様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月２４日現在

機関番号：１７１０２
研究種目：挑戦的萌芽研究
研究期間：２００９～２０１１
課題番号：２１６５００３１
研究課題名（和文）大量実テキストデータに潜む少量多品種な部分構造の統計的発見と利用に関する研究
研究課題名（英文）　Research on statistical discovery of a wide variety of patterns with low frequencies and its applications
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（和文）：
テキストデータの大部分ではなく、相対的に少ない部分に対して成立するパターン発見手法を構築するためにパターン発見への写像導入とパターン発見への距離空間導入に分けて研究を行った。前者に対し、近似文字列照合を用いたパターン発見手法を提案し、実験によりその有効性を示した。この際、文字の写像が本質的であることを示した。後者はパターンの抽出基準であり、「普通」の部分構造を組み合わせたパターン発見の有効性をゲノム配列やWeb上の文書による実験で示した。

研究成果の概要（英文）：

The goal of this research is to develop a framework to, given large text data, discover patterns which do not appear frequently. To achieve this goal, we review our existing researches from the following two viewpoints:
Mapping of letters for pattern discovery: Using an approximate pattern matching, we have proposed a pattern discovery and evaluated by experiments. In this method, we have found that mapping from several letters into one digit plays an important role.
Metric space for pattern discovery: The goal of this topic is to distinguish patterns from non-patterns. Instead of a rigid metric space, we first find usual substructures, and then we find a pattern as a combination of usual substructures. We have evaluated this method by experiments on genome sequences and Web documents.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,200,000	0	1,200,000
２０１０年度	900,000	0	900,000
２０１１年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,000,000	270,000	3,270,000

研究分野：総合領域
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、テキストマイニング、パターン発見

１．研究開始当初の背景

ICT 技術の普及により、様々なテキストデータが急激に増加しつつあり、これらのデータから有益な情報を抽出する手法の構築が必要である。ここでは、テキスト中に現れるパターンを抽出することを考える。

パターン発見の手法として、データマイニングにおける頻出パターンマイニングを用いることが考えられるが、データマイニングの手法は、厳密なスキーマ上に構築された（関係）データベース上のデータを対象に構築されており、そのままではテキストデータに適用できない。特に、テキストデータの場合、単語や部分文字列といったシンプルなパターンでは、パターンの頻度分布が Zipf の法則に従う、つまり、極端に頻出なパターンは数が少なく、低頻度のパターンが無数にある。頻出なパターンは一般的に不要語など無意味なものだが、頻度が低いものは無数にあるため、この中から意味のある低頻度パターンを峻別する手段がない。パターンの構文を複雑にすれば、有用なパターンが得られるかもしれないが、その定義は任意であり、データのドメインごとに異なると予想されるため、汎用性に欠ける。

２．研究の目的

大量のテキストデータから有用なパターンを発見するという大目標のもと、本研究の目的は、与えられたテキストデータの多くに対してマッチするパターンではなく、相対的に少ない部分データに対して成立する粒度の細いパターンを発見の一般的な手法を構築することである。

データマイニング分野の基本的な手法は、頻出なパターンの発見だが、意外なパターンの発見など少数のデータに対するルールの発見も行われており、本研究もそのような一つである。テキストを対象にした場合、検定ベースの意外なパターン発見が主に遺伝子情報処理の分野でよく研究されてきた。これは、あらかじめ仮定した確率モデルから計算されるパターンの頻度の期待値と、データ中で実際に観測された頻度との差により意外性を定義する。しかし、この場合、適切な確率モデルを定めることが難しい上に、スパースなデータ（多くのテキストはスパースである）をうまく扱えない。つまり、文字や短いパターンに確率を割りあて、より長いパターンの確率を計算することになるため、確率モデルから計算される期待値はどんなに長いパターンに対しても確率を付与してしまう。しかし、実際に観測される長いパターンの種類は少なく、同程度の長さで表されるパターンの多くはデータ中には観測されない。そのため、従来手法では、目的とする粒度の細いパターン発見はできない。

３．研究の方法

本研究の目的を達成するために、本研究の着想となった２つの研究テーマを新たな視点から再度検討し、これらのテーマの手法を発展させ有機的に連結させる方針で研究を行う。

一つ目は、パターンマッチの問題で、元の文字から写像した空間でパターンとテキストの畳み込みを計算することにより、高速にテキスト全体における近似照合のスコアを計算する。二つ目は、通常の文章がなす自然な分布(Zipf 則)からずれるものを特徴的なパターンとして発見する。

これらの研究テーマに共通するのは、パターンとして抽出しない部分を積極的に利用していることである。前者では、全ての写像に対し、パターンとテキストの不一致箇所で、一致・不一致が打ち消しあいパターンとの一致が見つかる。後者も同様に、パターンとして抽出しないテキストの情報を利用して、抽出すべきパターンを見つける。そこで、これらの研究展を「パターン発見への写像導入」と「パターン発見への距離空間導入」という観点で捉えなおし、これらを２つのアプローチで研究を進める。前者は文字やあるいは単語等をどう一つにまとめるかという観点であり、後者の距離空間が導入できれば、各パターンを評価する数値が得られるため、パターンとして抽出するかどうかの評価を行うことが可能になる。

４．研究成果

「パターン発見への写像導入」に対し、以下の成果が得られた。

(1) 写像を用いた近似文字列照合を行う乱択アルゴリズムの改良

最適な写像の生成方法の提案し、精度に関して、既存手法との実験的な評価を行った。また、従来高速フーリエ返還(FFT)を用いていた部分に実数値 FFT を用いることによる高速化を行った。これらの成果は、直接パターン発見を行うものではないが、厳密には一致していないパターンの高速発見のルーチンとして利用可能である。

(2) 近似文字列照合によるパターン発見

上述の近似文字列照合のための乱択アルゴリズムでは、畳み込み演算が重要であった。そこで、この演算を用いてパターン発見を行う着想を得て、剽窃検出の実験を行い、低頻度でも剽窃が適切に発見できることを示し

た。この際、数文字を一つにまとめて写像する処理が本質的であることを示した。
「パターン発見への距離空間導入」に対し、以下の成果が得られた。
(1) 背景集合を用いた例外的なテキストパターン発見手法の提案
距離空間が導入できれば、各パターンに対して評価値が定められることになるが、まず、条件を緩くして、パターンかそうでないかを峻別することにした。ここで、パターンかそうでないかを「普通」か「そうでないか」に分け、「普通」を定義するために背景集合を用いて例外的なテキストパターン発見手法を提案し、既存の例外的な指標(Z スコア)とを比較し、既存手法では見つけられないパタンが発見可能であることを示した。
(2) 部分構造の組み合わせによるパターンの発見の提案
上述の「背景集合を用いた例外的なパターン発見」を発展させ、「普通」を構成する部分構造（語など）を自動的に発見し、「部分構造の組み合わせによるパターンの発見」を提案し、この枠組みを自動生成されたスパム検出や、遺伝子配列に適用し、特徴的なパターンが発見できた。「部分構造発見」のアイデアは、テキストデータ以外にも適用可能であると考え、時系列データに対して適用し、突発的な事象（嵐や地震等）の予兆の発見が可能であることを示した。一般に予兆現象は微少であり、ノイズに埋もれやすく発見が困難である。そこで、データがオフラインで与えられると仮定し、大まかに突発現象の位置を特定してから、特異値分解を元にした手法により予兆を発見できることを示した。
本来は、距離を導入することを想定していたが、最終的には距離ではなく、「普通」を構成する部分構造（語など）の自動的に発見し、「部分構造の組み合わせによるパターンの発見」の提案となったが、距離として厳密に定めるよりシンプルで汎用性が高い手法を提案できた。

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 9 件）

(1). Kensuke Baba, Tetsuya Nakatoh, Yasuhiro Yamada, and Daisuke Ikeda. A Preprocessing for Approximate String Matching, 査読有, Communications in Computer and Information Science 252, 2011, 610-615.
(2). Tetsuya Nakatoh, Kensuke Baba, Yasuhiro Yamada, Daisuke Ikeda. Partial Plagiarism Detection using String Matching with Mismatches, 査読有, Communications in Computer and Information Science 254-6, 2011, 265-272.
(3). 徳永旭将, 池田大輔ら, 変化点検出を応用した時系列データからの突発現象の前兆検出アルゴリズム, 査読有, 情報処理学会論文誌 数理モデル化と応用 4 (No.3), 2011, 14-34.
(4). Takashi Uemura, Daisuke Ikeda, Takuya Kida and Hiroki Arimura, Unsupervised Spam Detection by Document Complexity Estimation with Maximal Overlap Method, 査読有, 人工知能学会論文誌 26(1), 2011, 297-306.
(5). Yuta Taniguchi and Daisuke Ikeda, Graph Clustering Based on Optimization of A Macroscopic Structure of Clusters, 査読有 Lecture Notes in Computer Science Vol. 6926, 2011, 335-350.
(6). Terumasa Tokunaga, Daisuke Ikeda 他 7 名, Onset Time Determination of Precursory Events in Time Series Data by an Extension of Singular Spectrum Transformation, 査読有, INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING 5(1), 2011, 46-60.
(7). Kensuke Baba, String Matching with Mismatches by Real-valued FFT, 査読有, Lecture Notes in Computer Science Vol 6019, 2010, 273-283.
(8). 中藤 哲也, 馬場 謙介, 池田 大輔, 森 雅生, 廣川 佐千男, FFT を用いた不一致を許す文字列照合アルゴリズムの精度評価, 査読有, 情報処理学会論文誌：データベース Vol.2, No.SIG 4(TOD 44), 2009, 1-8.
(9). Daisuke Ikeda and Einoshin Suzuki, Mining Peculiar Compositions of Frequent Substrings from Sparse Text Data Using Background Texts, 査読有, Lecture Notes in Artificial Intelligence Vol. 5781, 2009, 596-611.

〔学会発表〕（計 9 件）

(1). Kensuke Baba, Tetsuya Nakatoh, Yasuhiro Yamada, and Daisuke Ikeda. A Preprocessing for Approximate String Matching, International Conference on Informatics Engineering and Information Science, 2011/11/14, マレーシア
(2). Tetsuya Nakatoh, Kensuke Baba, Yasuhiro Yamada, Daisuke Ikeda. Partial Plagiarism Detection using String Matching with Mismatches, International Conference on Informatics Engineering and Information Science, 2011.11.14, マレーシア
(3). Yuta Taniguchi and Daisuke Ikeda,

Graph Clustering Based on Optimization of A Macroscopic Structure of Clusters, International Conference on Discovery Science, 2011/10/5, フィンランド
(4). Terumasa Tokunaga, Daisuke Ikeda 他７名, Detecting Precursory Events in TIme Series Data by an Extension of Singular Spectrum Transformation, 10th WSEAS International Conference on Applied Computer Science, 2010/10/4, 岩手.
(5). Kensuke Baba, String Matching with Mismatches by Real-valued FFT, International Conference on Computational Science and Its Applications, 2010年3月23日 福岡.
(6). Daisuke Ikeda and Einoshin Suzuki, Mining Peculiar Compositions of Frequent Substrings from Sparse Text Data Using Background Texts, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2009年9月9日 スロベニア

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
中藤 哲也 (NAKATOH TETSUYA)
九州大学・情報基盤研究開発センター・助教
研究者番号：20253502

山田 泰寛 (YAMADA YASUHIRO)
島根大学・総合理工学部・助教
研究者番号：50529609

(3)連携研究者
馬場 謙介 (BABA KENSUKE)
九州大学・附属図書館・准教授
研究者番号：70380681