プロービングによるテキストデータベースからの新規トピック文書抽出

Probing Text Databases to Extract New Topic Documents

毛利 隆軌♥ 北川 博之♦

Takanori MOURI Hiroyuki KITAGAWA

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．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抽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．

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 experiments on real text databases.

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ニュースデータを用いた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．

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[3]．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

♥ 学生会員 筑波大学大学院システム情報工学研究科
tmouri@kde.is.tsukuba.ac.jp
♦ 正会員 筑波大学電子・情報工学系
kitagawa@is.tsukuba.ac.jp

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．

Step1: 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

Step1 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 Step2 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[6]を用い，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．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.

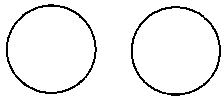

図 1　実験 1 に用いたトピックと文書数

実験 2

トピック“コーン”,“小麦”,“砂糖”に属する文書 640 件の文書からなるデータベースを db(t_1)とする．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．

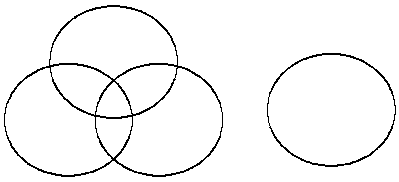

図 2　実験 2 に用いたトピックと文書数

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する．

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．

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 30 とした．

4.2　実験結果

図 3,4,5 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．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．

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 は，いずれの実験において新規文書の割合が 40～50%となっている．テキストデータベースをランダムにサンプリングした場合の 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．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．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ことを意味する．これは diff プロービングにより，新規文書であると判断される文書をより多く

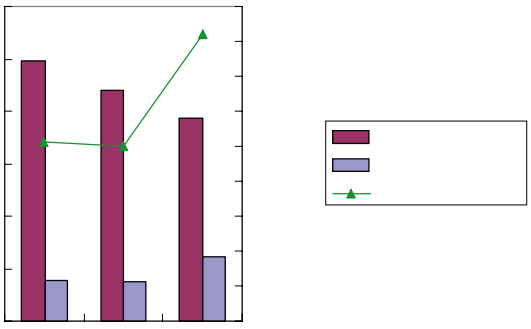

図 3　実験 1 の結果

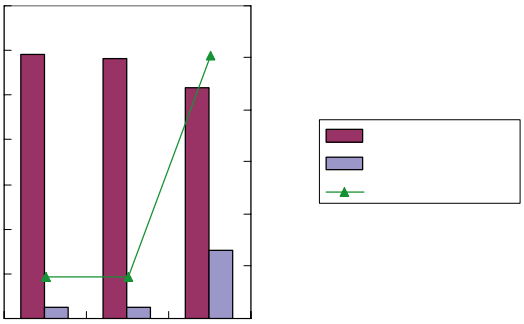

図 4　実験 2-1 の結果

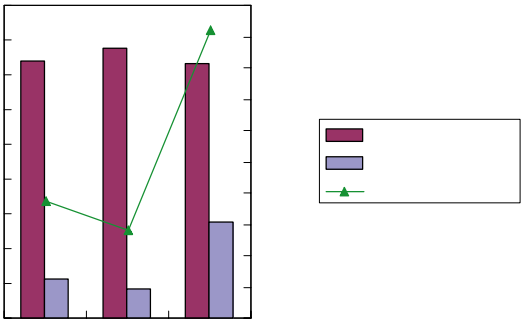

図 5　実験 2-2 の結果

取得でき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 1,2 の結果から 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いえる．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 と 2 では 2～3 倍精度が上がり，良い結果を示している．方法 3 においては少し精度が落ちたことになる．

5. CNN ニュースデータを用いた実験

5.1　実験内容

実験に利用したのは[7]のデータである．これは CNN Headline News 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 10 個のトピックと完全に適合するニュース記事を選び実験に用いた（表 1）．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，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 月までの記事である．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 $db(t_1)$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 $db(t_2)$とし

Topic ID	トピック名	初期文書数	追加数
TP_1	アジア経済危機	55	35
TP_2	アラバマ病院爆破事件	62	11
TP_3	ローマ法王のキューバ訪問	35	0
TP_4	長野オリンピック	81	0
TP_5	フロリダのトルネード被害	36	0
TP_6	Diane Zamora への有罪判決	23	0
TP_7	Oprah Winfrey に対しての訴訟	59	0
TP_8	Gene McKinney 軍曹の性的不品行に対する公判	91	0
TP_9	スーパーボール	33	0
TP_{10}	バイアグラ	0	48

表１　CNN データのトピックとデータ数

た(表 1). 本実験では $db(t_1)$には現れず $db(t_2)$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. TP_{10} の文書は 48 件であるので全体の 8.4%となる. 4 章より diff プロービングの方法 3 が優れているので，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，クラスタ毎に分類器を作成した場合について実験を行った.

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，分類器の生成時の特徴ベクトルの次元 d は 10，抽出文書数 n_2 の値は 30 とした.

5.2　実験結果

図 6 に実験結果を示す. 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を考えると，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っていることが確認できる.

6. 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した. CNN ニュースデータのような時系列的データに対しても有効性を示すことができた.

今後の課題として，より良いパラメータの設定方法，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

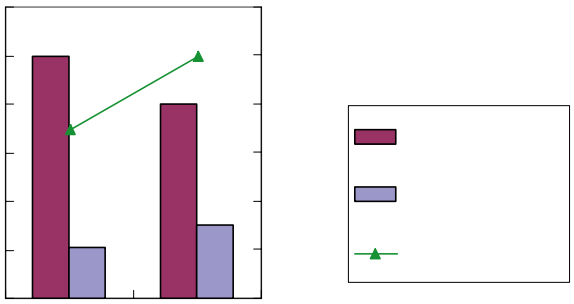

図 6　実験の結果

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.

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

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

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

毛利 隆軌　Takanori MOURI

筑波大学大学院システム情報工学研究科在学中. 2002 年筑波大学第三学群情報学類卒業. XML, WWW, 文書データベースに興味を持つ. 情報処理学会学生会員. 日本データベース学会学生会員.

北川 博之　Hiroyuki KITAGAWA

筑波大学電子・情報工学系教授. 1980 年東京大学大学院理学系研究科修了. 理学博士（東京大学）. 異種情報源統合，文書データベース，WWW の高度利用等の研究に従事. 著者「データベースシステム」（昭晃堂），「Unnormalized Relational Data Model」(共著，Springer-Verlag)等. ACM, IEEE-CS，日本データベース学会，情報処理学会，電子情報通信学会，日本ソフトウェア科学会，各会員.