災害記事データベースの構築および応用 —記事収集，全文検索，およびテキスト分析—

CONSTRUCTION AND APPLICATON OF DISASTER ARTICLE DATABASE: ARTICLE COLLECTION, FULL-TEXT SEARCH, AND CONTENT ANALYSIS

村川 猛彦[1]
Takehiko MURAKAWA

[1]システム工学部准教授

自然災害に対する意識の高まりとともに，災害発生時の備えが必要となっている．防災・減災や災害事例についての情報を入手する方法として，サーチエンジンや特定のサイトといった，インターネット上の記事からの取得が考えられる．しかしWeb上の情報は，日本語に限っても多数存在し，検索により発見した内容について，有益であるかの判断がすぐにできるとは限らない．そこで本研究では，各記事に出現する単語や，記事群で構成されるトピックに着目し，災害記事データベースの構築に取り組んできた．継続的な記事の収集，全文検索インタフェースの開発，および潜在的ディリクレ配分法を用いたトピック抽出のための調査・分析について述べる．

キーワード： *災害記事，文字情報，データベースシステム，全文検索，トピック抽出*

１．はじめに

近年の自然災害は増加傾向にある．また，防災・減災に対する意識も高まっており，災害発生時の備えが必要となっている．例えば，「ゲリラ豪雨」と呼ばれる1時間降水量が50mm以上の集中豪雨の回数は，1980年からの10年間の平均で198.5回，2000年からの10年間の平均で220.9回であり[1]，増加傾向にある．豪雨に限定することなく，災害が発生した際の避難方法や，避難時の装備などについて，個人的な対策が不可欠となっている．

防災・減災や災害事例に関する情報を入手する方法として，サーチエンジンやSNS（Social Networking Service），特定のサイトなどを通じ，インターネット上の記事を取得することが考えられる．しかしWeb上の記事は，日本語に限っても多数存在する．検索を行っても，必ずしも欲しい情報が手に入るとは限らず，その内容が自分にとって有益であるかの判断がただちに行えるとは限らない．

そこで記事内容を理解できているような分類や，記事として不適切なものを取り除くことができれば，効率よく情報の獲得・閲覧ができるのではないかと考え，災害記事データベースの構築に取り組んできた．収集した記事において，各記事に出現する単語や，記事群により構成されるトピックに着目し，潜在的ディリクレ配分法に基づくスコアリングおよび分析を行ってきた．本稿ではそのあらましを述べる．なお，本稿は既発表[2), 3)]をもとに，執筆時（2016年12月）までに得た情報などに適宜置き換えたものである．

２．災害記事

ブログを含むインターネット上の情報には，災害や防災に関するまとまった記事群が見られる．それらの取得方法として，GoogleやYahoo!といったサーチエンジンや，Twitter，FacebookといったSNSなどがあるが，サイト名や記事のタイトルから内容の理解が必ずしもできるとは限らない．

ブラウザ上に表示された災害記事の例[4)]を**図−1**に示す．この記事を得るため，サーチエンジン（Google）を用いていくつかのキーワードで検索した．本稿執筆の時点では，「津波」を検索すると約 21,400,000 件，「防災」だと約 156,000,000 件，また「津波 防災」とすると約 3,690,000 件のヒット数となり，その上位から選び，読み進めていくことで，最終的に記事に到達した．

一般には，情報取得[5)]の流れは「検索語の入力」「検索結果ページの一覧」「リンクをたどる」「記事を読む」

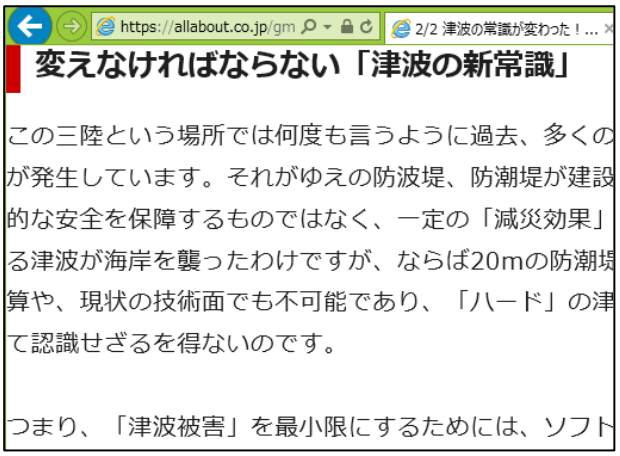

図-1　災害記事の例.

「望む記事かどうかの判断」で構成される．実際には前のステップに戻ることもよく行われ，満足できる情報にたどり着くまでには多くの時間がかかってしまう．

本研究でブログ記事を対象とした理由について述べる．安部ら[6]は被災者の体験談や意見などの被災経験が有用な情報として活用され各省庁や地方自治体がアンケート調査によって収集している．しかしアンケートの配布や集計は労力を要するだけでなく，コストに見合った情報が必ずしも得られるとは限らないことも指摘されている．また，この問題に対し，個人の経験や意見などが書かれているブログの活用によって不足する情報量を補おうとしており，収集した記事の中から地震の震度を自動抽出する手法について検討している．しかし地震という限定されたもののみを対象としており，また収集データも「地震」，「震度」というキーワードを含む記事のみを評価対象としている．

災害情報の収集は過去の災害分析や，今後の災害予測などに用いることができる．そして足りない情報をブログ記事から収集することにより，情報を効率的に集めることができると考えられる．

3．災害記事データベース

筆者らは和歌山大学独創的研究支援プロジェクトの一環である情報通信技術分野の災害関連記事自動収集システムとして「災害記事データベース」を構築中であり，学内外からPCによるアクセスが可能となっている．現在大量の情報が更新し続けているYahoo!ブログ（http://blogs.yahoo.co.jp/）から記事の取得を行い，記事分類プログラムを用いて分類し，インタフェースを通して利用者に記事を閲覧してもらう．構築する災害記事データベースでも同様に，個人の経験や意見などが書かれているブログを活用することによって公的機関からの情報だけでなく，それ以外の不足する情報を収集できることを目指している．

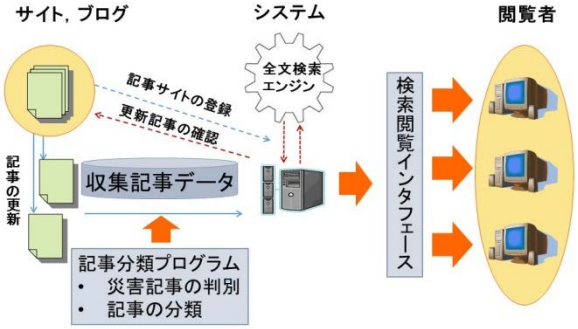

図-2　災害記事データベースのシステム構成.

しかしWebから災害記事を収集し，分析を行うと記事のデータ量が多くなる．また収集した記事を保存させることが必要であり，その中から必要な情報だけを抽出させなければならない．そこで，データベースを用いることでその問題を解決できないかと考えている．データベースを用いることで，収集記事の中から欲しい情報を抽出でき，必要となるたびにYahoo!ブログのサイトにアクセスするという手間を省くことが期待できる．

災害記事データベースのシステム構成を図-2に示す．災害記事データベースは，サーバ上に構築されたデータベースシステムであり，収集した記事をデータベースに保存しておく．記事テーブルは，主キーとなるID（機械的な番号），記事URL，タイトル，本文，登録日時（記事ページ内のメタデータより取得）からなる．

収集され，データベースに格納された記事群は，利用者がシステムの「検索閲覧インタフェース」を通して閲覧し，知識の収集に役立てることができる．その際，閲覧される記事は「ファイル名検索」や「単一ファイル内の文字列検索」ではなく「複数文書にまたがって，タイトルや本文を含む文書全体を対象とした検索」である全文検索によって，キーワードや分類項目などから検索することができる．そのために災害記事データベースでは全文検索エンジンを利用している．

4．記事収集の流れ

災害記事データベースの収集記事は，Yahoo!ブログ災害カテゴリから記事を収集することを想定している．Yahoo!ブログとは，Yahoo! Japanが提供するブログサービスであり，「趣味とスポーツ」や「ビシネスと経済」など多数のカテゴリから構成される．各カテゴリは階層化されており，そこから多数の個人ブログやタレントのブログ，官公庁や地方自治体の情報などを効率よく発見，閲覧することができる．Yahoo!ブログの「災害カテゴリ」へは，Yahoo!ブログトップページから「カテゴリ」タブを選択し，「生活と文化」という大きなカテゴリの中にある「災害」を選択することでアクセスができる．

「災害カテゴリ」には，災害に関する個人の体験や感

表-1 月ごとの収集記事数.

年月	記事数（件）
2015年6月	2,607
2015年7月	3,164
2015年8月	3,045
2015年9月	4,509
2015年10月	2,277
2015年11月	1,899
2015年12月	1,887
2016年1月	2,240
2016年2月	2,016
2016年3月	2,661
2016年4月	9,684
2016年5月	4,173
2016年6月	2,429
2016年7月	1,897
2016年8月	3,284
2016年9月	2,751
2016年10月	1,944
2016年11月	1,764

想などが多数投稿されている．2015年6月から2016年11月までにおける，月ごとの災害カテゴリの収集登録数を**表-1**に示す．2016年4月が突出しているが，これは熊本地震の発生が大きく関わっており，この月の収集記事を対象とした分析については第6章で述べる．2015年5月および2016年12月（執筆時まで），ならびに日付情報が取得できなかった記事を含め，収集記事数は合計55,971件である．

記事収集は，収集すべき記事のURLの獲得，URLに対応するHTMLの取得，そしてHTMLからの情報の取得に大きく分かれる．収集すべき記事は，Yahoo!ブログの災害カテゴリで最新の1,000件（20件×50ページ）から記事URLのみを取り出し，取得済のものを除外することで得られる．

個別の記事からの情報の取得に関する流れを述べる．各ブログ記事はHTMLで構成されている．そこでは<HTML>，<HEAD>，<BODY>といったタグを用いて構造化されているが，記事の本文を抽出する際には不要な情報が多い．

そこで不要なタグなどを除去し，タイトルと記事本文だけを抽出する際，スクリプト言語であるRubyのNokogiriライブラリを使用した．このライブラリはHTMLやXMLの構造を解析して，特定の要素を指定して抽出できる．例えば，取得し一時保存したHTMLファイルについて，そのファイル名を変数filenameに格納しておけば，Nokogiri::HTML(open(filename)).contentというRubyのプログラムコードにより，本文の文字列が獲得できる．

ある記事に対するブラウザ表示例を**図-3**，HTML（ソース）を**図-4**，抽出された本文を**図-5**にそれぞれ示す．

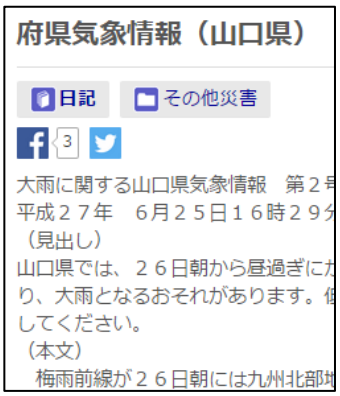

図-3　記事のブラウザ表示例.

図-4　HTML例.

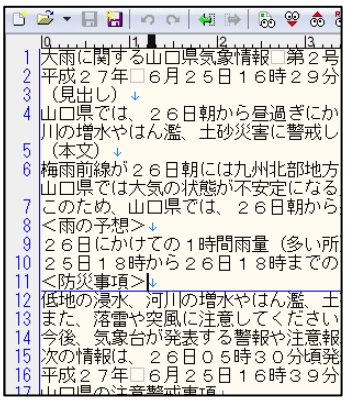

図-5　抽出された本文の例.

５．全文検索

本データベースで収集した記事本文を全文検索するためのインタフェースを構築した（http://fukai.sys.wakayama-u.ac.jp/~takehiko/sg/）．画面例を**図-6**に示す．

インタフェースでは，検索語のほか年月を指定できる．表内のリンクをクリックすると，該当記事（データベースの情報ではなく）にアクセスする．

開発にあたっては，Ajaxを活用している．これにより，ページ遷移をすることなく検索結果が表示される．サーバ側では，SQLを用いて瞬時に漏れのない全文検索を行

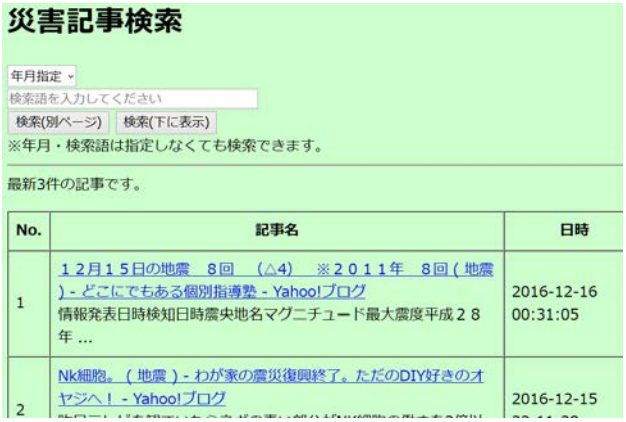

図-6　全文検索画面の例.

えるようにするため，PGroongaを導入している．

ブラウザ操作のほか，検索語や年月をURLのパラメータに指定することでも，検索結果を得ることができる．また通常のWebページ（HTML形式）のほか，RSS文書を生成することもでき，これにより，特定のキーワードを含む最新記事の一覧を，XML形式で取得することも可能としている．

６．LDAを用いた記事のトピック分析

収集した記事群の特徴などを定量的に把握するため，Latent Dirichlet Allocation（潜在的ディリクレ配分法，以下LDA）[7]を用いたトピック分析を試みた．

LDAとはトピックモデルの一つであり，1つの文書が複数のトピック（話題）から表現されるという仮定からの教師なし推定である．文書，単語，トピックの関連について，文献[7]の事例をもとに説明する．文書1には「国会」「審議」「首相」「選挙」といった単語が出現している．文書2では「五輪」「経済」「景気」「球場」，また文書3では「景気」「国会」「審議」「対策」といった単語が出現している．そして文書群より，「国会」「審議」「選挙」「内閣」といった単語で構成されるトピック1，「勝利」「五輪」「野球」「球場」で構成されるトピック2，「景気」「国会」「審議」「対策」で構成されるトピック3を獲得できたとする．

出現単語を比較したとき，文書1とトピック1はほぼ重なっているのに対し，文書2ではトピック2とトピック3が，また文書3ではトピック1とトピック3が混在しており，それらの文書には複数のトピックが含まれていると考えられる．このことにより各文書にはランダムなトピックの混合により表され，このトピックは単語の集合によって表される潜在的な意味に相当する．なお，トピックは明示的な話題（例えば，トピックの名称）を保持しているものではない点に注意する必要がある．

本研究では2016年4月の記事を3期間に区切った上で，それぞれ分析を行った．ここで，期間を区切った日付，およびその基準について述べる．まず，熊本県で最初に震度7が観測された日の前日である13日までを期間aとした．また，記事数を確認したところ13日までは1日あたりの登録記事数が100件を下回っていたが，744件を記録した14日を境に1日あたりの登録記事が増加し，16日には1日の登録記事が1,000件を上回っていることが確認できた．その後は徐々に登録記事数が落ち着きを見せ始め，22日には1日の登録記事数が500件を下回った．そこで14日から21日までを期間b，4月22日以降を期間cとして，分析を行った．

分析の流れを**図-7**に示す．はじめに各記事を，MeCabと呼ばれるソフトウェアで形態素に解析する．例えば「和歌山県でエレベーターに乗っているとき」という文を形態素解析したところ，「和歌山」，「県」，「で」，「エレベーター」，「に」，「乗っ」，「て」，「いる」，「とき」という9つの形態素に分解される．実際の出力には品詞をはじめ詳細情報が含まれており，これをもとに名詞のみを取り出して以降の分析に使用した．

ソフトウェア構成について簡単に説明する．分析そのものは，Pythonで書かれたプログラム[8]を使用した．ただしこのプログラムの入力は，1行が1文書（本研究ではブログ記事1件に対応）で，行は「語のID:出現回数」の並びで表現する必要がある．データベースに格納された記事に対し，この形式に変換してからプログラムを実行するなどの雑多な処理については，Rubyで独自に実装した．

抽出するトピック数は期間ごとに10ずつ，またスコア算出のためのキーワードはトピックごとに上位15個に，それぞれ限定した．

トピックを抽出したときの出力の例を**表-2**に示す．これは期間aのトピック1における単語分布である．上位のものからそのトピックに出現しやすいものが並んでおり，これらのような単語の集合がトピックの数だけ得られる．LDAにより各トピックの単語分布は得られる．そこで各トピックに関わりが高いと思われる記事の発見を，トピックごとに各期のスコアを算出することにより試みた．

スコアリングの具体的な方法は，トピックの単語出現確率と単語の出現頻度（トピックの単語が各ブログ記事の本文およびタイトルに何回出現したか）を掛け合わせることとし，式は(1)の通りとなる．ここでS_iは番号iの記事のスコア，w_jはトピックにおける番号jの単語の出現確率（記事に依存しない），f_{ji}は番号iの記事における番号jの単語の出現頻度を表す．

$$S_i = \sum_j w_j f_{ji} \qquad (1)$$

上記の手法に基づき，期間ごとの各トピックに対してスコアを算出した．最大スコアを**表-3**に示す．異なる表の同一番号（No.）のトピックには関連がない．またどの期間およびトピックの組み合わせにおいても，最小ス

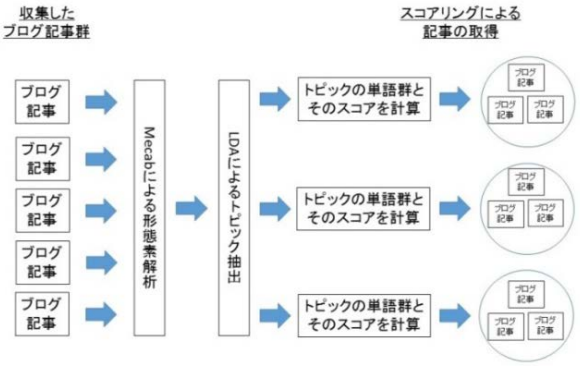

図-7　分析およびスコアリングの流れ.

表-2 単語分布（期間a，トピック1の上位10件）.

単語	出現確率
地震	0.040005
熊本	0.243347
余震	0.016335
被害	0.015950
何	0.009421
九州	0.008927
事	0.008771
震度	0.007252
人	0.008194
心配	0.008177

表-3 トピックごとの最大スコア.

期間a		期間b		期間c	
No.	最大スコア	No.	最大スコア	No.	最大スコア
1	2.0	1	4.4	1	5.4
2	2.8	2	2.9	2	8.9
3	4.0	3	10.8	3	11.7
4	1.5	4	11.3	4	55.6
5	3.6	5	5.9	5	101.8
6	6.3	6	13.5	6	63.5
7	16.2	7	14.2	7	6.0
8	38.8	8	206.3	8	12.0
9	39.3	9	46.9	9	14.5
10	17.9	10	186.1	10	116.1

注．表示の都合上，小数第1位までとしている.

コアは0となった．この表から，期間aの最大スコアは39.3，期間bの最大スコアは206.3，期間cの最大スコアは116.1となり，期間ごと，および各トピックの最大スコアに大きな差が生じていることが確認された.

7．考察

まず，LDAによる抽出トピックの比較を行う．熊本地震前となる期間aでは「地震」「震源」「噴火」などのトピックが確認できるものの，特定の地名が特にトピックとして頻出することはなく，「大阪」「茨城」「福島」など，複数の地名がトピックとして出現した．それに対し，熊本地震が発生後となる期間b，期間cのトピックを確認すると，「地震」「震源」に加えて「被災」「避難」などトピックが確認でき，地名では「熊本」が頻出した．このことから，日時指定を用いなくとも，「避難」や「被災」といったキーワードを用いることで，記事が災害の発生前後のどちらに書かれたものなのかをある程度判別することができるのではないかと考えられる.

次に，熊本地震発生後となる期間bと期間cの比較を行う．顕著な違いとして，期間bのトピック10において「八代」「宇城」「天草」という九州の地名が抽出された．これは期間cに対する分析では見られず，この結果により，期間を災害直後に限定した分析を行うことで，より詳細な事象に関したトピックの取得，運用が可能になるのではないかと考えられる.

スコアリング結果より，100を超えるスコアが得られた期間bのトピック8およびトピック10，期間cのトピック5およびトピック10においては，「市」「町」「村」「年」「月」「日」といった1文字からなるキーワードが上位に多く見られた．また，全体で最大スコアとなった記事[9]をブラウザで表示させたところ（図-8），地名や発生年月といった災害の情報の羅列が行われていることが確認できた．これにより，「市」「町」「村」「年」「月」「日」などの1文字からなるキーワードを用いたスコアが著しく高い記事は，「災害についての情報が羅列されている記事」として分類できるのではないかと考えられる.

期間aのトピック1，トピック2，トピック4，期間bのトピック2では，最大スコアがいずれも3未満となった．これらの記事についてトピックと記事を照らし合わせて確認した.

まず期間aのトピック1を確認すると，キーワードに「円」「建値」「決済」という，他と関連性が見られないものが見られた．このキーワードについて調査したところ，「建値」というキーワードが見られる記事8件すべてが同一ブログの別記事であった．また，これらの記事を確認したところ，いずれも「建値」と「決済」の金額に関する情報の羅列記事となっていた．このことから，ほかのものと関連性が見られない単語が出現した場合，今回と同様に災害と関連性の薄い類似記事をあぶり出すチェック機能を作ることができるのではないかと考えられる.

期間aのトピック2では，「日」「年」「者」「被災」などのキーワードが出現していた．またトピック4では，「地震」「噴火」「年」「回」などのキーワードが確認された.

期間bのトピック2では，「被災」「人」「支援」「災


以下は震度3を越える最近の地震（気象庁地震情報）

http://www.jma.go.jp/jp/quake/quake_singendo_index.html
情報発表日時 検知日時 震央地名 マグニチュード 最大震度

平成28年04月20日08時04分 20日08時00分頃 熊本県熊本地方 M3.6 震度3
平成28年04月20日03時57分 20日03時52分頃 熊本県熊本地方 M2.9 震度3
平成28年04月20日03時42分 20日03時39分頃 熊本県天草・芦北地方 M4.1 震度3
平成28年04月20日02時19分 20日02時16分頃 熊本県熊本地方 M4.0 震度3
平成28年04月20日01時00分 20日00時56分頃 熊本県熊本地方 M3.4 震度3
平成28年04月20日00時34分 20日00時29分頃 熊本県天草・芦北地方 M3.8 震度3
平成28年04月19日23時27分 19日23時23分頃 熊本県熊本地方 M3.2 震度3
平成28年04月19日22時30分 19日22時26分頃 熊本県熊本地方 M3.7 震度3
平成28年04月19日20時50分 19日20時47分頃 熊本県熊本地方 M4.9 震度5弱
平成28年04月19日20時37分 19日20時33分頃 熊本県熊本地方 M3.2 震度3
平成28年04月19日18時20分 19日18時14分頃 熊本県熊本地方 M3.3 震度3

図-8　期間b，トピック8の最大スコア記事.

害」などのキーワードが出現していた．また，最大スコアとなった記事を確認したところ，減災についての考察記事が得られた．この記事について「市」「町」「村」「年」「月」「日」などのキーワードが見られるトピック8およびトピック10によるスコアを確認してみたところ，6.750850や0.626421という平均的なスコアが得られた．このことより，この記事は上で述べた「災害情報が羅列された記事」とは異なる論述記事と言える．今後，この記事の類似例を調査することで，災害に関する論述記事抽出の助けとなるのではないかと考えられる．

8．おわりに

本研究では，記事収集機能，全文検索エンジン，検索閲覧インタフェースなどを持つ災害記事データベースの構築を実施してきた．また収集記事の分析に関して，2016年4月の災害記事群を3つの期間に区切り，それぞれの期間に対してLDA分析を行った．また，その結果をもとに記事のスコアリングを試みた．

今後もスコアリングに関する調査・考察を進め，災害記事データベースに格納されている記事の分類を行いながら，利用者に情報提供を行っていきたいと考えている．

謝辞：本研究を進めるにあたりシステム開発および分析に携わってきた硲石浩文氏，藤原史一氏の両名に深く感謝します．本研究はJSPS科研費 JP25242037の助成を受けたものです．

参考文献

1) 気象庁 | アメダスで見た短時間強雨発生回数の長期変化について，<http://www.jma.go.jp/jma/kishou/info/heavyraintrend.html>，2016年12月16日アクセス.
2) 硲石浩文，村川猛彦：防災・減災に関するWeb上の記事を対象とした分類の試み，情報知識学会誌，Vol.24，No.2，pp.184-188，2014.
3) 藤原史一，硲石浩文，村川猛彦：潜在的ディリクレ配分法を用いた平成28年熊本地震に関するブログ記事の分析，情報処理学会研究報告，Vol.2016-IFAT-123，No.9，2016.
4) 2/2 津波の常識が変わった！東日本大震災の現場を見る [防災] All About，<http://allabout.co.jp/gm/gc/379157/2/>，2016年12月16日アクセス.
5) Ingwersen, P. and Järvelin, K.（著），細野公男，岸田和明，緑川信之（訳）：情報検索の認知的転回―情報捜索と情報検索の統合―，丸善，2008.
6) 安部智也，安藤一秋：防災教育に向けた被災経験ブログの収集，2012年度JSiSE学生研究発表会，2012.
7) 岩田具治：トピックモデル，講談社，2015.
8) satomacoto ： Python で LDA を実装してみる，<http://satomacoto.blogspot.jp/2009/12/pythonlda.html>，2016年12月16日アクセス.
9) 地震頻発の阿蘇地方で地震活動が活発化そして南下? (地震) - 原典聖書研究 - Yahoo!ブログ，<http://blogs.yahoo.co.jp/semidalion/49259364.html>，2016年12月16日アクセス.

（2016. 12. 16受付）