Vol.2009-CE-99 No.4
2009/5/23

参照構造を用いた重要論文検索システム

井坂徳恭[†1]　藤本敬介[†1]　中山泰一[†1]

現在，論文検索においてはキーワード指定など重要論文を見つけるためにはいくつも難しい問題がある．そこで，論文サーベイを支援するために，論文の参照構造に着目した論文検索手法を提案する．論文の参照構造が Web 上のリンク構造に類似してることから，Web マイニングの知識を応用した．
本システムを実装し，実験を行ったところ，被引用数の多い重要な論文を発見し，ユーザの論文サーベイを手助けすることができた．

A system for searching important articles by using reference structure

Noriyuki Isaka,[†1] Keisuke Fujimoto[†1]
and Yasuichi Nakayama[†1]

This paper proposes a search technique which supports an article survey by using reference structure. The reference structure of articles can be considered as link structure of the Web, so we apply knowledge about the Web mining. Experimental results show our system is effective to search an important article referred to by many articles.

1. はじめに

近年，論文収集の方法のひとつとして Web 上での検索がある．現在では，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化および検索可能としたシステムが存在する．

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

これらを利用した論文サーベイを行うことができるようになっているが，ユーザが今まで扱っていなかった論文を調べようとするときには問題点がある．まず，参考文献の数が膨大である．論文に記載されている参考文献は，論文サーベイにおいて非常に有益な情報となるが，世代をさかのぼり調べるには数が膨大である．続いて検索ワードのの決定が困難となる問題がある．論文では，研究分野ごとに専門用語が多く存在する．また似通ったタイトルの論文も多く存在するため，検索によって見つかった論文の中から目当てのものを見つける作業がさらに必要となる．これらの問題から，ユーザが重要論文を発見することが難しくなっている．特に，新規の分野での論文サーベイを行おうとするユーザにとってはこれらが障害となりやすい．

そこで本論文では，ユーザの論文サーベイを手助けするシステムの設計と実装を行う．新規ユーザはある論文に興味を持ち，結果としてその研究分野の論文サーベイを行うことが多いと考えられる．そこで，本システムではユーザの興味をもった任意の論文から参照構造を利用することで重要論文の発見を行う．論文の参考文献による参照構造は，Web 上におけるリンクによる参照構造に類似している．そのため，本手法では Web マイニングのアルゴリズムのひとつである HITS アルゴリズムを応用することで，重要論文の発見を目指す．同時に研究の遷移を提示する．これにより，ユーザがこれまでその分野の研究がどのようにして行われてきたか，何が求められてきたかなどを理解することを手助けする．

実際に本手法を用い，ACM Portal の論文を対象として実験を行ったところ，有名な論文を数多く発見することに成功した．また，研究の遷移の提示においても，現在の研究がどのような研究，技術を元として行われているかを知ることに役立てられた．

以下，第 2 章では既存のシステムについて述べ，第 3 章では本システムで利用する HITS アルゴリズムについて述べる．第 4 章では本システムにおける論文の検索手法を，第 5 章では実装について述べる．第 6 章では，実装したシステムの実験と評価を行い，第 7 章では本論文をまとめる．

2. 既存のサーベイシステム

2.1 ACM Portal

ACM Portal[1)] は ACM(Association for Computing Machinery) の Web ベースのポータルサービスでる．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可能である．The Guide に

おいては，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．特徴として以下の点が挙げられる．

- ダウンロード数の表示
  過去 6 週間，1 年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，ユーザ間での情報の共有などを簡単に行うことができる．
- 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，刊行年など様々な条件により検索が行える．

2.2 Google Scholar

Google Scholar[2] 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となっている．タイトルや本文検索だけでなく，著者名や出版社，年代による検索を行うことができる．特徴として以下の点が挙げられる．

- 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簡単にわかる．
- Recent articles
  検索する際に近年の研究への重み付けを行って現在行われてる研究を見つけることができる．

2.3 問 題 点

これらの検索システムは目的とする論文がすでにわかっており，本文テキストを入手しようとしているユーザからは使いやすいが，論文サーベイの際には問題が存在する．新規分野において論文サーベイを行うユーザには，著者名や雑誌名などでの検索は便利ではあるが，その中から絞り込むこともやはり難しい．また，これらのシステムで参照情報を知ることができ，それを用い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，全てを調べることが難しい．

3. HITS アルゴリズム

HITS アルゴリズムは Kleinberg によって提案された，Web コミュニティ発見のための Web マイニングのアルゴリズムのひとつである[3]．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．

3.1 hub と authority

HITS アルゴリズムでは優秀な Web ページは優秀なリンク集からリンクされているという考えにもとづき，Web ページをランク付けしている．そのため，Web ページを authority と hub という 2 種類のページに分類している．authority とは，検索ワードに対して的確な情報をもつ Web ページであり，hub は多くの authority に対してリンクを持つような Web ページである．多くの hub からリンクされている Web ページはよい authority であり，多くの authority をリンクしている Web ページはよい hub である．

3.2 解析対象の行列の作成

はじめに検索キーワードを本文中に含むページによって集合を作成する．root は全文検索型のサーチエンジンによって得られた結果の上位数件を使うなどして収集する．これを root 集合とよぶ．次に root 集合の各ページからリンクしているページを収集する．また root 集合の各ページへのリンクを持っているページも同様に収集する．収集したページを root 集合に加える．作成された集合を base 集合 (図 1) とよぶ．base 集合を元にして隣接行列を作成する．ここで隣接行列 $A = [a_{ij}]$ はページ i からページ j へのリンクが存在する場合 $a_{ij} = 1$ とし，それ以外では $a_{ij} = 0$ となる行列である．

3.3 authority と hub の発見

base 集合の各ページを authority と hub に分類するために，ページ i に対して authority の値として a_i，hub の値として h_i を与える．各値は以下のように定義する．

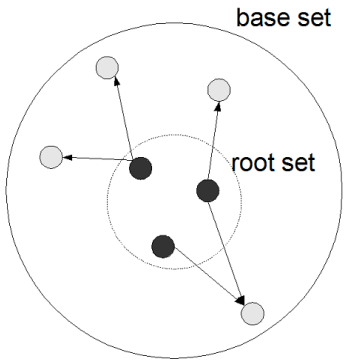

図 1 base 集合の作成

$$a_i = \sum_{j,j\to i} h_j \tag{1}$$

$$h_i = \sum_{j,i\to j} a_j \tag{2}$$

HITS アルゴリズムでは各ページに初期値 a_{i1}，h_{i1} を与えてこれらの式による反復計算を行うことで，各値を求める．

4. 論文の検索手法

4.1 論文の参照関係

本研究では，ある論文からその分野の重要論文を発見することを目標とする．そのために，論文における参照関係を利用する．学術論文においては，本文の最後に研究の際に参考とした論文や文献，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．そのため被引用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．

論文の参照関係は図 2 の通りである．これは Web ページにおけるリンク関係 (図 1) とよく似ている．これらの参照どちらも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 マイニングの技術を応用できると考えられる．

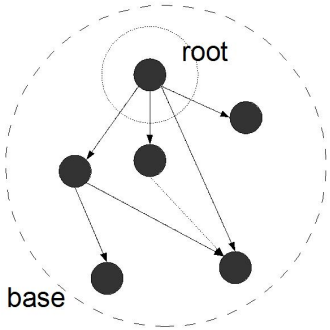

図 2 論文の参照構造における base 集合

4.2 論文の参照関係への HITS アルゴリズムの適用

4.2.1 HITS アルゴリズムの有用性

論文の参照関係を解析するために HITS アルゴリズムが適切であると考えられる理由として以下の三点があげられる．

- 構造を重視した解析
  HITS アルゴリズムで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．そのため，論文と Web ページの内容の差を考えずに適用できる．
- 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．これは論文においても参照構造から同様の収集方法が取れる．
- 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．

4.2.2 base 集合の作成

論文検索では検索ワードによって得られた上位数件の中でも，研究内容に大きな違いがあることが多い．よって本研究では root 集合にあたる部分をユーザが興味を持った論文を入力する．続いて root となった論文のリファレンス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リファレンス情報から論文を集合に追加していく．これを，論文のリファレンスが辿れなくなるか，集合に含まれる論文数が一定値にな

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．図 3 に動作を示す．

```
root：入力された論文
b_n：論文の集合
n：base 集合の最大数
A_nn：隣接行列

b_0 = root

for i=0···n
  b += b_i が参照している論文

for i=0···n
  for j=0···n
    if b_i が b_j を参照している
      A_ij = 1
    else
      A_ij = 0
```

図 3 BASE 集合の作成

4.2.3 authority と hub の算出

作成された隣接行列より各論文の authority と hub の値を求める．論文の参照構造は Web の参照構造と違い，半順序関係で表され，参照のループが存在しない．そのため，値のやり取りが一方通行になるため，数値がある一点に集中してしまう．そのため本手法では式 (1)(2) による反復計算を行い，十分な結果が得られた時点で計算を打ち切り，各論文の値とする．ここでは打ち切りの条件として，hub の最大値と，2 番目の値との比が 2:1 以上に開いたら終了としている．図 4 に動作を示す．

```
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

  for i=0···n
    for j=0···n
      hub_i += A_ij * auth_j

if k≥2 & hub_max/hub_max-1 >2
  end
```

図 4 論文の authority と hub の算出

5. 論文検索支援システムの実装

5.1 処理の流れ

本実装は第 4 章の手法に従い以下のような流れで処理を行う．

(1) ユーザの興味がある論文を入力し，root とする．
(2) root から参考文献を追加して，base 集合とする．
(3) base 集合から隣接行列を作成する．
(4) 隣接行列に対し HITS アルゴリズムを適用し，重要論文を発見，提示する．
(5) 重要論文までの参照の過程を図として出力する．

5.2 参照関係の抽出

まずは論文のデータベースから情報を取り出す．本実装では ACM Portal の論文データベースを利用する．ACM Portal では論文ごとの記事がひとつのページになっている．この

ページを解析し情報を取得する．今回は重要論文の発見するためにタイトル情報，発表された年代，リファレンスという３つの情報の取得を行った．

論文のページでは，タイトルやリファレンスなどの情報が規則的に並んでいるため，それらを順番に取得していく．

5.3 参照関係の解析

取得した論文情報を BASE 集合に入力していき，一定数になったところで HITS アルゴリズムによる解析を行う．ここで各論文の Authority 値と hub 値を求め，それぞれの値から重要論文を提示する．

5.4 論文の参照図の出力

提示された重要論文から root となる論文までにたどる過程の全てを木構造として出力する．隣接行列を転置するこで被参照に関する隣接行列を作成することができるため，重要論文から root となる論文までの参照を再帰的に木構造に格納していく．これを表示することによって研究の遷移を視覚化することができる．

6. 評　　価

6.1 重要論文の発見についての評価

ここではいくつかの分野の論文について本システムを適用し，実際に重要論文が発見できるかどうかの実験を行った．

実験では，ACM Portal における Citation Count が多い重要論文を利用し，その論文を含む BASE 集合を作成するような論文を数件選ぶ．選ばれた論文に対し本システムを適用することで，元となった重要論文を発見することができるかどうかを評価とした．

“Marching cubes: A high resolution 3D surface construction algorithm”

画像処理の分野の論文である “Marching cubes: A high resolution 3D surface construction algorithm” が発見できるか実験を行う．この論文は ACM Portal において引用数 750 件，年間ダウンロード数 2021 件と非常に重要な論文であると考えられる．実験では，ACM Portal における CITED BY を辿ることで発見できる論文のいくつかを対象とした．今回の実験では “3D content-based search using sketches”，“isosurface visualization of massive datasets on commodity off-the-shelf clusters”，“Construction in Any Dimension Using Convex Hulls”，“Describing shapes by geometrical-topological properties of real functions” の４つの論文に対して本システムを適用した．結果は表 1,2,3,4 のようになった．ここで cited は参照構造内での被引用数を表し，ACM cited は ACM Portal における被引用数を表している．

表 1 “3D content-based search using sketches” からの検索結果

タイトル	cited	auth 値	ACM cited
Global and local deformations of solid primitives	9	0.27	122
Marching cubes: A high resolution 3D surface construction algorithm	4	0.18	756
Computer Vision, 1st edition	6	0.17	251
Decimation of triangle meshes	3	0.14	195
The art of computer programming, volume 3	6	0.13	1200

表 2 “isosurface visualization of massive datasets on commodity off-the-shelf clusters” からの検索結果

タイトル	cited	auth 値	ACM cited
Marching cubes: A high resolution 3D surface construction algorithm	9	0.21	756
The Design and Analysis of Computer Algorithms, 1st edition	11	0.20	1349
The art of computer programming, volume 3	9	0.16	1200
Texture and reflection in computer generated images	9	0.14	14
3 A Sorting Classification of Parallel Rendering	3	0.13	105

表 3 “Construction in Any Dimension Using Convex Hulls” からの検索結果

タイトル	cited	auth 値	ACM cited
Marching cubes: A high resolution 3D surface construction algorithm	13	0.41	756
Volume rendering	10	0.27	240
Principles of interactive computer graphics (2nd ed.)	10	0.25	314
Fundamentals of interactive computer graphics	10	0.22	263
Scan line methods for displaying parametrically defined surfaces	6	0.10	46

これらの結果から，表 1,2,3 では目的の論文を発見することができた．さらに全ての結果において，ACM Portal での引用数 100 件以上の論文を発見できている．とくに表 1 においては，取得された参照構造内では参照数が少ないにもかかわらず，目的の論文を重要とし

表 4 “Describing shapes by geometrical-topological properties of real functions” からの検索結果

タイトル	cited	auth 値	ACM cited
The contour spectrum	11	0.12	47
Primitives for the manipulation of general subdivisions and the computation of Voronoi	5	0.11	150
Optimal surface reconstruction from planar contours	6	0.10	92
Representations for Rigid Solids: Theory, Methods, and Systems	2	0.09	156
On the definition and the construction of pockets in macromolecules	4	0,09	16

て発見することができた．この結果では，参照している論文のうち 2 つが hub のスコアの上位であったためだと考えられる．この場合 HITS アルゴリズムによって重要論文を発見することができたといえる．しかし表 4 の結果では目的の論文が発見できなかった．また上位として発見された論文のうち ACM Portal における参照数が少ないものもいくつかある．この結果では反復回数は 2 回で打ち切られ，その結果の hub のスコアは表 5 のようになっていた．ここで cite は参照している論文数を表している．

表 5 “Describing shapes by geometrical-topological properties of real functions” における hub 値

タイトル	cite	hub 値
Describing shapes by geometrical-topological properties of real functions	139	0.98
Voronoi diagrams & mdasha survey of a fundamental geometric data structure	57	0.16
Using extended feature objects for partial similarity retrieval	24	0.03
Three-dimensional object recognition	16	0.02
Feature-based similarity search in 3D object databases	19	0.02

最上位となっている論文はルートとして指定した論文となっている．この論文が参照している論文数が集合内で非常に多くなっているのが結果からわかる．そのため，2 回目の反復計算の結果ですでにスコアが集中してしまい，authority の算出に影響がでたと考えられる．このような構造においては hub と authority を計算するのが難しいため，別のアルゴリズムや収束の仕方を変換させるような手法が必要であると考えられる．

“Authoritative sources in a hyperlinked environment”

Web マイニングの分野の重要論文として “Authoritative sources in a hyperlinked environment” を発見できるか実験を行った．この論文を参照構造に含む論文として今回は “Discovering authorities in question answer communities by using link analysis”，“Semi-supervised conditional random fields for improved sequence segmentation and labeling”，“Categorization of web pages - Performance enhancement to search engine” という 3 つの論文について本システムを適用した．結果は表 6,7,8 となった．

表 6 “Discovering authorities in question answer communities by using link analysis” からの検索結果

タイトル	cited	auth 値	ACM cited
Information Retrieval, 2nd edition	6	0.16	919
The weighted majority algorithm	2	0.14	146
Anomaly hierarchies of mechanized inductive inference	3	0.12	5
The use of grammatical inference for designing programming languages	3	0.12	6
A Machine-Independent Theory of the Complexity of Recursive Functions	2	0,11	102

表 7 “Semi-supervised conditional random fields for improved sequence segmentation and labeling” からの検索結果

タイトル	cited	auth 値	ACM cited
A stochastic parts program and noun phrase parser for unrestricted text	10	00.25	247
Computational vision and regularization theory	4	0.21	80
Visual reconstruction	5	0.14	144
Hypertext	7	0.13	340
Automatic sense disambiguation using machine readable dictionaries	4	0,12	84
Authoritative sources in a hyperlinked environment	3	0,12	301

表 8 “Categorization of web pages - Performance enhancement to search engine” からの検索結果

タイトル	cited	auth 値	ACM cited
A locally adaptive data compression scheme	6	0.16	57
The Psychology of Human-Computer Interaction	7	0.14	630
Hypertext	7	0.13	340
Introduction to Modern Information Retrieval	5	0.13	1275
The Design and Analysis of Computer Algorithms, 1st edition	7	0.12	1349

この実験結果では表 7 の 6 番目としてだけ目的の論文が発見された．その他の結果では，

被引用数の多い重要な論文を発見できているが，目的の論文を見つけることができなかった．これは Web マイニングという分野の特徴がでていると考えられる．この分野は比較的新しいものであり，様々な研究が組み合わさっていると考えられる．そのため，画像処理の実験例と違い，参照構造内での被参照数も特出したものがほとんどなく，authority の値も比較的差が少なくなっている．hub の値についても，１回の反復計算による変化量が少なく，反復回数が多くなっている．このような分野では被引用数の多い，重要な論文は見つけることはできるが，多少想定と違う分野になってしまうことがある．

6.2 遷移図の出力に関する評価

研究の遷移を発見するための道筋の出力に関する評価を行う．まず，表 1 における “Marching cubes: A high resolution 3D surface construction algorithm” への参照構造の遷移図を出力する．結果は図 5 となる．タイトルの右の数値は各論文の hub 値を表示している．

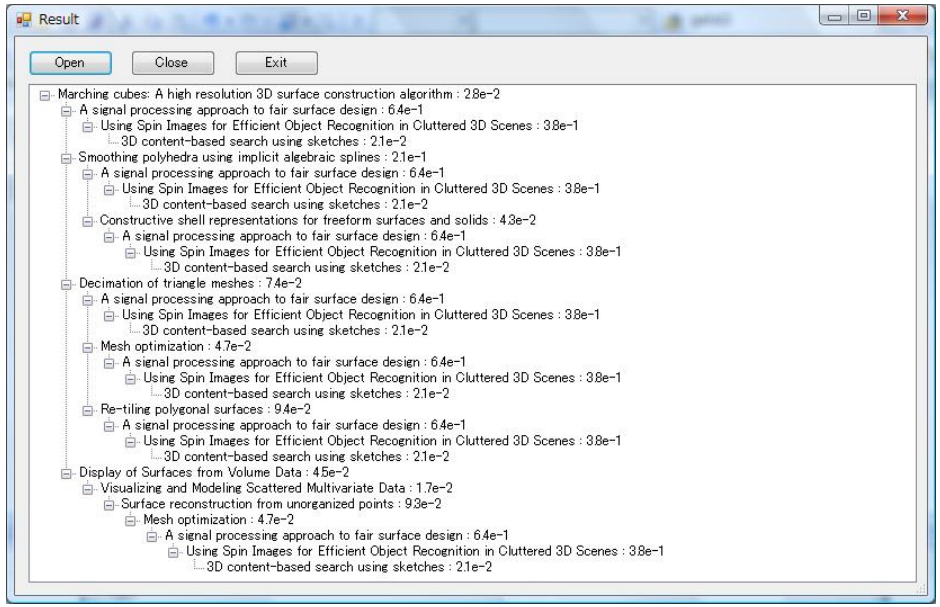

図 5 研究の遷移図

この結果から，“Marching cubes: A high resolution 3D surface construction algorithm” が上位に来た理由として，複数回登場している “A signal processing approach to fair surface design” が大きな影響を与えていると考えられる．またこの参照構造の中で，図 6 のような構造が見て取れる．これは論文の参照構造の中でよく見られる形であり，幅広い年代において参考される論文であることを示している．このような論文はその分野において非常に重要な論文である．本手法ではこのような参照をされている論文を数多く発見することができた．

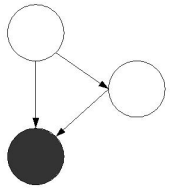

図 6 論文の参照構造の特徴形

続いて “Web usage mining with intentional browsing data” を扱う．図 7 では解析によって得られた結果の最上位であった “Virtual memory, processes, and sharing in MULTICS” に対しての道筋を表示している．

ここから Web マイニングはデータベースに関する研究が前身となっており，データベースの技術が計算機のメモリなどの有効利用のための技術ということで上位の論文を参照しているという研究の流れを理解することができる．このように，研究の遷移を知る上で本システムはとても有用であると考えられる．また，今回の例からわかるように，解析によって一見違う分野の論文が得られても，道筋を見ることにより研究の理解に貢献できると考えられる．

7. おわりに

本研究では論文の参照構造の解析によって重要論文を発見する手法を提案し，実装した．提案手法では，参照構造の解析のために Web マイニングのアルゴリズムである HITS アルゴリズムを応用した．これにより，ユーザの興味を持った論文のコミュニティ内での重要論文を提示することができた．また base 集合の作成手法により，論文検索の難点のひとつである検索ワード選択の問題を回避することができた．

近年の Web の拡大によって，検索システムや Web 上のデータベースはより大きく，よ

Vol.2009-CE-99 No.4
2009/5/23

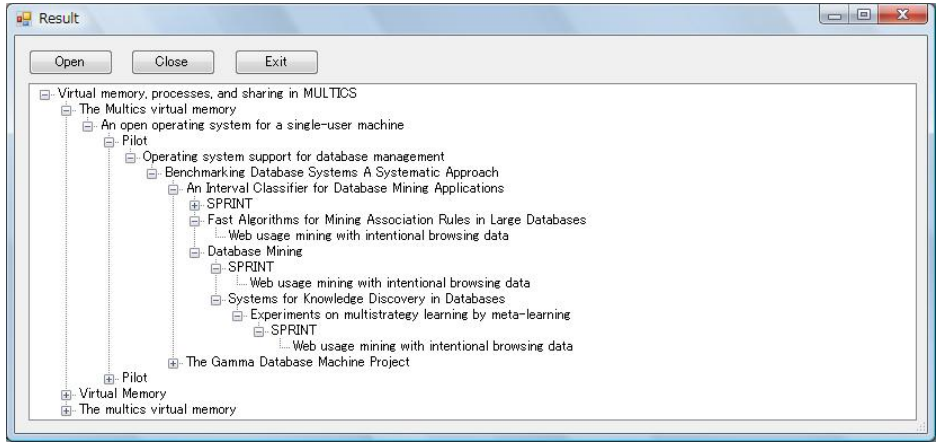

図 7 研究の遷移図

り便利になってきている．今後は今回利用した ACM Portal だけでなくより便利なデータベースが作られる可能性があると考えられる．それらのデータベースをうまく利用することでよりよい結果を得ることができると考えられる．

今後の課題として IEEE や Google Scholar などその他のデータベースへの適用がある．本実装では，ACM Portal からのデータの取得部分と，参照構造の解析部分は独立しているため，それぞれのデータベースに合わせた取得部分の作成を行うことで実装できると考えられる．

また，本手法では計算を打ち切ることで結果を取得している．これに，クラスタリングなど他の手法を利用するこで収束値を求められれば，よりよい結果が得られると考えられる．

参 考 文 献

1) The ACM Portal. http://portal.acm.org/
2) Google Scholar. http://scholar.google.co.jp/
3) Kleinberg, J.: Hubs, authorities, and communities, *ACM Computing Surveys*, vol.31, Issue 4es, Article no.5(1999).