DEIM Forum 2013 C2-6

検索エンジン API を用いたウェブページ収集におけるトピックの多様性

小池　大地† 　牧田　健作† 　宇津呂武仁†† 　河田　容英††† 　吉岡　真治††††

神門　典子†††††

† 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) ログワークス　〒 151-0053 東京都渋谷区代々木 1-3-15 天翔代々木ビル 6F
†††† 北海道大学大学院 情報科学研究科　〒 060-0808 北海道札幌市北区北 8 条西 5 丁目
††††† 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2

あらまし　本論文では，検索エンジン API を用いたウェブページ収集タスクにおいて，トピックモデルを用いたトピックの分布の観点から，できる限り多様なウェブページを収集する方式を提案する．提案方式においては，まず，特定のクエリに関連するニュース・ブログを収集した文書集合を対象としてトピックモデルを適用し，トピックの分布を求める．そして，各トピックとの関連が強いクエリを用いることにより，ニュース・ブログにおけるトピックとの関連が強いウェブページを収集する．提案方式により収集されたウェブページ群におけるトピックの分布を，ニュース・ブログにおけるトピック分布を考慮せず，特定のクエリのみを検索エンジン API に入力して収集されるウェブページ群におけるトピック分布と比較し，提案方式によって収集されたウェブページ群において，より多様性に富んだトピックの分布が観測できることを示す．

キーワード　ウェブページ収集, トピック, 多様性, 検索エンジン API, トピックモデル

Diversity of Topics in Collecting Web Pages using a Search Engine API

Daichi KOIKE†, Kensaku MAKITA†, Takehito UTSURO††, Yasuhide KAWADA†††, Masaharu YOSHIOKA††††, and Noriko KANDO†††††

† 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
††† Logworks Co., Ltd.　Tokyo 151-0053, Japan
†††† Graduate School of Information Science and Technology, Hokkaido University, Sapporo, 060-0808, Japan
††††† National Institute of Informatics, Tokyo 101-8430, Japan

Abstract　This paper proposes how to collect Web pages of diverse topics using a search engine API, where a topic model is employed to estimate distribution of topics. In the proposed framework, given a certain query, relevant news articles and blog posts are collected and the topic distribution within the collected document set is estimated by a topic model. Next, from each topic within the document set consisting of news and blogs, a few query terms for the search engine API are semi-automatically selected and are utilized in the procedure of collecting Web pages closely relevant to topics observed only in news or blogs. Experimental evaluation results show that the proposed procedure of collecting Web pages through queries collected from news and blogs achieve much more diverse distribution of topics compared with the baseline procedure of collecting Web pages through the search engine API without considering topics in news and blogs.

Key words　Web page collection, topic, diversity, search engine API, topic model

1. はじめに

本論文では，検索エンジン API を用いたウェブページ収集タスクにおいて，トピックモデルを用いたトピックの分布の観点から，できる限り多様なウェブページを収集する方式を提案する．提案方式においては，まず，特定のクエリに関連する

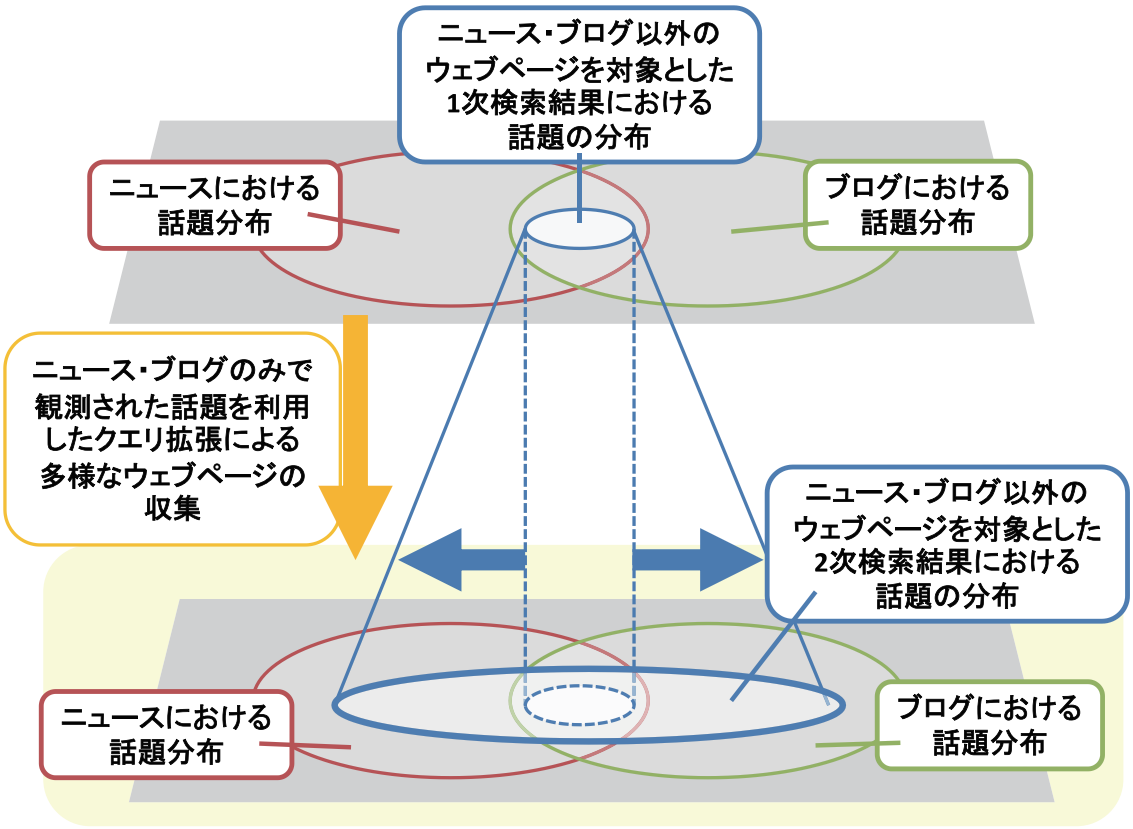


図 1　ニュース・ブログのみで観測された話題を利用したクエリ拡張による多様なウェブページの収集

ニュース・ブログを収集した文書集合を対象としてトピックモデルを適用し，トピックの分布を求める．そして，同一のクエリを対象としてニュース・ブログ以外のウェブページから収集した文書集合 (この文書集合を，ウェブページの 1 次検索結果と呼ぶ) を対象として，同様にトピックモデルを適用し，両者のトピック分布を比較する．その結果，図 1 上半分のように，ニュース・ブログにおける話題の分布と比較して，検索エンジン API を用いることによりニュース・ブログ以外のウェブページから収集された文書集合における話題の分布が相対的に小さいことを示す．

次に，提案手法においては，ニュースもしくはブログのみで観測された話題との関連が強いクエリを用いることにより，ニュース・ブログのみにおいて観測された話題との関連が強いウェブページを収集する (ここで収集された文書集合を，ウェブページの 2 次検索結果と呼ぶ)．そして，2 次検索により収集したウェブページ群に対してトピックモデルを適用することによって推定した話題の分布を，1 次検索結果における話題の分布と比較して，提案方式によって収集されたウェブページ群において，より多様性に富んだ話題の分布が観測できることを示す．特に，1 次検索結果において，ニュースあるいはブログのみにおいて観測された話題のウェブページだけでなく，ニュースあるいブログのいずれにおいても観測されなかった多様な話題のウェブページもあわせて収集可能であることを示す．

表 1　収集したウェブページ，ニュース，ブログの文書数

ウェブページ	ニュース	ブログ
1 次検索: 593	26,228 (朝日:7,541, 読売:6,568, 日経:12,119)	20,716
2 次検索: 20,374		

2. 文書収集の手順

本論文で分析対象としたウェブページ，ニュース記事，ブログ記事の収集手順を以下に述べる．本論文で分析対象とした文書数の一覧を表 1 に示す．

2.1 ウェブページの収集

ウェブページの収集においては，Yahoo! Search BOSS API(注1) を用い，検索エンジン API に対してクエリを指定することにより，日本語のサイトを対象として収集を行った．まず，初期クエリ t_0 を「東日本大震災」とし，一度に最大で 1,000 件のウェブページを取得した．その結果，720 件のウェブページが検索結果として得られた．その後，主要なニュース・ブログサイトのウェブページを除いたものを分析対象とした．その結果，1 次検索での分析対象のウェブページは，593 件となった．

次に，2 次検索として，4.2 節の手法により選定された語を 2 次クエリ t_1 とし，各 2 次クエリ，初期クエリ t_0 との AND

(注1)：http://developer.yahoo.com/search/boss/

検索によりウェブページの収集を行った．その結果，68,880 件が検索結果として得られた．その後，主要なニュース・ブログサイトのウェブページを除いた結果，2 次検索での分析対象のウェブページは，20,374 件となった．

2.2 ニュース記事の収集

ニュース記事としては，2011 年 3 月 11 日から 2012 年 7 月 10 日までの日付の記事を，日経新聞(注2)，朝日新聞(注3)，読売新聞(注4)の各新聞社のサイトから収集した 92,772 記事，63,906 記事，および，68,239 記事の合計 224,867 記事を用いた．その後,「東日本大震災」の 1 語がニュース記事中に出現するものだけを分析対象とした．その結果，各新聞社の記事数は，日経新聞が 12,119 記事，朝日新聞が 7,541 記事，読売新聞が 6,568 記事，合計 26,228 記事となった．

2.3 ブログ記事の収集

東日本大震災に関連するブログ記事の収集においては，東日本大震災との関連性が高い語として，人手で 26 個の語を選定し，その一つ一つを初期クエリ t_0 として，ブログ記事を収集した結果を用いた．初期クエリ t_0 を含む日本語ブログ記事の収集においては，Yahoo! Search BOSS API を利用し，日本語ブログホスト大手 6 社(注5)のドメインを対象として，2012 年 8 月下旬から 9 月上旬に，2011 年 3 月 11 日以降の日付の記事を対象として，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，ブログ記事集合を作成する．その後,「東日本大震災」の 1 語がブログ記事中に出現するものだけを分析対象とした．その結果，分析対象のブログ記事数は，20,716 記事となった．

3. トピックモデルを用いた話題分布の推定

3.1 トピックモデル

本研究では，トピックモデルとして潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [3] を用いる．LDA を用いたトピックモデルの推定においては，語 w の列によって表現された文書の集合と，トピック数 K を入力として，各トピック z_n $(n = 1, \ldots, K)$ における語 w の確率分布 $P(w|z_n)$ $(w \in V)$，及び，各文書 b におけるトピック z_n の確率分布 $P(z_n|b)$ $(n = 1, \ldots, K)$ を推定する．これらを推定するためのツールとしては，GibbsLDA++(注6)を用いた．LDA のハイパーパラメータである α，β には，GibbsLDA++の基本設定値である $\alpha = 50/K$，$\beta = 0.1$ を用いた．LDA ではトピック数 K を人手で与える必要があるが，本論文では，トピック数を 10 から 100 まで変化させてトピック推定を行い，得られたトピックを人手で見比べ，トピックの推定結果の性能がより高くなったトピック数を採用するという手順を採った．なお，このツールは推定の際に Gibbs サンプリングを用いているが，その反復回数は 2,000 とした．

表 3 トピックモデル推定時に指定したトピック数

ウェブページ	ニュース	ブログ
1 次検索: 15	70	60
2 次検索: 90		

3.2 文書に対するトピックの割り当て

本研究では，一つのニュース記事，あるいは，ブログ記事に対して，トピックを一意に割り当てる．文書集合を D，トピック数を K，1 つの文書を d $(d \in D)$ とすると，トピック z_n $(n = 1, \ldots, K)$ の記事集合 $D(z_n)$(ニュース記事・ブログ記事の和集合) は以下の式で表される．

$$D(z_n) = \left\{ d \in D \mid z_n = \underset{z_u\ (u=1,\ldots,K)}{\operatorname{argmax}} P(z_u|d) \right\}$$

これはつまり，文書 d におけるトピックの分布において，確率が最大のトピックに，文書 d を割り当てていることになる．

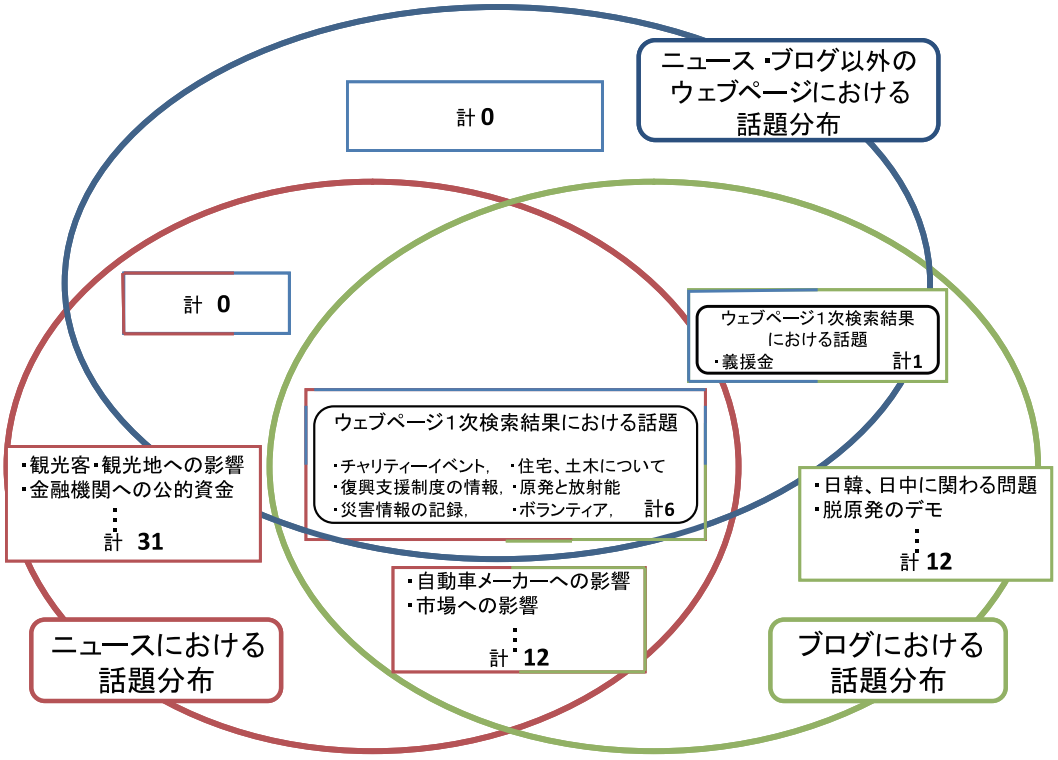
4. 多様なトピックのウェブページの収集

4.1 ニュースおよびブログを利用した多様なトピックの生成

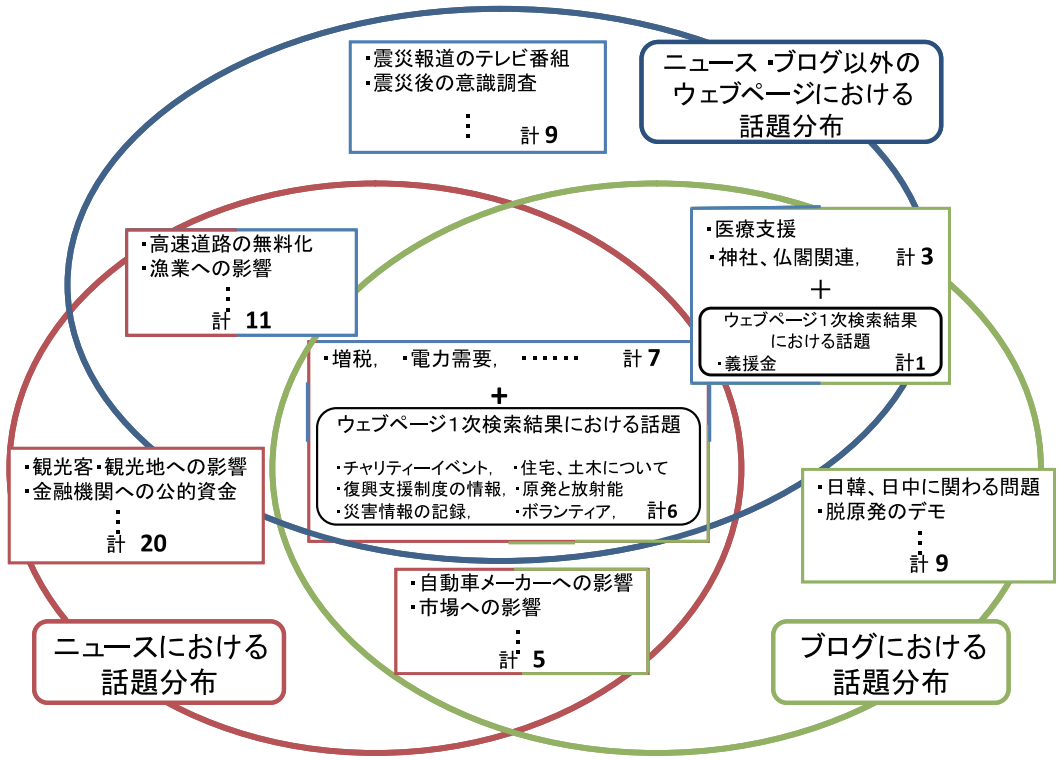
まず，2.1 節の 1 次検索の手順によって収集したウェブページ (ただし，ニュース・ブログを除く)，2.2 節の手順によって収集したニュース記事，および，2.3 節の手順によって収集したブログ記事に対して，それぞれ独立にトピックモデルの推定を行った．ただし，予備実験を経たうえで，それぞれ最も性能よくトピックモデルの推定が行えたトピック数として，表 3 に示すトピック数を用いた．このうち，ウェブページ集合から推定された 15 トピック，ニュース記事集合から推定された 70 トピック，ブログ記事集合から推定された 60 トピックのうち，東日本大震災に関連し，かつ，トピックに対応する文書集合において意味的まとまりのあるトピックの数は，それぞれ，8 トピック，67 トピック，40 トピックであった．次に，これらのトピックのうち，情報源となった文書がウェブページであるか，ニュース記事であるか，ブログ記事であるかの別を問わず，同一の話題と考えられるトピックの集約を行ったところ，合計 62 個の話題に集約された．この 62 個の話題が，ウェブページ，ニュース，ブログのそれぞれどの情報源において観測されたかの内訳を図 2(a) のベン図に示す．

この結果から分かるように，ニュース記事のみにおいて観測された話題が合計 31 話題，ブログ記事のみにおいて観測された話題が合計 12 話題，ニュース記事とブログ記事の両方で観測され，ニュース・ブログ以外のウェブページ (ただし，検索エンジン API の上位 1,000 位以内) では観測されなかった話題が合計 12 話題となり，全 62 話題のうちの約 90%を占めた．この結果から，ニュース・ブログ以外のウェブページを対象として，検索エンジン API の上位 1,000 位以内の範囲において，ニュー

(注2)：http://www.nikkei.com/
(注3)：http://www.asahi.com/
(注4)：http://www.yomiuri.co.jp/
(注5)：fc2.com, yahoo.co.jp, ameblo.jp, goo.ne.jp, livedoor.jp, hatena.ne.jp
(注6)：http://gibbslda.sourceforge.net/



(a) ウェブページが 1 次検索結果のみの場合の合計 62 話題



(b) ウェブページが 1 次検索結果 +2 次検索結果の場合の合計 71 話題

図 2　ウェブページ・ニュース・ブログにおける話題の分布

ス・ブログにおいて観測されるような多様な話題の文書を収集することが容易でないことが判明した.

4.2　各トピックに対応するウェブページの収集

次に，ニュース記事のみにおいて観測された 31 話題，ブログ記事のみにおいて観測された 12 話題，ニュース記事とブログ記事の両方で観測され，ニュース・ブログ以外のウェブページでは観測されなかった 12 話題に対応する各トピック z_n において，確率値 $P(w|z_n)$ が上位 20 位以内となる語 w のうち，各

表 2　ニュース・ブログのみで観測された話題を利用したクエリ拡張におけるクエリ

ニュースのみから選定したクエリ (計 49 個)	ブログのみから選定したクエリ (計 22 個)	ニュース・ブログ両方から生成したクエリ (計 15 個)
ビール，アンケート，エネルギー，デザイン，マンション，みずほ，ローン，営業利益，沿岸，学生，観光，給与，漁業，経営，経団連，原子炉，交渉，交付金，公演，公的資金，行方不明者，高速道路，国会，祭り，作品，市場，指紋，試合，事件，自動車，需要，就職，新幹線，申請，水準，政治，生徒，台風，大阪，地震保険，通信，統一地方選，避難所，百貨店，陛下，報道，貿易収支，防災，路線	韓国，イベント，フィギュアスケート，メディア，メーカー，医療，営業，雇用，子ども，小沢一郎，食品，脱原発，地震，中国，調査，天皇，東証，東電，日本人，被災者，放射線，神社	がれき，ボランティア，家族，義援金，経済，原発，自衛隊，住宅，世界，政府，選手，増税，津波，電力，福島県

トピック z_n における話題を適切に反映しており，初期クエリ t_0(=「東日本大震災」) との AND 検索を行う 2 次クエリ t_1 として適切であると考えられる語を 1 語，人手で選定し(注7)，検索エンジン API を用いたウェブページ検索により，各 2 次クエリごとに最大で 1,000 ページを取得した．2 次クエリの一覧を表 2 に示す．なお，2.1 節で述べたように，以上の 2 次検索の手順によって収集された分析対象ウェブページの総数は，20,374 件となった．

4.3 分　　析

次に，2 次検索によって収集されたウェブページを対象として，トピックモデルの推定を行った．ただし，トピック数をとしては，予備実験を経たうえで，最も性能よくトピックモデルの推定が行えたトピック数として，表 3 に示すトピック数 90 を用いた．このうち，東日本大震災に関連し，かつ，トピックに対応する文書集合において意味的まとまりのあるトピックの数は，50 トピックであった．さらに，前節において人手で集約した 62 個の話題に対して，同様に，50 トピックとの対応付けを行った結果，新たに 9 個の話題が追加され，合計 71 個の話題となった．この 71 個の話題が，ウェブページ，ニュース，ブログのそれぞれどの情報源において観測されたかの内訳を図 2(b) のベン図に示す．

このベン図から分かるように，新たに追加された 9 個の話題は，ニュースおよびブログにおいては観測されなかった，ウェブページ固有の話題である．また，図 2(a) と図 2(b) を比較すると分かるように，ウェブページの 1 次検索の結果において，ニュース記事のみにおいて観測された 31 話題，ブログ記事のみにおいて観測された 12 話題，ニュース記事とブログ記事の両方で観測され，ニュース・ブログ以外のウェブページでは観測されなかった 12 話題のうち，それぞれ，11 話題，3 話題，7 話題がウェブページの 2 次検索結果において観測可能となった．つまり，ウェブページ固有の話題である 9 個の話題とあわせると，合計 30 個の話題が 2 次検索によって新規に観測可能となったことが分かる．以上の結果から，検索エンジン API を用いた 1 次検索においては，ニュース・ブログ以外のウェブページからは収集困難であった多様な話題が，提案手法を用いることにより収集可能となることが示された．

また，トピックモデルによって各トピックに対応付けられたウェブページにうち，確率値 $P(z_n|b)$ の上位 5 ページを対象として，各トピックが対応する話題と照合する内容のページであるか否かの判定を人手で行った結果を表 4 に示す．この結果から，2 次検索によって収集されたウェブページの約 90% においては，各トピックが対応する話題と照合する内容のページであった．このことから，提案手法によって，話題が多様でかつ一定の情報量を持ったウェブページが収集可能となることが分かる．

5. 関 連 研 究

本論文に関連して，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 [6]，および，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 [1], [5], [9]，階層的なトピックの体系を推定する手法 [2] 等が提案されている．これらの手法においては，いずれも，閲覧対象の文書集合のみを用いて，ファセット体系およびファセットラベルに相当する情報を抽出している．また，メタ検索エンジンにおいてウェブページ検索結果の上位 200 記事程度を対象にして，検索結果のクラスタリングおよびラベル付けをした結果を提示するサービスとして，Yippy(注8) が知られている．一方，文献 [7] においては，与えられた文書集合の話題を俯瞰するタスクにおいて，Wikipedia を知識源として，検索された文書集合全体にわたる分野や話題の粒度にまで抽象化されたファセット体系を用いる手法を提案している．これらの先行研究においては，いずれも，与えられた文書集合における話題の広がりを俯瞰することに焦点が当てられている．

一方，本論文では，「検索エンジン API に対してクエリを与えることにより，ウェブページを収集する」というタスクを設定し，クエリを変更して複数回 API アクセスを行うことを許

(注7)：ここで，同一の話題に相当するトピックが複数存在する場合があるが，本論文において，各トピックから 2 次クエリを人手で選定する際には，他のトピックにおけるクエリ候補を参照せず，独立に 2 次クエリを選定した．このように，本論文における評価結果は，2 次クエリを人手で選定することによって，話題の広がりが最も大きくなる上限値を示したものと位置付けることができる．ただし，より上限値を上げる戦略として，各トピックからは，重複する 2 次クエリを選定しないようにすることも可能であり，この上限値の測定を行った結果については別途報告する．

なお，現在，2 次クエリの自動選定方式として，いくつかの方式について評価を行っており，それらの結果については別途報告する予定である．

(注8)：http://yippy.com/

表 4　各トピックごとのウェブページのまとまりと有用性の評価

(a) 評価対象トピック及びウェブページ数

	収集された ウェブページの数	LDA 推定時に指定した トピックの数	震災に関連し，かつ意味的まとまりのある トピックの数	評価対象となる ウェブページの数
1 次検索	593	15	8	40
2 次検索	20,374	90	50	250

(b) ウェブページ単位の評価結果

	各トピックに関連する ウェブページの数	各トピックに関連しない ウェブページの数	リンク切れなどで評価不可となった ウェブページの数
1 次検索	31 (77.5%)	7 (17.5%)	2 (5.0%)
2 次検索	226 (90.4%)	20 (8.0%)	4 (1.6%)

容し，かつ，ニュース・ブログといった外部言語資源も援用するという枠組みのもとで，できるだけ多様な話題のウェブページを収集する方式を提案している．

また，本論文に関連して，TREC の Web トラックの diversity タスク [4] や NTCIR の INTENT タスク [8] における文書ランキングタスクにおいては，ウェブ検索結果においてできるだけ多様な話題のウェブページを上位に順位付けすることを要求する仕様のもとで評価型タスクを行っている．同様に，NTCIR の INTENT タスク [8] におけるサブトピックマイニングタスクにおいては，クエリについてのサブトピックを列挙する課題を設定し，評価型タスクを行っている．

6. おわりに

本論文では，検索エンジン API を用いたウェブページ収集タスクにおいて，ニュースもしくはブログにおいてのみ観測され，ニュースあるいはブログ以外のウェブページからは，検索エンジン API によって収集されなかった話題の文書を選択的に収集することにより，検索エンジン API を用いてできる限り多様な話題のウェブページを収集する方式を提案した．

現在，各話題について，2 次検索によって収集されたウェブページ中の記載内容と，ニュース・ブログから収集された記事における記載内容の比較対照分析作業を行っている．この比較対照分析作業によって，ニュース・ブログでは取り上げられておらず，ウェブページ固有の情報とみなせる記載内容が，2 次検索によって収集されたウェブページ中にどの程度含まれているのかの検証を行う予定である．この検証を通して，提案手法にしたがって検索エンジン API を用いることにより，ニュース・ブログ以外のウェブページから，従来アクセスが困難であった情報が収集できるか否かが明らかになると考えている．

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] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *NIPS'03*, 2003.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[4] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC-2012 Web track. In *Proc. TREC-2012*, 2012.

[5] W. de Winter and M. de Rijke. Identifying facets in query-biased sets of blog posts. In *Proc. ICWSM*, pp. 251–254, 2007.

[6] 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.

[7] 牧田健作, 鈴木浩子, 小池大地, 宇津呂武仁, 河田容英. Wikipedia を知識源とする分野トピックモデルの推定と分析. 情報処理学会研究報告, Vol. 2012–DBS–155, , 2012.

[8] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT Task. In *Proc. 9th NTCIR Workshop Meeting*, pp. 82–105, 2011.

[9] 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.