キーフレーズに着目したブログ空間の口コミパターン抽出

平井　佑一*1　　熊野　雅仁*1　　木村　昌弘*1
Yuichi Hirai　　Masahito Kumano　　Masahiro Kimura

1. はじめに

近年、ブログはWeb上で誰もが容易に情報を発信できるメディアとして急速に普及し、ブログ空間は新たなコミュニケーションの場として活性化しており、ブログユーザ同士の情報伝搬現象を分析することが注目されている。本研究では、ブログ空間における口コミ的な情報伝搬現象分析システムの構築を目指し、キーフレーズに着目した分析システムを試作した。本システムでは、ブログ文書ストリームデータが与えられたとき、まず、Gruhlらの手法[1]を用いてそのデータからキーフレーズ群を抽出する。次に、各キーフレーズの口コミ的伝搬パターンを、ブログネットワーク情報とそのキーフレーズが記載されていたブログ文書の作成時間情報を用いて抽出する。さらに、得られた各キーフレーズの口コミ的伝搬パターンに対して、その伝搬経路長と伝搬速度を表示する。

2. 口コミパターン抽出機能

2.1 概要

本システムでは、まずブログ文書ストリームに対して形態素解析を行い単語に分解する。そして、活発に文書化されるトピックを特徴づける単語を抽出し、キーフレーズとする。次に、キーフレーズの記述があったブログ文書からキーフレーズの伝搬に関与した可能性のあるブログネットワークを同定し、その伝搬経路から口コミパターンを抽出する。

2.2 キーフレーズ抽出

キーフレーズの抽出法は、Gruhl らの手法を用いた。まず、i日目に単語tが出現した回数tf(i)を算出し、tf(i) > Tnを満たすtを抽出する。次に、抽出されたtの0日目からi-1日目までの出現回数$\sum_{j=0}^{i-1}$tf(j)とtf(i)からtfcidf(i) > Thを満たす単語をキーフレーズとし抽出する。

$$tfcidf(i) = (i-1)tf(i) / \sum_{j=0}^{i-1} tf(j)$$

2.3 伝搬経路抽出

キーフレーズが記述されたブログ文書群から、そのブログ文書を記述したユーザと時間を同定する。次に、ブログネットワークから同じキーフレーズの伝搬に関与した可能性のあるユーザのペア群を抽出する（図1参照）。このペア群の中から、まず、どのユーザからも伝搬を受けてないペアから順に、伝搬先がなくなるまで伝搬するペアを探索する。これにより、すべてのキーフレーズの伝搬経路を抽出する。

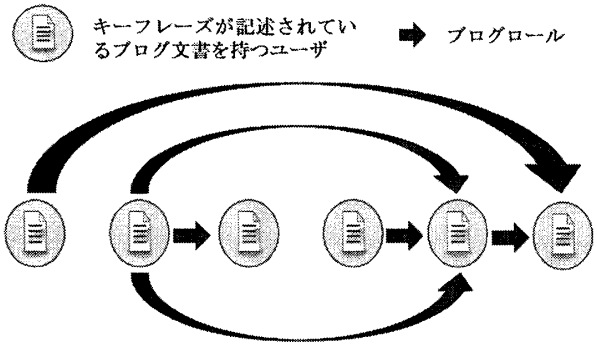

図1：キーフレーズ伝搬への関与可能性があるユーザペア

3. 伝搬分析機能

3.1 伝搬経路長

各キーフレーズの一つの伝搬経路ごとに、一番初めに情報伝搬させたユーザから、最後に伝搬を受けたユーザまでに経由したユーザの総数を算出する。

3.2 伝搬速度

各キーフレーズの一つの伝搬経路ごとに、一番初めに情報伝搬させたユーザから、最後に伝搬を受けたユーザまでの伝搬にかかった時間と、伝搬経路長に基づいて伝搬速度を算出する。

4. 実験

4.1 実験データ

実験では、Doblog※1のデータ※2を、分析対象の文書ストリームデータとして使用した。ブログ総記事数52,525、ブログロール総数（ブックマーク総数）115,552であった。

4.2 実験設定

キーフレーズ抽出における形態素解析エンジンには

*1 龍谷大学 理工学部 電子情報学科

※1 株式会社NTTデータ

※2（株）ホットリンクと（株）NTTデータの共同事業契約に基づき、（株）ホットリンクより提供。2003年10月から2005年6月のデータを利用

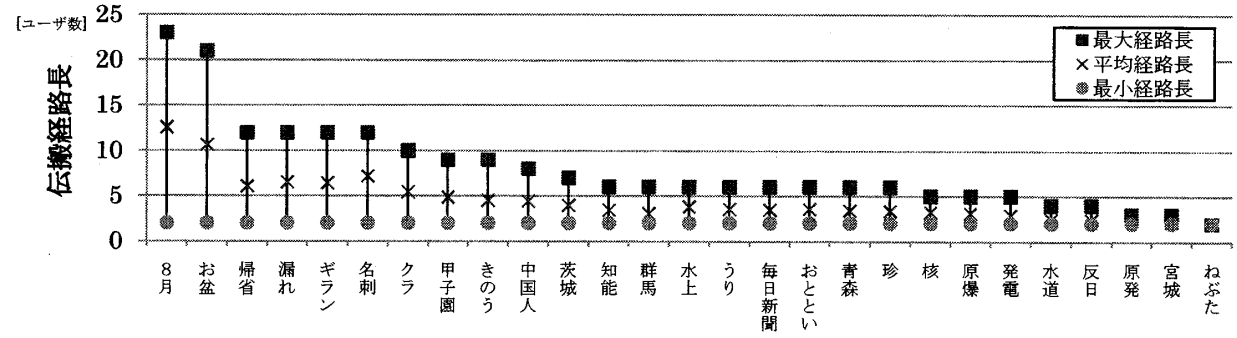

図2：伝搬経路長の最大経路長に基づくランキング（8月10日のキーフレーズ）

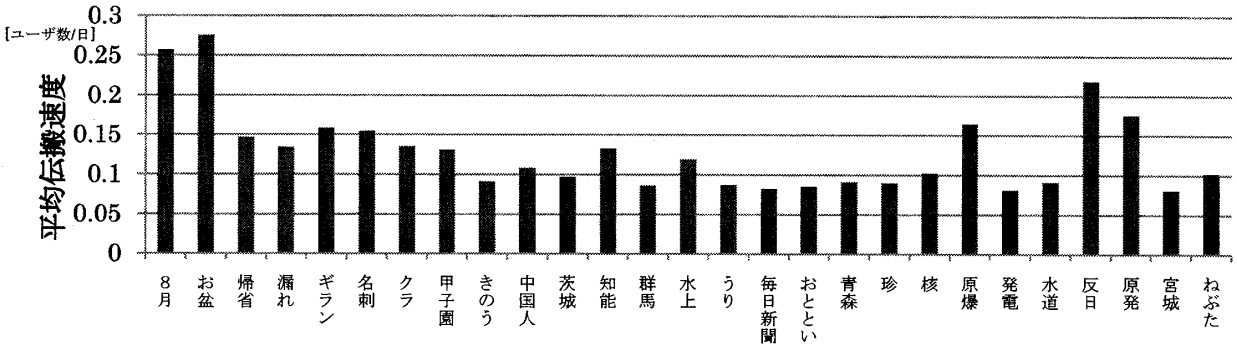

図3：平均伝搬速度に基づく分析結果（8月10日のキーフレーズ）

MeCabを使用した。ブログ文書ストリームには、Doblogの2004年6月1日から2004年9月10日までのデータを使用し、ブログネットワークとしては、ブログロールを用いた。キーフレーズ抽出のパラメータは、Gruhlらと同じTn=10とし、Th=3とした。キーフレーズとなる単語には、形態素解析した結果のうち名詞だけを用いた。

4.3 実験結果

キーフレーズ抽出には、2004年6月1日から2004年8月9日までのデータに基づいて、2004年8月10日のキーフレーズを抽出した。27個のキーフレーズが抽出された（図2, 図3参照）。本実験では、2004年7月10日から2004年9月10日までの期間におけるキーフレーズ伝搬を調べた。図2は、伝搬経路の最大経路長に基づいて、抽出されたキーフレーズをランキングした結果である。図3は、各キーフレーズの平均伝搬速度の出力結果である。キーフレーズ抽出では、「8月」や「お盆」のような口コミ情報との関連が薄いと考えられるものも抽出されたが、「ギラン」や「クラ」のような口コミ性の強い情報も抽出できた。口コミパターン抽出機能を高度化して、口コミ的情報のキーフレーズ群を抽出することは、今後の重要な課題である。

「ギラン」の情報伝搬パターンでは、総経路数2624、最大経路長12、平均経路長7.3、最小経路長2、平均伝搬速度0.15であった。平均伝搬速度からみると、あるユーザから別のユーザへ伝搬するのに、平均約1週間かかったことになる。よって、「ギラン」の情報伝搬では、多くの伝搬経路が存在し、伝搬経路長もある程度長く、ゆっくりとした伝搬であったことが示唆される。一方、「反日」の情報伝搬は、伝搬経路長は短いが、速い伝搬であったことが見て取れる。

5. まとめ

ブログ空間における口コミ的な情報伝搬現象分析システムの構築を目指し、キーフレーズに着目した分析システムを試作した。大規模ブログデータを用いた実験により、本システムの性能を調べた。キーフレーズの抽出法にはいくつかの問題点が示唆された。

謝辞

Doblogデータは（株）NTTデータおよび（株）ホットリンクより提供を受けた。記して感謝致します。

参考文献

[1] Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A.: Information diffusion through blogspace,in WWW'04, pp. 107–117 (2004)