頻出重み付き系列の
マイニングアルゴリズムに関する研究

和田・犬塚 研究室　　柴田 智幸
No. 16417557　　Shibata Tomoyuki

1 序論

テキストが単語の列で構成されるようにアイテムの列からなる集合から，頻出するパターンを抽出することは有用である．

この問題は，頻出系列マイニング問題として Agrawal らにより定式化 [1] されている．即ち，与えられた系列データベースにおいて，ユーザが定義する最小閾値以上の回数出現する部分系列をすべて抽出するタスクである．

本論文では，系列データとして，アイテムに対し重要さなどの重みが付いているとし，このとき，その重みを考慮して有用な系列を抽出することを考える．従来の系列マイニングの方法の多くは，ある系列が頻出するならば，その部分系列も頻出するというアプリオリの原理を利用している．しかし，アイテムが重みを持ち，これを考慮して頻出を定義したとき，この原理をそのまま適用することが難しい．

本論文では，重みを持つ系列も扱うことのできるマイニング手法を検討する．ここで，個々のアイテムの重みや，そこから決まる系列の重みはユーザの用途に応じて適切に定義するものとする．頻出重み付き系列マイニング問題を定義し，これを解く手法を二つ提案する．

2 頻出重み付き系列マイニング問題

$I = \{i_1, i_2, \ldots\}$ を，**アイテム**集合，I の部分集合を**エレメント**と呼ぶ．**系列**とはエレメントの列であり，$\langle e_1 e_2 \cdots e_l \rangle$ と書く．あるアイテム i がある系列 α 中のアイテムであることを，$i \lhd \alpha$ と書く．

定義 1 (部分系列)． 系列 $\alpha = \langle a_1 a_2 \cdots a_m \rangle$，$\beta = \langle b_1 b_2 \cdots b_n \rangle$ に対し，$a_1 \subseteq b_{j_1}, a_2 \subseteq b_{j_2}, \cdots, a_m \subseteq b_{j_m}$ となる整数 $1 \leq j_1 < j_2 < \cdots < j_m \leq n$ が存在する場合，α は β の部分系列といい，$\alpha \sqsubseteq \beta$ と書く．

系列データベースとは，タプル $\langle sid, s \rangle$ の集合である．sid は系列識別子，s は系列である．系列 α の系列データベース D における**サポート**とは，D 中のすべての系列のうち，系列 α を部分系列とする系列 s を含むタプルの数であり，$support_D(\alpha)$ と書く．

$$support_D(\alpha) = |\{\langle sid, s \rangle \in D | \alpha \sqsubseteq s\}|$$

アイテム i の重み$w(i)$ は，アイテム集合 I から正の実数への関数 w で与えられる．**系列 α の重み**$w(\alpha)$ は，$w : I$ の系列の集合 $\rightarrow R^+$ となる関数であり，アイテム i の重み $w(i)$ と系列 α の重みとしての $w(i)$ の値は一致するもので与えられる．この系列の重みの拡張の仕方は，ユーザによって与えられる．例えば，系列 α の重みとして，$w_{min}(\alpha) = \min_{i \lhd \alpha} w(i)$，$w_{max}(\alpha) = \max_{i \lhd \alpha} w(i)$ などが考えられる．

定義 2 (頻出重み付き系列)． 系列データベース D で系列 α が**頻出重み付き系列**であるとは，**サポート閾値**である任意の正の実数 min_supp に対し，$support_D(\alpha) \times w(\alpha) \geq min\_supp$ の場合をいう．

定義 3 (頻出重み付き系列マイニング問題)．頻出重み付き系列マイニング問題とは，系列データベース，サポート閾値，重み関数が与えられたとき，頻出重み付き系列をすべて抽出することである．

3 PrefixSpan

PrefixSpan[2] は，短い頻出系列からより長い頻出系列へと系列を拡張させていく手法である．

定義 4 (射影データベース)． 系列データベース D と系列 α に対し，$D|_\alpha$ を α-射影データベースという．

$$D|_\alpha = \{\langle sid, s \rangle \in D | \alpha \sqsubseteq s\}$$

図 1 に PrefixSpan の擬似コードを示す．

Input: 系列データベース D,
　　サポート閾値 min_supp
Output: すべての頻出系列の集合

call *PrefixSpan*$(\langle\rangle, D)$

procedure *PrefixSpan*$(\alpha, D|_\alpha)$
　// F は頻出するアイテムの集合
1. $F \leftarrow \{i \lhd s | (s \in D|_\alpha) \wedge (support_{D|_\alpha}(\alpha i) \geq min\_supp)\}$
2. **foreach** $i \in F$
3. 　$\alpha i, support_{D|_\alpha}(\alpha i)$ を出力
4. 　$D|_{\alpha i} \leftarrow \{\langle sid, s \rangle \in D | \alpha i \sqsubseteq s\}$
5. 　**call** *PrefixSpan*$(\alpha i, D|_{\alpha i})$

図 1．PrefixSpan の擬似コード

4 提案手法

4.1 Weighted-PrefixSpan

PrefixSpan を拡張した重みを利用できる手法を提案する．系列のサポートに重みを乗じた値がサポート閾値以上かどうか調べるように拡張する．

図 2 に Weighted-PrefixSpan の擬似コードを示す．

Input: 系列データベース D,
　　サポート閾値 min_supp，重み関数 w
Output: すべての重み付き頻出系列の集合

call *Weighted-PrefixSpan*$(\langle\rangle, D)$

procedure *Weighted-PrefixSpan*$(\alpha, D|_\alpha)$
1. $F \leftarrow \{i \lhd s | (s \in D|_\alpha) \wedge (support_{D|_\alpha}(\alpha i) \times w(\alpha i) \geq min\_supp)\}$
2. **foreach** $i \in F$
3. 　$\alpha i, support_{D|_\alpha}(\alpha i) \times w(\alpha i)$ を出力
4. 　$D|_{\alpha i} \leftarrow \{\langle sid, s \rangle \in D | \alpha i \sqsubseteq s\}$
5. 　**call** *Weighted-PrefixSpan*$(\alpha i, D|_{\alpha i})$

図 2．Weighted-PrefixSpan の擬似コード

手法の正当性を示すために以下の補題を示した．PrefixSpan の正しさは，主に再帰的な探索でもれなく系列が調べられていることにかかっている．提案手法においては，このことは次のように述べられる．

補題 1 (重み付きアプリオリ原理)． 系列 α，β に対して，$\alpha \sqsubseteq \beta$，系列の重み $w(\alpha) = w_{min}(\alpha)$ ならば，$support(\alpha) \times w(\alpha) \geq support(\beta) \times w(\beta)$ である．

この補題により w_{min} を採用した場合には，手法の正当性が示される．

4.2 OrderSpan

Weighted-PrefixSpan は，系列の必ず末尾にアイテムを付けて拡張するため，$\alpha \sqsubseteq \beta$ となる系列 α，β に対し，$w(\alpha) \geq w(\beta)$ が保証されない場合は使えない．この問題を解決するため，与えられるアイテム間の順序に基づいて頻出重み付き系列の拡張をおこなう OrderSpan を提案する．この拡張を順序拡張といい，アイテムを任意の位置に付けることを許している．

定義 5 (アイテム集合上の関係)． アイテム集合 I に対して，"$\preceq$" を I 上のある全順序関係とする．

定義 6 (順序拡張)． 系列 β が系列 α の**順序拡張**とは，以下の三つの条件を満たすときと定義する．

1. $\alpha \sqsubseteq \beta$
2. $len(\alpha) + 1 = len(\beta)$
3. α にアイテム i を付けて β になるとき，$j \lhd \alpha$ である任意の j に対し，$j \preceq i$ が成り立つ．また，i と同じアイテム j がある場合，i を付ける位置はその j の位置より後である．

expansion(α) とは，系列 α の順序拡張であるすべての系列の集合である．

OrderSpan の擬似コードを図 3 に示す．

Input: 系列データベース D,
　　サポート閾値 min_supp,
　　重み関数 w，全順序関係 $\preceq$

Output: すべての頻出重み付き系列の集合

call $OrderSpan(\langle\rangle, D)$

procedure $OrderSpan(\alpha, D|_\alpha)$

1. **foreach** $\beta \in expansion(\alpha)$
2. 　$D|_\alpha$ 中の β のサポートをカウント
3. 　**if** $support_{D|_\alpha}(\beta) \times w(\beta) \geq min\_supp$
4. 　　$\beta, support_{D|_\alpha}(\beta) \times w(\beta)$ を出力
5. 　　$D|_\beta \leftarrow \{\langle sid, s\rangle \in D | \beta \sqsubseteq s\}$
6. 　　**call** $OrderSpan(\beta, D|_\beta)$

図 3. OrderSpan の擬似コード

手法の正当性を示すために，次の補題を示した．

定義 7 (重みの単調性)． 系列 α，β に対して，$\beta \in expansion(\alpha)$ ならば，$w(\alpha) \geq w(\beta)$ であるとき，w は $\preceq$ に関し単調であるという．

補題 2 (重み付きアプリオリ原理)． w が $\preceq$ に関して単調であるとき，系列 α，β に対して，$\beta \in expansion(\alpha)$ ならば，$support(\alpha) \times w(\alpha) \geq support(\beta) \times w(\beta)$ である．

系列の重み w_{min} に対しては任意の順序が単調である．w_{max} についてもアイテム i，j に対し，$i \preceq j \Leftrightarrow w(i) \geq w(j)$ を定めれば単調となる．これらの $\preceq$ を使った OrderSpan の正当性を示した．

5 実験

英語のテキストデータを用いて，提案した二つの手法から同一の頻出重み付き系列の集合が得られることを確認した．また，OrderSpan の特徴を考察するため，データベースへのアクセス総数の変化に関して実験をおこなった．図 4 は英語のテキストデータで w_{min} を用いたとき，$\preceq$ をアイテムの重みの降順，昇順 (図中の Weight-Descending，Weight-Ascending)，データベース中に存在するアイテムの出現数の降順，昇順 (図中の Item-Descending，Item-Ascending) の四通りでアクセス総数を示している．最も影響を与える $\preceq$ に関しての要因は，データベース中に存在するアイテムの出現数である．

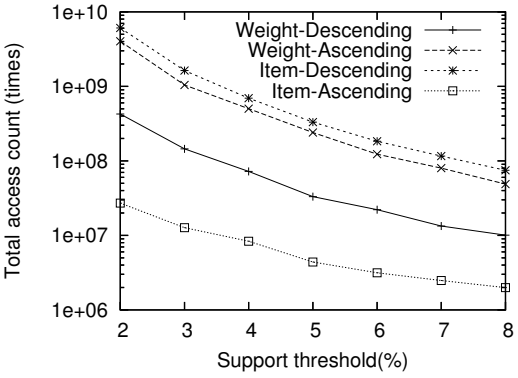

図 4. OrderSpan のアイテム間の順序に対するアクセス総数の変化

6 結論

頻出重み付き系列マイニング問題を定義し，これを解く手法を二つ提案し，その正当性を示した．

今後の課題としては，OrderSpan の特徴である $\preceq$ に対し効率が改善される一般的な順序があるかどうかの検討や，重みの付いている系列データに対し有用な知識を抽出可能かどうかの検証が考えられる．

参考文献

[1] R. Agrawal and R. Srikant. Mining Sequential Patterns. In *Proc. 11th Int'l Conf. Data Engineering, ICDE*, pp. 3–14, 1995.

[2] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu. Mining Sequential Patterns by Pattern-Growth: The Prefixspan Approach. *IEEE Trans. on Knowledge and Data Engineering*, Vol. 16, No. 11, pp. 1424–1440, 2004.