mixiのネットワーク分析

丸井　淳己*　加藤　幹生**　松尾　豊***　安田　雪****

東京大学工学部*

株式会社ミクシィ**

東京大学大学院工学系研究科***

関西大学社会学部****

ソーシャルネットワーキングサービス(SNS)は今や多くの人にとって欠かせないものとなっている．SNSを特徴付ける性質の一つにスモールワールド性があり，情報推薦や口コミによる広告媒体としての大きな可能性を秘めている．

我々は今回日本で最大のSNSであるmixiの分析をした．本研究で使用したデータは2009年5月時のものであり，ユーザ数は16,937,041，リンク数は414,250,844本であった．リンク先の友人をmixiでは「マイミクシィ」(以下略してマイミク)と呼称している．ユーザのデータは性別・年齢・アクティブ度(最終ログイン日からの日数)・都道府県で構成され，IDはシャッフルされているためユーザの特定は出来ないようになっている．さらに「あしあと」と呼ばれるユーザ間のアクセス履歴(2009/5/1-5/30,秒単位での時刻付) 1,522,157,737レコードの提供を受けた．

平均マイミク数は24.46人で，次数分布は以下のようになっている．

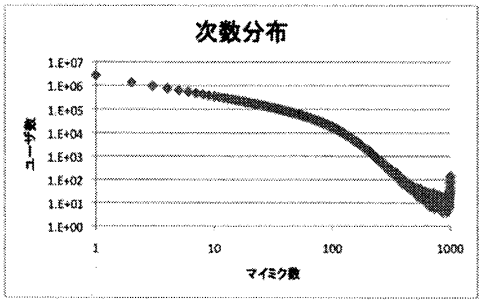

図1 次数分布

次数1000の付近でユーザ数が増えているのは，マイミク数が現時点で1000に制限されているためだと考えられる．

クラスタ係数は0.237であり2006年に行われたネットワーク分析での結果(0.328)と比較するとこの値は小さい[1]．平均経路長は全体の0.5%ほどの83250人に計算をしたところ5.45であった．この値は2006年での結果(5.53)と比べて大差ない結果である．

最初の分析として，ユーザの属性(年齢・性別)で切り分けてネットワーク分析を行った．同じ性別と年代である人々だけを構成員とした8つのネットワークは表1のようになっている．各ネットワークの平均次数は図2に示した．

	女性		男性	
年代	ユーザ数	リンク数	ユーザ数	ノード数
19-22	1947322	49660788	1600907	24804214
23-26	1933718	34867346	1635141	20643780
27-30	1485059	11567942	1308112	8432692
31-34	1132498	6279684	993353	4233516

表1 同質集団ネットワーク

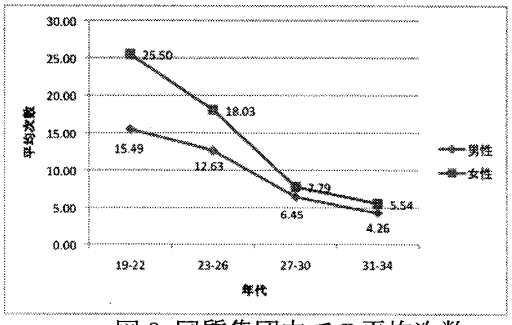

図2 同質集団内での平均次数

男女ともに年代を上がるほど平均次数が下がっていく．これを平均マイミク数と比較したグラフが図3であり，これを見ると年代が上がるほど自分以外の年代とのつながりが多くなる事がわかる．リンクの双方の次数の相関係数である Assortativity Coefficient(以下AC)についても計算を行った(図4)．これを見ると，女性が

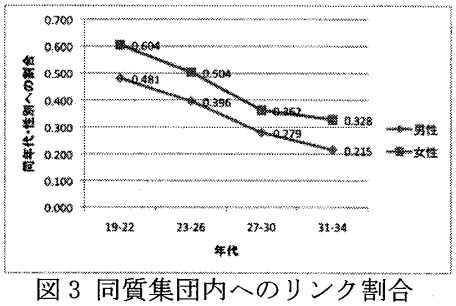

図3 同質集団内へのリンク割合

平均初婚年齢(28.3歳)を境に大きく変化している事がわかる．

以上の分析により年代が上がるほど相対的に他の年代・性別といった属性の違う人へのつな

Network Analysis of mixi
* Faculty of Engineering, The University of Tokyo
** mixi, Inc.
*** School of Engineering, The University of Tokyo
**** Dept. of Sociology, Kansai University

がりが多くなり，それぞれ人間関係が違う性質を持つことが推測出来る．そこで次の分析では同質に絞らずに全ての人間関係を種類分けした．

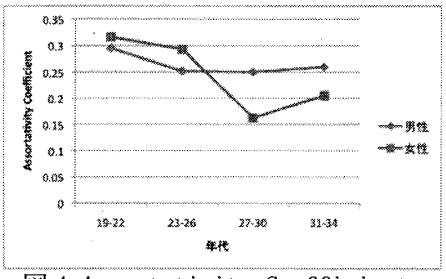

図4 Assortativity Coefficient

二番目の分析としてクラスタリングによるリンクの種類分けを行った．リンクの属性を両端のユーザの年齢差，性差(同性:0, 異性:1)，次数差，共通隣人数(全て絶対値)の4つで定義した．各属性をそれぞれの平均を引いて標準偏差で割る正規化後，k-means 法によるクラスタリングを行った．適切な k(クラスタ数)を設定するために次の指標を用いた．これは Bollegala らが階層化クラスタリングの際に用いたものである[2]．Γはクラスタ，Λはクラスタセットであり今回はkによって変わる値である．

$$C(\Gamma) = \begin{cases} 1 & |\Gamma| = 1 \\ \dfrac{2}{|\Gamma|(|\Gamma|-1)} \displaystyle\sum_{u\in\Gamma}\sum_{v\in\Gamma, v\neq u} sim(u,v) & \text{otherwise.} \end{cases}$$

$$\mathrm{IntCor}(\Lambda) = \frac{1}{k}\sum_{\Gamma\in\Lambda} C(\Gamma)$$

$$\mathrm{ExtCor}(\Lambda) = 1 - \frac{1}{|\Gamma_a||\Gamma_b|}\sum_{u\in\Gamma_a}\sum_{v\in\Gamma_b} sim(u,v)$$

$$(\Gamma_a, \Gamma_b) = \arg_{\Gamma_i,\Gamma_j\in\Lambda} \max C(\Gamma_i \oplus \Gamma_j)$$

$$Q(\Lambda) = \frac{1}{2}(\mathrm{IntCor}(\Lambda) + \mathrm{ExtCor}(\Lambda))$$

この値を k=3〜10 で計算すると図 5 のようになって k=5 が最大である．このときクラスタリングによって表 2 のような中心ベクトルが求まる．各クラスタの中心ベクトルの特徴から年齢差，次数差，仲良し，同性同年代，異性リンクとラベルをつけた．表 2 にのせたアクセス数はそのリンクを通る 30 日間にわたるアクセスの平均であるが，リンクの種類によって行き来の頻度が

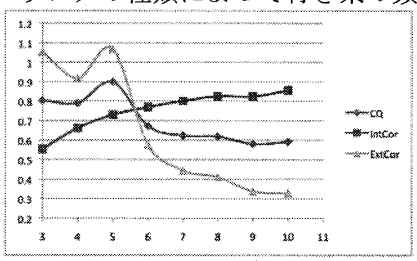

図5 クラスタリングクォリティ

リンク種別	年齢差	性差	次数差	共通隣人数	アクセス数
0. 年齢差	25.14	0.32	62.30	4.29	2.24
1. 次数差	5.76	0.45	403.95	4.92	1.43
2. 仲良し	1.58	0.16	65.58	24.45	2.62
3. 同性同年代	1.38	0.00	39.44	3.87	1.96
4. 異性	2.46	1.00	51.19	4.75	2.04

表2 クラスタの中心ベクトル

異なる．さらにそれぞれのリンクの性質からやり取りされる情報も違うと推察される．

より人間関係の性質を知るためにリンクの種類がどう共起するか調べた．具体的には三者関係を取り出してどの種類のリンクでお互いが繋がっているか，行き来しているかを調べた．

トライアド(2リンク)			
頻度	14.9%	14.6%	10.8%
トライアド(3リンク)			
頻度	2.54%	0.89%	0.83%

表3 トライアド(マイミク)

表 3 はトライアドを三者間にリンクが 2 本の場合と 3 本の場合で分けて頻度順に並べたものである．エッジに書かれた数字は表 2 と対応している．あしあとの場合は表 4 に示した．

トライアド(2リンク)			
頻度	10.0%	9.40%	7.29%
トライアド(3リンク)			
頻度	3.04%	0.64%	0.49%

表4 トライアド(あしあと)

この共起を見ると，マイミクでは次数差リンク二つというパターンが多く出現するのに対してあしあとでは同性同年代リンク 2 本をお互いに行き会うパターンが最も多かった．しかしながら必ずしも属性の似た人同士がつながっている訳ではなく，異性リンク二つといったパターンが多く，ある性に特化した情報は流れにくい事が推察される．

従来の推薦システムは属性情報を使った同質性を前提としたものだったが，以上の分析により同質なつながり以外の重要性を確認し，新しい情報推薦・伝播に有用な分析を提案した．

○参考文献

[1] 松尾豊，安田雪: SNS における関係形成原理 − mixi のデータ分析−人工知能学会論文誌，Vol. 22, No.5
[2] D. Bollegala, Y. Matsuo, and M. Ishizuka: Disambiguating Personal Names on the Web using Automatically Extracted Key Phrases, Proc. 17th ECAI-06, 2006