複数新聞記事サイトの横断検索とトピックのドリフト支援システム

AN IMPLEMENTATION OF CROSS-ARTICLE-SEARCH AND TOPIC DRIFTING AID SYSTEM FROM MULTIPLE NEWS SITES

山田 剛一[1]・大熊 耕平[2]・増田 英孝[3]・中川 裕志[4]

[1]東京電機大学　(E-mail:yamada@im.dendai.ac.jp)
[2]東京電機大学　(E-mail:ohkuma@cdl.im.dendai.ac.jp)
[3]東京電機大学　(E-mail:masuda@im.dendai.ac.jp)
[4]東京大学 / 社会技術研究システム 総括研究グループ (E-mail:nakagawa@dl.itc.u-tokyo.ac.jp)

本研究では，あるトピックについての文書群から，そのトピックに関連する別のトピックの文書群へとユーザを連続的にナビゲートするシステムを開発した．本システムでは複数の新聞記事サイトの横断検索を行い，その結果得られる同一トピックの記事間の差異をユーザに提示することによりナビゲーションを行う．本システムの適用範囲は同一トピックの新聞記事群に限らず，同一トピックについて議論しているドキュメント群一般に適用することができる．例えばある問題について議論しているドキュメント群が存在したとき，本システムを用いることにより，その問題に関連のある情報や，その問題への異なる視点を見つけることができる．

キーワード：トピックのドリフト，複数サイトの横断検索，ナビゲーション

1. はじめに

現在，インターネット上では主要な新聞社や出版社などによって記事が無料で公開されており，幅広く利用されている．これらの記事を公開しているサイト(新聞記事サイト)を横断的に検索することができれば，「複数のサイトを一度に調べたい」，「同じトピックの記事を重複して読みたくない」，「同一のトピックの記事が発信元によってどのように異なるのか知りたい」，「あるトピックの記事を読み，それに直接的，あるいは間接的に関係する記事をいもづる式に探索したい」といったユーザの要求に応えることが可能となる．本研究では，複数の新聞記事サイトを横断検索すると同時に上記の「いもづる式探索」をするシステムを試作した．

本システムで最初に各新聞社の記事の横断検索を行うと，ユーザは内容の類似する記事群を得ることになる．この類似する記事群の差異をユーザに提示することにより，ユーザは何がメイントピックで何がサブトピックなのか，あるいは情報源に固有の視点は何か，といったことを知ることができる．それらの情報をユーザが取捨選択して次回検索に反映させていくことにより，ユーザは上に述べたような「いもづる式」に新たなトピックへとナビゲートされる．このように，本システムはトピックのナビゲータの役割を果たすよう設計されている．

2. 社会技術としての本システムの位置づけ

我々が今回扱う対象としたのは複数の新聞社が公開している新聞記事群である．社会問題が取り上げられる媒体としては新聞は雑誌と並び主要な存在であり，マスメディアの動向を知る上で無視することはできない．一方，新聞は電子化データが一般に公開されている数少ない媒体の一つであり，これを情報源として活用することができれば広く利用がされるものと期待される．

新聞社がある社会問題を記事として取りあげる場合，それをどのような視点で取りあげるのか，どのようなサブトピックについて言及するのかは新聞社ごとに異なる．本システムではこのような新聞社間の差異をユーザに提示することができるため，ユーザは問題を多面的に捉えることができ，問題の全体像の把握に近づくことができる．本システムをナビゲータとすれば個々のサブトピックに関しても同様に多面的に捉えることができるため，これを繰り返すことによりユーザは複雑な問題の全体像を段階的に把握することが可能となる．

本システムの適用範囲は同一トピックの新聞記事群に限らず，同一トピックについて議論しているドキュメント群一般に適用することができる．例えばある問題について議論しているドキュメント群が存在したとき，本システムを用いることにより，その問題に関連のある情報

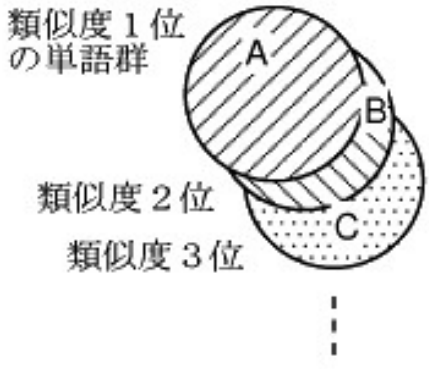

Fig. 1　類似記事群における単語の重なり

や，その問題への異なる視点を見つけることができる．コンセンサス形成過程においては，問題となっている話題だけで議論するのでは解決や妥協ができないことがあり，そのような場合に，関連する知識や意見はもちろんのこと，別の観点での意見や知見がコンセンサスを形成するために役立つものと期待される．

3. 複数新聞記事サイトのいもづる式ナビゲーションシステム

3.1. システムの概要

検索質問にマッチする記事を複数の新聞記事サイトから収集，検索を行うことで，類似記事を収集することができる. 類似する多数の記事には, 単独の記事に比べて, いろいろな観点からの情報，見方などが多数示されていると予想される．この類似記事を用いることで，ある話題について，単独の記事ではカバーしていなかったテーマへの糸口が見つかりやすくなると期待できる．

複数の類似記事はおよそ Fig. 1 のように内容がオーバーラップしている．質問と一番類似度が高い記事はユーザが全文を読むことを想定する．ここで，一番目の記事で扱っていない内容をいもづる式に手繰るという局面を想定してみよう．同じような問題設定は多文書自動要約に見られる [1]．多文書自動要約では，MMR(Maximal Marginal Relevancy)という考え方が使われる．すなわち，元の質問に類似していることと，既に選択した記事に類似していないことの両者を加味した基準によって記事を選択する．我々の目的では，いもづる式ナビゲーションの性質上，既に選択した記事に類似していないことにのみ重きをおくことになる．したがって，問題は類似度が 2 番目以降の記事に出ている内容のうち一番目の記事と重複しないものをどのようにブラウザ上に提示するかである．要約における MMR の場合，表示は記事単位であった．しかし，ここでは記事全部を提示してしまうと，いたずらに利用者に負担を強いる. そこで, 2 番目以降, 例えば n 番目に類似した記事内容を提示する方法は,

・n 番目の記事の内容のうち，既に選択した n-1 番目までの記事に出現していない内容を表す部分を表示する．

しかし，このような部分に対応する文の集合を探し出すことは，意味理解に近いことが必要であり，現実的ではない．より軽い処理で実現でき，かつ利用者にも indicative な情報を提示できるという観点からは単語を単位とする表示が現実的である．そこで，このシステムでは，Fig. 1 の重ならない部分を単語の集合とみなすことにした．単語を提示するもうひとつの利点は，単語には TF×IDF などの方法で重要度がつけられ，その重要度の順に表示するというコンパクトな表示ができる点である．

よって，提案するシステムでは，収集した記事とその記事の固有の単語を提示する．この単語を選択し再び検索を行うことでユーザをナビゲートする．

このシステムの流れを以下に示す.

1. 複数新聞記事サイトの記事を収集しインデックスを作成する.
2. ユーザが検索質問を入力する.
3. 検索質問の単語を含む記事群とその記事に含まれる単語群を取得する.
4. 検索質問と各記事との類似度を求める.
5. 検索質問に最も類似した記事と，他の記事との類似度を求め，この類似度順に記事を並び替える.
6. ユーザに記事群とその各記事に固有の単語を提示する.

複数の新聞社のサイトから，検索対象となる大量の記事をネットワーク経由で取得するのには時間がかかる．そのため，1 の記事の収集はユーザが検索を行う前にあらかじめ行っておく．この記事を収集する段階で，記事検索やナビゲーションを行う際に必要となるインデックスを作成する.

ユーザがシステムに検索質問を与えたとき，その検索の対象は，その時点までにシステムが収集したすべての新聞記事となる．検索質問は，単語一つでも，複数の単語を含んでいてもよい. システムは 1 で構築したインデックスから，検索質問に含まれる単語群を含む記事(の ID)と，その記事に含まれる単語群を取得する.

検索質問に含まれる単語群と，各記事に含まれる単語群から, 検索質問と各記事との類似度を求める. さらに, 検索質問との類似度が最も高い 1 記事を新たな検索質問

に見立て，この記事と他の記事との類似度を求める．この新たな類似度を使って記事を並び替えることにより，記事間の類似度の高い記事が上位に集まることになる．

検索質問に最も類似している記事はユーザに全文を提示する．その記事に類似している記事については，その類似度順に，より類似度の高い記事に含まれていない単語群を提示していく．

ユーザは提示された単語群を取捨選択し，次の検索の方向を定める．ここで 2 に戻る．このように，検索の方向をインタラクティブに変化させながら検索を進めていくのが，本システムのナビゲーション機能である．

3.2. 複数新聞社サイトからの記事収集

記事の収集は，新聞記事サイトのトップページを起点として行う．新聞記事サイトのページでは，記事本文のあるページへのリンクのほか，記事以外の様々なページへのリンクも同時に張られている. このような状況の下，サイトの中から記事本文のあるページを探し出し，そのページだけを取得する必要がある．

また，記事収集の対象とすべきサイトはユーザの好みによって異なり，さらに，記事を収集するサイトはいつでも新規に追加できることが望ましい．よって，サイト内から記事であるページを探し出す際に，特定のサイトに依存するような情報を用いて記事を収集する方法は使えない．そこで，新聞記事サイトから記事のページのみを探し出す方法として，新聞記事サイトに普遍的な特徴を用いて記事の収集を行う．

ここでは各ページの URL に注目した．記事本文を含むページの URL には，その記事が載った日付が含まれている. これを用いて記事ページに関する判断を行った．ほとんどの新聞記事サイトでは，記事収集を行った時点での, 最新の記事へのリンク（URL に日付を含んでいる）と，各ジャンル（社会，スポーツ，政治など）の記事一覧ページへのリンクが同一ページに存在するため，これを記事ページ探索の手がかりとして利用する．

まず，トップページに含まれるリンクの URL を全て取得し，その URL が日付を含むかを調べる．この時点で日付を持つ URL があれば，そのリンクが最新記事へのリンクであると判断することができる．トップページに日付を持つ URL のページへのリンクがなければ，見つかるまでリンクをたどっていく．

前述の性質から，最新の記事へのリンクのあるページには各ジャンルの記事一覧ページへのリンクがあるため, そこからさらに一階層リンクをたどれば各ジャンルの記事一覧ページを取得することができる．この記事一覧ページの中のリンクを調べ，特定の日付を持つ URL のページのみを取得することで記事であろうページだけを収集することができる．記事一覧のページのリンクがわかり記事であろうページを収集した時点で，記事収集を終了する．

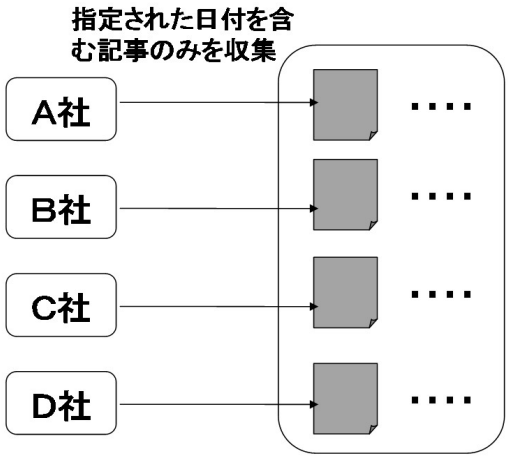

Fig. 2　複数サイトからの記事収集

しかし，この中には URL に日付を持っていたとしても記事でないページが含まれている．このページを除くために，その記事へのリンクがはってある，リンク元のページの文字列を手がかりとして使用する．記事へのリンクがはってあるリンク元の文字列は，リンク先のページでは記事の見出しとして含まれているので，記事へのリンクがはってある文字列がリンク先のページに含まれているか否かを調べる．これによって記事であろうページ群から記事のみを収集することができるようになる．

このような記事収集方法を行うことで，与えられたサイトのページ全てを調べることなしに，そのサイトの特定の日付の記事のみを収集することが可能となる．

以上の方法を，収集対象である複数サイトに対して行う．Fig. 2 の例では，A 社～D 社のサイトから特定の日付の記事だけを収集している．その収集したすべての記事が検索の対象となる．前述の方法を使えば，A 社～D 社だけでなく未知のサイトにも対応することができる．

3.3. 収集記事からの必要な部分の抽出

記事の検索のために，収集した記事から単語とその単語の TF 値(文書内出現頻度)と URL のデータベースを作成する．データベースを作成することで，後の記事の検索を行うことができる．

まず，収集した記事からその記事の見出しと本文を抽出する．ここでもサイト依存性を極力排除するため，リンク以外の HTML の要素は手がかりとして使用しない．まず見出しの抽出を行う際に，各記事のリンク元のページにある，この記事へのリンクがはってある文字列を手がかりとして使う．この文字列はその記事の見出しとほぼ同一の文字列であるので，これを用いて記事の見出しを探し出す. 見出しが発見された後, 見出し以降にある,

Fig. 3　ブラウザと連動した検索結果の表示

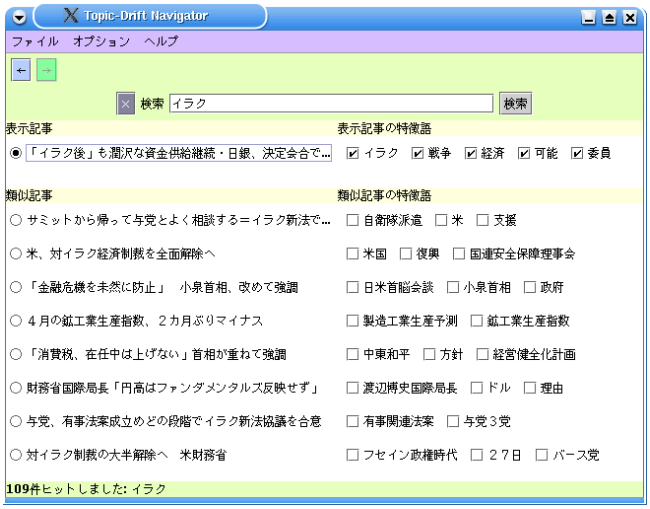

Fig. 4　検索例：イラク

句点がある文を探し出す．記事の本文には句点がついているので，句点がある文を探し出すことで本文を抽出することができる．次に，抽出した見出しと本文から，名詞を取り出す．これには，形態素解析システム茶筌 [3]を用いた．抽出した見出しと本文を茶筌にかけ形態素解析を行い，名詞である単語のみを取り出し，これをデータベース作成に用いた．また，ここで取り出した単語の TF 値も求める．これにより，単語をキーとして URL を返すデータベースと，URL をキーとしてその記事の単語と TF 値を返すデータベースを作ることができる．これを記事の検索に用いる．

3.4. 記事の検索

あらかじめ収集してある記事の中から検索質問に合う記事を検索する．検索質問の単語を記事の中に含んでいる記事だけをユーザに提示する対象の記事とし，収集してある記事の中から検索する．単語を含んだ記事かどうかを調べるときに，収集してある記事のソース全てを調べるのでは莫大な時間がかかってしまう．ここで，3.3 節で作成したデータベースを用いる．このデータベースを用いて，検索質問の単語を見出しまたは本文中に含んでいる記事を簡単に探し出すことができる．この探し出した記事を用いて類似度の算出をし，ユーザに記事を提示する．

3.5. 類似度の算出

検索質問に合致する記事を探し出すために，検索質問の単語群と各記事間の類似度を算出し，これを検索質問にマッチするかの判断に用いた．検索質問の単語群と各記事間の類似度には cosine 類似度を用いた．このときベクトルの次元は検索質問の単語および記事に含まれる単語の種類数となる．ベクトルの要素にはその単語の TF の値を用いている．

3.6. 類似度による記事の提示とナビゲーション

検索結果を表示するウィンドウの例を Fig. 3 に示す．検索質問との類似度が最大の記事については，見出しや含まれる単語のほか，Web ブラウザを用いて記事全文を提示する．その記事との類似度が高い記事群については類似度順に見出しを提示し，その見出しを選択することでブラウザに表示する記事の切り替えを行うことができる．ここで，それぞれの記事に現れたその記事固有の単語を次のナビゲーションのための検索語として提示する．この「記事固有の単語」とは，3.1 節でも示したようにその記事よりも類似度上位の記事に含まれていない，その記事の単語のことである．この単語は TF×IDF を計算し，その値の大きい順に提示している．ユーザはこの「記事固有の単語」を選択していくことでナビゲートされていく．

実際にはユーザに単語をそのまま提示するのではなく，単語が記事中で複合語として現れている場合にはその複合語を提示している．これにより，ユーザは各記事に現れる固有の概念が把握しやすくなっている．なおこの機能を実現するため，複合語の構成要素から複合語が得られる形のインデックスを作成している．

Fig. 4〜Fig. 6 に検索例を示す．Fig. 4 は，ユーザが検索質問として「イラク」と入力した場合の例である．検索で得られている記事はイラク関連のものであるが，それぞれ内容に異なりがあるため，ウィンドウ右側にその差異である単語群が示されている．イラク問題では「自衛隊派遣」「復興」「中東和平」「有事関連法案」といった関連トピックがあることがわかり，日本の国内問題としてのイラク問題という側面が強く現れている．これは，イラク問題について議論する際に，アメリカの行動だけでなく日本の姿勢についても目を向ける必要があることを示唆している．

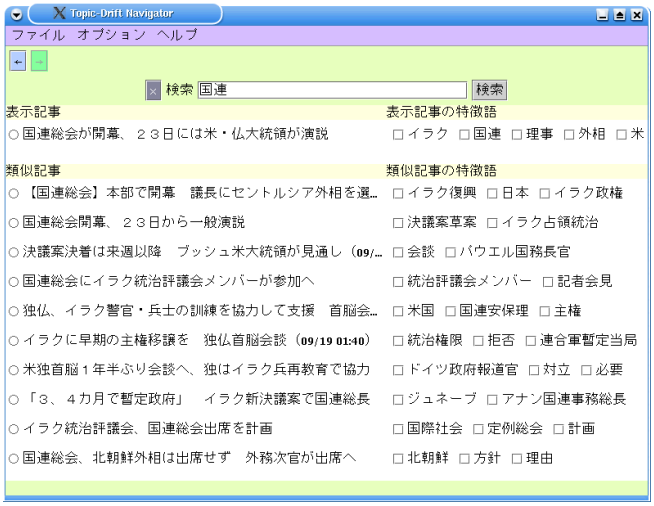

Fig. 5　検索例：国連

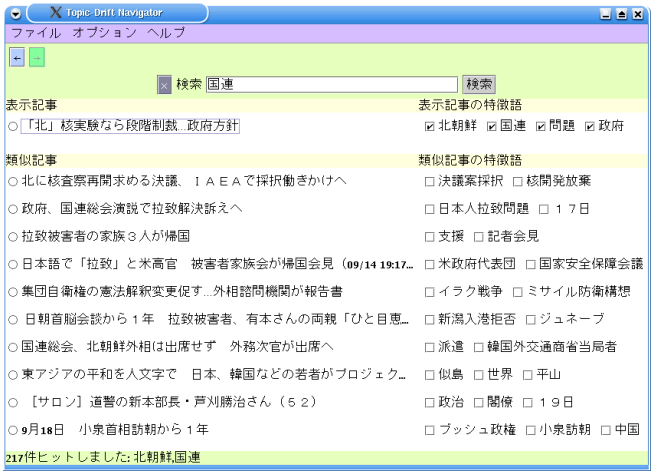

Fig. 6　検索例：国連，北朝鮮

各記事の見出しをクリックすればその関連トピックを含む記事をブラウザで見ることができるし，さらに詳しく知りたい場合には，関連トピックを表す語をチェックして検索ボタンを押せば，その関連トピックでの検索ができるため，さらなるサブトピックを見ることもできる．

Fig. 5 は，検索質問として「国連」を指定した場合の検索結果の例である．イラク戦争に関するトピック「イラク復興」「イラク占領統治」「統治評議会」「米国」などが主に現れているが，それらとはトピックの方向性が異なる「北朝鮮」もある．ここではユーザが興味を持ち深く知りたくなったのが北朝鮮関連のトピックであったと仮定し，チェックボックスにより「北朝鮮」を選択して，検索質問を「国連」「北朝鮮」の２語にして再度検索した．その検索結果が Fig. 6 である．イラク関連の話題は消え，北朝鮮問題のサブトピックが見て取れる．大きなトピックとしては核開発問題と日本人拉致問題があることがわかる．ここでユーザはさらに深く知りたいトピックを指定することにより，新たな方向へナビゲートされていく．

4.　関連研究

本研究のシステムはナビゲーションを主眼としているが，そのナビゲーションの方向がユーザの選択する単語群により決定づけられることから，関連性フィードバック(relevance feedback)[2]に似ている面がある．ただし，関連性フィードバックによる検索質問の拡張は，検索開始時のユーザの意図に近づくよう検索質問を修正していく仕組みであるのに対して，本研究のシステムでは，初回の検索結果と(関連はしているが)異なる方向へとナビゲートしていく．よって，ユーザからのフィードバックは，その内容も目的も異なっている．

本研究のシステムは類似記事群とその差異をユーザに提示するため，一般的なドキュメント空間の可視化システム [4]，あるいは Web の可視化システムとの類似性がある．それらに比べると本研究のシステムは単純なインタフェースであるが，ほぼ同一内容のドキュメントが多数存在するという状況下において，その相違点に着目しユーザに提示する場合には，単語そのものによって差異を提示するのが明解でわかりやすい．ただし，目的がナビゲーションではなく差異そのものの閲覧である場合には，異なる単語群を含む段落といった単位で抽出することが望ましいといえる．なお，Web 全体の拡大に伴い，「ほぼ同一内容のドキュメントが多数存在する」という状況も拡大しており，本研究のシステムの適用範囲も拡大の方向にあると考えている．

本研究のシステムはナビゲーションシステムであるため段階的にトピックがドリフトしていくが，その一回一回におけるドリフトの度合いは，すでに 3.1 節で指摘したように，多文書自動要約において MMR により定量化されている．本システムの目的は要約ではないが，トピックのドリフトを評価する際の参考としたい．

5.　おわりに

本研究のシステムの大きな特徴はトピックのドリフト支援であるが，このドリフトの履歴を可視化することにより，個々のユーザの視点に立ったトピック・サブトピックのマップが得られる．これはトピック・サブトピックの関連性の分析に役立つほか，ユーザが自分の履歴を鳥瞰することによって，ユーザの頭の中を整理することも可能となる．

本研究のシステムはまだ試作段階である．これから運用実験を行うことにより，トピックのドリフトがどのよ

うな条件下で効果を発揮するのか確認していきたい．また，このシステムがどのような目的・場面で役立つのか未知数な部分もあるため，トピックのドリフト履歴を可視化する機能を用いながら，このシステムの新たな可能性を明らかにしていきたい．

参考文献

1) 奥村学, 難波英嗣(2002)「テキスト自動要約に関する最近の話題」『自然言語処理』9(4), 97-116.
2) William, B, F., and Ricardo, B., (1992). *Information Retrieval --- Data Structures & Algorithms.*
3) 松本裕治, 北内啓, 山下善隆, 松田寛浅, 原正幸(1999)「日本語形態素解析システム 茶筌 version 2.0 使用説明書 第二版」『NAIST Technical Report』NAIST-IS-TR99012.
4) Earl, R. (1994). Galaxy of news: an approach to visualizing and understanding expansive news landscapes. *Proceedings of the 7th annual ACM symposium on User interface software and technology*, 3.-12.

本研究は, 社会技術研究システム ミッション・プログラム「安全性に係わる社会問題解決のための知識体系の構築」（2001～2002 年度は日本原子力研究所の事業, 2003 年度からは科学技術振興事業団の事業）の研究として行われた.

AN IMPLEMENTATION OF CROSS-ARTICLE-SEARCH AND TOPIC DRIFTING AID SYSTEM FROM MULTIPLE NEWS SITES

Koichi YAMADA [1], Kouhei OHKUMA [2], Hidetaka MASUDA[3], Hiroshi NAKAGAWA[4]

[1] School of Engineering, Tokyo Denki University. (E-mail:yamada@im.dendai.ac.jp)
[2] School of Engineering, Tokyo Denki University. (E-mail:ohkuma@cdl.im.dendai.ac.jp)
[3] School of Engineering, Tokyo Denki University. (E-mail:masuda@im.dendai.ac.jp)
[4] Information Technology Center, The University of Tokyo. (E-mail:nakagawa@dl.itc.u-tokyo.ac.jp)

In this research, we developed a system which navigates a user from a topic to the related other topics. This system retrieves articles of the specific topic from multiple news sites and shows differences between the articles. This system is for not only news articles but also general documents. By using this system, a user can find some related information and different points of view of the same topic.

Key Words: *Topic Drifting, Cross-Article-Search, Navigation.*