D-025

アイテムシーケンスデータからの 頻出否定シーケンシャルパターン抽出方式の検討

Mining frequent negative sequential patterns from item-sequene database

蘇 麗妍 † 新谷 隆彦 ‡ 大森 匡 ‡ 藤田 秀之 ‡
Liyan Su Takahiko Shintani Tadashi Ohmori Hideyuki Fujita

1. はじめに

シーケンシャルパターンマイニングはアイテムの発生順序を考慮して頻出パターンを抽出する手法である。様々なデータに適用されるようになり、発生すること (肯定) だけでなく、発生しないこと (否定) も順序を付けて並べたパターンである否定シーケンシャルパターン (以降、否定パターンと呼ぶ) が考慮されるようになった。大量の探索候補を調べなければならず、意味のない否定パターンが作成されるため、否定パターンの制約条件とその制約条件を満す否定パターンを効率良く抽出する方式の研究が進められてきた [1, 2]。これらでは否定を考慮するアイテムを肯定頻出となるアイテムに限定する手法、否定パターンを構成するアイテムをすべて肯定としたときに頻出となるパターンに限定する手法が提案された。しかし、頻出でないアイテムやパターンは考慮されないため、有用な否定パターンが抽出できない場合がある。

本研究では、イベントがアイテムで構成されるシーケンスデータから、意味のある否定パターンをすべて抽出するための制約条件を検討し、頻出肯定アイテムに否定アイテムを挿入することで否定パターンを作成し、頻出とならない候補のカウントを回避し、ビットマップを用いて頻度をカウントする手法を提案し、評価実験を行った。

2. 問題定義

肯定アイテムの集合を $\mathcal{I}p = \{i_1, \ldots, i_m\}$ とする。シーケンスデータベースはシーケンスデータ s の集合である。本研究では、発生した時刻の順に並べられたアイテムのリストからなるシーケンスデータ $s_i = \langle e_1, \ldots, e_s \rangle$ $(e_j \in \mathcal{I}p,\ 1 \leq j \leq s)$ を処理対象とする。これをアイテムシーケンスデータと呼ぶ。発生しないアイテムを否定アイテムと呼び、否定アイテムの集合を $\mathcal{I}n = \{\neg i_1, \ldots, \neg i_n\}$ とする。また、肯定アイテムと否定アイテムをアイテム $\mathcal{I} = \{\mathcal{I}p, \mathcal{I}n\}$ と呼ぶ。

アイテムを順に並べたリストがシーケンシャルパターンである。肯定アイテムのみからなるシーケンシャルパターンを肯定シーケンシャルパターン (肯定パターン) と呼ぶ。1 個以上の否定アイテムを含むシーケンシャルパターンを否定シーケンシャルパターン (否定パターン) と呼び、$NS = \langle c_1, \ldots, c_k \rangle$ $(\exists c_i \in \mathcal{I}n、1 \leq i \leq k)$ とする。否定パターンを構成するアイテムの数をサイズと呼ぶ。2 つの否定パターン NS_a と NS_b について、NS_a のすべてのアイテムが順序が保持されたままで NS_b に存在するとき、NS_a を NS_b のサブパターンと呼び、$NS_a \subseteq NS_b$ と表現する。ここで、否定パターン NS の先頭から k 個のサブパターンを先頭サブパターン NS_{pre}、NS のすべての肯定アイテムを順序を保持して取り出したパターンを最大肯定サブパターン $MPS(NS)$、NS のすべての否定アイテムを肯定に変更したパターンを肯定パートナー、NS のすべての肯定アイテムと 1 個の否定アイテムからなるサブパターンを 1- 否定サブパターン $1negNS$ と呼ぶ。

本研究では頻出する否定パターンをすべて見つけ出すために、フォーマットの制約として否定アイテムが連続しないこと、先頭末尾アイテムの制約として否定パターンの先頭と末尾のアイテムは肯定アイテムであることを制約条件とした。従来手法では、さらに肯定パートナーの制約として否定パターンを構成するアイテムと否定パターンの肯定パートナーが頻出でなければならないことも制約条件とされていた。肯定パートナーの制約は頻出する肯定パターンと関連のある否定パターンのみに限定するため探索する候補が少なくなるが、頻出とならないパターンは考慮されないことになり、多くの頻出する否定パターンの抽出を漏らすことになる。肯定パーナーが頻出でなくとも、否定アイテムを含むときに頻出となるパターンは無視できない。本研究の 2 つの制約条件により、発生しないことの順序を決定できない問題が解決され、開始と終了を確定できる否定パターンをすべて抽出できる。

ここで、アイテムシーケンスデータが否定パターンを含むことを定義する。アイテムシーケンスデータは肯定アイテムのみから構成され、否定アイテムは現われない。否定アイテムは発生しないことを意味するため、アイテムシーケンスデータに現れないことを考慮する必要がある。アイテムシーケンスデータ s が否定パターン NS を含むとは、NS のすべての肯定アイテムが s に順序を保持されたままで存在し、NS の各肯定アイテムの間にある否定アイテムの肯定パートナーが s のそれら肯定アイテムの間に存在しないとき $MPS(NS) \subseteq s$ かつ $\forall 1negNS \not\subseteq s$ のとき、s は NS を含む。

否定パターン抽出問題は、アイテムシーケンスデータベースからユーザが指定した頻度の最小値を満たし、フォーマットの制約と先頭末尾アイテムの制約を満たす否定パターンをすべて抽出することである。このような否定パターンを頻出否定パターンと呼び、否定パターン NS の頻度を $sup(NS)$ と表現する。

3. 提案手法

3.1. 否定パターンの作成

頻出肯定パターンに 1 個の否定アイテムを挿入すること、否定パターンに 1 個の否定アイテムを否定アイテムが連続しないように挿入することを繰り返して否定パターンを作成する。頻出肯定パターンは従来手法

†電気通信大学大学院情報システム学研究科
‡電気通信大学大学院情報理工学研究科

用いて抽出する。ここで、否定パターン NS の末尾に肯定アイテムを追加した否定パターンの頻度は、NS 以上となる可能性があるため、Aprior の性質による枝刈りができない。しかし、NS に 1 個の否定アイテムを挿入した否定パターン NS' の頻度は、NS の頻度より大きな値に成り得ないため、頻出でない NS に対する NS' は枝刈りできる。

3.2. 冗長なカウントの省略

カウントをせずに頻度を知ることができる否定パターンがある。頻出肯定パターン PS と PS の i 番目の肯定アイテムの次に 1 個の否定アイテムを挿入した否定パターン $NS1_i$ について、PS と $NS1_i$ の頻度が等しいとき、$NS1_i$ の $i+j$ 番目の肯定アイテムの次に 1 個の否定アイテムを挿入した否定パターン $NS2_{i+j}$ のカウントを省略できる。PS と $NS1_i$ の頻度が等しいので、PS の j 番目の肯定アイテムの次に否定アイテムを挿入した否定パターン $NS1_j$ と $NS2_{i+j}$ の頻度が等しくなるためである。同様に否定パターン $NS2$ と $NS2$ に 1 個の否定アイテムを挿入した否定パターン $NS2'$ の頻度が等しいとき、$NS2'$ にさらに否定アイテムを挿入した否定パターンのカウントを省略できる。

短い否定パターンの結果から、さらに否定アイテムを挿入した否定パターンをカウントするかを判定するため、冗長なカウントの省略では 2 個の肯定アイテムからなる頻出肯定アイテムセットに 1 個の否定アイテムを挿入したパターンから順に、より長い頻出肯定アイテムセットに 1 個の否定アイテムを挿入したパターン、否定パターンにさらに 1 個の否定アイテムを挿入したパターンの作成とカウントを続けることによって頻出否定パターンを抽出する。

3.3. 上限値による枝刈り

否定パターンの頻度から、その先頭サブパターンである否定パターンの頻度の上限値を計算できる。上限値が頻度の最小値を満たさない場合カウントを省略出来る。否定パターン NS と NS の先頭サブパターン NS_{pre} について、NS_{pre} カウントの対象となるアイテムシーケンスデータ $D^{NS_{pre}}$ は、$MPS(NS)$ を含むアイテムシーケンスデータ $D^{NS_{pre}}_{MPS(NS)}$ と $MPS(NS)$ を含まないアイテムシーケンスデータ $D^{NS_{pre}}_{\neg MPS(NS)}$ に分けられる。$D^{NS_{pre}}_{MPS(NS)}$ のうち、NS を含まないアイテムシーケンスデータは NS_{pre} を含まないため、NS_{pre} の上限値は

$$\text{上限値}\,(NS_{pre}) = sup(NS) + sup(MPS(NS_{pre})) - sup(MPS(NS))$$

で計算できる。上限値が頻度の最小値を満たさないとき、NS_{pre} のカウントを行わない。

長い否定パターンの頻度からその先頭サブパターンをカウントするかを判定するため、長いパターンから順により短い否定パターンの作成とカウントを続けることによって頻出否定パターンを抽出する。

3.4. 頻度のカウント

本研究ではビットマップを用いて否定パターンの頻度をカウントする。頻出肯定パターン抽出手法である SPAM[3] と同様に各アイテムのビット列を作成する。アイテムのビット列はアイテムシーケンスデータでそのアイテムが現れるシーケンスの位置のビットを 1、現われない位置を 0 とする。肯定パターン $PS = \langle i_1, \ldots, i_n \rangle$ の r 番目の肯定アイテムの次に否定アイテム $\neg x$ を挿入した否定パターンを $NS1$ とする。各アイテムシーケンスについて、PS の r 番目までの先頭サブパターン $\langle i_1, \ldots, i_r \rangle$ 最初に現れた位置と、$r+1$ 番目から末尾までのサブパターン $\langle i_{r+1}, \ldots, i_n \rangle$ が最後に現れた位置の間に $\neg x$ の肯定パートナーが現れるかどうかをビット列を用いて調べる。ビット列に 1 が残らないアイテムシーケンスの数を $NS1$ の頻度とする。$NS1$ が現れるすべての位置のビットを 1 とするビット列 (拡張用ビット列) を作成する。否定パターンに更に 1 個の否定アイテムを挿入した否定パターンをカウントするとき、拡張用ビット列を用いることによって元の否定パターンが現れるアイテムシーケンスデータを知ることができる。

4. 実験結果

提案する否定パターン抽出方式と冗長なカウントの省略と上限値による枝刈りの評価実験を行った。実験にはクリックストリームデータセットである MSNBC[4] を用いた。頻度の最小値を全体の 0.25% として否定パターンを抽出したとき、提案する否定パターン抽出方式の処理時間に対して、冗長なカウントの省略は約 56%、上限値による枝刈りは約 62% の削減を実現できた。また、この時に抽出できた否定パターンは約 4000 個であったが、肯定パートナーの制約を加えた手法では 5 個しか抽出できなかった。

5. おわりに

肯定パートナーの制約をなくすことによって頻出する否定パターンが抽出できない問題を解決する制約条件を設定し、必要のない否定パターンのカウントを回避して頻出否定パターンを抽出する手法を提案した。実験によって提案手法で多くの頻出否定パターンを抽出できることを確認した。

謝辞

本研究は JSPS 科研費 15H02696 の助成を受けたものです。

参考文献

[1] S.C.Hsueh,M.Y.Lin,C.L.Chen,”Mining Negative Sequential Patterns for E-Commerce Recommendations”, IEEE APSCC, pp.1213-1218,2008.

[2] X. Dong,Z.Zheng,Y.Zhao,”e-NSP:Efficient Negative Sequential Pattern Mining Based on Identified Positive Patterns Without Database Rescanning”, ACM CIKM,pp.825-830,2011.

[3] J.Ayres,J.Gehrke,T.Yiu,and J.Flannick,”Sequential PAttern Mining using A Bitmap Representatio”, ACM SIGKDD,pp.429-435,2002.

[4] UCI Machine Learning Repository: archive.ics.uci.edu