# テキストからの数値抽出による自動グラフ作成

A System for Generating Graphs from Documents

吉田 稔[*1] Minoru Yoshida　杉浦 隆博[*2] Takahiro Sugiura　山田 剛一[*2] Kouichi Yamada　増田 英孝[*2] Hidetaka Masuda　中川 裕志[*1] 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

In this paper, we describe our system of automatically generating graphs from given documents. The system extracts names of statistics by using sequence labeling with support vector machines (SVMs). Extracted names and their associated values are clustered to be used for drawing graphs. We also describe our participation in the MuST T2N task with this system by introducing the algorithm for calculating similarities between generated graphs and T2N problems.

## 1. はじめに

本稿では、与えられた新聞記事から、自動的に統計量を発見し、その動向情報をグラフ化するシステムの紹介を行う。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。また、我々は、本システムを用いて、MuST T2N タスクに参加した。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、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。

本研究では、機械学習の学習データとして、「動向情報の要約と可視化に関するワークショップ [4]（略称：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）を使用している。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. システム解説

提案システムは、与えられた新聞記事から自動的にグラフを作成する。システムへの入力は記事の集合であり[*1]、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。

### 2.1 統計量名の抽出

統計量名は、数値の属性名として現れる単語や複合語として捉える事ができる。例えば、「売上高」「内閣支持率」といった文字列が統計量名となる。実際には、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。本節では、機械学習を用いて、与えられた文書から自動的に統計量名を発見するアルゴリズムについて述べる。統計量名の抽出は、Support Vector Machine 二値分類器 [1] に基づく系列ラベリングを用いて行う。

本研究では、統計量名が、以下のような性質を持つという仮定を置く。

1. 数値の直前に出現する.
2. 数値と係り受け関係を持つ.
3. 複合名詞となる.

この仮定に基づき、SVM 分類器において用いる素性として「文字種」「文字」「文字の位置」[*2]「単語」「品詞」といった基本素性 [3][*3]に加え、以下のものを用いる。

**数値との共起頻度** 数値の直前（助詞は削除して考える）に出現する頻度。例えば,「累計出荷台数は 5000 台」といった文が対象となるとき,「台数」が数値直前に出現する単語となる。

**数値との構文関係** 数値と係り受け関係の有無や方向を表す番号。「係り元が数値」「係り先が数値」「数値と係り先が共通」「それ以外」の 4 種類。

**数値と動詞の組み合わせ頻度** 係り先の動詞と数値との共起頻度。(*円と, なる) 等のペアの頻度となる。

**複合名詞** 複合名詞に含まれているか否か

形態素解析には Chasen[5]，構文解析には Cabocha[6] をそれぞれ使用し，複合名詞抽出には言選 Web[7] を用いる．言選 Web は本来重要語抽出のために用いるが，本研究では複合名詞の抽出のみを対象としている．

アルゴリズムは、与えられた文の末尾から一文字ずつ「統計量名か否か」のラベルを付与していく。このとき、既に決定さ

連絡先: 吉田稔 mino@r.dl.itc.u-tokyo.ac.jp

*1 現在、記事集合は、MuST T2N のタスクとして与えられた記事集合（例えば「ガソリン価格」のタスクならば、それに関連した記事の集合）を与えている。人手での選別を経ない記事集合を入力とした実験は今後の課題である。

*2 [8] にある Start-End 法の B,E,I,S タグを用いている

*3 これらの素性は、現在の位置と、前後 2 文字の、計 5 文字を対象にして取得する。

れた文字のラベルも、素性として用いる。すなわち、$i$ 文字目のラベルを決定する際、$i+1$,$i+2$ 文字目に付与されたラベルも素性として用いる。

### 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。統計量名抽出システムの学習及びグラフ生成のためのルール作成（後述）には、T2N タスクで提供されたコーパスは使用していない。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 [9] 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。図 1 にシステムの概略を示す。

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。以下、これらのタスクとそれを解決するためのアルゴリズムについて解説する。

**統計量抽出**　システムはまず、与えられた統計量名に対応する数値（「統計量」）を抽出する。入力として統計量名の集合と単位の集合[*4]が与えられたとき、システムは、

- 各文を、CaboCha によって構文解析し、依存構造木を得る。
- 与えられた統計量名との類似度が高い文節 $x$ を見つける。
- 数字と単位文字列を含み、$x$ に一番近い（すなわち、依存構造木上において、$x$ へ到達するパスが一番短い）文節 $y$ を見つける。

という手順により、文節 $y$ を発見し、$y$ から統計量を抽出する。これに加えて、$x$ に直接係る文節を**統計量名修飾句**として、$y$ に直接係る文節を**統計量修飾句**として抽出する。

**時点抽出**　システムはまた、その統計量名がどの時点であるかを抽出する。時点の抽出には、人手によるパターンを用いた。パターンは、「*月」「*年」「昨年」「先月」といった、特定の文字列を含んでいれば時点と判定するという単純なものであり、前節で述べた「統計量名修飾句」または「統計量修飾句」がこのパターンに合致すればそれを時点として抽出する。

**統計量名クラスタリング**　システムはさらに、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。

システムはまず、(統計量名,　統計量) のペア同士の類似度を計算する。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、次のように定義される。

- $a_1$ と $a_2$ の共通文字数 $cc(a_1, a_2)$ を計算する。
- $cc(a_1, a_2) < th_{cc}$ であった場合、$sim((a_1, v_1), (a_2, v_2)) = 0$ となる。[*5]
- $cc(a_1, a_2) \geq th_{cc}$ の場合、ペア間類似度を、値の比率で定義する。（すなわち、$sim((a_1, v_1), (a_2, v_2)) = \frac{\min(v_1, v_2)}{\max(v_1, v_2)}$ となる。）

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 $th_{cl}$ を下回ったら停止する。クラスタ $C_1$ と $C_2$ の類似度は、

$$\max_{c_1 \in C_1, c_2 \in C_2} (sim(c_1, c_2))$$

として計算される。現在は、$th_{cl} = 0.5$ と設定している。

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。

## 3. 実験

### 3.1 統計量名抽出アルゴリズムの評価

本実験で用いる数値の統計量名抽出の正解データは，MuST で配布している MuST コーパスに含まれる統計量名（<name> タグが付与された表現）とする．これは，MuST コーパスの統計量名の多くが，新聞記事中の数値の統計量名と同一の表現となるためである．MuST コーパスに含まれる統計量名の要素は，毎日新聞の 98 年版と 99 年版に含まれる 581 記事を対象とし，人手でタグ付けされたものである．本実験では，MuST コーパスに含まれる 581 記事中の 518 記事を用いた．

#### 3.1.1 実験方法

実験では、記事集合をランダムに 5 分割した結果を用いる 5 分割交差検定に加え、**トピック分割交差検定**も行う。これは、あるトピックをテスト集合とするとき、トレーニング集合を**それ以外**のすべてのトピックとする手法である。これは、アルゴリズムが、未知のトピックに関してどの程度耐性を持つかを評価するための実験である。正解データと抽出した数値の統計量名を比較し，再現率及び適合率を算出する．再現率と適合率は以下のものとして定義する．

---

*4　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。

*5　$th_{cc}$ は閾値であり、現在は $th_{cc} = 3$ としている。

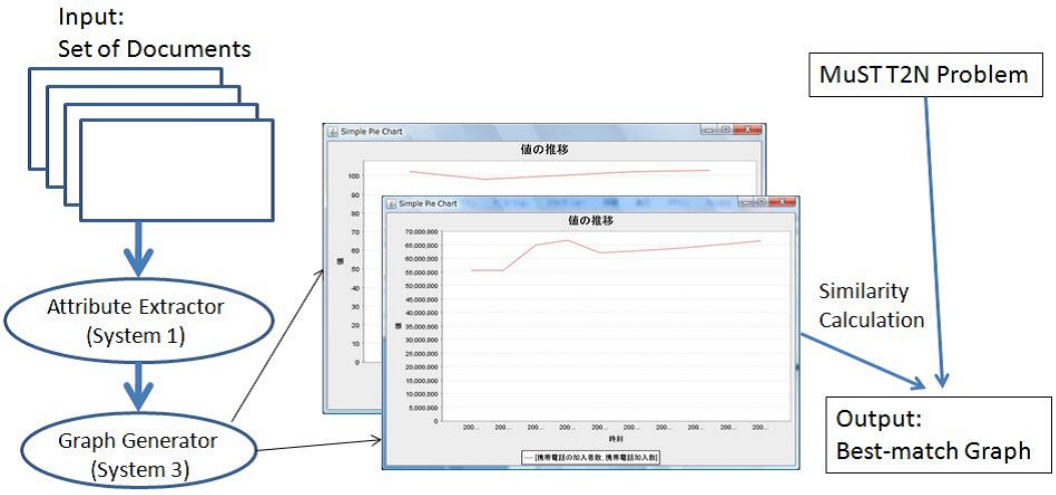


図 1: グラフ生成システムと MuST T2N タスクへの出力生成の概要

$$\text{再現率 } (R) = \frac{\text{抽出に成功した統計量名数}}{\text{正解統計量名数}}$$

$$\text{適合率 } (P) = \frac{\text{抽出に成功した統計量名数}}{\text{抽出した統計量名数}}$$

本実験で定義している正解データと同様の統計量名を完全に抽出した場合のみを○とし，不完全に抽出したものを△，抽出に失敗してしまったものを×とする．例えば，正解データが「完全失業者数」と統計量名であった場合,「完全失業者数」という文字列を全て抽出できたものを○とし,「失業者数」という一部の文字列しか抽出できなかった場合を△とし，それ以外を×とする．○および△となる抽出結果を正解として，再現率，適合率，F 値を算出する．本評価実験は，本研究で提案した SVM を用いた数値の統計量名抽出手法と，以前我々が行った依存構造木に基づく数値とその統計量名の対応パターンを用いたベースライン手法 [2] の二つの手法に対して行う．依存構造木に基く抽出手法では，複数の新聞記事を依存構造木に変換したものから，数値と統計量名の抽出パターンを作成し，作成したパターンに基き統計量名を自動抽出を行う手法である．

#### 3.1.2 記事全体の評価結果

表 1 は，それぞれの実験結果における，文書集合全体の F 値，再現率，適合率を求めたものである．ベースライン研究と比較し，トピック分割交差検定，5 分割交差検定の両方とも F 値，再現率，適合率が向上している．トピック分割交差検定は、5 分割交差検定と比べ、低い値となった。このことは、通常の 5 分割交差検定は、未知のトピックに対するアルゴリズムの耐性を測るのには適していないことを示唆している。

トピック毎の結果では、「日経平均株価」「為替レート」といったトピックが評価結果が低くなっており，特に「為替レート」に関しては完全に抽出を失敗し，再現率が 0 となった。「為替レート」に関する記事は「1 ドル= 132 円」といった文の「1 ドル」を数値の属性名として扱っている．このような特殊な表現に関して、構文関係や複合名詞といった、本研究で提案した素性がうまく働かなかったと推測される。

表 1: 統計量名抽出の実験結果

| Method | F-measure | Recall | Precision |
|---|---|---|---|
| ベースライン | 0.38 | 0.42 | 0.34 |
| トピック分割交差検定 | 0.65 | 0.53 | 0.82 |
| 5 分割交差検定 | 0.82 | 0.78 | 0.87 |

### 3.2 T2N タスク参加結果

表 2 に、T2N タスクの結果を示す。*6全体的傾向として、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。

表 3 に、トピック「ガソリン」の文書集合から抽出されたグラフ（統計量クラスタ）の上位 5 個を示す。なお、各グラフは、課題 010101 への類似度順にソートされている。最上位にランクされたクラスタは、すべて課題と適合する「ガソリン価格」の統計量で構成されていた。しかしながら、他の 5 つの「ガソリン価格」統計量が 2 位のクラスタへ含まれてしまっている。これは、比較的低い値のガソリン価格が、ノイズ「ディーゼル価格」に引きずられてしまったためである。このことから、現在の「文字類似度」「値の比率」のみを使う単純な手法は、まだ高い再現率を得るには改善の余地があると言える。例えば、統計量名を高い精度で判別できる特徴的な文字／文字列を統計的に発見する等の改善策が考えられる。また、2 位のクラスタでは、「昨年春」や「半年前」といった、（比較的安かった）過去のガソリン価格が多く含まれていたが、これらの時点情報の抽出には失敗していた。このことより、時点情報抽出に関しても改善の余地が多く残されていると考えられる。

*6 課題 010501–010503 には単位名が無いため、結果の出力を行っていない。

表 3: 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 |

表 2: T2N タスクの結果

| 問題 | Precision | Recall | F-measure |
|---|---|---|---|
| MuSTT2N010101 | 60.0 | 18.8 | 28.6 |
| MuSTT2N010102 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010201 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010202 | 33.3 | 13.3 | 19.0 |
| MuSTT2N010301 | 54.5 | 31.6 | 40.0 |
| MuSTT2N010302 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010303 | 11.1 | 5.9 | 7.7 |
| MuSTT2N010304 | 22.2 | 50.0 | 30.8 |
| MuSTT2N010401 | 28.6 | 7.4 | 11.8 |
| MuSTT2N010402 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010403 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010404 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010501 | – | – | – |
| MuSTT2N010502 | – | – | – |
| MuSTT2N010503 | – | – | – |
| MuSTT2N010601 | 50.0 | 11.5 | 18.8 |
| MuSTT2N010602 | 40.0 | 22.2 | 28.6 |
| MuSTT2N010603 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010604 | 20.0 | 20.0 | 20.0 |
| MuSTT2N010701 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010702 | 66.7 | 28.6 | 40.0 |
| MuSTT2N010801 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010802 | 80.0 | 40.0 | 53.3 |
| MuSTT2N010803 | 0.0 | 0.0 | 0.0 |
| MuSTT2N010804 | 0.0 | 0.0 | 0.0 |

### 3.3 結論と今後の課題

MuST T2N タスクへの参加を念頭に、新聞記事集合から自動的に統計量グラフを作成するシステムを開発した。システムは SVM 分類器を用いた系列ラベリングにより自動的に統計量名を抽出し、それを用いて類似統計量のクラスタリングとグラフの作成を行う。T2N への参加は、生成されたグラフと各課題の類似度を計算することで行った。しかしながら、T2N タスクへの参加の結果からは、現在の手法、特に自動グラフ作成のアルゴリズムには、まだ改善の余地が多く残されているという結論となった。

今後のシステム改善点としては、精度向上のためのアルゴリズムの開発のほか、単位名を使わずに統計量を抽出できるアルゴリズムへの変更と、相対値（「５%上昇」等）の抽出アルゴリズムの開発等が挙げられる。

## 参考文献


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

[2] 杉浦 隆博, 吉田 稔, 山田 剛一, 増田 英孝, 中川 裕志, 数値による新聞記事テキストマイニングシステムの提案, 情報科学技術フォーラム講演論文集, pp. 161–164, 2007

[3] 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.

[4] 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.

[5] 松本 裕治, 北内 哲, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸, 形態素解析システム『茶筅』version2.3.3 使用説明書, 奈良先端科学技術大学院大学, 2003.

[6] 工藤拓, 松本裕治, チャンキングの段階適用による係り受け解析, 情報処理学会論文誌, vol 43, no 6, pp. 1834–1842, 2002.

[7] 前田 朗, キーワード自動抽出システム「言選 Web」, 漢字文献情報処理研究会, vol 6, pp. 124.133, 2005.

[8] 内元 清貴, 馬 青, 村田 真樹, 小作 浩美, 内山 将夫, 井原 原 均, 最大エントロピー方と書き換え規則に基く日本語固有表現抽出, 言語処理学会誌, vol 7, no 2, pp. 63–90, 2000.

[9] JFreeChart, ”http://www.jfree.org/jfreechart/”