異なるカテゴリの嗜好情報に対する特徴解析及び評価

山下 翔† 鈴木 育男† 山本 雅人† 古川 正志†
†北海道大学大学院情報科学研究科

1 はじめに

人の趣味・嗜好は「文学」,「音楽」,「映画」など様々なカテゴリに分かれている．また，近年 Web 上では各カテゴリ内でのユーザの行動履歴や評価からそのユーザの嗜好を推定し，推薦するサービスが増加している．例えば音楽の SNS の「last.fm」や，オンライン DVD レンタル「Netflix」などが挙げられる．

一方，現実世界では「何かと趣味の合う人」，つまり複数のカテゴリにおいて共通の嗜好を持つ人が存在するものである．これは異なるカテゴリであってもカテゴリを越えた嗜好の類似性があることを示しているといえる．また，カテゴリ A と B は類似度が高いが，A と C は低いといったように，カテゴリによっては類似度も異なると考えられる．

そこで本研究では実データを用いて複数のカテゴリにおける人の嗜好情報の関係をネットワークとして捉え，特徴の解析及び評価を行う．また，嗜好情報に基づいたカテゴリ間の類似度を定義し，その値を算出し考察する．

2 嗜好ネットワーク

今回扱うネットワークはユーザの嗜好を表すコンテンツをノードとする．リンクに関しては，例えば，あるユーザがコンテンツ A と B を選択した場合，A と B は類似しているとする．このとき 2 人以上が A と B を選択した場合，ユーザ数を重みとし，A-B 間にリンクを張る．重み 1 のリンクはある特定ユーザのみの嗜好を反映したものであり，関係が弱いと考えたため省いた．

また，今回はカテゴリを越えた嗜好の関係に注目しているため，リンクは異なるカテゴリに属する嗜好間のみに張ることにした．このときネットワーク G は以下で表される．

$$G = (V, E) \quad : \text{ネットワーク}$$
$$V = \{v_i \mid i = 1, 2, \ldots\} \quad : \text{ノード集合}$$
$$E = \{e_{i,j} \mid v_i, v_j \in V, w_{i,j} > 1\} \quad : \text{リンク集合}$$

ここでリンク $e_{i,j}$ は重み $w_{i,j}$ を持つ．

An analysis of preference informations between different categories
Sho YAMASHITA Ikuo SUZUKI Masahito YAMAMOTO Masashi FURUKAWA
†Graduate School of Science and Technology, Hokkaido University
sho_y@complex.eng.hokudai.ac.jp

3 解析方法

3.1 コミュニティ分割

ネットワークのトポロジーからネットワーク内の密に結合した部分集合を抽出するアルゴリズムに関する研究が多くされている．このアルゴリズムを前述の嗜好ネットワークに適用し，ネットワーク内で関係の強い嗜好コミュニティを抽出する．

アルゴリズムは Clauset らによるアルゴリズム [1] を適用する．このアルゴリズムは Modularity という指標を用い，その値を最大化することでコミュニティを抽出する．本研究ではこのアルゴリズムをリンクに重みを付けたものに拡張し適用した．具体的には Modularity の初期値を式 (1)，更新の際に用いる a_i を式 (2) に変更した．

$$\Delta Q_{i,j} = \begin{cases} \dfrac{w_{i,j}}{2w} - \dfrac{w_i w_j}{(2w)^2} & : \ w_{i,j} > 1 \\ 0 & : \ otherwise \end{cases} \tag{1}$$

$$a_i = \frac{w_i}{2w} \tag{2}$$

ここで w はネットワーク全体の重みの総和，w_i はコミュニティ i 内の重みの総和を表す．

3.2 カテゴリ間類似度の計算法

カテゴリ間でのリンクの割合が大きいほどカテゴリ間の類似度は大きいといえる．よってコミュニティ C_i におけるカテゴリ CA_j と CA_k の類似度 S_{C_i,CA_j,CA_k} を式 (3) で定義する．

$$S_{C_i,CA_j,CA_k} = \frac{L_{C_i,CA_j,CA_k}}{|V_{C_i,CA_j}||V_{C_i,CA_k}|} \tag{3}$$

$$C = \{C_i \mid i = 1, 2, \ldots\} \quad : \text{コミュニティ集合}$$
$$CA = \{CA_k \mid k = 1, 2, \ldots\} \quad : \text{カテゴリ集合}$$

L_{C_i,CA_j,CA_k}，V_{C_i,CA_j} はそれぞれコミュニティ C_i におけるカテゴリ CA_j-CA_k 間のリンク数，カテゴリ CA_j のノード集合を表す．各コミュニティにおける類似度の平均をとることでネットワーク全体におけるカテゴリ CA_j と CA_k の類似度 S_{CA_j,CA_k} とする（式 (4)）．

$$S_{CA_j,CA_k} = \frac{1}{|C|} \sum_{i=1}^{|C|} S_{C_i,CA_j,CA_k} \tag{4}$$

4 実データ解析

4.1 対象データ

今回はAmazonリストマニアのデータを対象として解析した．リストマニアとはAmazonの利用者が自分の好きな商品やお勧め商品をリストし，公開するサービスである．

リストマニアには複数のカテゴリの商品をリストでき，さらに商用性のあるアソシエイトでもない．よって複数カテゴリにおけるユーザの真の嗜好情報が詰まったデータといえる．このリストマニアで，あるユーザにリストされている商品をそのユーザの嗜好としてネットワークを作成する．

カテゴリはAmazonの分類に従い，今回はリストされた商品の多い「文学・評論」「コミック」「音楽」「映画」の4つを対象カテゴリとした．また,「文学・評論」「コミック」のノードは著者,「音楽」のノードはアーティストにまとめた．

データは2006/11/14から2006/12/16の期間で取得したものを利用する．

4.2 解析手順

1. 嗜好ネットワークの作成
2. コミュニティ分割
3. カテゴリ間類似度の計算

4.3 解析結果・考察

嗜好ネットワーク　ネットワーク全体のノード数は9,569，リンク数は93,356であった．カテゴリ間に重み2以上のリンクが多数あることからも，カテゴリを越えた嗜好の類似性があるといえる．

また，ネットワークの次数分布はベキ乗則に従っていた（図1）．これは嗜好ネットワークがハブとなる少数の嗜好とそれに類似する多数の少次数の嗜好から構成されていることを示している．

具体的には次数上位のものには「村上春樹（文学）」，「Mr.Children（音楽）」,「ショーシャンクの空に（映画）」など有名なものが多かった．このことから次数が上位の嗜好は他のどの嗜好ともつながりやすい，つまり一般的に人気の高いものであると考えられる．

コミュニティ分割　コミュニティ分割した結果，ネットワークは61個のコミュニティに分割された．一例を図2に示す．コミュニティによってはノード数1のものや逆に1000を超えるものも存在した．

サイズの大きいコミュニティには複数のハブが存在していた．これは一般的に人気のある嗜好は他のどの嗜好ともリンクしやすく，ハブ同士がリンクしている場合や共通の嗜好にリンクしているハブが同じコミュニティに属されるためと考えられる．

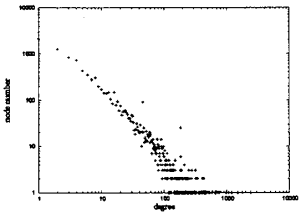

図1: 次数分布

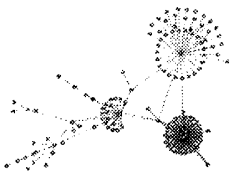

図2: コミュニティ例

表1: カテゴリ間類似度

カテゴリ	Similarity
文学-コミック	0.1819
文学-音楽	0.1567
文学-映画	0.1516
コミック-音楽	0.0601
コミック-映画	0.0731
音楽-映画	0.2441

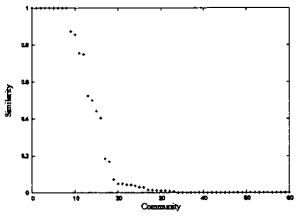

図3: 音楽-映画の類似度

カテゴリ間類似度　各カテゴリ間の類似度は表1のようになった．表からもわかるようにカテゴリの組み合わせによって類似度に差が見られる．今回は「音楽-映画」が最も類似度が高く,「コミック-音楽」が最も低い値となった．

また，図3は各コミュニティにおける「音楽-映画」の類似度を降順にソートした図である．図より，類似度が急激に低くなっていることがわかる．これはコミュニティによってカテゴリ間の類似度が高いか低いかがはっきり分かれていることを示している．

以上よりユーザにとって未知のカテゴリの嗜好を推定する際，別のどのカテゴリの嗜好から推定すればよいかを決定するための指標とするなどの応用が期待できる．

5 おわりに

本研究ではカテゴリを越えた人の嗜好関係を実データを用いてネットワークを形成し，コミュニティ分割により類似した嗜好の抽出を行った．さらに，カテゴリ間の類似度を定義・算出することでカテゴリによりつながりの強さが異なることを示した．

今後の展望としてはカテゴリ間類似度の妥当性の評価や，情報推薦などへの応用が挙げられる．

参考文献

[1] A.Clauset,MEJ.Newman and C.Moore: Finding community structure in very large networks. Phys. Rev.E,Vol 70,066111,2004