(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4254623号
(P4254623)
(45)発行日　平成21年4月15日(2009.4.15)　(24)登録日　平成21年2月6日(2009.2.6)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 220Z
G06F 17/30 170A

請求項の数 9　(全 16 頁)

(21)出願番号　特願2004-170612(P2004-170612)
(22)出願日　平成16年6月9日(2004.6.9)
(65)公開番号　特開2005-352613(P2005-352613A)
(43)公開日　平成17年12月22日(2005.12.22)
審査請求日　平成17年7月20日(2005.7.20)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100088812
弁理士　▲柳▼川　信
(72)発明者　森永　聡
東京都港区芝五丁目７番１号　日本電気株式会社内
(72)発明者　山西　健司
東京都港区芝五丁目７番１号　日本電気株式会社内

審査官　紀田　馨

最終頁に続く

(54)【発明の名称】トピック分析方法及びその装置並びにプログラム

(57)【特許請求の範囲】
【請求項１】
テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出するトピック分析装置であって、
トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習する学習手段と、
前記生成モデルを格納する記憶手段と、
前記記憶手段に格納された複数の候補となるトピックの生成モデルの中で、情報量基準に基づいて最適なトピックの生成モデルを選択して、その混合成分としてトピックを検出する手段と、
特定の時間のトピックの生成モデルの混合成分と、別の時間のトピック生成モデルの混合成分を比較して、新しいトピックの生成と既存のトピックの消滅を判定するトピック形成消滅判定手段と、
を含むことを特徴とするトピック分析装置。
【請求項２】
トピックの生成モデルの各混合成分に対応するトピックの特徴表現を、混合成分のパラメータに基づいて抽出して、各トピックを特徴付けるトピック特徴表現抽出手段を、更に含むことを特徴とする請求項１記載のトピック分析装置。
【請求項３】

テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出するトピック分析装置であって、
トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習する学習手段と、
前記生成モデルを格納する記憶手段と、
前記記憶手段に格納された複数の候補となるトピックの生成モデルの中で情報量基準に基づいて最適なトピックの生成モデルを選択してその混合成分としてトピックを検出する手段と、
トピックの生成モデルの各混合成分に対応するトピックの特徴表現を、混合成分のパラメータに基づいて抽出して、各トピックを特徴付けるトピック特徴抽出手段と、
を含むことを特徴とするトピック分析装置。

【請求項４】
テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出分析するコンピュータによるトピック分析方法であって、
前記コンピュータの学習機能により、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習して、記憶手段に記憶するステップと、
前記コンピュータのモデル選択機能により、前記記憶手段に記憶された複数の候補となるトピックの前記生成モデルの中で、情報量基準に基づいて最適なトピックの生成モデルを選択して、その混合成分としてトピックを検出するステップと、
前記コンピュータのトピック消滅判定機能により、特定の時間のトピックの生成モデルの混合成分と、別の時間のトピック生成モデルの混合成分とを比較して、新しいトピックの生成と既存のトピックの消滅を判定するステップと、
を含むことを特徴とするトピック分析方法。

【請求項５】
前記コンピュータのトピック特徴表現抽出機能により、トピックの生成モデルの各混合成分に対応するトピックの特徴表現を、混合成分のパラメータに基づいて抽出して、各トピックを特徴付けるステップを、更に含むことを特徴とする請求項４記載のトピック分析方法。

【請求項６】
テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出分析するコンピュータによるトピック分析方法であって、
前記コンピュータの学習機能により、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習して、記憶手段に記憶するステップと、
前記コンピュータのモデル選択機能により、前記記憶手段に記憶された複数の候補となるトピックの前記生成モデルの中で情報量基準に基づいて最適なトピックの生成モデルを選択して、その混合成分としてトピックを検出するステップと、
前記コンピュータのトピック特徴表現抽出機能により、トピックの生成モデルの各混合成分に対応するトピックの特徴表現を、混合成分のパラメータに基づいて抽出することにより、各トピックを特徴付けるステップと、
を含むことを特徴とするトピック分析方法。

【請求項７】
テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出する方法をコンピュータに実行させるためのプログラムであって、
前記コンピュータを、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習して、記憶手段に記憶する機能として動作させる処理と、

前記コンピュータを、前記記憶手段に記憶された複数の候補となるトピックの前記生成モデルの中で、情報量基準に基づいて最適なトピックの生成モデルを選択して、その混合成分としてトピックを検出する機能として動作させる処理と、
前記コンピュータを、特定の時間のトピックの生成モデルの混合成分と、別の時間のトピック生成モデルの混合成分とを比較して、新しいトピックの生成と既存のトピックの消滅を判定する機能として動作させる処理と、
を含むことを特徴とする、コンピュータにより読取可能なプログラム。
【請求項８】
前記コンピュータを、トピックの生成モデルの各混合成分に対応するトピックの特徴表現を、混合成分のパラメータに基づいて抽出して、各トピックを特徴付ける機能として動作させる処理を、更に含むことを特徴とする請求項７記載のプログラム。
【請求項９】
テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出する方法をコンピュータに実行させるためのプログラムであって、
前記コンピュータを、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習して、記憶手段に記憶する機能として動作させる処理と、
前記コンピュータを、前記記憶手段に記憶された複数の候補となるトピックの前記生成モデルの中で情報量基準に基づいて最適なトピックの生成モデルを選択して、その混合成分としてトピックを検出する機能として動作させる処理と、
前記コンピュータを、トピックの生成モデルの各混合成分に対応するトピックの特徴表現を、混合成分のパラメータに基づいて抽出することにより、各トピックを特徴付ける機能として動作させる処理と、
を含むことを特徴とする、コンピュータにより読取可能なプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明はトピック分析方法及びその装置並びにプログラムに関し、特にテキストマイニングや自然言語処理の分野において、時系列で追加されていくテキスト集合に対して、各時刻の主要なトピックを同定して各トピックの内容および変化を分析するトピック分析方式に関するものである。
【背景技術】
【０００２】
一括で与えられた時系列のテキストデータに対して、各時刻における主要な表現を抽出する方式としては、例えば、非特許文献１に示された方式が知られている。この方式では、テキストデータに現れる単語の中で、その出現頻度が特定の時間期間で上昇しているものが抽出され、その時間期間の開始時刻が主要トピックの出現時刻、期間の終了時刻がそのトピックの消滅時刻、その単語がトピックの内容を表現するものとされていた。
【０００３】
また、トピックの時系列的変化を可視化する方式としては、非特許文献２に開示の方式が知られている。しかし、上記２つの方式はいずれもデータが逐次的に与えられる語毎にオンラインでリアルタイムに処理することはできなかった。
【０００４】
ある特定の単語を含む文章の時系列の塊を検出する方式としては、非特許文献３に示された方式が知られているが、これは異なる単語を使っていても同一内容のトピックを表すようなトピックの分析には不向きであり、また、リアルタイムに分析できないという問題があった。
【０００５】
有限混合確率モデルを用いてトピックの同定や変化検出を行う方式としては、非特許文

献４に示された方式が知られているが、いずれもデータが逐次的に与えられる語毎にオンラインでリアルタイムに処理することはできなかった。

【０００６】

リアルタイムに有限混合確率モデルを学習する方式については、非特許文献５が知られているが、これはデータの時系列的順序を考慮するが、データの発生時間そのものを反映できないという問題があった。

【０００７】

【非特許文献１】R. Swan, J. Allan, “Automatic generation of overview timelines, ” Proc. SIGIR Intl. Conf. Information Retrieval, 2000. S.Harve, B.Hetzler, and L.Norwell: ThemeRiver: Visualizing theme changes over time,

【非特許文献２】Proceesings of IEEE Symposium on Information Visualization, 2000

【非特許文献３】J.Kleinberg: Bursty and hierarchical structure in streams,Proceedings of KDD2002, pp:91-101, ACM Press, 2003

【非特許文献４】X.Liu, Y.Gong, W.Xu, and S.Zhu: Document clustering with cluster refinement and model selection capabilities, Proceedings of SIGIR International Conference on Information Retrieval, 2002 やH.Li and K.Yamanishi: Topic analysis using a finite mixture model,Information Processing and Management, Vol.39/4, pp 521-541, 2003

【非特許文献５】K.Yamanishi, J.Takeuchi, G.Williams, and P.Milne: On-line unsupervised oultlier detection using finite mixtures with discounting learning algorithms,"in {\em Proceedings of KDD2000}, ACM Press, pp:320--324 2000

【発明の開示】

【発明が解決しようとする課題】

【０００８】

テキストデータが時間を追って追加されていくような状況で、随時、主要なトピックの内容同定をしたい場合には、従来の多くの方式は、多大な記憶容量と処理時間とが必要になるという問題があった。しかしながら、ＣＲＭ（Customer Relationship Management）やナレッジマネジメントおよびＷＥＢ監視などの目的で、時間的に追加されていくテキストデータのトピックを分析する際には、なるべく少ない記憶容量と処理時間でリアルタイムに分析を行う必要がある。

【０００９】

さらに上記の各方式においては、単一のトピックの内容が時間と共に微妙に変化していく場合に、「同じトピックだが内容が微妙に変化している」ことを知ることが出来ない。しかしながら、ＣＲＭやＷＥＢ監視目的のトピック分析などにおいては、「特定の商品に対する苦情内容の変化」の抽出などのように、単一トピックの内容変化を同定することによって得られる知見は大きい。

【００１０】

本発明の目的は、時間的に追加されていくテキストデータに対して、なるべく少ない記憶容量と処理時間で、随時、リアルタイムに主要トピックの個数および生成と消滅を同定すること、および主要トピックの特徴を抽出すること、それによって、単一トピックの内容が変化した場合にも、それを分析者が知ることが出来るようにしたトピック分析方法およびその装置並びにプログラムを提供することである。

【課題を解決するための手段】

【００１１】

本発明によるトピック分析装置は、テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出するトピック分析装置であって、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習する学習手段と、前記生成モデルを格納する記憶手段と、前記記憶手段に格納された複数の候補となるトピックの生成モデルの中で、情報量基準に基づいて最適なトピックの生成モデ

ルを選択して、その混合成分としてトピックを検出する手段と、特定の時間のトピックの生成モデルの混合成分と、別の時間のトピック生成モデルの混合成分を比較して、新しいトピックの生成と既存のトピックの消滅を判定するトピック形成消滅判定手段とを含むことを特徴とする。

【００１２】

本発明による他のトピック分析装置は、テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出するトピック分析装置であって、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習する学習手段と、前記生成モデルを格納する記憶手段と、前記記憶手段に格納された複数の候補となるトピックの生成モデルの中で情報量基準に基づいて最適なトピックの生成モデルを選択してその混合成分としてトピックを検出する手段と、トピックの生成モデルの各混合成分に対応するトピックの特徴表現を、混合成分のパラメータに基づいて抽出して、各トピックを特徴付けるトピック特徴抽出手段とを含むことを特徴とする。

【００１６】

本発明によるトピック分析方法は、テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出分析するコンピュータによるトピック分析方法であって、前記コンピュータの学習機能により、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習して、記憶手段に記憶するステップと、前記コンピュータのモデル選択機能により、前記記憶手段に記憶された複数の候補となるトピックの前記生成モデルの中で、情報量基準に基づいて最適なトピックの生成モデルを選択して、その混合成分としてトピックを検出するステップと、特定の時間のトピックの生成モデルの混合成分と、別の時間のトピック生成モデルの混合成分を比較して、新しいトピックの生成と既存のトピックの消滅を判定するトピック形成消滅判定手段とを含むことを特徴とする。

【００１７】

本発明による他のトピック分析装置は、テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出するトピック分析装置であって、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習する学習手段と、前記生成モデルを格納する記憶手段と、前記記憶手段に格納された複数の候補となるトピックの生成モデルの中で情報量基準に基づいて最適なトピックの生成モデルを選択してその混合成分としてトピックを検出する手段と、トピックの生成モデルの各混合成分に対応するトピックの特徴表現を、混合成分のパラメータに基づいて抽出して、各トピックを特徴付けるトピック特徴抽出手段とを含むことを特徴とする。

【００２１】

本発明によるプログラムは、テキストデータが時間とともに追加されていくような状況のもとで、該データを順次読み込みつつトピックを検出する方法をコンピュータに実行させるためのプログラムであって、前記コンピュータを、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習して、記憶手段に記憶する機能として動作させる処理と、前記コンピュータを、前記記憶手段に記憶された複数の候補となるトピックの前記生成モデルの中で、情報量基準に基づいて最適なトピックの生成モデルを選択して、その混合成分としてトピックを検出する機能として動作させる処理と、前記コンピュータを、特定の時間のトピックの生成モデルの混合成分と、別の時間のトピック生成モデルの混合成分とを比較して、新しいトピックの生成と既存のトピックの消滅を判定する機能として動作させる処理とを含むことを特徴とする。

【００２２】

本発明による他のプログラムは、テキストデータが時間とともに追加されていくような

状況のもとで、該データを順次読み込みつつトピックを検出する方法をコンピュータに実行させるためのプログラムであって、前記コンピュータを、トピックの生成モデルを混合分布モデルで表現し、データのタイムスタンプに応じて過去のデータほど激しく忘却しながら該トピックの生成モデルをオンラインで学習して、記憶手段に記憶する機能として動作させる処理と、前記コンピュータを、前記記憶手段に記憶された複数の候補となるトピックの前記生成モデルの中で情報量基準に基づいて最適なトピックの生成モデルを選択して、その混合成分としてトピックを検出する機能として動作させる処理と、前記コンピュータを、トピックの生成モデルの各混合成分に対応するトピックの特徴表現を、混合成分のパラメータに基づいて抽出することにより、各トピックを特徴付ける機能として動作させる処理とを含むことを特徴とする。

【００２６】

本発明の作用を述べる。本発明では、各テキストを文書ベクトルとして表現し、その生成モデルとして混合分布モデルを用いる。混合分布の一つのコンポーネントが一つのトピックに対応するとする。混合分布モデルはコンポーネントの個数等が異なる複数のものが保持される。新規テキストデータが追加されるたびに、学習手段によって各モデルのパラメータが追加学習され、モデル選択手段によって情報量基準に基づいて最も適切なモデルが選択される。選択されたモデルの各コンポーネントが主要なトピックを表している。また、モデル選択手段によってどのモデルが選択されるかが変化した場合には、トピック形成消滅判定手段により以前に選択されていたモデルと今回選択されたモデルの比較が行われ、どれが新たに形成されたトピックであるか、どのトピックが消滅したのかが判定される。

【００２７】

さらに本発明では、モデル選択手段によって選択されたモデルの各トピック、トピック生成／消滅判定手段によって判定された新たに生成されたトピック／消滅したトピックに関して，トピック特徴表現抽出手段が該当する混合分布のパラメータから、そのトピックの特徴表現を抽出し出力される。

【００２８】

複数の混合分布モデルを全て独立に学習し選択するのではなく、一つもしくは複数の上位モデルを学習し、学習された上位モデルからサブモデル生成手段によって複数のサブモデルを生成し、モデル選択手段によって、その中から適切なモデルを選択するのでも良い。さらに、サブモデルを独立に生成して保持するのではなく、サブモデル生成選択手段によって、特定のサブモデルの情報量基準を上位モデルから直接に計算し、最も適切なサブモデルを選択するのでもよい。

【００２９】

学習手段による各モデルのパラメータの追加学習においては、到着順の早いテキストデータに比べて到着順が後ろのテキストデータの内容を重視するようにしてもよい。さらに、テキストデータにタイムスタンプが付随している場合には、到着順のみならずタイムスタンプの内容を利用して、古いテキストデータに比べて最近のテキストデータほど内容を重視するようにしてもよい。

【００３０】

モデル選択手段およびサブモデル生成選択手段において適切なモデルを選択する際に、新たに入力されたテキストデータを用いて追加学習した前後の分布間の距離や、追加学習する前の分布において前記入力テキストデータが発生するのはどれくらい稀なのか、を各モデルに対して計算し、それを最小にするモデルを選択するのでもよい。さらに、これらを計算した結果をモデルの次元数で割ったものや、特定の時刻からの値の累積値、最近の値を重視するように重み付けした平均値などを計算するのでも良い。

【００３１】

トピック形成／消滅判定手段において、以前に選択されていたモデル（旧モデル）と今回選択されたモデル（新モデル）を比較する際に、旧モデルに含まれるコンポーネントと新モデルに含まれるコンポーネントの全ての組み合わせのペアについて類似度を計算し、

どの旧モデルのコンポーネントとも類似度が低い新モデルのコンポーネントを新たに形成されたトピックと判定、どの新モデルのコンポーネントとも類似度が低い旧モデルのコンポーネントを消滅したトピックと判定してもよい。コンポーネント間の類似度は、平均値間の距離や、分布の同一性検定におけるＰ値を用いてもよい。モデルが上位モデルから生成されたサブモデルである場合は、コンポーネント間の類似度として上位モデルにおける同一のコンポーネントから生成されているかどうかを用いてもよい。

【００３２】

トピック特徴表現抽出手段においては、各トピックを表すコンポーネントの確率分布に従ってテキストデータを発生させ、テキストデータを入力とする公知の特徴抽出技術を用いて各トピックの特徴表現を抽出してもよい。前記公知の特徴抽出技術で必要となるテキストデータの各種統計量が、コンポーネントのパラメータから計算できる場合は、その値を使って特徴抽出してもよい。サブ分布生成手段においては、上位モデルの幾つかのコンポーネントをコンポーネントとする混合分布をサブ分布としてもよい。

【発明の効果】

【００３３】

本発明の第一の効果は、時系列のテキストデータを複数の混合分布でモデル化し、忘却型逐次学習アルゴリズムによるパラメータ学習とモデル選択によって、主要トピックおよびその生成／消滅を、少ない記憶容量と処理時間で随時同定できることができるということである。この際、データのタイムスタンプを利用して、古いものほど、その効果を失いながらトピック構造を同定することができる。また、テキストデータが追加されるごとに新しい単語が出現して、その表現ベクトルの次元が上がっても、これに対応して、最適な主要トピックを同定することができる。

【００３４】

また、本発明の第二の効果は、学習された混合分布のパラメータから各トピックの特徴表現を同定することによって、トピックの内容を随時抽出できること、それによって、単一トピックの内容が変化した場合にも、それを分析者が知ることができるということである。

【発明を実施するための最良の形態】

【００３５】

以下に、図面を参照して本発明の実施の形態について詳細に説明する。図１は本発明の第一の実施の形態にかかるトピック分析装置の構成を示すブロック図である。本トピック分析装置は、全体としてコンピュータからなり、テキストデータ入力手段１、学習手段２１，……，２ｎ、混合分布モデル（モデル記憶手段）３１，……，３ｎ、モデル選択手段４、トピック形成／消滅判定手段５、トピック特徴表現抽出手段６、出力手段８を含んでいる。

【００３６】

テキストデータ入力手段１は、コールセンターのコンタクト内容や、Ｗｅｂから収集した監視対象ページの内容、新聞記事の内容などテキスト（文字情報）を入力する手段であり、対象とするデータを一括して入力するだけでなく、データが発生したり収集されたりする毎に、追加的に入力することも可能となっているものである。また、入力されたテキストは公知の形態素解析技術や構文解析技術によって分解され、さらに公知の属性選択技術や重み付け技術を用いることで、後記モデル３１～３ｎが対象とするデータ形式に変換される。

【００３７】

例えば、全ての単語のうち、名詞だけを取り出し、それらをw1，…，wN として、それらのテキスト中の頻度をｔｆ（w1），…，ｔｆ（wN ）としてベクトル（ｔｆ（w1），…，ｔｆ（wN ））をテキストデータの表現としたり、全体のテキストの数をM、単語wi を含むテキストの数をｄｆ（wi ）として、ｔｆ－ｉｄｆ値である、

ｔｆ－ｉｄｆ（wi ）＝ｔｆ（wi ）×ｌｏｇ（M／ｄｆ（wi ））

を各要素とするベクトル、

（ｔｆ－ｉｄｆ（ｗｉ），…，ｔｆ－ｉｄｆ（ｗＮ））
をテキストデータの表現としたりする。これらを構成する際に、予め頻度がしきい値に達しないものは最初から要素に入れないなどの前処理を行うこともあり得る。

【００３８】

本テキストデータ入力手段１は、テキストデータを操作入力するためのキーボードや、コールセンターデータベースの内容を逐次転送するプログラム、Ｗｅｂ上のテキストデータをダウンロードするアプリケーションなどの一般的な情報入力手段により構成される。

【００３９】

学習手段２１～２ｎは、テキストデータ入力手段１によって入力されたテキストデータに基づき、混合分布３１～３ｎを更新する手段である。混合分布３１～３ｎは、入力されるテキストデータの従う確率分布の候補として、テキストデータ入力手段１によって入力されたテキストデータに基づき推定されたものである。

【００４０】

一般に、確率モデルの考え方では、与えられたデータｘは、ある確率変数の実現値とみなされる。特に、この確率変数の確率密度関数が有限次元のパラメータａを持つ固定された関数形ｆ（ｘ；ａ）を持つと仮定すると、その確率密度関数族、

Ｆ＝｛ｆ（ｘ；ａ）｜ａ　ｉｎ　Ａ｝

をパラメトリック確率モデルという。なお、Ａはａのとり得る値の集合である。また、データｘに基づきパラメータａの値を推測することを推定という。例えば、ｌｏｇｆ（ｘ；ａ）をａの関数（対数尤度関数）とみなし、これを最大にするａを推定値とする最尤推定法などが一般的である。

【００４１】

また、複数の確率モデルの線形結合、

Ｍ＝｛ｆ（ｘ；Ｃ１，…，Ｃｎ，ａ１，…，ａｎ）
＝Ｃ１＊ｆ１（ｘ；ａ１）＋…＋Ｃｎ＊ｆｎ（ｘ；ａｎ）｜ａｉ　ｉｎ　Ａｉ，
Ｃ１＋…＋Ｃｎ＝１、　　Ｃｉ＞０　（ｉ＝１，…，ｋ）｝

によって与えられる確率モデルＭを混合モデル、その確率分布を混合分布、線形結合の対象となった元の各分布をコンポーネント、Ｃｉをｉ番目のコンポーネントの混合比率とよぶ。これは、ｙを１からｎまでの整数を値域とする確率変数とし、

Ｐｒ｛ｙ＝ｉ｝＝Ｃｉ，ｆ（ｘ｜ｙ＝ｉ）＝ｆｉ（ｘ；ａｉ）

を満たす確率変数ｚ＝（ｙ，ｘ）に対して、ｙを隠れ変数としてｘのみをモデル化したものと同じである。

【００４２】

ただし、ｙ＝ｉという条件の下でのｘの条件付密度関数をｆ（ｘ｜ｙ＝ｉ）としている。また、後の記述の簡単化のために、ｚ＝（ｙ，ｘ）の確率密度関数を、

ｇ（ｚ；Ｃ１，……，Ｃｎ，ａ１，……，ａｎ）

とする。

【００４３】

本発明においては、モデル３１～３ｎは、コンポーネント数やコンポーネントのパラメータが異なる混合モデルであるとし、各コンポーネントは特定の主要なトピックについて述べているテキストデータの従う確率分布であるとする。すなわち、与えられたモデルにおいて、コンポーネントの個数はテキストデータ集合の中の主要トピックの数を表し、各コンポーネントが各主要トピックに相当することになる。

【００４４】

混合モデルに対して、与えられたデータに基づいて最尤推定を行うことは非常に大きな計算量を必要とするが、計算量を節約して近似解を求める方法として、ＥＭ（Expectation Maximization）アルゴリズムがよく知られている。このＥＭアルゴリズムにおいては、対数尤度を直接に最大化するのではなく、隠れ変数ｙの値の事後分布の計算と、ｙの値で条件付けしたｘの対数尤度の前記事後分布による平均値Ｅｙ［ｌｏｇ　ｇ（ｘ｜ｙ）］の最大化を繰り返すことで、混合分布の各パラメータの推定が行われる。ただし、ｙの前記

事後分布による平均値をＥｙ［＊］としている。

【００４５】

さらに、データが一括で与えられるのではなく、逐次的に追加到着する状況で、混合分布のパラメータ推定結果をデータ追加時に更新していく逐次型のＥＭアルゴリズムも公知となっている。特に、非特許文献５では、データの到着順序が考慮され、最近到着したものが重要視され、昔に到着したデータの影響は徐々に軽くなっていく手法が記されている。これは、到着したデータの総数をＬ個としｌ番目のデータをｘｌ、そのときの隠れ変数をｙｌとした場合に、ｙｌの事後分布の計算と、最近到着したものの重みを大きくした対数尤度、

$$\Sigma E_{yl}[(1-r)^{(L-l)}\ r\log\ g(yl|xl)]$$

の最大化を逐次的に行うものである。

【００４６】

ただし、Σはｌ＝１～Ｌの和を表すとし、Ｅｙｌ［＊］はｙｌの事後分布による平均とする。上記の特別な場合としてｒ＝０としたものが、データの到着順序による重み付けをしない逐次型のＥＭアルゴリズムである。

【００４７】

本発明の学習手段２１～２ｎは前記の逐次型ＥＭアルゴリズムによって、テキストデータ入力手段１からデータが与えられるたびに、モデル３１～３ｎにおける混合分布の推定結果を更新する。さらに、テキストデータにタイムスタンプが付随している場合は、

$$\Sigma E_{yl}[(1-r)^{(L-tl)}r\log\ g(xl,yl|yl)]$$

を最大化するように逐次学習をおこなってもよい。ただし、ｌ番目のデータのタイムスタンプをｔｌとしている。こうすることによって、データの到着間隔が不ぞろいである場合にも、時間的に最近のデータを重要視し古いデータの影響を軽くするようにコンシスタントに推定が行われる。

【００４８】

例えば、混合モデルとして、各コンポーネントがガウス分布であるような場合を考えると、ｉ番目のコンポーネントは平均μ_i，分散共分散行列Σ_iをパラメータとするガウス密度関数として、

$$(1/(2\pi)^{d/2}|\Sigma_i|)\exp[-(1/2)(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)]$$

で表される。コンポーネントの数はｋ個あるとし、ｉ番目のコンポーネントの混合比率をξ_iとする。

【００４９】

この場合、t_{old}時刻のデータをx_nとし、t_{new}時刻に新しいデータx_{n+i}を入力としたとき、ｉ番目のコンポーネントの更新前の平均パラメータ、分散協分散行列パラメータ、混合比率をそれぞれμ_i^{old}，Σ_i^{old}，ξ_i^{old}とし、更新後のそれらをμ_i^{new}，Σ_i^{new}，ξ_i^{new}は、例えば以下のように計算することができる。ここで、ｄ、W_{ij}，S_iは助変数である。

【００５０】

【数１】

$$P=\frac{1}{\sum_{l=1}^{k}\exp\left\{\log\xi_l^{old}+\log\phi\left(x_{n+1}\middle|\mu_l^{old},\Sigma_l^{old}\right)-\log\xi_i^{old}-\log\phi\left(x_{n+1}\middle|\mu_i^{old},\Sigma_i^{old}\right)\right\}}$$

【数２】

$$W_{in+1}^{new} = WA\left(P, \frac{1}{k} \middle| 1, \alpha\right)$$

ここに、αはユーザ指定の定数である。

【００５１】

【数３】

$$\mu_i^{new} = WA\left(\mu_i^{old}, x_{n+1} \middle| \xi_i^{old} d^{old}, \lambda^{-(t_{new}-t_{old})} W_{in+1}^{new}\right)$$

ここに、λはユーザ指定の定数（忘却率）である。

【数４】

$$S_i^{new} = WA\left(S_i^{old}, x_{n+1}x_{n+1} \middle| \xi_i^{old} d^{old}, \lambda^{-(t_{new}-t_{old})} W_{in+1}^{new}\right)$$

【数５】

$$\Sigma_i^{new} = S_i^{new} - \mu_i^{new}\mu_i^{new}$$

【数６】

$$\xi_i^{new} = WA\left(\xi_i^{old}, W_{in+1}^{new} \middle| d^{old}, \lambda^{-(t_{new}-t_{old})}\right)$$

【数７】

$$d^{new} = \lambda^{-(t_{new}-t_{old})} d^{old} + 1$$

【００５２】

ただし、上記では、表記の簡単化のために、

（式１＊式３＋式２＊式４）／（式３＋式４）

と書くところを、

ＷＡ（式１，式２｜式３，式４）

として表している。

【００５３】

モデル選択手段４では、入力されるテキストデータの従う確率分布の候補であるモデル３１～３ｎのそれぞれに対し、テキストデータ入力手段１によって入力されたテキストに基づいて情報量基準の値が計算され、最も適切なモデルが選択される。例えば、Ｗをウイ

ンドウの大きさとし，ｔ番目のデータのベクトル表現の次元をｄｔとし、Ｐ（^{t}）（ｘ｜ｋ）をｋ個のコンポーネントからなる混合分布で、ｔ番目のデータが入力されてから逐次的にパラメータを更新したものであるとするとき、ｎ番目のデータを受け取ったときの情報量基準の値Ｉ（ｋ）は、例えば、以下のように計算できる。

$$I(k)=(1/W)\Sigma_{t=n-W}^{n}(-\log P^{(t)}(x_t|k))/d_t$$

【００５４】

この値を最小化するようなコンポーネント個数ｋが最適なコンポーネント数であり、それを構成するコンポーネントが主要トピックを表すコンポーネントであると同定することができる。この基準の値は、入力テキストデータが追加されるごとに新しい単語が出現して、その表現ベクトルの次元が上がっても、これに対応して計算できるものである。$P^{(t)}(x_t|k)$を構成するコンポーネントは、独立なコンポーネントであっても、上位の混合モデルのサブコンポーネントであってもよいものとする。

【００５５】

トピック形成／消滅判定手段５では、モデル選択手段４によって選択されモデルが変化した場合、新たに選択されたモデルのコンポーネントの中で、以前に選択されていたモデルには近いコンポーネントが存在しないものを、「新たに形成されたトピック」、逆に新しいモデルにおいて近いコンポーネントが存在しない古いモデルのコンポーネントを、「消滅したトピック」と判定し、出力手段７に出力する。コンポーネント間の近さの尺度としては、分布の同一性検定におけるＰ値や、二つの確率分布の近さを計る量として公知のＫＬ（Kullback Leibler）ダイバージェンス等を用いればよい。あるいは、さらに簡単に二つの確率分布の平均値の差などを用いても良い。

【００５６】

トピック特徴抽出手段６は、モデル選択手段４によって選択されたモデルに対して、各コンポーネントの特徴を抽出し、該当トピックの特徴表現として出力手段７に出力する。特徴表現を抽出するのには、単語の情報利得を計算して、その大きいものを抽出する方法を用いることができる。情報利得は、例えば、以下のように計算する。

【００５７】

ｔ番目のデータが与えられたときに、全体のデータの数をｔとして、全データの中で指定された単語ｗを含むデータの数をm_w、これを含まないデータの数をm'_w、ある指定したコンポーネント（かりにｉ番目とする）から発生したテキストの数をt_i　単語ｗを含むデータの中でｉ番目のコンポーネントから発生したデータ数をm_w^+、単語ｗを含まないデータの中でｉ番目のコンポーネントから発生したデータ数をm'^{+}_{w}とするとき、Ｉ（Ａ，Ｂ）を情報量尺度として、ｗの情報利得を、

$$IG(w)=I(t,t_i)-(I(m_w,m_w^+)+I(m'_w,m'^{+}_{w}))$$

のように計算する。

【００５８】

ここで、Ｉ（Ａ，Ｂ）の計算式としては、エントロピー、確率的コンプレキシティ、拡張型確率的コンプレキシティなどを用いることができる。エントロピーは、

$$I(A,B)=AH(B/A)=A(B\log(B/A)+(A-B)\log((A-B)/A))$$

で表されるものであり、確率的コンプレキシティは、

$$I(A,B)=AH(B/A)+(1/2)\log(A/2\pi)$$

で表されるものであり、拡張型確率的コンプレキシティは、

$$I(A,B)=\min\{B,A-B\}+c(A\log A)^{1/2}$$

で表されるものである。

【００５９】

また、ＩＧ（ｗ）の代わりに情報利得としてχ自乗検定量、

$$(m_w+m'_w)\times(m_w^+(m'_w-m'^{+}_{w})-(m_w-m_w^+)m'_w)\times((m_w^++m'^{+}_{w})\times(m_w-m_w^++m'_w-m'^{+}_{w})m_w\ m'_w$$

$)^{-1}$
を用いることもできる。
【００６０】
各ｉについて、ｉ番目のコンポーネントに対し、各ｗについて、上記情報利得を計算し、大きい順に指定された数の言葉を抽出することにより、特徴語を抽出することができる。また、しきい値を予め与えて、そのしきい値以上の情報利得を与える言葉を抽出することにより、特徴語を抽出することができる。上記情報利得を計算するのに必要な統計量は、ｔ番目のデータが与えられたときには、各ｉとｗに対し、ｔ，t_i ，m_w ，m'_w ，m_w^+ ，m'^+_w であるが、これらはデータが与えられる毎にインクリメンタルに計算できる。
【００６１】
本学習手段およびモデルは、ＣＰＵなどのマイクロプロセッサおよびその周辺回路と、モデル３１～３ｎを記憶している記憶装置、およびこれらの動作を統括するプログラムとが協働することにより構成されている。
【００６２】
図２は本発明の動作を示すフローチャートである。まず、ステップ１０１では、テキストデータ入力手段によってテキストデータが入力され、以降のステップでの処理の対象とするデータ形式に変換される。続いて、ステップ１０２では、前記変換されたテキストデータに基づき、学習手段によってモデルのパラメータ推定の更新を行う。これによって、各モデルにおいては今回入力されたデータの値を反映した新しいパラメータの値を保持することになる。
【００６３】
次に、ステップ１０３においては、保持されている複数のモデルの中から、これまでに入力されたテキストデータを鑑みて最も適切なモデルがモデル選択手段により選択される。選択されたモデルにおける混合分布の各コンポーネントが主要なトピックに対応している。
【００６４】
ステップ１０４においては、どのモデルが選択されたかが今回のデータ入力の結果、前回のそれと変化したかどうかが判定される。今回と前回で選択されるモデルが変わらなかった場合は、前回までのテキストデータにおける主要トピックに対して、今回のデータを入力することで新たに主要トピックの形成や消滅がおきなかったことを意味する。逆に、選択されるモデルが変化した場合は、一般に混合分布を構成するコンポーネントの数が変化しており、何らかの新規トピックの形成もしくは消滅が起きていることを意味する。
【００６５】
そこで、ステップ１０５においては、今回選択されたモデルのコンポーネントの中で、前回選択されていたモデルのコンポーネントのどれとも近いものがないものをトピック形成／消滅判定手段により同定し、新規に形成された主要トピックを表すコンポーネントであるとする。同様に、ステップ１０６においては、前回選択されていたモデルのコンポーネントの中で、今回選択されたモデルのコンポーネントのどれとも近いものがないものを同定し、主要でなくなったトピックを表すコンポーネントであるとする。
【００６６】
ステップ１０７では、今回選択されたモデルの各コンポーネントおよび新規形成/ 消滅したとされたコンポーネントの特徴がトピック特徴抽出手段により抽出され、該当するトピックの特徴表現とされる。新たにテキストデータが入力された場合は、ステップ１０１に戻り、一連の処理がなされる。また、ステップ１０３から１０７の処理は、入力される各テキストデータに対して毎回行う必要は無く、主要トピックの同定や新規形成／消滅トピックの同定を行うように、ユーザーなどから指示された場合やタイマーなどで指定された時刻にだけ行うようにしてもよい。
【００６７】
図３は本発明の第二の実施形態にかかるトピック分析装置の構成を示すブロック図であ

り、図１と同等部分は同一符号により示している。第一の実施形態との違いは、モデル選択手段でモデル選択する際の候補となるモデルが、上位モデルの複数のサブモデルである場合になっていることである。サブモデル生成手段９によって生成されたサブモデルに対して、第一の実施の形態と同様のモデル選択を行う。例えば、上位モデルとしては比較的多数のコンポーネントをもつ混合モデルを想定し、サブモデルとしてはそのコンポーネントを幾つか取り出して混合モデルを作った場合が相当する。

【００６８】

このような構成にすることで、並列に複数のモデルを保持する必要と、それぞれを学習手段によって更新する必要が無くなり、処理に必要な記憶容量や計算量を縮減することができる。また、トピック形成／消滅判定手段においても、二つのコンポーネントの間の近さの尺度として、「上位モデルで同一コンポーネントから生成されたかどうか」を採用することにより、確率分布間の距離等を尺度とする場合に比べて必要な計算量を縮減することが出来る。

【００６９】

図４は本発明の第三の実施形態にかかるトピック分析装置の構成を示すブロック図であり、図１と同等部分は同一符号にて示している。ここでも、モデル選択手段でモデル選択する際の候補となるモデルが、上位モデルの複数のサブモデルとして与えられるが、第二の実施形態との違いは、複数のサブモデルを並列に計算するのでなく、サブモデル生成選択手段４１によって、順番に情報量基準を計算し、最も適切なサブモデルを選択することにある。このような構成にすることで、サブモデル全てを保持しておく必要も無くなり、必要な記憶容量をさらに縮減することができる。

【００７０】

図５に本発明への入力データの例を示す。特定のタイプの電気製品に関して議論を行うＷＥＢ上の掲示板に対する監視データで、掲示板への書き込みが行われた日付時刻を付加された書き込み内容（テキストデータ）が１レコードを構成している。ＷＥＢ掲示板自体は投稿が随時行われるので、時間的にデータが随時追加されていくことになる。スケジュールに従って動くプログラムあるいは掲示板サーバー自体等により、新規に追加されたデータが本発明のトピック分析装置に入力され、各処理が行われるとする。

【００７１】

図６は、ある特定の時刻までデータが入力された場合の、本発明によるトピック分析の出力例である。各列が各主要トピックに相当し、モデル選択手段によって選択されたモデルにおける各コンポーネントに対して、トピック特徴表現抽出手段の出力を記載したものである。この分析例では、選択されたモデルには二つのコンポーネントがあり、一つ目のコンポーネントは、「商品ＸＸ」、「遅い」、「メール」などを特徴表現とする主要トピック、二つ目のコンポーネントは、「音」、「ＺＺ」、「よい」などを特徴表現とする主要トピックとなっている。

【００７２】

図７は、さらに特定の時刻までデータ入力が進んだ場合の、本発明によるトピック分析の出力例である。ただし、本出力例はこの時刻でモデル選択手段によってどのモデルが選択されたかが変化した場合を記載している。本出力例で、トピック形成／消滅判定手段により新規形成と判定されたトピックには「主要トピック：新規」、消滅したと判定されたトピックには「消滅トピック」、新しく選択されたモデルのコンポーネントで、以前のモデルのコンポーネントに近いものが存在するトピックには「主要トピック：継続」と列名がついている。

【００７３】

「商品ＸＸ」を特徴語とするトピックは、「主要トピック：継続」の列名を持つので、以前から主要であったトピックである。しかしながら、図６の「商品ＸＸ」のトピックと比較すると、「メール」の代わりに「ウイルス」が特徴語となっており、同じトピックでも内容が変化してきていることを分析者が見て取ることが可能となっている。

【００７４】

「音」や「ＺＺ」を特徴語としていたトピックは図６では主要トピックであったが、図７では「消滅トピック」として出力されている。図７の分析を行った時点で、このトピックが消滅したことが見て取れる。逆に、「新ＷＷ」などを特徴表現とするトピックは「主要トピック：新規」と同定されており、この時点であらたに主要トピックとなったことを分析者が見て取ることが出来る。

【図面の簡単な説明】

【００７５】

【図１】本発明の第一の実施形態に係るトピック分析装置の構成を表すブロック図である。

【図２】本発明の第一の実施形態に係るトピック分析装置の動作を示すフロー図である。

【図３】本発明の第二の実施形態に係るトピック分析装置の構成を表すブロック図である。

【図４】本発明の第三の実施形態に係るトピック分析装置の構成を表すブロック図である。

【図５】本発明への入力データ例である。

【図６】本発明の分析結果出力例（その１）である。

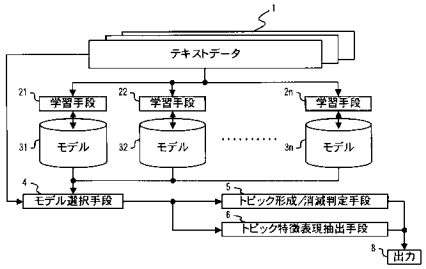
【図７】本発明の分析結果出力例（その２）である。

【符号の説明】

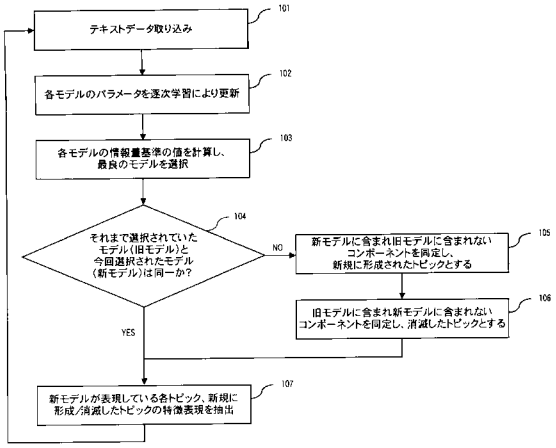
【００７６】

１　テキストデータ入力手段
２１～２ｎ　学習手段
３１～３ｎ　モデル（または上位モデル、サブモデル）
４　モデル選択手段
５　トピック形成／消滅判定手段
６　トピック特徴表現抽出手段
８　出力手段
９　サブモデル生成手段
４１　サブモデル生成選択手段

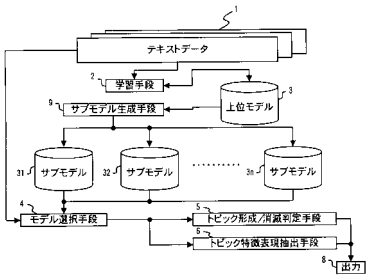
【図１】



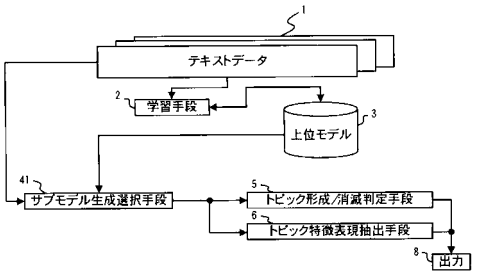
【図２】



【図３】



【図４】



【図５】

日付時刻	テキスト
○月○日○時○分	XXはかえって遅くなってしまった。
○月○日○時○分	テキストだけならモノクロのXXもよい。
○月○日○時○分	商品YYに役に立つようにがんばってみます。
○月○日○時○分	商品ZZ音がいいですよ。
○月○日○時○分	○○さんが惚れ込んでいる商品YYあたりいかがでしょうか。
○月○日○時○分	今度のYYは画面の色がいいですよね。
○月○日○時○分	入力はYYがいいのは分かっていますが高すぎます。
○月○日○時○分	XXは通好みすぎますね。
○月○日○時○分	XXが遅いのはどうしようもない。
○月○日○時○分	本当だ。XXは遅すぎで使えない。
○月○日○時○分	本当にメールだけならXXだけど、他は死ぬほど遅いよ。
○月○日○時○分	だれかZZの話もしてくれ。
……	……

【図６】

	主要トピック:継続	主要トピック:継続
特徴表現	商品XX	音
	遅い	ZZ
	メール	よい
	……	……

【図７】

	主要トピック:継続	消滅トピック	主要トピック:新規
特徴表現	商品XX	音	新WW
	ウィルス	ZZ	軽い
	遅い	よい	キーボード
	……	……	……

フロントページの続き

(56)参考文献 特開２００４－０５４３７０（ＪＰ，Ａ）
特開２００３－３３０９２２（ＪＰ，Ａ）
特開２００１－１０１１５４（ＪＰ，Ａ）
石川 佳治，忘却の概念に基づくインクリメンタルな文書クラスタリング手法，情報処理学会研究報告 ，日本，社団法人情報処理学会，２００１年 7月１８日，Ｖｏｌ．２００１ Ｎｏ．７０，第３１３頁乃至第３２０頁
H.Li, Topic Analysis Using a Finite Mixture Model, Information Processing and Management, ，２００３年，Vol.39/4, pp.521-543，以下のWebページより入手 http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VC8-4607XFJ-1&_user=1583244&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000053883&_version=1&_urlVersion=0&_userid=1583244&md5=dd4785d37b652bdb96b2c5a56f28ad1f
K.Yamanishi, On-line Unsupervised Oultlier Detection Using Finite Mixtures with Discounting Learning Algorithms, in Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD2000, ACM Press, ２０００年, pp.320-324 , 著者Webページより入手（http://www.nec.co.jp/rd/DTmining/members/yamanishi/index_e.html）
山西 健司，情報論的学習理論に基づくマイニング技術－外れ値検出とテキストマイニングを例に－，第４４回 人工知能基礎論研究会資料 ，日本，社団法人人工知能学会，２００１年 3月 8日，（ＳＩＧ－ＦＡＩ－Ａ００４），第３５頁乃至第４０頁

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩ）