DEWS2004 6-B-02

Hidden Webサイトからの 新規トピック文書抽出におけるプロービングの効率化

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フェー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. 我々は，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から，新たに追加された新規性の高いトピックを有する文書を抽出するための手法を提案してきた．しかし，従来の手法では一度の問合せで多くの新規トピック文書を抽出することが難しく，多くの新規トピック文書を抽出するためには，問合せ数が多くなってしまう問題があった．そこで本論文では，一度の問合せで多くの新規トピック文書を抽出し，効率よく新規トピック文書を抽出するための改善した問合せ手法について検討を行う．また，実テキストデータを用いた実験により，本手法の有効性を示す.

キーワード　トピック検出，テキストデータベース，知識発見，情報検索

Efficienct Probing for Extracting New Topic Contents from Hidden Web Sites

Takanori MOURI† and Hiroyuki KITAGAWA††

† Graduate School of Systems and Information Engineering, University of Tsukuba
†† Institute of Information Sciences and Electronics, University of Tsukuba
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. We proposed a method to automatically discover such content. The proposed method generates biased query probes using a classifier to be issued to a given text database with a keyword-based query interface. In this paper, we improve a method for more efficient probing. We evaluate its effectiveness with preliminary experiments.

Key words　Topic Detection, Text Database, knowledge discovery, Information Retrieval

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．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が，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．

我々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してきた [2]．その手法は，更新する前のコンテンツの情報を取得して，そのコンテンツの情報から分類器を生成し，データベースが更新された後に，その分類器と問合せを用いて新規性の高い文書を抽出するものである．しかし，従来の手法では一度の問合せで，多くの新規トピック文書を抽出することが難しかった．結果として，多くの新規トピック文書を抽出するためには問合せ数が多くなってしまうという問題があった．そこで本論文では，一度の問合せで多くの新規トピック文書を抽出し，効率よく新規トピック文書を抽出するための改善した問合せ手法を提案する．また，提案手法の有効性を実験を用いて評価する．

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] [3]．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．[1] [3] 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，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] [8] [9]．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．しかし，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フェー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 $db(t_1), db(t_2)$ 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，$db(t_2) - db(t_1)$ の文書内の新規トピックを有する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．なお以下の示すステップは基本的には [2] で示したものと同様であるが，Step 3 における問合せ語の選択方法と，問合せ結果の内容によって取得する文書数を変化させるという点が異なる．

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: クラスタの生成

Step 1 で取得した n_1 件の初期サンプル文書に対して階層的クラスタリング手法を用いてクラスタの生成を行う．

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．得られた文書と Step 2 で生成した各クラスタとの類似度を計算してどのクラスタ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] [3] 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．

1-1 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．

1-2 問合せ結果から上位 k_1 件の文書を取得する．

1-3 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 1-1 に戻る．

手順 1-1 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，取得した文書内の語からランダムに取り出す．

3.2 クラスタの生成

クラスタの生成手法として，本研究では階層的クラスタリング手法 [5] を用いる．アルゴリズムは基本的に以下の 3 つの手順で行う．

2-1 各文書だけから成るクラスタを生成する．すべてのクラスタの組の類似度を余弦尺度を用いて計算する．

2-2 もっとも類似度が高いクラスタの組を併合する．併合によってできたクラスタと他のクラスタの類似度を計算する．

2-3 すべてのクラスタ間の類似度が閾値 θ より小さくなるまで手順 2-2 を繰り返す．

初期プロービングにおいて取得した初期サンプル文書群に不要語除去や語幹抽出の処理を行った後，$TF \cdot IDF$ の重み付けに基づいてベクトルを生成する．ある文書 d における語 t の重み $w(d,t)$ は

$$w(d,t) = tf(d,t) \cdot idf(t)$$

$$tf(d,t) = \frac{f(d,t)}{\sum_{s \in d} f(s,t)}$$

$$idf(t) = \log \frac{n_1}{df(t)}$$

と与えられる．生成したベクトルを基に類似度を計算してクラスタを生成していく．

クラスタの併合時には，クラスタ c_i と c_j を併合したクラスタ c_{ij} とクラスタ $c_k(k \neq i, j)$ との類似度は次により計算する．

$$\theta_{ij,k} = \frac{1}{2}\theta_{i,k} + \frac{1}{2}\theta_{j,k} - \frac{1}{2}|\theta_{i,k} - \theta_{j,k}|$$

すなわち，2 つのクラスタの最も類似度が小さい文書間の類似度で 2 つのクラスタの類似度を近似する．

3.3 diff プロービング

diff プロービングは以下の 4 つの手順で行う．

3-1 語 w を選択し (3.4 節参照)，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．

3-2 問合せ結果から上位 k_2 件の文書 (**候補文書**) を取得する．

3-3 取得した k_2 件の候補文書を Step2 で作成した各クラスタとの類似度を調べる．どのクラスタに対してもその文書との類似度が閾値 θ 以上にならない場合，新規トピック文書と判断してその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．

3-4 抽出文書数が n_2 に達しない場合，取得した k_2 件の候補文書中の抽出文書の割合を調べる．その割合が ε 以上の場合は，問合せ結果の次の k_2 件の文書を候補文書として取得し手順 3-3 に戻る．しかし，ε に満たない場合は手順 3-1 に戻る．

手順 3-3 において，閾値によっては 1 つの文書からなるクラスタが多く存在することとなる．クラスタ内の文書との類似度のうち最も小さい値が閾値より高いことより，文書が 1 つのクラスタが多く存在すると，クラスタリングでの併合条件が緩くなる．すると新規トピック文書であってもいずれかのクラスタと併合する可能性が高くなる．よって類似度を計算する際，初期サンプル文書群より作成したクラスタ群の内，1 つの文書のみからなるクラスタは除く．

3.4 diff プロービングにおける語の選択方法

手順 3-1 における語 w の選択は，最初は初期サンプル文書内に含まれる語を除いた辞書からランダムに 1 語選ぶものとする．2 回目以降の選択方法は，従来の提案手法 [2] では，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といった方法をとっていた．しかし，この方法では新規トピックを特徴付けるような語でも初期サンプル文書内に 1 語でも含まれていると，問合せ語の候補から外れてしまうという問題や，新規トピック文書にだけ含まれる語であってもその語を含む文書数が少なければ，一度に多くの文書を抽出できないという問題がある．よって新規性の高い文書，すなわち抽出文書の多くの文書に含まれる語で，新規トピック文書ではない初期サンプル文書などにあまり含まれない語に注目すれば，新規性の高い文書だけを一度に多く抽出でき，プローブ数の効率化が図れると考えられる．そこで，本研究では情報利得 [4] を用いて語の選択を行う．語の選択方法を次に示す．

p 件からなる抽出文書の集合を P として，n 件からなる新規トピック文書でないと判断された候補文書と初期サンプル文書の集合を N とする．文書集合 P 内の文書が持つ語 w について情報利得を調べる．文書集合 P 内に語 w を含む文書数が p_i 件，含まない文書が p_j 件, 文書集合 N 内に語 w を含む文書数が n_i 件，含まない文書が n_j 件の時，語 w の情報利得 $Gain(w)$ は

$$Gain(w) = I(p, n) - E(w)$$

$$E(w) = \frac{p_i + n_i}{p + n} I(p_i, n_i) + \frac{p_j + n_j}{p + n} I(p_j, n_j)$$

$$I(p, n) = -\frac{p}{p + n} \log_2 \frac{p}{p + n} - \frac{n}{p + n} \log_2 \frac{n}{p + n}$$

で求められる．この時，情報利得が最も大きい語 w を次の問合せ語として選択する．

4. ニュースデータを用いた実験

4.1 実験データ

実験対象の文書データとして利用したのは 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 や New York Times 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，10 件以上のニュース記事を持つトピック 51 個を用いて選んで実験を行った (表 1)．実験では 1 つのニュース記事を 1 文書として扱う．

4.2 実 験 内 容

本手法を評価するのに実際の Hidden Web サイトを用いて実験を行うのが望ましいが，実際のサイトではコンテンツ全体を取得するのは困難であり，またどの文書が新規のものであるかは分からないといった問題があるので定量的で客観的な評価が難しい．よって文書データを用いて Hidden Web サイトを構築して本実験を行った．

TDT の文書を基に実験における $db(t_1)$ と $db(t_2)$ となる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，$TF \cdot IDF$ 法を用いた余弦尺度によるものとした．

4.2.1 実 験 1

実験 1 では小規模なデータベースを用いて，本手法の性能を評価する．実験では複数のトピックが混在するデータベースに，新たなトピックを 1 つ追加した場合について調べた．閾値を変化させた時の抽出文書に含まれる新規トピック文書の割合とプローブ数と分類器の精度について調べた．

データベースを構築する際に利用したトピックは文書数が多い上位 11 トピックである (表 2)．まずそのトピックのうち TP_1 から TP_{10} に属する文書を用いて $db(t_1)$ を構築した．$db(t_1)$ の文書数は全部で 1000 件とした．次に，新規トピック文書として，TP_{11} に属する文書のうち，元のデータベースの 8%にあた

Topic ID	トピック名	全文書数
TP_1	Current Conflict with Iraq	1322
TP_2	Asian Economic Crisis	1032
TP_3	Monica Lewinsky Case	969
TP_4	1998 Winter Olympics	540
TP_5	India,A Nuclear Power?	427
TP_6	Anti-Suharto Violence	324
TP_7	Israeli-Palestinian Talks(London)	203
TP_8	Pope visits Cuba	151
TP_9	GM Strike	138
TP_{10}	Sgt. Gene McKinney	126
TP_{11}	Violence in Algeria	125
TP_{12}	Unabomber	119
TP_{13}	McVeigh's Navy Dismissal & Fight	19
TP_{14}	Upcoming Philippine Elections	41
TP_{15}	Fossett's Balloon Ride	15
TP_{16}	Casey Martin Sues PGA	56
TP_{17}	Karla Faye Tucker	48
TP_{18}	State of the Union Address	42
TP_{19}	Babbitt Casino Case	20
TP_{20}	Bombing AL Clinic	99
TP_{21}	Cable Car Crash	110
TP_{22}	China Airlines Crash	36
TP_{23}	Tornado in Florida	53
TP_{24}	Diane Zamora	30
TP_{25}	Shevardnadze Assassination Attempt	38
TP_{26}	Oprah Lawsuit	70
TP_{27}	Mary Kay LeTourneau	12
TP_{28}	John Glenn	37
TP_{29}	Superbowl '98	84
TP_{30}	David Satcher confirmed	16
TP_{31}	Quality of Life, NYC	33
TP_{32}	Grossberg baby murder	26
TP_{33}	Asteroid Coming??	31
TP_{34}	Dr. Spock Dies	15
TP_{35}	Viagra Approval	93
TP_{36}	JJ the Whale	11
TP_{37}	James Earl Ray's Retrial?	49
TP_{38}	World Figure Skating Champs	20
TP_{39}	Bird Watchers Hostage	16
TP_{40}	Race Relations Meetings	12
TP_{41}	Rats in Space!	60
TP_{42}	Tony Awards	14
TP_{43}	Nigerian Protest Violence	50
TP_{44}	Denmark Strike	15
TP_{45}	World AIDS Conference	18
TP_{46}	NBA finals	79
TP_{47}	Anti-Chinese Violence in Indonesia	36
TP_{481}	Afghan Earthquake	23
TP_{49}	German Train derails	51
TP_{50}	Puerto Rico phone strike	12
TP_{51}	Clinton-Jiang Debate	68

表 1　TDT データのトピック名と全文書数

Topic ID	トピック名	実験 1-1	実験 1-2
TP_1	Current Conflict with Iraq	100	250
TP_2	Asian Economic Crisis	100	200
TP_3	Monica Lewinsky Case	100	190
TP_4	1998 Winter Olympics	100	100
TP_5	India,A Nuclear Power?	100	80
TP_6	Anti-Suharto Violence	100	60
TP_7	Israeli-Palestinian Talks(London)	100	40
TP_8	Pope visits Cuba	100	30
TP_9	GM Strike	100	25
TP_{10}	Sgt. Gene McKinney	100	25
TP_{11}	Violence in Algeria	80	80

表 2　実験 1 で用いるトピックとその文書数

る 80 件の文書を $db(t_1)$ に加えた．このように，全文書数 1080 件からなる $db(t_2)$ を構築した．

実験 1-1

$db(t_1)$ に存在する各トピックの文書数を同じにして実験を行った．TP_1 から TP_{10} に属する文書を各トピックごとに 100 件の文書を使用して $db(t_1)$ を構築した．

実験 1-2

$db(t_1)$ の各トピックの文書数の割合を実際の分布に合わせて実験を行った．すなわち，TP_1 から TP_{10} の各トピックの文書から同じ割合ずつ抽出して，合計 1000 件の文書を用いて $db(t_1)$ を構築した．

4.2.2 実　験　2

実験 2 は大規模なデータベースを用いて，本手法の性能を評価した．実験には全てのトピックを用いて行った．全てのトピックの文書数を同じにするには，各トピックの文書数が大きく異なるので難しい．よって実験 2 では存在する文書を全て用いてデータベースを構築した．ここでは追加するトピック数として，1 つ追加した場合と複数追加した場合において実験を行った．閾値を変化させて，抽出文書に含まれる新規トピック文書の割合とプローブ数と分類器の精度について調べた．また従来の手法と本手法の，プローブによって取得した候補文書中の新規トピック文書数についても調べた．また複数トピックを追加した場合の抽出した新規トピック文書のトピック毎の抽出文書数について調べた．

実験 2-1，実験 2-2 で，新規トピック文書として用いる 5 つのトピックに属する文書を除いた，46 トピックに属する 5600 件の文書を用いてデータベース $db(t_1)$ を構築した．次に，新規トピック文書として，元のデータベースの 8%にあたる 448 件の文書を $db(t_1)$ に加えた．このように，全文書数 6048 件からなる $db(t_2)$ を構築した．

実験 2-1

$db(t_1)$ に新規トピック文書として 1 つのトピックを追加して実験を行った．新規トピックとして追加するトピックは，$db(t_1)$ の 8%にあたる 448 件以上の文書を持つトピックが必要なので TP_4 を用いた．大規模なデータベースを構築するために必要な文書数を持つトピックのうちで，最も少ない文書数を持つ TP_4 を用いた．TP_4 に属する文書を新規トピック文書として追加して $db(t_2)$ を構築した．

実験 2-2

$db(t_1)$ に新規トピック文書として 4 つのトピックを追加して実験を行った．新規トピックとして追加するデータは，元のデータベースの 8%にあたる 448 件（各トピック 112 件）の文書である．追加する新規トピックとして，112 件以上文書があるトピック TP_5，TP_7，TP_9，TP_{10} を用いた．112 件以上文書を持つトピックのうち少ない文書数を持つ方からトピックを選択すると 110 から 150 件の文書を持つトピックが存在しなくなり，実際のトピックの文書数の分布と異なってしまう．実際のトピックの文書数の分布と 4 つのトピックに属する文書を除いた時のトピックの文書数の分布が近くなるように，上記の 4 つのトピックを選択した．

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_1 は 4，初期サンプル文書数 n_1 は 300 件とした．これらの k_1 や n_1 の値は文献 [3] における実験結果の考察に基づく．多くの Hidden Web サイトが問合せ結果を 1 ページ毎に 20 件表示することより，k_2 は 20 とした．抽出文書数 n_2 の値は 40 とした．類似度を調べるための閾値 θ を実験 1 では 0.02 から 0.14 まで，実験 2 では 0.04 から 0.14 まで 0.02 刻みで実験を行った．ε は 0.5 とした．

4.3 実 験 結 果

4.3.1 実験 1 の結果

実験の結果を図 1，2，3，4，5 に示す．図 1，2，3，4 の破線は本手法の結果を示し，実線は従来の手法の結果を示す．それぞれの図は 50 回の実験結果の平均を表している．図 1，3 は抽出文書 40 件を抽出するまでに発行したプローブ数を示している．図 2，4 は抽出文書 40 件中の新規トピック文書の割合を示している．図 5 は，実験 1-1，実験 1-2 で用いた分類器の精度を示している．分類器の精度は diff プローブで取得した候補文書全体を基に調べた．実線は新規でないトピック文書を新規トピック文書として誤った割合を示し，破線は新規トピック文書を新規でないとトピック文書として誤った割合を示している．

プローブ数

図 1，3 より，従来の手法と比べて本手法ではプローブ数を大幅に削減できたことがわかる．

図 1 より，従来の手法で最も問合せ数が少ない場合である $\theta = 0.02$ の約 26 回に対して，本手法は約 5 回である．従来の手法は本手法と比べて 5 倍以上のプローブ数が必要であることがわかる．図 3 においても，本手法の方がプローブ数が少ないことが分かる．

従来の手法では，新規トピック文書だけに含まれる語が選択されることが多いが，選択された語はその語を抽出した元の文書だけにしか存在しない語であったり，問合せ結果の件数が少ない場合が多く見られた．そのため多くのプローブ数が必要となってしまった．しかし本手法は，情報利得を用いるので，新規トピック文書の多くの文書に現れて，初期サンプル文書など従来からデータベースに存在する文書にあまり現れない語を選択することができる．よって 1 回のプローブで新規性の高い文書だけを多く取得できるような問合せをすることができる．実験結果からも，本手法は少ないプローブ数で文書の抽出をすることができると言える．

この実験において従来の手法は，閾値が大きくなると問合せ数が増加することが見られる．図 5 より，閾値を大きくすることで，新規でないトピック文書を新規トピック文書と誤って判断してその文書を抽出してしまうことがある．プローブの語は抽出した文書の語から選択されので，抽出文書中に新規でないトピック文書の多く存在すると，その新規でないトピック文書中の語から選択されることが多くなり，プローブの結果として新規トピックではない文書が得られる．そのため，新規トピック文書を抽出するためにプローブ数が多くなってしまう．閾値 $\theta = 0.02$ では抽出した文書が，ほとんど新規トピック文書のみであるので，その文書から選択される語は新規トピック文書を抽出する語であることが多く，プローブ数は少なくなる．

新規トピック文書の割合

図 2，4 より，新規トピック文書の割合については，いずれの閾値 θ においても本手法と従来の手法には大きな差はなく，高い割合で新規トピック文書を抽出できていることがわかる．閾値が大きくなると新規トピック文書の割合が下がるのは，図 5 から分かるように新規でないトピック文書を新規トピック文書であると誤る割合が高くなっているからである．

図 2 より，最も新規トピック文書の割合が高い，閾値 $\theta = 0.02$ では，新規トピック文書の割合は約 0.98 であり，ほとんどが新規トピック文書であると言える．また，新規トピック文書の割合が最も悪い場合である $\theta = 0.14$ でも新規トピック文書の割合は約 0.6 となっている．これはデータベース中の新規トピック文書の割合が全文書 1080 に対して 80 件で，約 0.07 であることを考えると，ランダムにサンプリングした場合の約 9 倍の精度で新規トピック文書が抽出できていると言える．従来の手法も同様に新規トピック文書の割合は高い．よって本手法も従来の手法も有効であると言える．

分類器の精度

図 5 より，初期サンプル文書から生成した分類器の精度は実験 1-1，実験 1-2 においても大きな違いはない．実験 1-2 では，データベース中のトピックで最も多くの文書を持つ TP_1 の文書数は 250 件，最も少ない文書を持つ TP_9，TP_{10} の文書数は 25 件で文書数は 10 倍も違う．しかし，全文書数 1000 件に対して，初期サンプル文書数 300 件であることから，初期サンプル文書としてそのトピックが抽出されるだけの割合を持っている．よって，いずれのトピックの情報も十分取得できており，精度が良い分類器を生成することができたと言える．図 5 からも実験 1-1，実験 1-2 では精度の良い分類器を生成できていると言える．

4.3.2 実験 2 の結果

実験の結果を図 6，7，8，9，10，表 3，4，5 に示す．実験結果は抽出文書 40 件を抽出するまで発行したプローブ数を図 6，8 に示し，抽出文書 40 件中の新規トピック文書の割合を図 7，9 に示す．それぞれの図は 50 回の実験結果の平均を表している．破線は本手法の結果を示し，実線は従来の手法での結果を示す．図 10 は分類器の精度を示している．分類器の精度は diff プローブで取得した候補文書全体を基に調べた．実線は新規でないトピック文書を新規トピック文書として誤って抽出した割合を示し，破線は新規トピック文書を新規でないとトピック文

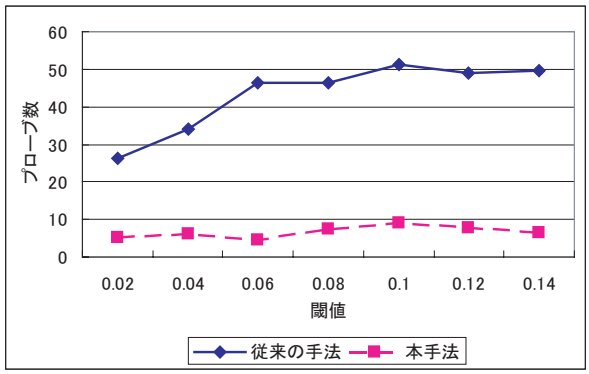

図 1　実験 1-1 の結果:プローブ数

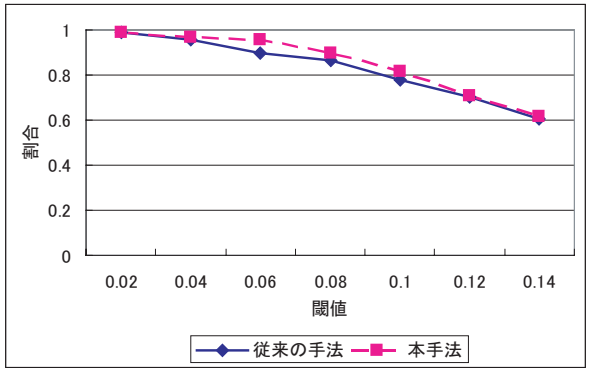

図 2　実験 1-1 の結果:新規トピック文書の割合

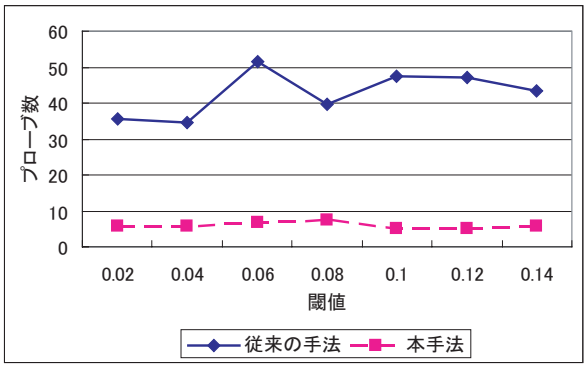

図 3　実験 1-2 の結果:プローブ数

書として誤った割合を示している．表 5 は diff プローブによって取得した候補文書中の新規トピック文書数を表している．この値も 50 回の実験結果の平均である．表 3，4 は閾値 $\theta = 0.04$ の場合の抽出文書の 40 件の中に含まれていた新規トピック文書のトピックの内訳について調べた結果である．スペースの関係上，表は 50 回行った実験のうちの 10 回の結果を示す．

プローブ数

図 6，8 より，本手法の方がプローブ数が少なくなっていることが分かる．図 6 の閾値 $\theta = 0.04$ では従来の手法の約 17.5 回に対して，本手法の約 9 回となっており，約半分になっている．それ以外の閾値においても，多くの場合でプローブ数が半分以下になっている．

図 6，8 から実験 2-1 と実験 2-2 では同様の傾向が見られる．追加したトピック数によらず，本手法では従来の手法よりプローブ数が少ないことが分かる．

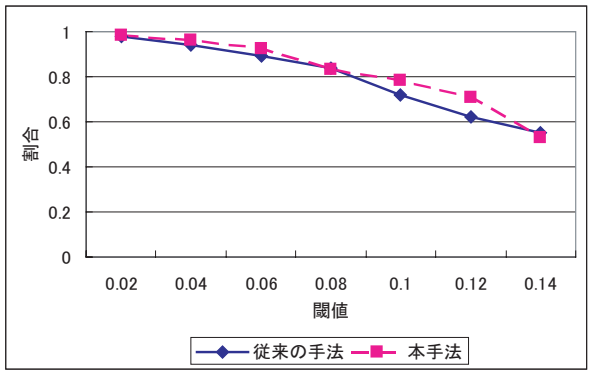

図 4　実験 1-2 の結果:新規トピック文書の割合

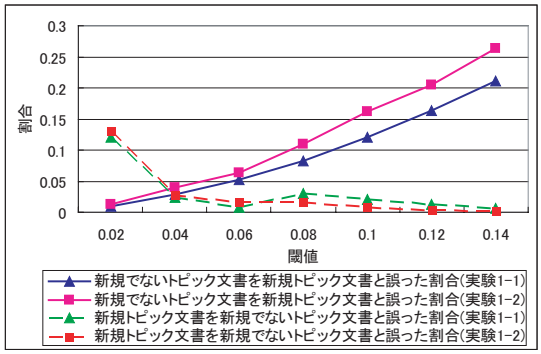

図 5　実験 1 の結果:分類器の精度

新規トピック文書の割合

抽出文書 40 件中の新規トピック文書の割合を図 7，9 に示す．図 7 より，実験 2-1 では従来の手法の場合が新規トピック文書割合が高い．図 9 より，実験 2-2 では本手法の場合が新規トピック文書割合が高い．本手法において最も良い場合は閾値 $\theta = 0.04$ の時で，新規トピック文書の割合は，実験 2-1 では約 0.4，実験 2-2 では約 0.45 の割合で取得できている．データベース中の新規トピック文書の割合が約 0.07 であることを考えると，ランダムにサンプリングして取得した場合の約 6 倍の割合で取得できていると言える．従来の手法も閾値 $\theta = 0.04$ の時が最も良く，新規トピック文書の割合は，実験 2-1 では約 0.51，実験 2-2 では約 0.42 の割合で取得できている．

実験 2-1 と実験 2-2 における従来の手法と本手法の新規トピック文書の割合の違いは次の理由からである．図 10 より，実験 2-1 と比べて実験 2-2 では，新規トピック文書を新規でないトピック文書と誤った割合が下がっている．本手法は新規トピックではないと判断されてた文書も考慮してプローブの語を選択するので，分類器が正しく働いている実験 2-2 の方がプローブによって新規トピック文書を取得できている．表 5 からも実験 2-2 は実験 2-1 と同数の文書数，またそれ以上の文書数の新規トピック文書を候補文書として取得できている．さらに実験 2-2 では分類器が実験 2-1 より正しく働いているので，新規トピック文書の割合が高くなる．

一方，従来の手法はプローブの語を選択する際に，新規トピックではないと判断された文書は考慮していないため，新規トピック文書を新規でないトピック文書と誤った割合だけが下

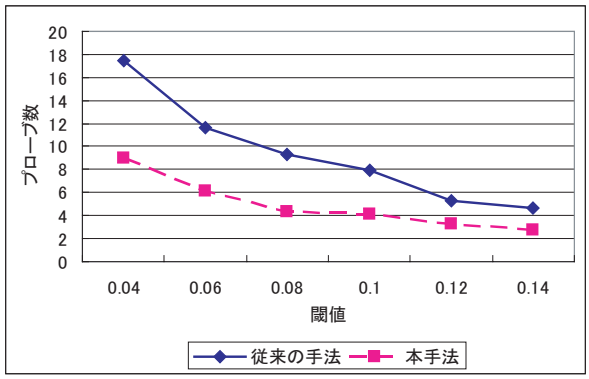

図 6　実験 2-1 の結果:プローブ数

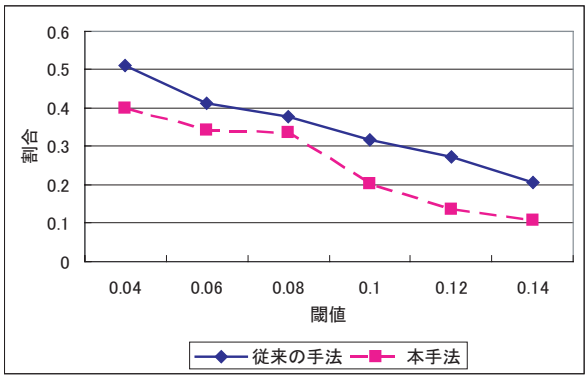

図 7　実験 2-1 の結果:新規トピック文書の割合

がっても，本手法と比べてプローブへの影響は少ない．さらに，従来の手法は 1 回のプローブで新規トピックを多く取得する方法ではないので，1 つのトピックに属する文書数が減った実験 2-2 では新規トピック文書を取得する数が少なくなっている (表 5)．結果として新規トピック文書の割合も少なくなってしまう．したがって，本手法は分類器の精度が良い場合に効率良く新規トピック文書の抽出が行え，精度が悪い場合は従来の手法の方が良いと言える．

抽出したトピック毎の数

実験 2-2 において，抽出した新規トピック文書の各トピックの抽出数を調べた．閾値 $\theta = 0.04$ の場合について調べた．50 回行った結果のうちの，10 回を表 3，4 に示す．各操作で抽出した各トピックに属する文書数と，抽出した各トピックに属する文書数の平均と標準偏差を示した．

実験においては新規トピック文書として加えた文書の数は各トピックについて同数の文書を追加している．しかし各操作において，ある 1 種類のトピックが多く抽出されて，平均的にトピックを抽出していない．その原因としては，従来の手法も本手法も問合せプローブの語が抽出された文書に含まれる語から選択されるからである．

表 3，4 より，従来の手法を用いた方が標準偏差が比較的小さい．従来の手法は，あるトピックに属する文書を多く抽出していても他のトピックに属する文書が 1 文書でも抽出されれば，ランダムにプローブの語を選択するので，複数のトピックが抽出される可能性がある．しかし，本手法はあるトピックに属する文書を多く抽出すると，そのトピックに属する文書を多く取得できるようなプローブの語を選択するので複数のトピックが抽出される可能性が低い．

しかし，従来の手法も本手法も複数のトピックを抽出できているとは言えない．よって今後はある 1 種類のトピックの文書を必要数抽出できたら，そのトピックを新規トピックではなく古いトピックとして扱うなどして分類器や，プローブの仕組みを変更するなどをして平均的に複数のトピックを取得する方法を検討する必要がある．

4.3.3　実験 1 と実験 2 の結果

プローブ数

実験 1 と実験 2 の結果より，本手法においては，いずれの場合においてもプローブ数は 10 回以下になっており，プローブ

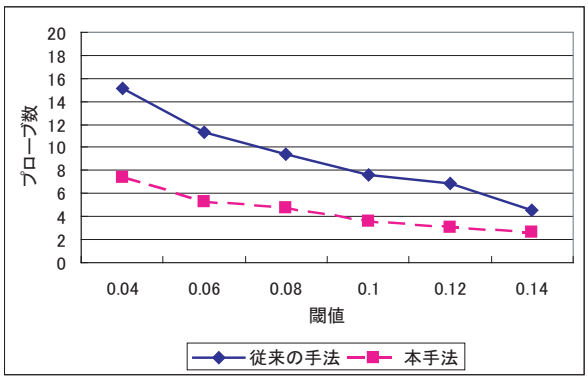

図 8　実験 2-2 の結果:プローブ数

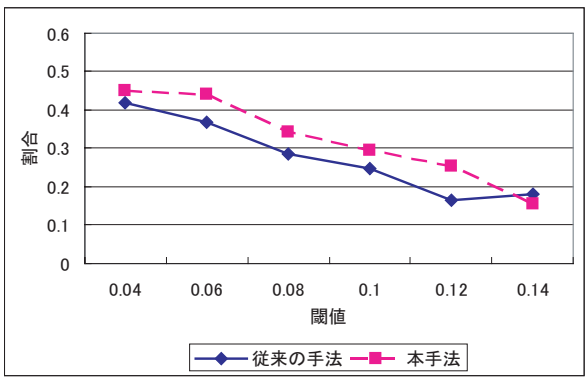

図 9　実験 2-2 の結果:新規トピック文書の割合

数の削減に関しては有効であることは言える．実験 1 と実験 2 での従来の手法におけるプローブ数の違いは次のようなことが言える．従来の手法については，ランダムに問合せ語を選択することから，1 回の問合せで多くの新規トピック文書を抽出すると言ったものではない．よって，全文書中の新規トピック文書の絶対数が少ない実験 1 では 1 回のプローブで取得できる新規トピック文書が少なく，結果としてプローブ数が増加してしまう．実験 2 では新規トピック文書の数が多いので，従来の手法でも多くの文書が抽出することができる．また実験 2 の方が分類器の精度が悪いため新規でないトピック文書を新規トピック文書と誤るので，少ないプローブ数で処理が終了する．

本手法を用いた場合は 1 回のプローブ数で多くの新規トピック文書を抽出するのでいずれの場合においても，少ない回数で抽出することが可能である．

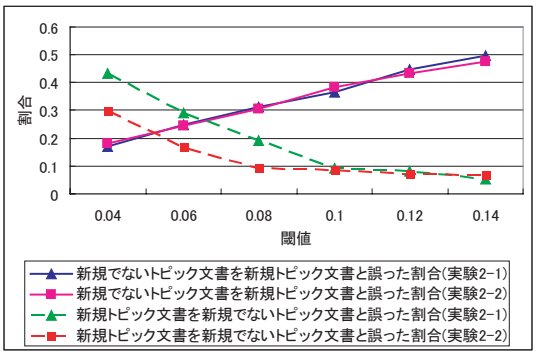

図 10　実験 2 の結果:分類器の精度

	TP_5	TP_7	TP_9	TP_{10}	合計
1 回目	5	0	0	4	9
2 回目	1	0	0	27	28
3 回目	3	5	1	0	9
4 回目	3	0	0	0	3
5 回目	1	0	0	2	3
6 回目	9	1	0	5	15
7 回目	3	0	10	11	24
8 回目	2	0	0	13	15
9 回目	0	0	14	0	14
10 回目	0	2	0	12	14
平均	5.6	0.8	2.5	7.1	13.4
標準偏差	2.6	1.5	4.8	8.1	7.7

表 3　実験 2-2 で従来の手法で抽出した各トピックの抽出数（$\theta = 0.04$）

	TP_5	TP_7	TP_9	TP_{10}	合計
1 回目	2	0	28	1	31
2 回目	0	0	1	0	1
3 回目	1	0	0	4	5
4 回目	0	2	29	0	31
5 回目	1	0	0	0	1
6 回目	0	3	2	0	5
7 回目	28	0	0	0	28
8 回目	6	2	0	0	8
9 回目	0	0	1	19	20
10 回目	34	2	0	0	36
平均	7.2	0.9	6.1	2.4	16.6
標準偏差	12.1	1.1	11.2	5.7	13.3

表 4　実験 2-2 で本手法で抽出した各トピックの抽出数（$\theta = 0.04$）

閾値	従来の手法 (実験 2-1)	本手法 (実験 2-1)	従来の手法 (実験 2-2)	本手法 (実験 2-2)
0.04	39.0	25.3	25.6	24.0
0.06	23.1	19.4	17.8	21.2
0.08	20.0	15.3	12.3	15.3
0.1	14.3	8.7	11.0	12.8
0.12	12.0	6.0	7.5	10.6
0.14	8.6	4.7	7.8	6.7

表 5　実験 2 で候補文書内の新規トピック文書数

新規トピック文書割合

実験 1 と実験 2 では，データベース中の新規トピック文書の割合は変わらないが，実験 1 と比べて実験 2 は，抽出した文書中の新規トピック文書の割合が低い．図 5，10 から，実験 2 における分類器の精度が悪いからであると言える．分類器の精度が落ちた原因は，実験 2 では初期状態のデータベースに文書数が少ないトピックが多く存在するからであると考えられる．文書少ないトピックは初期サンプル文書としてそのトピックの文書が抽出されず，初期サンプル文書が初期状態のデータベースを正しく表現することができない．したがって分類器の精度が悪くなってしまうと考えられる．この問題を解決するには，実験 2 のようなデータベースにも対応できるように，初期プロービングにおける初期サンプル文書の取得方法について全トピックを効率よく抽出する方法などの改善が必要である．

5.　まとめと今後の課題

本研究では，テキストデータベースを内包する Hidden Web サイトから新規性の高い文書を抽出する際，情報利得を用いてプロービングを行うことで効率よく新規トピック文書を抽出する方法について提案した．ニュースデータに対して実験を行い従来の手法より提案手法の効率が良いことを示すことができた．

今後の課題として，文中で述べたが，複数のトピックを効率よく抽出する方法への改善，初期プロービングにおいて多くのトピックをカバーする取得方法の検討，また対象データに日本語の文書を用い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抽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，科学研究費補助金基盤研究 (B)(#15300027), 特定領域研究 (2)(#15017207) による．

文　　献

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

[2] Takanori Mouri and Hiroyuki Kitagawa. Extracting New Topic Contents from Hidden Web Sites. IEEE International Conference on Information Technology, 2004.

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

[4] Quinlan J. R. *C4.5: Programs For Machine Learning.* Morgan Kaufmann Publishers, 1993.

[5] G.Salton. Automatic Information Organization and Retrieval, McGraw-Hill Book Company,1968.

[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] F. Walls, H. Jin, S. Sista, and R. Schwartz. Topic detection in broadcast news. In Proceedings of the DARPA Broadcast News Workshop, pages 193-198, San Francisco, CA, 1999. Morgan Kaufmann Publishers, Inc.

[9] Y. Yang, J. Zhang, J. Carbonell and C. Jin. Topic-conditioned Novelty Detection. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 688-693, 2002.