ニュースデータベースにおける新規事項の発見とその事後追跡

お茶の水女子大学 山下由美, 日本IBM(株) 東京基礎研究所 小林メイ 青野雅樹
E-mail: yumiy@imv.is.ocha.ac.jp, mei@jp.ibm.com, aono@jp.ibm.com

ニュースデータベースにおける新規事項の発見とその事後追跡のための新しい手法を用いたシステムを実装し、実験をおこなった。本手法はベクトルモデルに基づく手法であり、各ドキュメントは属性としてキーワードと発生時間（タイムスタンプ）をもつベクトルで表現される。従来の事後追跡法では、まずキーワードに基づくクラスタ化をおこない、その後時間に基づくクラスタ化をおこなうという2つのステップを要し、莫大な労力と計算コストがかかる。本手法ではキーワードとタイムスタンプの両方を同時に処理することで、即時的、かつ自動的な処理を可能とする。本稿では、本手法を用いたプロトタイプシステムによる実験結果を報告する。

Detecting and Tracking the Evolution of Events

Yumi Yamashita †, Mei Kobayashi ‡, Masaki Aono ‡
Ochanomizu Univ.†, IBM Research, Tokyo Research Laboratory ‡
E-mail: yumiy@imv.is.ocha.ac.jp, mei@jp.ibm.com, aono@jp.ibm.com

We present implementation studies of and a novel graphical user interface for a method to detect new events (or new classes) of documents in a very large dynamically changing news database and to track their evolution. In our prototype system, documents are modeled as vectors, the entries of which contain information on attributes, i.e., keywords and date stamps. The time stamp is treated as a keyword attribute, with two user specified parameters: (1) a weighted window around the date stamp for each document (or the anniversaries of the date stamp) to give importance to dates related to the document, and (2) the relative importance of the temporal information versus the keywords. Our prototype is novel in that it processes the temporal information and keyword information simultaneously. Other detection and tracking systems process documents in two steps: documents are clustered according to subjects/keywords, prior to processing the date stamp information.

1 はじめに

近年、膨大な電子文書がデータベースに蓄えられている。例えば、ニュースデータや顧客情報、在庫管理など、様々な活動において大規模なデータが扱われており、これらの情報を有効に活用することが必要不可欠になってきている。これらの情報の検索効率や精度を高めることは、様々な分野におけるユーザに対し、価値ある情報を与えることになる。データ検索効率を向上するために、これまでにも数多くの研究が進められてきた[3],[5],[6],[7]。検索効率を高める方法のひとつとして、データベース内の膨大なデータの中から新しい事項を発見し、そのフォローアップ事項を追跡してクラスタリングする方法が知られている[1]。

実際に検索エンジンとして用いられている検出及び事後追跡手法では、データに対してベクトルモデルを適用し[4]、クラスタ化している。従来法ではこのベクトルモデルにおいて、まずキーワードに基づくクラスタ化をおこなう。その後これらのクラスタに対し、データに付された日付／時間スタンプを用いた、新規事項の発見および事後追跡のためのクラスタ化をおこなう[2]。クラスタ化の手法としては主にベクトルの内積による分類が用いられており、各データに対応するベクトルを内積が所定の範囲となるいくつかのクラスタへ

ニュースデータベースにおける新規事項の発見とその事後追跡
山下 由美, 小林 メイ†, 青野 雅樹†
お茶の水女子大学 大学院 人間文化研究科
†日本IBM 東京基礎研究所
Detecting and Tracking the Evolution of Events
Yumi Yamashita, Mei Kobayashi†, and Masaki Aono†
Graduate School of Humanities and Sciences,
Ochanonizu University
†IBM Research, Tokyo Research Laboratory

分類する。従来の検索手法では、このように2段階のクラスタ化プロセスを用いているため、コストの高い作業となっている。

そこで、新規事項の発見およびその事後追跡を、自動的にリアルタイム処理できるシステムを提供することが必要とされている。本稿では、ユーザからの要求を受けてリアルタイムに実行処理するシステムのプロトタイプを実装し、検索／追跡処理の実験結果を報告する。

次章では、新しい手法におけるアルゴリズムを説明する。第3章では、本稿で提供するシステムの概要について述べる。第4章で、本システムによる処理の実行速度の計測結果と、新規な事項の発見および事後追跡結果を報告する。そして最後にまとめと今後の課題について述べる。

2 アルゴリズム

本稿で用いる手法はベクトル空間モデルに基づく手法である。ベクトル空間モデルでは文書および検索質問(クエリー)をベクトルとして表現し、各ベクトルの距離としてドキュメントの類似度を定義する。ベクトルの各要素には索引語（キーワード）および、ドキュメントの発生時間（タイムスタンプ）に対する重みが入る。キーワードの重みおよび文書ベクトル間の類似度としてはいくつか提案されているが、本稿では重みとしてTF-IDF [8] に基づいた正の値、類似度として余弦を使用する。TF-IDFは、単語頻度と文書頻度の逆数の積で単語の重みを決定する方法である。余弦は、ベクトル間の内積を正規化することでもとめられる。ベクトルモデルでは計算コストを削減するために次元削減処理をおこなう。本稿では次元削減アルゴリズムとして、共分散行列法を用いる。

本稿におけるドキュメントの検索および追跡処理は、以下の手順による。

- 検索対象データのベクトルモデリング
- クエリーのベクトルモデリング
- クエリーをみたすドキュメント群の検索
- 検索により抽出されたドキュメント群について
  新規事項の発見および事後追跡処理

[提案手法における処理手順]

以降に、ベクトルのモデリング手法および検索／追跡処理についての詳細を述べる。

2.1 ベクトルモデリング

検索対象データとユーザからの入力によるクエリーは、同じ手順によってモデリングされる。システムの実行時にはクエリーのみをモデリングし、あらかじめモデリングされたデータを参照して検索する。

ここで用いるベクトルモデルは、対象データ内の新規な事項を発見し、その事項に続くフォローアップドキュメントを特定することを前提としている。このため、時間情報をキーワードと同等な属性として扱い、次元削減の際には時間情報も同時に考慮される。つまり、キーワードだけでなく、時間による特徴づけも同時におこなわれている。本手法によるベクトルモデリングのフローチャートを図1に示す。

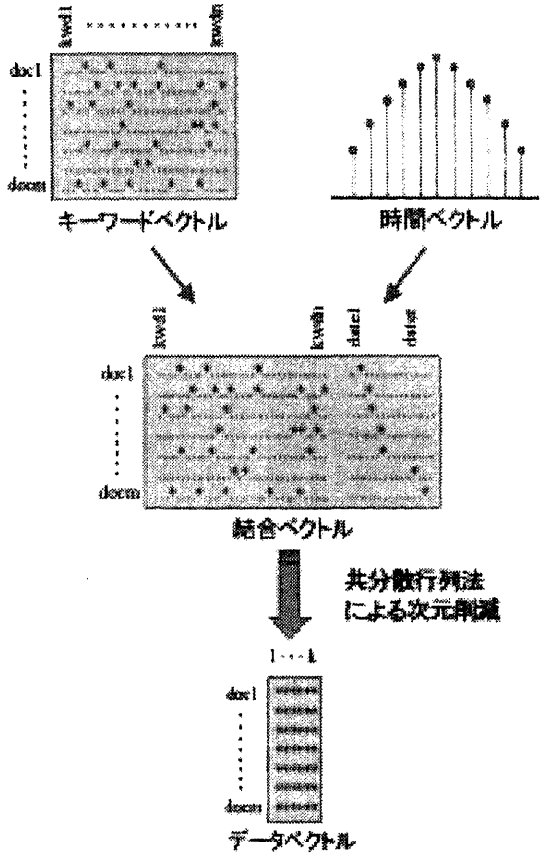
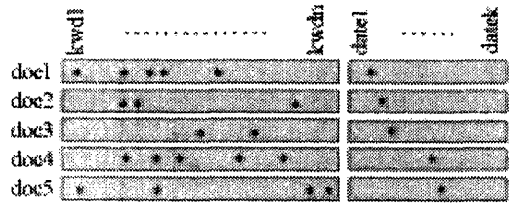


図 1: ベクトル作成手順

まず、対象データであるドキュメントからキーワードを抽出し、ベクトルを生成する。ドキュメント内に存在するキーワードに対応するベクトルの要素が、キーワードの重みを値としてもつ。

また、ドキュメントから発生時間を取得し、時間ベクトルを作成する。ドキュメントが発生した最初と最後の日付を基に、ドキュメントが生成された期間の日数分だけ要素をもつベクトルを定義する。対象データの時間ベクトルについては、各ドキュメントが生成した日付に対応する位置の要素にだけ値をもたせ、それ以外の要素を0とする。クエリーの時間ベクトルについては、各要素の重みをユーザが自由に決定できる。クエリーにおける時間の重みは、例えばある日付以降のドキュメントを優先的に検索したい場合や、特定の日付のドキュメントについて注目したい場合など、ユーザの意図を反映することができる。キーワードベクトルに時間ベクトルを付加し、一つのベクトルとして定義する。図2に例を示す。



* にはTF-IDFによる正の値が重みとして与えられる

図 2: ベクトル定義

本手法では、この時間情報を付加したベクトルに対して、次元削減をおこなう。次元削減には共分散行列法を用いる。この手法では、次元削減の過程で作成される行列のサイズが文書数に依存せず、属性数のみから決定されるため、文書数が膨大になった場合にも対応できる。

次元削減をおこなうことにより、データのサイズが縮小され、検索および追跡処理の速度が向上する。またドキュメント間の類似するキーワードの共起情報を取得できるため、同意語や多義語が適切に解釈されるという利点がある。

2.2 検索および追跡処理

ユーザからのクエリーをみたすドキュメント群を検索し、検索結果として抽出されたドキュメント群を追跡処理の対象とする。検索方法としては、ベクトルモデリングされたクエリーと、ベクトルモデリングされたデータとの余弦を計算し、余弦が閾値 ε をこえるドキュメントを検索結果として抽出する。このときの判定は膨大なデータに対しておこなわれるため、リアルタイムに結果を抽出するためには処理を高速化する必要がある。高速化する方法として、まずクエリーベクトル内で大きな値をもつ上位5つの要素を特定し、これらの位置に対応するデータベクトルの要素との間でのみ内積をとる。このときの内積値が一定値をこえた場合にだけ、全要素についての内積を計算し、余弦をもとめる。ここで、閾値 ε は0から1の間の値で、ユーザが任意に設定できるものとする。余弦の値が大きいほどベクトル間の角度は小さく、ドキュメント間の類似度が高いと判定できる。

ここで抽出したドキュメント群を発生時間順にソートし、追跡処理を開始する。追跡処理の手順は以下のとおりである。

1. まだ判定されていない最も古いドキュメント(N)を“新規”と判定
2. N以降に出現する、まだ判定されていない各ドキュメントとNの余弦を計算
3. 余弦が閾値 η 以上であり、1つ前に判定されたドキュメントの日付からdt以内に出現するドキュメントを“フォローアップ”と判定

以降、全ドキュメントに“新規”か“フォローアップ”かの判定がなされるまで1 - 3の処理を繰り返す。

閾値 η は0から1の間の値で、ユーザが任意に設定できるものとする。dtは時間についての閾値で、これもフォローアップドキュメントの判定に用いる。1つ前に判定されたドキュメントからの時間がdt以上離れている場合、そのドキュメントは“新規”と判定される。このdtの値は、対象とする事項の種類、ユーザによる検索の目的、必要性等を考慮して任意の値に最適化することができる。

3 システム概要

提案手法を用いたシステムのプロトタイプを実装し、これを本手法の評価実験に用いる。本システムでは、クエリーの入力および結果の表示をサポートするGUIを提供する。GUIを用いることで容易にクエリーを入力でき、またグラフによる表示から、結果を直感的に認識できる。特にタイムスタンプの入力にグラフウィンドウを用いることで、ユーザの意図を容易に反映させることができ、タイムスタンプを用いた検索／追跡処理の実行効率が向上される。

入力ウィンドウのスナップショットを図3に示す。ユーザはクエリーとして、キーワードとタイムスタンプを入力できる。また、各ドキュメントのクエリーに対する類似度を判定する際の閾値εおよび追跡処理の判定に用いる閾値ηを、スクロールバーを調節することで自由に設定できる。キーワードはダイアログボックスから直接入力し、タイムスタンプは図3に示すグラフから入力する。

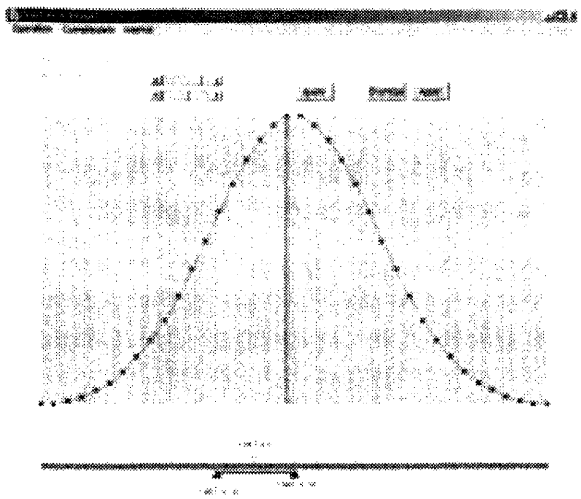


図 3: 入力ウィンドウ

タイムスタンプの設定にはいくつかのパラメータが用意されている。まず、タイムスタンプを適用する期間を設定し、この期間に対応するベクトル要素に重みづけする。タイムスタンプの重みはユーザがフリーハンドで自由に設定できるほか、いくつかの入力パターンを用意している；

- **正規分布：** 特定の日付に最も高い重みを設定し、この日からの距離にしたがって徐々に重みを減少させる。この形式のタイムスタンプは、ユーザが検索したいと望むドキュメントの発生日が幾分か不明確な場合や、ユーザが特定の日付周辺の事項を知りたいと希望する場合に有効である。
- **ステップ関数：** 特定の日付以降の重みを大きく設定し、それ以前の日付の重みは0とする。このタイムスタンプは、特定の日付以降に発生したドキュメントを優先的に検索／追跡するのに有効である。
- **ヘビサイド ステップ関数：** 特定の日付以前の重みだけを大きく設定し、その日付以降の重みを0に設定する。ステップ関数を用いる場合とは逆に、特定の日付以前に発生したドキュメントを優先的に検索／追跡処理するのに有効である。
- **スロープ関数：** 特定の日付以前の重みをゆるやかに増加させ，その日以降の重みを最大に設定する。この形式では、処理対象とするドキュメントの発生する境界線となる日付が幾分か不明確な場合に用いることが有効である。
- **デルタ関数：** 特定の1日だけの重みを大きく設定する。このタイムスタンプは、ユーザの処理対象とするドキュメントの発生日が明確に分かっている場合に有効である。
- **毎週／月／年：** 指定した日から，毎週／毎月／毎年といった一定の間隔毎に、タイムスタンプの重みを大きく設定することができる。この形式のタイムスタンプは、毎週もしくは毎月、毎年といった一定周期で発生するドキュメントを対象とする場合に有効である。

これらの入力から時間ベクトルを作成し、キーワードに付加してクエリーベクトルを生成する。

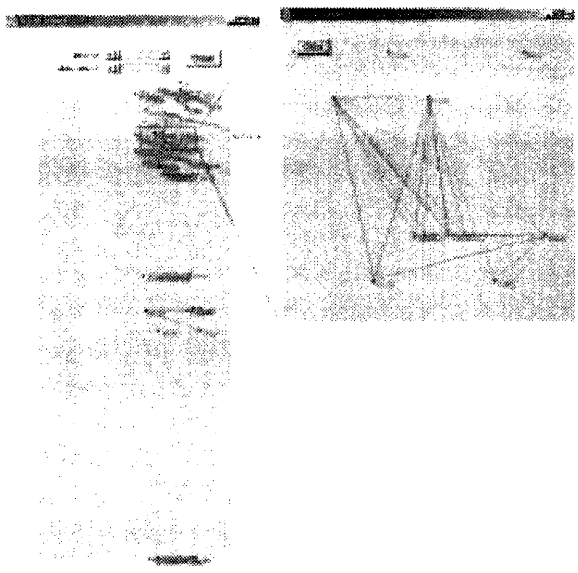


図 4: 出力ウィンドウ

以上の手続きにより生成されたクエリーベクトルに対して、本手法による検索／追跡処理を行な

い、結果をグラフとして出力する。出力ウィンドウのスナップショットを図4に示す。グラフの縦軸は時間軸であり、横軸がクエリーベクトルと各ドキュメントの余弦をあらわす。余弦が大きいほど、すなわちグラフの右側の領域に表示されているドキュメントほど、クエリーに対する類似度が高い。ここで、グラフ内のノードの色は追跡処理における判定結果を表している。赤色のノードは新規ドキュメント、ピンクのノードはフォローアップ事項を示す。ノード間を結ぶリンクはドキュメント間の関連を表し、あるドキュメントが別のドキュメントのフォローアップであれば、これらのドキュメント間はリンクで結ばれる。リンクの明度が低いほど強い関連をもっている。すなわちドキュメント間の類似度が高いことを示している。このグラフはドキュメントが生成された全時間の関連構造を1つの画面に表示するため、各要素を個別に認識することが困難となる。そこで、指定した範囲を拡大表示するオペレーションを追加した。拡大表示されたウィンドウ内では、マウスカーソルが近付くと近隣のノードのタイトルが表示される。また、特定のノードをクリックすると別ウィンドウが開き、ノードに対応するドキュメントの内容が表示される（図5）。

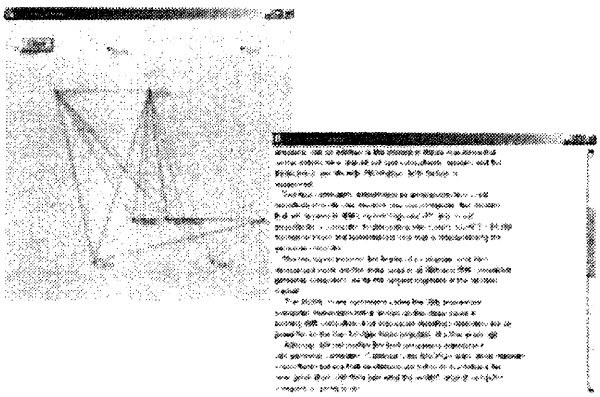


図 5: 詳細表示ウィンドウ

このプロトタイプシステムにより、処理の実行効率が向上され、結果の認識も容易になる。

4 実験

本システムを用いて新規事項の発見およびその事後追跡を行った結果と、検索／追跡処理についての実行時間を計測し、報告する。実験対象データはReutersのニュースドキュメント（21578件）で、次元削減後のベクトルの次元は200次元である。なお、実験環境として、IBMのIntelli Station (CPU:PentiumPro[§] 2GHz,RAM:2GB,OS:Windows 2000) を用いた。

4.1 実行速度

ここで計測する時間には、データの入出力処理にかかる時間を含まず、実際にユーザがクエリーを入力してから、検索／追跡し、結果が得られるまでの一連の処理にかかった時間だけを計測している。具体的な処理ステップとしては、

- クエリーのベクトルモデリング
- クエリーによるドキュメントの検索
- 検索結果ドキュメントの追跡処理

の3つである。

表1に示す計測時間は、同じ条件による検索および追跡を、それぞれ10回ずつ実行した結果の平均値である。上に示した3つの処理ステップそれぞれについての実行時間と、トータル時間を計測した。実験には、クエリーとする任意のキーワードを入力し、タイムスタンプを任意に設定した。

この結果にみられるように、各クエリーに対する実行時間は1秒を大きく下回り、この手法を用いることでリアルタイム処理を実現できる可能性があることを示唆している。

4.2 検索および追跡結果

ユーザからのクエリーに対する処理結果を検証する。評価項目として、クエリーに適したドキュメントを検索しているか、ドキュメント間の関連は適切に追跡されているかの2点を考える。まず、本手法でドキュメントの類似判定に用いる閾値 ε および、追跡処理に用いる閾値 η の最適値をもとめる。

4.2.1 検索結果の検証

クエリーベクトルとドキュメントベクトルの余弦が閾値 ε を越えている場合、そのドキュメントは検索結果として抽出される。

図6に、閾値 ε に対する再現率と適合率を示す。再現率とは、クエリーに関連する全ドキュメントのうち実際に抽出されたドキュメントの割合であり、適合率とは、抽出されたドキュメント群のうちクエリーに適合するドキュメントの割合で

[§] PentiumProはIntel社の商標である.

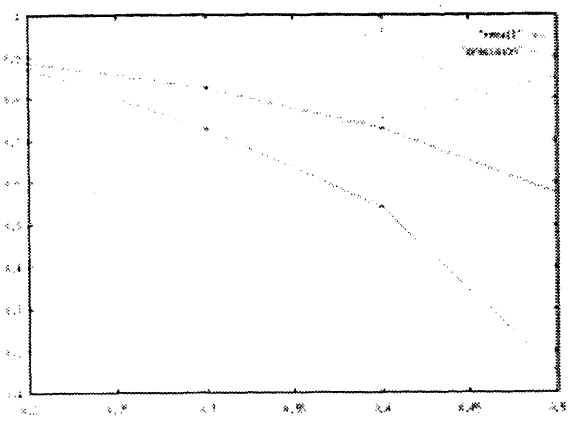


図 6: 再現率／適合率グラフ

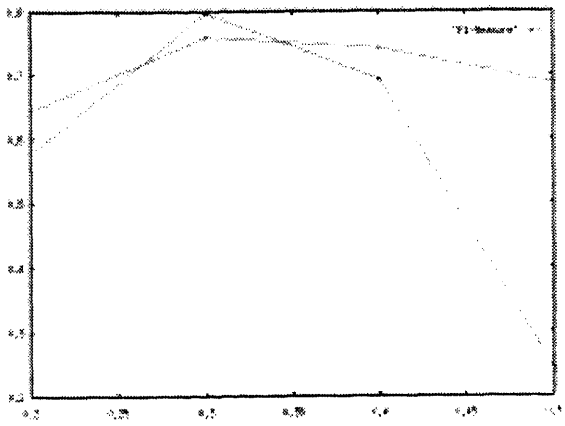


図 7: F 値グラフ

ある。グラフ（図 6）の x 軸は閾値 ε の値を表し、y 軸は再現率および適合率の値を表す。グラフにみられるように、閾値が小さいほど、検索結果のドキュメント数が多くなり再現率が高くなっている。一方、ドキュメント数が多いほど目的とするドキュメントの割合は小さくなり、適合率は小さくなっている。

一般に、適合率の値が大きくなるにつれて再現率の値が小さくなるというトレードオフの関係にあるため、2 つの値の両方を適度にみたす閾値を決定することが大切となる。しかし、2 つの値を同時に考慮することは困難である。これを一つの値で評価する方法として、F 値という評価式が知られている [9]。

$$\text{F 値} = \frac{2 \times (\text{再現率}) \times (\text{適合率})}{(\text{再現率}) + (\text{適合率})} \qquad (\text{評価式 1})$$

F 値は、再現率と適合率から決定される値である（評価式 1）。F 値の値が大きいほど、実行結果はより適切であるといえる。図 7 に、F 値を計測したグラフを示す。グラフ（図 7）の x 軸は閾値 ε の値、y 軸が F 値の値である。

このグラフから、ε =0.3 のときに F 値が最も大きくなり、システムの検索結果が最適となることがわかる。

以上の考察より、閾値としてもっとも適切な値は ε =0.3 といえる。これは今回の実験における最適値であり、対象データやクエリーの内容によって最適値が異なる。今回の実験結果による最適値を目安に、本システムではインタラクティブに閾値を設定して検索することができる。

4.2.2 追跡結果の検証

検索結果として抽出されたドキュメントについて、各ドキュメントが新規事項に関するドキュメントか、別のドキュメントのフォローアップ事項かの判定をおこなう際に、閾値 η を用いる。ドキュメント相互の余弦が閾値 η を超えている場合、これらのドキュメントは関連をもち、発生時間が後のドキュメントは、もう一方のドキュメントのフォローアップ事項であると判定される。ε =0.3 とした場合の検索結果に対して追跡処理をおこなった場合の実験結果を、以下に示す。

クエリーワード="iraq"、η =0.62 とした場合の結果をリスト 1 に示す。日付およびタイトルをリストアップし、追跡処理に対する判定を記述している。ここで、日付の前がインデントされているのものは、インデントされていない直前のドキュメントのフォローアップであることを示している。ここで、追跡結果の判定として、違っていると判断できるものに × と記述している。追跡結果の判定は、検索目的や、ユーザのドキュメントの見方に依存するため、一概に決定することはできないが、ここでは判定の一例を示している（リスト 1）。リスト 1 からも、高い精度で追跡処理を実現できていることが分かる。

ここでは、η =0.62 とした場合の追跡処理の一例を示したが、η の値はユーザの検索目的や判断基準に応じて自由に設定できる。

5 まとめと今後の課題

今回、新規事項の発見およびその事後追跡をおこなう新しい手法を用い、プロトタイプシステムによる実験結果を検証した。今回の実験により、本手法を用いてリアルタイム処理が実現可能であることがわかった。また、閾値およびタイムスタンプの決定をサポートするパラメータの最適値を経験的に模索し、特定した。

今後の課題としては、ベクトルモデルの次元削減後の次元数を調節し、ドキュメント集合の特徴を適切に捉えることのできる値を模索する。またGUIの整理を進め、より使いやすく、より認識しやすいシステムの構築を目指す。今回の実験にはReutersのニュースドキュメント(21578件)を用いたが、今後はさらに大規模なデータに適用し、検証する予定である。

参考文献

[1] S. Weiss et al., "Maximizing text-mining performance", In *Proc. IEEE Intelligent Systems*, 14(4):63–69, 1999.

[2] Y. Yang and J. Pedersen, "Intelligent information retrieval", In *Proc. IEEE Intelligent Systems*, 14(4):30–31, 1999.

[3] J. Allan et al., "Topic Detection and Tracking Pilot Study: Final Report", In *Proc. DARPA Broadcast News Transcription & Understanding Workshop.*, 194–218, 1998.

[4] G. Salton (ed.), "The SMART Retrieved System - Experiments in Automatic Document Processing", Prentice-Hall, Englewood Cliffs, NJ, 1971.

[5] A. Moffat, T. Bell, Managing Gigabytes, second ed., Morgan Kaufmann, San Francisco, CA, 1999.

[6] R. Baeza-Yates and B. Ribeiro-Neto (eds.), Modern Information Retreival, ACM Press, NY, 1999.

[7] W. Frakes, and B. Baeza-Yates (eds.), Information Retreival, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[8] R. K. Belew, C. J Van Rijsbergen, Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW, 2001.

[9] D. Lewis, and W. Gale, "A Sequential Algorithm for Training Text Classifiers", In *Proc. ACM SIGIR*, 3–13, 1994.

入力クエリー	ヒット件数	モデリング時間	検索時間	追跡時間	トータル時間
iraq, oil	13	259.3	34.5	0.2	294.0
gulf,war	15	303.2	28.0	0.1	331.3
ibm, computer	94	257.7	35.9	10.8	304.4
airplane	141	185.4	33.3	15.7	234.4
computer	182	179.7	41.7	20.8	242.2

時間の単位はミリ秒 (msec) である。

表1： 本手法による検索／追跡処理の実行時間

日付	タイトル	判定
03.01	Iraq says it repels Iranian attack	
03.11	Iraq says it crushed Iranian attack in north	
03.11	Iraq says it crushed Iranian attack in north	
03.11	Iraq says Iranians thrown back in south	
03.11	Iraq says Iran attack repulsed on southern front	
03.11	Iraq says Iranians thrown back in south	
03.11	Iraq says Iran attack repulsed on southern front	
03.11	Iraq says it crushed Iranian attack in north	
03.11	Iraq says Iranians thrown back in south	
03.11	Iraq says Iran attack repulsed on southern front	
03.11	Iraq says Iranians thrown back in south	
03.11	Iraq says Iran attack repulsed on southern front	
03.02	Iraq defers payments on 500 mls dlr euroloan	
03.02	UAL unit attacked on minority hiring	
03.03	Rostenkowski says he will oppose protectionist trade bill in U.S. house	
03.03	Iraq reports Iran attack repulsed east of Basra	
03.05	Iraq says it crushes new Iranian Basra offensive	
03.11	Iraq says it crushed Iranian attack in north	
03.18	Iraq says it attacks two ships in gulf	
03.23	Iraq reports attacks on supertanker, oil targets	×
04.07	Iraq says Iran offensive on southern front checked	
04.07	Iraq reports fierce fighting to remove Iranians	
04.08	Iraqi troops reported pushing back Iranians	
04.08	Iraqi troops reported pushing back Iranians	
04.07	Iraq CCC credit guarantees switched - USDA	
04.09	CCC guarantees to Iraq switched - USDA	

リスト1： クエリー "iraq" に対する追跡結果