名詞の連接情報を用いた関連文書検索手法

大竹清敬[†] 増山繁[†] 山本和英[††]

名詞の連接に着目した関連文書検索手法を提案し，実験により評価を行った．本論文で提案する手法は，ベクトル空間法に基づき，索引語の単位として名詞の連接を用いるという点が特徴である．情報検索において，ある1つの事象を示すために様々な名称を用いることによる精度の低下という問題がある．また，日本語の文書には多くの複合語が見られ，これが検索精度低下の一因となっている．本論文では，適合率を向上させるために，名詞の連接を索引の単位として用いることを提案する．また複合語における表記のゆれを吸収する経験則を導入することにより再現率が向上することを示す．日本経済新聞を対象として，単語のみに着目する従来手法との比較実験を行った．その結果，F 値の平均が，比較手法 76.2%，提案手法 85.9%となり，本手法の有効性を確認した．

A Retrieval Method for Relevant Documents Employing Connective Information of Nouns

KIYONORI OIITAKE,[†] SHIGERU MASUYAMA[†] and KAZUHIDE YAMAMOTO[††]

We propose a retrieval method of relevant Japanese documents by employing information on noun connections. The method is based on the vector space model, but the method employs noun connections as indexing terms. On the part of information retrieval, there is a problem that we can use various nouns to express a phenomenon. The problem causes a decline of precision in information retrieval. In addition, Japanese texts have many compound nouns and these nouns may be hindrance to retrieve, because these nouns may cause fluctuation. In this paper, we propose employing noun connections as indexing terms to improve the precision. And we show that the recall raised by heuristics solving fluctuation of compound nouns. We carried out experiments with comparing a word-based method and the proposed method for a Japanese newspaper (The Nihon Keizai Shimbun). The experimental results show that the proposed method attains 85.9% F-measure on the average which is approximately 10% higher than that of a conventional word-based method.

1. はじめに

近年，大量の機械可読文書（コーパス）が利用可能となっている．このような機械可読文書に対する情報検索手法は古くから研究され，また実用化されてきた[1]．従来用いられてきた手法の多くは，利用者からの検索質問に基づいて検索を行っている．

一方，情報検索において，検索要求に関連する文書を我々がすでに所持している場合や，キーワード検索の途中で関連する文書を発見した場合がある．このような場合，多くの従来手法では求める文書を検索するために，利用者が検索質問を入力しなければならない．つまり，「手元にある文書に関連する文書」という検索要求に対して，具体的な検索質問を作成しなければならない．これは利用者にとって煩雑であり，関連文書の要求は頻繁に起こるため，検索質問の入力は情報検索の効率を下げる要因になりうる．そこで，利用者がキーワードなどを入力することなく，所持している元文書をそのまま利用し，その関連文書を検索する手法が求められる．つまり，「手元にある文書に関連する文書」という検索要求に対して，検索を自動化する手法が必要である．

関連する文書を検索するための手法としては，ベクトル空間法やクラスタリングによる手法がある[1)~5)]．これらの手法は，文書の内容を索引語の集合でとらえる．そして，検索質問に含まれる索引語とそれらの索

† 豊橋技術科学大学知識情報工学系
Department of Knowledge-based Information Engineering, Toyohashi University of Technology
†† ATR 音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

引語集合を比較することによって，最も類似した文書を抽出する．これらの手法を日本語文書に対して導入する際の問題点は，分かち書きをしない日本語では，単語の定義を明確にできず，索引語として用いる単位を決定できない点である．

分かち書きをする英語において，索引語として単語のみを用いる場合と，名詞句を用いる場合の比較がEvansらによって行われている[6]．そこでは，名詞句に基づく検索が，適合率・再現率ともに，単語のみを用いる場合をわずかながら上まわるという結果が報告されている．また，文書を一定の長さに区切って扱うpassageを用いた研究がある[7]．passageを用いる場合は，passageをどのように定義するかが問題となる．文献8)では語彙的連鎖を用いてpassageを決定するアプローチをとっており，その有効性を検討している．しかしながら，日本語における一般的なpassage決定の手法は確立していない．

従来，日本語を対象としたベクトル空間法では，一定の索引語単位を扱ってきた．たとえば，形態素解析結果における形態素であるとか，文字種切りによる文字列などである．しかし，日本語では，複合語による様々な名詞が存在することから，これらを柔軟にとらえる必要がある．なぜならば，「自然言語処理」という複合語と，「ある言語におけるその処理は自然」という文には「自然」，「言語」，「処理」という3つの形態素が出現するが，それぞれ，異なる意味を形成するからである．つまり，単純にそれぞれの形態素をとらえるのではなく，その局所的な接続や語順などを考慮する必要がある．さらに，接続を考慮しても形態素の省略や，助詞の省略などによって，照合を行う際に不一致となる場合がある．そこで，本論文では名詞を中心とした形態素の連接を索引語の単位とし，形態素の省略などによる表記のゆれに対応した手法を提案する．

先行類似研究として，中川らの研究[9]がある．中川らは単名詞の前方連接数と後方連接数に着目した重要度計算法を提案し，複合語の索引語を自動抽出している．索引語を自動抽出した場合，その索引語の正当性が評価されなければならない．つまり，抽出した索引語が元の文書に対して適当であるかどうかが問題となる．我々は検索の際に手がかりとなる名詞の連接に着目するが，形式的に連接している事実のみを検索に用いる．そのため，形態素の連接が文書に対する索引語として適当であるかどうかを問題としない．

また，関連文書を検索するために，構文解析を用いる手法も考えられる．麻生ら[10]は，日本語を対象として係り受け解析を利用した手法を提案している．構文解析を利用すると，検索精度，特に適合率の向上が期待できる反面，曖昧性，頑健性，および解析時間などの点が問題となることが多い．そこで，本研究は大量文書を対象とし，高い頑健性を実現するために構文解析は用いない．

本論文では関連文書を検索するために，ベクトル空間法に基づき，索引語の単位として名詞を中心とした形態素の連接を用いる手法を提案する．さらに，関連文書検索に対する名詞の連接情報の有効性を確認するために，索引語として形態素のみを用いた手法[11]（以下比較手法）と本手法との比較実験を行い，評価した．その結果，本手法が比較手法に比べて適合率，再現率ともに良好な結果を示し，名詞の連接情報を用いることの有効性を確認した．

2. 関連文書検索手法

関連文書を検索するための手法について説明する．まず，各文書内の連接を抽出するために形態素解析を行う．以下，連接とは形態素の連接を指す．

2.1 名詞に着目した連接

名詞を中心とした連接として以下の4種類の型を導入する．本論文では，以下の4種類の連接に対して考慮する．

MN型 形容詞のあとに名詞が続く連接．この型の連接が出現する頻度は比較的少ないが，この連接が文書間で共起している場合，関連性が高いと考え導入する．

例：具体的な–措置，大規模な–援助

NN型 名詞のあとに名詞が続く連接．いわゆる複合語を扱うために導入する．4種類の中で最も頻繁に出現する連接である．

例：原子力–機関，大統領–選挙，環境–整備

NV型 名詞のあとに動詞が続く連接．サ変名詞が名詞として用いられているのか，動詞として用いられているのかを区別するために導入する．なお，動詞はその終止形を対応させる．

例：寄与–する，確認–する，開催–する

NP型 名詞のあとに句点「.」が続く連接．目的はNV型と同一である．名詞に連接する動詞が省略され，いわゆる体言止めとなる場合は，NV型の連接として扱えない．そのため，これをNV型と同様の目的で導入する．

2.2 連接の集合

各文書の形態素解析結果から名詞に着目したそれぞれの連接を抽出し，連接の頻度を求める．これにより，

各々の連接はその頻度を持つことになる．この文書から得られた連接の集合とその頻度をまとめて文書に対応する連接の集合と呼ぶ．

2.2.1 ヒューリスティックス

連接の集合を作成する際，検索精度を向上させるために，以下に示すヒューリスティックスを用いた．

(1) 名詞間に存在する「の」，読点「,」，なかぐろ点「・」を無視し，それらの両側にある名詞は連接しているとする．一般に，「の」が名詞間に存在する場合，名詞句を構成する．直観的に，このヒューリスティックスが扱うのは，影山[12]が言うところの語$^+$（word plus）と，名詞句の表現の違いの吸収である．なお「,」や「・」は実際に省略される例を観察したので，このヒューリスティックスとして導入する．

(2) 名詞が連続して3回出現する場合，“$N_1/N_2/N_3$”（/は形態素境界を示す）にはN_1がN_3にも連接しているとする．つまり，この3連接からはN_1–N_2，N_2–N_3という連接のみならず，N_1–N_3という連接も存在するとする．ただし，(1)の結果として出現した3連接に対しては適用しない．

(3) 括弧の「(」，「)」が出現する場合，たとえば“$M_1/(/M_2/)/M_3$”の場合には，別の文書でこれがM_1/M_3やM_2/M_3と表現されることがある．この観察より，括弧が出現した場合には「(」の直前の形態素が「(　)」で囲まれた部分と置き換え可能であるとして連接を構成する．つまり，上記の連接はM_1–M_3，および，M_2–M_3という2連接と見なす．

(4) 文書に見出しが存在する場合，見出しはその文書を検索する際の重要な手掛りとなる．そのため，見出しが存在する文書については，見出し中の名詞とその頻度を抽出する．この見出しから得た情報（以下，見出し情報）と，記事全体に対応する連接の集合は区別して扱う．

2.3 文書間の関連度評価

本節では，2文書間の関連度の評価方法について述べる．

2.3.1 連接の重みづけ

連接の集合内の要素にはそれぞれ頻度が与えられている．頻度が高い連接が重要であるとは限らず，頻度をそのまま重要性の指針として用いることはできない．そのため，従来から頻度に基づきつつも，重要度を考慮した数々の重みづけ手法が提案されている[3),11),13),14)]．本論文では，一般的な$TF \cdot IDF$法[4),13),14)]を用いて，連接に重みづけを行った．文書xにおける連接cの重みを，次の関数$W(x,c)$によって定義する．

$$W(x,c) = \frac{TF(x,c)}{\sum_{c_0 \in SN_x} TF(x,c_0)} \log\left(\frac{M}{af(c)}\right) \quad (1)$$

$TF(x,c)$：文書x中の連接cの出現回数
SN_x：文書xに対応する連接の集合
M：統計サンプルに含まれる文書の総数
$af(c)$：統計サンプル中で連接cを含む文書数

2.3.2 見出し中の名詞の重みづけ

2.2.1項で述べたように，見出し情報は，文書に対応する連接の集合とは区別される．連接に対する重みづけと同様に，この見出し中の名詞についても次の関数$H(x,h)$によって重みづけを行う．

$$H(x,h) = \frac{TF_h(x,h)}{num_h(x)} \quad (2)$$

$TF_h(x,h)$：文書xの見出し中の名詞hの見出し内での出現回数
$num_h(x)$：文書xの見出しに含まれる名詞の総数

2.3.3 関連度評価

文書x，yに対応する連接の集合ならびに見出し情報が与えられたときに，x，y間の関連度を式(3)の関数Rで求める．このRが閾値θを超えるものを関連文書とする．

$$R(x,y) = \frac{\sum_{c_{x\cap y}} W(x,c_{x\cap y}) + \beta \cdot CON(SN_x, SN_y)}{\sum_{c_x} W(x,c_x)} \cdot \frac{\sum_{c_{x\cap y}} W(y,c_{x\cap y}) + \beta \cdot CON(SN_x, SN_y)}{\sum_{c_y} W(y,c_y)} + \alpha\left(\sum_{h_{x\cap y}} H(x,h_{x\cap y}) \sum_{h_{x\cap y}} H(y,h_{x\cap y})\right) \quad (3)$$

x，y：文書
SN_x，SN_y：文書x，yに対応する連接の集合
c_x，c_y：$c_x \in SN_x$，$c_y \in SN_y$である連接
$c_{x\cap y}$：SN_xとSN_yに共通して含まれる連接☆
$CON(SN_x, SN_y)$：$(SN_x - SN_y)$と$(SN_y - SN_x)$

☆ 共通した連接とは連接を形成する2つの形態素が完全に一致した連接を指し，頻度は考慮しない．

に共通して含まれる名詞の数

$h_{x\cap y}$：文書 x の見出しと文書 y の見出しに共通して含まれる名詞

α：見出し情報の重要度を示す変数

β：$CON(SN_x, SN_y)$ の重要度を示す変数

以上の関数によって，2つの文書間の関連度を計算する．この関数の第1項は文書 x の連接の集合のうち，どの程度の連接が文書 y の連接の集合と共通しているかをその重要度について求め，同様に文書 y についても x と共通している連接の重要度の割合を求めたものを掛けたものである．また，第1項に含まれる $CON(SN_x, SN_y)$ によって，共起する連接は構成しないが，形態素単体で共通して含まれる名詞を評価するため，本手法は従来用いられてきた形態素のみを索引語とする手法を包含している．そして，第2項は文書の見出しに共通して含まれる名詞についてその重要度を掛けたものである．

3. 実　　験

本論文で提案した手法を計算機上に試作システムとして実装し，実験を行った．実験に用いた計算機はCPUがDual PentiumII 300 MHz，メモリが128 MBのPCである．試作システムはこのPC上にPerl言語を用いて実装した．

3.1 実験対象

新聞記事は，現代社会における大量情報の流通媒体であることから，非常に検索需要が高い．このことから，実験対象として日本経済新聞CD-ROM 92年度版1，2月の全記事（総数28,588記事）を用いた．

3.2 インデックスの作成

まず各記事をJUMAN[15]を用いて形態素解析し，形態素解析結果から連接の集合を作成した☆．

形態素の連接とその頻度を組としたリストを作成し，インデキシングを行った．インデックスはハッシュデータベースを用いて構成され，2種類存在する．1つは連接の集合を格納し，もう1つはそのインデックスに対する転置インデックス[4]である．連接の集合を格納するものを正規インデックスと呼ぶ．正規インデックスは記事IDをキーとして，記事に対応する連接の集合と見出し情報を格納する．転置インデックスは，1つの連接をキーとし，その連接が含まれる記事ID群をデータとして格納する．実験対象に対して両インデックスを作成したところ，それらの容量は正規インデックス18.5 MB，転置インデックス19.9 MBである．

3.3 関連記事検索手順

実際の関連記事検索手順を以下に示す．

入力：　元記事ID

出力：　関連記事群ID

Step 1　与えられた記事IDに対応する連接の集合を正規インデックスから得る．

Step 2　Step 1で得た連接の集合に含まれるすべての連接について転置インデックスを参照し，連接の集合中の各連接が含まれる記事ID群を得る．

Step 3　Step 2で得られた記事ID群から1つを選択し，その記事IDに対応する連接の集合を正規インデックスから得る．

Step 4　Step 1，Step 3で得た連接の集合による記事間の関連度を計算し，関連度が閾値 θ を超えた場合，Step 3で選択した記事IDを関連記事として出力する．

Step 5　まだ計算していない記事IDがあればStep 3へ，そうでなければ終了する．

図1にシステムの概要図を示す．

また，本手法は，元文書として複数の文書を入力することが可能である．複数文書が入力された場合は，対応する連接の集合の和集合を求める，このとき，同一の連接が存在する場合はその頻度を加算する．こうして得た連接の集合を用いて関連文書検索を行う．

3.4 実験データ

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．また，本手法における変数 α，β を予備実験により決定し，それぞれ，$\alpha=5$，$\beta=2$ とした．本実験の際にはこの変数を用いるが，さらに見出しによる重要度を無視した場合を検討するために，$\alpha=0$，$\beta=2$ の場合についても実験を行った．

実験対象の元記事として10記事を設定し，これらの元記事に対応する関連記事群を人手で抽出した．記事群は以下に示すとおりである．

記事群A　「たか号」の遭難について，事故原因や生

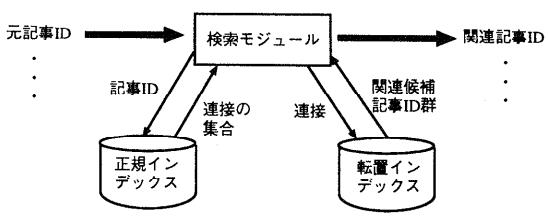


図1　システムの概要

Fig. 1　The system overview.

☆ 連接の集合作成時の名詞はJUMANが標準で使用する辞書中の名詞を指すが，形式名詞，時相名詞，数詞，副詞的名詞は除く．

存者の回復などに関する記事（元記事長：766文字，検索範囲内記事数：7,881，関連記事数：29）

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（元記事長：664文字，検索範囲内記事数：6,142，関連記事数：26）

記事群C ある企業の汚職事件に関する記事（元記事長：852文字，検索範囲内記事数：8,584，関連記事数：62）

記事群D 従軍慰安婦問題ならびに首相訪韓に関連する記事（元記事長：804文字，検索範囲内記事数：7,896，関連記事数：31）

記事群E 国連本部での安保理サミットに関する記事（元記事長：1,167文字，検索範囲内記事数：7,721，関連記事数：36）

記事群F 脳死臨調答申についての議論や，対応に関する記事（元記事長：931文字，検索範囲内記事数：8,436，関連記事数：44）

記事群G 医師誘拐事件で，人質解放から犯人指名手配，逮捕に関する記事（元記事長：1,423文字，検索範囲内記事数：8,436，関連記事数：37）

記事群H ある企業の汚職事件のうち，政治家逮捕に関する記事（元記事長：1,494文字，検索範囲内記事数：7,896，関連記事数：77）

記事群I サハリン沖石油開発に関する，ある企業やロシア共和国の動きに関する記事（元記事長：215文字，検索範囲内記事数：7,771，関連記事数：28）

記事群J ウルグアイ・ラウンドと関係したコメ関税化問題に関する記事（元記事長：1,471文字，検索範囲内記事数：7,929，関連記事数：44）

3.5 比較手法

本論文で提案する名詞の連接情報を用いた手法と比較するために，新谷らの手法[11]を比較手法として採り上げる．新谷らの手法は，形態素に対して重みづけを行い，それらの上位25%を用いて関連度評価を行う手法である．それに対し，本手法は形態素の連接に重みを付加するので，新谷らの手法が比較の対象として適していると考えた．この手法で使用する変数は文献11)で用いられているものに対して変更を加えずに使用した．また，使用した形態素解析器は，本手法で用いたものと同一のJUMAN Version 3.4[15]（標準辞書に変更を加えず使用）である．比較手法においても，本手法と同様にインデックスを作成した．比較手法におけるインデックスの容量は正規インデックス6.5 MB，転置インデックス9.7 MBである．

3.6 評価関数

本論文では記事群 X において閾値 θ で検索した場合の精度を F 値（F-measure）を用いて評価する．F 値の定義を以下に示す．

$$F(X,\theta) = \frac{(\gamma^2+1)\times P(X,\theta)\times R(X,\theta)}{\gamma^2\times P(X,\theta)+R(X,\theta)} \quad (\%)$$

ここで，変数 γ は適合率の再現率に対する相対的な重要性である．本論文ではこの変数 $\gamma=1$ とする．また，P は記事群 X において閾値 θ で検索した場合の適合率（precision）であり，R は再現率（recall）である．以下に，適合率と再現率の定義を示す．

$$\text{適合率：}P(X,\theta) = \frac{\text{検索した中で正解の記事数}}{\text{検索した記事数}}\times 100 \quad (\%)$$

$$\text{再現率：}R(X,\theta) = \frac{\text{検索した中で正解の記事数}}{\text{正解の記事数}}\times 100 \quad (\%)$$

3.7 閾値の決定

比較手法ならびに本手法とも，関連度計算結果の値に対し関連記事であると判断する閾値を決定しなければならない．そのため，10記事群を5記事群ずつの学習グループとテストグループに分けた．学習グループは記事群A，B，C，D，Eから構成され，テストグループは記事群F，G，H，I，Jから構成される．閾値 θ は学習グループにおいて θ を変化させながら実験を行い，グループにおける $F(X,\theta)$ の平均が最も大きい θ を閾値として設定する．

まず，比較手法の学習グループにおける結果を**表1**に示す．比較手法において θ の刻み幅を0.01として，0.01から0.13の範囲で実験を行った結果，閾値 $\theta=0.02$ のときに $F(X,\theta)$ の平均が最大となった．

次に，本手法（$\alpha=5$）において θ の刻み幅を0.1として，0.1から1.5の範囲で学習グループに対して実験を行った結果，$\theta=0.5$ のときに $F(X,\theta)$ の平均が最大となった．本手法の学習グループにおける結果を**表2**に示す．同様にして，本手法（$\alpha=0$）において実験を行った結果，こちらも $\theta=0.5$ のときに $F(X,\theta)$ の平均が最大となった．記事群AからEにおける $F(X,\theta)$ の平均は78.7%であった．

3.8 実験結果

テストグループを対象に，決定した閾値を用いて実験を行った．比較手法による結果を**表3**に示す．

次に本手法（$\alpha=5$：見出し情報利用）のテストグループにおける結果を**表4**に示す．

そして，本手法（$\alpha=0$：見出し情報無視）のテストグループにおける結果を**表5**に示す．

さらに，比較手法と本手法において，閾値を変化さ

表1 比較手法の検索精度（学習グループ）
Table 1 Retrieval precision by Araya's method (for training group).

記事	適合率	再現率	F 値
A	78.6%（22/28）	75.9%（22/29）	77.2%
B	66.7%（22/33）	84.6%（22/26）	74.6%
C	97.4%（37/38）	59.7%（37/62）	74.0%
D	78.1%（25/32）	80.6%（25/31）	79.4%
E	43.5%（30/69）	83.3%（30/36）	57.1%
平均	72.8%	76.8%	72.5%

表2 本手法の検索精度（学習グループ）
Table 2 Retrieval precision by our method (for training group).

記事	適合率	再現率	F 値
A	70.7%（29/41）	100.0%（29/29）	82.9%
B	80.0%（24/30）	92.3%（24/26）	85.7%
C	87.7%（57/65）	91.9%（57/62）	89.8%
D	57.4%（31/54）	100%（31/31）	72.9%
E	64.6%（31/48）	86.1%（31/36）	73.8%
平均	72.1%	94.1%	81.0%

表3 比較手法の検索精度（テストグループ）
Table 3 Retrieval precision by Araya's method (for test group).

記事	適合率	再現率	F 値
F	91.5%（43/47）	97.7%（43/44）	94.5%
G	96.3%（26/27）	68.4%（26/38）	80.0%
H	88.6%（31/35）	40.3%（31/77）	55.4%
I	76.9%（20/26）	71.4%（20/28）	74.1%
J	82.1%（32/39）	72.7%（32/44）	77.1%
平均	87.1%	70.1%	76.2%

表4 本手法（$\alpha = 5$）の検索精度（テストグループ）
Table 4 Retrieval precision by our method (for test group, $\alpha = 5$).

記事	適合率	再現率	F 値
F	97.5%（39/40）	88.6%（39/44）	92.9%
G	90.6%（29/32）	76.3%（29/38）	82.9%
H	88.2%（60/68）	77.9%（60/77）	82.8%
I	84.8%（28/33）	100.0%（28/28）	91.8%
J	82.9%（34/41）	77.2%（34/44）	80.0%
平均	88.8%	84.0%	85.9%

表5 本手法（$\alpha = 0$）の検索精度（テストグループ）
Table 5 Retrieval preciesion by our method (for test group, $\alpha = 0$).

記事	適合率	再現率	F 値
F	100.0%（37/37）	84.1%（37/44）	91.4%
G	93.1%（27/29）	71.1%（27/38）	80.6%
H	89.8%（44/49）	57.1%（44/77）	69.8%
I	96.4%（27/28）	96.4%（27/28）	96.4%
J	81.0%（34/42）	77.3%（34/44）	79.1%
平均	92.1%	77.2%	83.5%

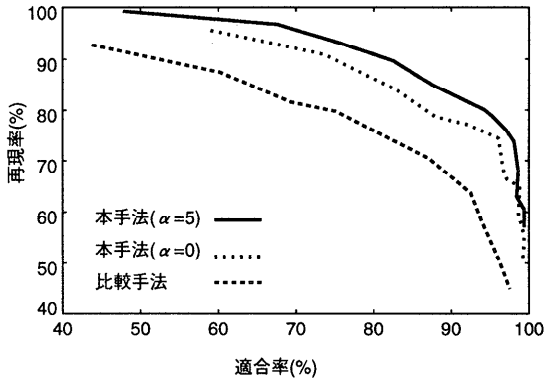


図2 適合率と再現率の平均の関係
Fig. 2 Relation between precision and recall average.

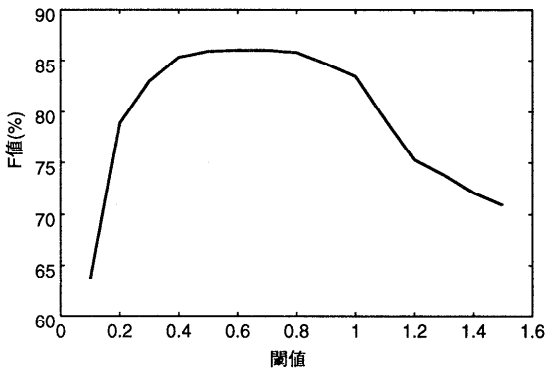


図3 閾値と F 値の関係
Fig. 3 Relation between threshold value and F-measure.

せた場合の精度の違いを見るために，テストグループに対して閾値を変化させて実験した．この場合の適合率と再現率の平均の関係を図**2** に示す．

また，本手法（$\alpha = 5$）における，テストグループに対する閾値と F 値の平均の関係を図**3** に示す．

比較手法と本手法の 10 記事群についての平均検索時間を計測した結果，比較手法 0.44 秒，本手法 4.96 秒であった．計測の方法は A から J のそれぞれの記事群の検索時間を計測し，これを 10 回繰り返しその全体の平均をとった．

4. 考　　察

4.1 手法の特性

実験結果である表 3，表 4 および図 2 から，本論文で提案した手法が，見出し情報を用いる，用いないにかかわらず，形態素のみに重みづけを行う比較手法と比べて高い精度を達成している．このことによって，ベクトル空間モデルにおける索引語として形態素のみを用いるよりも，形態素の連接を用いる方が有効であ

るといえる．

図2から，適合率と再現率の間にトレードオフの関係があることと，その変化が，本手法，比較手法とも類似していることが分かる．閾値を上げると，適合率が増加し，再現率が減少するが，適合率が90%を超えて，100%に近づくと，再現率の減少が激しくなる．また，$\alpha = 0$ と $\alpha = 5$ における違いから，見出し情報を用いたほうが精度が高くなるといえる．ただし，$\alpha = 5$ の場合の精度の平均が85.9%であり $\alpha = 0$ の場合の精度の平均は83.5%であるところから，見出し情報が精度に大きくは寄与しない．

図3から，本手法（$\alpha = 5$）における閾値はある幅を持って安定した結果を出すことが分かる．具体的には，0.4から0.8の間で，記事によって良い結果の場合とそうでない結果を示す場合がある．また閾値が0.7の場合に F 値が最高となる．そのときの適合率の平均は94.2%であり，再現率の平均は80.1%であり，再現率より適合率が大きくなっている．閾値が0.4の場合は適合率の平均は82.4%であり，再現率の平均は89.8%であり，この場合には逆に適合率より再現率が大きくなる．このように，本手法では適合率，再現率がともに80%以上で F 値が安定した値を出す閾値は0.4から0.8であるといえる．

4.2 検索結果の分析

実験結果中の記事Dにおいて，比較手法が本手法よりも優れた F 値を与えている．これは本手法においては適合率が低いためであるが，この適合率低下の原因となっているのは，重要性の低い複合名詞である．特に記事Dにおいては「官房長官」や「首脳会談」などが数多く出現する．また，その他の記事群に見られる例としては「第三管区海上保安本部」などの長い複合名詞がある．これらの複合名詞は本手法で用いているヒューリスティックスによって，多くの連接を構成し，かつ，それらが記事間で共通した連接となるので記事間の関連度が過大評価されてしまう．

記事Hに代表されるように，ヒューリスティックスが効果的に機能している場合は再現率が向上している．本手法と，比較手法とで，結果が大きく異なるのは再現率であり，これは表4と表5から判断して，見出しのヒューリスティクスの効果である．また重要な複合語「○○□□元長官」が「○○元長官」と照合したり，「○○□□代議士」と「○○代議士」が照合するなど，本手法の特徴が発揮され，比較手法より優れた再現率を示した．

本手法はベクトル空間法における索引語の単位を形態素の連接としているところから，一般的には適合率の向上が期待できる．しかし実際には再現率向上のために導入した，表記のゆれを吸収するヒューリスティックスによって，適合率が低下する場合もあることを観察した．以上のように，本手法と比較手法とを比較すると，本手法は再現率の向上に対し，それほど適合率が低下していないと見ることができる．

4.3 検索時間

本手法と比較手法とで，検索時間に大きな差が生じた．考えられる要因は2つある．まず，本手法にて扱うデータ量が比較手法に比べて大きい．そのため，記事間の関連度を計算する際に比較手法に比べ時間がかかる．これに付随して，本手法でのインデックスの容量が比較手法のインデックスに比べ大きくなってしまうが，近年の計算機の急速な発展により，この点はそれほど問題とならなくなりつつある．次に，比較手法はあらかじめ各形態素の重みを計算してインデキシングしてあるのに対し，本手法は複数記事の入力に対応させるために，インデックスには頻度情報のみを含ませ，連接の重みを検索時に求める．そのため計算時間が比較手法に比べ，余分にかかることになる．しかしながら，実際の検索の時間は十分実用的な範囲であると考えている．

5. まとめ

本論文では，ベクトル空間法に基づいて，索引語の単位として名詞を中心とした形態素の連接を用いた関連文書検索手法を提案した．本手法の特徴は，名詞を中心とした連接を用いることにより，複合語などに柔軟に対処した点である．本手法を計算機上に試作システムとして実装し，日本経済新聞を対象に実験を行った．この際，形態素のみに対して重みづけを行う新谷らの手法[11]との比較実験を行い，本手法を用いることでより高い精度で検索できることを確認した（F 値の平均値：86.0%）．このことから，従来の形態素のみに重みづけを行う手法に比べ，形態素の連接に重みづけを行う本手法が，より高精度な関連記事検索を実現している．

今後の課題として，不要な連接の削除があげられる．そのために，

- 複合名詞の係り受け解析の検討
  複合名詞の構造の理解により不要な連接の生成を抑制できる可能性があること，
- ヒューリスティックスの見直し
  考察で述べたように，現状のヒューリスティックスでは，検索上重要ではない連接を生成するルールを含んでいるので，より効果的なものとすること，

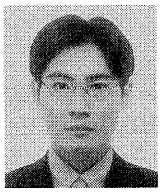
などを検討する必要がある．

謝辞 本研究の一部は文部省科学研究費特定領域研究B（2）および（財）国際コミュニケーション基金の援助を受けて行った．

参考文献

1) Belkin, N.J. and Croft, W.B.: Retrieval Techniques, *Annual Review of Information Science and Technology* (*ARIST*), Vol.22, pp.109–145 (1987).
2) Salton, G. (Ed.): *THE SMART RETRIEVAL SYSTEM/Experiments in Automatic Document Processing*, Prentice-Hall (1971).
3) Salton, G. and Buckley, C.: Term-Weighting Approaches In Automatic Text Retrieval, *Information Processing & Management*, Vol.24, No.5, pp.513–523 (1988).
4) 長尾 真（編）：岩波講座ソフトウェア科学15 自然言語処理，岩波書店 (1996).
5) 岩山 真，徳永健伸：確率的クラスタリングを用いた文書連想検索，自然言語処理，Vol.5, No.1, pp.101–117 (1998).
6) Evans, D.A. and Zhai, C.: Noun-Phrase Analysis in Unrestricted Text for Information Retrieval, *Proc. 34th Annual meeting of Association for Computational Linguistics*, Santa Cruz, California, pp.17–24 (1996).
7) Callan, J.P.: Passage-Level Evidence in Document Retrieval, *Proc. 17th annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.302–310 (1994).
8) 望月 源，本田岳夫，奥村 学：語彙的連鎖を利用した文書検索，情報処理学会研究報告，97-NL-117, pp.137–144 (1997).
9) 中川裕志，森 辰則，松崎知美，川上大介：日本語マニュアル文における名詞間の連接情報を用いたハイパーテキスト化のための索引語の抽出，情報処理学会論文誌，Vol.38, No.10, pp.1986–1994 (1997).
10) 麻生和昭，峯 恒憲，雨宮真人：単語の係り受け構造を利用したWWW上での日本語テキスト検索システム，言語処理学会第3回年次大会発表論文集，pp.253–256 (1997).
11) 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).
12) 影山太郎：文法と語形成，ひつじ書房 (1993).
13) 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).
14) Luhn, H.P.: A Statistical Approach to Mechanized Encodeing and Searching of Literary Information, *IBM Journal of Research and Development*, Vol.1, No.4, pp.309–317 (1957).
15) 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 3.4 使用説明書 (1997).

(平成10年7月1日受付)
(平成11年2月8日採録)

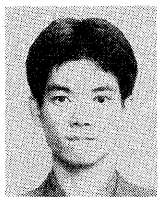


大竹 清敬（学生会員）

1998年豊橋技術科学大学大学院修士課程知識情報工学専攻修了．現在同大学院博士後期課程電子・情報工学専攻在学中．自然言語処理，特に情報検索，自動要約の研究に従事．言語処理学会，人工知能学会各学生会員．

増山 繁（正会員）

1977年京都大学工学部数理工学科卒業．1979年同大学院修士課程修了．1982年日本学術振興会奨励研究員．1983年同大学院博士後期課程修了．1984年京都大学工学部数理工学科助手，1989年豊橋技術科学大学知識情報工学系講師，現在，同教授．アルゴリズム工学，特に，グラフ・ネットワーク，組合せ最適化のアルゴリズム，並列アルゴリズム，および，自然言語処理，特にテキスト自動要約等の研究に従事．工学博士．



山本 和英（正会員）

1996年豊橋技術科学大学大学院博士後期課程システム情報工学専攻修了．博士（工学）．同年よりATR音声翻訳通信研究所客員研究員，現在に至る．1998年中国科学院自動化研究所国外訪問学者．自然言語処理，特に要約処理，機械翻訳，韓国語および中国語処理の研究に従事．1995年NLPRS'95 Best Paper Awards．言語処理学会，ACL各会員．