移動ウィンドウ方式に基づくテキストデータからのトピック検出

濱本　雅史[†]　　北川　博之[†,††]　　Jia-Yu Pan[†††]　　Christos Faloutsos[†††]

† 筑波大学 システム情報工学研究科　〒 305–8573 茨城県つくば市天王台 1–1–1
†† 筑波大学 計算科学研究センター　〒 305–8573 茨城県つくば市天王台 1–1–1
††† School of Computer Science, Carnegie Mellon Univeristy
E-mail: †hamamoto@kde.cs.tsukuba.ac.jp, ††kitagawa@cs.tsukuba.ac.jp, †††{jypan,christos}@cs.cmu.edu

あらまし　テキストストリームから与えられる大量のテキストデータより主要な話題を検出する技術は確立されておらず重要な研究課題の一つであるといえる。本論文ではテキストストリームに対し移動ウィンドウ方式でトピック検出を行う。ウィンドウとは一定の語数単位で区切られたテキストデータの断片であり、本論文では検出されたトピックとウィンドウ間の相互情報量を考慮することで最適なウィンドウ幅とトピック数を同時に求める手法を提案する。これら提案手法に対し実験を行いその有効性を検討する。

キーワード　トピック検出, テキストマイニング, 知識発見, テキストストリーム

Topic Detection from Text Data Using Moving Windows

Masafumi HAMAMOTO[†], Hiroyuki KITAGAWA[†,††], Jia-Yu PAN[†††], and Christos FALOUTSOS[†††]

† Graduate School of Systems and Information Engineering, University of Tsukuba, Tennohdai 1–1–1, Tsukuba, Ibaraki, 305–8573 Japan
†† Center for Computational Sciences, University of Tsukuba, Tennohdai 1–1–1, Tsukuba, Ibaraki, 305–8573 Japan
††† School of Computer Science, Carnegie Mellon Univeristy
E-mail: †hamamoto@kde.cs.tsukuba.ac.jp, ††kitagawa@cs.tsukuba.ac.jp, †††{jypan,christos}@cs.cmu.edu

Abstract　Reliable techniques are needed to detect the main topics in lots of text data given via text stream. This paper gives a topic detection method for text stream by moving window method. A window is a text fragment that is split from text data by some constant number of words. In this paper we propose a method to decide adequate window size and the number of topics by mutual information between detected topics and given windows. We examine effectiveness of our proposed method by experiments.

Key words　Topic Detection, Text Mining, Knowledge Discovery, Text Stream

1. はじめに

チャットや文字放送など、テキストストリームと呼ばれる継続的にテキストデータの配信を行うサービスが近年増加している。テキストストリームからは大量のテキストデータが与えられるが、それらを逐一読むことはほぼ不可能であり、テキストデータ中における主要な話題のみを提示することが有用である。ここで特定の話題のことをトピックと定義し、各トピックに対応する特徴的な単語群とトピックに合致するテキストデータの断片を発見することをトピック検出と定義する。

本論文ではテキストストリームに対し移動ウィンドウ方式でトピック検出を行う。移動ウィンドウ方式とは、データストリームから与えられるデータを一定量ごとに分割する手法である。この一定量をウィンドウ幅、分割されたデータをウィンドウと呼ぶ。本論文では移動ウィンドウ方式をテキストストリームに対して適用し、得られるウィンドウの集合からトピックを検出する手法を検討する。

移動ウィンドウ方式を用いる場合、ウィンドウ幅を定める必要がある。またトピック検出を行う場合、予めトピック数をパラメータとして設定する必要がある。本論文では検出されたトピックとウィンドウ間の相互情報量を考慮することで最適なウィンドウ幅とトピック数を求める手法を提案する。

我々の研究グループは、これまで各種トピック検出手法を比較検討してきた [8]。この結果独立成分分析 [10] を用いた手法が他の手法に比べ有効性が高いことが示唆された。本論文では今までと異なった尺度で評価を行い、独立成分分析を用いた手法の有効性を確かめる。その上でトピック検出手法の評価結果と最適ウィンドウ幅を決定する提案手法の実験結果が整合する

ことを示し、提案手法が有用であることを述べる。

本論文は以下のように構成される。2 章において関連研究について述べ、それらと比較した本研究の位置づけを行なう。3 章では既存のトピック検出手法を、移動ウィンドウ方式の文脈で比較検討する。4 章では相互情報量を用いたウィンドウ幅およびトピック数の決定手法を提案する。また人工データと実データを用いて実験を行い、提案手法の有効性を明らかにする。最後にまとめと今後の課題について述べる。

2. 関連研究

2.1 トピック検出に関する研究

トピック検出に関する既存の研究の中で最も主要なものとして、米 NIST 主催の Topic Detection and Tracking(TDT) がある [1]。これは主にトピック検出とトピック追跡（ユーザにより与えられたトピックがどの記事中に現れるかを提示すること）に関するコンペティション形式のプロジェクトである。この成果として、インクリメンタルなクラスタリングを用いる手法 [21] や、単連結の階層的クラスタリングを用いる手法 [20] などが提案された。この他にもこれまでに提案されたトピック検出手法の多くはクラスタリングをベースにしたものとなっている。

TDT 以外のトピック検出に関する研究では、単語の出現確率から作成された統計モデルを用いる手法 [15] や、自己組織化ニューラルネットワークにおけるクラスタリングの問題として扱う手法 [19] など、様々な手法が提案されている。そのうち独立成分分析を用いた手法として、文書集合からトピックとそのキーワードを見つける研究がある [12]。この研究では文書と単語で表される行列に対し独立成分分析を用い、トピックと単語で表された行列とトピックと文書で表された行列に分割する手法が提案されている。トピックと単語で表された行列では複数の文書が一つのトピックに圧縮されていると考えられ、これによりあるトピックを表す文書群からなるクラスタが発見される。逆にトピックと文書で表された行列によってあるトピックを表す単語群が発見される。この考えは本研究でも用いている。またその応用としてチャットログからのトピック検出の研究 [4] [13] がある。これらの研究はテキストデータの時間相関性に着目し、データに特化した独立成分分析アルゴリズムを用いることでトピック検出を行なっている。

これに対し、本研究は特徴軸という観点でトピック検出手法を一般化し、データ解析の分野で用いられている比較的汎用的なアルゴリズムを適用した。この各解析手法の相互比較を行い、各手法の性質を分析する。

2.2 ウィンドウ幅とトピック数に関する研究

本研究では、テキストストリームから与えられるテキストデータを固定の単語数ごとに分割する移動ウィンドウ方式 [3] を用いている。移動ウィンドウ方式は一般的なデータストリーム分析において用いられる手法である。一方 TDT の研究ではテキストストリーム中の話題の区切りを推定し、テキストの断片からトピックを検出する手法を用いている [2]。しかし区切りを推定するには学習データが必要であったり、断片間の類似度を定める必要がある [9]。

テキストデータ中のトピック数を推定する直接的な研究は見当たらないが、主成分分析で累積寄与率を考慮して主成分数を推定する手法 [11] や、クラスタリングにおけるクラスタ数の推定手法 [17] の利用が考えられる。しかし主成分数を推定する手法の場合、どの程度の累積寄与率がトピック数に当たるのかを予め定めておく必要がある。またクラスタ数の推定では、クラスタの形状が Gauss 分布であることが仮定されている。

これらに対し本研究で提案する相互情報量を用いた手法は、事前学習やパラメータの設定を必要としないだけでなく、トピックの内容に関する仮定も必要としない。また最適なウィンドウ幅とトピック数を同時に推定することが可能である。

3. トピック検出手法

3.1 移動ウィンドウ方式

移動ウィンドウはデータストリーム分析において用いられる手法のひとつである [3]。データストリームを分析する際、全てのデータを主記憶中に保持することは一般的に不可能である。しかし移動ウィンドウ方式では予め決められた一定量ごとにデータを分割することで分析を行う。この一定量をウィンドウ幅と呼び、ウィンドウ幅ごとに区切られたデータストリームの断片をウィンドウと呼ぶ。テキストストリームにおけるウィンドウは、テキストストリームから与えられたテキストデータの断片である。ただし単語の出現状況の前後関係を保つため、隣接するウィンドウは半分程度重ねる。

このようにして作成されたウィンドウの集合から、どのようにトピックを検出するかを次節において説明する。一方ウィンドウ幅をどのように定めるかについては次章で説明する。

3.2 特徴軸によるトピック検出

移動ウィンドウ方式により作成されたウィンドウの集合から、我々は特徴軸を発見することでトピック検出を行うことを検討してきた [8]。全テキストデータ中に m 個の語彙が用いられている場合、各ウィンドウは m 次元のベクトルとして表現できる。このとき同じトピックのウィンドウならばウィンドウ中の単語の出現状況は似ており、これらのウィンドウを表すベクトルはあるひとつの特徴的な軸の周りに分布していると考えられる。よって m 次元空間中におけるベクトルの分布より特徴軸を抽出する問題としてトピック検出の問題を捉えることができる。このとき抽出する特徴軸の個数を、検出するトピック数として予め設定する必要があるが、この決定手法については次章で説明する。

m 次元空間中に分布したウィンドウから特徴軸を抽出する手法として、我々は特異値分解、クラスタリング、独立成分分析の 3 種類を検討している。これらについて以下で具体的に説明する。

特異値分解

特徴的な軸を計算する古典的な手法として、特異値分解 (SVD: Singular Value Decomposition) を用いる方法がある。応用分野によって主成分分析 (PCA: Principal Component Analysis)、潜在的意味インデキシング (LSI: Latent Semantic Indexing)、KL 変換 (Karhunen-Loeve transformation) とも呼ばれるがすべて本質的

には同じ手法である。ここで見つかる軸にはデータの分散を最大にするという性質を持っている。具体的には行列を特異値分解して得られる主成分ベクトルを特徴軸とする方法で、考え方は非常にシンプルであるものの見つかった軸は文書中で共起する単語の関係を表すという性質を持つことが知られている [6]。ただしこれらの軸の間では直交するという制約を持つ。

クラスタリング

クラスタリングは似ているデータオブジェクト同士を同じグループ (クラスタ) に分類することである。各クラスタの性質を調べることで、データ中にどのような性質を持つオブジェクトがあるのかがわかる。クラスタリング手法は非常に様々な種類があるが、最も単純な手法と考えられるのが k-means 法 [16] である。本論文でもこれを用いる。

トピック検出の文脈で考えると、クラスタリングをウィンドウの集合に適用することで同じトピックのウィンドウを同じクラスタに分類することができる。このとき各クラスタの重心を示すベクトルがそのトピックを表す特徴軸であるとみなせる。同じトピックのウィンドウであるかを測る尺度として余弦尺度 (2つのベクトルの内積を各ベクトルのノルムで割ったもの) を用いる。各単語の出現割合が似ているとこの尺度の値が大きくなり、逆に異なる場合は値が小さくなる。

独立成分分析

独立成分分析 [10] は信号処理の分野で発展した、混ざり合った信号から元の信号を復元する手法の一つである。これは源信号が独立に発生するという仮定のもと、m 個の混合信号 $D = (\mathbf{d_1}, ..., \mathbf{d_m})^T$ から源信号 $S = (\mathbf{s_1}, ..., \mathbf{s_m})^T$ と混合係数行列 $A = (\mathbf{a_1}, ..., \mathbf{a_m})$ を推定する方法である。ここで $\mathbf{d}, \mathbf{s}, \mathbf{a}$ はそれぞれ列ベクトルとする。また各源信号 $\mathbf{s_i}$ は独立成分とも呼ばれる。これらの間には $D = AS$ という関係が成り立つ。推定には様々な手法があるが、最も一般的なのが非ガウス性を最大にする手法である。直観的には多くの信号が混ざりあった信号はよりノイズに近くなるということであり、ノイズは一般的にガウス分布で表されるので、非ガウス性が高い信号が求める信号であるということである。

独立成分分析を用いたトピック検出手法の手順は図 1 に挙げた通りである。第 1 段階 (図 1(1),(2)) では次元削減より予め推定されたトピック数 (ここでは k) の次元数に各ウィンドウを射影する。第 2 段階で実際に独立成分分析を行ない源信号と混合係数行列に分解する。このとき得られる源信号 S は、各特徴軸とウィンドウの合致度合を示す。第 3 段階では次元削減で用いた主成分ベクトルおよび混合係数行列を用いて特徴軸を計算する。

3.3 特徴軸抽出手法の比較検討

前節で述べた 3 種類の特徴軸抽出手法を、実データを用いた実験により比較検討する。実験には MATLAB を用いた。特異値分解とクラスタリングについては MATLAB 付属のパッケージを、独立成分分析については JADE [5] パッケージを用いた。

実験に使用したのは TDT2 [7] のデータである。これは CNN Headline News や New York Times News Service など 6 種類の配信源における 1998 年 1 月から 6 月までのニュース記事を収録

入力：X={x_1, ..., x_n}, k(特徴軸数)
(1) 特異値分解 $X = U\Lambda V$ を計算
(2) 左特異ベクトル U のうち対応する特異値が最大の k 個 U_k を用いて $X_k = U_k^T X$ と次元削減
(3) 独立成分分析により $X_k = AS$ を計算
(4) 特徴軸を $U_k A$ と計算

図 1 独立成分分析を用いた特徴軸の計算

Topic ID	トピック名
TP_1	アジア経済危機
TP_2	Monica Lewinsky
TP_3	長野オリンピック
TP_4	対イラク衝突
TP_5	スーパーボウル
TP_6	タバコ会社に対する健康被害訴訟
TP_7	インドの核疑惑
TP_8	イスラエルとパレスチナの対話
TP_9	インドネシアの反スハルト暴動
TP_{10}	爆弾犯 Theodore Kaczynski への判決
TP_{11}	ローマ法王のキューバ訪問
TP_{12}	アラバマ病院爆破事件
TP_{13}	イタリアのケーブルカー事故
TP_{14}	フロリダのトルネード被害
TP_{15}	Oprah Winfrey の狂牛病報道問題
TP_{16}	Gene Mckinney 軍曹の性的不品行事件
TP_{17}	バイアグラ
TP_{18}	Jonesboro での少年の銃乱射事件
TP_{19}	スペースシャトルでの生物実験
TP_{20}	General Motors のストライキ

表 1 使用した記事のトピック

したコーパスである。収録されたニュース記事の一部には、トピック付けの情報および記事とトピックとの適合具合 (完全に適合するか一部のみ適合するかの 2 種類) の情報が付加されている。以下に述べる 2 つの実験のうち実験 1 では表 1 に挙げた 20 個のトピックと完全に適合する CNN の記事を、実験 2 では表 1 の TP_1 から TP_{10} までの 10 個のトピック完全に対応する New York Times の記事を用いた。各トピックについてランダムに一定数の記事を選び使用した。この記事数は、実験 1 では 30 件、実験 2 では 20 件である。各記事に対し不要語の除去と語幹抽出を行い、ランダムに並べた記事を連結させたテキストデータを実験対象とした。ランダムに並べる理由は、この手法が対象とするデータがニュース記事に限るものではないことから、各トピックに関する時間相関性の影響を少なくするためである。

またこの実験ではトピック数は何らかの手法により推定できていると仮定する。

評価手法

論文 [8] で行っている評価実験では、各特徴軸が最も近い 1 トピックに合致するとみなし、その合致度合の平均値を評価値とした。しかしこの評価方法では、複数の特徴軸が同一のトピッ

クと合致することを許しており、多くの特徴軸が同一のトピックと合致する場合にも高い値を与えてしまう。そこで本論文では各トピックと各特徴軸が１対１に対応するという制約の元で評価を行う。

まず各トピック検出手法により得られた k(k は推定されたトピック数) 個の特徴軸 $v_1, ..., v_k$ それぞれについて各トピックを表すベクトル $t_1, ..., t_k$ との余弦の絶対値を計算する。その後各トピックと各特徴軸の余弦の絶対値の和が最大となるよう１対１に対応させる。この対応は Hungarian 法 [14] によって求めることができる。ここでトピックと特徴軸を対応させたときの余弦の絶対値の和の最大値 C_{max} に対し、トピックの検出率を $100(c_{max}/k)$ として定義する。

各トピックを表すベクトルは次のようにして得られる。各トピック i の全記事を連結した単語列を D_i ($1 \leqq i \leqq k$, k はトピック数) とする。一方全記事中の語彙 (語数 m) の語を $w_j (1 \leqq j \leqq m)$ とする。各 D_i について、D_i 中に含まれる w_j の頻度を tf_{ij}、w_j を含むトピックの数を df_j とする。このときトピック i について以下の m 次元ベクトルを与える。これはある特定のトピックのみに頻出する単語に対応する次元の値が大きくなるベクトルである。

$$t_i = (log(1+tf_{i1})log(\frac{k}{df_1}), \cdots, log(1+tf_{im})log(\frac{k}{df_m}))^T$$

実験１:CNN Headline News データ

この実験では TDT2 のデータのうち CNN Headline News データを用いた。全記事を連結した単語列の語数は 31,787 語、語彙数は 4550 語である。また元の 1 記事あたりの平均語数は 53 語である。このデータに対し、ウィンドウ幅を 16、32、64、128、192、256(語) と設定し実験を行なった。

評価結果は図 2 である。図の横軸はウィンドウ幅、縦軸はトピックの検出率を表している。クラスタリングに関しては試行の度出力が変化するので、10 回の試行における平均値と標準偏差を示している。この図から、ウィンドウ幅が極端に小さな場合を除き独立成分分析を用いた手法が他の手法よりも検出率が高くなっていることがわかる。特にウィンドウ幅が広くなるにつれ、クラスタリングを用いた手法は極端に検出率が下がっているが、独立成分分析を用いた手法は検出率が高く保たれていることがわかる。この理由として、独立成分分析はウィンドウ幅が広く様々なトピックが１ウィンドウ中に混在している状況に対し、比較的堅牢であるということが挙げられる。

実験 2:New York Times News Service データ

この実験では New York Times のデータを用いた。実験 1 で用いた CNN の記事と大きく異なるのは全体の語彙数と 1 記事当たりの平均語数である。語彙数は 8609 語であり、元の 1 記事当たりの平均語数は約 411 語である。また各記事を連結した単語列の語数は 82,104 語となる。このデータに対し、ウィンドウ幅を 64、128、256、384、512、640(語) と設定し実験を行なった。

評価結果は図 3 である。図の横軸と縦軸は実験１と同様である。この図を見ると、どの手法も最適なウィンドウ幅を持つという全体的な傾向は実験１と同じであることが分かる。また

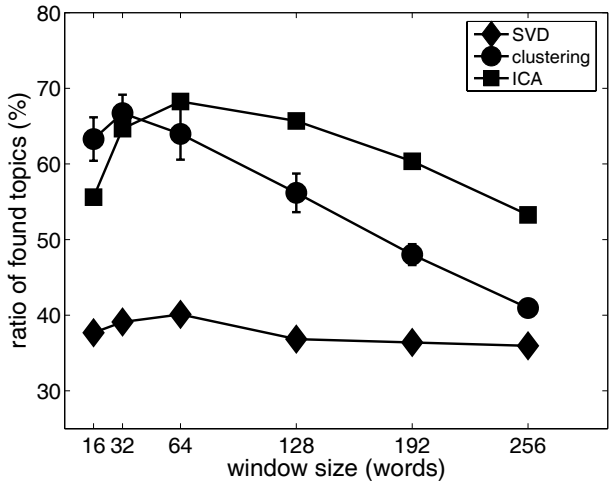

図 2　実験１におけるウィンドウ幅とトピックの検出率

個々の手法についてもその傾向は実験１と類似している事が分かるが、特に独立成分分析が他の手法よりも高い検出率を安定して与えている。これは次のような理由だと考えられる。実験１と比べ、ウィンドウ幅が広い場合でも１ウィンドウに含まれるトピック数は少なくなるが語彙数が多くなっている。このとき各ウィンドウ中には、各トピックと直接関連していない語が多数含まれるが、独立成分分析はこのような状況であっても適切な特徴軸を発見できていると考えられる。

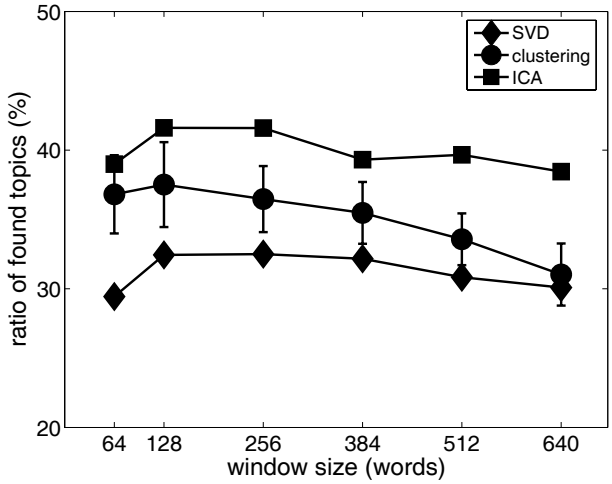

図 3　実験２におけるウィンドウ幅とトピックの検出率

4. ウィンドウ幅とトピック数の決定手法

前章ではトピック数がすでに分かっているという仮定の元で、特徴軸を用いたトピック検出手法を定量的に比較した。その結果いずれの手法でも最適なウィンドウ幅が存在し、その幅より狭くても広くても評価値が下がることがわかった。また、独立成分分析を用いた手法は他の手法よりもウィンドウ幅の変化に対し比較的堅牢であることがわかり、有用な手法であることが示された。

そこで本章では独立成分分析を用いたトピック検出手法について、最適なウィンドウ幅とトピック数を決定するための手法を提案するとともに、実験によりその有効性を評価する。

4.1 提案手法

移動ウィンドウ方式においてウィンドウ幅をどう定めるかが大きな問題となる。ウィンドウ幅が狭すぎる場合、ウィンドウの数が膨大になるだけでなく同じトピックを表す単語群が共起する頻度が低くなる。逆にウィンドウ幅が広すぎる場合、複数のトピックがひとつのウィンドウに含まれてしまうために適切なトピックを検出することができない。

前章で説明した独立成分分析を用いたトピック検出手法では、検出された各トピックを表す特徴軸と各ウィンドウの対応度合を表す行列が源信号として得られる。もしウィンドウ幅が最適な値のとき、すなわち前章の実験においてトピックの検出率が最も高くなる場合、検出されたトピックのいずれか１つのみに対応するウィンドウが増加すると考えられる。一方ウィンドウ幅が適していない場合、トピックをうまく捉える事ができていないので、複数のトピックと対応してしまうウィンドウが増加すると考えられる。

具体的な例として、前章の実験１により得られた源信号のうち表 1 の $TP_7, TP_8, TP_9, TP_{12}, TP_{15}$ に対応する信号を図 4,5,6 に示す。ただし各源信号は平均 0、分散 1 に正規化後、絶対値を取ってある。また図のスケールを合わせるため、分析対象の単語列の先頭から約 12,800 語までのウィンドウのみを示した。各図の横軸は先頭からのウィンドウの番号を表し、縦軸は源信号の絶対値、すなわち特徴軸とウィンドウの対応度合を表す。また各特徴軸に対応する単語群を表 2,3,4 に示す。各特徴軸から単語群を抽出する手法であるが、ここでは特徴軸の成分の絶対値が最大のもの 10 個に対応する単語群を抽出している。

図 4 はウィンドウ幅が 64 語の場合であるが、どの源信号も一定数のウィンドウで対応度合が高くなっている。また各ウィンドウにおいて、いずれかの特徴軸との対応度合が高くなる場合には、他の特徴軸との対応度合は低くなっている。実際に表 2 の単語群を見ると妥当な単語群が選ばれていることがわかる。

一方図 5 に挙げたウィンドウ幅が 16 語の場合、TP_9 以外の源信号はウィンドウ幅 64 語の場合と非常に似た傾向を持っている事が分かる。しかし TP_9 に対応する源信号は非常に多数のウィンドウで対応度合が高くなっている。この特徴軸から得られた単語群を表 3 で見るとウィンドウ幅が 64 語の場合と比べ dai, week, time など、比較的一般的な語が多く得られている事が分かる。これはウィンドウ幅が狭すぎるために起こった現象だと考えられる。

ウィンドウが 256 語の場合の源信号は図 6 である。この図をウィンドウ幅が 64 語の場合と比べると、各源信号のピークの位置は類似しているが、ピークであると明確に識別できるウィンドウが少なくなっている。すなわちウィンドウ中にあるトピックのテキストデータが含まれているにも関わらず、そのトピックの対応度合が高くならないという現象が多数起こっている。例として TP_7 に関する源信号では先頭から 10 番目付近に１つのピークがあるが、他のウィンドウについては明確にピークであると識別できない。しかしウィンドウ幅が 64 語の場合 (図 4) や 16 語の場合 (図 5) を見ると多くのピークがある事が分かる。これはウィンドウ幅が広すぎるために起こった現象だと考えられる。

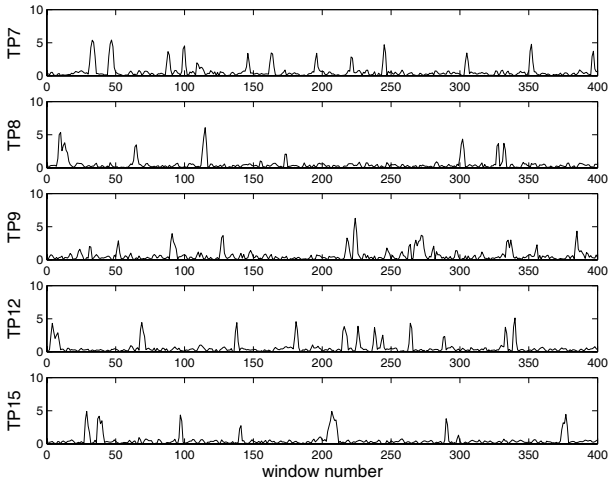

図 4　ウィンドウ幅 64 語に得られた源信号

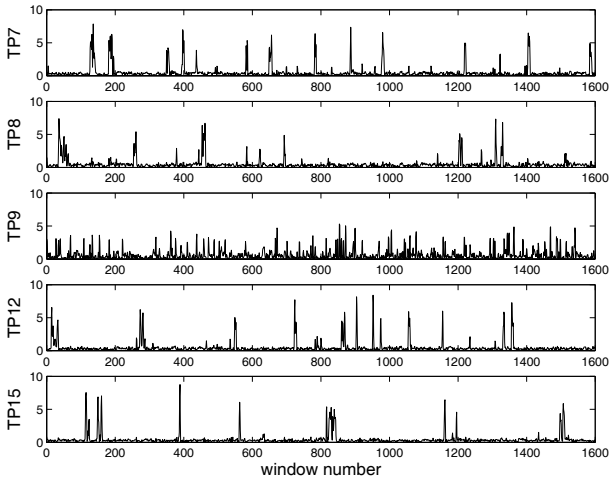

図 5　ウィンドウ幅 16 語に得られた源信号

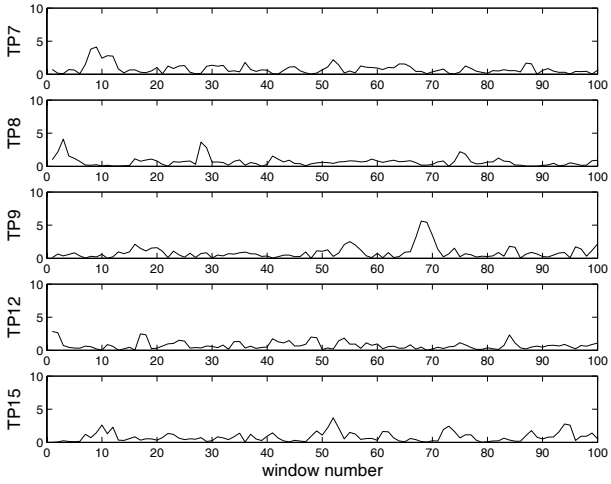

図 6　ウィンドウ幅 256 語に得られた源信号

我々は検出されたトピックとウィンドウの間の相互情報量を用いてウィンドウ幅を決定する手法を提案する。これは検出さ

TP_7	TP_8	TP_9	TP_{12}	TP_{15}
test	isra	presid	bomb	winfrei
india	netanyahu	suharto	clinic	beef
nuclear	palestinian	indonesia	rudolph	oprah
pakistan	peac	student	birmingham	texa
countri	israel	jakarta	women	talk
peopl	minist	govern	eric	cattl
conduct	prime	protest	alabama	rancher
treati	benjamin	habibi	fbi	cow
clinton	talk	indonesian	suspect	diseas
nation	arafat	dai	wit	mad

表 2　ウィンドウ幅 64 語の場合に抽出された単語群

TP_7	TP_8	TP_9	TP_{12}	TP_{15}
test	isra	dai	bomb	winfrei
india	netanyahu	week	clinic	oprah
nuclear	minist	suharto	rudolph	beef
pakistan	palestinian	student	women	talk
countri	prime	talk	birmingham	texa
conduct	peac	presid	alabama	rancher
unit	talk	time	eric	cattl
peopl	israel	clinton	wit	cow
secur	albright	countri	suspect	mad
weapon	benjamin	indonesia	robert	diseas

表 3　ウィンドウ幅 16 語の場合に抽出された単語群

TP_7	TP_8	TP_9	TP_{12}	TP_{15}
pakistan	isra	student	bomb	winfrei
india	palestinian	indonesia	rudolph	beef
nuclear	netanyahu	jakarta	clinic	oprah
test	israel	forc	eric	texa
winfrei	peac	clinton	birmingham	talk
texa	minist	fan	women	india
oprah	prime	violenc	alabama	test
beef	talk	demonstr	fbi	nuclear
rancher	arafat	viagra	presid	cattl
tobacco	yasser	militari	suspect	cow

表 4　ウィンドウ幅 256 語の場合に抽出された単語群

れたトピックと各ウィンドウがどれだけ 1 対 1 の対応に近いかを相互情報量により示す方法である。本論文では、相互情報量が区間 [0,1] で表されるよう正規化された不確定性係数 [18] を用いる事でウィンドウ幅を決定する。不確定性係数は、2 つの確率変数 X, Y に対し以下の式で表される。ここで $H(X)$ は確率変数 X のエントロピーを表す。

$$U(X,Y) = 2\left[\frac{H(X)+H(Y)-H(X,Y)}{H(X)+H(Y)}\right]$$

この値がより 1 に近い程 X と Y は 1 対 1 に近い事を示す。逆に X と Y に全く相関性が見られないとき、この値は 0 に近づく。従ってこの不確定性係数が高い値を与えるウィンドウ幅およびトピック数が適切な値であると考えられる。

具体的に、n 個のウィンドウに対して m 個の独立信号 $S = (\mathbf{s_1}, \cdots, \mathbf{s_m})^T, \mathbf{s_i} = (s_{i1}, \cdots, s_{in})^T$ が与えられた場合を考える。確率変数 $X = 1, 2, \cdots, m$ を検出されたトピック、確率変数 $Y = 1, 2, \cdots, n$ をウィンドウとすると、$H(X), H(Y), H(X,Y)$ は次のように計算される。

$$H(X) = -\sum_{p=1}^{m}\left[\frac{\sum_j |s_{p,j}|}{\sum_{i,j}|s_{i,j}|} log(\frac{\sum_j |s_{p,j}|}{\sum_{i,j}|s_{i,j}|})\right]$$

$$H(Y) = -\sum_{q=1}^{n}\left[\frac{\sum_i |s_{i,q}|}{\sum_{i,j}|s_{i,j}|} log(\frac{\sum_i |s_{i,q}|}{\sum_{i,j}|s_{i,j}|})\right]$$

$$H(X,Y) = -\sum_{p=1}^{m}\sum_{q=1}^{n}\left[\frac{|s_{p,q}|}{\sum_{i,j}|s_{i,j}|} log(\frac{|s_{p,q}|}{\sum_{i,j}|s_{i,j}|})\right]$$

4.2 実　　験

本節では、人工データと前章で用いた 2 種類の実データの計 3 種類のデータで提案手法を実験し、その有効性を確かめる。

4.2.1 人工データによる実験

ここでは人工的に生成された疑似テキストストリームを用いた実験を行い、適切なウィンドウ幅とトピック数が決定できるか検討する。

実験では、トピック数が 10 個であり、各トピックがトピック間で重ならない 100 語の語彙を持つテキストストリームを想定する。また各トピックセグメントはそれぞれのトピックの語彙からランダムに選択された 100 単語から成る。トピックセグメントの出現の仕方は、全 10 トピックをランダムに並べたものを、同じトピックが連続しないよう 100 回連結させた順とする。このような実験データに対し、ウィンドウ幅を 20 語から 300 語まで 20 語単位で、抽出する特徴軸数を 5 個から 50 個まで 5 個単位で変化させたときの不確定性係数を計算した。

実験結果は図 7 である。図の中心から左奥方向の軸はウィンドウ幅、右奥方向の軸は抽出した特徴軸数、縦軸は不確定性係数を表す。また図における各マスの幅はウィンドウ幅が 20 語、特徴軸数が 5 個である。各マスの色は、各マスの頂点のうちウィンドウ幅が狭く特徴軸数が少なくなる頂点における不確定係数の値を表している。この値が全体の最小値に近い程黒、全体の最大値に近い程白になる。

この図を見るとウィンドウ幅が 100 語かつ特徴軸数が 10 個の地点に 1 つのピークがあることがわかる。一方ウィンドウ幅が非常に狭いときには特徴軸の数が増加するにつれ不確定性係数が高い値を示している。この原因は前節で述べたとおり、ウィンドウ幅が狭すぎるためである。

しかしウィンドウ幅が極端に狭くトピック数が多い場合、不確定性係数が高くなるという現象が起こっている。ウィンドウ幅が狭い場合 1 ウィンドウ中に含まれる語数が極端に少ないため、トピックという大域的な観点ではなく、個々のウィンドウにおいてごく少数の単語群が出現するかという局所的な観点になってしまう。よって検出されるトピックがごく一部のウィンドウのみに対応するということになり、結果として不確定性係数が高くなる。

4.2.2 実データによる実験

本節では 3.3 節の実験 1 と実験 2 で用いたデータに対し、ウィ

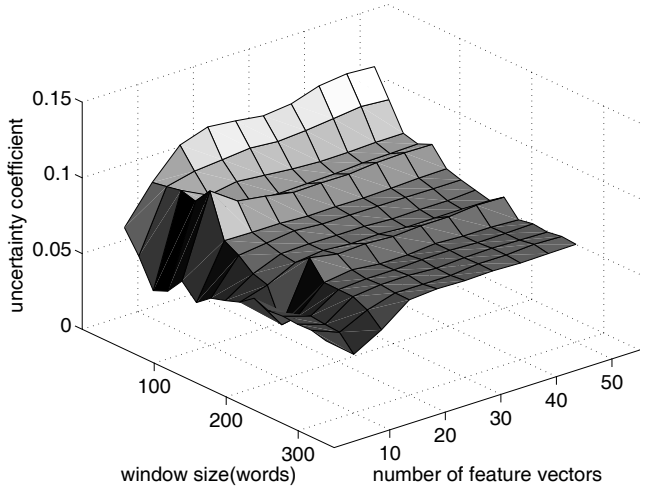

図 7　人工データによる実験のウィンドウ幅・特徴軸数・不確定性係数

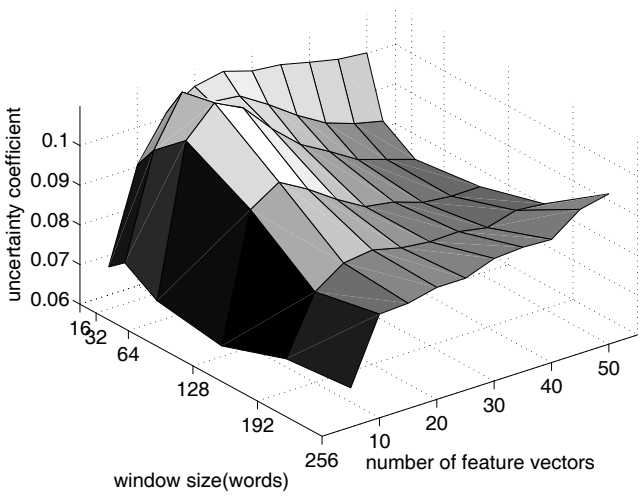

図 8　実験 1 におけるウィンドウ幅・特徴軸数・不確定性係数

ンドウ幅と抽出される特徴軸数を変化させて不確定性係数を計算する。

実験 1: CNN Headline News データ

ここでは 3.3 節の実験 1 、すなわち表 1 の 20 トピックに関する CNN の記事を用いた実験データに対し実験を行った。この実験データに対しウィンドウ幅を 16, 32, 64, 128, 192, 256 語の 6 種類、特徴軸数を 5 個から 50 個まで 5 個きざみで 10 種類の計 60 通りについて不確定性係数を計算した。

実験結果は図 8 である。図の各軸の意味は人工データと同様である。また図のセルにおけるウィンドウ幅の間隔は、左奥から手前に向けて 16,32,64,64,64 語である。この図を見ると、特徴軸数が 15 個でウィンドウ幅が 64 語の場合に不確定性係数が最大となっている。ウィンドウに関しては 3.3 節の実験 1 の結果 2 と整合し、特徴軸の数は実験データに含まれるトピック数と多少異なるものの、10 から 20 付近という指標を与えることができており、本提案手法が有効であると考えられる。一方でウィンドウ幅が 16 語の場合に不確定性係数が高い現象はこの実験でも見られるが、人工データの場合よりも極端ではないことがわかる。これは語彙数が人工データに比べ非常に増えたことで、ごく局所的に単語が共起する確率が低下したためだと考えられる。

実験 2: New York Times News Service データ

ここでは 3.3 節の実験 2、すなわち表 1 の TP_1 から TP_{10} までの 10 トピックに関する New York Times の記事を用いた実験データに対し実験を行った。この実験データに対しウィンドウ幅を 64, 128, 256, 384, 512, 640 語の 6 種類、特徴軸数を 5 個から 30 個まで 5 個きざみで 6 種類の計 36 通りについて不確定性係数を計算した。

実験結果は図 9 である。図の各軸の意味は人工データおよび実験 1 と同様である。また図のセルにおけるウィンドウ幅の間隔は、左奥のみ 64 語で残りは 128 語である。この場合も 3.3 節の実験 2 の結果である図 3 と整合し、ウィンドウ幅が 128 語から 256 語でかつ特徴軸数が 10 個のところでピークとなっている。

一方人工データおよび CNN データの場合と異なり、ウィンドウ幅が非常に広い場合、特徴軸数が多くなるにつれ不確定性係数の値も大きくなっている。一般的にウィンドウ幅が広いとき、全体のウィンドウ数も減少するため、特徴軸数とウィンドウ数が近づくにつれ 1 対 1 の対応により近くなる。実験 1 ではウィンドウ幅が広い場合、ウィンドウ中に多くのトピックの記事が含まれていたため、特徴軸とウィンドウの対応づけが適切でなく、結果として不確定性係数が低いままだったと考えられる。しかし実験 2 では元の New York Times の 1 記事当たりの語数が多いため、ウィンドウ幅が広くても特徴軸とウィンドウの対応づけが比較的適切に行うことができ、不確定性係数が高い値を与えたと考えられる。

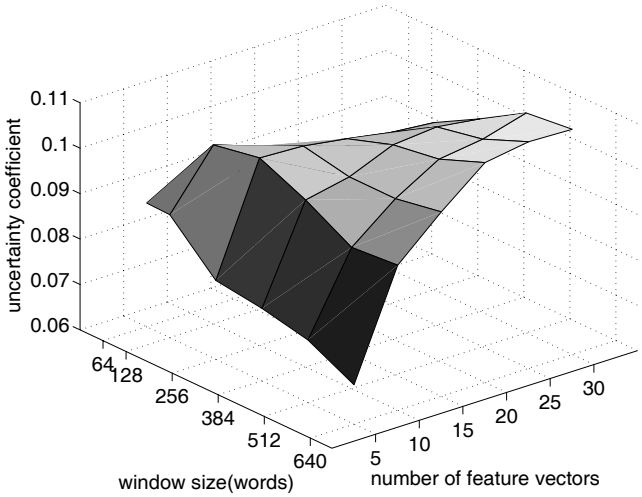

図 9　実験 2 におけるウィンドウ幅・特徴軸数・不確定性係数

5. まとめと今後の課題

本研究ではテキストストリームから与えられるテキストデータに対し、移動ウィンドウ方式を用いて得られるウィンドウ群から特徴軸を抽出する事でトピックを検出する方法を検討した。

特徴軸を抽出する手法としては特異値分解、クラスタリング、独立成分分析の 3 種類を検討し、これらの間で比較実験を行な

うことで各手法の性質を明らかにした。この結果いずれの手法も最適なウィンドウ幅を持つことがわかった。各手法を比較すると、特異値分解はいずれの手法よりも有効とは言えず、クラスタリングはウィンドウ幅が狭い場合に、独立成分分析を用いた手法はウィンドウ幅が広い場合に比較的有効であることがわかった。またウィンドウ幅の変化に対し独立成分分析は比較的堅牢であるが、クラスタリングは変化に敏感であることがわかった。

一方独立成分分析を用いた手法について、検出されたトピックとウィンドウ間の相互情報量から最適なウィンドウ幅とトピック数を決定する手法を提案した。この提案手法について人工データと実データを用いた実験から有効性を確かめた。実データを用いた実験では特徴軸の抽出手法の比較実験と整合する結果が得られた。

今後の課題として、各特徴軸抽出手法や最適ウィンドウ幅とトピック数に関する提案手法の性質をより詳細に分析を進めることが挙げられる。また本研究ではウィンドウの集合に対するトピック検出を行ったが、ウィンドウはテキストストリームからテキストデータが与えられる都度作成される。そこで本研究の内容を元にウィンドウが与えられる度に分析を行う、インクリメンタルなトピック検出手法も今後の課題として挙げられる。

謝辞

本研究の一部は、日本学術振興会日米科学協力事業・共同研究、科学研究費補助金基盤研究 (B)(#15300027)、特定領域研究 (2)(#16016205) による。

文　献

[1] Topic Detection and Tracking (TDT) http://www.nist.gov/speech/tests/tdt/

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic Detection and Tracking Pilot Study Final Report. *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, pp.194-218, 1998.

[3] B. Babcock, M. Datar, and R. Motwani. Sampling From a Moving Window Over Streaming Data. In Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2002), pp. 633–634, 2002.

[4] E. Bingham. Topic Identification in Dynamical Text by Extracting Minimum Complexity Time Components. *3rd International Conference on Independent Component Analysys and Blind Signal Separation*, San Diego, California, pp. 546-551, 2001.

[5] J.-F. Cardoso, and A. Souloumiac. Jacobi Angles for Simultaneous Diagonalization. *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 1, pp. 161-164, 1996.

[6] S. Deerwester, S.T. Dumais, G.W. Furnas, and T.K. Landauer. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[7] J. Fiscus, G. Doddington, J. Garofolo, and A. Martin. NIST's 1998 Topic Detection and Tracking Evaluation (TDT2). *Proc. of the DARPA Broadcast News Workshop*, Hemdon, Virginia, pp. 19–24, 1999.

[8] 濱本雅史, 北川博之, Jia-Yu Pan, and Christos Faloutsos. 独立成分分析を用いたテキストデータからのトピック検出. 電子情報通信学会第 15 回データ工学ワークショップ (DEWS2004), 2004 年 3 月.

[9] M. Hearst and C. Plaunt. Subtopic Structuring for Full-Length Document Access. *Proc. of the 16th Annual International ACM/SIGIR Conference*, Pittsburgh, PA, pp. 59-68, 1993.

[10] A. Hyvärinen, J. Karhunen, and E. Oja. Independent Component Analysis. John Wiley, New York, 2001.

[11] I. T. Jolliffe. Principal Component Analysis. Spriger Verlag, New York, 1986.

[12] A. Kabán, and M. Girolami. Unsupervised Topic Separation and Keyword Identification in Document Collections: A Projection Approach. *Tech. Rep. 10, Dept. of Computing and Information Systems*, Univ. of Paisley, 2000.

[13] T. Kolenda, L. K. Hansen, and J. Larsen. Signal Detection Using ICA: Application to Chat Room Topic Spotting. *3rd International Conference on Independent Component Analysys and Blind Signal Separation*, San Diego, California, pp. 540-545, 2001.

[14] H. Kuhn. The hungarian method for the assignment problem. Naval Research Logistics Quarterly 2, pp. 83–97, 1955.

[15] H. Li, and K. Yamanishi. Topic Analysis Using Finite Mixture Model. *Information Processing and Management*, vol. 39, pp.521-541, 2003.

[16] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. *Proc. 5th Berkeley Symposium in Mathematics, Statistics and Probability*, 1:281-297, 1967.

[17] D. Pelleg and A. Moore. X-means: Extending kmeans with efficient estimation of the number of clusters. *Proc. 17th International Conf. on Machine Learning*, San Francisco, CA, pp.727–734, 2000,

[18] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. ニューメリカルレシピ・イン・シー [日本語版] C 言語による数値計算のレシピ. 技術評論社, 1993 年.

[19] K. Rajaraman and A. Tan. Topic Detection, Tracking and Trend Analysis Using Self-Organizing Neural Networks. In *Proc. 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2001)*, pages 102–107, Hong Kong, 2001.

[20] J. M. Schultz, and M. Liberman. Topic Detection and Tracking Using idf-Weighted Cosine Coefficient. *Proc. DARPA Broadcast News Workshop*, Hemdon, Virginia, pp. 189-192, 1999.

[21] F. Walls, H. Jin, S. Sista, and R. Schwartz. Topic Detection in Broadcast News. *Proc. DARPA Broadcast News Workshop*, Hemdon, Virginia, pp. 193-198, 1999.