自然言語処理 122－12
(1997. 11. 20)

名詞を中心とした連接に着目した新聞の関連記事検索手法

大竹清敬* 増山 繁* 山本 和英**
{otake, masuyama}@smlab.tutkie.tut.ac.jp yamamoto@itl.atr.co.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

概要

本稿では，名詞を中心とした連接に着目した新聞の関連記事を検索する手法を提案する．従来の検索手法では名詞を中心にキーワードとなる可能性のある単語を抽出し，重みをつけ，その重みを点数化して1記事の特徴を表現するものが多い．そして，記事間で共起した単語の点数の総和により記事間の関連度を評価する．本稿で提案する手法は，単語単位だけではなく，名詞を中心としてその前後の単語の連接を考慮する．連接は名詞を中心とした局所有向グラフ上の有向辺として捉えられる．各連接にTF·IDF法を用いて重みづけをし，記事の特徴を捉える．本手法を用いて実験を行なった結果，元記事以後16日間の範囲に存在する関連記事を適合率86%，再現率84%で検索できた．

A Retrieval Method of Relevant Newspaper Articles Focusing on Noun Connections

Kiyonori OHTAKE* Shigeru MASUYAMA* Kazuhide YAMAMOTO**

* Dept. of Knowledge-based Info. Eng., Toyohashi Univ. of Tech.
** ATR Interpreting Telecommunications Research Laboratories

Abstract

This paper proposes a retrieval method of relevant Japanese newspaper articles by focusing on consecutive co-occurrences of nouns. Most of these retrieval methods proposed up to the present employed word weighting approach. In this paper, we consider not only word weighting but also noun connections. We employ oriented graphs as a data structure for expressing characteristics of an article in order to capture relevancy among articles. We have implemented a prototype system of retrieving relevant articles useing this data structure. This method retrieves relevant articles within 16 days since the original article was printed. As a result, its precision attains 86%, and its recall attains 84%.

1 はじめに

現在，大量の機械可読文書(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，検索需要が高い．しかし，読者が興味を持った記事の関連記事を検索すること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．

このような手法として，単語ベクトルを用いたベクトル空間モデルに基づく手法や，利用者が選択した適合情報を検索質問に反映させる関連性フィードバックに基づく手法が研究されている．これらはいずれも何らかの検索質問を入力し，その質問に最も適合する文書を検索する手法である．しかし，本研究が目的としているのは明示的な検索要求文を入力または利用することなしに，元記事に関連する記事を検索する手法の提案である．また，本研究に直接的に類似した研究として[1]がある．以上の研究のうちのほとんどが，扱う単位として形態素を用いている．しかし，[2]において，日本語は，複合語が多く現れる言語であり，語が複合しているという情報の利用がより高精度な検索を実現するために必要であるとしている．そして，複合語を意識した場合の方が再現率，適合率とも改善されることが確認されている．

この結果は，従来手法が形態素を単位としてモデリングされていることに大きく起因する．日本語のような膠着言語においては，単語の区切りを判断することは困難であり，そのため形態素が処理単位として用いられてきた．しかし，形態素を単位として扱う手法にも限界がある．テキスト内の形態素に対していかなる重みづけを行なっても，それらは別個の単位として扱われる．そのため，検索において再現率を上げようとする場合には余計な形態素を多く抽出し，適合率が大きく下がることが考えられる．

そこで，形態素より大きく，そのテキストをさらに特徴づけることが可能な単位を考える必要がある．本研究では従来からキーワードの候補とされてきた名詞に着目すると共に，その前後に存在する形態素との連接を一つの単位として扱う手法を提案する．この手法を用いて [1] との比較実験を行なったので報告する．

2 名詞の連接に着目した関連記事検索手法

各記事の特徴を抽出するために形態素解析を行なう．形態素解析には現在広く利用されているJUMAN[3] を用いる．この時，複合語は個々の構成要素である形態素へと分割される．どのような検索モデルを用いるにしろ，個々の形態素をその単位として検索を行なう場合，ノイズが大きくなることが予想できる．そのため，本手法では記事内の名詞とその前後の形態素の連接に着目し，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，その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. 中心となる節点は各局所有向グラフにただ一つ存在し，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．

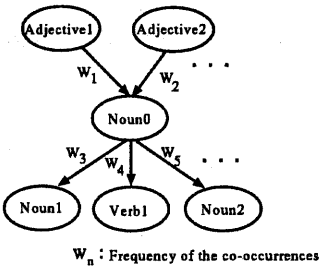

図 1: 局所有向グラフ

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，ファイル内の各名詞に着目する．それらの名詞の前後の形態素からその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．また各局所有向グラフを識別するものは中心節点の名詞である．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．

2.2.1 ヒューリスティックス

局所有向グラフ集合を作成する際，検索精度をより向上させるために用いたヒューリスティックスは以下のとおりである．

1. 省略
   名詞間に存在する名詞接続助詞の「の」，読点「、」,「・」は存在しないものとして扱われる．
2. 名詞の 3 連接
   名詞が連続して 3 回出現する場合 (“A/B/C”)(/ は連接を表わす) には A と C も連接していると考える．そして “A/C” によるグラフも作成する．
3. 括弧
   括弧の「(」,「)」が出現する場合，“A / (/ B /) / C ”，この時 “A / C” および “B / C” としてグラフを作成する．
4. 見出し
   見出しは記事を検索する際の重要な手掛りとなる．そのため見出し内の名詞をその頻度とともにリストとして各記事のデータとする．

2.3 局所有向グラフ集合間の関連度

ここでは，2つの局所有向グラフ集合間の関連度をどのように評価するのかについて説明する．

2.3.1 有向辺の重みづけ

局所有向グラフ上の有向辺には有向辺が表わす連接の頻度が与えられている．重みづけの手法としてH.P. Luhn[4] の TF 法や K.S. Jones[5] の IDF 法などがある．本手法において関連度を評価する際には，記事 d 中の連接 c の重みを $TF \cdot IDF$ を応用した，次の関数 $W(d,c)$ で計算する．

$$W(d,c) = \frac{TF(d,c)}{\sum_{c_0 \in d} TF(d,c_0)} \times \log(M/df(c)) \quad (1)$$

$TF(d,c)$: 記事 d 中の連接 c の出現回数
M : 統計サンプルに含まれる文書の総数
$df(c)$: 統計サンプルの中で連接 c を含む文書数

2.3.2 見出しの名詞の重みづけ

次に記事 d の見出しに出現する名詞 h を次の関数 $H(d,h)$ を用いて重みづけする．

$$H(d,h) = TF(d,h)/num_h(d) \quad (2)$$

$TF(d,h)$: 記事 d の見出し中の名詞 h の見出し内の出現回数
$num_h(d)$: 記事 d の見出しに含まれる名詞の数

2.3.3 関連度評価

記事 d_x, d_y 間の関連度関数 $R1$ を式 (3) で定義し，$R1$ が閾値 b を越えるものを関連記事とする．

$$R1(d_x,d_y) = \frac{\sum_{c_{x\cap y}} W(d_x,c_{x\cap y}) + \sum_{ON(G_x,G_y)} \beta}{\sum_{c_x} W(d_x,c_x)} \times \frac{\sum_{c_{x\cap y}} W(d_y,c_{x\cap y}) + \sum_{ON(G_x,G_y)} \beta}{\sum_{c_y} W(d_y,c_y)} + \alpha \left(\sum_{h_{x\cap y}} H(d_x,h_{x\cap y}) \times \sum_{h_{x\cap y}} H(d_y,h_{x\cap y}) \right) \quad (3)$$

G_x : d_xの局所有向グラフ集合
G_y : d_yの局所有向グラフ集合
c_x : $c_x \in G_x$である連接，　c_y : $c_y \in G_y$である連接
$c_{x\cap y}$: $c_{x\cap y} \in (G_x \cap G_y)$
$h_{x\cap y}$: $h \in (d_x \cap d_y)$ の見出しの名詞
$ON(G_x,G_y)$: G_x, G_yのうち中心節点のみが同一であるグラフの数
α : 見出しの重要度を表わすパラメータ
β : 中心節点だけが同じグラフを評価するためのパラメータ

2.4 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存在する有向辺 (形態素の連接) 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．

2.5 関連記事検索

ここでは，実際の関連記事検索の手順を示す．

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 から一つを選択し，その局所有向グラフ集合を正規インデックスより得る．

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 (b) 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．

図 2 にシステムの概要を示す．

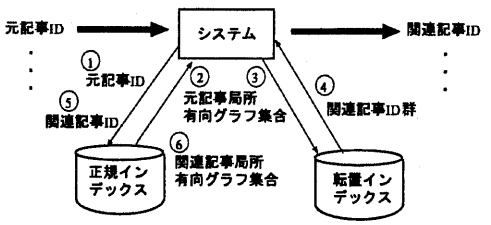

図 2: システムの概要

3 実験

日本経済新聞 92 年度 1，2 月に掲載された記事から 8 記事 (元記事) を選び，それぞれに関連する記

事 (記事群) を検索する実験を行なった．関連記事は，それぞれの元記事の掲載日以後 16 日以内の範囲の全紙面から検索した．本手法におけるパラメータ α, β をそれぞれ，$\alpha = 5, \beta = 2$ として実験した．

3.1 比較手法

語の連接を意識した本手法と従来手法とを比較するために [1] で提案されているフェルマー型曲線による重みづけを用いた手法を比較手法とする．比較手法は具体的には次の 3 つの重みづけ関数を組み合せることにより実現される．

3.1.1 出現回数による単語の重みづけ

記事中に出現する単語[1] ほどその記事の内容をよく表わすので，そのような単語に大きな重みを与える重みづけを行なう．重みづけ関数 $F(d, t_0)$ を次式で定義する．

$$F(d, t_0) = \frac{TF(d, t_0)}{\sum_{t \in d} TF(d, t)}$$

$TF(d, t)$: 記事 d 中の単語 t の出現回数

3.1.2 出現位置による単語の重みづけ

新聞記事では重要な内容ほど最初に記述されることを考慮して,「記事の最初の方に出現する単語ほど重要であり，記事の最後の方になって初めて出現する単語ほど重要でない」と見なす重みづけを行なう．記事 d 中の単語 t_0の出現位置による重みづけ関数 $L(d, t_0)$ を次のように定義する．

$$L(d, t_0) = 1.0 - \frac{x_{t_0} - 1}{X}$$

X : 記事 d の本文 (見出しとリード除く[2]) 中の文数
x_{t_0} : 名詞 t_0の初出現文の位置

3.1.3 フェルマー型曲線による重みづけ

新谷ら [1] は単語の統計サンプル中の出現回数そのもので評価するのではなく，統計サンプル中で最も出現回数の高い (最もありふれた) 単語の出現回数に対する比率によって評価することを考案した．単語 t_0の一般的な使用頻度による重みづけ関数 $W_a(t_0, a)$ は次式で定義される．

$$W_a(T_0, a) = \left\{ 1 - \left(\frac{TF(N, t_0)}{TF(N, t_{max})} \right)^a \right\}^{1/a}$$

N : 統計サンプルとなる大量の文書
t_{max} : N中で最大の出現回数の単語
$TF(N, t)$: N中の単語 t の出現回数
a : 曲線次数 (実験により決定[3])

3.1.4 統計サンプル

単語 t の一般的な使用頻度による重みづけ回数を計算するには，統計サンプル N中の単語 t の出現回数 $TF(N, t)$ が必要である．そこで，91 年度の日経新聞の記事 1 年分を日本語形態素解析システム JUMAN[3] で解析し，各単語の出現回数を計算し，$TF(N, t)$ とした．その中で名詞群の総数は 16,887,379 であり，最も出現回数の多かった単語 t_{max} は “日本 (固有名詞)” で回数は 95700 だった．

以上，3 つの関数を組みあわせた，比較手法は

$$\begin{aligned} V_a(d, t_0, a) &= F(d, t_0) \times L(d, t_0) \times Wa(t_0, a) \\ &= \frac{TF(d, t_0)}{\sum_{t \in d} TF(d, t)} \times \left(1.0 - \frac{(x_{t_0} - 1)}{X} \right) \\ &\quad \times \left\{ 1 - \left(\frac{TF(N, t_0)}{TF(N, t_{max})} \right)^a \right\}^{1/a} \end{aligned}$$

という関数によって各単語の重みを計算する．

3.1.5 特徴語リスト

一記事ごとに各出現単語とその点数化された重みとを記録したリスト (以下，特徴語リストとする) を作成し，それを用いて記事間の関連度を評価する．その作成方法はまず，形態素解析によって得られた名詞群から上述の関数 $V_a(d, t_0, a)$ によって各単語の点数が求まる．各記事ごとに点数の高い上位 25%のみ[4]を記録しこれを特徴語リストとする．

3.1.6 関連記事評価基準

与えられた 2 つの記事が関連記事か否かの判断を，それぞれの特徴語リスト中に含まれる共起単語の点数の総和により行なう．記事 d_x，d_y間の関連度関数 $R2$ を式 (4) で定義し，$R2$ が閾値 b を越えるものを関連記事と見なすことにする．ただし，s_x，s_yはそれぞれ d_x，d_yの特徴語リストである．

$$\begin{aligned} R2(d_x, d_y) &= \frac{\sum_{t_{x \cap y}} V_a(d_x, t_{x \cap y}, a)}{\sum_{t_x} V_a(d_x, t_x, a)} \\ &\quad \times \frac{\sum_{t_{x \cap y}} V_a(d_y, t_{x \cap y}, a)}{\sum_{t_y} V_a(d_y, t_y, a)} \end{aligned} \tag{4}$$

$t_x = t \in s_x, \quad t_y = t \in s_y, \quad t_{x \cap y} = t \in (s_x \cap s_y)$

[1] ここでの単語は「サ変名詞」「普通名詞」「固有名詞」「未定義語」の形態素である．
[2] 見出しとリード部では $L(d, t_0) = 1.0$ とする．
[3] 今回の報告では [1] にならい 0.3 とした．
[4] [1] によると予備調査の結果，記事を特徴づける形態素の大部分が上位 25%以内に含まれることが分かったとある．

3.2 実験対象記事

8記事群のうち，パラメータを決定するための4つの記事群(以下，学習グループとする)と，決定したパラメータを用いて実際に評価するための4つの記事群(以下，テストグループとする)とに分けた．

3.2.1 学習グループ

記事群A 「たか号」の遭難について，生存者の証言による事故原因や生存者の回復などに関する記事

記事群B 朝鮮半島の非核化にむけての南北朝鮮の交渉に関する記事

記事群C 「＊＊急便」が絡んだ汚職事件に関する記事

記事群D 第二次世界対戦中の日本植民地下での従軍慰安婦問題，ならびに宮沢首相訪韓による慰安婦問題の動きに関連する記事

3.2.2 テストグループ

記事群E ニューヨークの国連本部で開催された安保理サミットに関する記事

記事群F 脳死臨調の答申が発表されてからの議論や各大学の対応に関する記事

記事群G 東京都＊＊区の医師が誘拐された事件で，人質の解放から犯人の指名手配，逮捕に関する事件

記事群H 鉄骨会社「＊＊」に絡む汚職事件のうち，＊＊元長官逮捕に関する記事

3.3 評価方法

各手法の検索精度を，その手法を用いた検索結果と，あらかじめ筆者が人手で抽出しておいた各記事群ごとの関連記事との整合性により評価した．なお人手で抽出する際は，筆者の主観に従い各々の記事について最も重要なトピックが元記事のそれと一致する記事を関連記事と判断し，抽出した．抽出する際には日経CD-ROMに附属のキーワードデータを用いたキーワード検索ならびに全文検索を用いて，なるべくもれのないようにした．

3.3.1 評価関数

本稿では[1]にならい，記事群Xにおいて閾値bで検索した場合の精度を以下のように評価する．

$$\text{適合率}: pre(X,b) = \frac{\text{検索した中で正解の記事数}}{\text{検索した記事数}}$$

$$\text{再現率}: rec(X,b) = \frac{\text{検索した中で正解の記事数}}{\text{正解の記事数}}$$

表1: 比較手法の検索精度(学習グループ)

記事	適合率	再現率	平均
A	100%(22/22)	76%(22/29)	88%
B	87%(20/23)	77%(20/26)	82%
C	98%(40/41)	65%(40/62)	81%
D	100%(25/25)	81%(25/31)	90%
平均	96%	74%	85%

表2: 比較手法の検索精度(テストグループ)

記事	適合率	再現率	平均
E	55%(23/42)	64%(23/36)	59%
F	98%(42/43)	95%(42/44)	97%
G	100%(22/22)	59%(22/37)	80%
H	89%(48/54)	62%(48/77)	76%
平均	85%	70%	78%

また，それらの平均$val(X,b)$を定義する．

$$val(X,b) = (pre(X,b) + rec(X,b))/2$$

3.4 結果

3.4.1 比較手法の結果

比較手法の学習グループにおける検索精度を表1に示す．閾値$b = 0.06$のとき$val(X,b)$の相加平均が最大となった．

次に比較手法のテストグループにおける検索精度を表2に示す．

3.4.2 本手法の結果

本手法の学習グループにおける検索精度を表3に示す．閾値$b = 0.5$のとき$val(X,b)$の相加平均が最大となった．

次に本手法のテストグループにおける抽出精度を表4に示す．

4 考察

比較実験の結果から本手法が比較手法に比べて高い再現率を実現しているにもかかわらず，適合率がそれほど低下しない手法であることが分かる．それは形態素を単位として扱う手法の場合，再現率をあげようと閾値を変更した場合，目的の検索にとってノイズとなる形態素を多く抽出する．そのため適合率の低下率が再現率の上昇率より大きくなることが考えられる．それに比べて本手法では形態素が連接しているという情報を用いているため，再現率を高くする閾値を設定しても，それほど適合率が下がら

表 3: 本手法の検索精度 (学習グループ)

記事	適合率	再現率	平均
A	73%(29/40)	100%(29/29)	86%
B	80%(24/30)	92%(24/26)	86%
C	91%(57/63)	92%(57/62)	91%
D	62%(31/50)	100%(31/31)	81%
平均	76%	96%	86%

表 4: 本手法の検索精度 (テストグループ)

記事	適合率	再現率	平均
E	63%(33/52)	92%(33/36)	78%
F	98%(39/40)	89%(39/44)	93%
G	90%(28/31)	76%(28/37)	83%
H	91%(62/68)	81%(62/77)	81%
平均	86%	84%	85%

ない．つまり連接を単位とすることにより，ノイズとなる情報を抽出しないからである．

今回の評価基準 (適合率と再現率の平均) によると，比較手法は再現率に比べて適合率が高くなる傾向があり，本手法はその反対であるといえる．それは学習グループにおける結果に顕著にあらわれている．このことから本手法が再現率を上げるために検索の範囲を広げる際，極端に拡大しないという特性を持つことがが確認できる．

両者を比較するために，図 3にテストグループにおいて閾値を変化させた場合の適合率と再現率の平均の関係を示す．図からも分かるように，本手法の方がより高い適合率，再現率を示している．

また，本手法の難点としては，処理が繁雑になるため，処理時間がかかるという点がある．それに，局所有向グラフを収めるインデックスのサイズが大きいということがある．処理時間に関しては，精度を多少犠牲にするが，転置インデックスを引いたときに，いくつの連接が共起しているかによって足切りすると検索速度が著しく改善することが分かっている．さらに，現状では，実験した記事数が少ないので，今後はこれを多くし，様々な記事について，その検索結果にどのような違いがあるか検討する必要がある．

5 まとめ

我々は名詞を中心としてその前後の連接に着目した関連記事検索の一手法を考案した．元記事以後 16 日間の範囲で検索する実験を行なった結果，その精度は適合率 86%，再現率 84%であった．本手法を形態素のみに対する重みづけを行なう手法と比較した結果，連接に着目する本手法がより高い精度で検索できる手法であることを確認した．

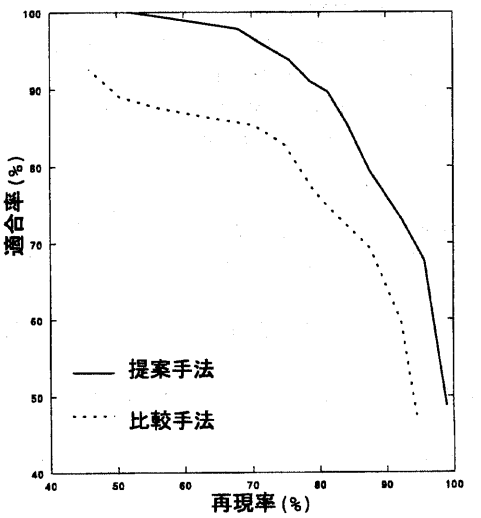

図 3: 適合率と再現率の平均の関係

参考文献

[1] 新谷研, 角田達彦, 大石巧, 長尾眞：単語の共起頻度と出現位置による新聞関連記事の検索手法, 情報処理学会論文誌, Vol. 38, No. 4, pp. 855–862 (1997).

[2] 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).

[3] 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).

[4] Luhn, H. P.: A Statistical Approach to Mechanized Encodeing and Searching of Literary Information, *IBM Journal*, No. October, pp. 309–317 (1957).

[5] Jones, K. S.: A Statistical Interpretation of Term Specificity and its Applilcation in Retrieval, *Journal of Documentation*, Vol. 28, No. 1, pp. 11–21 (1972).