社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MPS－72 (48)
2008－BIO－15 (48)
2008／12／18

ブログ記事における話題の流行に注目したリコメンドシステム

田中 秀和 †, 大原 倫太朗 †, 佐々木 一洋 †,
須貝 友亮 †, 山口 紀之 †, 渡邉 真也 †

† 室蘭工業大学情報工学科

本研究では，入力されたブログ記事に対して類似したブログ記事を時系列で示し，話題の流行を視覚的に提示するリコメンドシステムを開発した．本システムは，ブログ記事の類似度推定と話題変遷の可視化という2つの機能からなっており，ブログ記事の類似度推定では文章中からの特徴語抽出に基づく方法を利用しており，話題の変遷の可視化ではグラフやワードクラウドといった手法を用いている．幾つかの入力例に対する出力結果例を通して，本システムにおける類似度推定の妥当性および可視化の有用性について検証を行った．

Development of a blog recommendation system that represents a trend of a topic in blog articles

Hidekazu TANAKA † Rintaro OHARA † Kazuhiro SASAKI †
Yusuke SUGAI † Noriyuki YAMAGUCHI † Shinya WATANABE †

† Department of Computer Science & Systems Engineering, Muroran Institute of Technology

In this paper, we propose a new blog recommendation system that visualizes a trend of a topic in a article of the entered blog. This system is based on mainly two features: similarity estimation between blog articles and visualization of topic transition. The former is based on feature extraction means for detecting distinctive word and the latter is implemented in "word crowd" and "transition graph of frequency". Through the results of some examples, we clarified the validity of similarity estimation and the effectiveness of the visualization of topic transition in the proposed system.

1 はじめに

本研究ではこのような状況において,「(社) 情報処理学会数理モデル化と問題解決研究会」と「ソネットエンタテインメント (株)」主催のリコメンデーションサービスコンテスト[1] への応募を目的として，話題の流行に着目したリコメンデーションシステム「トレンドブログ (以下，トレブロ)」の開発を試みた．

本システムでは，単に入力されたブログ記事と類似したものを提示するだけでなく，同一の話題を扱った記事の数を時系列グラフで表現し，記事を月別に提示する機能を持たせた．この機能により，ある記事の話題について書かれた他の記事を時系列順に整理することが可能となり，関心のある話題の変遷を時間に沿ってユーザに提示することができる．

本研究では開発したシステムをココログなどのブログサービスに適用し，その結果から，最適なシステムの構造とブログを提示する際に用いるアルゴリズムの妥当性について検討を行った．

2 システムの概要

本章ではシステムの概要について述べる．提案するリコメンデーションシステムの特徴は，入力した記事と類似した記事を出力するための類似度推定と，グラフやワードクラウドを用いた話題の変遷の可視化である．

2.1 ブログ記事の類似度推定

入力された記事に類似した記事を出力するためにブログ記事の類似度推定を行う必要がある．本システムでは後述する戸田氏らが提案した TF-IDF 法に基づく方法[2]を採用した．

2.2 話題の変遷の可視化

本システムでは，入力された記事における話題とその話題の時間的な変遷に注目し，グラフやワードクラウドといった手法による可視化を行っている．

可視化において，話題の変遷を示すために次の3つの機能を実装している．

1. 検索結果のブログ記事を1ヶ月ごとに分けて表示する．
2. 月ごとの話題への注目度の変化を折れ線グラフとして表示する．
3. ある月の検索結果（ブログ記事群）でどのような単語が使われているのかをワードクラウドとして一覧表示する．

2.2.1 ワードクラウド

ワードクラウドでは単語の「色」「大きさ」を変えることでその出現傾向を示す．これによって各時期において固有な話題を知ることができる．ワードクラウドの詳細については4章で述べる．

2.3 システムの構成

トレブロの一連の機能は次の4つの Step から構成されている．

Step1 検索結果の候補群を選び出す
入力されたブログ記事（入力記事）に含まれる単語が含まれるブログ記事（候補記事群）を記事データベースから複数選び出す．

Step2 検索結果のランク付けを行う
候補記事群のそれぞれについて，入力記事との類似度を計算しランク付けを行う．

Step3 話題の変遷の抽出
候補記事群から「話題の変遷」に関する情報を抽出する．

Step4 表示
Flash などで構成されたクライアントによって情報を表示する．

3 類似度推定のアルゴリズム

本章では，ブログ記事から特徴的な語の抽出および「類似度推定」について述べる．

3.1 出力するブログ記事の選定

データベースに登録されているブログ記事数が膨大であるため，全てのブログ記事に対し次節で挙げる類似度推定を行うことは現実的ではない．そのため，入力されたブログ記事から抽出された特徴語をある一定以上含むブログ記事のみをデータベースから取得するようにした．具体的な選定手法を以下に示す．

Step1
入力されたブログ記事の名詞を出現頻度順に任意の数だけ抽出する．ここで n に，抽出した名詞数を代入する．

Step2
抽出した名詞を n 個含むブログ記事を選ぶ．

Step3
選定した記事がある一定数以上を超えた場合は終了する．そうでなければ，Step4 に進む．

Step4
n を1減らして，Step2 へ戻る．

3.2 ブログ記事間の類似度推定

入力されたブログ記事と，選定されたブログ記事間の類似度を求めるために，文書の特徴ベクトルを用いる．ここでの特徴ベクトルは，ブログ記事から抽出された複数の名詞から構成される．

類似度の算出においても，戸田らが提案する手法を採用した[2]．

3.2.1 特徴ベクトル

各記事間の類似度を算出するために，ブログ記事 E_i の特徴ベクトルを次のように定義する．

$$E_i = \{w_i^1, w_i^2, \cdots, w_i^m\} \tag{1}$$

ここで，w_i はブログ記事 E_i に含まれる各名詞の記事中の重要度である．

3.2.2 TF-IDF 法

各名詞の重要度を算出するため，TF-IDF 法を使用した．あるブログ記事 E_i における名詞 t の重要度 w_i^t は式 (2) にて求める．

$$w_i^t = \frac{\log(\mathrm{tf}(t, E_i) + 1)}{\log(M)} * \log(\frac{N}{\mathrm{df}(t)}) \tag{2}$$

ここで，$\mathrm{tf}(t, E_i)$ はブログ記事 E_i 中に名詞 t が出現する回数，$\mathrm{df}(t)$ はデータベースに登録したブログ記事中での名詞 t が出現する記事数，N はデータベースに登録したブログ記事総数，M はブログ

記事 E_i より抽出された全ての名詞の出現回数の総和を示す．式 (2) から，名詞の重要度は以下の特徴を持つことが分かる．

- 対象となるブログ記事中での全ての名詞の出現回数に対して，その名詞の出現回数の割合が高い時，大きな値を取る．
- 選定対象とするブログ記事の総数に対して，その名詞が含まれているブログ記事の総数の割合が低い時，大きな値となる．

3.2.3 類似度の算出

前節で算出した各名詞の重要度を用いて，類似度を求める．ブログ記事 E_i，E_j の類似度 $\mathrm{sim}(E_i, E_j)$ は式 (3) にて求める．

$$\mathrm{sim}(E_i, E_j) = w_i^1 w_j^1 + \ldots + w_i^m w_j^m \qquad (3)$$

この式で求める類似度によって，ブログ記事のランク付けを行う．

4 ワードクラウド

4.1 単語のサイズの決定

描画する各単語のサイズは，月内の各単語の出現回数を比較することにより決定する．例えば，1月における単語 A の出現回数が，1月に出現する他の単語の出現回数より多い場合，単語 A は大きく描画される．単語サイズは5段階設けており，各単語が描画されるサイズは，各単語の出現回数から月内での偏差値を求めることで決定する．

j 月における単語 i の出現回数の標準偏差 dev_i，偏差値 ss_i^j は，以下の式 (4)，(5) で求める．

$$\mathrm{dev}_i = \sqrt{\frac{\sum_{i=1}^{n}(m_j - w_i^j)^2}{12}} \qquad (4)$$

$$\mathrm{ss}_i^j = 10 \times \frac{w_i^j - m_j}{\mathrm{dev}_i} + 50 \qquad (5)$$

ここで，m_j は j 月に含まれる全単語の出現回数の平均，n は i 月に含まれる全単語数，w_i^j は単語 i の j 月における出現回数を示す．

式 (5) で求めた偏差値を元に，各単語を5段階に評価する．偏差値が 70 以上の単語は，非常に高い出現回数となるという仮定に基づき，j 月における単語 i の大きさは以下の式 (6) で求める．

$$\mathrm{dev}_i = \mathrm{round}(4 \times \frac{dev_i}{70}) \qquad (6)$$

ここでの $\mathrm{round}(x)$ は x を四捨五入する関数を示す．

4.2 単語の色の決定

また，描画する各単語の色（白黒グラデーション）は，その単語のある月の出現回数と，全月の出現回数の分散を比較することにより決定する．例えば，2月から 12 月における単語 A の出現回数が均等であり，1月における単語 A の出現回数が他の月と比べて非常に多くなっている場合，1月の単語 A は白く描画される．色は白から黒のグラデーションで4段階設けており，各単語が描画される色は，各単語の分散の偏差値を求めることで決定する．

m_j を単語 j の各月の出現回数の平均，n を 12，w_i^j を単語 j の i 月における出現回数と置き換えた時，i 月における単語 j の分散の標準偏差 dev_i，偏差値 ss_i^j は，式 (4)，(5) で求めることが出来る．ここで求めた偏差値は，単語のサイズの評価と同様，式 (6) を用いた評価を行う．

5 検証実験

本実験では，2007 年に書かれたココログ[1]の記事データからランダムに選択した 40 件の記事データを入力としてシステムの実験を行った[2]．

類似度の算出に対する評価は，出力記事と入力記事との関連を主観に基づき比較することで行った．40 件の入力に対する各記事の出力から，ランク順に上位 10 件，下位 10 件を抽出し評価することで，類似度の算出に対する妥当性を評価した．ここで各出力記事の評価には，「同一の内容について書かれた記事」，「同じジャンルについて書かれた記事」，「無関係な記事」，「スパムブログ」「現在削除されている記事」の5つを用意した．

話題の変遷の可視化については，時事要素を含んだブログを入力して，グラフやワードクラウドが適当な結果を示しているかを主観に基づき評価した．

5.1 パラメータの設定

本実験では，ブログ記事の選定段階において入力記事から抽出する名詞数は最大 20 個，選定を終了する時のブログ記事数の下限は 100 件とする．

[1] http://www.cocolog-nifty.com/

[2] なお，入力に用いるブログ記事には，スパムブログや存在しないページといった，実験の入力として不適切なものは除いた．

5.2 類似度の算出に関する実験

入力記事と出力記事の関係の総計を Table 1 に示す.

Table 1 実験結果の統計

評価	内容一致	同ジャンル	無関係	スパム	削除	合計
上位	76	103	107	101	13	400
下位	17	55	167	159	2	400

Table 1 の結果から，下位の結果は上位の結果に対して，関係する記事数が少なく，無関係の記事，スパムブログが多くなっていることが分かる．これより，類似度に基づいたランク付けが正しく行われていると言える．

また，上位，下位共に，スパムブログが多く含まれていることが分かる．このスパムブログには無意味なキーワードだけが大量に含まれているため，スパムブログが多く選定されてしまっていると考えられる．この事は，ランキング上位での関係した記事の割合が 4 割程度となっている事に強く関連しており，本システムにおける大きな課題であると考えている．

5.3 話題変遷の可視化に関する実験

話題の変遷がグラフによって効果的に視覚化されているかの検証を行うために,「ネットゲーム[3]の相手に向けて新年の挨拶をする」という内容を持つブログ記事に対する出力結果の考察を行った．得られた結果のうち，12 月のものを Fig. 1 に，6 月のものを Fig. 2 に示す．各図では，ブログ記事に出現する単語の傾向と，お勧めのブログ記事のタイトル，および各月のブログ記事の数を折れ線グラフで表示している．

Fig. 1 におけるグラフより，1 月，12 月で記事数が大幅に増えていることが分かる．入力記事の内容から年末年始に記事数が増えるのは妥当な結果と考えられるため，話題の変遷がうまく反映されていると言える．

次に Fig. 1 のワードクラウドの結果をみると「年」という単語が大きく，白い文字で表示されている．文字の大きさからは出現回数が多く，文字の色から 12 月特有の単語であると判断できる．

また，Fig. 2 でのワードクラウドの結果では,「年」という単語は小さく，黒い色で表示されていることが分かる．このことから 6 月では「年」の出現回数が少なく，また，その月固有の特徴的な単語ではないことが読み取れる．12 月と比べると 6 月の「年」という単語に対する評価が低いのは妥当な結果と言える．

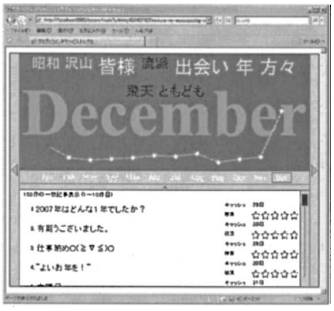

Fig. 1 出力結果:12 月のワード

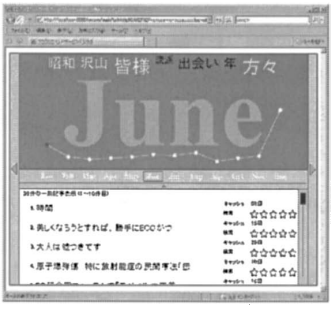

Fig. 2 出力結果:6 月のワード

6 まとめ

本研究では文章中の単語の重要度に基づいたブログ記事の類似度推定の手法を用いて，入力されたブログ記事と類似の記事を推薦するブログリコメンドシステムの開発を試みた．

実験結果より，関係のある記事は全体の 4 割近くであるのに対して．スパムブログも全体の 4 割近く含まれていることが分かった．

話題変遷の可視化に関する実験では，著者らの目的に沿った良好な結果が出力され，適当な可視化がなされていることが確認できた．

今後は，類似度推定の精度の向上のために，スパムブログの除去に取り組む予定である．

参考文献

1) リコメンドサービスコンテスト
http://www.so-net.ne.jp/web2/compe2008/contest.html
.

2) 戸田 智子, 福田 直樹, 石川 博, Blog 記事のクラスタリングに基づいたカテゴリ別話題変遷パタンの抽出. 電子情報通信学会 第 18 回データ工学ワークショップ論文集, 2007.

[3] 飛天，流派といった単語がネットゲーム内で用いられる単語である