テキストマイニングを用いたウェブデータからのキーワード獲得

有村, 博紀
九州大学大学院システム情報科学研究院

安部, 潤一郎
九州大学大学院システム情報科学研究院

坂本, 比呂志
九州大学大学院システム情報科学研究院

有川, 節夫
九州大学大学院システム情報科学研究院

https://doi.org/10.15017/4782066

出版情報：九州大学情報基盤センター年報. 1, pp.25-27, 2001-10. 九州大学情報基盤センター
バージョン：
権利関係：

情報基盤センター年報第1号（2001年）
九州大学情報基盤センター

テキストマイニングを用いたウェブデータからのキーワード獲得

Discovery of Important Keywords From the Web Using Text Mining

有村博紀 †,‡　　安部 潤一郎 †　　坂本 比呂志 †　　有川 節夫 †

† 九州大学大学院システム情報科学研究院
† Depertment of Informatics, Kyushu University
‡ さきがけ研究21, ‡ PRESTO, JST

要旨　われわれの研究グループでは，大規模テキストデータを対象に，最適パターン発見の枠組みにもとづいて，高速かつ頑健なテキストマイニング手法を開発してきた．本稿では，われわれのテキストマイニング手法を紹介し，ウェブデータからのキーワード発見への応用について述べる．

Abstract　In this talk, we consider text mining from Web. We present fast and robust text mining algorithms that finds simple class of string patterns, called proximity association patterns, from large collections of unstructured text data. We also report experiments of keywords extraction from Web.

1　はじめに

1990年代半ばから現在まで，高速なネットワークと大容量記憶装置の発達を背景として，大量かつ多様なテキストデータの利用が急速に進んできた．例えば，ネットワーク上に分散したウェブページの全体は，現在，最も大規模なテキストデータベースとみなせる．また，ファイルシステム上に蓄積されたビジネス文書の集積や，XMLアーカイブも大規模テキストデータの例である．そこで，これらの大規模テキストデータに対して，従来からの情報検索手法を超える新しいアクセス手法の開発が急務となっている．

このための有力な候補の一つが，データマイニングである．データマイニングは，大量データから自明でないパターンや規則を発見するための方法である．しかし，ウェブデータに代表される大規模テキストデータは，

(1) 明示的な構造をもたない，
(2) 非均質で多様な，
(3) 膨大なテキストの集積

という特徴をもつ．そのため，関係データベースを対象としている従来のデータマイニング手法を，テキストデータに直接適用することはむずかしい．

そこで，われわれの研究グループでは，構造をもたないテキストデータを対象に，最適パターン発見と呼ばれる枠組みにもとづいて，一連のテキストマイニング手法を開発してきた．本チュートリアルでは，われわれが開発したテキストマイニング手法を紹介し，実際の大規模テキストデータへの応用について解説する．

2　フレーズ相関パターン

発見対象のパターンとして，相関ルールを任意長の文字列を属性としてもてるように拡張し，

$$(\langle \texttt{attack on} \rangle, \langle \texttt{oil platforms} \rangle; 2)$$

のような形をした**語相関パターン**という単純なクラスのパターンを導入した [2].

語相関パターンは，テキスト中の複数の文字列のならびと，各文字列の出現の間の距離（近接度）の組である．上の語相関パターンは，テキスト中にはじめに ⟨attack on⟩ が出現し，つづいて2単語以内に ⟨oil platforms⟩ が出現するようなパターンを表している．

図1に実際の英文テキストにおける語相関パターンの出現例を示す．これは，英文新聞記事データから，後述の最適化パターン発見を用いて実際にみつかったパターンである．このように，近接語相関パターンは，

- Monday's attack on two Iranian oil platforms by American forces in the Gulf.
- ... attack on two Iranian oil platforms in retaliation for an Iranian attack last Friday on a Kuwaiti ship ...
- ... the action would involve an attack on an oil platform.
- ... the United States' attack on an Iranian oil platform on Monday and said it should not worsen the Gulf crisis.
- The attack on the oil platform was the latest example of a U.S. ...
- ... attack on an Iranian oil platform in the Gulf on Monday appeared to be ...
- One source said of the attack on the oil platform: ...
- A top Iranian military official said America's attack on an Iranian oil platform on Monday had involved the United States in full-scale war ...
- Weinberger was asked why the United States had chosen to attack an oil platform rather than Iranian Silkworm missile platforms ...

図 1: 語相関パターン (⟨attack on⟩, ⟨oil platform⟩; 2) の Reuters-21578 データでの出現.

任意長の複数の文字列と近接度の組み合わせで，さまざまな文に共通するパターンと文脈情報を表現可能である．

3 最適パターン発見

それでは，与えられたテキスト集合を特徴付けるためには，マイニングアルゴリズムは，どのようなパターンを発見すれば良いだろうか？

本研究では，パターン発見の枠組みとして，**最適パターン発見**を採用する．これは，テキストマイニング問題を，正負例の有限集合 $S \subseteq \Sigma^* \times \{0,1\}$ が与えられたとき，例に対する**分類誤差** (empirical classification error)

$$Error_S(H) = \sum_{\langle x,b\rangle \in S} [H(x) \neq b]$$

を最小化するようなパターン H を見つける最適化問題として定式化しようというものである [1,3]．ここで，分類例 $\langle x,b\rangle \in S$ において，分類ラベル $b \in \{0,1\}$ は，そのテキスト x が興味のあるカテゴリに属するか否かを示す．目的関数としては，情報エントロピーやカイ 2 乗指標など，一般的な統計的な尺度を利用可能である．

この最適パターン発見は，1970 年代の統計的決定理論における経験誤差最小化にその起源をもち，データ中の雑音に頑健で，未知パターンのクラスがわからない場合にも，うまく働くことが理論的にわかっている．最近，データマイニングと計算学習理論分野で，最適パターン発見が独立に再発見され，理論と応用の両面で盛んに研究されている [3].

4 高速発見アルゴリズム

パターン中の文字列の最大数を定数 d としたとき，全解探索アルゴリズムでは，$O(n^{2d+1})$ 時間を要し，実用に耐えない．そこで，固定した d について，入力サイズ n に対して，線形に近い時間で最適パターンを発見する高速なアルゴリズム Split-Merge を開発した [2].

テキストマイニングの難しさは，仮説空間の膨大さである．わずか 1.2MB の英文テキストが 700,000 個以上の異なる文字列を含み得る．ここでは，接尾辞配列というデータ構造を用いて，テキストに出現するすべての部分文字列をうまく管理し，パターンの枚挙と出現位置を同時に計算する．

5 ウェブマイニングへの応用

最適パターン発見を用いて，以下のように，ウェブからのテキストマイニングが実現可能である．従来のテキストマイニング方式の多くは，高頻度パターンの抽出に基づいている．しかし，これではテキストに特徴的なパターン（それらの多くは中頻度である）が高頻度パターンに隠ぺいされるという問題が生じる．

そこで，利用者が興味をもっているテキストを正例とし，それ以外のテキスト全体を負例として，最適パターン発見アルゴリズムを適用し，特徴的なパターンを発見する [1].

図 2 に，ウェブ検索エンジンとウェブロボットを用いて収集したウェブページ集合を対象にしたテキストマイニング実験の結果を示す．ここでは，情報エントロピーを用いて，自動車会社 HONDA 関連のページを

(a) HONDA vs. SOFTBANK

Rank	Pattern	Rank	Pattern
1	⟨honda ⟩	11	⟨miles ⟩
2	⟨prelude ⟩	12	⟨bike ⟩
3	⟨i ⟩	13	⟨motorcycle ⟩
4	⟨car ⟩	14	⟨racing ⟩
5	⟨parts ⟩	15	⟨black ⟩
6	⟨engine ⟩	16	⟨si ⟩
7	⟨99 ⟩	17	⟨me ⟩
8	⟨rear ⟩	18	⟨tires ⟩
9	⟨vtec ⟩	19	⟨fuel ⟩
10	⟨exhaust ⟩	20	⟨my ⟩

(b) HONDA vs. TOYOTA

Rank	Pattern	Rank	Pattern
1	⟨honda ⟩	11	⟨honda prelude ⟩
2	⟨prelude ⟩	12	⟨98 ⟩
3	⟨vtec ⟩	13	⟨valkyrie ⟩
4	⟨si ⟩	14	⟨99 time ⟩
5	⟨bike ⟩	15	⟨honda s ⟩
6	⟨99 ⟩	16	37 ⟨rims ⟩
7	⟨motorcycle ⟩	17	⟨looking ⟩
8	⟨the honda ⟩	18	⟨looking for ⟩
9	⟨prelude si ⟩	19	⟨scooters ⟩
10	⟨civic ⟩	20	⟨black ⟩

図 2: (a) 自動車会社 HONDA 関連とインターネットビジネス会社 SOFTBANK 関連のウェブページを，それぞれ正例と負例として，最適パターン発見をおこなった．(b) 同じ自動車会社 HONDA 関連と TOYOTA 関連のウェブページで同じ実験をおこなった．

特徴付けるパターンを発見するのが目標である．データは，正例と負例ともに，タグを除いて 5MB 前後である．

図 2 で，自動車に関連しないページと比較した左側 (a) では，一般的な自動車用語が抽出されている．同じ自動車会社 TOYOTA と比較した右側 (b) では，HONDA が生産している具体的な車種名が抽出されていることがわかる．非常に多い負例に対してもうまく働く．最適パターン発見が，語彙や内容に関する事前知識なしに，適切なキーワードを見つけていることに注意されたい．

参考文献

[1] 安部, 藤野, 下薗, 有村, 有川, テキストデータからの高速データマイニング, 人工知能学会誌, Vol.15, No.4, pp.618–628, 2000.

[2] H. Arimura, S. Shimozono, S. Arikawa, Efficient discovery of optimal word-association patterns in large text databases, *New Generation Computing*, Vol.18, pp.49–60, 2000.

[3] S. Morishita, On classification and regression, In Proc. the 1st Discovery Science (DS'98), LNAI 1532, Springer-Verlag, 1998.