4K-2

つながりを重視したWebコミュニティからの知識発見手法

福島 健吾* 今泉 忠† 出原 至道‡
§多摩大学大学院 経営情報学研究科

1 はじめに

近年，インターネットを介した双方向コミュニケーションが盛んである．無料で提供される電子掲示板サービスやSNS‖の興隆は，その最たる例である．これらの中で行われる膨大な量の「発言」から得られる知見は，企業や投資家だけではなく，参加者にとっても有益である場合が多い．本研究では，場 [4] の概念を念頭においたリンク構造だけではない「つながり」を重要視した知識発見手法を提案する．

2 手法の提案

2.1 概要

本手法で対象とするWebコミュニティは，電子掲示板上の一つのスレッドとする．一般的なスレッドは，つけられたタイトルに関連する複数の書込み（発言）の集合から成り立っている．また，一つの発言には**参照した発言（リンク），発言順（番号），タイムスタンプ，ハンドルネーム**（ID），**発言の内容**等を含んでおり，半構造化データセットである．

コミュニティの特徴は，図1のように，リンクによるつながりと場による概念に支配されるつながりにより 構築されると考えられる．そこで，リンク構造と概念構造を併せ持った**コミュニティ行列**を構築し，その固有ベクトルの成分を求めることで，重要な発言を求める．また，ある固有値における固有ベクトルの成分が突出している発言の集合を求めることで，発言の流れを読み取れるようにする．

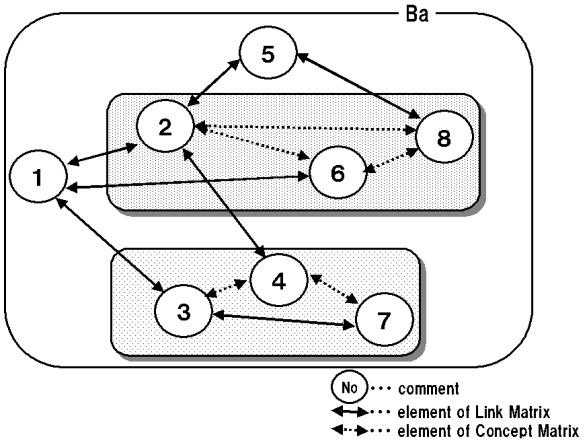

図 1: 概念をもつコミュニティ

2.2 リンク構造の抽出

スレッドからリンク関係を抽出した行列を**リンク行列**Lと定義する．Lの各成分l_{ij}は，スレッドにおける発言数をnとしたときに式 (1) のように定義する．Lは$n \times n$の対称行列となる．

$$l_{ij} = \begin{cases} 1 & \text{発言 } i(j) \text{ が発言 } j(i) \text{ の参照を明記した場合} \\ 0 & \text{上記以外の場合} \end{cases} \tag{1}$$

2.3 概念構造の抽出

場に存在する共通概念を概念ベクトルgで表現する．gは，t次元のベクトルであり，それぞれの次元は場に含まれている概念語をあらわし，その値としては重要度を表しているとする．発言ベクトルv_iはgと同じt次元のベクトルであり，発言された語の出現頻度を基にした重要度から構成される．このとき発言iと発言jは，概念ベクトルを介して接続されているものとし，そのの距離x_{ij}を式 (2) で定義する．また，このとき，発言iと発言jの概念がなす角$\theta_{x_{ij}}$を式 (3) で定義する．

$$x_{ij} = \cos\theta_i \cdot \cos\theta_j \qquad (0 \leq \theta_i, \theta_j \leq \pi/2) \tag{2}$$

$$\theta_{x_{ij}} = \cos^{-1}(x_{ij}) \qquad (0 \leq \theta_{x_{ij}} \leq \pi/2) \tag{3}$$

2.4 コミュニティ行列

コミュニティ行列Cの要素c_{ij}は，式 (4) で定義され，場における発言の構造をあらわす．

$$c_{ij} = \cos\theta_{c_{ij}} \tag{4}$$

ただし，θc_{ij}は，発言iと発言jのリンクがなす角と概念がなす角から構成することとし，(5) で定義する．このとき，リンクがなす角$\theta_{l_{ij}}$は式 (6) と定義する．よって，Cもまた$n \times n$の対称行列となる．

$$\theta_{c_{ij}} = max\left(0, \left(\alpha\theta_{l_{ij}} + \beta\theta_{x_{ij}} - \frac{\pi}{2}\right)\right) \quad (0 \leq \alpha, \beta \leq 1) \tag{5}$$

$$\theta_{l_{ij}} = \cos^{-1} l_{ij} \qquad (\theta_{l_{ij}} = 0, \pi/2) \tag{6}$$

Method of Knowledge Discovery from Web community that valued connection

* FUKUSHIMA Kengo (ken5@pc5.so-net.ne.jp)
† IMAIZUMI Tadashi (imaizumi@tama.ac.jp)
‡ IDEHARA Norimichi(idehara@tama.ac.jp)
§ Graduate School of Management and Infomation Sciences, Tama University

‖ Social Networking Site

3　手法の検証

3.1　データセット

検証のために用いたデータはインターネット上の掲示板から著者が任意に選択した 12 のスレッドであり，2.1 で示した一般的なスレッドの条件を満たしている．ここからスレッドのリンク情報を抽出し，リンク行列を作成した．また，タイトルと発言内容を形態素に分解し，出現頻度の高い形態素から順にのべ出現形態素数が全体の 70 %となるように選択し，概念ベクトルの次元とした．概念ベクトルは単位ベクトルで構成し，発言ベクトルは形態素の出現頻度の TF·IDF 値を求めて構成した．また，コミュニティ行列を作成する際のパラメタは $\alpha = 1, \beta = 1$ とした．

3.2　結果・考察

一例として，図 2 にスレッドのリンク行列の第 8, 第 9 固有値の固有ベクトルが持つ値，図 3 に同じスレッドのコミュニティ行列の第 9 固有値の固有ベクトルが持つ値を示す．いずれも横軸は発言順（番号）である．また，図 4 には着目した固有ベクトルの相関係数を示し，図 5 に，図 2 と図 3 で突出した発言の実リンクの構造を抜き出したものを示す．

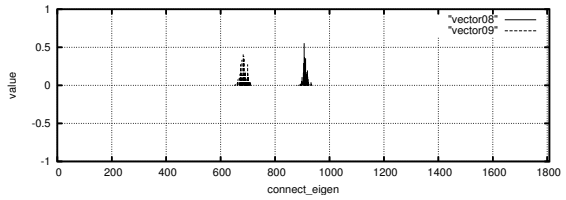

図 2: リンク行列の固有ベクトル（第 8 固有値, 第 9 固有値）

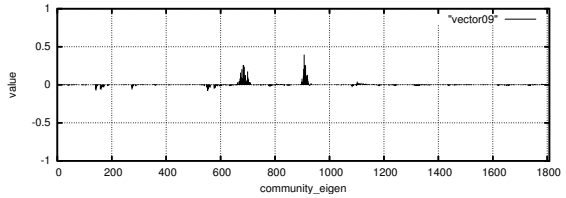

図 3: コミュニティ行列の固有ベクトル（第 9 固有値）

		L_Vector08	L_Vector09	C_Vector09
L_Vector08	correlation coefficient	1	-.010	.713**
	significance probability		.678	.000
L_Vector09	correlation coefficient	-.010	1	.630**
	significance probability	.678		.000
C_Vector09	correlation coefficient	.713**	.630**	1
	significance probability	.000	.000	

**. The correlation coefficient is significant in 1% level (both sides).

図 4: 固有ベクトルの相関係数

リンク行列の固有ベクトルを見ることで，発言の構造を抜き出せていることが確認できる．また，コミュニティ行列には，リンク行列から抽出した複数のコミュニティを含んでいることがわかる．また，それぞれの固有ベクトルの値が最も高いところが話題の起点であることが確認できる．

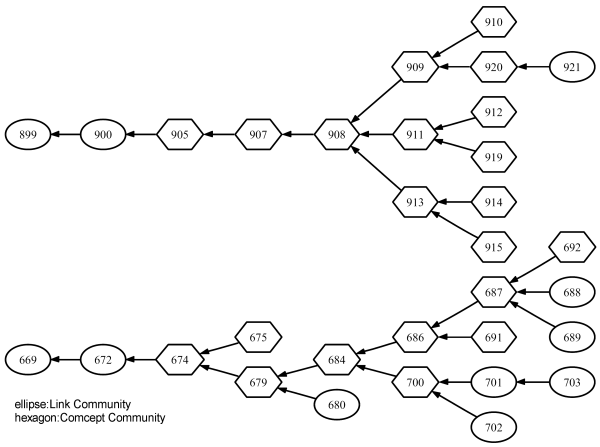

図 5: 抽出したコミュニティの連結 (番号は発言順)

4　まとめ

リンク行列の複数の固有ベクトルに着目することで，複数のつながりをもった発言の構造を抽出することが可能となり，抽出した発言の集合から個々の発言や単語単位の分析では得られない発言の流れを理解することが容易となった．また，コミュニティ行列の固有ベクトルを求めることで，直接リンクが無くても，同じ概念をもった発言の構造群を結びつけたコミュニティを抽出することが可能となった．

本研究では概念ベクトルを機械的に生成したが，分析者の意図を持って構築することも可能である．調べたい概念が存在する場合は，概念ベクトルを適宜作成することで，分析者の意図を汲んだ知識発見が可能となろう．

参考文献

[1] 松村 真宏, 大澤幸生, 石塚満. テキストによるコミュニケーションにおける影響の普及モデル. 人工知能学会論文誌 Vol.17, No.3, pp.59-267, 2002.

[2] 松尾豊, 大澤幸生, 石塚満. 電子掲示板における会話からのトピックの発見と要約. 人工知能学会全国大会 3D1-07.2002.

[3] 小林四一. コンピュータ・コミュニケーションにおける電子コミュニティの構造分析. 筑波大学大学院修士論文.1996.

[4] NONAKA Ikujiro , KONNO Noboru . The Concept of "Ba": Building a Foundation for Knowledge Creation. California Management Review,40(3):pp. 1 - 15, 1998.

[5] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. 1998.