論文のメタ情報を利用した研究者の研究履歴の自動生成

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 工学部 情報工学科
††† 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り，論文情報から，研究者の研究履歴を自動的に生成する（リサーチマイニング）研究が行われている．既存の研究では，論文の引用情報を用いて関連論文集合を発見する．本研究では，引用情報に加えて，共著者，発表年，関連プロジェクトなどのメタ情報を利用して論文をクラスタリングし，研究者の研究履歴を自動的に生成する手法を提案する．

キーワード 研究履歴, クラスタリング, K-平均法

Automatic Generation of a Researcher's Research History using Meta Informations of Research Papers

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
†† Department of Computer Science, School of Engineering, Tokyo Institute of Technology
††† 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

Key words Research mining, Clustering, K-means

1. はじめに

近年，ネットワーク技術の発達，情報インフラの普及に伴い，電子的に閲覧可能な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．一般に公開された研究成果をもとに，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．しかし，研究者はある研究テーマに関する活動を比較的長期間行うため，同一テーマについての論文などの研究成果は数が多い．また，同時に複数の研究テーマや研究プロジェクトに携わることもあり，研究者の研究分野や研究活動の履歴を人手で把握したり分析したりするには多大なコストが必要である．

このため，論文間の類似関係を自動的に計算し研究動向を分析する研究が行われている．論文の引用関係を利用した書誌結合 (bibliographic coupling) や共引用分析 (co-citation analysis) がその例である．書誌結合は，参照・被参照関係にある論文は同じ主題を扱っているという仮定のもと，2 つの論文間の関連度を参照論文の重複数を基に計算する．難波らは，参照の仕方を考慮した書誌結合による論文の類似度計算手法を提案している [3]．一方，引用分析はある論文が他の論文に共に引用されている回数を論文の類似度として分析を行う．また，論文の付与されたキーワードを利用した研究動向分析手法も提案されている [1]．

日本をはじめ世界では，大学などの研究機関における研究活動成果を一般に公開することが主流になりつつある．そのような活動の一つに研究レポジトリがある．研究レポジトリは論文や特許といった研究者の研究活動の成果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 [10] という研究レポジトリを公開・開発している．しかし，現在の研究レポジトリの多くは研究成果のアーカイブと情報発信に特化しており，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考慮されてい

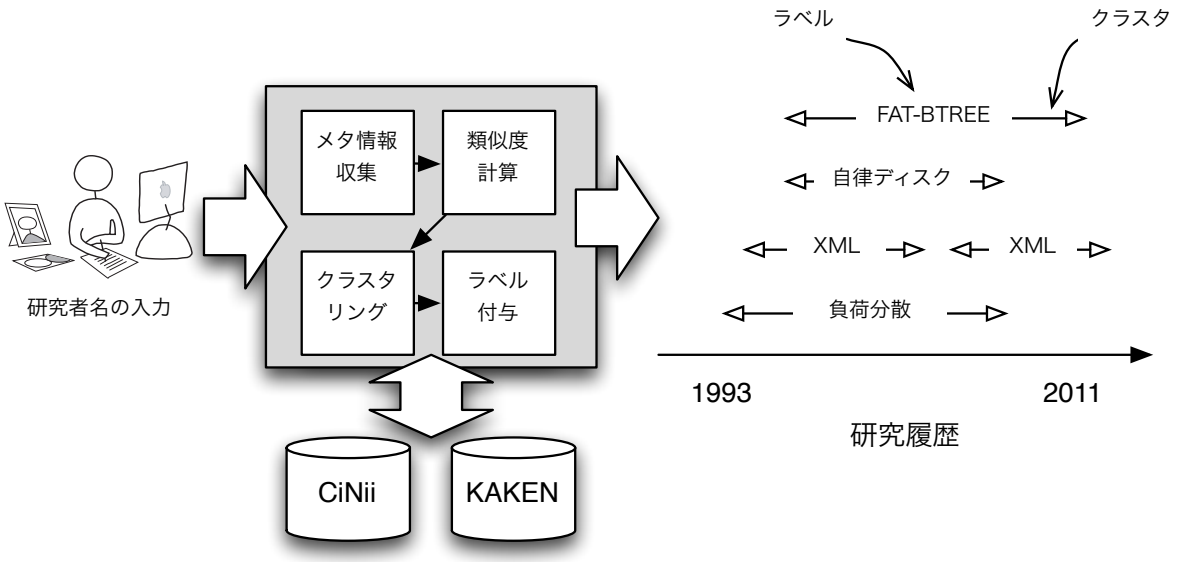

図 1　提案手法の概要図

ない．

論文などの研究成果から，研究動向や研究者の研究履歴を自動的に生成する研究が行われている．このような研究はリサーチマイニングと呼ばれている．引用情報を利用して科学技術の動向を可視化する手法 [4], [11]，研究者の研究の経緯発展を発見する手法 [7], [8] が提案されている．

従来の研究では，引用情報のみを用いて論文の類似度を計算している．しかし，論文間の関係を発見するには，引用情報の他に著者名，出版年，キーワードなどのメタ情報も利用できると考えられる．本研究では，引用情報に加えて，共著者，出版年，キーワード，関連プロジェクトなどの論文のメタ情報を利用して，論文をクラスタリングし，生成されたクラスタを時系列に可視化することにより，研究者の研究履歴を明らかにする手法を提案する．

以下，2. 節で関連研究について述べ，3. 節では提案手法の全体の処理について説明する．そして，4. 節において複数のメタ情報を利用した論文間の類似度の計算方法について述べ，5. 節と 6. 節では論文のクラスタリング手法と初期クラスタについて説明する．次に，7. 節では提案手法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 8. 節においてまとめと今後の課題について述べる．

2. 関 連 研 究

論文などの研究成果から，科学技術の動向や研究者の研究履歴を自動的に生成する研究が行われており，このような研究はリサーチマイニングと呼ばれている．

研究分野を特定して科学技術動向に関する情報を抽出し可視化する手法が提案されている [4]．難波らは，引用情報を用いて特定分野の論文を収集し，収集した論文の表題から要素技術用語に関する情報を抽出する．そして，要素技術の変化を年次表示することにより，特定の分野や研究テーマにおける科学技術の動向を可視化している．

また，特定の研究者の研究動向を可視化する手法も提案されている [7], [8]．吉田らは，論文の引用情報を用いて論文の発展経緯をグラフ化し，そのグラフをクラスタリングすることで，研究者の研究履歴を可視化する．

これらの 2 つの手法の共通点は，論文間の関係を発見するには引用情報のみを利用していることである．しかし，論文間の関係性を表す情報としては引用情報以外にも，著者，出版年，キーワード，関連プロジェクトなどのメタ情報がある．これらのメタ情報を利用することで，これまで発見できなかった論文の関連性を見つけることが可能になると考えられる．

3. 論文のメタ情報を利用した研究履歴の生成

この研究では，特定の研究者の論文情報から研究履歴を自動的に生成することを目的とする．研究履歴に生成方法についての概要図を図 1 に示す．ユーザは研究者名を入力し，システムは入力された研究者を著者に含む論文を収集する．次に，収集した論文をクラスタリングし，生成されたクラスタに研究テーマのラベルを付与する．最後に，クラスタを時系列に可視化することにより，研究者の研究履歴を生成する．

研究者の研究履歴の生成は，以下の 5 つのステップである．

Step 1. 特定研究者の論文のメタ情報を収集する．
Step 2. 収集した論文に対して，論文間の類似度を計算する．
Step 3. Step 2. の類似度をもとにクラスタリングを行う．
Step 4. Step 3. で作成されたクラスタにラベルをふる．
Step 5. Step 4. の結果を時系列に可視化する．

本論文では，Step 3 の論文のクラスタリング手法の詳細と手法の評価実験について述べる．

4. 論文の類似度

論文は，次の 5 つの属性により表現する．

- 著者情報

- 発表年
- キーワード
- 引用情報
- 関連プロジェクト情報

論文の類似度は各属性の類似度の線形結合と定義する（式 1）.

$$\begin{aligned} Sim(P_a, P_b) = {} & \alpha\ Sim_a(P_a, P_b) + \beta\ Sim_y(P_a, P_b) \\ & + \gamma\ Sim_k(P_a, P_b) + \delta\ Sim_r(P_a, P_b) \\ & + \epsilon\ Sim_p(P_a, P_b) \\ & \text{ただし,} \alpha + \beta + \gamma + \delta + \epsilon = 1 \end{aligned} \tag{1}$$

$Sim(P_a, P_b)$ は論文 P_a と P_b の類似度，$Sim_a(P_a, P_b)$ は著者情報の類似度，$Sim_y(P_a, P_b)$ は発表年の類似度，$Sim_k(P_a, P_b)$ はキーワードの類似度，$Sim_r(P_a, P_b)$ は引用情報の類似度，$Sim_p(P_a, P_b)$ は関連プロジェクトの類似度を表す．複数の類似度を線形結合することで，類似度の重要度をマニュアルで調整することが容易になるという長所がある．また，自動的にパラメータを調整した場合にも，考察する際にその値を直感的に理解しやすいという長所もある．

4.1 著者類似度

論文の著者情報は，著者を次元とし値を 1 としたベクトルとして表現し，これを著者ベクトルと呼ぶ．著者情報の類似度 Sim_a は，各論文の著者ベクトルのコサインを類似度と定義する．

例えば，著者情報が次のような 2 本の論文の場合，

$\boldsymbol{P}_a$: 加藤 大智，橋本 泰一，横田 治夫
$\boldsymbol{P}_b$: NGUYEN MANH CUONG，橋本 泰一，横田 治夫

著者類似度 $Sim_a(P_a, P_b)$ は，

$$Sim_a(P_a, P_b) = \frac{2}{\sqrt{3}\sqrt{3}} = \frac{2}{3}$$

となる．

4.2 発表年類似度

発表年の類似度 Sim_y は，2 本の論文の発表年の近さで定義する．

$$Sim_y(P_a, P_b) = \begin{cases} 1 & y \text{ 年以内} \\ 0 & y \text{ 年よりも大きい} \end{cases} \tag{2}$$

この論文では，$y = 2$ として実験を行った．本論文では非常に単純な手法を用いた．この定義以外の計算方法として，発表年の差の逆数を用いるなども考えられる．

4.3 キーワード類似度

論文のキーワード情報は，著者情報と同様に，キーワードを次元とし値を 1 としたベクトルとして表現し，各論文のキーワードベクトルのコサインを類似度と定義する．

例えば，次のようなキーワードが付与された 2 本の論文の場合，

$\boldsymbol{P}_a$: XML，省電力，データベース
$\boldsymbol{P}_b$: XML, データベース, Multi Processors

キーワード類似度 $Sim_k(P_a, P_b)$ は，

$$Sim_k(P_a, P_b) = \frac{2}{\sqrt{3}\sqrt{3}} = \frac{2}{3}$$

となる．

4.4 引用類似度

論文の引用情報は，著者情報，キーワード情報と同様に，引用文献を次元とし値を 1 としたベクトルとして表現し，各論文の引用文献ベクトルのコサインを類似度と定義する．

例えば，次のような引用情報を持った 2 本の論文の場合，

$\boldsymbol{P}_a$: 文献 1，文献 2，文献 3，文献 4，文献 7，文献 12
$\boldsymbol{P}_b$: 文献 1，文献 3，文献 4，文献 7，文献 9，文献 12

引用類似度 $Sim_a(P_a, P_b)$ は，

$$Sim_r(P_a, P_b) = \frac{5}{\sqrt{6}\sqrt{6}} = \frac{5}{6}$$

となる．

4.5 関連プロジェクト類似度

関連プロジェクト類似度は，同一プロジェクトの研究成果であるかどうかという観点で類似度を定義する．日本の代表的な研究費である学術研究振興会科学研究補助金では，採択された研究課題の成果報告の一部として，学術論文や発表文献について研究課題の報告書に記載する必要がある．このような同じ報告書に記載された論文を同一プロジェクトの研究成果であるとみなす．

$$Sim_p(P_a, P_b) = \begin{cases} 1 & \text{同一のプロジェクトの研究成果である} \\ 0 & \text{そうでない} \end{cases} \tag{3}$$

5. 論文のクラスタリング

クラスタリングには，K-means 法 [9] を用いた．K-means 法とは，クラスタに含まれる各データを距離が最も近いクラスタに再配分しこれを繰り返すことでクラスタを生成するアルゴリズムである．アルゴリズムの概略を下記に示す．

Step 1. すべてデータを初期クラスタとして K 個のクラスタに配置する．
Step 2. 各データに対して，クラスタに含まれるデータとの距離の平均を求める．
Step 3. Step 3. で求めた距離が最も短いクラスタにデータを再配分する．
Step 4. Step 2,3 の操作をクラスタが変化しなくなるまで繰り返す．

この研究では．データは論文とし，データ間の距離は論文の類似度の逆数とした．論文類似度が 0 の場合，距離の平均の計算には無視される．

6. 初期クラスタの生成

K-means 法において，Step1 の K 個の初期クラスタの配置がクラスタリングの性能に大きく影響を与える．そのため，初期クラスタはできるだけ類似したデータが集まっていることが望ましい．論文においては，共著者やキーワードが共通すると，その論文同士は非常に類似性が高いことが経験的に分かっている．そこで，このヒューリスティックを利用して，初期クラスタ時に類似性の高いと思われる論文を同じクラスタに集めておくようにすることで，クラスタリングの精度を向上させる．評価実験においては，ランダム，著者，キーワードの２種類の初期クラスタの生成方法を用いた．各アルゴリズムの詳細は次のとおりである．

6.1 ランダム

ランダムに K 個のクラスタに論文を配置する．

6.2 著　　者

初期クラスタにおいて，なるべく同じ著者を含む論文を同じクラスタに配置されるようにする．そのアルゴリズムを次に示す．

Step 1. ランダムに K 本の論文を選び，それぞれ別のクラスタに配置する．
Step 2. 残った論文に対して，クラスタに含まれる論文と共通する著者が少なくとも一人いるクラスタに配置する．
Step 3. Step 2. で配置できなかった論文をランダムに配置する．

6.3 キーワード

初期クラスタにおいて，なるべく同じキーワードを含む論文を同じクラスタに配置されるようにする．そのアルゴリズムを次に示す．

Step 1. ランダムに K 本の論文を選び，それぞれ別のクラスタに配置する．
Step 2. 残った論文に対して，クラスタに含まれる論文に共通するキーワードが一人いるクラスタに配置する．
Step 3. Step 2. で配置できなかった論文をランダムに配置する．

7. 評価実験と考察

7.1 論文情報の取得方法

提案手法の有効性検証のために，プロトタイプシステムを実装した．プロトタイプシステムからアクセスする外部情報源として，NII 論文情報ナビゲータ CiNii [5] と科学研究費補助金データベース KAKEN [6] を用いた．CiNii 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．また KAKEN 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．これらのデータベースからメタ情報を取得した．

表 2　人手により分類した評価実験データ

研究テーマ名	論文数	期間
負荷分散	40	1993 - 2008
自律ディスク	28	1999 - 2007
FAT-BTREE	26	1997 - 2007
e-ラーニング	24	2002 - 2008
Web	19	2002 - 2008
アクティブデータベース	8	1994 - 2008
並列論理型言語	6	1994 - 1998
冗長ディスクアレイ	5	1993 - 1997
XML	5	2003 - 2006
リサーチマイニング	5	2004 - 2005

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，レスポンスを受信する．このレスポンスメッセージを解析し，タイトル，共著者，出版年，キーワード，引用情報を取得する．

同様に 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．

7.2 評価実験

前述の方法より「横田治夫」を著者として含む論文（194 本）を収集した．収集した論文は，人手により 10 の研究テーマ（クラスタ）に分類した（表 2）．

この人手によるクラスタと提案手法により作成したクラスタを比較して，提案手法の評価を行う．評価尺度は，エントロピー（Entropy）と純度（Purity）を用いた．エントロピー（$Entropy$）と純度（$Purity$）の定義 [2] は次のとおりである．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．

$$Entropy = \sum_{r=1}^{p} \frac{n_r}{n}\left(-\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}\right) \tag{4}$$

$$Purity = \sum_{r=1}^{p} \frac{n_r}{n}\left(\frac{1}{n_r}\max_i(n_r^i)\right) \tag{5}$$

ここで，c_r は r 番目のクラスタ，p はクラスタ数，q はカテゴリ数を表す．また，n は総文書数，n_r はクラスタ c_r に含まれる文書数，n_r^i はクラスタ c_r に含まれるカテゴリ i の文書数を表す．

評価実験では，クラスタ数 k は 10 と固定し，論文類似度の線形結合のパラメータ $\alpha, \beta, \gamma, \delta, \epsilon$ と初期クラスタの生成

表 1　評価実験結果 (E: Entropy, P: Purity)

	著者のみ		発表年のみ		キーワードのみ		引用のみ		プロジェクトのみ	
	E	P	E	P	E	P	E	P	E	P
ランダム	0.532	0.272	0.718	0.249	0.695	0.253	0.706	0.252	0.718	0.249
著者	0.528	0.295	0.661	0.276	0.646	0.279	0.627	0.281	0.627	0.281
キーワード	0.522	0.326	0.648	0.309	0.632	0.311	0.652	0.310	0.648	0.309

	均等		著・キ・引を重視					
			小		中		大	
	E	P	E	P	E	P	E	P
ランダム	0.548	0.315	0.509	0.307	0.488	0.296	0.480	0.288
著者	0.560	0.327	0.533	0.319	0.504	0.311	0.497	0.305
キーワード	0.554	0.343	0.508	0.348	0.473	0.344	0.489	0.335

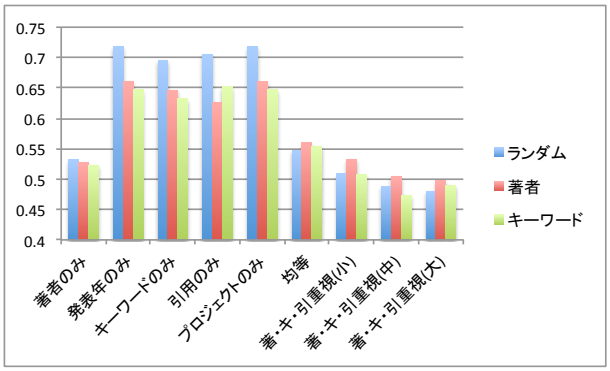

図 2　評価実験結果 (エントロピー)

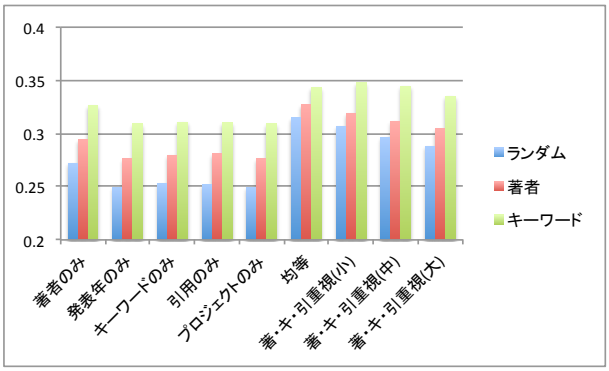

図 3　評価実験結果 (純度)

方法（ランダム，著者，キーワード）を変えて行った．パラメータは，著者類似度（α），キーワード類似度（γ），引用類似度（δ）の重みを変化させた．その比を下記に示す．

	$\alpha:\beta:\gamma:\delta:\epsilon$
著者のみ:	1 : 0 : 0 : 0 : 0
発表年のみ:	0 : 1 : 0 : 0 : 0
キーワードのみ:	0 : 0 : 1 : 0 : 0
引用のみ:	0 : 0 : 0 : 1 : 0
プロジェクトのみ:	0 : 0 : 0 : 0 : 1
均等:	1 : 1 : 1 : 1 : 1
著・キ・引を重視 (小):	2 : 1 : 2 : 2 : 1
著・キ・引を重視 (中):	5 : 1 : 5 : 5 : 1
著・キ・引を重視 (大):	10 : 1 : 10 : 10 : 1

各類似度を単独で使用した場合（著者のみ，発表年のみ，キーワードのみ，引用のみ，プロジェクトのみ），類似度を均等にした場合（均等），著者，キーワード，引用の各類似度を少し，中くらい，大きく重視するように変化させた場合（著・キ・引を重視 (小，中，大)）である．著者，キーワード，引用の類似度の重みを変化させた理由は，経験的に著者，キーワード，引用情報が共通する論文は同一の研究テーマであることが多いためである．発表年や関連プロジェクトとの重視する度合いを変化させることで，クラスタリングへの影響を実験により確認する．

初期クラスタの設定にはランダム的な要素を含むため，パラメータと初期クラスタが同一の設定で 10 回の施行を行い，エントロピーと純度の平均で評価する．

7.3　考　　察

評価実験結果（表 1，図 2，図 3）より，メタ情報をそれぞれ 1 つずつだけ使った場合（著者のみ，発表年のみ，キーワードのみ，引用のみ，プロジェクトのみ）のエントロピーと純度と，複数のメタ情報を合わせて使った場合（均等，著・キ・引を重視 (小，中，大)）を比較すると，各メタ情報を単体で利用するよりも，合わせて利用した方がエントロピが低く，純度が高いことが分かった．

均等と著・キ・引を重視 (小，中，大) のエントロピーと純度を比較すると，パラメータを均等に配分するよりも，著者類似度，キーワード類似度，引用類似度の重みを大きくすると純度とエントロピーが低下する傾向にある．つまり，論文のクラスタリングを行う上では，重視する評価指標によって，類似度の重要度のパラメータを調整する必要があることがわかった．

次に，初期クラスタの生成方法について比較する．ランダムに初期クラスタを生成した場合（ランダム）と著者をもとに初期クラスタを生成した場合（著者）では，ランダムの方が，エントロピーが低く，著者の方が純度が高い．一方，キーワードをもとに初期クラスタを生成した場合（キーワード）は，他の 2 手法に比べ，エントロピーがより低く，純度もより高い．キーワードをもとにして初期クラスタを生成する方法が最もクラスタリングの性能がよいことがわかった．

しかし，今回の評価実験では，純度は最も高くて 0.35 程度であり，クラスタに含まれる論文の半分以上が異なる研究テー

表 3　研究テーマごとの評価実験結果（著・キ・引を重視 (中), キーワード）

研究テーマ名	論文数	平均クラスタ数	第一クラスタ平均論文数	第ニクラスタ平均論文数	割合 (%)
負荷分散	40	6.0	14.0	9.9	59.7
自律ディスク	28	4.1	15.5	7.6	82.6
FAT-BTREE	26	3.7	11.0	7.0	72.0
e-ラーニング	24	4.0	12.4	7.5	82.8
Web	19	4.7	6.5	5.0	60.5
アクティブデータベース	8	3.2	4.0	2.8	84.4
並列論理型言語	6	2.5	3.6	1.9	91.6
冗長ディスクアレイ	5	2.1	3.9	1.0	97.5
XML	5	2.7	2.3	2.0	85.0
リサーチマイニング	5	1.1	5.0	0	100.0

マである．そのため，クラスタ数（K）や類似度のパラメータの調整を行い，純度の向上を目指す必要があると考えられる．

研究テーマごとにクラスタの分析を行った．最もエントロピーが低かった評価実験（パラメータ W3，キーワード）の結果について各研究テーマごとに以下の数値を算出した（表 3）．

平均クラスタ数: 10 クラスタのうち研究テーマの論文が属したクラスタ数

第一クラスタの平均論文数: 最も多く同じ研究テーマの論文が属したクラスタでの，その研究テーマの論文数

第ニクラスタの平均論文数: 二番目に多く同じ研究テーマの論文が属したクラスタでの，その研究テーマの論文数

割合: 第一クラスタと第二クラスタでの論文数の全体の論文数に対する割合

「負荷分散」「Web」は，平均クラスタ数が 4.7 以上と複数のクラスタに分散する傾向にある．しかし,「自立ディスク」や「e-ラーニング」も平均クラスタ数が 4 前後と高いが，第一クラスタの平均論文数と第二クラスタの平均論文数が大きく，複数のクラスタに分散しているが，主に 2 つのクラスタに論文が集中している．

また，論文数が少ない「並列論理型言語」「冗長ディスクアレイ」「リサーチマイニング」は第一クラスタの平均論文数が大きく，主に一つのクラスタに集中している．

しかし,「アクティブデータベース」「XML」は，第一クラスタの平均論文数が大きくなく，平均的に複数のクラスタに分散している．

第一および第二クラスタの平均論文数が全体の約 80%前後である研究テーマが多いことから，研究テーマの論文がほぼ 2 つのクラスタに集中する傾向にある．しかし，表 1 において，純度は 0.375 であるため，クラスタは少数の大きな研究テーマが混在しており，玉石混淆なクラスタを形成してはいない．研究テーマが一つのクラスタに統合されないようにするために，クラスタリングアルゴリズムの改善が必要である．

8. まとめと今後の課題

本論文では，論文のメタ情報を利用してクラスタリングを行い，研究者の研究履歴を自動的に生成する手法を提案した．提案手法では，論文の著者，発表年，キーワード，引用情報，関連プロジェクトを利用して，個々の類似度を線形結合することにより論文の類似度を定義する．そして，クラスタリング手法の一つである K-means 法を用いてクラスタリングを行うことで，研究者の研究テーマを自動的にクラスタリングし，研究履歴を生成する．

論文のメタ情報を CiNii および KAKEN を用いて収集し，特定の研究者の論文を収集し評価実験を行った．評価実験の結果，エントロピー約 0.5，純度が約 0.35 程度であった．類似度のパラメータを変化させることにより，クラスタリング結果のエントロピーや純度が変化することが確認できた．特に，著者，キーワード，引用情報の類似度を重要視することでエントロピーは減少するが，純度も減少する傾向にあることがわかった．そして，研究テーマごとにクラスタを分析した結果，同一研究テーマの論文は主に 2 クラスタ程度に集中し比較的まとまりやすい傾向にあるが，まとまった研究テーマが複数統合されクラスタを形成する傾向にある．

本論文では，著者，キーワード，引用を重視したパラメータの場合のクラスタリングの評価を行った．しかし，各パラメータごとに重要度は異なるはずである．今後，各類似度の重要度のパラメータをどのように最適化するかということについて検討する．また，評価実験では純度が比較的低くく，クラスタリングの性能向上も今後の課題である．さらに，論文のキーワードの出現頻度などを考慮して，クラスタに研究テーマラベルを付与する手法についても検討し，研究レポジトリの新たな機能としてシステムの開発を目指す．

謝　　辞

本研究の一部は，文部科学省科学研究費補助金特定領域研究 (#21013017)，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] M. Callon, J. P. Courtial, and F. Laville. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Sientometrics*, Vol. 22, pp. 155–205, 1991.

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

2003.
[3] 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, Vol. 42, No. 11, pp. 2640–2649, 2001.
[4] 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 回年次大会ワークショップ「言語処理と情報可視化の接点」, 2006.
[5] 国立情報学研究所. NII 論文情報ナビゲータ CiNii. http://ci.nii.ac.jp/.
[6] 国立情報学研究所. 科学研究費補助金データベース KAKEN. http://kaken.nii.ac.jp/.
[7] 吉田誠, 小林隆志, 横田治夫. リサーチマイニング手法におけるクラスタリング閾値設定指針の考察. 情報処理学会データベース・システム研究会 (2004-DBS-134(II)), pp. 553–560, 2004.
[8] 吉田誠, 小林隆志, 横田治夫. 公開されている論文 DB からのマクロ情報抽出に対するリサーチマイニング手法と他手法の比較. 情報処理学会論文誌: データベース, Vol. 45, No. SIG7(TOD22), pp. 24–32, 2004.
[9] 宮本定明. クラスター分析入門 ファジィクラスタリングの理論と応用. 森北出版株式会社, 1999.
[10] 東京工業大学. 東京工業大学リサーチレポジトリ T2R2. http://t2r2.star.titech.ac.jp/.
[11] 近藤友樹, 難波英嗣, 奥村学, 新森昭宏, 谷川英和, 鈴木泰山. 論文データベースからの研究動向情報の抽出. 言語処理学会第 13 回年次大会, 2007.