人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第1回)
SIG-AM-01-01

価値観に基づく情報推薦におけるレビュー分析の利用可能性に関する検討

服部 俊一[1*]　毛 中杰[1]　高間 康史[1]
Shunichi Hattori[1], Zhongjie Mao[1], Yasufumi Takama[1]

[1] 首都大学東京
[1] Tokyo Metropolitan University

Abstract: レビューサイトに記載されたアイテムの評判情報を収集・分析することで，アイテムやユーザの特性を推論する技術が注目されている．本研究ではユーザの価値観に基づく情報推薦を行うための，レビュー分析手法の利用可能性について検討する．レビューサイトから取得した情報のマイニングによりアイテムの各評価属性に対するユーザの価値判断を推論し，その結果に基づき新たな観点によるユーザモデリングを行うことを目指す.

1　はじめに

本稿ではアイテムに対するユーザの価値観に基づく情報推薦を目的とした，ユーザのこだわりに着目したレビュー分析手法を提案する．利用者にとって有用な情報を見つけ出す情報推薦システムが情報フィルタリングの一手法として注目されているが，多くの既存手法ではユーザの嗜好に近いアイテムや似たような嗜好を持つユーザが好むアイテムを推薦対象として扱っている．そのため，推薦されたアイテムはユーザにとって既知のものであることが多く，満足な推薦結果を得ることができない場合が多い [1]．一方で，マーケティングの分野では個人の嗜好や消費行動を推定する要素として「価値観（Personal Values）」が注目され，広く活用されている．価値観はアイテムの属性から独立した要素であることから，これを用いることで従来手法とは異なる新たな観点からの推薦が可能になると考える．本稿では価値観と繋がりの深い要素としてユーザの「こだわり」を推論することで新たな観点によるユーザのモデリングを行う手法について提案する．また，ショッピングサイトにおける商品レビューから構文解析によるレビュー分析を行い，その結果からのユーザモデリングの可能性および考慮すべき課題について考察する．

*連絡先：首都大学東京大学院
システムデザイン研究科情報通信システム学域
〒 191-0065 東京都日野市旭が丘 6-6
E-mail: shattori@krectmt3.sd.tmu.ac.jp

2　関連研究

2.1　情報推薦手法

既存の情報推薦手法の多くは内容ベースフィルタリングと協調フィルタリングに分類することができる [2]．内容ベースフィルタリングはアイテムの属性とユーザの嗜好を比較して推薦アイテムの推論を行う．協調フィルタリングは多くのユーザの嗜好情報を過去の行動という形で記録し，そのユーザと嗜好の類似した他のユーザの嗜好情報を用いてユーザの嗜好を推測する [3]．協調フィルタリングの利点は，アイテムの属性情報がなくても推薦が行えること，および処理が手軽であることであり，これらの理由から商用サイト含め現在最も広く利用されている手法である．これらの手法を改良し推薦の精度を向上させる研究は広く行われてきたが，精度向上によって推薦されたアイテムがユーザにとって既知のものであったり，似たようなアイテムばかり推薦されてしまったりといった問題も発生している．この問題を解決するために Novelty および Serendipity という概念がユーザの満足度向上のために重要な指標として注目されている [4]．これらの指標に着目した研究はいくつか行われており，Ziegler らは，アイテムの分類情報を用いて推薦リストを拡張し推薦アイテムの多様化を試みている [1]．清水らは，ユーザがアイテムを知っているかどうかを表す「発見性」という指標を用いることで，ユーザの知らない好みのアイテムを推薦する手法を提案している [5]．また，秋山らは，ユーザが Serendipity があると感じる情報をアンケートにより収集し，その情報を用いてユーザモデルを構築する推薦手法を提案している [6]．このように，Novelty や Serendipity といった概念に基づく推薦アイテムの多様

化は現在の情報推薦が満たすべき条件として研究が進められている．

2.2 価値観に基づく嗜好・消費行動の推論

価値観は消費者の嗜好や行動に強く影響を及ぼすと考えられており，マーケティングの分野では古くから利用されている．Rokeach は消費者の嗜好に関わる価値観を 18 の要素に分類した Rokeach Value Survey [7] と呼ばれる調査方法を提案し，多くの調査で利用されている．Vinson らは，保守的な価値観を持つ大学と革新的な価値観を持つ大学，それぞれに所属する学生の間に有意な嗜好の差があることをアンケート調査により明らかにしている [8]．近年でも，Holbrook が消費・購買行動に影響を与える価値観を 8 つに分類する [9] など，消費者の嗜好と価値観は関連の深いテーマとして研究および調査が進められている．

従来の内容ベースフィルタリングでは，例えば映画であればジャンルや出演俳優など，アイテムの属性値に対する好き嫌いを元に推論を行うため，既知アイテムが推薦されたり，同一アイテムが何度も推薦されたりする傾向が強い．このような内容ベースの推薦と比較して，本稿における価値観はユーザ指向の属性であり，アイテムが持つ属性に対しメタ属性的な性質を持つ．そのため，価値観に基づく推論を行うことで従来手法とは異なる観点からの推薦が可能になると期待できる．

2.3 評判情報の分析

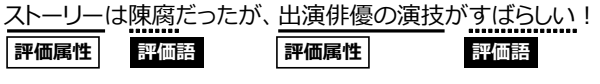
ユーザや商品の特性を分析するため，テキストマイニング技術を用いてレビューサイト等に掲載された評判情報を自動的に抽出・解析する研究も広く行われている．代表的な手法はレビューの評価文から評価属性（アイテムの機能や特徴を表す項目）とそれに対する評価に使われている語を抽出することにより評価属性と評価語を 1 つの組として収集するものである．小林らは，評価対象・評価属性・評価表現の共起パタンから評価属性・評価表現を半自動収集する方法を提案している [10]．また，平山らは係り受け解析を用いて評価属性およびそれに対する評価極性を抽出することで商品に対する評価を表形式に可視化する手法を提案している．これらの手法はユーザのアイテム選択やユーザに対する情報推薦，企業の商品開発などへの応用が期待されている．

3 ユーザのこだわりに着目したレビュー分析手法

本稿では，価値観と繋がりの深い要素としてユーザの「こだわり」に着目したレビュー分析手法，およびその結果に基づくユーザモデリング手法について提案する．本稿では情報推薦における価値観を「どの評価属性を重視してアイテムの評価を決定するか」を判断するための基準と定義し，それが各評価属性に対する「こだわり」の強さとして表れると考える．レビューから評価属性に対する評価とアイテムに対する評価の関係を分析していくことで，どの評価属性がアイテムへの評価に影響を与えたか（ユーザはどの評価属性にこだわりを持っているか）を推論することができると考える．

3.1 レビュー分析によるユーザモデリング

アイテムに対するユーザのレビューには様々な内容が含まれているが，その中からアイテムの評価値（星の数など）とレビュー文を用いて分析を行う．レビュー分析を行うため，レビュー文から評価属性と評価語を抽出する．評価属性とはアイテムの特徴を表し，評価の基準となる項目である．例えば映画であればストーリーや演出，出演俳優などが評価属性となる．また，評価語は評価属性に対する評価を表す語で，評価属性に係る「(ストーリーが) 好き」や「(演出が) つまらない」といった表現となる．提案手法では，レビュー文から係り受け解析を行うことで評価属性および評価語の組を抽出する．抽出したそれぞれの組について評価表現辞書を用いて評価語の極性判定（好評・不評）を行う．レビュー文から評価属性・評価語の組を抽出し，その極性を判定した例を図 1 に示す．



評価属性	評価語	極性
ストーリー	陳腐	− (不評)
(出演俳優の) 演技	すばらしい	+ (好評)

図 1: 評価尺度と評価語および極性の抽出

この結果を元に，評価属性毎にアイテムの評価に与える影響度を推論する．アイテムの評価はレビューに付けられている星の数などを利用し，例えば星 1 つから 5 つの 5 段階評価である場合，星 1 つおよび星 2 つを不評，星 4 つおよび星 5 つを好評として判断する．本研究ではこの影響度を属性スコアと呼ぶ．ある評価属性における属性スコアは以下に示す極性一致率 [12] を用いて算出する．

$$\text{極性一致率} = \frac{\text{好評頻度}}{\text{好評頻度}+\text{不評頻度}} \quad (1)$$

好評頻度はある評価属性が使用されているレビュー文の中でそのアイテムが好評と評価された文の割合を表す．提案手法では，これらの値を極性毎に分けて求める．ある評価属性の好評に関する属性スコアは，その属性が好評を表す評価語と共に使用された場合に着目して式 (1) により求める．不評に関する属性スコアは不評を表す評価語と共に使用された場合に着目し，式 (1) において好評頻度と不評頻度を入れ替えて求める．

以上のように求めた各評価属性の属性スコアを用いてユーザモデルを作成する．あるユーザが書いた全てのレビュー文から好評，不評それぞれの属性スコアを計算し，評価属性ごとに保持する．属性スコアが好評・不評の少なくとも一方について高い評価属性はユーザが強いこだわりをもっており，アイテムの評価に影響を与える「推薦時に重要度の高い属性」であると推論される．一方でスコアの低い評価属性はアイテムの評価にそれほど影響を及ぼさず,「推薦時に重要度の低い属性」であると推論される．

4　レビュー分析結果

本研究ではユーザモデリングを行うためのレビュー分析対象として，楽天データ公開[1]にて利用可能となっている「みんなのレビュー・ロコミ情報」を用いる．これらのレビューは「楽天みんなのレビュー[2]」にて公開されているユーザが投稿した商品レビューであり，今回はこの中からジャンル「DVD」および「Blu-ray」に属するレビュー 20,576 件を抽出し分析対象とした．

4.1　分析対象とするレビューの絞り込み

ユーザモデリングを行うためには対象となるユーザがある程度の数のレビューを投稿している必要があることから，それぞれのユーザが何件のレビューを投稿しているかを集計しまとめたものが図 2 である．「DVD」および「Blu-ray」に属するレビュー 20,576 件を投稿したユニークユーザ数は 11,462 名であり，その中でレビューを 1 件しか投稿していないユーザは 7,888 名存在した．ユーザモデル作成に必要なレビュー数の閾値については別途検討する必要があるが，仮に 10 件とした場合，その条件を満たすユーザは 147 名存在した．しかし，この 147 名が投稿したレビュー文を見ると図 3 に示すようにその多くは同じ内容が記載された，いわゆる「コピペ」に相当するレビューであった．さらに，多数のレビューを投稿しているユーザの多くは全てのレビューにおいて星 5 つ（5 段階評価）を付けているなど，これらのレビューについては商品に対して適切な評価が行われているとは言い難い．

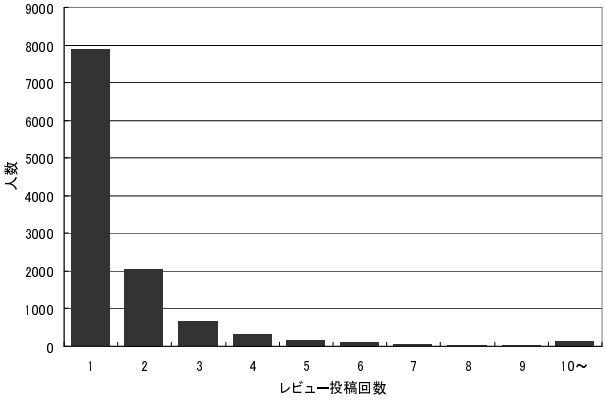


図 2: レビュー投稿回数毎のユーザ数

投稿者	レビュータイトル	レビュー内容
user78164	こんなに清楚な	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	可愛すぎる♪	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	レベル高いね♪	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	妄想世界へ	僕も妄想世界へ旅立ちました！安く買えてよかったです。迅速対応と丁寧な
user78164	これくらいの体型	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	まあ可愛い♪	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	買えて良かった	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	可愛いね	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	大好き♪	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	可愛いね♪	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	可愛いね♪	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	可愛いね♪	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状
user78164	可愛いね♪	安く買えてよかったです。迅速対応と丁寧な梱包に感謝しています。商品状

図 3: 同じ内容が記載されたレビュー例

このようなレビューが投稿される理由のひとつとして，楽天市場に出店している店舗の多くが商品レビューを投稿することで特典を付与する（送料を無料にする・ポイントを加算する）等の施策を行なっていることが挙げられる．これらの特典を受けるために必要な条件はレビューを投稿することのみでありレビューの質は精査されないため，この特典が同一のユーザによって同内容のレビューや最高評価のレビューが大量に投稿される大きな要因となっていると考えられる．

そこで，上記のようなレビューを除外して，適切なレビューのみを分析対象として利用するため下記 3 つの条件に基づきレビューの絞り込みを行ったところ，適合したのは 20,576 件中 342 件のレビューとなった．

1. 10 件以上のレビューを投稿しているユーザが書いたレビューであること
2. 複数のレビューに同内容の表現（コピペ）を用いていないこと
3. レビューを書いたユーザが最高評価（5 つ星）以外の評価を 1 つ以上の商品に対して付けていること

[1] http://rit.rakuten.co.jp/rdr/
[2] http://review.rakuten.co.jp/

4.2 評価属性・評価語の抽出

以上のように絞り込んだ 342 件のレビューを対象として係り受け解析器 cabocha[13] を用いて解析を行った結果，係り受け関係をなす 433 の組を取り出すことができた．しかし，その中から評価属性と評価語の組み合わせとして適切であるかどうかを手動で分類したところ，適切と考えられる組み合わせは 54 組しか存在しなかった．抽出した（またはできなかった）係り受け関係の例を図 4 に示す．

レビュー文	係り受け関係
個人的に、女優さんが好みです。	女優さん => 好み
ストーリー自体は単純明快！	ストーリー自体 => 単純明快
身の毛のよだつような恐怖感は一切感じられない。	恐怖感 => 感じる
ミュージカルはマイナス要素でしたが、全体では傑作と思いましたので星五つで。	なし

図 4: レビュー文より抽出した係り受け関係の例

図 4 において，1 番目と 2 番目の例は適切に評価属性・評価語の抽出を行えたケースである．一方で，3 番目の例においては「恐怖感は一切感じられない」という記述に対して「恐怖感 ⇒ 感じる」という語の組み合わせが抽出されており，抽出結果のみを用いてしまうと本来記載されている意味とは逆の表現になってしまう．そのため，このようなケースでは評価属性・評価語の抽出に加えて文章全体の極性を判定し，その結果を評価語に反映させていく必要がある．また，4 番目の例では「ミュージカル ⇒ マイナス要素」という係り受け関係が評価属性・評価語の組として抽出されることが期待されたが，今回の構文解析では抽出することができなかった．

4.3 検討すべき課題

以上に述べたレビュー分析の結果から，今後検討していく必要があると考えられる課題を 3 点挙げる．

4.3.1 評価属性に用いられる語の分類

評価属性として，映画であれば「ストーリー」や「出演俳優」「演出」等が挙げられるが，レビュー分析によって抽出されるキーワードには無数のパターンが存在する．具体例として,「出演俳優」に関するものであれば「俳優」「女優さん」「主演俳優」「キャスト」といった表現であり，多くの表現がレビュー文の中で用いられている．これらは本来同じ評価属性についての評価を記述したものであることから，有用なユーザモデルを構築するためには類義語辞書や類似度計算等の手法を用いてこれらの表現を自動で分類していく仕組みが必要になると考える．

4.3.2 アイテムのモデリング

今回の分析結果のように，レビュー文からユーザモデリングに必要な評価属性・評価語を抽出する精度はそれほど高くないことから，多くのレビューを投稿していないユーザについては十分なユーザモデルを構築できない可能性がある．ユーザの持つこだわり・価値観はユーザモデルとして表現されるべきであるが，ある評価属性がアイテムの評価にどの程度影響を与えるかはアイテムによって異なるケースも多いと考えられる（例：ストーリーの評価は低いが全体的な評価は高い映画）．ユーザモデリングと併せてアイテムに関しても同様にモデリングを行うことで，評価に高い影響を与える評価属性をより高い精度で推論できるのではないかと考える．

4.3.3 他の情報源を用いたモデリング

今回の分析結果を踏まえレビュー文から評価属性・評価語を抽出する精度を上げていくことも必要だが，それと同時に他の情報源から得られる評判情報を分析していくことも併せて必要であると考える．例えば，価格.com[3]に投稿されるレビューでは，各商品は図 5 に示すようにジャンル毎にあらかじめ定められている評価属性について評価が与えられている．この各評価属性に対する評価値をユーザモデリングに用いることができれば，分析対象となる評価属性は限定されるもののレビュー文の構文解析を行うことなくユーザのこだわりを推論できる可能性がある．

図 5: 価格.com の評価属性別レビュー

5 おわりに

本稿では価値観に基づく情報推薦を実現するための，ユーザのこだわりに着目したレビュー分析手法について

[3]http://kakaku.com/

提案した．提案手法を用いて「楽天みんなのレビュー」に投稿されたレビュー文の分析を行い，その結果を踏まえ検討すべき課題について考察した．今後は考察内容に基づくレビューの分析，および他の情報源を用いたアイテム・ユーザモデリングを行なっていく予定である．レビュー分析にあたっては情報通信研究機構（NICT）が公開している意見（評価表現）抽出ツール[4]が有用であり，文章全体の極性判定等に活用できると考えている．

参考文献

[1] C. N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, “Improving Recommendation Lists Through Topic Diversification,” WWW ’05 Proceedings of the 14th international conference on World Wide Web, pp.22-32, 2005.

[2] 神嶌 敏弘, 推薦システムのアルゴリズム (2), 人工知能学会誌 23 巻 1 号, pp.89-103, 2008.

[3] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, pp.175-186, 1994.

[4] J. L.Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, “Evaluating collaborative filtering recommender systems,” ACM Transactions on Information Systems, Vol.22, No.1, pp.5-53, 2004.

[5] 清水 拓也, 土方 嘉徳, 西田 正吾, 発見性を考慮した協調フィルタリングアルゴリズム, 電子情報通信学会論文誌 D, Vol.J91-D, No.3, pp.538-550, 2008.

[6] 秋山 高行, 小原 清弘, 谷崎 正明, Serendipity のある推薦システムの方式提案と検証, 電子情報通信学会技術研究報告 109(272), pp.81-87, 2009.

[7] M. Rokeach, “The Nature of Human Values,” New York: The Free Press, 1973.

[8] D. E. Vinson, J. E. Scott, and L. M. Lamont, “The role of personal values in marketing and consumer behavior,” The Journal of Marketing, Vol. 41, No. 2, pp. 44-50, 1977.

[9] M. B. Holbrook, “Consumer value: a framework for analysis and research,” Routledge, 1999.

[10] 小林 のぞみ，乾 健太郎，松本 祐治，立石 健二，福島 俊一, テキストマイニングによる評価表現の収集, 情報処理学会研究報告, 2003-NL-154, pp. 77-84, 2003.

[11] 平山 拓央, 湯本 高行, 新居 学, 高橋 豐, 属性評価モデルに基づく商品評価の抽出と提示, 第 9 回日本データベース学会年次大会, F2-5, 2011.

[12] 金山 博, テキストを用いた評判と嗜好の分析, 情報処理, 48 巻 9 号, pp.1001-1007, 2007.

[13] 工藤 拓，松本 裕治, チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, 43 巻 6 号, pp.1834-1842, 2002.

[4] http://alaginrc.nict.go.jp/opinion/

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第1回)
SIG-AM-01-02

主体的な自己探求のためのテキストマイニングツールと web スクレイピングのためのフレームワーク

The text-mining tool for an active self-quest and the framework for web scraping

伊藤貴一、[1]　熊坂賢次 [2]

Takaichi Ito[1],Kenji Kumasaka [2]

[1] 慶応義塾大学院政策・メディア研究科
[1] Graduate School of Media and Governance, Keio University
[2] 慶応義塾大学環境情報学部
[2] Faculty of Environment and Information Studies, Keio University

Abstract: This paper describes two tools. One is making the network of the relation of language by oneself, and it is a tool which visualizes the self-recognition which nestled up to data rather than analysis of objective text data. Another is a framework for the data acquisition from a web. Although it is already known that web has a lot of data, acquiring it needs special skill. It is facilitated. For text mining, since it is a required tool, it states.

0.はじめに

この論文では二つのツールについて述べる。一つは、自分自身で言葉の関係のネットワークを作ることで、客観的なテキストデータの分析というよりも、データに寄り添った自己認識を可視化するツールである。もうひとつは、ウェブからのデータ取得のためのフレームワークである。web に大量のデータがあることは既に知られているが、それを取得するのは専門の技能を必要とする。それを簡便化するものである。テキストマイニングのためには、必要なツールであるので述べる。

1.主体的な自己探求のためのテキストマイニングツール

現在多くのテキストマイニングツールが開発されているが、それらのツールを利用するユーザは、ツールが解析した客観的で絶対的な結果を前提に、その結果を懸命に解釈する受動的な他者でしかない。そのため、テキストマイニングツールを魔法の道具だと思って、その結果をただ受け入れるという考えない人たちを作りだしてしまう。しかし、テキストには文脈や背景知識といった暗黙知が含まれており、機械的には分析しにくいものを多分に含んでいるため、考えずただ受動的にその結果を受け取るという態度は望ましくない。むしろ、ユーザは解読する主体として自らの問題意識に従って、ツールが合理的に判断した素案と対話しながら自分なりに納得する成果を導き出す、という「自己探求的で対話的な関与」を可能にするツール開発が必要とされる。このようなコンセプトに基づいて開発した、柔らかい構造化ツール『Hipparu-McS：ヒッパルーマックス）』[1]である。

そのため、テキストの客観的な事実を可視化するというよりも、客観的事実と主観的実感をすり合わせていく作業をするためのツールであるといえ、むしろ自己認識を可視化するという性質をもつ。分析者の頭を働かせるようにすることがこのツールの目的である。

また、テキストには潜在的に、いつ、どこで、だれが、書いたのかという５W1H のような情報は付随する。例えば、男が書いた文章なのか、女が書いた文章なのかという情報は、分析の手かがりになる。他にも、テキスト全体が肯定的か、否定的かという情報や、テキストの分類の結果、文章に付随する画像の情報もあるだろう。これらの情報をテキストの中に入れて分析すればいいという小技が存在するが、性質が違うものを混ぜるのはよくないので別処理した方がよい。それら文章に付随する情報を扱うための仕組みがあると分析に深みがますので、その仕組みも実装した。

2.実装

2.1 基本機能

Hipparu-McS は，手軽に使ってもらうために，特殊な形式のデータファイルを作る必要がない，ただのテキストデータを読み込むだけで，処理をしてくれるように実装した（Fig.1）。ツールの画面領域はメニューバー，サイドバー，キャンバス，データビュー，単語ランキングの 5 領域からなる（Fig.1-A）。メニューバーは入力データの指定や SAVE/LOAD 機能，サイドバーはノードの削除や表示の調整などの機能，キャンバスはノードとリンクを描く場所，データビューは文章データを見る場所，単語ランキングは注目単語を指定する場所である。

以下，挙動プロセスを示す。なお，挙動プロセスの番号は Fig.1-B 以降に対応する。

(1) メニューバーの「入力データ」「テキストファイル」を選択し，単語の頻度ランキングを作る。さらに自身の問題意識に基づいて，ランキングから適切な単語を選択し，キャンバスにノードをおく。初期状態が真っ白なのは真っ白な気持ちでデータに向

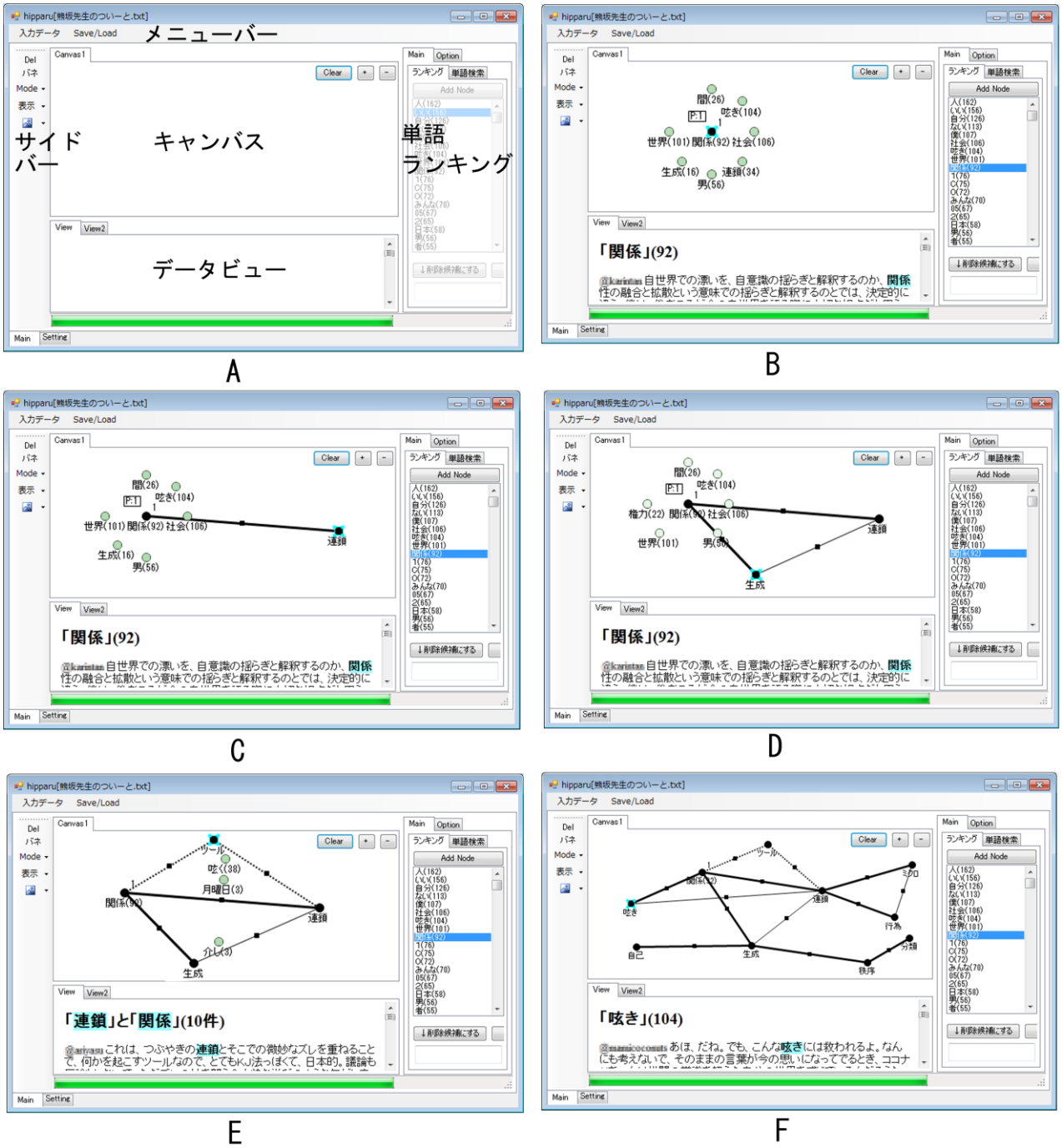


Figure 1　Hipparu-McS の挙動プロセス

かい合ってほしいためである。ユーザ自身で言葉のネットワークを探求するために，ノードはマウスで自由に動かせるようにしてある。そしてノードを選択したとき，その周辺に 7 個の関係の強い語が時計回りで表示され，データビューにその単語を含む文章が表示される。関係の強さは Jaccard（Jaccard= |A∩B|/|A∪B|）指標を使い，基本設定では一行分の文章での単語の関係を計算している。単語の後にある数字は単語の頻度である。また 7 個だけの表示では探求する単語にたどり着けないことがあるので,「P」のところをクリックすると次の候補 7 つを出すことができる。（Fig.1-B）。

(2) 候補語の中で適切だと思った単語をマウスでつかんで「引っ張る」とノードとして採用され，黒リンク（図上では太線）が作られる（Fig.1-C）。この新たに引っ張られたノードを選択すると，その周囲に関係の強い語が表示され，同様に引っ張ることができる。

(3) ネットワークを探求する過程で，すでにノードとした単語と関係の強い単語を選択したとき，新たに赤リンク（図上では細線）が自動的に引かれる。黒リンクが自己探求の線であるのに対して，赤リンクはツールが合理的判断に基づいて引いた線である（Fig.1-D）。

(4)リンク上にある黒いボタンを選択すると，リンクで結ばれる 2 つのノード単語と関係が強い単語の頻度ランキングの上位 4 つが表示される。この 4 つの候補ノードも選択すると緑リンク（図上では点線）が引かれる。この単語は，頻度は低いが，ユーザが探求したリンクと両ノードに共通する単語なので，さらに深い探求を誘発する機能をもつ（Fig.1-E）。

(5) 上記の探求行為の繰り返しによって，複雑で多様な構造図が作成される。（Fig.1-F）。

2.2 文章の属性処理

文章に属性をつけることを考えた時、理想的な形式は、XML のような構造化された形式であろう。しかし、先に述べたように、簡単に使ってもらうために、シンプルなテキストを入力データとしてもらうことを目指している。そのため、簡易的な記法を作った。

テキスト<分類カテゴリ:属性, 分類カテゴリ 1:属性 2,・・・>

このような形で、文章の後に<>で括ったものを付け加えることで認識されるようにした。＜＞の中は、カンマ区切りで、属性の数は可変である。また分類カテゴリをつけることで、属性の関係を定義する。例えば、ただの「男」「女」だけの情報では男と女に排他関係があることが機械にはわからない。そこで、「性別:男」「性別:女」とすることで他の多数の属性があったとしてもその関係がわかるようになる。

そして、属性のデータは、分析時に文章の生データを見るとき読めるようにすることや、その集計、また、にこの属性によるフィルタ機能を使えるようした。

3.分析事例

女子学生が自分の食べ歩きブログ 98 記事を素材に，Hipparu-McS を利用して，彼女自身の食意識を探求した柔らかい構造化の成果を取り上げる[2]。

おいしいという言葉に注目してグラフを作っていく中で、漠然としていた自分の食に対する意識が構造化されていった（Fig2）。その中で、自身が注目しているのは、味、香り、食感の三要素であるということに気付いた。そして、さらに、その「味」という言葉に注目することで、さらなる食に対しての変数を気づくことができた（Fig3）。自身が書いた文章をツールを通じて分析することで、自分でも思っていなかった関係性や構造が発見され,「自分はそんなことを考えていたのか」という新鮮な気づきがもたらされた。

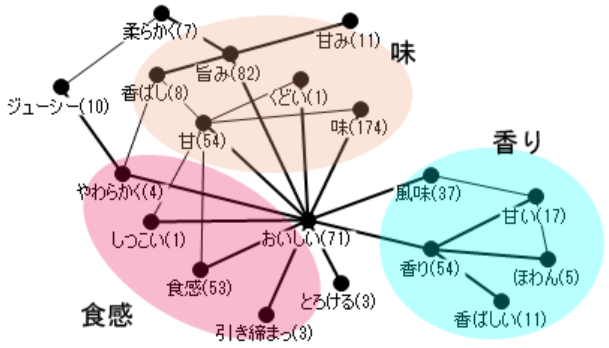


Fig. 2 「おいしい」の単語と共起関係

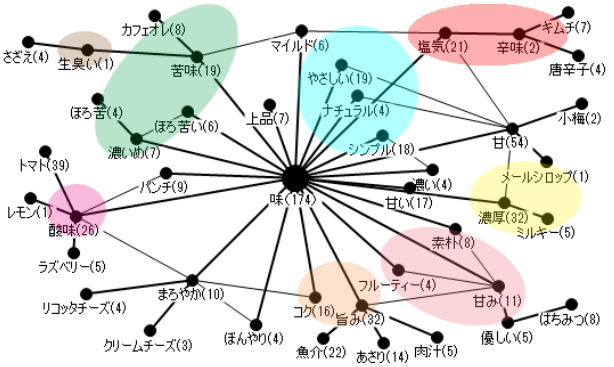


Fig. 3 「おいしい」のさらなる探究

4. Hipparu-McS の課題

このツールでは、同じテキストを分析しても、分析者が違えば、分析者のその時の関心が違えば、作られてくるグラフは全く違うものになってしまう。データに沿って作られているのだから、これをいけないことと否定的に見る必要はない。そもそもテキストを分析する視点は多様であり、そしてその解釈も多様である。しかし解釈は多様だ、というだけで終わっては単なる相対主義（主観主義）であり、その多様性を統合する視点をいかに仕組みとして組み込むか、という重要な問題が残る。そのため、同一のデータから作られる多様なグラフを統合する仕組みが必要である。また、主観重視といっても、客観的に提供される情報が貧弱であるという課題もある。今は単語のランキングと単純な単語の２者関係でしかなく、全体像として客観性というものを認識できる情報が弱い。その点も強化しなくてはならない点だろう。

あと、属性を分析するための仕組みを用意しているが、その機能もただのビューワとフィルタでしかなく、貧弱である。違う属性との比較や、同じグラフでも、属性を変えるとリンクが変化していく様子を表現する機能が必要だと考えている。

5.Web スクレイピングのためのフレームワーク

１９９０年代半ばの商用インターネット開始から、多くのサイトが立ち上がり、web 上に情報が整理され、大量の情報が置かれることになった。Web 上に大量に情報があり、それを分析したいといっても、アクセスするには、プログラムをできる人には、web アクセスのプログラミングを組めばいいのだが、一般の人では、一つ一つ web ページをブラウザでみるなどの方法しかない。大量のデータを扱うことのできるためには、個々人がプログラミングスキルを身につけるべきというは、賛同したいことであるが、現実には習得しなくてはいけないスキルが増えている現代では難しい。そのため、欲しい web 上のデータの大量取得（Web スクレイピング）を容易にするためのツールが必要である。

しかし、web 上の特定の HTML タグを取ってくるツール（例えば、掲示板の書き込み部分のみを抽出する）というのは簡単に作れるが、欲しいのは構造化されたデータ（例えば、掲示板の書き込み、書いた日付、書いた人の名前などの情報が一つの塊としてあるデータ）であり、web の多種多様性に適応するような仕組みでなくてはならない。そのために、機能を小分けし、モジュール化して、それをユーザが対象の web サイトに合わせて組み合わせるという方法がいい。そのため、ツールというよりも、フレームワークという形にした。

プログラマの立場からすると、Web スクレイピングは、今、Web 関係のライブラリはメジャーなプログラミング言語には揃っているので、自分の習得した言語で、容易に作れるようになってきている。しかし、それでもプログラミング言語のもつ決まりごと（変数の宣言、変数の一時格納、ループ処理、後処理など）に多くの労力を削がれ、書くコード量が増え、重要なパラメータがあちこちに散らばり、見通しが悪くなってしまう。そのため、必要なパラメータだけを記述して、機能を組み合わせるだけのフレームワークにするのは、全体の見通しがよくなり有益である。

機能を組み合わせるというのは、データフローを作るということに他ならず、web からのデータ取得をデータフロープログラミングするためのフレームワークでもある。これは、プログラミング言語のパラダイムで言う、手続き型言語から関数型言語に変形させることであるともいえる[3]。

このような変化は、web スクレイピングは対象のウェブサービスに依存し、そのデザインが変更された時に作り直さないといけないものなので、後に変更されることを前提とし、見通しのよさを作るのは必要なことである。

このフレームワークは、http://rawler.codeplex.com/ にて Rawler フレームワークとしてオープンソースで公開している。

似たような web スクレイピングのためのツールとして Java で作成されたオープンソースプロジェクトの Web-Harvest[4]がある。これとの違いは、Web-Harvest は XML でのプログラミングに念頭が置かれているが、Rawler フレームワークは、ビジュアルプログラミングしようという構想があったことに起因する関数だけで記述しようとしていることであると思われる。

6.仕様と実装

実装は C#で行い、使うときには XAML で記述できるようにした。XAML とは、Extensible Application Markup Language の略で、マイクロソフトの XML をベースとした拡張である。特徴としては、XML のタグが C#のクラス名であり、記述することでインスタンスが生成されるため、オブジェクトの状態と関

係を記述できる仕組みである。主にアプリケーションの外観のデザイン（ボタンの配置など）に使われている（例：WPF／Silverlight)。XAML はあくまでテキストなので、そのルールがわかっていれば、コピーアンドペーストに優れ、別のところで書いたコードがそのまま使える。また、エディタの機能で折りたたみができ、そうすると見通しがよいものになる。そのため採用した。

6.1 基底クラスの設計

XAML の仕組みを使い記述するために、すべての機能は、RawlerBase という基底クラスを継承したものになっており、オブジェクト指向言語のポリモーフィズム(多態性)を利用したものになっている。RawlerBase には、主にオブジェクトの親子関係を格納する Parent,Childern プロパティと、オブジェクト自身が持つ Text プロパティ、そして Run()メソッドがある。XAML は XML なので木構造であり、タグの入れ子関係で記述できるように親子関係の情報を持っている。Run()メソッド実行すると親の Text プロパティを参照して Run を実行し自身の Text を作り、そして子の Run()を実行するというものになっている。

そのため、XAML で作られた親子関係の木構造に沿って、深さ優先探索のように実行されていく。

木構造でのデータの流れの記述は時として、深い階層になってしまい、可読性を落とすことにつながる。例えば、部分の抽出→タグの消去→改行の削除→空白削除といった処理などをすると一気に階層が深くなってしまう。そのため、RawlerBase には PreTree プロパティがあり、そこでは、RawlerBase で前処理を記述することができ、それは、その命令が行われる前に実行される。行数を必要とするので長くなるのだが､多くの XML のツールでは折りたたみができるので、折りたたんでしまえば、苦にならない。そして、コードの見通しはよくなる。

6.2 複数処理（繰り返し処理）

HTML での Link の取得のように複数になるものもある。上に書いた方法では単数しか処理ができないので、RawlerBase を継承した RawlerMultiBase クラスをつくり子をリストに入っているデータの数だけ複数回実行する、複数処理に対応させた。このようにすることで単数複数を気にせず配置できるようになっている。命名規則として単複を意識させるため、主要な複数のものは複数形にしている。(例 Links、Tags、ReadLines など)

また、繰り返し処理の制御のためのクエリ機能の RawlerQuery プロパティがある。これは C#の LINQ (Language Integrated Query： 統合言語クエリ)に影響を受けたもので、その部分的なラッパーである。これを使うことで、初めの要素の抽出、最後の要素の抽出や、指定した条件に適合するものだけを抽出するなどの得られた複数のテキストに対する処理ができる。

6.3 継承されたクラス群

個々の機能は、RawlerBase、RawlerMultiBase を継承したものである｡Page クラスは親テキストを URL として web ページにアクセスする。Tags クラスは html を解釈して指定したタグを抽出する。Links クラスはリンクを抽出する。このような形で web ページからのデータ取得はできる。データの処理としては、Data クラスはデータを蓄積し、DataWrite クラスは Data クラスに属性を付けて Text を書き込み、NextDataRow クラスで今まで書いたのを一つの塊として確定させる。この一連の処理で構造化されたデータの取得が可能になる。

また、制御として、IF 文相当の Contains（指定した文字列が含まれていると実行される）や Equal（指定した文字列と同じ場合に実行される）Switch 文相当の Switch クラスがある。これで、構造化プログラミングの順次、反復、分岐の三要素をすべて満たすことになる。順次は XAML の木構造、反復は RawlerMultiBase クラス、分岐は、RawlerQuery や Contains クラスなどの命令群である。

そのほか、３０近くのクラスがある。ログインが必要なページにアクセスするためのログイン機能も存在する。これらを使うことで様々な種類の web ページからデータ取得だけではなく、ファイルの読み書き、さまざまな繰り返し処理、テキストの変形、エラー報告までできるものとなった。

6.4 拡張性

C#のプロジェクト内で使えば、WPF アプリの作成のようにコードビハインドとして、オブジェクトに対して追加のイベント処理ができるため、データベースとの連携処理や、さらに細かい処理をすることができる。

また、RawlerBase を継承したクラスを作れば、このフレームワークに乗っかることができる。テキストを変数として、テキスト（単数複数）を返す関数はいかようにも作れる。著者自身、再利用する可能性が高い必要な処理はその都度作っている。

他にも色々な可能性を考えることができる。たとえば、グーグルやツイッター、フェイスブックなどの WebAPI を使いデータを取得することも、継承したクラスを作ればいい。形態素分析を行うことや、

テキストの分類し、そのクラス名を返すこと、テキストがポジティブかネガティブか判定するといったテキスト分析も継承したクラス作ればいいだけである（形態素解析するクラスはすでに存在する）。他のテキスト処理と組み合わせることができるため、メリットは大きい。このようにさまざまなことを扱うことが本来的にできるため、ソースを非公開にせず、オープンソースにして公開している。

7.具体なコード例

RawlerTool は、XAML を入力し、実行できる環境を提供する(Fig.4)。XAML の作成もできるが、入力補助がある VisualStudio を使うことが望ましい。(code:1)はブログのコメントを取得する例である。Data タグにある xmlns は、使用する DLL の指定、決まり文句である。Page のところで URL を指定すると取得開始し、Tags でコメント部分だけを抽出し、そして、DataWrite で Data に対して、これは comment であるという情報を付けて書き込む。PreTree でタグを削除する前処理をしている。同様に、名前部分を取得する。ClipText は始まりと終わりを指定してそこに挟まれるテキストを取得する。

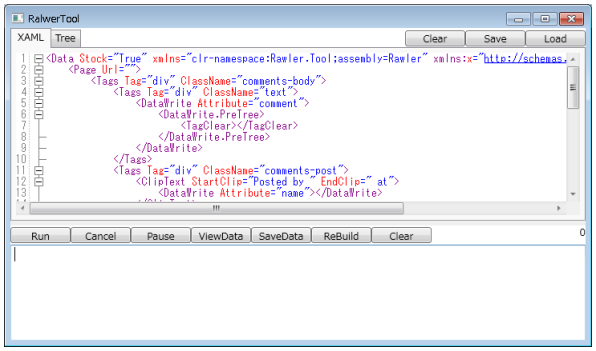


Fig. 4　RawlerTool の画面

そして、NextDataRow でそれが一つの塊であると確定し、次の塊に移る。

ページ送りに対応していて、「次のコメント」というリンクを探し、NextPage を実行すると、Page にその URL を読み込む命令をし、同じことが繰り返される。これによりすべてのコメントの取得ができることになる。

Blog の URL のところを書き変えて、これを実行すると、その記事のすべてのコメントと書いた人の名前のペアが Data に蓄積される。このように必要な記述だけで、ページの取得、繰り返し処理、データの取得が行うことができる。

```
<Data
   xmlns="clr-namespace:Rawler.Tool;assembly=Rawler"    xmlns:x="http://schemas.microsoft.com/winfx/2006/xaml" >
   <Page Url="ブログ記事の URL">
      <Tags Tag="div" ClassName="comments-body">
         <Tags Tag="div" ClassName="text">
            <DataWrite Attribute="comment">
               <DataWrite.PreTree>
                  <TagClear></TagClear>
               </DataWrite.PreTree>
            </DataWrite>
         </Tags>
         <Tags Tag="div" ClassName="comments-post">
            <ClipText StartClip="Posted by " EndClip=" at">
               <DataWrite Attribute="name"></DataWrite>
            </ClipText>
         </Tags>
         <NextDataRow></NextDataRow>
      </Tags>
      <Links LabelFilter="次のコメント" IsSingle="True">
         <NextPage  ></NextPage>
      </Links>
   </Page>
</Data>
```

Code.1 Rawler のサンプル。ブログのコメント取得

これだけでは、毎回 Blog の URL を入れないとだめなので不便である。その時は、Page の直近の親で対象とする URL のリストつくり繰り返しを行わせればいい。繰り返す内容を直接書いてもいいし、ファイルから読み込ませるのでもいい。そのようなクラスは用意されている。

Page は親にテキストがあれば、それを URL としてアクセスし、HTML を取得する。そのため、Page→Link→Page というようにすれば、ブラウザでページ移動する感じで複数のページにアクセスして取得することが簡単に記述できる。このため、一覧ページと詳細ページがあるような構成のサイトでのデータ取得では威力を発揮する。

8.活用事例

すでにこのフレームワークを使い、１０程度のサイトからのデータ取得の実績がある。比較的容易にデータ取得ができるため、その分、分析する対象が広がる。

一例として、AKB48 の分析を挙げる。国民的アイドルとなった AKB48、ファンとのコミュニケーションのためにブログをかいている。アイドルのブログの記事そのものはアイドル相応なもので特筆することではないが、コメント欄がすごい。総選挙での上位陣は、コメント数が万単位であり、投稿時直後だけではなく継続的にコメントが書かれている。このテキストを取得するために、このフレームワークを使いデータを取得し（先ほど例示したコード）、前述の Hipparu-McS を使い分析を行った。いろいろデータを探索していく中で、それぞれのメンバーでの「ブログ、テレビ、握手」の使われ方の違いに気付き、ちょうど３つなので、それを三角グラフにプロットした。（Fig.5）

AKB について特に知らない人にはどのメンバーも同じことをしているように見えるかもしれない。しかし、このように可視化すると、ファンの反応として、AKB48 のメンバーそれぞれのメディア戦略が違うことがうかがえる。また、「僕」「私」「俺」「みんな」といった人称代名詞の使い方にも差があり、円グラフを作ってみると、それぞれで全然違うファンであることが考察できる。詳細は「アイドルブログのコメント欄から見る、「君と僕の関係」」というタイトルのブログで公開してある[4]。この記事はソーシャルブックマークサービスのはてなブックマークでは、７００近くブックマークを集める人気記事となり、ブックマークコメントでは、面白い分析だ、ブログのコメント欄に注目することが面白い、ということが多数書かれている。

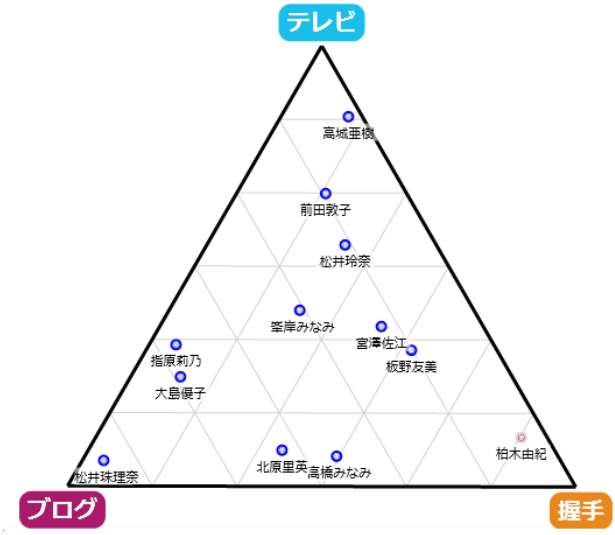


Fig. 5 AKB48 のメディア戦略の違い

このような高い評価になったのは、分析した彼女のセンスによるところが非常に大きい。このようなセンスある人に、分析できるデータをすばやく用意できるというのがすごく重要なことである。

つまり、重要なことは AKB48 の分析をしたということではなく、すばやくデータ取得し、分析できるツールを揃えることによって、分析できる世界が広がるということである。分析はトライアンドエラーの繰り返しなので、そのサイクルが速いことがいいに決まっている。

9. Rawler フレームワークの課題

著者であれば、通常のサイトのクロールを、３０分から１時間程度でこのフレームワークを使って作ることができる。しかし、現在、オープンソースで公開しているが、ドキュメントが圧倒的にすくなく、誰もが使えるというものになっていない。そのため、サンプルの数を増やして、ドキュメントを整備することが必要であろう。

また、作成のためのツールも必要である。VisualStudio での XAML の作成は完成度が非常に高いが、無償で配布されているとはいえ、多くの人に VisualStudio のインストールを必須にするのは酷であろう。結局のところ半ばむき出しのプログラミングなところがあるので、プログラマには相性が良いところがあるだろうが、初心者にも扱いやすいビジュアルプログラミングをできるようなこともできるようにする必要があるだろう。

Rawler フレームワークの価値は、データの成型の命令を柔軟に記述することができるということである。そのため、前述した、Hipparu-McS でも、現状

では一時的にテキストファイルに書き出せば分析できるが、フレームワークで記述することでデータの入力をできるようにすれば、web からの情報を直に分析するということが可能になる。データの入力のインターフェースになりうる。このようなことをできるようにしていきたい。

参考文献

[1] 伊藤貴一, 熊坂賢次, 諏訪正樹, 花房真理子「自己探求する柔らかい構造化ツール(HIPPARU-MCS)の実装と評価」, COMPUTER & EDUCATION VOL.029,2010

[2] 花房 真理子, 熊坂賢次, 伊藤貴一,「おいしさの探求ーブログのテキスト解析によるおいしさの意味世界の可視化ー」情報処理学会, 第 8 回ネットワーク生態学シンポジウム, 神奈川, 2012 年 3 月

[3] J. Hughes, Why Functional Programming Matters, In D. Turner, editor, Research Topics in Functional Programming, Addison Wesley, 1990

[4] http://web-harvest.sourceforge.net/

[5] http://d.hatena.ne.jp/haruna26/20120204/1328351411

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第1回)
SIG-AM-01-03

複数の時系列データの比較に基づく グラフの言語表現生成手法

Generating Linguistic Expression of Charts Based on Comparison of Multilpe Time-Series Data

末吉 れいら [1]　松下 光範 [1*]　白水 菜々重 [2]
Reira Sueyoshi[1]　Mitsunori Matsushita[1]　Nanae Shirozu[2]

[1] 関西大学
[1] Kansai University
[2] 奈良先端科学技術大学院大学
[2] Nara Institute of Science and Technology

Abstract: This paper proposes a method for generating linguistic expressions from a time-series data. The proposed method takes differences and similarities among multiple time-series data into consideration: The method generates linguistic expressions by executes three processes sequentially. First, a characteristic such as “rise,” “drop,” and “stable” is evaluated in each data point of the data series. Second, for each data point in a data series, a weight is assigned by calculating a degree of attention, which is estimated by comparison with another time-series data. Finally, the most pertinent expression is selected.

1 はじめに

現在、インターネットを介してさまざまな時系列データや統計データを得ることが出来るようになってきた。しかし「ここ一週間で急落した株は?」や「価格が緩やかに上昇している商品は?」といった言語による検索要求を通じてユーザの意図や関心に合致した区間・粒度のデータを得ることは困難である。このような検索要求から、その条件に見合った変動をしている時系列データを特定したり、特定の時系列データから該当する時期を見つけたりすることができれば、ユーザの時系列データに対するアクセス性の向上が期待できる。本研究のゴールはこのような情報アクセスを可能にする技術を実現することであり、現在そのひとつのアプローチとして、時系列情報を予め自然言語表現で記述しておき、それとユーザの検索要求とのマッチングによって適切な範囲・粒度の時系列情報を特定し、視覚化する手法の実現を目指している [1, 2]。

我々は、このような情報アクセス技術の実現に必要な要素技術として、(1) 時系列データに基づく言語表現の生成、(2) 自然言語で表現された質問の解釈、(3) これらふたつのマッチング方法の定式化、が必要であると考えている [1]。本稿ではこのうちの（1）に焦点をあて、時系列データの持つ解釈の多様性を考慮した言語表現の生成について検討する。

(1) に関して最も効率的・効果的な方法は、時系列データとそれを説明したテキスト（新聞記事など）を対応づけて、時系列データの特徴を適切に表現している文を抽出することであるが、このようなテキストが常に得られる保証はない。そのため、時系列データのみが与えられた状態でも、そこからその時系列データを適切に表現する言語表現を生成する技術が必要になる。

ここで注意すべきは、同じ振る舞いのデータであっても状況や文脈によって解釈が異なる場合があるという点である。例えば、ある企業の株価が変動した場合、同じ値幅の下落であっても特定の銘柄だけ下落していればその下落に注目がいくが、多くの銘柄が下落していればあまり注目に値せず、他の特徴に注目するだろう。すなわち、人は探索の文脈や状況に応じて時系列データの注目点を変えることで、適応的なデータの解釈を行なっていると言える [3]。

本研究ではこのような、ユーザが行う複数の時系列データの比較行為に着目し、それをモデル化することで、より人の直感に沿った時系列データの探索・アクセスを可能にすることを試みる。この方針の下、本稿では、状況や文脈によって変化する解釈を取り扱うた

*連絡先: 松下 光範 関西大学総合情報学部 〒569–1095 大阪府高槻市霊仙寺町 2–1–1 Tel: (072) 690–2437 Fax: (072) 690–2491 e-mail: mat@res.kutc.kansai-u.ac.jp

め、時系列データに最大値や最小値、上昇・下降・安定といった特徴を付与し、特徴の重要度に応じて重み付けを行う手法を提案する。この手法では、異なる種類のデータ、あるいは、同じ種類のデータの異なる期間のデータといった異なる複数の時系列データを比較する場合に、それぞれの特徴の差異に応じて動的に重みを変更する。

2 提案手法

前節で述べたように、複数の時系列データを比較して分析することで複数の事象に跨った包括的な知見を獲得し、より深いデータの理解が可能になると期待される。しかしこの場合、比較対象に応じて時系列データの持つ値の「意味」が相対的に扱われるため、文脈による解釈の変化が生じる。

そこで、本研究では状況や文脈によって変化する解釈を取り扱うため、値の上昇や下降、安定といった時系列データの変化傾向を特徴として捉え、特徴の重要度を算出して各言語ラベルに重み付けを行う。付与された重みは、ユーザの要求に合った複数の時系列データを比較する場合に用いる。特徴の類似性や特異性とユーザの要求への合致度を加味して動的に重みを変更することで、時系列データの相対的な評価を考慮した言語表現を生成する。

図 1 に提案手法の概要を示す。この手法では、予め時系列データに対する特徴の付与、付与した特徴に対する重み付けを行う。続いて、ユーザの要求によって絞り込まれた時系列データを対象に複数の時系列データを比較することで動的な重みの変更を行い、ユーザへ視覚的に提示する。

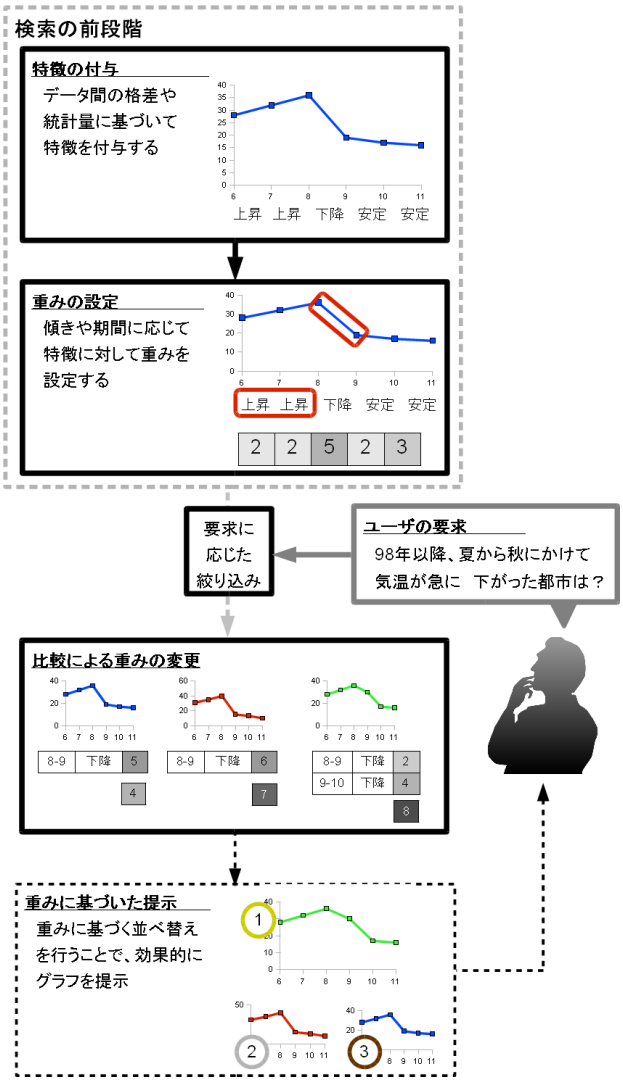


図 1: 提案手法の概要

2.1 時系列データに対する特徴の付与

時系列データに対する重み付けに際し、時系列データの数値情報から重み付けの指標や基準となる上昇や下降などの特徴を言語表現として付与する特徴や特徴の算出方法について述べる。時系列データに対して特徴の付与を行うため、まず時系列データから得られる特徴を整理した。本研究では、時系列データから得られる特徴を基礎統計量、数値間の関係性、異なる統計量比較のための指標といった 3 つの観点に着目し、整理を行った。

基礎統計量については、最大値、最小値、平均値、標本数、データ範囲、標準偏差を対象とした。

時系列データの数値間に生じる関係性については、傾斜、傾斜の寄与度、傾斜傾向、傾斜傾向の持続期間を特徴とし、傾斜の度合いと向きから得られる傾斜傾向を主な特徴とした。このうち、傾斜傾向 (gap tendency) に関しては、時系列データのとる範囲における傾斜の割合である寄与度 (contribution degree)[4] を基に算出した。時系列データ X の時点 $t \in T$ (T は時点の全体集合) における要素を $x_t \in X$ とすると、傾斜に対する x_t の寄与度 $cd(x_t)$ は式 (1) で求められる。

$$cd(x_t) = \frac{x_{t+1} - x_t}{\max(X) - \min(X)} \tag{1}$$

ここで、 $\max(X)$ は X の要素の最大値、$\min(X)$ は X の要素の最小値を各々示している。この $cd(x_t)$ に基づき、時系列データ X の時点 t における傾斜傾向 (現在の実装では「上昇」「下降」「安定」の 3 つ) を付与する。判定の基準は、閾値パラメータを $\tau(> 0)$ とすると、$|cd(x_t)| < \tau$ の場合に「安定」、$cd(x_t) \geq \tau$ の場合に「上昇」$cd(x_t) \leq -\tau$ の場合に「下降」とした。なお、現在の実装では $\tau = 0.05$ としている。

2.2 特徴に対する重み付け

次に提案手法では、2.1 節で求めた特徴を基に重み付けを行う。重み付けを行うにあたり、傾斜傾向と期間を重み付けの対象とし、傾斜や傾斜の寄与度といった特徴は対象に対するパラメータとして扱う。パラメータを元に算出されたそれぞれの重みを元に、対象とする時系列データへの重みを決定する。本研究における重み付けの段階では、比較による動的な重みの評価はユーザの要求を得た際に行うため、言語表現に対する定量的な評価として直接評価法に基づく重み付けを行う。

重み付けの手法については、標本数や傾斜の大小を用いた傾斜傾向に関する重み付けと、傾斜傾向の持続に関する期間の長短を用いた重み付けのふたつを検討している。本稿ではこのうち、前者について述べる。

傾斜や傾斜傾向を用いた重み付けの手法に関しては、以下で述べる 3 手法を検討することとした。

2.2.1 手法 1

この手法は全体的特徴を重視する手法である。すなわち、時系列データ全体において、上昇・下降・安定の各傾斜傾向ごとに標本数を算出し、傾斜傾向の標本数が多い場合に重くなるように重み付けの係数を設定する。

また、上昇・下降に関しては傾斜が大きい場合に重みを増やし、安定に関しては傾斜が 0 に近いほど重みが増加するという基準を設けた。

単一の時系列データにおいて、特徴として付与された傾斜傾向と傾斜を用いて重み付けを行った。重み付けには上昇・下降・安定の各傾斜傾向ごとに標本数に応じて設定した係数を算出し、各傾斜傾向に定められた基準に応じたポイントの付与を行い、それらを併用して重み付けを行う。具体的な方法は以下のとおりである。

係数 (coefficient) は、時系列データの各傾斜傾向の標本数に基づき、標本数の多いものについて値が大きくなるよう設定した。ある傾斜傾向 $gt_i \in GT, GT = \{$ 上昇, 下降, 安定 $\}$ に属する標本数を $num(gt_i)$ とすると、上昇・下降・安定の各傾斜傾向 gt_i に対する係数 $coe(gt_i)$ は、式 (2) によって算出される。

$$coe(gt_i) = \frac{num(gt_i)}{N} \tag{2}$$

ただし $N = \sum_{j=1}^{3} num(gt_j)$ である。 ポイントの設定では、まず、上昇・下降では傾斜が大きいもの、安定では傾斜が 0 に近いもの、という基準に応じて各傾斜傾向内で順位付けを行い、順位による得点を割り当てた。例えば、傾斜傾向「上昇」の標本数が 6 の場合、傾斜の大きなものから順に、1 位には 6 点、2 位には 5 点、3 位には 4 点といったように、その傾斜傾向の標本数に応じた得点を割り当てた。次に、各得点を傾斜傾向内における全ての得点を加算したものにより除算を行い、ポイントとした。時点 x_t の傾斜を $gap_t = x_{t+1} - x_t$ とし、x_t の属する傾斜傾向 gt_i 内の順序を考慮して与えられる得点を $gtp_i(x_t)$ とした場合、与えられるポイント $Point(x_t)$ は式 (3) によって求めた。

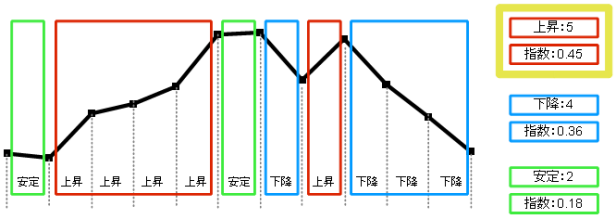


図 2: 手法 1 による重み付け

$$Point(x_t) = \frac{100}{\sum_{j=1}^{num(gt_i)} gtp_i(x_j)} \times gtp_i(x_t) \tag{3}$$

例えば、傾斜傾向の標本数が 6 の場合、傾斜傾向内における全ての得点の合計は 21 となり、傾斜傾向に応じたポイントは順位が 1 位の場合 28.57、順位が 2 位の場合 23.80 となる。

上述したように傾斜傾向ごとに算出した係数とそれぞれの傾斜傾向に与えられたポイントを積算して重み付けを行った。

図 2 に手法 1 による重み付けの例を示す。

2.2.2 手法 2

この手法は大きな変化がある局所的傾向を重視する手法である。全体的な流れや算出方法に関しては手法 1 と同様であるが、重要視する観点として上昇・下降・安定の各傾斜傾向ごとの標本数が少ない場合に重みを増やす。

この手法では、手法 1 と同様に、単一の時系列データにおいて、特徴として付与された傾斜傾向と傾斜を用いて重み付けを行う。重み付けには上昇・下降・安定の各傾斜傾向ごとに標本数に応じて設定した係数を算出し、各傾斜傾向に定められた基準に応じた得点の付与を行い、それらを併用して重み付けを行う。係数は、時系列データの各傾斜傾向の標本数に基づき、標本数の少ないものついて値が大きくなるよう設定する。したがって、傾斜傾向 gt_i に対する係数 $coe(gt_i)$ は、式 (4) のようになる。

$$coe(gt_i) = \frac{N - num(gt_i)}{N} \tag{4}$$

重みに関しては、手法 1 と同じく式 (3) によってポイントを算出し求めた。

図 3 に手法 2 による重み付けの例を示す。

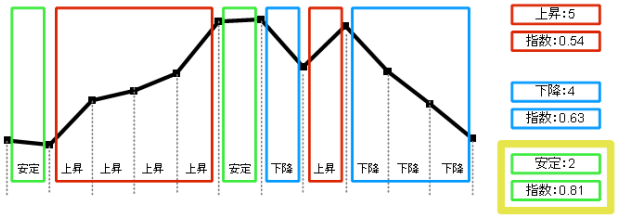


図 3: 手法 2 による重み付け

2.2.3 手法 3

この手法では、算出された乖離を傾斜傾向の重みとして用いる。すなわち、時系列データの全体的な傾向であるトレンドと個々のデータの持つ傾向の乖離が著しい箇所について顕著な特徴とし重みを設定する。

トレンドに関しては、開始点と終了点を線で結んだ 1 本の直線で求める手法や開始点・終了点・最大値・最小値の 4 点を用いて 3 本の直線で求める手法、2 次式の当てはめによる手法、単純移動平均を用いた不規則変動 (ノイズ) の除去によって求める手法などがある。本研究では、トレンドとして単純移動平均を採用し、時系列データの数値から単純移動平均を用いた乖離を算出する。

時系列データの全体的な傾向であるトレンドと個々のデータの持つ傾向の乖離が著しい箇所について顕著な特徴とし重みを設定する。

この手法では、まずトレンドの算出のために単純移動平均によってデータの平滑化を行う。平滑化には、時系列データの非系統的な誤差部分を互いに除去するために局所的に平均を取る単純移動平均法を用いる。単純移動平均法では、各時点のデータをその周辺の n 個のデータの平均によって置き換えることで平均を取る。この幅 n を「ウィンドウ幅」と呼ぶ [4]。

本研究では、対象データからグラフの概形を求める際に単純移動平均法で必要とされるウィンドウ幅を設定するため、対象データのうち、サンプル数が 12 のものと 30 のものに関しては時系列データとサンプル数を 3、5、10 で除算した値のウィンドウ幅で単純移動平均をとった場合のグラフをそれぞれ用意し、比較・検討を行った。ただし、サンプル数が 5 以下のものに関しては、3、5、10 の除算では優位性のある結果が得られなかったため、1、2、3、4 のウィンドウ幅で単純移動平均をとった場合について検証を行った。

この検証の結果、サンプル数が 12、30 のものに関してはウィンドウ幅を データ数/5 に設定することで効果的な値が得られることが判った。しかし、日経平均株価のデータでは、データ数が少なかったために全ての検証において効果的な値は得られなかった。そのため、データ数が少ない場合におけるトレンドの算出やサンプルの取り方に関して考慮する必要性がある。本研究では、上記の結果を基に単純移動平均によるトレンド算出のウィンドウ幅 n について 5 と設定した。

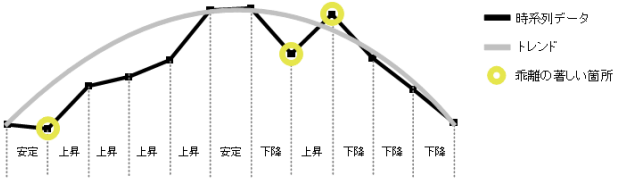


図 4: 手法 3 による重み付け

この単純移動平均により得られたトレンドを用いて、トレンドにおける傾斜を算出した。そして、トレンドの傾斜を時系列データから得られた傾斜によって除算し、傾斜傾向の乖離を求めた。

図 4 に手法 3 による重み付けの例を示す。

3 対象データ

本研究では、複数の時系列データの比較による相対的な評価を得るため、時系列データの中でも、年 (サンプル数: 12) 、月 (サンプル数: 28〜31) 、週 (サンプル数: 2〜5) の区切りで構成された 3 種類の異なるデータを対象とし、検証を行った。

年単位で構成されたデータとして、気象庁の気象統計情報より大阪・札幌の 2010 年から過去 15 年分の月別累計降水量のデータ (サンプル数: 12 件/年) を用いた。

月単位で構成されたデータとして、気象庁の気象統計情報より大阪の 2010 年から過去 3 年分の日別平均気温のデータ (サンプル数: 28〜31 件/月) を用いた。

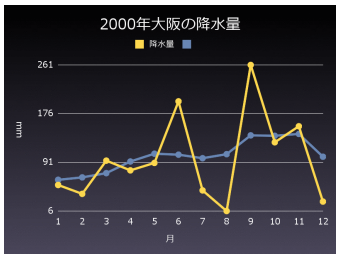
週単位で構成されたデータとして、Yahoo!ファイナンスより 2011 年の 4 月頭から 9 月末までの週別日経平均株価の始値と終値のデータ (サンプル数: 2〜5 件/週) を用いた。

検証では、上記のデータから特徴表現とグラフの生成に必要な形式に変換したテキストファイルを人手で用意した。

4 検証

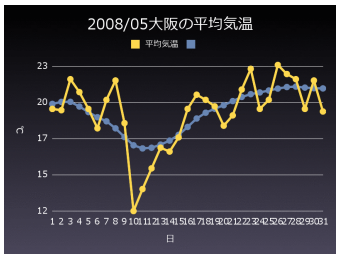
提案した重み付けの手法によって得られた重みについて、外れ値や変化点、グラフ特徴などの観点の下、評価を行った。図 5 から図 7 に各時系列データによって得られた重みを示す。

その結果、上昇・下降の傾斜傾向に対する重み付けに関して、(1) の傾斜傾向の標本数が多数で傾斜が特徴的な場合の手法では、傾斜傾向の標本数が最多であっても標本数に大きな差がない場合、傾斜の大きい傾向が優先された。しかし、傾斜傾向の標本数に大きな差



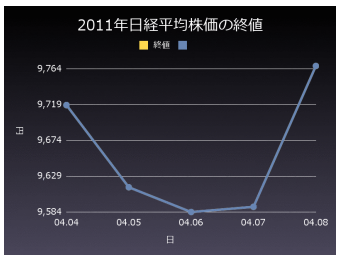
種類	内容							
文字情報	2000 年 大阪の月別累計降水量 (1 月〜8 月抜粋)							
数値情報	52.0	36.5	94.5	77.5	90.5	198.0	42.5	6.5
傾斜傾向	下降	上昇	下降	安定	上昇	下降	下降	
手法 (1)	7.27	10.91	14.55	3.64	15.58	12.99	2.60	
手法 (2)	2.27	19.09	4.55	6.36	6.06	10.82	1.01	
手法 (3)	19.75	50.58	37.33	0.61	109.22	149.30	43.06	

図 5: 年単位のデータ　サンプル数が 12 の場合 (1 月〜8 月抜粋)



種類	内容							
文字情報	2008 年 5 月 大阪の平均気温 (1 日〜8 日抜粋)							
数値情報	20.2	20.1	22.5	21.5	20.2	18.7	20.9	22.4
傾斜傾向	安定	上昇	下降	下降	下降	上昇	上昇	
手法 (1)	5.83	4.58	2.08	3.75	2.50	2.50	1.25	
手法 (2)	2.03	4.58	0.73	3.75	0.87	2.50	0.44	
手法 (3)	0.26	2.40	0.64	0.87	1.10	2.51	2.04	

図 6: 月単位のデータ　サンプル数が 30 の場合 (1 日〜8 日抜粋)



種類	内容				
文字情報	2011 年 4/4 から 4/8 までの日経平均株価 (終値)				
数値情報	9718.89	9615.55	9584.37	9590.93	9768.08
傾斜傾向	下降	下降	安定	上昇	
手法 (1)	25.0	16.67	33.33	25.00	
手法 (2)	8.33	7.14	33.33	8.33	
手法 (3)	0.00	0.00	0.00	0.00	

図 7: 週単位のデータ　サンプル数が 5 の場合

が見られた場合、傾斜の大きさに関わらず重要視されるため、傾斜が小さくても標本数が多いものが重要視された。(2) の傾斜傾向の標本数が少数で傾斜が特徴的な場合の手法では、傾斜傾向の標本数が最少となる場合、今回の対象データでは安定の傾斜傾向が最少となる場合が 9 割を占めたが、(3) の手法と比較した場合に重くすることが可能となった。この手法では、標本数の多いものが極端に軽視されるなどの問題も生じたため、係数の算出手法について検討する必要性がある。

(3) の全体的な傾向 (トレンド) との乖離が著しい場合の手法では、時系列データの傾斜とトレンドの傾斜との差をそのまま重みに用いているため、グラフ特徴と外れ値という点では最も基準に合致していた。しかし、 (3) の単純移動平均を用いたトレンドの算出に関して、月別降水量と平均気温のデータについては平均を取る幅を データ数/5 にすることで効果的な値が得られたが、日経平均株価のデータでは、サンプル数が少なかったため効果的な値は得られなかったという問題点も見られた。

以上のことから、上昇・下降に対する重み付けに関しては本研究における観点から見た場合、 (3) の手法が最も適していると判断した。また、安定に対する重み付けに関しては、 (1) 傾斜傾向の標本数が多数で傾斜が特徴的な場合と (2) 傾斜傾向の標本数が少数で傾斜が特徴的な場合について検証を行ったが、今回対象としたデータを用いた場合、安定の標本数が上昇や下降の標本数に比べて圧倒的に少なく、手法の違いによる有意差は見られなかった。このことから、上昇・下降に関しては (3) の手法による重み付けは有益であるが、サンプルの取り方と安定に対する基準の設定を考慮する必要性が明らかになった。さらに、安定に関する基準を設定していなかったため、安定に対する重み付けが低くなってしまうことも問題として挙げられる。また、単一の手法による重み付けではなく、複数の手法を併用することでより観点に沿った重み付けが可能だと考えられる。以上のような検証の結果に基づき、検索の前段階にあたる特徴の付与と重み付けを行うプログラムについて実装を行った。実行結果の例を図 8 に示す。

本システムの動作環境について、OS は Microsoft

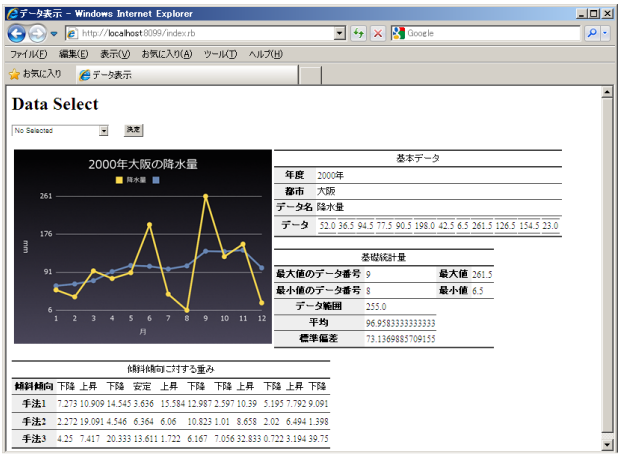


図 8: 実行結果の表示例

Windows XP Home Edition を用い、Web ブラウザは Internet Explorer 8 を用いた。プログラミング言語は Ruby1.8.7 を用いた。プロトタイプシステムの実装に際して対象としたデータは、3 章で提示したデータを用いた。

5 考察

本研究では単一の時系列データから得られる特徴を基に、その特徴に対して重み付けを行い、ユーザの要求に応じた複数の時系列データを比較した場合に要求との合致度を考慮し重みの変更を行うことで、相対的な重み付けを行う手法について提案した。

そのうち、提案手法で述べた 3 つの観点に基づき単一の時系列データから特徴を算出し、算出した特徴に対して傾きに関する重み付けの段階を対象として検証を行った。今後、提案手法で述べた期間に関する重み付けを行った場合、それぞれの重み付け手法で得られた値の取り扱いについて、単一で取り扱い 2 種の重みとすべきか、複合して 1 種の重みとすべきか検討を行う必要性がある。それぞれ独立の重みとして取り扱う場合、ユーザの要求が期間に関するものであれば期間の重みに焦点をあてることで、容易に特徴を捉えることが可能になり、要求に応じた判断が可能になると考えられる。しかし、ユーザの要求が傾斜と期間を複合したものであった場合には、それぞれ独立で重み付けられた傾斜と期間に関連性を持たせる必要がある。上記の考察を踏まえた上で、今後期間に関する重み付けの検証を行った後、検討を行う必要性がある。

今後、現段階で実装を行った傾斜傾向に対する重み付けに際する問題に関して、検証の結果でも述べたように取得するサンプル数に対する制約や重み付けの手法について検討を行い、更なる検証を行う必要性がある。加えて、重み付けの対象に期間が追加された場合、提示する情報について考慮する必要がある。また、重み付けの観点に対する整理と重み付けや提示する情報に関する被験者実験を実施し、人間の認識により合致した重み付けの値や効果的な情報提示について知見を得る必要があると考えている。

6 おわりに

本研究では、複数の時系列データの比較により変化する解釈を取り扱うため、時系列データに対する動的な重み付け手法の枠組みを提案した。提案した手法における 3 段階のうち、検索の前段階として必要である時系列データに対する特徴の付与と特徴に対する重み付けに関して検証を行い、検証により得た知見を踏まえ実装を行った。今後、本研究における検証や実装により明らかとなった問題点について、更なる検証と重み付け手法の再考や提示手法の改善を行う必要性がある。また、複数の時系列データを比較し重みを変化させることで、ユーザの要求に応じたグラフ解釈の変化に対応した方式への拡張について検討する。

7 謝辞

本研究は科学研究費補助金基盤研究（C）（課題番号:22500209）の助成を受けた。記して謝意を表す。

参考文献

[1] 松下光範, 末吉れいら: 言語表現による時系列データ検索のための基礎検討, 第 19 回 Web インテリジェンスとインタラクション研究会, pp. 31–32 (2011).

[2] 末吉れいら, 田中和広, 白水菜々重, 松下光範: 比較対象に着目したグラフの言語表現の生成, 第 21 回 Web インテリジェンスとインタラクション研究会, pp. 37–38 (2011).

[3] 小泉尚之, 松下光範, 松田昌史, 馬野元秀: 言語情報と統計グラフの相互変換に関する基礎検討, 人工知能学会全国大会, 2H5-6 (2007).

[4] 熊原啓作, 渡辺美智子: 身近な統計, 放送大学教育振興会 (2007).

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第1回)
SIG-AM-01-04

広電沿線観光情報提示システムの構築

Construction of a System for Providing Travel Information along Hiroden Streetcar Lines

石野亜耶[1]　難波英嗣[1]　竹澤寿幸[1]

Aya Ishino[1], Hidetsugu Nanaba[1], and Toshiyuki Takezawa[1]

[1]広島市立大学大学院　情報科学研究科

[1] Graduate School of Information Sciences, Hiroshima City University

Abstract: In this paper, we propose a method for identifying Hiroshima Electric Railway (Hiroden) blogs in a blog database. Hiroden blogs are defined as travel journals that provide regional information along Hiroden streetcar stations. To investigate the effectiveness of our method, we conducted some experiments. From the experimental results, we obtained precision of 82.4% and recall of 64.5% in automatic identification of Hiroden blogs.

1. はじめに

2007 年 1 月に「観光立国推進基本法」が施行され、2008 年 10 月には国土交通省の外局として観光庁が設置されるなど、日本では「観光」を 21 世紀の基幹産業と位置付け、観光を支援する多様な取り組みが積極的に推進されている。現在、広島県では、NHK 大河ドラマ「平清盛」に関するイベントが行われている。また、2013 年 4 月からは、日本最大の菓子業界の展示会である全国菓子大博覧会(ひろしま菓子博 2013)が開催されるなど、観光客を集める様々な取り組みが行われている。そこで、本研究では、広島の特徴のひとつである、広島電鉄の電車（広電）を使用した観光を支援するための枠組みの一つとして、広電の電停に関する旅行ブログ（電停ブログ）を収集し、路線図にマッピングし旅行者に提示する広電沿線観光情報提示システムの構築を行う。広電沿線観光情報システムを作成することで、ガイドブックに掲載されていない、地域に基づいた情報を発信することができると考えられる。また、近年ブログ著者の属性（性別、年齢、居住域など）を文体や記載内容から自動的に推定する研究が進んでいる[1、2、3]。このような技術を利用し、ブログ著者の属性と、システムの利用者の属性を照らし合わせることで、例えば「女性に人気のレストラン」や「若い人に人気の観光名所」など、利用者に適した観光情報を推薦することができると考えられる。

本論文の構成は以下の通りである。2 節ではシステム動作例、3 節では関連研究、4 節では提案手法、5 節では実験結果と考察について述べ、6 節で本稿をまとめる。

2. システム動作例

本研究で構築した広電沿線観光情報提示システムについて、その動作例を紹介する。図 1 は、広電沿線観光情報提示システムを、iPad 上で動作させたときの画像である。図 2 は、広電沿線観光情報提示システムの画面である。広電の電停および主要な観光名所が描かれている。図 2 の路線図の一部をクリックすると、拡大路線図を表示することができる。図 3 は、図 2 の紙屋町エリア（図中①）をクリックした際の拡大路線図である。図 3 の電停をクリックすると、その電停に関する電停ブログのリンク集を閲覧することができる。図 4 は、図 3 の“原爆ドーム前”という電停をクリックした際に、閲覧することができる電停ブログのリンク集である。本研究では、広電沿線観光情報提示システムで提示する電停ブログを収集する手法を提案する。

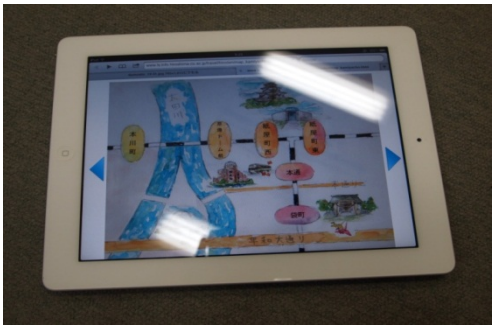


図 1: 広電沿線観光情報提示システムの動作例

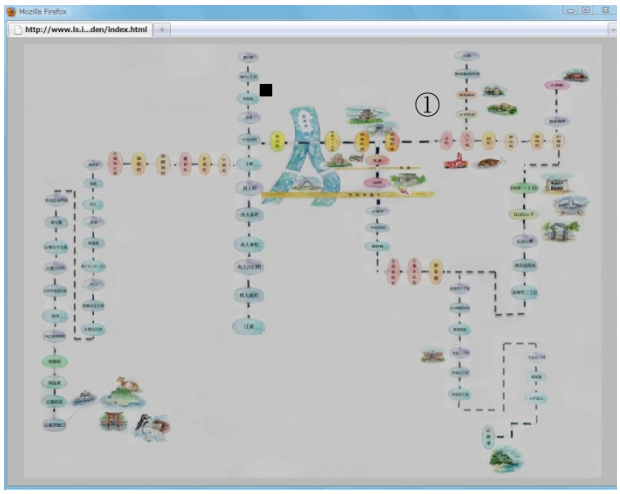


図 2: 広電沿線観光情報提示システムの路線図

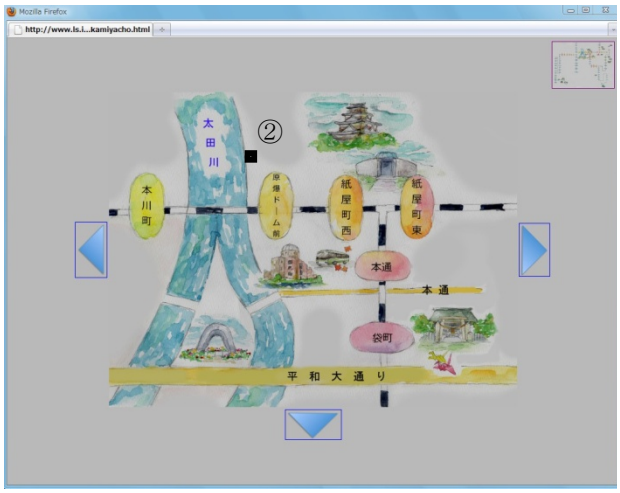


図 3: 紙屋町エリアの拡大路線図

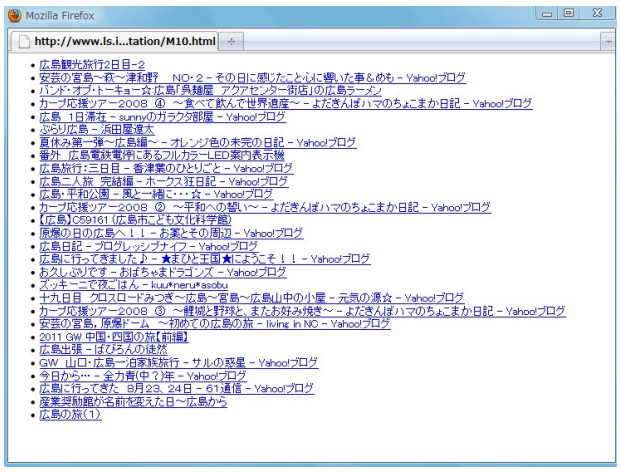


図 4: 電停ブログのリンク集

3. 関連研究

本研究では、広電を使用した観光を支援するための枠組みの一つとして、広電の電停に関する旅行ブログ（電停ブログ）を Web から収集する手法を提案する。本研究と同様に、Web から地域情報を自動収集する研究がある。大槻ら[4]は、地域情報ウェブディレクトリを自動編集するシステムを提案している。地域情報ウェブディレクトリは地域情報検索に利用される。大槻らは、地域情報として自治体が提供する地域情報サイトと、そのリンク先の地域サイトを対象としているが、本研究では、ブログを対象としている点で異なる。

本研究と同様に、ブログを情報源とし、地域情報を自動抽出する研究がある。岡本ら[5]は、一般のブログ検索エンジンを利用することで、地名を含むブログエントリを収集し、それらのブログエントリから、地域イベント情報を抽出する手法を提案している。本研究では、電停ブログを収集することを目的としているため、岡本らの研究とは異なる。

Web から観光情報を収集する研究として徳久ら[6]の研究がある。徳久らは、ブログから、観光開発のヒントとなる文を抽出する手法を提案している。石野ら[7]は、ブログデータベースから、機械学習を用いて旅行ブログを検出する手法を提案している。石野らは、“旅行”、“観光”、“ツアー”などの旅行ブログによく出現する単語の有無を素性に使用している。石野らは旅行ブログの収集を目的としているが、本研究では、電停ブログの収集を目的としている点で異なる。

旅行ブログやそのエントリを登録したポータルサイトとしては、“Travel Blog”[1]、“旅行・観光ブログ村”[2]、“フォートラベル”[3]などがある。これらのポータルサイトでは、ブロガーが自身のブログを旅行ブログとして登録することで、旅行ブログの集積を行う。しかし、ブログ空間にはたくさんのブログが存在しており、このようなポータルサイトに登録されていない一般ブログの中にも電停ブログが多数存在する。電停ブログのように、ある地域に限定したブログは、有名観光地に関連するブログと比較すると、ブログの件数が少ないと考えられる。よって、本研究では、一般ブログを対象として、電停ブログかどうかを自動判定することで、より多くの電停ブログの収集を行う。

郡ら[8]は、ブログからユーザの行動時の代表的な

[1] http://www.travelblog.org/
[2] http://travel.blogmura.com/
[3] http://4travel.jp/

経路とその文脈を抽出し、それらを地図上にマッピングすることにより、集約して提示するシステムを提案している。また、Davidov[9]は、Web から交通手段や経路の地理的なネットワークを見つける手法を提案している。Ishino ら[10]は、旅行ブログから、機械学習を用いて、旅行者の行動経路を抽出する手法を提案している。これらの研究と、本研究で構築した広電沿線観光情報提示システムを組み合わせることで、旅行者に最適なモデルルートを推薦することができると考えられる。

寺西ら[11]は、観光情報雑誌に、旅行ブログや、Yahoo!知恵袋付けることで、より網羅性の高い観光情報を提示する手法を提案している。本研究では、電停ブログを収集し路線図にマッピングし提示するシステムを構築することを目的としている。今後は、Yahoo!知恵袋など他のコンテンツも、広電沿線観光情報提示システムで提示できるよう改良していく予定である。

4. 広電沿線観光情報提示システムの構築

本節では、広電沿線観光情報提示システムで提示する電停ブログの収集手法について説明を行う。電停ブログの収集手法は、以下の 2 つのステップに分かれている。(1)については 4.1 節、(2)については 4.2 節で説明を行う。

(1) ブログの収集
(2) 電停ブログの判定

4.1. ブログの収集

電停ブログの収集のためには、各電停に関連する情報が記載されたブログが必要になる。そこで本研究では、各電停の名称(78 件)をクエリとして Yahoo! 検索（ブログ）で検索を行い、ブログの収集を行った。その結果、1,748 件のブログが収集された。

4.2. 電停ブログの判定

本研究では、広電の電停に関するブログや、ブログ著者が広電の電停で下車、観光を行ったブログを電停ブログと定義する。

4.1 節で収集したブログには、広島電鉄の電停と同一の地名や他の交通機関の駅名に関する情報が記載されているブログや、観光に関連しないブログが含まれる。本研究では、4.1 節で収集されたブログに対し、電停ブログかどうかを、機械学習を用いて自動判定する。

図 5 は、人手で電停ブログであると判定されたブログの一例である。図 5 に示すブログのように、電停ブログには、広電の電停名や、“市電”、“電停”などの広電に関連する単語が頻出する傾向がある。ブログ著者が観光を行った際には、“観光”、“散策”などの単語がよく使われる。また、撮影した写真を掲載する傾向がある。よって本研究では、機械学習に以下の素性を使用することで、電停ブログの自動判定を行う。

- 電停名の出現頻度 （78 件）
- 広島電鉄関連の単語（広島電鉄、広電、市電、電停など）の出現頻度 （5 件）
- 広島電鉄の電停に関連しない単語（JR、新幹線、フェリーなど）の出現頻度 （6 件）
- 旅行関連の単語（観光、散策、撮影など）の出現頻度 （9 件）
- 写真の有無

ココに車を停めて後の移動は市内電車**広電**
市電と宮島行の船が一日中乗り放題の一日乗車乗船券を買ってまずは**原爆ドーム前**へ
昨今の原発問題もあって、そこらじゅうで署名活動してました
・・・・・・・・・・(略)・・・・・・・・・・
周辺のテキヤを満喫し**広電**で宮島口まで移動

図 5: 電停ブログの一例

5. 実験

本研究で行った実験について説明する。

データセット

実験用データには、4.1 節で収集した Yahoo ブログ 1,748 件に対し、人手で電停ブログかどうかの判定を行った結果を用いる。人手で電停ブログの判定を行った結果を表 1 に示す。

表 1：電停ブログの人手での判定結果

電停ブログ（件）	その他（件）	合計（件）
568	1,180	1,748

比較手法

提案手法の有効性を確認するため、4.1 節で収集したブログ 1,748 件を、全て電停ブログとして判定した場合を比較実験とした。

機械学習と評価尺度

電停ブログの判定の機械学習には TinySVM を用いた。2 次の多項式カーネルを使用し、4 分割交差検

定を行った。評価尺度として、精度・再現率・F 値を用いた。

実験結果と考察

実験結果を表 2 に示す。表 2 の実験結果より、比較手法に比べ、提案手法が精度・F 値ともに高い数値を記録した。よって提案手法の有効性を示せたといえる。

表 2: 電停ブログの自動判定結果

	精度(%)	再現率(%)	F 値(%)
提案手法	82.4	64.5	72.4
比較手法	32.5	100.0	49.1

提案手法では、精度に比べ、再現率が低い結果であった。本論文では、再現率低下の原因について考察を行う。

再現率の低下の原因は、手掛かり語の不足であった。本研究では、電停ブログの判定を、4.2 節で示した手掛かり語を用いて機械学習により行った。使用した手掛かり語は、広電関連の単語が大部分を占めている。しかし、電停ブログには、電停で下車した後、観光や、食事した状況が詳しく記述される場合がある。この場合、本研究で使用した、広電関連の手掛かり語が、あまり出現しない傾向がある。この問題を解決するためには、収集した電停ブログを解析し、各電停の電停ブログによく出現する観光名所や、レストラン名、土産物の名前などを収集し、手掛かり語として追加することが考えられる。

6. まとめ

本研究では、電停ブログを収集する手法を提案した。電停ブログの収集手法は、(1)ブログの収集、(2)電停ブログの判定の 2 つのステップに分かれる。電停ブログの判定では、精度 82.4%、再現率 64.5%を得られており、提案手法の有効性を示すことができた。また、収集した電停ブログを路線図にマッピングし、旅行者に提示する広電沿線観光情報提示システムの構築を行った。

今後の課題としては、収集した電停ブログを、“観光”や“食事”などの観点で分類し、旅行者が電停ブログを効率的に閲覧することができるようにすることが挙げられる。また、電停に関連する Yahoo!知恵袋や、ニュースなど、様々なコンテンツを自動で収集することで、より網羅性の高い広電沿線観光情報提示システムを構築することが考えられる。

参考文献

[1] Yasuda, N., Hirao, T., Suzuki, J., and Isozaki, H.: Identifying Bloggers' Residential Areas, Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp.231-236, (2006)

[2] Ikeda, D., Takamura, H., and Okumura, M.: Semi-supervised Learning for Blog Classification, Proceedings of the 23rd AAAI Conference on Artificial Intelligence, pp.1156-1161, (2008)

[3] Schler, J., Koppel, M., Argamon, S., and Pennebaker, J.: Effects of Age and Gender on Blogging, Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs, pp.199-205, (2006)

[4] 大槻 洋輔, 佐藤 理史: 地域情報ウェブディレクトリの自動編集, 情報処理学会論文誌, Vol.42, No.9, pp.2310-2318, (2001)

[5] 岡本 昌之, 菊池 匡晃: ブログからの地域イベント情報抽出, 情報処理, Vol.51, No. 1, pp.14-17, (2010)

[6] 徳久 雅人, 奥村 秀人, 村田 真樹: 観光開発支援のためのブログ記事からの評判分析, 観光と情報, Vol.7, No.1, pp.85-98, (2011)

[7] 石野 亜耶, 難波 英嗣, 竹澤 寿幸: 旅行ブログからの観光情報の自動抽出, 日本知能情報ファジィ学会誌, Vol.22, No.6, pp.667-679, (2010)

[8] 郡 宏志, 服部 峻, 手塚 太郎, 田島 敬史, 田中 克己: ブログからのビジターの代表的な経路とそのコンテキスト抽出, 情報処理学会研究報告データベースシステム研究会, Vol.2006, No.78, pp.35-42, (2006)

[9] Davidov, D.: Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.267-175, (2009)

[10] Ishino, A., Nanba, H. and Takezawa, T.: Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries, Proceedings of the 18th international Conference on Information Technology and Travel & Tourism (ENTER2011), (2011)

[11] 寺西 拓也, 野村 達二, 平山 智子, 石野 亜耶, 難波 英嗣, 竹澤 寿幸: 観光ガイドブックへの旅行ブログエントリと質問応答コンテンツの対応付け, 言語処理学会 第 18 回年次大会, (2012)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第1回)
SIG-AM-01-05

電子カルテテキストデータに関する一考察

A Consideration of Text Data Within Electronic Medical Records

串間宗夫　田之上光一　荒木賢二　鈴木斎王　荒木早苗　仁鎌照絵　山崎友義

Muneo Kushima, Koichi Tanoue, Kenji Araki, Muneou Suzuki, Sanae Araki, Terue Nikama and Tomoyoshi Yamazaki

宮崎大学医学部附属病院医療情報部
Section on Medical Information, Faculty of Medicine, University of Miyazaki

Abstract: An Electronic Medical Record (EMR) records information on patients by computers instead of by paper. Many medical documents, including EMRs that describe the treatment information of patients, are text information. Such text information is complicated. The data arrangement and retrieval of such text parts become difficult because they are often described in a free format; the words, phrases, and expressions are too subjective and reflect each writer. In the present study, we considered the text data of the nursing record within Electronic Medical Records of the University of Miyazaki Hospital.

1. はじめに

昭和 63 年 5 月 6 日付けの厚生省通知により OA 機器による診療録の作成が認められた[1]。その後、平成 6 年 3 月 29 日付けの厚生省通知でエックス線写真等の電子保存が認められ[2]、医療情報を電子媒体に保存することが容認された[3]。平成 11 年 4 月 22 日に各都道府県知事宛に「診療録等の電子媒体による保存について」の通知がされ、初めていわゆる電子カルテの運用が可能になった。電子カルテは紙カルテに比べより多くの者に閲覧される可能性が高く、プライバシー保護が適切に行われているかどうかが大切である[4]。電子媒体による保存では、真正性の確保、見読性の確保、保存性の確保の 3 条件を満たす必要がある。医療現場における IT 化が進展し、電子カルテの導入を図る病院が増加している[5]。医療現場で導入されている医療情報システムのうち、電子カルテに蓄積されているデータのほとんどは構造化されていないテキストデータであると言われており、テキストデータから医学的に興味深く重要な知識を発見することは大切なことである。電子カルテとはカルテ(診療録)に書いていた診療情報を、電子的に記録し保存したものである。また、病院情報システムは、診療に関する正確で伝達可能な情報を集めることができ、病院全体の診療業務の効率化に関わるリスクマネージメントを推進する上で有用なツールである[6]。病院情報システムによって集められた情報は院内の医療従事者間だけでなく、病診連携など地域の医療従事者間との医療ネットワークにおいても広範囲の活用が期待される[7]。大学医学部附属病院としては、比較的早く電子カルテを導入しその運用を開始した。本院の病院情報システムは、厚生労働省の電子的保存条件を満たしており、医師や看護師、各部門のカルテ記載を電子的に行い、CT 画像などを各端末から参照できるほか、地域との医療情報連携機能や経営分析機能などを有してする。

2. 病院情報システム（電子カルテを含む）

2.1 特徴

特徴として、以下のようにまとめられる。

(1) オーダリング・医事会計システムと連動した診療業務の効率化(ペーパーレス, 診療情報の多目的利用)が図れる。

(2) 患者情報や医学知識が共有でき、チーム医療を支援することにより、医療の質を向上させることが可能である。

(3) 病院情報システムに蓄積されたデータをデータベース化し病院管理、経営分析、疫学などに後利用が可能である。

(4) 複数の医療機関をネットワークでつなぐことにより、地域、全国間で情報が共有化でき、患者への情報開示の手段などに有用である。

(5) 病院情報システムは診療科だけでなく医事課、薬剤部、看護部、放射線部、検査部など各部門から入力された情報が収められ、その情報はリアルタイムにどの端末からも参照できる。

図 1(a)(b)に紙カルテと電子カルテのイメージ図を示す。

2.2 電子カルテ稼働後の状況

2006 年 5 月からの本附属病院電子カルテ稼働後の状況について述べる。

1) 2006年5月より、ペーパーレス、フィルムレスによる診療業務が定着した。

・ ナースステーションの資料等が整備された。

・ 2010 年 5 月からの新外来棟では、古い紙カルテ貸出が原則禁止となった。

・ 重いフィルムを移動させる必要がなくなり、研修医の業務が軽減された。

・ 従来使用していた紙のカルテを知らない医師が現れた。

2) 稼働後の要望は約 2000 あった。

・ 要望に対して順次対応した。

・ 持続的な開発、バージョンアップが、利用者の満足度を大きく向上させている。

2) 特に喜ばれている機能

(運用に乗っていることが最大の成果である)。

・ ベッドサイド携帯端末(看護師)。

・ SBC(メタフレーム)により個人の端末でフル機能が使える(医師)。

・ さまざまな文書(診断書等)を、気軽に電子カルテに載せることができる(医師)。

2.3 宮崎大学医学部附属病院

宮崎大学医学部附属病院では、比較的早く病院情報システム(IZANAMI)を導入し、電子カルテの運用も早期に開始している。医師や看護師、各部門のカルテ記載を電子的に行い、CT などの画像も各端末から参照できる他、「はにわネット」と呼ぶ地域との医療情報連携機能や経営分析機能を有する。

宮崎大学医学部附属病院は、以下の経緯で病院情報システムを開発している。

- 平成 13 年 はにわネット運用開始
- 平成 14 年 はにわネット理事会「中核病院への普及の必要性」
- 平成 15 年 IZANAMI 開発開始
- 平成 18 年 5 月 宮崎大学病院 IZANAMI 稼働
- 平成 18 年秋 IZANAMI パッケージ版普及開始
- 平成 19 年 3 月 IZANAMI 診療所版稼働
- 平成 19 年 8 月 IZANAMI 病院版稼働

平成 18 年 5 月の医療情報システム更新に際し、地元の IT 企業とのコラボレーションにより開発した病院情報システムを導入した。IZANAMI の表示画面を図 2 に示す。神話の国宮崎、伊邪那美命(いざなみのみこと)にちなんで名付けたられた IZANAMI は、他の多くの大学病院で稼動しているものとは異なる以下に述べるようなユニークな特徴を持っている。

(1) これまで大学病院クラスの病院情報システムは、全て大手の医療システムベンダーのものが中心であったが、IZANAMI は、地元企業とともに開発したもので極めて稀なケースと言える。

(2) パフォーマンス(画面が開く速さ)を徹底的に重視した点である。世界最速と謳われている Cache'というデータベースを採用することにより、仮想的に 5 年分の患者情報を作成し 100 台の端末から同時にアクセスを行い、基本的な画面展開を試験し、ほぼ 3 秒以内のパフォーマンスを可能にした。

(3) 法人化後の大学病院に強く求められている経営改善に真に役立つシステムを目指したことである。さらに、

- クリニカルパス(スケジュール表)
- 自動作成機能
- バリアンスのリアルタイム自動集計

を持たせた。また、宮崎の地域連携プロジェクトである「はにわネット」に完全に対応し、患者数増や在院日数短縮に不可欠な地域連携機能を高いレベルで実現した。

現在、稼動して約5年数か月になるが、新開発システムゆえに危惧された大きな障害や混乱はなく、医師をはじめとする現場の医療従事者からも高い評価を得るとともに、院外からも多くの見学者が訪れている。今年度、同じ電子カルテシステムが久留米大学病院に導入される。

3. 電子カルテテキストデータ

3.1 看護記録

看護記録とは、看護実践の一連の過程を記述したものであり、医療の担い手である看護師が記載した看護記録は、診療記録の一部に含まれるものである。

日本看護協会「看護記録および診療情報の取り扱いに関する指針」によると、看護記録の特徴は、

- 必要なことは漏れなく記述する。
- 必要でないことは一つも書かない。
- 無防備な看護記録の現実を改める。
  ○個人的感情の記載

  ○感想、憶測, 個人的見解

- [大原則]重大医療事故発生時には、記録方式を経時的記録に変える。

と述べてある。

更に、具体的には、看護記録は、生活歴や検査歴、更に、予約などのちょっとしたメモなどにも使用されている。実際にはテキストにはこういうことを記録するというルールがないので、あやふやな感じを印象として持つのが現状である。看護師は、患者が述べた言葉を覚え、更にメモを取り、最後にまとめて電子カルテに入力している。以下の S.O.A.P として、実際の患者の状況が記録されている。

- S=患者が直接提供する主観的な状況・患者が話した内容。
- O=客観的事実・医療スタッフの目から見た患者の様子や認識の状態。

- A=それらの情報から導き出される評価・判断。
- P=今後の計画・実際に行ったケアである。

患者・電子カルテ利用者に焦点を当てたコラム形式の記録方法であり、また、医療安全上で経時記録方法としてのフォーカスチャーティングという書き方もある。図 3 に看護記録について示す。

3.2 経過記録

経過記録とは、医療を必要とする人の問題の経過や治療・処置・ケア・看護実践とその結果を記載したものである。また、経過記録は、医師が記入したものであり、主観的・客観的・対応・計画等を示したものである。具体的な特徴としては、

- 診察したときの状態
- 生活歴、既往症、検査歴等
- 患者や家族に説明した内容等
- メモ・出来事
- S.O.A.P として入力されている

医師が初診以降に記録したものを一般に経過記録と呼ぶ。紙カルテでは左側に S、O、A が書かれ、プランとして実際にオーダされた薬剤や処置などの医療行為が右側に書かれる。図 4 に医師経過記録について示す。

3.3 フローシート

記録における疾患特有の観察所見は、フローシートに登録して所見を入力している。フローシートの特徴としては、

- 体温や血圧などのバイタルサイン、疼痛や出血・しびれなどの観察項目、化学療法 の副作用チェック、教育や指導の実施。
- ベッドサイド端末(スマートフォン)からも入力可能。
- 記録を時系列に表示したもの(約 5800 項目)である。

宮崎大学では観察所見、診察所見、看護処置等のうち、患者ごとにフローシートと呼ばれる記録欄に記録している。S.O.A.P では O の部分であり、これをベッドサイドで入力することにより、記録の転記などの労力を低減するようにしている。フローシートでは項目(コード化されている)と値が対になって入力されるシステムとなっており、電子カルテのテキストデータからは分離されている。フローシートは一般的な用語とは言えない。図 5 にフローシートについて示す。

4. 電子カルテ記録

4.1 具体的にどのように入力しているのか

入力用画面と、閲覧用画面が異なっており、入力用画面では、項目と入力カラムがあり、項目(コード化はされていない)ごとの入力になっている。項目により入力様式は、

- チェック
- 選択
- 自由入力
- データ自動取得
- シェーマ(画像等)

の設定ができる。項目と値がセットになっているため、データとして活用しやすい。ただし、自由入力やシェーマは活用することに問題がある。

4.2 入力についての問題点

- 見た目に合わせた入力ではない。
- 文字の強調やフォントの変更ができない。
- キーボード入力に慣れないスタッフがいる。
- 正確な入力がなされていない。

→データとして活用できない。

4.3 どのような手だてがあれば効率よくなるか

すべての文章では不可能であるが、一部の文章では専用の入力画面を作るべきである。特に、タブレット端末を利用することで、効率化できるところは利用したい。手術や検査の申し送り票などは良い対象である。

4.4 入力側(臨床現場)としては

- 臨機応変に入力することができるので、仕事がしやすい。
- 曖昧な表現が可能。

4.5 活用側(研究者・事務)としては

- スペースや半角全角等が混ざり、データがとりにくい。
- 文字量も多くなり時間がかかる。(手作業になりやすい)
- 統一性がないため比較・分類がしにくい。

4.6 テキスト入力で行うこと

- ケアを行う前と行ったケアを記録する前に、他のケア提供者が何を書いてあるかよく読む。
- 問題点として挙げられたものがケアされずに放置されていないかどうか確認する。
- ケアを行った後は出来るだけ早い時点で記録するようにする。
- 患者の行動や言葉を直接引用し、患者に何が起こったか、どのようなケアを誰がいつ実施したのか、また、その反応等の事実を正しく記録する必要に応じて、絵や写真を張るようにして具体的に示すようにする。
- 読みやすいように書く、決められた記録の様式で記入する。
- 略語を用いるときは、各施設のマニュアルに記載され認められている略語のみを用いる。
- 記載していない場所がないか確認する。

4.7 テキスト入力での禁止事項

- 事前にこれから行う処置やケアを記入しない。
- 自分が実際見ていない患者の記録はしない。

- 意味のない語句や、患者のケアおよび観察に関係のない攻撃的な表現をしない。
- 患者にレッテルをはったり、偏見による内容を記録してはならない。
- 「～と思われる」「～のように見える」といった曖昧な表現はしない。
- 施設で認められていない略語を使用しない。
- イニシャルや簡略化した著名は用いない。

テキスト入力は、入力側にとってはメリットが多いが、活用側にとっては問題なことが多い。上手く整理して決めた値で入力をできるか。システムの機能を充実させ、欲しい部分だけ抽出できるか、等が今後の課題である。どの大学病院でも、テキストデータをどのようにとるか悩ましいところである。一般的に、なるべく統一してテキストデータ入力を行いたいと考えている。

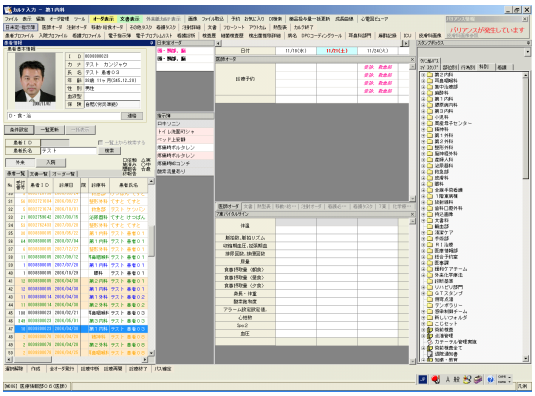
5. まとめ

宮崎大学医学部附属病院で稼働している電子カルテの医療記録であるテキストデータについて、具体的にどのように入力しているのか、入力についての問題点、どのような手立てがあれば入力の効率がよくなるのかについて考察した。

参考文献

[1] 厚生省: 診療録の記載方法について, 昭和 63 年 5 月 6 日.
[2] 厚生省: エックス線写真等の光磁気ディスク等への保存について, 平成 6 年 3 月 29 日.
[3] 厚生省: 診療録等の電子媒体による保存について, 1999 年 4 月 22 日.
[4] 鈴木斎王、荒木賢二、吉原博幸: 総合医療情報システムの監査実施経験, 医療情報学, 22(4)、pp.347-353、2002.
[5] 紀ノ定保臣: 電子カルテ時代の医療情報学, 医療情報学, 23(5) , pp.397-405、2003.
[6] 松村泰志: 電子カルテと病院情報システム-診療情報の包括的管理と利用-, 医療情報学, 21(3) , pp.211-222、2001.
[7] 松村泰志、中野裕彦、楠岡英雄、その他: ネットワーク型電子カルテによる病院・診療所連携情報システム, 医療情報学, 22(1) , pp.19-26、2002.

(a) 紙カルテ



(b) 電子カルテ起動画面

図 1 紙カルテと電子カルテ

図 2 電子カルテトップ表示画面

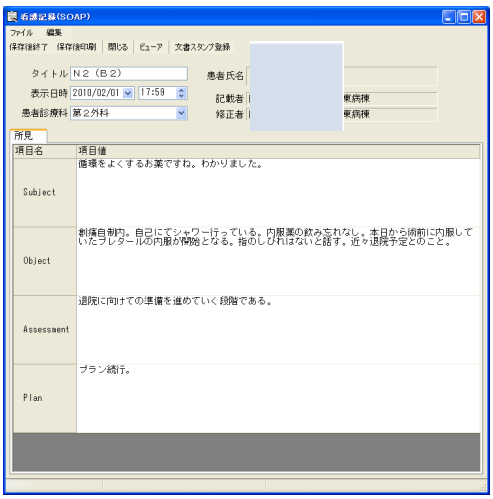


図 3　看護記録

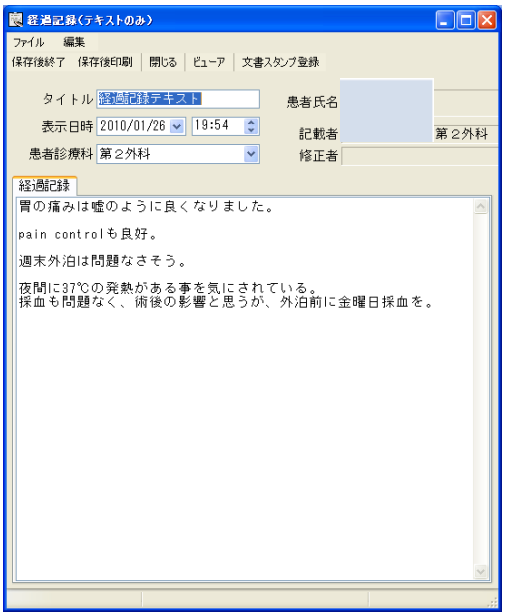
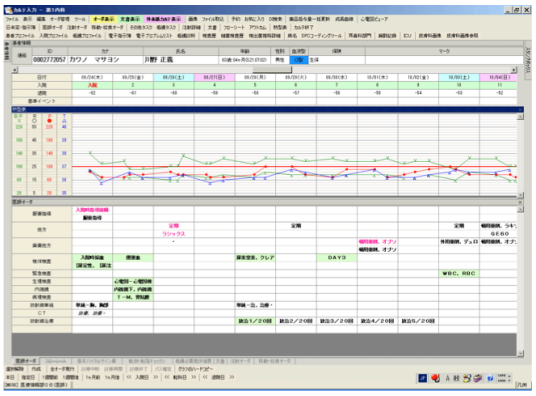


図 4　医師経過記録



(a)　オーバービュー

(b)　Android の入力画面

図 5　フローシート