次元圧縮法を導入したクエリに基づくバイクラスタリング—情報推薦への応用—

次元圧縮法を導入したクエリに基づくバイクラスタリング—情報推薦への応用—

次元圧縮法を導入したクエリに基づくバイクラスタリング
—情報推薦への応用—

○武内充　三浦功輝　岡田吉史（室蘭工業大学）

概要　以前，我々はクエリに基づくバイクラスタリングを用いた情報推薦手法を提案した．本研究では，新たに推薦スコアが非常に良く似たユーザまたはアイテムを融合する次元圧縮法を導入した．実験として，縮減前と縮減後のデータセットのサイズとバイクラスタ計算時間の比較を行う．
キーワード: 情報推薦,バイクラスタリング

1　はじめに

インターネットの普及により，莫大な情報を得られるようになった反面，自分の目的に合った情報のみを見つけ出すことは非常に困難になっている．そのためにユーザの意図や嗜好に合った情報を推定して提示する情報推奨技術の研究が盛んに行われている．

以前，我々はクエリに基づくバイクラスタリングを用いた情報推薦手法(以下，前手法)を提案した [1)]．この手法は，推薦要求ユーザが高評価しているアイテムをクエリとし，そのクエリを高評価しているユーザのみを用いてバイクラスタリングを行う．この手法によりバイクラスタリングに要する計算時間を劇的に縮減することが出来た．しかし，この方法では，同じアイテムを含む膨大なバイクラスタが生成されるため，推薦アイテムの選定において冗長な計算が行われていただけでなく，アイテムのランキングが適切に行われなかった可能性がある．

本研究では，評価スコアが非常によく似たユーザ，またはアイテムを融合する次元圧縮法を導入したアイテム推薦の手法を提案する．本稿では，本手法によるデータセット縮減効果の結果に加え，縮減前後のデータセットからのバイクラスタリング生成時間の結果を報告する．

2　本手法の概要

2.1 データセット

本研究では, GroupLensで公開されている情報推薦システムのベンチマークデータセットMovieLensを用いる[2)]．MovieLensは1682本の映画に対して943人のユーザが1から5までの5段階評価を行ったデータが10万件格納されている．また，ユーザ一人当たりの最低評価件数は10件である. データセットに含まれる評価値は，大きいほど高評価であることを意味し, 4以上を高評価，それ以外を低評価と定義する．

2.2 トランザクションデータベース作成

使用するデータセットから, トランザクションデータベースを構築する. 本実験で使用するトランザクションデータベースは，行にユーザ，列にアイテム，要素に評価値が格納されている．評価値は高評価を1，低評価を0とする. 構築したトランザクションデータベースをユーザに関して3分割し，その内の1つをテストデータセット，残りの2つを訓練データセットとする．訓練データセットのみを縮減することで圧縮データセットを作成する．

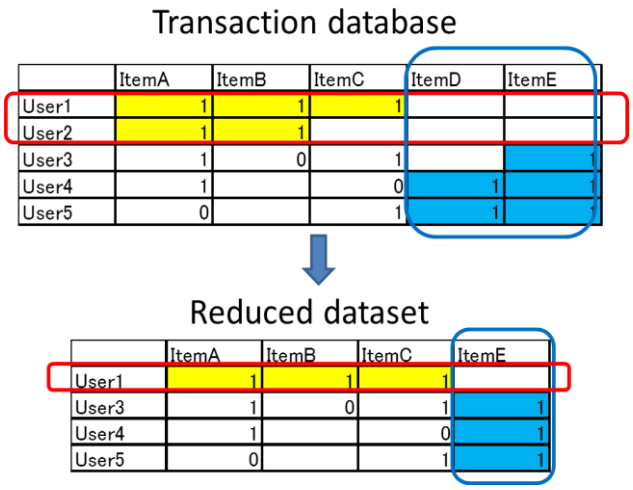

Fig 1:　Reduction procedure

2.3 圧縮データセット作成

Fig. 1は訓練データセットから評価スコアが非常によく似たユーザ，またはアイテムを融合して圧縮データセットを作成する方法を示している．

ユーザの融合手順は，以下のとおりである：

1) ユーザをアイテムの評価件数が多い順にソートする．
2) 評価件数が最も多い未融合のユーザを親ユーザとし，それ以外を子ユーザとする．
3) 親ユーザと各子ユーザの間でアイテムに対する評価を比較し，以下の条件を全て満たす子ユーザを削除する．
   - 未融合である．
   - 親ユーザと子ユーザが共に評価しているアイテムの評価値が全て一致している．
   - 親ユーザが評価しているアイテムの内，α%以上のアイテムを子ユーザが評価している．
4) 2)から4)を全てのユーザが親ユーザになるまで繰り返す．

Table 1: dataset and Reduction Rate

	BeforeReduction	IU	UI
UserValue(ReductionRate)	627	161(74%)	478(24%)
ItemValue(ReductionRate)	1682	1046(38%)	264(84%)

アイテムの融合手順は，以下のとおりである．

1) アイテムを被評価件数が多い順にソートする．
2) 被評価件数が最も多い未融合のアイテムを親アイテムとし，それ以外を子アイテムとする．
3) 親アイテムと各子アイテムの間でユーザからの評価を比較し，以下の条件を全て満たす子アイテムを削除する．
   - 未融合である．
   - ユーザが親アイテムと子アイテムを共に評価している場合，それらの評価値が一致している．
   - 親アイテムを評価しているユーザの内，α%以上のユーザが子アイテムを評価している．
4) 2)から4)を全てのアイテムが親アイテムになるまで繰り返す．

本実験では$\alpha = 50$とした．圧縮データセットは先にユーザを融合した場合と先にアイテムを融合した場合の2種類が存在する．

2.4 バイクラスタリングとアイテム推薦

圧縮データセットから，飽和集合マイニングに基づくバイクラスタリング法を用いて，バイクラスタの抽出を行う[3]．この方法は，LCM(Linear time Closed itemset Miner)[4]と呼ばれる飽和集合列挙アルゴリズムを用いており，指定された最小サポート数(最小ユーザ数)と最小アイテム数のもとで網羅的なバイクラスタ探索を行うことが出来る．抽出したバイクラスタに含まれるアイテムiのスコアは1)式で定義される[5]．

$$Score(i) = \sum_{b_i} \frac{|I_q \cap I_{b_i}|}{|I_{b_i}|} \times |U_{b_i}| \quad \cdots 1)$$

ここで，I_qは推薦要求ユーザがクエリとして入力したアイテムの集合，I_{b_i}はバイクラスタb_iに含まれるアイテムの集合，U_{b_i}はバイクラスタb_iに含まれるユーザの集合である．

3 実験

本稿では本手法でどの程度データセットを縮減できたのか純粋に調べるため，縮減前後のデータセットのサイズを比較する．また，縮減前後のデータセットでそれぞれバイクラスタを生成し，生成に要した時間の比較を行う．この時，クエリに基づくバイクラスタリングは行わず，本手法のみで縮減されたデータセットにおける結果を求める．バイクラスタ生成のパラメータは最小ユーザ数を 20，最小アイテム数を 5 とする．計算機環境は Intel Xeon Processor X5680 ，3.33GHz ,24GB RAM を搭載した PC であり，OS は Ubuntu 14.04.1 LTS である．

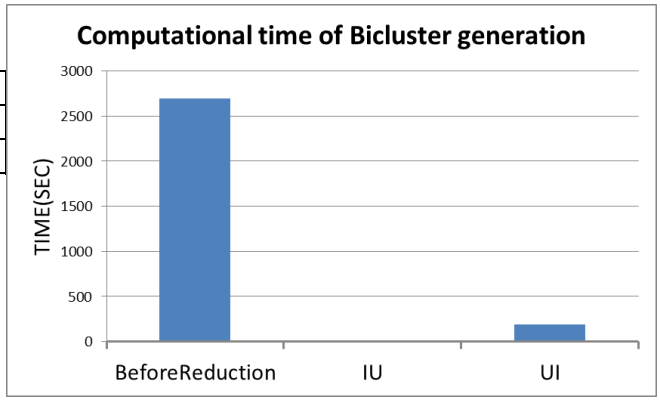

Fig. 2: Computational time of Bicluster generation

4 結果・考察

Table. 1 は縮減前と縮減後のデータセットのサイズを示している．ここでのサイズとは，ユーザ数×アイテム数の評価値行列の要素数を指す．この Table における IU はアイテム→ユーザの順に縮減を行ったときの結果を，UI はユーザ→アイテムの順に縮減を行った結果を示している．また，Table 内の数値は，縮減前の要素数，および縮減後の要素数と縮減率を表している．Table. 1 から IU，UI ともにデータセットが劇的に縮減されていることが分かる．また，アイテムを先に縮減した IU ではユーザの縮減率がより大きく，ユーザを先に縮減した UI ではアイテムの縮減率がより大きいということが分かる．これは，先にユーザを融合すると，ユーザの次元が圧縮されアイテム間で融合が起きる可能性が高くなり，先にアイテムを融合するとアイテムの次元が圧縮されユーザ間で融合が起きる可能性が高くなった等が原因として考えられる．Fig. 2 はバイクラスタ生成時間を示している．Fig. 2 から 縮減前と比べてバイクラスタ総生成時間は大幅に減少したことが分かる．

5 まとめ・今後の課題

本稿では次元圧縮法を導入しデータセットの評価スコアが良く似たユーザ，アイテムを融合する手法を提案した．訓練データセットを縮減し，バイクラスタ生成時間を削減することが出来た．今後は，縮減データセットに対しクエリを用いたバイクラスタリングを行い，抽出されたバイクラスタを用いて推薦精度を算出する．また，他の異なるデータセットにも適用し，本手法の問題点の吟味とさらなる改善へ向けた検討を行っていく．

参考文献

1) 横山直也，岡田吉史，"クエリに基づくバイクラスタリングを用いた協調フィルタリング法"，日本感性工学会生命ソフトウェアシンポジウム 2014, 2014.
2) GroupLens_Research: http://www.grouplens.org

3) Y . Okada, W . Fujibuchi, and P . Horton,“A biclustering method for gene expression module discovery using a closed itemset enumeration algorithm”, IPSJ Trans. on Bioinformatics, 48(SIG 5(TBIO2)), pp.39-48, 2007.
4) 宇野 毅明,有村 博紀:“飽和集合列挙アルゴリズムを用いた大規模データベースからのルール発見手法”,統計数理,vol.53, no.2, pp.317-329, 2005.
5) P.Symeonidis, A. Nanopoulos, and Y.Manolopoulos, “Near est Biclusters Collaborative Filtering,” 2006.