

## 日本語新聞記事を対象とした関連記事検索の一手法

2 B-1

大竹 清敬

山本 和英

増山 繁

{otake,yamamoto,masuyama}@smlab.tut.kie.tut.ac.jp

豊橋技術科学大学 知識情報工学系

## 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。これらの新聞記事の中から必要とする情報を検索するために、従来はキーワードや日付を組み合わせた検索式による検索が主に用いられてきた。

いま、ある検索式を用いて必要とする情報を含む記事を見つけることができたとする。さらにその記事に関連する記事の検索を考えると、関連記事を機械的に検索することはできないので、再度、検索式を用いて検索する。この場合、必要としている情報を有する記事だけに適切に絞りこむには限界がある。また適当な検索式そのものを作成できない場合もある。そこで、本研究では着目している記事に関連する記事を効率良く検索すること、ならびに関連性を基準として記事を順序付けするための一手法を提案する。

関連した研究としては [1] [2] などの類似用例検索があるが、新聞記事のような大量のテキストを対象とした研究はなされていない。また大量のテキストを対象としたテキスト分類の研究として [3] [4] などがある。

類似テキスト検索の手法としてはいくつかのアプローチがあるが、[1] [2] では共にシソーラスを用いている。また、[3] においてはシソーラスを使用してテキストの特徴ベクトルを作成し、関連性の指標としている。[4] においてはテキスト分類を2つの方法、シソーラスを用いた特徴ベクトルと単語間共起による特徴ベクトルを用いてそれぞれ実験している。その結果、精度の点においてシソーラスよりも単語間共起を用いた場合のほうが、良い結果が得られたと報告されている。

本研究では、新聞記事における名詞に着目し、名詞を中心とした単語共起頻度を用いて記事間の関連度を評価する一手法を考案し実験を行った。

## 2 名詞に着目した関連度評価法

## 2.1 評価のためのデータ構造

関連度評価のためのデータ構造について説明する。現実世界の事象と一対一に対応する品詞は名詞であることに着目した。新聞記事において、関連のある記事間では同一の名詞が数多く用いられているという予想が

立つ。そこで新聞記事内の名詞とその前後の語に着目し、名詞を中心としてその前後にどれだけ同じ語が共起しているか、その共起頻度を評価し、記事間の関連度として用いることにする。

まず各新聞記事（見出しも含む）を形態素解析ツール JUMAN を使用して形態素解析しておく。次に形態素解析結果中の名詞に着目し、名詞とその前後の形態素から図 1 に示すような局所有向グラフを以下の手順により作成する。

1. 名詞の前 / 後の形態素が名詞接続助詞（「の」、「や」など）であった場合には、それらを無視し、さらにその前 / 後の形態素を用いる。
2. 名詞を中心とした、その前後の形態素との共起回数を有向辺の重みとして付加する。
3. 必ず名詞である中心の節点（以下中心節点と呼ぶ）に出入りする節点（以下周辺節点と呼ぶ）の形態品詞（JUMAN の辞書による品詞）が「指示詞」、「助詞」、「特殊」（句読点、記号など）であった場合はその周辺節点を削除する。
4. 3つの連続した形態素の品詞が全て名詞だった場合は、1番目の形態素の節点から3番目の形態素の節点へ有向辺を設け、通常の1/2の重みを付加する（図1右）。

検索の元記事として複数の記事を指定することが可能である。このとき指定された記事の局所有向グラフ集合の和集合をとることにより、複数の記事の局所有向グラフ集合を作成する。

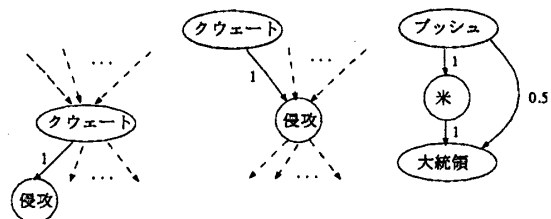


図 1: 名詞の共起情報を利用した局所有向グラフ

この局所有向グラフ集合を用いて関連度を評価する。

## 2.2 関連度評価のアルゴリズム

関連度評価のためのアルゴリズムについて説明する。アルゴリズムを以下に示す。

A Method on Retrieval of Relevant Japanese Newspaper Articles

Kiyonori OHTAKE, Kazuhide YAMAMOTO, and Shigeru MASUYAMA

Department of Knowledge-based Information Engineering, Toyohashi University of Technology

**Step1** 元記事と対象記事グラフ集合において同一の中心節点を持つ局所グラフをそれぞれ選択する。

**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。

**Step3** 元記事のグラフに1番目(これをS1)から3番目の節点(これをS3)への有向辺が存在したならば、対象記事グラフ集合中にS1と同一の中心節点(これをT2)をもつグラフが存在するか調べる。存在したならば、有向辺の向きを考慮しT2の周辺節点にS3と同一の節点の存在を調べる。存在すればそれぞれの有向辺の重みを加算し、評価値に加える。

**Step4** 対象記事のグラフについてもStep3と同様の処理をする。

**Step5** 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。

元記事(複数個でも可)と検索対象の各記事の局所有向グラフ集合に対して、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。

また、計算機実験においては以下の2つのデータをパラメータとしてプログラムに渡す。

1. 記事の長さを先頭(見出しを含む)から文単位で制限するための、文の数。
2. 中心節点が固有名詞の場合にそれに入出力する有向辺の出現1回につき余分に付加する重み。

1は検索の元記事に対して、対象記事の長さによらずにあった場合でも公平に評価するためである。2は中心節点が固有名詞である場合、その名詞は関連性評価上重要でありうるためそれぞれプログラムへ渡すことにした。

### 3 実験

本手法の有効性を確認するために計算機実験を行った。実験のためのプログラムはPerl言語を用いてSun SPARC station I上に実装されている。実験に使用したデータは日本経済新聞CD-ROM 90年版から以下の条件により記事を選択した。

- 期間:7月1日-7月31日
- 各記事に振られたキーワード: 外交

以上により271の記事に絞りこまれ、その中から次の見出しの記事を検索の元記事として選択した。

- 中国・李鵬首相、円借款の凍結解除を歓迎。

この記事に対して、記事の長さ6文、固有名詞に対する重み2で残りの270記事に対し検索した結果上位10記事を表1に示す。

### 4 考察

実験結果から順位付けの正当性の評価はできないが、検索元記事に関連した記事が上位に集中して挙げられていることがわかる。関連記事検索を目的とし、記事の

| 評価値  | 見出し                                   |
|------|---------------------------------------|
| 40   | 日米首脳会談、米大統領、コメ開放決断促す<br>——対中円借款再開を容認。 |
| 35   | 日仏首相会談、仏、厳しい対中認識。                     |
| 34   | 円借款再開、月末にも訪中団<br>——サミット後、解除伝える。       |
| 33   | 円借款再開容認、米、対中改善の糸口に<br>——対象、民生用に限定。    |
| 32   | 対中円借款、独自解除も、首相意向<br>——首脳会談で米に伝達。      |
| 31   | 李鵬首相、インドネシア訪問来月6日から4日間。               |
| 31   | 事務折衝では進展ない、北方領土で海部首相<br>——加首相に表明。     |
| 30.5 | 中国、外交で久々の得点、次はシンガポールに照準。              |
| 23   | 首相、渡辺氏に表明、対中円借款再開へ最大限の努力。             |
| 22   | インドネシア外相きょう訪中——人民日報、歓迎の論評。            |

表1: 関連記事検索結果上位10記事の見出しと評価値

表層的情報を用いる本手法の有効性を確認することができた。しかし現段階では順位付けの正当性の評価を行うことは困難である。

また、評価値による関連記事の順序付けは可能だが、検索元記事に対する関連の有無を機械的に判断することはできない。しかし、その他の数々の実験結果より評価値の減衰状況によってそれが判断できるのではないかと推測している。

### 5 まとめ

日本語新聞記事における関連記事検索の一手法を考案し、実験を行いその有効性を確認した。今後はより高精度な検索のためにパラメータの与え方を再検討する。順位付けの正当性を検証するためにアンケートによる調査を行なう。考察でも述べたが、関連の有無の判別も重要であり、これを早急に実現する予定である。これが、実現されると本手法をクラスタリングによる自動文書分類へ応用することが容易となる。また、現段階では元記事が複数個の場合の検索も可能であり、検索結果をフィードバックし、さらに検索することが可能である。しかし、マンマシンインタフェースが不十分であるため、その機能を十分に活用できない。そこで、検索・閲覧が容易なGUIを構築し、評価を順次行う予定である。

### 参考文献

- [1] 岡本育史, 佐藤健, 堀順吉, 松尾勝美: 類似事例検索システムー通信ソフト故障診断問題への適用ー, 情報処理学会第51回全国大会講演論文集, Vol. 3, pp. 219-220 (1995).
- [2] 宇津呂武仁: 類似度テンプレートを利用した高速類似用例検索, 情報処理学会研究報告 94-NL-103, pp. 33-40 (1994).
- [3] Yamamoto, K., Masuyama, S. and Naito, S.: Automatic Text Classification Method with Simple Class-Weighting Approach, *Proc. of NLP'95*, Vol. 1, pp. 498-503 (1995).
- [4] 湯浅夏樹, 上田徹, 外川文雄: 大量文書データ中の単語間共起を利用した文書分類, 情報処理学会論文誌, Vol. 36, No. 8, pp. 1819-1827 (1995).