F-054

ネットワークモデルを用いた異なるカテゴリ間の嗜好性の解析

An analysis of preference information between different categories using network model

山下翔† 鈴木育男† 山本雅人† 古川正志†

1 はじめに

近年、様々な企業がE-commerceを導入し、各ユーザに対してお勧め商品を提示し、売上の向上を図っている。このためにはユーザの嗜好情報の解析が重要となっている[1]。これらは「音楽」「映画」「ファッション」等、あるカテゴリ内でのユーザの嗜好情報を基にコンテンツを推薦するものがほとんどである。

しかし、現実には複数のカテゴリにおいて嗜好の近い共通の人が存在する。このことから、あるカテゴリで嗜好の近いユーザ群は別のカテゴリでも嗜好が近いと仮定できる。つまり、カテゴリを越えた解析を行うことでより幅広いコンテンツ推薦を実現できる。また、異なるカテゴリのコンテンツを提供する企業間でもターゲットとするユーザ群が共通であると発見できればビジネスの幅が広がることも期待できる。

そこで本研究では、まずコンテンツとユーザ間の嗜好関係を表現するネットワーク理論モデルを構築する。また、構築した理論モデルに基づき異なるカテゴリ間におけるユーザの嗜好性を実データを利用して解析する。

2 理論モデル

本研究ではカテゴリごとにユーザとコンテンツ間の嗜好関係をネットワークで表現する。各カテゴリにおけるネットワークを解析することで嗜好の近いユーザ群と、そのユーザ群の嗜好を表すコンテンツ群をコミュニティとして抽出する。抽出したコミュニティ間の類似度を定義し、カテゴリの異なるコミュニティ間の類似度を算出し、カテゴリを越えた嗜好性の解析を行う。

2.1 ネットワークモデル

各カテゴリにおいてユーザとコンテンツをノードとし、ユーザがコンテンツを好む場合にユーザとコンテンツ間にリンクを貼り、二部グラフを作成する。これによりユーザとコンテンツの嗜好関係を表すネットワークを作成できる。カテゴリ i に関するネットワーク N_i は以下のように表現する。

$$N_i = (P_i, U, E_i) \tag{1}$$

$$P_i = \{p_{i1}, p_{i2}, \ldots, p_{in_{p_i}}\} \tag{2}$$

$$U = \{u_1, u_2, \ldots, u_{n_u}\} \tag{3}$$

$$E_i \subset P_i \times U \tag{4}$$

ここで、n_{p_i} はカテゴリ i のコンテンツ数、n_u はユーザ数、P_i は N_i のコンテンツ集合、U はユーザ集合、E_i は N_i のリンク集合である。

また、$P_i \cap P_j = \emptyset$、$E_i \cap E_j = \emptyset$ $(i, j \in \{1, \ldots, m\}, i \neq j)$ とし、あるコンテンツやリンクは複数のカテゴリに存在しない。ここで m はカテゴリ数である。

2.2 コミュニティモデル

各ネットワークからリンクが密である二部グラフをコミュニティとして抽出する。N_i から抽出したコミュニティ集合を $C_i = \bigcup_j C_{ij}$ $(j \in \{1, \ldots, n_{C_i}\})$ とし、C_{ij} を以下のように表す。ただし、n_{C_i} はカテゴリ i のコミュニティ数である。

$$C_{ij} = (P_{ij}, U_{ij}, E_{ij}) \tag{5}$$

$$P_{ij} = \{p_{ij,k} \mid p_{ij,k} \in P_i\} \tag{6}$$

$$U_{ij} = \{u_{ij,k} \mid u_{ij,k} \in U\} \tag{7}$$

$$E_{ij} \subset P_{ij} \times U_{ij} \tag{8}$$

ここで P_{ij}、U_{ij}、E_{ij} はそれぞれ C_{ij} のコンテンツ集合、ユーザ集合、リンク集合を表す。また、$P_{ij} \cap P_{ik} = \emptyset$、$U_{ij} \cap U_{ik} = \emptyset$ $(j, k \in \{1, \ldots, n_{C_i}\}, j \neq k)$ であり、同一カテゴリ内において、あるノードは複数のコミュニティに属さないとする。U_{ij} は P_{ij} を好むユーザの集合であり、嗜好が近いユーザ集合となる。

2.3 コミュニティ間類似度

カテゴリ i におけるコミュニティ C_{ij} とカテゴリ k におけるコミュニティ C_{kl} 間の類似度 $S(C_{ij}, C_{kl})$ を定義し、算出する。算出結果によりカテゴリを越えたコミュニティ間の関連性を解析する。

† 北海道大学大学院情報科学研究科

3 実験

これまで述べた理論モデルに基づき、実データを利用して異なるカテゴリ間の嗜好性を解析する。

3.1 対象データ

今回は Amazon のリストマニアのデータを対象として解析する。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りユーザの嗜好性を表すデータである。さらにリストマニアには複数のカテゴリの商品でもリストできるため今回の実験に適したデータである。

カテゴリは Amazon の分類に従い、「文学・評論」「コミック」「映画」「音楽」の 4 つを対象とした。また、「文学・評論」「コミック」のコンテンツを著者、「音楽」のコンテンツを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。

3.2 コミュニティ抽出手法

密な二部グラフを抽出する手法として、沈ら [2] の提案するウェブコミュニティ抽出アルゴリズムを利用する。このアルゴリズムにより抽出されるコミュニティ内では、任意の 2 つのコンテンツは必ず N 人以上の共通ユーザとリンクしている。つまりネットワークからより密に連結した二部グラフを抽出できる。今回は $N = 3$ としてコミュニティを抽出する。

3.3 コミュニティ間の類似度

今回、2 つのコミュニティ間で共通ユーザが占める割合をコミュニティ間の類似度と定義した。コミュニティ C_{ij} と C_{kl} 間の類似度は以下の式で表す。

$$0 \leqq S(C_{ij}, C_{kl}) = \frac{|U_{ij} \cap U_{kl}|}{|U_{ij}| + |U_{kl}| - |U_{ij} \cap U_{kl}|} \leqq 1 \quad (9)$$

4 結果・考察

今回、コミュニティ抽出において $N = 3$ としたため、3 人以上のユーザにリストされたコンテンツのみを対象としたネットワークを作成した。ユーザ数は 4,701 であり、各カテゴリにおけるコンテンツ数、リンク数は**表 1** に示した。

表 1: 各カテゴリのネットワーク

カテゴリ	コンテンツ数	リンク数
文学・評論	2,217	20,428
コミック	1,276	12,989
映画	3,306	25,814
音楽	5,191	44,438

4.1 コミュニティ抽出結果

各カテゴリにおけるコミュニティ抽出によって得られたコミュニティの中にはサイズが 100 を超える大きなものが存在した。しかし、半分以上は 10 未満の小さいコミュニティとして抽出された (図 1)。サイズの大きいコミュニティは、幅広く様々なコンテンツを好むユーザや一般的にどのユーザからも人気のあるコンテンツの影響によりできたと考えられる。逆に小さいコミュニティは少数のユーザがマニアックなコンテンツを好んでいることを表していると考えられる。

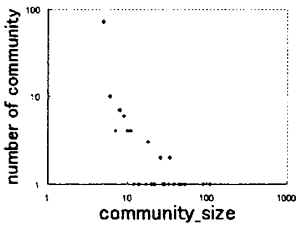

図 1: 「音楽」のコミュニティサイズ分布

表 2: 最大類似度

カテゴリ	類似度
文学-コミック	0.3333
文学-音楽	0.2
文学-映画	0.2093
コミック-音楽	0.2
コミック-映画	0.2
音楽-映画	0.2

4.2 類似度の計算結果

今回定義した類似度の算出結果は最大でも 0.3333 と大きなものは存在しなかった (**表 2**)。その理由として、サイズの大きく異なるコミュニティ間の類似度は共通ユーザ数が少なく小さくなってしまう点が考えられる。また、特定のカテゴリのみにリンクのあるユーザも多くいることから、そもそもコミュニティサイズに比べ共通ユーザが少ない可能性も考えられる。

5 まとめ

本研究ではカテゴリを越えた人の嗜好性を解析するため、ユーザとコンテンツの嗜好関係を表すネットワーク理論モデルを構築し実データを利用し解析を行った。問題点として二部グラフのコミュニティ抽出手法の決定や類似度の定義が挙げられる。これらは手法や定義により結果が異なるため、どの手法や定義を選択するか吟味する必要がある。また、類似度がどの値から 2 つのコミュニティが近いか判断するための閾値の決定なども今後の課題に挙げられる。

参考文献

[1] J.B Schafer,J Konstan,J Riedl.Recommender Systems in E-Commerce,1999

[2] 沈 垣甫, 田浦 健次朗, 近山 隆. ウェブコミュニティ抽出アルゴリズムの改良,2007