文書集合の話題俯瞰手法に関する分析*

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¶
筑波大学大学院 システム情報工学研究科† 筑波大学 システム情報系‡
(株) ナビックス§ 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¶

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，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を提案した [7]．具体的には，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方式を提案した．

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はなく，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．

2 文書集合の話題俯瞰のためのクラスタリング手法

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 [4] に基づく方法を用いる [8][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」,「スキー」などの話題ラベルが付与されている．同様にして，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．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．各手法の説明を以下に述べる．

(ID=1) 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.

(ID=2) 話題ラベル数による違いの比較，分析のため，(ID=1) において話題ラベル数を 1 とする．

(ID=3) 純度項 (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) の有無による違いの比較，分析のため，(ID=1) において，純度項を考慮しない．

*An Analysis on Methods for Overview of Topics in a Document Set

[1] 文献 [7] 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 [6] 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．

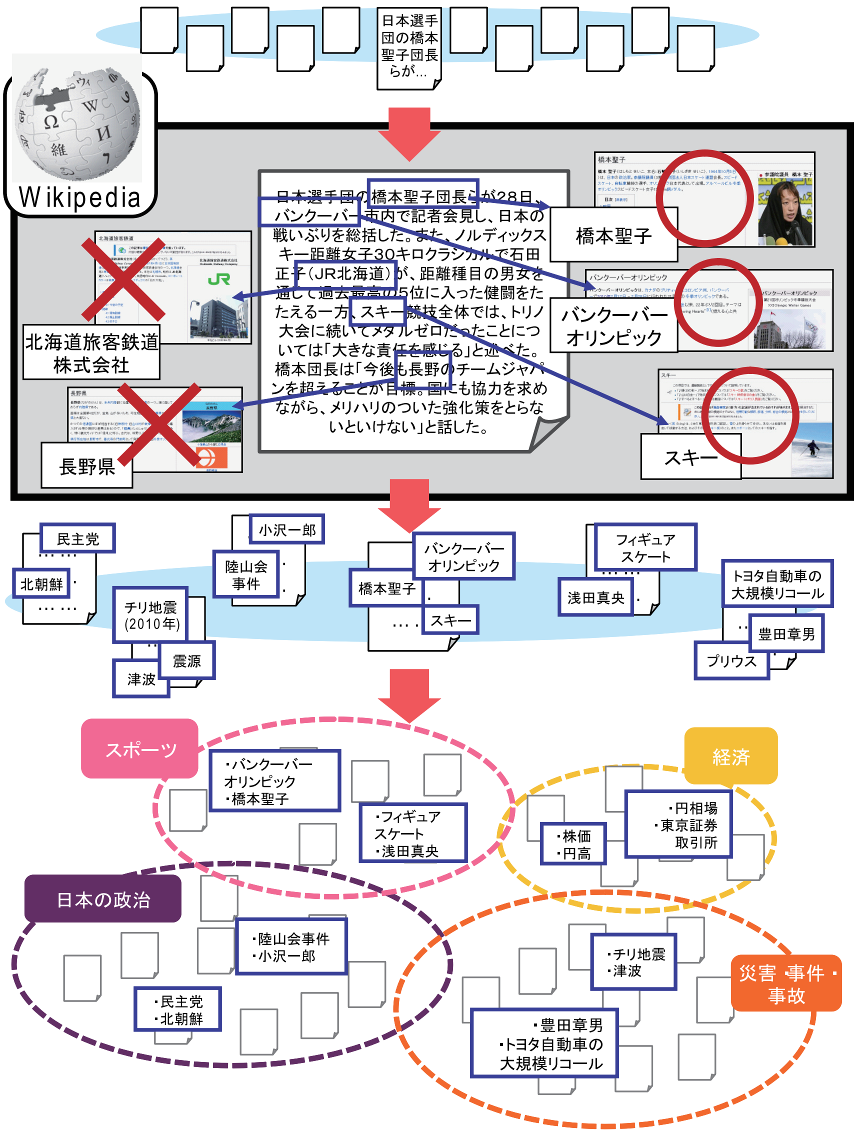

図 1: 文書への話題ラベルの付与および文書集合の話題俯瞰の枠組み

(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，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，文献 [3] の手法を模倣し，確率の差を用いて話題ラベルの候補集合を作成するとともに，純度項は用いない．

3 評価

3.1 評価手順

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[2] を利用し，日本語ブログホスト大手 6 社[3]のドメインを対象としてブログ記事の収集を行った．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．

これらのブログ記事集合を対象文書集合として，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．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

[2] http://developer.yahoo.com/search/boss/
[3] fc2.com，yahoo.co.jp，ameblo.jp，goo.ne.jp，livedoor.jp，hatena.ne.jp

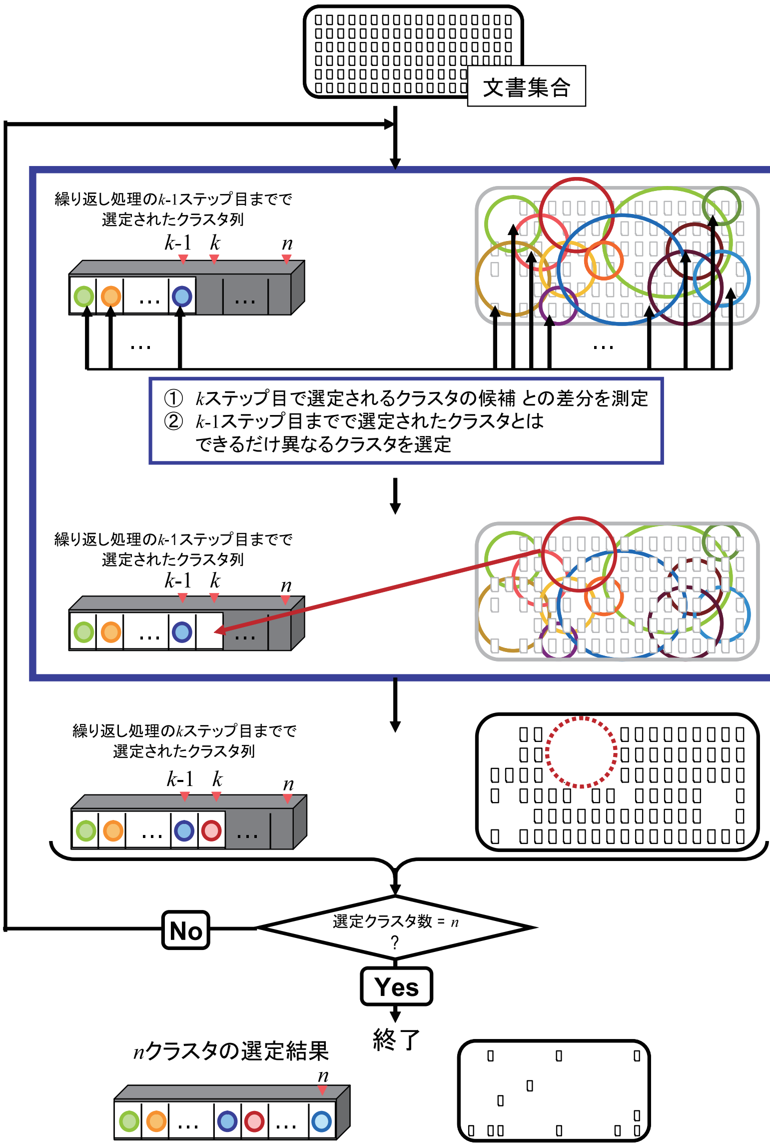

図 2: クラスタ間の冗長性を考慮した話題俯瞰のためのクラスタリングアルゴリズムの考え方

率を求めた．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．

3.2 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．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ったが，評価結果から分かるように，既存研究 [3] を一部模倣した手法，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．

4 関連研究

文献 [3] 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

表 1: 評価対象手法の一覧および評価結果

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)	被覆率 (%)	重複率 (%)
1 上限値	**80.0**	1	4	10	8.8	7.6
1	**75.6**	16	2	10	8.8	7.6
2	75.3	1	4	-	46.3	28.0
3	60.0	12	5	14	40.5	17.6
4	59.3	7	0	9	8.2	12.9
5	61.3	2	2	-	96.6	75.4

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

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 [3] 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 [5, 1] 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 [7] 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．

5 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した．本論文の枠組みを，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．

参考文献

[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] 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.

[4] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. 21st SIGIR*, pp. 275–281, 1998.

[5] 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.

[6] 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.

[7] 横本大輔, 林東權, 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.

[8] 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. 文書集合の話題俯瞰のためのクラスタリング手法. 第 4 回 DEIM フォーラム論文集, 2012.