DEIM Forum 2013 C10-4

ラベル付きクラスタリングにおける 情報利得を用いたキーワード選定結果の適用手法

加藤　大智† 　橋本　泰一†† 　渡辺　陽介††† 　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
†† グリー株式会社
〒 106-6190 東京都港区六本木 6-10-1 六本木ヒルズ森タワー
††† 東京工業大学 学術国際情報センター
〒 152-8550 東京都目黒区大岡山 2-12-1
E-mail: †{kato,watanabe}@de.cs.titech.ac.jp, ††taiichi84@gmail.com, †††yokota@cs.titech.ac.jp

あらまし　動画や画像・論文など，キーワードやタグが付与されているデータが多く存在し，そうしたデータを分類する要求が高まっている．本稿では，キーワードが付与されたデータと付与されていないデータが混合したデータをクラスタリングすることを目的とする．キーワードはすべてのデータについているとは限らず，またキーワードの中には分類に適さないものもあるため，単純にキーワードのみを用いてすべてのデータを分類することは難しい．そこで本研究では，より精度の高いクラスタリングを行うため，情報利得によってデータに付けられたキーワードの中からラベルとして適するものとそれに関連するデータを選定し，それを教師データとして半教師付きクラスタリングを行う手法を提案する．実際の論文データに対して適用して比較を行う．
キーワード　クラスタリング, クラスタラベリング，SVM，情報利得，リサーチマイニング

Clustering Methods Combining Keyword Selection Based on Information Gain

Daichi KATO†, Taiichi HASHIMOTO††, Yousuke WATANABE†††, and Haruo YOKOTA†

† Department of Computer Science, Graduate School of Information Science and Engineering,
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo, 152-8552 Japan
†† GREE, Inc.
Roppongi Hills Mori Tower, 6-10-1 Roppongi, Minato-ku, Tokyo, Japan
††† Global Scientific Information and Computing Center, Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo, 152-8550 Japan
E-mail: †{kato,watanabe}@de.cs.titech.ac.jp, ††taiichi84@gmail.com, †††yokota@cs.titech.ac.jp

1. はじめに

近年，さまざまなところで情報量が爆発的に増加している．そこで，大規模データから傾向や新たな知見を得るための手法として，似たデータをグループ分けするクラスタリングと呼ばれる手法がよく用いられている．代表的なものとしては，K-means 法や階層的クラスタリングなどがある [1].

クラスタリングの特徴の一つとして，教師付きデータをあらかじめ与えない，教師無しの手法であることがあげられる．このため，事前にどのような分類があるか分からない状態からデータの傾向を見るようなときに用いられるが，データ集合の類似性のみを用いて判定を行うため，利用者の意図とは異なる分類となることも多い．

これを解消するため，クラスタリングの際に教師付きデータを部分的に与えてクラスタリングを行う手法もあり，これは半教師付クラスタリングと呼ばれている．しかし教師付きデータ

をあらかじめ与えられないことも多く，適用可能な局面が限られてくる．

そこで，厳密な教師データを人手で与える以外のアプローチとして，本研究ではデータに付けられたキーワードやタグを用いることを考える．キーワードがつけられているデータとしては，動画や画像・書籍・論文などがあげられる．しかしキーワードはすべてのデータについているとは限らず，またキーワードの中には分類に適さないものもあるため，単純にキーワードのみを用いてすべてのデータを分類することは難しい．

この問題を解決するため，キーワードの情報を選別し有用なものだけを残して疑似教師データを作り，これとクラスタリングを組み合わせて，精度の高いクラスタリングを行うことを提案する．ここで技術的に問題となるのが，疑似教師データを作るためのキーワード選定手法，キーワードがないデータをクラスタリングする手法，クラスタリングにおいてキーワード選定結果をどのように活用するかである．

そこで本研究では，次のような手法を提案する．まず，データに付けられたキーワードの中からラベルとして適するものとそれに関連するデータを選定する．選定には，情報利得から各キーワードのスコアを求め，最も高いものをラベルとして採用する．そして情報利得によって得られたラベルを制約として半教師付きクラスタリングを行う．半教師付きクラスタリングにラベル選定結果を適用することで，キーワードがつけられていないデータに関しても精度の高いクラスタリングを行えることが期待される．本研究では，サポートベクターマシン (SVM) で分離超平面を求め，その結果を初期平面としてマージン最大化クラスタリングを行う手法と，情報利得によって得られたラベルを教師データとして，半教師付の SVM を実行する手法を提案する．

本論文の概要は以下のとおりである．まず第 2 章で本研究で使用するクラスタリング手法について述べる．第 3 章で今回我々が提案するクラスタリング手法を述べる．第 4 章で評価実験の手法と実験結果の提示・考察を行う．第 5 章で関連研究を紹介し，第 6 章でまとめと今後の課題を述べる．

2. 前提とするクラスタリング手法

まず，本研究のクラスタリングで用いるサポートベクターマシン (Support Vector Machine, 以下 SVM)，マージン最大化クラスタリング (Maximum Margin Clustering, 以下 MMC) [2]，半教師付きサポートベクターマシン (Semi-Supervised Support Vector Machine, 以下 S3VM) [3] の概要を説明する．

2.1 SVM

SVM とは，教師データのクラスが判明している場合に，その 2 つのクラスを線形分離するような超平面を求める手法である．図 1 の左では，丸と三角で表した教師データを分離するような超平面を求めている．このような超平面はいくつも考えられるが，SVM では，最も近いサンプルとの距離（マージン）が最大となるような超平面を解とする．このとき，超平面にもっとも近いサンプルをサポートベクターと呼ぶ．しかし教師データを線形分離するような超平面を求めることができない場合もある．その場合，マージンの逆数+識別の誤りを最小化するような分離超平面を求める，ソフトマージンと呼ばれる手法が使われる．

SVM では，教師データから求めた分離超平面を用いて，クラスが未知のデータを分離する．ただし，その際の未知データ集合の分布に合わせて分離超平面を微調整することはできない．

2.2 MMC

SVM の考え方に基づいて，教師無しクラスタリングの手法として利用した，マージン最大化クラスタリング (Maximum Margin Clustering, MMC) [2] と呼ばれる手法が提案されている．MMC では，データが何らかのクラス分類を持つと考え，マージンが最大となる分離超平面を求める．データがとり得るすべてのクラス分類について分離超平面を求め，その中でマージンが最大のものを，データを分離する超平面とする（図 1 の右）．

データを $x_1, \ldots, x_n$，x_i のラベルを $y_i \in \{-1, +1\}$[1] とすると，MMC は，次のようにマージンの逆数 $\mathbf{W}$ を最小化する問題として定式化することができる．

$$\min_{y_1,\ldots,y_n} \min_{\mathbf{W},b,\xi_i} \quad \frac{1}{2}\mathbf{W}^T\mathbf{W} + \frac{C}{n}\sum_{i=1}^{n}\xi_i \qquad (1)$$
$$s.t. \quad y_i(\mathbf{W}^T\mathbf{x}_i + b) \geqq 1 - \xi_i,$$
$$\xi_i \geqq 0, \forall i = 1, \ldots, n$$

ここで，C は識別誤りの許容度を調整するパラメータである．C が大きいほど，識別の誤りを許容しなくなる．

本研究では，MMC を高速に解く手法である CPMMC(Cutting Plane MMC) [4] を用いる．CPMMC では，ランダムに初期平面を与え，その周辺のデータのみを用いて超平面を求めることで，計算量の削減を図っている．このため，結果が初期平面に左右されるという問題点が知られている．

2.3 S3VM

データによっては，ラベルがついた教師データとついていないデータの両方を得られるものがある．そのようなデータをクラスタリングする手法として，半教師付きサポートベクターマシン (Semi-Supervised Support Vector Machine, 以下 S3VM) と呼ばれるアルゴリズムが存在する [3]．S3VM では，ラベルがついていないデータがとり得るすべてのクラス分類について分離超平面を求め，マージンが最大のものを，データを分離する超平面とする（図 1 の中央）．

ラベル付きデータを $x_1, \ldots, x_l$，ラベルなしデータを $x_{l+1}, \ldots, x_n$，ラベル付きデータのラベルを $y_i \in \{-1, +1\}$[1] とすると，S3VM は，次のようにマージンの逆数 $\mathbf{W}$ を最小化する問題として定式化することができる．

$$\min_{y_{l+1},\ldots,y_n} \min_{\mathbf{W},b,\xi_i} \quad \frac{1}{2}\mathbf{W}^T\mathbf{W} + \frac{C_l}{n}\sum_{i=1}^{l}\xi_i + \frac{C_u}{n}\sum_{j=l+1}^{n}\xi_j \qquad (2)$$
$$s.t. \quad y_i(\mathbf{W}^T\mathbf{x}_i + b) \geqq 1 - \xi_i, \forall i = 1, \ldots, l$$
$$y_i(\mathbf{W}^T\mathbf{x}_j + b) \geqq 1 - \xi_j, \forall j = l+1, \ldots, n$$
$$\xi_i \geqq 0, \forall i = 1, \ldots, l$$

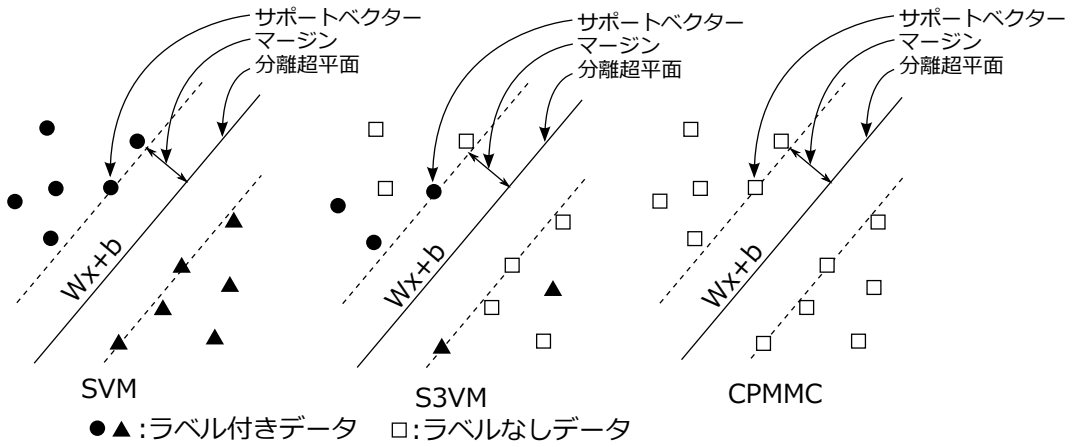

図 1　SVM,S3VM,MMC における分離超平面の求め方

$$\xi_j \geq 0, \forall j = l+1, \ldots, n$$

ここで，C_l はラベル付きデータの，C_u はラベルなしデータの識別誤りの許容度を調整するパラメータである．C_l, C_u が大きいほど，識別の誤りを許容しなくなる．

S3VM では，教師データのみから求められた超平面をそのまま用いるのではなく，ラベルが未知のデータの分布に合わせて超平面を動かすことができる．この部分が SVM との違いである．

本研究では，これを高速に解く手法である CutS3VM [5] を用いてクラスタリングを行う．CutS3VM では，CPMMC 同様，分離超平面の周辺のデータのみを用いて新たな超平面を求めることで，計算量の削減を図っている．

3. 提 案 手 法

本研究で扱う問題は，複数のキーワード情報が付与されたデータと何も付与されていないデータが混在するようなデータ集合からクラスタを生成することである．本研究では，次のような手法でデータのクラスタリングを行うことを提案する（図 3）．

まず，データ集合をキーワードがあるものとないものに分割する．次に，データに付けられたキーワードの中からラベルとして適するものとそれに関連するデータを選定する．選定には，3.1 節で述べる方法で情報利得からキーワードのスコアを求め，最も高いものを採用する．そして情報利得によって得られたラベルを制約として半教師付きクラスタリングを行う．本研究では，情報利得によって得られたラベルを教師データとして，半教師付の SVM を実行する手法（提案手法 1）と，サポートベクターマシンで分離超平面を求め，その結果を初期平面としてマージン最大化クラスタリングを行う手法（提案手法 2）を提案する．

3.1 情報利得を用いたラベル選定

まず，情報利得 (Information Gain,IG) を用いてキーワード選定を行う手順を説明する．

データ集合とキーワードの集合の関係において最良な状態とは，複数のデータ集合に共通するキーワードがない状態である

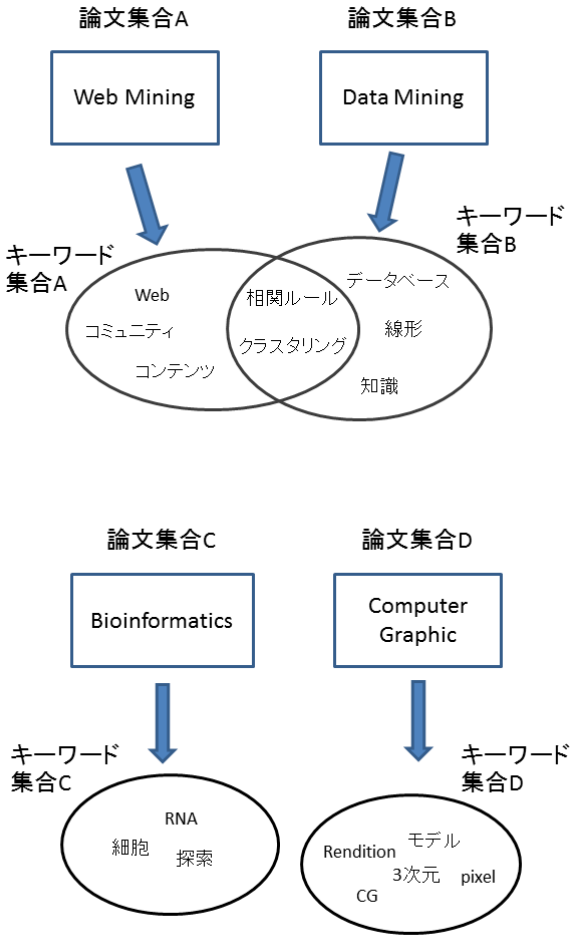

図 2　データ集合とキーワードの関係性（論文の分類への適用）

と仮定する．たとえば図 2 は論文データの例であるが，データ集合 A と B のように共通するキーワード「相関ルール」や「クラスタリング」があるようなデータ集合を形成することは望ましくなく，データ集合 C と D のように，共通するキーワードがないデータ集合が形成されることが望ましい．

そこで，複数のキーワードを含むデータの集合 X が与えられたとき，なるべくこの条件を満たすような 2 つのデータ集合

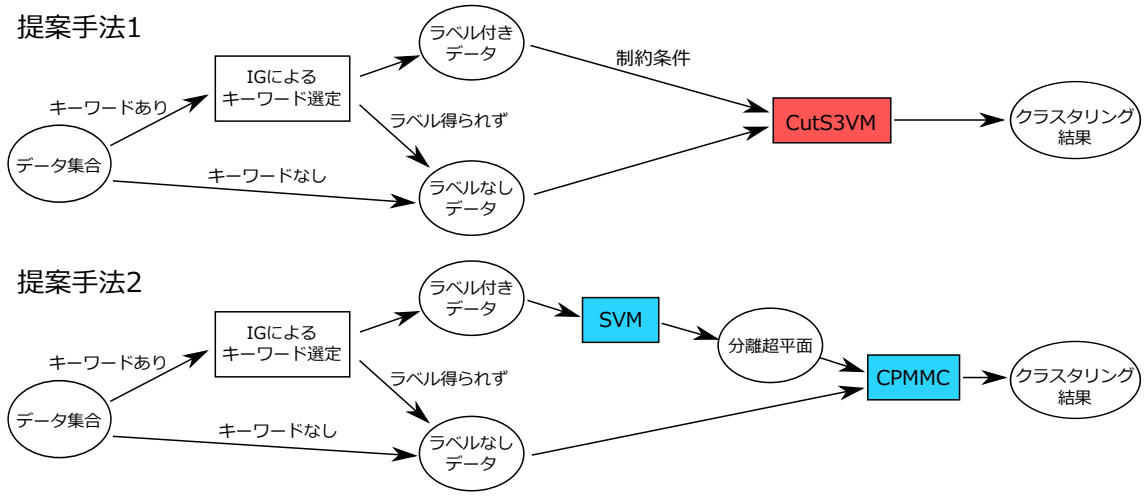

図 3　提案手法のフロー図

$X_k, X_{\overline{k}}$ へ分割する．データ集合 X に含まれるデータに付与されたキーワード集合 K のうちあるキーワード k に注目し，キーワード k を持つデータの集合 X_k とキーワード k を持たないデータの集合 $X_{\overline{k}}$ へ分割する．この分割したデータ集合 $X_k, X_{\overline{k}}$ に対して，キーワード $k' \in K - \{k\}$ の情報利得 $IG_k(k')$ を計算する．

キーワード k と k' が共起しやすいもしくは共起しにくい場合，$IG_k(k')$ は高い値をとる．したがって，$X_k, X_{\overline{k}}$ において，他のキーワード k' のどれにおいても，$IG_k(k')$ が高い値を示すのであれば，データ集合とキーワード k の集合の関係が良い状態であるといえる．

つまり，データ集合 X が与えられたとき，先の条件を満たす最も良いキーワード k は，

$$\operatorname*{argmax}_k \min_{k'} IG_k(k')$$

により求めることができる．こうしてできた集合 X_k に含まれるデータを，ラベル k を持つラベル付きデータと呼ぶ．また k を含まない集合 $X_{\overline{k}}$ を別のキーワードでさらに再帰的に分割することにより，複数のラベルとラベル付きデータを選定することができる．

3.2　クラスタリング

続いてクラスタリングを行う．3.1 節で得られたデータのラベルが変わらないような制約を加えて多値クラスタリング問題を解くと，データをキーワード別に分類することができる．

SVM では，サンプルのクラスがすべて判明している状態から分類器を作成する．一方 MMC では，データのクラスが全く判明していない状態からクラスタリングを行う．本研究では，キーワード選定によってデータのクラスが一部判明しているため，どちらとも異なる問題を解かなければならない．具体的には，キーワード選定で選ばれたラベルのうち 1 つを選び，そのラベルがついたデータを正例，それ以外のラベルがついたデータを負例とした半教師付きの 2 値クラスタリングを実行する．それぞれのラベルに対しクラスタリングを実行することで，多値クラスタリングが実現できる．

本研究では，この半教師付き 2 値クラスタリングについて 2 つの手法を提案する．1 つはキーワード選定の結果から得られる制約を加えた CutS3VM を解く手法（提案手法 1），もう 1 つはキーワード選定の結果を用いて分離超平面を作り，これを初期平面として CPMMC を解く手法（提案手法 2）である．それぞれの手法のフロー図を，図 3 で示す．

3.2.1　提案手法 1:キーワード選定結果を用いた CutS3VM

提案手法 1 として，キーワード選定の結果から得られる制約を加えて CutS3VM を解く手法について説明する．

まず，情報利得を用いてキーワードとそれに対応するデータを選定する．次に，正例となるデータのラベルを $+1$，負例となるデータのラベルを -1 として CutS3VM を解く．

なお CutS3VM では，ラベルの偏りを避けるため，$\sum_{j=l+1}^{n} \mathbf{x}_i = 0$ となるように $\mathbf{x}_i$ を変換し，b を $b = 2r - 1$（r はデータ全体に占める正例の割合）という定数と定義している．本研究では，正例とされたラベル付きデータの割合と実際の正例の割合が同じであると仮定し，r = 正例とされたラベル付きデータの割合 とした．

3.2.2　提案手法 2:キーワード選定結果を用いた CPMMC

次に，キーワード選定で得られたラベル付きデータを用いて分離超平面を作り，これを初期平面として CPMMC を解く手法について説明する．CPMMC の問題点として，クラスタリング結果が初期平面に左右されるというものがある．オリジナルの手法では初期平面をランダムに与えていたが，この方法では安定して精度を得られない．そこで，キーワード選定の結果を用いて得られたラベル付きデータだけを最もよく分離できる（マージンが最大となる）分離超平面をまず求め，それをラベルなしデータを含めた CPMMC における初期平面とみなす．

この考えに基づいた提案手法 2 を以下で説明する．

まず，提案手法 1 と同様に，情報利得を用いてキーワードとそれに対応するデータを選定する．得られたラベル付きデータのみを用いて SVM を実行すると，分離超平面を得ることができる．この分離超平面を初期平面として CPMMC を実行し，クラスタリング結果を得る．

CPMMC で得られた分離超平面は，ラベル付きデータとラ

表 1　人手により分類した正解セット

No.	研究テーマ名	論文数	発表期間（年）
1	負荷分散	41	1993 - 2008
2	自律ディスク	28	1999 - 2007
3	e-ラーニング	27	2001 - 2010
4	Fat-Btree	26	1997 - 2007
5	web	19	2002 - 2008
6	アクティブデータベース	11	1994 - 2008
7	並列論理型言語	6	1994 - 1998
8	冗長ディスクアレイ	7	1993 - 1997
9	XML	5	2003 - 2006
10	リサーチマイニング	5	2004 - 2005

ベルなしデータの両方を含めて最もよく分離できる（マージンを最大化する）ものとなっており，一般的には初期平面からは変化する．

4. 評価実験

4.1 実験の目的

提案手法の性能を確かめるため，評価実験を行う．

評価実験では，キーワード付きのデータとして論文を用い，クラスタリングでは論文をメタ情報でベクトル化したものを用いる．ここで用いたメタ情報は，論文の著者・発表年・論文に付けられたキーワード・引用している論文・関連プロジェクト(注1)である．

4.2 評価実験環境

評価実験では，CiNii Articles [6] から「横田治夫」を著者として含む論文 201 本のメタ情報を収集した．CiNii では論文の情報を RDF で取得することが可能である．ここから，著者・発表年・引用論文・キーワードを収集した．また，科学研究費補助金のデータベースである KAKEN [7] からプロジェクトおよび関連する論文の情報を収集した．

4.3 正解セットと評価尺度

「横田治夫」を著者として含む論文 201 本を人手で分類したところ，そのうち 175 本の論文を 10 個のテーマ（カテゴリ）に分けることができた（表 1）．そこで，この人手による分類を正解セットとし，これにどれだけ近いキーワードやクラスタを求められたかで評価を行う．

クラスタリングの評価尺度として，正解率を用いた．定義は次のとおりである．

$$Accuracy = \max\left(\frac{n_1^1 + n_2^2}{n}, \frac{n_1^2 + n_2^1}{n}\right) \quad (3)$$

ここで，n は論文数，n_r^i はクラスタ r に含まれる正解セット i の論文の数である．情報利得によって得られたラベルも評価に入れるべきであると考え，ラベル付き論文も評価の対象とする．

4.4 2 値クラスタリング

本手法の性能を確かめるため，2 値クラスタリングの実験を行う．この実験では，表 1 から 2 つの研究テーマを選び，それらに含まれる論文を混在させた論文集合に対してクラスタリングを行う．

また，提案手法はどちらもキーワード選定の結果に依存しており，キーワード選定の性能とクラスタリングの性能を分けて確かめるため，次のような 3 種類の実験を行う．

情報利得を利用　すべての論文から，情報利得を用いてラベルとそれに対応する論文を選定する．その中から，実験で用いる研究テーマに対応するラベルを 2 つ選び，そのラベルをキーワードとしてもつ論文をラベル付き論文として実験を行う．この実験はランダム性がないため，1 回のみ実験する．

正解セットから 20%をランダムに選択　情報利得を用いて生成した疑似教師データが人手によって作成した本物の教師データと比べてどの程度の品質であるのかを調べるため，情報利得を用いて疑似教師データを生成する代わりに，正解セットから教師データを生成して実験を行う．選んだ 2 つの正解セットをそれぞれランダムに 5 分割し，そのうち 1 つずつをラベル付き論文，それ以外をラベルなし論文として実験する．これをすべての分割に対して行う（5 回）．正解セットの分割を変えて 5 回行い，その平均の正解率で評価する．このような実験では，5 分割したうちの 4 つをラベル付きデータ，残り 1 つをラベルなしデータとして実験するのが一般的だが，本研究では情報利得によって得られるラベル付き論文が全体に対して非常に少ないため，このような手法で実験を行う．

クラスタリングでは提案手法 1 と同様の方法で教師データを与えた CutS3VM，提案手法 2 と同様の方法で初期分離超平面を与えた CPMMC を用いる．

ラベル不使用　比較対象として，論文のラベルを生成せず，CPMMC のみを用いてクラスタリングを行う．

今回の実験で選んだ研究テーマ，論文数，その 20%の論文数，情報利得で得られた対応するキーワード，情報利得で得られたラベルが付されている論文数は表 2 のとおりである．

4.5 実験結果

実験結果は表 3 の通りである．情報利得を用いて疑似教師データを生成し 2 値クラスタリングを行った場合，提案手法 1 では 10 通りの組み合わせのうち 8 通り，提案手法 2 では 10 通りの組み合わせのうち 9 通りでラベル不使用の CPMMC より正解率が向上した．また，提案手法 1 と提案手法 2 を比較すると，提案手法 1 のほうが正解率で勝っているのが 5 通り，提案手法 2 のほうが勝っているのが 3 通り，同率が 2 通りであった．

提案手法 1 が提案手法 2 より勝っているのは，提案手法 2 では CPMMC の初期分離超平面を与える際にしかラベルの情報を用いていないため，CPMMC を起動した後は，情報利得で得られたラベルと異なるクラスタリング結果を生成してしまうことがあるためである．

また，正解セットから 20%をランダムに選択し 2 値クラスタリングを行った場合と比較すると，提案手法 1 では情報利得を用いた場合に 10 通りのうち 4 通りで正解率が上がり，提案手法 2 では 8 通りで正解率が上がった．情報利得を用いて作った疑似教師データには誤った分類結果も含まれているため，一般的には正解セットから与えた教師データに比べて正解率が落ち

(注1)：科学研究費補助金（科研費）の，同じ研究成果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する．

表 2　2 値クラスタリングで利用する論文のデータ

No.	研究テーマ名	論文数	教師データ（論文数の 20%）	IG によるキーワード名	IG によるラベル付き論文数（疑似教師データ数）
1	負荷分散	41	8	負荷分散	8
2	自律ディスク	28	5	autonomous disks	10
3	e-ラーニング	27	5	e-learning	11
4	Fat-Btree	26	5	Fat-Btree	12
5	アクティブデータベース	11	2	アクティブデータベース	5

表 3　実 験 結 果

		正解セットの 20%を教師データに選択		情報利得を利用		ラベル不使用
正解セット 1	正解セット 2	CutS3VM	CPMMC	提案手法 1	提案手法 2	CPMMC
負荷分散	自律ディスク	0.905	0.773	0.855	0.913	0.601
負荷分散	e-ラーニング	0.886	0.705	0.574	0.838	0.659
負荷分散	Fat-Btree	0.829	0.559	0.507	0.687	0.510
負荷分散	アクティブデータベース	0.800	0.584	0.615	0.596	0.584
自律ディスク	e-ラーニング	0.952	0.919	0.982	0.982	0.738
自律ディスク	Fat-Btree	0.969	0.911	0.907	0.741	0.589
自律ディスク	アクティブデータベース	0.940	0.816	0.974	0.846	0.902
e-ラーニング	Fat-Btree	0.822	0.747	1.000	1.000	0.696
e-ラーニング	アクティブデータベース	0.954	0.941	0.842	0.711	0.638
Fat-Btree	アクティブデータベース	0.891	0.702	0.946	0.757	0.649

ると考えられる．それでも正解率が上昇したケースがあるのは，表 2 にあるように情報利得で得られたラベル付き論文数が正解セットの論文数の 20%を上回っており，(疑似) 教師データの数が増えたことによるものと考えられる．

教師データを与えた場合に比べれば正解率が低いこともあるが，ラベル不使用の CPMMC よりも正解率は向上し，この手法の有用性が確認された．

5. 関 連 研 究

Nguyen らは，論文のクラスタリングを行うために MMC を用いている [8] [9]．まず，生成対象とする研究者が執筆した論文のメタ情報を収集する．続いてクラスタリングを行うため，論文に付されたメタ情報を，それぞれのメタ情報の出現頻度を値で表したベクトルに変換する．このベクトルデータを 2 分割するような分離超平面を CPMMC で求め，2 値クラスタリングを行う．得られた 2 つのクラスタのうち，論文数が多い方を再度クラスタリングし，多値クラスタリングを行う．ラベリングは，各クラスタで出現頻度が最も高いキーワードを採用している．この手法の特長として，パラメータによる類似度の重みの調整が不要であり，クラスタリングに K-means 法を用いた場合と比較して精度が高いということがあげられる．一方，CPMMC は初期平面をランダムに置いて計算を行うため，クラスタリング結果が初期平面に左右されることが問題点として挙げられる．

われわれの過去の研究 [10] では，本研究と同様に情報利得を用いてラベルの選定を行い，K-means 法を用いてクラスタリングを行った．まず，データに付されたメタ情報の類似度を求める．各メタ情報に対し類似度を定義し，それらを重み付きで線形結合したものを，データ間の類似度とする．この類似度を用いて K-means 法を実行するが，各データを中心が最も近いクラスタに再配分する際に，ラベル付きデータは再配分を行わない．このクラスタリング手法の問題点として，ラベル選定の結果得られたラベルが誤っていた場合でもクラスタリングで修正されないことや，類似度の重みづけのパラメータを手動で設定しなければならないという問題がある．

6. まとめ・今後の課題

本論文では，クラスタリングを行う前に，あらかじめデータに付けられたキーワードの中からラベルとして適するものを情報利得を用い選定する手法を提案した．また，精度の高いクラスタリングを行うため，情報利得によって得られたラベルを制約として SVM で分離超平面を求め，その結果を初期値としてマージン最大化クラスタリング (MMC) を行う手法と，情報利得によって得られたラベルを教師データとして，半教師付の SVM を実行する手法を提案した．

提案手法の正当性を確かめるため，論文のメタ情報を収集して評価実験を行った．その結果，2 値クラスタリングにおいては，前者の提案手法の正解率は 61%から 100%，後者の提案手法の正解率は 59%から 100%で，どちらも既存研究より良い結果を得ることができた．しかし，情報利得により得られるラベルを用いたクラスタリングでは一部のデータで性能が落ちることも確認されたため，精度の向上が課題となる．

今後の課題として，現在の手法を多値クラスタリングに拡張する手法を提案する必要がある．また大規模なデータセットでの実験や，論文以外のデータへの適用手法の検討が必要になると考えられる．

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] 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, 1999.

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

[3] Kristin P. Bennett and Ayhan Demiriz, "Semi-supervised support vector machines", In *Proceedings of Neural Information Processing Systems*, 1998.

[4] Bin Zhao, Fei Wang, and Changshui Zhang, "Efficient maximum margin clustering via cutting plane algorithm", In *The 8th SIAM International Conference on Data Mining (SDM 08)*, pp. 751–762, Atlanta, Georgia, USA, 2008.

[5] Bin Zhao, Fei Wang, and Changshui Zhang, "CutS3VM: a fast semi-supervised SVM algorithm", In *KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 830–838, New York, NY, USA, August 2008. ACM.

[6] 国立情報学研究所, 「CiNii articles - 日本の論文をさがす」, http://ci.nii.ac.jp/.

[7] 国立情報学研究所, 「Kaken - 科学研究費補助金データベース」, http://kaken.nii.ac.jp/.

[8] Manh Cuong Nguyen, Daichi Kato, Taiichi Hashimoto, and Haruo Yokota, "Research history generation using maximum margin clustering of research papers based on metainformation", In *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*, iiWAS '11, pp. 20–27, New York, NY, USA, 2011. ACM.

[9] Manh Cuong Nguyen, 加藤大智, 橋本泰一, 横田治夫, 「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」, 第 4 回データ工学と情報マネジメントに関するフォーラム (deim Forum 2012), F6-3, 2012.

[10] 加藤大智, Manh Cuong Nguyen, 橋本泰一, 横田治夫, 「論文のラベル付きクラスタリングのための情報利得を用いたキーワード選定」, 第 4 回データ工学と情報マネジメントに関するフォーラム (deim Forum 2012), E10-1, 2012.