相関ルールを利用したSNSのコミュニティ分析

利光　由加子　　(**指導教員**: **増永 良文**)

1　はじめに

近年，ウェブで SNS（ソーシャルネットワーキングサイト）という友人・知人間のコミュニケーションを円滑にする手段や場を提供したり,「友人の友人」といったつながりを通じて新たな人間関係を構築する場を提供するサービスが増えてきている．このサービスの中で日本で最もユーザ数が多いのが mixi[1] であり，現在登録しているユーザは 800 万人を超えているといわれている (2007 年 1 月現在). この mixi にはコミュニティという趣味や嗜好，居住地域，出身校などに関して掲げられたテーマに同調する者が集まる仕組みがあり，その総数は 100 万以上ともいわれる．ユーザはコミュニティに入ることによって自分の興味や関心事を主張でき，その中でいろいろな情報交換や自分と嗜好が似た人を見つけたりもできるのである．そこで本研究ではこのコミュニティを利用することで自分が興味のある分野の情報を集めたり，同じような嗜好の人の動向を知ることができることを期待し，データマイニングで知られる相関ルールと mixi のコミュニティを利用してコミュニティ内での傾向分析を行うことにした.

2　相関ルール抽出

2.1　データマイニング

コミュニティ分析のためにデータマイニング技術の代表的なものである相関ルールを利用する．データマイニングとは巨大なデータの集合やデータベースからパターンや規則を探し出す技術であり，それにより得た情報を，販売戦略や，商品企画など，実世界の身近なところに生かすことが期待されている.

2.2　相関ルール抽出法

2.2.1　相関ルールとは　相関ルールとは「パンを購入する人の多くはミルクも購入する」 というような規則であり，パンを X，ミルクを Y とした時 X ⇒ Y と記述される．以下にルール抽出に必要なサポート度と確信度の説明をする.

I=i1,i2, ... ,iN をアイテム全体の集合とし，D をトランザクション集合データベースとする．あるアイテムセット X について，D の内の s %のトランザクションが X を含むとき，アイテムセット X は s のサポート度 (support) を持つという．また，相関ルール X ⇒ Y については， アイテムセット X ∩ Y のサポート度を相関ルール X ⇒ Y のサポート度と定義する．相関ルール X ⇒ Y について，X を含むトランザクションの内の c %のトランザクションが Y も含むとき，相関ルール X ⇒ Y は c の確信度（confidence）を持つという.

2.2.2　FP-growth 法　従来，サポート度の閾値である最小サポート度以上を満たすアイテム集合である頻出アイテム集合を求め，相関ルール抽出するのにアプリオリアルゴリズム [2] というアルゴリズムが使われてきたが，これは計算量が多く効率が悪い．そこで本研究では FP-growth アルゴリズム [3] を利用して相関ルールを抽出することにした．FP-growth 法は特殊なデータ構造である FP-tree を参照するだけで全ての頻出パターンを数え上げることができ，検索コストを大幅に減らすことができる.

図 1 の例を使って FP-tree 作成法を説明する．図 1 の左上の表は 9 つのトランザクションが入ったデータベースを表している．これが客の購買データを表しているとすると T100～T900 のトランザクションは客の購買，アイテム ID は商品をそれぞれ表しており，一行目のトランザクション T100 はある客が商品 I1,I2,I5 を購入していることを表している.

＜ FP-tree 作成法＞

データベースをスキャンし，それらのサポート数（頻出度）を引き出し，サポート度の降順で分類してリスト，L=[I2:7,I1:6,I3:6,I4:2,I5:2] を作り，図 1 の左下の図の node-link を作る．次に、null で表記される木の根を作り，データベース D をスキャンする．各々のトランザクションの中のアイテムは L 順の中で処理され，木はそれぞれのトランザクションから作られる．例えば第 1 のトランザクション “T100:I1，I2，I5”は L の中の（I2,I1,I5）の 3 つのアイテムを含み、〈(I2:1),(I1:1),(I5:1)〉の 3 つのノードで tree の最初の枝の構造を導く，このノード I2 は根の子として関連づけられ I1 は I2 と関連づけられ，I5 は I2 と関連づけられる．このようにトランザクション中のアイテムを順番にいれて図 1 の右のような tree を作る．枝がトランザクションのために付け加えられると，共通の先頭に沿った各々のノードの数は 1 増加させられ，先頭の後のアイテムのためのノードはそれに応じて作られる．木の走査を容易にするめ、各々のアイテムは node-links を経て木で発生をする．このようにして木を作ることによりデータベースで頻出パターンをマイニングする問題は FP-tree をマイニングする問題に置き変えられる.

＜ FP-tree のマイニング (最小サポート=2) ＞

FP-tree をマイニングすることにより表 1 に示すような頻出パターンが生成される．頻出アイテム集合の作り方としてここではれ例として I5 を考える．I5 は図 1 の FP-tree の 2 つの枝で発生する.（I5 の発生は node-links の連鎖によって容易に見つけられることができる.）これらの枝によって作られたパスは〈(I2 I1 I5:1)〉と〈(I2 I1 I3 I5:1)〉である．したがって，末尾として I5 を考えると，対応する 2 つの先頭パス〈(I2 I1:1)〉と〈(I2 I1 I3:1)〉で，それが表 1 の条件つきパターンベースを作る．ここで後ろの数字（ここでは 1）はパターンベースのサポート度を示している．この条件付きパターンベースから条件付 FP-tree を生成する．例の場合＜ I2,I1,I3 ＞のサポート度 1 は最小サポート度 2 よりも

Community Analysis of SNS using Association Rules by Yukako TOSHIMITSU

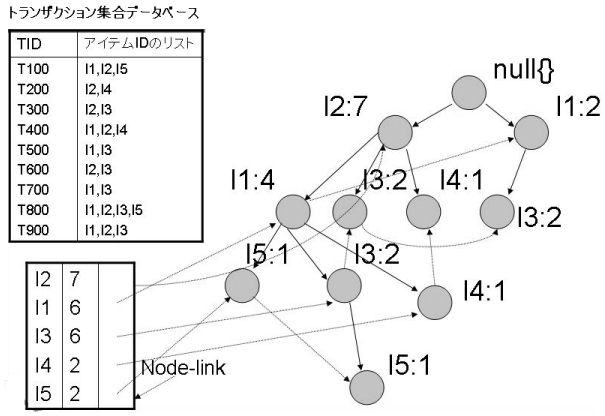

図 1: **データベースのトランザクションを tree に入れたときの様子**

小さいため I3 は含まれず条件付 FP-tree は< I2:2,I1:2 >となる．これから頻出パターンを生成すると，シングルパス全ての頻出パターンの組み合わせは I2 I5:2,I1 I5:2,I2 I1 I5:2 となる．よってこれからできる相関ルールは I2 ⇒ I5（サポート度=2/9, 確信度=2/7)I1 ⇒ I5(サポート度=2/9 確信度=2/6=1/3) I2,I1 ⇒ I5(サポート度=2/9 確信度=2/4)I2 ⇒ I1,I5(サポート度=1/2 確信度=2/9=2/7) となり，このようにして頻出アイテム集合を Node-link の降順で生成していく．

表 1　アイテムごとの頻出パターンの生成

アイテム	条件付きパターンベース	条件付き FP-tree	頻出パターン生成
I5	(I2 I1:1),(I2 I1 I3:1)	<I2:2,I1:2>	I2I5:2,I1I5:2,I2I1I5:2
I4	(I2 I1:1),(I2:1)	<I2:2>	I2I4:2
I3	(I2 I1:2),(I2:2),(I1:2)	<I2:4,I1:2><I1:2>	I2I3:4,I1I3:4,I2I1I3:2
I1	(I2:4)	<I2:4>	I2I1:4

3　実装と検証

データの収集では mixi に簡単にアクセスするためのモジュールである WWW::Mixi モジュールを利用して mixi サーバからユーザ ID とコミュニティID を抽出し，MySQL を利用してデータベースに格納した．データを収集する際に

手法 1．ランダムにユーザを 1000 人選んでそのユーザが入っているコミュニティの ID を集める

手法 2．あるコミュニティを基点にそのコミュニティに入っているユーザを集め，それぞれのユーザが入っているコミュニティを集める．

以上 2 つの手法で収集したデータを使い，FP-tree のプログラムを C 言語で作成して実際にマイニングを行った．結果を表 2 と表 3 に示す．

表 2　手法 1 と 2 でとったデータ数と最小サポートごとに得られる相関ルールの数

	手法 1	お茶大	大分市	cancam 好き	Disney マニア	Disney 嫌い
ユーザ数	1000	1311	1209	1912	1635	1388
コミュニティ数	18538	34887	31385	38340	43591	79086
support=15	0	33	49	373	203	189
support=20	0	12	27	221	137	42
support=25	0	4	14	182	82	30
support=30	0	1	9	125	69	8

表 3(コミュニティごとの代表的な相関ルール)

コミュニティ名	相関ルール	サポート度	確信度
お茶大	空を見る人⇒星空好き★	3.5 %	34.8 %
	よく物をなくす⇒期限ギリギリまで行動できない	3.8 %	32 %
大分市	よだきぃを標準語にする会⇒大分弁を話そう	3.0 %	62.5 %
	大分県のラーメン⇒大分トリニータ	8.3 %	24.8 %
cancam	物欲が止まらない⇒いい女になる秘訣	11.7 %	30.3 %
	ラブワンピ⇒スカート好きなんだもん	6.5 %	60.0 %
Disney マニア	Disney 最新耳より情報 ⇒ TOKYO Disney RESORT	13.9 %	39.9 %
	東京ディズニーランド⇒東京ディズニーシー	5.3 %	67.0 %
Disney 嫌い	フォント⇒ Photoshop	5.3 %	43.2 %
	笑える画像⇒資料になりそうなウェブサイト	6 %	37.3 %

表 2 は手法 1 と手法 2 で使ったコミュニティごとでのデータ収集で得られたユーザ数，ユーザの入っているコミュニティの総数，最小サポートを X 人とした時にそれぞれ抽出できたルールの数を示している．例えばお茶大の結果を例とするとユーザ数が 1311 人でありそのユーザは全部で 34887 種類のコミュニティに入っており, サポート値が 15 人の場合は 33 通り,20 人の場合 12 通り, 25 人の場合は 4 通り,30 人の場合は 1 通りの相関ルールが抽出されたことを表している．また表 3 は手法 2 で実際に抽出できた相関ルールとそのサポート度と確信度を表している．例として表の一番上の段で説明すると,「空を見る人」と「星空が好き★」の両方のコミュニティに入っている人は「お茶大」コミュニティに入っている人のうちの 3.5 %であり「空を見る人」のうちの 34.8 %は「星空が好き★」のコミュニティに入っていることを表している．

データをランダムに選ぶ場合，サポートの値を 10 まで小さくしないとルールを抽出することができなかったが，コミュニティごとに選ぶとコミュニティの特徴を反映するルールを得ることができた．また選ぶコミュニティによって得られるルールの数や性質に違いがみられた．地域や大学で選ぶよりも趣味や嗜好に沿ったコミュニティのほうがより多くの有益なルールを得ることができた．これは趣味や嗜好で選んだコミュニティに集まっているユーザはそれをもとに他のコミュニティに入るが，地域や学校で選んだコミュニティに集まっているユーザははその他のコミュニティを選ぶときにはそれぞれ個人の趣味や嗜好で選んでいるからだといえる．

4　まとめと今後の課題

本研究では相関ルールを利用して mixi のコミュニティによる傾向分析を行った．コミュニティごとにデータを集めて相関ルールを抽出するとそのコミュニティに入っている人の傾向を分析することができた．またその傾向分析は趣味や嗜好に偏ったコミュニティで行うほどコミュニティの特徴が大きく現れ有益なルールが得られた．この分析は社会科学にも役立てることができるのではないかと思う．今後は社会科学における活用についても考えていきたい．

[謝辞]

本研究を進めるにあたりご助言・ご指導いただいた本学情報科学科講師の渡辺知恵美先生に深く感謝致します．

参考文献

[1] mixi, http://mixi.co.jp/

[2] R.Agrawal and R.Srikant. “Fast algorithms for mining association rules.”In proceedigs of VLDB 1994, pp. 487–499, Santiago,Chile,Sept. 1994.

[3] J.Han, J.Pei, and P.S.Yu, “Mining Frequent Ptterns without Candidate Generation,” In Proceedings of the SIGMOD Conference 2000, pp. 1-12,2000.