特徴間の類似性を考慮した特徴量集約手法の検討

大野　成義† 　太田 学†† 　片山　薫‡ 石川　博‡‡

†職業能力開発総合大学校情報システム工学科 〒229-1196 相模原市橋本台 4-1-1
††岡山大学大学院自然科学研究科 〒700-8530 岡山市津島中 3-1-1
‡首都大学東京大学院システムデザイン研究科 〒192-0397 八王子市南大沢 1-1
‡‡静岡大学情報学部情報科学科 〒432-8011 浜松市城北 3-5-1

E-mail: †ohno@uitec.ac.jp

あらまし 膨大な数のデータを扱う際にクラスタリングは有効な手段である．ユーザがクラスタリング結果から求める情報を得るためにはクラスタ間の関係を把握する必要があり，クラスタの特徴量を集約しなければならない．本稿では，集約する特徴量にはなんらかの類似性が存在すると考えて，この類似性を特徴間のなす角度とする．角度，つまり斜交の概念を取り入れて特徴量を集約する方法を提案する．また，その提案方法の有効性を検討するために特徴量を Web ページの語に限定し適合の正解がある NTCIR のデータを使い実験を行った．

キーワード Web とインターネット，クラスタリング，特徴量集約

1. はじめに

大量のデータを効率的に扱う手段として，クラスタリングは有用な手段である．特に，ユーザが漠然とした検索要求を持ち，様々な話題の中から目的の情報について調べようとしている時，データがクラスタリングによってあるカテゴリーに分類されていれば，欲しい情報を持つクラスタに容易に問い合わせることが可能となる．

ユーザがクラスタを適切に選択するためには，各クラスタに含まれるデータの特徴を表すラベルがあると便利である．このラベルは各クラスタに含まれるデータの特徴量を集約することで求めることができる．ただし，ラベルは分割されたクラスタのそれぞれの特徴を知ることには有効ではあるが，個々のクラスタにのみ注目しているのでクラスタ間の関係までは分からない．各クラスタの関係性を知ることは個々の話題についての類似性や差異性を明らかにすることにもつながり，非常に重要であると考えられる．

すでに Scatter/Gather[3]のような同じカテゴリーに属する文書について差異を比較する方式は研究されているが，カテゴリーの違いに関わらず差異の比較検討が可能である手法を検討する．

また，中心的な話題を表すための特徴量の集約手法として一般的であるのは tf-idf[1][2]重み付けに基づく集約手法であるが，その集約手法に斜交基底を用いたメタ検索におけるランクリストの統合方法[11]を利用する．これは，クラスタに含まれるデータの特徴量にはなんらかの類似性があると考えられるからである．そこで，各データの特徴量の相関を計算し，各データのなす角度として特徴量の集約を行う方法を提案する．

本稿の構成は次の通りである．第 2 章では関連研究について述べる．第 3 章では提案手法について，第 4 章では実験方法とその結果について述べ，最後に第 5 章でまとめと今後の課題について述べる．

2. 関連研究

関連研究として Web 検索結果のクラスタリング方法と差異を比較する Scatter/Gather を紹介する．

2.1. Web 検索結果のクラスタリング

Web 検索結果のクラスタリングは大きく二つに分類できる．コンテンツ・マイニングを利用したクラスタリングと，ストラクチャ・マイニングを利用したものである．前者はコンテンツを分析し，特徴語を抽出し，その特徴語に着目して似通ったページを一つのクラスタにまとめる[6][14]．従って，コンテンツに書かれている言語に依存する．また，特徴語を抽出していることから，コンテンツの内容，意味を分析していることになる．コンテンツ・マイニングによるクラスタリングは研究が進んでおり，商用の Vivisimo[7] や WSM[8] のようなシステムが存在する．後者はコンテンツ・マイニングだけでなくログの分析も利用してより精度を上げようとしているようである．しかし，コンテンツ・マイニングによるクラスタリングは，動画や画像データなどで構成されていてテキストの少ないページには適用が難しい．

一方，Web ページはリンク情報を持っていることから，このリンク情報を分析することで有用な情報を取りだすことをストラクチャ・マイニング[9][15]と呼んでいる．クラスタリングでもこのリンク情報を使う方法が考えられる．このような方法によるクラスタリン

グはあまり研究が進んでおらず，商用システムも知られていない．また，ストラクチャ・マイニングであれば，テキストの少ない動画や画像・音声データで構成されているページにも適用できる．

リンク情報を用いたクラスタリングとして，例えば，Y.Wang ら[10]の研究がある．コンテンツ・マイニングによるクラスタリングでは語や句に着目し，これらを共有するページを同一のクラスタに分類することから，語や句の代わりにリンクに着目し，同じリンクを持つページを同一のクラスタに分類する方法を提案している．

また，Max Flow アルゴリズムを用いた Web ページのクラスタリング方法[12]もリンク情報を用いたクラスタリング手法の一つである．この方法はリンク情報のみで，コンテンツは全く分析していないにもかかわらず，良好な結果を出している．しかし，コンテンツを分析しないためラベルを持たない，クラスタのサイズが小さいといった問題がある．そこで，このクラスタリング方法に提案手法を適用して特徴量集約を行えば，クラスタにラベルをつけ，小さいクラスタの統合を行うことも可能となる．

2.2. Scatter/Gather

Web 文書の検索結果などの多量な情報を扱う手法として，クラスタリングとユーザによるクラスタの選択の繰り返しによって欲しい情報に誘導する Scatter/Gather がある．これは，入力された情報の集合を 10 のクラスタに分類し，各クラスタの持つ文書に頻出する単語を話題単語として各クラスタに付加し，表示する．ユーザは興味のあるクラスタを選択すると，今度は選択されたクラスタに分類された情報のみを対象として再びクラスタを生成する．これを繰り返すことでユーザは漠然とした興味からでも得たい情報にたどり着くことができるという手法である．

クラスタの再構成という意味では提案手法の考え方と共通している部分がある．しかし，Scatter/Gather では選択されたクラスタのサブクラスタを扱うのに対し，提案手法では特徴空間を変化させることで異なるクラスタの情報をも含んだ新たなクラスタの生成を行う点が異なっている．

3. 提案手法

提案する特徴量集約手法は小さいクラスタがすでに存在することが前提となる．ステップ１として，クラスタに含まれる各文書の特徴量を抽出し，文書単位でその特徴量（語）の順位を決める．ステップ２として，その順位づけした特徴量（語）のリストを使って文書間の相関を計算し，余弦をとって文書間の類似性を表す角度とする．ステップ３として，ステップ２で求めた角度から斜交軸を構成し，クラスタを表す特徴量を計算する．

3.1. 特徴量の抽出

「茶筌[4]」を利用して各文書の形態素解析を行い，得られた品詞情報を利用して各文書に含まれる名詞・未知語を特徴語候補として抽出する．さらに数詞・代名詞などの不要語の除去を行い，名詞・未知語が連続して出現する時には語の連結を行いひとつの特徴語として扱う．この M 個の特徴語の集合を $F:\{f_1,f_2,\ldots,f_M\}$ とする．

文書 d_j の特徴量を表現するため，特徴語 f_i の文書 d_j における重要度 $w_{i,j}$ を tf-idf による重み付けで表す．ここで，特徴語の頻度 $tf_{i,j}$ は文書 d_j における f_i の出現頻度，文書頻度 df_i は f_i が出現する文書数である．文書の総数を N とすると，逆文書頻度 idf_i は df_i の逆数であり，log $\frac{N}{df_i}$ で表される．このとき $w_{i,j}$ は，

$$w_{i,j} = tf_{i,j} \cdot idf_i$$

で表される．

先に求めた $w_{i,j}$ から文書 d_j に含まれる特徴語 f_i の各文書における重要度の順位を決め，$R_j(f_i)$ で表す．さらに，文書 d_j における特徴語 f_i の種類数 num_j を使って

$$r_j(f_i) = \left(num_j + 1 - R_j(f_i)\right) / num_j$$

と規格化する．

例えば，あるクラスタには 7 つの文書が含まれており，7 つのうちの一つの文書には表 1 のような 4 つの特徴語（「物理学」，「天文学」，「ガリレイ」，「宇宙」）が見つかったとする．文書中の出現頻度 tf と文書頻度 df からその文書における各特徴語の重要度 w，順位 R，規格化された順位 r は表 1 のとおりになる．クラスタに含まれる文書数は 7 としたが，クラスタが複数あり，文書の総数は 200 と仮定して計算した．文書頻度もあるクラスタに限定せず，200 の全文書中に特徴語を含む文書がいくつあるかで評価する．クラスタは既に類似の文書が集められているという前提にあり，クラスタの特徴を表す語は各文書に共通している可能性が高い．そこで，重要度 w を計算するとき，文書の総数（N）や文書頻度をクラスタに限定せず，全文書を対象とする．

表 1 特徴量抽出の例

特徴語	tf	df	w（重要度）	R（順位）	r
物理学	3	5	15.97	1	1.0
天文学	2	16	7.29	2	0.75
ガリレイ	1	7	4.84	3	0.5
宇宙	1	21	3.25	4	0.25

ここまでの説明でクラスタリングされるのは文書(Web ページ)であると仮定しているが，これ以降は対象をより一般的なデータ，画像や音声，動画としても適用可能である．クラスタリングされるデータの特徴を抽出し，各データにおける特徴の順位だけを特徴量としているからである．

例えば，あるクラスタには 3 つのデータ(Web ページ)A, B, C が含まれていたとする．それぞれのデータに対して特徴を抽出し順位づけを行う（表 2）．この特徴量からデータ A, B, C を含むクラスタの特徴量を決める．メタ検索エンジンで用いられているように，単純に順位を点数化し平均を求めると，各特徴は(a, b, c, f, e, d, g)=(0.93, 0.67, 0.47, 0.33, 0.27, 0.2, 0.13)となる．この計算方法は，全ての特徴，データ（Web ページ）が完全独立で直交であるという仮定が基になっている．しかし，データ A とデータ B は比較的似ており，データ C は少し異質である．これらを同じように扱うのは不自然である．そこで各データ(Web ページ)の相関を計算し，各データのなす角度とする．完全独立な直交系でなく斜交の概念を取り入れて特徴量の集約を行う．

表 2 各データ(Web ページ)の特徴量

R（順位）	r	A	B	C
1	1.0	特徴 a	特徴 a	特徴 f
2	0.8	特徴 b	特徴 c	特徴 a
3	0.6	特徴 c	特徴 b	特徴 b
4	0.4	特徴 d	特徴 e	特徴 g
5	0.2	特徴 e	特徴 d	特徴 e

3.2. 相関係数による斜交角度の決定

相関を調べる方法として一般的なピアソンの相関係数の他にケンドールの順位相関係数がある．前節の文書 i における特徴 a の規格化された順位 $r_i(a)$ は整数ではないが，元々は整数の順位 $R_i(a)$ であり，これを使って相関を求めるために，本提案手法ではピアソンでなくケンドールの順位相関係数を使用する．同じクラスタに含まれる文書 i と j の特徴の相関を表すケンドールの相関係数 r_k は，前節で求めた $r_i(a)$ をもちいて以下の式で求められる．

$$r_k = \frac{\sum P_{ab} - \sum Q_{ab}}{m(m-1)}$$

ここで

$r_i(a) > r_i(b)$ かつ $r_j(a) > r_j(b)$ なら $P_{ab} = 1$

$r_i(a) > r_i(b)$ かつ $r_j(a) > r_j(b)$ なら $Q_{ab} = 1$

$r_i(a) > r_i(b)$ かつ $r_j(a) > r_j(b)$ なら $Q_{ab} = 1$

であり，$\sum$ は文書 i と j に含まれる特徴量 $a, b(\neq a)$ について和をとる．m は文書 i と j に含まれる特徴量の総数であり，文書 i に含まれる特徴の種類数 num_i や文書 j についての num_j に等しいとは限らない．これは，文書 i に含まれる特徴で文書 j には含まれない特徴もありえるためである．

文書 i と j に含まれる特徴とその順位が全く同じ場合，相関係数 r_k は 1 となる．逆に同じ特徴を含むが順位が全く逆の場合は-1 となる．文書 i と j に含まれる特徴に共通するものがない場合，つまり，文書 i に含まれる特徴は文書 j に全く含まれず，文書 j の特徴が文書 i に全く含まれない場合，相関係数 r_k は 0 となる．ただし，文書 i に特徴 a が含まれない場合は $r_i(a) = 0$ とする．各文書を多次元空間上の直線とし，相関係数をその直線間のなす角の余弦だとすれば，相関係数 r_k が 0 ということは直交を意味する．つまり，共通する特徴を持たない文書は類似性がなく，多次元空間で直交する．なお，この多次元空間の次元数は特徴の数ではなく文書の数になる．

特徴量として各文書における特徴（語）の順位を使用することから，提案手法ではケンドールの順位相関係数を使用したが，一般的なピアソンの相関係数の定義をそのまま使うと意味の面でも不都合が生じる．上記のように，文書 i に含まれる特徴は文書 j に全く含まれず，文書 j の特徴が文書 i に全く含まれないような場合，ピアソンの相関係数は 0 にならないからである．

3.3. 特徴量集約によるクラスタの特徴量

相関係数が求まれば，集約する文書の特徴量を表す軸上の基底ベクトルを計算する．これは，「斜交基底を用いたメタ検索におけるランクリストの統合方法」のランクリストの代わりに 3.1 節で計算した各文書の特徴量を使えばよい．

各文書の特徴量は多次元空間上の各特徴語の射影

であると考え，各文書の特徴量から各特徴語の位置を求める．各軸上の順位の値からこの軸に垂直な超平面を考え,全ての超平面の交わる点が求める位置となる．求める多次元空間上の特徴語 a の位置を表すベクトルを $\vec{v}_a$ とすると，全ての i について $(\vec{v}_a - r_i(a)\vec{w}_i)\cdot\vec{w}_i = 0$ を満たす $\vec{v}_a$ を求めればよい．

ここで、・はベクトルの内積を表し，$r_i(a)$ は規格化されたランキングの値，$\vec{w}_i$ は検索エンジンの検索結果を表す軸の単位ベクトルである．

各特徴語は空間上に分布しており，一つの直線上に並んでいるとは限らない．集約してクラスタの特徴量を得るために空間上に分布している点の集合を一つの直線で近似する必要がある．多次元空間上の各特徴語の重心を求め，原点とその重心を通る直線をその近似直線とする．従って，この近似した直線を表す単位ベクトルは $\vec{V} = \frac{\sum_a \vec{v}_a}{|\sum_a \vec{v}_a|}$ で表現できる．ここで $\sum_a$ は集約されるクラスタに含まれる文書の持つ特徴語の全てに関して和をとることを意味し，これらの特徴語は全ての文書に含まれている必要はない．

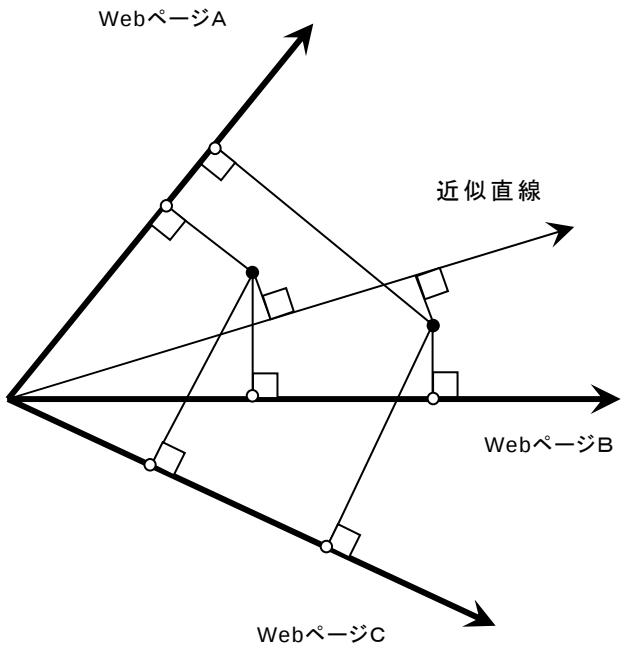

図１ Web ページＡ，Ｂ，Ｃとその特徴量を示す点

各ページからこの近似直線に下ろした垂線の足が最も本当の特徴量に近い特徴量であると期待できる．従って，$|(\vec{V}\cdot\vec{v}_a)|$ が特徴語 a の集約した特徴量としての値 $r(a)$ となる．

4. 評価実験

4.1. 実験方法

提案方法を定量的に評価するために，NTCIR-4 の Web タスクDのデータを使って実験を行った．Formal Run で使われたのは11のtopicについての検索結果である.各 topic についての検索結果はそれぞれ 200 位までWeb ページがリストされている．これをクラスタリングする．

クラスタリングの評価方法も NTCIR-4 の Web タスクDと同じ方法を採用する．それは，利用者が明確な検索要求を抱いてブランジングするような場面を想定して検討された評価方法である．ここでは，分類結果における適合ページの分布を分析することとし，適合ページの多いクラスタに含まれるページ群のランキングに関する精度と再現率を算出する．具体的には適合ページを多く含むクラスタを順番にソートする．上位20個のページを取りだし，平均適合率，20位までの適合率(表では適合率と表記),20位までの再現率(表では再現率と表記)を計算する．

実験では，Max Flow アルゴリズムを用いた Web ページのクラスタリング方法でクラスタリングした結果に対して提案手法を用いて特徴量集約を行った．集約を行う前では，クラスタにラベル名をつけることができなかったが，集約後はもっとも高いランキングの特徴語をそのクラスタを表す特徴語としてラベリングする．さらに，集約したクラスタの特徴量を用いてクラスタ間の差異を定量的に評価することが可能となる．この差異を評価して小さいクラスタをまとめて大きなクラスタを構成することができる．

4.2. 実験結果

検索結果は高適合(S)，適合(A)，部分的適合(B)，不適合(C)の4段階で評価される．表3の rigid は高適合と適合のみを適合ページと判断したことを表しており，表4の relaxed は不適合以外を全て適合ページと判断したことを表している．

各表には，比較のためコンテンツ・マイニングによって非排他的にクラスタリングしたMETAL[13]の結果も並べて表示する．NTCIR-4 の Web タスク D に参加したチームは全てコンテンツ・マイニングによるクラスタリングを行っていた．そのなかで METAL は最も良い結果を記録しているため，比較することにした．また，表の「集約前」は Max Flow アルゴリズムを用いて行ったストラクチャ・マイニングのみによるクラスタリングの結果である.表5で確認できるようにコンテ

ンツ･マイニングによるMETALに比べてクラスタのサイズが小さい．「特徴量集約1」は提案手法によって特徴量集約を行い，特徴量の差異の小さいクラスタを結合してできた結果を評価している．「特徴量集約2」は，集約前の小さいクラスタに含まれる文書をひとつのデータとみなしてMETALと同じ手法で再クラスタリングした結果を評価した．

表 3　rigid 判定による結果(%)

	平均適合率	適合率	再現率
METAL	36.0	44.9	75.4
集約前	21.4	44.9	50.3
特徴量集約 1	17.6	28.7	45.5
特徴量集約 2	32.0	42.5	69.2

表 4　relaxed 判定による結果(%)

	平均適合率	適合率	再現率
METAL	30.0	48.0	53.2
集約前	33.5	52.4	60.5
特徴量集約 1	28.7	46.2	50.1
特徴量集約 2	29.8	47.5	52.1

表 5　クラスタサイズの比較

	クラスタ数	最大サイズ
METAL	48.5	51.4
集約前	27.1	9.5
特徴量集約 1	33.6	23.4
特徴量集約 2	45.4	55.5

<NUM>0003</NUM>
<TITLE>コペルニクス，地動説，キリスト教
</TITLE>
<DESC>コペルニクスの地動説がキリスト教社会でどのように受容されていたかを調べたい。
</DESC>
<BACK>史料を忠実に追うとコペルニクスの地動説（太陽中心説）は 16 世紀にすでにキリスト教に認められていた、という科学史家もいるようだ。この問題について事実関係を知りたい。
</BACK>
<RELE>文書内にて地動説に関するコペルニクス本人の著作や発言などとキリスト教の関係に言及されていれば適合文書とする。コペルニクス本人とは関係なく、地動説とキリスト教の関係を述べるに留まった文書、またはコペルニクス本人とキリスト教の関係を述べたのみで地動説に触れていない文書は部分的適合とする。コペルニクスの地動説の説明のみでキリスト教との関連が言及されていないものは不適合とする。</RELE>

図 2　クエリ 3 の検索課題

表 5 から提案手法による特徴量集約によってクラスタが大きくなっていることがわかる．これは，集約したクラスタの特徴量を用いてクラスタ間の差異を定量的に評価することが可能になったことの一つの証拠である．クラスタ数も増加しているが，これはクラスタに分類されなかったページも提案手法によってクラスタリングされて，新しいクラスタができていることを意味している．しかし，適合率や再現率に関して，表 3 と表 4 から提案手法による特徴量集約は有効とはいえない．

全ての topic（検索課題）で有効でなかったわけではない．例えば，クエリ 3（図 2）のクラスタリング結果を表 6 と表 7 に示す．

表 6 は METAL によるクラスタリング結果であり表 7 は提案手法を用いて特徴量集約を行い小さいクラスタを結合させた結果できたクラスタの一部である．このようにクラスタの再クラスタリングに成功する場合もある．しかし，ほとんどの topic で再クラスタリングに失敗している．

表 3 や表 4 を見る限り「特徴量集約 2」の方法の方が提案手法より良い結果になっている．しかし，これは Max Flow アルゴリズムを用いて行うストラクチャ・マイニングによるクラスタリングではクラスタに分類できなかったページがクラスタリングされただけである．全ての項目において METAL を超えない．クラスタの結合ではノイズを集めるだけであり，表 7 のような特定の topic に限っても METAL より良い結果を得ることはできなかった．Max Flow アルゴリズムを用いて行うクラスタリングと比べてもクラスタのサイズが大きくはなるが，必ずしも良い結果になるとは限らない．

表 6 クラスタリング結果の例(METAL)

ページ ID	順位	ラベル名	適合判定
NW011452773	1	宇宙	A
NW008449606	2	宇宙	B
NW011452772	3	宇宙	B
NW011452783	4	宇宙	B
NW001673028	5	宇宙	C
NW013290383	6	宇宙	C
NW006585020	7	宇宙	C
NW011452812	8	宇宙	B
NW011452811	9	宇宙	B
NW012004625	10	宇宙	C
NW012004595	11	宇宙	C
NW008449608	12	宇宙	A
NW013290356	13	宇宙	C
NW003387031	14	宇宙	C
NW002658037	1	天文学	A
NW009614525	2	天文学	C
NW001443580	3	天文学	C
NW002280945	4	天文学	S
NW011649170	5	天文学	S
NW009382688	6	天文学	A

表 7 特徴量集約して結合したクラスタリングの例

ページ ID	順位	特徴語	適合判定
NW011452773	1	宇宙	A
NW008449520	2	宇宙	B
NW011452772	3	宇宙	B
NW008449606	4	宇宙	B
NW003295673	5	宇宙	C
NW011452783	6	宇宙	B
NW008449505	7	宇宙	A
NW008450014	8	宇宙	C
NW008449782	9	宇宙	C
NW011452812	10	宇宙	B
NW011452811	11	宇宙	B
NW003295675	12	宇宙	C
NW008449480	13	宇宙	C
NW008449608	14	宇宙	A
NW008449609	15	宇宙	A
NW002951265	1	天文学	C
NW007125920	2	天文学	A
NW002935649	3	天文学	A
NW007125730	4	天文学	A
NW007125770	5	天文学	B

4.3. 考察

提案手法による特徴量集約は図 2 の検索課題については表 7 のように成功する例もあるが，表 3 と表 4 からはほとんどの検索課題について失敗したといえる．小さいクラスタを結合させてクラスタのサイズを大きくすることはできても，適合文書を多く含むクラスタは適合文書を多く含むクラスタとのみ，不適合文書ばかりのクラスタは不適合文書ばかりのクラスタとのみ，別々に結合させることはできていないためである．

原因として 2 つのことがあげられる．一つは，集約するのは文書の特徴を代表するような一つもしくは少数の特徴語ではなく，文書に含まれる特徴語全てとしていることにある．文書における特徴語の重要度に順位をつけており全ての特徴語を同じ重みで扱ってはいないとはいえ，順位が下位の特徴語も意味集約時に影響を与える．実験で比較対象にした METAL や「特徴量集約 2」は文書の特徴を代表するような一つもしくは少数の特徴語を代表特徴語とし，これだけでどのクラスタに所属するか決定している．従って，重要度の低い特徴語まで特徴量集約に使用したことが，悪影響を及ぼし不適合文書，ノイズも集めてしまい，これが再クラスタリング失敗の原因の一つになっている．

再クラスタリング失敗の原因のもう一つは，集約した特徴量を決定する方法にある．提案方法は，クラスタに含まれる文書の特徴量，つまり特徴語の順位を比較する．特徴量の類似性を考慮して集約を行い，単純な多数決は行っていない．類似な特徴を単純に加算するのではなく，類似でない，差異のある特徴を強調して集約する．クラスタに適合文書だけでなく不適合文書も含まれていると，不適合文書の特徴が集約時に強調されてクラスタの特徴に影響を与え，再クラスタリング失敗の原因の一つになっている．

提案方法は，適合文書を集める本実験のように，類似のものを集めるには有効でないかもしれないが，ニュース記事の違いを見つけるなど，文書間の差異を見つけるには効果を発揮する意味集約方法であると期待できる．

5. まとめと今後の課題

クラスタの特徴量を集約する方法を提案した．集約する特徴量にはなんらかの類似性が存在すると考えて，この類似性を特徴間のなす角度とし，角度，つまり斜交の概念を取り入れて特徴量を集約する方法を提案した．提案方法を定量的に評価するために NTCIR-4 の Web タスク D のデータを使って実験を行った．Max Flow アルゴリズムを用いて行ったクラスタリング結果に提案手法を適用して特徴量集約を行い，クラスタサイズを大きくすることに成功した．また，ストラクチャ・マイニングではつけられなかったクラスタの特徴を表すラベルも得られた．更に，特定の topic では適合率,再現率ともに良くなる場合もあった.しかし，平均的にはノイズを拾ってきてしまい適合率，再現率ともに低下してしまった．

ノイズを拾ってこないクラスタの結合方法を検討する必要がある．また，定量的に評価するために NTCIR-4 の Web タスク D のデータを使ったが，他の実験で評価することも検討に値する．提案した特徴量集約手法は検索結果のクラスタリングだけでなく，ニュース記事の集約など，いろいろな分野での利用が考えられる．更に，特徴量は順位で表せるものであるため，特徴語，つまり文書だけでなく動画や画像・音声データに対する適用も検討していく．

文　献

[1] Salton, G. and Buckley, C.: “Term-weighting approaches in automatic text retrieval,” Information Processing and Management, 24, pp.513-523, 1988d.

[2] Salton, G. and Buckley, C.: “Improving retrieval performance by relevance feedback,” Journal of the American Society for Information Science, 41(4), pp.288-297, 1990.

[3] Marti A. Hearst, Jan O. Pedersen, “Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results”, in Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval (SIGIR'96), pp. 76-84, 1996.

[4] 茶筌 ：http://chasen.naist.jp/hiki/ChaSen/ Accessed 2005.

[5] 渡辺匡，"文書間の差異に着目したクラスタリン

グ手法”, 第 67 回情報処理学会全国大会,2005.

[6] Zamir, O, et al, "Grouper: A Dynamic Clustering Interface to Web Search Results", Proc. WWW8, 1999.

[7] Vivisimo http://vivisimo.com/

[8] WSM http://wsm.directtaps.net/

[9] J.Kleinberg, Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, Vol.46, No.5, pp.604-632, September 1999.

[10] Y.Wang and M.Kitsuregawa, Use Link-based Clustering to Improve Search Results, Proceedings of the 2nd International Conference on Web Information System Engineering, IEEE Computer Society, Dec. 2001.

[11] 大野成義, 太田学, 片山薫, 石川博, “斜交基底を用いたメタ検索におけるランクリストの統合方法の提案,” 電子情報通信学会論文誌 D-I, Vol. J88-D1, No. 3, pp.657-667, March 2005.

[12] 大野成義, 渡辺匡, 片山薫, 石川博, 太田学, “Max Flow アルゴリズムを用いた Web ページのクラスタリング方法とその評価,” 情報処理学会論文誌（データベース）, Vol.47 No.SIG4 (TOD29) pp.65-75, March 2006.

[13] M. Ohta, H. Narita and S. Ohno, “Overlapping Clustering Method Using Local and Global Importance of Feature Terms at NTCIR-4 Web Task,” Working Notes of the 4th NTCIR Meeting, Supplement volume 1, pp.102-110, June 2004.

[14] Zeng, H.J. et al. Learning to Cluster Web Search Results, ACM SIGIR04.

[15] Kumar, R. et al. Trawling the Web for Emerging Cyber-communities, Proceeding of the 8th international conference on World Wide Web, pp.1481—1493, 1999.