数値域推定を用いた時系列統計情報の抽出

Extracting Time Series Statistics using Value Range Estimation

井上 裁都[*1] 鈴木 宏哉[*1] 斎藤 博昭[*1]
Tatsukuni INOUE Hiroya SUSUKI Hiroaki SAITO

[*1]慶應義塾大学大学院 理工学研究科
Keio University Faculty of Science and Technology

The goal of our research is to constract the trend information extraction system that is focused on value range estimation for extracting time series statistics. In this paper, in order to constract the system, we propose the trend information extraction system focused on value range estimation, for our first step.

1. はじめに

近年注目されている自動要約や情報可視化のような情報アクセス技術の研究課題の一つに「動向情報の要約と可視化」がある．動向情報とは「幾つかの統計量の時系列データを基として，その変化を通時的にとらえつつ，それらを単に羅列するのではなく，総合的にまとめ上げることで得られるもの」[加藤 04]であり，ある商品の価格や売上の状況，内閣や政党の支持状況などがその典型となる．この動向情報の要約と可視化に対し，共通の素材を用いて協調的かつ競争的に取り組むことを目的として,「MuST：動向情報の要約と可視化に関するワークショップ」（以下，MuST）において研究が進められている．

動向情報の要約と可視化における基礎的な研究課題の一つに「テキスト群からの時系列統計情報の自動抽出」がある．MuSTにおいては，この研究課題は “T2N” 課題と呼ばれ，T2N は，テキスト群中で，ある統計量のどの時点 (日付) のどの値が話題になっているのかを明らかにし，抽出された時系列統計情報をグラフ化することで，テキスト群の関心に従った統計量の可視化を可能にする．T2N は動向情報の要約と可視化のためには必要不可欠な技術ではあるが，抽出における十分な適合率・再現率を達成する手法は未だ確立されていない．

本論文では，時系列統計情報抽出におけるより高い適合率を達成するため，統計情報の構成要素の一つである値情報の取りうる範囲，すなわち数値域に着目し，値情報の数値域を推定する手法を提案する．

2. 時系列統計情報

本論文においては，時系列統計情報は「統計量名」「値情報」「日付情報」の三つ組からなる情報であると定める．また，動向情報はこの時系列統計情報の集合より構成されるものと考える．ここで定義した時系列統計情報の具体例を以下に示す．

統計量名 レギュラーガソリンの全国平均店頭価格
値情報 104 円
日付情報 2000 年 10 月 10 日

上記の統計量名は抽出対象の統計情報を識別するクエリの主要素でもある．しかし，クエリはその他にも「統計量の単位」(単数または複数),「統計量名の別名」（任意個）を含む．クエリの具体例は以下の通りである．

統計量名 レギュラーガソリンの全国平均店頭価格
統計量の単位 円
統計量名の別名 ガソリン価格

連絡先: 井上 裁都，慶應義塾大学大学院，〒 223 − 8522 神奈川県横浜市港北区日吉 3-14-1，Tel: (045)-563-1141，e-mail: inoue@nak.ics.keio.ac.jp

3. 関連研究

難波らは文書横断文間関係を考慮し動向情報を抽出している [難波 05]．文書横断文間関係とは，衛藤らが文間および段落間に定義している 14 種類の関係である [衛藤 05]．これらの関係の中から，動向情報と関連のある「推移」と「更新」という二つの関係を用いている．

曽我らはテキストに含まれる比較表現を考慮した動向情報抽出手法を提案している [曽我 06]．比較表現はテキスト中の統計情報と組み合わせることにより，テキスト中では直接述べられていない間接的統計情報の抽出を可能にする．曽我らはこれを利用し，いくつかの誤抽出した統計情報について検出・除外することに成功している．

4. 値情報の数値域

本論文では値情報が取り得ると考えられる値の範囲を値情報の「数値域」と呼ぶ．例えば統計量名が「レギュラーガソリンの全国平均店頭価格」の場合,「80 円から 200 円まで」が数値域の一例となる．この数値域を自動推定しこれを時系列統計情報抽出に応用することを考える．

数値域に着目した理由として，先行研究の動向情報可視化の結果における図 1 に例示されるような事例の存在が挙げられる．図 1 は「レギュラーガソリンの全国平均店頭価格」を毎日新聞の記事 2 年分から抽出し，ある期間についてその動向をグラフ化した結果である．

図 1 には，明らかに誤抽出と思える情報が数箇所含まれている．具体的に示せば，価格が 20 円以下のプロットである．このため，図 1 からは実際のガソリン価格の動向が読み取りづらくなっている．

このような誤情報が抽出されてしまう原因は様々であるが，誤抽出の原因の一例として，以下に示す自然言語文の存在がある．下記の例文に含まれる「ガソリンは４円」という節はシステムが「４円」をガソリン価格として抽出する要因となる．

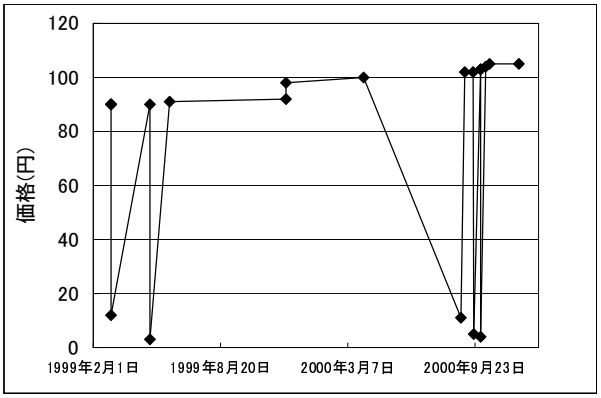

図 1: 動向情報の可視化結果例

原油価格が上昇し始めた昨年３月と比べ、原油は１６円上がったが、ガソリンは４円、軽油は９円が店頭価格に未転嫁となっており、石油業界は「努力はもう限界」と話す。

難波らは可視化の結果，図 1 のようなグラフが得られることに対し,「グラフを見れば抽出に失敗していると思われる点が容易に推測できるため，このような点をインタラクティブに削除したり，数値情報を抽出した記事にリンクしたりする機能をシステムに備えれば，ユーザは比較的容易に正しいグラフを得ることができる」と考えていた．

人間は自然とガソリン価格は 20 円を下回ることはないと判断する．しかし，従来のシステムではこのような判断は為されていない．システムにこの判断をさせることができれば，前述のような誤抽出情報の除外が自動的に可能となる．システムが上記の判断をするためには，抽出対象の統計情報に対し，前述の「数値域」を推定する必要があると考えられる．

5. 数値域推定手法

本論文では，筆者らが過去に提案した手法 [Inoue 08] を発展させた手法を用いる．入力としてはクエリと新聞記事コーパスの二入力を想定し，クエリは 2. 節で述べた「統計量名」「統計量の単位」「統計量名の別名」を含むものである．

5.1 前処理

前処理は，クエリ解析，記事抽出，パラグラフ抽出，値情報抽出の四段階から成る．

クエリ解析では MeCab を用い,「統計量名」および「統計量の別名」から名詞の形態素を抽出し，これらの IDF を算出する．ここで抽出された名詞は以後「キーワード」と呼称する．キーワードの IDF は下記の式より算出する．

$$\mathrm{IDF}(w) = \log_{10} \frac{|D_a|}{f_d(w, D_a)}$$

ここで w はキーワード，D_a は文書の集合，$|D_a|$ は文書の総数，$f_d(w, D_a)$ は D_a 中の w を含む文書の数である．D_a としては入力の新聞記事コーパスを用いる．

キーワードは「統計量名」と各「統計量名の別名」それぞれから抽出されることになるが，これらのキーワードをまとめて一つの集合と見なすと，以後の前処理の各段階において悪影響を及ぼす可能性が高い．よって，キーワードは「統計量名」別に抽出する．以後，一つ一つのキーワード集合を W，W の集合を U_W と呼ぶ．

記事抽出・パラグラフ抽出では，クエリに関連した記事・パラグラフをコーパスより抽出する．抽出には記事抽出・パラグラフ抽出両方で同じ手順を用いる．これは抽出する文書の大きさを変えて同じ手順を二度繰り返すことを意味する．具体的な手順は以下のようになる．

1. U_W から W を，全文書から文書 d を一つずつ取りだす．
2. d と W に共通のキーワードの IDF の総和を s_1 とする．
3. W 中の全キーワードの IDF の総和を s_0 とする．
4. $s_1/s_0 \geq 0.6$ であるならば d を関連文書として抽出する．
5. 全ての W と d の組合せについて手順 1〜4 を繰り返す．

上記の手順中の文書 d が記事・パラグラフである．手順 4 の $s_1/s_0 \geq 0.6$ という条件は経験的に決定した．

値情報抽出では前述の抽出パラグラフからクエリに対応した値情報を取り出す．値情報の抽出は，150 程度のルールに基づき実行されるが，紙面の都合上，ここでは重要なもののみを述べるに留める．

まず，値情報は二つのキーワードと組にして抽出する．二つのキーワードとは，例えば統計量名が「レギュラーガソリンの全国平均店頭価格」の場合,「ガソリン」「価格」が一例となる．この二キーワードの組合せは，U_W 中の全ての W に対し，W から選ぶことが可能な二キーワードの組合せとする．ただし，二キーワードの IDF の和が閾値，ここでは 3.0 を下回った場合，これらは重要度の低いキーワードの組合せとして，除外する．このキーワードの組合せは以後 P と呼び，この集合を U_P とする．

この U_P を用い，下記の手順より値情報を抽出する．ここで値情報とは「数値」と「統計量の単位」(例: 円) が連結した表現かつ「比較表現」と係り受け関係 (解析には CaboCha を使用) にないものとする.「比較表現」「日付表現」の抽出については [Inoue 08] を参照されたい．

1. U_P から P を取りだす．
2. 抽出パラグラフからパラグラフ d_p を取りだす．
3. d_p において最初に P 中の全キーワード w および日付表現の出現を確認できる位置以降の値情報を探索．もし値情報が見つからなければ，手順 8 に進む．
4. 発見した値情報を抽出する．
5. 値情報発見位置以降から P 中の w，日付表現，値情報を探索．もしこれらが発見できなければ，手順 8 に進む．
6. もし w か日付表現が発見されず，値情報が見つかったならば，手順 8 に進む．
7. w または日付表現の発見位置以降の値情報を探索．もし値情報が見つからなければ，手順 8 に進む．
8. 手順 4 に戻る．
9. 全てのパラグラフについて手順 2〜7 を繰り返し，抽出した値情報と P の組を作る．
10. U_P の全要素について手順 1〜8 を繰り返す．

上記の手順を用いることで，キーワードの組が「ガソリン」「価格」であれば，下記の例文から「１０３円」のみを抽出することが可能となる．

日本国内でも原油価格上昇を反映し、今月２日現在のガソリンの店頭価格（１リットル当たり）が全国平均で１０３円、軽油が８３円と、先月に比べそれぞれ１円値上がりした。

値情報抽出の結果，U_P の各要素 P に対応した値情報の集合 S が得られる．この S の集合を以後 U_S と呼ぶ．

5.2 数値域推定

5.1 節で抽出した U_S を用い値情報の数値域を推定する．本手法ではまず U_S の各要素 S に対し，5.2.1 節で述べるアルゴリズムを適用し，数値域推定を行う．その結果，複数の推定数値域が得られることになるが，これらを 5.2.2 節で述べる手法より統合することで最終的な数値域推定結果とする．

5.2.1 数値域推定アルゴリズム

本論文では，値情報の集合 S の分布を考えたとき，密度の高い領域と低い領域の両方が存在することを仮定する．さらに密度の高い領域の内，その領域が含むデータの個数が最も多い領域は値情報の数値域を表していると考える．このような領域を S の分布から抽出するため，本手法では S に対し階層的クラスタリングを実行する．具体的なアルゴリズムは以下のようになる．

1. $S = \{x_i | i \in \mathbb{N}, i < n\}$ とする．ここで x_i は値情報，n は S の要素数である．
2. 全ての i に対し，x_i を含む 1 要素クラスタ C_i を生成する．
3. 式 (1) を用い，最短距離法を適用して各クラスタ間距離を算出する．
$$d_X(x_i, x_j) = |\log_{10} x_i - \log_{10} x_j| \quad (1)$$
4. もし $d_X(C_i, C_j) < t$ を満たす i, j $(i \neq j)$ が存在しないのであれば，手順 6 に進む．(t: 閾値; 算出方法は後述．)
5. クラスタ間距離が最短のクラスタをクラスタリングし，手順 3 に戻る．
6. クラスタ中の要素数が最大のクラスタを C^* とする．
7. $\min C^*$, $\max C^*$ をそれぞれ数値域の下限 β，上限 α とする．

上記のアルゴリズムの具体的な適用結果は図 2 より理解ができる．

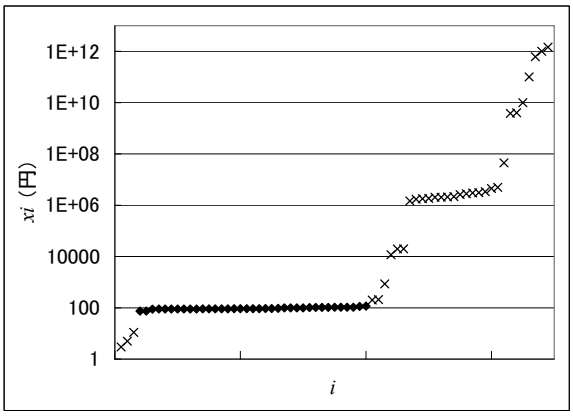

図 2: 数値域推定アルゴリズムの適用例

図 2 はキーワード「ガソリン」「価格」と組となる値情報の集合に対し数値域推定アルゴリズムを適用した様子を示したグラフである．図中においては分かりやすさのため，値情報をソートし，縦軸を対数スケールとした上で値情報をプロットしている．図中の◆のプロットは，アルゴリズム適用の結果，手順 6 の段階でクラスタ C^* に属することになる値情報を示す．ガソリン価格は 4. 節で述べたように 100 円前後の値を取るため，アルゴリズムによる数値域推定の結果はおおよそ正しくなることが図 2 より読み取れる．

図 2 中の隣り合うプロットの縦軸方向の距離に着目してみると，◆のプロット同士の距離は非常に小さく，反対に×同士または◆と×のプロット間の距離は比較的大きいことが分かる．式 (1) の意味を考えると，数値域推定アルゴリズムにおいてはこの縦軸方向の距離が小さいプロット同士，言い換えれば大体横に並んでいるプロット同士が同じクラスタに属することになる．すると図 2 からは，アルゴリズムの適用によって図中の◆のプロットが属するクラスタが最大となり，適切な数値域推定が為されることが把握できる．

5.2.2 推定数値域の統合

まず，5.2.1 節において保留していた数値域推定アルゴリズム中の閾値 t は次の手順で決定する．

1. S の要素を昇順に整列し，$S = \{x_i | i \in \mathbb{N}, i < n\}$ とする．x_i は i 番目の要素 (値情報)，n は S の要素数である．
2. $0 \leq i \leq n-2$ を満たす i に対し，式 (1) を用いて次式で定義する δ_i を算出し，これらの集合を D とする．
$$\delta_i = d_X(x_i, x_{i+1}) = |\log x_i - \log x_{i+1}|$$
3. D から値が 0 の要素を削除する．
4. D の中央値 $\mathrm{med}(D)$ を算出する．ただし，D の要素数が 1 以下の場合，$\mathrm{med}(D) = 0$ とする．
5. $t = c_1 \mathrm{med}(D) + c_2$ とする．(c_1, c_2: 経験的に決定する値)

次に，U_S の各要素に対し 5.2.1 節のアルゴリズムを適用した結果得られた各推定数値域は以下の手順で統合する．

1. U_S の各要素に対し，$c_1 = 2.0$, $c_2 = 0.1$ とし，前述の閾値決定手法と 5.2.1 節の数値域推定アルゴリズムを用いて数値域を推定する．
2. $A = \{\alpha_k | k \in \mathbb{N}, k < m\}$，$B = \{\beta_k | k \in \mathbb{N}, k < m\}$ とする．α_k, β_k は U_S の要素 S_k に対する推定数値域の上限，下限，m は U_S の要素数がである．
3. A と B に対して，$c_1 = 1.0$, $c_2 = 0.1$ とし，前述の閾値決定手法と 5.2.1 節のアルゴリズムを用いて数値域を推定する．
4. A に対して得られた推定数値域の上下限を α_A, β_A，B に対して得られた上下限を α_B, β_B とする．
5. α_A を「統合上限」，β_B を「統合下限」とする．

最後に上記の統合上下限を補正する．これは統合上下限をそのまま最終的な値情報の上下限とすると，実際の数値域より狭い範囲となることがあるためである．数値域はもし実際より広めに推定したとしても，数値域推定をせずに時系列統計情報を抽出した場合と比較して，適合率・再現率が低下することはない．しかし，逆に実際より狭く数値域を推定した場合，再現率が低下するため大きな問題となる．

補正に用いる式は以下の通りである．これらの式では統合上限 α_A と統合下限 β_B に加え，統合手順で述べた β_A, α_B も用いる．式 (2) が上限の補正式，式 (3) が下限の補正式，α^* が最終的な推定数値域の上限，β^* が最終的な下限となる．

$$\alpha^* = \alpha_A^2 / \beta_A \quad (2)$$

$$\beta^* = \beta_B^2/\alpha_B \quad (3)$$

5.3 数値域推定の具体例

5.1 節で抽出した U_S の各要素に対し，5.2.1 節のアルゴリズムと 5.2.2 節の推定数値域統合手法を適用し，値情報の数値域を推定する具体例を以下に述べる．ここでは統計量名「ＰＨＳの加入台数」に対する値情報の数値域推定を考える．

表 1 は「ＰＨＳの加入台数」についての U_S の抽出結果例および U_S の要素別の数値域上下限推定結果を示す．ただし，表 1 は一例であり，実際の U_S の抽出結果は表 1 に示したデータ数より多くの値情報が抽出される．

表 1: 「ＰＨＳの加入台数」の数値域推定例 (単位: 台)

「ＰＨＳ」「加入」		「ＰＨＳ」「台数」		「加入」「台数」	
1735100		227000		1160000	
3486600	β_1	1735100		5634500	
4937000	*	4000000	β_2	7174000	
5634500	*	4937000	*	17190000	β_3
5857000	*	5634500	*	25000000	*
6088000	*	6472000	*	34973000	*
6652000	*	6568000	*	47300000	*
7068000	α_1	6568000	*	63800000	*
10000000		7067000	α_2	85260000	*
39788000		10000000		116000000	α_2

表 1 において右側の欄が空白の数値はその値が推定数値域外の値であることを示す．表 1 は 5.2.2 節の推定数値域統合手法の手順 1 の適用結果を示すが，さらに手順 2〜4 を適用すると手順 4 の四つの値は次のようになる．

$$\alpha_A = \alpha_1 = 7068000, \quad \beta_A = \alpha_2 = 7067000$$

$$\alpha_B = \beta_2 = 4000000, \quad \beta_B = \beta_1 = 3486600$$

そして式 (2), (3) から最終的な数値域の上下限を算出すると，「ＰＨＳの加入台数」の値情報の推定数値域の上限は 7069000，下限は 3039094 と求めることができる．

6. 実験と考察

6.1 実験課題

本論文においては，実験課題 (クエリ) として NTCIR-7 ワークショップにおける MuST タスクの一つ T2N サブタスク [Kato 08] の評価課題 25 件から以下の 8 件を選出し，これを提案手法の評価に用いた．

1. レギュラーガソリンの全国平均店頭価格
2. ドバイ原油価格
3. 携帯電話の加入者数
4. ＰＨＳの加入台数
5. 固定電話の加入台数
6. 鉱工業生産指数
7. 鉱工業出荷指数
8. 鉱工業在庫指数

6.2 実験結果と考察

6.1 節の課題に対する 5. 節の手法を用いた値情報の数値域推定結果ならびに正解データ値情報の上下限値を表 2 に示す．

表 2 を見ると，まず全ての推定上限が正解上限より小さい，または推定下限が正解下限より大きい値となっていないことが分かる．これより，時系列統計情報の抽出結果から値情報が推定数値域外の情報をフィルタリングすることによって，6.1 節に示した課題においては抽出再現率が低下することはないことが言える．また，課題#3 の推定下限，#5 の推定上限は正解上下限から比較的大きく離れた値となっているが，その他の推定上下限については正解上下限との比が 2 以下であり，推定値と正解値が近い値になったと言うことができる．

表 2: 推定数値域と正解値情報の上下限値の比較

#	推定下限	推定上限	正解下限	正解上限
1	61	204	91	105
2	9.7	44.9	10.0	40.0
3	17,958,915	116,000,000	48,475,000	66,390,200
4	3,039,095	7,069,000	5,634,500	5,870,000
5	47,300,000	222,855,250	55,520,000	61,530,000
6	88.2	142.5	98.9	108.3
7	79.5	182.9	103.9	108.4
8	78.8	174.6	93.0	95.9

7. 終わりに

本論文では，時系列統計情報の抽出におけるより高い適合率を達成するため，新聞記事から抽出される統計値情報を用い，値情報の数値域を推定する手法を提案した．評価実験の結果，推定数値域は，数値域外の情報をフィルタリングすることによって統計情報抽出の再現率を低下させることなく，適合率を向上させる可能性が高いものとなることが分かった．

今後の課題は，抽出統計情報フィルタリングによる適合率の向上について調査を行う．

参考文献

[Inoue 08] Inoue, T., Yamamoto, T., Toriyabe, M., Shimizu, E., Susuki, H., Saito, H.: Extraction of Chronological Statistics Using Domain Specific Knowledge, Proceedings of NTCIR-7 Workshop Meeting, pp. 494-501, 2008.

[Kato 08] Kato, T., Matsushita, M.: Overview of MuST at the NTCIR-7 Workshop - Challenges to Multi-modal Summarization for Trend Information -, Proceedings of NTCIR-7 Workshop Meeting, pp. 475-488, 2008.

[衛藤 05] 衛藤 純司, 奥村 学: 文書横断文間関係タグ付きコーパスの構築, 言語処理学会第 11 回年次大会発表論文集, pp. 482-485, 2005.

[加藤 04] 加藤 恒昭, 松下 光範, 平尾 努: 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, Vol. 2004-NL-164, No. 15, pp. 89-94, 2004.

[曽我 06] 曽我 真也，斎藤 博昭: 動向情報提示システムの構築，言語処理学会第 12 回年次大会ワークショップ「言語処理と情報可視化の接点」論文集, pp. 5-8, (2006).

[難波 05] 難波英嗣，相澤輝昭，国政美伸，福島志穂，奥村学: 文書横断文間関係を考慮した動向情報の抽出と可視化，電子情報通信学会技術研究報告，Vol. 105，pp. 67-74，2005.