情 報 学 基 礎 51－6
自 然 言 語 処 理 127－6
（1998. 9. 17）

語彙的連鎖に基づくパッセージ検索

望月 源†, 岩山 真‡, 奥村 学†
†北陸先端科学技術大学院大学 情報科学研究科
‡(株) 日立製作所 基礎研究所
{motizuki,oku}@jaist.ac.jp iwayama@harl.hitachi.co.jp

[概要]
計算機上の文書データの増大に伴い，膨大なデータの中からユーザの求める文書を効率よく索き出す文書検索の重要性が高まっている．最近の文書検索では，ユーザの入力したクエリーと関連の高い文書の一部分を取り出して類似度を計算するパッセージレベルの検索が注目されている．パッセージ検索におけるパッセージとは，文書中でクエリーの内容と強く関連する内容を持つ連続した一部分のことを言う．パッセージ検索では，このパッセージをどのように決定するかが問題となる．良いパッセージを決定するためには，パッセージ自体が意味的なまとまりを形成し，パッセージの位置やサイズがクエリーや文書に応じて柔軟に設定される必要があると考えられる．本稿では，文書中の文脈情報である語彙的連鎖を利用し，クエリーと文書の適切な類似度を計算できるパッセージ決定手法について述べる．また，このパッセージを使用し，検索精度を向上させる検索手法について述べる．

[キーワード] パッセージ，文書検索，語彙的連鎖，共起単語

Passage-Level Document Retrieval Using Lexical Chains

†MOCHIZUKI Hajime, ‡IWAYAMA Makoto, †OKUMURA Manabu
† School of Information Science, Japan Advanced Institute of Science and Technology
‡Advanced Research Lab., Hitachi Ltd.

Abstract
The importance of document retrieval systems which can retrieve relevant documents for user's needs is now increasing with the growing availability of full-text documents. The passage-level document retrieval has been received much attentions in the recent document retrieval task. The definition of a passage is considered as a sequent part of document which contain a relating content to a content of a query. In the passage retrieval, it is a problem how to decide the passages. It is considered that the passages which form meaning coherent units are effective in the improvement of the accuracy. Furthermore, it is also effective that the size and location of each passage is calculated flexibly with each query and document. In this paper we describe a definition of a passage calculation which can be able to calculate a similarity between each content of user's query and each part of document, using lexical chains which denote local document contexts. We also present a passage-level document retrieval method which improve the accuracy.

Key Words passage, document retrieval, lexical chain, cooccurrence word

1 はじめに

文書検索システムは，クエリーの内容と各文書の内容との類似度を計算し，値の高い順に文書を並べて表示する．この類似度は，一般にクエリー内のタームとマッチするタームの文書中の重要度を基に計算される．各タームの重要度は，タームの文書中の出現頻度 (tf) および，文書集合全体での分散 (idf) に基づいて計算する場合が多い．

伝統的な検索手法では，文書全体を1つのまとまりとし，文書中の各タームの重要度を文書全体における重要度として計算する．しかし，実際の文書，特に長い文書は様々な話題を含むため，文書中の部分により話題が異なる場合も多い．話題の違いは，その話題が述べられる部分に出現するタームの違いとして現われる．例えば，あるタームが文書中の一部分では頻出し，他の部分ではほとんど出現しないという状況もある．このような文書では，文書全体を1つの単位とするだけでは各タームの重要度計算には充分でなく，各話題を表わす部分を別々に扱って重要度を計算する必要がある．こうした点から，最近の研究では，パッセージを用いた検索が注目されている [9, 1, 3, 4, 5]．

パッセージ検索では，良いパッセージを決定することが精度向上の上で重要になる．パッセージとは一般的には文書中の連続した一部分のことを言うが，パッセージ検索においては，単なる連続部分では充分ではなく，文書中でクエリーの内容と強く関連する内容を持つ意味的なまとまりを形成する必要がある．また，パッセージは，入力されたクエリーに応じて動的に計算される方が望ましく，パッセージのサイズがクエリーや文書に応じて柔軟に設定されることも良いパッセージの決定につながると考えられる．

本研究ではクエリーが入力された時点でクエリーと各文書に応じて意味的なまとまりを持つパッセージを動的に決定する手法を示す．また，語彙的連鎖の使用によりクエリーに応じた良いパッセージが抽出でき，そのパッセージを使用することで検索精度が向上することを示す．

2 パッセージレベルの文書検索

パッセージの単位には大きく分けて，文書の章，節や形式段落のような形式的な情報に基づくもの [9, 4]，固定長や可変長のウィンドウに基づくもの [1, 5]，形式によらない意味的なまとまりに基づくもの [3] の3種類がある．

形式的な情報に基づくパッセージ抽出は，著者が決定した構造に従う手法であり，インデキシング時に決定できるため処理が容易という利点がある．しかし，現実の文書では同じ話題が複数の段落や章，節にまたがる場合もある．また，日本語においては，形式的な切れ目と内容の切れ目が一致しない場合もある [10]．このような場合，検索精度に悪影響を与える可能性がある．また，クエリーに関係なくパッセージを抽出するため，検索精度向上の条件として，どのようなクエリーに対しても適切な類似度を計算できるパッセージを決定する必要がある．しかし，このようなパッセージの決定は困難であり，この点が問題点として指摘されている [1, 5]．

固定長や可変長ウィンドウによるパッセージは，クエリーの入力時にウィンドウをスライドさせながら各文書を走査し，クエリーとの類似度が高いウィンドウを関連の高いパッセージとして決定する．これは，クエリーに応じてパッセージを決定できる利点がある．しかし，有意なウィンドウサイズを決定しなければならないという問題がある．実際には効果的な検索のために，文書の長さや種類によってウィンドウサイズを調整する方法がよくとられている．また，ウィンドウ境界にまたがる部分の影響で検索精度が悪くなる可能性がある．Callan はこの問題に対処するために，クエリーと最初に一致する場所から走査を開始し，以降のウィンドウは前のウィンドウの中間点から開始する手法を提案している [1]．しかし，有意なウィンドウサイズを決定する問題は依然として残る．またこのタイプのパッセージは，文書の意味的な要素を反映していないという問題がある．

形式によらない意味的なまとまりに基づくパッセージ抽出は，文書の内容に基づいているため最も望ましい方法である．このタイプでは，文書を談話セグメント (意味段落) に分割し，インデキシング時にパッセージを決定する方法と，クエリーが入力された時点でクエリーと関連の強い意味的なまとまりとしてパッセージを抽出する方法が考えられる．

Hearst と Plaunt は，前者の方法として，文書を固定長のブロックに区切り，文書中の語の結束性を計算し，結束性を持つ語がまたがる割合の多いブロックどうしをまとめ，談話セグメント，すなわちパッセージを形成する [3]．しかし，形式的な手法と同様に，クエリーに関係なくパッセージを決定するため，どのよ

うなクエリーにも適切なパッセージを仮定しているという問題がある．また，文書の談話セグメントへの分割は多くの研究者によって行われている [2, 6] が，現在のところ充分な精度での談話セグメントへの分割は達成されていない．

しかし，厳密に談話セグメントを計算する代わりに，比較的浅い処理を用いて文書中の意味的にまとまった部分を取り出すことは可能である．例えば，語彙的連鎖 [7] を計算すると，文書中の語彙ごとにまとまった部分を計算できる．この語彙的連鎖の情報を使用することで，意味的なまとまりに基づくパッセージ抽出の後者の方法が実現できる．クエリーと一致する語彙に関する語彙的連鎖が出現する部分をまとめることでクエリーと関連するパッセージを動的に取り出せる．

次節では，我々がパッセージの決定に使用する語彙的連鎖の計算方法について説明する．

3 語彙的連鎖の計算

語彙的連鎖とは文書中で互いに意味的な関係を持つタームの連続のことである．連鎖の形成基準として，同一タームの反復，シソーラス上の同一概念に属するタームの連続，共起しやすいタームの連続を使用する．なお，語彙的連鎖を構成するタームは，各文書を形態素解析 [13] して取り出した名詞・動詞・形容詞とする．

3.1 同一タームの反復に基づく語彙的連鎖

このタイプの語彙的連鎖は，最も単純な手法である．表層形式の同じタームは互いに意味的に関連のあるタームと考え連鎖を構成する．

3.2 シソーラス上の同一概念に基づく語彙的連鎖

意味的な関連の基準を，シソーラス上の同一概念に属するタームと考え連鎖を構成する．シソーラスを用いる場合，1つのタームが複数の概念に含まれる場合には，語義曖昧性の問題が発生する．そのため，語義曖昧性を解消しつつ語彙的連鎖を生成する手法 [8] により語彙的連鎖を生成する．なおシソーラスとして角川類語新辞典 [11] の小分類を使用する．

3.3 タームの共起関係に基づく語彙的連鎖

文書コーパスからタームの共起の強さである共起スコアを計算し，この共起スコアを用いて関連するタームのクラスタを構成する．1つのクラスタ内のタームの連続が1つの語彙的連鎖となる．

ターム X と Y の共起スコアを式 (1) のコサイン距離によって計算する．

$$coscr(X,Y) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n}(x_i^2)} \times \sqrt{\sum_{i=1}^{n}(y_i^2)}} \quad (1)$$

ここで，x_i と y_i はそれぞれ文書 i におけるターム X と Y の出現数 (tf)，n はコーパスの全文書数を表わす．

このタイプの語彙的連鎖の計算は，次のように行なう．文書の先頭から順に1文を取り出し，1文内のタームの共起スコアを計算する (式 (1))．1ターム1クラスタから開始し，クラスタ間の類似度をタームの共起スコアを基に式 (2) により計算し，類似度の高い順に，閾値以上の類似度を持つクラスタをマージし，クラスタリングを行なう．1文内での処理が終了した後に，その時点までに作成された文書全体でのクラスタと今計算した1文内でのクラスタとの2段階目のクラスタリングを1文内の場合と同様に行なう．これを文書内の文がなくなるまで繰り返す．

クラスタ間の類似尺度には最短距離法を用いる．クラスタ C_i と C_j の間の類似度を以下のように計算する．

$$sim(C_i, C_j) = \max coscr(X \in C_i, Y \in C_j) \quad (2)$$

ここで X, Y はそれぞれクラスタ C_i 内，C_j 内のタームである．

3.4 有意な連鎖の選択

計算された語彙的連鎖の中には重要と考え難いものも含まれる．連鎖を構成するターム数が少ない場合や，タームの出現位置が互いに離れている場合には，その連鎖を有意な連鎖と認めない方が良い．一方，連鎖を構成するタームの数が多く，文書のある部分で高密度でタームが出現する連鎖は有意な連鎖と考えられる．また，1つの連鎖の中で長い間タームの出現しない部分 (ギャップと呼ぶ) がある場合は，そのギャップの範囲では連鎖に関連する話題が述べられていないと考えられる．そのため，ギャップで連鎖を切り離して別の連鎖とした方が良い．

そこで，連鎖およびギャップの長さを考慮した以下の制約を設け，有意な連鎖を選択する．

- ギャップ長の閾値を設定し，閾値以上の間出現しない場合は，連鎖を切り別々の連鎖とする．
- 連鎖長の閾値を設定し，連鎖の覆う範囲が閾値以上の長さをもつ連鎖だけを有意な連鎖とする．

4 語彙的連鎖に基づくパッセージ検索

4.1 語彙的連鎖の重要度

各連鎖の重要度は一般的な$tf*idf$に基づいて計算する．文書d内の語彙的連鎖C_dの重要度w_{C_d}を次のように定義する．

シソーラスおよび同一タームの反復を使用する場合，

$$w_{C_d} = \mid C_d \mid \times log(N/n_{C_d}) \qquad (3)$$

共起タームを使用する場合，

$$w_{C_d} = \mid C_d \mid \times log(N/\max_{c \in C_d} n_c) \qquad (4)$$

ここで，$\mid C_d \mid$は語彙的連鎖C_dを構成するタームの総数，Nはデータベース中の全文書数，n_{C_d}は連鎖C_dと同一の概念に属する連鎖が出現する文書の数，$max_{c \in C_d} n_c$は，連鎖C_dを構成するタームの出現する文書数の中で最大の値をそれぞれ示している[1]．

4.2 語彙的連鎖のインデキシング

検索時にタームの属する語彙的連鎖の出現文書と連鎖の出現位置および重要度を取り出すため，語彙的連鎖のインデキシングを行う．インデックスは，

(連鎖を構成するターム，文書ID，連鎖ID[出現範囲]，重要度)

という形式からなる．図1にインデックスの例を示す．

(星，00001，A1[1-36]，xxxx)
(星雲，00001，A1[1-36]，xxxx)
(星，00001，A2[93-116]，yyyy)
(太陽，00001，A2[93-116]，yyyy)
(月，00001，A2[93-116]，yyyy)

図 1: インデックスの例

図1のインデックスを用いるとターム『星』から文書00001の連鎖A1とA2が検索でき，連鎖A1の範囲は1語目から36語目まで，連鎖A2の範囲は93語目から116語目までであることがわかる．

[1]共起タームを使用する場合，連鎖を構成するタームの種類が文書ごとに異なり，各連鎖の出現する文書数を計算できない．そのため，連鎖の分散情報は連鎖を構成するタームの内，最も多くの文書に出現するタームの情報を使用することにした．

4.3 クエリーと関連するパッセージの計算

語彙的連鎖のインデックスを使用して，クエリーと関連の強いパッセージを計算する手法について述べる．パッセージの計算は以下の手続きで行う．

1. クエリーを形態素解析し，クエリータームとして名詞・動詞・形容詞を選択する．
2. 各クエリータームごとにインデックスを索き，語彙的連鎖の出現位置の情報を得る．
3. クエリータームにマッチした連鎖の含まれる文書ごとに，出現する各連鎖をまとめパッセージの範囲を決定する．
4. 各パッセージとクエリーとの類似度を計算する．

手続き3のパッセージの範囲は次のように決定する．文書中の各語彙的連鎖のうち，出現位置に重なりのある連鎖どうしを1つのパッセージ候補としてマージする．出現位置に重なりのある連鎖がなくなるまでマージを繰り返し，最終的に残ったパッセージ候補をパッセージとする．1つのパッセージの範囲は，パッセージ内の最初の語彙的連鎖が始まるタームから最後の語彙的連鎖が終了するタームまでである．

手続き4では，次の条件をより多く満たすパッセージほどクエリーと強く類似していると考え，パッセージ内の連鎖の数，重要度および連鎖どうしの重なり度合を考慮してパッセージとクエリーの類似度を計算する．

- 多くのクエリータームと関連する語彙的連鎖を含む．
- 重要度の高い語彙的連鎖を多く含む．
- 各語彙的連鎖の出現位置の重複部分が多い．

まず，式(5)により，クエリータームq_kと対応する語彙的連鎖c_kとの類似度cw_kを計算する．

$$cw_k = (tf_{q_k} \times log(N/n_k))^2 \times w_{C_k} \qquad (5)$$

ここで，tf_{q_k}はクエリータームq_kのクエリー内の頻度，Nは全文書数，n_kはタームq_kの出現する文書数，w_{C_k}はタームq_kに対応する語彙的連鎖C_kの重要度である．

次に，cw_kを連鎖の長さcl_kで割り，連鎖の範囲内の各タームjのスコア$cw_{k,j}$を計算する(式(6))．

$$cw_{k,j} = cw_k/cl_k \qquad (6)$$

最後にクエリーターム毎の各スコア$cw_{k,j}$をパッセージの開始位置から終了位置まで足しながら各位置での

連鎖の重なりに応じて重みをかける (式 (7)).

$$sim(Q,P_i)=\sum_{j=begin_i}^{end_i}(\sum_{k=1}^{|Q|}cw_{k,j})\times(\mid q_j\mid^2/\mid Q\mid^2) \quad (7)$$

ここで，Qはクエリー内のターム列，P_iはパッセージiのターム列であり，$begin_i$，end_iはそれぞれパッセージP_iの開始するタームの位置および終了するタームの位置を表す．$cw_{k,j}$は語彙的連鎖kの位置jにおけるスコアである．$\mid Q\mid$はクエリータームの数，$\mid q_j\mid$はクエリータームに対応する連鎖の内，パッセージP_i内のj番目の語を範囲に含んでいる連鎖の数である．

5 実験

我々のパッセージ検索の有効性を調べるため，文書全体によるキーワード検索 (以下，キーワード検索) と，形式段落，ウィンドウ，我々の語彙的連鎖のそれぞれに基づくパッセージ検索との比較実験を行う．また，各パッセージ検索単独の場合とキーワード検索との組み合わせによる場合の比較も行う．

実験には，『情報検索システム評価用テストコレクション BMIR-J2』[12] を使用する．BMIR-J2 は，対象文書 5080 件 (1994 年の毎日新聞から選択した経済および工学，工業技術一般に関連する記事)，クエリー 50 種からなる．クエリーは 5 種類の機能分類がされている．正解判定は A，B の 2 ランクがあり，A はクエリーを主題とする記事，B はクエリーの内容を少しでも記述している記事をそれぞれ表わす．

パッセージ検索は，文書中で強く関連する部分を取り出すという性質から，正解判定には，主題を表わす A ランクを使用する．また，パッセージ検索は特に長い文書で有効であると考えられるため，全 5080 件の内比較的長い文書として 1600 バイト以上の文書 904 件を選択して使用する．クエリーは，数値・レンジ機能[2]および知識処理機能[3]を必要としないクエリーから，正解文書数が 5 文書を越えるクエリーをすべて選択して使用する[4]．また，各文書には見出しが付いているが，本文と合せて 1 つの文書として扱う．テストコレクションの特徴を表 1にまとめる．

表 1: テストコレクションの特徴

文書数	904
クエリー数	8
正解文書数	52
平均正解文書数	6.5
平均文書長 (バイト数)	1785.8
平均文書長 (ターム数)	199.5
平均文数	19.1

なお，すべての検索手法において，文書を形態素解析し，名詞，動詞，形容詞をインデキシング用タームとする．パッセージ検索では同一文書内でクエリーとマッチするパッセージが複数ある場合には，類似度が最大のものをクエリーと文書の類似度として使用する．

5.1 キーワード検索

1 つの文書内のターム t の重みは一般的な $tf*idf$ の式 (8) により計算し，文書をタームの重みつきベクトル Dで表現する．

$$w_t=tf_t\times log(N/n_t) \quad (8)$$

ここで，tf_tは文書内でのタームtの出現頻度，Nは全文書数，n_tはタームtが出現する文書の数である．

クエリーベクトルQと文書ベクトルDとの間の類似度は次式で計算する．

$$sim(Q,D)=\sum_t(tf_{q_t}/log(N/n_t))^2\times w_t \quad (9)$$

ここで，tf_{q_t}はクエリータームq_tのクエリー内の頻度である．

5.2 形式段落に基づくパッセージ検索

見出しと各形式段落をそれぞれ 1 つの文書として扱う．パッセージ毎の各タームの重要度は式 (8) と同様であり，クエリーとパッセージの類似度は式 (9) と同様であるが，tf_tはパッセージ内に出現するタームtの数，Nは全段落数であり，n_tはタームtの出現する段落の数となる．なお，本実験では 1 文書平均 8.8 段落であり，1 段落当り平均 28.2 タームである．

[2]システムに求められる機能として『数の数え上げや，数値などの範囲を正しく解釈する．数値の大小比較や単位の理解・変換なども含む』が要求される．

[3]システムに求められる機能として『世界知識を利用する．常識的な判断や，蓄積された事実からの推論などを含む』が要求される．

[4]本研究では，後述するように評価尺度として再現率と適合率を使用する．このような評価尺度を用いる場合，正解文書数の少ないクエリーを使用すると統計的な信頼性が低下する．そのため，本実験で使用するテストコレクションにおいて信頼性の基準とされている 5 文書以上の正解を持つクエリーだけを使用することにした．

5.3 ウィンドウに基づくパッセージ検索

サイズlの固定長ウィンドウによりパッセージを作る．Callanの手法[1]と同様に，クエリーにマッチするタームの最初の出現位置から走査を開始し，$\frac{l}{2}$ずつウィンドウをずらしながら類似度を計算する．

ウィンドウ内のタームの重要度は式(8)と同様であり，各ウィンドウとクエリーの類似度は式(9)で計算されるが，tf_tはウィンドウ内に出現するタームtの数となる．ウィンドウのサイズを，20〜300の範囲で20刻みに設定しそれぞれのサイズで実験を行う．

5.4 語彙的連鎖に基づくパッセージ検索

3節の各手法により語彙的連鎖を計算し，4節の手法によりパッセージとクエリーの類似度を計算する．有意な連鎖を選択するために，ギャップ長と連鎖長の制約を課す．本実験では，両制約とも閾値を文書内のターム数の1/4, 1/8, 1/16, 1/32の4通りに変化させ16通りの組み合わせを用いる．また，タームの共起に基づく語彙的連鎖の計算では，連鎖の決定に共起スコアの閾値を用いる．本実験では，0.2, 0.25, 0.3, 0.35, 0.4の各閾値で連鎖を計算する．なお，共起スコアの計算では大規模コーパスが必要となるため，テストコレクションと同じ毎日新聞94年の記事1年分(約10万記事)を使用する．

5.5 キーワードとパッセージ検索の統合

各パッセージ検索とキーワード検索とを統合した検索を行う．数多くの統合手法が考えられるが，本研究では次の統合手法を実装して実験を行う．

文書iのキーワード検索，パッセージ検索におけるスコアを，それぞれの最大スコアで割って正規化し足し合わせた値をiのスコアとする．

$$scr_i = \frac{kscr_i}{kscr_{max}} + \frac{pscr_i}{pscr_{max}} \tag{10}$$

ここで，$kscr_i$は文書iのキーワード検索によるスコア，$pscr_i$はパッセージ検索によるスコア，$kscr_{max}$は$kscr_i$の中で最大のスコア，$pscr_{max}$は$pscr_i$の中で最大のスコアを表わす．

この統合手法では，パッセージとキーワード検索の両スコアが相対的に高い文書が上位にランクされる．

5.6 比較実験

前節の4つの検索手法を使用して以下の組み合わせで実験を行う．

1. キーワード検索 (document)
2. 形式段落に基づくパッセージ検索 (formpara)
3. ウィンドウに基づくパッセージ検索 (window)
4. 語彙的連鎖に基づくパッセージ検索
   - 4-a. 同一タームの反復による語彙的連鎖 (repetition)
   - 4-b. シソーラスによる語彙的連鎖 (thesaurus)
   - 4-c. 共起タームによる語彙的連鎖 (cooccurrence)
5. 1と2の組み合わせ (formpara_doc)
6. 1と3の組み合わせ (window_doc)
7. 1と4の組み合わせ
   - 7-a. 1と4-a (repetition_doc)
   - 7-b. 1と4-b (thesaurus_doc)
   - 7-c. 1と4-c (cooccurrence_doc)

評価尺度には，再現率(*recall*)と適合率(*precision*)を使用する．

$$\text{再現率} = \frac{\text{システムにより検出された正解文書数}}{\text{全ての正解文書数}} \tag{11}$$

$$\text{適合率} = \frac{\text{システムにより検出された正解文書数}}{\text{システムが検出した文書数}} \tag{12}$$

但し，各クエリーに対するシステムの出力を上位M位までとする．Mを上位2位から26位まで2文書刻み($M = 2, 4, \cdots, 26$)で各Mの時点での適合率，再現率の平均を計算する[5]．

図2に各パッセージ検索単独とキーワード検索の結果を示し，図3に各パッセージ検索とキーワード検索の統合結果を示す．なお，各検索実験にはさまざまなパラメータが存在するが，図2と図3には，最も良かった場合の結果を示している．最も良い結果が得られた各パラメータの値と，その際の平均パッセージサイズおよびパッセージサイズの標準偏差を表2に示す．

5.7 考察

パッセージ検索単独(図2)では，ウィンドウ型がキーワード検索単独とほぼ同じであり，形式段落，語彙的連鎖のパッセージの中では共起による語彙的連鎖型が一番良いものの，キーワード検索単独を上回る精度を得られなかった．一方，キーワード検索との統合(図3)では，共起による語彙的連鎖型が多くの部分でキーワード検索単独を上回る良い適合率，再現率を得た．形式段落型と他の語彙的連鎖型のパッセージでは，

[5] クエリーによっては出力数がM個に満たない場合が存在するが，式(12)の右辺の分母をMにして計算している．

表 2: 結果の良かったパラメータの値

	共起	同一ターム	シソーラス	ウィンドウ	形式段落
共起スコア閾値	0.25	-	-	-	-
連鎖閾値 (単語数/*)	8	32	32	-	-
ギャップ閾値 (単語数/*)	4	8	8	260	-
平均パッセージサイズ	96.0	81.9	64.9	260	43.4
標準偏差	52.3	68.4	46.5	-	19.3

単独よりも統合で適合率，再現率が向上したがキーワード検索単独の結果を上回ることができなかった．ウィンドウ型は統合によってもキーワード検索単独の結果とほぼ同じであった．

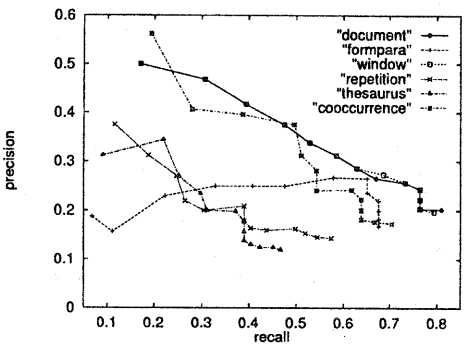

図 2: パッセージ検索単独

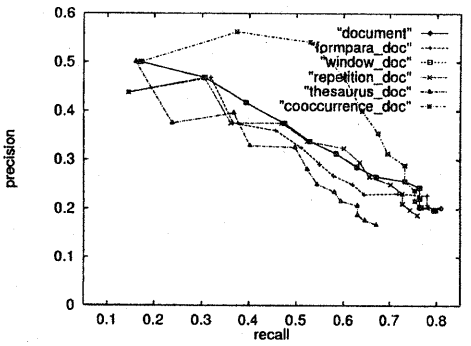

図 3: パッセージとキーワード検索の統合

結果的に共起による語彙的連鎖に基づくパッセージ検索とキーワード検索を統合することで，最も高い適合率，再現率が得られた．特に検索文書数が少ない段階での差は顕著であり，上位 10 位までの文書を見ると，キーワード検索単独では適合率約 34%，再現率約 53%であるのに対し，統合により適合率約 40%，再現率約 64%を得ている．

今回の実験では，パッセージ検索単独の結果がキーワード検索を上回っていないが，この結果がパッセージ検索の優劣を直接示すことにはならない．パッセージ検索とキーワード検索では，有効に働く文書が異なると考えられるため，キーワード検索と異なる正解文書の順位を上げ，不正解の文書の順位を下げることができていれば，統合によって全体的な精度を向上させる可能性が高いからである．

各パッセージ検索単独結果の上位 (10 位以内)，下位 (11 位〜26 位以内)，圏外 (27 位以下) にランクされる文書をキーワード検索の結果と比較すると次のような傾向が見られた．

- 上位の文書集合がキーワード検索結果と異なる割合いは，シソーラス，共起，同一ターム，形式段落の順に高い．ウィンドウ型は，キーワード検索とほぼ同じ文書を同じ順位で選択している．
- 下位，圏外の正解文書を上位へ上げる割合 (正の働き) は，共起，同一ターム，シソーラス，形式段落の順に高い．逆に，上位の正解文書を下げる割合 (負の働き) は，形式段落，シソーラス，共起，同一タームの順に高い．
- 上位の不正解文書を下位，圏外へ下げる割合 (正の働き) は，シソーラス，共起，同一ターム，形式段落の順に高い．逆に，下位の不正解文書を上げる割合 (負の働き) は，シソーラス，形式段落，同一ターム，共起の順に高い．

ウィンドウ型は検索結果がキーワード検索とほとんど同じである．形式段落型は比較的狭い範囲で文書の順位が入れ換わっており，上位，下位の文書集合はキーワード検索とそれほど変わらない．3 タイプの語彙的連鎖型は，上位にランクされる文書集合が異なる割合が高が，シソーラスと同一タームによる連鎖型は，不正解文書が上位に含まれる割合も高い．一方，共起に

よる連鎖型では，上位に正解文書が含まれる割合が他の 2 つの語彙的連鎖型に比べて高い．以上の特徴から，共起による語彙的連鎖型のパッセージ検索が，キーワード検索との統合によって精度を向上する可能性がもっとも高い手法であるといえる．

実際に，統合結果である図 3において，ウィンドウ型の統合にはほとんど変化が見られない．また，形式段落型も統合によって精度が向上しているとは言えない．また，シソーラスと同一タームによる語彙的連鎖型の場合も，キーワード検索単独の場合を上回っていない．一方，共起による語彙的連鎖型では，統合によって良い精度を示すことができている．

以上のことから，共起による語彙的連鎖に基づくパッセージ検索がもっとも優れたパッセージ検索であるため，文書全体の検索との統合によって，高い精度を得ることができたと言える．

6 おわりに

本稿では，文書中の語彙的連鎖を利用して，クエリーと対象文書の適切な類似度を計算するパッセージ検索手法について述べた．他のパッセージ検索手法との比較により，タームの共起により作られた語彙的連鎖が，より優れたパッセージの決定ができ，キーワード検索との統合によって，高い適合率を得ることができることを示した．

本稿で述べたパッセージ検索手法には，パラメータとして考えられる要素が多数存在する．例えば，共起スコアの計算では別の方法も考えられる．また，同一文書内で複数のパッセージが存在する場合に，上位 N 位のパッセージの合計スコアとする方法なども考えられる．今後パラメータとして考えられる要素を明らかにし，検索精度への影響について検討していく必要がある．

謝辞

本研究では，(社) 情報処理学会・データベースシステム研究会が，新情報処理開発機構との共同作業により，毎日新聞 CD-ROM'94 データ版を基に構築した情報検索システム評価用テストコレクション BMIR-J2 を利用させていただきました．感謝致します．また，「角川類語新辞典」の使用を許可して下さいました株式会社 角川書店に感謝致します．本研究を進めるにあたり貴重な御助言を下さいました高野明彦氏，丹羽芳樹氏をはじめとする日立製作所基礎研究所ソフトウェア研究プログラムグループの皆様に感謝致します．また，共起計算プログラムの提供およびシステム実装に関する御助言を頂きました同グループの西岡真吾氏に感謝致します．

参考文献

[1] J. P. Callan. Passage-Level Evidence in Document Retrieval. In *Proc. of 17th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*, pp. 302–310, 1994.

[2] M.A. Hearst. Multi-Paragraph Segmentation of Expository Texts. In *Proc. of the 32nd Annual Meeting of the Association for Computati onal Linguistics*, pp. 9–16, 1994.

[3] M.A. Hearst and C. Plaunt. Subtopic Structuring for Full-Length Document Access. In *Proc. of 16th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*, pp. 59–68, 1993.

[4] D. Knaus, E. Mittendorf, and P. Schäuble. Improving a Basic Retrieval Method by Links and Passage Level Evidence. In *Proc. of the third Text REtrieval Conference*, pp. 241–246, 1994.

[5] M. Melucci. Passage Retrieval: A Probabilistic Technique. *Information Processing & Management*, Vol. 34, No. 1, pp. 43–63, 1998.

[6] H. Mochizuki, T Honda, and M. Okumura. Text Segmentation with Multiple Surface Linguistic Cues. In *Proc. of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 881–885, 1998.

[7] J. Morris and G. Hirst. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, Vol. 17, No. 1, pp. 21–48, 1991.

[8] M. Okumura and T. Honda. Word sense disambiguation and text segmentation based on lexical cohesion. In *Proc. of the 15th International Conference on Computational Linguistics*, pp. 755–761, 1994.

[9] G. Salton, J. Allan, and C Buckley. Approaches to passage retrieval in full text information systems. In *Proc. of 16th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*, pp. 49–56, 1993.

[10] 所一哉. 現代文レトリック読解法. 匠出版, 1987.

[11] 大野晋, 浜西正人. 角川類語新辞典. 角川書店, 1981.

[12] 木谷ほか. 日本語情報検索システム評価用テキストコレクション BMIR-J2. 情報処理学会研究会資料 DBS-144-3, pp. 15–22, 1998.

[13] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明. 形態素解析システム『茶筌』version 1.5 使用説明書, 1997.