情報処理学会論文誌　データベース　Vol.6　No.3　1–12 (June 2013)

特徴抽出を目的とした文書クラスタからの一貫性阻害要素除去

佐藤 進也[1,a)]　高橋 公海[1,b)]　松尾 真人[1,c)]

受付日 2012年12月20日, 採録日 2013年4月5日

概要：クラスタリングにより文書集合を意味的に分類し，それぞれの特徴を表す情報（特徴語）を抽出するという目的のため，クラスタリング結果を改善する方法を考案した．本手法では，各クラスタから，いわゆるノイズと呼ばれるような，クラスタを構成する文書集合の意味的一貫性を阻害する要素を除去する．除去する文書を決定するためには，別のアルゴリズムで得たクラスタリング結果を利用する．これにより，従来埋もれていた特徴語の発見が可能になる．本手法の有効性を確認するため，Q&A サイトのページを集めて文書集合を作成し，そのクラスタリング結果から特徴的な場所を抽出する（たとえば，バーベキューに関する質問のクラスタから「公園」といった場所を抽出する）実験を行った．10 個の文書集合を作成し，それぞれに対して提案手法を適用した結果，延べ百数十の場所が新たに得られた．また，本手法は質問に対して意外な関連性のある場所を抽出する傾向があることが分かった．

キーワード：文書クラタスタリング，特徴語抽出，凝集型階層的クラスタリング，k 平均法，コミュニティ抽出，LDA

Document Cluster Purification for Feature Extraction

Shin-ya Sato[1,a)]　Masami Takahashi[1,b)]　Masato Matsuo[1,c)]

Received: December 20, 2012, Accepted: April 5, 2013

Abstract: For effectively extracting features from document clusters, we developed a technique for improving the quality of the clustering results, which purifies original clusters (i.e., eliminates unwanted elements in each cluster) by using the outcome from another clustering algorithm. For verifying the effectiveness of the proposed approach, we conducted an experiment to discover associations between document clusters and their characteristic places using pages in a social Q&A site (e.g., associate “park” with a document cluster of questions about barbecues). We obtained a hundred and several tens of places in total by applying the proposed approach to 10 document sets. Futhermore, we observed a tendency that the approach discovered unexpected associations.

Keywords: document clustering, feature term extraction, agglomerative clustering, K-means, community detection, LDA

1. はじめに

クラスタリングは，データをその特徴の類似度からグループ分けするデータ処理の一手法である [1]．その文書データへの適用についても長年にわたって研究され，実用システムにも応用されてきた．その一例として，Web 検索結果を分類して提示するシステムをあげることができる [2]．文書データのクラスタリング（文書クラスタリング）では，多くの場合，文書を内容の類似性に基づいて分類することを目的にしている．別な言い方をすると，同一クラスタ内は意味的に一貫性を有するような分類を目指している．クラスタの意味的一貫性（厳密には「クラスタを構成する文書の意味的一貫性」であるが，「クラスタの意味

[1] NTT 未来ねっと研究所
NTT Network Innovation Laboratories, NTT Corporation, Musashino, Tokyo 180–8585, Japan
a) shin-ya.sato@acm.org
b) t.masami@lab.ntt.co.jp
c) matsuo.masato@lab.ntt.co.jp

的一貫性」と略記する）が保証されれば，利用者は，情報の取捨選択を文書単位ではなくクラスタ単位で行うことが可能になり，情報検索・探索の効率化がもたらされる．各クラスタの内容を端的に表現することを目的として特徴語抽出などが行われるが，クラスタの意味的一貫性が高ければ，この処理においてもより適切な結果が期待できる．すなわち，クラスタリングを用いた情報探索や情報抽出の品質は，クラスタリング結果の品質に依存する．

今までに様々なクラスタリングアルゴリズムが開発されてきているが，一般に，高い意味的一貫性を持つようなクラスタリングを実現するのは容易ではない．まず，クラスタリングのためには文書間の類似性を測る尺度を導入する必要があるが，意味的な関連性を十分に反映した尺度を定義することは難しい．さらに，クラスタリング結果にデータの特徴だけでなくクラスタリングアルゴリズムの性質までもが反映されてしまうという問題や，データ量増加にともなう計算コストの問題などが指摘されている [1]．

本論文では，クラスタリングにより文書データを意味的に分類しそれぞれの特徴を表す情報（特徴語）を抽出するという処理の品質向上を目的として，クラスタリング結果を改善する方法について議論する．確かに，クラスタリングアルゴリズムを精緻なものにすることにより，より適切なクラスタリング結果が期待できる．しかし，特徴抽出という目的のためにはそれが唯一の方法ではなく，クラスタから一貫性を阻害する要素を除去するという方法がある（4.2 節）．本論文では，あるクラスタリングアルゴリズムで得られたクラスタの一貫性阻害要素を，もう 1 つのクラスタリングアルゴリズムの結果を用いて除去する手法を示す．この手法によれば，既存のクラスタリングアルゴリズムを組み合わせることで，単一のクラスタリング結果からは得られなかった特徴語を抽出することができる．

以下，本論文の構成は以下のとおりである．まず，2 章で特徴語抽出とクラスタリング結果の改善という 2 つの概念の関係を明らかにした後，3 章で関連研究を紹介し本研究の位置づけを明らかにする．4 章で一貫性を阻害する要素の除去というアイデアと，その具体的方法について説明する．提案手法の評価実験の内容と結果を 5 章で示し，6 章で考察を加える．

2. クラスタからの特徴語抽出

特徴語の抽出を目的としたクラスタリング結果の改善手法について議論するのにあたり，本章では，まず，特徴語抽出とクラスタリング結果の改善という 2 つの事柄の関係を明らかにする．

2.1 想定しているタスク

まず，提案手法の効果が期待できる問題（タスク）の例を示し，その中で，二者の関係を説明する．

一般には，対象（たとえば文書）の集合を部分集合に分割する問題（以下，分割問題と呼ぶ）は，対象を帰属させるべきクラスが事前に決まっているか否かにより，クラス分類とクラスタリングに大別されるが，分類クラスとして想定している分類基準はあるものの，それを事前に把握し尽くすことが難しい場合もある．その場合，クラスタリング手法を用いることでボトムアップにクラスを把握していくという，いわば，クラスをクラスタで近似するアプローチが考えられる．

このようなアプローチによる問題解決の具体例としては，文書中の人物の識別（人名の曖昧性解消）をあげることができる．人名の曖昧性解消の基本的なタスクとして，ある人名を含む文書を，各クラスタがそれぞれ異なる（同姓同名の）人物に対応するようにクラスタリングするというものがある [3]．この場合，個々のクラスタがそれぞれ異なる人物に対応していると考えられる．そして，クラスタから抽出された特徴語は，個々の人物を識別するための情報（人物の属性，特徴）と考えることができる．クラスタが同一人物に関する文書で占められる割合が高くなれば，その文書集合から得られる情報は当該人物に関するものである確率も高くなり，その人物に関して適切な情報をより多く，精度良く得られると考えられる．

このような，クラスタでクラスを近似するアプローチにおいて，本論文で提案するクラスタリング結果の改善手法は，近似の精度を上げる工夫として位置づけることができる．そして，近似精度を向上させた結果として，本来の目的であるクラスの特徴をクラスタからより多く，正確に抽出することを狙う．

2.2 クラスタリング結果改善により期待される効果

本節では，クラスタリング結果の改善が特徴語の抽出にどのように影響するのかを具体例を通して説明する．ここでは，語の固有性に基づいて特徴語としての採否を決定することを考える．具体的には，文書集合全体とクラスタ内とで語の出現確率を計算し，前者より後者が高い場合，その語は当該クラスタにおける固有性が高いと見なし，特徴語として採用する．

図 1 は，文書集合 D をクラスタリングした結果，クラスタ C_i が得られたという状況をベン図と同様な方法で表

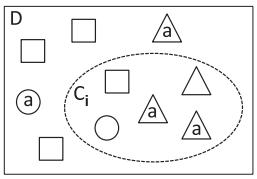

図 1　文書，クラスタ，文書の特徴の関係を表現した模式図

Fig. 1 Schematic diagram representing relationships among documents, a cluster and a characteristic feature of documents.

現したものである．この例を用いて，クラスタ C_i としてまとめられた内容（トピック）を表す特徴語を抽出する問題を考える．◯，□，△ はそれぞれ文書であり，形はそれらの内容（トピック）を表している．つまり，◯，□，△ がそれぞれがまとめられて分類されている状態が理想的なクラスタリング結果である．また，文書を表す記号の中にある「a」は，それらの文書に語 a が出現することを示している．△ のトピックに属する文書の 75%に語 a が出現しており，語 a はこのトピックと関連性があると考えられる．一方，C_i は 60%が △ で構成されており，△ の 75%が C_i に含まれている．よって，C_i と △ のトピックとの間に対応関係があると見なすことができる．

図では文書とトピックとの対応関係が明らかにされているが，これは人手で各文書の内容を調べた結果はじめて分かるものである．以下に述べるデータ処理に際してはこの情報は得られていないと仮定する．トピックと文書の対応関係は未知なので，代わりにトピックに対応するクラスタを使ってトピックの特徴を把握すること，具体的には，トピック △ と語 a の関連性を，C_i と語 a との関連性としてとらえることを考える．文書に a が出現する確率を調べると，D 全体では $4/10 = 0.4$ である．トピック △ に対応するクラスタ C_i 内ではこの確率が高くなり，それにより a との関連性が示されることが期待される．しかし，実際には，C_i における a の出現確率は D 全体と変わらない 2/5 である．これは，C_i に他のトピックの文書が混在しているためである．もし，これらを除去できれば，a の出現確率は 2/3 となり，0.4 よりも大きい値が得られる．△ と語 a の関連性を導き出すという目的のためには，クラスタ内の一貫性を阻害する要素（上記例では，C_i における ◯ や □）を除去することも有効であることをこの例は示している．

2.3　サブトピックに焦点をあてた特徴抽出

一般に，あるトピック全般に関する文書集合から特徴語を抽出しようとすれば，出現頻度の高い，そのトピックを概説するような特徴語が抽出される．一方，そのサブトピックに焦点をあてれば，サブトピック固有の視点からもとのトピックをより詳細に説明する特徴語の発見が可能になる．ここで「焦点をあてる」とは，あるサブトピックを選び，そのサブトピックに関連する文書からなる集合を構成し，そこから特徴語を抽出することを指す．このとき，サブトピックに固有な語の出現頻度がその文書集合中で相対的に高くなり，特徴語として抽出される可能性が高くなる．2.1 節で述べた人物の特徴を抽出する例でいえば，サブトピックとしてはその人物の仕事や趣味などをあげることができる．サブトピックに焦点をあてて特徴抽出することにより詳細な人物の把握が期待できる．

一貫性阻害要素除去において，正確な処理，すなわち，阻害要素のみを除去することは困難であり，多くの場合，目標としているトピックに関連する文書も一部除去されてしまうことが予想される．そして，結果的に，除去されずに残った文書の集合がもともとのトピックのサブトピックを形成している状況が起こりうる．このとき，前述の議論から，新たな特徴語の発見が期待できる．実際，評価実験の結果を調べてみると，サブトピックに焦点をあてた特徴語抽出と同等な処理を提案手法でも（結果的に）行っている可能性が示唆される（詳しくは 6 章参照）．

3.　関連研究

本章では，分割問題を解く手法（分割手法）を整理し，提案手法の位置づけ，新規性を明らかにする．まず，分割手法を機械学習の観点から，教師あり学習と教師なし学習に分け，それぞれについて前章で示したタイプの問題を解決する手段としての適否，計算量の観点からの実用性などについて議論する．

3.1　教師あり学習

分類問題に限らず，一般に，教師あり学習では訓練データ作成コストの高さが問題として指摘されている．この問題に対し，co-training を含む半教師あり学習（分割問題を解くものとしては，たとえば文献 [4], [5], [6] など）では比較的少数の訓練データから学習を進めていくことができるため，上記問題を部分的に解決している．しかし，2 章でも述べた，文書中に記されている人物を識別するタスクのような学習対象が多数で全体数も未知であるような場合には，訓練データをあらかじめ用意することはやはり難しい．

3.2　教師なし学習

2 章で述べたような問題に対して教師なし学習の分割手法であるクラスタリングを用い，クラス分類の近似として用いる場合，それぞれのクラスタへのノイズの混入を防ぎ，いかに目標とするクラスに近づけるかが問題となる．

3.2.1　クラスタリングアルゴリズムの高度化

1 章でも述べたように，単一のアルゴリズムで高い意味的一貫性を持つようなクラスタリングを実現するのは容易ではない．本項では，クラスタリングアルゴリズム高度化の難しさを示す目的で，n 次元空間内の点の集合をクラスタリング対象とするという前提のもと，クラスタリングアルゴリズムの持つ問題点とその解消を目指して進められてきた研究の例を簡単に紹介する．

k 平均方法 [7] は，現在でも広く用いられているクラスタリング手法であるが，クラスタどうしが線形分離できない状況（n 次元空間で入り組み合っているような状況）では適切な分類ができないことが知られている [16]．この問題に対して，空間中の分布状況をもとに複雑な構造を持つデータでもクラスタリング可能な DBSCAN [8] が考案された．また，カーネル関数により非線形構造を変換しクラス

タリングを行う方法 [9] も開発されている．しかし，この方法にも，複雑な構造を適切に扱うため人手によるパラメータ調整を要するという問題がある．

さらに，前章でも述べたように，クラスタリングアルゴリズム自体が高度化されても，それで文書分類という具体的な問題のすべてが解決されるわけではない．実際に文書分類の精度を向上させるためには，文書間の（意味的）類似性を測る方法（たとえば，各文書を n 次元空間の点に対応させる方法）の高度化も必要である．

3.2.2　クラスタアンサンブル

クラスタアンサンブルは異なる複数のクラスタリング結果（弱クラスタリング）をまとめて良いクラスタリング結果（強クラスタリング）を得る手法の総称である [10]．強クラスタリングを構成する方法としては，弱クラスタリングにおける共起性を利用するもの，すなわち，同一クラスタに属する割合から推定されるデータ間の類似性を利用するものなどがある．

複数のクラスタリング結果ができる限り整合するような最適なクラスタリングを見つけ出す処理は組合せ最適化問題としてとらえることができ，その計算量は NP 完全であることが知られている [11]．そこで，実用性の観点から，厳密解を求める代わりにメタヒューリスティックなどを用いて近似解を見つけ出す方法などが開発されてきた（文献 [12], [13] など）．たとえば，文献 [13] ではメタヒューリスティックとして遺伝的アルゴリズムを用いて最適解を探している．N をクラスタリングの対象となっているすべての要素の数としたとき，その計算量は $O(N^2)$ である [15]．クラスタアンサンブルは，ノイズ混入が少ないクラスタの集合を得るための一手段となりうるが，一般に，すべてのクラスタリング（すなわち，データ集合を部分集合に分割するすべてのやり方）を探索空間とする探索のコスト（計算量）は，メタヒューリスティックなどを用いてもなお高い．

クラスタアンサンブルと比較したとき，本論文で提案する手法は，探索方法を工夫したものではなく，探索空間を別のものに置き換えたものと考えることができる．提案手法は，与えられた各クラスタから関連性の低いデータを除去することでノイズの少ないデータ集合の作成を目標とする．つまり，複数あるデータ除去のやり方の中から適切なものを選び出す探索であると考えられる．本論文で提案する手法では，この探索を，2 つのクラスタリング結果の重なり具合いを調べる（4.2.2 項の式 (2)）という，クラスタアンサンブルの探索に比べると簡単な処理（計算量としては $O(N)$）で実現している．

4.　一貫性阻害要因除去

4.1　クラスタリング品質を評価する指標

一貫性阻害要素除去の方法を説明する前に，クラスタリング結果の品質を評価するための指標について述べる．

まず，説明の出発点として，情報検索の場合，すなわち，文書集合から検索要求を満たす文書を選び出す場合を考える．検索結果の文書集合を C，正解の文書集合を A とすると，precision，recall はそれぞれ

$$\begin{aligned}\mathrm{Precision}(C,A) &= \frac{|C\cap A|}{|C|},\\ \mathrm{Recall}(C,A) &= \frac{|C\cap A|}{|A|}\\ &= \mathrm{Precision}(A,C)\end{aligned}$$

と定義される．また，F 値はこれらの調和平均として定義される．

これに対し，purity，inverse purity [17] はクラスタリング結果の評価においてそれぞれ precision，recall に対応する指標の 1 つと考えられ，クラスタリングの結果得られた各クラスタを $\{C_i\}$，正解の分類を $\{A_i\}$ としたときに，precision/recall を使って次のように定義される．

$$\mathrm{Purity} = \sum_i \frac{|C_i|}{N}\max_j \mathrm{Precision}(C_i,A_j)$$

$$\text{Inverse purity} = \sum_i \frac{|A_i|}{N}\max_j \mathrm{Recall}(C_j,A_i)$$

ここで $N=\sum_i|C_i|$ である．purity はクラスタ内の一貫性，すなわち，クラスタが単一の A_i で占められている度合いを評価するものである．一方，inverse purity は，単一の A_i が 1 つのクラスタに集中している度合い（集中性）を評価するものである．情報検索の F 値同様，クラスタリング結果を総合的に評価する指標として，purity と inverse purity の調和平均である F 値が広く用いられている [17]．本論文でも，クラスタリング結果の品質を測る指標として，purity，inverse purity と（こららの調和平均である）F 値を用いる．

なお，分類対象の文書ごとに Precision，Recall 相当の数量を計算し，その平均によりクラスタリング結果を評価する precision BCubed，recall BCubed という指標も提案されている [17]．文書 d に対し，$\{C_i\}$ のうち d が属しているクラスタを $C(d)$，$\{A_i\}$ のうち d が属すべき正解文書集合を $A(d)$ と書くと，precision BCubed，recall BCubed はそれぞれ以下のように定義される．

$$\text{Precision BCubed} = \sum_d \frac{1}{N}\mathrm{Precision}(C(d),A(d))$$

$$\text{Recall BCubed} = \sum_d \frac{1}{N}\mathrm{Recall}(C(d),A(d))$$

ちなみに，5 章の実験で得られたクラスタリング結果を使って purity と precision BCubed の関係を調べてみたところ，強い正の相関が認められた（Pearson 相関係数が 0.995 であった）ことを付け加えておく．

4.2　一貫性阻害要素除去の方法

4.2.1　基本方針

いま，ある観点から次に示すような 4 つのグループ A_1〜A_4 に分類されるべき 18 個の文書 $d_1, \ldots, d_{18}$ からなる文書集合があるとする（A_i への帰属関係が分かりやすいように d_j に色を付けてある）．2 つの文書が同一のグループに属する場合，これらは同種であるということにする．

$$A_1 = \{d_1, d_2, d_3, d_4, d_5\}$$
$$A_2 = \{d_6, d_7, d_8, d_9, d_{10}, d_{11}\}$$
$$A_3 = \{d_{12}, d_{13}, d_{14}, d_{15}, d_{16}, d_{17}\}$$
$$A_4 = \{d_{18}\}$$

この文書集合に対して，前節で述べたような知識を得る目的であるクラスタリング手法を適用した結果，以下のようなクラスタ $\{C_i\}$ を得たとする．

$$C_1 = \{d_1, d_2, d_3, d_9, d_{10}, d_{15}\},$$
$$C_2 = \{d_4, d_6, d_7, d_8, d_{16}, d_{17}\},$$
$$C_3 = \{d_5, d_{11}, d_{12}, d_{13}, d_{14}, d_{18}\}$$

それぞれのクラスタにおいて，そのほぼ半分は同種の文書で占められており（たとえば，C_1 では，その半分が A_1 の文書である），残りの半分によりクラスタ内の一貫性が阻害されている．この阻害要素を除去することで一貫性を向上させるというのが本手法のポイントである．一貫性阻害要素を除去する方法としては，たとえば，$\hat{C}_i$ を

$$\hat{C}_i = C_i \cap A_k, \quad k = \operatorname*{argmax}_j |C_i \cap A_j| \tag{1}$$

と定義すれば，

$$\hat{C}_1 = \{d_1, d_2, d_3\}, \hat{C}_2 = \{d_6, d_7, d_8\}, \hat{C}_3 = \{d_{12}, d_{13}, d_{14}\}$$

という一貫性の高いクラスタを得ることができる．

4.2.2　補助的クラスタリング手法の利用

式 (1) には A_i が使われているが，実際にはこれをあらかじめ知ることはできない（既知であればクラスタリングの必要がない）．そこで，本手法では，$\{C_i\}$ を導き出したクラスタリング手法とは異なる，もう 1 つのクラスタリング手法（補助的クラスタリング手法）により別なクラスタ $\{Q_i\}$ を作り出し，これを $\{A_i\}$ の代わりに用いる．すなわち，次の式により一貫性の高いクラスタ $\{I_i\}$ の導出を狙う．

$$I_i = C_i \cap Q_k, \quad k = \operatorname*{argmax}_j |C_i \cap Q_j| \tag{2}$$

なお，上式は，クラスタリング結果 $\{C_i\}$ を $\{Q_i\}$ を利用して改善する方法を与えるものであるので，$\{C_i\}$ と $\{Q_i\}$ に関して対称でないことに注意を促しておきたい．

ここで，この方法の効果を具体的に示すため，以下に示す例を用いて一貫性阻害要因の除去を実際に行ってみる．いま，ある補助的クラスタリング手法により次のような $\{Q_i\}$ が得られたとする．

$$Q_1 = \{d_1, d_2, d_3, d_9\}, \quad Q_2 = \{d_4, d_5\},$$
$$Q_3 = \{d_{10}, d_{12}, d_{13}, d_{14}\}, \quad Q_4 = \{d_6, d_7, d_8, d_{16}, d_{18}\},$$
$$Q_5 = \{d_{11}, d_{15}\}, \quad Q_6 = \{d_{17}\}$$

このとき，式 (2) に基づいて I_i を計算すると次のようになり，各クラスタの一貫性が向上しているのが分かる．定量的に比較すると，$\{C_i\}$ では P，IP，F の値がそれぞれ 0.5，0.56，0.53 であったのに対し，$\{I_i\}$ ではそれらの値は等しく 0.82 となっている．

$$I_1 = C_1 \cap Q_1 = \{d_1, d_2, d_3, d_9\}$$
$$I_2 = C_2 \cap Q_4 = \{d_6, d_7, d_8, d_{16}\}$$
$$I_3 = C_3 \cap Q_3 = \{d_{12}, d_{13}, d_{14}\}$$

なお，どのような補助的クラスタリング手法でも必ず一貫性を向上させられるわけではない（$F(\mathcal{I}(\mathcal{C}, \mathcal{Q})) \leq F(\mathcal{C})$ となってしまう可能性もある）．よって，一般に，それぞれのクラスタリングの結果に対して，一貫性を改善するために適切な補助的クラスタリング手法を選ぶ必要がある．ただし，通常クラスタリングによく用いられる手法については，おおむね補助的クラスタリング手法として有用であることは，実データを用いた実験により確認している．また，本項ではもっぱら purity を基準とする一貫性の向上について議論したが，後に示すように，本手法により inverse purity で評価される集中性も改善され，その結果として F 値が向上することが分かっている．

5.　評価実験

5.1　実験内容

一貫性阻害要素除去の効果を調べるため，クラスタリングにより文書集合を分類しそれぞれから特徴的な情報を抽出するタスクを行った．具体的には，文書集合として Q&A サイトのページを選び，そのクラスタリング結果から特徴的な場所の抽出を試みた．Q&A サイトとは，ユーザ同士がお互いの質問に答え，疑問を解決するウェブサイトのことであり，それぞれの質問に対する（多くの場合複数の）回答が 1 つのページにまとめられている（これを QA ページと呼ぶことにする）．QA ページをクラスタリングすることは，質問，あるいは質問の背景にある問題を分類することであると考えられる．そして，場所の情報を抽出することは，それぞれの問題に関わる場所（問題が発生しやすい場所，解決策を見つけられる可能性が高い場所，など）を発見することに対応すると考えられる．

実際の実験に際しては，処理にかかるコストも考慮し上記の課題を次のように具体化した．まず，Q&A サイトには数多くの様々な質問が投稿されており，それを 1 度にま

とめて分類するのは計算コストと評価作業コストの双方の観点から困難である．よって，語 x を選び，その語に関連する QA ページの集合 D をクラスタリングの対象とした．たとえば，x として「かぼちゃ」を選んだ場合，QA ページが家庭における調理に関わる問題，栽培関係の問題などに分類され，それぞれに「台所」や「畑」などが対応付けられることが期待される．D の要素を収集する手段として検索エンジンを利用した．具体的には，x とともに検索対象とするサイトを指定するオプション（たとえば，“site:”）をクエリに用いて指定 Q&A サイト内で x に関連するページを検索した．

場所に関する情報の抽出については，場所を表す語の集合 L をあらかじめ用意しておき，その中から適切なものを選び出す方法を選んだ．L を作成するため，公園，交差点といった場所，学校や劇場といった建造物，書店や肉屋といった店舗，それに居間や台所といった部屋の種類を，辞書や Web ページ中の表などから抽出した．その結果，99 語からなるリストを得た．L の要素 y を文書集合 $S \subset D$ の特徴語として採用するか否かは，χ^2 検定により判断した．具体的には，「語 y の出現」と「S への帰属」という 2 つの条件に関する D の分割表（条件への適合/不適合の組合せに該当するものの数を示した表）

	y	$\neg y$
S	a	b
$\neg S$	c	d

を作成し，χ^2 値とその p 値を計算する．p 値が有意水準を下回るとき，y を S の特徴語として採用する．本実験では有意水準を 0.05 とした．なお，χ^2 検定の p 値の代わりに相互情報量を基準にした実験も行った．その内容と結果については 5.4.2 項で述べる．

特徴語抽出の方法としては，tf・idf [14] のような指標によりクラスタに属する文書に含まれるすべての語を順位付けし，その上位 n 語を採用するというアプローチも考えられる．この場合，採否の基準となる n を決めるための妥当な方法（根拠）が必要である．さらに，処理結果（抽出した特徴語）を比較する際，語間の意味的関連性も考慮しなければならない（たとえば，「バーベキュー」というトピックに関する文書集合から「炭」が抽出された場合と「木炭」が抽出された場合とを同等な結果と判断する，など）．これらの問題を回避するため，本論文では，意味的な独立性が比較的高い語の集合 L をあらかじめ用意し，各語の特徴語としての適否を判定するという方法をとった．

改めて評価実験の手順をまとめると以下のようになる．

(1) x に関連する QA ページを収集する（これを D とする）．
(2) 既存のクラスタリングアルゴリズムにより D を分類し，クラスタ $\{C_i\}$ を得る．
(3) 各 C_i から一貫性阻害要素を除去したクラスタ I_i を作成する．
(4) L の中から C_i，I_i それぞれの特徴語を選出し，比較する．

上記 (4) において，一貫性阻害要素除去の効果を確認することが本実験の最終的な目的であるが，その途中の段階においてクラスタリング品質の改善状況も確認した．具体的には，(2) で得たクラスタリング結果の品質を評価し，(3) の一貫性阻害要素除去後の結果と比較した．それぞれの結果を 5.2 節と 5.3 節で，特徴語抽出における効果の評価を 5.4 節で示す．

5.2　クラスタリング品質の評価

本節では，上記手順の (2) で得られたクラスタリングの品質の評価結果を示す．これは，（質問者が問うている）問題の分類という観点から，既存のアルゴリズムが QA ページをどれだけ適切に分類できているかを評価したものである．評価の結果について述べる前に，評価対象としたアルゴリズムの概要を説明する．

5.2.1　既存のクラスタリングアルゴリズム

凝集型階層的手法

凝集型階層的手法は，広く使われている基本的なクラスタリング手法の 1 つである [1]．この手法では，まず個々のデータだけからなるクラスタを生成して初期状態とし，逐次，最も距離の近いクラスタを併合していく．クラスタ間の距離はデータ間の距離（非類似度）を使って定義される．たとえば，単連結法や群平均法におけるクラスタ間距離は次のように定義される：

$$\ell_s(C_1, C_2) = \min_{d_1 \in C_1, d_2 \in C_2} \mathrm{dist}(d_1, d_2),$$

$$\ell_a(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{d_1 \in C_1} \sum_{d_2 \in C_2} \mathrm{dist}(d_1, d_2),$$

ここで，$\mathrm{dist}(d_1, d_2)$ はデータ d_1 と d_2 の距離である．

凝集型階層的手法では，クラスタを併合していく時点（ステップ）ごとにそれぞれクラスタの集合が得られる．単連結法および群平均法で k ステップまで併合したときに得られるクラスタの集合をそれぞれ $\mathrm{ACS}(k)$，$\mathrm{ACA}(k)$ と書くことにする．

分割最適化手法

分割最適化手法では，データ分割の適切さを評価する関数を定め，その最適解を探索する．本手法の代表的な例である k 平均法では，各クラスタに属するデータとクラスタの重心との距離の平方和を評価関数とし，それが最小となるようにデータを k 個のクラスタに分割する．k 平均法で得られたクラスタ集合を $\mathrm{KM}(k)$ と書くことにする．

生成モデルベースの手法

生成モデルベースの手法では，データ生成に関する確率モデルを仮定し，観測値であるデータを当該モデルにあてはめて分類を行う．その一例として，ここでは，潜在ディリ

クレ配分法（latent Dirichlet allocation; LDA）[18] を使った方法を示す．

LDA では，各文書を複数の潜在トピックの確率分布としてとらえ，各潜在トピックを単語の確率分布としてとらえる．つまり，文書における語の出現は，これらの分布の組合せにより確率的に起きると考える．この仮定のもと，文書における単語の出現頻度から，それぞれの確率分布を推定する．なお，LDA では潜在トピック数 k をあらかじめ決めておく必要がある．LDA により推定される確率分布をもとに，文書とその主たる潜在トピックを対応付けることができ，その結果として文書の分類が得られる．文書 d_j に i 番目の潜在トピックが現れる確率を θ_{ji} としたとき，i 番目の潜在トピックに対応する文書クラスタは次のようにして得られる：

$$C_i = \{d_j | \theta_{ji} \geq \theta_{jm} \text{ for } m \neq i\}$$

この結果得られるクラスタ集合を LDA(k) と書く．

なお，LDA を用いたクラスタリングにおいては，次に示すように，トピックと文書の関係を一貫性阻害要素除去に利用することも考えられる．まず，クラスタ C_i に属する文書をトピック分布に基づいて順序づける：

$$C_i = \{d_{j_1}, d_{j_2}, \ldots, d_{j_n}\},\ \theta_{j_1 i} \geq \theta_{j_2 i} \geq \cdots \geq \theta_{j_n i}$$

ここで $\theta_{j_k i}$ は i 番目の潜在トピックが文書 d_{j_k} に生起する確率である．このシーケンスの前方にある文書ほどトピックとの関連性が高いと考えられるので，後方にある文書を一貫性阻害要素と見なし除去する．より具体的には，$0 \leq p \leq 1$ である p に対して上記シーケンスを $p : (1-p)$ に分け，後方を除去する．除去の結果得られた文書クラスタを LDA^p(k) と表記する．

ネットワークのコミュニティ抽出を利用した手法

データにネットワーク構造を定義できる場合，すなわち，個々のデータをノードとしデータ間の関係を示すリンクが定義できる場合には，コミュニティ分割アルゴリズムを利用してデータを分類することができる [20]．ネットワークにおけるコミュニティとは，内部のつながりが密であり，その外部とのつながりが疎であるようなノードの集合として定義される [19]．つながりの疎密を定量的に測る方法は今までに複数提案されており，コミュニティ分割も様々な手法が開発されている [21]．手法間の性能比較も行われており，walktrap 法 [22] がデータへの依存性が低く他のものより良い性能を示すという結果が得られている [23]．

文書集合に対しては様々なネットワーク構造が定義できるが，ここでは，以下に示す手順で構成される語の共有関係を表すネットワークを考える．すなわち，(1) 各文書から（たかだか）k 個の特徴語を抽出し (2) m 個の特徴語を共有する文書間に重さ m の無向リンクを生成する．このネットワークに対して walktrap 法を適用して得られたクラスタリング結果を CD(k) と書く．

5.2.2　既存アルゴリズムのクラスタリング品質

上記アルゴリズムによるクラスタリング品質を評価するため，x として 1 つの語を選んで D を作成し，実際にクラスタリングを行い，purity，inverse purity，F 値を測った．なお，LDA などの確率的アルゴリズムについては，乱数生成シードを含む初期値を変えて複数回クラスタリングを実行して評価指標を計算し，その平均をもって当該アルゴリズムを評価した．

今回の実験では x として「かぼちゃ」を選び，728 文書からなる D を作成した．D の各文書の内容を人手で調べた結果，ハロウィンや冬至などのイベントに加え，離乳食やダイエットなど食品としての利用に関するもの，かぼちゃの切り方などの調理方法に関するもの，農作物として栽培に関するもの，小動物の飼育に関するもの（種が餌になる）など 43 種類の問題に分類することができた．問題の分類にあたっては，その状況の属性（時間，場所），特徴的な構成要素（人，もの），人の行為，その目的，構成要素間の相互関係の違いを基準とした．たとえば，ダイエット食と離乳食に関する質問は，双方ともかぼちゃの食べ方に関するものであるが，食事の主体者や目的が異なるため異なるものとして区別した．本実験では，この人手による分類を正解とし，クラスタリングの品質を評価した．この正解の客観性には議論の余地があると思われるが，本実験の目的は一貫性除去による改善の効果を確認すること，言い換えると除去前後の比較評価することであるので，多少の恣意性は許容できると判断した．人手により得られた問題のクラスのうち，いくつかを**表 1** に示す．

図 2 は，ACS(k)，KM(k)，LDA(k)，CD(k) それぞれのクラスタリング品質を示したものである．いずれも，パラメータ k に対して定まるクラスタリング結果の purity，inverse purity および F 値をプロットしたものである．4 者中，LDA が最も安定して（パラメータへの依存性が低く）高い分類品質を示している．しかし，その LDA においても purity が 0.6 を超えることはなかった．すなわち，この

表 1　「かぼちゃ」に関連する状況の例
Table 1　Examples of situations related to “pumpkins”.

問題の内容・状況	状況におけるかぼちゃの役割/特徴
あがり症の克服	「人をかぼちゃと思う」という対策
かぼちゃの栽培	農作物としてのかぼちゃ
かぼちゃの切断	硬くて切り難いかぼちゃの皮
健康食	食物繊維が多く栄養価の高い食材
しみの除去	頑固なかぼちゃのしみ
小動物の飼育	小動物の餌（種子）
シンデレラ	物語に登場するかぼちゃの馬車
バーベキュー	バーベキューの材料
ハロウィン	ランタンの材料
離乳食	離乳食の材料

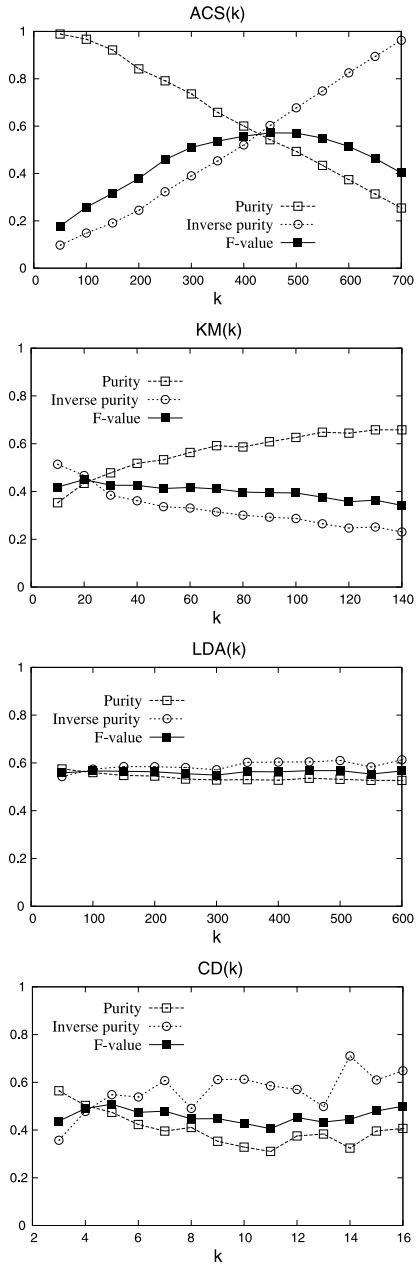

図 2　ACS(k)，KM(k)，LDA(k)，CD(k) のクラスタリング品質
Fig. 2　Clustering qualities of ACS(k), KM(k), LDA(k) and CD(k).

結果は，おおむねクラスタの 4 割強が一貫性阻害要素で占められていることを示している．

5.3　クラスタの一貫性向上における効果

既存アルゴリズムのクラスタリング品質を確認したところで，本節では一貫性阻害要素除去手法を実際に適用した結果を示し，その効果を確認する．具体的には，他のクラスタリングアルゴリズムに比べ安定して高い F 値を示していた LDA によるクラスタリング結果から，その他のアルゴリズムを利用して一貫性阻害要素を除去する．

まず，LDA(m) によるクラスタリング結果を ACA(k) を補助的クラスタリング手法として用いて一貫性阻害要素除去を試みた．ここでは，$m = 400$ の場合を示す．**図 3** は除去後の purity，F 値（縦軸左目盛り），それに文書数（縦軸右目盛り）を k ごとにプロットしたものである．本手法は，クラスタ内の一貫性を損なう要素を除去するという点が特徴的であった．いわば，データの量を犠牲にして質の向上を狙う手法であるといえる．図 3 では実際に，k の値が小さくなるのに従い，文書数は減少し purity は増加していることが見て取れる．LDA によるクラスタリングでは purity はおよそ 0.5 であったが，一貫性阻害要素除去によりその値が増加している．たとえば，補助的クラスタリング手法として ACA(500) を採用すると，purity の値がおよそ 0.8 である約 200 文書からなるデータが得られる．また，purity の増加にともない F 値も増加しているのが分かる．これは，本手法により一貫性と集中性が総合的に改善されていることを示している．

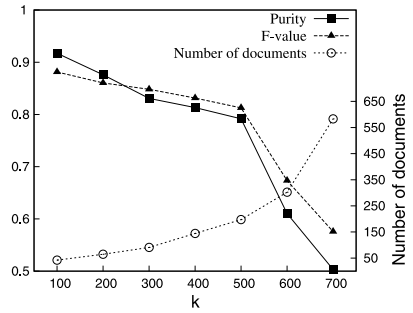

図 3　ACA による一貫性阻害要素除去の効果
Fig. 3　Effects of cluster purification using ACA.

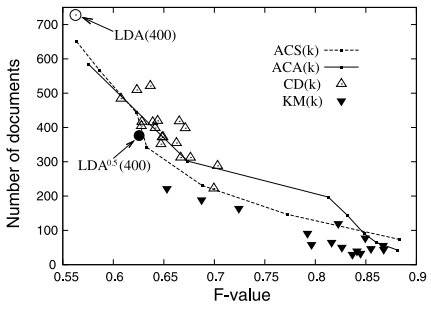

図 4　一貫性阻害要素除去の効果
Fig. 4　Effects of cluster purification techniques.

さらに，ACS(k)，CD(k)，KM(k) を補助的クラスタリング手法に用いた場合の効果，さらに LDA のトピック分布を利用した手法である LDA^p(k) による効果を調べた．LDA^p(k) については，LDA(k) の purity が 0.5 程度であったことから，$p = 0.5$ とした．**図 4** は，それぞれの手法を用い k を変化させて一貫性阻害要素除去を行い，得られた結果ごとに F 値と除去後残った文書の数をプロットしたものである．ACS と ACA に関しては，k の変化にともなって F 値，文書数はそれぞれ単調に減少/増加する．この変化傾向を示すため，プロットした点を線で結んで示した．

図中，LDA(400) のクラスタリング結果に対応する点は

左上に位置している．もし，文書を取りこぼしなくより正確に分類する理想的なクラスタリングがあるとすれば，図の右上の部分に位置することになる．一方，提案手法による結果は，左上から右下にかけて分布している．これは，データの量を犠牲にして質を向上させているということであり，いずれの一貫性阻害要素除去手法においても期待される効果が得られていることを示している．

さらに，この結果をもとに手法どうしを比較することができる．先にも述べたように，プロットされた点がより右上方向に分布していることが望ましい．この観点からすると，LDA$^p(k)$ あるいは KM(k) は補助的クラスタリング手法として他の手法より劣っているといえる．

5.4　特徴抽出における効果

最後に，実験のステップ (4) の結果を示す．具体的には，LDA(k) で得られた各クラスタ C_i から抽出された特徴と，C_i より一貫性阻害要素を除去して得られた I_i から抽出された特徴の違いを明らかにする．本節でも 5.3 節と同様に $k = 400$ の場合を示す．なお，本実験では，抽出すべき特徴として「場所」を選んだ（5.1 節）ことを思い出してほしい．C_i とその内容的一貫性を向上させた I_i は，質問あるいは質問の背景にある問題を状況の類似性に基づいてまとめた問題のクラスに対応しており，場所の情報を抽出することは，それぞれの問題に関わる場所を発見することに対応すると考えられることを，ここで改めて言及しておく．

5.4.1　実験内容と評価方法

C_i から一貫性阻害要素を除去するための補助的クラスタリング手法として CD(m) を用いた．実際の実験では m として様々な値を試したが，本論文ではその中から $m = 5$, 13 の結果を示す．

一貫性阻害要素除去の効果を評価する方法としては，C_i から抽出した場所の集合と，I_i から抽出した場所の集合とを比較し，前者になく後者でのみ得られた場所について，その量と質（適/不適）を調べるという方法をとった．質の評価のため，抽出された場所は，問題に直接関係するもの，直接関連しないが間接的な関連性が認められるもの，関連性が認められないもの，のいずれかに主観により分類した．関係が直接的であるか否かは，その場所で問題が発生しているか，あるいは問題を解決するためにその場所が使われているか否かで判断した．たとえば，「家計のやりくり」という問題に対して「駐車場」が抽出されたが，これは間接的に関係する場所に分類した．なぜなら，駐車場（を借りること）にかかる費用は家計に関係する（実際，質問に対する回答の中でそういった内容のコメントが寄せられている）が，駐車場でその問題が発生しているわけではなく，駐車場でその問題が解決されるわけでもないからである．

また，この実験ではステップ (1) の x として「かぼちゃ」以外の語も試した．その語を選ぶ手段として，Web 上にあるピクトグラムに関する情報を利用した．ピクトグラムは（日常で関わることの多い）事物を指し示す絵文字である．そのキャプションを集め，その中から無作為に 10 語を選んで実験に用いた．

5.4.2　評価結果

まず，抽出された場所の数をクラスタ単位で集計した．その結果を**表 2** に示す．本表は，語 x ごとに，何らかの場所が抽出された C_i の数 $\#C$ と I_i の数 $\#I$，問題と直接的関係のある場所が抽出された I_i の数 $\#I^+$ を示したものである．表中の a，b そして ab はそれぞれ，補助的クラスタリング手法として CD(5)，CD(13) を用いた場合と，それらの結果をマージした場合の数を示している．たとえば，「灯り」に関する文書集合の場合，23 個の C_i から何らかの場所が抽出され，CD(5) で一貫性阻害要素除去した場合には，8 個の I_i から適切な場所が抽出できている．一貫性阻害要素除去のおおまかな効果は $\#I/\#C$ と $\#I^+/\#I$ で把握できる．これらはそれぞれ，何らかの場所と関連付けられている C_i のうち一貫性阻害要素除去により新たな場所との関連性が見い出されたものの割合と，そのうち適切な場所が抽出されたものの割合である．a，b，ab それぞれについて計算した結果を**表 3** に示す．表から，たとえば ab の場合，何らかの場所と関連付けられている C_i の 4 割強から一貫性阻害要素除去により新たな場所との関連性が見い出され，そのうち 8 割弱が妥当なものであったことが分かる．

表 2　場所が抽出されたクラスタの数
Table 2　The number of clusters from which places were extracted.

x	$\#C$	$\#I$			$\#I^+$		
		a	b	ab	a	b	ab
灯り	23	9	5	11	8	5	10
お皿	52	16	19	27	15	14	22
かぼちゃ	17	7	4	9	7	3	9
きのこ	29	8	7	11	8	4	9
小包	31	5	8	11	1	4	5
シャワー	65	24	18	28	17	14	21
台所	76	23	17	31	17	13	23
時計	72	25	19	32	21	15	27
封筒	67	16	13	21	10	10	14
ローソク	17	1	4	4	1	4	4
	449	134	114	185	105	86	144

表 3　クラスタ数で評価した一貫性阻害要素除去の特徴抽出における効果
Table 3　Effect of the cluster purification technique on feature extraction evaluated on the basis of the number of clusters.

	a	b	ab
$\#I/\#C$	0.2873	0.2561	0.4120
$\#I^+/\#I$	0.7519	0.7565	0.7784

表 4　抽出された場所の数
Table 4　The number of extracted places.

x	L^+			$L^\pm$			L^-		
	a	b	ab	a	b	ab	a	b	ab
灯り	11	7	16	2	0	2	1	0	1
お皿	21	15	30	3	4	7	2	5	7
かぼちゃ	9	4	14	4	1	5	0	1	1
きのこ	10	5	13	1	1	2	2	3	4
小包	1	5	6	2	3	4	2	2	4
シャワー	28	22	47	5	3	7	9	4	11
台所	22	18	39	10	5	14	5	3	7
時計	34	22	52	7	3	10	4	5	8
封筒	14	15	25	7	5	11	6	2	6
ローソク	3	6	9	1	0	1	0	1	1
	153	119	251	42	25	63	31	26	50

表 5　特徴の数で評価した一貫性阻害要素除去の特徴抽出における効果
Table 5　Effect of the cluster purification technique on feature extraction evaluated on the basis of the number of features.

	a	b	ab
L^+/L^*	0.6770	0.7000	0.6896
$(L^+ + L^\pm)/L^*$	0.8628	0.8471	0.8626

次に，抽出された場所の数を集計した結果を**表 4** に示す．L^+，$L^\pm$，L^- はそれぞれ問題に直接的関係のある場所，間接的関係のある場所，関係のない場所（つまり誤抽出）の数であり，a，b，ab は表 2 と同じく補助的クラスタリング手法を表している．この結果をもとに，補助的クラスタリング手法ごとに，抽出されたすべての場所の数 $L^* = L^+ + L^\pm + L^-$ に対して問題と直接的関係ある場所占める割合 L^+/L^*，直接あるいは間接的関係のあるものの割合 $(L^+ + L^\pm)/L^*$ を計算し，**表 5** にまとめた．前者は，問題と直接関係ある場所を特徴抽出の正解としたときの精度であり，後者は間接的関係のあるものまで正解に含めた場合の精度である．間接的関係まで含めれば，補助的クラスタリング手法の選択にかかわらず 85%前後の精度が得られている．

さらに，誤抽出の原因を把握するため，実際の Q&A ページを参照し，それぞれの場所が言及されている文脈を調べた．その結果，誤抽出は 5 種類に大別できることが分かった．以下，それらを多い順に列挙する（括弧内は誤抽出全体に占める割合）：問題に関連して語られているエピソードの中に出現するもの（0.32），語の多義性に起因し，同一の表記だが異なった意味で使われているもの（0.26），商品や芸術作品など，ものの名前の中に出現しているもの（0.16），「～に比べて」など修飾（説明）のために使われているもの（0.14），そもそも，問題の主旨と関連しない文書に含まれていたため関連性が認められないもの（0.12）．

表 6　質問内容（問題）に対応付けられた場所の例
Table 6　Examples of places associated with the questions (problems).

x	質問内容	対応付けられた場所
灯り	天体観測	空，田舎，パチンコ屋
お皿	TV アンテナ	ベランダ，屋根，電気屋
かぼちゃ	バーベキュー	公園，肉屋，100 円ショップ
きのこ	がん治療	病院，薬局，温泉
小包	宅配便	会社，玄関，酒屋
シャワー	結婚式	教会，ホテル，階段
台所	下水処理	市役所，道路，風呂
時計	身だしなみ	レストラン，百貨店，クリーニング店
封筒	銀杏の処理	100 円ショップ，川，ベランダ
ローソク	誕生祝い	レストラン，花屋，遊園地

表 7　質問内容（問題）に対応付けられた場所の例（相互情報量を用いた場合）
Table 7　Examples of places associated with the questions (problems) on the basis of mutual information.

x	質問内容	対応付けられた場所
灯り	結婚式での撮影	レストラン
お皿	膝のトラブル	電車
きのこ	原発事故（食材の汚染）	台所
小包	荷物の発送	コンビニエンスストア
シャワー	髪と頭皮のトラブル	美容院
台所	介護施設（住宅）	風呂

本実験で実際に抽出された場所の例を**表 6** に示す．下線は一貫性阻害要素除去により得られた場所であることを表している．全体的に，一貫性阻害要素除去により得られた場所は，一見すると問題に無関係のようだが実際には関係がある「意外な場所」である傾向が見て取れる．

最後に，χ^2 検定の p 値の代わりに相互情報量 [24] を基準にした場合でも提案手法が有効であることを確認した結果を示す．一般に，文書集合 D において，$d \in D$ における語 y の出現と，d が $S \subset D$ に帰属するという 2 つの事象の相互情報量 $\mathrm{MI}(y, S)$ は

$$\mathrm{MI}(y, S) = \log \frac{\Pr(y, S)}{\Pr(S) \cdot \Pr(y)}$$

で与えられる．ここで，$\Pr(y)$，$\Pr(S)$ はそれぞれの事象の生起確率であり，$\Pr(y, S)$ は同時確率である．2 事象が独立であればこの値は 0 になる．y を C_i の特徴語として採用する条件を $\mathrm{MI}(y, C_i) > 0$ とし（I_i についても同様），χ^2 検定の場合と同じデータを用いて実験を行った．その結果，やはり，一貫性阻害要素除去により関連性のある場所を発見できることを確認した．新たに発見された場所の例を**表 7** に示す．

6. 評価結果のまとめと考察

5.4.2 項で示した結果をもとに，特徴抽出における一貫性阻害要素除去手法の量・質的効果をまとめる．まず，量

について，CD(5) を補助的クラスタリング手法に使った場合（表 2〜5 における a の場合）を例にとると，関係のある場所が新たに抽出されたクラスタの数は合計で 105 個であり，抽出された場所の数は 153 カ所であった．つまり，一貫性阻害要素除去の手法を用いた結果，1 文書集合あたり平均して 10.5 個のクラスタからそれぞれに特徴的な場所を 1.5 個程度新たに発見できたことになる．CD(13) を利用した場合も同程度な結果が得られている．一方，抽出の質（精度）に関しては 5.4.2 項の繰返しになるが，一貫性阻害要素除去により新たに得られたすべての場所のうち 7 割程度が直接的関係を持つものであった．さらに，間接的関係を持つものを合算するとその全体に占める割合は 8 割を超えた．今回の実験の目的は，一貫性阻害要素除去手法により妥当な特徴（場所）を新たに発見できることの確認であったが，これらの結果はその期待に応えるものである．

実際に抽出された場所（表 6）のうち一貫性阻害要素除去により得られたものは問題と意外な関連性を有する傾向にあることを指摘したが，ここに提案手法の 1 つの特徴が現れていると考えられる．一貫性阻害要素除去のそもそもの目的は，目標とする文脈に合致しない文書を除去することでクラスタの一貫性を向上させることであった．これに対し，提案手法では，2 つのクラスタリング結果の共通部分をとることにより，もともとの文脈を特化（サブトピックを抽出）し，その結果として一貫性を向上させていると考えられる．たとえば，「灯り」に関連する「天体観測」の話題では，提案手法により「パチンコ屋」が抽出されたが，これは異なるクラスタリング手法による結果の共通部分をとることにより，天体観測を難しくしている光害という話題に焦点をあてたクラスタが生成された結果である．提案手法において異なる補助的クラスタリング手法を用いれば異なるサブトピックが得られ，その結果として抽出される特徴も変わると考えられる．実際，実験ではこの考えを支持する結果が得られている．表 5 の L^+ の欄に a，b として示した異なる補助的クラスタリング手法による結果を比較すると，a，b に共通するものの総数 21（$= 153 + 119 - 251$）は，これらをマージしたものの総数 251 の 1 割にも達しておらず，二者の共通性が低いことが分かる．

7. むすび

本論文では，クラスタリングにより文書データを意味的に分類しそれぞれの特徴を表す情報（特徴語）を抽出するという目的のため，各クラスタから一貫性阻害要素を除去するいうクラスタリング結果の改善方法を提案し，実験によりその効果を確認した．提案手法は，いわば，データの量を犠牲にして質の向上を狙う手法であるといえる．ただし，実際に特徴を抽出する際には，データ量もある程度確保する必要がある．よって，本手法の効果的な適用のためには，データ量とクラスタリング品質のバランスを調整する必要がある．調整の指針など，適用方法について今後検討を進めていく．

参考文献

[1] Jain, A.K., Murty, M.N. and Flynn, P.J.: Data clustering: A review, *ACM Computing Surveys*, Vol.31, No.3, pp.264–323 (1999).

[2] Clusty, available from ⟨http://clusty.com⟩.

[3] Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S. and Amigó, E.: WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks, *CLEF 2009 Working Notes* (2009).

[4] Yarowsky, D.: Unsupervised word sense disambiguation rivalling supervised methods, *Proc. 33rd Annual Meeting on Association for Computational Linguistics*, pp.189–196 (1995).

[5] Blum, A. and Mitchell, T.: Combining labeled and unlabeled data with co-training, *Proc. 11th Annual Conference on Computational Learning Theory*, pp.92–100 (1998).

[6] Goldman, S.A. and Zhou, Y.: Enhancing Supervised Learning with Unlabeled Data, *Proc. 17th International Conference on Machine Learning*, pp.327–334 (2000).

[7] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations, *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp.281–297 (1967).

[8] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, *Proc. 2nd International Conference on Knowledge Discovery and Data Mining*, pp.226–231 (1996).

[9] Girolami, M.: Mercer Kernel-Based Clustering in Feature Space, *IEEE Trans. Neural Networks*, Vol.13, pp.780–784 (2002).

[10] Strehl, A. and Ghosh, J.: Cluster ensembles – A knowledge reuse framework for combining multiple partitions, *The Journal of Machine Learning Research*, Vol.3, pp.583–617 (2003).

[11] Barthélemy, J.-P. and Leclerc, B.: The median procedure for partition, *Partitioning Data Sets*, Cox, I.J. et al. (Eds.), AMS DIMACS Series in Discrete Mathematics, Vol.19, pp.3–34 (1995).

[12] de Amorim, S., Barthélemy, J.-P. and Ribeiro, C.: Clustering and clique partitioning: Simulated annealing and tabu search approaches, *Journal of Classification*, Vol.9, No.1, pp.17–41 (1992).

[13] Analoui, M. and Sadighian, N.: Solving cluster ensemble problems by correlation's matrix and GA, *IFIP International Federation for Information Processing*, Vol.228, pp.227–231 (2006).

[14] Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).

[15] Ghaemi, R., Sulaiman, M.N., Ibrahim, H. and Mustapha, N.: A Survey: Clustering Ensembles Techniques, *World Academy of Science, Engineering and Technology*, Vol.38, pp.644–653 (2009).

[16] Guha, S., Rastogi, R. and Shim, K.: CURE: An efficient clustering algorithm for large databases, *Proc. 1998 ACM SIGMOD International Conference on Management of Data*, pp.73–84 (1998).

[17] Amigó, E., Gonzalo, J., Artiles, J. and Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints, *Information Retrieval*,

Vol.12, No.4, pp.461–486 (2009).
[18] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet allocation, *The Journal of Machine Learning Research*, Vol.3. pp.993–1022 (2003).
[19] Newman, M.E.J.: *Networks: An Introduction*, Oxford University Press (2010).
[20] Oliveira, T.B.S., Zhao, L., Faceli, K. and de Carvalho, A.C.P.L.F.: Data clustering based on complex network community detection, *IEEE Congress on Evolutionary Computation*, pp.2121–2126 (2008).
[21] Fortunato, S.: Community detection in graphs, *Physics Reports*, Vol.486, pp.75–174 (2010).
[22] Pons, P. and Latapy, M.: Computing communities in large networks using random walks, *Journal of Graph Algorithms and Applications*, Vol.10, No.2, pp.191–218 (2006).
[23] Orman, G.K. and Labatut, V.: A Comparison of Community Detection Algorithms on Artificial Networks, *Proc. 12th International Conference on Discovery Science*, pp.242–256 (2009).
[24] Fano, R.: *Transmission of Information*, MIT Press (1961).

佐藤 進也　（正会員）

1986 年東北大学理学部数学科卒業．1988 年同大学大学院修士課程修了．同年日本電信電話株式会社入社．以来，協調作業支援，Web 情報検索・マイニング，複雑ネットワーク等の研究に従事．現在，NTT 未来ねっと研究所主任研究員．博士（情報理工学）．訳書「スモールワールド」（ダンカン・ワッツ著，東京電機大学出版局，共訳）．ACM，ISOC，電子情報通信学会，人工知能学会各会員．

高橋 公海

2008 年筑波大学図書館情報専門学群卒業．2010 年同大学大学院図書館情報メディア研究科博士前期課程修了．同年日本電信電話株式会社入社．現在，NTT 未来ねっと研究所所属．主にデータマイニング，機械学習等の研究に従事．

松尾 真人　（正会員）

1986 年京都大学工学部精密工学科卒業．1988 年同大学大学院修士課程修了．同年日本電信電話株式会社入社．以来，適応型ネットワークサービス技術，ユビキタスコンピューティング技術の研究に従事．現在，NTT 未来ねっと研究所主幹研究員．電子情報通信学会会員．

（担当編集委員　小西 修）