複数新聞記事サイトの横断検索システムの試作

大熊耕平 †, 山田剛一 †, 増田英孝 †, 中川裕志 ‡
東京電機大学工学部 † 東京大学情報基盤センター ‡

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巡りたい」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システムを試作した。

本システムで横断検索を行うと、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 横断検索システムの実装

2.1 システムの概要

検索要求にマッチする記事を複数のサイトから検索、収集し、その記事と固有の単語を提示する。そして、これを用いて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. ユーザが検索語を入力する。
3. 検索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. 検索語と、取得したそれぞれの記事の単語群から、検索語と各記事の類似度を算出する。
5. 類似度で記事群を並べ換え、各記事固有の単語を提示する。

複数新聞社から大量の記事をネットワーク経由でダウンロードするには時間がかかるので 1 の記事の収集は検索を行う前にあらかじめ行っておく。検索を行う際にはそれ以前に収集した全ての記事が検索の対象となる。

2 でユーザが入力する検索語はひとつ、又は複数である。ここで入力された検索語は茶筌 [1] を用いて分割し、名詞のみを取りだし、検索語として用いる。

5 で提示される単語を検索語としてユーザが選ぶことで、2 の所に戻り繰り返し検索を行うことができる。ここで、ユーザを新たなトピックへとナビゲートできる。

2.2 記事情報収集と類似度の算出

新聞記事サイトから目的の記事を取得するには記事本文のあるページとその日付を判別する必要がある。各新聞社サイトの記事には URL に日付が含まれているので、ここに注目し、URL に日付を持つものを記事、持たないものを記事以外と判断し、記事情報の収集を行う。 ここで得る記事情報は記事であると判断された URL とその見出し、そしてその記事の持つ名詞である。

記事からの名詞の抽出には茶筌 [1] を用いた。

類似度は名詞の TF からベクトルをつくり、そのベクトルを用いた cosine 類似度から求めた。

2.3 類似度による記事の提示とナビゲーション

算出した類似度を用いて、検索語と類似度が一番高い記事をユーザの望む記事として提示する。同時にこ

An Implementation of Cross-article-search System from Multiple News Site
†Kouhei OHKUMA,†Koichi YAMADA, †Hidetaka MASUDA and ‡Hiroshi NAKAGAWA
†School of Engineering,Tokyo Denki University, ‡Information Technology Center,The University of Tokyo

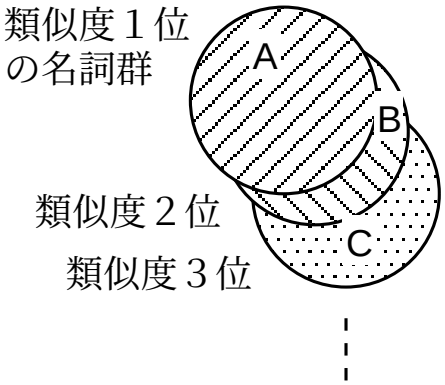

図 1: 固有名詞の抽出

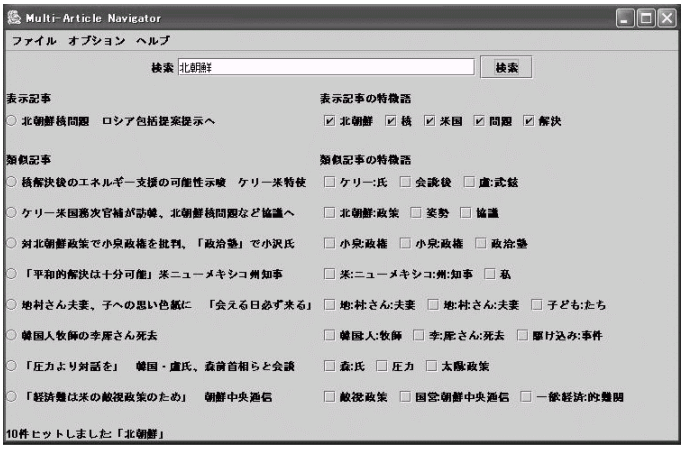

図 2: 検索結果表示画面

の記事に現れた名詞を次のナビゲーションのための検索語として提示する。また、類似度が二番目以下の記事に含まれる、その各記事よりも類似度が上位の記事に含まれない名詞を TFIDF 値を用いて並べ換え、TFIDF が高い名詞を提示する。図１に名詞の表示の説明図を示す。まず表示するものは類似度１位の名詞群である。これは、類似度１位なのでそのまま全ての名詞 (A) を TFIDF 値で並べ換え表示する。次に類似度２位の名詞群から名詞を表示するが、このとき記事中にある名詞の中で類似度１位の記事にある名詞は破棄し、残りの名詞 (この記事固有の名詞)(B) を TFIDF で並べ換え表示する。以下同じように、類似度３位の記事は１位と２位の記事に含まれない名詞 (C) で並べ換え表示、と繰り返す。これを行うことで２位以下の記事はその記事固有の名詞を提示させることができ、この名詞をユーザが選択することで、新たな方向へとユーザをナビゲートできる。

図２に検索結果を示す。この図では検索語「北朝鮮」を入力し検索を行っている。表示記事は類似度１位の記事、類似記事は上から類似度２位の記事を表し以下類似度が高い順に表示している。このとき、類似記事の中で類似度が高い記事は表示記事と同じトピックの他社の記事を表している。この結果から、表示記事の特徴語 (北朝鮮、核、米国など) を選択して検索を行うと、トピックのドリフトを抑えて検索を行うことができ、また類似記事固有の特徴語 (小泉政権、駆け込み事件など) を選択して検索を行うと新たな方向へとナビゲートすることができる。

ここで、名詞をそのまま提示するのでは内容がつかめないので、表示には実際の記事でのその名詞の前後を用いて複合名詞として表示する。

3　おわりに

本システムを実装することで、ユーザをある話題について関連性のある方向へ、また、新たな方向へとナビゲートすることができる。

今後はユーザが読みたい新聞記事サイトを追加できるようにする。そのために、いろいろなサイトでの記事検出の検証を行う。

参考文献

[1] 奈良先端科学技術大学院大学自然言語処理学講座 形態素解析システム「茶筌」, http://chasen.aist-nara.ac.jp/