【レギュラー論文】

ウィキラベルによる報道争点の検出

野　本　忠　司[†1]

本研究では国内外の新聞・放送メディアの注目対象と注目度，また動向を客観に比較可能にするため，米シンクタンク PEJ (Pew Center for Excellence in Journalism) の報道指標である News Coverage Index (NCI) を自動化することを考える．本研究は，特に自動化問題を文書クラスタへのラベル付与という観点で捉え，記事クラスタをウィキペディアの概念体系にマップし，ラベルを生成するというアプローチを展開する．TDT5 データセットを用いて評価した結果，ラベル付与については従来手法に比べて顕著に優位であることが確認された．さらに，実際にニュースサイトから採取したデータに基づいて本手法がどの程度 NCI のモデル化に成功しているのか検証する．

Mapping Media Agenda with WikiLabel

TADASHI NOMOTO[†1]

In this paper, we are interested in exploring how to automate PEW's NCI, a measure of how much coverage the media have given to a particular topic, as it would enable an objective comparison of media outlets − print or broadcast − on their coverage of news topics. We pursue a particular approach we call WikiLabel, which works by detecting topical clusters in news articles and labeling them with the title of a Wikipedia page that closely matches that cluster. The approach has an obvious advantage over past approaches in that it is able to produce fluent labels, as they are hand-written by human editors. We carried out some experiments on the TDT5 dataset, which found that the approach works rather robustly for an arbitrary set of documents in the news domain. Comparisons were made with some baselines, including the state of the art, with results strongly in favor of our approach.

†1 国文学研究資料館
National Institute of Japanese Literature

1. はじめに

News Coverage Index (NCI) は，現在米国内でインターネット，地上放送，ケーブル TV を通じて発信されている報道情報をもとに，米シンクタンク PEJ (Pew Center for Excellence in Journalism) [⋆1] がコンテンツ分析の専門スタッフを動員して 1 週間単位で決定しているメディア注目度を表す指標であり，米主要メディアの動向を俯瞰する上で貴重な情報源となっている．災害，感染症，環境汚染，テロなど様々な問題がグローバル化している現在，国内だけではなく海外メディアとの比較が情報の信憑性，事態の深刻さを判断する上で重要であることは言うまでもない．NCI はメディアの動向を分かり易く示す指標ではあるものの，(1) 米メディアが対象で他国（英語以外）への適用が困難，そのためニュースの国際比較ができない，(2) 人手に依存するため集計に時間がかかり即時性がない，などの問題があるのも事実である．このような問題意識のもと，本研究では報道争点（メディアの注目トピック）の自動検出，それに基づく報道量の集計を行うことを目的にする．報道争点を機械的に抽出するという試みは，報告者の知る限り前例がない [⋆1]．

NCI は，文字の大きさ，文字数，報道記事の紙面上の配置，またメディアの影響力など様々なファクターを考慮して決定されている数値であるため，これ自体をモデル化するのは現実的ではない．本研究では，このため問題をやや単純化し，インターネットでアクセス可能なニュース報道から主要トピックの自動検出とトピックに関連した記事総数の計量により NCI 相当の指標を算出することとし，実際の NCI との整合性をもって指標の有効性を判断することにした．

報道争点の検出は，記事のクラスタリングと生成クラスタへのラベル付与という問題として捉え直すことができる．本研究では NCI でのトピックラベルとウィキペディアのページタイトルが類似していることに着目し，ウィキペディアを使ったラベル付与という，従来にはない新しい観点からアプローチすることにした．クラスタリングは K 平均法（K means）と呼ばれる標準的なクラスタリング法を採用した．クラスタのラベル付与は，近年データ・マイニングの分野で盛んに研究されている「トピックモデル」と密接に関連する．LDA, pLSA など代表的なトピックモデル[1),5)] では，通常コーパスに出現した単語の分布としてトピックを表現するため，その解釈が常に問題になってきた[2),8)]．

⋆1 http://www.journalism.org
⋆1 報道争点の問題は，これまでメディア報道の受け手側への効果という観点から、計量政治学、マスコミ研究を中心に「アジェンダ設定機能仮説」の文脈で議論されてきた[7),11)–14)]．

図 1　ウィキページ「アイスランド」

本研究は，基本的に文献[2),10)] に着想を得ているが，文献[2)] が，テキスト本文を操作してラベルを生成するのに対して，本アプローチはウィキペディア自体を操作してラベルを生成することを目標にする．一方，文献[10)] は，少数の生物系論文の分類をウィキペディアのオントロジーをそのまま用いて行うものであり，本稿の様にウィキペディアの積極的な再構成を目指すものではない．

2. アプローチ

前節でも述べたように，本稿は報道争点の検出を記事のクラスタリングと生成クラスタへのラベル付与という観点からアプローチする．クラスタリングは標準的な K means 法を利用することとし，以下ではウィキペディアを用いたクラスタへのラベル付与を中心に述べる．

図 1 は「アイスランド」のウィキペディア・ページである．ウィキペディアのページは，インフォボックス，目次，図表，脚註，文献，関連項目，カテゴリーやナビゲーション用リンクなど，本文の他にも様々な要素から構成されている．本手法では，まず，本文とセクションタイトルを（以下，ページ本文）を取り出し，テキスト構造に従って，ページ本文を切り取っていく．これは，ページを細分化し，ページの主題（メインタイトル）に関連したより細かい概念を形成することで，クラスタの意味解釈をより精密にすることを意図している．ページ本文は形式的には**図 2** のように主題目を根とする木構造を成していると考えることができる．

2.1 ウィキペディアの解体

ここで，各ノード（節点）はセクションを表していると考える．根を始点として，葉を含めたノードに至るパス（経路）を列挙すると**表 1** のようになる．

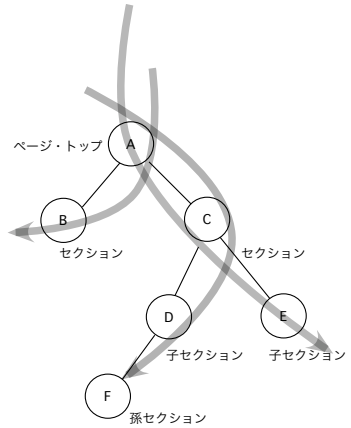


図 2　ウィキページの解体

表 1　パスの例

$p_1 = \langle \mathrm{A} \rangle$
$p_2 = \langle \mathrm{A\ B} \rangle$
$p_3 = \langle \mathrm{A\ C} \rangle$
$p_4 = \langle \mathrm{A\ C\ D} \rangle$
$p_5 = \langle \mathrm{A\ C\ E} \rangle$
$p_6 = \langle \mathrm{A\ C\ D\ F} \rangle$

次にパスに従って，ページを切り取る．例えば，p_2 からは，ページトップ（ルート）[⋆1] とセクション B，p_4 からはページトップとセクション C，D よりなるページが出来上がる．このように作られたページをミニページと呼び，その見出しを当該ノードからページトップまでの経路に現れたノードの見出しの結合とする．（なお，以下では便宜的にセクションのノードをその見出し（タイトル）で言及することにする．）

形式的には，ミニページは以下のように構成する．$p_1, p_2, ..., p_6$ はパスを表す．

$$\mathcal{D}(p_1) = \mathcal{C}^+(A)$$
$$\mathcal{D}(p_2) = concat(Text(A), \mathcal{C}^+(B))$$
$$\mathcal{D}(p_3) = concat(Text(A), \mathcal{C}^+(C))$$
$$\mathcal{D}(p_4) = concat(Text(A), Text(C), \mathcal{C}^+(D))$$
$$\mathcal{D}(p_5) = concat(Text(A), Text(C), \mathcal{C}^+(E))$$
$$\mathcal{D}(p_6) = concat(Text(A), Text(C), Text(D), \mathcal{C}^+(F))$$

ここで，C^+ を

$$\mathcal{C}^+(n) = \bigcup_{i \in D(n)} \{\mathcal{C}^+(i), Text(n)\}$$

⋆1 図 1 では，メインタイトル「アイスランド」とその定義箇所に当たる．

表 2 仮想ウィキページ

アイスランド

アイスランド共和国（アイスランドきょうわこく），通称アイスランドは，北ヨーロッパの北大西洋上に位置する共和制国家．

1 政治

「もっとも汚職が少ない国」と言われている．

2 軍事

NATO の原加盟国であるが自国軍は所持しておらず，世界でも希少な「常備軍を持たない国」である．

と定義する．n はノードを表し，$D(n)$ は n に直接支配された子ノードとする．$Text(n)$ は n に付随するテキスト，つまり，n の見出しに直接支配されたテキスト部を表す．また，$concat(x, y, z)$ は文字列 x, y, z の連結とする．

また，ミニページのタイトルを以下のように構成する．

$$\mathcal{T}(p_1) = H(A)$$
$$\mathcal{T}(p_2) = concat(H(A), H(B))$$
$$\mathcal{T}(p_3) = concat(H(A), H(C))$$
$$\mathcal{T}(p_4) = concat(H(A), H(C), H(D))$$
$$\mathcal{T}(p_5) = concat(H(A), H(C), H(E))$$
$$\mathcal{T}(p_6) = concat(H(A), H(C), H(D), H(F))$$

ここで $H(n)$ はノード n に付随するセクションタイトルとする．パス p_1 から生成されるミニページのタイトルは，$\mathcal{T}(p_1)$，その本文は $D(p_1)$ となる．本稿では，このような手順で生成されたページから成るウィキペディアを「再生ウィキ」と呼ぶ．例えば，**表 2** をウィキペディアの原ページであるとすると，上述の手順で**表 3**, **表 4** のミニページが生成される．ミニページでは見出しが「アイスランド≫政治」「アイスランド≫軍事」のように拡張されていることに注意されたい．(見出しの連結を “≫”で明示している．)

2.2 再生ウィキ

ウィキペディアから生成される再生ウィキは 1 つとは限らない．パスの深さをコントロールすることによって，様々な粒度を持ったウィキペディアを構成することが可能である．例えば，**図 4** は，図 2 のページからパスの長さを頂点から 1 に制限して再生成したページである．トピック「A」に関して,「B」と「C」というテーマ別に分割された「A」のページが出来上がる．**図 5** はパスの長さを無限大にした場合である．この場合，テーマの粒度を最も細

表 3 ミニページ (1)

アイスランド≫政治

アイスランド共和国（アイスランドきょうわこく），通称アイスランドは，北ヨーロッパの北大西洋上に位置する共和制国家．

「もっとも汚職が少ない国」と言われている．

表 4 ミニページ (2)

アイスランド≫軍事

アイスランド共和国（アイスランドきょうわこく），通称アイスランドは，北ヨーロッパの北大西洋上に位置する共和制国家．

NATO の原加盟国であるが自国軍は所持しておらず，世界でも希少な「常備軍を持たない国」である．

図 3 0 階ウィキページ

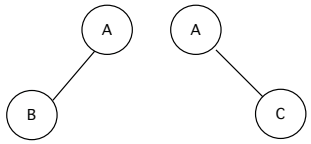


図 4 1 階ウィキページ

かくして「A」を分割したことになる．$\mathcal{D}$ の構成法から分かるようにミニページでは，頂点までのパス上のノードに付随するテキストをすべて含む．このため，ミニページはトピック「A」の記述をある観点からの詳細化していることに他ならない．例えば，表 3 はアイスランドの政治に注目したページ，表 4 はアイスランドの軍事に注目したページと解釈できる．

以下では，深度 1 のパスから生成されたウィキペディアを 1 階再生ウィキ，無限大のパスから構成されたウィキペディアを無限階再生ウィキと呼ぶ．図 4 では，1 階再生ウィキは p_2, p_3，無限階再生ウィキは p_2, p_5, p_6 から生成される．また，頂点ノード p_1 のみから再生されるウィキを 0 階再生ウィキと呼ぶ．これは，オリジナルのウィキペディアと本文上同一になる（**図 3**）．

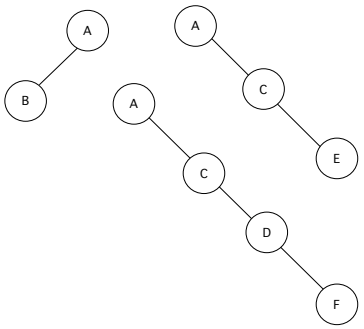


図 5 無限階ウィキページ

2.3 サーチによるラベル生成

ラベルの生成は，以下のステップで行う．まず，適当なクラスタリング法によって，記事集合をいくつかのクラスタに分割する．(実験では，記事を単語の重み (TFIDF) ベクトルに変換後，K means 法を用いてクラスタリングを行った．) 得られたそれぞれのクラスタについて，クラスタに出現している単語のうち，重要度の高いもの上から N 個を選択する．選択した単語をクエリとして，再生ウィキを検索する[⋆1]．検索された上位 K ページの見出しの列を当該クラスタのラベルとして採用する．この手法の大きなメリットは，人手によって作成された見出しをそのまま利用することによって，従来法では生成が困難であった可読性の高い，自然なラベルが得られるという点である．クラスタ中の単語重要度は TFIDF を採用した．なお，クラスタリングはオープンソースの Apache/Mahout[⋆2] を用いて行った．

ここで，ラベル生成モデルを以下で定義する．

$$l^*_\theta = \underset{l:p[l]\in\mathcal{W}^M}{\arg\max}\ Score(p[l], \theta|_N) \tag{1}$$

但し，$p[l]$ は，見出しが l であるミニページ，特に $T(p) = l$ とする (表 2 参照)．$\theta|N$ は上位 N 個の重要単語に要素を制限したクラスタ，$\mathcal{W}^M$ は，M 階再生ウィキを表す．また，関数 $Score$ を以下で与える．

$$Score(p[l], \theta|_N) = \lambda \mathsf{COS}(p[l], \theta|_N) + (1-\lambda)\mathsf{OVL}(l, \theta|_N) \tag{2}$$

⋆1 本研究では，Apache/Lucy を用いて再生ウィキのインデキシングと検索を行った (http://incubator.apache.org/lucy/)．
⋆2 http://mahout.apache.org

ここで，COS, OVL を以下で定義する．

$$\mathsf{COS}(\mathbf{r}, \mathbf{q}) = \frac{\langle \mathbf{r}, \mathbf{q} \rangle}{\sqrt{\langle \mathbf{r}, \mathbf{r} \rangle}\sqrt{\langle \mathbf{q}, \mathbf{q} \rangle}} \tag{3}$$

$$\mathsf{OVL}(l, \mathbf{v}) = \frac{\sum_{i\in\mathcal{U}(l)} \sum_{j\in\mathbf{v}} w_j \cdot \delta(i, j)}{\sum_{j\in\mathbf{v}} w_j} \tag{4}$$

$\mathsf{COS}(\mathbf{r}, \mathbf{q})$ は，ベクトル $\mathbf{r}, \mathbf{q}$ の規格化内積，$\mathsf{OVL}(l, \mathbf{v})$ は，l（見出し）とベクトル $\mathbf{v}$（クラスタ）の文字列上の類似度を表す．$U(l)$ は l 中に現れるユニークな単語の集合を表す．OVL は l と $\mathbf{v}$ との間で強い類似性を要求することで意味的ドリフトを抑制することを企図している．また，δ はクロネッカーのデルタ関数である．λ は COS と OVL の重みを配分するための係数で，経験的に設定する[⋆3]．

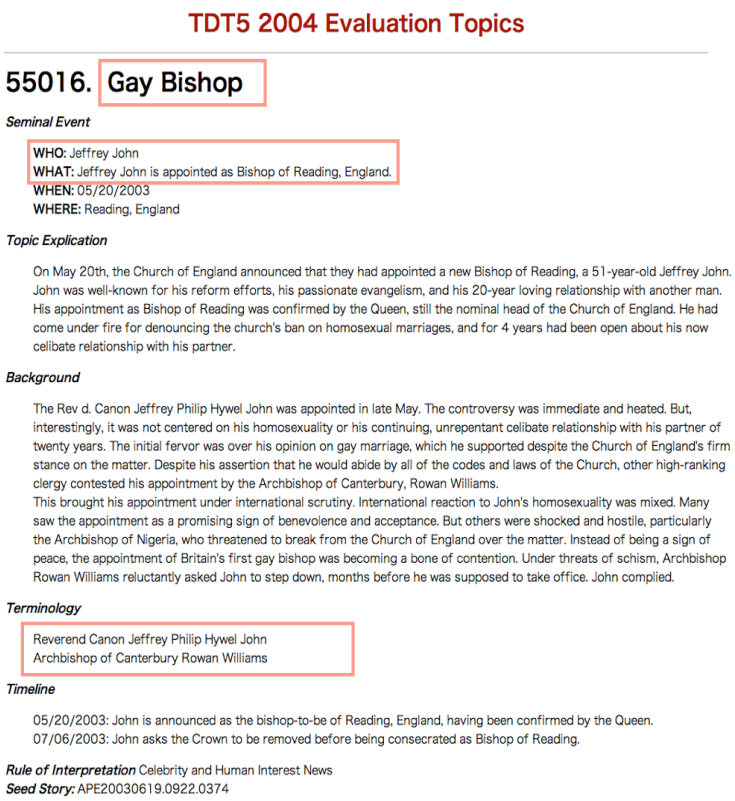
3. 結　　果

以下では，ラベル付与の精度を TDT-5 と呼ばれるデータセットを用いて評価する．TDT-5 は，米国標準技術研究所 (National Institute of Standards and Technology) が，米国内大学と共同で作成したオンラインニュースのトピック検出・追跡技術のベンチマークデータである．TDT-5 はラベル付与の評価が本来の目的ではないが，ここではデータを再構成して，ラベル付与のベンチマークとして利用することを考える．

TDT-5 は，北京語，アラビア語，英語の約 40 万のニュース記事と，記事のトピック同一性などについてアノテーションを施したデータより成る．250 のトピックについて記事がすべて人手でクラスタリングされている．それぞれのトピックについて，概略と背景，キーワード，タイトルなどがトピック・プロファイルという形で提供されている[4]（**図 6** 参照）．本研究では TDT-5 の英語記事に付されたトピック 100 件について，プロファイルに記されている，タイトル，登場人物，キーワードをクラスタのラベルと見なし，それらをいかに正確に復元できるかというシナリオで評価することにした．トピック・クラスタの規模，関連文書数については**表 5** を参照されたい．

以下では，このようなラベルを擬似ラベルと呼ぶことにする．図 6 の赤枠で囲った部分，つまり，(1) Gay Bishop, (2) Jeffrey John, (3) Jeffrey John is appointed as Bishop of Reading, England, (4) Reverend Canon Jeffrey Philip Hywel John, (5) Archbishop of Canterbury Rowan

⋆3 実験では 0.5.



TDT5 2004 Evaluation Topics

55016. Gay Bishop

Seminal Event

WHO: Jeffrey John
WHAT: Jeffrey John is appointed as Bishop of Reading, England.
WHEN: 05/20/2003
WHERE: Reading, England

Topic Explication

On May 20th, the Church of England announced that they had appointed a new Bishop of Reading, a 51-year-old Jeffrey John. John was well-known for his reform efforts, his passionate evangelism, and his 20-year loving relationship with another man. His appointment as Bishop of Reading was confirmed by the Queen, still the nominal head of the Church of England. He had come under fire for denouncing the church's ban on homosexual marriages, and for 4 years had been open about his now celibate relationship with his partner.

Background

The Rev d. Canon Jeffrey Philip Hywel John was appointed in late May. The controversy was immediate and heated. But, interestingly, it was not centered on his homosexuality or his continuing, unrepentant celibate relationship with his partner of twenty years. The initial fervor was over his opinion on gay marriage, which he supported despite the Church of England's firm stance on the matter. Despite his assertion that he would abide by all of the codes and laws of the Church, other high-ranking clergy contested his appointment by the Archbishop of Canterbury, Rowan Williams.
This brought his appointment under international scrutiny. International reaction to John's homosexuality was mixed. Many saw the appointment as a promising sign of benevolence and acceptance. But others were shocked and hostile, particularly the Archbishop of Nigeria, who threatened to break from the Church of England over the matter. Instead of being a sign of peace, the appointment of Britain's first gay bishop was becoming a bone of contention. Under threats of schism, Archbishop Rowan Williams reluctantly asked John to step down, months before he was supposed to take office. John complied.

Terminology

Reverend Canon Jeffrey Philip Hywel John
Archbishop of Canterbury Rowan Williams

Timeline

05/20/2003: John is announced as the bishop-to-be of Reading, England, having been confirmed by the Queen.
07/06/2003: John asks the Crown to be removed before being consecrated as Bishop of Reading.

Rule of Interpretation Celebrity and Human Interest News
Seed Story: APE20030619.0922.0374

図 6 TDT5 トピック・プロファイル．枠内が擬似ラベル．

Williams が擬似ラベルを表す．

復元の精度の計測には，ROUGE-W と呼ばれる尺度を利用する[6]．これは，2 つの任意の文字列について，その類似度を「最長共通部分列」(LCS, Longest Common Subsequence) という概念で定義するものである．例えば，(1) ABCDEF と (2) BDEF, (3) DBFG を比較すると (1-2) の最長部分列は BDEF であるが，(1-3) の最長部分列は D あるいは B である．よって，(2) が (3) より (1) に近いということになる．ROUGE-W は 0（完全不一致）から 1（完全一致）の範囲の実数値を取る．具体的には，$\mathcal{R}(\mathcal{C}|_k, \mathcal{L}) = \max_{i \in \mathcal{C}|_k} \mathsf{ROUGE}(i, \mathcal{L})$ と指標とする．ただし，$\mathcal{C}|_k$ は，$Score$ 値においてトップ k 個の候補ラベル．$\mathcal{L}$ は擬似ラベルの

表 5 TDT5 トピック・クラスタの概要

トピック数	トピック文書数
100	4,501

トピック・クラスタ		
最小文書数	最大文書数	平均文書数
3	244	45.01

表 6 再生ウィキの規模

ウィキペディア（英語）4 月 6 日版	
階層	ページ数
$\mathcal{W}^0$	3,914,761
$\mathcal{W}^1$	5,674,922
$\mathcal{W}^\infty$	7,583,458

表 7 ロング・ラベルとショート・ラベル

	1 次タイトル　2 次タイトル
T1	アイスランド ≫経済
T2	アイスランド ≫経済 ≫国家経済
T3	アイスランド ≫経済 ≫国家経済 ≫金融危機以前

集合．さらに，$\mathsf{ROUGE}(i, \mathcal{L}) = \max_{j \in \mathcal{L}} \mathsf{ROUGE\text{-}W}(i, j)$ とする．$\mathsf{ROUGE\text{-}W}(i, j)$ は文字列 i, j の最長共通部分列の正規化長を表す[6]．従って，$\mathcal{R}$ は，候補ラベルと擬似ラベルを組み合わせて得られる最も高い ROUGE-W スコアを返すことになる．

表 6 に実験に用いた再生ウィキのページ総数を示す．表中の $\mathcal{W}^0$ は，0 階，$\mathcal{W}^1$ は 1 階，$\mathcal{W}^\infty$ は，無限階の再生ウィキである．無限階は 0 階に比べ約 2 倍の規模になっている．ウィキペディアは 2011 年 4 月 6 日付英語版ウィキダンプを用いた．なお，再生ウィキへの検索については各トピッククラスタの単語を TFIDF で重み付けした上で，上位 50 語を検索キーワードとして用いた．

ここで，ミニページのショート・ラベルとロング・ラベルについて説明する．ショートラベルとはミニページの 1 次タイトルを使って生成したラベルを意味する．これに対してミニページに付与されている完全なタイトル，つまり，$\mathcal{T}$ をそのまま用いて生成したラベルをロング・ラベルと呼ぶ．**表 7** の例では，T1 のショート・タイトルが「アイスランド」，ロング・タイトルが「アイスランド≫経済」，T3 ではショート・タイトルが「アイスランド」，ロング・タイトルが「アイスランド≫経済 ≫国家経済 ≫金融危機以前」となる．但し，

表 8 ショート・ラベルを用いた結果（0 階再生ウィキ）

k	TFIDF	WikiLabel	T-score	Mei et al. (2007)
1	0.1240	**0.3103**	0.1485	0.0761
3	0.1858	**0.4456**	0.2155	0.1063
5	0.1999	**0.4860**	0.2320	0.1153

表 9 ショート・ラベルを用いた結果（1 階再生ウィキ）

k	TFIDF	WikiLabel	T-score	Mei et al. (2007)
1	0.1520	**0.3757**	0.1683	0.0971
3	0.1756	**0.4515**	0.2084	0.1119
5	0.1879	**0.4789**	0.2212	0.1194

表 10 ショート・ラベルを用いた結果（無限階再生ウィキ）

k	TFIDF	WikiLabel	T-score	Mei et al. (2007)
1	0.1474	**0.3325**	0.1694	0.0907
3	0.1691	**0.4148**	0.2037	0.1074
5	0.1846	**0.4501**	0.2195	0.1188

表 11 ロング・ラベルを用いた結果（1 階再生ウィキ）

k	TFIDF	WikiLabel	T-score	Mei et al. (2007)
1	0.1035	**0.2041**	0.1569	0.0704
3	0.1290	**0.2620**	0.2098	0.1012
5	0.1411	**0.2720**	0.2163	0.1089

ロング・タイトルの長さは，ウィキペディアの解体深度に依存することに注意されたい．

表 8 に 0 階再生ウィキでの結果を示す．サーチ結果の上位 k タイトルをラベルとして採用している．表中の WikiLabel（ウィキラベル）は提案手法， T-score[3], Mei et al[8] は従来法である．また，タームを TFIDF でソートした上位 k 個をラベル候補とするベースライン (TFIDF) の結果も載せた．表中の数字は 100 トピックで平均した $\mathcal{R}$ を示している．k に依らず，提案手法が優位であることが分かる．なお，0 階再生ウィキでは，ショート・ラベルとロング・ラベルが一致する．

表 9 に 1 階再生ウィキを用いたショート・ラベルの結果を示す．表 8 と同様，他の手法に比べウィキラベルが優位であるが，注目すべきは，精度が 0 階再生ウィキより優れている点である．これは，ウィキペディアの解体が精度に貢献していることに他ならない．**表 10** は無限階再生ウィキを使ったショート・ラベルの結果である．精度が 1 階再生ウィキに比べ劣化している．検索精度の低下が原因と考えられるが，これはウィキペディアの過度な分割によって関連性判定に悪影響が生じたものと推測される．

表 12 ロング・ラベルを用いた結果（無限階再生ウィキ）

k	TFIDF	WikiLabel	T-score	Mei et al. (2007)
1	0.0817	0.1409	**0.1470**	0.0572
3	0.1026	**0.1821**	0.1706	0.0843
5	0.1186	**0.2026**	0.1930	0.0969

次にロング・ラベルの結果を見ていく．ロング・ラベルとは，ミニページに付与されたタイトルをすべて使って，ラベルを生成する手法である．

表 11 は 1 階再生ウィキによるロング・ラベル生成の結果である．提案手法のウィキラベルは他の手法に比べて良好であるが，ショートと比較すると精度が低下していることが分かる．

一方，**表 12** は，ラベルを最大限に拡張した場合の結果を表している．1 階再生ウィキのロング・ラベルの結果より精度がさらに劣化しており，$k = 1$ で T-score のほうが優勢になっている．これはラベルが長くなるにつれ無駄な部分が増えることを示唆している．

以上，TDT-5 をベンチマークとして本手法を評価した．ラベルを長くすると精度の落ち込みが目立つようになるが，従来手法に比べて，顕著に優位であることが確認された．以下では，ウィキラベルを使って，日米注目トピックの報道量とその推移を実際に見ていくことにする．

4. 報道争点のマッピング

図 7 は，東日本大震災が発生した翌週 2011 年 3 月 14 日から同年 3 月 20 日までの 1 週間に日米の主要メディア 25 社のインターネットサイト（**表 13**）から採取した報道記事約 15,000 件をもとにウィキラベルが自動生成した上位 5 トピックとその量である[⋆1]．なお，米メディアについては，英語ウィキペディアから 1 階再生ウィキを，国内メディアは日本語ウィキペディアから無限階再生ウィキを形成し報道量の計測を行った．横軸は報道件数を示す．クラスタリングを 20 回実行し生成した 400 個のクラスタをトピック（ロング・ラベル）別に集約，そのうち報道件数の多いもの上位 5 件を提示している．(ここでは分かり易さ

⋆1 米メディアについては, PEJ の調査方法 (http://www.journalism.org/about_news_index/methodology) を参考に選択した．国内メディアについては，主要地方紙，全国紙，関東キー局を中心にアクセス可能なサイトを選んだ．

表 13 記事収集対象メディア

米メディア	The New York Times, Yahoo.com, CNN, MSNBC, Fox, USATODAY, Washington Post, ABC, NBC, BBC, Reuters
国内メディア	朝日新聞，中日新聞，北海道新聞，JCAST，時事コム，河北新報，毎日新聞，読売新聞, 日本経済新聞，産経新聞，東京新聞，NHK, TBS ニュース， TV 朝日

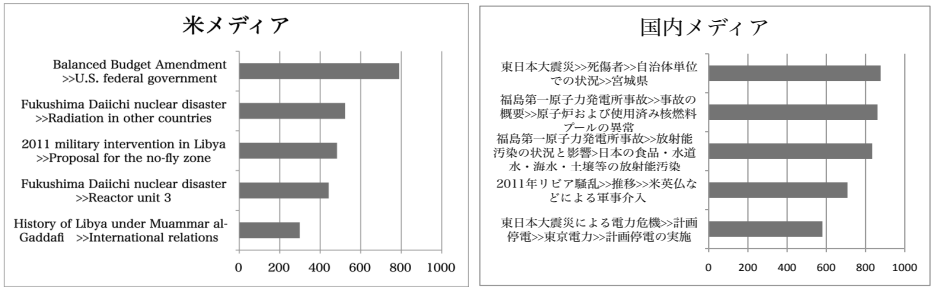


図 7 ウィキラベルよる 2011 年 3 月 14 日〜20 日の日米トップニュース

のためロング・ラベルを採用している.）また，再生ウィキの検索は，クラスタ内の単語を TFIDF で重み付けした上で，英語クラスタについては上位 50 語，日本語クラスタについては上位 100 語を検索キーとして用いた.

図 7 を見ると，この週，福島原発関連トピックが米メディアを圧倒していたことが分かる．一方，国内メディアは，大震災関連報道と並んで福島原発事故に強い関心を示していることが分かる.

なお，結果の妥当性であるが，国内についてはレファレンスになり得る指標が存在しないため検証できないが，米メディアについては，NCI 値を参考にすることができる．**図 8** が同時期の NCI である.

この週のトップは東北震災関連，2 位は中東の騒乱となっている．PEJ の解説によると震災ニュースの 3 分の 2 は原発事故関連であり，中東騒乱のメイン・トピックはリビアに対する国連の飛行禁止区域の設定との解説がある[⋆1]．従って，NCI の 1 位と 2 位はウィキラベルの米メディアトップ項目の 2 位と 3 位に対応していると考えられる.

次に，福島原発事故，特に放射能汚染関連の報道が日米でどのように推移したかをウィキ

⋆1 http://www.journalism.org/index_report /pej_news_coverage_index_march_1420_2011

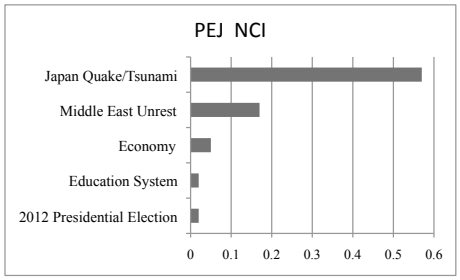


図 8 2011 年 3 月 14 日〜20 日の NCI

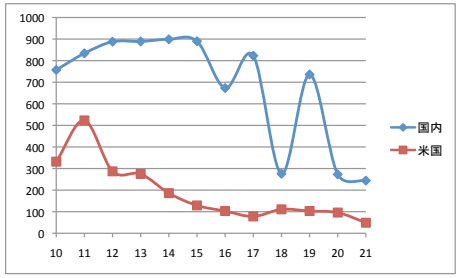


図 9 福島原発放射能汚染報道の推移

ラベルを使って見てみることにする．図 9 は，2011 年の第 10 週目から 21 週目まで（3 月 7 日〜5 月 29 日）の日米の放射能汚染関連の報道量の動きである．但し，国内報道については「福島第一原子力発電所事故>>放射能汚染の状況と影響」，米の報道については，‘Fukushima Daiichi nuclear disaster’ をラベルに持つクラスタを選び，その大きさを持って放射能汚染関連の報道量とした．トピック解析は週単位で行った．日本語ウィキペディアは，2011 年 5 月 22 日版（無限階再生ウィキ），英語ウィキペディアは，同年 5 月 27 日版（1 階再生ウィキ）を用いた．再生後のページ総数は日本語が 2,920,403，英語が 5,798,483 となった.

青い線が国内メディアでの報道量推移，赤い線が米メディアの推移である．米メディアでは，福島原発事故が起こった翌週（11 週目）に事故関連報道がピークに達してその後急激に減少している．この傾向は NCI の結果とも合致している．PEJ によれば，NATO 軍のリビア介入を契機に米メディアの関心が大きくシフトしたことが原因とされる.

一方，国内メディアは 15 週目（4 月 11 日〜17 日）にようやくピークを迎え，その後振動しながら退潮している．18 週目（ゴールデンウィーク）に極端に報道量が落ちているこ

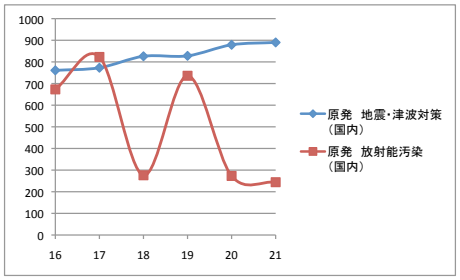


図 10 福島原発観点別報道の推移

とが調味深い．ところで，海外メディアに関して，3 月末から 4 月初めにかけて国内でその過熱報道ぶりが再三批判されたが，国内メディアでも放射能汚染に関する報道が極めて多かったことを示唆する結果となった．

図 10 は，国内メディアにおいて放射能汚染に対する報道が減少していく中，福島原発の地震・津波対策問題への関心が 16 週以降でも高いことを示している．同じトピックでも観点によって動きに違いが見られる．青のラインが「福島第一原子力発電所事故≫地震と津波の対策」赤のラインが「福島第一原子力発電所事故≫放射能汚染の状況と影響」をラベルに持つクラスタの推移を表している．

5. 結　　論

以上，ウィキラベルによる報道争点の検出について，その手法の詳細と実際のデータへの応用例を示した．英語についてのみであるが TDT-5 コーパスを用いて従来法との客観的な比較を行い，その優位性を確認した．本研究のメインテーマである，米 PEJ の NCI の自動化については直接的なモデリングが不可能であるため，独自の計測モデルを設計し，NCI との整合性を実例に基づいて検証した．本手法は，ニュースドメインにおいて，クラスタの内容了解性が従来法に比べ概ね優れていると判断されるが，課題も明らかになっている．

その 1 つはラベル生成をすべてウィキペディアに委ねているため，ウィキペディアに存在しないラベルを生成することができない．また，イベントの粒度とウィキペディアの項目の粒度がずれることがある．例えば，ウィキペディアは鉄道会社それぞれの運行についての記述はあるものの，地震による鉄道の運休という概念は存在しない．このため，適切なラベルを生成できない．(因にこのような問題の対処法として，文献[9] のようにオンデマンドでウェブからページ概念を作り出す方法が考えられる．) 一方，実装レベルの問題としては，再生ウィキが巨大になるにつれ検索効率が著しく低下するという点も挙げられる．さらに大きな課題として，国内メディアについてウィキラベルの妥当性が確認できていないという問題も残っている．今後はこれらの課題について検討を進めるとともに，本手法の中国，韓国への拡張も視野に入れていく予定である．

謝辞　本研究は，(財) 放送文化基金（平成 21 年度研究助成）の支援を受けて行った．

参　考　文　献

1) Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
2) Carmel, D., Roitman, H. and Zwerding, N.: Enhancing Cluster Labeling Using Wikipedia, *Proceedings of SIGIR'09*, pp.139–146 (2009).
3) Church, K., Gale, W., Hanks, P. and Hindle, D.: Using Statistics in Lexical Analysis, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon* (Zernik, U., ed.), Lawrence Erlbaum Associates, Hillsdale, NJ (1991).
4) Glenn, M., Strassel, S., Kong, J. and Maeda, K.: TDT5 Topics and Annotations (2006). Linguistic Data Consortium, Philadelphia.
5) Hofmann, T.: Probabilistic Latent Semantic Analysis, *Proceedings of Uncertainity in Artificial Intelligence (UAI'99)* (1999).
6) Lin, C.-Y.: ROUGE: a Package for Automatic Evaluation of Summaries, *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)* (2004).
7) McCombs, M.E. and Shaw, D.L.: The Agenda-Setting Function of Mass Media, *Public Opinion Quarterly*, Vol.36, No.2, pp.176–187 (1972).
8) Mei, Q., Shen, X. and Zhai, C.: Automatic Labeling of Multinomial Topic Models, *Proceedings of KDD'07*, pp.490–499 (2007).
9) Sauper, C. and Barzilay, R.: Automatically Generating Wikipedia Articles: A Structure-Aware Approach, *Proceedings of the 47st Annual Meeting of the Association for Computational Linguistics and the 4th IJCNLP of the AFNLP*, pp.208–216.
10) Syed, Z.S., Finin, T. and Joshi, A.: Wikipedia as an Ontology for Describing Documents, *Proceedings of the Second International Conference on Weblogs and Social Media*, AAAI Press, pp.136–144 (2008).
11) 李　光鎬：ふたつの「北朝鮮」，慶應義塾大学メディア・コミュニケーション研究所紀要，pp.59–71 (2006).
12) 竹下俊朗：メディアの議題設定機能，学文社 (2008).
13) 岡田直之：マスコミ研究の視座と課題，東京大学出版会 (1992).
14) 小林良彰（編）：政治過程の計量分析 (RFP 叢書)，芦書房 (1991).