2C2-03

Weblogネットワークの特徴とユーザの行動に関する分析

Analysis of Network and User's Activity in Weblogs

古川 忠延[*1] 松澤 智史[*2] 松尾 豊[*3] 内山 幸樹[*4] 武田 正之[*2]
Tadanobu Furukawa Tomofumi Matsuzawa Yutaka Matsuo Koki Uchiyama Masayuki Takeda

[*1]東京理科大学大学院 理工学研究科
Graduate School of Science and Technology, Tokyo University of Science

[*2]東京理科大学 理工学部
Tokyo University of Science

[*3]産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

[*4]株式会社ホットリンク
hottolink,Inc.

We propose extracting information from relations in a blog network for using a recommendation system. We use the bookmark, comment, trackback, reading activity, and the similarity of contents as the relation of blogs, and define the strength and type as the measure for relations. We analyze the correlation between those measure and users' reading activity. We attempt to determine the relations on which users regularly read the blogs.

As a result, we understand the interesting blog for one user is also interesting to other user who regularly read one's blog. And the similarity of contents does not influential really by our algorithm in this paper.

1. はじめに

近年，Web における個人の情報発信の形態として Weblog（以下 Blog）が注目を集めている．Blog においては，ユーザは自身の Blog をに記事を投稿するとともに，他の Blog を訪れて閲覧し，興味のある記事を見つけてはコメントを書いたり，リンクやトラックバックを張ったりできるという特徴がある [谷口 04]．このようなユーザの活動は Blog 間に“ 繋がり ”を形成し，どの Blog とどの Blog の関係が強いのか，またどのユーザとどのユーザの関係が深いのかといった関係性を把握する根拠となる．

本稿では Blog 間の関係性に着目し，それがユーザの閲覧行動にどのくらい影響を与えているかを分析する．すなわち，ユーザが閲覧している Blog を，Blog 間の関係性から判別することを目標とする．このような判別が可能であれば，Blog の関係性に基づく有効な推薦サービスを構築することができると考えられる．Blog の関係性からユーザの閲覧行動を把握しようとする分析として，[古川 05] ではユーザの行動のみを関係性として行っているが，本稿では Blog 間の内容にも注目して繋がりの一つとして用いる．すなわち，Blog 間の類似度はユーザの閲覧行動に何らかの影響を与えているのではないかという仮定の下で，分析を行う．

なお分析には，Blog ホスティングサービスである Doblog [*1]のデータを用いる．Doblog では，ユーザがログインした上で Blog の書き込みやコメント書き込みを行うため，閲覧行動を含めたユーザの行動が取得可能である．また「ブックマーク」と呼ばれるお気に入り Blog を登録する機能があり，これも関係性の 1 つとして利用する．この Doblog 上の限定的なユーザのデータを利用し，ユーザの行動分析を行っていく．

2. Blog 間の関係性

本稿では，Blog 間の繋がりとして以下の 5 つを考える．

連絡先: 古川 忠延，東京理科大学大学院 理工学研究科 情報科学専攻，千葉県野田市山崎 2641，04-7124-1501，tektf@mt.is.noda.tus.ac.jp

*1 http://www.doblog.com/．データは 2003 年 10 月から 2004 年 12 月までの期間のもので，その中でもアクセス数上位の 3300 名（全ユーザ数の 10%）分のデータのみを使用した

- ブックマーク
  Doblog 独自の機能．ユーザが自身の Blog から，お気に入りの他の Doblog へ張っているリンク．
- コメントとトラックバック
  Blog の標準的な機能
- 定期的閲覧関係
  本稿では，ユーザ A からユーザ B の Blog への訪問が平均して 1 週間に 1 度以上の割合で行われているとき，それを A から B への定期的閲覧関係と定義する．
- 内容類似性
  Blog 内で記述している内容の近さ．2 つの Blog の内容がある程度より近い場合，それらは繋がっている，と解釈する．詳細は次章で述べる．

これらの繋がりを組合せて用いることで，ユーザが自分の Blog からどのような関係性にある Blog に対して興味を持って閲覧しているのかを分析する．関係性としては，ユーザが直接の繋がりで辿り着けない対象のうちで，閲覧するとした場合にユーザからの繋がりの影響が最も大きく現れると考えられるいずれかの繋がりの 2 ホップを考え，その強さ・タイプの 2 つの面について，閲覧行動との関係の調査を行う．

2.1 関係性の強さ

関係性の強弱の指標として，ノード間をある 1 種類の繋がりの 2 ホップで繋ぐことを考えた場合の経路数を用いた．例として，図 1 において A → B 間の経路数は 3 である（A → 1 → B / A → 2 → B / A → 4 → B）．関係性の強さに関する分析では，Blog 間において 5 つの各繋がりについてのこの経路数とユーザの閲覧行動との関係を調べる．

2.2 関係性のタイプ

ある起点となる Blog から 2 ホップで辿ることのできる範囲は，図 2 において塗り潰して示した 25 通りある．関係性のタイプに関する分析では，各 Blog 間にこの内のどの関係性が成立しているのかのデータを抽出する（複数の関係性が成立していることもある）．

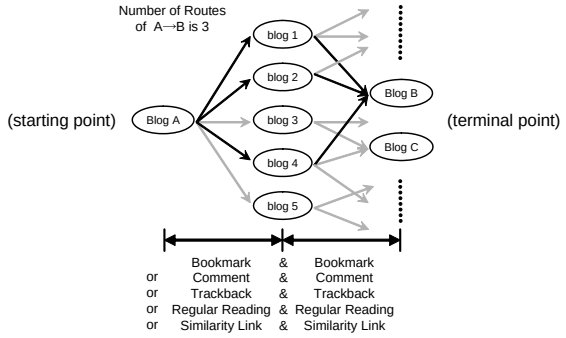

図 1: ある関係性の 2 ホップで繋がる Blog 間の経路数

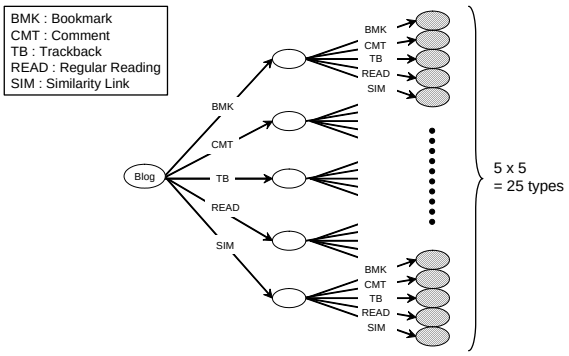

図 2: 25 種類の関係性

3. 内容類似性

Blog 間の繋がりの一つとして用いる Blog 間の内容の類似度には，各 Blog を出現する語の特徴ベクトルで表した場合に，Blog 間で計算される余弦を 1000 倍して整数化した値を用いる．そして類似度の値が閾値[*2]を超えた場合，2 つの Blog 間に“ 類似性 ”による繋がりが存在すると判定することとする．

3.1 特徴ベクトルの作成

本稿における各 Blog の特徴ベクトルの作成手法を示す．

1. 3300 ユーザ分の Blog の集合 D 内の各 Blog から，出現頻度上位 1%の語（1 つの Blog 当り平均約 20 語）を抽出し，全てを合わせて語集合 W（重複もあるため約 1 万語になる）とする．ここから，各 Blog における各語の出現頻度の対応関係を表す以下のような行列 X を得る．

$$X = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{cccc} w_1 & w_2 & \cdots & w_m \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \end{array}$$
$$n = 3300, w_j \in W$$

2. 1. により得られた行列 X に対し，LSI を用いて列（語の次元）を縮退 [中川 lsi]．縮退後の行列の各行を，各 Blog の特徴ベクトルとする．[*3]

*2 閾値は 400 とした．これは，Doblog の検索データベースにおいて，同一のカテゴリーに分類されている Blog 間で計算した類似度の平均値が，およそ 400 程度であったためである

*3 特徴ベクトルの作成では他に，1. の行列をそのまま使用した場合や，理論値との差を利用したの χ^2 値を用いた方法などで実験を行ったが，今回はその中で他のリンクとの相関が最も強く表れた LSI を採用した．

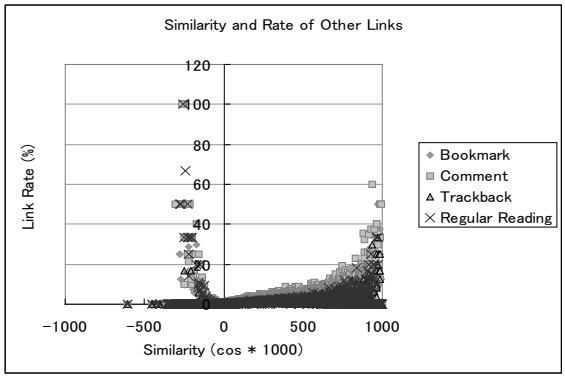

図 3: 内容の類似度と他の繋がりとの関係

3.2 類似度と他の繋がりの関係

図 3 は，2 つの Blog 間の類似度と，ブックマーク・コメント・トラックバック・定期的閲覧関係の成立割合との関係を表したものである．本稿では，類似性はユーザの行動に影響している，という仮定の下で分析を行っていくので，このグラフのような関係となる類似度の算出方法はある程度妥当であると考えられる．

4. 関係性と定期的閲覧行動の相関

4.1 アルゴリズム

本稿は，ユーザはある Blog を一度でも訪問したことがある場合に，どのような関係性にあれば定期的に閲覧するのかを分析することを目標とする．したがって，一度以上訪問したことのある Blog の組 (訪問した側, 訪問された側) についてそれぞれ，関係性の強さとタイプ，そして定期的閲覧関係に発展しているかどうかを調べる．この一度以上訪問したことのある Blog の組合せは 165922 通りあり，すなわち 165922 通りの (関係性の強さ, 関係性のタイプ, 定期的閲覧しているか否か) のセットを学習データとして，機械学習を行う．その結果，関係性の強さおよび関係性のタイプそれぞれについて，その中でどの要因がユーザの定期的閲覧行動に強く影響しているのかを抽出できる．なお，機械学習による決定木構築には C4.5 アルゴリズム [Quinlan 93] を用いた．

4.2 結果と考察

4.21 関係性の強さと定期的閲覧行動

まず．関係性の強さと定期的閲覧行動の成立割合の関係を図 4 に示す．図より，経路数が多い関係になればなるほど，一度でも訪れれば定期的閲覧関係に発展しやすいことが分かる．(グラフにおいて，右上がりの傾向から大きく逸脱して成立割合が 0%や 100%になっている部分は，データ数の少なさによる誤差と考えられる.) ただし，5 つの繋がりは同時に成立していることも多く，独立とは言えないため，このグラフだけではどの繋がりによる経路数が，より強く定期的閲覧行動と関係しているかは判別できない．そこで，図 5 に機械学習の結果 (上位 5 階層) を示す．

図中の丸いノードは繋がりの種類とその経路数を，四角のノード（リーフ）は直前までの判別によって分類されるデータを表し，YES/NO は定期的閲覧関係が成り立っているか否か，

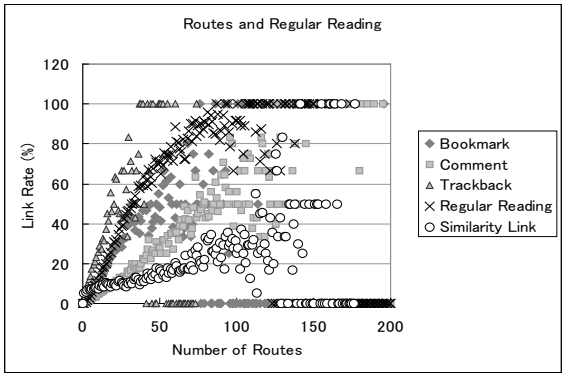

図 4: 経路数と定期的閲覧行動成立割合の関係

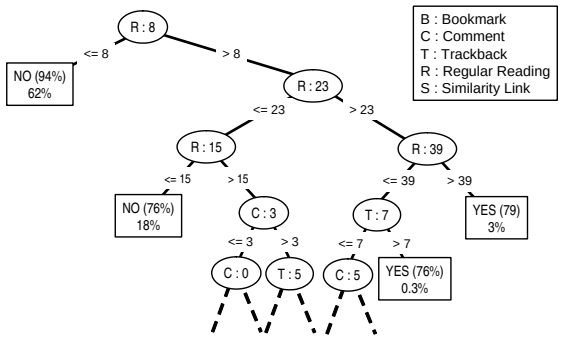

図 5: 関係性の強さからの定期的閲覧行動の判別

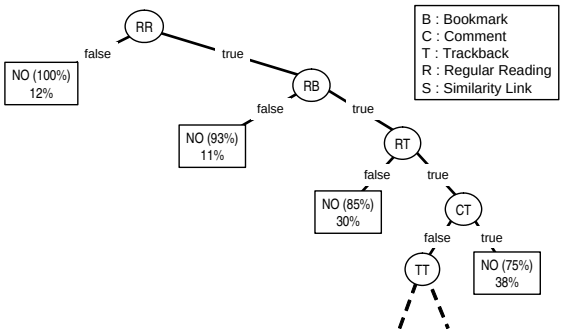

図 6: 関係性のタイプからの定期的閲覧行動の判別

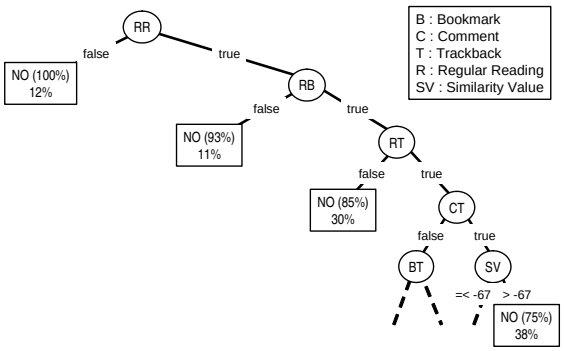

図 7: 関係性のタイプからの定期的閲覧行動の判別（属性として内容類似度の値を組み込んだ場合）

括弧内の数値はその分類が真である確率，下の数値は分類されるデータ数の全体における割合を表している．定期的閲覧関係に対する影響力の大きい属性から順に，上位のノードとして並んでいる．

この決定木より，最も影響しているのは定期的閲覧関係による経路数であることが分かる．この経路数が 8 以下となる関係性は全体の 62%を占め，94%の確率で定期的閲覧に発展ないことを表している．以降も定期的閲覧の経路数が判別に大きく貢献しており，すなわち,「自分が定期的閲覧しているBlog のユーザたちのうち，その多くが定期的閲覧しているような Blog」は自分にとってもまた定期的閲覧の対象になりやすい Blog であることを示している．

また，内容の類似性による繋がりの経路数は 5 階層までに出現しておらず，その影響あまり大きくないという結果になった．

4.22 関係性のタイプと定期的閲覧行動

関係性のタイプからの，定期的閲覧行動の判別のための決定木を図 6 に示す．決定木の見方は図 5 と同様であるが，ノード内の“ RB ”などの 2 文字のアルファベットは「自分が定期的閲覧（R）している Blog のユーザがブックマーク（B）している」というタイプの 2 ホップの関係性を表す．

関係性のタイプに関する分析においても，ユーザの定期的閲覧行動に最も関係しているのは「自分が定期的閲覧しているBlog のユーザが定期的閲覧しているような Blog」という結果が現れた．さらに下位の分類においても“ RB ”,“ RT ”といった「自分が定期的閲覧している Blog」を起点とした繋がりが現れており，その影響力の強さを示している．

また，この分析においても内容類似度による繋がりを含むような関係性は上位階層に出現しなかった．そこで,「類似度による繋がりを含む 2 ホップの関係性」の属性の代わりに，2 つの Blog 間における類似度の数値をそのまま属性として含めた分析を行ってみた（図 7）.

結果として，図 6 の 5 階層目に類似度の影響が現れる形となった．ただし直前の“ CT ”までの判別で全体の 38%のデータを 0.75 の確率で分類できるのに対し（図 6），その下で“ Sim ”の判別を行った場合（図 7）でも分類できる量とその精度の差は 1%未満であり，この判別は影響力のほとんどないものである．

なお，同様の操作を関係性の強さの分析においても行ってみたが，そちらでは上位 5 階層までに類似度による分類は出現せず，やはり決定木に大きな変化は見られなかった．

5. まとめ

本稿では，Blog 間においてコメントやトラックバックといったユーザの行動による繋がりと同時に，文書内容の類似性による繋がりを用意し，それらによる関係性とユーザの閲覧行動の間に相関を見出すための分析を行った．

その結果，自身が興味を持って閲覧している Blog のユーザが閲覧している Blog というのは，自身にとっても興味深いものであるという傾向が得られた．

さらに，ユーザ自身が書いている内容と閲覧している Blog に書いている内容の類似度にはほとんど関係がないという結果が現れた．ただし，Blog 間の類似度の算出方法，特に特徴ベクトルの抽出方法には検討の余地があるため，今後このことはさらに吟味していく必要がある．

参考文献

[Quinlan 93] J.R.Quinlan.: C4.5: Programs for Machine Learning (1993), (邦訳 : 古川康一 訳, トッパン (1995)).

[谷口 04] 谷口智哉，松尾豊，石塚満: Blog コミュニティの抽出と分析，第 6 回セマンティックウェブとオントロジー研究会（2004）

[中川 lsi] 中 川 裕 志: Latent Semantic Indexing，(URL) http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/infoDB/ir-lsi.pdf.

[古川 05] 古川忠延，松澤智史，松尾豊，内山幸樹，武田正之: Weblog ネットワークにおけるユーザ間の関係と閲覧行動の分析，情報処理学会第 67 回全国大会（2005）