

SNSにおけるユーザの関係の視覚化

吉田 英樹†

小柳 滋‡

立命館大学大学院

1. まえがき

近年、Web の普及により SNS を利用する人が増えてきている。日本最大手の mixi では会員数が 100 万人を突破し、今後も拡大が続けると思われる。その一方でそれだけの人々が存在する空間では自分と親しい人や趣味趣向の合う人を短時間で探し出すのは容易ではない。そこで本研究では SNS 空間のリンク構造と blog などのテキストを用いてユーザの関係を視覚化する GPS 表示の提案、実装を行う。それによって自分の目的に合った人を探すことを支援する。このシステムの実現のためにはユーザ間の類似度を検出する必要があり、その計算には比較を行うユーザ間と、指標となるユーザとのリンク距離を用いる手法と、テキスト解析を行い各ユーザの類似度を求めるという手法を用いる。それによって GPS のように目的地を定め、それに近づいていく形が再現できると考える。

2. SNS

SNS とは参加者が互いに友人を紹介しあって、新たな友人関係を広げることがを目的に開設されたコミュニティ型の Web サイトである。SNS のサービスの中には自分のプロフィールや写真を公開する機能の他に blog、コミュニティの登録など、趣味を公開する機能がある。自己申告のプロフィールに加え、友達からのコメントという外部の評価が加わるのでプロフィールに信用性を与えることができる。こうやって作られたプロフィールには、友達の一覧が掲載されているので、友達の友達をたどっていくことができ、そこをたどりながら自分に合った友達を見つけていくこともできる。

3. 類似度の検出

3. 1 GPS 表示

本研究ではユーザの類似度を GPS 表示による 2 次元マップ上にユーザを配置する。GPS 表示を行うためには中央に配置する基準点が必要となるが、基準点には自分自身ではなく、現在閲覧しているページのユーザ（以下、ユーザ C）を用いることとする。これはユーザが目的を持ってリンクを辿っていることを表現できると考えたためである。

3. 2 リンク構造からの類似度

ユーザ間の類似度を求める方法の一つにユーザ間のリンク距離を用いる。その手法として本研究では相関係数を用いる。相関係数とは、2 つのデータ列の間の相関を示す統計学的指標である。-1 から 1 の間の実数値をとる、1 に近いときは 2 つのデータ列には正の相関がある

といい、-1 に近ければ負の相関があるという。0 に近いときはもとのデータ列の相関は弱いといえる。

相関係数の求め方は 2 つのデータ列 $x = \{x_i\}$, $y = \{y_i\}$ ($i = 1, 2, \dots, n$) が与えられたとき、式 1 を用いて求めることができる。なお、本論文では x に自分自身、 y にユーザ C とリンク関係にあるユーザに関するデータを扱う。

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (\text{式 1})$$

この時、分子にあたる部分は変数 x 、 y の共分散となり、分母にあたる部分は変数 x の標準偏差と変数 y の標準偏差をかけた値となる。また、 $\bar{x}$ $\bar{y}$ はそれぞれの データ $x = \{x_i\}$, $y = \{y_i\}$ の期待値である、期待値 $\bar{x}$ $\bar{y}$ の求め方は

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (\text{式 2})$$

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad (\text{式 3})$$

である。単純なリンク距離を用いて類似度を求めた際、多くのリンクを持つユーザが存在すると、正確な類似度を求めることが困難になる恐れがある。そのため、周りのユーザとの関係を利用することのできる相関係数を用いる。具体的な x の $\bar{x}$ $\bar{y}$ は、 x , y それぞれと他のユーザとのリンク距離、からユーザ C と他のユーザとのリンク距離の差分を取った値、その総和から求める。これによりユーザ C と比較して、 y がどれだけ自分自身(x)に類似しているかを求めることができる。また、自分自身からリンク距離が 6 以内のユーザのみを他のユーザとする。これは「6 次の隔たり」という 6 人の知人の連鎖を介せば世界中のどんな人にもたどり着けるという理論を採用している。

3. 3 テキスト情報からの類似度

本研究においてテキストとは、SNS 上でのプロフィール、blog のことを指す。まず、テキストの品詞解析を行い、数詞及び非自立の名詞を除く名詞のみを抽出する。既出の名詞であればその名詞 ID に、新規の名詞であれば新しい名詞 ID を割り当てる。そして各ユーザ毎に使用した名詞を行列に登録する。ユーザが使用した名詞 ID の列に 1 を、そうでなければ 0 を挿入する形で行列に格納する。その結果は表 1 のようになる。協調フィルタリングを行い、自分自身と他のユーザとの類似度を算出する。類似度は 0 から 1 の間の実数値を取り、全く同じであれば 1、似ていなければ 0 となる。例えばスポーツに関する blog を書いているユーザ同士は、共通の趣味のないユーザ同士ユーザよりも類似度が高く、類似度が 1 に近くなる傾向にある。

Visualization of user relation in SNS

† Hideki Yoshida Graduate School of Science and Engineering, Ritsumeikan University

‡ Shigeru Oyanagi Ritsumeikan University

表 1 ユーザごとの名詞行列

ユーザ id \ 名詞 id	1	2	3	...	n
1	1	0	1	...	1
2	0	1	0	...	0
...	...	...	...	...	...
m	1	0	0	...	1

4. 類似度の視覚化

4. 1 視覚化手法

3章で挙げた2つの手法を用いて、SNSでのユーザ間の類似度の視覚化を行う。

4. 2 ユーザの配置

4. 2. 1 リンク構造によるユーザの配置

前章で述べたように、ユーザCの相関係数は0となるので、相関係数をそのまま用いることで自分自身と他のユーザとの類似度を表すことができる。

4. 2. 2 テキスト情報によるユーザの配置

テキスト情報からの類似度では、ユーザCとの類似度（以下、類似度C）が0から1の間の様々な値をとる。そのため、ユーザCを中央に配置するためには類似度Cを中間値である、0.5の値に修正する必要がある。その手法として、本研究では(0, 0)、(1, 1)、(類似度C, 0.5)をx,yとする2次方程式を用いる。その式を用いて、他のユーザの類似度を修正する。

4. 2. 3 マップ上への配置

上記の方法によって得られた値によって、ユーザを配置する。X軸の値にはテキスト情報による類似度の値を、Y軸にはリンク距離による類似度の値を用いる。配置図は以下の図1ようになる。

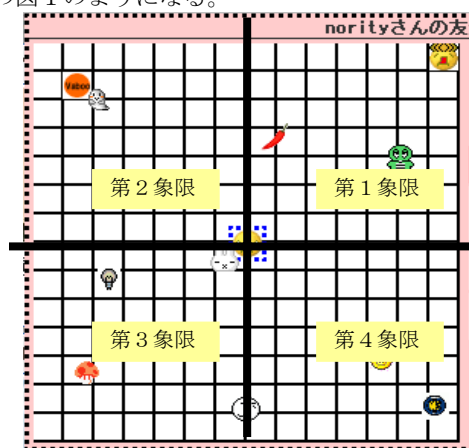


図 1 ユーザの配置

図1ではX軸、Y軸ともに正の値ほど類似度が高く、負の値ほど類似度が低いことを示している。つまり、第1象限はユーザCよりもリンク距離、テキスト情報ともに類似度が高く、第2象限では相関係数のみ類似度が高い、という見方をすることができる。なお、本論文では実験データにインディーズ音楽配信 SNS サイト BUZZWIRE (<http://www.buzzwire.jp>)のユーザデータを利用した。なお、

本論文ではテキスト情報にはプロフィールのテキストのみを利用した。

5. 実験

5. 1 実行結果

3種類のデータセットを用いて実験を行った。

1. 理想的なリンク構造、及び趣味を持つユーザ群
2. 実 SNS 上でのコミュニティを重視したユーザ群
3. 実 SNS 上での筆者を中心としたユーザ群

1. のデータセットに関しては予想される結果とほぼ同等の結果が得られた。このことから、本論文で用いたアルゴリズムや手法が正しい事が証明できたと考える。また、2. 3. に関しては次節でユーザによるアンケート結果の表を記す。

5. 2 評価

表 2 2. に関する5段階評価の平均

総合評価	リンク構造	テキスト情報	使い易さ
4.2	4.2	3.9	3.5

表 3 3. に関する5段階評価の平均

総合評価	リンク構造	テキスト情報	使い易さ
3.9	4.5	3.4	3.5

表2,3ともにユーザ50人に対してアンケートを取った結果を表している。表の項目である総合評価に関しては、このツールによって SNS を利用する楽しみが増えた、新しい友達を見つけるのに役に立ったなどの評価基準により算出したものである。

5. 3 考察

実行結果よりリンク距離による類似度の精度には一定の評価が得られたが、名詞による類似度の検出には十分に満足のいく結果は得られなかった。今後の課題として、名詞間の関連度の抽出および利用、一人称などの一般的過ぎる名詞の削除、名詞の頻出度を利用した多値の協調フィルタリングの利用などが挙げられる。さらには blog の書き込み日付を利用した重み付けなどの検討も挙げられる。また、評価に対しても、ユーザ個々の主観に頼る部分が多いため、明確な評価基準を策定し、客観的な評価が得られるよう工夫する必要がある。

6. おわりに

本論文では、SNS 上でのユーザ同士の関係を類似度の観点から視覚化することを行った。その結果、十分なツールとして活用できると考えられる。

参考文献

- [1] 松原望: 入門確率過程, 東京図書, 2003
- [2] 加藤 千恵子, 石村 貞夫: 相関係数と回帰直線, 東京図書, 2003
- [3] S.Oyanagi, K.Kubota, A.Nakase :“Mining WWW Access Sequence by Matrix Clustering” International Workshop on Web Mining for Usage Patterns and User Profiles, (WEBKDD-2002) pp.91-99, (2002.8) .