自然言語処理 119－4
音声言語情報処理 16－10
（1997. 5. 26）

トレンド・トラッキング型テキスト自動分類の試み

巖寺 俊哲　菊井 玄一郎
ＮＴＴ情報通信研究所

概要

本稿では、まず、ネットニュース記事データを例に投稿される記事の情報内容が時間とともに変化していることを定量的、定性的に示す。さらに、分類対象の情報内容の変化に追従しつつテキストを分類するトレンド・トラッキング型テキスト分類手法について報告する。本手法を用いた実験システムを試作し、ネットニュース記事を対象に分類実験を行なった。その結果、従来の分類手法と比較して、トレンド・トラッキングにより、分類精度が向上することを確認した。

An Automatic Text Classification Using Trend-Tracking

Toshiaki IWADERA, and Gen'ichiro KIKUI
NTT Information and Communication Systems Laboratories

Abstract

In this paper, we describe novel text classification method using trend-tracking. It classifies texts as tracking the transition of information in them. First we have analyzed the transition of information content by using internet news articles. Then we developed an experimental system and conducted text classification experiments for internet news articles. The experimental results show that our method using trend-tracking leads to the improvement of classification accuracy in comparison with conventional methods.

1 はじめに

情報通信基盤の整備によって大量かつ多様なテキスト情報に容易にアクセスできるようになってきた現在、大量のテキストの中から有用な情報をいかに効率的に探索するかが問題となっている。この問題に対処するための有効なツールの一つが、与えられたテキストをその意味内容に基づいて、あらかじめ決められたカテゴリに自動的に分類する「テキスト自動分類（技術）」であり、今までに様々な手法が提案されている。本稿では特にネットニュースや新聞記事、あるいは、更新頻度の高い Web ページ（群）などのように時間的に内容が徐々に推移していくようなテキストを自動分類する手法について検討する。

既存のテキスト自動分類手法は、

1. 表記等の統計情報を用いる方法 [1, 2, 3, 4]
2. 分類体系に依存した知識を用いる手法 [5]

の２種類に大別できる。前者は、予め分類済のテキスト集合を用意し、このテキスト集合中に出現する単語等の表記の頻度等の統計情報を用いる手法である。この手法は、処理が簡単で汎用性が高く、近年、様々な手法が提案されている [1, 2, 3, 4]。後者は、予め人手によりキーワードと分類分野間の階層的な対応表を作成し、この対応表をたどることによりテキストを分類する手法である。

しかし、これらの手法では、分類対象の情報内容の変化が考慮されていない。このため、変化した場合には、分類精度が低下すると考えられる（後で実験データを示す）。分類精度を維持するためには、変化に対応して、統計情報を抽出するための分類済のテキスト集合を変更したり、人手により対応表を変更する必要がある。さらに、情報内容の変化を自動的には検出し、適応することができない。このため、人間が分類結果をモニタすることによって情報内容の変化が起きたか否かを判断する必要があり、起きたと判断した時点でテキスト集合や対応表の変更を行なう必要がある。

そこで本研究では、分類対象情報の単語やキーワードの出現分布の変化（トレンド）に自動的に追従しながらテキストを分類する「トレンド・トラッキング型テキスト自動分類」を提案し、実験システムを試作した。さらに、このシステムを使用して、ネット・ニュース記事を例題として自動分類実験を行ない、トレンド・トラッキングの有効性を検討した。

以下、まず2節で、ネット・ニュース記事データを例として情報内容の変化について定量的、定性的に述べる。3節では、今回試作した実験システムで用いた手法について述べる。4節では、今回行なった実験内容と結果について述べる。5節では、本手法の適用結果について考察する。6節では、まとめと今後の課題を述べる。

2 情報内容の変化

情報内容が変化した場合、変化に応じて使用される語彙も変化すると考えられる。ここでは、今回、例題としたネット・ニュース記事中の使用語彙の変化を定量的、定性的に見ることによって情報内容の変化を調べる。調査対象としたニュース・グループを、表 1に示す。以下、これらのニュース・グループを対象に投稿日別に分割し、投稿日と情報内容の変化の関係を定量的、定性的に示す。

表 1: 調査対象としたニュース・グループ

ニュース・グループ名	ニュース・グループ名
fj.fleamarket.autos	fj.rec.food
fj.fleamarket.books	fj.rec.games
fj.fleamarket.comp	fj.rec.outdoor
fj.fleamarket.misc	fj.rec.rail
fj.fleamarket.tickets	fj.rec.sports.ski
fj.rec.autos	fj.rec.travel
fj.rec.autos.sports	fj.rec.travel.air
fj.rec.autos.wagon	fj.rec.travel.japan
fj.rec.bus	fj.rec.travel.world
fj.rec.drink	fj.rec.wine
fj.rec.drink.liquor	fj.wanted

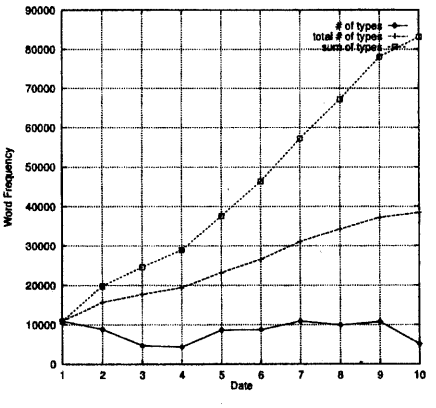

図 1: 異なり語数と投稿日の関係

2.1 定量的変化

まず、各投稿日毎に出現する異なり語数の変化を用いて情報内容の変化を調べた。ここで対象にした記事中の異なり語は、各記事を juman2.0 を使用して形態素解析し、得られた見出し語のうち品詞が名詞と付与されていたものである。異なり語数と投稿日の関係を図 1に示す。図 1において、“# of types” は、各投稿日毎に出現した異なり語数を、“total # of types” は、各投稿日以前に出現した異なり語数を、“sum of types” は、各投稿日毎に出現した異なり語数を単純に累積した数をそれぞれ表している。

図 1において、異なり語数は、時間の経過とともに一様に増加している。これに対して、各投稿日毎の異なり語数は、約 10000 語〜約 5000 語の範囲で推移している。これは、新しい語が毎日出現していることを示す。また、毎日の出現数を単純に累積した数と異なり語数を比較すると、単純に累積した数に比べ、異なり語数が少ないことを示している。もし、毎日出現する語がすべてそれ以前に出現したことがない初出語の場合、異なり語数と日々の異なり語数を単純に累積した数は一致する。これは、毎日出現する語のすべてが初

出語ではなく、それ以前に出現した語と共通の語も出現していることを示している。

2.2 定性的変化

次に、各ニュース・グループ毎に各グループを特徴つける語を投稿日毎にみることによってどのように情報内容が変化しているかを調べた。グループを特徴つける語は、後述する実験システムでも用いている χ^2 値を用いて計算されるスコアを用いることにより得られる。このスコアが高いほど、各グループをより特徴つける語となる。

表2に各ニュース・グループ毎にスコア上位語と投稿日の関係の数例を示す。表2から、実際に情報内容が変化していることが見てとれる。たとえば、表2から、ニュースグループ fj.rec.food では、議論の主要な内容が、カツ丼→ワイン→蕎麦→ラーメン→ホヤ→コーヒーと遷移していることが推測できる。

2.3 情報内容の変化のまとめ

以上のことをまとめると、次のようになる。

- 情報内容は、完全に別の内容に遷移するのではなく、ある時点と次の時点の情報内容は、その一部が重なっており、徐々に変化している
- 語彙という観点から見ると新出語彙と既出語彙が共起しつつ遷移している

これは、我々の直観にも合致している。

3 トレンド・トラッキング型テキスト分類手法

ここでは、今回、試作した実験システムで用いたトレンド・トラッキング型テキスト自動分類手法について述べる。

本手法は、2節で得られた、「新出語彙と既出語彙が共起しつつ遷移している」という知見に基づいている。この知見は、既出の情報を用いて新出の情報を獲得できることを示している。すなわち、情報内容の変化に自動的に追従してテキストを分類することが可能であることを示している。

本手法は、以下の3つのプロセスから構成される。

1. 学習プロセス
2. 分類プロセス
3. トレンド・トラッキング・プロセス

ここで、1、2は、それぞれ従来の統計情報を用いる分類手法の学習プロセス、分類プロセスと基本的に同様である。

統計情報を用いる手法は、そこで用いるモデルあるいは考え方によって、

1. ベクトル空間モデルを用いる手法 [4, 6]
2. χ^2 検定の考え方を用いる手法 [1, 2]

の2種類の手法に大別できる。また、これらの2つの考え方を組み合わせた手法も提案されている [7]。本手法では、分類処理の核となる手法として後者の χ^2 検定の考え方を用いる手法を採用した。

3のトレンド・トラッキング・プロセスは、本手法に特徴的なプロセスである。このプロセスでは、分類対象としているテキストに含まれている既出語彙を利用し、新出語彙を獲得し、分類スコアテーブルを更新するための再学習データとして学習プロセスにフィードバックする。これにより情報内容の変化に追従することが可能となる。

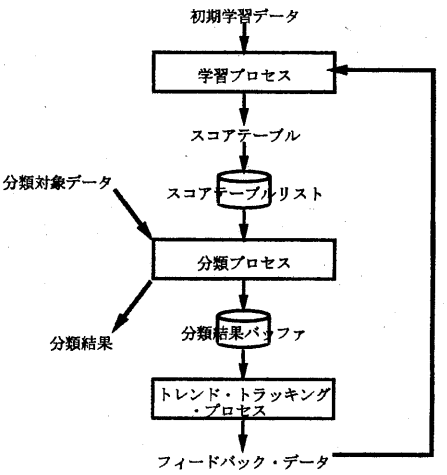

図 2: トレンド・トラッキング型テキスト分類手法の概観

本手法の処理の流れを図2に示す。

まず、初期学習データとして正解分類先が付与されているテキスト集合を与えることによって処理が開始される。学習プロセスは、与えられた学習データを基に各分類先毎のテキストの特徴を分類用スコア・テーブルとして出力する。分類プロセスは、分類対象テキストが入力されると、学習プロセスにより作成された分類用スコアテーブルを使って入力されたテキストを分類し、その結果を出力するとともに、分類結果バッファに蓄積する。トレンド・トラッキング・プロセスは、定期的に分類結果バッファから分類結果を読みだし、以降の処理に有用なデータを選択し、学習プロセスにフィードバックする。

学習プロセスは、このフィードバックされた分類結果を用いて、新たな分類用スコア・テーブルを作成する。

以下、各プロセスについて述べる。

3.1 学習プロセス

学習プロセスでは、学習データ中のテキストの各単語の分類用スコアを計算し、分類用スコアテーブルを作成する。

分類用スコアテーブルは、以下に示す手順で作成される。まず、学習データ中の各テキストを形態素解析して、単語を抽出し、これらの分類先毎の出現頻度を計算する。次に出現頻度から分類用スコアを計算する。分類用スコアは、文献 [1] の計算式に従って、χ^2 値を用いて計算される。ここで計算される分類用スコアは、特定の分類先にのみ多く出現する単語ほど高スコアに、また、各分類先に均等に出現する単語ほど低スコアとなる。学習データから抽出されたすべての異なり単語について、上記の手順で分類用スコアを計算し、分類用スコアテーブルを作成する。

表 2: 投稿日毎の各ニュース・グループを特徴つける語

投稿日	fj.rec.bus	fj.rec.food	fj.rec.sports.ski	fj.rec.travel.japan	fj.rec.wine
1	市 バス 市営	カツ 丼 ソース	スキー 板 製品化	温泉 レンタカー 北海道	the your and
2	市 バス 市営	味 ワイン フジツボ	サングラス 板 スキー	水族館 同僚 跡	ソムリエ 試験 合格
3	駅 新 小	(au) a	スキー 事業所 連盟	テント 彼 公園	コルヴォ ワイン 選択肢
4	入口 投	卵 蕎麦	／	駒ヶ根	ワイン ブドウ
5	駅前 一 鹿児島	／	グリップ 調整 膝	バス 湖	／
6	乗 継 釣銭	ラーメン 味 店	レース スキー 大会	庭 寺	ワイン ピアノ 協奏曲
7	回数券 地下鉄 バス	ラーメン " カレー	スキー 大会 二村	峠 工事 津別	ワイン NewsGroup 通
8	柿 路線 奈	ラーメン 味 吉野	スキー 上級 膝	伊賀 上野 御坊	ワイン ショット バー
9	引率 回数券 バス	店 屋 ホヤ	板 スキー ワックス	伊賀 上野 野市	ワイン Vollum 石松
10	運賃	コーヒー	／	通路 納札 峠	ワイン コルク

学習プロセスの結果得られる分類用スコアテーブルは、分類先数が m 個、学習用データから抽出された異なり単語数が n 個のとき、$m \times n$ のテーブルとなる。作成された分類用スコアテーブルは、作成に使用した学習データがどの時点での学習データかを示すタイム・インデックスとともに分類用スコアテーブル・リストに記録される。作成に使用した学習データが初期学習データの場合は、タイム・インデックスは、ゼロとする。

3.2 分類プロセス

分類プロセスでは、学習プロセスで作成された分類用スコアテーブルを用いて以下の手順で、分類対象テキストを分類する。

まず、分類対象テキスト d を形態素解析し、各単語の出現頻度を計算する。

次に、分類対象テキスト d のスコア・ベクトル $\boldsymbol{S}=(s_1,\cdots,s_j,\cdots,s_m)$ を計算する。ここで、s_j は、テキスト d の分類先 j に対するスコアである。m は、分類先数である。$\boldsymbol{S}$ は、式1により計算する。

$$\boldsymbol{S}=\sum_{t=0}^{T} c_t \boldsymbol{S}_t \qquad (1)$$

ここで、$\boldsymbol{S}_t$ は、タイム・インデックスが t であるスコアテーブルを用いて計算されるスコア・ベクトルであり、各成分は、各分類先を示し、各成分の値は、各分類先のスコアを表す。これは、式2で与えられる。

$$\boldsymbol{S}_t=\sum_{i=1}^{n} tf_i \boldsymbol{W}_{ti} \qquad (2)$$

ここで、$\boldsymbol{W}_{ti}$ は、タイム・インデックス t のスコアテーブルから読み出された単語 i の単語スコア・ベクトル $(w_{t1},\cdots,w_{tj},\cdots,w_{tm})$ であり、、w_{tj} は、単語 i の分類先 j に対するスコアである。また、tf_i は、単語 i の出現頻度、n は、分類対象テキストから抽出された単語数を示す。もし、単語 i が参照したスコアテーブル中に存在しない場合は、単語スコア・ベクトル $\boldsymbol{W}_{ti}=\boldsymbol{O}$ とする。以上により、タイム・インデックス t であるスコアテーブルを用いたスコア・ベクトル $\boldsymbol{S}_t$ が計算される。

また、式1中の c_t は、$\boldsymbol{S}_t$ の組み合わせ方を制御する係数であり、現在、処理開始前に人手で任意に設定するようになっている。また、T は、スコアテーブル・リストに記録されている中で最新のスコアテーブルのタイム・インデックスである。

次に、出力する分類先を決定する。以上のプロセスで得られたスコア・ベクトル $\boldsymbol{S}$ は、分類先候補を表している。この中から出力する分類先を選択する。一般に、スコアが最も高い分類先を1つだけ出力する方式と、各分類先のスコアの大きさを考慮して分類先を複数出力する方式がある。本手法では、現在、スコアが最も高い分類先を1つだけ出力する方式をとっている。

また、以降のトレンド・トラッキング・プロセスで分類結果を利用するために、1つに決定された分類先だけでなく、スコア・ベクトルと分類対象となったテキストを分類結果バッファに記録する。

3.3 トレンド・トラッキング・プロセス

従来の統計情報を用いた分類手法では、、分類対象データの情報内容が変化した場合、変化に対応した正解分類先付与済みの学習データを収集、あるいは、人手により作成し、再度、学習していた。トレンド・トラッキング・プロセスは、この作業を自動化するためのプロセスである。

トレンド・トラッキング・プロセスは、以下に示す手順で進行する。定期的に分類結果バッファから記録されている分類結果を読みだす。分類結果を読み出す毎に、タイム・インデックスをインクリメントし、分類結果に付与する。読み出した分類結果をテキストを単位として、後述する分類結果評価法を用いて評価し、フィードバック適性度を付与する。フィードバック適性度は、フィードバックするデータとしての適切さを表す点数であり、点数が高いほどフィードバックデータとしてより適していることを示す。このフィードバック適性度が予め設定されたある閾値以上である分類結果が、再学習用データとして、付与されているタイム・インデックスとともに学習プロセスにフィードバックされる。（フィードバックされた分類結果は、学習プロセスによってタイム・インデックス t の分類用スコア・テーブルが作成される。）

分類結果評価法

学習プロセスにフィードバックし、再学習用データとして使用する分類結果は、より良質なデータであることが望ましい。すなわち、正しい分類先が出力されているデータがよい。分類結果評価法は、フィードバックするデータとしての分類結果の適切さを評価する方

法である。

本手法で用いている分類用スコアの計算方法では、分類対象テキストがある特定の分類先にのみ多く出現する単語を多く含んでいるほど、その分類先のスコアが高くなり、他の分類先とのスコアの差が大きくなる。反対に、様々な分類先に出現する単語を多く含んでいるほど分類先間のスコアの差が小さくなる。すなわち、前述した分類プロセスにおいて得られスコア・ベクトルにおいて、分類先として出力した成分の値が、他の分類先候補の成分の値に比較して突出して高スコアであるほどその分類先が正解である可能性が高い。本分類結果評価法では、この性質を利用して、フィードバック適性度を求める。

フィードバック適性度は、出力分類先のスコアと他の分類先候補のスコアの差の大きさ、すなわち出力分類先スコアの突出の大きさを表す指標である。そこで、最適な突出をあらわす参照ベクトル $\boldsymbol{R}$ を用意し、このベクトルと分類結果のスコア・ベクトルの類似度を計算することによって、フィードバック適性度を算出する。

まず、参照ベクトル $\boldsymbol{R}$ を用意する。このベクトルの大きさ $|\boldsymbol{R}|$ は、スコア・ベクトルの大きさ $|\boldsymbol{S}|$ と同じであり、出力分類先に対応する成分の値が1であり、他の成分の値は0であるベクトルである。次に参照ベクトル $\boldsymbol{R}$ とスコア・ベクトル $\boldsymbol{S}$ の類似度を計算する。この類似度が、フィードバック適性度 FI である。フィードバック適性度 FI は、式3により計算する。

$$FI = \frac{\boldsymbol{R} \cdot \boldsymbol{S}}{|\boldsymbol{R}||\boldsymbol{S}|} \qquad (3)$$

ここで、$\boldsymbol{R} \cdot \boldsymbol{S}$ は、参照ベクトル $\boldsymbol{R}$ とスコア・ベクトル $\boldsymbol{S}$ の内積を表す。

4 実験

トレンド・トラッキングの効果を調べるために前述した手法を用いた分類実験を行なった。

実験は、トレンド・トラッキングを用いた実験と対照実験とし従来方式による実験（トレンド・トラッキングを用いない実験）の2つを行なった。従来方式による実験によって、本手法で核として用いている分類手法の性能と実験対象データの性質も同時に得ることができる。

実験で用いている形態素解析は、JUMAN2.0である。また、学習データテキストおよび分類対象テキストから抽出した単語は、形態素解析結果で品詞が名詞と付与されている形態素のみである。

4.1 実験対象データ

実験対象データは、ネットニュース記事である。対象とした記事は、表1に示すニュースグループに1996年9月1日～10日の10日間に投稿されたものである。各記事の分類先は、その記事が投稿されたニュースグループ名を用いた。

これらの実験対象記事を各記事の投稿日毎に分割した。すなわち、10日間に投稿されたデータを対象にしていることから全データを10個の記事群に分割した。これらの各記事群を1個の単位として、学習データまたは分類対象データとして用いた。

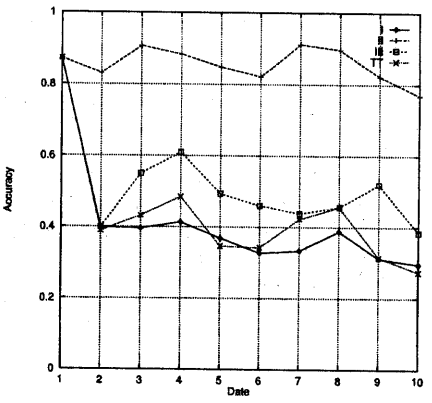

図 3: 分類精度の推移

4.2 実験方法

実験は、分類方式毎、学習データ毎に行なった。各実験では、各投稿日毎の記事を分類対象データとして分類精度を測定し、これが投稿日とともにどのように推移するかを調べた。

分類精度をみる評価尺度としては、一般に、再現率、適合率が用いられる。しかし、本手法では、分類先を1個だけ出力するように設定しているため、常に、再現率と適合率は同一である。したがって、今回の実験では、式4により算出される値を分類精度とする。

$$\text{分類精度} = \frac{\text{正しく分類された記事数}}{\text{分類対象とした記事数}} \qquad (4)$$

4.3 従来方式による分類実験

従来方式による分類実験、すなわち、トレンド・トラッキングを用いない実験では、学習データとして次に示す3種類をそれぞれ用いた3つの実験を行なった。

I: 投稿日が最も古い一日分の記事群

II: 分類対象と同一投稿日の記事群

III: 分類対象データとした記事群の投稿日の前日を投稿日とする記事群

図3に実験結果を示す。図中の各グラフ、I、II、IIIは、それぞれ上記の学習データI、II、IIを用いた場合の実験結果を表している。

本実験では、学習データIIを用いた場合（グラフ：II）、最もよい分類精度で、常に80%以上を得ている。これは、理想的な実験条件の設定であるが、従来方式でもよりよい学習データが得られれば高い精度で分類できる可能性を示している。次に分類精度が高かったのは、学習データIIIを用いた場合（グラフ：III）であり、分類精度40%～60%で推移している。最も精度が低かったのは、学習データIとした場合（グラフ：I）である。第1日目のみ高い精度を得ているが、これは、この日だけが実験条件の設定から分類対象データと学習データが同一であるからである。第2日目以降は、精度が約40%から徐々に低下し、10日目では、約30%になっている。

4.4 トレンド・トラッキングを用いた分類実験

次にトレンド・トラッキングを用いた分類実験を行なった。初期学習データとしては、上記の学習データIを用いた。

この実験では、分類時に、スコアテーブル・リストに記録されている各スコアテーブルから得られるスコアを組み合わせるための式1の組み合わせ方を制御する係数を、$c_T = 0.5$、$c_0 = 0.5$、$c_t = 0 (t = 1 \ldots T-1)$ としている。すなわち、次式によりスコア・ベクトルを算出している。

$$S = 0.5S_T + 0.5S_0 \qquad (5)$$

これは、スコア・ベクトルの算出に初期学習データと直前にフィードバックされたデータから算出されたスコアのみを用いていることを示す。また、フィードバックデータを選別するために用いるフィードバック適性度の閾値は、0としている。

図3に実験結果を示す。図中で、グラフTTが、トレンド・トラッキングを用いた場合の実験結果のグラフである。

本実験の結果、トレンド・トラッキングを用いた場合の精度は、従来方式で学習データIIIを用いた場合の精度と学習データIを用いた場合の精度の間でほぼ推移している。

5 考察

まず、従来方式による分類実験結果を情報内容の変化の観点から検討する。実験結果から学習データをIIからIIIへ変更することで、分類精度が約５０～３０ポイント低下している。学習データのこの違いは、分類対象データと同一日のデータか前日のデータかの違いである。これは、closed dataからopen dataへ変わったことによる精度の低下とともに、今回の実験では、分類対象データであるニュース記事の投稿日の違いによるすなわち情報内容の相違とみることができる。この相違が小さいほど精度の低下は、小さいと考えられる。また、学習データIを用いた場合の実験結果をみると、分類対象データ記事の投稿日が学習データとして用いたものから離れるにしたがって精度が徐々に低下している。これは、情報内容が徐々に変化しており、時間が経るにしたがって学習データとして用いたデータの内容と分類対象としたデータの内容の相違が大きくなっていることを示している。

次に、トレンド・トラッキングを用いた分類実験結果を検討する。この実験は、従来方式で学習データIを用いた場合と外部から与えた学習データ、分類対象データはすべて同一である。実験条件の差異は、トレンド・トラッキング手法の使用の有無のみである。実験結果からトレンド・トラッキングを用いている本実験結果のほう高い分類精度を得ている。このことからトレンド・トラッキングが分類精度の向上に対して効果があることがわかる。

さらに、分類するのに使用した学習データの量の観点から比較する。従来方式で、学習データIIIを用いた場合は、常に分類対象データの前日のデータを学習データとして使用している。すなわち、分類対象日数分の学習データを使用している。これに対して、トレンド・トラッキングを用いた場合は、最も古い投稿日の一日分の学習データのみを使用している。この２つの結果を分類精度で比較した場合、トレンド・トラッキングを用いた結果は、０～１５ポイントの低下にとどまっている。

以上のこのことからトレンド・トラッキングを用いることで、学習データ量が同一ならば、精度が向上し、より少ない学習データ量である程度の分類精度が得られる効果がある。

6 おわりに

本稿では、ネットニュース記事データを例にその情報内容が投稿日の経過とともに変化していることを定量的、定性的に示した。さらに、情報内容の変化に追従しつつテキストを分類する、トレンド・トラッキング型テキスト自動分類手法について報告した。また、この手法を用いた実験システムを試作し、ネットニュース記事を例題に分類実験を行なった。その結果、トレンド・トラッキングを用いることで、学習データ量が同一ならば、精度が向上し、より少ない学習データ量である程度の分類精度が得られる効果があることが確認できた。

本稿で報告した手法では、分類結果を再評価することにより情報内容のトレンドを捉えた。この考え方は、様々な統計的分類手法に適用することができる。今後は、ここで用いた χ^2 値を用いる手法以外の手法に対する分類結果のフィードバックの精度との関係を実験により検討する。また、初期学習データとして与えるデータの質、量について本稿では触れなかったが、これについての調査、検討も必要である。

参考文献

[1] 田村淳, 渡辺道枝, 原良憲, 笠原裕. 統計的手法による文書自動分類. pp. 1305–1306. 情報処理学会, March 1988. 第３６回全国大会.

[2] 河合敦夫. 意味属性の学習結果にもとづく文書自動分類方式. 情報処理学会論文誌, Vol. 33, No. 9, pp. 1114–1122, 1992.

[3] Makoto Iwayama and Takenobu Tokunaga. A probabilistic model for text categorization: Based on a single random variable with multiple values. 1994. ANLP'94.

[4] 湯浅夏樹, 上田徹, 外川文雄. 大量文書データ中の単語間共起を利用した文書分類. 情報処理学会論文誌, Vol. 36, No. 8, pp. 1819–1827, 1995.

[5] 亀田弘之, 藤崎博也. テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム. 情報処理学会論文誌, Vol. 28, No. 11, pp. 1103–1111, 1987.

[6] 徳永健伸, 岩山真. 重み付きＩＤＦを用いた文書の自動分類. 情報処理学会自然言語処理研究会資料 NL106-3, 1994.

[7] 渡辺靖彦, 竹内雅人, 村田真樹, 長尾真. χ^2 法を用いた重要漢字の自動抽出と文献の自動分類. 電子情報通信学会技術研究報告 NLC94-25, Oct. 1994.