電気学会論文誌 C（電子・情報・システム部門誌）
IEEJ Transactions on Electronics, Information and Systems
Vol.132 No.4 pp.623–624　DOI: 10.1541/ieejeiss.132.623

研究開発
レ タ ー

Web 情報からの自動タギングのデータマイニングへの利用

学生員　杉村　博*　正　員　松本　一教*

Utilization of Automatic Tagging Using Web Information to Datamining

Hiroshi Sugimura*, Student Member, Kazunori Matsumoto*, Member

（2011 年 9 月 16 日受付，2011 年 12 月 1 日再受付）

This paper proposes a data annotation system using the automatic tagging approach. Although annotations of data are useful for deep analysis and mining of it, the cost of providing them becomes huge in most of the cases. In order to solve this problem, we develop a semi-automatic method that consists of two stages. In the first stage, it searches the Web space for relating information, and discovers candidates of effective annotations. The second stage uses knowledge of a human user. The candidates are investigated and refined by the user, and then they become annotations. We in this paper focus on time-series data, and show effectiveness of a GUI tool that supports the above process.

キーワード：自動タギング，アノテーション，メタデータ，アノテーション時系列データ
Keywords: Automatic tagging, Annotation, Meta data, Annotated time series data

1. はじめに

時系列データは，幅広い分野で応用されるため，この形式のデータからのデータマイニング技術は重要である。時系列データだけを用いてマイニングを行う様々な手法が開発されているが[1]，実際は記録されたデータには重要な背景情報が内在しており，人間の解釈や記録に関与したイベント情報を組み合わせることによってさらなる知識抽出が行えると考えられる。データに対して関連する情報をタグとして付与して利用する方法が研究されており，文書や医療の電子カルテなどへ付与したタグの利用方法が提案されている[2][3]。このようにタグを付与することをタギング（アノテーション）という。タギングにかかるコストは非常に高く，作業の自動化や分散化が必要であるが，その様な研究はいまだに十分ではない。そこで本論文では，フィナンシャル時系列データを対象として Web クローラを応用した自動タギングについて提案し，一部を実験により検証する。

2. 時系列データへのタギング

本論文で扱う時系列データ S は実数のシーケンス $S = (s_1, s_2, \cdots, s_n)$ である。この時 n はこのシーケンスの長さである。Fig. 1 に示すように，タグはある一点の時点に対して与えられるものと，時区間に対して与えられるものの 2 種類が考えられるので両者を取り扱う必要がある。

* 神奈川工科大学 工学研究科 情報工学専攻
〒 243-0292　神奈川県厚木市下荻野 1030
Course of Information and Computer Sciences, Graduate School of Engineering, Kanagawa Institute of Technology
1030, Shimo-ogino, Atsugi-shi, Kanagawa 243-0292, Japan

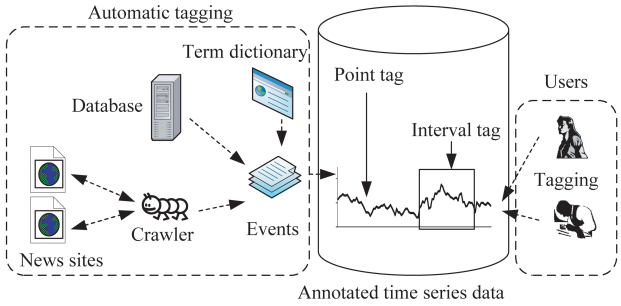

Fig. 1.　Automatic tagging for time series data

```
Input: URL database UDB,
       Tag dictionary Tdic.
for each Webpage x in UDB
  for each term w∈ Tdic which appears in x
    infer the time t of w;
    tag w at t;
  end
end
```

Fig. 2.　Crawling Algorithm

タグを自動的に付与する機能の最初のステップは，Fig. 2 に大まかな流れを示すように Web クローラによる Web 上の情報の自動収集を利用する。Web 空間すべてを収集対象とする方法では探索範囲が広すぎるために適切な方法で絞り込む必要がある。本論文では，対象とする Web ページの情報（URL）を事前にデータベースに与えておく方法を採用している。その上でさらに，各ページから収集する情報を，Table 1 に示すタグ辞書を利用することで絞り込む。

タグ辞書とは，有効な情報となると期待できる単語を登

Table 1.　Tag dictionary

Tag	単語
人事	人事，再編，就任，異動，改編
開発	開発，実用化，低価格化，実現，受注，完成，完工
経営	販売，発売，提供，決算，実務報告，会社設立，売り上げ
通知	イベント，開催，対応，お知らせ，ご案内
謝罪	リコール，お詫び，恐れ，無料交換，障害

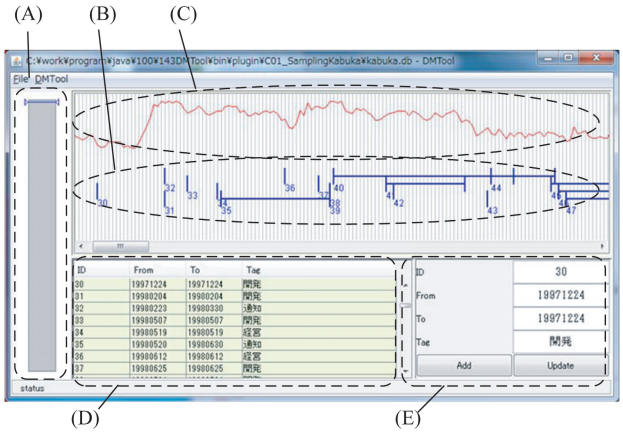

Fig. 3.　GUI of the system

録した一種のオントロジーである。単語そのものではなく，定められたタグに統一して付与することで，ページ毎の標記の違いを吸収できる。この辞書の作成は予備実験により効果が得られる見通しのついた語を登録することで行った。この様なタグに対して，その生起した時刻を推定する。本論文の実験では，ニュース記事のページや企業 PR のページが主であるため，その発表時間を取り出して，この時点に対するタグとする。

次のステップでは，自動的に与えらえたタグをユーザの判断で調整する。Fig. 3 はそのためのインタフェースを示しているが，時点についているタグを時区間上のものに修正することも行える。

3.　実装と実験

実際にタギングされた時系列データを表示するシステムを Java によって実装した。Fig. 3 に実装したシステムの画面を示す。この画面では大きく 5 つの機能を提供している。(A) は時系列データ全体からみた現在の表示位置を示している。(B) はタグの位置と範囲，そしてタグの ID を示している。タグのついた時刻の修正や，時点につけられたタグの時区間上への拡大や，その逆を行うことができる。(C) は時系列データを折れ線グラフで示す。(D) にはデータベースに登録されているタグの一覧が表示されていて，(B) のタグ ID と対応付けられている。(E) は選択したタグの内容を修正したり，新しいタグを追加する機能を提供する。

時系列データには電機，通信，自動車の業界から選んだ 3 社の株価データを使用する。クローリング対象とする Web は各社のサイトおよびニュースサイト 48 個を設定した。Table 1 に示すタグ辞書を用いて，ニュースタイトルに単語が含まれている場合に対応する Tag を付ける。収集したタグは，1996 年 10 月 22 日から 2011 年 10 月 22 日までの

Table 2.　Total number of tags

Tag	人事	開発	経営	通知	謝罪	Total
Electronics industry	198	135	23	118	5	479
Telecommunications	142	465	203	107	0	917
Car manufacturer	238	291	138	50	4	421

Table 3.　Experimental result

	Tag			All		
Company	Series	Size	Accuracy(%)	Series	Size	Accuracy(%)
Electronics industry	108	18	74.9	329	95	63.3
Telecommunications	117	13	76.5	392	77	62.7
Car manufacturer	127	11	84.9	317	69	67.9

間に合計 2117 個となった。各タグの数を Table 2 に示す。本実験ではクローラで集めるタグを絞っているため，タグ付けは比較的に疎な状況である。また，特定の時期に集中しておらず，全体的にほぼ均等に分散している。

タグを基にして収集したデータから得られる知識の一例として，株価の売り買い予測のための決定木を作成する。タグを含む期間のすべてのデータを，データサイズ 25 で切り抜き，前から 20 データを過去データ，後ろの 5 データを未来予測データとする。未来予測データは上昇，下降，停滞に分類して学習することにより，過去の 20 データを基にして売り買い予測を行う決定木を導出する。

比較実験として，タグのない部分も含め，時系列データから得られる全ての部分時系列データを用いて決定木学習を行う。Table 3 には，タグに含む期間のデータだけを収集した場合を Tag，すべてのデータを用いた場合を All として，獲得した部分時系列データ数と決定木のサイズ，分類精度を示す。分類精度は 10 分割交叉検定で求めた。

4.　おわりに

本論文ではニュース情報を基にして時系列データに自動的にタギングを行うシステムについて提案した。そして，収集されたタグを有効に使うことで未来予測の精度を向上出来ることを実験によって示した。タグ辞書に登録した程度の小規模な情報を自動収集することだけでも十分有効に活用できることが明らかになった。今後の課題は本システムをベースとしたソーシャルシステム化と，タグと時系列データの連携による新たな知識抽出方法の検討である。

文　献

（1） G. Dong and J. Pei: “Sequence Data Mining”, Springer (2007)
（2） 大谷　淳・井上　潮：「コンテンツ変化を考慮した時系列 Web アノテーションシステム」，第 19 回データ工学ワークショップ，第 6 回日本データベース学会年次大会 (DEWS2008), Vol.B1-3 (2008)
（3） M. Takaai and H. Masuichi: “An Annotation Method of Clinical Texts using a Medical Ontology”, IEICE Technical Report NLC206-80, Vol.106, No.517, pp.42–48 (2007) (in Japanese)
鷹合基行・増市　博：「臨床テキストに対する医学オントロジーに基づくアノテーション手法に関する研究」，信学会技術研究報告，NLC，言語理解とコミュニケーション，Vol.106, No.517, pp.43–48 (2007)
（4） 柳本豪一・吉岡理文・大松　繁：「ソーシャルブックマークを用いた Web ページのクラスタリング」，日本データベース学会論文誌，Vol.8, No.1, pp.149–154 (2009)