2P-6

病院ブログ記事からの行動分析の試み

竹内 希史 † 徳久 雅人 † 村田 真樹 † 村上 仁一 †
鳥取大学 工学部 知能情報工学科 †

1 はじめに

病院内での行動支援を行うためには，病院内での人々の行動体系を知る必要がある．行動体系を知るために，病院内での行動を記述した文書集合の取得，および，その分析が一つの方法として存在する．本稿では，分析者がブログ記事をソースとして行動分析を行うための支援ツールの作成を試みる．

先行研究では，ブルーベリー狩りにおける行動分析に，ブルーベリー狩りのブログ記事の部分抽出と，そのクラスタリングを行った [1]．しかし病院ブログに適用するとノイズが多すぎる問題が生じた．理由は，ブルーベリー狩りの記事は旅行記として時系列に沿って作文されるが，病院記事はそうした傾向が見られないためである．したがって，クラスタリングを行う前に，分析に不要な文を削除する必要性が高い．ここで，最終的にツールにより選出されたブログの文を分析者が読むことを想定するならば，クラスタリングの前にわずかな量の文を読み，不要な文にアノテーションしておくことが許される．

そこで，本稿では，アノテーションした文を元に，多量の文をフィルタにかけて，不要と推定される文を除去する．そして必要と推定された文のクラスタリングにより，行動の分析を支援することを提案する．こうして得られたクラスタごとに，分析者が行動のタイプを分析することで，行動の体系化が容易になることを目指す．

2 支援ツールの処理の流れ

以下の順で処理を進める．

処理 1．病院文集合の抽出: 2008 年 7 月 31 日から 2013 年 4 月 31 日までのブログ記事から「病院で」を含む文を抽出する．以降は病院文集合と呼ぶ．

処理 2．特徴語の取得: まず，一般ブログ記事から無作為に 5,000 文を抽出し，100 文を 1 文書にして 50 個の文書を用意する．次に「病院」を含む文をブログ記事から無作為に 100 文抽出して 1 つの文書とする．病院の文書における各語の $tf \cdot idf$ 値を算出し，スコアの高い語を特徴語とする．例えば，「入院」,「注射」,「治す」，などが得られる．

処理 3．不要文の除去: 分析者に病院文集合の一部を表示し，不要文にアノテーションをしてもらい，学習データとする．これを学習した SVM により，残りの病院文集合から不要文を除去する．用いた素性を表 1 に示す．

処理 4．クラスタリング: 処理 3 の結果および学習データの各文を Repeated Bisection 法でクラスタリングする．ツールは bayon を用いる（`-l 1.0` のオプションを指定）．クラスタリングにおける素性は特徴語の unigram を用いる．

処理 5．クラスタの選択: 学習データで不要文としたものが含まれるクラスタを，破棄する．以上の結果，得られるクラスタを推定クラスタと呼び，E_j と記す（$j \in [1, M]$ はクラスタ番号，M はクラスタ数）．

表 1 不要文除去の SVM に用いる素性

種別	説明
f1	文中に特徴語が出現したか否か．
f2	文中の自立語の unigram．
f3	文中の格要素と述語のペアを日本語語彙大系の意味属性コードで表したもの．

3 クラスタリングの評価

3.1 正解クラスタの作成

クラスタリングの性能を評価するために，正解クラスタを人手で作成する．正解クラスタは，病院文集合に記述された病院内での人々の行動に基づき作成する．行動の種類で分けられたクラスタを正解クラスタと呼び，A_i と記す（$i \in [1, N]$ はクラスタ番号，N は正解クラスタ数）．行動を表さない文は不要文であるので，1 つのクラスタにまとめる．このクラスタを形式上，正解不要クラスタと呼び，A_0 と記す．

3.2 評価式

必要なクラスタだけを表示すること，かつ，得られるべき行動を逃さないことが，支援ツールへの評価基準となる．

A prototype system of behavior analysys from blog entries of hospital

†Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University

3.2.1 正解文密度とその行動

推定クラスタにおける正解文の密度を D_{ij} とする（式 (1)）.

$$D_{ij} = \frac{|A_i \cap E_j|}{|E_j|} \tag{1}$$

特定の正解クラスタだけで D_{ij} が高くなる E_j があるならば，分析者は，その E_j の文を読むことで，その正解クラスタ A_i に対応する行動を思い付きやすくなる．そこで推定クラスタ E_j から得られると推定される行動を $\widehat{i_j}$ とする（式 (2)）.

$$\widehat{i_j} = \arg \max_{i \in [0,N]} D_{ij} \tag{2}$$

3.3 信頼性

並べられた推定クラスタを分析者が信頼する割合 B を式 (3) で評価する.

$$B = \frac{1}{M}\left|\{j \in [1,M] \mid \widehat{i_j} \neq 0\}\right| \tag{3}$$

3.4 網羅性

正解クラスタを分析者が思いつく割合 C は，高い密度で推定クラスタが対応した正解クラスタの割合であり，式 (4) で評価する.

$$C = \frac{1}{N}\left|\{\widehat{i_j} \mid j \in [1,M], \widehat{i_j} \neq 0\}\right| \tag{4}$$

4 実験

4.1 条件

2008 年 7 月 31 日から 2013 年 4 月 31 日までのブログ記事から，病院文集合を作成した．テストデータは，病院文集合から無作為抽出した 500 文とする.

分析者は，1 名とする．分析者は，不要文の正例と負例が 10 文ずつになるまで，テストデータから文の抽出とアノテーションを行なった．学習データは 20 文である.

一方，評価実験用にテストデータから正解クラスタを手作業で作成した．正解クラスタ数 $N = 20$ となった.

4.2 結果

不要文の除去について，480 文のうち 376 文が必要な文と判定された．クラスタリングは，学習データを加えた 396 文が対象であり，この時点では，推定クラスタ数 $M' = 46$ となった.

クラスタ選択の結果，8 件のクラスタを破棄し，最終的には，推定クラスタ数 $M = 38$ となった.

結果を表 2 に示す．表には，比較のため，クラスタ選択を行なわない場合 (case1)，ならびに，不要文除去およびクラスタ選択を行なわない場合 (case2) も示す．なお，F 値=$2BC/(B+C)$ である.

表 2 実験結果

手法	信頼性 B	網羅性 C	F 値	M	文数
提案	0.39	0.75	0.51	38	386
case1	0.37	0.85	0.51	46	396
case2	0.25	0.65	0.36	53	500

5 考察

正解クラスタに割り当てた病院内行動を表 3 に示す．行動名称および系列分けは手作業で定めた．○印は，推定クラスタから得られると推定されることを表す．系列 4 は病院に特化した行動ではないが，病院内の人々の行動支援を想定すると発見しておきたかった行動である.

表 3 正解クラスタに対応する行動

系列 1	系列 2	系列 3
待機する	○入院する	○薬を貰う
熱を測る	○手術する	○書類を貰う
○診察を受ける	○リハビリ	
○検査を受ける		
○身体測定をする		
○治療する		
○注射する		
○点滴を受ける		
○カウンセリング		

系列 4	系列 5	系列 6
買い物をする	○実習を受ける	○出産
見つける	○働く	
人に会う		

6 おわりに

病院内での行動をブログ記事から分析するための支援ツールを作成した．病院ブログの文は，行動についてクラスタリングを行うと，そのままでは不要なクラスタが多く生じるという問題があったが，本ツールでのわずかな学習データによる SVM でのフィルタリングは，不要なクラスタの削減に効果があることが確認された.

参考文献

[1] Tokuhisa, M., Yamamoto, T., Fukui, T., Murata, M., Murakami, J.: Extracting and Clustering Blog Texts to Investigate Experiences of Tourists, ICAL 2013, Vol.1, pp.268–273, 2013.