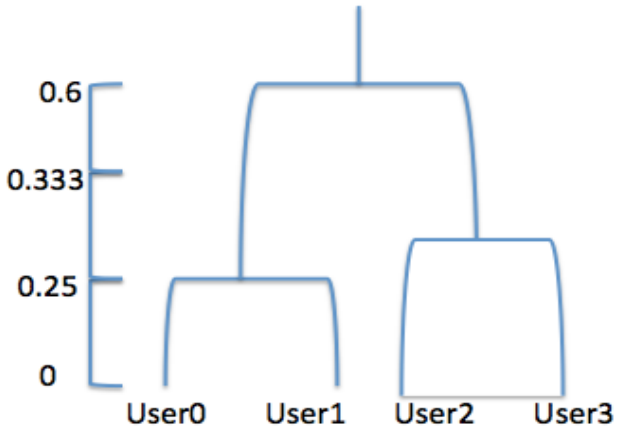

図 3　生成されたデンドログラム
Figure 3　Generated dendrogram.

4. プロトタイプシステムの実装

本章では、実装したプロトタイプシステムについて述べる。

プロトタイプシステムは、Apache 上で動いているプログラムにブラウザからアクセスする形式で利用する。システム構成を以下の図に示す。（図４）

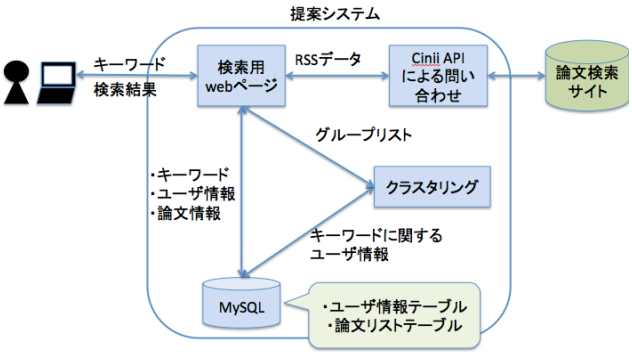

図 2　システム構成図
Figure 3　System Configuration

本システムは以下の 3 つの機能を持つ。

1. Cinii からの検索
2. キーワードに対して、類似ユーザグループ内のユーザのダウンロードした論文リストの提示
3. グループのユーザが利用したキーワードの提示

Cinii からの検索機能は、Cinii が提供している opensearch API を利用して Cinii からの検索を実現している。検索画面を図に示す。（図５）検索画面には、Cinii からの検索結果、Cinii の論文情報ページへのリンクを持つ論文ダウンロードボタン、さらに、opensearch API から取得できる論文の書誌情報を表示する。表示する書誌情報は以下の通りであ

る。

- 研究タイトル
- abstract
- 著者
- 出版者
- 刊行物名
- ISSN
- 出版年月日

直接 Cinii へのリンクを提示せず検索結果に論文情報を全て提示する理由は、検索結果を見るだけで論文が必要か不要かを判断できるようにするためである。本システムでは実際に論文データを持っている訳では無いため、論文情報サイトへのリンクをダウンロードボタンに割り当てる事で、ユーザが論文を必要としたかどうかの判断を可能にした。

キーワードに対する類似ユーザのグループの提示機能は、前述した階層的クラスタリング等の手法を利用してグループを抽出し、グループを検索結果に反映させる機能である。ユーザがキーワードを入力して検索した際に、蓄積データがあれば黄色い枠内にグループが提示される（図６）。この時、階層的クラスタリングは３ユーザ以上いないと比較対象が無くクラスタリングができないため、キーワードに対して蓄積されているユーザのデータが２件以内の時には、ユーザの検索履歴をそれぞれグループとして提示する。

グループ内で利用されたキーワードの提示機能は、機能２で提示した各グループの項目の最下行に提示する。各グループはユーザのダウンロードした論文を基に抽出されたため、これらのキーワードは各グループが暗黙的に持つユーザごとの目的に沿ったキーワードとなっている。

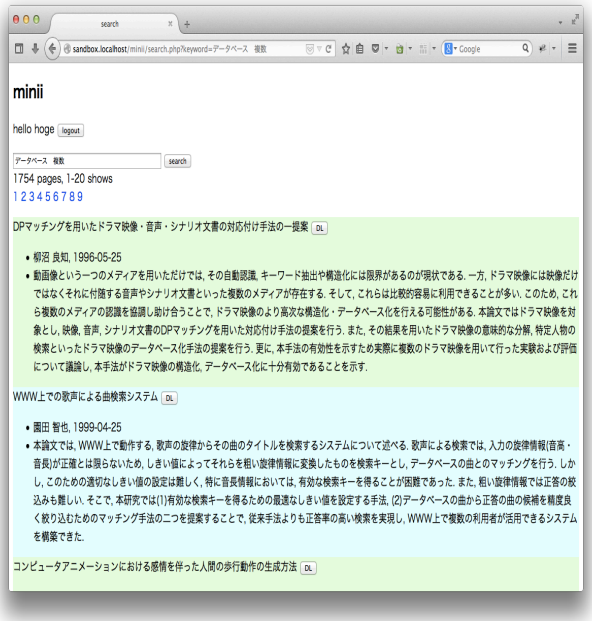

図 5　論文検索インタフェース
Figure 5　Interface for the search paper

図 6　キーワードに関連したグループの提示
Figure 6　Suggest groups related keyword

5. まとめ

本研究では、論文検索において発生するキーワード選定の困難さや、膨大な検索結果から深く辿らなければならないという問題を解決するために、他者の検索履歴を利用する事で容易に論文検索ができるシステムを提案・実装した。本システムでは、システム上での論文検索機能に加えて、同一キーワードを利用したユーザ間の検索履歴を元にクラスタリングし、抽出したグループでダウンロードされている論文と、そのグループでよく利用されるキーワードを推薦する。これにより、既に情報が蓄積されているキーワードであれば、深く検索結果を辿る事無く論文を取得でき、関連キーワードを得られるようになる。

今後は提案システムの評価実験を行いシステムの改善を行っていく。さらに、蓄積データの増加に伴う計算量の増加への対策や、階層的クラスタリングにおけるクラスタの生成方法について、最短距離法では１つのクラスタが肥大化してしまう問題があるため、実験を通じて使い適切な手法へと変えていく事等が課題である。

参考文献

1) http://ci.nii.ac.jp/
2) http://scholar.google.co.jp/schhp?hl=ja
3) http://ci.nii.ac.jp/info/ja/api/a_opensearch.html
4) 林佑磨、奥野峻弥、山名早人:意味概念に基づいた関連論文検索システム、deim2014、2014
5) 難波英嗣、神門典子、奥村学:論文間の参照情報を考慮した関連論文の組織化、情報処理学会論文誌 42(11), 2640-2649,2001
6) 吉田誠、小林隆志、難波英嗣、奥村学、横田治夫:研究論文デー

タベースからの研究のマクロな流れの抽出、DEWS2003、2002