統計値の分布密度推定に基づく動向情報の抽出

井上裁都, 鈴木宏哉, 斎藤博昭
慶應義塾大学大学院 理工学研究科
Email: {inoue,susuki,hxs}@nak.ics.keio.ac.jp

1　はじめに

ここ数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」[3] 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た MuST ワークショップ[1]がある．

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「テキスト群からの統計動向情報の自動抽出」がある．これはテキスト群中で，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ことを目的とする．

本論文では，より高い統計動向情報の抽出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統計値情報 (以後，値情報) の取りうる範囲 (以後，数値域) に着目し，カーネル密度推定法による統計値の推定分布密度に基づいた値情報数値域の推定手法，及びその動向情報抽出への適用を提案する．

2　統計動向情報

本論文において，統計動向情報は「統計量名」「値情報」「日付情報」の三つ組を一要素とする情報の集合であると定める．今定義した「三つ組」の具体例を示す．

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

また，統計動向情報抽出における入力クエリとしては，三つ組の一要素の「統計量名」(1 個),「統計量名の別名」(任意個),「統計量の単位」(1 個以上) を想定する．クエリの具体例を示す．

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

[1]http://must.c.u-tokyo.ac.jp/

3　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数値域と呼び，これを推定して，テキストから誤抽出された値情報から推定数値域外の値を除外することで統計動向情報抽出の適合率を向上させることを考える．これにより，例えばクエリ (統計量名) が「レギュラーガソリンの全国平均店頭価格」であれば，数値域を「80 円から 200 円まで」と推定することで，下記の文中から「4 円」が値情報として誤抽出されたとしてもこの除外が可能となる．

> 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。

もし上述のような誤抽出情報を除外しない場合，統計動向情報を可視化しようとすると，図 1 に示されるような不適切な可視化となる．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．

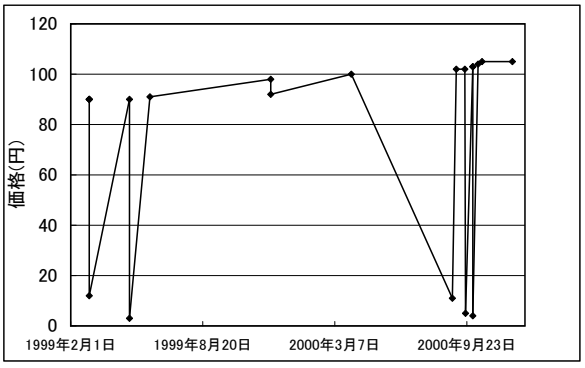

図 1: 動向情報の可視化結果例

図 1 を見ると，価格が 20 円以下のプロットが複数存在していることがわかるが，人間であれば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．もしシステムが同様の判断をすることができれば，図 1 中の誤抽出情報を除外でき，より動向の分かりやすいグラフの出力が可能と考えられる．

4 従来手法

筆者らは以前より統計動向情報抽出のための数値域推定の手法を提案している [1][2]．以前の手法は，テキストから抽出された値情報間の数値比を距離とみなし，距離が閾値以下の値情報同士をクラスタリングして，最大のクラスタの最小値，最大値に補正値を乗算したものを推定数値域とする．この手法は，クラスタリングのための閾値，ならびに最大クラスタの最小値、最大値の補正値を適切に設定する必要があり，これらの設定方法が根拠に乏しくアドホックであるという問題点がある．

本論文では，数値域推定の手法を一新し，カーネル密度推定法による統計値の推定分布密度に基づいた数値域推定の手法を提案する．本手法もまた以前の手法同様，設定困難なパラメータが存在するが，本手法ではこのパラメータの設定法に意味付けが存在し，数値域推定のアドホックさが軽減される利点がある．

5 提案手法

入力として，新聞記事コーパスと 2 節で述べた「統計量名」,「統計量名の別名」,「統計量の単位」から成るクエリを想定する．

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，記事抽出，パラグラフ抽出は以前に提案した手法で用いたものと全く同じ手法を用いる．詳細は [2] を参照されたい．なお [2] にて,「統計量名」および「統計量の別名」から抽出された名詞の形態素を「キーワード」と呼称しているが，本論文でもこの用語を用いる．また本論文では,「統計量名」「統計量の別名」から抽出されるキーワードの集合を W_0 とする．

値情報抽出は，[2] の手法と異なり，値情報 x_i を x_i と共起するキーワード集合 W_i と組にして抽出する．x_i および共起するキーワード集合 W_i はパラグラフ抽出で得られたパラグラフに下記の値情報抽出アルゴリズムを適用することで抽出する．ここで値情報とは「数値表現」と「統計量の単位」が連結した文字列 (例: 1 万円) かつ「比較表現」と係り受け関係にないものである.「比較表現」とは「1 円値上がり」の「値上がり」のような値情報の変動量を示す表現を示す．また，下記アルゴリズム中の「日付表現」とは三つ組中の一要素「日付情報」に変換可能な表現である．数値表現，比較表現，日付表現は人手で作成した辞書に基づき抽出する．さらにここでは，キーワードとは前述の W_0 に含まれるものとする．

1. $i = 1$ とする．
2. 抽出パラグラフ中の未探索パラグラフを d_p とする．未探索パラグラフがなければ終了する．
3. d_p 先頭から値情報を探索．発見できなければ手順 2 に戻る．
4. 値情報を発見した場合，これを x_i とし，d_p 先頭から x_i の間に出現するキーワードの集合を W_i とする．
5. x_i の位置以降からキーワード，値情報，日付表現を探索する．
6. 値情報を発見するか，キーワード，日付表現を発見できない場合，手順 2 に戻る．
7. キーワード，日付表現を発見した場合，$i = i + 1$ とし，その位置以降から値情報を探索．発見できなければ手順 2 に戻る．
8. 手順 4 に戻る．

上記アルゴリズム手順 4 のキーワードの出現の判定においては,「携帯電話とＰＨＳ」のような「並立表現」中など，クエリにより単純に「キーワード文字列」の出現を「キーワード」の出現と判定すべきではない場合がある．これに対処するため，手順 4 ではある種のクエリに特化した数個のルールが実際には適用されるが，本論文では説明を割愛する．

例として，クエリが「レギュラーガソリンの全国平均店頭価格」のとき，下記の例文に対して値情報抽出アルゴリズムを適用すると，値情報「１０３円」と組のキーワード集合として「ガソリン」「全国」「平均」「店頭」「価格」が抽出される．このとき,「８３円」や「1 円」が値情報として抽出されることはない．

> 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。

5.2 カーネル密度推定

本論文では，5.1 節の手法で抽出される値情報の数値の分布を考えたとき，正解情報はこの分布の中で密度の高い領域に集中すると仮定する．これは，誤抽出情報は新聞記事中に現れる任意の値を取ると考えられるのに対し，正解情報は 3 節で述べたような取り得る値の範囲が存在すると考えられるためである．そこで本手法ではまず，5.1 節の手法で得た値情報を標本としてカーネル密度推定法 [5] を適用し，分布密度推定を行う．カーネル密度推定の定義式 $\hat{f}_h(x)$ を (1) に示す．ここで，N は標本数，h はバンド幅 (平滑化パラメータ, $h > 0$)，K はカーネル関数，x_i は標本を示す．

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{x - x_i}{h}) \tag{1}$$

密度推定する上でまずは確率変数を考える必要がある．新聞記事に現れる統計値は一桁から十桁以上 (億，兆) まで非常に幅が広い．これは統計値をそのまま確率変数とした場合，カーネル密度推定のパラメータ h の設定が難しくなるという問題を起こす．そこで本手法では，統計値 X の対数，つまり $Y = \log X$ を確率変数とする．これにより h は固定の値でも経験的に十分な精度で密度推定が可能になる．

またカーネル密度推定ではカーネル関数として平均 0，分散 1 のガウス関数が一般的に用いられるが，統計値 X の分布が左右対称な分布とすると，Y の分布は左裾の重い分布となる．このため，本手法ではガウス関数 K_1 から K_2，K_3 へと変数変換を行い，K_3 をカーネル関数として用いる．

$$K_1(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$$

$$K_2(x) = \begin{cases} \sqrt{\frac{2}{\pi}} \exp(-\frac{x^2}{2}), & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$K_3(y) = \sqrt{\frac{2}{\pi}} \exp(y - \frac{e^{2y}}{2}),\ y = \log x \tag{2}$$

さらに本手法では，抽出値情報 x_i に対し，組にして抽出された W_i を利用して，x_i に重み付けをする．具体的には，W_i の各要素 w に対し式 (3) より IDF を求め，式 (4) より重み m_i を算出する．式 (3)(4) において N は抽出値情報の総数，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．

$$\mathrm{IDF}(w) = \log \frac{|D_a|}{f_d(w, D_a)} \tag{3}$$

$$m_i = \frac{\sum_{w \in W_i} \mathrm{IDF}(w, D_a)}{\sum_{i=1}^{N} \sum_{w \in W_i} \mathrm{IDF}(w, D_a)} \tag{4}$$

従って，本手法の分布密度推定式は式 (1)(2)(4) を用いて，式 (5) のようになる．

$$\hat{f}_h(y) = \frac{1}{h} \sum_{i=1}^{N} m_i K_3(\frac{y - \log x_i}{h}) \tag{5}$$

クエリが「レギュラーガソリンの全国平均店頭価格」のとき，$h = 0.5$ とし式 (5) を用いて分布密度推定した例を図 2 に示す．図 2 は縦軸が推定分布密度，横軸が統計値を示す．

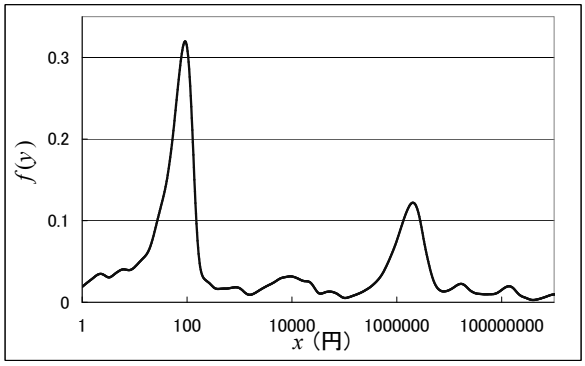

図 2: カーネル密度推定の例

図 2 を見ると，実際の「ガソリン価格」に近い 100 円付近が最も密度の高い領域になっていることがわかる．

5.3 数値域推定

数値域は式 (5) より得られる推定分布密度を用いて推定する．ここではまず数値域推定のアルゴリズムを以下に示す．

1. $\hat{f}_h(\log x)$ の値が最大となる x を $x_{\max}$ とする．
2. 閾値パラメータ k を下式より算出する．ここで，σ は値情報の対数値の標準偏差，a，b は経験的に決まる値であり，いずれも正の値とする．

$$k = \exp(-a\sigma - b)$$

3. $x < x_{\max}$ かつ $\hat{f}_h(\log x) < k\hat{f}_h(\log x_{\max})$ を満たす最大の x を数値域下限値 x_α とする．
4. $x > x_{\max}$ かつ $\hat{f}_h(\log x) < k\hat{f}_h(\log x_{\max})$ を満たす最小の x を数値域上限値 x_β とする．

手順 1 では密度が最大となる x を求めているが，これは手順 3, 4 で求まる数値域上下限は密度最大の x 前後の値と考えられるためである．

手順 2 の「値情報の対数値の標準偏差」は学習データの存在を想定したものである．つまり，正解値情報を人手でいくつか抽出済みであり，この対数値の標準偏差 σ を利用して数値域推定することを考えている．学習データの存在を想定した理由は，学習データが存在しない場合，複雑なルールを用意する必要が存在し，そのルールの汎用性が低い可能性が考えられたためである．よって本手法は学習データを用いた半教師ありの学習手法となる．

パラメータ k 算出式に σ を導入した理由は，標準偏差の大きさと数値域の広さの間の相関が考えられたことによる．また，k 算出式の b は k の上限を調整するパラメータであり，これは σ が 0 に近い値であっても数値域が極度に狭くならないよう導入されている．見方を変えれば，b は数値域上下限の補正項とも言うことができる．

クエリが「レギュラーガソリンの全国平均店頭価格」のときの数値域推定の具体例は以下のようになる．まず数値域推定アルゴリズム手順 2 の σ は 6.1 節で述べるデータ (1) を用いると，$\sigma = 0.051$ と求まる．これより $a = 1.2, b = 0.5$ とすると $k = 0.57$ となる．この k を用い，図 2 に示した分布密度推定の結果に対し手順 3, 4 を適用すると，$x_\alpha = 47.9$, $x_\beta = 131.8$ となる．

6 実験と考察

6.1 実験システムと実験課題

実験システムは，ルールベースに基づいた統計動向情報抽出システムであるベースラインシステム，5 節の手法により数値域を推定する数値域推定システムの二システムから成る．提案手法の評価は，ベースラインシステムより抽出された統計動向情報 (出力 A とする) に対し，数値域推定システムにて 3 節で述べたように推定数値域外の値情報を含む三つ組を除外した統計動向情報 (出力 B とする) を求め，出力 A と出力 B の適合率・再現率・F 値を比較することで行う．

実験システム内では形態素解析，係り受け解析に CaboCha[2]を利用した．数値域推定に用いる新聞記事コーパスとしては 1998 年から 2001 年までの毎日新聞の記事 4 年分を使用した．また，数値域推定におけるパラメータは経験的に $h = 0.5$, $a = 1.2$，$b = 0.5$ と設定した．

実験課題 (クエリ) としては，第 7 回 NTCIR ワークショップ[3] における MuST タスクの一つ T2N サブタスク[4][4] の評価課題 25 件から 14 件を選出し，これを提案手法評価に用いた．また，課題の正解データとしては T2N 課題参加者に対して配布されたデータを使用した．課題 14 件 (統計量名) は以下の通りである．

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

[2]http://chasen.org/~taku/software/cabocha/
[3]http://research.nii.ac.jp/ntcir/
[4]1 節で述べた「テキスト群からの統計動向情報の自動抽出」する課題

6. 鉱工業生産指数
7. 鉱工業出荷指数
8. 鉱工業在庫指数
9. ｉモード加入者数
10. ＥＺｗｅｂの加入者数
11. Ｊ―スカイ加入者数
12. ネット接続可能な携帯電話の総加入台数
13. センター試験の志願者数
14. センター試験の志願倍率

5.3 節の手法で σ の算出に必要な学習データとしても上述の T2N 課題の配布データを利用した．なお，正解データ数はクエリ 1 件に対し平均約 12 件となっている．実験は正解データを二分割し，一つを学習データとして利用して，正解データ全体を抽出対象とする実験を学習データを変え二回行った．分割は数値で昇順ソートし上位半分をデータ (1)，下位半分をデータ (2) とし二分割した．ソートした理由は，この状態のときデータの分散が最小になり，数値域推定の上で最も不利な状況になると考えられたためである．

6.2 実験結果と考察

6.1 節で述べた出力 A と出力 B の適合率・再現率・F 値を，学習データとしてデータ (1) を用いて評価した結果を表 1 に，データ (2) を用いて評価した結果を表 2 に示す．表 1，2 にて「出力」はシステムが出力した三つ組の総数,「正解」は「出力」の中の正解の個数を示す．また，全抽出対象記事中の総正解数は 160 である．

表 1: データ (1) で学習時の評価結果

	出力	正解	適合率	再現率	F 値
出力 A	172	98	0.570	0.613	0.590
出力 B	151	98	0.649	0.613	0.630
B/A	-	-	1.139	1.000	1.068

表 2: データ (2) で学習時の評価結果

	出力	正解	適合率	再現率	F 値
出力 A	172	98	0.570	0.613	0.590
出力 B	144	96	0.667	0.600	0.632
B/A	-	-	1.170	0.980	1.070

表 1 を見ると，出力 B は再現率の低下がなく，適合率が 13.9%，F 値が 6.8%向上していることがわかる．また表 2 を見ると，出力 B の再現率が 2%低下しているものの，適合率が 17.0%，F 値が 7.0%向上し，数値としては表 1 よりも良い結果となっていることがわかる．

データ (2) を学習データとしたとき，出力 B の再現率が低下したのは具体的にはクエリが「ｉモード加入者数」と「Ｊ―スカイ加入者数」の出力である．これらのデータは正解データの分布に偏りがあり，いずれも数値の大きいデータが数多く存在するという特徴がある．データ (2) は全データを昇順ソートしたときの下位半分，つまり大きい数値を含む方のデータであり，すると前述のデータ分布の特徴からデータ (2) はデータ (1) と比較して 5.3 節で述べた σ が小さな値を取ることになる．σ が小さいと 5.3 節で述べた通り，数値域は狭く推定される．もし，推定数値域が正解の数値域よりも狭ければ，推定数値域を外れ，正解数値域内にある正解の抽出情報が除外されるため，再現率が低下してしまうことになる．

逆にデータ (1) を学習データとすると,「ｉモード加入者数」「Ｊ―スカイ加入者数」の再現率の低下がなくなる．これはデータ (2) と逆に算出される σ が大きな値となり，数値域が広めに推定されるためである.「ｉモード加入者数」「Ｊ―スカイ加入者数」は時間の経過とともに統計値が単調増加する傾向にあり，かつ抽出対象の新聞記事においては古い情報より新しい情報の方が多いという傾向がある．これはつまりデータ (1) には古いデータが多く含まれることを意味する．古いデータを学習データとすれば再現率低下の危険性が低くなるかは検証の余地があるが，もしこれが正しいとすると，システムの目的を学習データとして古いデータを用い新しいデータを自動抽出することとすることで，再現率低下の危険性の低い統計動向情報抽出ができる可能性も考えられる．

7 終わりに

本論文では，より高い統計動向情報の抽出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からカーネル密度推定法によって統計値の分布密度を推定し，さらに値情報数値域を推定する手法の提案，及びその動向情報抽出への適用の評価を行った．評価実験の結果，提案手法は数値域外の情報のフィルタリングより情報抽出の再現率低下を抑えつつ，適合率を大きく向上させることがわかった．

今後の課題としては，数値域推定におけるパラメータ h，a，b の最適化方法の確立，または閾値パラメータ k の決定方法の改善，数値域推定の前処理手法の改善，MuST T2N タスクのクエリ 25 件中の未対応クエリへの対応，さらにその他のクエリに対する提案手法の有効性の検証が挙げられる．

参考文献

[1] 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).

[2] 井上 裁都, 鈴木 宏哉, 斎藤 博昭: 数値域推定を用いた時系列統計情報の抽出, 人工知能学会第 23 回全国大会論文集, 3F2-NFC3-4, (2009).

[3] 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).

[4] 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).

[5] Parzen, E.: On estimation of a probability density and mode, Annals of Mathematical Statistics, Vol. 35, pp. 1065-1076, (1962).