DEIM Forum 2011 A6-3

ニュース記事中の話題に関連するブログ記事の収集手法

佐藤　由紀† 　横本　大輔† 　牧田　健作†† 　宇津呂武仁† 　福原　知宏†††

† 筑波大学大学院システム情報工学研究科 知能機能システム専攻　〒 305-8573 茨城県つくば市天王台 1-1-1
†† 筑波大学理工学群工学システム学類　〒 305-8573 茨城県つくば市天王台 1-1-1
††† 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　本論文においては，ニュース記事中において話題を表すキーワードを自動選定し，それらのキーワードに関して詳細な記述をしているブログ記事を収集する手法を提案する．特に，Wikipedia を知識源としてニュース記事中の話題に密接に関連するキーワードを選定する手法，および，ニュース記事のタイトルおよび冒頭部分からキーワードを選定する手法の比較を行う．また，検索エンジン API を利用して，ニュース記事との間で類似度の高いブログ記事を選定する．

キーワード　情報検索，Wikipedia，ニュース，ブログ，トピック分析

Collecting Blog Posts related to Topics in a News Article

Yuki SATO†, Daisuke YOKOMOTO†, Kensaku MAKITA††, Takehito UTSURO†, and Tomohiro FUKUHARA†††

† Graduate School of Systems and Information Engineering,University of Tsukuba, Tsukuba, 305-8573, Japan
†† College of Engineering Systems, School of Science and Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan
††† Center for Service Research, National Institute of Advanced Industrial Science and Technology, Tokyo, 135-0064, Japan

Abstract　This paper studies complementary navigation of news and blog, where, given a news article, keywords indicating topics in the news article are automatically selected, and then, blog posts that are closely related the topics are collected. In this framework, we compare two approaches. In the first approach, *Wikipedia* entries are utilized as fundamental knowledge source for linking news articles and blog posts. In the second approach, keywords extracted from the beginning part of the news article are selected. Finally, in our framework, we utilize a search engine API for collecting blog posts that are closely related to those extracted keywords.

Key words　information retrieval, Wikipedia, news, blog, topic analysis

1. はじめに

本研究では，検索エンジン等を用いた検索行動のうちでも，特に，客観的かつ恒久的な事実を記載した Wikipedia，詳細な事実情報を報道するニュース，および，個人の主観的意見や経験などを豊富に記載したブログの検索に焦点を当てて，利用者の検索行動を支援する枠組みを提供することを目的とする．本研究では，これらの三種類の情報源の間で，密接に関連する項目や記述部分の間を相互にナビゲートする機能を実現し，利用者の検索行動を支援する (図 1).

Wikipedia，ニュース，ブログの三者を比較すると，Wikipedia は，インターネット上の最大規模の百科事典として，近年，様々な研究分野において利用されている (例えば，文献 [1], [8]). 日本語では，約 72 万 5,000 のエントリ (2011 年 1 月現在) が収録されており，しかも，多くの人が自由にエントリを書くことができるため，ニュースやブログで話題となる事項のエントリが，迅速に作成されるという特徴を持っている．Wikipedia を利用した研究事例としては，図書館の分類体系と Wikipedia カテゴリの対応付けを行う研究 [8] や，Wikipedia の言語間リンクを利用して多言語対訳辞書を作成するという研究 [1] などがある．

ニュースとブログを比較すると，ニュースは，従来より，日々の報道を閲覧するという形で利用されてきた．一方，ブログについても近年，世界中でブログサービスやブログツールが普及し，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評判を発信することが可能になるのに伴って，様々な情報がブログに記載され，また，商用ブログ検索サービスを利用す

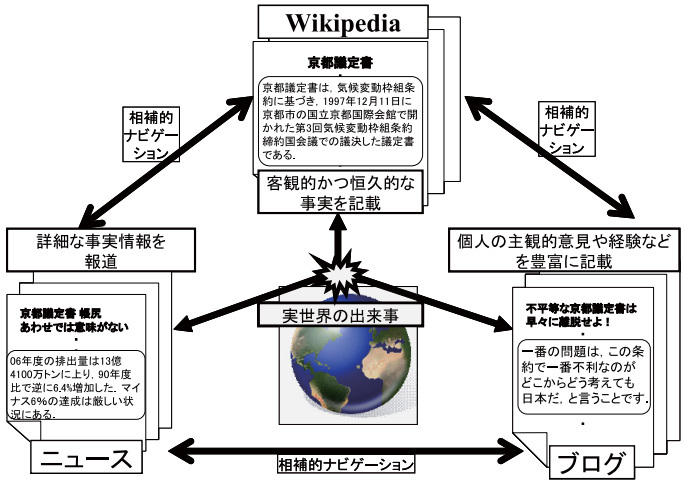

図 1　Wikipedia，ニュース，ブログ間の相補的ナビゲーションの枠組み

ることでそれらの情報を取得することが出来るようになった．具体的なサービスの例として，*Technorati*(注1)，*BlogPulse*(注2)，*kizasi.jp*(注3) などが挙げられる．これらの検索サービスは，巨大なブログ空間の索引付けという観点から見ると，キーワードや評判，時系列変化や人手によって作成されたカテゴリ情報などを索引として用いて，利用者の求めるブログ記事やブログサイトを検索する．

本研究において，Wikipedia，ニュース，ブログの三種類の情報源の間で，密接に関連する項目や記述部分を相互にナビゲートする機能を実現するにあたっては，まず，あるトピックについて，Wikipedia のエントリから関連する用語を抽出し，これらの用語を知識源として，関連するニュース記事，ブログサイト，ブログ記事を検索する．この検索のうち，特にブログサイトの検索においては，我々はすでに，文献 [6] において，Wikipedia エントリの記述内容をトピックとする有用なブログサイトを検索する方式を確立している．この方式においては，Wikipedia エントリタイトルを検索クエリとして，商用検索エンジン API により上位のブログサイトを収集し，これを，当該キーワード，および Wikipedia エントリから抽出した関連語の出現数順に順位付けするという要素技術を用いている．

一方，本論文においては，本研究におけるこれまでの成果をふまえて，ニュース記事中において話題を表すキーワードを自動選定し，それらのキーワードに関して詳細な記述をしているブログ記事を収集する手法を提案する．本論文における「ニュース記事に関連するブログ記事収集の枠組み」の模式図を図 2 に示す．本論文では，特に，Wikipedia を知識源としてニュース記事中の話題に密接に関連するキーワードを選定する手法，および，ニュース記事のタイトルおよび冒頭部分からキーワードを選定する手法の比較を行う．また，検索エンジン API を利用して，ニュース記事との間で類似度の高いブログ記事を選定する手法を用いる．評価実験を通して，ニュース記事中において話題を表すキーワードを自動選定する手法としては，Wikipedia を知識源とする手法とニュース記事中に位置を用いる手法を併用する (4.1.3 節) ことにより，それぞれの手法が相補的に機能し，評価対象の全ニュース記事に対する総合的な性能が改善できることを示す．

2. 関 連 研 究

ニュース記事に対して関連するブログ記事を対応付ける方式に関する関連研究は，大別すると，ニュース記事およびブログ記事のテキスト情報の間の関連性に基づく手法 [3]，および，ブログ記事からニュース記事へのリンクによる引用情報を用いる手法 [2], [4], [5] に分けられる．このうち，本論文の手法は，文献 [3] と同様に，ニュース記事およびブログ記事のテキスト情報の間の関連性に基づく手法に相当する．

本論文の手法と文献 [3] の手法との間の主要な違いは以下の通りである．

(a)　ニュース記事のトピックを表すキーワードの選定手法

(b)　ブログ記事の収集方式 (独自のクローリングか既存の検索エンジン API か)

(c)　ブログ記事の日付の扱い

(d)　ニュース記事ベクトルおよびブログ記事ベクトルの次元

(a) に関しては，文献 [3] においては，ニュース記事のトピックを表すキーワードを選定する手法として，その候補を，タイトル，および，一文目に含まれる語に限定するという手法 (4.1.2 節「ニュース記事中の位置を用いた方法」) を採っている．一方，本論文では，ニュース記事における記述内容との間で高い類似度を持つ Wikipedia エントリを選定し，これらのエントリのタイトルによってニュース記事のトピックを表すという手法 (4.1.1 節「Wikipedia を用いた方法」) を提案する．なお，評価実験の結果では，両手法を併用する (4.1.3 節) ことにより，それぞれの手法が相補的に機能し，評価対象の全ニュース記事に対する総合的な性能が改善できることがわかった．(b) に関しては，文献 [3] では，ブログ記事の収集において，既存の検索エンジン API は用いず，独自にブログのクローリング

(注1)：http:/technorati.com/
(注2)：http://www.blogpulse.com/
(注3)：http://kziasi.jp/（日本語のみ）

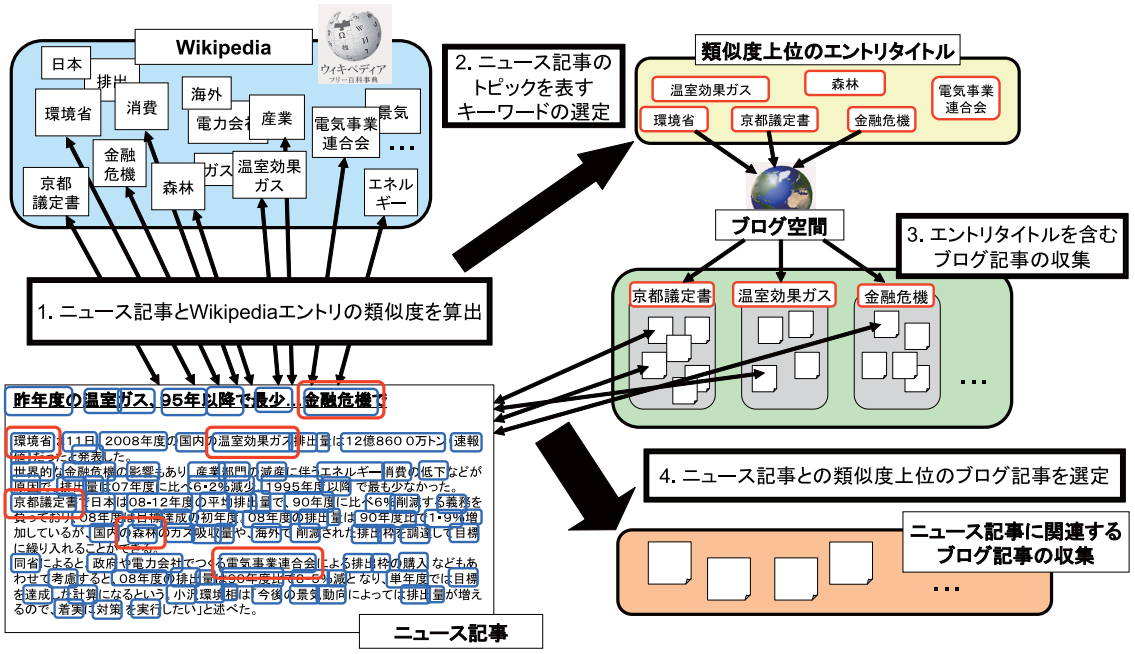

図 2　ニュース記事に関連するブログ記事収集の枠組み

を行っている．一方，本論文では，独自のクローリングは行わず，より容易に利用可能な既存の検索エンジン API を用いる．特に，ニュース記事のトピックを表すキーワードを自動選定し，検索エンジン API に与えることにより，ニュース記事に関連するブログ記事を収集するアプローチを採用している．(c) に関しては，文献 [3] においては，ニュース記事の日付以降，7 日以内の日付のブログ記事に限定して，関連ブログ記事の収集を行なっている．特に，手法面においては，ニュース記事とブログ記事との間の文書類似度において，ニュース記事の日付の直後における語の出現頻度の推移を考慮した重み付けを用いることにより，ニュース記事に関連したブログ記事の対応付けの性能を改善している．一方，本研究では，ニュース記事の日付以降，1 週間以内の日付のブログ記事に限定して関連ブログ記事の収集を行った場合，および，ブログ記事の日付に対する制限は設けず，ニュース記事の日付の前後の任意の日付のブログ記事を収集対象とした場合の比較を行った．その結果，ニュース記事のトピックによって，特にニュース記事の日付直後に関連ブログ記事が投稿される度合いの大きいトピックと，ニュース記事の日付の前後の任意の日付にわたって関連ブログ記事が広く分散するトピックの両方が存在することが分かった．(d) に関しては，文献 [3] においては，ニュース記事ベクトル，ブログ記事ベクトルとも，茶筌によって形態素解析を行った後の名詞，アルファベット，未知語を次元としているが，本論文では，Wikipedia エントリのタイトルを次元としており，ニュース記事，および，ブログ記事のトピックに相当する可能性の高い固有表現等の名詞句を次元とできている点が異なっている．

一方，ブログ記事からニュース記事へのリンクによる引用情報を用いる手法 [2], [4], [5] のうち，例えば，文献 [5] においては，ブログ記事中で参照しているウェブサイトやニュース記事をそのユーザの興味の対象として，ブロガーの嗜好を利用したウェブ情報推薦システムを提案している．具体的にはニュースサイトとブログの対応付けを行い，ユーザの嗜好にあったニュース記事を推薦するということを行っている．また，文献 [9] では，同じ事象について，複数の情報源の情報の伝え方の異なりかたを分析することを目的として，複数の国の代表的なメディアが発信するニュースを情報源として，各々の国の世論がどのように事象を分析しているのかを把握する方式を提案している．

3. Wikipedia を用いた関連語の収集

3.1 Wikipedia

Wikipedia とは多くの人が自由に書くことができるインターネット上の巨大な辞書のことであり，日本語で約 72 万 5,000 のエントリ (2011 年 1 月現在) がある．さらに，10 個程度の主要カテゴリ以下にサブカテゴリ，エントリが連なる，巨大なグラフ構造になっている．また，カテゴリがグラフ構造の節にあたり，エントリが節内に列挙されている．他の知識源と比較した場合の Wikipedia の最大の利点として，日常的に，新たなエントリの作成と記述の更新が行われており，ブログにおける分析対象となり得る主要なトピックが網羅されている点が挙げられる．

3.2 Wikipedia 関連語の収集

トピック名がタイトルである Wikipedia エントリを知識源として，トピック名に密接に関連する Wikipedia 関連語を収集する．特に，本論文においては，各エントリのリダイレクト，各エントリ本文中の太字，各エントリの段落タイトル，および，本文中における名詞句(注4)を Wikipedia 関連語として収集する．

(注4)：茶筌 (http://chasen-legacy.sourceforge.jp/) および IPAdic (http://sourceforge.jp/projects/ipadic/) を用いて形態素解析を行った後，名詞，接頭詞，未知語のいずれかが連続する形態素列を連結したものを名詞句とした．

4. ニュース記事に関連するブログ記事の収集

4.1 ニュース記事のトピックを表すキーワードの選定

本節では，ニュース記事のトピックを表すキーワードを選定する手法について述べる．なお，本論文では，2009 年 6 月 1 日 ~2010 年 5 月 31 日の 1 年間の期間において収集したニュース記事集合(注5) から選定したニュース記事を評価対象として用いる．また，ニュース記事集合に対する DF(文書頻度)，IDF(逆文書頻度) の統計量を算出する際には，この記事集合の全ニュース記事を用いる．

4.1.1 Wikipedia を用いた方法

本論文の手法においてニュース記事 d のトピックを表すキーワードを選定するにあたっては，ニュース記事 d 中に出現する Wikipedia エントリのタイトルをキーワードの候補とする．そして，各 Wikipedia エントリの本文 e とニュース記事 d との間で類似度を測定し，この類似度の高いエントリのタイトルを選定する．

本手法の詳細を以下で述べる．まず，Wikipedia エントリの本文 e とニュース記事 d との間の類似度の定式化においては，両文書に共通する次元として，3.2 節において e から収集した Wikipedia 関連語の集合 $R(e)$ を用いる．具体的には，Wikipedia エントリの本文 e は，e から収集した Wikipedia 関連語集合 $R(e)$ の要素 r を次元とし，各次元の値が 1 であるベクトルとする．ニュース記事 d も同様に，e から収集した各 Wikipedia 関連語集合 $R(e)$ の要素 r を次元とし，ニュース記事 d 中における関連語 r の頻度 $\mathrm{TF}(d,r)$ を値とするベクトルとする．そして，ベクトル e と d の内積によって，Wikipedia エントリの本文 e とニュース記事 d との間の類似度 $Sim_{n,w}(e,d)$ を定式化する．

$$Sim_{n,w}(e,d) = \sum_{r \in R(e)} 1 \times \mathrm{TF}(d,r)$$

次に，予備実験(注6) において，以下のエントリ集合の評価を行った．

- ニュース記事 d との間の類似度 $Sim_{n,w}(e,d)$ の高い上位 10 エントリの集合．
- 類似度 $Sim_{n,w}(e,d)$ の下限を満たすエントリのうち，類似度の高い上位 10 エントリの集合．ただし，類似度の下限としては，以下の二通りを評価した．
  - $Sim_{n,w}(e,d) > 10$
  - $Sim_{n,w}(e,d) > 15$
- 4.2 節においてブログ記事収集の対象とする 10 ブログホストにおけるヒット数の上限を満たすエントリのうち，類似度の高い上位 10 エントリの集合．ただし，ヒット数の上限としては，以下の二通りを評価した．
  - ヒット数が 100 万未満
  - ヒット数が 1,000 万未満
- 類似度 $Sim_{n,w}(e,d)$ の下限，および，10 ブログホストにおけるヒット数の上限の両方の条件を満たすエントリのうちの，類似度の高い上位 10 エントリの集合．

そして，予備実験において性能の高かった以下の二種類のエントリ集合を評価対象とした．

- E_w^1 ··· 10 ブログホストにおけるヒット数が 100 万未満のエントリのうち，類似度 $Sim_{n,w}(e,d)$ の高い上位 10 エントリの集合．
- E_w^2 ··· 類似度の下限 $Sim_{n,w}(e,d) > 15$ を満たし，10 ブログホストにおけるヒット数が 1,000 万未満のエントリのうち，類似度 $Sim_{n,w}(e,d)$ の高い上位 10 エントリの集合．

4.1.2 ニュース記事中の位置を用いた方法

文献 [3] においては，ニュース記事のトピックを表すキーワードは，ニュース記事のタイトルおよび一文目に集中しているとして，ニュース記事ベクトルの次元の候補をタイトル，および，一文目に含まれる語に限定している．また，それらの語に対して，ニュース記事集合全体を用いて TFIDF の値を算出し，この値の上位の語をキーワードとして選定している．本論文でも，上述した Wikipedia を用いる方法と併用する目的で，このニュース記事中の位置を用いる方法を採用する．

具体的には，まず，上述した Wikipedia を用いる方法の場合と同様に，ニュース記事 d のタイトルおよび一文目に出現する Wikipedia エントリのタイトルをキーワードの候補とする．そして，これらのキーワード候補に対して，ニュース記事集合全体を用いて TFIDF の値を算出し，この値の上位 10 エントリの集合を E_{pos} として，以降の手続きにおいて用いる．

4.1.3 両手法を併用した方法

上述した Wikipedia を用いる方法，および，ニュース記事中の位置を用いる方法を併用するために，両者のエントリ集合の和集合をとり，これを用いる．以上をふまえて，以降の評価実験においては，各手法単独での Wikipedia エントリ集合，および，両手法による Wikipedia エントリ集合の和集合として，以下の 4 通りを評価対象とする(注7)．

（1） E_w^1 ··· 10 ブログホストにおけるヒット数が 100 万未満のエントリのうち，類似度 $Sim_{n,w}(e,d)$ の高い上位 10 エントリの集合．

（2） $E_w^1 \cup E_{pos}$ ··· E_w^1 とニュース記事中の位置を用いて選定したエントリ集合 E_{pos} の和集合．

（3） $E_w^2 \cup E_{pos}$ ··· 類似度の下限 $Sim_{n,w}(e,d) > 15$ を満たし，10 ブログホストにおけるヒット数が 1,000 万未満のエントリのうち，類似度 $Sim_{n,w}(e,d)$ の高い上位 10 エントリの集合 E_w^2 と E_{pos} の和集合．

（4） $E_w^1 \cup E_w^2 \cup E_{pos}$

次節以降では，説明のため，上記の 4 通りのいずれかの手法に

(注5)：日経新聞 (http://www.nikkei.com/)，朝日新聞 (http://www.asahi.com/)，読売新聞 (http://www.yomiuri.co.jp/) の各新聞社のサイトから収集した 56,503 記事，38,758 記事，および，62,684 記事の合計 157,945 記事．

(注6)：2008 年 1 月 1 日 ~9 月 29 日の期間に収集したニュース記事 10 記事を対象として行った．

(注7)：E_w^2 および E_{pos} においては，4.3 節のブログ記事順位付けの結果，ニュース記事とブログ記事との間で共有される Wikipedia エントリタイトルの数が少なく，類似度が同点となるブログ記事が多数収集される場合が多かった．そこで，5. 節においては，これらの評価結果は割愛した．

より選定された Wikipedia エントリ集合を E と記述する．そして，ニュース記事のトピックを表すキーワードの集合としては，集合 E の各要素 e のタイトル $t(e)$ の集合を用いる．

4.2 ブログ記事の収集

前節の手順によりニュース記事 d から選定された Wikipedia エントリ集合 E に対して，ニュース記事 d に関連するブログ記事の候補集合 $P(E)$ を収集する．本論文では，ブログ記事の収集においては，Yahoo!Japan 検索 API(注8)を利用する．ただし，ブログホスト大手 10 社(注9)のドメインに限って検索を行った．検索の際には，Wikipedia エントリ集合 E の要素 e のタイトルを検索クエリとして，複数のブログホスト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，Wikipedia エントリ e のタイトル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 Wikipedia エントリ e のタイトルを含むブログ記事を収集し，ブログ記事集合 $P(e)$ を作成する．最終的に，Wikipedia エントリ集合 E 中の全要素 e に対して収集されたブログ記事集合 $P(e)$ の和集合を，ニュース記事 d に関連するブログ記事の候補集合 $P(E)$ とする．

$$P(E) = \bigcup_{e \in E} P(e)$$

なお，本論文において評価対象としたブログ記事については，2010 年 8 月 ~11 月の期間に収集した．

4.3 ブログ記事の順位付け

本論文の手法においては，ニュース記事 d，および，前節で収集した関連ブログ記事候補集合の各要素 p との間で類似度を測定し，この類似度の降順にブログ記事を順位付けて，上位のブログ記事を選定する．

本手法の詳細を以下で述べる．ニュース記事 d とブログ記事 p との間の類似度の定式化において，両文書に共通する次元としては，4.1 節において，ニュース記事のトピックを表すキーワードの集合として選定した Wikipedia エントリ集合 E の各要素 e のタイトル $t(e)$ を用いる．具体的には，ニュース記事 d，および，ブログ記事 p はいずれも，Wikipedia エントリ集合 E の各要素 e のタイトル $t(e)$ を次元とするベクトルとして表現する(注10)．

ここで，文献 [3] において，ニュース記事とブログ記事の類似度を測定する際には，ニュース記事ベクトルの各次元の値として，ニュース記事集合全体を用いて算出した TFIDF の値を用いている．また，ブログ記事ベクトルの各次元の値としては，TFIDF と IDF とを比較評価した結果，ブログ記事における TF を考慮せず，ブログ記事集合中から算出した IDF の値のみを用いる方が高性能であるとしている．さらに，ニュース記事とブログ記事の類似度としては，ニュース記事ベクトルとブログ記事ベクトルの内積および余弦を比較評価し，内積を用いる方が高性能であるとしている．

以上をふまえて，本論文においても，ニュース記事ベクトルの各次元の値としては，ニュース記事集合全体を用いて算出した TFIDF の値を用いる．具体的には，次式の $\mathrm{TFIDF}_n(d, t(e))$ を用いる．

$$\mathrm{TFIDF}_n(d, t(e)) = \mathrm{TF}(d, t(e)) \times \log \frac{N_n}{\mathrm{DF}_n(t(e))}$$

ただし，上式において，$\mathrm{TF}(d, t(e))$，N_n，$\mathrm{DF}_n(t(e))$ は，それぞれ，ニュース記事 d 中における $t(e)$ の頻度，ニュース記事集合全体の記事数，ニュース記事集合全体における $t(e)$ の文書頻度である．また，ブログ記事ベクトルの各次元の値としては，前節においてブログ記事収集の対象とした 10 ブログホストにおける IDF の値を推定したものを用いる．具体的には，総ブログ記事数 N_b，および，$t(e)$ の文書頻度 $\mathrm{DF}_b(t(e))$ の推定値を用いて算出した次式の $\mathrm{IDF}_b(t(e))$ を用いる．

$$\mathrm{IDF}_b(t(e)) = \log \frac{N_b}{\mathrm{DF}_b(t(e))}$$

ただし，総ブログ記事数 N_b の推定値としては，10 ブログホストにおけるヒット数が十分に大きい値を示す一般語のヒット数を用いた(注11)．また，$t(e)$ の文書頻度 $\mathrm{DF}_b(t(e))$ の推定値としては，10 ブログホストにおける $t(e)$ のヒット数を用いた．さらに，ニュース記事 d とブログ記事 p との間の類似度 $Sim_{n,b}(d, p)$ は，ベクトル d と p の間の内積によって定義する(注12)．

$$Sim_{n,b}(d, p) = \sum_{e \in E} \mathrm{TFIDF}_n(d, t(e)) \times \mathrm{IDF}_b(t(e))$$

5. 評　　価

5.1 評 価 手 順

2009 年 6 月 1 日 ~2010 年 5 月 31 日の 1 年間の期間において収集したニュース記事集合から，20 記事を選定し，評価対象

(注8)：http://www.yahoo.co.jp/

(注9)：fc2.com, yahoo.co.jp, rakuten.co.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, yaplog.jp, webry.info.jp, hatena.ne.jp

(注10)：ニュース記事とブログ記事との間の類似度の次元として，ニュース記事，および，ブログ記事中に出現する全 Wikipedia エントリをベクトルの次元として類似度を測定する方式の評価も行ったが，エントリ集合 E の各要素 e のタイトル $t(e)$ を次元とする方式の方が高い性能であった．

(注11)：予備実験において，総ブログ記事数 N_b の推定値を変動させて，ニュース記事に関連するブログ記事の順位付け性能を比較したが，順位付け性能の変動は見られなかった．

(注12)：ニュース記事との関連性が最も高い Wikipedia エントリ上位 10 個を用いて収集したブログ記事の集合に対して，ニュース記事との間の関連性の強さによってブログ記事の順位付けを行う方式として，文献 [7] においては，各 Wikipedia エントリとブログ記事の間の関連性の強さ (文献 [6] において我々が提案した尺度) を考慮する方式を提案している．この方式では，Wikipedia エントリから抽出したエントリタイトル，リダイレクト，各エントリ本文中の太字，および，本文中における他エントリへのリンクのアンカーテキストの情報を用いることにより，まず，各 Wikipedia エントリとブログ記事の間の関連性の強さを測定する．そして，ニュース記事との関連性が最も高い Wikipedia エントリ上位 10 個との間の関連性の強さの総和をとることにより，ニュース記事とブログ記事との間の関連性の強さを測定している．しかし，本論文で述べる評価実験の予備実験として，2008 年 1 月 1 日 ~9 月 29 日の期間に収集したニュース記事 10 記事を対象として行った評価実験においては，本節で述べる手法の性能が，文献 [7] の手法の性能を上回ったため，5. 節で述べる評価実験においても，本節で述べる手法の評価結果のみを示す．

表 1　ニュース記事・関連ブログ記事間の共有キーワード数 / 関連ブログ記事収集結果の正解率

キーワード選定手法		ニュース記事から選定されたキーワード数の平均値	ニュース記事・関連ブログ記事間の共有キーワード数の平均値		ブログ記事収集結果の正解率 (%)
			ニュース記事全文から選定されたキーワード	ニュース記事 2 文目以降から選定されたキーワード	
Wikipedia を用いる方法	E_w^1: 10 ブログホストにおけるヒット数が 100 万未満のエントリのうち，類似度 $Sim_{n,w}(e,d)$ の高い上位 10 エントリの集合	9.8	8.1	4.3	57.5
	E_w^2: 類似度の下限 $Sim_{n,w}(e,d) > 15$ を満たし，10 ブログホストにおけるヒット数が 1,000 万未満のエントリのうち，類似度 $Sim_{n,w}(e,d)$ の高い上位 10 エントリの集合	6.4	—	—	—
E_{pos}: ニュース記事中の位置を用いて選定したエントリ集合		9.1	—	—	—
両手法を併用した方法	$E_w^1 \cup E_{pos}$	15.6	12.2	4.1	**61.5**
	$E_w^2 \cup E_{pos}$	12.0	10.1	2.0	53.5
	$E_w^1 \cup E_w^2 \cup E_{pos}$	16.4	12.8	4.7	**62.0**

とする．各ニュース記事の話題は,「年金」に関するものが 5 記事,「地球温暖化」,「医療」,「喫煙」に関するものがそれぞれ 4 記事,「北朝鮮」に関するものが 2 記事,「振り込め詐欺」に関するものが 1 記事となっている．各ニュース記事の文字数は，約 200 文字から 600 文字程度であり，平均で約 400 文字である．

評価においては，ニュース記事のトピックを表すキーワードの選定法としては，4.1.3 節において述べた 4 通りの手法を用いる．また，一つのニュース記事に対して収集されたブログ記事集合に対して，4.3 節で述べた類似度を用いてブログ記事の順位付けを行い，上位 10 ブログ記事を評価対象とする．評価の際には，収集されたブログ記事がニュース記事に関連するか否かを人手で判定し，正解率を算出する．ただし,「関連ブログ記事」の判定においては，比較的狭い範囲の話題のブログ記事に絞って「関連ブログ記事」と判定した．例えば，図 2 に示すニュース記事「昨年度の温室ガス、95 年以降で最少 . . . 金融危機で」の場合には，日本国内における温室効果ガスの排出量の変移を話題とするブログ記事までを「関連ブログ記事」と判定し，その他の，環境・エネルギー問題対策による雇用の創出や，温室効果ガス削減方法といった周辺の話題についてのブログ記事は,「関連ブログ記事」とは判定しなかった．

5.2　評価結果: ブログ記事の日付を考慮しない場合

表 1 に，各手法ごとに，ニュース記事から選定されたキーワード数，ニュース記事および関連ブログ記事間で共有されたキーワード数，および，関連ブログ記事収集結果の正解率を示す．なお，ニュース記事および関連ブログ記事間で共有されたキーワード数については，ニュース記事全文から選定されたキーワード分とあわせて，ニュース記事の 2 文目以降から選定されたキーワード分も示す．この結果から分かるように，Wikipedia を用いる方法とニュース記事中の位置を用いる方法の両手法を併用した場合に，正解率が最大になる．また，ニュース記事中の 2 文目以降から選定されたキーワードのうち，ニュース記事と関連ブログ記事の間で共有されたものが一定数存在することが分かる．

次に，選定されたキーワード集合が包含関係にあり，正解率の差が相対的に大きい手法の組に対して，各手法間の正解率の差を詳細に分析する．具体的には，$E_w^1 \cup E_{pos}$ と E_w^1 の組，および，$E_w^1 \cup E_w^2 \cup E_{pos}$ と $E_w^2 \cup E_{pos}$ の組が対象となる．

まず，$E_w^1 \cup E_{pos}$ と E_w^1 の組についての分析結果を示す．$E_w^1 \cup E_{pos}$ と E_w^1 の組について,「$E_w^1 \cup E_{pos}$ の正解率 $> E_w^1$ の正解率」,「$E_w^1 \cup E_{pos}$ の正解率 $< E_w^1$ の正解率」,「$E_w^1 \cup E_{pos}$ の正解率 $= E_w^1$ の正解率」となるニュース記事の数は，それぞれ，8 記事，4 記事，8 記事であった．このうち，まず,「$E_w^1 \cup E_{pos}$ の正解率 $> E_w^1$ の正解率」の場合について，ニュース記事，および，関連ブログ記事として判定されたブログ記事の例を表 2(a) に示す．両者の性能の差は，ブログ記事収集に用いられた Wikipedia エントリタイトルの集合の違いに起因するため，表中には，$E_w^1 \cap E_{pos}$ に含まれるエントリ，E_w^1 のみに含まれるエントリ，および，E_{pos} のみに含まれるエントリをそれぞれ示す．このうち，実際に両者の性能差の原因となっているのは，E_{pos} のみに含まれる，比較的一般語に近い 5 エントリである．表 2(a) に例として示した関連ブログ記事は，E_w^1 およびブログ記事の両方に含まれる 8 エントリのタイトルを用いたブログ記事の収集段階においては，収集されなかった．このように，表 2(a) に例として示したニュース記事の場合は，E_{pos} のみに含まれる，比較的一般語に近いエントリを併用することにより，関連ブログ記事収集の正解率が改善されることが分かる．

次に,「$E_w^1 \cup E_{pos}$ の正解率 $< E_w^1$ の正解率」の場合について，ニュース記事，および，関連ブログ記事として判定されたブログ記事の例を表 2(b) に示す．この場合，表 2(a) とは逆に，表 2(b) に例として示した関連ブログ記事は，エントリ集合 E_w^1

表 2　キーワード選定手法 $E_w^1 \cup E_{pos}$ と E_w^1 の間における 10 位以内の関連ブログ記事の比較の例 (ブログ記事の日付を考慮しない場合)

(a) 「$E_w^1 \cup E_{pos}$ の正解率 $>$ E_w^1 の正解率」となるニュース記事の例

ニュース記事		ブログ記事収集結果の正解率 (%)	
日付	タイトル	$E_w^1 \cup E_{pos}$	E_w^1
2009 年 11 月 11 日	「昨年度の温室ガス、95 年以降で最少 . . . 金融危機で」	90	70

$E_w^1 \cup E_{pos}$ のみにおいて 10 位以内に順位付けされた関連ブログ記事の例				
ブログ記事の日付・要約	ブログ記事との間で共有されたエントリ数 (エントリ)			E_w^1 における ブログ記事 の順位
	$E_w^1 \cap E_{pos}$ に 含まれる エントリ	E_w^1 のみに 含まれる エントリ	E_{pos} のみに 含まれる エントリ	
2010 年 3 月 26 日「2008 年度の排出量は低下したが . . . 努力した結果ではなく . . .」	4 エントリ (環境省，金融危機，温室効果ガス，温室効果)	4 エントリ (京都議定書，議定書，景気，森林)	**5 エントリ (温室，年度，環境，排出，ガス)**	収集されない

(b) 「$E_w^1 \cup E_{pos}$ の正解率 $<$ E_w^1 の正解率」となるニュース記事の例

ニュース記事		ブログ記事収集結果の正解率 (%)	
日付	タイトル	$E_w^1 \cup E_{pos}$	E_w^1
2009 年 6 月 25 日	「患者取り違え事故、後絶たず　05 年以降で 85 件」	10	30

E_w^1 のみにおいて 10 位以内に順位付けされた関連ブログ記事の例				
ブログ記事の日付・要約	ブログ記事との間で共有されたエントリ数 (エントリ)			$E_w^1 \cup E_{pos}$ おける ブログ記事 の順位
	$E_w^1 \cap E_{pos}$ に 含まれる エントリ	E_w^1 のみに 含まれる エントリ	E_{pos} のみに 含まれる エントリ	
2009 年 11 月 22 日「「医療版事故調査委員会 (事故調)」創設の議論がストップ . . .」	1 エントリ (医療事故)	4 エントリ (横浜市大，横浜市，医療機関，義務)	**4 エントリ (患者，医療，事故，過去)**	15 位

のみを用いたブログ記事の順位付けにおいてのみ上位に順位付けされている．一方，このブログ記事は，E_{pos} のみに含まれる，比較的一般語に近い 4 エントリを含むエントリ集合 $E_w^1 \cup E_{pos}$ を用いたブログ記事の順位付けにおいては，15 位にまで順位を下げている．これは，この 4 エントリ (あるいは，E_{pos} のみに含まれるその他のいずれかのエントリ) を記事中に含む「当該ニュース記事には関連しないブログ記事」が順位付けの上位を占めるためである．このように，表 2(b) に例として示したニュース記事の場合は，E_{pos} のみに含まれる，比較的一般語に近いエントリを併用しない方が，関連ブログ記事収集の正解率が改善されることが分かる．

5.3　ブログ記事の日付の有無の比較

次に，ニュース記事中のトピックを表すキーワードの選定手法として $E_w^1 \cup E_w^2 \cup E_{pos}$ を用いた場合について，ニュース記事およびブログ記事の日付を考慮する場合，および，日付を考慮しない場合の比較を行う．

具体的には，ニュース記事およびブログ記事の日付を考慮する場合については，ニュース記事の日付以降，7 日以内の日付のブログ記事に限定して関連ブログ記事の収集を行った(注13)．日付の考慮の有無の比較結果を表 3 に示す．評価対象の全 20 ニュース記事全体での評価においては，日付を考慮しない方が高い正解率となった．なお，日付を考慮しない場合について，上位に順位付けされ，関連ブログ記事として判定されたブログ記事のうち，ニュース記事の日付から 8 日以降の日付となるブログ記事の割合は約 80%であった．逆に，日付を考慮する場合について，上位に順位付けされ，関連ブログ記事として判定されたブログ記事のうち，日付を考慮する場合にのみ判定対象となったブログ記事の割合は，約 65%であった．

ここで，ニュース記事ごとの傾向を分析するために，全 20 記事を，i) 日付の有無とは無関係に中程度の正解率 (50%) 以上となる 5 記事，ii) 日付の有無とは無関係に低い正解率 (20%以下) となる 4 記事，iii) 日付を考慮する場合に比べて，日付を考慮しない方が (20%以上) 高い正解率となる 10 記事，iv) 日付を考慮しない場合に比べて，日付を考慮する方が高い正解率となる 1 記事，に分類し，平均正解率を比較するとともに，ニュー

(注13)：各ブログホストについて，ブログ記事の HTML ファイルから日付を抽出するパターンを記述することにより，ブログ記事が投稿された日付の抽出を行った．

表 3　関連ブログ記事収集性能におけるブログ記事の日付の有無の比較
(キーワード選定手法: $E_w^1 \cup E_w^2 \cup E_{pos}$ の場合)

分類		ニュース記事数	平均正解率 (%)		ニュース記事の日付・タイトル (典型例) および 話題の特徴
			日付無	日付有	
日付の有無とは無関係	中正解率 (50%) 以上	5	84.0	76.0	2009 年 7 月 7 日「北朝鮮ミサイル発射を非難，国連安保理が「議長談話」」··· 報道直後の時期を含めて，ブロガーが幅広い期間に渡って関心を持つ． 2009 年 12 月 2 日「たばこ 1 箱千円に、がん対策協議会長が提言」··· ブロガーの関心が報道直後に集中し，日付を考慮した場合に上位に順位付けられる関連ブログ記事の半数前後は，日付を考慮しない場合も上位に順位付けられる．
	低正解率 (20%以下)	4	12.5	7.5	2009 年 6 月 3 日「有害微小物質、たばこの煙こもる店の 3 分の 1 で基準超え」··· 話題性が低いため，時期を問わず，ブロガーに関心を持たれない．
日付無の方が (20%以上) 高い正解率		10	76.0	23.3	2009 年 8 月 26「たばこの死者、世界で年 600 万人 . . . 米推計」··· ブロガーは，とりたてて報道直後の時期に集中して関心を持つというわけではなく，幅広い期間に渡って関心を持つ．
日付有の方が 高い正解率		1	10	80	2009 年 6 月 25 日「患者取り違え事故、後絶たず　05 年以降で 85 件」··· 日付無の場合でも近い話題のブログ記事が上位に順位付けられているが，ニュース記事における直接の話題 (「医療事故」) から少しずれるため，日付有の場合の方が密接に関連するブログ記事が多く上位に順位付けられる．
全体		20	62.0	36.1	—

ス記事の日付・タイトルの典型例，および，話題の特徴を表 3 に示す．

この結果から，i)∼iv) の分類ごとに，それぞれニュース記事の話題，および，ブログ記事における関心に動向の傾向が大きく異なることが分かる．例えば，i) においては，特にニュース記事の日付直後に関連ブログ記事が投稿される度合いが大きく，日付の考慮の有無によらず，比較的高い正解率で関連ブログ記事が収集できる．一方，ii) においては，ニュース記事の話題性が低いため，時期を問わず，ブロガーに関心を持たれない．また，全体の半数を占める iii) においては，ニュース記事の日付とは無関係に，幅広い期間に渡って関連ブログ記事が投稿されるため，日付を考慮せずブログ記事を収集した方が，正解率が高くなる．

6. おわりに

本論文においては，ニュース記事中において話題を表すキーワードを自動選定し，それらのキーワードに関して詳細な記述をしているブログ記事を収集する手法を提案した．特に，Wikipedia を知識源としてニュース記事中の話題に密接に関連するキーワードを選定する手法，および，ニュース記事のタイトルおよび冒頭部分からキーワードを選定する手法の比較を行った．評価実験を通して，ニュース記事中において話題を表すキーワードを自動選定する手法としては，Wikipedia を知識源とする手法とニュース記事中に位置を用いる手法を併用することにより，それぞれの手法が相補的に機能し，評価対象の全ニュース記事に対する総合的な性能が改善できることを示した．

謝辞　本論文で使用したニュース記事に関して協力して頂いた，北海道大学大学院情報科学研究科 吉岡真治准教授に感謝する．

文　献

[1] 新井嘉章, 福原知宏, 増田英孝, 中川裕志. Wikipedia を用いた多言語ブログ検索のための訳語抽出. 情報処理学会第 70 回全国大会講演論文集, 第 5 巻, pp. 55–56. 情報処理学会, 2008.

[2] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. Konig. Blews: Using blogs to provide context for news articles. In *Proc. ICWSM*, pp. 60–67, 2008.

[3] 池田大介, 藤木稔明, 奥村学. blog とニュース記事の自動対応付け. 言語処理学会第 11 回年次大会論文集, pp. 1030–1033, 2005.

[4] 石崎諒, 青野雅樹. Web ニュースに対するブログ意見の分析ツール. 電子情報通信学会技術研究報告, WI2-2008-52, pp. 11–12, 2008.

[5] 小原恭介, 山田剛一, 絹川博之, 中川裕志. Blogger の嗜好を利用した協調フィルタリングによる Web 情報推薦システム. 第 19 回人工知能学会全国大会発表論文集, 2005.

[6] 中崎寛之, 川場真理子, 横本大輔, 宇津呂武仁, 福原知宏. 多言語 Wikipedia エントリを知識源とする特定トピックの日英ブログサイト検索と日英対照ブログ分析. 人工知能学会論文誌, Vol. 25, No. 5, pp. 613–622, 2010.

[7] 佐藤由紀, 横本大輔, 宇津呂武仁, 福原知宏. Wikipedia を介した関連ニュース・ブログの対応付けにおける複数トピックの統合方式. 第 24 回人工知能学会全国大会論文集, June 2010.

[8] 田村悟之, 清田陽司, 増田英孝, 中川裕志. 図書館における自動レファレンスサービスシステムの実現 Web 上の二次情報と図書館の一次情報の統合. 情報処理学会研究報告, Vol. 2007, No. (2007–FI–179), pp. 1–8, 2007.

[9] 吉岡真治. 複数のニュース源の差異を考慮したニュース分析の研究. 言語処理学会第 13 回年次大会「大規模 Web 研究基盤上での自然言語処理・情報検索研究」 ワークショップ論文集, pp. 27–20, 2007.