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　 ９ 日現在

機関番号：１６１０１
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００１６２

研究課題名（和文）WWW上の多種メディア情報利用のための数値情報解析

研究課題名（英文）Mining Numbers in Text for Various Kinds of Text Data

研究代表者
吉田　稔（Yoshida, Minoru）

徳島大学・ソシオテクノサイエンス研究部・講師

研究者番号：４０３６１６８８

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：テキスト文書中の数値情報を対象とした検索を目的とし、そのために必要な処理、特に、数値の属性・話題の抽出問題に取り組んだ。様々なメディアの文書の解析を可能にするために、非構造的テキスト（文章など）と構造的テキスト（表形式など）どちらにも適用可能な数値およびその文脈の抽出手法を提案した。確率モデルを用いた教師なし学習を軸とし、Web文書レイアウト解析と表構造解析による属性や単位の抽出を行ったほか、数値を含む特徴的文字列の抽出手法の開発も行った。また、数値をコード化することで、単語と同様の確率モデルの構築、および、数値表現の索引付けが可能となり、メディアの形式に依らない検索が可能となった。

研究成果の概要（英文）：We studied a method for extracting contexts (i.e., attributes or topics) of numbers written in text. Our goal is to develop a system that accept numbers as queries and returns appropriate data from the various kinds of text data such as Wikipedia, Twitter, etc. To achieve this goal, we proposed a method for extracting numbers and their contexts applicable both to unstructured texts (e.g., sentences) and semi-structured texts (e.g., tables). Our method uses unsupervised learning algorithms based on probabilistic generative models for texts to extract attributes and hierarchical topics from Web documents. We also proposed a method to extract corpus-specific number expressions from any kind of text data. For number expressions, we found a coding scheme that can be used both for indexing and probabilistic generative models.

研究分野：テキストマイニング

キーワード：数値情報抽出　レイアウト解析

１．研究開始当初の背景

近年の電子的文書の増大に伴い、テキスト情報を計算機で統計的に処理し、そこから知識を取り出す「テキストマイニング」の研究が様々な研究者により精力的に行われている。内外の研究動向としては、文章を対象とした「言語的解析」（単語の意味や構文の抽出）が主流である。しかしながら、ある種の話題（科学的知見に関わるものなど）においては、数値情報に情報が集約され、その意味理解を行うことが必須であることがある。例えば、放射線の観測値やその人体への影響等に関するテキストを理解するには、テキスト中に示された測定値の正確な取得が不可欠である。この問題に対し、代表者は、テキスト中に文字列の形で書かれた数値情報（数値的情報としてデータベースに格納されていない情報）を数値として適切に処理するための研究を行なってきた。（若手研究(B)「テキスト中の数値表現からの知識発見に関する研究」）

しかしながら、既存研究では、数値情報が解析中のテキストに出現した場合、その数値そのものを取り出すことはできても、「その数値が、より広い文脈の中で、どのような意味を持っているのか」、の解釈は未解決のままであった。すなわち、例えば、同じ「1.0ppm」という数値でも、「コメのカドミウム濃度が1.0ppm」なのか、「水中のフッ化物濃度が1.0ppm」なのかを判別することができなかった。本研究提案は、代表者の上記既存研究をより実用的な技術へ発展させるため、「WWW上のさまざまな文書から、クエリとして与えられた数値情報に関するテキストを、属性や話題まで考慮して取り出す」というタスクの解決を目指す。特に、近年の WWW 上の電子的文書のトレンドとして、ブログ、マイクロブログ（Twitter 等）、ユーザ編集型事典（Wikipedia 等）、といった、様々な形態（メディア）の情報が流通しているため、一つのメディアに偏らず、様々なメディアの文書を横断的に解析するシステムを実現する。これらの各種メディアには、用途の向き・不向きがある（例：Twitter は、流行りの話題を知るのに有効、Wikipedia は、網羅的に知識を得るのに有効、等）ため、例えば、解析中のテキストに「5μシーベルトの放射線が検出された」という文があったとき、Twitter 上の関連テキストを取得しその事実を検証する、あるいは、Wikipedia の関連文書を取得し、観測された放射線量の人体への影響を知る、といった、様々な観点からのテキストの利用が可能になる。

この際に問題となるのが、Wikipedia や HTML ページ、専門的なブログ等、多くのメディアにおいて、数値に関連する情報が、表形式や箇条書き形式により提示されることである。このような対象においては、従来自然言語処理が主な対象としてきた、「整った文章」からの情報抽出だけでなく、「構造的テキスト表現（表形式など）」の解析を行うことが必要となる。この「構造的テキスト表現」は、代表者が従来より研究を行なってきた分野であり（若手研究(B)「ＨＴＭＬ文書からの論理構造自動推定に関する研究」）、そこで得られた技術を発展させることで、構造的テキスト・非構造的テキスト両者へ横断的に適用できる、数値の文脈推定アルゴリズムを構築する。

２．研究の目的

本研究は、クエリとして与えられた数値と、テキスト中の数値の同義性（類似性）の判定という視点から見ることができる。この処理は、「数値そのものの同義性」の判定と、「文脈の同義性」の判定に大別することができる。前者の「数値そのものの同義性」については、従来の単語同義語抽出アルゴリズムが応用できる。また、同じ数値が違う表記で表現される問題もあるが、漢数字とアラビア数字、コンマの有無など、基本的な異表記については、これも代表者の既存研究によるシステムにおいて認識可能である。

後者の「文脈の同義性」を判定するための技術として、確率的生成モデルによる教師なし学習手法を用いる。確率的生成モデルは、近年、「トピックモデル」として注目を集め、テキストを自動的に分析し、意味の似た単語や意味の似た文書を確率的に特定できるモデルとして、テキスト解析の分野で広く使われ、その有用性が認められている。このような確率的教師なし手法を用いることで、「『意味の同じ語』同士のマッチング」や「話題（トピック）の同じ文書の判定」、さらには、「属性名の判定」も可能となる。確率的生成モデルは、単語の出現を確率的にモデル化するものであり、単語の出てくる文字列ならば様々な対象に適用可能のため、これを、文章、箇条書き、表形式それぞれのモデル化に適用し、その相互利用を目指す。このさい、以下の３つの課題を解決する必要があり、その研究を行う。

（１）「数値属性名獲得」の問題：数値単独だけでは、その意味を捉えることは難しい。その数値が「何の数値なのか」を表す情報（「属性名」）を、文章ならば構文情報、表形式なら表の構造等を用いたモデル化を行うことにより抽出する。

（２）文章と文章以外の統合的モデルの構築法：表形式、箇条書き形式、文章を統一的に扱えるモデルの提案。さらに、同一文書内の、表形式と文章の関係などの参照関係をモデル化する手法についても研究を行う。

（３）「数値のトピックモデル」の研究：数値情報のような連続値を、単語情報のような離散値と統一的にモデル化するためのモデルの構築。

３．研究の方法

上記目的を達成するため、既存のトピックモ

デルを拡張することを検討したものの、検討の結果、特に、LDA 等の既存のトピックモデルを、そのまま表形式中の数値情報を対象として適用した場合、問題が多いという知見を得た。これは、表形式が限られた構造の中で、ある種のテンプレートに沿って情報を提示する形式であるのに対し、LDA 等の文章を対象としたモデルは、単語が比較的自由に配置されることを念頭に置いているため、自由度が高く、決まりきった配置を分析する目的には適していないこと、また、属性名が、表形式中に通常一回のみ表記されるという性質が、確率的生成モデルに適さないという理由による。そこで、トピックモデルに拘らず、各形態に適した確率モデルを模索することとした。
まず、コーパスの取得を行い、その後、上記課題について、
（1）数値属性名獲得：レイアウト解析による文脈抽出、表形式解析による属性抽出、Distant Supervision を用いた単位推定
（2）文章と文章以外の統合的モデル：数値を適切に抽象化することによる、数値の文字列としての取り扱い
（3）「数値のトピックモデル」の研究：Polya Tree 確率モデルを応用した、テキスト中の数値に適した確率モデルの構築
という方針で研究を進めることとした。
コーパス取得は、Wikipedia に関しては、提供されているアーカイブを用いる。また、Twitter のデータに関しては、代表者の所属する研究室でも蓄積されたデータが存在するため、それの利用も検討するほか、Twitter 社が提供する API による独自データの取得も行う。
（1）数値属性名獲得に関する研究：
レイアウト解析による数値属性名獲得を行う。
レイアウトの特徴（類似する意味の単語には類似するレイアウトが与えられる）を捉えるのに適している確率モデルを開発し、これを利用して、教師なし学習により、大量の Web 文書から文書構造を取り出す。これにより、テキスト中の数値について、それがどのような位置づけであるかの把握が容易となる。また、表形式についても、同様に適切なモデルを検討する。
また、データを観察した結果、数値の単位が省略されるケースが多数存在し、この省略された単位の推定も行う必要があることがわかった。上記手法により得られた文脈情報を利用し、表形式中の各セルで省略された単位名を推定する手法を開発することとした。
（2）文章と文章以外の統合的モデル
データ中の数値表現を観察する中で、テキスト中に表記された数値表現は、「桁数」と「上位桁の数字（有効数字）」による、浮動小数点表現が適しているという結論に至った。適切な桁数を設定し、上記手法により、数値をコード化することで、数値の索引化が可能となり、文章と文章以外、さらに異なるメディア間でも容易にマッチングが可能となる。
このさい、数値そのもののみならず、単位等、それと連接する文字列が、数値の意味特定に重要となってくる。実際のデータを観察したところ、Wikipedia 表形式、Twitter それぞれで、各メディアに特徴的な定形表現が存在し、数値データの多くがこれら定形表現に基いて表現されていることがわかった。そこで、任意のテキストから頻出する定形表現を高速に取得する手法を開発する。
（3）「数値のトピックモデル」の研究
上記の数値のコード化により、実数値である数値情報を、単語と同様の離散的情報として取り扱うことが可能となる。この離散的表現に適したモデルとして、Polya Tree 確率モデルを応用した「テキスト中の数値表現のための確率モデル」を構築する。このモデルを用い、数値クエリと表形式中の数値表現の適合性評価や、表形式中で省略された単位の推定等の応用を試みる。
また、研究成果として得られるアルゴリズムを実際に利用し、従来のテキストからの数値検索システムを高度化するなどの、応用システムを構築し、その実用性について検討する。

４．研究成果

（1）数値属性名獲得に関する研究（レイアウト解析による数値属性名獲得）
①Web 文書の構造解析
Wikipedia 以外の、明示的にカテゴリ情報が与えられていない Web 文書に関してカテゴリ情報の付与を行う「レイアウト解析」に関して、これに適したモデルとして、「階層ベイズモデル」という確率モデルが適していることがわかった。階層ベイズモデルを用いることにより、「同一文書中で似た意味の単語は、似たレイアウトを伴う」という性質と、「属例や見出し等、単語の役割によって、使われるレイアウトに特徴がある」という性質を、同時にモデル化できる。このモデルによる Web 文書の構造推定を、Collapsed Gibbs Sampling の実装、さらにその並列化により高速に行う手法を開発した。実際のデータを用い、ヒューリスティクスを利用したベースライン手法との比較を行ったところ、ベースライン手法から精度を改善させることを確認した。本研究成果は、アジア地区のデータマイニングに関する定評ある国際会議 PAKDD2014 にて発表を行った。（学会発表４）
また、Wikipedia 表形式について、それを含むテキストの見出し抽出（構造解析）を行い、さらに、記事のタイトルと、そのタイトルの上位語を Wikipedia のリンク構造から取得し、上記手法を利用した表形式からの属性抽出と組み合わせることで、各セルの意味的位置づけを、効率的に表現する手法を開発した。
②表形式の構造解析
表形式に対するベイズ的確率モデルに関して検討を行った。いくつかのモデルを検討し

た結果、セル間の依存関係を、表形式全体で統一するのではなく、各セル毎に独立させるという方針で、Pachinko Allocation に基づくモデルを作成することで、表形式の属性・属性値構造を推定できるモデルを考案し、Collapsed Gibbs Sampling を行うことによって、実際にある程度構造推定が行えることを確認した。しかしながら、今回用いた Wikipedia のデータでは、比較的単純な表形式が多かったため、実際には、代表者の既存研究を利用した、より単純なモデルを利用して、高速に構造推定を行っている。
また、上記文脈情報を利用し、表形式内の各セルを、独立したセルとして取り出すことが可能となった。しかしながら、特に Wikipedia では、表形式中の数値について、その単位が省略されることが少なくない。この問題に対し、単純なヒューリスティクス以上の手法が必要であることがわかった。単位の省略されたセルについて、その文脈情報および数値そのものの情報をモデル化し、最も確率の高い単位を推定する手法を開発した。
（２）文章と文章以外の統合的モデル
数値を適切にコード化し処理することで、サンプリング等の技術を用いずに高速な検索が可能になることを発見し、実際に検索システムに応用することを試みた。システム試作の結果、問題なく数値の高速検索が行えることがわかったため、これを数値モデル化に応用することとした。
また、この「数値データをコード化して文字列検索可能にする手法」を用い、表形式と Twitter テキストを、同一のコード化を通じて統合的に取り扱い可能にするという目標のもと、Wikipedia 表形式と、Twitter 中の数値を同時にコード化することを試みた。
実際のデータに適用したところ、Wikipedia 表形式、Twitter それぞれで、各メディアに特徴的な定形表現が存在し、数値データの多くがこれら定形表現に基いて表現されていることがわかった。そこで、任意のテキストから頻出する定形表現を高速に取得する手法を開発した。この成果は、アジア地区の情報検索で定評ある国際会議 AIRS 2014 にて発表を行った。（学会発表２）この手法は、先に開発した数値のコード化手法と組み合わせることで、テキスト中の数値データにも適用可能である。特に、表形式中の行・列毎に定形表現を抽出することで、各行および列を少数のパターンで記述する文字列の集合を取り出すことができた。（例えば、Wikipedia における日本の最高気温ランキングの表を与えたときに、「気温」の列の値から、「4*°C」というパターンを自動的に抽出する等。）同様に、Twitter のデータに適用することで、数値を含む定形表現を自動的に抽出できることがわかった。（地震速報や、マンションの宣伝等）
（３）「数値のトピックモデル」の研究
また、このコード化が、Polya Tree と呼ばれる確率モデルと相性が良いことを発見し、これを応用した表形式の確率モデルを設計し、サンプリングにより表形式の構造を推定するプログラムを開発した。また、表の構造推定のみならず、上記の単位推定手法においても、この確率モデルを利用している。そのほか、最終的な検索システムの出力形式として、カテゴリ名のみならず、表形式そのものを出力するという「数値→表形式検索システム」のシステムの試作も行った。これは、前述の数値のコード化を用い、連続するセルについて、「数値の並びの尤もらしさ」をモデル化することにより、入力された数値、あるいは数値の列に対し、最も適すると判定した表をランキングで出力するものである。
（４）その他関連する研究
テキスト外にメタデータとして存在する数値情報の活用の可能性に関しても検討を行った。具体的には、テキストに数値情報や位置情報が紐付けられていた場合に、その情報がテキスト中の単語とどのように関連付けられているかをマイニングする手法について研究を行った。（学会発表５）その他、数値列が複数存在する表に対し、その相関を抽出するシステムの試作も行った。例えば、人口と面積の間の相関がそれほど強くない事、野球の記録において、勝利数と敗戦数の関係に緩い相関があること等が観察できた。
（５）テキスト中の数値マイニングシステムの高度化
上記研究成果を、数値表現検索システム Qiwi として実装した。システムは、上記の数値コード化を用いて高速化を行ったシステムを実装した。上記モデル化を利用し、テキスト中の数値を、コード化された形に変換して解釈し、並べ替え・索引付けを行うシステムを実装した。これにより、従来問題となっていた、数値で開始するクエリで検索時間が膨大になる問題に対し、ほぼ解決することが可能となった。併せて、テキスト中に存在する数値の傾向を調査するためのツールとして、数値表現の分布を高速にマイニングし、その分布グラフを表示する機能を実現した。（図１）

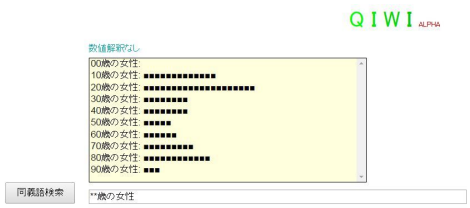

図１：数値表現のテキスト中分布の高速表示

また、これとは別に、数値クラスタリングを行い、それぞれのクラスタ毎に特徴文字列表現を実装し、速度よりも抽出内容を重視したシステムも実装した。（図２）

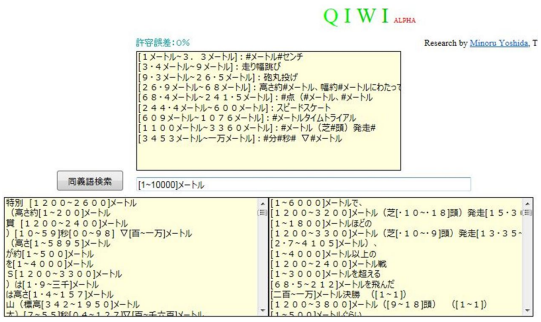

図２：数値の自動クラスタリング、及び特徴的な文字列の抽出

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　２　件）

(1) Kazuyuki Matsumoto, Kyosuke Akita, Xielifuguli Keranmu, Minoru Yoshida and Kenji Kita : Extraction Japanese Slang from Weblog Data Based on Script Type and Stroke Count, Procedia Computer Science, Vol.35, No.2014, pp.464-473, **査読有**, (2014).

(2) 藏本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎, 吉田稔, 中川裕志, “新聞記事のテキストマイニングによる長期市場動向の分析”, 人工知能学会論文誌 Vol. 28, No. 3, pp.291-296, **査読有**,(2013).

〔学会発表〕（計　１６　件）

(1) Kazuyuki Matsumoto, Sasayama Manabu, Qingmei Xiao, Fujisawa Akira, Minoru Yoshida and Kenji Kita : Reranking the Search Results for Lyric Retrieval Based on the Songwriters' Specific Usage of Words, The proceedings of the 4th international conference on electronics, communications and networks (CECNet2014), **査読有**, Beijing (China), 2014 年 12 月 14 日.

(2) Minoru Yoshida, Kazuyuki Matsumoto, Qingmei Xiao, Xielifuguli Keranmu, Kenji Kita and Hiroshi Nakagawa : Extracting Corpus-Specific Strings by Using Suffix Arrays Enhanced with Longest Common Prefix, Proceedings of the 10th Asia Information Retrieval Society Conference (AIRS 2014), Kuching (Malaysia), **査読有**, LNCS 8870, pp. 360-370, Kuching, 2014 年 12 月 5 日

(3) Kazuyuki Matsumoto, Fuji Ren, Qingmei Xiao, Minoru Yoshida and Kenji Kita : Emotion Predicting Method Based on Emotion State Change of Personae according to the Other's Utterance, Proceedings of the 3rd IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS2014), **査読有**, Hong Kong (China), 2014 年 11 月 29 日

(4) Minoru Yoshida, Kazuyuki Matsumoto, Kenji Kita and Hiroshi Nakagawa : Unsupervised Analysis of Web Page Semantic Structures by Hierarchical Bayesian Modeling, Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2014, Part II, LNAI 8444, pp.572-583, Tainan (Taiwan), **査読有**, 2014 年 5 月 16 日.

(5) 加藤宏紀, 荒牧英治, 宮部真衣, 吉田稔, 佐藤一誠, 中川裕志, “ソーシャルメディアからの地域固有表現の抽出”, 電子情報通信学会, 言語理解とコミュニケーション, NLC2012-38, pp.29-34, 査読無, 東京工業大学（東京都・目黒区）, 2012 年 12 月 19 日

(6) 吉田 稔, 杉浦 隆博, 廣川 敬真, 山田 剛一, 増田 英孝, 中川 裕志：テキスト中の数値情報マイニングと情報編纂：MuST 参加から見えてきたもの, 人工知能学会第 26 回全国大会, 3B3-NFC-4-1, 山口県教育会館（山口県・山口市）, 査読無, 2012 年 6 月 14 日

６．研究組織

(1)研究代表者

吉田　稔　　（YOSHIDA MINORU）

徳島大学・大学院ソシオテクノサイエンス研究部・講師

研究者番号：40361688