何日目日記
時間経過を揃えたソーシャル日記システムと時間特徴マイニング

粟飯原俊介 †　　中谷洸樹 ‡　　田中久美子 †
† 九州大学大学院システム情報科学研究院　　‡ 九州大学工学部
{aihara, nakatani, kumiko}@cl.ait.kyushu-u.ac.jp

1　はじめに

個人ブログを発信するユーザが増え、Twitter や Facebook の社会的影響力も増す中、そこからのマイニングの研究が盛んに行われるようになった。その一環として、特に時間の観点からの知識獲得が注目を浴びている。これを受け、機械学習や言語処理分野では、関連する機械処理のセッションか特別に設けられるなどしている。本稿では、時間の中でも、相対的な「経過時間」に着目した知識マイニングのためのソーシャルシステムと時間特徴マイニングシステム「何日目日記」の提案を行う。

かねてより時間の観点からのマイニングは、ニュース記事などの文書の作成時間 (タイムスタンプ) が付与され、時系列順に整列されているコーパスを対象に研究が行われてきた。ある期間の重要な話題 (バースト) 推定の研究から始まり [1, 2]、話題の移り変わりや追跡を行う研究が、Topic Detection and Tracking(TDT) プロジェクト [3] の中で数多く行われている。

近年ではより細かい要素技術として、タイムスタンプが付与されていない文書から、データが構築された時間の推定 [4] や、複数データの時間的関連 (前後、同時など) の判定 [5] などが研究されている。しかし、一番易しいはずの「タイムスタンプ推定」問題であっても、文章の内容に関する時間的推論を要するため、実用にはほど遠いのが現状であり、ひいては繰り返し行われる事態の中における相対的な時間関係に基づいたマイニングはさらに難しい問題となっている。フォーマットの揃っていない個別のウェブログの集合から、「レーシックの手術を受けてから 5 日経過したが、痛みが継続している。痛みはどれくらいの期間継続するのだろうか?」などといった、単純にバーストだけではなく、期間と増加減少等の傾向を考慮しなければならない問に対して、言語的な情報だけでなく、集合知的な統計的知見を元に解答することは非常に難しいと言えるだろう。

マイニングの性能の限界は、時間に限らず機械学習手法の限界としてどの知識抽出においても見られる。この問題に対し、機械学習手法の性能向上とは異なる新しい解決策としてソーシャルシステムを活用するアプローチかが近年盛んであり、数多くのユーザにメタデータを記述し、情報検索等の分野で応用が進められている。同様に、時間の観点からの知識獲得の限界に対しても、集合知を再構築しやすいような「知識マイニング向け」ソーシャルシステムを導入することが考えられる。

本稿で提案するシステム「何日目日記」は、経過時間について予めユーザにそろえて記述してもらう事で前述の「タイムスタンプ」推定問題 や時間的関連判定が不要な場を構築し、その中で集合知として、ある事態が発生し継続する期間や傾向などの時間特徴を抽出し可視化することで、結果を社会に還元することを目的とするものである。

2　何日目日記

ここでは、提案システム「何日目日記」の詳細について述べる。「何日目日記」システムは大きく分けて以下のようなモジュールから構成されている。

図 1: 何日目日記のシステム概要

ソーシャル日記システム　ユーザの管理や日記の記入・閲覧を行うための Web システム

データベース　記入された日記を経過時間を揃えて保存するデータベース

図 2: 何日目日記のスクリーンショット

時間特徴マイニング 経過時間を揃えたデータベースから時間特徴を抽出するモジュール

時間特徴提示 UI 集合知的な時間特徴を可視化して提示するモジュール

おけるソーシャル日記システムは、あるカテゴリ (例: 医療) の中のテーマ (例:レーシック) ごとに複数のユーザ (例: 患者) が経過としての日記を記載するソーシャルシステムである。対象となるテーマとしては以下の様なものを想定としている。

- 医療分野 (レーシック、歯列矯正、投薬・闘病など)
- 健康管理分野 (禁煙、ダイエット、筋トレなど)
- 趣味分野 (熱帯魚飼育日記、植物栽培日記、学習や練習の記録など)

経過時間を揃えた日記に関するデータベースは以下のような構造を持っている。

ユーザはカテゴリとテーマ、基準時間を設定し、日記を作成する。日記を作成後、その下に一回分のエントリーを登録する。経過時間は、基本的には投稿時間と開始日との差分を元に計算されるが、エントリー作成時に任意に設定することも可能である。エントリーには画像やタグ、地理情報等のメタデータも付与が可能となっている。

登録された日記は、公開されているものに関して、ユーザごと、カテゴリごと、テーマごと、時間粒度ごとに閲覧が可能であり、§3 で詳細を述べる「時間特徴マイニング」の結果である時間特徴提示 UI と合わせて提示する (図 2)。

3 時間特徴マイニング

ここでは、提案する時間特徴マイニングについて述べる。これまでの時間に着目したテキスト処理の分野では、主にいつ何が起きたか、という話題のバーストにフォーカスが向けられていた。それは対象とするテキストが多くはニュース記事のような数多くの話題が常に移り変わっていくものが対象であったからといえるだろう。翻って、「何日目日記」のような一つのテーマに関して複数の人が経過時間を揃えて記述していく場合に、バーストのような「いつ何が起こるか」だけに着目するのは不十分であり、事態がどのように変化をしていくかまでを捉える必要がある。事態の変化を捉えるにあたって重要となるのが、時間的変化の傾向とその傾向が持つ意味 (良し悪し) とその継続期間である。言及の増減が表す意味を考えるには、語そのものの評価極性 (ポジネガ) を踏まえて考えなければならない。当然、それがいつまで継続するのかも重要であろう。

そこで、「何日目日記」システム上において、各テーマごとの特徴語、傾向、評価極性、期間の 4 つ組を時間特徴として定義し、自然言語処理手法と時系列分析の手法を用いた抽出手法に関して詳細を述べる。

3.1 特徴語抽出

テーマごとの特徴語の抽出には、χ^2 値を用いた。特徴語の候補となる単語は、mecab を用いて形態素解析を行い数詞以外の自立語のみを取り出しその基本形を用いる。テーマごとに単語の χ^2 値を計算する為、比較用のデータとして毎日新聞の 2009 年版のデータ集を用いた。テーマごとに単語が出現するエントリー数と毎日新聞データ内の単語が出現する記事数を計算し、χ^2 値の上位語をテーマごとの特徴語とする。

χ^2 値により得られたテーマごとの特徴語に対して、評価極性を付与する。評価極性の付与には、乾・岡崎研究室で公開中の日本語評価極性辞書 [6][1]を用いた。評価極性辞書で評価極性が判定出来ないものは特徴語から取り除いている。

3.2 評価極性を考慮した時系列データの作成

特徴語に対する言及の時間変動を見るために経過日数ごとに言及数を集計し、時系列データを作成する。「何日目日記」のようなシステムに集積されていくデータは、基準時間から離れれば離れるほどエントリーがまばらになるような不均衡なデータであると考えられる為、固定の時間間隔ごとに集計を行わずスライディングウィンドウ方式を用いて集計し時系列データを作成する。

ウインドウごとのスコアは、傾向を正しく捉える為に、特徴語の単純な出現頻度を用いるのではなく、特徴語の評価極性と言及の肯定・否定の極性を考慮して計算する。

エントリーの中に対象とする特徴語が含まれている場

[1] http://www.cl.ecei.tohoku.ac.jp/index.php?公開資源/日本語評価極性辞書

合に、エントリーに対して表 1 の基準で評価ラベルを割り当てる。言及がない場合の評価ラベルは 0 となる。ウィンドウごとのスコアは、ウィンドウ内のエントリーの評価ラベルの合計値をウィンドウサイズで割ったものを用いる。

言及の極性は、本稿では周辺共起語と係り受け共起語を用いて、シンプルなルールーベースの判定を行なっている。エントリー内に特徴語に対する複数の言及がある場合は、多数決で評価ラベルを決定する。エントリーレベルで評価ラベルを付与する理由は、エントリー長のばらつきや、繰り返し等のバイアスを除去するためである。

表 1: 評価ラベルの付与

言及の極性＼評価極性	ポジティブ	ネガティブ
肯定	1	-1
否定	-1	1

以下に特徴語ごとの時系列データ作成の手順を示す。

1. テーマ内のエントリー e を経過時間順にソートした系列 $E = [e_0, e_1, ..., e_{n\ \ 1}]$ から、ウィンドウ幅 l、移動量 p のスライディングウィンドウ方式で k 個のウィンドウ $w_i = [e_{i\ p}, e_{i\ p+1}, ..., e_{i\ p+l}]$ の系列 $W = [w_0, w_1, ..., w_{k\ \ 1}]$ を得る。
2. w_j 内のエントリーのそれぞれに対して特徴語の評価極性と言及の極性を考慮して評価ラベルを付与し、その合計値をウィンドウサイズで割ったスコア s_j を計算する
3. すべてのウィンドウに対してスコアを計算し時系列データ $X = [s_0, s_1, ..., s_{k\ \ 1}]$ を作成する。

それぞれのウィンドウの経過時間はウィンドウ内に含まれるエントリーの経過時間の平均値 t_j を付与し、X と対応する経過時間データ $T = [t_0, t_1, ..., t_{k\ \ 1}]$ を作成する。

3.3 期間と傾向の特定

特定の出現傾向を持つ期間の特定の為に特徴語の言及に関する時系列データの分割を行う。時系列データの分割に関する手法は古くから研究されているが、本稿では増減などの一次的傾向を持った部分系列に分割を行うため、セグメント回帰 [7] を用いる。セグメント回帰を適用する際の時系列データの分割位置推定には、回帰分析に基づいた再帰分割法 [8] を採用する。[8] では分割位置推定のための評価基準として線形回帰モデルの残差二乗和を用いているが、本稿では線形回帰モデルの AIC 値 [9] を評価基準として用いることで、分割位置推定だけではなく、分割数の推定も自動化する。

以下に回帰分析に基づいた再帰分割法の手順を示す。

1. 分割を行う前処理として、X に対して 3 次の移動平均フィルタを用いて平滑化を行う。
2. T を用いて X を回帰するモデルを最小二乗法で推定し、その場合の AIC 値 AIC_{all} を計算する。
3. X と T を任意の位置 $j(0 < j < |X| - 1)$ で二分分割し、分割位置の左右の領域にそれぞれ対して回帰モデルを推定し、AIC 値を計算する。
4. 左右の領域の AIC 値の和が最小となる位置 bp を分割点候補とし、その際の左右の領域の AIC 値の和を AIC_{seg} とする。
5. $AIC_{all} > AIC_{seg}$ となった場合、位置 bp で X と T を分割し、分割位置の左右の領域それぞれに対して再帰的に手順 2 以下を適用する。$AIC_{all} <= AIC_{seg}$ の場合は分割を行わず、停止する。

再帰分割法によって得られた部分系列に対して、回帰係数に対する t 検定を実行し、$p < 0.05$ で有意であった部分系列を増加・減少などの一次的傾向を持った期間であるとする。

4 実験

4.1 実験データ

「何日目日記」システムは現在構築中であり、一般に公開されているものではないため、データが無いのが現状である。そこで、時間特徴マイニングの実験と評価の為、「何日目日記」の内部構造に合わせたデータを web から人手で構築した。レーシックの手術後の経過について記述されている 101 個の日記を収集し、1271 個のエントリーを集めた。この実験データは、経過時間の粒度は一日単位となっている。また、時系列データの作成は、移動量を 17、ウィンドウ幅を 200 として行った。

4.2 実験結果

実験データから抽出された評価極性を持つ特徴語の例を表 2 に示す。χ^2 値の上位で評価極性を持つ語はユーザが関心を持つとおもわれる語が得られているといえる。

また、評価の為に、特徴語「ドライアイ」と「痛み」で作成した時系列データに対して、期間と傾向の分析を行った結果を図 3 に示す。灰色の区間が傾向が無いと判定された期間、赤色の区間が、ネガティブな言及が増えている、つまり悪化している区間、青色の区間がネガティブな言及の減少、つまり収束を表していると言える。特徴語「痛み」においては、術後直後に最もネガティブな言及が多いが、約 7 日目にはネガティブな言及がなくなっているという事が分かる。論文等の出展は見つけられなかった

表 2: 抽出された特徴語例

単語	χ 二乗値	極性
視力	46476.11	ポジティブ
ドライアイ	15622.73	ネガティブ
近視	15057.49	ネガティブ
乱視	10671.28	ネガティブ
綺麗	5653.63	ポジティブ
痛み	4302.68	ネガティブ

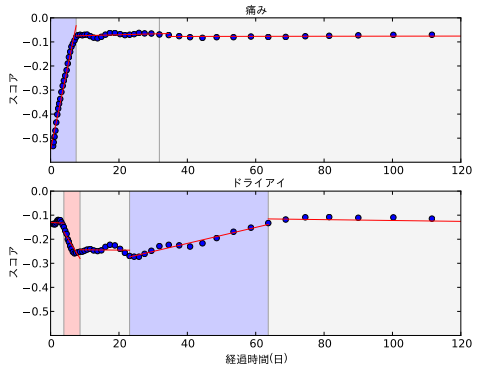

図 3: 傾向と期間の分割例

が、レーシック後に目の痛みを感じるのは大体術後 5 日目程度で、ほとんどの場合は一週間以内に痛みが治まると言われていること[2]と整合がとれていると言える。ドライアイに関しても、最初一週間にドライアイの症状が現れ、概ね術後約 3ヶ月で症状が治まるという報告 [10] とある程度整合性はとれている結果となっている。

5 まとめ

本稿では、経過時間に着目したソーシャル日記システム「何日目日記」と、そこからの時間特徴マイニングに関する提案を行い、その有効性の一端を示した。今後の研究の方針としては、以下のようなものが考えられる。本稿では扱う傾向として一次的傾向ののみを対象としたが、周期性など他の傾向も扱って行く必要があるだろう。またデータ量の問題から、特定のテーマと経過時間にのみ基づいた時間特徴マイニングとなっているが、今後は、記述内容の個人差やカテゴリの情報を用いて、より精密な時間特徴マイニングが可能になると考えられる。

現在、「何日目日記」システムの公開の準備を進めているが、当面は web 上から人手でレーシック以外の複数のテーマについてのテキストを収集して研究を進めて行く必要がある。人手で収集したテキストを元に、web からのテキストの収集の自動化なども今後の課題となる。

[2] http://1paraguay.net/2008/11/post_20.php

参考文献

[1] R. Swan and J. Allan. Automatic generation of overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 49–56, 2000.

[2] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 91–101, 2002.

[3] J. Allan. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic, 2002.

[4] N. Chambers. Labeling documents with timestamps: Learning from their time expressions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 98–106, 2012.

[5] O. Kolomiyets, S. Bethard, and M. Moens. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 88–97, 2012.

[6] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2005.

[7] V. McGee and W. Carleton. Piecewise regression. *Journal of the American Statistical Association*, Vol. 65, No. 331, pp. 1109–1124, 1970.

[8] M. Last, A. Kandel, and H. Bunke. *Data Mining In Time Series Databases*. World Scienti c, 2004.

[9] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pp. 267–281, 1973.

[10] C. De Paiva, Z. Chen, D. Koch, M. Hamill, F. Manuel, S. Hassan, K. Wilhelmus, and S. Pflugfelder. The incidence and risk factors for developing dry eye after myopic lasik. *Am J Ophthalmol*, Vol. 141, No. 3, pp. 438–45, 2006.