共起語に基づいた階層型文書クラスタリング手法

村上 浩司　　橋本 泰一　　乾 孝司
内海 和夫　　石川 正道

概要：本研究は，クラスタ毎の重要な数文書のみを効率よく提示することを目的とし，文書-単語マトリクスの代わりに単語-文書行列を入力として，まず単語をクラスタリングし，それらの単語を含む文書を間接的にクラスタリングする文書クラスタリング手法を提案する．提案手法は文書クラスタリングの際に，入力行列に用いた単語の tf·idf 値に基づいたスコアリングにより各文書クラスタに所属する文書をランキングし，その上位の文書のみを出力とすることで，クラスタの特徴を表す文書を同定できる利点がある．提案手法によって得られた各クラスタの上位ランクの文書は，他クラスタの文書とは排他的な特徴を持ちつつ，高い精度でクラスタリングされていることが確認された．

キーワード：共起語，単語-文書行列，階層型文書クラスタリング，文書ランキング

A Hierarchical Document Clustering Method Based on Co-occurence Words

MURKAMI Koji　　HASHIMOTO Taiichi　　INUI Takashi
UTSUMI Kazuo　　ISHIKAWA Masamichi

Abstract： In this paper, we report the results of our investigation of the new document clustering approach which is based on term-document matrix instead of document-term matrix. In our clustering approach, the terms are clustered by considering co-occurence words among documents. At the time, the documents are also clustered indirectly because the documents include the classfied terms. The documents in each cluster are ranked by weighting of terms. This process is able to identify the documents which charactelize the cluster. In this paper, we show higher clustering performance than general document clustering approach.

Keywords： Co-occurence words, Term-document matrix, Hierarchical Document Clustering, Document ranking

1　はじめに

電子化文書の増加に伴い，膨大な文書を俯瞰的，効率的に分類，整理することが非常に重要になっている．これを実現する 1 つの技術として文書クラスタリングがある．結果として得られたクラスタに属する文書は，出現する単語の分布が類似していることから，クラスタ別に文書を分析することで，それぞれのクラスタが示すトピックなどを大まかに捉えることができる．

基本的にクラスタリングアルゴリズムは，対象のデータをその特徴の類似度から内的なまとまりと外的な分離を達成することのみを目的としている．文書クラスタリングの場合，アルゴリズムは各文書クラスタに所属する文書の重要性を考慮しないために，クラスタの特徴を示す文書を同定しない．しかしながら実際に，文書クラスタリング結果から文書群の俯瞰的な分析を行うには，各クラスタのトピックを表現するような中心的な文書をまず読む必要がある．ところがクラスタリングそのものでは，代表

東京工業大学 統合研究院
Integrated Research Institute, Tokyo Institute of Technology
連絡先：murakami@iri.titech.ac.jp

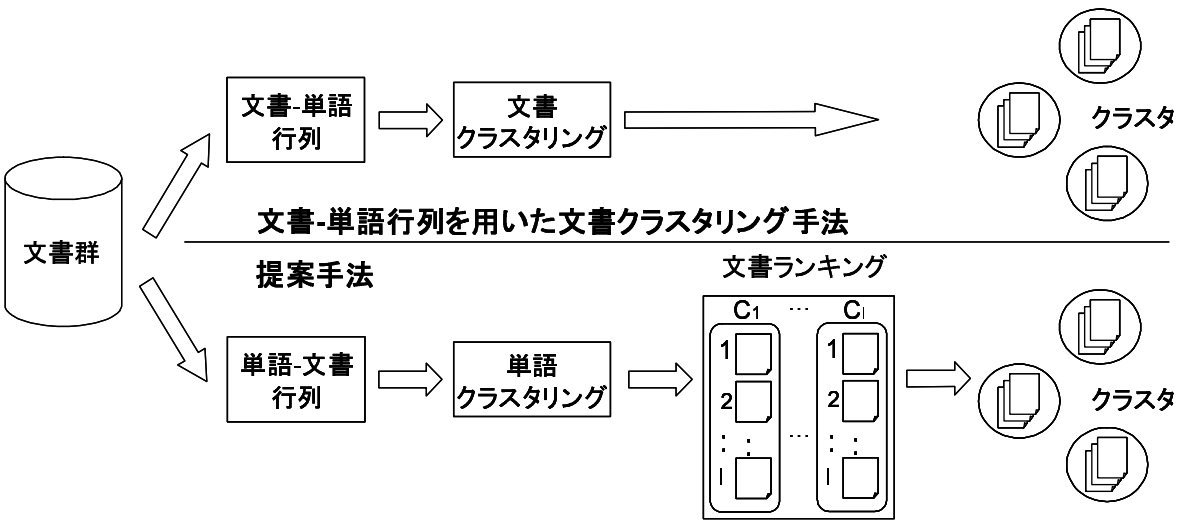

図 1　提案手法の構成と一般的なクラスタリング処理との比較

的な文書が同定されないために，分析者自身がクラスタ内のすべての文書に目を通し，中心的な文書を特定しなければならない．対象の文書群の規模が大きくなった場合には，こうした作業には多大な時間と労力が必要となる．我々は現在，新聞記事からの社会課題抽出 [10] を検討しているが，対象とする新聞記事群の規模は小さくはないことから，こうした問題に直面している．

そこで我々は，文書クラスタリング結果の俯瞰的な分析を行うために，クラスタ毎に大まかなトピックを表す代表的な文書の効率良い提示を目的とした文書クラスタリング手法について検討した結果を報告する．

本論文の構成は以下の通りである．第 2 章では関連研究について述べ，第 3 章では提案手法の具体的なアルゴリズムについて説明する．

第 4 章で評価実験，第 5 章で結果についての考察を行い，第 6 章でまとめを行う．

2　関連研究

対象データの全体の俯瞰的情報や傾向などを効率よく提示するといった分析者の実作業を意識したデータマイニング手法としては，例えば藤井らの研究 [9] がある．この研究ではグラフ構造データを扱い，クラスタリングすることで大まかにデータを分けて，そのクラスタの少数のサンプルを提示するというアプローチである．しかしながら，クラスタ内の代表的なサンプルを同定し提示するまでには至っていない．

単語の頻度情報に加えて共起単語情報を文書クラスタリングの入力ベクトルとして用いた研究は，小熊ら [8] により行われた．その結果，共起単語の情報は文書クラスタリングに対して大きく寄与しないという結果が得られているが，クラスタリングアルゴリズムや対象の文書の規模などにおいて，十分に検証されていない．

共起情報に着目し，文書-単語行列ではなく単語-文書行列を適用した研究には，例えば Javier らによる語義曖昧性解消 [1] があり，単語頻度のみの情報を用いた場合と比較して，共起情報の有効性を示した．

また，文書-単語行列の代わりに単語-文書行列を用いた研究には，汎用連想計算エンジンの GETA で用いられている連想検索 [5] などがある．ここでは，入力文書群に近い単語集合を検索し，次にその単語集合から関連する文書群の検索を行う際に，単語-文書行列を利用している．関連文書を同定するために，単語集合の共起情報を利用する点では，本研究においての単語-文書行列の導入と関連するが，本研究では単語-文書行列を，文書クラスタリングの入力として利用する．

3　共起語に基づいた階層型文書クラスタリング手法

3.1　概要

本研究で提案する文書クラスタリング手法の概要を図 1 に示す．図 1 では，上段が文書-単語行列に基づいた文書クラスタリング手法，下段が提案手法の処理の流れである．提案手法はまず，対象の文書群から “単語-文書行列” を作成し，この行列を入力

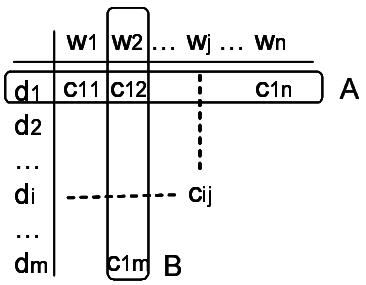

図 2　入力行列の生成

として単語クラスタリングを行う．単語クラスタリングを行うことで，入力に使われた単語を含む文書を間接的にクラスタリングする．このとき各単語クラスタにおいて，これらの単語を含む文書をランキングすることでその単語クラスタの特徴を表す代表的な文書を特定する．提案手法はクラスタリングアルゴリズムではなく，前処理として文書クラスタリングへの入力を文書-単語行列から単語-文書行列へ変換し，文書クラスタリングの後処理として文書クラスタに属する文書をランキングするものである．

以下 ,3.2 で単語-文書行列について説明する．3.3 でクラスタリングアルゴリズム，そして 3.5 で文書ランキングについてそれぞれ説明する．

3.2　単語-文書行列の生成

提案手法は，文書クラスタリングで一般的に用いられる文書-単語行列を入力とするのではなく，単語-文書行列を入力として利用するところが大きな特徴である．単語-文書行列を生成するために用いる素性情報は，文書-単語行列と同一であり，対象文書群からその他の情報を抽出する必要はない．文書-単語入力行列と単語-文書行列における，文書群から抽出した文書情報 d_i と単語情報 w_j と値 c_{ij} の関係を図 2 に示す．

文書-単語行列を入力とした場合は，図 2 中の A のように，行を一つの単位としてクラスタリングする．つまり文書中の単語の出現分布の類似性により，文書をグループ化する．これに対して提案手法で用いる単語-文書行列を入力とする場合は，図 2 の B のように列に着目してクラスタリングを行う．これは対象の文書を，文書間における単語の共起性に基づき分類することとなる．

3.3　クラスタリングアルゴリズム

提案手法は，文書-単語行列を転置した単語-文書行列にして入力しクラスタリングを行うが，3.1 で述べたように本報告ではクラスタリングアルゴリズムそのものの改良は行わない．ここでは，ボトムアップ型の階層型クラスタリングとして広く用いられている UPGMA アルゴリズム [7] および，トップダウン型の Partitioning アルゴリズム [6] を用いる．本報告では，ミネソタ大学の Karypis が公開しているクラスタリングツールキット，CLUTO[2] を用いてクラスタリングを行う．

3.4　文書クラスタリング

単語-文書行列を入力としたクラスタリングは単語クラスタリングであるが，このとき同時に，単語クラスタに属する単語を含む文書を共起語に基づいて集約していることになる．これは間接的な文書クラスタリングであるが，このとき，文書-単語行列を用いた文書クラスタリングと違い，文書クラスタと同時にクラスタの特徴語も得ることができる．

このとき必ずしも 1 文書が 1 クラスタに所属するとは限らず，複数のクラスタに同一文書が属することもある．

3.5　文書ランキング

文書ランキングは図 1 下段右に示される処理である．これは，前節の文書クラスタリングによって得られた文書クラスタ毎に，クラスタを特徴づける代表的な文書の検索を目的とするものである．

文書ランキングは，素性として利用した単語をキーワードとして，それぞれのクラスタ内で行う．

クエリに対して類似する文書を検索するモデルには，確率モデル [12] や概念間の関連度を用いたもの [11] , Okapi[4, 3] の BM25 式などがあるが，本研究では tf/idf 値に基づいた手法を用いた．以下にその検索モデルを説明する．

クラスタ集合 C 中のクラスタ c_i $(c_i \in C, 1 \leq i \leq |C|, |C|$ はクラスタ集合の大きさ$)$ に所属する単語を含む文書群を D_i とすると，c_i に所属する文書を d_{ij} $(d_{ij} \in D_i, 1 \leq j \leq |D_i|)$ と表現できる．

以下の式は，クラスタリングによりまとめられたそれぞれの単語 w の tf·idf 値を計算し，それぞれの文書に含まれる単語の tf·idf 値の合計をスコアとして文書ランキングを行うことを示す．

$$Sc(d_{ij}) = \sum_{w \in d_{ij}} tfidf(w) \tag{1}$$

$$tf \cdot idf(w) = freq(w) \cdot (log \frac{|D|}{df_w} + 1) \tag{2}$$

このとき，$|D|$ は文書全体の大きさを表す．このスコア Sc により文書群中で高頻度に出現し，かつ特定のクラスタに多く出現する単語を含む文書が上位にランキングされ，クラスタを特徴付ける文書として出力される．

表 1　データセットの詳細

セット名	文書数	異なり単語数	クラス数
fbis	2463	2001	17
hitech	2301	126373	6
reviews	4069	126373	5
la1	3204	31472	6
tr31	927	10128	7
tr41	878	7454	10
re0	1504	2886	13
re1	1657	3758	25
k1a	2340	21839	20
k1b	2340	21839	6
wap	1560	8460	20

4　評価実験

提案手法を評価するために，文書-単語行列を用いた文書クラスタリングと，提案手法である共起単語に基づいた，単語-文書行列を用いた文書クラスタリングの分類精度の比較する．

4.1　実験データ

実験のためのクラスタリングデータには，Karypis らが公開しているデータセット[*1]を用いた．このデータセットは 24 種類のデータからなり，それぞれのセットは，文書-単語行列，単語-素性 ID のインデクス，各文書の正解クラス名のファイルで構成される．このうち，Karypis らが CLUTO の性能評価 [6] に用いた 11 種類のデータを用いることとした．表 1 にそれぞれのデータの詳細を示す．各データの行列では，単語の出現頻度がベクトルの値である．

4.2　評価尺度

本実験では，ある正解クラスの文書がそれぞれのクラスタにどれくらい適切に所属するかに着目する必要がある．そこで評価尺度として，Entropy，Purity， Recall， Precision の 4 種類を用いた．Recall と Precision は正解が与えられるときに一般的に用いられる評価尺度であるため，ここでは説明を割愛する．Entorpy は各クラスの文書がどのようにそれぞれのクラスタに割り当てられたかを表し，Purity は，クラスタ C_r に属する文書が，クラスタ内の多数を占める正解クラスに分類されたと考えたときの正解率である．Entropy，Purity ともに範囲は [0,1] であり，最適なクラスタリングが行われた場合，ただ 1 つのクラスの文書が，1 つのクラスタに所属することになる．この場合は Entoropy は 0，Purity は 1 となる．

クラスタ C_i の大きさを n_i としたときの Entropy は，

$$E(C_i) = -\frac{1}{log(c)} \sum_{j=1}^{c} \frac{n_i^j}{n_i} log \frac{n_i^j}{n_i} \qquad (3)$$

で定義され，c はクラス数，n_i^j は j 番目のクラスの文書のうち，i 番目のクラスタに所属する文書の数を表す．クラスタリング全体の Entropy は，下の式で示すように，重みを付けた各クラスタのエントロピーの合計となる．

$$Entropy = \sum_{i=1}^{k} \frac{n_i}{n} E(C_i) \qquad (4)$$

また，クラスタ S_i および，全体の Purity は以下のように定義される

$$P(C_i) = \frac{1}{n_i} \max_j (n_i^j) \qquad (5)$$

$$Purity = \sum_{i=1}^{k} \frac{n_i}{n} P(C_i) \qquad (6)$$

4.3　クラスタリング

提案手法は，クラスタリングの前処理としてクラスタリングへの入力を単語-文書行列へ変換し，後処理としてクラスタ内の文書をランキングするものである．そのため，基本的にクラスタリングアルゴリズムに依存しない．そこで，Partition[6] および UPGMA[7] の 2 種類のクラスタリングアルゴリズムに対して提案手法を適用して評価を行う．また，クラスタリング中で用いる文書間の類似度には，Cosine 距離を用いた．

提案手法は 1 文書が複数のクラスタに所属する可能性のあるソフトクラスタリングであるが，ここでは比較する 2 つのアルゴリズムの評価に合わせるためにランキング結果をハード化する必要がある．もし複数のクラスタに同一文書が所属する場合，ランキングスコアがもっとも高くなるクラスタで評価した．

4.4　実験結果

表 2 に Partition アルゴリズム，表 3 に UPGMA アルゴリズムを用いた文書クラスタリング実験の結果を示す．表中の n は提案手法における，文書ランキングによる上位文書の数を示し，5 ，10 ，20 ，50,100 の 5 種類である．また “Ent.”，“Pur.”，“Prec.”，“Rec.” はそれぞれ，Entropy， Purity， Recall，

[*1] http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/chameleon-data.tar.gz

表 2　Partition アルゴリズムを用いた文書クラスタリング結果の比較

	fbis			
method	Ent.	Pur.	Prec.	Rec.
n=100	0.391	0.632	0.377	0.421
n=50	0.328	0.668	0.407	0.410
n=20	0.268	0.708	0.458	0.458
n=10	**0.243**	**0.718**	**0.476**	**0.476**
n=5	0.180	0.753	0.458	0.458
normal	0.343	0.677	0.425	0.390

	hitech			
	Ent.	Pur.	Prec.	Rec.
	0.826	0.372	0.306	0.306
	0.826	0.377	0.323	0.323
	0.763	0.425	0.333	0.333
	0.649	0.483	0.383	0.383
	0.492	0.533	0.399	0.399
	0.619	**0.607**	**0.441**	**0.412**

	reviews			
	Ent.	Pur.	Prec.	Rec.
	0.198	0.900	0.602	0.602
	0.141	0.928	0.604	0.604
	0.150	0.920	0.610	0.610
	0.081	**0.960**	**0.600**	**0.600**
	0.062	0.960	0.640	0.640
	0.426	0.754	0.511	0.559

	la1			
method	Ent.	Pur.	Prec.	Rec.
n=100	0.425	0.727	0.518	0.518
n=50	0.363	0.767	0.573	0.573
n=20	0.350	0.767	0.575	0.575
n=10	**0.321**	0.767	0.566	0.566
n=5	0.289	0.733	0.600	0.600
normal	0.383	**0.796**	**0.654**	**0.672**

	tr31			
	Ent.	Pur.	Prec.	Rec.
	0.581	0.557	0.359	0.385
	0.523	0.600	0.391	0.395
	0.439	0.643	0.364	0.364
	0.307	0.757	0.400	**0.400**
	0.245	0.800	0.457	0.457
	0.220	**0.850**	**0.479**	0.383

	tr41			
	Ent.	Pur.	Prec.	Rec.
	0.488	0.568	0.448	0.668
	0.413	0.640	0.522	0.631
	0.286	0.690	0.575	0.577
	0.257	0.720	**0.630**	**0.630**
	0.187	0.760	0.639	0.639
	0.238	**0.789**	0.485	0.544

	re0			
method	Ent.	Pur.	Prec.	Rec.
n=100	0.508	0.542	0.225	0.337
n=50	0.449	0.597	0.253	0.291
n=20	0.309	0.681	0.276	0.284
n=10	**0.218**	**0.746**	**0.284**	**0.284**
n=5	0.207	0.754	0.261	0.261
normal	0.397	0.620	0.250	0.274

	re1			
	Ent.	Pur.	Prec.	Rec.
	0.593	0.434	0.225	0.449
	0.450	0.541	0.308	0.401
	0.314	0.620	0.348	0.353
	0.222	0.692	0.367	0.367
	0.145	0.768	0.383	0.383
	0.288	**0.719**	**0.381**	**0.459**

	k1a			
	Ent.	Pur.	Prec.	Rec.
	0.580	0.411	0.275	0.331
	0.502	0.466	0.313	0.320
	0.384	0.573	0.374	0.378
	0.273	0.655	0.379	0.379
	0.203	0.700	0.389	0.389
	0.340	**0.688**	**0.394**	**0.340**

	k1b			
method	Ent.	Pur.	Prec.	Rec.
n=100	0.297	0.757	0.445	0.445
n=50	0.301	0.740	0.476	0.476
n=20	0.206	0.833	0.449	0.449
n=10	**0.170**	**0.883**	0.483	0.483
n=5	0.186	0.867	0.500	0.500
normal	0.190	0.839	**0.577**	**0.704**

	wap			
	Ent.	Pur.	Prec.	Rec.
	0.467	0.564	0.303	0.422
	0.306	0.688	0.393	0.429
	0.172	0.808	0.470	0.478
	0.134	**0.830**	**0.475**	**0.480**
	0.096	0.860	0.489	0.489
	0.312	0.704	0.423	0.394

Precision である．“normal” は，文書-単語行列による文書クラスタリングの結果である．提案手法での $n = 10$ の場合と，“nomal” の結果を比較し，高精度の値を太字で強調している．どちらのアルゴリズムにおいても，提案手法では n の値を小さくすることで精度が上昇していることから，クラスタ内の文書ランキングにより適切なクラスの文書が上位にランクされていることが分かる．文書-単語行列によるクラスタリングの結果を Partition，UPGMA アルゴリズムで比較すると，Partition アルゴリズムの方が全体的に良好な結果となった．

表 2 の結果では $n = 10$ の場合と normal の結果を比較すると，11 種類のデータセットでの Entropy と Precision の平均が，提案手法ではそれぞれ 0.261，0.458 なのに対して，normal では 0.341，0.456 と，Precision に大きな差はないが，Entropy が高いことが分かる．また，表 3 の結果では，$n = 10$ の場合を同様に比較すると平均 Entropy と平均

表 3　UPGMA アルゴリズムを用いた文書クラスタリング結果の比較

	fbis			
method	Ent.	Pur.	Prec.	Rec.
n=100	0.561	0.497	0.325	0.385
n=50	0.433	0.587	0.391	0.399
n=20	0.318	0.662	0.435	0.435
n=10	**0.250**	**0.688**	0.447	**0.447**
n=5	0.174	0.788	0.494	0.494
normal	0.427	0.624	**0.469**	0.421

	hitech			
	Ent.	Pur.	Prec.	Rec.
	0.850	0.387	0.266	0.101
	0.804	0.481	0.276	0.113
	0.582	0.614	0.300	0.141
	0.472	**0.708**	0.150	0.150
	0.455	0.643	0.300	0.166
	0.922	0.271	**0.364**	**0.183**

	reviews			
	Ent.	Pur.	Prec.	Rec.
	0.772	0.496	0.302	0.302
	0.753	0.500	0.304	0.304
	0.754	0.450	0.290	0.290
	0.647	**0.520**	0.380	**0.380**
	0.408	0.720	0.600	0.600
	0.872	0.349	**0.435**	0.204

	la1			
method	Ent.	Pur.	Prec.	Rec.
n=100	0.840	0.324	0.215	0.044
n=50	0.750	0.436	0.229	0.066
n=20	0.613	0.600	0.250	0.091
n=10	**0.405**	**0.733**	0.266	0.116
n=5	0.372	0.700	0.233	0.100
normal	0.939	0.297	**0.462**	**0.168**

	tr31			
	Ent.	Pur.	Prec.	Rec.
	0.571	0.561	0.281	0.276
	0.489	0.631	0.364	0.364
	0.377	0.708	0.414	0.414
	0.323	0.729	0.414	0.414
	0.123	0.886	0.428	0.428
	0.357	**0.758**	**0.461**	**0.506**

	tr41			
	Ent.	Pur.	Prec.	Rec.
	0.676	0.435	0.221	0.320
	0.579	0.512	0.306	0.344
	0.255	0.655	0.385	0.385
	0.300	0.710	0.370	0.370
	0.191	0.780	0.419	0.419
	0.350	**0.722**	**0.380**	**0.415**

	re0			
method	Ent.	Pur.	Prec.	Rec.
n=100	0.691	0.370	0.135	0.219
n=50	0.667	0.381	0.161	0.224
n=20	0.580	0.419	0.230	0.243
n=10	**0.482**	0.468	**0.284**	**0.284**
n=5	0.359	0.569	0.353	0.353
normal	0.606	**0.475**	0.130	0.168

	re1			
	Ent.	Pur.	Prec.	Rec.
	0.667	0.333	0.148	0.295
	0.590	0.368	0.192	0.285
	0.451	0.450	0.289	0.306
	0.356	0.510	0.359	0.359
	0.274	0.576	0.399	0.399
	0.404	**0.617**	**0.412**	**0.380**

	k1a			
	Ent.	Pur.	Prec.	Rec.
	0.678	0.347	0.197	0.235
	0.606	0.397	0.239	0.252
	0.469	0.496	0.300	0.300
	0.355	**0.581**	**0.324**	**0.324**
	0.247	0.656	0.390	0.390
	0.464	0.536	0.304	0.284

	k1b			
method	Ent.	Pur.	Prec.	Rec.
n=100	0.595	0.555	0.275	0.275
n=50	0.563	0.586	0.336	0.336
n=20	0.420	0.683	0.433	0.433
n=10	**0.331**	0.745	**0.533**	**0.533**
n=5	0.167	0.846	0.566	0.566
normal	0.303	**0.856**	0.454	0.486

	wap			
	Ent.	Pur.	Prec.	Rec.
	0.672	0.354	0.232	0.322
	0.581	0.426	0.264	0.292
	0.446	0.503	0.332	0.344
	0.349	**0.565**	0.380	0.380
	0.253	0.640	0.399	0.399
	0.449	0.540	**0.403**	**0.344**

Precision が提案手法ではそれぞれ 0.388，0.355 であり，normal では 0.554，0.389 と若干 Precision が下がったが Entropy が大きく改善された．

5　考察

5.1　Entropy の改善

クラスタ C_i に分類された文書が，様々な正解集合に属する場合にそのクラスタの Entropy である $E(C_i)$ は増加する．そのため Entropy が高いクラスタは，クラスタが示す特徴の傾向が捉えにくい．従って，こうしたクラスタから代表的な文書を見つけることは，分析者の負担になると考えられる．実験結果が示す Entropy の改善は，提案手法が文書-単語行列を用いた文書クラスタリングに比べて，クラスタの特徴を表す代表的な文書を効率よく選別できることを示している．

また表 3 中の，分類精度が非常に低い結果 Entropy が高くなった ($Ent. > 0.8$) データである

hitech，reviews，la1 に対しても，高い Entropy を維持しながら分類が行われていることがわかる．これらのデータセットは，UPGMA などのボトムアップアプローチの階層型クラスタリングがその性質上，適切な分類が難しい特徴を持つと考えられるが，文書-単語行列ではなく単語-文書行列を用いることで，こうした問題を回避し，適切にクラスタリングできると考えられる．

Entropy の改善は，特にクラスタ数が多い ($|C| > 15$)4 種類のデータセット (fbis,re1, k1a, wap) で大きい．Patition アルゴリズムの場合，normal の 0.321 に対して提案手法では 0.218 であった．UPGMA アルゴリズムでは，クラスタリングが適切に行われなかったデータ (hitech, reviews, la1) を除いた 8 データの平均 Entropy が normal で 0.420，提案手法で 0.343 であった．その中で上記 4 種類のデータに関しては，normal の 0.436 に対して 0.328 であった．

5.2 課題

このように提案手法は文書-単語行列を用いた文書クラスタリングにおいて同等の Precision で，かつ低い Entropy でクラスタリングできる．しかしながら行列作成，クラスタリングアルゴリズム，文書ランキング，デンドログラムが示すクラスタ階層，の各処理の中で議論がまだ深く進んでない点もあり，今後の課題として議論する．

5.2.1 行列作成

提案手法は単語-文書行列を入力として，単語の共起性により文書をクラスタリングする手法であるが，本報告中の実験で用いた入力行列中のベクトルの値は，文書群に出現する単語の頻度であり，単語の共起に関する情報は入力行列に含まれていない．そこで共起単語の出現分布の類似性から文書をクラスタリングするために，文書中で共起する単語対をベクトルの要素として行列を作成することが考えられる．また，単語と共起単語対の両情報を行列の要素とすることも考えられる．この場合，行列が更にスパースになることが考えられるが，本実験で用いたデータセットでデータ密度が 1% 以下のデータ (k1a, k1b, la1, reviews) でも提案手法は良好なクラスタリング結果を示しているため，大きな問題とはならないと考えられる．

5.2.2 クラスタリングアルゴリズム

本実験で用いた UPGMA では，クラスタを結合させるための類似度の計算には各文書の単語のベクトル値のみが用いられている．行列作成の場合と同様，共起単語の情報付加を考えることができる．5.2.1 のように，入力行列に共起単語対の情報を付加してクラスタリングすることも考えられるが，クラスタを結合させるための計算において，ベクトルの要素を均等に計算するのではなく，共起する単語の要素に対して重みを与えるなど共起単語を考慮するよう，アルゴリズム自体の改良も考えられる．

5.2.3 文書ランキング

本実験で用いたクラスタ内文書ランキングには tf·idf に基づいたモデルを採用した．しかしながらこれは種々の手法を比較したわけではないため，他のモデルと比較する必要がある．比較する対象としては 3.5 節でも述べた，確率モデル [12] や概念間の関連度を用いたモデル [11]，Okapi[4, 3] の BM25 式などがある．また，クラスタリングアルゴリズムと同様に，共起単語の情報を考慮してランキングできるようなアルゴリズムの開発も検討する必要がある．

5.2.4 デンドログラムが示すクラスタ階層

文書-単語行列を用いた階層的文書クラスタリングを行った場合，クラスタは単語の出現分布の類似性を反映する．クラスタの結合に関してもクラスタ間の単語出現分布の類似性に依存する．そのため，結合する 2 つのクラスタで共通した単語が高い頻度を持つことが多く，こうした単語は tf·idf などのスコアリングにより特徴的な単語として抽出が可能である．これに対して単語-文書行列を用いた文書クラスタリングにおいては，単語が出現する文書分布の類似性からクラスタを構成するため，同様のスコアリングでは，結合したクラスタの特徴語は元の 2 つのクラスタの上位スコアの単語になりにくいことが考えられる．そこで，単語-文書行列に基づいた文書クラスタリングにおいては，単語の出現分布を捉えるだけではなく文書分布も考慮したスコアリングを用いて，デンドログラム上のクラスタから特徴語の抽出を行う必要がある．

6 まとめ

実際にクラスタリング結果を俯瞰的に分析する際，クラスタの代表的ないくつかの文書のみを優先的に分析する必要があるにも拘らず，クラスタリングアルゴリズムはこうした文書の重要性を考慮しない．本研究は，クラスタ毎の重要な数文書のみを効率よく提示することを目的として，単語-文書行列を入力として単語をクラスタリングすることで，それらの単語を含む文書を間接的にクラスタリングする手法を提案した．評価実験では，文書-単語行列を入力とした文書クラスタリングの精度と比較した．その結果，Precision はそのままで Entropy が

大幅に改善されたことから，提案手法はクラスタ内の代表的な文書の効率的な同定が可能であることを確認した．

本手法では共起単語の情報を明示的に用いていないが，こうした情報を提案手法の処理において利用することを検討した．また単語クラスタリングの結果を文書クラスタリングに適用するアプローチとして，共クラスタリングがある．提案手法と関連する技術として，調査する必要がある．今後の予定として，共起単語の情報を利用した処理および文書ランキング手法の比較を行う予定である．

参考文献

[1] Javier Artiles Picon, Anselmo Penas, and Felisa Verdejo. Word sense disambiguation based on term to term similarity in a context space. In *Proc. Senseval-3, ACL-SIGLEX*, pp. 58–63, 2004.

[2] Geoge Karypis. *CLUTO A Clustering Toolkit Release2.1.1*. University of Minnesota, Department of Computer Sciense, 2003.

[3] S.E. Robertson and S. Walker. Okapi/keenbow at trec-8. In *Proceedings of the Eighth Text REetrieval Conference (TREC-8)*, pp. 151–161, 2000.

[4] S.E. Robertson, S. Walker, and M. Beaulieu. Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive. In *Proceedings of the Seventh Text REetrieval Conference (TREC-7)*, pp. 253–264, 1999.

[5] A. Takano, Y. Niwa, S. Nishioka, M. Iwayama, T. Hisamitsu, O. Imaichi, and H. Sakurai. Information access based on associative calculation. *SOFSEM 2000, LNCS*, Vol. Vol.1963, pp. 187–201, 2000.

[6] Ying Zhao and Geoge Karypis. Hierachical clustering algorithms for document datasets. Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 2003.

[7] 宮本定明. クラスター分析入門. 森北出版, 1999.

[8] 小熊淳一, 内海彰. 語の共起情報を用いた文書クラスタリング. 人工知能学会全国大会資料, pp. 2E1–01, 2005.

[9] 庄田良助, 松田喬, 吉田哲也, 元田浩, 鷲尾隆. 構造的類似度に基づくグラフクラスタリング. 人工知能学会全国大会資料, pp. 3F1–02, 2003.

[10] 大熊和彦. 新しい大学研究「ソリューション研究」の意義と課題. 研究・技術計画学会 第 21 回年次学術大会予稿集, pp. 88–91, 2006.

[11] 藤井啓彰, 小島一秀, 渡部広一, 川岡司. 概念間の関連度に基づく情報ランク付けを用いた知的検索手法. 人工知能学会誌, 17 巻 6 号 D, pp. 684–689, 2002.

[12] 徳永健伸. 情報検索と言語処理. 東京大学出版会, 1999.