DEWS2007 A8-4

TrackBackと特徴語に基づくBlogクローリングとBlog記事の推薦

鎌田　基之† 　福田　直樹† 　石川　博†

† 静岡大学情報学部情報科学科　〒 432–8011 静岡県浜松市城北 3–5–1
E-mail: †cs3029@s.inf.shizuoka.ac.jp, ††{fukuta,ishikawa}@inf.shizuoka.ac.jp

あらまし　最近注目されている Blog には，関連する Blog 記事同士をリンクするための TrackBack というメカニズムが存在し，複数の関連する Blog が TrackBack によって緩やかなコミュニティを形成している．本研究では，TrackBack が形成する緩やかなコミュニティを対象とした，TrackBack に基づく Blog 記事のクローリング手法を提案する．本研究では，TrackBack を辿るだけでは話題の混濁が起こるのを避けるため，特徴語によるフィルタリングを行うことで，ある特定の話題に限定した Blog クローリングを実現する．収集した Blog 記事集合に対して，TrackBack 数を用いたスコアリングを行うことで，ユーザへの有用な Blog 記事の推薦を行う．提案手法の有効性を確認するために，2006 年 11 月から 2007 年 2 月までに Blog で実際に起きた 4 つの話題に対して本手法 を適用し，その有効性を示す．
キーワード　Blog，TrackBack，クローリング，ランキング，推薦

Blog Crawling and Blog Entry Recommendation based on TrackBack and Characteristic Words

Motoyuki KAMADA†, Naoki FUKUTA†, and Hiroshi ISHIKAWA†

† Department of Computer Science, Faculty of Informatics, Shizuoka University　3–5–1 Jouhoku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
E-mail: †cs3029@s.inf.shizuoka.ac.jp, ††{fukuta,ishikawa}@inf.shizuoka.ac.jp

Abstract　Blogs have TrackBack mechanisms which can make links to relevant entries each other. Some related blog entries form lax blog communities by TrackBacks. In this paper, we propose a crawling method based on TrackBacks for finding such lax communities. First, we realize a crawling mechanism that is focused on gathering certain topics by characteristic words to avoid topic contamination occured in generic TrackBack-based crawling mechanisms. Second, we realize a recommendation mechanism to find out useful blog entries by TrackBacks-based scoring. We show the effectiveness of our proposed method in four recent popular topics on the blogs.
Key words　Blog, TrackBack, Crawling, Ranking, Recommendation

1. はじめに

1.1 背　　景

近年，Web2.0 というキーワードとともに個人の情報発信の場が個々のウェブサイトから Blog へと変化してきている．平成 18 年 4 月の総務省による発表によると，Blog 登録者数が平成 18 年 3 月末現在で 868 万人に達している [1]．この貴重な情報源である Blog からの効果的な情報の収集は重要な課題となっている．

Blog には，関連する Blog 記事同士をリンクするための，TrackBack という Blog 固有の機能がある（図 1）．TrackBack では，相手の Blog 記事に自分の記事から参照リンクを張った際に，自分の Blog 記事から相手の Blog 記事へ TrackBack ping を送信することで，参照リンクを張ったことを相手へ通知し，その結果として相手の Blog 記事からのリンクを得ることができる．本論文では，以降，TrackBack ping の送信元である Blog 記事を **TrackBack 元の Blog 記事**，TrackBack ping の送信先である Blog 記事を **TrackBack 先の Blog 記事**と表現することとする．初期の Blog では，TrackBack ping を送信する際には，TrackBack 先の Blog 記事への参照リンクを含めることが通常であった．しかし，最近では，参照リンクの有無に関わらず，関連する内容を書いているという意味でも TrackBack が利用されている．TrackBack 先の Blog 記事は，自らの Blog 記事と関連する内容の Blog 記事を知ることができ，TrackBack 元の Blog 記事は TrackBack 先からのリンクを得ることができる．以上のことから TrackBack 先の Blog 記事との関係について以下の 2 点が仮定できると考えられる．

- TrackBack 元の Blog 記事の著者から，自分の Blog 記

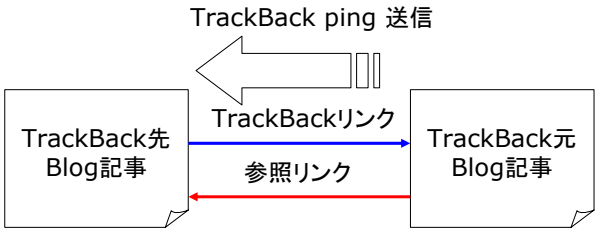

図 1　TrackBack

事との関連性の高さが認められている

- TrackBack 元の Blog 記事の著者から，自分の Blog 記事がリンクを得たい Blog 記事であると認められている

TrackBack ping を TrackBack 元の Blog 記事の著者からの推薦と捉えることで，Blog 記事著者間での重要度を測ることができると考えられる．すなわち,「他の Blog 記事著者からみた重要度」は TrackBack によって評価することができると考えられる．

Blog から情報を収集する手段としては，Google ブログ検索 [2] などの Blog サーチエンジンが挙げられる．Blog サーチエンジンを利用することで，検索キーワードに関連性のある Blog 記事を検索することができる．しかし，Blog 固有の性質を考慮し，TrackBack によって形成される相互評価のメカニズムを最大限に活用した情報の収集や，TrackBack によって形成される緩やかな構造そのものを効果的に収集するためのメカニズムの実現には，まだ改良の余地がある．

1.2　目　　的

本研究では，関連した Blog 記事の間で利用されている TrackBack に基づいたクローリングをすることで，ある特定の話題に限定した Blog 記事集合を収集するための手法を提案する．話題の特徴語を用いたフィルタリングを行うことで話題に限定したクローリングの精度を上げる．

また,「他の Blog 記事著者からみた重要度」を TrackBack より算出することで Blog 記事のスコアリングを行い，その結果を用いた Blog 記事の推薦を実現する．

2.　関 連 研 究

既存の Blog 記事の収集やクローリングは，検索エンジンに利用することを目的としているものが多く，日々多くの Blog 記事を収集している．収集の対象となる Blog 記事の発見には，主に以下の方法が用いられている．

- Blog サービスの提供する Blog 記事更新情報からの抽出
- Ping server の更新情報からの抽出

また，Blog 記事の収集には，主に以下の手法が用いられている．

- RSS 内の情報に基づく収集
- HTML の解析に基づく収集

RSS では，最新の記事から 10 記事程度のみに限定して 1 つの RSS として配信されている場合が多い．このため，RSS に基づく収集手法では，RSS が配信されるごとの記事収集処理コストは低いものの，過去の Blog 記事は収集の対象とならない．

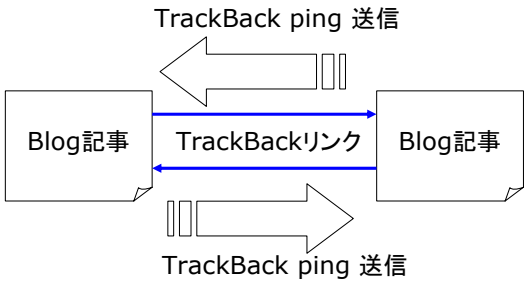

図 2　相互 TrackBack

一方，HTML から解析した情報に基づく手法は，トップページからリンクを辿って，ディレクトリ以下に存在する HTML 文書をすべて収集することで，全ての Blog 記事を収集することが可能であるが，対象となるページが膨大になるため各ページごとの解析のための処理コストが高い．

南野らは，Blog 記事を網羅的に収集し，監視するシステムの提案を行っている [3]．南野らの手法では，Blog 記事の発見については，WWW 全体を対象としたクローリングや Blog リンク集，ping サーバの更新情報を利用したクローリングによって Web ページを得てから，得られた Web ページを Blog であるか個別に判定することで Blog 記事の発見を行っている．Blog 記事の収集については，HTML を直接解析することで行っている．

井原らは，画像情報を含む Blog 記事の収集とそれらを検索するシステムを構築している [4]．井原らは，Blog 記事の発見を，Blog サービスの提供する Blog 記事更新情報の Web ページを巡回し，Blog トップページの URL を抽出することで行っている．Blog 記事の収集については，初回のみ HTML を直接解析して収集し以降は RSS を利用し，新しい Blog 記事のみを取得している．

本研究では，Blog 記事の発見を，TrackBack 元の Blog 記事を辿ることで行うことで，ある特定の話題に限定した Blog 記事収集を行う．また Blog 記事の収集は，HTML を直接解析することで行うことで，RSS には通常記述されない TrackBack 情報の収集を行う．

Blog 記事の TrackBack に関する研究として，中島らは，TrackBack 利用状況の調査を行うことで，TrackBack リンクで繋がった Blog 記事の関係について考察している [5]．中島らは，参照リンクを伴わない TrackBack を空 TrackBack と定義し，空 TrackBack が，解析対象とした TrackBack のうち全体の 99%以上であったことを明らかにしている．また，Blog 記事の一時的なコミュニティ形成とみなすことができるブログスレッドを提案し，ブログスレッド形成における TrackBack の重要性を明らかにしており，また，相互の空 TrackBack によるコミュニティ形成が行われていることも明らかにしている．

本研究では，中島らの指摘した TrackBack によるコミュニティ形成能力に着目し，TrackBack を辿ることで，ある話題に限定した Blog 記事のクローリングを行う．さらに，参照リンクを考慮しない相互 TrackBack（図 2）に着目し，片方向の

図 3　処理の概要

TrackBack より相互 TrackBack による繋がりに重みを置いた Blog 記事のスコアリングを行うことで Blog 記事の推薦を行う．

一般に，Web のクローリングにおいては，フォーカストクローリングという手法が存在する．Chakrabarti らは，Web クローリングにおいて話題に関連のあるページを優先的に選択し収集する Web 資源の探索方法を提案している [6].

本研究では，Web を対象とするこの考えを Blog に適用することで，話題に限定した Blog クローリングを行う．着目した話題に限定するための特徴語を指定し，その特徴語に基づきクローリングを行う．

Web におけるリンク構造を解析することで Web ページの重要度を測定するものとして，PageRank アルゴリズム [7] や HITS アルゴリズム [8] などが存在する．PageRank アルゴリズムでは，リンクを支持投票とみなし，リンクを受けたページの重要度を測定する．

本研究では，通常のリンクではなく，Blog 固有の機能である TrackBack を Blog 記事への支持投票とみなすことで，Blog 記事へのスコアリングを行う．

3. 提案アプローチの概要

本論文で提案する，TrackBack に基づく Blog クローリングと Blog 記事の推薦の手順を図 3 に示す．

最初に，TrackBack に基づく Blog 記事のクローリングを行う．そのクローリングの過程で同時に，話題の限定と TrackBack スパム対策のためのフィルタリングを行い，対象とする話題に限定された Blog 記事の収集を行う．続いて，収集された Blog 記事の集合に対して TrackBack 数に基づいたスコアリングを行う．最後に，収集された Blog 記事集合内でのスコアが高い Blog 記事を推薦する．

3.1 TrackBack に基づく Blog クローリング

従来の Blog クローリングでは，Ping server や Blog サービスの更新情報を用いた Blog 記事の発見と収集が行われている．しかし，本研究でのクローリング対象の Blog 記事は，TrackBack 先か，あるいは TrackBack 元の Blog 記事に限定している．そのため，本手法では TrackBack リンクを利用することで Blog 記事のクローリングを行う．また，本手法では Blog 記事の TrackBack 情報を必要とするため，Blog 記事の HTML を解析することで TrackBack 情報を抽出する．通常，TrackBack は Blog 記事間で使用されるものであるため，ある Blog 記事の TrackBack 元である TrackBack リンク先もまた Blog 記事であると仮定できる．本クローリング手法ではクローリング中の Web ページが Blog 記事であるかどうかを判定しない．

```
crawling ( Blog b ) {
  b の HTML を解析し，すべての TrackBack 元を抽出する;
  解析した情報を格納する;
  foreach tb_origin ( TrackBack 元集合 ) {
    if ( tb_origin が未解析 ) {
      crawling( tb_origin );
    }
  }
}
```

図 4　クローリングアルゴリズム（話題絞り込みなし）

表 1　Blog 記事の内容 (500 件)

内容	いじめ	政治国際	MNP（携帯）	スポーツ	その他
件数	17	131	171	112	69

クローリング手法として，図 4 に，話題による対象絞り込みを含まない，TrackBack に基づく一般的なクローリングのアルゴリズムを示す．最初に元となる Blog 記事の HTML を解析し，TrackBack 情報を抽出する．TrackBack 元 URL や TrackBack 数などを格納した後，続けて TrackBack 元のクローリングを行う．ここで，TrackBack がない場合や対象が既に解析済みの TrackBack 元であった場合はクローリングを行わず前の処理へと戻る．対象が初めて解析する TrackBack 元であった場合は，その HTML を解析し，抽出した TrackBack 元を対象としたクローリングを続けて行う．これを再帰的に繰り返すことで，TrackBack に基づいた Blog クローリングは実現される．ここで収集する情報は，各 Blog 記事の URL，TrackBack 数，TrackBack 元 URL である．

話題による対象絞り込みを含まない，TrackBack に基づく一般的なクローリングの効果の検証のための予備実験として，図 4 のアルゴリズムを用いた Blog クローリングを行った（実験 A）．Seed Blog 記事（クローリングの元となる Blog 記事）には，「いじめ，自殺」に関連する Blog 記事を用いた．この Seed Blog 記事の TrackBack 数は 4 件であった．この Seed Blog 記事を起点に，TrackBack に基づくクローリングを行った．本予備実験では 2000 件を収集した時点でクローリングを停止し，そのうち 500 件の Blog 記事の内容を調査した．その 500 件の Blog 記事の実際の内容の内訳を表 1 に示す．Seed Blog 記事が「いじめ，自殺」に関連するものであったにも関わらず，特に「政治」や「MNP（携帯）」や「スポーツ」に関連する Blog 記事が多くみられた．また「その他」の一部には TrackBack スパムと思われる Blog 記事もみられた．

ここで，図 4 に示したクローリング方法では，Seed Blog 記事の話題だけでなく，他の話題を扱う Blog 記事の多くをクローリング対象とする場合があることがわかった．しかし，話題の混濁が起こっているものの，各話題はある程度のまとまりとなって収集されている．

TrackBack 元を起点としたクローリングは，収集した対象が Blog 記事であるかどうかの判定が省略可能であることと，話題のまとまりごとの収集が可能であることの，2 つの利点があ

ると考えられる．

3.2 フィルタリング

表 1 では 2000 件中 500 件の結果を示しているが，Seed Blog 記事の話題「いじめ，自殺」からは全く関連性のみられない携帯電話の番号ポータビリティ（MNP）やスポーツの話題までも収集されている．この原因として考えられることは，Blog 記事では 1 つの記事内で複数の話題を取り上げることが少なくないということである．複数の話題を扱った TrackBack 元を起点とした収集を行っているうちに話題の混濁が起きることがわかった．本研究では，着目した話題に限定した Blog 記事集合を収集することを目的としているため，話題の混濁した Blog 記事集合が収集されることは好ましくない．したがって，クローリング対象を，着目した話題に限定するため，収集の対象とする Blog 記事を,「ユーザが設定した特徴語を Blog 記事の本文に含む Blog 記事」と定義する．この定義に当てはまる Blog 記事のみを収集の対象に含めることでフォーカストクローリングを実現する．なお，特徴語は複数個の設定を可能とし，設定された特徴語がすべて含まれている Blog 記事のみを収集の対象とする．

TrackBack 元を辿る際の問題として，TrackBack スパムの存在が挙げられる．TrackBack スパムとは，相手の Blog 記事の内容とは無関係な TrackBack を送信することで，自分の Blog 記事へのリンクを得るものである．本論文では TrackBack スパムを,「特徴語が Blog 記事の本文に含まれない，TrackBack 先の Blog 記事の内容とは無関係な内容を持つ TrackBack 元の Blog 記事」と定義する．

TrackBack スパムの送信は自動的にリンクを得られることから，読者の多い著名な Blog 記事などに対して行うことで，自分の Blog 記事へと読者を誘導することを目的としている．これは TrackBack に基づいたクローリングを行う上で排除すべき存在である．その理由は以下の 2 つである．

- 話題の混濁が起こり，話題に限定した収集が不可能
- 不必要な Blog 記事のクローリングによる，クローリング時間の増加

この TrackBack スパムには，前述したフォーカストクローリングを行うことで対処できる．TrackBack スパムは多くの場合，TrackBack 先とは記事の内容が全く関連しないため，TrackBack スパムの本文には特徴語が含まれていないと考えられる．したがって，前述の特徴語を用いたフォーカストクローリングを行うことで，同時に TrackBack スパムの排除も可能である．

図 5 に図 4 のアルゴリズムを改良し，フィルタリングを追加したフォーカストクローリングアルゴリズムを示す．

特徴語によるフォーカストクローリングの効果の検証のための予備実験として，先ほどの実験 A で用いたものと同じ Seed Blog 記事を用い，特徴語（「いじめ」と「自殺」）を用いたフィルタリングを行い，クローリングを行った（実験 B）．収集した Blog 記事は 158 件であり，すべて特徴語（「いじめ」と「自殺」）が含まれた Blog 記事が収集された．その内容を表 2 に示す．表 2 中の「いじめ・政治」は，主に教育基本法改正といじめ問題についての Blog 記事,「いじめ・マスコミ」は，主にマスコミによるいじめ・自殺報道の問題点についての Blog 記事,「自殺・履修問題」は，主に高校の履修不足問題による自殺についての Blog 記事である．また,「一日のニュース」とは，その日に起こった複数の出来事を 1 つの Blog 記事にまとめて書かれているもので,「いじめ・自殺」の内容だけでなく，それ以外の内容も含まれている．収集した Blog 記事には「いじめ」や「自殺」という特徴語が含まれてはいるものの,「自殺・履修問題」や「1 日のニュース」のように，内容のすべてが「いじめ・自殺」でないものもいくつか見られたが，大部分については「いじめ・自殺」の話題を扱う Blog 記事が収集された．

```
crawling( Blog b , Topic t ){
  if ( b の本文に t が含まれる ) {
    b の HTML を解析し，すべての TrackBack 元を抽出する;
    解析した情報を格納する;
    foreach tb_origin ( TrackBack 元集合 ) {
      if ( tb_origin が未解析 ) {
        crawling( tb_origin, t );
      }
    }
  }
  else { b を解析済みとする; }
}
```

図 5　特徴語によるフォーカストクローリングアルゴリズム

表 2　Blog 記事の内容 (フィルタあり) (158 件)

内容	いじめ・自殺	いじめ・政治	いじめ・マスコミ	自殺・履修問題	1 日のニュース
件数	93	43	4	13	5

特徴語を用いたフィルタリングを行うことは，着目した話題に限定した Blog 記事のクローリングを行う上で，着目した話題の Blog 記事集合の収集が可能であることと，TrackBack スパムの排除が可能であることの，2 つの利点があると考えられる．

3.3 Blog 記事のスコアリング

TrackBack は一方的に作成可能であり，TrackBack 元の Blog 記事が TrackBack ping を送信すれば，TrackBack 先の Blog 記事から自動的にリンクを得ることができる．この性質を利用し TrackBack は，TrackBack 元の Blog 記事にとって，TrackBack 先の Blog 記事から自らへのアクセスを増やすという目的のためにも用いられている．すなわち，TrackBack 先の Blog 記事が人気で有名であればあるほど TrackBack されやすいと仮定できる．人気で有名であるということは，内容もそれに従って充実していると考えられる．この仮定に基づき，本手法では，TrackBack を Blog 記事著者による推薦と捉えることでスコアリングを行う．

TrackBack の使い方の 1 つとして，現在，TrackBack を受けた TrackBack 先の Blog 記事から，TrackBack 元の Blog 記事へ TrackBack することによって，お互いの記事を TrackBack により自動的に作られたリンクで相互にリンクし合うという

表 3　スコアリング（相互 TrackBack 数: α=1.0, β=0）

識別子	$BTB(n)$	$TB(n)$	$Score(n)$	内容
105	13	9	13	いじめ，自殺
76	11	16	11	教育全般
138	9	7	9	いじめ，教育再生会議
55	8	77	8	タウンミーティング
52	7	56	7	教育基本法強行採決
14	7	7	7	いじめ，自殺予告
31	6	20	6	いじめ，教育基本法
33	6	12	6	履修問題，校長自殺
21	6	7	6	いじめ
140	6	5	6	いじめ，教育再生会議
137	6	3	6	いじめ，教育再生会議

ことが行われている．これを本論文では相互 TrackBack と呼ぶ．相互 TrackBack は，Blog 記事著者がお互いの Blog 記事を推薦し合っているという状態である．本論文では，TrackBack を，相互 TrackBack とそれ以外の片方向 TrackBack の 2 種類に分けて考えることとする．片方向 TrackBack とは，相互 TrackBack を除いた Blog 記事が受けている TrackBack と定義する．すなわち，相互 TrackBack 数と片方向 TrackBack 数の合計が，Blog 記事の受けている TrackBack 数となる．

相互 TrackBack が存在するということは，著者同士が互いの Blog 記事の関連性が高いことを認め合っていると解釈できる．本手法では，この仮定に基づき，相互 TrackBack は片方向 TrackBack に比べ，重要度を高くする．本手法では，片方向 TrackBack よりも相互 TrackBack 数が多いほどスコアが高くなるように重みづけを行う．本手法でのスコア計算の式を式 (1) に示す.

$$Score(n) = \alpha \times BTB(n) + \beta \times TB(n) \quad (\alpha > \beta) \qquad (1)$$

式 (1) における α，β は，重みづけ係数である．$BTB(n)$ は Blog 記事 n の持つ相互 TrackBack 数，$TB(n)$ は Blog 記事 n の持つ片方向 TrackBack 数とする．求められた $Score(n)$ が高い Blog 記事ほど，その重要度が高い記事とする．

式 (1) における重みづけ係数の検討のための予備実験として，先程の実験 B で得られた Blog 記事集合に対して，式 (1) を用いスコアリングを行った．ここで重みづけ係数である α，β については，α=1.0，β=1.0 の TrackBack 数を用いた場合，α=1.0，β=0 の相互 TrackBack 数のみを用いた場合，α=0，β=1.0 の片方向 TrackBack 数のみを用いた場合，の 3 パターンで実験を行う．

この 3 パターンで得られた結果のスコア 10 位までの Blog 記事をそれぞれ表 3，表 4，および表 5 に示す．表 3，表 4，および表 5 中の「識別子」は，収集した Blog 記事に付けられた一意な番号である.「$BTB(n)$」は相互 TrackBack 数,「$TB(n)$」は片方向 TrackBack 数であり,「$Score(n)$」は各パターンでの α，β の値を用いて式 (1) により算出した値である．また,「内容」という項目は Blog 記事のタイトル・本文から判断した最もその Blog 記事で語られていた話題を示している．

実験 B で得られた Blog 集合に対して，片方向 TrackBack 数のみを用いた場合（表 5）と TrackBack 数を用いた場合（表 4）については大きな違いが見られなかった．これは，片方向 TrackBack 数が，相互 TrackBack 数に比べ，非常に多かったためである．したがって，表 4，表 5 それぞれに現れている Blog 記事は 9 位までは同じである．しかし，TrackBack 数を用いた場合の第 10 位には，片方向 TrackBack 数のみを用いた場合には現れなかった,「いじめ，自殺」を扱った Blog 記事が現れている．TrackBack 数を用いた場合において,「いじめ，自殺」を扱った識別子 146 の Blog 記事の順位は 3 つ下がっているものの,「いじめ，自殺」を扱った Blog 記事が第 10 位までで 2 つ現れたことは好ましい結果であると考えられる．片方向 TrackBack 数のみを用いるのと比較して，相互 TrackBack 数も含めた場合のほうが良い結果が得られると考えられる．

また，相互 TrackBack 数のみを用いた場合（表 3）については，片方向 TrackBack 数を無視しているため，今回の実験対象では非常に多かった片方向 TrackBack 数の影響を受けていない．TrackBack 数を用いた場合には第 10 位であった識別子 105 の Blog 記事が，表 3 では第 1 位となっている．識別子 105 の Blog 記事は「いじめ・自殺」を扱ったものである．さらに，TrackBack 数を用いた場合（表 5）と比べ,「いじめ・自殺」を扱った Blog 記事が上位に位置していると考えられる．特徴語を「いじめ」と「自殺」として収集した Blog 記事集合において,「いじめ・自殺」を扱っていた Blog 記事が上位に選ばれたことは好ましい結果であると考えられる．そのため単純に TrackBack 数を用いる方法と比較して，相互 TrackBack 数を用いた場合のほうが良い結果が得られると考えられる．

表 4　スコアリング（TrackBack 数: α=1.0, β=1.0）

識別子	$BTB(n)$	$TB(n)$	$Score(n)$	内容
92	2	90	92	教育基本法強行採決
55	8	77	85	タウンミーティング
52	7	56	63	教育基本法強行採決
63	0	56	56	教育基本法強行採決
56	2	31	33	教育基本法強行採決
76	11	16	27	教育全般
31	6	20	26	いじめ，教育基本法
142	3	21	24	教育基本法，安部内閣
146	1	23	24	いじめ，自殺
105	13	9	22	いじめ，自殺

表 5　スコアリング（片方向 TrackBack 数: α=0, β=1.0）

識別子	$BTB(n)$	$TB(n)$	$Score(n)$	内容
92	2	90	90	教育基本法強行採決
55	8	77	77	タウンミーティング
52	7	56	56	教育基本法強行採決
63	0	56	56	教育基本法強行採決
56	2	31	31	教育基本法強行採決
146	1	23	23	いじめ，自殺
142	3	21	21	教育基本法，安部内閣
31	6	20	20	いじめ，教育基本法
76	11	16	16	教育全般
100	2	15	15	履修問題

4. 評価実験

4.1 指　　針

評価実験として，次に挙げる話題に限定した Blog 記事のクローリングを行い，収集された Blog 記事集合に対して TrackBack に基づいたスコアリングを行った．本評価実験で用いる Seed Blog 記事は，livedoor NEWS [9] に TrackBack を送信し，リンクを得ている Blog 記事とした．livedoor NEWS では，日々配信されるニュース記事に対して，Blog 記事から TrackBack を受け付ける機能を提供している．

それでは今回の評価実験に用いた Seed Blog 記事の話題は，以下の 3 つである．

（1） 不二家 洋菓子販売を全面休止

- 2007/01/15 時点での TrackBack: Blog 記事 16 件
- 2007/01/11 livedoor NEWS - ライブドア・ニュース

（2） フジ『発掘！あるある大事典 II』の納豆特集で捏造

- 2007/01/22 時点での TrackBack: Blog 記事 40 件
- 2007/01/20 livedoor NEWS - PJnews

（3） 宮崎知事選：そのまんま東氏が初当選

- 2007/01/26 時点での TrackBack: Blog 記事 97 件
- 2007/01/22 livedoor NEWS - 毎日新聞

それぞれの Seed ニュース記事に対して TrackBack している Blog 記事すべてを，Seed Blog 記事として利用した．利用した Seed Blog 記事には，重複したものや TrackBack スパムと言えるようなものが存在する場合がある．特徴語によるフィルタリングの効果を確かめるため，これらのものもそのまま用いることとした．

スコアリングの式については，前節の予備実験の結果等から検討した結果，重みづけ係数 α を 4.0，β を 1.0 とした式 (2) で行った．

$$Score(n) = 4.0 \times BTB(n) + 1.0 \times TB(n) \qquad (2)$$

評価にあたって，本手法によって得られた Blog 記事の内容を著者らが主観的に判断することも可能であるが，実験結果の公平性を高めるため，できるだけ客観的かつ計測可能な指標として，以下の 5 つを実験結果の有効性の評価指標として用いることとした．

- オフィシャルサイト（公式サイト）へのリンク
- ニュース記事へのリンク数
- ニュース記事の引用数
- その他関連する情報へのリンク数
- 掲載画像数

それぞれプレーンテキスト以上に情報量を持っており，Blog 記事の内容を測る指標として挙げられるものと考える．特に，今回扱った話題ではオフィシャルサイトと呼べるものが存在するため，別途オフィシャルサイトへのリンクを取り上げることとした．ただし，上記の 5 つの指標はあくまでも結果の評価の目安として用いるためのものであり，実験に用いたフォーカストクローラ自体はこれらの指標を一切利用していない．これら 5 つの指標で高い評価値を持つものを，本論文では「情報量の多い」Blog 記事と呼ぶことにする．

表 6　不二家：Blog 記事の内容 59 件

内容	不二家 洋菓子	不二家 その他全般	日記
件数	23	35	1

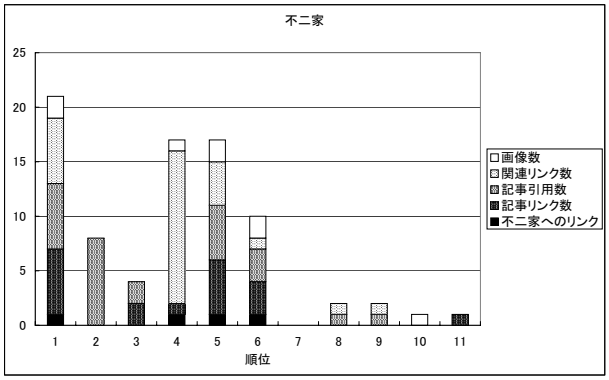

図 6　不二家 片方向 TrackBack 数

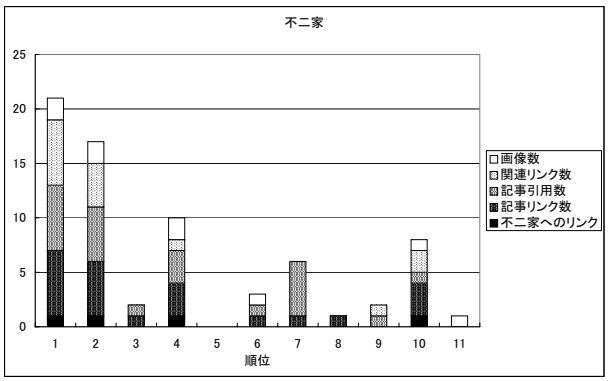

図 7　不二家 提案方式

4.2 話題：不二家 洋菓子販売を全面休止

クローリングの際のパラメータは，以下のように定めた．

- Seed Blog 記事: 16 件
- 特徴語: 不二家

このクローリングで得られた Blog 記事は 59 件であった．収集された Blog 記事の内容を表 6 に示す．

表 6 の,「不二家 洋菓子」は Seed ニュース記事について書かれた Blog 記事,「不二家 その他全般」は「不二家 洋菓子」以外の一連のニュースについて書かれた Blog 記事,「日記」は本文のほとんどが日記であって，この話題については軽く触れている程度の Blog 記事を，それぞれ示している．

収集された Blog 記事は，すべて「不二家」の話題について触れたものであった．TrackBack スパムと思われるような Blog 記事は含まれておらず，着目した話題に限定したクローリングが行われたと言える．

次に収集した Blog 記事に対して，式 (2) を用いて算出したスコアに基づいて順位付けた．その第 10 位までの Blog 記事の内容を表すグラフを図 7 に，比較対象として片方向 TrackBack 数で順位付けたものを図 6 に示す．

図 7 の提案方式のものと図 6 の片方向 TrackBack 数のものでは，10 位までに現れる Blog 記事に違いが見られた．第 1 位

表 7　あるある大事典捏造：Blog 記事の内容 127 件

内容	あるある 納豆・捏造	ダイエット	マスメディア	1 日の ニュース	日記
件数	117	2	4	2	2

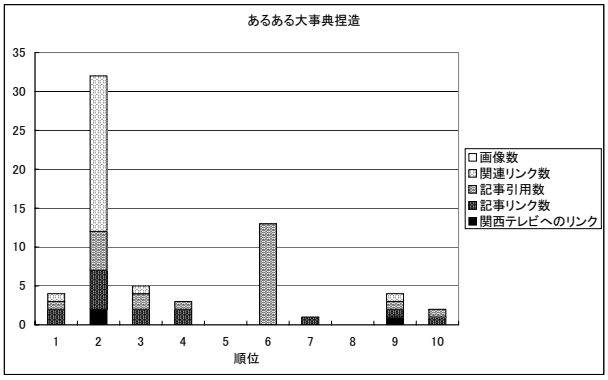

図 8　あるある大事典 片方向 TrackBack 数

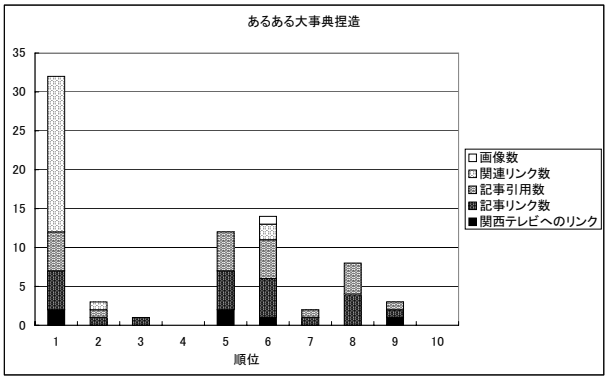

図 9　あるある大事典 提案方式

に現れる Blog 記事は同様のものであったが，それ以外は順位が異なっているか片方では現れなかった Blog 記事が現れている．片方向 TrackBack 数のものでは第 5 位，第 6 位に比較的情報量の多い Blog 記事が現れているが，提案方式のものではそれぞれが順位を上げ，第 2 位，第 4 位に現れている．また，第 7 位〜第 10 位の Blog 記事においても，片方向 TrackBack 数のものよりも提案方式のものが情報量の多い Blog 記事が現れており，良い結果となっている．

4.3　話題：フジ『発掘！あるある大事典 II』の納豆特集で捏造

クローリングの際のパラメータは，以下のように定めた．

- Seed Blog 記事: 40 件
- 特徴語: あるある，納豆，造（ねつ造，捏造）

このクローリングで得られた Blog 記事は 127 件であった．その収集された Blog 記事の内容を表 7 に示す．

表 7 の,「あるある 納豆・捏造」は Seed ニュース記事について言及した Blog 記事,「ダイエット」はこの話題について触れつつダイエットについて書かれた Blog 記事,「マスメディア」はマスメディアの捏造について書かれた Blog 記事,「1 日のニュース」はこの話題も含めた複数のニュースが書かれた Blog 記事,「日記」はこの話題について軽く触れている程度の Blog 記事である．

ここでも，収集された記事は，すべて「あるある 納豆・捏造」の話題について触れたものであった．TrackBack スパムと思われるような Blog 記事は含まれておらず，着目した話題に限定したクローリングが行われたと言える．

次に収集した Blog 記事に対して，式 (2) を用いて算出したスコアに基づいて順位付けた．その第 10 位までの Blog 記事の内容を表すグラフを図 9 に，比較対象として片方向 TrackBack 数で順位付けたものを図 8 に示す．

この話題においても，図 9 の提案方式のものと図 8 の片方向 TrackBack 数のものでは，10 位までに現れる Blog 記事に違いが見られた．片方向 TrackBack 数のものでは第 2 位，第 6 位に比較的情報量の多い Blog 記事が現れている．ここで提案方式のものでは，片方向 TrackBack 数のもので第 2 位にあった Blog 記事が順位を上げ第 1 位となっている．加えて，提案方式のものでは，片方向 TrackBack のものに比べて第 5 位〜第 9 位にかけて，ある程度情報量の多い Blog 記事が現れ，良い結果となっている．

4.4　話題：宮崎知事選：そのまんま東氏が初当選

クローリングの際のパラメータは，以下のように定めた．

- Seed Blog 記事: 97 件
- 特徴語: そのまんま，宮崎，知事，当選

このクローリングで得られた Blog 記事は 124 件であった．その収集された Blog 記事の内容を表 8 に示す．

表 8 の,「そのまんま東氏 当選」は Seed ニュース記事について言及した Blog 記事,「当選後」はこの話題の翌日のニュースについて書かれた Blog 記事,「出馬」はこの話題の 1ヶ月ほど前に書かれた Blog 記事,「1 日のニュース」はこの話題も含めた複数のニュースが書かれた Blog 記事,「日記・その他」はこの話題について軽く触れている程度の日記や主にアイドルやその他の情報を扱っている Blog 記事などである．

ここでも，収集された記事は,「そのまんま東氏 当選」の話題を含んだものであった.「そのまんま東氏が当選した」程度の情報しか書かれていない Blog 記事が「日記・その他」に存在したが，全く話題に触れていない Blog 記事ではなかった．したがって，着目した話題に限定したクローリングが本手法により効果的に行われたといえる．

次に,「そのまんま東氏 当選」の話題において収集した Blog 記事に対して，式 (2) を用いて算出したスコアに基づいて順位付けた．その第 10 位までの Blog 記事の内容を表すグラフを図 11 に，比較対象として片方向 TrackBack 数で順位付けたものを図 10 に示す．

この話題においても，図 11 の提案方式のものと図 10 の片方向 TrackBack 数のものでは，10 位までに現れる Blog 記事に違いが見られた．片方向 TrackBack 数のものでは，第 5 位，第 7 位に比較的情報量の多い Blog 記事が現れている．提案方式のものでは，片方向 TrackBack 数のもので第 7 位にあったものが順位を上げ，第 5 位となっている．加えて，提案方式のものでは，片方向 TrackBack 数のもので第 5 位にあった Blog 記事が第 10 位までには現れていないが，第 2 位にはそれ以上に

表 8　そのまんま東氏当選：Blog 記事の内容 124 件

内容	そのまんま東氏 当選	当選後	出馬	知事選	1 日のニュース	日記・その他
件数	104	6	1	2	4	7

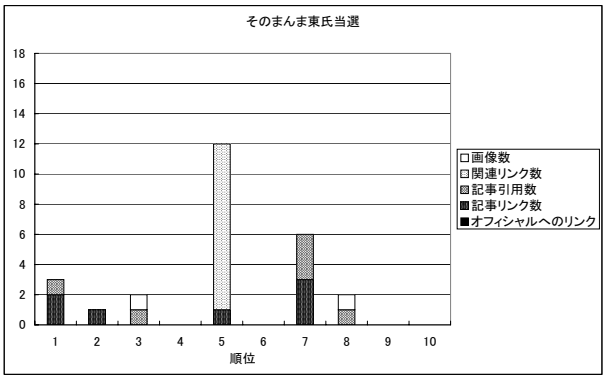

図 10　そのまんま東氏当選 片方向 TrackBack 数

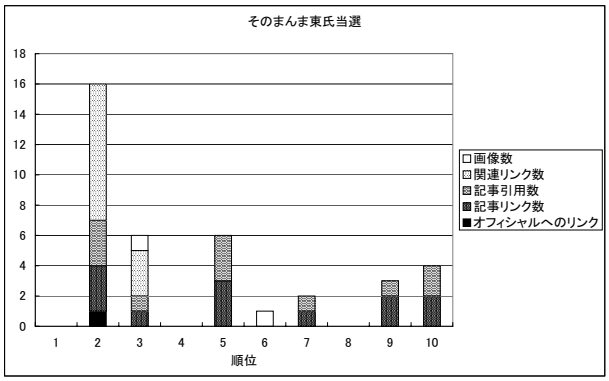

図 11　そのまんま東氏当選 提案方式

情報量の多い Blog 記事が現れている．さらに，提案手法のものでは第 3 位に片方向 TrackBack 数のものにはなかった Blog 記事が現れ，良い結果となっている．

5. おわりに

TrackBack に基づく Blog クローリングの手法とその Blog 記事のスコアリングを提案した．TrackBack を利用した Blog 記事のクローリングを可能とし，フィルタリングを行うことで，着目した話題に限定してクローリングを行う手法を提案した．また，3 つの話題を対象に評価実験を行い，相互 TrackBack 数に重みを置いた TrackBack に基づくスコアリングの有効性を示した．

今後の課題としては以下のことが挙げられる．

- クローリング精度の向上

本研究のクローリングにおいては，TrackBack の抽出は Blog サービス毎に HTML の解析を行うことで TrackBack 元 Blog 記事の URL や TrackBack 数の抽出を行っている．またフィルタリングにおける本文の特定においても同様である．このままの方法では Blog サービスの仕様変更に逐一対応しなければならない．ある程度の変更に耐えうる柔軟な TrackBack の抽出方法が必要である．そのため，Blog サービスに依存しない抽出方法の確立も検討する．

- 重みづけとスコアリングの評価

重みづけ係数については，α=4.0，β=1.0 で評価実験を行った．今回のこの結果は今回扱った 3 つの話題に関する Blog 記事集合における結果であるため，他の Blog 記事集合の場合における実験を行い，さらなる検討を行っていく．また，今回の評価実験においては本研究で定義した片方向 TrackBack 数のものとの比較を行った．今後は従来 Blog サーチエンジンでの結果の提示に用いられている TF・IDF ベースのスコアリングとの比較も検討する．

- スコアリング方法の改善

本論文で提案したスコアリング方法では，片方向 TrackBack の意味について，深い考慮をしていない．例えば，「TrackBack は受けたが相手に TrackBack はしない」といったようなマイナス評価のような Blog 記事著者の意思も考えられる．そのような点も考慮したスコアリングの改善が考えられる．

参考文献

[1] ブログ及び SNS の登録者数（平成 18 年 3 月末現在），総務省報道資料（平成 18 年 4 月 13 日），http://www.soumu.go.jp/s-news/2006/060413_2.html，accessed 2006.12.11.

[2] Google ブログ検索，http://blogsearch.google.co.jp/

[3] 南野朋之，鈴木泰裕，藤木稔明，奥村学，“blog の自動収集と監視”，人工知能学会論文誌，Vol.19，No.6，pp.511-520，2004.

[4] 伊原 伸介，林 貴宏，尾内 理紀夫，“もぶろげっと: 画像情報を含む Blog 記事検索システム”，インタラクティブシステムとソフトウェアに関するワークショップ (WISS2005) 論文集，pp. 69-74，2005.

[5] 中島伸介，舘村純一，原良憲，田中克己，植村俊亮，“ブログ空間におけるトラックバック利用状況の調査および考察”，DEWS2006 1B-i6，2006.

[6] Soumen Chakrabarti，Martin van den Berg，Byron Dom，“Focused crawling: a new approach to topic-specific Web resource discovery”，Proc. 8th International World Wide Web conference，1999.

[7] L. Page，S. Brin，R. Motwani，T. Winograd，“The PageRank Citation Ranking: Bringing Order to the Web”，Stanford Digital Libraries Working Paper，1998

[8] Jon M. Kleinberg，“Authoritative Sources in a Hyperlinked Environment”，Journal of the ACM，vol.46，no.5，pp.604-632，1999

[9] livedoor NEWS，http://news.livedoor.com/