

文書集合の話題俯瞰のためのクラスタリング手法

横本 大輔[†] 鈴木 浩子[†] 牧田 健作[†] 宇津呂武仁^{††} 河田 容英^{†††}

福原 知宏^{††††}

[†] 筑波大学大学院システム情報工学研究科 〒 305-8573 茨城県つくば市天王台 1-1-1

^{††} 筑波大学システム情報系知能機能工学域 〒 305-8573 茨城県つくば市天王台 1-1-1

^{†††} (株) ナビックス 〒 141-0031 東京都品川区西五反田 8-3-6

^{††††} 独立行政法人 産業技術総合研究所 サービス工学研究センター 〒 135-0064 東京都江東区青梅 2-3-26

あらまし 本論文では、複数の話題の間の冗長性を考慮して、文書集合における最適な話題俯瞰を実現するための文書クラスタリング手法を確立することを目的とする。本論文の枠組みにおいては、まず、Wikipedia を知識源として、各文書の内容と Wikipedia 中の記述を照合しながら、各文書の内容に密接に関連した話題ラベルを複数個付与する。次に、話題ラベルが付与された文集集合に対して、複数の話題の間の冗長性を考慮して、文書集合における最適な話題俯瞰を実現するための文書クラスタリングを行う。以上の枠組みを、特定のクエリに対して関連するブログ記事を収集した文書集合に対して適用し、その有効性を評価した。

キーワード クラスタリング, 話題, 俯瞰, 冗長性

A Clustering Method for Overview of Topics in a Document Set

Daisuke YOKOMOTO[†], Hiroko SUZUKI[†], Kensaku MAKITA[†], Takehito UTSURO^{††}, Yasuhide

KAWADA^{†††}, and Tomohiro FUKUHARA^{††††}

[†] Grad. Sch. of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan

^{††} Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba 305-8573 Japan

^{†††} Navix Co., Ltd. Tokyo 141-0031, Japan

^{††††} Center for Service Research, National Institute of Advanced Industrial Science and Technology, Tokyo, 135-0064, Japan

Key words clustering, topic, overview, redundancy

1. はじめに

現代の情報社会においては、情報の氾濫の問題が顕著であり、このことは、いわゆる情報爆発の問題を引き起こしている。そして、そのように爆発的に増大する情報の集約や、俯瞰をするための技術の開発が強く望まれている。中でも、情報爆発が最も顕著に現れているのはウェブである。ウェブ上の情報の一例として、近年、一般個人が自由に情報を発信するツールであるブログが世界中で普及し、各地域の人々がそれぞれインターネット上で個人の意見や主観を発信することが可能になった。それに伴い、様々な情報がブログに記載され、様々な人々の意見や主観が Web 上に氾濫している。

このような状況を鑑みて、本論文の前段として、我々は、ブログ空間における多種多様な話題を俯瞰的に閲覧する方式の研究を提案した [3], [4], [7], [10]。具体的には、基本的な方式 [10]

として、Wikipedia を知識源として話題の体系を構築し、この Wikipedia の体系を元に、ブロガーのブログ記事集合に対して話題を対応付ける (ここで、話題として対応付けられた Wikipedia エントリを観点と呼ぶ) 方式を提案した。また、そのほか、複数の言語間で話題の分布を比較分析する方式 [3], [7],あるいは、時系列方向の話題の分布を分析する方式 [4] 等を提案した。

しかし、文書集合を効率よく俯瞰するためには、文書を話題に分類するだけでなく、複数の話題の間の類似関係を把握し、類似した冗長な話題を省き、代表的な話題に集約した上で閲覧する必要がある。この点において、我々のこれまでに提案した手法では、複数の話題の間の関連性を考慮した枠組みとなっていなかった。そこで、本論文では、複数の話題の間の冗長性を考慮して、文書集合における最適な話題俯瞰を実現するための文書クラスタリング手法を確立する。

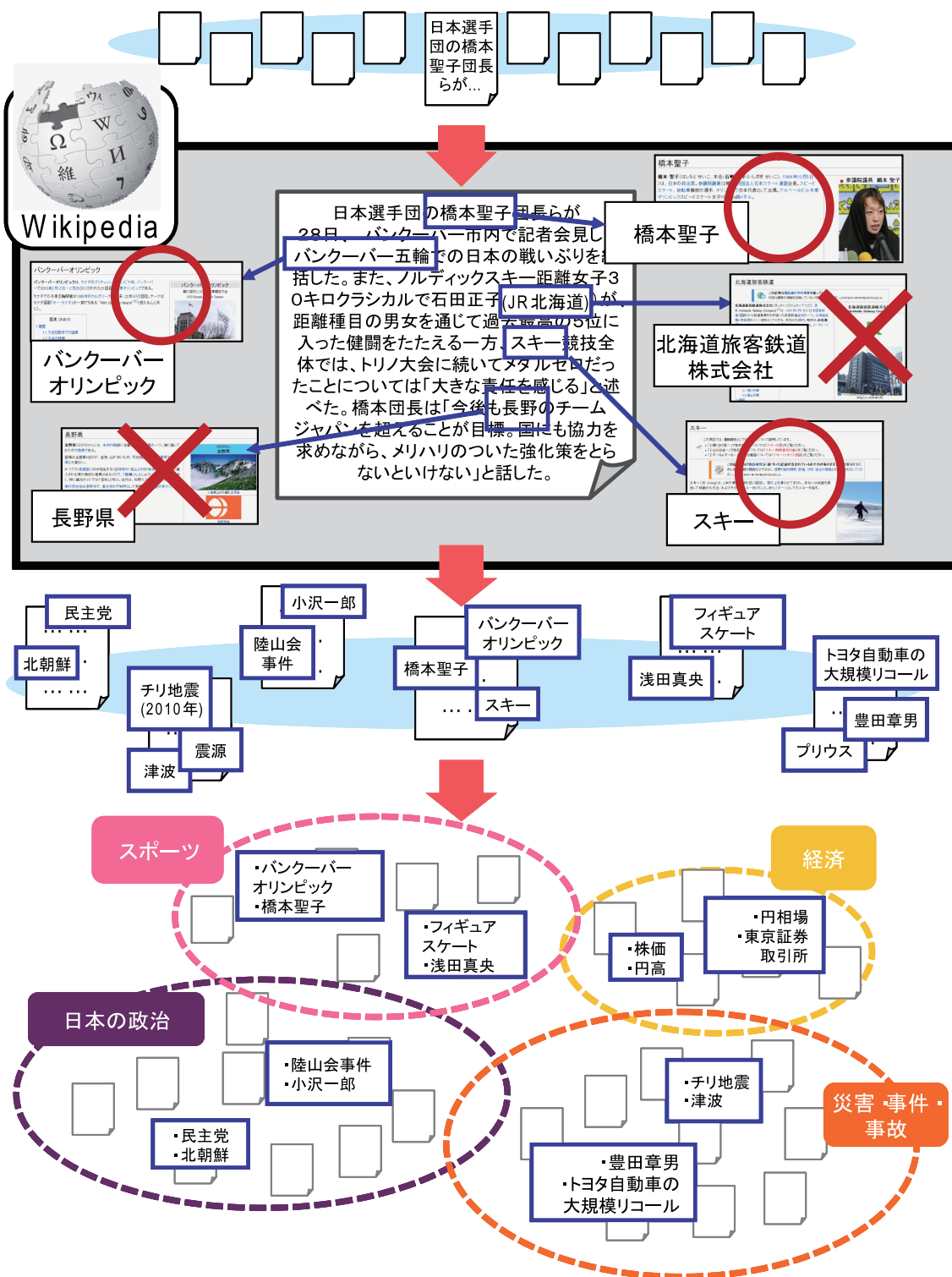


図1 文書への話題ラベルの付与および文書集合の話題俯瞰の枠組み

本論文における文書集合の話題俯瞰の全体的な枠組みを図1に示す。この枠組みにおいては、まず、個々の文書に対して、話題ラベルを複数個付与する。この話題ラベル付与の処理においては、図1に示すように、Wikipediaを知識源として、各文書の内容とWikipedia中の記述を照合しながら、各文書の内容に密接に関連した話題ラベルを複数個付与する。この話題ラベル付与の処理においては、各文書をクエリとして、Wikipediaのエントリを順位付けする問題として定式化し、特に、理論的な

枠組みとして、クエリ尤度モデル[6]に基づく方法(2.節)を用いる^(注1)。一例として、図1では、バンクーバー五輪の橋本聖子選手団長の記者会見の記事に対して「橋本聖子」、「バンクーバーオリンピック」、「スキ」などの話題ラベルが付与されて

(注1): 文献[10]においては、文書の内容とWikipediaエントリ中の記述の間の関連性を測定するために、文書類似度に基づく方法を用いていたが、文献[9]の結果においては、本論文で用いるクエリ尤度モデルに基づく手法の方が高い性能を示したため、本論文ではクエリ尤度モデルに基づく手法を用いる。

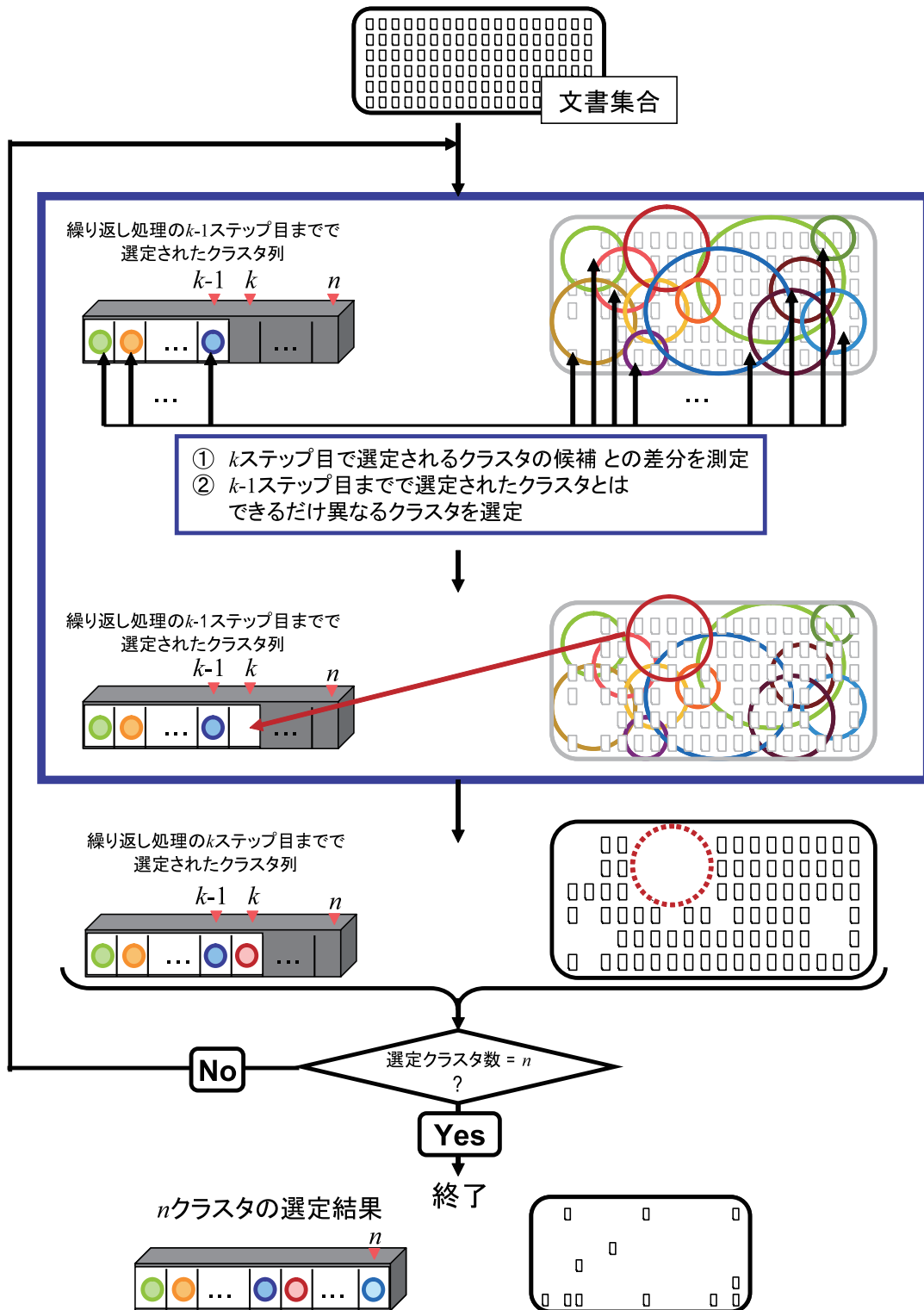


図2 クラスタ間の冗長性を考慮した話題俯瞰のためのクラスタリングアルゴリズムの考え方

いる。同様に、すべての対象文書に対して話題ラベルの付与を行う。

次に、話題ラベルが付与された文集集合に対して、複数の話題の間の冗長性を考慮して、文書集合における最適な話題俯瞰を実現するための文書クラスタリングを行う(3.節)。このクラスタリングにおいては、図2に示すように、各クラスタに属する文書に付与された話題ラベルに基づいて、クラスタ間の冗長性を測定するとともに、各クラスタの代表性を測定し、できる

だけ代表的で、かつ、すでに選定された上位のクラスタとは類似しないクラスタを順に選定する。

本論文では、以上の枠組みを、特定のクエリに対して関連するブログ記事を収集した文書集合に対して適用し、その有効性を評価した。本論文の範囲では、特に、選定された上位のクラスタの正解率、および、上位クラスタ間の冗長性の除去に重点を置いてアルゴリズムの調整を行い、一部、既存研究[5]を模倣した手法、および、語の頻度を用いた tf-idf による手法等の

性能を上回る性能を達成した。

2. 文書への話題ラベルの付与

本節では、提案手法、比較対象として用いた tf-idf による手法、および、一部、既存研究 [5] を模倣した手法について述べる。

2.1 Wikipedia を知識源とする文書への話題ラベルの付与

2.1.1 クエリ尤度モデル

本節では、情報検索の手法の 1 つであり、本研究で用いているクエリ尤度モデル [6] について説明する。

クエリと文書集合が与えられ、クエリに適合する度合いに従って文書を順位付けしたい場合、文書 D ($\in$ 文書コレクション C) がクエリ Q に適合する確率 $P(D | Q)$ を求めることができれば、これに従って順位付けを行うことができる。このとき、ベイズの定理を用いることにより、次式を得ることができる。

$$P(D | Q) = \frac{P(Q | D)P(D)}{P(Q)} \quad (1)$$

$P(Q)$ は文書 D に依存しないので定数とみなせる。また 文書 D に関しての何らかの事前知識がない限り、 $P(D)$ は一様であるとみなす。このとき、式 (1) を次のように簡略化できる。

$$\begin{aligned} P(D | Q) &= \frac{P(Q | D)P(D)}{P(Q)} \\ &\propto P(Q | D)P(D) \\ &\propto P(Q | D) \end{aligned}$$

上式に基づく情報検索手法をクエリ尤度モデルと呼ぶ。

次に、文書中の単語は独立に生起すると仮定して、文書をユニグラム言語モデルとして表現する。このとき、文書 D におけるユニグラム言語モデル θ_D からクエリ Q が生成される尤度、すなわちクエリ尤度 $P(Q | \theta_D)$ は、語彙 $w \in V = \{w_1, \dots, w_{|V|}\}$ のクエリ Q における出現頻度 $c(w_i, Q)$ を用いて次式のように定義される。

$$P(Q | \theta_D) = \prod_{w_i \in V} P(w_i | \theta_D)^{c(w_i, Q)} \quad (2)$$

上式を求めるためには、文書モデルのパラメータ θ_D を推定しなければならない。そのための推定方法の 1 つとしては、文書 D における最尤推定値 $P_{ML}(w_i | \theta_D)$ と、文書コレクション C における最尤推定値 $P_{ML}(w_i | \theta_C)$ を線形補間するという手法がある。まず、文書 D における最尤推定値 $P_{ML}(w_i | \theta_D)$ は、相対頻度を用いて以下のように定義される。

$$P_{ML}(w_i | \theta_D) = \frac{c(w_i, D)}{|D|} \quad (3)$$

ここで、 $c(w_i, D)$ は文書 D において語 w_i が出現する頻度、 $|D|$ は文書 D の文書長さすなわち総語数を示す。一方、文書コレクション C における最尤推定値 $P_{ML}(w_i | \theta_C)$ は次のようにして推定できる。

$$P_{ML}(w | \theta_C) = \frac{\sum_{D \in C} c(w_i, D)}{\sum_{D \in C} |D|} \quad (4)$$

ここで、文書コレクション C は、検索対象の文書集合である。

式 (3)、および、式 (4) と、補間の度合いを表現する定数 $\lambda \in [0, 1]$ を用いて、線形補間法による推定値は次のように定義される。

$$P(w_i | \theta_D) = \lambda P_{ML}(w_i | \theta_D) + (1 - \lambda) P_{ML}(w_i | \theta_C) \quad (5)$$

上式を用いることで、式 (2) のクエリ尤度 $P(Q | \theta_D)$ を求めることができる。

2.1.2 文書への話題ラベルの付与

本節では、前節で述べたクエリ尤度モデルの考え方をを用いて、対象文書集合の個々の文書に対して話題ラベルを付与する手法について説明する。なお本節以降では、文書を表す記号として A を用いる。

本研究では、文書中に出現する Wikipedia エントリタイトルから、文書の話題ラベルとして相応しいものを自動選定する。そのために、文書をクエリとみなして、文書中にエントリタイトルが出現した Wikipedia エントリ集合のランキングを行う。

はじめに、対象文書集合 $\mathbb{A}$ の個々の文書 $A \in \mathbb{A}$ は、文書中に出現した Wikipedia エントリタイトルの集合として表現される。

$$A = \{t(E_1), \dots, t(E_n)\}$$

まず、対象文書集合 $\mathbb{A}$ において、エントリタイトル $t(E)$ が 10 個以上の文書に出現した Wikipedia エントリを集めて、Wikipedia エントリ集合 $\mathbb{E}(\mathbb{A})$ を作成する。

$$\mathbb{E}(\mathbb{A}) = \left\{ E \mid \text{df}(\mathbb{A}, t(E)) \geq 10 \right\}$$

次に、文書 A にエントリタイトル $t(E)$ が出現し、かつ、Wikipedia エントリ集合 $\mathbb{E}(\mathbb{A})$ に含まれる Wikipedia エントリ E を抽出し、文書 A に対する話題ラベルの候補集合に対応する Wikipedia エントリ集合 $\mathbb{E}(A)$ を作成する。

$$\mathbb{E}(A) = \left\{ E \in \mathbb{E}(\mathbb{A}) \mid t(E) \in A \right\}$$

そして、クエリ尤度モデルに基づいて、文書 A をクエリとみなして、Wikipedia エントリ集合 $\mathbb{E}(A)$ の順位付けを行う。具体的には、 $Q = A$ 、 $D = E \in \mathbb{E}(A)$ 、 $C = \mathbb{E}(\mathbb{A})$ として、式 (2) の $P(A | \theta_E)$ を求める。

以上のように推定した $P(A | \theta_E)$ を用いることで、文書 A に付与する話題ラベル集合 $L(A)$ を以下の手順で決定する。具体的には、Wikipedia エントリ集合 $\mathbb{E}(A)$ における $P(A | \theta_E)$ の最大値 $\max_{E' \in \mathbb{E}(A)}$ に対して、その α 倍以上の $P(A | \theta_E)$ を持つ Wikipedia エントリのタイトル $t(E)$ を、文書 A の話題ラベルとして抽出する。また、1 文書に付与する話題ラベルの数は最大 10 個とし、 $P(A | \theta_E)$ の大きいものから順に上位 10 個までエントリを選定し、そのタイトルを $L(A)$ の要素とした。以上より、 $t(E) \in L(A)$ となる Wikipedia エントリ E の満たすべき条件は以下となる。

- $E \in \mathbb{E}(A)$

- $P(A | \theta_E) \geq (\alpha \times \max_{E' \in \mathbb{E}(A)} P(A | \theta_{E'}))$
- $P(A | E)$ の大きいものから順に 10 個 E を選ぶ。

なお、本論文では $\alpha = 0.6$ とした。

2.2 比較手法：tf-idf

本節では、比較手法として、Wikipedia の知識を用いず、単語頻度と文書頻度を用いた tf-idf により話題ラベルを付与する手法について説明する。

具体的には、 $\text{tf}(A, E)$ を、文書 A においてエントリタイトル $t(E)$ が出現する頻度、 $\text{df}(\mathbb{A}, E)$ を、文書集合 $\mathbb{A}$ においてエントリタイトル $t(E)$ が出現する文書数、 $|\mathbb{A}|$ を総文書数として、tf-idf の値が下限値以上となるエントリタイトル $t(E)$ を最大 10 個選定する。以上より、話題ラベル集合 $L(A)$ の要素 $t(E)$ の満たすべき条件は以下となる。

- $E \in \mathbb{E}(A)$
- $(\text{tf}(A, t(E)) \cdot \log_2 \frac{|\mathbb{A}|}{\text{df}(\mathbb{A}, t(E))}) \geq k$
- $(\text{tf}(A, t(E)) \cdot \log_2 \frac{|\mathbb{A}|}{\text{df}(\mathbb{A}, t(E))})$ の大きいものから順に 10 個まで選ぶ。

また、本論文では、tf-idf の下限値を $k = 10.0$ とした。

2.3 比較手法：話題ラベルの出現確率の差を用いる手法

最後に、先行研究 [5] で提案されている手法について説明する。この手法では、対象文書集合における出現確率が、一般の文書集合における出現確率に比べて大きい語が、対象文書集合において話題になっている語であると考え、そのような語を話題ラベルして抽出する。

具体的には、まず語 w の出現確率の距離 $\delta_w(p \| q)$ として、次式で定義される pointwise KL ダイバージェンスを用いる。

$$\delta_w(p \| q) = p(w) \log \frac{p(w)}{q(w)}$$

次に、対象文書集合と一般的な文書集合における出現確率の pointwise KL ダイバージェンスを求める。本論文では、一般的な文書集合として、Wikipedia の全エントリの集合を $\mathbb{W}$ として、 $\mathbb{W}$ の全文書集合を用いた。また、式 (4) の $P_{ML}(w_i | C)$ において、文書コレクション C を、それぞれ、 $C = \mathbb{A}$ 、 $C = \mathbb{W}$ として推定した最尤推定値を $P_{ML}(t(E) | \mathbb{A})$ 、および、 $P_{ML}(t(E) | \mathbb{W})$ として、以下の pointwise KL ダイバージェンスを求める。

$$\delta_w(P_{ML}(t(E) | \mathbb{A}) \| P_{ML}(t(E) | \mathbb{W})) (= \phi \text{ とおく})$$

次に、 $\mathbb{E}(\mathbb{A})$ 中の Wikipedia エントリのタイトルのうち、 ϕ の大きいものから降順に上位 300 個のエントリタイトル $t(E)$ を収集し、文書集合 $\mathbb{A}$ に対する話題ラベルの候補集合 $L_0(\mathbb{A})$ を作成する。

$$L_0(\mathbb{A}) = \left\{ t(E) \mid E \in \mathbb{E}(\mathbb{A}), t(E) \in L_0(\mathbb{A}) \right\}$$

そして、各文書 A に対する話題ラベルとして、文書中出现したエントリタイトル $t(E)$ のうち $L_0(\mathbb{A})$ に含まれるものを収集し話題ラベル集合 $L(A)$ とする。

$$L(A) = \left\{ t(E) \in A \mid t(E) \in L_0(\mathbb{A}) \right\}$$

3. 話題俯瞰のためのクラスタリング手順

3.1 定式化

まず、話題ラベル l を持つ文書の集合 $\mathbb{A}(l)$ を以下で定義する。

$$\mathbb{A}(l) = \left\{ A \in \mathbb{A} \mid l \in L(A) \right\}$$

さらに、これを拡張して、話題ラベル集合 L の要素であるラベルを全て持つ文書の集合 $\mathbb{A}(L)$ を、各要素であるラベルを持つ文書の集合の積集合として、以下で定義する。

$$\mathbb{A}(L) = \bigcap_{l \in L} \left\{ A \in \mathbb{A} \mid l \in L(A) \right\}$$

この表記を用いれば、ラベル l を持つ文書の集合は、 $\mathbb{A}(\{l\})$ と表現される。同様に、ラベルの組 l_1 および l_2 を持つ文書の集合は、 $\mathbb{A}(\{l_1, l_2\})$ と表現される。

そして、文書に付与するラベルとして、10 個以上の文書に付与される必要があるという条件を課すと、話題ラベルの候補集合としては、一つの話題ラベルの場合は、以下の $\mathbb{L}_1$ 、二つの話題ラベルの組の場合は、以下の $\mathbb{L}_2$ で記述される。

$$\mathbb{L}_1 = \left\{ \{l\} \mid l \in \bigcup_{A \in \mathbb{A}} L(A), |\mathbb{A}(\{l\})| \geq 10 \right\}$$

$$\mathbb{L}_2 = \left\{ \{l_1, l_2\} \mid l_1, l_2 \in \mathbb{L}_1, |\mathbb{A}(\{l_1, l_2\})| \geq 10 \right\}$$

ここで、 $\mathbb{L}_1$ 、あるいは、 $\mathbb{L}_2$ は、実際には、前節で述べた各手法によって文書 A に付与された話題ラベルの集合を $L(A)$ として、それらの $L(A)$ を上記の式に適用することにより求める。

本研究のタスクは、これらの話題ラベルで定義されるクラスタから、できるだけ文書集合を被覆し、かつ、できるだけクラスタ間の重複が少なくなるように、クラスタを抽出することである。そこで、本論文においては、以下に述べる各項の重み付き線形和をコストとして、貪欲アルゴリズムによりコストをできるだけ小さくするラベルを順に選定する^(注2)。

- 各文書におけるラベルと文書の関係の強さの平均を表す純度項

$$C_{rel}(L) = - \frac{\sum_{A \in \mathbb{A}(L)} \sum_{l \in L} rel(A, l)}{|\mathbb{A}(L)|}$$

$$rel(A, l) = \begin{cases} \log_2 P(A | E) & (\text{クエリ尤度モデル}) \\ \text{tf}(A, l) \cdot \log_2 \frac{|\mathbb{A}|}{\text{df}(\mathbb{A}, E)} & (\text{tf-idf}) \end{cases}$$

- クラスターの大きさを表す項

$$C_{size}(L) = - \frac{|\mathbb{A}(L)|}{|\mathbb{A}|}$$

- 候補クラスター集合 $\mathbb{L}$ 内における重複度を表す項

$$C_{dup-rst}(L) = \frac{\sum_{L' \in \mathbb{L} - \{L\}} J(\mathbb{A}(L), \mathbb{A}(L'))}{|\mathbb{L}| - 1}$$

(注2)：この項のうち、「各文書におけるラベルと文書の関係の強さの総和を表す純度項」を除いては、文献 [5] で提案されたものを参考に行っているが、符合の正負などは本論文において経験的に設定した。

- 既に選択したクラスタ集合に対する重複度を表す項

$$C_{dup-slt}(L) = \sum_{L' \in \mathbb{L}_{slt}} J(A(L), A(L'))$$

ただし, $J(S_1, S_2)$ は次式で定義される Jaccard 係数を示す.

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

また, 各項の重み付き線形和を次式で定義する.

$$cost(L) =$$

$$\lambda_1 \cdot C_{rel} + \lambda_2 \cdot C_{size} + \lambda_3 \cdot C_{dup-rst} + \lambda_4 \cdot C_{dup-slt}$$

本論文では, コストができるだけ小さくなるような順にラベルを選定する手順として以下の「話題ラベル選択手順」を用いる.

「話題ラベル選択手順」

手法ごとに決められている話題ラベルの候補集合 $\mathbb{L}_i$, あるいは, $\mathbb{L}_2$ を $\mathbb{L}$ の初期値とし, 選択された話題ラベルの集合 $\mathbb{L}_{slt}$ の初期値を空集合とする.

(1) 各ラベル候補 $L \in \mathbb{L}$ について, $cost(L)$ を計算

(2) $score(l^n)$ が最大のラベルを選択

$$L_{min} = \underset{L \in \mathbb{L}}{\operatorname{argmin}} cost(L)$$

(2a) 選択したラベル L_{min} を $\mathbb{L}_{slt}$ に追加

(2b) 選択したラベル L_{min} を候補ラベル集合 $\mathbb{L}$ から除外

(2c) L_{min} の順位 $rank(L_{min}) = |\mathbb{L}_{slt}|$ を記録
 $|\mathbb{L}_{slt}| \geq N$ を満たすまで, (1), (2) の処理を繰り返す.

3.2 評価対象手法

評価手法の一覧を表 1 に示す.

(ID=1) 2.1 節の手法により, Wikipedia を知識源として, クエリ尤度モデルに基づいて話題ラベルの付与を行う. 話題ラベル数は 2. コスト $cost(L)$ の各項の重みは経験的に $\lambda_1 = \lambda_2 = \lambda_3 = 0.3, \lambda_4 = 0.1$ とした.

(ID=2) 話題ラベル数による違いの比較, 分析のため, (ID=1) と同様の手法で, 話題ラベル数のみ 1 とする.

(ID=3) 純度項の有無による違いの比較, 分析のため, (ID=1) におけるコスト $cost(L)$ において, 純度項を考慮しないように $\lambda_1 = 0$ とする. それ以外の重みは, 暫定的に, $\lambda_2 = 0.7, \lambda_3 = 0.2, \lambda_4 = 0.1$ とした.

(ID=4) 文書に対する話題ラベルの付与において, Wikipedia の知識を用いることの有効性を確認するため, Wikipedia の知識を用いない話題ラベルの付与手法を用いる. 具体的には, 2.2 節の手法により, tf-idf に基づいて話題ラベルを付与する.

(ID=5) Wikipedia の知識を用いずに話題ラベルの付与を行う手法として, 2.3 節の手法により, 確率の差を用いて話題ラベルの候補集合を作成する. また, 文献 [5] の手法を模倣し, 純度項は用いない.

4. 評価

4.1 データセット

実験に用いるデータセットとしては, クエリ t_0 に対して, 関連するブログ記事集合を収集した結果を用いた. クエリ t_0 を含む日本語ブログの収集においては, Yahoo! Search BOSS API^(注3) を利用し, 日本語ブログホスト大手 6 社^(注4) のドメインを対象としてブログ記事の収集を行った. 検索の際には, 複数のドメインを一度に指定して検索し, 1,000 件の記事を取得する. 次に, ブログ記事検索後, 検索結果の URL をブログサイト単位にまとめる. その結果, 一つの検索クエリあたり約 200 前後のブログサイトが取得される. 次に, 各ブログサイトをドメイン指定し, 初期クエリ t_0 を検索クエリとすることにより, 各ブログサイト中において初期クエリ t_0 を含むブログ記事を収集し, ブログ記事集合 A を作成する.

本論文においては, 「原子力発電所」をクエリ t_0 とし, 8,051 記事を収集した.

4.2 評価手順

前節で作成したブログ記事集合を対象文書集合 $\mathbb{D}$ として, 3.2 節で述べた手法により, 50 個のクラスタを抽出した結果を評価した. 評価手順としては, ラベルを人手により評価し, 「適切」, 「やや適切」, 「やや適切でない」, 「適切でない」の 4 段階の適合度を付与した. そして, 「適切」な場合の適合度を 100% とし, 以下, 段階的に 66%, 33%, 0% とし, 上位 50 個のクラスタにおけるラベルの適合度の平均として, 正解率を評価した.

- ラベル数が 1 の場合

$$\text{正解率} = \frac{\sum_{i=1}^{50} \text{適合度}(l_{i1})}{50}$$

- ラベル数が 2 の場合

$$\text{正解率} = \frac{\sum_{i=1}^{50} (\text{適合度}(l_{i1}) + \text{適合度}(l_{i2}))}{50 \times 2}$$

ここで, l_{ij} は i 番目に抽出されたクラスタの話題ラベルを表す. 話題ラベル数が 1 の場合, i 番目のクラスタの話題ラベルは l_{i1} のみである. 話題ラベル数が 2 の場合, i 番目のクラスタは 2 つの話題ラベル l_{i1}, l_{i2} を持つ. そして, 上位 50 個のクラスタが持つすべての話題の適合度の平均として, 正解率を定義した.

また, 50 個のクラスタによって被覆した文書の割合と, 収集された文書の重複率を求めた.

$$\text{被覆率} = \frac{\left| \bigcup_{i=1}^{50} A(L_i) \right|}{|A|}$$

$$\text{重複率} = 1 - \frac{\left| \bigcup_{i=1}^{50} A(L_i) \right|}{\sum_{i=1}^{50} |A(L_i)|}$$

(注3) : <http://developer.yahoo.com/search/boss/>

(注4) : fc2.com, yahoo.co.jp, ameblo.jp, goo.ne.jp, livedoor.jp, hatena.ne.jp

さらに、適合度とは別に、ラベルの粒度が大きすぎるもの(適合度は「やや適切でない」)、小さすぎるもの(適合度は「やや適切でない」)の数を集計し、ラベル数2の場合については、2つのラベルが冗長かどうかについても判定を付与した。

ここで、2個の話題ラベルを用いた際には、一方のラベルの粒度が大きい、または小さい場合でも、もう一方のラベルが適切であれば、ユーザはクラスタの内容を把握することができると考えられる。そこで、提案手法における正解率の上限値として、一方のラベルの粒度が不適切な場合でも、もう一方のラベルと同じ適合度を付与した場合の正解率を求めた。

4.3 評価結果

評価の結果を表1に示す。また、各手法において抽出された上位5個の話題ラベルを表2に示す。

まず、手法1および手法2と手法4の比較を以下に述べる。手法1および手法22のWikipediaを知識源として話題ラベルの付与を行い、コスト $cost(L)$ において純度項を用いた手法は、ラベル数によらず、約75%という高い正解率を示した。これに対して、手法4のWikipediaを知識源として用いず、tf-idfに基づいて話題ラベルの付与を行った場合、正解率は59.3%と、15%程度の差がみられた。tf-idfによる手法は、単語の出現頻度のみに基づいて話題ラベルを付与しているため、「時事通信社」、「出版社」といった、その新聞社のニュースを引用していれば出現する単語や、ブログ記事のヘッダ部、フッター部に出現しているだけの単語などを話題ラベルとして抽出してしまっており、これらが正解率を大きく下げた原因となっていた。また、抽出された適切な話題ラベルの傾向として、tf-idfによる手法は「首相、野田」や「東電、損害賠償」といった、ある程度一般的で、文書において出現しやすいWikipediaエントリタイトルが選ばれる傾向にあった。これに対して、Wikipediaを知識源として用いる手法では、「国際原子力事象評価尺度、原子力安全委員会」、「非常用炉心冷却装置」など、具体性の高いWikipediaエントリタイトルが話題ラベルとして多く選ばれていた。これらの話題ラベルは、文書の主題になっていても、文書中の出現頻度があり多くならず、そのため単語頻度の情報だけを用いている手法では収集できなかったと考えられる。

次に、手法1および手法2と手法5の比較を以下に述べる。手法5の、確率の差を用いて話題ラベルの候補を制限する手法も、正解率は61.3%であり、手法1、2のWikipediaを知識源として用いる手法に比べ、15%ほど低い正解率を示した。正解率を下げる要因としては、科学者や反原発活動家の人名といった、粒度の小さい話題ラベルが収集される傾向にあることが挙げられる。それらの粒度が小さいWikipediaエントリタイトルは一般的な文書集合における出現確率が小さく、確率の差が大きくなりやすいため、話題ラベルの候補として抽出されやすい。そのため、そのような粒度の小さい話題ラベルが抽出される傾向が強くなっていると考えられる。また、話題ラベルを1つしか提示していないため、「報酬」、「牛肉」、「ロボット」といった、それだけではどのような話題なのか把握できないものが多く見られた。これに対して、Wikipediaを知識源とする手法では具体性の高い話題ラベルが選ばれているため、話題ラベルが1つ

の場合であっても、内容の把握は容易になっている。

手法1および手法2と手法3の比較を以下に述べる。手法3の、純度の項が用いない手法の場合も、正解率は60.0%であり、手法1および手法2のWikipediaを知識源として用いる手法に比べ、15%ほど低い正解率を示した。正解率を下げる要因としては、「日本 政治」や「原子炉 システム」など、粒度の大きい話題ラベルが多く抽出されていることが挙げられる。これは、純度項を考慮せずに、重複の少ないものから、できる限り大きいクラスタを選ぶことになるため、多くの文書を持つ粒度の大きな話題ラベルが収集する傾向が強くなっているものと考えられる。これに対して、純度項を用いた場合、話題ラベルのまとまりの良さを捉えることで、そのような話題ラベルが選ばれる問題を解決できていると考えられる。

最後に、手法1と手法2の比較を以下に述べる。手法1の、ラベル数2の場合において、「東海地震、地震予知」、「再生可能エネルギー、ドイツ」など、2つの話題ラベルがあることで文書集合の内容が的確に表現されているものが確認できた。しかし、両手法の正解率はほぼ同程度という結果になっている。ラベル数2の場合では、2つのラベルのうち一方が適切であっても、それと包含関係にある粒度の大きいラベルが、もう一つのラベルに選ばれているものがあり、それらが正解率を下げる原因となっていた。2つのラベルの冗長性を考慮して、そのようなラベル組を選ばないようにすることで、精度の改善が期待できると思われる。そのための手法の1つとして、Wikipediaの知識源を用いてラベル間の上位下位関係を判定するといった方法が考えられる。また、手法1の正解率の上限値として、一方のラベルが適切であれば、もう片方に粒度が不適切なラベルが選ばれることを許容するという基準で評価を行った結果、正解率は80.0%となり、単独のラベルを用いる手法2よりも高い正解率を示した。

5. 関連研究

文献[5]では、本論文と同様に文書集合をできるだけ冗長性を排してクラスタリングするタスクを集合被覆問題として定式化している。対象文書集合と一般的な文書集合における単語の出現確率の差に基づいて、話題ラベルとなる n グラムを抽出し、貪欲法によって集合被覆問題を解く手法を提案し、確率の差のみに基づいてクラスタリングする場合に比べて高い性能を発揮することを示している。この手法に対して、本論文では、Wikipediaを知識源として文書に話題ラベルを付与することで、より適切な話題ラベルを抽出できることを示した。また、文献[5]では単一の話題ラベルを用いているのに対し、本論文ではこれを拡張して複数のラベルを付与することを提案した。

一方、文献[2]は、Webページの検索結果を分類し、各分類に対して適切な要約文を付与するという手法を提案している。また、文献[1],[8]では、検索された個々のWebページに対してラベルの付与を行い、付与されたラベルに基づいて分類を行う手法を提案している。これらの手法では、我々のこれまでの手法[10]と同様に、複数のクラスタ間の冗長性を陽に検出し、抑制する方式を採用していない点が本論文の方式とは異なる。

表 1 各評価対象手法の評価結果

手法 ID	話題ラベルの正解率 (%)	粒度大 (個数)	粒度小 (個数)	冗長 (個数)	被覆された述べ文書数	被覆率 (%)	重複率 (%)
1	75.6	16	2	10	710	8.8	7.6
2	75.3	1	4	-	3,727	46.3	28.0
3	60.0	12	5	14	3,257	40.5	17.6
4	59.3	7	0	9	661	8.2	12.9
5	61.3	2	2	-	7,776	96.6	75.4
1 上限値	80.0 (+4.4)	16	2	-	710	8.8	7.6
3 上限値	64.0 (+4.0)	12	5	-	3,257	40.5	17.6
4 上限値	66.7 (+7.4)	7	0	-	661	8.2	12.9

表 2 各手法において抽出された上位 5 個の話題ラベル

手法 ID	話題ラベル
1	新党日本, 政党
	原子力事故, チェルノブイリ原子力発電所事故
	津波, リスク
	石炭, 再生可能エネルギー
	発電所, 揚水発電
2	河内原子力発電所
	原子力
	広野火力発電所
	新党日本
	高速鉄道
3	原子力発電, 原子力
	福島第一原子力発電所, 東京電力
	放射能, 放射性物質
	原子炉, 原子力
	津波, 地震
4	復興, ボランティア
	歯科, インプラント
	時事通信社, 小説
	検察官, 小沢一郎
	岩手県, 宮城県
5	原子力発電
	原子炉
	原子力
	放射能
	東京電力

6. おわりに

本論文では、複数の話題の間の冗長性を考慮して、文書集合における最適な話題俯瞰を実現するための文書クラスタリング手法を確立することを目的とした。本論文の枠組みにおいては、まず、Wikipedia を知識源として、各文書の内容と Wikipedia 中の記述を照合しながら、各文書の内容に密接に関連した話題ラベルを複数個付与する。次に、話題ラベルが付与された文集集合に対して、複数の話題の間の冗長性を考慮して、文書集合における最適な話題俯瞰を実現するための文書クラスタリング

を行った。以上の枠組みを、特定のクエリに対して関連するブログ記事を収集した文書集合に対して適用し、その有効性を評価した。今後は、被覆率を改善するために、LDA 等の既存のトピックモデルによって推定されたトピックと提案手法を併用し、より俯瞰性の高いクラスタリング手法を実現する。その手法においては、文献[11]等とは異なり、LDA 等の既存のトピックモデルによって推定されたトピックにおいても一定の冗長性が存在するという設定のもとで、トピック集合の俯瞰を実現する。

文 献

- [1] 馬場康夫, 黒橋禎夫. キーワード蒸留型クラスタリングによる大規模ウェブ情報の俯瞰. 情報処理学会論文誌, Vol. 50, No. 4, pp. 1399–1409, 2009.
- [2] 原島純, 黒橋禎夫. PLSI を用いたウェブ検索結果の要約. 言語処理学会第 16 回年次大会論文集, pp. 118–121, 2010.
- [3] 牧田健作, 横本大輔, 鈴木浩子, 宇津呂武仁, 河田容英, 福原知宏. Wikipedia を多言語知識源とするブログ集合の話題分析. 電子情報通信学会技術研究報告, NLC2011-18, pp. 95–100, 2011.
- [4] 牧田健作, 横本大輔, 宇津呂武仁, 福原知宏. トピックに関する話題の時系列分布に着目したブログ分析. 第 3 回 DEIM フォーラム論文集, 2011.
- [5] P. Muthukrishnan, J. Gerrish, and D. R. Radev. Detecting multiple facets of an event using graph-based unsupervised methods. In *Proc. 22nd COLING*, pp. 609–616, 2008.
- [6] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. 21st SIGIR*, pp. 275–281, 1998.
- [7] 鈴木浩子, 横本大輔, 牧田健作, 宇津呂武仁, 河田容英, 福原知宏. Wikipedia を知識源とする日英ブログ記事集合の観点分類と言語間対照分析. 情報処理学会研究報告, Vol. 2011, No. (2011-DBS-153), 2011.
- [8] 戸田浩之, 中渡瀬秀一, 片岡良治. 特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案. 情報処理学会論文誌: データベース, Vol. 46, No. SIG 13(TOD 27), pp. 40–52, 2005.
- [9] D. Yokomoto, K. Makita, H. Suzuki, D. Koike, T. Utsuro, Y. Kawada, and T. Fukuhara. LDA-based topic modeling in labeling blog posts with Wikipedia entries. In *Proc. 1st IDP*, 2012.
- [10] 横本大輔, 林東権, 牧田健作, 宇津呂武仁, 河田容英, 福原知宏, 神門典子, 吉岡真治, 中川裕志, 清田陽司. 特定トピックに関するブログ記事集合の観点分類における Wikipedia の利用. 第 3 回 DEIM フォーラム論文集, 2011.
- [11] Y. Yue and C. Guestrin. Linear submodular bandits and their application to diversified retrieval. In *Advances in Neural Information Processing Systems 24*, pp. 2483–2491, 2011.