情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-CE-105 No.1
2010/7/10

Web アプリケーションを用いた 重要論文サーベイシステムの設計

井 坂 徳 恭[†1]　中 山 泰 一[†1]

論文検索の手法のひとつとして Web 検索が挙げられる．しかし，論文の Web 検索では，ユーザが重要な論文を発見することが難しいという問題があった．そこで，我々は，知識の少ない，初めて論文サーベイを行うユーザを支援するために，ユーザの興味のある分野における重要な論文を発見するシステムの研究を行ってきた．

本論文では．多くのユーザが簡単に利用でき，論文サーベイの方法を効率的に理解することを目的として，Web アプリケーションによるシステムの実装を行う．

Design of a Survey System for Searching Important Articles by Web Application

Noriyuki Isaka[†1] and Yasuichi Nakayama[†1]

This paper proposes a survey system for searching important articles to support novice users. We have so far researched a search technique by using reference structure, therefore, we apply this technique to our survey system.

In this paper, we discuss how to design the Web application and ranking algorithms for searching important articles.

1. はじめに

学術論文を収集する方法のひとつとして Web 上での検索がある．現在，論文のための検索エンジンや，学会のポータルサイトによる，論文のデータベースを検索可能としたシステムが存在する．これらのシステムによる検索によって論文サーベイができる．しかし，ユーザがあまり知識を持っていない分野での論文サーベイには，重要な論文を見つけることが難しいという問題がある．

学術的な研究では，年々新しい分野の研究が現れており，分野の中でも研究が細分化している．また，いくつもの分野を統合した研究などもある．このように，研究は複雑に関係しあい，膨大な数の文献が存在し，今後さらに複雑になっていく．そのため，学術論文の検索を行い，適切に必要な分野の論文を絞り込むことは難しい．適切な絞込みを行うには，ユーザが，その分野における専門用語などの知識を適切に持っていなければいけない．

ユーザの論文検索を手助けする研究として，拡張クエリを扱う方法[1)] がある．しかし，この方法では，ユーザが興味のある論文を探すことができるが，その分野における重要な論文を発見するには知識が必要となり，難しい．そこで，膨大な数の参考文献や検索結果を調べる必要があり，ユーザの負担が大きくなる．

そこで我々は，ユーザの論文検索を支援するシステムの設計，実装を行ってきた[2)]．ユーザの指定した論文を基点とし，参考文献から参照構造を作り，HITS アルゴリズムを応用し，重要な論文を発見する．参照の集合を適切に作ることで，必要な分野での論文の発見の精度を上げることができる．同時に，得られた参照構造の中で，重要とされた論文までの参照の経路の提示する．これらを行うことで，ユーザが自分の知りたい分野での論文サーベイを容易にし，その分野の研究の遷移を簡単に知ることができる．このシステムにより，ユーザの論文サーベイの支援に対し，一定の成果を上げることができた．

本論文では，この論文検索支援システムを用いて，論文サーベイを初めて行うような初心者ユーザに対して，論文サーベイについて理解させるためのシステムを設計，実装する．論文サーベイは，多くの論文を調べ，その参考文献を辿りながら重要な論文を発見していき，ある分野の研究について理解していく．論文検索支援システムでは，参考文献を辿り，重要な論文を発見する手法をとっており，同時に経路の表示を行っている．これらを利用し，適切なユーザインタフェースの設計をし，表示することで，初心者ユーザに対して，論文サーベイの方法を理解させることを目標とする．

同時に，論文検索支援システムの精度の向上を行う．HITS アルゴリズムを応用し，成果を出しているが，Web 上における問題点と同様の問題により，検索結果が想定していないものになる問題が確認できている．そこで，HITS アルゴリズムのように，他のいくつもの Web ページのランク付けアルゴリズムを応用し，実験を行うことで，最適なアルゴリズムの設計を行う．

これら２つを行うことで，初心者に対しての教育と同時に，新しい分野の研究を行う，多

†1 電気通信大学情報工学科
Department of Computer Science, The University of Electro-Communications.

くの研究者に対しても，有益なシステムとなると考えられる．

以下，本論文では，第２章で既存の論文検索システムやその研究について述べる．第３章では，これまで行ってきた論文検索支援システムの成果について述べる．第４章，第５章では今後進めていく研究のうち，それぞれインタフェース，検索アルゴリズムについて述べ，第６章では今後の予定について述べる．

2. 既存の論文検索システム

2.1 ACM Portal

ACM Portal[3] は ACM(Association for Computing Machinery) の Web ベースのポータルサービスである．オンラインジャーナルのデータベースである Digital Library と書籍データベースである The Guide の２つから構成されている．Digital Library では，ACM が刊行している学会誌などに記載されたジャーナルを検索することができる．The Guide では，コンピュータ分野における書籍情報を検索することができる．以下の特徴が挙げられる．

- ダウンロード数の表示
  過去 6 週間，1 年間でのダウンロード数を見ることができるため，記事の引用以外にも研究の注目度がわかる．
- 各論文の記事へ URL リンクが可能
  URL によって論文の記事を参照できるために，ユーザ間で情報を簡単に共有できる．
- 参照情報
  各論文ごとに参照，被参照の情報が掲載されており，記事へのリンクがされているため，参考文献などを簡単に辿ることができる．
- 様々な検索機能
  キーワード，著者名，ISBN/ISSN，雑誌名，出版社名，刊年など様々な条件により検索ができる．

2.2 Google Scholar

Google Scholar[4] は Google が収集しているデータから，特に学術文献を抽出し，それらを検索しやすいように機能を追加した検索システムである．タイトルや本文検索だけでなく，著者名や出版社，年代による検索ができる．以下の特徴が挙げられる．

- 大規模なデータベース
  Google のキャッシュを利用しているため，非常に大規模なデータベースである．
- 著者名，年代，出版社による検索
  単純なタイトルや本文以外の条件による検索が行える．
- 引用元の情報
  その論文から派生した研究を簡単に見つけられる．また引用数によりその論文の注目度がわかる．
- Recent articles
  検索する際に，近年の研究への重み付けを行い現在行われてる研究を見つけることができる．

2.3 関 連 研 究

論文のランク付けを行う研究は数多く行われている．その中でも，Web ページのランク付けアルゴリズムを論文のランク付けに利用されており，よい結果が出ている[5]．

また，論文検索にランク付けアルゴリズムを応用した研究として，HITS アルゴリズムを利用し，サーベイ論文を発見する方法[6] がある．これは論文データベースに対して HITS アルゴリズムを適用し，サーベイ論文を発見する手法である．

2.4 問 題 点

これらの検索システムは目的とする論文に対し知識を持っているユーザならば，目的の論文紙や，著者，キーワードを利用して必要な論文を見つけることができる．しかし，ユーザの知識が少ない分野において論文サーベイを行う際，著者や論文紙の情報は使えず，キーワードの指定も曖昧になるため，重要な論文を見つけることが難しい．また，これらのシステムを用いた参照情報によって論文サーベイを行うこともできるが，非常に多くの論文が参照の中に現れるため，その全てを調べることは難しい．

既存の論文のランク付けを行う研究では，Web と同様，巨大な論文のデータベースの領域全体に対して行っている．1 章で述べたように，自ら重要な論文を発見できるユーザに対しては有効であるが，知識の少ないユーザに対しては必要のない論文を数多く提示してしまう可能性が高い．

3. 論文サーベイ支援システム

本章では，我々がこれまで行ってきた論文検索支援システムに関して，論文の検索手法の設計について述べる．

２章で述べたように，多くのシステムや研究が，検索における利便性の向上を目的としている．しかし，論文サーベイになれてないユーザでは，どの論文が興味のあるものである

か，重要であるかを見極める知識が少ないため，検索における利便性の向上だけでは不十分であると考えられる．

そこで我々は，初心者ユーザでも簡単に重要な論文を検索できる手法として，参照構造を用いた論文サーベイ支援システムを設計，実装を行ってきた．

3.1 論文の参照関係

本研究は，ある学術論文からその分野の重要論文を発見することを目的とする．そのために，学術論文における参照関係を利用する．学術論文は，研究の際に引用した論文や文献，Web ページなどを参考文献として記載している．参考文献は研究中で利用した技術などを記載しているため，その研究の分野を理解する上で非常に重要である．そのため被参照数はその論文の注目度を表す指標のひとつとなっている．

論文の参照関係は，論文をノードとすると，図 1 のような無閉路有向グラフで表わされる．これは，Web ページにおけるリンク関係と類似しており，同様に支持投票としての性質を持っている．よって参照関係の解析に Web のランク付けアルゴリズムを利用できる．

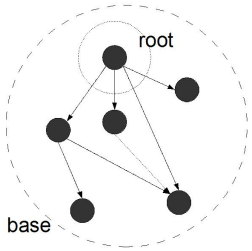

図 1 論文の参照構造における base 集合

3.2 論文の検索アルゴリズム

本システムでは，論文の参照関係を解析するために Web ページのランク付けに使われるランク付けアルゴリズムを利用する．ランク付けアルゴリズムとは，Web ページの集合などにおいて，各ページの重要度を計算し，ランク付けを行い，重要なページを発見するアルゴリズムである．Google の検索エンジンなどで使われる PageRank[7] や Yahoo!の検索エンジンで使われる HITS アルゴリズム[8]，HITS にコンテンツ内容の考慮を加えた IMP アルゴリズム[9] などがある．

今回の実装では，HITS アルゴリズムを利用した．これには，以下の理由が挙げられる．

- 構造を重視した解析
  HITS アルゴリズムはページの内容ではなく，リンク構造そのものを重要視している．よって，Web ページと論文の内容の違いを考える必要がない．また，論文は PDF ファイルで保存されていることが多く，内容解析には時間がかかってしまう．
- 集合の作成方法
  解析対象の集合を集める際に，まず root 集合を用意して，そこから集合を作成している．本手法では，ユーザの興味のある論文のみを root とすることで，領域を限定した base 集合を作成する．
- hub と authority
  論文検索においては，重要な論文だけでなく，それらを多数参照している論文を見つけることも重要である．HITS アルゴリズムにおける hub と authority の概念により，両方を同時に求めることができる．

以下に重要な論文の発見手法についての詳細を述べる．

3.2.1 base 集合の作成

一般的な論文検索では，上位数件の検索結果の中でも，研究の分野に違いがでてしまうことがある．そのため，知識の無いユーザに，興味を持っている分野の重要な論文を提示するためには，検索対象となる論文の分野を限定する必要がある．よって本手法では，root 集合にあたる部分をユーザが興味を持った論文のみとする．続いて root となった論文の参照情報から論文を集合に追加して base 集合を作成する．さらに base 集合に含まれる論文の参照情報から論文を集合に追加していく．これを，論文の参照が辿れなくなるか，集合に含まれる論文数が一定数になるまで繰り返す．ここから隣接行列 A を求める．論文の参照関係における隣接行 $A = [a_{ij}]$ は論文 i が論文 j を参照している場合 $a_{ij} = 1$ とし，それ以外では $a_{ij} = 0$ とする．これにより，ユーザの興味のある分野に限定した base 集合が作成される．

3.2.2 authority と hub の算出

作成された隣接行列から，HITS アルゴリズムの反復計算を応用し，各論文の authority と hub の値を求める．auth 値，hub 値は初期値として，全て 1 で初期化する．本手法はランク付けを行う範囲を限定しているため，中には base 集合中の，過半数の論文を参照しているサーベイ論文が存在する可能性がある．ここで，HITS アルゴリズムをそのまま適用してしまうと，このような論文が，最も重要な hub であると判断され，それ以外の論文に hub としての価値がなくなってしまう．この状態では，そのサーベイ論文のみに参照されて

Vol.2010-CE-105 No.1
2010/7/10

いる論文が，他の多くから参照されている論文より重要であると判断される．本システムは，その分野に知識の無いユーザに重要な論文を提示することが目的であるため，特定の論文の hub 値が突出する前に，反復計算を終了する条件を設定し，この問題に対応する．本システムにおける authority と hub の算出アルゴリズムは，図 2 となる．

n :論文数
$auth_n$:論文の authority 値の集合
hub_n :論文の hub 値の集合
A_{nn} :隣接行列

for k=0···
  for i=0···n
    for j=0···n
      $auth_i$ += A_{ji} * hub_j
normalization of $auth$

  for i=0···n
    for j=0···n
      hub_i += A_{ij} * $auth_j$
normalization of hub

if Termination condition == true
  end

図 2 論文の authority と hub の算出

3.3 システムの実装と評価

これらのアルゴリズムの実装と評価を行った．C#を用いたアプリケーションとしてシステムを実装し，参照関係の習得のために ACM Portal を利用した．

入力はユーザが興味を持った任意の論文とし，そこから参照構造を取得，解析し，重要度の高い順に参照構造内の論文を出力する（図 3）．

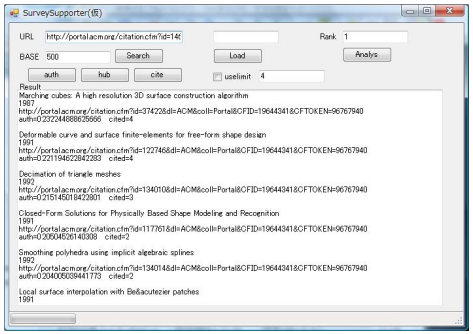

図 3 検索結果の出力

また，出力された任意の論文に対し，研究の遷移図を出力する（図 4）．これは，別ウィンドウで出力される．

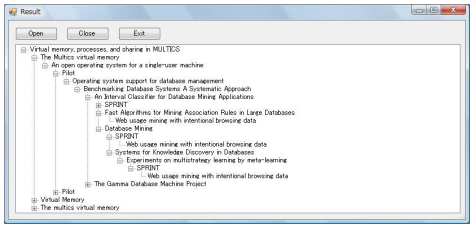

図 4 遷移図の出力

このシステムを用いて，実際にいくつかの例に対し論文サーベイを実行した結果，同じ分野の重要な論文を発見することができた．また，遷移図では，研究の流れや，どんな研究を基礎としているかを，ユーザが理解することを助けることができる．

3.4 問 題 点

問題点として，以下の 2 点が挙げられる．

- ユーザインタフェース
  システムは，論文サーベイ支援のための手法を評価するために，簡単なアプリケーションとして実装している．本研究の目的は，初心者のユーザに，わかりやすく伝えることである．また，実際に初心者に利用してもらうためには，簡単にシステムを利用できる

環境を構築する必要がある．

- 検索の精度
  実際に評価を行ったところ，多くの結果では，同じ分野の重要な論文を提示することができた．しかし，分野や研究内容によっては，多くの分野の複合であったり，比較的新しいため，十分な参照を分野内で得られないなどの問題から，予想と違う結果になるという問題が見られた．

4. ユーザインタフェースの改良

本章では，既存のシステムの問題点であるユーザインタフェースの改良を行う．初心者ユーザのサーベイ支援を効率的に行うため，利用しやすく分かりやすいインタフェースの設計を目指す．

4.1 Web アプリケーションを用いた実装

本研究では，Web アプリケーションとしての実装を行う．Web アプリケーションとして実装することにより，アプリケーションのインストールや更新などが必要なく，プラットフォームに依存しないため，多くのユーザが簡単に利用できる．

また，同時に ACM 以外の論文データベースでの探索の実装を行う．CiNii[10] では，ACM Portal 同様，参照文献のリンクが存在するため，実装が容易である．

4.2 参照構造取得の可視化

初心者ユーザを対象として，参照構造の取得に関する可視化を行う．論文サーベイを行う際には，参考文献に挙げられている論文を見ていくことは重要である．論文サーベイ支援システムでは，大規模ではあるが，実際にこれらの作業を行っており，視覚化することで，教育に役立つと考えられる．

4.3 遷移図の出力

現状では，重要な論文への経路を全て表示している．おおよその研究の流れを理解する手助けにはなるが，論文の経路は非常に多くなり，全てを見ようとすると，ユーザの負担が大きくなってしまう．そこで，ここでも Web 上で使われている可視化技術を応用することで，ユーザにわかりやすい表示を行えるように設計する．

そこで，グラフ分割によるクラスタリング法である Betweenness Clustering[11] などのクラスタリング手法を利用し，複雑な経路を複数に分割し，重要度などで順序付け表示する．図 5 に例を示す．ここでは，直線的な経路とリンクが密な経路に分類された想定をしている．これにより，重点的に参照されている部分などを発見することができると考えられる．

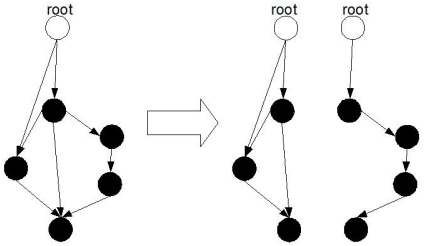

図 5 遷移図の分割例

また同様に，取得した参照構造全体にも同様のクラスタリングを行い，複数の研究が関わっていることをユーザが理解することができる考えられる．図 6 に，クラスタリングの例を示す．図のように，参照構造をいくつかのコミュニティに分類することで，リンクが密になっていると考えられる同様の集合を発見することができる．そのようなコミュニティの大きさや重要度により，ユーザが興味をもった分野の位置づけを表わすことも可能であると考えられる．

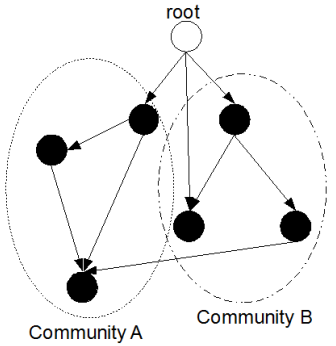

図 6 クラスタリングにおけるコミュニティの発見例

5. アルゴリズムの改良

本章では，既存システムの問題点であった検索の精度の問題について改良を行う．

5.1 HITS アルゴリズムの問題点

HITS アルゴリズムとは Kleinberg によって提案された，Web コミュニティを発見するためのランク付けアルゴリズムである．Web ページ間のリンク構造を解析することにより，Web ページの内容を解析せずに有益な Web ページを探し出すことができる．

しかし，TKC 効果と呼ばれる問題点がある．TKC 効果とは，Tighty-Knit Community Effect の略称である．これは，小さいが，内部リンクが密に張られているようなコミュニティが上位になってしまう．また，ランク付けを行う対象ないに，複数のコミュニティが存在する場合，リンクが密に張られているコミュニティのスコアが高くなってしまい，その他の重要なコミュニティを見つけることができなくなってしまう．

5.2 SALSA

HITS アルゴリズム以外のランク付けアルゴリズムとして，SALSA[12] がある．HITS アルゴリズムと同様に Hub と Authority の概念を利用している．HITS アルゴリズムが Hub と Authority を相互に強化する関係として処理している．しかし，SALSA ではマルコフ連鎖に基づいたランダムウォークの確立的特性を利用し，Authority と Hub，それぞれのマルコフ連鎖を分析することでスコアを求める．これにより，HITS アルゴリズムより，TKC 効果を受けにくく，複数のコミュニティが関係している論文の参照構造で良い効果が得られると考えられる．

5.3 システムへの適用

HITS アルゴリズムの問題点のひとつの解決として，SALSA アルゴリズムを挙げたが，ランク付けアルゴリズムは非常に多くの研究がされている．また，Web ページにおける問題点が，論文の参照構造では問題にならない場合も考えられる．そのため，実際にシステムへと適用し，運用していき，ユーザの意見を聞くことで，効率のよいアルゴリズムの発見，及び応用を行う．

6. 今後の予定

今後の予定として，４章，５章で述べた設計を実装するとともに，Web アプリケーションの公開を行い，実際にユーザの意見として，結果の出力に関する見やすさ，分かりやすさ，及び検索アルゴリズムに対する検索結果の精度について聞いていく予定である．

これらにより，初心者ユーザに対しても，論文サーベイを支援でき，同時にサーベイの方法や，研究の遷移の理解を助けるシステムを実装していく予定である．

参 考 文 献

1) 榊 剛史, 石塚 満: Web を活用した論文サーベイシステムに関する研究, 東京大学大学院 情報理工学系研究科 修士論文(2005).
2) 井坂 徳恭，藤本 敬介，中山 泰一: 参照構造を用いた重要論文検索システム, 情報処理学会 コンピュータと教育研究会 *99* 回研究発表会(2009).
3) The ACM Portal, ACM(online), available from <http://portal.acm.org/> (accessed 2009-11-20).
4) Google Scholar, Google(online), available from <http://scholar.google.co.jp/> (accessed 2009-11-20).
5) Y, Sun. and C. L, Giles.: “Popularity Weighted Ranking for Academic Digital Libraries”, *Lecture Notes ic Computer Science*, 4425, pp.605-612(2007)
6) 難波 英嗣, 奥村 学: 多言語論文データベースを用いたサーベイ論文検出 : サーベイ論文自動作成の実現に向けて, 電子情報通信学会技術研究報告. *NLC,* 言語理解とコミュニケーション, vol.102, No.199,pp.35-41(2002).
7) Sergey, B. and Lawrence, P.: “The anatomy of a large-scale hypertextual Web search engine”, *Computer Networks and ISDN Systems*, vol.30, Issues1-7,Pages107-117(1998).
8) Kleinberg, J.: “Authoritative sources in a hyperlinked environment”, *ACM Computing Surveys*, vol.31, Issue 4es, Article no.5(1999).
9) Krishna, B. and Monika, R.H.: “Improved algorithms for topic distillation in a hyperlinked environment”, *SIGIR ’98*, Pages:104-111(1998).
10) CiNii - NII 論文情報ナビゲータ, CiNii(online), available from <http://ci.nii.ac.jp/> (accessed 2010-4-1).
11) Michelle, G. and Mark, N.: “Community structure in social and biological networks”, *Proceedings of the National Academy of Sciences of the United States of America*, vol.99, Pages7821-7826(2002).
12) Ronny, L. and Shlomo, M.: “The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect”, *ACM Transactions on Information Systems*, Pages387-401(2000).