3B3-NFC-4-1

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

吉田 稔[*1] 杉浦 隆博[*2] 廣川 敬真[*2] 山田 剛一[*2] 増田 英孝[*2] 中川 裕志[*1]
Minoru Yoshida Takahiro Sugiura Takamasa Hirokawa Kouichi Yamada Hidetaka Masuda Hiroshi Nakagawa

[*1]東京大学情報基盤センター Information Technology Center, University of Tokyo
[*2]東京電機大学 Tokyo Denki University

We participated in MuST to discuss about research on relations between texts and numbers, and also participated in the MuST T2N subtask with the system that automatically generates graphs from texts. In this paper, we summarize the graph generation system along with other systems we presented at MuST, and also report the subsequent development of research.

1. はじめに

本稿では，我々の「動向情報の要約と可視化に関するワークショップ」(MuST)[1] への参加，特に，NTCIR-7 で行われた MuST-T2N タスクへの参加を通じて得られたものと，その後の関連研究の発展について述べる．

我々は,「テキストと数値の関連」というテーマ，特に，**テキスト中に書かれた数値情報**に着目し，その活用に関する研究を行なっている．研究の目的としては，テキスト集合を，**数値という切り口**からまとめることを主眼としており，その具体的な手段として,「情報抽出」と「情報検索」の二つの視点から取り組んでいる.「情報抽出」という視点からは，テキスト集合から (統計量名, 統計量) ペアを自動抽出し，さらに，それを同一の統計量名でまとめることで，テキストからグラフを自動抽出する，という研究に取り組んだ.（詳細については，JSAI2009 における予稿 [2] を参照されたい.）この研究は，MuST が対象としている「動向情報」の抽出と共通するものであり，MuST へは，主にこの研究を中心に，テキスト中における数値情報の扱いを議論できる場として参加した.「情報検索」という視点からは，**数値をクエリとした文書ランキング** [3] の研究をもとに，現在,「数値をクエリとした全文検索，およびそれに基づくテキストマイニング」の研究に取り組んでいる．

以下，本稿では，まず「グラフ自動作成システム」について解説し，その後,「数値検索」に関する研究を紹介し，MuST で得られた知見がどのように活かされているかを述べる．また,「テキスト中の数値」以外の研究テーマとして参加した,「テキストと経済指標の関連」に関する研究についても，その後の発展について簡単に触れる．

2. テキストからの自動グラフ作成

本研究は，与えられた新聞記事から自動的にグラフを作成することを目的とする．システムへの入力は記事の集合であり，システムは，記事集合から統計量のグラフを作成する.（作成されるグラフは１つとは限らない.）本システムは，MuST T2N タスクに参加した．

このシステムは「与えられた新聞記事から，統計量を発見するシステム」「発見した統計量をまとめ，グラフを作成するシステム」の２つのサブシステムに分けられている．MuST T2N タスクは，与えられた同一の話題の複数の記事から統計量名のグラフを作成するというものである．T2N タスクにおいては，グラフを作成するための統計量名が予め与えられていたが，これに対し，我々のグラフ作成システムは，**自動的に統計量名を発見し，グラフを作成する．**

システムは，統計量名の判断において機械学習を行なっており，その学習データとして，MuST における研究用データセットにある動向情報コーパス（通称：MuST コーパス）を使用した．このコーパスは，毎日新聞社の毎日新聞 98 年版と毎日新聞 99 年版の 2 年間の新聞記事に対して，記事中に存在する統計量の名前や値，日付などの要素を抜き出し，値に関してはどの統計量のものか，日付に関してはその絶対表現はいつか，といったタグ付けを人手で記述したものである．また，タグを付与した新聞記事の集合を「ガソリン」「日経平均株価」「商業販売統計」といったトピック毎に分類している．

2.1 統計量名の抽出

文書中には様々な統計量名が出現するため，単純に統計量名の辞書を用いるだけでは，未知の記事中の統計量名を網羅性良く特定することは難しい．このため，統計量名の抽出は，機械学習（具体的には，Support Vector Machine 二値分類器 [4]）による系列ラベリングにより行った．

我々は,「文字種」「文字」「文字の位置」「単語」「品詞」といった基本素性 [5] に加え,「統計量名の多くが複合名詞であること」（例：「内閣支持率」「ガソリン価格」）や,「統計量は数値であり，従って，統計量名は，統語的に数値と関連の深い位置にあること」といった，統計量名の特徴を機械学習の素性（具体的には,「単語の，数値との連接頻度」「数値との構文関係」等）として取り入れた．さらに，複合名詞の主辞が最後尾の単語であり，後ろの単語ほど重要であるという性質から，系列ラベリングの際，文の後方からラベルを決定し，前方のラベルを決定する際に，直後の文字のラベルを素性として加えるという逐次的アルゴリズムを採用した．

2.2 グラフ生成アルゴリズム

前節で解説した統計量名抽出システムを用い，新聞記事から統計量名を抽出する．システムは，統計量名とその値を抽出し，それらをクラスタリングした後，グラフを Java JFreeChart クラス [6] を用いて生成する．グラフの生成自体はライブラリを用いて自明に行えるため，以降，グラフを「三つ組 (統計量名, 時点, 統計量) の集合」と定義する．

連絡先: 吉田稔 mino@r.dl.itc.u-tokyo.ac.jp

抽出された統計量名を用いグラフを生成するために必要なタスクとして,「統計量の抽出」「時点の抽出」「統計量名クラスタリング」がある．このうち,「統計量の抽出」は，入力として統計量名の集合と単位の集合[*1]が与えられたとき，与えられた統計量名に対応する数値（「統計量」）を抽出するタスクであり，各文を構文解析した後,「与えられた統計量名と類似度が高い文節 x の発見」および,「構文構造上，x に一番近い，数値・単位を含む文節の発見」を行うことによって実現する．また,「時点の抽出」は，人手によるパターンを記述することにより行った．これにより，文の集合から,「(統計量名, 統計量, 時点)」という３つ組が複数抽出されることになる．以下，これら３つ組の集合から，グラフを生成する「統計量名クラスタリング」について解説する．

統計量名クラスタリング　システムは，複数の記事から同一の統計量を表現したペアを集め，グラフを生成する．このとき，同一の統計量でも，異なる表現で表わされることが頻繁に起こるため，同一の統計量であることを判定する処理が必要となる．すなわち，抽出された (統計量名, 統計量) のペア集合をクラスタリングし，類似度の高いペア同士をまとめるという処理を行う．生成された各クラスタ内のペアを，同一のグラフに含める．

クラスタリングには，類似度に基づくボトムアップ型の手法を用いる．すなわち，１つのペアを１つのクラスタと見なした状態からはじめ，類似度の一番高いクラスタ同士を併合して行き，閾値を下回ったら停止する．

このときの (統計量名, 統計量) のペア同士の類似度を計算する際，統計量の特徴を活用した．具体的には，ペア類似度は，統計量名の類似度と統計量の類似度の両者を用いることにより計算される．ペア (a_1, v_1) とペア (a_2, v_2) の類似度は，文字列 a_1 と a_2 の類似度が閾値未満ならゼロとし，閾値以上のときは，値（統計量）の類似度から類似度を定義する．具体的には,

$$sim((a_1, v_1), (a_2, v_2)) = \frac{\min(v_1, v_2)}{\max(v_1, v_2)}$$

となる．

また，実際には,「同一時点の複数の統計量は，別々の統計量である」という仮定に基づき,「同一クラスタ内に同一の時点から複数のペアが入ることはない」という制約を課している．もしもクラスタ A,B を併合する際，A,B が同一時点のペアを含んでいた場合は，併合を行わない．

2.3　MuST T2N タスクへの参加

MuST T2N は,「入力として統計量名（と単位）が与えられたとき，記事中の該当する統計量を抽出してグラフを描く」というタスクである．我々のシステムは，記事中から統計量名を自動的に抜き出しグラフを描くものであるが，ここから，T2N の各課題に対して最も適したものを選択する必要がある．このため，システムは，課題が与えられたとき，**グラフと各課題の類似度**を計算し，最も類似度の高いグラフを出力する．

T2N の各課題は，想定される統計量名の集合 PA を持つ．グラフと課題の類似度は，PA 中の統計量名と，各グラフの持つ統計量名の類似度をもとに,

$$\sum_i \max_j sim(pa_j, a_i)$$

と計算される．ここで pa_j は，PA 中の j 個目の統計量名，a_i はグラフ中の i 個目の統計量名である．類似度の平均でなく和を取る理由は，平均をとった場合,「少ない統計量名で構成されるグラフで，偶然に課題との類似度が高いもの」が選択されるというノイズを避ける為である．

*1　現在，システムは統計量の単位（「人」「円」等）を入力として必要とする．

2.4　結果

2.4.1　統計量名抽出アルゴリズムの評価

提案した統計量名抽出アルゴリズムの正解率を，MuST コーパスで測定したところ，再現率 0.78, 適合率 0.87 (F 値 0.82) という高い抽出精度を達成した.（ただし，正解文字列の一部しか抽出しなかった場合も正解と見なした場合.）しかしながら，これは，交差検定の際のコーパス分割を完全にランダムにしており，訓練データとテストデータに同一のトピック（話題）の記事が含まれている場合である．訓練データの記事を，テストデータと別の話題に限定した場合，再現率 0.53, 適合率 0.82 (F 値 0.65) と，正解率は大幅に落ち込んだ．

そのほか，トピック毎の正解率の傾向を見ると，特に，イレギュラーな統計量表記（「1 ドル= 132 円」等）をしている場合に，提案手法が対応しきれていない等の興味深い問題点が明らかになった．

2.4.2　T2N タスク参加結果

T2N タスクにおける結果は，比較的精度の良い課題でも F 値 53 %となり，F 値が 0 となった課題が半数近くに上るなど，それほど良い成績を残すことはできなかった．特に，同一の文書集合における複数の課題において，比較的精度の高い課題と，精度の低い課題が混在していた．例えば，課題 010401 の精度は比較的高いが，それと同じ文書集合に対する課題である 010402,010403,010404 はいずれも正しい結果を抽出することができなかった．この文書集合に対しては,「内閣支持率」「内閣不支持率」「自民党支持率」「民主党支持率」という，統計量名・統計量ともに類似度の高い複数の課題が混在していたため，正しいグラフを選択することができなかった．また，ガソリン価格のうち，値の低いものが「ディーゼル価格」のクラスタに併合されてしまう等，数値の範囲を利用したクラスタリング手法の限界も見られた.（表 1）このように，T2N タスクに参加することにより，提案手法の「精緻な統計量名の分別能力の欠如」という課題が明らかになったといえる．

3.　数値検索と数値クラスタリング

我々は，現在，より数値に特化したシステムとして，テキスト中に書かれた数字を，数値として検索できる全文検索システムの開発に取り組んでいる．本節ではこの紹介を行う．詳細については [7] を参照されたい．

テキスト情報の中には,「25 歳」「10000 円」等，多くの数値表現が含まれている．通常，サーチエンジンやテキストへの索引付けツールでは，数値情報は，そのまま数字の文字列として取り扱うか，#や 0 などの数値を表す特殊記号に置き換えられるかであった．しかしながら，数値は，異なる数値間に順序を持つ（例えば，“200” は “100” より大きく，“101” は “100” と近い）等，他の文字列と異なる性質を持つ．こういった数値の特性を活用した検索を提供することは，テキスト中の数値からの知識発見に大きく貢献することが期待できる．

例えば，数字を全て同一に扱うシステムでは，“1”，“2”，“213”，“215” という異なる数字文字列の区別をつけることができない．これに対し,「**類似する**数字（ここでは例えば “1” と “2”，“213” と “215”）を同一に扱い，そうでない数字（ここでは例えば “1” と “213”）を別に扱う」ことができれば，数値の性質をテキストマイニングに応用することができる．これは，**数字の範囲表現**（上の例では，例えば “[1..2]” と “[213..215]”

表 1: ガソリン価格の記事集合から抽出された，単位「円」に関する統計量のクラスタ．ここで “dupli” は，より上位のクラスタに同一時点の同一統計量が含まれていることを示す．

順位	統計量名	統計量	サイズ	類似度
1	ガソリン価格	98.0–103.0	6	36.0
2	ディーゼル価格 (83.0), ガソリン価格（最安値）	83.0–92.0	6	28.0
3	ガソリン価格内の税金 (60.0), ガソリン価格 (dupli.)	60.0–103.0	3	18.0
4	ガソリン価格の上昇	2.0–2.5	3	17.0
5	電気料金 (200.0), ガソリン価格	105.0–200.0	3	10.0

という 2 つの異なる数値範囲）を用いることにより実現できる．我々は，与えられた数値コレクションから適切な**数値範囲**を導く機能，さらに，そのような数値範囲を**検索クエリ**として用いる検索機能を備えた新たなシステムを提案する．例えばユーザは,「[100..200] 円」といったクエリを用い,「“円” の左側に 100 以上 200 以下の数値を含む文字列」を取り出すことができ，これにより，ユーザは数値の範囲を，通常の文字列と同様に用いることができるようになる．

このような柔軟な数値の取り扱いは，様々なテキストマイニングタスク，特に，数値文字列を，**文脈情報**として用いる場合に有用である．例えば，単語の類似度を測定する際，一般には，各単語 w に関連する語（w の周辺に出現する語，w と係り受け関係にある語，等）の出現分布情報を単語 w の**意味**を表す情報として捉え，それらの類似度を計算することにより単語同士の類似度を測る．我々のシステムでは，この文脈情報に,「数値の類似度」という観点を導入することができるようになる．例えば,「コーヒー」「紅茶」「定食」といった 3 つの単語の類似度を測るとき，その周辺にそれぞれ「200 円」「180 円」「1000 円」といった数値文字列が出現していた場合,「“200” と “180” を同一視し，“1000” は別に扱う」という処理が可能になることにより,「コーヒー」と「紅茶」が,「コーヒー」と「定食」よりも似ている，という知識を得ることができるようになる．

前者の「数値範囲を発見する」問題に対して，システムは，確率モデル DPM (Dirichlet Process Mixture, ディリクレ過程混合分布)[10] を利用したクラスタリングを行う．DPM クラスタリングを用いることにより，クラスタ数を事前に指定せず，自動的に適切なクラスタ数が推定される．

後者の**数値範囲による検索**の実現のため，我々は，基本的な索引構造として，全文検索を可能にするデータ構造 Suffix Array（接尾辞配列）を採用し，Suffix Array を利用して数値を検索するための手法を提案した．具体的には，Suffix Array での文字列検索の際，通常の二分探索を，下記の「数値範囲クエリ」で検索できるように拡張したものであり，本稿では，我々の提案システムを，**Number Suffix Array** と呼ぶ．具体的には，次のことができるようになる．

数値範囲クエリ 通常の文字列検索に加え，Number Suffix Array では，**数値範囲**をクエリに使用することができる．例えば,「`[1000..2000]`」というクエリ*2により，1000 以上 2000 以下の数値をすべて取得することができる．また,「`[30..50]` 人」のように，数値範囲と通常の文字列を混在させたクエリも可能である．このクエリに対し,「頻度計測」と「連接文字列取得」という 2 種類の出力を行うことができる．

検索結果の数値範囲による表示 Number Suffix Array はさらに，複数の検索結果を，数値範囲を用いてまとめる機能を持つ．例えば,「300 円」「330 円」「350 円」という検索結果があったとき，これらをまとめて「`[300..350]` 円」という結果として表示することができる．どの範囲の数値をまとめて表示するかは，クラスタリングにより自動的に判別する．これにより,「1」と「1000000」をまとめて「`[1..1000000]`」となる，といった，不自然な数値範囲を避けることができる．

3.1 Number Suffix Array によるテキストマイニング

Number Suffix Array の応用として,「用例検索」と「同義語抽出」の二つのテキストマイニングに Number Suffix Array を適用した例を示す．

3.1.1 用例検索システム Kiwi

用例検索 Kiwi[11] は，与えられたクエリに「連接しやすい」文字列を提示するシステムである．ここで「連接しやすさ」は，Kiwi スコアと呼ばれるスコア関数で定義される．例えば,「*処理」というクエリ*3に対して,「情報処理」「知識処理」「自然言語処理」等,「処理」を右に含む文字列を検索結果として返す．Kiwi アルゴリズムの特徴は,「頻度と長さを考慮したスコア付け」と,「分岐数に基づく適切な切れ目の発見」にある．これらは，連接文字列 trie(木構造) を探索する問題であり，この木構造探索は，Suffix Array を用いてシミュレートすることができる．

Kiwi アルゴリズムを Number Suffix Array 上で実装することにより，数値範囲を用例の一部として抽出したり，数値範囲に連接し易い用例を取得したりすることが可能となった．毎日新聞コーパスを利用した用例検索の例を図 1 に示す．例えば「`[3..8]` 歳」というクエリに対して「息子」「娘」といった連接語が取得されるのに対し,「`[20..40]` 歳」というクエリに対しては「男性」「女性」「若者」といった連接語が取得されている．

3.1.2 同義語抽出

我々は，Suffix Array を用いたオンデマンド同義語抽出を提案している [12]．この手法では，クエリ q の文脈（q に連接する文字列）を Suffix Array を用いて取得*4，さらにその後，取得された文脈に連接する文字列を逆方向に検索して取得することで，q に類似した言葉を取得している．我々は，このアルゴリズムを，Number Suffix Array を用いて拡張した．まず，文脈語として，数値範囲を利用できるようにする．たとえば,「10 名」「15 名」「20 名」といった文脈語を自動的に「`[10..20]`

*2 [と] で囲まれた部分は,「左の数値から右の数値までの数値範囲」を示す．

*3 ここで*は，検索したい文字列の位置を示す．

*4 このさい，左右両方向の文脈を取得するため，順方向と逆方向 (Reversed Suffix Array) の 2 つの Suffix Array を用いている．

クエリ：「[3..8] 歳の」
(1)[4..8] 歳のとき／ (2)[3..8] 歳のころ／ (3)[3..8] 歳の時に／ (4)[3..5] 歳の息子／ (5)[6..8] 歳の少女／ (6)4 歳のときから／ (7)[3..5] 歳の女の子／ (8)4 歳のときからピアノを習／ (9)[3..4] 歳のころから母／ (10)5 歳の外国産馬

クエリ：「[20..40] 歳の」
(1)[20..35] 歳の女性／ (2)[20..37] 歳の時に／ (3)[21..40] 歳の男性／ (4)[21..40] 歳の若さで／ (5)[26..36] 歳のベテラン／ (6)[20..40] 歳のとき／ (7)[20..35] 歳のころ／ (8)[20..25] 歳の若者／ (9)[20..37] 歳の誕生日／ (10)[20..31] 歳のころ

図 1: Number-Kiwi の出力例.（2 つのクエリに対する右側連接文字列の違い.）

名」のようにまとめることで，文脈および取得できる同義語候補のカバー率を向上させる．

「数値範囲」をテキストの類似度測定に用いるというアイデアは，前述のグラフ作成システムで用いた「統計量名クラスタリング」と同様のものであり，MuST の参加を通じて得られたアイデアを発展させたものと考えている．

4. テキストによる経済指標予測

本稿の主題である「テキスト中の数値表現の解析」とは異なるが，MuST への参加では，テキストと数値の関連を扱う研究として，テキストから株価予測を行うシステムの開発も行った．[13] これに関しても，その発展形として，経済指標と株価の関連分析として，現在研究を進めている．MuST 参加時のシステムは，単語の極性推定と，それに基づく株価上昇・下落の予測を目指していたが，現在は，より予測のし易い取引高に着目し，トピックモデルを用いた文書クラスタリングにより文書を話題（トピック）別に分類し，「取引高を上昇させる話題（記事）の発見」と，それを利用した「取引高の予測」について，研究を行なっている．実験の結果，確信度の高い上位１０記事について 56.3 %の精度で取引高上昇の予測が行えるという結果を得ている．詳細については [14] を参照されたい．

5. おわりに

我々は,「数値とテキストの関連性」という視点からの研究を議論できる場として MuST へ参加した．MuST へは,「情報抽出によるグラフ作成」「数値による情報検索」「株価とテキストの関連」という 3 つのテーマで参加を行ったが，MuST への参加を通じて，多くの類似の研究を行う研究者からのフィードバックを得ることができ，ここで発表した研究は，いずれも，継続的に発展させることができていると考えている．

参考文献

[1] 加藤 恒昭, 松下 光範, 平尾 勉, 動向情報の要約と可視化に関するワークショップの提案, 情報処理学会自然言語処理研究会研究報告, 2004-NL-164, pp. 89–94, 2004.

[2] 吉田 稔, 杉浦 隆博, 山田 剛一, 増田 英孝, 中川 裕志, テキストからの数値抽出による自動グラフ作成, 人工知能学会第 23 回全国大会, 3F2-NFC3-2, 2009

[3] 杉浦 隆博, 吉田 稔, 山田 剛一, 増田 英孝, 中川 裕志, 新聞記事の数値による情報検索システムの提案と実装, 第 21 回人工知能学会全国大会, 2H5-9, 2007

[4] V.Vapnik, Tha Nature of Statistical Learning Theory, Springer, 1995.

[5] 中野 桂吾, 平井 有三, 日本語固有表現抽出における文節情報の利用, 情報処理学会論文誌, vol 45, no 3, pp.934–941, 2004.

[6] JFreeChart, "http://www.jfree.org/jfreechart/"

[7] Minoru Yoshida, Issei Sato, Hiroshi Nakagawa, and Akira Terada, Mining Numbers in Text Using Suffix Arrays and Clustering Based on Dirichlet Process Mixture Models, Proceedings of PAKDD-2010, pp. 230-237, 2010.

[8] Manber , U. and Myers, G. Suffix Arrays: A New Method for On-line String Searches, Proceedings of the first ACM-SIAM Symposium on Discrete Algorithms, pp. 319–327, 1990.

[9] Marcus Fontoura and Ronny Lempel and Runping Qi and Jason Y. Zien, Inverted Index Support for Numeric Search, Internet Mathematics, 3(2), pp.153–186, 2006

[10] Charles E Antoniak, Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems, The Annals of Statistics, 2(6), pp.1152–1174, 1974.

[11] Kumiko Tanaka-Ishii and Hiroshi Nakagawa, A Multilingual Usage Consultation Tool based on Internet Searching —More than search engine, Less than QA, Proceedings of the 14th International World Wide Web Conference (WWW2005), pp. 363–371, 2005.

[12] 吉田稔, 中川裕志, 寺田昭, コーパス検索支援のための動的同義語候補抽出, 人工知能学会論文誌, 25(1), pp. 122–132, 2010

[13] 廣川 敬真, 吉田 稔, 山田 剛一, 増田 英孝, 中川 裕志, 新聞記事のテキスト情報と株価動向の関係の解析, MuST(動向情報の要約と可視化に関するワークショップ) 2007 年度成果進捗報告会, 2008

[14] 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.