様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　　８　日現在

機関番号：　１７１０２
研究種目：　基盤研究(B)
研究期間：　2012～2014
課題番号：　２４３０００５９

研究課題名（和文）大規模テキストデータ中の部分構造と稀少な複合パタンの階層的な発見

研究課題名（英文）Hierarchical Discovery of Sub-structures and Rare Patterns of Them in Large Text Data

研究代表者

池田　大輔（Ikeda, Daisuke）

九州大学・システム情報科学研究科（研究院・准教授

研究者番号：００２９４９９２

交付決定額（研究期間全体）：（直接経費）　　7,100,000円

研究成果の概要（和文）：本研究では、テキストデータ中の頻出な部分構造を組み合わせた非頻出なパタン発見の研究を行う。テキストにはZipf則があり、非頻出なパタン候補は無数にあるが、頻出なパタンを組み合わせたパタンの頻度が相対的に多い（絶対的には少ない）ものを発見することで、意味のある非頻出なパタン＝稀少パタンを発見する。このため、既に構築した例外文字列発見の枠組みを拡張と、新たに提案した「純度が高いパタン(pure pattern)」の枠組みで研究を行った。両者とも、細菌のゲノム配列におけるパタン発見での有効性を確認し、さらに、位置情報を持つブログデータやコンテキストの表現、学術論文への関連語発見等への適用も行った。

研究成果の概要（英文）：This research is devoted to finding infrequent patterns of frequent sub-patterns from large text data. Because the text data follows Zipf's law, there exist so many infrequent patterns. Therefore, the goal is quite challenging. Among so many candidates of infrequent patterns, we try to find relatively many, but absolutely few, composite patterns of frequent patterns.
To do so, our two basic approaches are to extend the framework of peculiar patterns we have already developed and to create a new framework based on pure patterns. For both approaches, we evaluated their effectiveness using bacterial genome sequences. In addition to them, we developed mining methods for data in various fields, such as clustering geotagged blogs, context-aware information retrieval, and query expansion for academic theses.

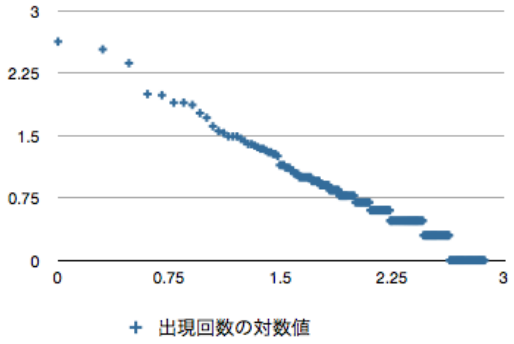
研究分野：　テキストマイニング

キーワード：　例外文字列パタン　純度の高いパタン　purity measure

１．研究開始当初の背景

様々なデータの爆発的な増加に対し、マイニングや機械学習等の様々な手法が開発されている。データベースを対象としたデータマイニングでは、データが属性の組で表現されており、頻出パタン発見問題など頻度ベースの一般的な枠組みが確立されている。一方、テキストマイニングの対象はゲノム配列や自然言語、さらにマルチメディアデータをテキストとみなしたものなど様々で、統一的な枠組みはない。例えば自然言語処理では、単語や品詞などの背景知識を用いて、その分野特有の処理を行うが、同様の手法は別ドメインのデータには適用できない。一部、データマイニングの列挙アルゴリズムを直接テキストに応用した研究もあるが、アルゴリズムのスケーラビリティが主眼で、パタンの有用性についは述べられていない。

テキストデータには広く Zipf 則が成立するため、少数の高頻度パタンと無数の低頻度パタン(ロングテール)がある。下図は、楽曲のデータをテキスト化したものの部分文字列の頻度分布である。



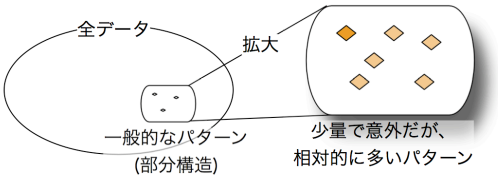
一般に、高頻度パタンは不要語等で、中頻度の多くはそのドメインで常識的なパタン、例えば定型的なフレーズである。

そのため、データマイニングの頻出パタンマイニングや統計的検定ベースの頻度(出現確率)に代表されるような頻度を基準とした有用性の定義では、不要語など役に立たないパタンしか得られない。一方で、不要語などを削除する方法は汎用性に欠ける。低頻度には意外で有用なパタンが隠れている可能性があるが、これを汎用的に見つける手段はなく、稀少パタンを無数の低頻度パタンから峻別する研究は意義深い。

２．研究の目的

本研究では、頻出なパタンを組み合わせた非頻出なパタン発見の研究を行う。非頻出なパタン候補は多いが、頻出なサブパタンを組み合わせたパタンの頻度が相対的に多い(絶対的には少ない)ものを発見することで、意味のある非頻出なパタン=稀少パタンを発見する。

一般に、サブパタンの単位を長くとると、同じ長さで可能な文字列の種類が極端に増えるが、データスパースネスにより、データ中に実際には、観測されるものはごく一部になる。例えば、DNA 配列において長さ 3 の文字列は 64 種類しかないが、長さを 6 にとると 4 千以上になり、実際には現われないものも多くある。さらに、サブパタンを組み合わせたパタン候補はほとんど存在しなくなる。逆に、サブパタンを組み合わせたパタンが存在すれば、非常に稀な現象であり、 低頻度でも抽出すべきと判断できる。つまり、サブパタンにより低頻度部分を拡大していると考えることができる(下図参照)。



３．研究の方法

申請者らのこれまでの研究で、頻出文字列のペアを連接させたパタンを例外文字列として定式化し、これを見つける高速なアルゴリズムを実装し、ゲノム配列から従来の手法では見つからなかったパタンが見つかることを実証した。パタンを見つける文書集合(ターゲット)に加え、背景集合を仮定し、パタンの絶対的な頻度ではなく、ターゲットと背景での頻度の比によりパタンを発見することで、無数にある低頻度文字列から有用なパタンを選別できた。この枠組みはシンプルで、サブパタンやパタンの組み合わせを拡張することで汎用的な枠組みになると考え、この手法を拡張することを基本的な方針とする。

また、有用性を確認する手段として、ゲノム配列を用いて研究を進める。特定の細菌ゲノムを用いた予備実験では、例外文字列の手法により見つけたパタンが、RNA 等の特定の機能を持つ部分配列とよく一致することが分かった。これらの部分配列は低頻度だが重要という意味で、本研究の手法を評価するのに好適である。本研究が目指すのは汎用的なテキストマイニングの手法の確立であり、ゲノム配列以外のデータへ適用し、その有効性を評価する。

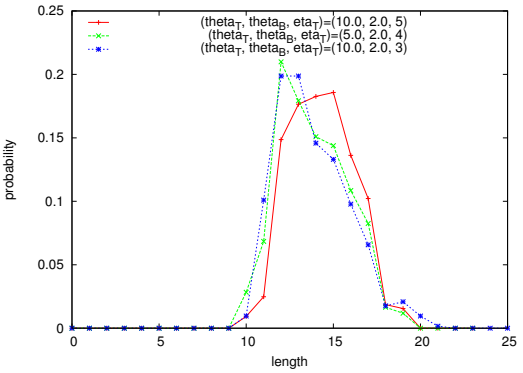
４．研究成果

(1)例外文字列発見の枠組み

特定の細菌ゲノムを用いた予備実験をさらに拡張し、様々な GC コンテンツやゲノムサイズを持つ最近の DNA 配列を対象に、例外文字列を発見する実験を行った。予備実験の結果を受け、tRNA, rRNA をはじめとする RNA と、Transposon, Phage 等を正解の特徴とし、これらとの一致を情報検索の分野でよく用いられる再現率と精度で評価し、よく一致す

ることが分かった(雑誌論文 7, 9)。これらは、従来手法として z-score(平均値からのずれを標準偏差で正規化したもの)と比較し、従来手法では見つけられない稀少パタンであることも示した。

また、興味深い現象として、見つかったパタンの長さは非常に狭い範囲に集中していることが示された(下図参照、雑誌論文 7)。



サブパタンの単位を長くとると、同じ長さで可能な文字列の種類が極端に増え、データスパースネスが起こりだすが、短い場合にはある程度どのパタンも一様に出現する。2 つのパタンの連接が 10〜15 程度であるので、長さが 6, 7 程度のサブパタンが見つかっていることになるが、この長さくらいからデータスパースネスが顕著になる。つまり、低頻度で意味のあるパタンを見つけるために、ある程度長いパタンを組み合わせる必要があるが、あまりに長すぎるとほとんどパタン自体が存在しなくなる。その長さの調整が自動的に行われていると考えることができる。

このように、有用かつ興味深い知見が得られた一方で、実際に利用する際に背景集合と多くのパラメータ設定が必要なことが欠点として認識された。そこで、単一集合のみを用い、パラメータを少なくした枠組みを試してみたが、この場合、得られたパタンはゲノム配列全体にランダムに配置されることが多く、生物学的に意味のある部位にほとんど合致しないことが分かり、汎用的な枠組みとしては使えないことが分かった(雑誌論文 8)。

(2)purity measure による純度の高いパタンの枠組み

我々は、ゲノム配列とは別に、通常の文章からの剽窃検出も行ってきた。剽窃された部分は、ある種のコピーであり、低頻度ながら長い部分が一致するという特徴を持つ。一方で、剽窃する側は、検出手段をかいくぐるために、完全にコピーするのではなく、単語を類義語に変換する、文の順番を変えるなど、様々な編集作業を行い、コピーの度合いを低くするため、重複を単純に検出するだけでは不十分である。このように編集を行なったとしても、もとの文章の部分の大部分は残っており、それを検出できれば剽窃検出ができるのではないか、というアイデアのもとに、ある部分 T における、T の部分パタンがどの程度 T と一緒に出現するかを定量化し、purity とい指標として提案し、剽窃検出におたける有効性を確認した(雑誌論文 10 番)。

この指標は、ある塊 T に着目したときに、その部分が他にはあまり出現せずに、どれだけ純粋に T にのみ出現するかを定量化したものである。これは、例外文字列より必要なパラメータも少なく、シンプルであるが、例外文字列より正確に特徴的な部分配列を捉えていることが分かった。さらに、合致する配列部分は水平伝播と関係する遺伝子によく合致することが分かった。水平伝播は、親から子へと伝播する垂直伝播に対し、よりダイナミックな進化に関連していると考えられており、非常に重要な発見である。実際、上述した Transposon や Phage は水平伝播に重要な役割りを果たすことが知られている。

さらに、ドメイン固有の知識として、純度の高いパタンの発生に関する簡単な仮定を置き、このようなパタンの頻度分布を予測した。具体的には、大きなパタンが部分パタンを含むかどうかの二項分布と考え、ある長さ m 以上であれば常に含まれるというパラメータを用意する。これにより、長さが与えられると、ある部分文字列の純度が測れる。15 種の細菌のゲノムデータを用い、予測の分布と実際の頻度分布と比較することで、予測の有効性を示した(学会発表 1)。

(3)他のデータへの適用

他の分野のドメイン固有の知識として、位置情報付きのマイクロブログからの意味のあるクラスタの発見を行った。位置情報のみからなるクラスタリングに加え、固有の知識としてマイクロブログのテキストを用い、これにより、粒度の細かい階層的なクラスタが得られることを示した (雑誌論文 1)。また、低頻度だが重要なパタンを、論文検索など、専門用語が多く頻度が低くても重要な情報が多い情報検索における専門用語ととらえ、低頻度だが関連の深い検索語の拡張(query expansion)の手法を提案した(雑誌論文 2 等)。他に、コンテキストや分野固有の知識を単語のベクトルとして表すことで、情報検索におけるコンテキストに依存した検索が容易に行える枠組みを提案した(雑誌論文 3 等)。これらの取り組みは、それぞれの分野における稀少なパタン発見や階層的クラスタリング等ではあるものの、統一的な枠組みとして組み込むには至らなかった。

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１０件）

1. Yuta Taniguchi, Daiki Monzen, Sari Ariestien Lutfiana, Daisuke Ikeda: ``Discover Overlapping Topical Regions by Geo-semantic Clustering

of Tweets", Proceedings of the Eighth International Symposium on Mining and Web, pp. 552--557, Mar. 2015.

2. Takehiro Shiraishi, Toshihiro Aoyama, Kazutsuna Yamaji, Takao Namiki, and Daisuke Ikeda: ``Discovering Unpredictably Related Words from Logs of Scholarly Repositories for Grouping Similar Queries", Roger Y. Lee (Ed): Applied Computing and Information Technology, Studies Computational Intelligence Vol.553, pp.48--60, May 2014.

3. Yusuke Hosoi, Yuta Taniguchi, and Daisuke Ikeda: ``Replacing Log-Based Profiles to Context Profiles and Its Application to Context-aware Document Clustering", WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS, Vol. 11, pp.51--60, 2014

4. Yusuke Hosoi, Yuta Taniguchi, and Daisuke Ikeda: ``Generalization User Profiles to Context Profiles and Its Application to Context-aware Document Clustering", Proceedings of the 8th International Conference on Communications and Information Technology, pp.262--270, Jan. 2014.

5. 谷口雄太、池田大輔: ``テキストに対する Purity 尺度の適用と改良", システム情報科学紀要第１９巻第１号, pp.1--6, Jan. 2014.

6. Yuta Taniguchi, Yasuhiro Yamada, Osamu Maruyama, Satru Kuhara, and Daisuke Ikeda: ``The Purity Measure for Genomic Regions Leads to Horizontally Transferred Genes", Journal of Bioinformatics and Computational Biology, 11(6):1343002, Dec., 2013. DOI: 10.1142/S0219720013430026

7. Daisuke Ikeda and Einoshin Suzuki: ``Finding Peculiar Compositions of Two Frequent Strings with Background Texts", Journal of Knowledge and Information Systems, Vol.~41, Issue~2, pp.~499--530, Sep. 2013 DOI: 10.1007/s10115-013-0688-9

8. Daisuke Ikeda: ``Mining Infrequent Patterns of Two Frequent Substrings from a Single Set of Biological Sequences", Proceedings of the 2013 International Conference on Parallel and Distributed Processing Techniques and Applications, Volume I, 136--142, July 22--25, 2013.

9. Daisuke Ikeda, Osamu Maruyama and Satoru Kuhara: ``Infrequent, Unexpected, and Contrast Pattern Discovery from Bacterial Genomes by Genome-wide Comparat ive Analysis", Proceedings of the 4th International Conference on Bioinformatics Models, Methods and Algorithms, pp. 308--311, Feb. 11--14, 2013.

10. Yasuhiro Yamada, Tetsuya Nakatoh, Kensuke Baba and Daisuke Ikeda, ``Mining Pure Patterns in Texts", Proceedings of the 2012 IIAI International Conference on Advanced Applied Informatics, pp.285--290, Sep. 2012.

〔学会発表〕（計　５件）

1. Y. Taniguchi, R. Masui, T. Aoyama and D. Ikeda: ``Probabilistic Model for Purity Values of Bacterial Genome Sequences", 3rd International Conference on Bioinformatics and Computational Biology. Hong Kong, Mar. 2015.

2. 中藤 哲也, 山田 泰寛, 馬場 謙介, 池田 大輔, 廣川 佐千男: 近似文字列照合を用いた剽窃検出手法の評価, 平成 25 年度電気関係学会九州支部連合大会（第 66 回連合大会）講演論文集 No.10-1A-09, 鹿児島大学 2014/9/18.

3. Takehiro Shiraishi, Toshihiro Aoyama, Kazutsuna Yamaji, Takao Namiki, and Daisuke Ikeda: ``Preliminary Results for Discovering Related Words from Logs of Scholarly Repositories", Proceedings of IIAI International Conference on Advanced Information Technologies (CDROM), 30th Nov. 2013.

4. Yuta Taniguchi, Yasuhiro Yamada, Osamu Maruyama, Satru Kuhara, and Daisuke Ikeda: ``The Purity Measure for Genomic Regions Leads to Horizontally Transferred Genes", International Conference on Genome Informatics, Dec. 16--18, 2013.

5. Tetsuya Nakatoh, Kensuke Baba, Yasuhiro Yamada, and Daisuke Ikeda: ``Speed Improvement of the Plagiarism Detection Method'', Proceedings of IIAI International Conference on Advanced Information Technologies (CDROM), Nov. 2013.

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・大学院総合理工学研究科・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