D-044

Blog クラスタリングのための関連ページ選択方法
Selection methods of relevant Web pages for blog clustering

戸田 雄士† 太田 学†
Yuji Toda Manabu Ohta

1．はじめに

近年注目されている情報源の一つに Blog（Weblog）がある．速報性，リアルタイム性のある新鮮な情報が発信されるという点で通常の Web ページとは異なり，有用な情報源として考えられるようになっている．我々は Blog 独自の機能である Trackback を利用することでこの Blog の検索結果に特化したクラスタリング手法の提案を行った[1]．この手法では「Trackback によって繋がっているページは内容的に関連がある場合が多い」ことを利用して，話題の抽出，クラスタリング結果の改善を行う．

関連ページとしてリンクで繋がるページを利用する研究は多く行われている．杉山らはリンク先ページの内容を文書の特徴付けに利用する研究を行っている[2]．

これまで，こうした研究と提案手法との比較が十分行えていなかった．そこで本研究では Trackback と通常のリンクの比較を行い，その結果を Blog クラスタリングシステムで利用する方法について述べる．

2．Blog クラスタリングシステム

我々は[1]で提案したクラスタリング手法実現のため，Blog クラスタリングシステムの実装を行った．本システムは大きく分けて次の 4 ステップからなる．

(1) Blog 検索
(2) 関連ページ URL 抽出
(3) 特徴語抽出
(4) クラスタへのエントリ割り当て

2.1 Blog 検索

Blog 検索エンジンを利用して Blog 記事（エントリ）の検索を行う．得られた検索結果ページから HTML タグのパターンマッチングによりエントリのタイトル，サマリ，URL を抽出する．

Blog 検索エンジンには goo ブログ Search[3]を利用した．

2.2 関連ページ URL 抽出

得られた検索結果の各エントリについて HTML タグ解析を行うことで TrackbackURL，本文中の通常リンク URL を関連ページ URL として抽出する．

抽出した URL の中にはスパムのように内容的に関連のない URL も含まれる．そのため，ここでは「ページ中に検索語を含まないページはスパムとする」というルールによってこれを除去する．

2.3 特徴語抽出

得られたエントリ集合から特徴語の抽出を行う．特徴語とは，エントリの内容を代表する語であり，クラスタラベルの候補でもある．

特徴語抽出の流れは以下のようになっている．

形態素解析

形態素解析器 Sen[4]を利用してタイトル，サマリから単語を抽出する．「形態素/解析/器」のように，複数の名詞が連続して現れる場合にはそれらを連結して「形態素解析器」のように一つの単語として処理を行う．

単語の重要度計算

各単語についてその出現回数，出現エントリ数からエントリ集合全体における重要度を計算し，重要度が高いものをその検索結果の特徴語とする．重要度の算出には TFIDF 法を採用した．

関連ページによる重み

エントリと関連ページの両方に共通して出現する単語は重要な単語である可能性が高い．そこで，関連ページに出現する単語の重要度が高くなるように関連ページの情報を重みとして与える．

最終的にこの重要度が高い上位 n 件をエントリ集合の特徴語とする．

2.4 クラスタへのエントリ割り当て

第 2.3 節の方法で抽出された特徴語をラベルとするクラスタを生成し，検索結果の各エントリを適切なクラスタへ割り当てる．

エントリのクラスタへの所属可否判定には，各エントリとその関連ページにおける特徴語の TFIDF 法に基づく重要度を利用する．この重要度が閾値より大きければエントリをそのクラスタへと割り当てる．また，どのクラスタへも割り当てられなかったエントリは「その他」クラスタへ割り当てる．

3．Trackback とリンクに関する実験

関連ページの選択方法として Trackback と本文中の通常リンクの比較実験を行う． 具体的には，抽出した特徴語によって検索結果と Trackback/リンク先ページ間の話題の関連性を調べる．

3.1 実験データ

実験で使用するデータは，2007 年 4 月 25 日～5 月 29 日の間に収集した Blog 検索結果 500 件(100 件/検索語)である．検索語と各検索結果からプログラムにより抽出した関連ページ数を表 1 に示す．

表 1: 検索語と関連ページ数

検索語	Trackback	リンク
統一地方選	5	3
toto	7	12
Vista	6	32
はしか	8	6
イチロー	17	2

† 岡山大学 大学院 自然科学研究科

3.2 特徴語抽出実験

Blog 検索結果集合，Trackback 集合，リンク集合のそれぞれがどのような話題を含んでいるかを調べる．検索語“統一地方選”による Blog 検索結果から TFIDF 法による重要度計算に基づき抽出した特徴語上位 10 件を表 2 に示す．表中で網掛けされているものは，適合特徴語で，これは検索語との関連が容易に推測できる特徴語のことである．

「首相」，「法案」，「政治」といった特徴語は検索語“統一地方選”と直接的な関連を持たないと判断したため，今回は非適合とした．しかし、判定基準によっては適合とみなすことも考えられる．他には Trackback 集合の結果において「人たち」，「あと三ヶ月」といった関連がない単語が見られるが、それ以外は適合となっており，それぞれの集合が検索語と一定の関連性を持っていると見なせる．

他の検索語についても同様に実験を行った．その結果については表 3 にまとめる．

表 3 を見ると，“toto”の検索結果集合，“Vista”“イチロー”のリンク集合において適合特徴語数が少なくなっている．しかしながら平均すると，どの集合も 5 割前後の適合特徴語をもつことから，関連ページ集合と見なすことが可能といえる．また，関連ページである Trackback 集合とリンク集合を比較すると，適合特徴語数の平均では前者が勝っているが，検索語ごとの結果はまちまちであり関連ページとしてどちらか一方が優れているとは一概には言えない．

3.3 Trackback/リンク先ページの分類

続いて，Trackback 集合，リンク集合に含まれるページの分類実験を行った．

実験では著者が関連ページの内容を確認して，①意見ページ，②情報ページ，③スパム，④その他の 4 種類に分類した．それぞれ書き手の意見が反映されたページ，ニュースサイトや公式ページなどの情報源となるページ，広告などのスパム，それ以外のページとなっている．

対象は表 1 の検索結果集合より人手によって抽出した Trackback59 件，本文中リンク 140 件である．結果を図 1 に示す．

図 1 を見ると，Trackback 集合に含まれるページの 81.4%が意見ページという結果が得られた．また 11.9%が情報ページ，5.1%がスパム，残り 1.6%がその他のページとなっている． その他のページに含まれたページとしては，Trackback を受け付けることによって、特定の話題に関する Blog の記事を集積するトラックバックセンターある．一方リンク集合に含まれるページで最も多かったのは情報ページで 62.1%，続いて意見ページが 18.6%，スパムが 16.4%，残り 2.9%がその他となっている．その他のページには YouTube などの動画サイトや地図サイトのページが含まれた．

この結果から，Trackback によって繋がるページの多くは意見ページ，すなわち書き手の意見を強く反映しているページであるといえる．またこうしたページはリンクによって繋がるページにもある程度含まれる．しかしリンクの場合は情報ページ，すなわち記事中で扱う話題の根拠、あるいは情報源となるページが多くの割合を占めているといえる．以上から，Trackback とリンクによる関連ページはそれぞれ違った特性をもつといえ，ユーザの要求に合わせて両者を併用することでより効果的なクラスタリングが可能と考えられる．

表 2: “統一地方選”の特徴語

	検索結果	Trackback	リンク
1	市長	白票	県議選
2	知事	国民	市議選
3	格差	人たち	現職
4	現職	あと三ヶ月	国民
5	市議	法案	候補
6	議員	補選	市長
7	首相	野党	親知事派
8	県議	与党	首相
9	都知事	首相	新駅
10	全国	政治	地域情報

表 3: 適合特徴語数

検索語	検索結果	Trackback	リンク
統一地方選	7	5	9
toto	3	5	6
Vista	8	7	1
はしか	5	4	4
イチロー	7	6	2
平均	6.0	5.4	4.4

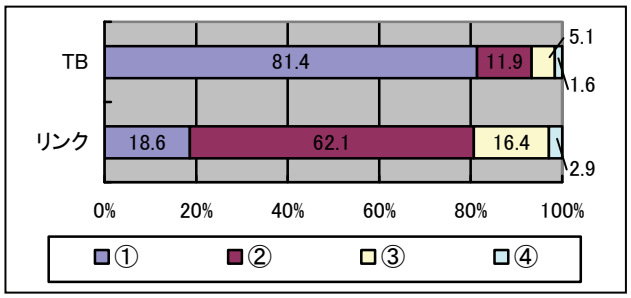

図 1:Trackback/リンク先ページの分類

4．関連ページの選択方法

第 3 節の結果から，検索結果と関連ページの話題の関連性，そしてそれぞれの特性が明らかとなった．ここではクラスタリングに上記の結果を反映させる関連ページの選択方法について述べる．

4.1 関連ページの特性を考慮した重要度計算

一般的に検索を行うユーザの検索意図はその時々によって異なる．こうした検索意図の変化に柔軟に対応できるシステムがユーザにとって有用なシステムであるといえる．しかしながらユーザの検索意図をシステムが自動で推測することは通常困難であるため，ここでは利用する関連ページの割合をユーザに指定してもらう．具体的には，記事の書き手の意見を強く反映させたクラスタリングを行いたい場合には Trackback の重みの割合を大きく，話題の根拠を強く反映させたい場合にはリンクの重みの割合を大きく設定する．この情報を元に，Blog 検索結果のクラスタリングを行う．

表 4:“イチロー”の特徴語

関連ページなし	$r = 0.0$	$r = 0.25$	$r = 0.5$	$r = 0.75$	$r = 1.0$
安打	安打	安打	安打	安打	安打
先頭打者本塁打	チーム	チーム	外野手	外野手	外野手
外野手	打率	外野手	選手	選手	選手
打席	お金	打率	打席	打席	打席
林檎	レッドソックス	打席	メジャー	メジャー	メジャー
メジャー	打数	選手	チーム	情報	先頭打者本塁打
マルチ	先頭打者本塁打	レッドソックス	打率	打数	野球
野球	外野手	打数	レッドソックス	投手	情報
選手	情報	情報	情報	先頭打者本塁打	打数
打数	打席	メジャー	打数	野球	投手

単語 t_i の検索結果集合における重要度を $tfidf_i$，Trackback 集合における重要度を $TBtfidf_i$，リンク集合中における重要度を $LINKtfidf_i$，関連ページ重みの割合を $r(0 \le r \le 1)$ としたとき，重要度 $Score_i$ を以下のように求める．

$$Score_i = tfidf_i * W_i$$

$$W_i = \begin{cases} 1 & if\,(R_i = 0) \\ 1 + w * (R_i / R_{\max}) & otherwise \end{cases}$$

$$R_i = r * TBtfidf_i + (1-r) LINKtfidf_i$$

ただし $TBtfidf_i$，$LINKtfidf_i$ は集合の大きさの影響を無視するように，それぞれの最大値で割ることで[0,1]に正規化した値を使用する．また，$R_{\max}$ は R_i の最大値，w は重みの大きさを決める係数である．この w については $tfidf_i$ の分布から，その最大値 $tfidf_{\max}$ と上位 100 件の平均値 $tfidf_{ave100}$ を用いて以下のように定めた．

$$w = \frac{tfidf_{\max}}{tfidf_{ave100}}$$

$Score_i$ が高い上位 n 件を検索結果の特徴語とし，関連ページを考慮したクラスタリングのクラスタラベルに利用する．

4.2 関連ページを利用した特徴語抽出実験

上記の手法によってどのような特徴語が抽出されるか実験を行った．検索語“イチロー”による Blog 検索結果についての結果を表 4 に示す．r の値が小さいときはリンクの影響が大きく，Trackback の影響は小さい．r の値が大きくなるにつれ逆になる．

表 4 を見ると，リンクの影響が大きい場合に特徴語として挙がっている特徴語「レッドソックス」から，レッドソックスに関するニュース記事が関連記事として含まれていると予想できる．記事の内容を確認してみたところ，レッドソックスの松坂投手との対決に関する話題が確認できた．また，Trackback の重みを大きくした場合に上位にきている「外野手」，「選手」，「打席」といった特徴語は記事の書き手が“イチロー”をどう捉えているか，どこに興味をもっているかを反映していると考えられる．

5．おわりに

関連ページを利用する Blog 検索結果クラスタリングのために，Trackback とリンクの 2 種類の関連ページについて実験を行った．この実験から「Trackback 集合は書き手の意見を反映した記事が多く，リンク集合は話題の根拠となる情報を含む記事場合が多い」という両者の性質の違いを明らかにした．また，それぞれの特性を考慮し，ユーザの要求に合わせたクラスタリング手法の提案を行った．この手法により，「ニュースで扱われた話題に関する記事を読みたい」，「他の人が興味を持っている話題を知りたい」といったユーザの要求に合わせたクラスタを提供することができると考えられる．

今後の課題としては関連ページ抽出精度の向上，クラスタへの要素割り当て方法の改良、そしてクラスタリング結果の総合的な評価が挙げられる．

文献

[1] 戸田雄士，太田学“Trackback を利用した Blog クラスタリング”，DBWeb2006 論文集 pp.337-344
[2] 杉山一成，波多野賢治，吉川正俊，植村俊亮“リンク先ページの内容を反映させた Web ページの特徴ベクトル改良法”，DEWS2002 論文集，2002 年 03 月
[3] goo ブログ Search，http://blog.goo.ne.jp/
[4] Sen Project，http://ultimania.org/sen/