類似文書間の差異の提示によるトピックドリフト支援システム

山田 剛一[1]・大熊 耕平[1]・増田 英孝[1]・中川 裕志[2]

[1]東京電機大学
[2]東京大学 / 社会技術研究システム

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（図 1）．

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. トピックドリフト支援システム

本システムの基盤は検索エンジンであり，その検索対象はユーザが指定する複数の新聞記事サイトから収集した記事群である．そのため，検索結果には類似記事が多く並ぶことになる．類似記事はおよそ図 2 のように内容がオーバーラップしている．質問と最も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，すで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

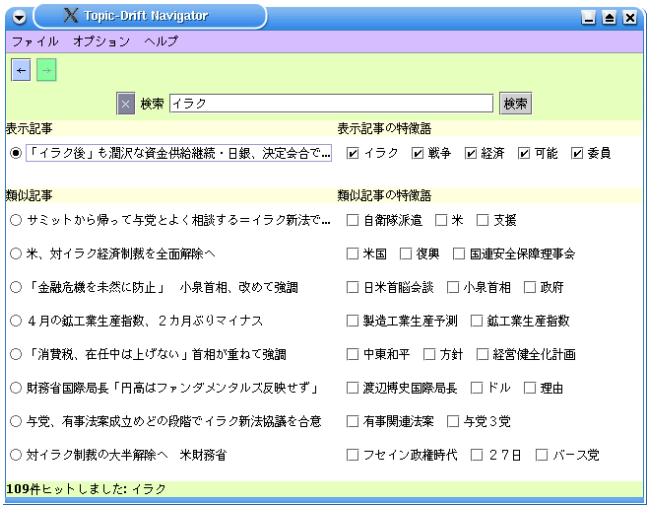

図 1 実行画面（「イラク」で検索）

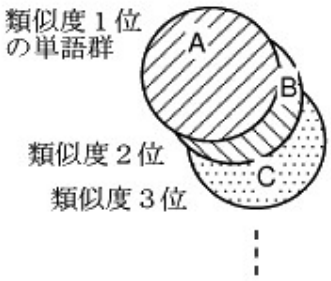

図 2 類似記事群における単語の重なり

Topic Drifting Aid System
Koichi Yamada[1], Kouhei Ohkuma[1], Hidetaka Masuda[1], Hiroshi Nakagawa[2]
[1]Tokyo Denki University
[2]The University of Tokyo / Research Institute of Science and Technology for Society

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，図2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．ユーザはこの単語を選択し再び検索を行う．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．

このシステムの流れを以下に示す．

1. ユーザが指定した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．指定されたサイトから，特定の日付の記事ページを自動収集する．
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ベクトル空間法により求める．
5. 検索質問に最も類似した記事を新たな検索質問に見立て，他の記事をこの類似度順に並び替える．これにより，記事間の類似度の高い記事が上位に集まる．
6. ユーザに記事群とその各記事に固有の単語群を提示する．単語群は重み（TF×IDF）順に並べる．検索質問に最も類似している記事は連動しているWebブラウザで表示する．他の記事も必要であればWebブラウザで表示させることができる．
7. 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．

図1と図3で検索例を示す．図1は検索質問を「イラク」とした例である．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る．

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．

図3は検索質問として「国連」を指定した場合の例である．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，それらとは方向性が異なる「北朝鮮」もある．「北朝鮮」を選択し検索質問を「国連」「北朝鮮」の2語にして再度検索をすれば，北朝鮮問題のサブトピック，

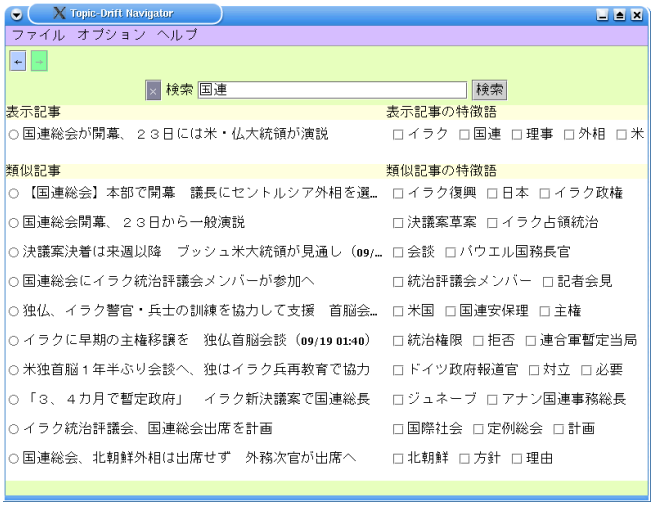

図3　実行画面（「国連」で検索）

核開発問題と日本人拉致問題が見えてくる．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．

3. 関連研究

本システムは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，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ため，ユーザからのフィードバック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，ドキュメント空間の可視化システム[3]との類似性がある．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といえる．

4. おわりに

類似記事間の差異の提示によるトピックドリフト支援システムを構築した．今後はドリフトの履歴を分析することにより，ドリフトの有効性を確認していきたい．

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)，pp.97-116.
2) William B. Frakes, Ricardo A. Baeza-Yates (Eds.) (1992). *Information Retrieval: Data Structures & Algorithms.*
3) 武田浩一，野美山浩 (2000)「テキスト情報の可視化を利用した情報検索」『情報処理』41(4)，pp.343-350.

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．