様式 C-19

科学研究費補助金研究成果報告書

平成２１年　５月２９日現在

研究種目：若手研究(B)
研究期間：2007～2008
課題番号：19700150
研究課題名（和文）言語のスケールフリー性に着目した大規模テキストからの特徴的なパターン発見

研究課題名（英文）Pattern Discovery from Large Text Data Based on the Property of Languages Being Scale-Free

研究代表者 池田 大輔（Ikeda, Daisuke）九州大学・大学院システム情報科学研究院・准教授
研究者番号：00294992

研究成果の概要：

本研究の大目標は、スケールフリー性を利用し、言語や対象領域に依存しないテキストマイニングの手法を確立することである。これに対し、可変長の文字列の組み合わせでパターンを発見する手法を 2 つ提案し、その有効性を実験により示した。最初の手法で用いるパターンは、複数の可変長部分文字列が重複を持って重なっている。この手法により、従来は困難だったワードサラダと呼ばれる人工的に生成されたスパムを検出できるようになった。この手法は、普通の頻度分布と異なる部分を抽出するという意味で従来よく用いられてきた標準正規分布からのずれ(z-score)を用いた手法に近い。一方で、データマイニングの分野で研究されてきた例外パターン発見の枠組みをテキストに応用し、z-score では見つけられなかったパターンを発見できることを、DNA 配列を用いた実験により示した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,800,000	0	1,800,000
2008 年度	1,500,000	450,000	1,950,000
年度			
年度			
年度			
総　計	3,300,000	450,000	3,750,000

研究分野：計算機科学
科研費の分科・細目：情報学・知能情報学
キーワード：知識発見とデータマイニング

１．研究開始当初の背景

Web 上の情報やコンテンツは従来提供者から利用者への一方向的な提供が主だったが、ブログや掲示板など情報の利用者によるコメントのように付加的な情報が蓄積・提供されつつある。このような情報は集合知として価値があり、マーケティングや購買前の調査など、企業や個人が次の行動を決定する時に利用される。しかし、これらの情報は玉石混淆であり、有用な情報を見つけるために多くの時間を割かなければならない。そこで、このような情報を解析して有用な情報や知識を抽出する必要性が増加している。

テキストではないデータマイニングはこ

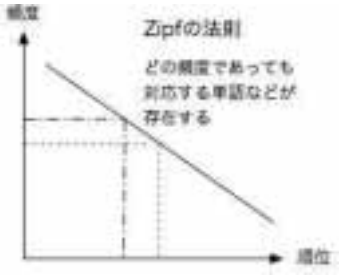
の 10 年ほどよく研究されており、基本的に高頻度なパターンを抽出する枠組みが存在する。また、単に高頻度というだけでなく、特徴的なパターンについても研究されており、極大パターンや閉パターンなどの概念が提案されている。

一方、データマイニングと比較するとテキストマイニングの研究は十分に進んでいるとは言い難い。自然言語処理や機械学習などの手法を取り入れた研究はあるものの、基本的な枠組みはデータマイニングと同じで高頻度なパターンを発見している。しかし、後述するように、スケールフリー性が成立することが多いテキストデータに対してこの枠組みでは十分ではないと思われる。また、これらの研究では対象とする言語を固定したり、発見すべきパターンを固定しており、対象とする言語やドメインがある程度限定されている。よって、言語やドメインに限定せず、言語のスケールフリー性に着目した手法の確立が求められている。

２．研究の目的

言語やドメインに依存しないためには、抽出すべきパターンや知識を言語やドメインに依存せずに定義しなければならない。従来の（テキスト）マイニングでは主に頻度を用いて定義しており言語に依存しないが、適切な頻度のしきい値（最小サポート値）はドメインに依存する。本研究では言語やドメインに依存しない共通する特徴としてスケールフリー性に着目する。

スケールフリー性は、単語の頻度分布に対するベキ分布や Zipf の法則、最近ではロングテールとも呼ばれる性質で、単語の使用頻度がベキ分布に従うというものである（下図）。



スケールフリー性により、どの順位付近においても同じ形の頻度分布を形成していることがわかる。これは、頻度によるしきい値では、意味のあるパターンが発見できないことを意味している。

そこでこの分布を背景分布と考え、分布間の距離などにより抽出すべきパターンかどうかを判断する。つまり、本研究の目的は、大量のテキストデータの背景分布から乖離する部分が抽出すべきパターンであるとの仮説を検証し、この枠組みにおける高速なマイニングアルゴリズムを開発することである。

３．研究の方法

上述の目的に対し、以下の 3 項目のマイルストーンを設けて研究を進める。

●確率分布によるパターンの特徴づけ：本研究は言語のスケールフリー性、つまり、単語の頻度分布に着目しており、確率分布を用いることは自然である。そこで、特徴的なパターンと背景分布（特徴的でない他のテキスト部分）の距離による問題の定式化などを行い、頻度のしきい値ではないテキストパターンマイニングアルゴリズムを構築し、実験よりその有効性を確認する。

●文字列データ構造による高速なアルゴリズムの構築：接尾辞木などのデータ構造の特徴を用いた高速なアルゴリズムを構築する。具体的にどのようなデータ構造になるかは、次のパターン拡張も考慮しつつ検討する。

●部分文字列パターンからより複雑なパターンへの拡張：最も簡単なパターンである部分文字列パターンに対するアルゴリズムを拡張し、より広いパターンクラスに適用可能にし、実データを用いてパターンの有効性を示す。つまり、厳密に一致する部分を抽出するだけでなく、例えば、近似的に一致する部分であったり、文法的な構造を持ったパターンを抽出可能にする。

４．研究成果

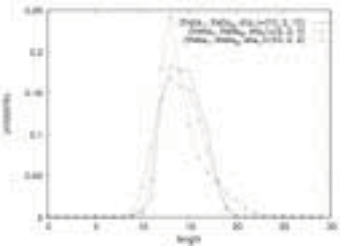
上述のマイルストーンごとに得られた成果を説明する。

●パターンの特徴づけ：パターンとして、複数の可変長の部分文字列が重なって連接したものを提案し、パターンの確率分布からの

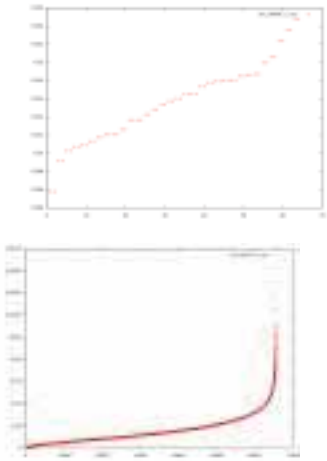
ずれにより特徴的なパターンとして発見する手法を提案した。重なりを持つため文脈情報を維持でき、また、可変長文字列であるためあらかじめ長さを指定する必要はない。しかし、用いた確率分布は正規分布であり、スケールフリー性とは異なる。また、正規分布からのずれという意味では、z-score などの手法に近い。そこで、データマイニングの分野で用いられていた例外パターン発見の枠組みをテキストデータに対し適用した。

具体的には、一般的な語の分布の推定と同時に、特徴的なパターンの発見を行う「例外文字列発見問題」という問題を定式化した。この枠組みでは２つの文書集合に対し、片方に頻出する２つの部分文字列パターンを見つけ、これを連接してできた新たな部分文字列パターンがもう一方に頻出する時にこれを「例外的」なパターンとして抽出する。確率分布を陽に用いる方法では、実際には正規分布を仮定しており、例外的なパターンはどのような入力に対しても同程度存在する。一方で、提案手法は比較対象としたい集合を背景集合として用いることで、様々な例外パターンを対比的に抽出できることを示した。さらに、従来の手法では見つけることが困難なパターンを発見できることを、DNA 配列を用いた実験により示した。

上述したように、この枠組みでは確率分布を陽には用いないが、得られたパターンの長さがよく揃っている（下図参照）。このグラフはゲノム配列から抽出した意外なパターンの長さであり１５前後に集中している。



次の２つは長さ３と８の固定長部分配列の分布であり、3 の時は直線的だが、8 の時はべき分布となり、大きな偏りが見られる。つまり、偏りが現れ始めた長さを連接したもの（長さ１５から１６程度）が意外なパターンとして見つかっている。これは言語のスケールフリー性が現れ始めたところを発見していることを意味する。あらかじめ機械学習等で最適な長さや分布を学習せずに、頻度や確率に関する簡単な条件だけで分布が偏りはじめるところを探しだしていることを意味していると考えられ、大変興味深い現象である。



図：例外パターンの長さの分布と長さごとの文字列の分布(長さ 3 と 8)

●アルゴリズムの構築と実装：上述の例外文字列パターンの発見アルゴリズムを、文字列に対するデータ構造である接尾辞木を用いたアルゴリズムを構築し、実装した。実装には効率的な枝刈りの仕組みが備えられており、理論的には O(N^2)の計算量だが、多くの実際のデータに対してはほぼ線形時間で動作することを確認した。また、効率的なメモリ使用により、DNA 配列全体を使った実験が可能であることも確認した。

●より複雑なパターンの検討：例外文字列パターンは頻出な文字列の連接で定義されるが、より複雑なパターンへの拡張を視野にいれ、部分列への拡張を検討した。そこで、可変長の文字列を重ねて複雑なパターンを構成する手法も検討した。これらの頻度からパターンの頻度を推定する手法を提案した。この手法により、従来は不可能だったワードサラダと呼ばれる文章の中の単語がランダムに変えられた特殊なスパム(ワードサラダ)を検出することが可能になった。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）
Takashi Uemura, Daisuke Ikeda and Hiroki

Arimura,”Unsupervised Spam Detection by Document Complexity Estimation”, Proceedings of the 11th Inernational Conference on Discovery Science, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol. 5255, pp.319--331, Oct. 2008.

〔学会発表〕(計３件)

徳永旭将、中村和幸、樋口知之、池田大輔、大久保翔、藤本昌子、吉川顕正、湯元清文、MAGDAS/CPMN グループ湯元清文:``時系列データマイニングによる動的ヘテロなシステムからの知識発見 -宇宙天気研究における大規模帰納処理システム構築へ向けて'', 日本地球惑星科学連合 2009 年大会、May 2009.

稲田泰裕, 池田大輔, 鈴木英之進:“CF-Suffix Trie を用いた頻出移動パターンマイニング手法”, 第9回データマイニングと統計数理研究会, 京都, 3/3 2009.

Takashi Uemura, Daisuke Ikeda and Hiroki Arimura,”Unsupervised Spam Detection by Document Complexity Estimation”, Proceedings of the 11th Inernational Conference on Discovery Science, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol. 5255, pp.319--331, Oct. 2008.

〔図書〕(計０件)

〔産業財産権〕

○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕

ホームページ等

６．研究組織

(1)研究代表者

池田 大輔（Ikeda, Daisuke）

九州大学・大学院システム情報科学研究院・准教授

研究者番号：00294992

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：