DEWS2003 7-C-03

プロービングを用いた動的テキストデータベースからの新規トピック文書検出

毛利　隆軌†　　北川　博之††

† 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
†† 筑波大学電子・情報工学系　〒 305–8573 茨城県つくば市天王台 1–1–1
E-mail: †tmouri@kde.is.tsukuba.ac.jp, ††kitagawa@is.tsukuba.ac.jp

あらまし　Hidden Web サイトをはじめとして，内包するデータベースコンテンツを問合せインターフェイスを介して外部の利用者に提供する情報源が増加している．多くの情報源では，そのコンテンツは時間と共に動的に追加更新される．データベースコンテンツの変化内容を知ることは，新規トピック検出やトレンド分析等の情報利用において重要である．しかし，上記のような情報源においては，利用者がコンテンツ管理者からの特別な手助けなしに問合せインタフェースのみを用いてその変化傾向を知ることは一般に困難である．本論文では，キーワードに基づく問合せインタフェースを有しそのコンテンツが動的に追加更新されるテキストデータベースを対象に，問合せプローブと分類器を用いて，新たに追加された新規性の高いトピックを有するコンテンツを検出するための手法を提案する．また，実テキストデータを用いた実験により，本手法の有効性を評価する．

キーワード　トピック検出，分類，テキストデータベース，知識発見

Probing Dynamic Text Databases to Extract New Topic Documents

Takanori MOURI† and Hiroyuki KITAGAWA††

† Graduate School of Systems and Information Engineering, University of Tsukuba　Tennohdai 1–1–1, Tsukuba, Ibaraki 305–8573,Japan
†† Institute of Information Sciences and Electronics, University of Tsukuba　Tennohdai 1–1–1, Tsukuba, Ibaraki 305–8573,Japan
E-mail: †tmouri@kde.is.tsukuba.ac.jp, ††kitagawa@is.tsukuba.ac.jp

Abstract　There are many information sources which provide their database contents through query interfaces. Hidden Web sites are typical examples. Usually, their database contents dynamically change, new documents on emerging topics being appended. In applications like topic detection and trend analysis, we want to discover newly emerging contents in the databases. However, it is very difficult for ordinary users to detect them only through the query interfaces without support by the database contents administrators. In this paper, we propose a novel method to automatically discover such contents. The proposed method generates biased query probes using a classifier to be issued to a given text database with a keyword-based query interface. They are focused to extracting documents on newly emerging topics. We evaluate its effectiveness with preliminary experiments.

Key words　Topic Detection, Classification, Text Database, Knowledge Discovery

1. はじめに

現在，インターネット上には問合せインタフェースを介して様々なデータベースコンテンツを提供する情報源が存在している．Hidden Web サイト等はそのような情報源の代表的な例である．インターネットが情報流通の基盤となった今日では，これらの情報源が内包するコンテンツは，社会における関心事や情報ニーズを分析する際の手がかりとなる貴重な資源である．特に，新規性の高いトピックの検出やトレンドの分析等の知識発見応用においては，そのコンテンツの時間的変化傾向を知ることが重要となる．

しかし，一般の利用者がそのコンテンツアクセスに利用可能な手段は，通常，キーワードに基づく問合せインタフェース等の単純なものに限られており，利用者自身が問合せ条件を工夫

して新規性の高いコンテンツを抽出することは一般に非常に困難である．データベースコンテンツ全体をダウンロードできるような状況の場合には，以前のスナップショットと現在のスナップショットを直接比較分析することで変化傾向を知ることが可能である．しかし，このような手段が全ての情報源に適用できる訳ではない．また，それが可能であっても，大量のコンテンツをダウンロードし比較分析するための効率的な手段が必要となる．

本論文では，テキストデータベースが提供する通常のキーワードに基づく問合せインタフェースのみを利用して，新規性の高いコンテンツ（文書）を重点的に抽出するための手法を提案する．提案方式のポイントは，新規性の高い文書を抽出するのに向いた問合せ用のキーワードをどのように特定するかという点である．本論文では，３種類の方法について実験により比較検討を行い，その中の１つが特に有効性が高いことを示す．

以下の 2. 章において関連研究について述べる，3. 章では本研究における提案手法を述べる．4. 章では Reuter の記事データを用いた実験について述べ．5. 章では CNN ニュースデータより時間的順序を考慮した実験について述べる．6. 章では，より実際のデータに近いものとして毎日新聞の記事を用いた実験を行いその実験について述べる．最後にまとめと今後の課題について述べる．

2. 関 連 研 究

本研究が対象とする Hidden Web サイト等のコンテンツの概要を，キーワードに基づく問合せインタフェースのみを用いて抽出するための研究が最近いくつか行われている [1] [2]．これらの方法では，情報源に対して問合せプローブ (query probe) と呼ぶ問合せを多数発行し，サンプル文書を獲得する．これらのサンプル文書から情報源が内包するデータベースのコンテンツを推定する．また，サンプル文書に出現した語やその出現頻度をまとめたものをコンテンツサマリと呼び，当該データベースコンテンツの一種のプロファイルとして用いる．これらの研究は，情報源のコンテンツのある時点でのスナップショットのプロファイルを問合せプローブを用いて獲得することを目的としている．本論文で提案する手法では，3. 章に述べるように，初期プロービングと diff プロービングの２段階のプロービングを行う．初期プロービングは，基本的には上記の手法に基づくものであるが，diff プロービングにおいては，新規性が高い文書を抽出するための問合せである diff プローブを発行する点が特徴である. [1] [2] で提案されているようなプロービングを２回行い，それぞれで得られるサンプル文書やコンテンツサマリを比較することでコンテンツの変化傾向を分析する方法も考え得る．しかし，本提案の diff プロービングに比べて従来のプロービングには多くの問合せプローブの発行が必要なことや，コンテンツの部分的な変化を多数のサンプル文書や全体的なコンテンツサマリの中から見出すのは容易でないといった問題点がある．

新規性の高いトピックの検出に関しては，これまでトピック検出等の領域で多くの研究が行われている [6]．これらでは，ニュースストリーム等から新規性の高いトピックを自動的に検出する方法が検討されている．これらの研究では到着するデータコンテンツを全て直接的に分析対象とすることが可能な状況を想定している．本研究は，Hidden Web サイト等，問合せインタフェースを介してのみコンテンツの抽出が可能な情報源を対象としており，この点で従来のトピック検出等に関する研究が想定している環境とは大きく異なる．

3. 提 案 方 式

文書群をコンテンツとし，キーワードに基づく問合せインターフェイスをもつテキストデータベース db が存在するものとする．問合せ結果は何らかの基準でランク付けされて返されるものとする．２つの時刻 t_1, t_2 $(t_1 < t_2)$ における db のスナップショットを $\mathrm{db}(t_1)$, $\mathrm{db}(t_2)$ とする．本論文では，db が処理可能な問合せを発行することにより，$\mathrm{db}(t_2)$-$\mathrm{db}(t_1)$ の文書をより多く抽出するための手法を提案する．

提案手法は，次の３つのステップからなる．

Step 1: 初期プロービング

時刻 t_1 において実行される．**初期プローブ**と呼ぶ問合せを情報源に発行することを，n_1 件のサンプル文書 (**初期サンプル文書**) を取得するまで繰り返す．

Step 2: 分類器の作成

Step 1 で取得した n_1 件の初期サンプル文書を正例として分類器を作成する．この分類器は，与えられた文書が正例の文書群とどの程度類似しているかを判定し，正例の文書群と同じクラスに属するか属さないかを判定可能なものとする．

Step 3: diff プロービング

時刻 t_2 において実行される．**diff プローブ**と呼ぶ問合せを情報源に発行する．得られた文書を Step 2 で作成した分類器にかけ，正例と同じクラスに属さないと判定された文書のみを**抽出文書**とする．抽出文書数が n_2 件となるまで，この操作を繰り返す．

以下に，各ステップのより詳細について説明する．

3.1 初期プロービング

初期プロービングの手法は, [1] [2] で用いられているプロービング手法と同様である．辞書データが利用可能であるものとし，次の 3 つの手順で行う．

（1） 語 w を選択し (詳細は下記)，データベースに w のみをキーワードとする問合せを発行する．

（2） 問合せ結果から上位 k 件の文書を取得する．

（3） 取得した文書数が n_1 に達した場合終了する．それ以外の場合は手順 (1) に戻る．

手順 (1) での語 w の選択の方法は，最初は辞書からランダムに 1 語を取り出す．2 回目以降は，辞書からランダムに取り出す方法 (RS-Ord) と，取得した文書内の語からランダムに取り出す方法 (RS-Lrd) が挙げられており，一般的に後者の方が有効であるが示されている [2]．本研究では RS-Lrd を用いる．

3.2 分類器の作成

分類器として，本研究では One-Class Support Vector Machine(SVM) [4] を用いる．初期プロービングにおいて取得した

初期サンプル文書に不要語除去や語幹抽出を行った後，d 次元の特徴ベクトルを作成する．特徴ベクトルには，初期サンプル文書群全体において出現頻度が高い d 個の語を用いる．初期サンプル文書から得られた全ての特徴ベクトルを正例として SVM に学習させることで，分類器の作成を行う．

3.3 diff プロービング

diff プロービングは以下の 3 つの手順で行う．

（1） 語 w を選択し (下記参照)，データベースに w のみをキーワードとする問合せ (diff プローブ) を発行する．

（2） 問合せ結果から上位 k 件の文書 (**候補文書**) を取得する．

（3） (2) で取得した k 件の候補文書を Step2 で作成した分類器にかけ，正例と同じクラスでないと判定された文書を抽出文書に加える．抽出文書数が n_2 に達した場合終了する．それ以外の場合は手順 (1) に戻る．

手順 (1) における語 w の選択は，最初は初期プロービングと同様に，辞書からランダムに 1 語選ぶものとする．2 回目以降は，新規性の高い文書を選択するため以下の 3 つの方法を考える．

方法 1．抽出文書に含まれる語からランダムに取得する．

方法 2．抽出文書に含まれる語からランダムに選択するが，分類を行った際，抽出文書に含めるべきでないと判断された候補文書に含まれていた語は除く．

方法 3．抽出文書に含まれる単語からランダムに選択するが，初期サンプル文書に含まれていた語は除く．

3.4 異種性の高いトピックを複数含む場合の分類器作成法

上記の分類器を作成する際，初期サンプル文書群全体を 1 クラスとして扱っている．このため，db(t_1) に複数のトピックが混在している場合には分類の精度が落ちる可能性がある．その場合の対処法として次の方法が考えられる．

クラスタリングにより，初期サンプル文書群をトピック毎に分類する．次に各クラスタに対して上記の分類器を作成する．diff プロービングにおける分類時は，候補文書を各クラスタの分類器にかけ，どのトピックにも属さないと判断された文書のみを抽出文書とする．

4. Reuter の記事データを用いた実験

4.1 実 験 内 容

実験対象の文書データとして Reuters-21578 [5] を使用した．このデータはあらかじめいくつかのトピックに分類されている．このトピック分けされているデータから 2 種類のデータベースを構築した．また，テキストデータベースの問合せ処理は，tf・idf 法を用いた余弦尺度によるものとした．分類に使用する SVM のライブラリとして [3] を用いた．SVM のカーネルやパラメータに関してはデフォルトの設定をそのまま使用した．

実験 1 トピック “小麦” に属する文書 306 件を db(t_1) とし，それにトピック “コーヒー” の文書 30 件を新規文書として加えて db(t_2) とした (図 1)．db(t_2) 内の新規文書の割合は 8.9%となる．抽出文書数 n_2 を一定とし，候補文書数，抽出文書中の新規文書 (db(t_2)-db(t_1) 中の文書) の割合を調べた．diff プロービングに関しては先に挙げた 3 種類の方法に対して評価を行った．

実験 2 トピック “コーン” に属する文書 251 件，“小麦” に属する文書 306 件，“砂糖” に属する文書 176 件の文書からなるデータベースを db(t_1) とする．この文書群は “小麦” と “コーン” の両方に属する文書が 83 件，“小麦” と “砂糖” の両方に属する文書が 3 件，“コーン” と “砂糖” の両方に属する文書 7 件となっており，合計で 640 件からなる．それにトピック “コーヒー” に属する文書 68 件を新規文書として加えて db(t_2) とした (図 2)．db(t_2) 内の全文書数は 708 件であり，新規文書の割合は 9.6%となる．これらに対して実験 1 と同様の測定を行った．ただし，本実験では db(t_1) に 3 種類のトピックの文書が混在するため，以下の 2 つの場合について実験した．

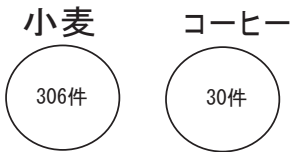

図 1　実験 1 に用いたトピックと文書数

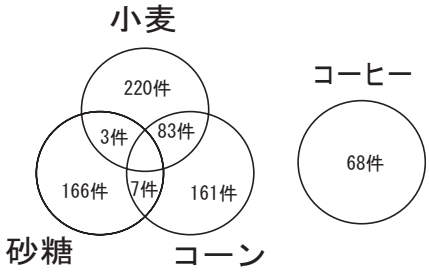

図 2　実験 2 に用いたトピックと文書数

実験 2-1 初期サンプル文書群全体を 1 クラスとして扱い分類器を作成した．

実験 2-2 初期サンプル文書群を 3.4 節の手法を用いてトピック毎に分類し各トピックに対して分類器を作成する．本来なら初期サンプル文書群をクラスタリングして，トピック毎に分類する必要があるが，本実験ではクラスタリングが理想的に行われると仮定し，初期サンプル文書のトピックラベルに基づき 3 つのクラスタに分類する．分類した 3 つのクラスタに対してそれぞれの分類器を作成した．

実験 3 実験 2-1 で diff プロービングの方法 3 を用いた場合について，抽出文書数 n_2 の値を変化させてみた．

パラメータの設定

プロービングを行う際に取得する文書数 k は 4，初期サンプル文書数 n_1 は実験 1 では 100，実験 2 では 300 とした．これらの k や n_1 の値は文献 [2] における実験結果の考察に基づく．また，分類器の生成時の特徴ベクトルの次元 d は 10 とした．これは文献 [4] における実験結果において，d=10 において最も良い結果が得られていることによる．抽出文書数 n_2 の値は，実験 1，2 では 30，実験 3 では 10，30，50 とした．

4.2 実 験 結 果

図 3，4，5，6，7 に実験結果を示す．各図は，10 回実験した結果の平均を表している．図 3 から図 6 までは，棒グラフにおける左側は候補文書数，右側は抽出文書中の新規文書数を表している．折れ線グラフは抽出文書数中の新規文書数の割合を示

している．図 7 は実験 1 と実験 2 で，初期サンプル文書を n_1 件取得するために発行した初期プローブの回数と抽出文書を n_2 件取得するために発行した diff プローブの回数を示した．

diff プロービングの 3 つの方法の中では，方法 3 が最も良い結果を示している．方法 3 における新規文書の割合は，実験 1 では 41%，実験 2-1 では 50%，実験 2-2 では 46%となっている．テキストデータベース内に存在する新規文書の割合を考えると，ランダムにデータベースをサンプリングした場合の約 4〜5 倍の割合で新規文書を取得することができていると言える．実験 3 において抽出文書数を変更した場合も，新規文書の抽出割合は抽出文書数 10 件では 55%，30 件では 50%，50 件では 55%であるので，大きな変化は見られなかったと言える．また方法 3 の候補文書数が方法 1 と方法 2 と比べて少ないことがわかる．一番違いが大きいのは実験 1 であり，方法 1 での候補文書数は 49.6 件，方法 2 では 44.1 件であるのに対して，方法 3 の候補文書数は 38.7 件である．取得した候補文書の中により多く抽出文書と判断される文書つまり新規文書であると判断される文書が多くが含まれ，抽出文書に含まれるべきでないつまり新規文書ではないと判断される文書が少ないということである．これは diff プロービングによって取得する文書がより新規文書であると判断される文書を取得していることを示している．さらにその抽出文書中に多くの新規文書が含まれているのでより効率が良いことも言える．

一方，方法 1 と 2 は実験 1 の結果では新規文書割合が 25.7%と 25%となっている．実験 2-1 では 8%と 8%，実験 2-2 では 14%と 19%となっている．全体の文書に対する新規文書の割合はもともと約 10%であるので，データベースからランダムに文書をサンプリングした場合でも，新規文書が含まれる割合は約 10%である．方法 1 と 2 は 1 つのトピックに対しては有効であるが，複数のトピックが混在する場合にはランダムにサンプリングした場合とあまり変わらないと言える．

1 つの分類器を作成する場合と複数の分類器を作成する場合では，方法 1 では 8%から 19%に，方法 2 では 8%から 14%にあがった．方法 1 と方法 2 では良い結果を示している．方法 3 においては 50%から 46%と少し精度が落ちたことになる．

図 7 より，プローブの発行回数に関しては，実験 1 では初期プローブが 81.1 回，diff プローブの方法 1 では 37.7 回，方法 2 では 25.9 回，方法 3 では 89.2 回である．実験 2 での初期プローブは 361.4 回である．実験 2-1 では diff プローブの方法 1 では 29.1 回，方法 2 では 28.8 回，方法 3 では 163.4 回である．実験 2-2 では diff プローブの方法 1 では 19.7 回，方法 2 では 18.7 回，方法 3 では 118.7 回である．いずれの場合でも，従来のプロービングより diff プロービングを用いることでプローブ数を削減できていると言える．ただし，方法 1 と方法 2 と比べて方法 3 では回数が大幅に増えてしまっている．候補文書数が少ないのに，プローブ数が増えていることを考えると，方法 3 ではすでに取得した文書を取ってくることが多いと考えられる．よって問合せ語 w の選択方法について考える必要がある．

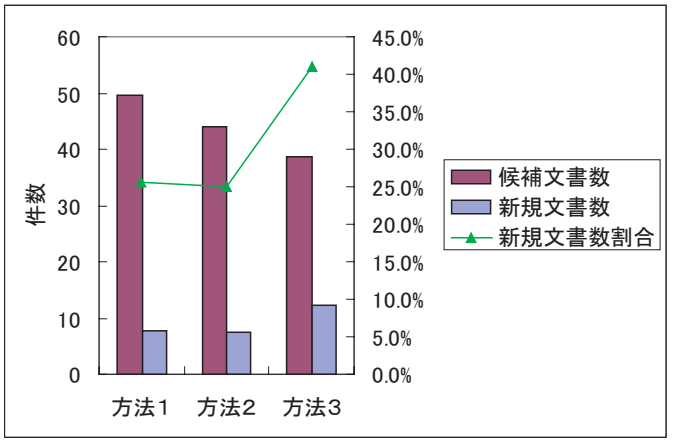

図 3　実験 1 の結果

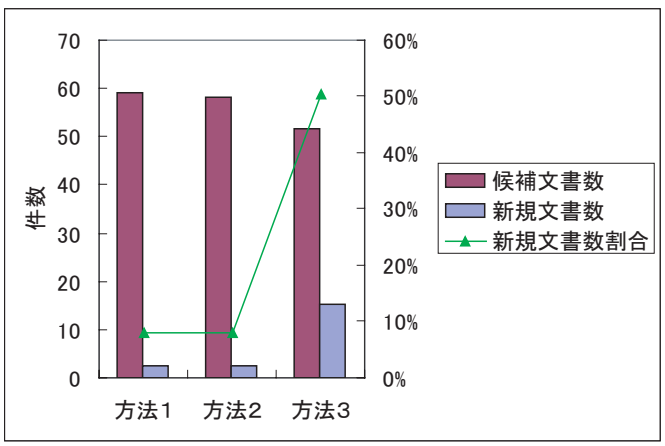

図 4　実験 2-1 の結果

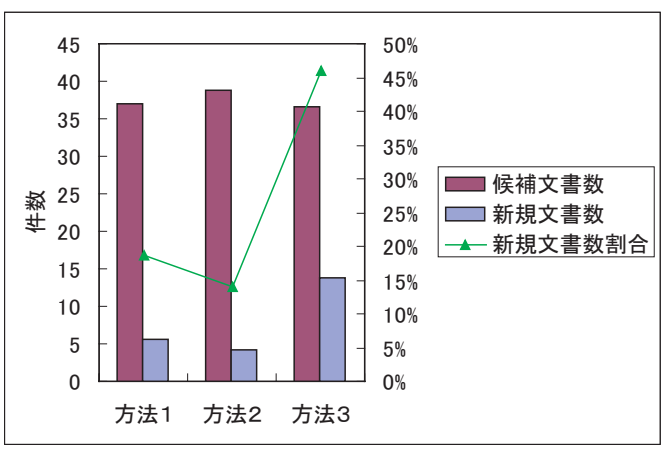

図 5　実験 2-2 の結果

5.　CNN ニュースデータを用いた実験

5.1　実験内容

利用したのは 1998 年の Topic Detection and Tracking(TDT)Phase 2 [7] で使われたデータであり，これは CNN Headline News のほか New York Times や AP 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集録したコーパスである．集録されたニュース記事の一部にはト

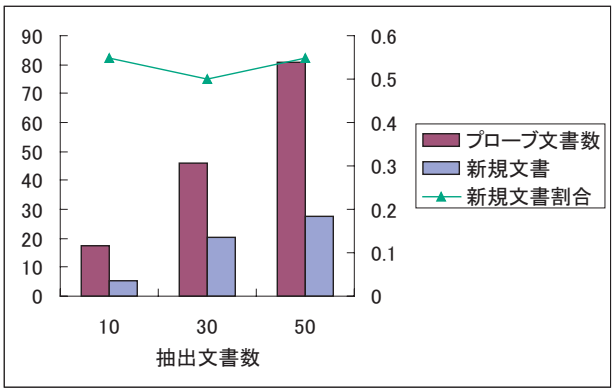

図 6　実験 3 の結果

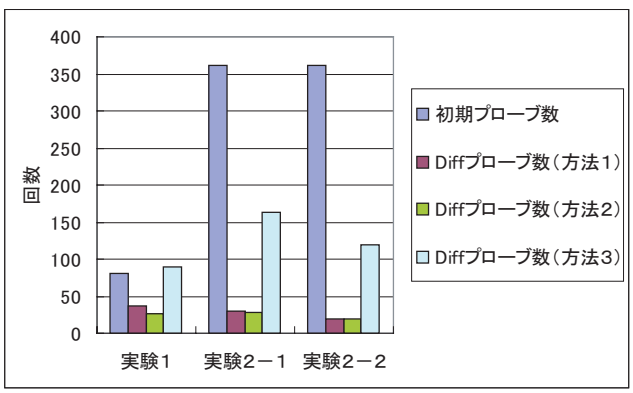

図 7　プローブ回数

Topic ID	トピック名
TP_1	アジア経済危機
TP_2	アラバマ病院爆破事件
TP_3	ローマ法王のキューバ訪問
TP_4	長野オリンピック
TP_5	フロリダのトルネード被害
TP_6	Diane Zamora への有罪判決
TP_7	Oprah Winfrey に対しての訴訟
TP_8	Gene McKinney 軍曹の性的不品行に対する公判
TP_9	スーパーボール
TP_{10}	バイアグラ

表 1　CNN データのトピックとデータ数

ピック付けおよび記事とトピックとの適合の具合（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．ここではトピックと完全に適合するニュース記事を選び，その中から 10 個のトピックを選んだ（表 1）．

またこれらの記事には日付が付けられており，TP_1 と TP_2 の全ての記事は 1 月から 6 月までの範囲にあり．TP_3 から TP_9 までの全ての記事は 1 月から 3 月までの範囲にある．また．TP_{10} は 4 月から 6 月までの記事．各記事の日付の情報を基にデータベースの構築を行った．テキストデータベースの問合せ処理や，分類器の作成方法は 4. 章と同様である．

実験 1 つの記事を 1 文書として扱い，これらの文書のうち 1 月から 3 月の記事 475 件の文書を $\mathrm{db}(t_1)$ とし，4 月から 6 月までの記事 94 件の文書を加えた計 569 件の文書を $\mathrm{db}(t_2)$ とした (表 2)．本実験では $\mathrm{db}(t_1)$ には現れず $\mathrm{db}(t_2)$ のみに現れる TP_{10} に属する文書だけを新規文書として扱うこととする．よって TP_{10} の文書は 48 件であるので新規文書は全体の 8.4%となる．4. 章より diff プロービングの方法としては方法 3 が優れるので，方法 3 についてのみを実験対象とする．初期サンプル文書群全体を 1 クラスとして扱い 1 つの分類器を作成した場合と，トピックラベルに基づき 9 つのクラスタに分類し，各クラスタに対して分類器を作成した場合について実験を行った．

Topic ID	更新前の文書数	更新後の文書数
TP_1	55	90
TP_2	62	73
TP_3	35	35
TP_4	81	81
TP_5	36	36
TP_6	23	23
TP_7	59	59
TP_8	91	91
TP_9	33	33
TP_{10}	0	48

表 2　実験に用いたデータ数

パラメータの設定

4. 章の実験 2 と同様に，プロービングを行う際に取得する文書数 k は 4，初期サンプル文書数 n_1 は 300 件とした．分類器の生成時の特徴ベクトルの次元 d は 10 とした．抽出文書数 n_2 の値は 30 とした．

5.2　実 験 結 果

図 8 に実験結果を示す．棒グラフにおける左側は候補文書数，右側は抽出文書中の新規文書数を表している．折れ線グラフは抽出文書数中の新規文書数の割合を示している．このグラフは 10 回実験した結果の平均を表している．

新規文書割合は，1 つの分類器を作成した場合は 34.7%，複数の分類器を作成した場合は 50%である．新規トピックの文書 TP_{10} の全体に対する割合は 8.4%なので，ランダムにサンプリングする場合と比べて，1 つの分類器を作成した場合は約 4 倍，複数の分類器を作成した場合は約 6 倍の精度で新規文書を抽出できている．

複数のトピックを含む場合の分類器の作成方法については，次のような違いが見られた．複数の分類器を作成した方が，候補文書数が少なく，新規文書の割合が高い．問合せ方法が同じであることを考慮すると，分類器の精度が上がり，候補文書中に存在する本来新規文書であるものを正例と間違って判断することなく正確に新規文書を負例と判断して抽出していると言える．また 4. 章の実験 2 と比べると，トピック数が多い場合には，1 つの分類器を作成するより複数の分類器を作成した方が新規文書数の割合が上がると言えることが確認できる．

6.　毎日新聞データを用いた実験

6.1　実 験 内 容

対象データは毎日新聞（2001 年発行分）の国際面に記載されていた日本語の記事の 1 月から 9 月までのデータを使用した．

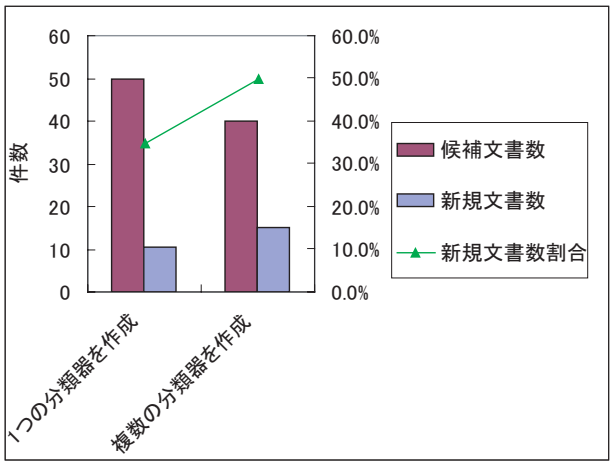

図 8　実験の結果

テキストデータベースの構築は Namazu [9] を用いた．問合せ処理は Namazu のスコア値によるランク付けとした．Namazu では問合せ結果を日付情報でソートでき，それを用いると新規文書の検出には有効であると考えられる．しかし Hidden Web が常に日付情報でソートできるとは限らないため，本実験ではスコア値によるランク付けを用いた．問合せ語の抽出や特徴ベクトルの作成には，日本語形態素解析システム Chasen [8] を用いて単語分割，形態素解析を行った．本実験では日本語を扱うため，不要語除去や語幹抽出などは行っていない．diff プロービングには方法 3 を用いた．特徴ベクトルの次元数や，特徴ベクトルや問合せキーワードの作成に用いる語の抽出のための品詞の選択方法を変化させて評価した．

実験 1 2001 年の 1 月から 8 月までのデータ 6751 件の文書からなるデータベースを db(t_1) とし，db(t_1) に 2001 年 9 月のデータ 780 件を入れた 7531 件の文書からなるデータベースを db(t_2) とした．初期サンプル文書群全体を 1 クラスとして扱い分類器を作成した．特徴ベクトルの生成に用いる品詞を固定して，特徴ベクトルの次元数を変化させた．さらにいくつかの次元数について抽出された新規文書の内容について調べた．

実験 2 実験 1 と同様のデータベースを用いた．初期サンプル文書群全体を 1 クラスとして扱い分類器を作成した．特徴ベクトルの次元数を固定し，扱う品詞の種類を変化させた．さらに抽出された新規文書の内容について調べた．

パラメータ設定

プロービングを行う際に取得する文書数 k は 4，初期サンプル文書数 n_1 は 1000 件とした．抽出文書数 n_2 の値は 50 とした．

実験 1 で扱う品詞は「名詞・動詞・形容詞」とし，次元数 d を 10，50，100，300，500 と変化させた．次元数 d が 10，100，500 について新規文書の内容について調べた．実験 2 では扱う品詞を「名詞」,「名詞・動詞」,「名詞・動詞・形容詞」とし，次元数 d を 100 とした．

6.2　実 験 結 果

図 9，10，11，12 に実験結果を示す．それぞれ 10 回実験した結果の平均を表している．

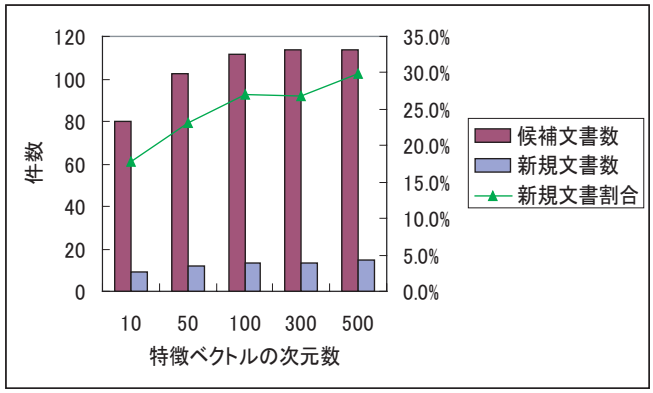

図 9　特徴ベクトルの次元を変化させた実験結果

図 9，10 においては棒グラフにおける左側は候補文書数，右側は抽出文書中の新規文書数を表している．折れ線グラフは抽出文書数中の新規文書数の割合を示している．図 9 より特徴ベクトルの次元数を増加させると，新規文書の割合も 17.8%，23.2%，27%，26.8%，29.8%と増加している．データベースにおける新規文書数の割合が 10.4%であることを考えると，どの場合においても，ランダムにサンプリングする場合より優れている．品詞の選択方法を変化させた場合の新規文書の割合は「名詞」の時が 13.4%,「名詞・動詞」の時が 12.8%,「名詞・動詞・形容詞」の時が 13.5%であったのでそれほど大きな変化が見られなかった．

図 11，12 においては棒グラフにおける左側は抽出文書中の新規文書数，右側は抽出文書中の「アメリカ同時多発テロ」や「タリバン政権」に関する文書数を表している．折れ線グラフは抽出文書数中の「アメリカ同時多発テロ」や「タリバン政権」に関する文書の割合を示している．実験 1 において抽出文書のうち「アメリカ同時多発テロ」と「タリバン政権」に関連する記事の割合が次元数が 10 次元では 4.4%，100 次元では 14.4%，500 次元では 18.8%となっている．これらの記事が実際に取り上げられる 9 月 12 日以降の文書数は 494 件である．また他のトピックの文書も含まれることを考えると,「アメリカ同時多発テロ」や「タリバン政権」に関する文書はデータベース全体の 6.6%以下であると考えられる．よって次元数 500 の場合は,「アメリカ同時多発テロ」や「タリバン政権」に関する文書の抽出に関しては，ランダムにサンプリングを行う場合の，約 3 倍以上の効率で該当文書を抽出できていると言える．また，実験 2 においても抽出文書中の「アメリカ同時多発テロ」や「タリバン政権」に関する文書の割合は「名詞」では 15.4%,「名詞・動詞」では 15.8%,「名詞・動詞・形容詞」では 14.4%であった．このように品詞を変えた場合でも，抽出文書の内容に大きな変化は見られなかった．

7.　まとめと今後の課題

本研究では，動的にコンテンツが追加更新されるテキストデータベースから新規性の高い文書を抽出するための手法を提案した．また 3 種類の diff プロービングを実験により比較し，方法 3 の有効性が高いことを示すことができた．時間的順序を

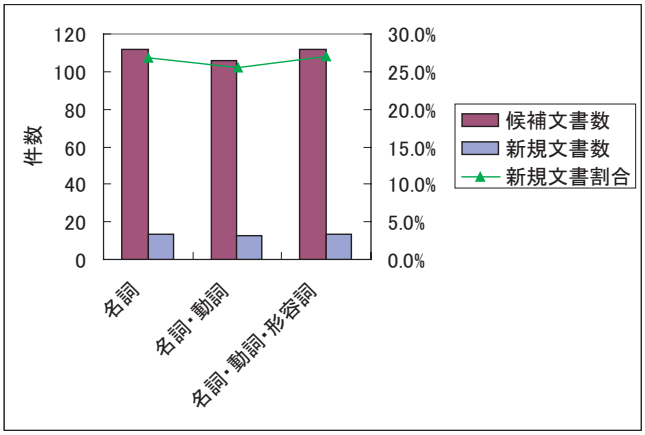

図 10　品詞を変化させた実験結果

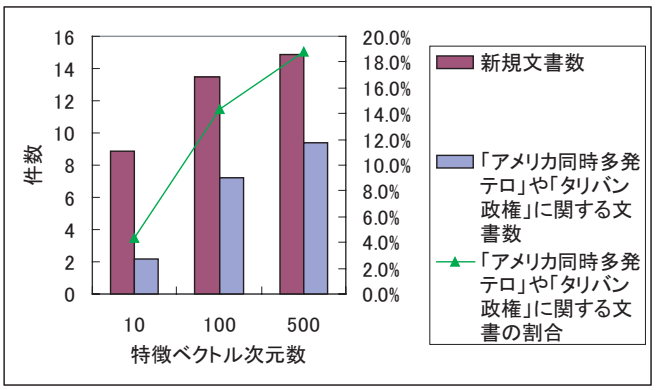

図 11　特徴ベクトルの次元を変化させた場合における「アメリカ同時多発テロ」や「タリバン政権」に関する文書

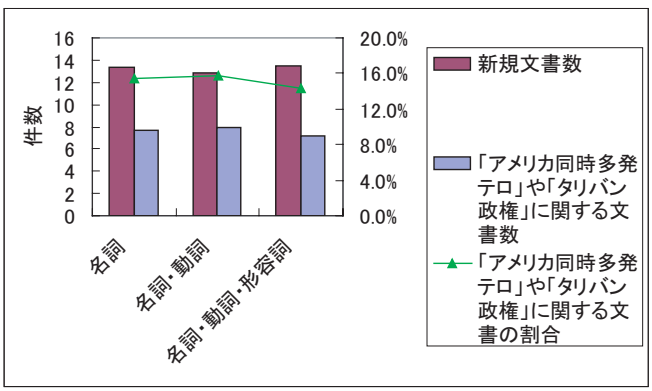

図 12　品詞を変化させた場合における「アメリカ同時多発テロ」や「タリバン政権」に関する文書

考慮したデータに対しても有効性を示すことができた．新聞記事を用いた大規模なデータに対して，新規トピックを含む文書の抽出を行うことができた．

今後の課題として，問合せ語の選択方法，分類器の精度の向上がある．また実在する Hidden Web サイトを対象とした実験が挙げられる，さらに本手法で得た抽出文書から新規トピックそのものを抽出する方法についても検討が必要である．

また本手法は，複数のテキストデータベースコンテンツの差分情報の検出等にも用いることができると考えられる．そのような視点からの検討も今後必要である．

謝　　辞

本研究の一部は，日本学術振興会科学研究費基盤研究 (B)(12480067) による．

毎日新聞 2001 年版の使用に関して，記事データの研究利用許諾をいただいた毎日新聞社に感謝いたします．

文　　献

[1] J. Callan and M. Connell. Query-Based Sampling of Text Databases. *ACM TOIS*, 19(2) 2001

[2] Panagiotis G. Ipeirotis and Luis Gravano. Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection. *Proc. 28th VLDB Conf.*, 2002.

[3] LIBSVM – A Library for Support Vector Machines http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[4] Larry M.Maevitz and Malik Yousef. One-Class SVMs for Document Classification.

[5] D.Lewis. Reuters-21578 text categorization test collection. http://www.research.att.com/ lewis,1997.

[6] Topic Detection Task. http://www.nist.gov/speech/tests/tdt/tasks/detect/htm.

[7] 1998 Topic Detection and Tracking Project (TDT-2) http://www.nist.gov/speech/tests/tdt/tdt98/

[8] Morphological Analyzer Chasen. http://chasen.aist-nara.ac.jp/

[9] 全文検索システム Namazu. http://www.namazu.org/