自然言語処理 140−17
（2000. 11. 22）

分野連想語の出現位置に基づく話題分野の特定手法

獅々堀 正幹　　岡田 真　　池田 俊彦　　青江 順一
徳島大学 工学部 知能情報工学科
E-mail : bori@is.tokushima-u.ac.jp

複数の話題が混在する文書から，各話題のまとまり部分（パッセージ）を特定し，各パッセージの話題分野を決定する技術は，情報検索の分野に限らず，様々な分野で重要な役割を担う技術である．本稿では，事前に定義された分野体系に従って構築した分野連想語を用いて，パッセージを特定する手法を提案する．分野連想語とは，特定の分野を的確に連想できる単語のことで，分野体系に従って分類された文書データから構築することができる．本手法では，話題の継続性に着目し，分野連想語の水準（連想分野の範囲）や連続出現性から算出した継続度により，パッセージのまとまりを形成する．また，話題の転換性を考慮したアルゴリズムにより，パッセージ間の区切りを明確にし，各パッセージの話題分野を特定する．
キーワード：分野連想後，テキスト分割，パッセージ検索，テキストタイリング

A Retrieval Method of Relevant Passages Using Field Reminding Word's Locations

Masami SHISHIBORI, Makoto OKADA, Toshihiko IKEDA and Jun-ichi AOE
Dpt. of Information Science & Intelligent Systems, Faculty of Engineering,
Tokushima University
E-mail : bori@is.tokushima-u.ac.jp

As the length of the document becomes long, it tends to contain various topics. For such the document, the passage retrieval method, which specifies the set of the coherent sentences about each topic, is the very useful technique for various natural language processing systems, such as information retrieval systems, spoken dialogue systems and Kana-to-Kanji conversion systems. This paper proposes the method to decide the passages using the field reminding words. Field reminding words can be directly related to the field, which must be constructed by human beforehand. This method decides the range of the passage based on five association levels and positions of field reminding words in the document, and the topic field of each passage is specified by the field name which these words are associated with. Moreover, the algorithm proposed in this paper can be avoid the overlap between neighboring passages.
Keywords: Field Reminding Word, Text Segmentation, Passage Retrieval, Text Tiling

1．はじめに

複数の分野の話題が混在する文書から，話題のまとまり部分（パッセージ[5]）を特定し，パッセージの話題分野を決定する技術は，様々な研究分野で重要な役割を担う技術である．例えば，文書検索の分野では，文書全体を検索対象とするのではなく，検索要求に合致した断片のみを検索するパッセージ検索技術[1-5]が注目されている．また，文書分類の分野でも，パッセージという単位を使用し，パッセージ間の類似度により文書を分類するパッセージ分類[6]という手法が用いられている．更に，情報抽出の分野でも，文書内に混在する個々の話題部分を特定する技術は，抽出精度を大きく向上させる要因となる[7]．

また，上記のような情報検索の分野以外にも，広い範囲でパッセージ特定技術は有効である．例えば，音声対話システム[8]において，対話が行われている話題分野が特定できれば，その分野に

応じた対話モデルの推定が可能になる．また，仮名漢字変換システム[9]でも，話題分野を利用することで，同音異義語等の変換候補を絞り込むことができる．更に，図表とその説明箇所（パッセージに相当）との対応付けを行う技術[10]により，文書構造をより詳細に構築することができる．このように，パッセージの特定技術は幅広い研究分野で応用されているが，従来の手法では，パッセージを特定するための手がかり語となる単語が持つ分野特定の強さや範囲を考慮していないため，検出精度に問題を残していた．

一方，辻ら[11]は，事前に定義した分野体系に従って分類された文書データから各分野特有の分野連想語を構築する手法を提案している．この分野連想語とは，“投手”や“選挙”のような単語を見るだけで，＜野球＞[注1]や＜政治＞という常識的分野を認知することができる単語または複合語のことである．

本稿では，パッセージをある分野について書かれた文章のまとまりとしてとらえ，分野連想語を用いてパッセージの範囲を決定し，そのパッセージの分野を特定する手法を提案する．分野連想語をパッセージの特定に用いる場合，分野連想語が現れる部分の分野は特定可能だが，分野連想語が現れない部分の分野を如何に決定するかが問題となる．そこで，本手法では話題の流れの特徴を検証し，分野連想語の連続出現率から算出した話題の継続度に従ってパッセージのまとまりを形成する．また，話題の転換性[12]を考慮したアルゴリズムにより，隣接したパッセージ間の区切りを明確化する．

以下，**2．**では，従来のパッセージ特定手法について説明し，本研究の位置づけを行う．**3．**では，本手法で用いる分野体系，及びその体系に従った分野連想語について説明する．**4．**では，話題の流れの特徴，及び，その特徴を考慮したパッセージ特定アルゴリズムを示す．**5．**では，実際の文書データを用いた実験により，本手法の有効性を確認し，最後に**6．**で，まとめと今後の課題に触れる．

[注1] 通常の単語の表記方法と区別するため，分野名を＜＞内に記す．

2．本研究の位置づけ

従来法としては，特定すべきパッセージの範囲が文書内の章節や段落のような形式的な文書構造に基づくもの［1］，固定長や可変長のウィンドウに基づくもの［2,3］，意味的なまとまりに基づくもの［4,5,10］の大きく3種類に分けることができる．

まず，文書構造に基づく手法は，段落や章節にまたがる話題が存在する場合に問題となる．また，固定長ウィンドウによる手法[2]は，ウィンドウをスライドさせ，検索要求と最も類似度の高いウィンドウをパッセージとする．可変長ウィンドウによる手法[3]は，一定の範囲内でウィンドウ幅を変化させることにより，パッセージの範囲を調整することができる．しかしながら，パッセージの意味的なまとまりを形成する上での柔軟性と処理コストの実用性から判断しても，両手法とも有効な手法とはいえない．

意味的なまとまりに基づく手法として，Hearstら[4]は，文書を固定長のブロックに分割し，ブロック内に出現する単語の結束性から類似したブロックをまとめてパッセージを形成する手法を提案しているが，表層的な単語情報しか用いていない点に改善の余地がある．また，望月ら［5］は，検索要求文の単語以外に，同一概念語，及び共起語の出現位置から語彙的な連鎖を計算し，パッセージを抽出する手法を提案している．しかしながら，語彙的連鎖を求める上で，各単語の連鎖長やギャップ長はある一定のしきい値で設定されており，それら単語の連続出現性や出現分布などは考慮されていない．更に，水野ら［10］は，パッセージ検索の応用として，文書中の各図表に対する説明箇所の特定を行う手法を提案している．この手法では，図表中やキャプション内で用いられている単語をキーワードとし，文書内でのキーワードの出現密度分布が偏って高い部分を説明箇所としている．ここで，各キーワードの出現の偏り具合の算出には，ハニング窓関数で計算された出現密度[13]を使用している．この手法は，キーワードの出現分布に基づいているため，高い精度でパッセージのまとまりを形成することができるが，パッセージの境界部分

の精度に問題が残る[注2]．つまり，説明内容の転換性については考慮しておらず，各説明箇所の区切りを明確化するアルゴリズムについては言及されていない．そのために，文書内で複数の図表が密集し，それらの説明箇所が隣接している場合，説明箇所が重複してしまうおそれがある．

本手法は，文献[1]-[3]と異なり，意味的なまとまりから柔軟にパッセージを形成することが出来る．また，文献[4]では，文書内に現れるすべての単語情報を用いているが，本手法では，分野特定に有効な分野連想語に関する情報を用いパッセージを特定する．更に，話題の継続性を考慮し，同一の分野を連想させる分野連想語が連続して出現する度合いを出現分布情報として導入している点が文献[5]と異なる．また，話題の転換性を考慮し，複数の分野のパッセージが連続している場合にも，パッセージ間の重複を避け，各パッセージの区切りを明確化するアルゴリズムを導入している点が文献[10]と異なる．

3．分野連想語について

3．1　分野体系

分野体系とは，各分野の上位下位関係を木構造で表現したものである．以後，分野体系を分野木，分野木の葉に相当する分野を終端分野，終端分野以外は中間分野と呼ぶ．本研究では，用語辞書[14,15]を基にし，分野木を構築した．この分野木の全分野数は200個，大分野数は10個，中間分野数は18個，終端分野172個である．また，直接の上位分野，下位分野をそれぞれ親分野，子分野と呼ぶ．分野の指定は分野名のパス<S>で記述するが，根に相当する＜全体分野＞は省略する．また，特に矛盾が生じない場合はパス指定を省略して終端分野のみで説明する．**図1**に，分野木の例を示す．例えば，分野パス<S>=＜スポーツ¥相撲＞は＜スポーツ＞の下位の終端分野＜相撲＞を表す．

このような分野木に従って文書データを分類し，形態素解析を施した後，各文書内に存在する単語を切り出す．その後，各分野に属する文書データ内に出現する単語の集中率[11]を計算することで，各分野の分野連想語を決定する．また，ここで求まる分野連想語は，形態素辞書に登録されている単語（短単位語）であり，複合語に対する分野連想語は，短単位語の分野継承に基づき半自動的に構築することができる．尚，形態素解析の結果，未登録語となる単語は，分野連想語の対象とはしない．図1の分野木内に，文書データから抽出した短単位語の分野連想語の例を示す．

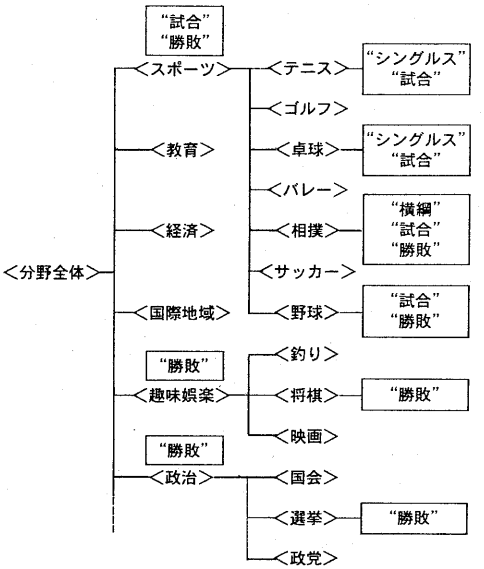


図1　分野木の例

3．2　分野連想語における水準

コーパスデータから決定した分野連想語には，連想する分野の広さに違いがあるため，分野連想語wの水準を次のように定義している[11]．

水準1：唯一の終端分野のみを連想する．

水準2：同じ親分野をもつ終端分野の中で限られた複数の終端分野のみを連想する．

水準3：唯一の中間分野を連想する．

水準4：複数の中間，終端分野を連想する．

図1で例示した分野連想語に従うと，まず，「横綱」のように終端分野＜相撲＞を一意に連想する単語が水準1，「シングルス」のように同じ親分野内の複数の終端分野＜テニス＞，＜卓球＞を特定する単語が水準2，「試合」のように，一つの中間分野＜スポーツ＞を特定する単語が水準

[注2] 再現率は高いが，適合率が低い．

3，「勝敗」は，複数の終端分野＜趣味￥将棋＞，＜政治￥選挙＞や中間分野＜スポーツ＞を特定するので水準４となる．また，「場合」のように分野を特定しない単語は連想語ではない．

４．パッセージ特定法

本手法では，１文毎に処理を進め，各分野のパッセージを特定する．以下，説明に用いる各変数を定義する．まず，解析対象文書を$d=\{s_1, s_2, \ldots, s_i, \ldots, s_M\}$とする．但し，$s_i$は文書$d$内の$i$番目の文を表す．次に，分野木を$F=\{F_1, F_2, \ldots, F_k, \ldots, F_N\}$，但し，$F_k$は個々の分野パス名とする．また，$Freq(s_i, F_k)$を文$s_i$内に存在する分野$F_k$の分野連想語に対する（連想の強さを表す）ポイント数，$Passage(F_k)=\{P_{k1}, P_{k2}, \ldots, P_{kl}, \ldots\}$を文書$d$内に存在する分野$F_k$のパッセージの集合とする．但し，$P_{kl}$は文書$d$内に存在する分野$F_k$のパッセージを示す．

４．１　分野連想語のポイント集計

本手法では，まず，文書内に存在する分野連想語を１文毎に検出する．また，検出された分野連想語は，水準により分野を特定する強さが異なるので，本手法では，各水準毎にポイントを設定し，各文に対する分野のポイントを集計する．

しかし，単にポイントを集計しただけでは，分野連想語が出現した文に対する分野は決定できるが，出現しなかった文の分野は不明である．このように，ポイント集計段階では，下記のような二つの問題点を生じる．

問題点（１）：分野連想語が連続して出現しなければ，パッセージが散発的になり，まとまりがなくなる．

問題点（２）：一文内に複数の分野の分野連想語が同じポイントで出現した場合，複数の分野にまたがったパッセージが形成されてしまい，パッセージ間の重複を生じてしまう．

このような問題を解決するために，本手法では，話題の流れの特徴を利用し，分野毎に（散発的ではなく）まとまりがあり，かつ，分野間の重複のないパッセージを特定する方法を提案する．

４．２　話題の継続性と転換性

新聞や雑誌記事などの一般的な文書内に記述される話題の流れには，以下のような二つの特徴があると考えられる．

特徴（１）：一連の話題は継続性を持ち，一つの話題が散発的に行われることはない．

特徴（２）：話題の流れには転換性[12]があり，複数の話題が平行的に同時進行しない．

特徴（１）から，散発している小規模のパッセージを継続性という尺度からまとめ上げる必要がある．そこで，話題の継続性を計るため，継続度α_{ij}（文s_iでの分野F_jの継続度）を計算する．α_{ij}は，分野F_jの分野連想語が連続して出現すると高まり，連続性が途絶えると衰退するように設定する．計算方法の詳細は，次節にて説明する．

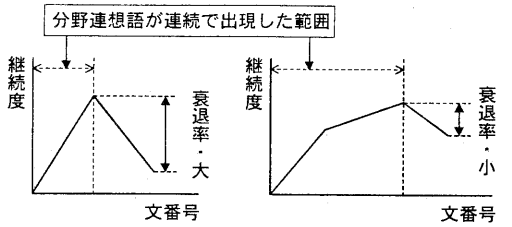
また，特徴（２）より，一つの文は一個以下の話題に対応するため，処理対象の文が話題としている分野を話題分野と定義し，F_{theme}で表す．尚，この話題分野F_{theme}は，話題の転換と共に変化する．

４．３　継続度の計算方法

本手法では，継続度α_{ij}を計算する際に，話題の継続性が衰退する度合いを表す衰退率を定義する．本研究では，図２(a)に示すように，狭い範囲で突発的に出現した分野連想語を含む文は，話題の継続度は少なく話題が衰退し易いと考える．逆に，図２(b)のような広範囲で絶え間なく出現した分野連想語を含む文は，話題の継続性が高く話題が衰退しにくいと考える．そこで，分野連想語の連続出現度を考慮して，文s_iでの分野F_jの衰退率（Dec_{ij}）を以下の式で定義する．

$$Decij = -1 \times \left\{ \frac{\sum_{sk \in Ci-1} Freq(sk, Fj) + Freq(si, Fj)}{num(Ci-1)+1} \right\}$$

但し，C_{i-1}は，文s_{i-1}から遡って，分野F_jの分野連想が連続出現した文集合，つまり，$C_{i-1}=\{s_{i-n}, \ldots, s_k, \ldots, s_{i-1}\}$に対して，$Freq(s_k, F_j)\neq 0$，$Freq(s_{i-n-1}, F_j)=0$を満たす文集合である．また，$num(C_{i-1})$は文集号$C_{i-1}$の要素数$n$を表す．



(a) 分野連想語が狭い範囲で突発的に出現した場合　(b) 分野連想語が広い範囲で絶え間なく出現した場合

図２　分野連想語の連続性と衰退率の関係

この衰退率により，継続度は，解析が新たな文に移行すると衰退し，その後，分野連想語が現れると上昇すると考え，以下の手順で継続度α_{ij}を求める．

【 継続度α_{ij}の計算手順 】

手順１：$\alpha_{ij} = \alpha_{i-1j} + \rho \times Dec_{ij}$；

但し，$\alpha_i < 0$の場合は，$\alpha_i = 0$ とする；

手順２：$\alpha_{ij} = \alpha_{jj} + Freq(s_i, F_j)$；

[手順終了]

尚，パラメータρ（$0 < \rho < 1$）は，衰退率が継続度に影響する度合いとして定義され，ρの値が高いときは衰退率の影響も大きくなり，話題の継続度が低くなる．そのため，話題の変化が多い文書に対してρを高くすると有効である．逆に，ρを小さくすると，話題の継続度が高くなるため，話題の変化が少ない文書に対して有効である．

４．４　パッセージ特定アルゴリズム

４．４．１　全体アルゴリズムの概要

本手法では，１文毎にパッセージ特定アルゴリズムを適用し，パッセージを形成する．本アルゴリズムは，話題出現判定処理，話題転換処理，話題継続処理に分かれており，分野F_jの継続度α_{ij}と話題分野F_{theme}の継続度α_{itheme}の値によって，いずれかの処理に分岐される．

まず，F_{theme}が一意に定まっていない[注3]間は，各分野に対してα_{ij}を算出し，話題出現判定処理を行う．また，F_{theme}が何だかの分野に決定された後はα_{itheme}も算出し，$\alpha_{itheme} < \alpha_{ij}$となる分野$F_j$が現れると，話題転換処理を行う．逆に，$\alpha_{itheme} > \alpha_{ij}$ならば，話題継続処理を行う．以下，個別に各処理の内容を説明する．

[注3] このような場合を＜分野不定＞と定義する．

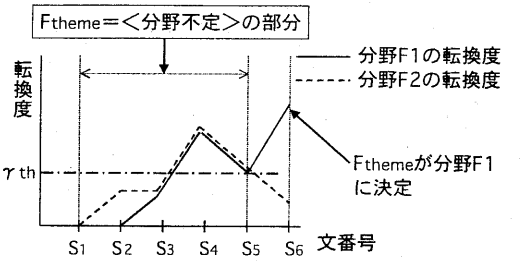


図３　話題出現判定処理の例

４．４．２　話題出現判定処理

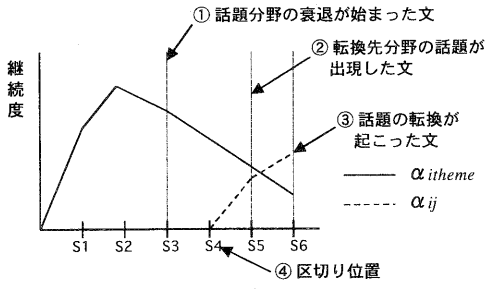
本処理では，α_{ij}が閾値γ_{th}を越えない，または，α_{ij}が最大となる分野F_jが２分野以上存在した場合，F_{theme}＝＜分野不定＞とし，解析文をパッセージ候補として，スタックに格納する．逆に，α_{ij}が閾値を越え，かつ，最大となるF_jが１分野に絞られた場合，F_jをF_{theme}とし，パッセージ候補から$Freq(s_j, F_{theme}) = 0$の文s_jを取り除く．図３に示す例では，文s_6で$F_{theme} = F_1$となり，文s_1, s_2が取り除かれた文s_3～s_6でパッセージ候補が形成される．

４．４．３　話題転換処理

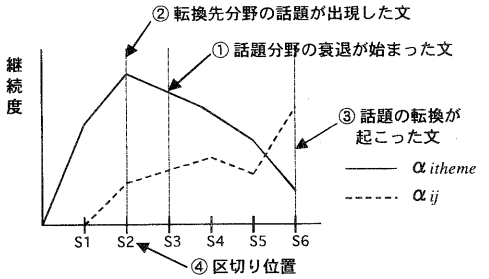
話題の転換が起こった場合，隣接したパッセージ間の区切りを明確にする必要がある．例えば，転換が起こった文s_iで初めて話題が発生しているのならば，s_{i-1}を区切りにすればよい．しかし，転換先の分野が文s_i以前から話題を継続している可能性もあるため，処理のバックトラックを行い，区切り位置$s_{i'}$を決定する．

まず，文s_iから１文ずつ遡り，転換先分野F_jの継続度α_{ij}が最初に0となる文を$s_{i'}$とする．但し，α_{itheme}が増加している間は，F_{theme}の話題が継続していると考え，s_iはα_{itheme}が最後に衰退し始めた文以降とする．そして，文$s_{i'}$までを転換前の話題分野のパッセージとし，文$s_{i'+1}$～s_iを転換後の話題分野のパッセージ候補とする．

例えば，図４(a)の場合，文s_6で転換が起こっているが，α_{ij}が0となる文s_4を区切り位置とする．また，図４(b)の場合，α_{ij}が0となる文はs_1であるが，文s_2まではα_{itheme}が増加しているので，区切り位置は文s_2となる．



(a) 話題分野の衰退した後で，転換先の話題が出現した場合



(b) 転換先の話題が出現した後で，話題分野が衰退した場合

図４　区切り位置の決定説明図

尚，図４(b)の文s_2～s_4のように，α_{itheme}が減少し始めた部分とα_{ij}が増加し始めた部分のどちらを優先してパッセージと認定するかは，より深い議論が必要になるが，今回の研究では，転換度の増加部分を優先した．

４．４．４　話題継続処理

本処理では，α_{itheme}が閾値γ_{th}以上の場合，話題が継続しているとみなし，文s_iをパッセージ候補に追加する．逆に，γ_{th}よりも低い場合は，話題が終了したとみなし，パッセージ候補から$Freq(s_j,F_{theme})$=0の文s_jを取り除いた残りの文を$Passage(F_{theme})$に追加する．

５．評　価

５．１　実験データの内容

本実験では，3,248個の分野連想語を準備した．図５に各分野に対する水準毎の連想語の数を示す．尚，＜スポーツ＞に対する分野連想語が他の分野とくらべて多いが，これは，＜スポーツ＞の文書データ数の多さから，分野連想語を収集しやすく，また，分野を特定できる単語も多く存在したためである．

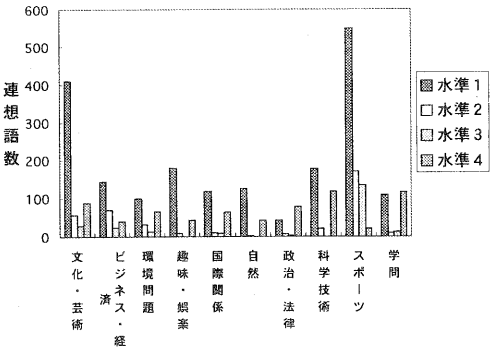


図５　各分野毎の分野連想語数

実験に用いた文書データは，主としてCD-毎日新聞’95データ集より収集し，予め分野木に人手で分類した．この文書データから，図５に示す各分野毎に無作為にN行選び，ランダムに5分野まとめたファイルを作成した．このように，５つの話題分野が混在したファイルをNの値を5，10，15，20と変化させながら各30ファイル作成し，評価用データセットを準備した．尚，分野連想語を構築する際に用いた文書データから作成したクローズドデータセットと分野連想語を構築には用いなかった文書データから作成したオープンデータセットの２種類を用意した．

５．２　パッセージの特定精度

本手法より決定したパッセージと，予め人間により決定したパッセージとがどの程度一致しているかを適合率と再現率を求めて比較した．精度評価に用いる適合率Ｐと再現率Ｒは，出力パッセージと正解パッセージとが一致する文字数をP_{con}，出力パッセージの文字数をP_{out}，正解パッセージの文字数をP_{ans}とすると，$P=P_{con}/P_{out}$，$R=P_{con}/P_{ans}$となる．

また，比較実験として，語彙的連鎖により手法[5]を応用したものを比較手法として用いた．文献[5]で提案されている手法は，検索質問文内に出現する単語，及びその同一概念語，共起語等がある一定の閾値（連鎖閾値）以内で出現する部分をパッセージのまとまりとし，かつ，ある一定の閾値（ギャップ閾値）以上で出現しない部分をパッセージの区切りとしている．この手法に基づき，各水準の分野連想語の連鎖からパッセージを特定する手法を比較手法とした．

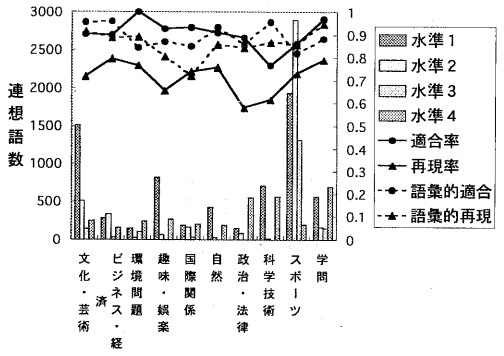


図6　クローズドデータに対する特定精度

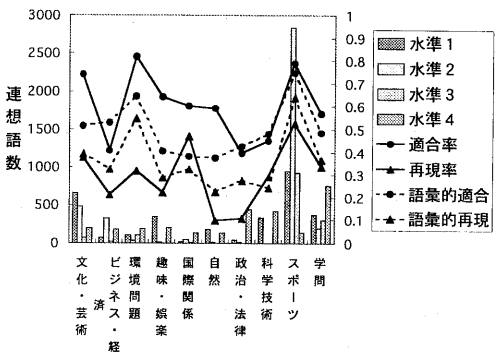


図7　オープンデータに対する特定精度

尚，本手法で用いるパラメーターとして，各水準毎の分野連想語の出現ポイントは，水準1を10，水準2を5，水準3を3，水準4を2と，全分野で一律のポイントを設定した．また，減衰率の影響を左右するパラメーターρは，0.8とした．更に，比較手法で用いるパラメーターとして，連鎖閾値を全単語数／32，ギャップ閾値を全単語数／4に設定した[注4]．

まず，クローズドデータに対する実験結果を図6に示す．但し，図6内の棒グラフは，評価用文書内で検出された（各水準毎の）分野連想語数を表す．また，適合率・再現率に関しては，実線のグラフが本手法，破線が比較手法の結果である．

図6より，水準1の分野連想語の検出率が高い＜文化・芸術＞，＜スポーツ＞において，高い精度の適合率，再現率が得られている．＜学問＞についても，高い精度が得られているが，これは，水準1だけでなく他の水準の分野連想が均等に検出されていたためと思われる．また，＜趣味・娯楽＞，＜科学技術＞については，水準1の分野連想語は多く検出されているが，他の水準の検出率が低いため，パッセージが途切れてしまい，再現率が低くなっている．

比較手法と比べてみると，本手法の平均適合率が0.92，平均再現率は0.71であるのに対し，比較手法は，それぞれ0.89，0.86であった．本手法の方が比較手法よりも，若干，適合率が向上しているのは，本手法が継続度を用いて話題転換処理を行っているため，パッセージ間の境界部分での精度が比較手法よりも勝っていたためである．

[注4] 文献[5]で紹介されている二組の閾値で実験を行った結果，精度の良かった方を本文内で用いる．

一方，再現率に関しては，比較手法が本手法よりも良い結果を得ている．これは，図6からも分かるように検出される分野連想語数は，各分野毎に大差があるにも関わらず，本手法では，分野連想語の出現ポイントを全分野で同じ値に設定したためである．つまり，＜政治・法律＞などの分野連想語数が少ない分野では，＜スポーツ＞などの分野に比べて，継続度が低くなり，パッセージが途切れやすい．これに対し，比較手法では，ポイントを考慮していないため，このような現象は見られなかった．更に，このポイントの問題は，＜文化・芸術＞，＜スポーツ＞の適合率が平均値レベルで，検出された分野連想語数が多い割に適合率が伸びていないことにも影響している．つまり，これらの分野の継続度が高くなり過ぎ，分野連想語数が少ない分野の継続度を覆ってしまうケースが見られた．

次に，オープンデータセットに対する実験結果を図7に示す．クローズドデータの精度に比べ，全体的に再現率が低下している．これは，検出される分野連想語の数，特に，水準1の連想語数が極端に減少したためである．しかしながら，＜スポーツ＞のように，質・量共に整った分野連想語が構築できている分野に関しては，精度の低下はさほど見受けられなかった．一方，＜ビジネス・経済＞，＜政治・法律＞のように，検出された水準1の分野連想語数に比べて，水準1以外の連想語数が極端に多い場合には，同じ大分類内の他の分野に誤認識される場合があり，適合率の低下を招いた．特に，＜学問＞では，クローズドデータ

に比べて，検出された分野連想語数が増加したにも関わらず，適合率再現率とも低下している．これも，水準１以外の連想語が水準１に比べて多く検出されたことが原因である．この点に関して，ポイントの問題にも帰着するが，水準３，４の影響を軽くするポイントの改良が必要と思われる．

また，比較手法と比べてみると，本手法の平均適合率が0.60，平均再現率は0.30であるのに対し，比較手法は，それぞれ0.50，0.36となり，適合率では，本手法が大きく上回っている．これは，水準１以外の連想語が多く検出されたことにより，比較手法での語彙的連鎖が他の分野まで続いてしまい，適合率が低下したことによる．

結論として，分野連想語が構築し易い分野や水準１の分野連想語が多く存在する分野に対しては，本手法はかなり高い精度でパッセージを特定できると言える．また，語彙的連鎖に基づく手法と比べて，本手法では，各水準毎に異なるポイントを設定し，各分野毎の継続度を考慮しているため，パッセージ間の区切りは明確化し，適合率では上回った．一方，再現率に関しては，全分野で一律のポイントを設定したため，連想語数の少ない分野ではパッセージのまとまりを形成できず，比較手法の方が勝っていた．

この問題点に対して，各分野毎，水準毎に与える分野連想語の出現ポイントに差を持たせることにより，精度の向上を図る．具体的には，図５の事前に構築した分野連想語数と図６に示される検出語数とには，各分野間での比率が殆ど同じであるため，事前に構築した分野連想語数から各分野のポイントを算出する．

６．まとめ

本論文では，パッセージをある分野の話題について書かれたまとまりとしてとらえ，話題の継続性と転換性を考慮したパッセージ決定手法を提案した．今後の研究の方向として，各分野毎でのポイント設定方法の変更方法，また，分野連想語間の共起情報や格情報を付与することによって，水準１以外の連想語が複数出現した場合にも，一意に特定の分野を連想させる手法を考案中である．

参考文献

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

[2] J. P. Callen: Passage-Level Evidence in Docu-ment Retrieval, *Proc. of 17th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Develop-ment in Information Retrieval*, pp.302-310, 1994.

[3] M. Melucci : Passage Retrieval : A Probabilistic Technique, *Information Processing & Manage-ment*, Vol. 34, No. 1, pp.43-63, 1998.

[4] M. A. Hearst and C. Plaunt: Subtopic Struc-turing for Full-Length Document Access, *Proc. of 16th Annual International ACM Special Interest Group on Information Retrieval Conference on Re-search and Development in Information Retrieval*, pp.59-68, 1993.

[5] 望月源，岩山真，奥村学：語彙的連鎖に基づくパッセージ検索，自然言語処理，Vol.6，No.3，pp.101-126，1999.

[6] 岩山真，徳永健伸：確率モデルに基づくパッセージ分類とその応用，自然言語処理，Vol.6，No.3，pp.181-198，1999.

[7] 西野文人，落合亮：抽出情報の実体あいまい性の解消，言語処理学会第６回年次大会ワークショップ論文集，pp.41-48，2000.

[8] 北研二，福井義和，永田昌明，森本逞：発話タイプつきコーパスを用いた確率的対話モデルの自動生成，自然言語処理，Vol.4，No.4，pp.73-85，1997.

[9] M. Yasutake, Y. Koyama, K. Yoshimura and K. Shudo: Kana-to-Kanji Conversion Systems Based on Large-Scale Collocation Data, *Proc. of the 18-th International Conference on Computer Processing of Oriental Languages*, pp.479-484, 1999.

[10] 水野浩之，黄瀬浩一，松本啓之亮：単語の出現密度分布と偏出度を用いた図表と説明テキストの対応付け，情報処理学会論文誌，Vol.40，No.12，pp.4400-4403，1999.

[11] 辻孝子，泓田正雄，森田和宏，青江順一：複合語の分野連想語の効率的決定法，自然言語処理，Vol.7，No. 2，pp.3-26，2000.

[12] 内元清貴，小作浩美，井佐原均：対話型ネットニュースグループにおける話題転換記事の推定，言語処理学会第３回年次大会発表論文集，pp.377-380，1997.

[13] 黒橋禎夫，白木伸征，長尾真：出現密度分布を用いた語の重要説明箇所の特定，情報処理学会論文誌，Vol.38，No.4，pp.845-854，1997.

[14] 現代用語の基礎知識，自由国民社，1997.

[15] 戸澤忠彦（編）：情報・知識imidas，集英社，1998.