

Blog コミュニティにおける話題抽出に関する研究

藤島 浩二 宮本衛市

公立はこだて未来大学 情報アーキテクチャ学科

1. はじめに

現在、インターネット上には blog と呼ばれる Web サイトが数多く存在する。blog にはコメントやトラックバック等、他者からのフィードバックを容易に得ることができる機能が備わっているうえ、文章を手軽に書いて更新できることから、インターネット上での発表や主張を行うに適している。これによってひとつの記事を中心としてその話題が広まり、blog 間での議論が展開されるケースも多い。これはトラックバック等、blog が相互の繋がりに特化した機能を有していることが要因であると考えられる。このような特性から、blog は現在インターネットにおいて blog 独自の領域を築くまでに至っている。

blogの情報を検索するためには、通常の検索エンジンの他、blog専用の検索エンジンも使用される。NAMAAN¹、blog search²などはblog検索エンジンのごく一部である。これらの検索エンジンは、blogが通常のWebサイトと比べて更新頻度が高いことに着目し、blogに合わせてデータベースの更新頻度を高くしていることを特徴としている。

その他の例としては、Blukfeeds³というblogやWebサイトのメタ情報を表すRSSを扱った検索エンジンや、評判情報検索やパースト度の解析など、多くの機能を有しているBlogWatcher⁴[1]が挙げられる。

一方、前述したとおり blog 間では議論を展開することが多い。それにも関わらず、上記の検索エンジンで出力される情報はあくまでぶつ切りの情報であり、このような議論に着目した検索結果を提示してくれる検索エンジンは存在しない。このため、途中から議論に参加しようとして流れを把握することや、目的の情報を探そうとした際、議論の規模が大きいとたくさんの blog を閲覧しなければならない。また blog 間では必ずしもリンクが双方向に張られているわけでないので、重要な情報を見落とす可能性もある。

このような問題に対処するため、本研究では blog 間 -

つまり blog コミュニティで行っている議論を探索し、その内容をアウトライン化したうえで出力するシステム、“BlogReader”の開発を行なった。これにより、blog 間の議論や話題の流れ、規模などを容易に把握することができ、また情報の見落としを防ぐこともできる。

2. 話題の検索

本システムは、blog の代表的な機能であるトラックバックによって接続されている blog 群を一種のコミュニティであると見なし、それらの中からユーザによって指定された話題を扱うコミュニティをアウトライン化した上でユーザに提示するというものである。

なお、このコミュニティでは blog そのものを使用するのではなく、その blog を構成する個々の記事（以下エントリー）をコミュニティにおける構成要素として取り扱っている。

本検索システムでは、主に以下に挙げる 2 つの機能から成り立っている。

2.1 コミュニティの検索

まずユーザは、自分が知りたいと思う話題を取り扱っているコミュニティを探さなければならない。

話題の検索を行なううえで重要なことは、単なるキーワードによる検索では話題を探すことにはならないという点である。本来、話題の中心にあるものは、ニュースサイトの記事など文章として存在するものであり、単なるキーワードのみを指定するのでは本来の話題に関するコミュニティを検索するという目的は達成されない。このため、あるサイトからの引用などの文章を検索クエリにし、各々のコミュニティからエントリーを検索できるよう、本システムでは検索に汎用連想検索エンジンであるGETA[2]を採用している。この関連文書検索で検索された文書が各コミュニティにどの程度の割合で存在するかにより、画面上に提示するコミュニティの決定を行なっている。

またこれによって、コミュニティ内にトラックバックスパム等によって生じてしまった関連しないエントリーを

¹ NAMAAN <http://www.namaan.net/>

² blog search <http://blog.threetree.jp/>

³ Blukfeeds <http://bulkfeeds.net/>

⁴ BlogWatcher <http://blogwatcher.pi.titech.ac.jp/>

排除することができる。さらにコミュニティにおいて全体的に関連性の薄いエントリが集合している場合、このようなコミュニティも排除することができる。

その他にも、ユーザが提示されたコミュニティの中から選択する際、判断の材料とするための各コミュニティにおいて重要と思われる単語 トピックワード等を容易に解析することが可能となる。

2.2 コミュニティ内のアウトライン化

ユーザがコミュニティを選択すると、そのコミュニティにおける内容をアウトライン化したものが表示される。基本的な内容としてはコミュニティに含まれるエントリの一覧を表したものだが、以下のような手法を適用したうえでユーザがコミュニティにおける話題を把握しやすいようにしてある。

・コミュニティ内における各エントリの整列

エントリの並びに関しては、トラックバックの構造が分かるよう、最もトラックバックを受けているエントリを最上位に位置するものとして、そこから各エントリを階層構造で表示するようにしてある。これによって、エントリ同士の関係が上位に存在するか下位に存在するかで、各エントリがトラックバックの送り主か受け手であるか、という関係が分かるようになってある。

・エントリから抽出した要点のみを表示

表示されるエントリは、そのエントリのタイトルと共に、そのエントリにおける要点が表示されるようになっていく。ユーザはこれら各エントリにおける要点を閲覧することにより、直接そのエントリを見に行かずとも、そのエントリの内容を端的に把握することができ、且つコミュニティにおけるそのエントリの位置付けなどを知ることが可能となっている。

なお、ここでいう要点とは、そのエントリにおいて特徴的であり、かつコミュニティ全体において主軸となっている話題に背かない部分のことを指す。抽出の際には、この要点の定義に沿ったベクトルを各々の文に設定し、クラスタリングを行なうことにより要点が否かに仕分けされている。

これらの手法により生成されたアウトラインを図 1 に示す。これを読むことにより、ユーザはコミュニティを掲示板のように手軽に読むことができ、コミュニティ全体における話題の流れを容易に把握することが可能となっている。ユーザはさらに詳しい情報を求める場合のみ、

blog へと赴いて改めてその内容の確認を行えば良い。

3. まとめ

本研究では、トラックバックによって接続されている blog 群を一種のコミュニティと見なし、そこに含まれる各 blog の要点と見られる文を抽出、アウトライン化して提示するシステムの提案を行った。

今後の課題としては、本システムの中核となっている要点抽出における精度の改善や、インタフェースなどの改良などを行なっていき、更なる完成度の向上を目指したいと考えている。

参考文献

[1] 奥村 学, 南野 朋之, 藤木 裕明, 鈴木 泰裕 : blog ページの自動収集と監視に基づくテキストマイニング, 人工知能学会, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-01, 2004

[2] 高野明彦, 丹羽秀樹, 西岡真吾, 岩山真, 久光徹, 今一修, 櫻井博文, 徳永健伸, 奥村学, 望月源, 野本忠司 : 汎用連想計算エンジンの開発と大規模文書分析への応用, 第 19 回 IPA 技術発表会 2000 年 10 月

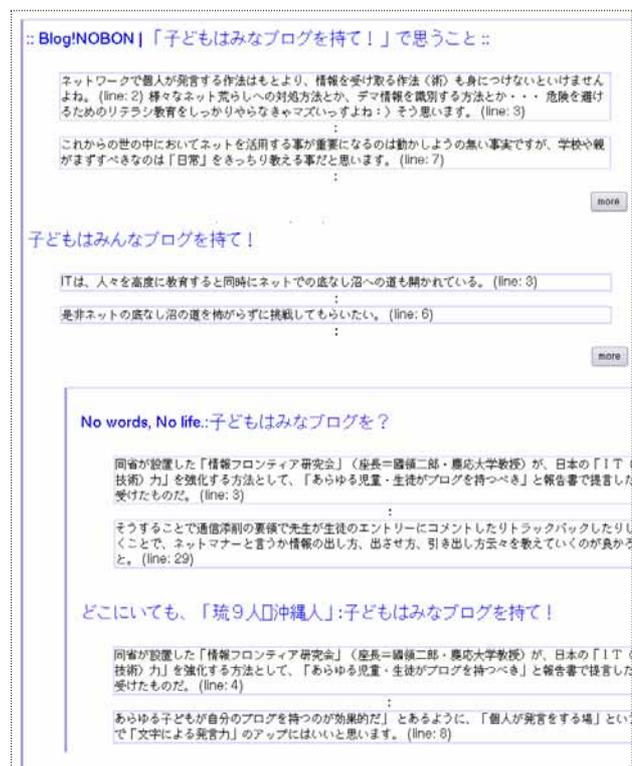


図 1. アウトラインの表示