名詞の連接に着目した日本語新聞記事の関連記事検索手法

大竹 清敬† 山本 和英‡ 増山 繁†
otake@smlab.tutkie.tut.ac.jp, yamamoto@itl.atr.co.jp, masuyama@tutkie.tut.ac.jp
†豊橋技術科学大学 知識情報工学系
‡ATR 音声翻訳通信研究所

1 はじめに

現在大量の機械可読文章 (コーパス) が存在している．中でも，新聞は現代社会の大量情報の流通媒体であるため，検索需要が多い．しかし，読者が興味を持った記事の関連記事を検索する事は大変手間がかかる．そこでこのような関連記事を効率良く検索できる手法が必要とされる．そのためには，膨大かつ多様な情報から読者が欲する情報を得るために，各記事の一定の内容を捉える事が必要となる．そのための手法として通常の構文解析を用いているものもあるが，対象が非均質であるため汎用性が問題になる．このため，より浅いが頑強な自然言語処理，例えば統計的な処理やテンプレートを併用する方法，共起関係を利用した方法などがある [1]．

本研究では記事間の関連度を「同一あるいは類似した表現の共有度合」と定義し，名詞とその連接に着目して記事内容を捉える手法を提案する．我々は本手法が関連記事検索において関連度の定義通りの検索を有効に行なう事を [2] にて確認した．関連記事検索において重要なのはユーザの視点である．一つの記事の中に複数のトピックが存在するような場合はユーザがどのトピックに関連する記事を求めているのかを判断する必要がある．しかし本手法ではそのようなことを行なわず，上述の関連度の定義に従って検索を行なう．ユーザはその事を踏まえた上で，検索結果の中から自分の求める記事を見つけ，必要ならばさらに元記事とその記事をあわせて再度関連記事検索を行なえばよい．我々は検索においてインタラクションは必須のものであると考えており，そのインタラクションにおいて見通しよく検索できることが重要だと考えている．結果，検索中におこるユーザの視点移動に対しても柔軟に対応することが可能となる．インタラクションにおいて重要なのはその速度である．検索速度が遅くてはユーザを満足させる検索が行なえない．

A Retrieval Method of Relevant Japanese Newspaper Articles by Focusing on Noun Connections
Kiyonori OHTAKE†, Kazuhide YAMAMOTO‡, and Shigeru MASUYAMA†
†Department of Knowledge-based Information Engineering, Toyohashi University of Technology
‡ATR Interpreting Telecommunications Research Laboratories

我々は望むインタラクションを実現するために [2] を発展させ，インデックスを用いて検索速度を向上する手法を考案した．これをWWW上のインタフェースを通して使用した実験の結果，ユーザの視点から見通しよく関連記事検索を行なえる手法である事を確認したので報告する．

2 名詞の連接に着目した関連記事検索手法

日本語において，名詞は複合して一つの名詞を構成することが多く，また [3] に示されるように，そのような複合名詞は記事を検索する場合の手掛りとなりやすい．記事を検索するための手掛りとなる情報を得るために形態素解析を行なうことが多い．現在広く利用されている JUMAN[4] を用いて形態素解析を行なった場合，複合名詞は個々の構成要素である名詞の形態素へと分割される．個々の名詞を手掛りとして検索を行なった場合，ノイズが大きくなりすぎ，検索に適さないと予想できる．そのため，本手法では記事内の名詞とその連接に着目し，各記事の特徴を表すデータを形態素解析結果から抽出し，抽出したデータを比較することにより関連度を求める．

2.1 データ構造

記事の特徴を表現するために本手法で用いるデータ構造は，形態素を節点とする有向グラフの集合である．以下これを局所有向グラフならびに局所有向グラフ集合と呼ぶ．定義を以下に挙げる．

1. 中心となる節点は一つの局所有向グラフに一つのみ存在し，その品詞は名詞である．但し，時相名詞，形式名詞，副詞的名詞，数詞は除く．
2. 中心節点である名詞を終点として持つ有向辺の始点となる節点の品詞は形容詞である．
3. 中心節点である名詞を始点として持つ有向辺の終点となる節点の品詞は名詞，動詞のいずれかである．名詞は中心節点と同一の条件をみたすもの．また，動詞はその活用語幹を格納し，サ変動詞もそれに含むものとする．
4. 有向辺はそれが示す連接の頻度を重みとして持つ．

局所有向グラフの例を図1に示す．

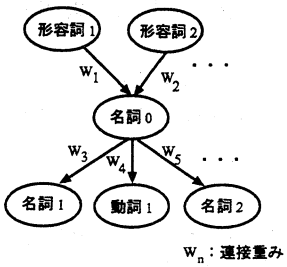

図 1: 局所有向グラフ

2.2 局所有向グラフ集合の作成

記事の形態素解析結果をファイルとして用意し，ファイル内の名詞に着目する．その名詞の前後の形態素から名詞に対応する局所有向グラフを作成する．有向辺には始点と終点の連接の出現頻度を表す重みが付加される．また局所有向グラフを識別するものは中心節点の名詞となる．記事内の全ての名詞に対して，局所有向グラフを作成し，記事に対応する局所有向グラフ集合を得る．この局所有向グラフ集合がその記事の特徴をあらわし，関連記事検索の手掛りとなる．

2.3 ヒューリスティクス

ここでは，記事間における表記の揺れを吸収するために局所有向グラフ作成時に用いたヒューリスティクスについて説明する．

1. 名詞接続助詞の「の」など
   名詞間に存在する名詞接続助詞の「の」は無視する．つまりテキスト中に "A/の/B" (/は連接を表す) という形態素の連接が出現した場合には "A/B"が出現した場合と同様に扱われる．また，読点や「・」が名詞にはさまれている場合も同様に扱う．
2. 3つの隣接する名詞
   名詞が連続して3個出現した場合 ("A/B/C")，には1番目と3番目が連接しているとも考え ("A/C")，そのようなグラフを作成する
3. 括弧の取り扱い
   記事中に括弧の「(」が出現した場合，具体的には A / (/ B /) / C のようなパターンを考える．このとき，A は B で置きかえられることもあるとして処理する．つまり A→C ならびに B→C であることも考慮して有向辺を設定する．また B の中身が $b_1/b_2/b_3$ であるような場合でも A が b_3 で置きかえられることもあるとして処理する．このとき $b_1/b_2/b_3$ の連接によるグラフも作成する．
4. 見出しの取り扱い
   見出しは記事を表現する重要な手掛りとなることから，本手法においても重要視している．そこで，見出しに出現する形態素のうち名詞に関して，その出現頻度を局所有向グラフとは別のデータとして持つ事とする．

2.4 2つの局所有向グラフ集合の関連度の評価方法

関連度評価のためのアルゴリズムを以下に示す．

Step1 元記事と対象記事との局所有向グラフ集合において同一の中心節点 (名詞) を持つ局所有向グラフをそれぞれ選択する．そのようなグラフがなければ終了．

Step2 両者のグラフにおいて同一の始点と終点を持つ有向辺の重みを加算し，評価値へ加える．もし，そのような有向辺が存在しない場合 (中心節点のみが同一) には評価値へ CNW(Center Node Weight, 中心節点重み) を加算する．

Step3 両者のグラフ集合において同一の中心節点を持つ局所グラフが他にも存在すれば **Step2** へ，そうでなければ終了．

以上のアルゴリズムを適用し，元記事と対象記事間の関連度の評価値を算出する．

2.5 インデックス

ハッシュデータベースの2種類のインデックスを使用する．まず，作成した局所有向グラフ集合を格納する正規インデックスを用意する．正規インデックスには記事IDをキーとして，局所有向グラフ集合と見出しに出現した名詞とその出現頻度をリスト形式で表現したデータが格納される．次に局所有向グラフ集合内における名詞を始点とする節点の連接データ (A→B の形式で，A として名詞のみをとる) がどの記事に含まれるかを示すリストを格納した転置インデックスを作成する．

ハッシュデータベースを使用することにより高速な検索が望める．

2.5.1 インデックスサイズの最適化

正規インデックスはその記事の内容を表現する局所有向グラフ集合からなるので，基本的にはそれ以上サイズを小さくすることはできない．しかし，転置インデックスはある形態素の連接がどの記事に含まれているかを示しているものなので，多くの無駄が含まれていることが容易に理解できる．まず，関連記事を検索する上では手掛りとならない一つの記事にしか含まれない連接を転置インデックスから削除する．次にIIOCV(Inverted Index Optimize Cut-off Value, 転置インデックス最適化足切り値) 以上の記事に含まれている連接も転置インデックスから削除する．また一つの記事にのみ含まれる連接

を示す有向辺を正規インデックスからも削除し，正規インデックスのサイズを小さくする．IIOCV 以上の記事に含まれる連接を正規インデックスから削除しないのは関連度評価の際にスコアリングが正しく行えないからである．

2.6 関連記事検索

ここでは，実際の関連記事検索がどのように行われるかを示す．

Step1 与えられた記事 ID に対応する局所有向グラフ集合を正規インデックスより得る．

Step2 転置インデックスを用いて局所有向グラフ集合内の有向辺，つまり形態素の連接を含む記事 ID の集合を得る．

Step3 **Step2** で得られた記事 ID のうち，ある一つの記事の局所有向グラフ集合を正規インデックスより得る．このときその記事 ID の頻度が stcv 以上のものを対象とする．

Step4 2 つの局所有向グラフ集合の関連度を計算する．関連度の評価値が足切り値 (vcv) を上回ったらその記事 ID と評価値を出力する．

Step5 他にまだ関連度を計算していない記事があれば **Step3** へ，そうでなければ終了．

vcv : Valuation Cut-off Value, 評価値足切り値
stcv : Search Target Cut-off Value, 検索対象候補足切り値

図 2 にシステムの概要を示す．

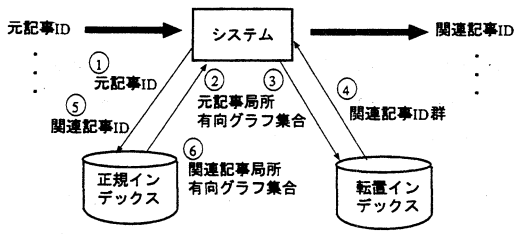

図 2: システムの概要

3 実験

3.1 概要

本手法を Perl 言語を用いて実装し，WWW 上で動作するインタフェースを作成した．その動作の説明ならびに，本手法で提案する関連記事検索と全文検索の比較を行なう．システム動作例を図 3に示す．

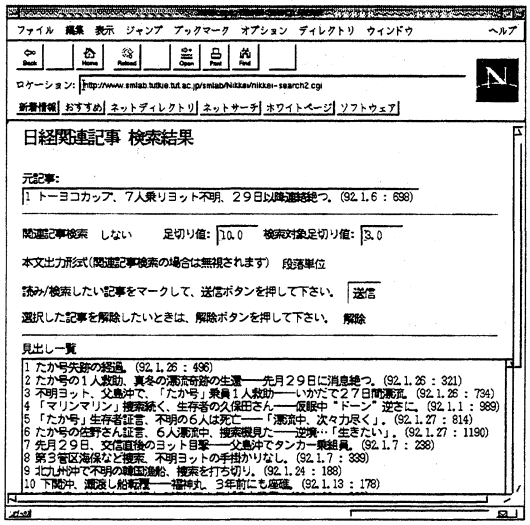

図 3: システム動作例

3.2 実験データ

実験に用いるデータは日本経済新聞 CD-ROM の 92 年版である．このうち 1 月から 4 月の全記事[1]60,017 記事を実験対象データとする．

3.3 手順

1. 正規インデックスの作成
   実験データを JUMAN[4] で形態素解析し正規インデックスを作成した (総エントリ数 60,017)．このとき CNW=0.5 である．
2. 転置インデックスの作成
   正規インデックスと同様に転置インデックスも作成した．このとき IIOCV=200 でインデックスサイズを最適化した (総エントリ数 278,169)．転置インデックスから削除された一つの記事にしか含まれない連接を正規インデックスからも削除した．
   インデックスのサイズは正規 =23.85MB，転置 =40.29MB である．
3. 検索実験
   ユーザからの指定がない場合は vcv=10, stcv=3 で動作するように設定し実験を行なった．

3.4 動作例

作成したインタフェースは日経 CD-ROM 版のキーワードデータを用いる通常のキーワード検索の機能も備えている．このキーワード検索を用いてある記事群を絞

[1]JUMAN で処理可能な記事のみ

り込み，その中から特定の記事(複数でもかまわない)を指定し関連記事検索を行なう．複数の記事が検索の元記事として指定された場合はそれらの記事をひとまとまりとして扱う．関連記事検索の際には2つのパラメータvcv, stcvを指定することができる．vcvは検索結果の表示のための足切り値であり，表示する結果の数に影響を与える．stcvは検索の対象とする記事を決定する足切り値であることから，検索結果に影響を与え，検索時間に与える影響も大きい．

3.5 全文検索との比較

関連記事検索がどの程度漏れなく検索できるかを検討するために全文検索との比較を行なった．手順を以下に示す．

1. 特定の1記事を選択し，関連記事検索を行なう．このとき最大再現率を求めるため，2つのパラメータはvcv=0, stcv=0で検索を行なった．
2. その記事の内容から想定されるキーワードを選び全文検索を行なう．
3. 全文検索結果を正解とし，関連記事検索結果の再現率を求める．このとき，指定した元記事は正解から除く．

3.5.1 記事1

日付：1992.1.1　　文字数：637文字

見出し：南北朝鮮、非核化宣言に合意――査察協定、「北」早期調印を約束。

キーワード	再現率(%)
非核化	91.57
南北朝鮮	97.33

3.5.2 記事2

日付：1992.1.6　　文字数：698文字

見出し：トーヨコカップ、7人乗りヨット不明、29日以降連絡絶つ。[2]

キーワード	再現率(%)
たか号	67.57
トーヨコカップ	100.00

3.5.3 記事3

日付：1992.2.9　　文字数：540文字

見出し：アルベールビル五輪開幕――最多の64カ国・地域から参加。

キーワード	再現率(%)
アルベールビル	50.00
五輪開幕	40.00
冬季五輪	25.64

[2]見出しからでは判断しにくいが「たか号」遭難に関する記事である．

4 考察

結果から判断して記事1のように比較的他の話題と関連しないような記事である場合は検索漏れが少ない．逆に記事3のように非常に多肢にわたる話題の場合は正解とする記事の中に元記事とはそれほど関係ない記事が多く含まれ再現率が低下してしまう．しかし，本手法による通常の使用において最大再現率を求めるような検索を行なうことは考えにくく，また時間もかかる(記事1で約30秒)．通常の利用を想定したパラメータ(vcv=10, stcv=3)では再現率が低下する(記事1の非核化で68.67%)ものの検索時間は短縮される(記事1で約5秒)．このときユーザを満足させる記事が上位に存在しているかが評価基準となる．

本手法を評価する際の最大の問題点は客観的な評価を行ないにくいということである．インタラクションを含めた検索を行ない最終的なユーザの満足が得られるかどうかが重要であり，この評価は今後の課題である．

この実験により本手法がある程度の再現率を保ちつつ，見通しよく関連記事検索を行うものであることがわかった．今回の実験では検索手法の質的面を検証するために検索時間に対して大きな注意をはらわなかったが通常の使用を想定したパラメータでは数秒のオーダーであり，十分実用的であると考える．

5 まとめ

我々は名詞の連接に着目した関連記事検索の一手法を考案し，計算機上へ実装した．WWW上で動作するインタフェースを作成し実験を行なった．実験結果から我々が提案する手法が関連記事を検索するためにある程度の再現率を保ちつつ，見通しよく検索できるものであることを確認した．

謝辞

本研究で，日本経済新聞の一部記事について本稿への引用許可をいただいた(株)日本経済新聞社に深謝する．

参考文献

[1] 武田英明：ネットワークを利用した知的情報統合，人工知能学会誌，Vol. 11, No. 5, pp. 680–688 (1996).

[2] 大竹清敬，山本和英，増山繁：日本語新聞記事を対象とした関連記事検索の一手法，情報処理学会第52回全国大会講演論文集，Vol. 3, pp. 19–20 (1996).

[3] 山田剛一，森辰則，中川裕志：情報検索のための複合語マッチング，情報処理学会研究報告 96-NL-115, pp. 91–97 (1996).

[4] 松本裕治，黒橋禎夫，山地治，妙木裕，長尾真：日本語形態素解析システム JUMAN version 3.1 使用説明書 (1996).