

クラスター内変動最小アルゴリズムに基づくトピックセグメンテーション

別所 克人

日本電信電話株式会社 NTTサイバースペース研究所

bessho.katsuji@lab.ntt.co.jp

テキストをトピック単位に分割するトピックセグメンテーションは、コンテンツを構造化するメタ情報自動生成技術にとって重要な要素技術となっている。本稿では、テキストを、特異値分解によって得られた単語の意味表現である概念ベクトルの系列ととらえ、ベクトルの系列を分割するクラスター列で、クラスター内変動が最小となるものをトピック区間列とする手法を提案する。新聞記事を用いた評価実験の結果、提案手法は、局所的な範囲内でトピック境界を判断する手法よりも高精度をもつことを検証した。

Topic Segmentation Based on Minimum Within-Cluster Fluctuation Algorithm

Katsuji BESSHO

NTT Cyber-Space Laboratories, NTT Corporation

bessho.katsuji@lab.ntt.co.jp

Topic segmentation dividing a given large text into a consecutive topic sequence is one of key technologies to generate meta data of the text. We propose a novel segmentation algorithm based on the minimum within-cluster fluctuation criteria. Generating a conceptual vector sequence from the given text by using singular value decomposition, the algorithm derives the most likely topic sequence based on the criteria from the whole conceptual vector sequence, where the topic sequence is equivalent to the cluster sequence and a cluster is a partial conceptual vector sequence. The segmentation experiments using several articles on the newspaper show that the proposed method achieved far better performance than a conventional method which determines topic boundaries by using the local conceptual vector sequence.

1. はじめに

日々膨大に生成されていく映像・音声・テキスト情報からなるコンテンツに高速に的確にアクセスできるようにするためには、あらかじめ検索に適した形にコンテンツを編集・構造化し、検索のためのメタデータを付与しておく必要がある。メタデータ作成には多大のコストを要するので、自動的に生成できることが望まれる。

コンテンツは一般に複数のトピック区間から構成されている場合が多い。ユーザが多数あるコンテンツ群の中から所望する情報を含んだトピック区間のコンテンツを取得するために、例えば要望を表す検索キーワードをシステムに入力し、当該キーワードのコンテンツ中の出現位置を特定するというやり方が考えられる。しかしながら、コンテンツが映像や音声の場合、そのキーワードを含むトピック区間の開始・終了位置が分からなければ、トピック区間のコンテンツの的確な再生ができない。コンテンツがテキストであっても、該当するトピック区間をユーザに目で見て判断させることになり、該当するトピック区間を瞬時に取得することは困難である。そこで、メタデータ自動生成にあたり、コンテンツをトピック単位に分割するトピックセグメンテーションが重要な要素技術となってくる。

音声やテロップを認識させた結果である文字情報も含め、テキスト情報は検索にとって非常に有用な情報である。本稿では、トピックセグメンテーションの対象としてテキスト情報を考え、さらに研究の初段階として新聞記事等の書き言葉を対象とした。

2. 従来手法の問題点

トピックセグメンテーションの手法の一つとして、テキスト中の単語の結束性に着目するものがある。従来手法の一つである Hearst による方法（本稿では Hearst 法と呼ぶことにする）[1][2]は、テキスト中の単語列における各単語境界の前後に一定の単語数の窓を設定し、各窓ごとに、単語の窓内の出現頻度ベクトルをとり、そのベクトル間の余弦測度を当該単語境界における結束度とする。次に、結束度の微弱な振動を除去するため、単語

境界の結束度を、当該単語境界とその前後一定数の単語境界の結束度の平均に変換する（結束度の平滑化）。トピック境界では、この結束度が極小となっていると期待されるので、結束度が極小となる単語境界（極小点と呼ぶ）で、結束度の谷の深さ（depth score と呼ぶ）の大きいものからトピック境界として出力する。

また、筆者はこれまで単語の意味表現の一つである概念ベクトルを用いたトピックセグメンテーションの研究を行ってきた[3]。[3]において提案した概念ベクトル結束度法は、Hearst 法における窓に対応するベクトルとして、窓に含まれる単語の概念ベクトルの重心をとるものである。

これらの手法では、一定のサイズの窓をスライドさせていくので、窓幅よりも短いトピック区間が窓に含まれているとき、対応するベクトルは当該トピック区間の意味を適切に表していず、結果として算出される結束度も適切なものとならない。このため、窓幅よりも短いようなトピック区間の検出が困難である。また、テキストの局所的な範囲においてトピック境界を判断しているため、非常に長いトピック区間は細分され、細かいトピック区間が検出され、非常に長いトピック区間は検出が困難である傾向がある。

これらの課題を解決するため、本稿では、概念ベクトルを用いたトピックセグメンテーションの手法として、ベクトルの系列を分割するクラスター列で、クラスター内変動が最小となるものをトピック区間列とする手法を提案する。任意の区間を考慮し、広域的、局所的な範囲双方でトピック区間としての妥当性を判断するので、様々な長さのトピック区間の検出に有効であると考えられる。

以下、3 節において、概念ベクトルを格納する概念ベースの生成及びそれを用いたクラスター内変動最小アルゴリズムの説明を行う。4 節で評価実験結果を述べ、5 節でまとめを述べる。

3. 提案手法

3.1. 概念ベース生成

本手法では、あらかじめ学習用コーパスを基に、学習用コーパス中の各単語にその共起パターンをベクトル化して得られる意味表現（概念ベクトル

表1 共起行列の例

	...	国会	...	園芸	...
...	...	...	...	...	...
選挙	...	287	...	3	...
...	...	...	...	...	...
苗木	...	2	...	93	...
...	...	...	...	...	...

と呼ぶ)を対応付け、単語とその概念ベクトルの対の集合である概念ベースを生成しておく。ある2単語に対応するベクトル値が近ければ、共起パターンが似ているので、この2単語は意味的に近いということが推測される。

概念ベースの生成では、まず学習用コーパスを形態素解析した後、各自立語間の1文中に共起する頻度をカウントした共起行列を作成する(表1参照)。共起行列の各行をベクトルと見立てると、各自立語にその共起パターンを表すベクトルが対応付けられる。但し、このままではデータのスパースネスが生じることを始めとして、テキストデータから抽出される単語の情報には常に欠落があると予想されるため、ベクトル間の類似度の精度は低いと考えられる。また、一般にベクトルの次元数は非常に大きなものとなるため、計算量も無視できないものとなる。このため、[4][5]の手法に従い、共起行列を特異値分解により、次元数を縮退させた行列に変換する。

$n \times p$ の行列 X を共起行列としたとき、特異値分解により共起行列 X は、以下のように分解できる。

$$\begin{matrix} X &= & U & S & V^t \\ n \times p & & n \times a & a \times a & a \times p \end{matrix} \quad (1)$$

ここで、 $a = \text{rank } X \leq \min(n, p)$ 、 $U^t U = V^t V = I$ (I : 単位行列)であり、 $S = (c_{ij})$ としたとき、 $c_{ii} \geq c_{jj} > 0$ ($1 \leq i \leq j \leq a$)である。 c_{ii} ($1 \leq i \leq a$)を X の特異値と呼ぶ。 $1 \leq b \leq a$ に対し、

$$\begin{matrix} X' &= & U' & S' & V'^t \\ n \times p & & n \times b & b \times b & b \times p \end{matrix} \quad (2)$$

をとるとき、 X の第 i 行目の行ベクトルは、 p 次元空間中のある b 次元部分空間にある、 $U'S'$ の第 i 行目の行ベクトルに射影変換される。 $U'S'$ の各行ベクトルは、 U' の対応する行ベクトルを、各座標ごとに対応する特異値の割合で伸縮したもので、本研究では、変換後のベクトルを $U'S'$ ではなく、 U' の行ベクトルをその長さで割って単位ベク

トルに正規化したものとする。変換後のベクトルを概念ベクトルと呼び、単語とその概念ベクトルの対の集合を概念ベースと呼ぶ。

3.2. クラスタ内変動最小アルゴリズム

セグメント対象テキスト中の各単語に、概念ベース中のベクトルを対応付けて得られるベクトル列の変化は、単語の意味の変遷を表していると考えられるので、このベクトル列の変化を利用してテキストの分割が行えることが期待できる。

本手法では、セグメント対象テキストを形態素解析して単語に分割し、得られた各単語のうち、自立語のみに概念ベース中の対応するベクトルを付与する。

本研究では、トピック間の境界は文間の境界と仮定する。単語間の境界としてもアルゴリズム上、本質的な違いはない。

テキストを分割するトピック区間列(各区間は文集合)において、同一区間に含まれる概念ベクトルは、比較的類似性が高い(幾何学的距離が近い)と予想される。異なるトピックを表す区間のそれぞれで、内部に含まれる概念ベクトルの集合をとったとき、2つの集合はクラスター群としてよく分離されていると推測される。このことから、テキストを分割する区間列で、クラスター群として妥当である(よく分離されている)ものがトピック区間列である可能性が高いと考えられる。以後、任意の区間を考えたとき、区間そのものをクラスターと呼ぶことにする。

クラスター群としての妥当性の定義について述べる。これは、文献[6]に基づく。テキスト中の概念ベクトルの系列を

$$w_1, w_2, \dots, w_x \quad (3)$$

とし、テキスト中の文の系列を

$$s_1, s_2, \dots, s_g \quad (4)$$

とする。テキストを分割する任意のクラスター列は、文番号 $1 \leq h \leq g$ の分割

$$\begin{aligned} T_1 &= (1, 2, \dots, n_2 - 1), T_2 = (n_2, n_2 + 1, \dots, n_3 - 1), \dots, \\ T_i &= (n_i, n_i + 1, \dots, n_{i+1} - 1), \dots, T_j = (n_j, n_j + 1, \dots, g) \end{aligned} \quad (5)$$

という形になる。各文 s_h は、概念ベクトルの系列 $w_{m_h}, w_{m_h+1}, \dots, w_{m_{h+1}-1}$ から成り立っているものと

する。

概念ベクトルの系列(3)の全変動 A を、

$$A = \sum_{k=1}^x \|w_k - \bar{w}\|^2 \quad (6)$$

$$\bar{w} = \frac{1}{x} \sum_{k=1}^x w_k \quad (7)$$

と定義する。 $\bar{w}$ は全概念ベクトルの重心であり、 A は、該重心ベクトルと各概念ベクトルとの間のユークリッド距離の自乗の和である。

クラスター列(5)のクラスター間変動 $B(T_1, T_2, \dots, T_j)$ を、

$$B(T_1, T_2, \dots, T_j) = \sum_{i=1}^j (m_{n_{i+1}} - m_{n_i}) \|\bar{w}_i - \bar{w}\|^2 \quad (8)$$

$$\bar{w}_i = \frac{1}{m_{n_{i+1}} - m_{n_i}} \sum_{k=m_{n_i}}^{m_{n_{i+1}}-1} w_k \quad (9)$$

と定義する。

全変動の定義より、クラスター $T_i = (n_i, n_i + 1, \dots, n_{i+1} - 1)$ の全変動 $C(n_i, n_{i+1} - 1)$ は

$$C(n_i, n_{i+1} - 1) = \sum_{k=m_{n_i}}^{m_{n_{i+1}}-1} \|w_k - \bar{w}_i\|^2 \quad (10)$$

と定義される。クラスター列(5)のクラスター内変動 $e(T_1, T_2, \dots, T_j)$ を

$$e(T_1, T_2, \dots, T_j) = \sum_{i=1}^j C(n_i, n_{i+1} - 1) \quad (11)$$

と定義する。

計算により、任意のクラスター列(5)に対し、関係式

$$A = B(T_1, T_2, \dots, T_j) + e(T_1, T_2, \dots, T_j) \quad (12)$$

が成立することが証明できる。

全変動 A は一定であるので、クラスター内変動 $e(T_1, T_2, \dots, T_j)$ が小さいほど、クラスター間変動 $B(T_1, T_2, \dots, T_j)$ は大きくなり、各クラスター間にはよく分離されているといえる。そこでクラスター内変動の小ささが、クラスター群としての妥当性を示す指標となる。

関係式(12)から分かることは、あるクラスター列をさらに細分割して得られるクラスター列のクラスター内変動は、必ず分割前のクラスター列のクラスター内変動以下となるということである。クラスター内変動は、クラスター数一つの場合、

最も大きく、各クラスターが一文の場合、最も小さくなる。従って、クラスター群として妥当かどうかは、クラスター数を固定した場合のクラスター列の集合内で意味がある。

以下、クラスター T_i の全変動 $C(n_i, n_{i+1} - 1)$ を、クラスター T_i のコストと呼び、クラスター列(5)のクラスター内変動 $e(T_1, T_2, \dots, T_j)$ を、クラスター列(5)のコストと呼ぶことにする。

任意のクラスター数 j に対し、テキストを j 個に分割するクラスター列のコストの最小値 E_j 及び最小コストをとるクラスター列を求める。

文の系列 $s_1, s_2, \dots, s_h$ ($1 \leq h \leq g$) を q 個に分割するクラスター列で最小のコストをとる分割を $P(h, q)$ と表すことにする。

$$\begin{aligned} P(h, q) : T_1 &= (1, 2, \dots, n_2 - 1), \dots, \\ T_{q-1} &= (n_{q-1}, n_{q-1} + 1, \dots, n_q - 1), \\ T_q &= (n_q, n_q + 1, \dots, h) \end{aligned} \quad (13)$$

としたとき、クラスター列のコストは各クラスターのコストの和なので

$$\begin{aligned} P(n_q - 1, q - 1) : T_1 &= (1, 2, \dots, n_2 - 1), \dots, \\ T_{q-1} &= (n_{q-1}, n_{q-1} + 1, \dots, n_q - 1) \end{aligned} \quad (14)$$

となる。この性質を用いて、任意のクラスター数 j に対し、 $P(g, j)$ 及び $e(P(g, j)) (= E_j)$ を、以下のダイナミック・プログラミング[7]で効率的に求めることができる。

【アルゴリズムの開始】

step1

$1 \leq r \leq s \leq g$ なる全ての r, s に対して、クラスター $(r, r + 1, \dots, s)$ のコスト $C(r, s)$ を計算する。

step2

$e(P(h, 2))$ ($2 \leq h \leq g$) を、

$$e(P(h, 2)) = \min_{2 \leq t \leq h} [C(1, t - 1) + C(t, h)] \quad (15)$$

として求める。

$$t_{h,2} = \arg \min_{2 \leq t \leq h} [C(1, t - 1) + C(t, h)] \quad (16)$$

として記憶しておく。

step3

クラスター数 $3 \leq q \leq g$ に対し、 $e(P(h, q))$ ($q \leq h \leq g$) を、

$$e(P(h, q)) = \min_{q \leq t \leq h} [e(P(t-1, q-1)) + C(t, h)] \quad (17)$$

として求める。

$$t_{h,q} = \arg \min_{q \leq t \leq h} [e(P(t-1, q-1)) + C(t, h)] \quad (18)$$

として記憶しておく。

step4

クラスター数 $2 \leq j \leq g$ に対し、クラスター列 $P(g, j)$ を求める。

$$P(h, 2) = (1, \dots, t_{h,2}-1, t_{h,2}, \dots, h) \quad (19)$$

$$P(h, q) = P(t_{h,q}-1, q-1) \cup (t_{h,q}, \dots, h) \quad (20)$$

より、求めようとするクラスター列の一番最後のクラスターの最初の文番号を取得できることから、クラスター列 $P(g, j)$ を得ることができる。

【アルゴリズムの終了】

関係式(12)より、クラスター数 j が増えるにつれ、 E_j は単調減少していく。また最小コスト値間の比 $R_j = E_j / E_{j-1}$ も一般に、 j が増えるにつれ単調減少していく傾向がある。クラスター列が正解のトピック区間列とほぼ一致しているとき、それ以上クラスター数が増えてもコストは殆ど減少することなく変わらず、比の値は 1 に近くなる。そこで R_j の閾値 α をとり、 $R_j \geq \alpha$ となる j で、最大の j に対応するクラスター列 $P(g, j)$ をトピック区間列と認定する。

4. 評価実験

4.1. 実験データ

概念ベースを生成するための学習用コーパスとして、CD - 毎日新聞 2000 年版 1 年分の約 106,000 記事の見出しと本文の部分を用いた。共起行列として、各行に対応する単語群として高頻度語 30,000 語をとり、各列に対応する単語群として、頻度順位が上位 51 番目以降の 1,500 語をとった。これを特異値分解により 100 次元の概念ベ

スに変換した。なお概念ベース生成には、NTT が開発した発想誘導型情報検索システム [8] [9] を用いた。

セグメント対象テキストとしては、CD - 毎日新聞 2001 年版から、経済、社会、科学のカテゴリの記事を 20 個ずつ計 60 個とり、異なるカテゴリの記事が隣接するように記事の本文部分を接続したデータを用意した。精度評価においては、各記事を 1 トピックと仮定する。このようなデータを、共通の記事がないように 2 つ用意した。各データ A、B の情報は表 2 のとおりである。ここでベースラインとは、ランダムに選んだ文境界が正解境界である確率を意味する。

表 2 データ A、B の情報

	データ A	データ B
記事数	60 記事	60 記事
総文数	1109 文	830 文
記事の最大長	68 文	74 文
記事の最小長	2 文	2 文
記事の平均長	18.5 文	13.8 文
ベースライン	5.3%	7.1%

4.2. 提案手法の評価実験

データ A、B を形態素解析し、得られた各単語のうち、自立語のみに概念ベース中の対応するベクトルを付与する。得られた概念ベクトルの系列に対して、クラスター内変動最小アルゴリズムを適用し、任意のクラスター数 j に対し、最小コスト E_j を求める。 E_j は j が増えるにつれ、図 1 のように単調減少していく。また最小コスト値間の比 $R_j = E_j / E_{j-1}$ も、図 2 のように j が増えるにつれ単調減少していく傾向がある。図 1、2 とともにデータ A に関するものである。別データに対する予備実験により R_j の閾値 α として $\alpha = 1.0005$ をとり、データ A、B に対し、 $R_j \geq \alpha$ となる j で、最大の j に対応する、最小コストをとるクラスター列 T を出力した。また、クラスター数が正解の 60 をとる、最小コストをとるクラスター列 N も出力した。各場合の精度は表 3 のとおりとなった。

なおここで、

再現率 = 正解の出力境界数 / 正解境界数

適合率 = 正解の出力境界数 / 出力境界数

F 値 = $(2 \times \text{再現率} \times \text{適合率}) / (\text{再現率} + \text{適合率})$

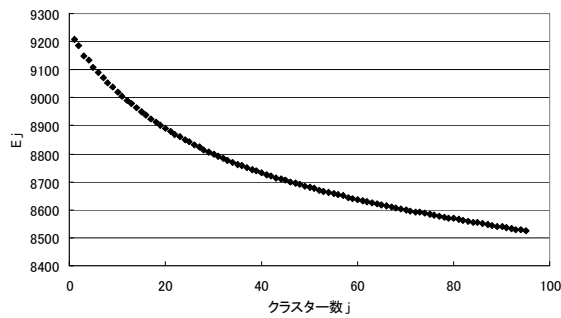


図1 クラスタ数と最小コストの関係

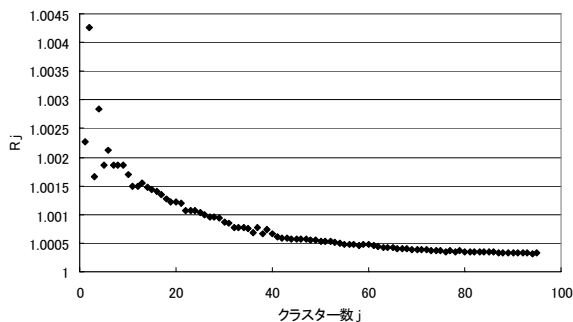


図2 クラスタ数と最小コスト値間比の関係

表3 提案手法の精度

データA					
	照合度合	再現率	適合率	F 値	誤り率
T(クラスタ 数: 53)	完全一致	69.5%	78.8%	73.9%	10.6%
	± 1 文	72.9%	82.7%	77.5%	
N(クラスタ 数: 60)	完全一致	76.3%	76.3%	76.3%	10.6%
	± 1 文	79.7%	79.7%	79.7%	

データB					
	照合度合	再現率	適合率	F 値	誤り率
T(クラスタ 数: 60)	完全一致	84.7%	84.7%	84.7%	9.4%
	± 1 文	86.4%	86.4%	86.4%	
N(クラスタ 数: 60)	完全一致	84.7%	84.7%	84.7%	9.4%
	± 1 文	86.4%	86.4%	86.4%	

である。照合度合が完全一致とは、正解境界と完全一致する出力境界のみでの精度を意味し、± 1 文とは、正解境界から前後 1 文だけの範囲にある出力境界での精度を意味する。誤り率は、一定の距離にある 2 単語の組で 1 記事内にあるものが別クラスターにあたり、異なる記事にあるものが同一クラスターにある割合であり、低いほど精度がよい。2 単語間の距離として、1 記事の平均長の半分程度の単語数をとった。

表4 概念ベクトル結束度法の精度

データA					
	照合度合	再現率	適合率	F 値	誤り率
T(クラスタ 数: 75)	完全一致	64.4%	51.4%	57.1%	19.4%
	± 1 文	67.8%	64.9%	66.3%	
N(クラスタ 数: 60)	完全一致	59.3%	59.3%	59.3%	16.1%
	± 1 文	62.7%	74.6%	68.1%	

データB					
	照合度合	再現率	適合率	F 値	誤り率
T(クラスタ 数: 51)	完全一致	67.8%	80.0%	73.4%	11.1%
	± 1 文	71.2%	90.0%	79.5%	
N(クラスタ 数: 60)	完全一致	74.6%	74.6%	74.6%	12.0%
	± 1 文	78.0%	84.7%	81.2%	

表5 Hearst 法の精度

データA					
	照合度合	再現率	適合率	F 値	誤り率
T(クラスタ 数: 69)	完全一致	22.0%	19.1%	20.5%	48.7%
	± 1 文	25.4%	22.1%	23.6%	
N(クラスタ 数: 60)	完全一致	20.3%	20.3%	20.3%	48.5%
	± 1 文	22.0%	22.0%	22.0%	

データB					
	照合度合	再現率	適合率	F 値	誤り率
T(クラスタ 数: 43)	完全一致	22.0%	31.0%	25.7%	41.9%
	± 1 文	22.0%	42.9%	29.1%	
N(クラスタ 数: 60)	完全一致	30.5%	30.5%	30.5%	37.9%
	± 1 文	35.6%	49.2%	41.3%	

4.3.他手法との比較

提案手法と他手法の精度を比較するため、2 節で述べた概念ベクトル結束度法と Hearst 法による評価実験を行った。ともに、窓幅は 1 記事の平均長の 3 分の 1 程度の単語数を取り、結束度の平滑化では、各単語境界とその直前、直後の単語境界の結束度の平均をとった。極小点となる単語境界の depth score を求め、depth score の大きい単語境界から、単語境界を直近の文境界に変換した上で、同じ文境界は重複することなく出力した。ここで depth score の閾値を実験的に、概念ベクトル結束度法では $\mu+1.3\sigma$ (μ :平均、 σ :標準偏差)とし、Hearst 法では $\mu+1.5\sigma$ として出力した(これで得られるクラスター列を T と表す)。また別に、正解境界数である 59 だけ境界を出力した(これで得られるクラスター列を N と表す)。各手法の精度は表 4、5 のようになった。

同じ概念ベクトルを用いる手法でも提案手法は概念ベクトル結束度法に比べ、F 値が 5.2% ~ 17.0% 上回り、誤り率が 1.7% ~ 8.8% 削減される。また、提案手法は Hearst 法に比べ、F 値が 45.1% ~ 59.0% 上回り、誤り率が 28.5% ~ 38.1% 削減され

る。これにより、提案手法はここで挙げたような他手法よりも高性能であることがいえる。

また提案手法ではデータAやBにおいて、74 文からなる長い記事や 2 文からなる短い記事で完全に再現しているものがある。これに対し、他手法では全般的に、非常に長い記事や短すぎる記事は再現できていない傾向がある。このことは、提案手法において任意の区間を考慮し、広域的、局所的な範囲双方でトピック区間としての妥当性を判断していることが、様々な長さのトピック区間の検出に有効であることを示している。

5. おわりに

テキストを概念ベクトルの系列ととらえ、ベクトルの系列を分割するクラスター列で、クラスター内変動が最小となるものをトピック区間列とする手法を提案した。本手法は、局所的な範囲内でトピック境界を判断する手法よりも優れた精度をもっていることを実験により検証した。

提案手法では任意のクラスターのコストを計算するため、文数が多い場合の処理時間は非常に長くなる。文の系列を 2 分割するコスト最小のクラスター列を求め、以降、これまで求めた分割位置を固定したまま、1 回分割してコスト最小となるクラスター列を求めていけば、最適ではないが、良いクラスター列が比較的短時間で得られる。各文をクラスターとし、以降、隣接するクラスター同士を結合してコスト最小となるクラスター列を求めていった場合も同様である。このような準最適解の精度についても今後検証していきたい。

また、クラスターの全変動の式(10)は、トピック区間において概念ベクトルが正規分布に従って分布していると仮定した場合の対数尤度に負の符号をつけたものに該当する。今後さらなる精度向上のため、トピック区間における概念ベクトルの分布の特性を精査し、クラスター群としてのよりよい妥当性の尺度を考究していきたい。

提案手法を現在構築中の映像コンテンツのインデキシングシステム[10][11]に組み込み、大語彙連続音声認識結果のテキストに対するトピックセグメンテーションの精度評価も行っていく予定である。

参考文献

- [1]Hearst, M.A.: Multi-Paragraph Segmentation of Expository Text, 32nd Annual Meeting of the Association for Computational Linguistics, pp.9-16(1994).
- [2]Hearst, M.A.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, Computational Linguistics, Vol.23, No.1, pp.33-64(1997).
- [3]別所克人: 単語の概念ベクトルを用いたテキストセグメンテーション, 情報処理学会論文誌, Vol.42, No.11(2001).
- [4]Schütze, H.: Dimensions of Meaning, Proc. Supercomputing '92, pp.787-796(1992).
- [5]Schütze, H. and Pedersen, J.O.: A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, Proc. RIAO '94, pp.266-274(1994).
- [6]奥野忠一, 芳賀敏郎, 久米 均, 吉澤 正: 多変量解析法, 日科技連(1971).
- [7]John, A.H.: Clustering Algorithms, John Wiley & Sons(1975).
- [8]Kato, T., Shimada, S., Kumamoto, M. and Matsuzawa, K.: Idea-Deriving Information Retrieval System, Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.187-193(1999).
- [9]熊本 睦, 島田茂夫, 加藤恒昭: 概念ベースの情報検索への適用—概念ベースを用いた検索の特性評価, 情報処理学会研究報告, Vol.SIG-ICS 115, pp.9-16(1999).
- [10]林良彦他: 映像コンテンツのインデキシングのための音声・言語処理, 第 65 回情報処理学会全国大会, (2003).
- [11]大附克年他: 大語彙連続音声認識を用いた音声・映像コンテンツのインデキシング, 日本音響学会春季研究発表会, (2003).