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　６月　１２日現在

機関番号：１２６０１
研究種目：若手研究(B)
研究期間：2010～2011
課題番号：２２７００１３７
研究課題名（和文）　テキスト中の数値表現からの知識発見に関する研究

研究課題名（英文）　Knowledge Discovery from Numbers in Text

研究代表者
吉田　稔　　（ YOSHIDA MINORU ）
東京大学・情報基盤センター・助教
研究者番号：40361688

研究成果の概要（和文）：
テキスト中の数値表現を適切に取り扱い，数値と言語の統合的なマイニングを行うための基盤技術の研究を行った.具体的な方針として,テキストを接尾辞配列により索引付けし，そこで数字列に対し，数値としての検索が行えるように拡張を行った．このシステムを，大規模なテキストに適用できるよう高速化し，これにより，文字列と数値の関係を対話的に取得できる基盤を構築できた．また，応用先として，数値を多く含む業務レポート等に対するテキストマイニングの研究を行った.

研究成果の概要（英文）：
We studied a method for processing numbers written in text to discover relations between words and numbers. We indexed texts using suffix arrays augmented with functions for searching digits as numbers with the queries being able to include range of numbers. The search function can be performed in reasonable time for large text, which enabled us to obtain the relations between words and numbers interactively from such texts. We also studied methods for mining the texts that contain many numbers.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,400,000	420,000	1,820,000
２０１１年度	800,000	240,000	1,040,000
年度			
年度			
年度			
総　計	2,200,000	660,000	2,860,000

研究分野：自然言語処理
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理、数値情報、テキストマイニング、接尾辞配列

１．研究開始当初の背景

近年、文書の電子化の進展、組織からのWWWによる情報発信の一般化に伴い、企業や大学等、様々な組織において、膨大な量の文書データが電子的に蓄積されている。これら蓄積された文書集合をどのように活用するかは、テキストマイニングの問題として、現在盛んに研究が行われている。文書の中に記述された事象においては、「価格」や「年齢」、「時刻」など、多くの数値に関する情報が含まれており、これら数値情報は、そのテキストにとって非常に重要な意味を持つこ

とが少なくない。例えば、航空会社の業務レポートでは、飛行機の高度や燃料の量・乗客数等に関する記述が多く見られ、これらが数値情報として表現されている。これらを適切に解析することで、トラブルの起こりやすい高度等、重要な情報のマイニングが行えるようになる。
しかしながら、一般的な文書集合では、数値情報が独立してデータベース等に格納されることは稀であり（「文書作成日」など、特に定形化された例外を除く）、多くの場合は、文章の形で、テキストの中に埋め込まれている。そのため、データマイニングの分野での、数値・テキストの同時マイニング手法である「データベースに与えられた数値を、テキストと併せて分析する」という手法は、ここでは適用できない。また、現状の自然言語処理やテキストマイニングの分野では、数値は「単語として解釈できないノイズ」として扱われる場合が多い。例外として情報抽出の分野では、日付や価格等、特定の数値情報についてパターンを用いた抽出を行う研究が多く存在するが、数値情報一般の取り扱いや、抽出された数値情報の他のタスクへの活用についての研究は少ない。これに対し本研究では、テキストデータの中に埋め込まれた数値「テキスト中の数値表現」を、適切に数値として解釈し、検索や分析を可能とし、さらに、抽出された数値の、意味解析等への応用を可能とする、テキスト中の数値表現を最大限に活用する新たなシステムの実現を目指す。
数値情報は、通常の単語と異なり、数値間の順序（大小関係）や類似度（値の近さ）等、特有の性質を有している。本研究では、これらの数値の性質を具体的に扱え、しかも通常の単語と同等の可用性を備えた離散的表現として、「数値の範囲表現」に着目する。例えば「4 歳」「5 歳」「6 歳」という数値表現の集合を「[4~6]歳」という範囲表現により扱うことで、複数の数値の集合を一つの単語のように扱うことが可能となり、網羅性の向上につながる。
数値範囲をテキストマイニングに用いるためには、まず文書中の数値表現を取り出し（後述「数値語彙の獲得」）、その後類似する数値表現をまとめる（後述「数値異表記の正規化」「数値範囲の推定」）ことによる、「数値範囲の抽出」、また、抽出された数値範囲を利用するための、「数値範囲の検索」の手法が必要であり、本研究では、これらのタスクについて個別に研究開発を行う。また、具体的応用として、「同義語獲得（単語意味類似度測定）」「共起語取得（単語意味ベクトル取得）」という 2 種類のテキストマイニングタスクにおいて、数値範囲表現が扱えるように拡張を行う。

２．研究の目的
数値と言語のシームレスな取り扱いを目指し、具体的な研究目的として、以下の４つを設定した。特に、サイズの大きいコーパスに対しても適用可能なアルゴリズムの研究を行う。
（１）数値語彙の獲得：
数値の周辺文字列を、適切に切り出す。そのために、「４歳」「午前５時」等、数値を含む文字列を、どこで切り出せば意味を持つ単位となるかの判定を行う。
（２）数値異表記の正規化：
「二千円」と「2000 円」、「10 キロメートル」と「10000 メートル」、「1 時間 30 分」と「90 分」のような、「同じ意味だが違う表現」となる数値表現を正規化する。
（３）数値範囲の自動推定：
数値の集合が与えられたとき、そこから範囲表現のために適切な数値の部分集合を推定する。文書集合では明示的に数値範囲が与えられているわけではないため、そこから自動的に数値範囲を取り出す手法が必要である。
（４）高速数値範囲検索：
ユーザが任意に与えた、範囲表現を含む文字列を検索し、頻度等の情報を取得する機能の実現。本提案では、接尾辞配列(Suffix Array)を文書集合への索引構造として用いることで、オンデマンドな頻度計測や連接文字列取得を実現する。
　さらに、これらの基盤技術をもとに、数値範囲を含んだ文字列を対象にした、同義語抽出や共起語抽出システムの開発を行う。
提案システムは、「ユーザー（人間）」からの検索要求の他、「他のシステム」からの検索要求（呼び出し）も扱えるようなモジュールとして提供する。このため、他のシステムの計算の一部として使えるような、高速なシステムの実現を目指す。
「他のシステム」からの呼び出しでは、数値範囲ではなく、文書中の数値文字列（例えば「17 歳」）がそのまま用いられることが想定されるため、網羅性向上のため、類似する数値範囲（例えば「[15~19]歳」）を返す機能も実現する。

３．研究の方法
（１）コーパス取得
提案システムが対象とする文書集合として、Wikipedia、東京大学 Web ページを取得し、これを用いる。これらは、数百メガかそれ以上というサイズを持つコーパスであり、単純なアルゴリズムでは、実時間で数値情報マイニングを行うことが困難である。これらのコーパスに対して高速に数値テキストマイニングを行うことを目標とする。
（２）基盤技術の開発

研究目的において挙げた以下の４つの問題について、それぞれ研究・開発を行った。
①数値語彙の獲得：
数値を含む文字列パターン「数値複合語」を獲得する。「数値範囲を含む文字列」に対するスコア付けを行い、スコアの高い文字列パターンを取得するという、自動獲得手法の開発を行う。このさい、頻度計測等に数値範囲検索機能を用いる。また、連接文字パターンの分岐数測定等、既存の単語境界推定で用いられる指標を使い、数値範囲語彙の切り出しに応用する。
②数値異表記の正規化：
漢数字や、コンマを含む数値等、同じ数値が様々な表現をとることに対処するため、人手により数字文字列を解釈し、数値に直すルールを記述する。これを検索に組み込むことで、「異なる文字列だが同じ数値」を同じものとして検索できるようになる。
③数値範囲の自動推定：
数値範囲推定のための「数値クラスタリング」について研究を行う。この場合、テキストマイニングへの応用を考え、できる限り高速であることが望ましい。数値クラスタリングの際に問題となるのは、適切なクラスタ数が与えられる数値集合によって様々であるということである。このため、クラスタ数を自動的に決定できる確率モデルとしてディリクレ過程混合モデル（DPM）を採用し、DPMに基づく高速クラスタリング手法を研究する。さらに、大規模コーパスにおいて、クラスタリングを行う数値のリストが膨大になることを想定し、これに対するクラスタリングを高速化するため、数値リストのサイズに対し超線形の計算量でクラスタリングを行う手法についても検討する。
④高速数値範囲検索：
本研究では、テキストを扱うための索引構造として、接尾辞配列（Suffix Array）の採用を検討する。接尾辞配列は、高速かつ省メモリな文字列検索のための索引構造である。このデータ構造は、文字列出現パターンを表す木構造として用いることができ、本研究が課題とする文脈文字列抽出に対して、「連接文字列パターンの高速な取得」という形で大きな貢献が期待できる。接尾辞配列は、文書中の部分文字列に対し「辞書順ソート」を行ったものであるが、例えば、これを「数値順ソート」（数値の部分は数値として解釈してソート）に変えることで「数値範囲」を含む文字列についても、通常の文字列と同様の検索（二分探索による高速検索）が可能となる。また、接尾辞配列は、全文検索の索引構造であるため、提案システムのように、予めどのような検索語が用いられるのかが確定しない場合に特に適している。
■「数値範囲」以外の数値取り扱い方針についての検討
実際には、数値集合を「範囲」を用いて扱うことが適切でない場合が存在することも想定される。このため、例えば、ある分布（正規分布や、その他指数分布族等）に基づき、与えられた数値に近い順に高いスコアを付す検索手法等、「範囲」以外の数値取り扱い手法についても検討を行う。

（３）テキストマイニングへの応用・システム開発および評価
数値的文脈を、様々なテキストマイニングに応用し、数値的情報の有効性を調べる。例えば、開発した基盤技術を、共起語の取得、同義語の取得等のテキストマイニングタスクへ応用する。また、実際にユーザーや他のシステムからそれらの機能を使用できるシステムの開発を行う。

４．研究成果
（１）テキスト中の数値をマイニングするための基盤技術の開発
①高速クラスタリング手法の開発：
テキスト中に言葉とともに出現する数値の範囲を自動的に推定するために、確率モデルに基づき、クラスタ数を自動的に推定する高速クラスタリング手法の開発を行い、さらに、数値リストのサイズが膨大になった場合にも対処できるよう、二分探索を応用し、高速にクラスタを発見する手法の開発にも取り組んだ。この結果、Wikipedia のような膨大なテキストを対象とした場合でも、問題なく数値クラスタリングを行えることを確認した。
②高速検索のための索引付け手法の開発：
数値範囲を用いた検索を行えるような、テキスト全文検索を拡張した索引付けを行う手法の研究を行った。提案手法は、与えられたテキストを改変することなく、少量の追加索引構造を用いるだけで、高速な数値範囲検索機能を実現する。さらに、クエリの種類に応じた場合分け、特に、数値で開始するクエリについて、「数値のみの場合」と「それ以外の場合」に場合分けを行い、追加データ構造を用意することにより、従来よりも高速に連接文字列を取得することに成功した。これにより、様々な数値範囲クエリに対しリアルタイムに反応することが可能となった。また、数値集合を「範囲」を用いて扱うことが適切でない場合に対処するため、使用したモデル（無限混合ガウス分布）のパラメータ（分散）の調整による検索結果の数値範囲の変化の調査や、パラメータの自動推定手法についての調査を行い、適切なパラメータを与えることで、数値範囲ではなく数値そのものを文脈として取り出せることを確認した。
③数値語彙の獲得：

この索引構造を用いることで、既存の連接文字列マイニングアルゴリズムを、数値範囲を含む文字列のマイニングに拡張することができる。これにより、「数値語彙」（数値範囲を含む語彙）の獲得が可能となった。
④数値異表記の獲得：
異なる文字列が同じ数値を表す「数値異表記」の問題に関しては、コーパスを改変せず、追加索引構造の作成アルゴリズムを拡張することで、柔軟に対処するアルゴリズムの開発を行った。これにより、全角数字、漢数字など、様々な形式での数値の取り扱いが可能となった。コーパスとしては Wikipedia、東京大学 Web ページに関して、上記アルゴリズムが適用できることを確認した。その他、新聞記事，企業の業務レポート等のテキストへの本手法の適用も試みた。

（２）テキストマイニングへの応用
開発したシステムは、数値範囲という概念を用いることで、数字文字列を通常の文字列と同様に扱いつつ数値としての性質も利用できるという汎用性の高いシステムとなっており、他のシステムからの利用も容易である。本システムのテキストマイニング応用への精度評価として、同義語抽出の文脈情報に数値範囲を用いる手法に関して詳細な実験を行い、適切な閾値を用いて数値範囲を用いるか否かの切り替えを行うことにより、同義語抽出の精度を向上させることができることを確認した。
並行して、このような技術の応用先として、数値データを含むテキストに対するテキストマイニングの研究を行った。具体的には、機器異常診断に際して蓄積された業務レポートを対象とした。この業務レポート中のテキストは、機器に関する様々な数値情報を含み、これに対して、レポートのクラスタリングや要約を行う手法についての研究も行った。クラスタリングや要約に際しては、数値の単位情報等を特徴量として用いることを行った。
また、これと同時に、比較対象として、株価や風邪薬販売量等、テキスト外の数値と言語を対応づけるための研究も行った。
その他のテキストマイニングタスクの一例として、「検索における同姓同名問題」にも取り組み、文脈情報を適切に重み付けするための手法に関する研究を行った。

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４件）
①吉田稔、池田雅紀、小野真吾、佐藤一誠、中川裕志．二段階クラスタリングを単語重み付与に応用した人名曖昧性解消，日本データベース学会論文誌，査読有，Vol.9, No.2, 2010, 19-24.

②吉田稔，中川裕志： テキストマイニングの活用，情報の科学と技術，査読無， 60 巻 6 号，2010，230-235.

③Minoru Yoshida, Masaki Ikeda, Shingo Ono, Issei Sato, and Hiroshi Nakagawa. Person Name Disambiguation by Bootstrapping, Proceedings of SIGIR-2010 (the 33rd Annual ACM SIGIR Conference)，査読有，2010, 10-17.

④ Minoru Yoshida, Issei Sato, Hiroshi Nakagawa, Akira Terada. Mining Numbers in Text Using Suffix Arrays and Clustering Based on Dirichlet Process Mixture Models, Proceedings of PAKDD-2010 (The 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining), Springer (LNAI 6119)，査読有，2010，230-237.

〔学会発表〕（計１１件）
①谷田和章，荒牧英治，佐藤一誠，吉田稔，中川裕志．ソーシャルメディアによる風邪流行の予測，言語処理学会 第 18 回年次大会，広島，2012 年 3 月 15 日.

②吉田稔，中川裕志，渋谷久恵，前田俊二．テキストマイニングによる機器異常診断支援の試み，第 4 回データ工学と情報マネジメントに関するフォーラム(DEIM 2012)，F5-4，神戸，2012 年 3 月 4 日

③吉田稔，中川裕志，石田智也，中嶋啓浩，松井藤五郎，和泉潔，池田翔，本多隆虎，ニュース記事クラスタリングによる取引高予測の試み，人工知能学会第 25 回全国大会，2H1-OS18-7，盛岡，2011 年 6 月 2 日

④Minoru Yoshida, Hiroshi Nakagawa, Web People Search: Person Name Disambiguation and Other Problems (Tutorial), the 2nd Asian Conference on Machine Learning (ACML 2010)，2010 年 11 月 8 日

⑤Minoru Yoshida, Shin Matsushima, Shingo Ono, Hiroshi Nakagawa. ITC-UT: Tweet Categorization by Query Categrization for On-line Reputation management. WePS-3, CLEF 2010 Labs, 2010 年 9 月 23 日

〔図書〕（計１件）
①Minoru Yoshida, Hiroshi Nakagawa, Akira

Terada. On-demand Synonym Extraction Using Suffix Arrays, Chapter in Book: ″Information Extraction from the Internet″, Nan Tang (editor), iConcept Press, 2011, 73-87.

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

６．研究組織
(1)研究代表者
吉田　稔　　(YOSHIDA MINORU)
東京大学・情報基盤センター・助教
研究者番号：40361688

(2)研究分担者

(3)連携研究者