潜在トピックに着目したTwitter上のユーザ推薦システムの構築

古賀　裕之[*1]　谷口 忠大[*2]

Developing a recommendation engine to recommend users who share same latent topics on Twitter

Hiroyuki Koga[*1] Tadahiro Taniguchi[*2]

Abstract – **In this paper, we describe a following recommendation algorithm on a web communication tool, Twitter. Twitter is a famous web service through which users can send short messages containing less than 140 letters to their followers. If users on Twitter select other users as their followings, they can read their messages, called tweets. This means that how to choose their followings is an important topic to enrich the online informal communication. Therefore, we developed a novel method utilizing latent dirichlet allocation algorithm which is usually adopted in studies about document clustering. We also evaluated the effectiveness of the proposed method through an experiment.**

Keywords : **LDA, Twitter, Information Recommendation**

1. はじめに

インフォーマルコミュニケーションは娯楽のためのみならず，日常の定型業務におけるルーチンを逸脱し，社外での繋がりを形成したり，自らの関わる情報の流れに有益な外乱を生み出す視点からも重要である．たとえば，喫煙室や談話室，給湯室などで行われるコミュニケーションはただ単に心理的な安らぎといった効果だけでなく，重要な情報や発想が得られるなど様々な効果が期待される．近年，オンライン，オフラインを問わず，インフォーマルコミュニケーション支援を行う種々のサービスの開発，研究などが行われている．オンラインのインフォーマルコミュニケーションにおいては，旧来，掲示板，ブログ，WEB チャットなど発信者，閲覧者の関係性に個別的な制約を設けないものが主流であったが，国内での mixi [14] の人気などを契機に，SNS を始め，ユーザ自身が自らの発言を公開もしくは閲覧するユーザの範囲を管理することのできるコミュニケーションツールが好まれるようになってきている．Twitter はそのようなサービスの中でも最も注目されているツールの一つである．そこでは，如何に自らに関わりのある情報を発信するユーザを発見しフォローする，つまり，閲覧を開始するということが重要となる．そこで，本稿ではユーザに関わりの深いユーザを効果的に発見し，推薦する人工知能の開発を目指し，ユーザ推薦アルゴリズムの提案を行う．そのために，Latent Dirichlet Allocation(LDA) という文章クラスタリングの手法を応用する．また，実験を通してその有効性について検証する．

*1: 立命館大学大学院　理工学研究科
*2: 立命館大学情報理工学部 知能情報学科
*1: Graduate School of Engineering, Ritsumeikan University
*2: Department of Human & Computer Intelligence Ritsumeikan University

2. 研究背景

2.1 インフォーマルコミュニケーション支援

情報技術を用いたインフォーマルコミュニケーション支援については既に様々な研究がある．中野らの Traveling Cafe [12] では，コーヒーを飲む場所全てをインフォーマル空間と捉え，コーヒーサーバーにコーヒーを注ぎに来た人に他の人にコーヒーを注ぎに行くことを促すことでコミュニケーションの支援を行っている．椎尾らのアンビエント表示によるコミュニケーション支援 [8] ではコーヒーなどを飲むためのお茶のみスペースをインフォーマル空間として捉え，この空間にあるコミュニティ内の人々を集めることでインフォーマルコミュニケーションの支援を行っている．これらは何らかのきっかけを情報技術により構成し実世界でのコミュニケーションを支援するという研究と言える．松原らはこのような，インフォーマルコミュニケーションの契機となる存在を「言い訳オブジェクト」と呼んでいる [9]．

一方で，空間的に同じ場所に存在し得ない状態でのインフォーマルなコミュニケーションを如何に育むかということも重要である．岩渕らは周囲の会話のざわめきを感じさせるインスタントメッセンジャー RippleDesk [10] により，周囲のざわめきの感覚を取り入れることで，会話に参加するきっかけをユーザに提供している．商業的な WEB サービスとなるが，コミュニティや友だち機能で人と人の関係性情報を参加者自体に編集させることで，オンラインのコミュニケーショ

ンの場を形成する mixi をはじめとした種々の SNS や Twitter もその一種であると言えよう．本稿では Twitter に着目し，Twitter 上でのインフォーマルコミュニケーションをさらに有用にするために，そのユーザに関連の深いユーザ，興味を抱くであろうユーザを種々の情報から統計的に予測し推薦する手法を提案する．

2.2 Twitter

Twitter はつぶやきのような発言に基づく雑談 WEB 上で行う場を提供する WEB サービスである．Twitter には既存の SNS やメールなどに比べ，より気軽に多様な人へ発話できるという特徴がある．ユーザは自らがフォローするユーザを自由に選択することができ，その人々の発言のみを閲覧することが出来る．一方で，自らがフォローしていない人の発言はタイムライン上に表示されず，そのユーザの発話を契機としたインタラクションは起こりにくくなる．逆に，自らの興味の無いユーザをあまり多くフォローするとタイムラインは自らにとって不要な発言で溢れ，効果的なインフォーマルコミュニケーションが生まれにくくなる．よって，自らにとって有益な情報を発信する，もしくは，自らと実世界でも関係性のあるようなユーザを発見し，選別しフォローすることが，利用者にとっては Twitter の仕組み上重要となる [1]．

しかし，Twitter 上においては各ユーザの発話自体が「つぶやき」程度のものでしかないために，Web 上で最も代表的な検索手法であるキーワード検索によるユーザ検索機能は有効な検索方法にはなりにくく，リンクをクリックして辿ることでユーザを探すというアドホックな方法がその代替となっているのが実情である．その際，following の following, RT, list もしくは mention を辿ることが比較的有効な関連ユーザの発見手法として多くのユーザが用いている．これは，ユーザの following の following や自らに mention してきた人，自らの興味のあることを RT した人とは潜在的に共通の興味や話題を持っている場合が多いと考えられるためである．ここで mention とは，図 2 に示すように，発言が誰に対してのものかを明確にして発言することを指し，RT とは他のユーザの発言を引用して発言することを指す．また，Twitter にはユーザが独自に同じグループに属すると各ユーザ自身が判断して分類を作る仕組みがあり，これは list と呼ばれる．これも関連ユーザを探す上で重要な手段となる．

各ユーザが整備する following や list といった集合が完全であれば，これを見ることで，自らに関連するユーザを捜すこともできる．しかしながら，これらは人手により整備されるために常に不完全である．そこで，本稿では，これらの情報を確率モデルを用いて扱うことで統計的にフォロー関係や list を生成している潜在的なトピックを抽出する．共通の潜在トピックを共有するユーザを新たな following として推薦することで，効果的なユーザ推薦アルゴリズムの構築を目指す．

3. 潜在トピック抽出手法

本稿ではユーザのフォロー関係などに潜在するトピックを抽出する手法として LDA (latent dirichlet allocation) [1] を用いる．LDA は文書をクラスタリングするための手法であり，文書はトピック z の多項分布であり，トピック z が決まると多項分布に基づき単語 w が生成されるという生成モデルの仮定のもと単語と文書のクラスタリングを行う手法である [1]．これを本提案では LDA における文書を Twitter のユーザ，単語をユーザの following やユーザが含まれる list 名などと読み替え適用することで，Twitter のユーザのクラスタリングを行う．LDA のグラフィカルモデルを図 3 に示す．α は多項分布 θ を作るディリクレ分布のハイパーパラメータで，β はトピック毎の多項分布の情報を保持する行列である．

このためのコーパスの作成を行う際に，Twitter ユーザに関する 4 種類のデータを LDA における単語として用いる．以下でこれらについて述べる．

3.1 単語データの種類

以下に挙げる 4 種類のデータを用いてユーザのコーパスを作成した．このコーパスを以下で述べる LDA に適用することで潜在トピックモデルの構築を行う．

1. following
2. RT [2]
3. mention
4. list

1 については，ユーザの following のユーザ名の一覧にユーザ自身のユーザ名を付け加えたものを単語群として扱う．2 については，ユーザの発言のうち RT 記号を含むもののみを抽出し，その引用先の発言をラベル付けしそのラベルを一つの単語として扱う [3]．3 については，取得したユーザの発言から他のユーザへの mention を行ったもののみを抽出し，その mention を行った先のユーザ名を単語として扱う．4 については，ユーザが含まれている list 名を単語としてあつかう．

1: 以降，あるユーザがフォローしているユーザをそのユーザの following，あるユーザをフォローしているユーザをそのユーザの follower と呼ぶ

2: RT には非公式のものと公式のものがあるが，ここでは非公式のものを扱う．

3: ここで注意すべき点は，RT された発言を文書として扱い，その中の単語を単語として扱うのではなく，その発言自体を一つの単語として扱うということである

図 1　RT の例：ユーザ kogame5 が yoichiw のつぶやきを RT している

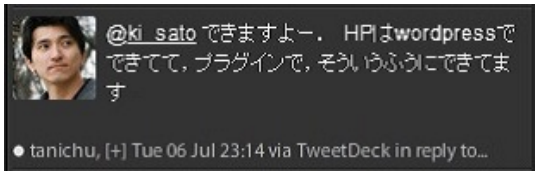

図 2　mention の例：ユーザ tanichu が ki_sato に発言している

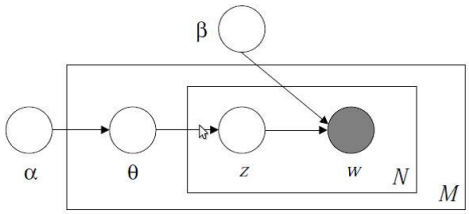

図 3　LDA のグラフィカルモデル

ただし，これらの 4 種類のデータにおいて，対象ユーザがデータを持たない場合が存在するため，全ユーザの 4 種類のデータにベースラインとして *hoge* という無意味文字列の追加を行った [4]．

上記の 1〜4 の内，一種類もしくは複数種類のデータを並べたものをそのユーザの文書と呼ぶことにする．そして，複数人分のユーザの文書を集めたものをコーパスとする．尚，実験ではこれらのデータは Twitter API [6] を利用して取得した．

3.2　Latent Dirichlet Alloction [1]

LDA は，文書は k 個のトピックに応じて発生した単語で構成されていると仮定し，各単語のトピック推定を行う手法である．Griffiths らは Gibbs Sampling を用いて LDA によるトピックの推定を行う手法を提案している [4]．これは，Gibbs Sampling により LDA のトピック推定結果を直接サンプリングするものであり，局所解に陥ることなく大域最適解の周辺のサンプルを効率的に得ることが出来る．

Gibbs sampling による LDA の定式化では最終的には以下のような式が得られ，この確率に従って順次サンプリングを行うことで各単語のトピックが得られる．

$$P(z_i = j \quad | \quad \boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \quad (1)$$

ここで，$\boldsymbol{z}_i$: i 番目のトピック，$\boldsymbol{w}$：単語集合，$n_{-i,j}^{(w_i)}$: i 番目以外でトピック j に割り当てられた単語 w_i の数，$n_{-i,j}^{(d_i)}$: i 番目以外でトピック j に割り当てられた文書 d_i の数，α, β: ハイパーパラメータ
W：単語数，$j = 1 \sim T$：トピック数，$d = 1 \sim D$: 文書数となる．ここで式中の $\cdot$ 記号はその添え字全てについての和をとることを意味している．右辺第一項がトピック j での w_i の確率，第二項が文書 d_i でのトピック j の確率を表している．

Gibbs Sampling アルゴリズム

1. $\{\boldsymbol{X}_i : i = 1...M\}$ の初期化
2. $l = 1...T$ に対して以下を行う

$$\boldsymbol{X}_1^{(l+1)} \sim f_1(x_1 | x_2^{(l)}, ..., x_M^{(l)})$$
$$\boldsymbol{X}_2^{(l+1)} \sim f_2(x_2 | x_1^{(l+1)}, x_3^{(l)}, ..., x_M^{(l)})$$
$$\vdots$$
$$\boldsymbol{X}_M^{(l+1)} \sim f_M(x_M | x_1^{(l+1)}, ..., x_{M-1}^{(l+1)})$$

サンプリングしたい確率分布 $f(\boldsymbol{x}) = f(x_1, ..., x_M)$ を考る．ギブスサンプリングの各ステップでは，1 つの変数の値が置き換えられる．その際，残りの変数の値を固定した条件での，対象の変数の条件付き分布に従って抽出した値を用いる．すなわち，x_i を分布 $f(x_i \mid x\_i)$ から抽出された値で置き換える．この手続きは，各ステップで更新する変数をある決まった順序で循環するか，ある分布に従ってランダムに選択することで繰り返される．ここで M はデータ数，T は繰り返すステップ数を表す．

このアルゴリズムを指定回数繰り返すことで，各単語をトピック分類することができる．

3.3　推薦ユーザの選出手法

ユーザの潜在トピックモデル間の距離を比較することで，ユーザ間の興味の近さの比較を行い，対象ユーザに推薦するユーザを選出する．潜在トピックを表す多項分布間の距離の比較には，KL ダイバージェンスを用いた．50 人から選出した 1 人のユーザを対象ユーザ P とする．対象ユーザ P から比較ユーザ Q の KL ダイバージェンス $D_{KL}(P||Q)$ を以下のように定義する [5]．

$$D_{KL}(P||Q) = \sum_{x=0}^{T} p(x)\{\log(p(x)) - \log(q(x))\} \quad (2)$$

4: つまり全てのユーザが必ず一度 *hoge* と言ったという設定

5: ここでユーザとそのユーザに対応する多項分布を同一視している

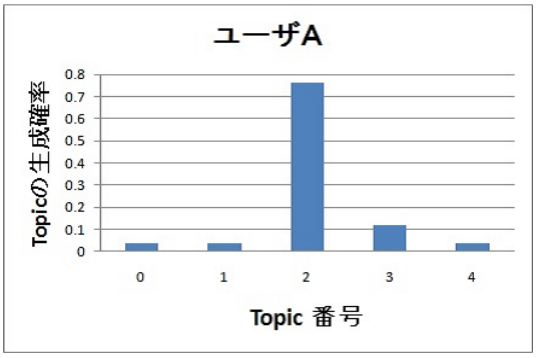

図 4　各ユーザ毎に得られる潜在トピック分布を示すグラフの例

ただし，x:topic 番号，T:topic 数，$p(x)$:対象ユーザの topic x の生成確率，$q(x)$:比較ユーザの topic x の生成確率 である．$D_{KL}(P||Q) < 1.5$ となる比較ユーザでかつ，対象ユーザの following でないユーザを対象ユーザへの推薦ユーザとする [6].

4. 実験

4.1 実験条件

2010 年 6 月 28 日から 2010 年 7 月 1 日の間の発言と，2010 年 7 月 1 日時点での following, list を 50 人のユーザを対象として取得した．LDA を用いて 50 人のユーザの潜在トピックの抽出を行った．その際，list, following，RT，mention，それぞれのデータを用いてコーパスを作成した．図 4 に LDA の結果として出力されるユーザに相当するトピック分布の例を示す．図はユーザ A の持つ 5 つのトピックとそれらの生成確率を示している．ここでは topic2 は約 0.75 の確率で生成される多項分布となっていることがわかる． mention データと RT データについては単体では潜在トピックを抽出する十分なデータ量とならなかったために，それら単体のコーパスは用いないことにした．

潜在トピック間の距離の近さに基づく対象ユーザへのユーザの推薦と，それに基づく推薦ユーザについてのアンケートを行う検証実験について述べる．被験者には上記 50 人のユーザの中から，8 人を選んだ．被験者には提示ユーザを提示する．提示ユーザは，前節で述べた推薦ユーザ (距離の近いユーザ) と，$D_{KL}(P||Q)$ が最大から 2 人のユーザ (距離の遠いユーザ) と，最大の $D_{KL}(P||Q)$ の半分近辺の距離にいる 3 人のユーザ (距離の中くらいのユーザ) とした．

被験者には，このようにして選出した提示ユーザの Twitter のホーム画面からユーザの発言, following, list などを自由に閲覧してもらったのちに以下の三種類の質問に対する解答を順に答えて貰った．質問項目は

6: ここで 1.5 という数字自体は発見的に定めたものである．

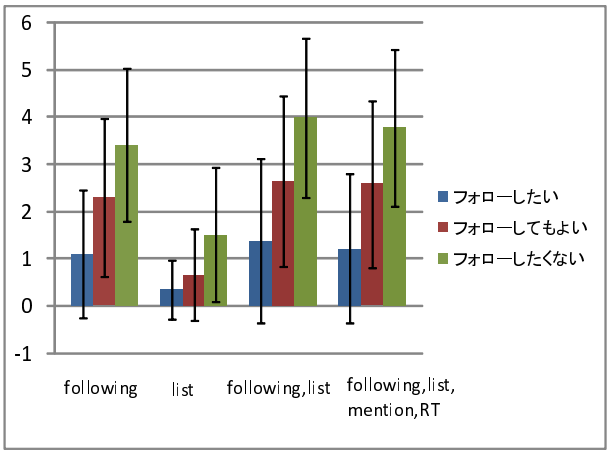

図 5　各条件に対するアンケート結果と KL ダイバージェンスの関係

1. 提示ユーザをフォローしたいか
2. 提示ユーザと知り合いか
3. 提示ユーザを以前フォローしていたか

質問 1,2 に関しては，本推薦手法の有効性の尺度とした．また，Twitter では知人であるが好き嫌いなどの理由でフォローを断ち切る例がしばしばみられ，この確認のために質問 3 を行った．今回の実験では，8 人のユーザに対して平均 24 人の提示ユーザについてのアンケートを行った．

実験では (1)following のみから作ったコーパス (2)list のみから作ったコーパス (3)following と list から作ったコーパス (4)following, list, RT, mention から作ったコーパスの 4 種類を用意し，それぞれの実験条件ごとに推薦ユーザを選出し，これらに対する回答を被験者からを得た．

なお，LDA の計算には Hieu らにより開発され，公開されている LDA implementation in C++ using Gibbs Sampling を用いた [7].

4.2 実験結果

本節では，アンケートによる検証実験の結果について述べる．質問項目 1 の「フォローしたいかどうか」には (1) フォローしたい，(2) フォローしても良い，(3) フォローしたくない　の三段階評価で答えてもらった．それぞれの実験条件毎にアンケートの結果を集計したものを図 5 に示す．各条件と各アンケート結果に対して，KL ダイバージェンスの平均値とその標準偏差を誤差棒で示している．　これらの結果から，全ての実験条件で距離の近いユーザであればあるほど，フォローしたい，もしくはフォローしても良いという答えがえられる傾向が読み取れる．また，図 6 には実験における回答の数を各条件ごとに示している．list のみをコーパスとした場合が一番結果は良くなかったが，一方で微妙な差ではあるが following のみより follow-

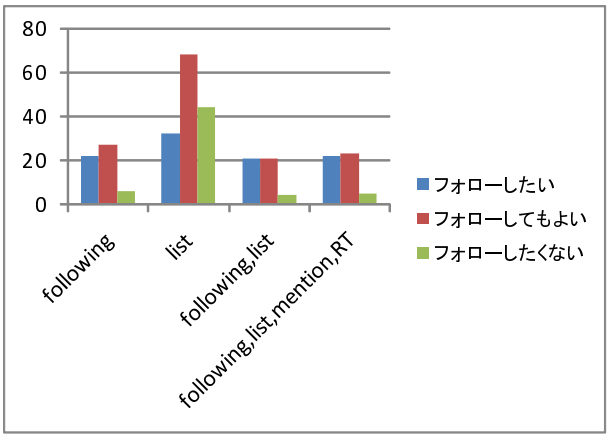

図 6　各条件による推薦ユーザに対する回答数

ing と list を両方扱った場合の方が全体に対する「フォローしたい」と答える割合は高かった (follwing のみ 40%に対して，following+list で 46%程度)．被験者数の関係から，確定的な結論は導けないが，少なくとも用いるデータを混ぜた際に必ずしも内分点的な結果が得られるわけでもないことがわかる．これより，より上手く複数の情報源を組み合わせる手法を研究することも有益であることが示唆される．

各実験条件での推薦結果の間には有意な差はみられなかったが，following を元にした推薦では，推薦候補自体がすでにフォローされているという事例が散見され，推薦ユーザ数が少なくなる傾向がみられた．

質問項目 2 の「提示ユーザと知り合いか」には (1) 知り合いである，(2) 知り合いでない　の二択で答えてもらった．詳細な結果は省略するが following 情報のみをコーパスに用いた場合「知り合いである」が 27%であったのに対して，list 情報のみをコーパスに用いた場合では 11%と大きな差が開いた．これより，これら二つのコーパスから得られる推薦候補の間には質的に大きな差があると考えられる．質問項目 3 の結果については特段の結果が得られたかったため説明を省略する．

5.　まとめ

本稿ではインフォーマルコミュニケーション支援のためのツールとしての Twitter に着目し，その中でのユーザ推薦手法の構築を行った．following や list によるタグとユーザの関係を単語と文書の関係に見立てて，クラスタリングを行うことで，潜在トピックを抽出した．その潜在トピックの共通性という形でユーザの類似性を検出した．この類似性に基づき，共通のトピックを持っているユーザを推薦することで，ユーザ推薦について良好な結果を得ることが出来た．

しかし，mention や RT といった，動的なユーザ行動から推薦につなげる点においては，良好な結果が得られて居らず今後の課題である．また，本来ならば質的に異なる list や following という情報を一つの文書の中の同等な単語として導入し扱ったが，これらの情報を質の違いを考慮しながら如何に有機的に統合するかも今後の課題であると言える．中村らはロボットの概念獲得のために複数の質的に違う情報を同時に扱うマルチモーダル LDA [2] [3] を開発しており，このような手法を上手く用いることも今後の課題である．また，オンラインとオフラインを上手く繋ぐことでインフォーマルコミュニケーション支援をより効果的に行うメディアのあり方について研究を進めていきたい．

参考文献

[1] David M.Blei,Andrew Y.Ng,Michael I.Jordan. Latent Dirichlet Allocation The Journal of Machine Learning Research, 2003
[2] 中村 友昭, 長井 隆行, 岩橋 直人, ロボットによる物体のマルチモーダルカテゴリゼーション電子情報通信学会論文誌 D Vol.J91-D No.10 pp.2507-2518 (社) 電子情報通信学会 2008
[3] 中村 友昭, 長井 隆行, 岩橋 直人, 複数のマルチモーダル LDA を用いた抽象的概念の形成 The 24th Annual Conference of the Japaneses Society for Artificial Intelligence,2010
[4] T. L. Griffiths, and M. Steyvers (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101 (suppl. 1) , 5228 5235. T. Hofmann (1999). Probabilis
[5] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian Data Analysis, 2nd Edition. Chapman & Hall CRC, 2003.
[6] Twitter API 仕様書 http://watcher.moe-nifty.com/memo/docs/twitterAPI20.txt
[7] Xuan Hieu Phan,Cam Tu Nguyen LDA implementation in C++ using Gibbs Sampling http://gibbslda.sourceforge.net/
[8] 椎尾一郎, 美馬のゆり アンビエント表示によるコミュニケーション支援インタラクション 2001 論文集, 情報処理学会シンポジウムシリーズ, Vol. 2001, No. 5, pp. 163-164, (2001)
[9] 松原孝志, 臼杵正郎, 杉山公造, 西本一志 言い訳オブジェクトをサイバー囲炉裏：共有インフォーマル空間におけるコミュニケーションを触発するメディアの提案情報処理学会論文誌，Vol. 44, No.12, 2003.
[10] 岩淵志学, 久松孝臣, 高橋伸, 田中二郎 周囲の会話のざわめきを感じさせるインスタントメッセンジャー RippleDesk ヒューマンインタフェースシンポジウム 2005, vol.2, pp.977-980, ヒューマンインタフェース学会，2005.
[11] 松田完, 西本一志　HuNeAS:　大規模組織内での偶発的な出会いを利用した情報共有の促進とヒューマンネットワーク活性化支援の試み情報処理学会論文誌,Vol43,No12,pp3571-3581 (2002)
[12] 中野利彦, 亀和田慧太, 杉戸準, 永岡良章, 小倉加奈, Traveling Cafe：分散型オフィス環境におけるコミュニケーション促進支援システム, インタラクション 2006 論文集，pp.227-228 (2006)
[13] 辻田眸，塚田浩二，椎尾一郎 遠距離恋愛支援システム 第 14 回インタラクティブシステムとソフトウェアに関

するワークショップ (WISS 2006), No. 43, pp.17-22, Dec. 6-8 2006. 日本ソフトウェア科学会研究会資料シリーズ.
[14] ソーシャルネットワーキングサービス [mixi], http://mixi.jp/