Blogデータのクラスタリングと分析

力 規晃† 越村 三幸‡ 藤田 博‡ 長谷川 隆三‡
徳山工業高等専門学校 情報電子工学科† 九州大学 大学院システム情報科学研究院‡

1．はじめに

現在、無数のBlogデータがインターネット上にある。これらのBlogの著者には、類似の嗜好を持つ人も多く存在する。これらの特定の嗜好を持つBlogを分類し、そこで成り立つWebページの閲覧動向や興味のある事柄などを抽出できれば、推薦システム等へ有効に利用できると思われる。本研究では、インターネット上にあるBlogデータから、クラスタリング手法により特定の嗜好を持つBlogを抽出し、このデータから帰納論理プログラミング（Inductive Logic Programming, ILP）[1]を用いて、特定の嗜好を持つBlog著者グループで成り立つ特徴の抽出を行う。また、この特徴の推薦システムへの応用についても考察を行う。

2．Blog

Blogとは Web 上の日記などの個人が記述した記事が時系列に配置されている Web コンテンツであり、厳密な定義は存在しないが、次のようないくつかの特徴的な機能を持っている。

- **Trackback**
  ユーザーがリンク先の相手に対して、リンクを張ったことを通知する機能である。リンク元のタイトル、URL、内容の要約などが送信される。
- コメント
  通常の掲示板と類似の機能であるが、Blog 上のどの記事に対するコメントであるかが明確となっている。
- **RSS配信**
  RSS 配信とは XML 形式でブログの記事の情報を配信する仕組みである。これによってより明確に Blog の内容や更新の情報を第三者が取得できる。

3．Blogのクラスタリング

インターネット上の Blog サイトから Blog データを収集した後、形態素解析を行い、名詞を取り出す。これを用いて特徴ベクトルを作り、クラスタリングを行う。

3．1 Blogデータの収集

Blog サイトの RSS ファイルにより、あるカテゴリの記事を持つ Blog のトップページを集め、トラックバックを抽出する。さらに、その Blog 上の記事を RSS ファイルにより、各記事のカテゴリ、タイトル、本文、リンクを取り出す。

3．2 形態素解析

形態素解析器 Sen[4]を利用して、タイトルと記事のテキストを形態素に分割し、名詞の単語のみを取り出す。その際に名詞が連続して現れる場合は結合して一つの単語として処理をする。

3．3 特徴ベクトル

この取り出した各単語について出現回数を計算し、全ての Blog での出現回数の上位 n 個の単語 $t_i\ (i=1,\ldots n)$ を用いて、特徴ベクトル

$$\begin{pmatrix} w(t_1,d_j) \\ w(t_2,d_j) \\ \vdots \\ w(t_n,d_j) \end{pmatrix}$$

を計算する。ここで d_j は j 番目の Blog であり、Blog の総数は m とする。それぞれの $w(t_i,d_j)$ については TF・IDF、エントロピー、分散の 3 つの手法により算出する。

- **TF・IDFによる算出**

$$w(t_i,d_j)=tf_{ij}\cdot idf_i$$

tf_{ij} は t_i が d_j 中で出現する出現頻度であり、idf_i は t_i の出現する Blog の全ての Blog の中に占める割合の逆数である。

- エントロピーによる算出

$$w(t_i,d_j)=\frac{f_{ij}}{F_i}\cdot(\log m-H_{ij})$$

ここでエントロピー H_{ij} は

$$H_{ij}=-\sum_j\frac{f_{ij}}{F_i}\log\frac{f_{ij}}{F_i}$$

Clustering and analysis for blog data.
†Noriaki Chikara ・
Department of Computer Science & Electronic Engineering, Tokuyama College of Technology
‡Miyuki Koshimura, Hiroshi Fujita, Ryuzo Hasegawa
Faculty of Information Science and Electrical Engineering, Kyushu university

であり、F_iは全 Blog でのt_iの出現頻度であり、f_{ij}はd_j内でのt_iの出現頻度である。

- 分散による算出

$$w(t_i,d_j)=\frac{1}{m}\sum_j(\bar{v}_i-v_{ij})^2$$

ここで、v_{ij}はd_jでのt_iの出現頻度であり、$\bar{v}_i$はt_iの出現頻度の平均値である。

3．4　クラスタリング手法

2種類のクラスタリング手法を用いる。

- Ward 法[5]

階層的クラスタリング手法の一手法である。階層的クラスタリング手法では最初に各データをクラスタとして置き、その中で近いクラスタのペアを選び、併合することを繰り返す。Ward 法ではクラスタ間の距離としてクラスタの重心から各データへの距離の自乗の和の増加分を用いる。

- K-means 法[5]

非階層的クラスタリング手法の一手法である。

最初に各クラスタで任意の代表点を選ぶ。次に最も距離が近い代表点のクラスタに各データは分類し、分類されたデータを元に新しい代表点を選び、さらに各データを分類しなおす。この一連の作業を繰り返す。

4．帰納論理プログラミング

帰納論理プログラミングとは一階述語論理に基づいた機械学習の手法である。ILP の一般的な枠組みは、正例E^+と負例E^-,背景知識Bが節集合で与えられ、

$$\begin{cases} H\cup B\models \mathrm{E}^+ \\ H\cup B\cup E^- \not\models \square \end{cases}$$

を満たす仮説Hを見つけることである。
本研究での正負例、背景知識で用いる知識表現の例を示す。

creator(Blog1,Name1).
Name1 がブログ Blog1 の著者である。
link(Content1,Content2).
記事 Content1 から記事 Content2 へリンクが張られている。
contain(Blog1,Content1).
ブログ Blog1 に記事 Content1 が含まれる。
date(Content1,Time).
記事 Content1 の記述された日時は Time である。
next(Content1,Content2).
記事 Content1 は記事 Content2 の直後の記事である。
prev(Content1,Content2).
記事 Content1 は記事 Content2 の直前の記事である。
after(Content1,Content2).
記事 Content1 は記事 Content2 よりも後の記事である。
before(Content1,Content2).
記事 Content1 は記事 Content2 よりも前の記事である。
category(Content1,Category1).
記事 Content1 は分類 Category1 に含まれる。

以上の知識表現から、本研究では ILP システム Aleph[2]を用いてリンクと記事の時系列に関連した仮説をルールとして導き出す。

導出したルールは、ある特徴を持った Blog グループで成り立つ、Blog 記事の時系列とリンクに関するルールであり、ある特定の嗜好をもった Blog の著者グループで成り立つ Web ページ閲覧に関するルールとも見なすことができる。

このルールを用いれば、類似した別の Blog 著者への Web ページの推薦に利用が可能だと思われる。また、この場合述語表現のルールによって推薦を行うため、推薦理由も容易に示すことができる。

5．おわりに

本研究ではインターネット上にある Blog データから、クラスタリング手法により特定の嗜好を持つ Blog を抽出し、このデータから ILP を用いて、特定の嗜好を持つ Blog 著者グループで成り立つルールの抽出を行った。

今後、様々な Blog データから導出したルールを用いて、Web 推薦の実証や実験を行っていく。

本研究は科研費(20240003)の助成を受けたものである。

参考文献

[1] 古川康一,尾崎知伸,植野研:帰納論理プログラミング,共立出版,2001
[2] A.Srinivasan:The Aleph Manual, http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/, 1999
[3]戸田雄士,太田学:関連ページ特性に基づいた Blog クラスタリング,DEWS2008 B4-4,2008
[4] Sen Project, http://ultimania.org/sen/
[5]新納浩幸:R で学ぶクラスタ解析,オーム社,2007