論文のメタ情報を利用した研究履歴自動抽出・可視化システム

NGUYENMANH CUONG† 加藤 大智† 橋本 泰一†† 横田 治夫†

† 東京工業大学 大学院情報理工研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1
†† 東京工業大学 総合プロジェクト支援センター
E-mail: {cuong, kato}@de.cs.titech.ac.jp, hashimoto.t.ab@m.titech.ac.jp, yokota@cs.titech.ac.jp

あらまし 近年, インターネットを通して多くの論文が公開されている．そして, 公開された研究成果をもとに学術の研究動向を把握したいというニーズがある．しかし, 一人の研究者でも論文数が多く, かつ同時に複数の研究テーマを研究する傾向があるため, 人手で論文情報を分析し研究活動を把握するのは非常に困難である．本研究では研究者の研究履歴を自動的に抽出して可視化するシステムを提案する．研究者がインターネットに公開している論文メタ情報を収集し, それに基づいてクラスタリングを行い同じ研究テーマの論文をグループ化する．各グループに対して研究テーマを抽出し, 時系列で研究者毎に各研究テーマと研究期間を可視化する．

キーワード 研究履歴, クラスタリング

Automatic Research History Generation-Visualization System based on Metainformation

MANH CUONG NGUYEN†, Daichi KATO†, Taiichi HASHIMOTO††, and Haruo YOKOTA†

† Department of Computer Science, Tokyo Institute of Technology
2-12-1 Oookayama, Meguro-ku, Tokyo, 152-8552 Japan
†† The Research Project Support Center, Tokyo Institute of Technology
E-mail: {cuong, kato}@de.cs.titech.ac.jp, hashimoto.t.ab@m.titech.ac.jp, yokota@cs.titech.ac.jp

Abstract Recently, analyzing research papers to understand research trends has attracted attention. However, it is difficult to analyze the research papers manually. We propose a system that extract and visualize a researcher's research topics automatically from metainformation of research papers published on the internet. As an researcher's name is input, the system retrieves metainformation of research papers by that author. Based on the metainformation, it groups papers which have the same research topic, and retrieve the topic name of each group. Then it visualize the topic names and research period of those topics in a time series graph.

Key words Research History, Clustering

1. はじめに

近年，インターネットの普及に伴い，インターネットにおける電子的に公開される研究論文の数が増大し，研究者の論文や研究成果を容易に手に入れることができるようになってきた．大学や研究所などの研究機関では研究レポジトリを構築し，研究活動成果を社会に公開することが主流になりつつある．研究レポジトリは研究者の研究活動の成果である論文や特許などを配信するシステムやサービスであり，東京工業大学では T2R2 [2] という研究レポジトリを開発・公開している．しかし，現在の研究レポジトリの多くは主に研究成果のアーカイブと情報発信という機能だけを持っている．保持している研究成果を分析したり，研究者の研究の特徴や研究履歴などの情報を発信したりする機能についてはほとんど考慮されていない．

今まで入手が困難であった研究論文を網羅的に集め，俯瞰的に分析することで，学術研究の動向を把握したいというニーズも増加している．研究者が公開している論文などから，研究動向や研究経緯を抽出する研究がおこなわれていた．難波らは，引用情報を利用して科学技術の動向を可視化する [3], [4]. 吉田らは引用情報を利用して研究者の研究の経緯発展を発見する [9], [10]．しかし，これらの研究では引用情報だけを利用して研究行動や経緯を抽出している．引用情報に加えて，他の論文メタ情報も使う方が効果的だと考えられる．また，一人の研究者でも論文数が多く, かつ同時に複数の研究テーマを研究す

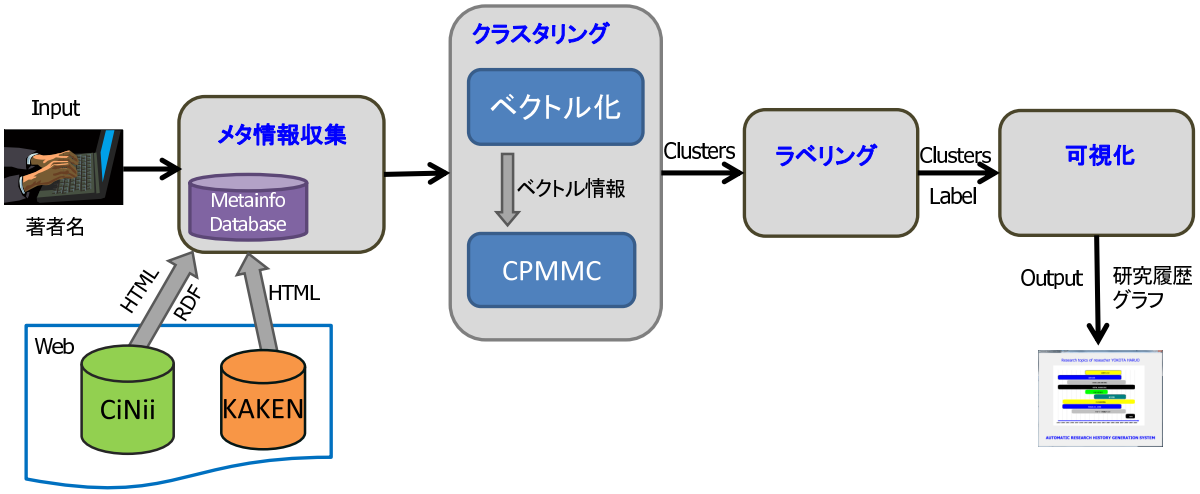

図 1　提案システムの処理フロー

る傾向があるため, 人手で論文情報を分析し研究活動を把握するのは非常に困難である.

本研究では研究者の研究履歴を自動的に抽出して可視化するシステムを提案する. このシステムは, 研究者がインターネットに公開している論文メタ情報を収集し, それに基づいてクラスタリングを行い同じ研究テーマの論文をクラスタ化する. そして, 各クラスタに対して研究テーマを抽出し, 時系列で研究者毎に各研究テーマと研究期間を可視化する. 論文のメタ情報は共著者情報, キーワード, 出版年, 引用情報, そして関連プロジェクトを利用する. これらのメタ情報を NII 論文情報ナビゲータ CiNii [1] と科学研究費補助金データベース KAKEN [7] をから取得する. そして, 同じ研究テーマの論文をクラスタ化するのクラスタリング処理はマージン最大化クラスタリング (Maximum Margin Clustering, MMC) [8] を利用する. MMC を論文クラスタリングに適用するために, 論文メタ情報のベクトル化方法と分離超平面の自動初期化手法を提案している [5], [6].

以下, 節では提案システムの概要と各処理モジュールの詳細ついて述べる. 次に, 3. 節では提案システムに対する評価実験について述べ, 実験結果について考察する. 最後に 4. 節においてまとめと今後の課題について述べる.

2. 提案システム

2.1 システム全体

研究者の研究履歴抽出・可視化システムの概要を図 1 に示す. 対象研究者の氏名を入力とし, 研究履歴を出力とする. 以下の 5 つのモジュールからシステムを構成する.

1. 論文メタ情報データベース：取得した論文のメタデータを格納するデータベース.

2. メタ情報収集モジュール：インターネットに公開されている論文のメタ情報を取得するモジュール. 取得した情報を論文メタ情報データベースに格納する.

3. クラスタリングモジュール：同じ研究テーマの論文をクラスタ化するモジュール. 論文のメタ情報を利用して論文をベクトルで表現し, MMC でクラスタリングを行う.

4. ラベリングモジュール：各論文クラスタの研究テーマ名を抽出するモジュール.

5. 可視化モジュール：研究者の研究履歴を可視化し, 出力する.

ユーザにより著者名が入力されてから研究履歴を表示するまでのシステムの処理フローは次のとおりである（図 1）. ユーザが著者名を入力した後, データベースにはその著者の論文データが存在しているかどうかを確認する. CiNii への問い合わせコストを削減するために, クラスタリング結果をキャッシュに保存する. 入力された著者名が既存データに該当する場合, 既存のクラスタリング結果を取得して, 各論文クラスタの研究テーマ名を抽出して, 研究テーマ名と研究期間を図化する. 該当しない場合は, 改めて CiNii と KAKEN からメタ情報を取得し, データベースに格納する. 次に, データベースのメタ情報を利用して, 論文をベクトル化してクラスタリングを行い, 論文をクラスタ化する. 各論文クラスタの研究テーマ名を抽出し, 最後は各論文クラスタのテーマと研究期間をグラフで可視化する.

以下, 各モジュールの詳細について説明する.

2.2 メタ情報収集モジュール

このモジュールは CiNii と KAKEN から著者の論文メタ情報を取得し, 論文メタ情報データベースに格納する. 論文メタ情報を保持するデータベースシステムは PostgreSQL を利用し

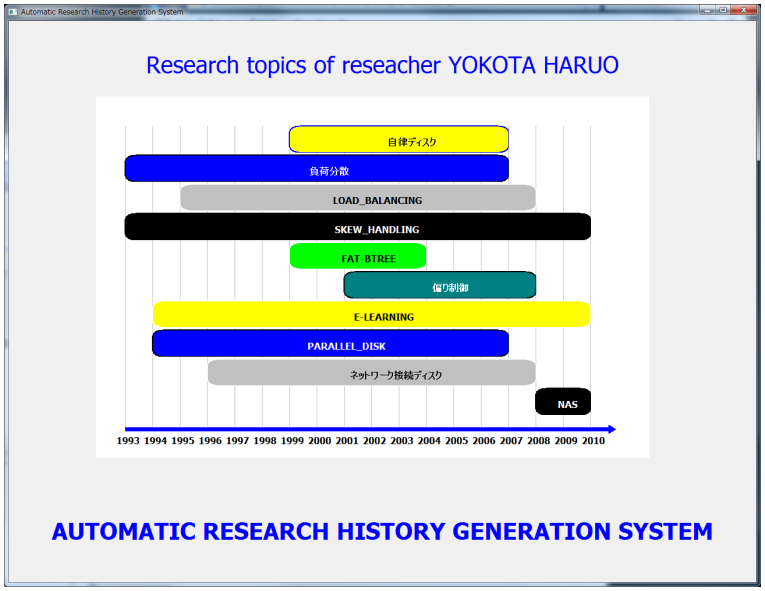

図 2　スクリーンショット

ている．データベースにはキャッシュ用の既存データと，新しい著者名が入力された毎に書き換えるテンポラリデータがある．各著者に対して論文情報と関連プロジェクト情報のテーブルがある．

CiNii [1] は，学協会刊行物・大学研究紀要・国立国会図書館の雑誌記事索引データベースなど，学術論文情報を検索の対象とする論文データベースである．KAKEN [7] は，文部科学省及び日本学術振興会が交付する科学研究費補助金により行われた研究の採択課題，研究実績報告，研究成果概要を収録したデータベースである．

メタ情報の共著者情報，キーワード，出版年，引用情報は CiNii から取得する．CiNii はウェブサービスとして提供されており，文献・研究者などについての検索が可能となっている．論文情報を取得するために，まず CiNii に HTML リクエストを送信し，著者の論文リスト情報の HTML レスポンスを受信する．このレスポンスメッセージに対してパタンマッチング解析を行い，研究者の論文 ID リストとタイトルを取得する．次に，CiNii の RDF(Resource Description Framework)API を通して論文の共著者情報，出版年とキーワード情報を取得する．各論文 ID を指定して CiNii の RDF にリクエストを送って，返されたリソースからプロパティーの共著者情報，出版年とキーワード情報を抽出する．メタ情報収集モジュールの RDF 通信部分では Apache Jena Framework を利用して実装した．

また，各論文 ID を使って CiNii にその論文紹介の HTML 情報をリクエストして，受信したこのレスポンスメッセージに対してパタンマッチング解析を行うことにより引用情報と出版年を取得する．

関連プロジェクト情報は KAKEN から取得する．KAKEN もウェブサービスとして提供されている．KAKEN から HTML レスポンスを受信し，研究課題とそれに対応する発表文献（論文）を取得する．このうち CiNii に収録されている論文については，研究実績報告や研究成果概要のページの「発表文献」セクションにその URL が記載されているため，その論文に関連するプロジェクトとして研究課題を登録する．なお，1 つの論文に対し複数の研究課題が対応づけられている場合，そのうち 1 つを任意に選択する．

また，CiNii への問い合わせコストを削減するために，クラスタリング結果をキャッシュに保存する．

2.3　クラスタリングモジュール

データベースから研究者の論文集合のメタ情報を取得し，各論文をベクトル化する．そして，ベクトル情報をテキストファイルに保存する．MMC クラスタリングエンジンはそのテキストファイルを読み込んで，クラスタリングエンジンではベクトル情報を基づいてマージン最大化クラスタリングを行い，研究テーマの同じ論文をクラスタ化する．クラスタリング結果となった論文クラスタの集合はラベリングモジュールに渡される．

2.3.1　ベクトル化

論文は，著者情報，発表年，キーワード，引用情報，そして関連プロジェクト情報の 5 つの属性により表現する．メタ情報だけに注目して使うのは，インターネットからすべての論文の内容を取得するのは困難である．なぜならば，著作権問題により公開されていない論文が数多くあるためである．

論文は，以下のベクトル v によって表現する．

$$v = \{k_1, ..., k_j, c_1, ..., c_k, y_1, ..., y_l, p_1, ..., p_m, as_1, ..., as_n\}$$

ここで，k_i はすべての論文における各キーワードの出現頻度とする．同様に，c_i はすべての論文における各引用論文の出現頻度とし，p_i はすべての論文における各プロジェクトの出現頻度とする．著者情報の素性には，3 名の共著者セットを含む論文数を as_i とする．ここで，3 名の共著者セットとは，ある論文で共著関係にある三名の著者の三つ組である．著者が三名以上の論文の場合には，すべての三組を計算する．著者が二名以下の論文場合は，as_i の値はすべて 0 とする．

出版年情報の素性は，すべての論文の中で，最初に出版された論文の出版年をオフセット Y_0 とし，次のように定義する．

$$y_i = Y_i - Y_0 \quad (1)$$

2.3.2 マージン最大化クラスタリング

クラスタリング処理ではベクトル集合上で学習して, 近いベクトル同士をクラスタ化する．ここで, 複数クラスタに分けるのマルチクラスクラスタリングを行うためには二値クラスタリングを繰り返す．その際, 結果のクラスタ集合の中から論文数が一番多いのクラスタを次の二値クラスタリングの対象とする．

二値クラスタリングでは CPMMC を利用する．CPMMC とは、Bin Zhao ら [12] が提案した Cutting Plane アルゴリズムを使ってマージン最大化クラスタリング（MMC）を高速化した手法である．CPMMC では分離超平面に近いデータのみに基づいて新しい分離超平面を計算する．計算する際に Constrained Concave-Convex Procedure (CCCP) [11] を用いて最適化問題を解くことにより新しい分離超平面を求める．CPMMC は，Cutting Plane アルゴリズムと CCCP を組み合わせることにより従来よりも高速に分離超平面を求めることが可能になっている．

しかし，CPMMC を利用して論文データをクラスタリングするには分離超平面を適切に初期化する必要がある．通常では分離超平面がランダムに初期化されるが，初期状態で分離超平面に近いデータ（制約）が存在しない，または初期状態においてすべてのデータが制約となったという 2 つの場合には分離超平面を正しく探索できない．初期状態で制約となるデータの割合を制御することにより分離超平面を自動的に初期化する [5], [6].

CPMMC プロセスは Matlab で実装した．クラスタリングエンジンでは Java から Matlab の CPMMC プログラムを起動す．CPMMC は以前ベクトル化処理で保存したベクトル行列のファイルを読み込んで，学習プロセスを動かす．CPMMC が終了したとき，分類結果をテキストファイルに保存し，Matlab を終了する．クラスタリングエンジンではその分類結果のテキストファイルを読み込んで，論文集合を実際にクラスタに分ける．

2.4 ラベリングモジュール

ラベリングモジュールではクラスタリングの結果となった各論文クラスタに対して，そのクラスタの研究テーマ名を決定する．各クラスタに対して，クラスタに含まれた論文のキーワード情報をデータベースから抽出する．すべての論文のキーワード集合の中から出現頻度が一番高いキーワードをそのクラスタの研究テーマ名にする．そして、論文のクラスタ情報と各クラスタの研究テーマ名を可視化モジュールに渡す.

2.5 可視化モジュール

このモジュールは図 2 のように, 抽出できた各研究テーマとその研究期間を表示する．まず，分類された各論文クラスタに対して，そのクラスタに含まれた論文の出版年をメタ情報データベースからを取得する．そして，そのクラスタの最初と最後の出版年を計算し，それをそのクラスタの研究期間とする．この研究期間と渡されたそのクラスタの研究テーマを可視化する．

可視化する際に，Java の SWT(Standard Widget Toolkit) と Jface Tookit を利用した．各論文クラスタには色別のブロックで表現する．ブロックの始まりと最後の位置をそのクラスタの最初と最後の出版年に応じて描く．

3. 評価実験

「横田治夫」と「喜連川優」を著者として含む論文（それぞれ論文数が 201 本, 430 本）を CiNii [1] より収集した．収集した論文に対して人手により研究テーマ別に論文を分類し，システムで実際に分類を行った結果と比較した．人手による研究テーマの分類をより正確にするために，本研究の著者の研究分野と近いの 2 人の著者を対象とした．クラスタリング結果を評価するにはエントロピー（*Entropy*）と純度（*Purity*）を用いた．エントロピーと純度の定義は [13] の定義を利用している．エントロピーはクラスタリング結果の同一クラスタに対する複数の研究テーマの混ざり具合を表し，低ければ低いほど複数の研究テーマが混在しないクラスタが多いことを表す．純度はクラスタ内で最も多い研究テーマの論文の割合を表し，1 に近ければ近いほど単一の研究テーマのクラスタが多いことを表す．

分類結果を分析すると，研究テーマ抽出結果には以下の 4 つのタイプがある．1，著者名が指定されたとき，その研究者の研究テーマの中，実際にラベルと期間を抽出することができている．2，ラベルを抽出ことができたが，期間が間違っている．3，正解セットの研究テーマの中，ラベルが抽出されなかった．4，同一の研究テーマのが複数のクラスタに分けられる．以下は「横田治夫」と「喜連川優」の分類結果においてこの傾向を詳しく分析する．

3.1 著者「横田治夫」に対する実験

「横田治夫」著者に対する人手によって分類した研究テーマ集合とシステムが出力した研究テーマ集合を表 1，2 に示す．また，各クラスタにおける研究テーマの論文数を表 3 に示す．分類タイプ 1 の研究テーマ数が 1，タイプ 2 のテーマ数が 2，タイプ 3 のテーマ数が 6，タイプ 4 のテーマ数が 1 となっている．「横田治夫」著者に対する分類結果ではエントロピーが 0.412 で，純度が 0.583 となった．

正解セットの研究テーマの中,「自律ディスク／“ 自律ディスク ”(#9)」,「e-ラーニング／“ E-Learning ”(#8)」のテーマが実際に抽出されている.「自律ディスク」に対してはラベルと研究期間が正しいが,“ E-Learning ”は研究期間が間違っている．分類結果におけるクラスタ 8（#8）では「e-ラーニング」の論文数の割合が多く，キーワード出現頻度が一番高いため，クラスタ 8 のラベルが「e-ラーニング」に関連するラベルとなった．しかし，このクラスタには他のテーマの論文も混在し，これらの論文の出版年が「e-ラーニング」の期間に入らないため，クラスタ 8 の研究期間が「e-ラーニング」の期間と異なった．

また，ラベルが抽出されなかったものが存在する．例えば,「並列論理型言語」,「Web」,「XML」がある.「並列論理型言語」の論文はすべてクラスタ 1 に集中しているが，クラスタ 1 では「負荷分散」の論文からの“ 負荷分散 ”というキーワードが多く，ラベルが“ 負荷分散 ”となった．テーマ「Web」は主にク

表 1 「横田治夫」の人手による分類結果

研究テーマ名	期間	分類タイプ
負荷分散	1993 - 2008	4
自律ディスク	1999 - 2007	1
FAT-BTREE	1997 - 2007	2
e-ラーニング	2002 - 2008	2
Web	2002 - 2008	3
並列論理型言語	1994 - 1998	3
アクティブデータベース	1994 - 2008	3
冗長ディスクアレイ	1993 - 1997	3
XML	2003 - 2006	3
リサーチマイニング	2004 - 2005	3

表 2 「横田治夫」のシステム出力

研究テーマ名	期間
負荷分散 (#1)	1993 - 2007
自律ディスク (#9)	1999 - 2007
Fat Btree(#10)	1999 - 2004
E-Learning(#8)	1994 - 2010
Skew Handling(#4)	1993 - 2010
Parallel Disk(#7)	1994 - 2007
偏り制御 (#5)	2001 - 2008
Load Balancing(#6)	1995 - 2008
ネットワーク接続ディスク (#3)	1996 - 2008
NAS(#2)	2008 - 2010

表 3 「横田治夫」のクラスタリング結果

研究テーマ	クラスタ									
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
冗長ディスクアレイ	4	1	0	0	0	0	0	0	0	0
負荷分散	12	2	5	1	0	4	0	0	6	10
並列論理型言語	6	0	0	0	0	0	0	0	0	0
アクティブデータベース	0	1	1	0	0	2	3	1	0	0
FAT-BTREE	1	0	4	2	0	0	0	1	0	18
自律ディスク	0	0	0	0	0	0	1	0	26	1
e-ラーニング	0	3	0	3	1	8	0	9	0	0
Web	0	1	0	0	5	13	0	0	0	0
XML	0	0	0	0	0	5	0	0	0	0
リサーチマイニング	0	0	0	0	2	0	3	0	0	0
Others	12	3	0	3	6	3	2	3	0	2

ラスタ 5（#5）とクラスタ 6（#6）で集中しているが，クラスタ 5 ではキーワード“ 偏り制御 ”が一番多く，クラスタ 6 ではキーワード“ Load Balancing ”が一番出現頻度が高く，クラスタ 5 と 6 のラベルが“ 偏り制御 ”と“ Load Balancing ”となり，「Web」のラベルを抽出できなくなってしまう．同様の理由で，テーマ「XML」のすべての論文がクラスタ 6 にまとまったが，「XML」のラベルを抽出できなくなってしまう．

そして，システムの出力には同一の研究テーマのが複数のクラスタに分けられる場合もある．例えば,「負荷分散」の研究テーマが“ 負荷分散 ”,“ Load Balancing ”,“ Skew Handling ”と“ 偏り制御 ”の 4 つのクラスタに分けられる．クラスタ 1（#1）ではキーワード“ 負荷分散 ”の出現頻度が一番高く，このクラスタのラベルがそのキーワードになった．クラスタ 4（#4）では“ Skew Handling ”，クラスタ 5 では“ 偏り制御 ”，クラスタ 6 では“ Load Balancing ”の「負荷分散」に関連するキーワードの出現頻度が高く，これらのクラスのラベルがそれぞれのキーワードとなった．つまり，クラスタリングの結果が適切ではなかったため，正しく抽出されるテーマも複数存在し，期間が間違ったものやラベルが抽出されないもの，複数のテーマとして抽出されたものも存在する．

3.2 著者「喜連川優」に対する実験

著者「喜連川優」に対する人手によって分類した研究テーマ集合とシステムが出力した研究テーマ集合を表 4，5 に示す．また，各クラスタにおける研究テーマの配分状況を表 6 に示す．著者「横田治夫」の分類結果と同様に，著者「喜連川優」の分類結果においても同様な傾向が見られる．分類タイプ 1 の研究テーマ数が 0，タイプ 2 のテーマ数が 4，タイプ 3 のテーマ数が 7，タイプ 4 のテーマ数が 2 となっている．著者「喜連川優」に対する分類結果ではエントロピーが 0.420 で，純度が 0.520 となった．

正解セットの研究テーマの中,「DB Structural Degradation ／“ Structural Deterioration ”」,「Load Balancing ／“ Load Balancing ”」,「Database, Data mining ／“ Data Mining ”」,「Large Scale PC Cluster ／“ PC Cluster ”」のテーマが実際関連するラベルが抽出されている．しかし，これらのテーマの研究期間が間違っている．これは，それらのラベルとなったクラスタ 1，5，11，12 では他のテーマの論文も混在するため，計算するとき研究期間が違ってくるからである．

また，ラベルが抽出されなかったテーマが存在する．分類結果では，テーマ「Power Saving」,「IP-SAN」のラベルが抽出されなかった．テーマ「Power Saving」はクラスタ 1（#1）とクラスタ 5（#5）に主に分散しているが，クラスタ 1 では「DB Structure Degradation」の論文数が圧倒的に多く，クラスタ 1 のラベルがこのテーマのキーワードである“ Structural Deterioration ”となった．また，クラスタ 5 では「Parallel Operation」の方の論文数が多く，出現頻度の高いキーワード“ Load Balancing ”がこのクラスタのラベルになった．このため,「Power Saving」のラベルを抽出できなかった．また，クラスタ 8（#8）にも“ 負荷分散 ”のキーワードが多く出現しこのクラスタのラ

表 4　著者「喜連川優」の人手による分類結果

人手による研究テーマ名	期間	分類タイプ
DB Structure Degradation	2004 - 2007	2
Load Balancing	1993 - 2011	2
Database, Datamining	1989 - 1995	2
Large Scale PC Cluster	1997 - 2001	2
Natural Langage Operation	2006 - 2010	3
Web Mining	2001 - 2010	4
Functional Disc System	1986 - 1991	3
Earth Environment Data Archive	1987 - 2010	3
IP-SAN	2001 - 2007	3
Parallel Database	1994 - 2004	3
Parallel Operation	1989 - 2005	4
Information Explosion	2006 - 2011	3
Power Saving	2007 - 2011	3

表 5 「喜連川優」のシステム出力

システムの出力	期間
Structural Deterioration(#1)	1996 - 2011
Load Balancing(#5)	1986 - 2008
Data Mining(#11)	1987 - 2011
PC Cluster(#12)	1988 - 2011
PC クラスタ (#13)	1997 - 2010
Web Community(#6)	1993 - 2010
ウェブコミュニティ(#7)	1995 - 2006
データマイニング (#2)	1986 - 2011
相関ルール (#3)	1986 - 1997
Database(#4)	1990 - 1995
負荷分散 (#8)	1989 - 2011
Association Rule(#9)	1995 - 2006
2 次記憶装置 (#10)	2000 - 2010

表 6 「喜連川優」のクラスタリング結果

研究テーマ	クラスタ												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13
Functional Disc System	1	0	0	0	0	0	0	0	0	7	0	0	0
Earth Environment Data	0	1	3	12	0	4	5	0	0	6	0	0	0
Database, Datamining	4	1	8	0	2	0	0	0	0	2	13	0	0
Parallel Operation	4	4	5	5	8	0	3	6	4	9	2	1	9
Load Balancing	6	6	1	2	2	0	1	0	3	0	0	0	0
Parallel Database	2	0	2	2	2	0	0	0	0	0	0	0	2
Large Scale PC Cluster	0	0	0	0	0	0	5	1	0	0	0	6	5
IP-SAN	0	1	0	0	0	8	7	0	0	0	0	3	4
Web Mining	0	9	5	2	0	9	16	0	4	0	0	13	7
DB Structure Degradation	10	0	0	0	4	0	0	0	0	0	0	0	0
Natural Langage Operation	0	0	0	0	0	0	6	2	0	0	0	0	1
Information Explosion	1	0	1	3	0	0	1	1	8	0	0	0	0
Power Saving	3	0	0	1	2	0	0	0	0	0	0	0	0

ベルが“ 負荷分散 ”となった．同様に，テーマ「IP-SAN」は主にクラスタ 6（#6）とクラスタ 7（#7）に集中するが，両方とも論文の数は「Web Mining」より少ないため，キーワードの出現頻度がより低く，ラベル抽出ができない．

また,「Web Mining」が“ Web Community ”,“ ウェブコミュニティ ”,“ 相関ルール ”と“ Association Rule ”の 4 つのクラスタに分けられるように，システムの出力には同一の研究テーマのが複数のクラスタに分けられる場合もある．分類結果において,「Web Mining」の論文が複数のクラスタに分散している．クラスタ 6 と 7 のでは他のテーマの論文も混在するが,「Web Mining」の論文数の方が多いためキーワードの出現頻度が高く，ラベルが「Web Mining」のキーワードである“ Web Community ”と“ ウェブコミュニティ ”となった．クラスタ 9（#9）ではキーワード“ Association Rule ”の出現頻度が高く，このクラスタのラベルが“ Association Rule ”となった．クラスタ 4（#4）では「Earth Environment Data」の論文数が多く，そのテーマのキーワードである“ 相関ルール ”のラベルとして抽出された．

3.3　実験のまとめ

評価実験では，著者「横田治夫」と「喜連川優」の分類結果にはすべて同じ傾向がみられる．著者名が指定されたとき，その研究者の研究テーマと研究期間を実際に抽出することができている．しかし，ラベルが抽出されるが期間が間違うテーマも存在する．これは，クラスタリング処理では比較的に綺麗にクラスタされ，ある程度論文が 1 つのクラスタに集中するが，他のテーマの論文も多少混在して，研究期間が異なるからである．また，研究テーマの期間は論文クラスタの最初と最後に出版された論文の年を利用してそれを研究期間としているため，その最初と最後の年の間に研究活動が行われない期間がある場合にはそれを再現できていない．

そして，正解セットの研究テーマの中，ラベルが抽出されなかったものが存在する．これは，クラスタリングを行うとき，研究テーマが 1 つのクラスタに固まったがそのクラスタにより大きいの別のテーマも混在するため，ラベル抽出する際により大きいの研究テーマのキーワード出現頻度がより高く，そのテーマのラベルになってしまう．このため，論文の数が少ない研究テーマのラベルが抽出できない．

また，システムの出力には同一の研究テーマが複数のクラスタに分けられる場合もある．これは，クラスタリングを行うとき，研究テーマが複数のクラスタに分散したからである．1 つ

の研究テーマが複数のクラスタに分散されたが，各クラスタにおいてその研究テーマが他に混在しているテーマより論文の数が多く，自分を持っているキーワードの出現頻度がより高いため，複数のクラスタのラベルがそのテーマに関連するものとなる．

これらの問題点を改善するには，クラスタリング精度をさらに向上する必要である．また，研究テーマ名をより正確に抽出するためにラベリング方法を改善する必要である．

4. まとめと今後の課題

本研究では研究者の研究履歴を自動的に抽出して可視化するシステムを提案した．研究者の名前を入力してから, 自動的にインターネットから論文情報を取得し, クラスタリングを行い研究者の研究履歴を可視化することができている．しかし，研究テーマが抽出できないテーマもある．これは, クラスタリングの精度とラベリングの抽出方法として単純にキーワードの出現頻度だけを考慮しているため, 最終的の結果に影響している．

今後の課題の 1 つとしてはクラスタリング精度を向上することである．キーワードを含まない論文にコンテンツからキーワード情報を抽出することにより，クラスタリング精度を上げることが考えられる．また，メタ情報のベクトル化方法を改善することが考えられる．

もう 1 つの課題としてはラベリングの抽出する方法を考慮することである．現在，研究テーマ名の抽出にはキーワード出現頻度だけを考慮しているが，同意味の別の用語を異なるキーワードとして扱っている．辞書などを用意して同意味の用語を同一のキーワードとして扱うことにより，研究テーマ名をより正しく抽出できるかどうかを確認することが今後の課題である．

謝辞 本研究の一部は，日本学術振興会科学研究費補助金基盤研究 (A)(#22240005) の助成により行われた．

文　献

[1] Scholarly and Academic Information Navigator CiNii. http://ci.nii.ac.jp/.

[2] Tokyo Institute of Technology Research Repository T2R2. http://t2r2.star.titech.ac.jp/.

[3] Yuki Kondo, Hidetsugu Nanba, Takanori Okurmura, Akihiro Nimori, Hidekazu Yatsukawa, and Yasuyama Suzuki. Retrieve research trend information from research database. In *The Association for Natural Language Processing 13th Forum*, 2007.

[4] Hidetsugu Nanba and Yoshiko Yaguchi. Retrieve and visulize research trend information from research database. In *The Association for Natural Language Processing 12th Forum*, 2006.

[5] Manh Cuong Nguyen, Daiichi Kato, Taiichi Hashimoto, and Haruo Yokota. Automatic generation of a researcher's research history using meta informations of research papers. In *The 3rd Forum on Data Engineering and Information Management*, 2011.

[6] Manh Cuong Nguyen, Daiichi Kato, Taiichi Hashimoto, and Haruo Yokota. Research history generation using maximum margin clustering of research papers based on metainformation. In *The 13th International Conference on Information Integration and Web-based Applications and Services*, 2011.

[7] National Institute of Information. KAKEN. http://kaken.nii.ac.jp/.

[8] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In *Neural Information Processing Systems*, 2004.

[9] Makoto Yoshida, Takashi Kobayashi, and Haruo Yokota. Comparison of the research mining and the other methods for retrieving macro-information from an open research-paper db. *Information Processing Society of Japan: Database*, Vol. 45, No. SIG7(TOD22), pp. 24–32, 2004.

[10] Makoto Yoshida, Takashi Kobayashi, and Haruo Yokota. Consideration of the clustering threshold in the research mining method. In *Information Processing Society of Japan: Database Management System (2004-DBS-134(II))*, pp. 553–560, 2004.

[11] Alan L. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, Vol. 15, pp. 915–936, 2003.

[12] Bin Zhao, Fei Wang, and Changshui Zhang. Efficient maximum margin clustering via cutting plane algorithm. In *SIAM International Conference on Data Mining*, pp. 751–762, 2008.

[13] Ying Zhao and Geoge Karypis. Criterion function for document clustering. Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 2003.