2B1-R-3-6

背景知識を利用したデータマイニング

On datamining with Relating Background Information

杉村　博[*1]　　松本　一教[*1]
Hiroshi SUGIMURA　Kazunori MATSUMOTO

[*1]神奈川工科大学大学院　工学研究科　情報工学専攻
Graduate School of Engineering, Kanagawa Institute of Technology

This paper shows the effectiveness of automated information, which is collected over the Web, in the process of time-series datamining. The system proposed in this paper consists mainly of two mechanisms. First, we search the Web space for relating information, which is expressed as keywords, to the current mining goal, and convert them into standard tags using a predefined dictionary. Second, association rules are extracted using both the collected tags and time-series data. We introduce the concept of maximum allowance length to discard noise and to focus only on promising information.

1. はじめに

本論文では数値のシーケンスである時系列データ解析のために，メタデータを組み合わせて知識獲得を行うための手法を提案する．メタデータを有効活用するための研究は文書 [1]，ビデオ [2]，オーディオ [3]，医療電子カルテ [4] 等に対して行われているが，時系列データに対しての研究はまだ不十分である．データ記録に関わる背景情報を組み合わせることで，数値のシーケンスだけを取り扱う手法では得られなかった新たな知識が得られると考えられ，例えば株価チャートとニュース，血液中の成分量と投薬情報，電力消費とユーザの操作などといった知識獲得が想定できる．

本論文では二つの技術で構成される．一つ目は時系列データへの自動的なアノテーションである．人間がすべての背景情報をアノテーションするならばマイニング対象のデータサイズに依存して膨大な人的コストを必要とする．ソーシャル化による人的コストの分散化か機械による自動化が考えられるが，ソーシャル化は多人数特有の問題を新たに生み出してしまう．そこでマイニング対象を株価データとし，Web ニュースサイトからクローラを用いて獲得した情報をタグ化し，自動的にアノテーションを行う．

二つ目はアノテーションされた時系列データからの知識抽出方法である．一つのアノテーションは時系列データの一部分を指し示し，その時点におけるデータの背景情報を与える．この様なアノテーション時系列データから，知識獲得を行うためのアノテーション相関ルールについて提案する．

2. 自動的なアノテーション

図 1 は自動アノテーションシステムの概要を示している．Web 上の全てのデータを用いると処理時間やコストに対して良い情報が得られるとは限らない．そこで会社の PR 用の Web サイトやニュースサイトの URL を登録した URL 辞書を用いる，効率よく解析対象の株価に対しての情報を収集する．ニュースグループを表すタグとそのタグに関連付けられた単語を組にした単語辞書に基づいて，ニュースをタグにしてその発表時間とともにアノテーションデータベースに格納する．この様にして作成されたアノテーション時系列データの概要を図 2 に示す．

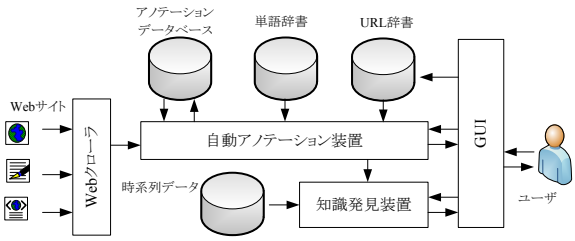

図 1: 自動アノテーション装置

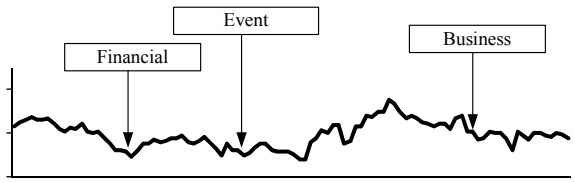

図 2: アノテーション時系列データ

タグ辞書はタグと重要な用語の組であり，一種のオントロジである．収集された Web ページは同等のニュースであっても表現や表記の揺れがあり，そのままアノテーションデータと用いても，類似ニュースを基にした株価データの動きを機械的に抽出することは難しいため，ニュースのグループ分けが必要となる．表 1 に示すタグ辞書を使用することで，Web ページの中の表記ゆれを抑える．有効な単語は，予備実験を行うことで選択した．

3. アノテーション相関ルール

システムは時系列データのアノテーションとその周辺の形状に基づく知識を抽出する．図 3 にアノテーション時系列データから抽出されるルールとその操作の概要を示す．

連絡先: 杉村 博，神奈川工科大学大学院工学研究科情報工学専攻，神奈川県厚木市下荻野 1030，TEL/FAX:046-291-3199，hiroshi.sugimura@gmail.com

表 1: タグ辞書

分類	単語
人事	人事，再編，就任，異動，改編，役員
開発	開発，実用化，低価格化，実現，受注，完成，完工，着工，営業
イベント	イベント，ショー，展示会
販売	販売，発売，提供
決算	決算，財務報告，株主総会，会計
報告書	実務報告，会社設立，経営計画，売り上げ，売上，設立，開発，報告

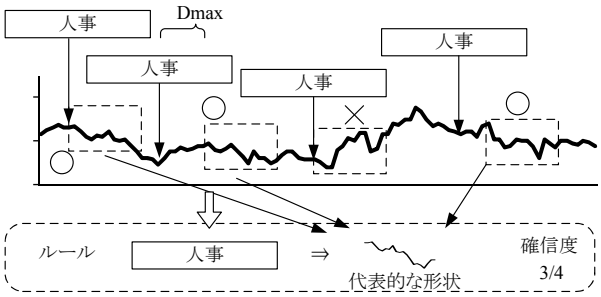

図 3: 概要

ニュース発生時間を基にして，決められた長さの部分時系列データを切り出す．切り出された時系列データはいくつかの代表的な形状にグループ分けされる．このグループ分けには距離関数を DTW 距離とした k-means によって行う．次に，ここで抽出した形状とアノテーションとの関係性について調査する．

形状とアノテーションの関係性は 2 つの閾値によって調査する．アノテーションと形状の出現位置が大きく離れている場合において，その二つの間には関係性が薄いとみなす．この関係性を調査するためにの新たな閾値，最大許容長 $D_{\max}$ を定義する．さらに，ルールが成立する確率について調査する．このために Apriori といった代表的な相関ルールマイニングの確信度の枠組みを拡張する．提案手法の確信度の求め方は，X の出現位置から $D_{\max}$ 以内に Y が存在する確率とする．つまり確信度は

$$\frac{X \text{ の出現位置から } D_{\max} \text{ 以内に } Y \text{ が存在する場合の数}}{X \text{ の出現数}}$$

で計算される．

4. 実験

実験では実際の日本の株価データを用いる．15 の会社について機械，運送設備，サービスと貿易の 3 カテゴリから，それぞれ 5 件づつ証券コード順に従って選択した．合計 5947 の会社の PR 用 Web ページとニュースサイトについて URL 辞書に登録した．単語辞書には表 1 と同じ 6 種類のタグと 33 種類の単語を保存しているテーブルを用いた．実際に得られたニュースデータは Web サイトの公開時期やニュースの記録時期，上場タイミングの差によって各社で用いるデータの期間はそれぞれ異なるが，平均で 8 年 3ヶ月分のデータ，全株価データレコードは 29001 個となった．最終的に，5950 個のニュースを使用し，アノテーションのタグの数は 2443 となった．保存されたアノテーションは均等に全体に存在し，比較的疎な状態となった．

図 4 に獲得したルールの一例を示す．図 4 の (a) の会社ではイベントで次製品の展示を主に行っており，展示後の株価は 72%の確率で上昇する傾向があることが分かった．また，図 4 の (b) の会社では貿易関連の業務を行っており，日本から取引先の会社に対しての異動のときに大きく期待感のある株価の動きとなっていたことが分かった．

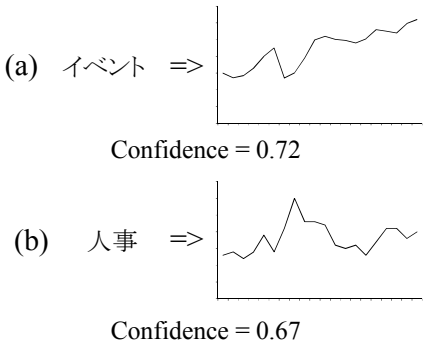

図 4: 獲得したルール

5. 結論

本論文は時系列データと背景情報から知識を抽出する手法について二つの機能を提案した．一つ目は人間のアノテーションにかかる作業コスト削減のために，Web サイトから自動的にタグを獲得する機能である．URL 辞書に従ってニュースを収集し，単語辞書に従ってタグにする．これによって，効率よく解析対象の株価に対しての情報を収集し，ニュースのグループ分けを行う．二つ目はアノテーションを基にして時系列データの動きを予測するためのルール獲得方法として，アノテーション相関ルールを提案した．最大許容長と最小確信度の二つの閾値にしたがって効果的な知識を獲得できることを実験によって示した．

参考文献

[1] R Wei Wu, Bin Zhang et al: Automatic generation of personalized annotation tags for twitter users. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 689-692 (2010).

[2] Emily Moxley, Tao Mei, Xian-Sheng Hua, et al: Automatic video annotation through search and mining. Multimedia and Expo, 2008 IEEE International Conference on, pp. 685-688 (2008).

[3] Hung-Yi Lo, Shou-De Lin, et al: Audio Tag Annotation and Retrieval Using Tag Count Information. Advances in Multimedia Modeling, pp 339-349, Springer (2011).

[4] Dimitrovski, I., Kocev, D., et al: Hierarchical annotation of medical images, Pattern Recognition, Elsevier (2011).