F-054

時間間隔を考慮した営業日報からの系列パターン抽出

Extracting Sequential Patterns from Daily Sales Reports with Time Intervals

植野 研† 櫻井 茂明† 折原 良平†
Ken Ueno Shigeaki Sakurai Ryohei Orihara

1．背景

近年，データマイニング技術の普及により，複雑な構造をもった知識の抽出が求められている[3]．特に，時系列テキストデータからの知識発見は，医療データ，営業日報データ[2,4,7]など，多くの領域で盛んに研究されている[9]．この中でも営業日報データの解析は，営業マネージャが営業担当者の行動を把握する上での正確性，迅速性を向上させることができると予測される．しかしながら，日報から行動パターンを解析している研究はほとんど見られない[5,6]．日報・日誌の量が多くなれば，解釈精度の低下や分析時間コストの増大が予測される．時間間隔を考慮したパターン抽出[2]が営業活動支援システム(SFA)として確立すれば，よりきめ細かな営業支援が可能であると考えられる．ここでは，日報から，時間間隔付き頻出行動パターンを抽出した結果を報告する．

2．時系列日報分析システム

2.1 報告文からの時系列データ生成方法

近年，SFA システムは，自由記述部分をなるべく少なくし，報告すべき部分はリストから選択する傾向が見られる[8]．しかしながら，想定されたリストから選択するだけでは細かな顧客情報を得ることが難しく，自由記述文で情報を補う必要があると考えられる．

ところが，自由記述で書かれた日報は，多様な記述表現などの理由により解析が困難になりがちである．日報データは，日付，顧客名，担当者名，所属，案件名，報告文の６項目から成っている．報告文は自由記述である．我々はこの問題をテキストデータを形態素解析した後，キー概念辞書を用いることにより解決した（**図１**）．キー概念辞書を用いた情報抽出により表層表現のゆれを吸収し，異表記でも同一概念な言葉を概念クラス *cclass* とキー概念 *kc* からなる概念タプル $ct=(cclass,kc)$ に置き換えることができる．この方法により日報中の「手ごたえを感じた」「反応がよい」などの表層表現は(評判, 良い)という概念タプルに吸収される．

各日報データの報告文を概念タプルに変換した後，顧客名と案件名をキーにして日付でソートし，時系列データを生成する．この処理により，顧客名と案件名の異なり分だけ，概念タプル集合のイベント列から成る系列データが生成される．最後に，各概念タプル集合の日付 *dt* から集合間の日数差を離散化し，日数間隔イベント *dd* として集合間に挿入した．頻出系列パターンの各タプルには，元の日報を辿れるように日報テキスト番号を付与しておく．このテキスト番号でタプルと文書を紐付けておけば，営業マネージャが興味深いパターンを見つけたときに，即座に該当する日報の詳細を読むことができる．

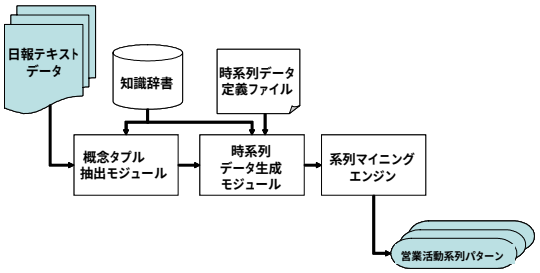

図１：時系列日報分析システムの構成

2.2 営業担当者の行動パターン抽出方法

営業マネージャは，営業担当者がよく実行する活動の系列を注意深く観察することで，アドバイスを与え，活動の成功事例，失敗事例を知識として蓄積していると考えられる． この分析作業を効率化するため，時系列データ生成から得られた概念タプル集合系列から，頻出する系列パターンを抽出する．本システムでは，系列パターンマイニングの基本的なアルゴリズムである *AprioriAll*[1]を拡張し，実装した．アルゴリズムの概要を述べる．

2.2.1 日数間隔を考慮した部分系列の定義

まず，活動イベントの集合を $E=\{e_1,\ldots,e_Z\}$ とする．ただし，活動イベント *e* は概念タプル *ct* と *dt* のタプルであり， $e=(ct,dt)$ である．ここで，系列 s_1 における概念タプル集合を $A_i \in E$ とし，離散化された日数間隔イベントを ad_j とすると，系列 $\alpha =< A_1 ad_1 \cdots ad_{n-1} A_n >$ と定義できる．ここで他の系列を $s_2 =< B_1 bd_1 \cdots bd_{m-1} B_m >$ としたとき，$A_1 \subseteq B_{k_1}, A_2 \subseteq B_{k_2}, \ldots, A_n \subseteq B_{k_n}$ を満たす整数 *k* が存在する場合，ある系列 s_1 は s_2 の部分系列であるといい，$s_1 \sqsubseteq s_2$ と表す．ただし，$1 \le k_1 < \ldots < k_n \le m$ である．

2.2.2 制約によるパターン数の削減

系列パターン *p* のサポート $Sup(p)$ を以下に定義する．すべての系列の集合を $S=\{s_1,s_2,\ldots,s_N\}$ とすると，サポート $Sup(p)$ は $Sup(p)=\left|\{p \mid p \sqsubseteq s, s \in S\}\right| / N$ と定義できる．つぎに，頻出する系列パターン *fp* を，すべての系列集合 *S* のうち生成されたパターンが予めユーザにより与えられた最小サポート ξ 以上の全ての系列パターンとする．すなわち，頻出する系列パターン集合 *FP* は，$FP=\{fp \mid Sup(fp) \ge \xi\}$ と定義される．ここである頻出系列パターンを fp_1，他の頻出系列パターンを fp_2 としたとき， $fp_1 \sqsubseteq fp_2$ であるような fp_1 は fp_2 に吸収する[1].

†(株)東芝 研究開発センター 知識メディアラボラトリー
{ken.ueno, shigeaki.sakurai, ryohei.orihara}@toshiba.co.jp

パターンの重複を吸収するほかに，以下の３つ制約を導入することでパターンを抽出する．（１）最大系列長制約 *MAXL* （２）最大要素組合せ数 *MAXC* （３）パターン制約集合 $PCTR$ を導入する．*MAXL* は候補系列生成において $m \le MAXL$ とするものである．ただしここでの系列長とは，GSP[1]などでの系列長とは異なり，ひとつの集合を１系列長と数える．系列候補生成時に生成される各概念タプル集合から完全な部分集合を生成する場合，概念タプル集合に含まれるタプルの種類 *kt* から *u* 個の要素を含む ${}_{kt}\mathrm{C}_{u}$ 個の部分集合を生成しなければならないため，莫大な計算コストがかかる．そこで，分析のスピードが要求される営業日報分析に応用するため，完全な部分集合を全て生成して検査する代わりに最大組合せ数 *MAXC* をパラメータとして設定し，${}_{kt}\mathrm{C}_{MAXC}$ 個のみを候補系列の部分集合とすることとした．ただし，*MAXC* を導入しても概念タプルは全種類考慮されることになる．また，系列パターンに対する制約 *pcr* を $PCTR$ として定義した．さらに，頻出する系列パターン集合 *FP* に対して制約を加えるように $FP = \{ fp \mid fp \sqsubseteq pcr, pcr \in PCTR, Sup(fp) \ge \xi \}$ とした．

3．実験方法

社内の営業部の５部署で約半年間に渡り収集された営業日報テキストデータから，営業担当者の行動パターンを抽出した．$\xi = 0.02$ とし，パターン抽出の制約は，最大系列長 $MAXL = 5$，最大要素組合せ数 $MAXC = 4$ とした．日付は $1 \le d_1 < 7$， $7 \le d_2 < 14$， $14 \le d_3 < 21$， $21 \le d_4 < 28$， $28 \le d_5 < 35$， $35 \le d_6 < 42$ に離散化した．パターン制約 *PCONSTR* を{<(一般，要望) *INTVL* (不評, 怒り・不満)>, <(一般, 要望) *INTVL* (不評, 価格)>, <(一般, 要望) *INTVL* (不評, 困難)>, <(一般, 要望) *INTVL* (不評,クレーム)>, }>, <(一般, 要望) *INTVL* (不評,宿題)>}のように定義し，頻出系列パターンとして行動パターン集合を求めた．ただし $INTVL \in \{d_1, d_2, \ldots, d_6\}$ である．

3.1 時間間隔導入効果の評価方法

定量評価として，パターン削減率 *PR* を計算する．時間間隔なしで生成されたパターン数を α，時間間隔ありのパターン数を β とし，各部署毎にパターン削減率 $PR = 1 - (\alpha / \beta)$ を求め，これらの平均 $\overline{PR}$ を計算する．また，日報原文より，パターンが正しいかを確認する．

4．結果と評価

５部署中４部署で生成パターン数が削減された．この４部署のパターン削減率は平均 40%に達した．のこり１部署に関してはパターン数が増加したが増加率は約 14%にとどまった．これは最小支持度を超える日数間隔の種類が増えるためだと考えられる．これらの結果より，時間間隔の導入により日数間隔が明らかになるだけでなく，パターンを削減する効果があることが分かった．また，これらのパターンが元の日報文書の内容とほぼ一致していることが分かった．たとえば，パターン１では「価格設定で困難にぶつかったが何とか対処し受注にこぎつけた」,「機器導入のデメリットの改善を検討する必要が生じたが，検討の結果最終的に契約できた」ことを，パターン２では「製品の最終テストで問題が発生した場合に対応が厳しいとの懸念があったが，最終的に問題なく開発を完了した」「システム導入を全社展開するスケジュールが短く困難だったがスケジュールを再考し問題なく作業を完了できた」ことを示していることが理解できた．

●パターン１（ $Sup(fp) \cong 2.241\%$ ）
{(一般, 意見), (一般, 要望)}－[not less than 1 week]→
{(一般, 要望), (不評, 困難)} →{(好評, 内定・受注)}

●パターン２（ $Sup(fp) \cong 2.521\%$ ）
{(一般, 要望)}－[not less than 1 week]→
{(不評, 困難)}→{(好評, 関心が高い), (製品, 保守)}

5．まとめ

時間間隔を考慮した営業活動の系列パターンを抽出できた．また大幅なパターン削減率を実現でき，パターンがほぼ正しいことを確認した．これらのパターンは営業活動の危機管理や機会損失を防ぐために有用であると考えられる．時間間隔を導入することで，商談の進み具合に応じて迅速な対応が必要な行動や後になってクレームが付きがちな製品カテゴリなども容易に特定できる．得られたパターンは，過去にどのような過程で営業活動の壁に直面しどのように乗り越えたのかといった現場の営業知識を一人一人の営業活動に直接的に生かすことが可能である．パターン削減率からみて，人手でパターンを選別するのは大変困難であり，本システムが実際の営業日報から知識を発見するのに有用であると示唆された．

参考文献

[1] R. Agrawal et.al.: “Mining Sequential Patterns”. pp.3-14, Proc. 11th Int. Conf. Data Engineering, ICDE, 1995.

[2] C.Bettini et.al.: “Mining Temporal Relationships with Multiple Granularities in Time Sequences”. IEEE Data Engineering Bulletin Vol.21 No.1 pp.32-38 ,1998 .

[3] J.F.Roddick et.al. : “A Survey of Temporal Knowledge Dicovery Paradigms and Methods”. IEEE Tran. on Knowledge and Data Engineering, Vol.14, No.4, , pp.750-767, 2002 .

[4] 市村 由美ほか．日報分析システムと分析用知識記述支援ツールの開発電子情報通信学会論文誌 Vol.J86-D-II, No.2, pp.310-323, 2003.

[5] 植野 研ほか．時系列共起アルゴリズムによる営業知識獲得 信学技報 AI2003-73 (2004-1) pp.61-66, 2004.

[6] 櫻井 茂明ほか．テキストデータからの時系列パターンの発見 第 21 回ファジィワークショップ,pp.4-2, 2003.

[7] 柴田 親男ほか．企業における非定型文書の活用促進事例 情報処理 Vol.44, No.10, pp.1022-1027, 2003.

[8] ソフトブレーン株式会社 e セールスマネージャー ホームページ http://www.e-sales.jp/

[9] 高野 洋ほか．大規模な時系列テキストデータからのイベント時系列パターンの発見 人工知能学会研究会資料 SIG-KBS-A304, pp.233-238, 2004.