自 然 言 語 処 理　135－2
（2000．1．27）

窓関数を用いた部分テキスト検索 –ベクトル空間法と出現密度法の比較–

水野 浩之　黄瀬 浩一　松本 啓之亮
大阪府立大学 工学部 情報工学科

文書検索分野では，従来の文書単位での検索に加えて検索結果を文書中の該当する部分で提示できる手法（部分テキスト検索）が必要になっている．本稿では，窓関数を用いた部分テキスト検索としてベクトル空間法と出現密度法を利用した手法について述べる．ベクトル空間法を利用する手法は，従来法の固定長で区切ったテキストを検索単位とする手法を拡張したものである．出現密度法を利用する手法は，検索質問で使われている語に対してテキスト中での出現を調べ，密集している部分を抽出するものである．両手法を理論的，実験的な面から比較し，双方にそれほど大きな差異がないことを示す．

Passage Retrieval Methods Using Window Functions –A Comparison between the Vector Space Model and the Density Distribution method–

Hiroyuki Mizuno, Koichi Kise, Keinosuke Matsumoto
Dept. of Computer and Systems Sciences, College of Eng., Osaka Prefecture University

In recent years, passage retrieval which is to retrieve appropriate parts of documents has received attention in the field of document retrieval. In this report, we focus on passage retrieval methods based on the simple vector space model and the density distribution of keywords, both of which employ a window function to obtain passages. Methods based on the vector space model are an extension of the existing fixed length passage. On the other hand, methods with the density distribution are to extract passages which densely contain terms in a query. We compare these two types of methods from theoretical and experimental viewpoints, and show that they have no significant difference.

1　はじめに

膨大な文書から欲しい情報を効率良く手に入れるために，部分テキスト検索 (passage retrieval) が注目されている．部分テキスト検索とは，検索質問に対する検索結果を文書単位ではなく，文書から検索要求に該当する部分 (部分テキスト) を単位として提示するものを指す．該当する部分のみを出力することにより，文書全体に目を通す手間を省くことができる．

部分テキスト検索は，これまでに様々な手法が提案されている．従来法としては，論理構造法 [2]，固定長法 [3]，集約法 [4]，語彙的連鎖法 [5]，出現密度法 [1, 8, 9]，全パターン法 [6] がある．ここでは，特に固定長法と出現密度法に着目する．

固定長法は，テキストを一定の単語数で区切り，区切られたテキストをベクトルで表現し，検索質問との関連度を測って該当する部分を抽出する手法である．この手法においてテキストを一定の単語数で区切ることは，幅と高さ一定の窓 (矩形窓) をかけ，その中に含まれる語を取り出すことに相当するといえる．このことから，固定長法は矩形窓を用いた部分テキスト検索手法であるといえる．

ここで，固定長法の拡張として矩形窓以外の窓関数も利用してテキストを区切り，テキストをベクトルで表現する際に窓関数の重みを考慮するものを考える．これは，窓関数を用いた部分テキスト検索手法としてベクトル空間法に基づくものである．

また，窓関数を用いた部分テキスト手法の１つとして，出現密度法に基づくものについても述べる．出現密度法は，まず，検索質問の語に対して文書に

おける出現の密集度合を窓関数を用いて求め, 出現密度 [1] という値で表す. そして, 出現密度の高い部分を該当する部分として抽出する手法である.

本稿では, 窓関数を用いた部分テキスト検索の中からベクトル空間法と出現密度法に基づくものについて説明し, 異なるプロセスをもつ両手法の間にあまり差異がないことを理論的, 実験的に明らかにする.

以下, 2章で部分テキスト検索の従来法について述べ, 3章では窓関数を利用した部分テキスト検索について述べる. 4章では比較実験について述べ, 5章でまとめを述べる.

2 部分テキスト検索

文書検索において, 検索結果を文書単位でなく, 該当する部分を単位として提示するものをここでは部分テキスト検索と呼ぶ. 従来法には, 論理構造法 [2], 固定長法 [3], 集約法 [4], 語彙的連鎖法 [5], 出現密度法 [1, 8, 9], 全パターン法 [6] がある.

従来法は, 部分テキストの決定基準の違いによって分類することができる. 分類の観点としては, 検索質問を考慮するか否か, また, テキストの内容を考慮するか否かがある. 検索質問を考慮しない手法は, 検索対象となるテキストが与えられたとき, テキストのみの特徴であらかじめ決定した部分テキストと検索質問を比較する. 一方, 検索質問を考慮する手法は, 検索質問が入力された後, 検索質問に適するように部分テキストが編成される. また, テキストの内容を考慮しない手法は, テキストを決められた規則に沿って機械的に区切ることで部分テキストを得る. 一方, テキストの内容を考慮するものは, テキストを意味的なまとまりを考えて区切る. これらの観点から従来法を分類したものを表 1に示す.

論理構造法は, 章や節, 段落といった明示的な文書の切れ目でテキストを区切る. そして, 区切られたテキストを部分テキストとみなし, 検索質問との類似度を測り該当する部分テキストを出力する.

固定長法は, ある一定の単語数で機械的にテキストを区切り, それぞれを部分テキストとする. 検索方法は, 論理構造法と同様である.

集約法は, あらかじめテキストを固定長でブロックに区切り, 隣接するブロック同士の類似度を測る. そして, 似ているものをまとめて部分テキストとする. 検索方法は, 論理構造法と同様である.

語彙的連鎖法は, テキストにおいてある語彙*が続いている部分をその語彙に対する部分テキストとする. そして, 検索質問に含まれる語彙に対応した部分テキストを抽出し, まとめ直して検索結果として出力する.

*ここでいう語彙とは, 単語, その単語のシソーラス展開語や共起語を指す.

表 1: 従来法の部分テキスト決定基準による分類

検索手法	検索質問	テキストの内容
論理構造法	考慮なし	考慮なし
固定長法	考慮なし	考慮なし
集約法	考慮なし	考慮あり
語彙的連鎖法	考慮あり	考慮あり
出現密度法	考慮あり	考慮なし
全パターン法	考慮なし	考慮なし

出現密度法は, 検索質問の語の出現をテキストに対して調べ, 密集している部分を検索質問に該当している部分テキストとして出力する.

全パターン法は, 1単語から文書全体まで抽出する単位を変化させ, 重なりを許してテキストを区切ることであらゆるパターンの部分テキストを得る. 検索質問と全ての部分テキストを比較し, 類似度の高いものを出力する. 全パターン法は, 膨大なメモリ量や計算時間を費す必要があるが, これまで述べてきた手法から得られる全ての部分テキスト含むものである.

3 窓関数を利用した部分テキスト検索

ここでは, 窓関数を用いた部分テキスト検索手法としてベクトル空間法と出現密度法に基づくものについて説明する. ベクトル空間法に基づく手法は, 固定長法を拡張したものである. 従来の固定長法は, 矩形窓を考え, その中に含まれる語を用いてベクトル化し, 検索質問と比較して検索する手法であると考えられる. ベクトル空間法に基づく手法は, 固定長法において矩形窓を含む窓関数一般を用いて部分テキストを取り出し, 窓関数から与えられる重みを考慮して部分テキストをベクトル化するものである.

一方, 出現密度法は, 検索質問で使われている語を入力とし, 文書においてそれらの語の出現が密集している部分を抽出する手法である. 語の出現の密集度合は, 適当な範囲を設定してその中にどれだけ対象となる語が出現しているかを求めなくてはならない. 適当な範囲として章や節, 段落といったものを利用することが考えられるが, 文字数による正規化の問題があり正確な密集度合が求まらない. そこで, 窓関数を用いることによって適当な範囲を定め, その中で語の出現の密集度合を求める.

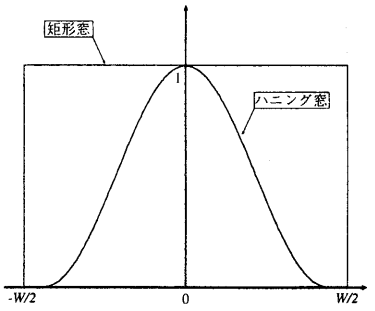

図 1: 矩形窓とハニング窓

本稿では, 窓関数 $f(i)$ として図 1に示す矩形窓 $f_R(i)$ とハニング窓 $f_H(i)$ を用いる. それぞれの関数は窓幅を W とすると以下のように定義される.

$$f_R(i) = \begin{cases} 1, & (|i| \leq W/2) \\ 0, & (\text{otherwise}) \end{cases} \quad (1)$$

$$f_H(i) = \begin{cases} \frac{1}{2}(1+\cos 2\pi \frac{i}{W}) & (|i| \leq W/2) \\ 0 & (\text{otherwise}) \end{cases} \quad (2)$$

3.1 節ではベクトル空間法に基づく手法, **3.2** 節では出現密度法に基づく手法について説明する. **3.3** 節では両手法を計算式から比較する.

3.1 ベクトル空間法

ここで説明する手法は, 固定長法を拡張し, 一般化したものである. ある窓幅をもった窓関数をテキストにかけ, そこに含まれるテキストを用いてベクトルで表現する. ベクトルで表現する際, 窓関数から与えられる重みを考慮する. 窓に含まれるテキストから得たベクトルを部分テキストベクトルと呼ぶ. 窓を 1 単語ずつずらしていくことで, 任意の位置のテキストから部分テキストベクトルを得ることができる.

処理のながれを以下に示す.

(1) 索引語の設定

テキストを形態素解析 (juman3.6)[7] を用いて単語に分け, 単語 $a(l)(1 \leq l \leq L)$ の列として表現する. ここで L は, テキストの全単語数であり, l は単語の出現位置を表す. 次に, 検索に利用する語 (索引語) としてテキストに含まれる名詞 (n 種類) を $t_j(1 \leq j \leq n)$ と設定し, 索引語の集合を T とする.

(2) 検索質問ベクトルの設定

検索質問もテキストと同様に形態素解析で単語に分け, その中の名詞のうち T に属するものをキーワードとし, キーワードの集合を K とする. そして, K から検索質問ベクトル

$$\boldsymbol{Q} = (q_1, q_2, \cdots, q_n) \quad (3)$$

を設定する. ここで, q_j は $t_j \in K$ のとき 1, それ以外のとき 0 とする.

(3) 部分テキストベクトルの設定

部分テキストベクトル $M(l)$ は, 窓関数 $f(i)$ を単語列 $a(l)$ にかけることによって得られる部分テキスト P_l から求める. 部分テキスト P_l は, 窓の中心を l としたときに, $[a(l-W/2), \cdots, a(l+W/2)]$ で表される. まず, 索引語 t_j に対する重み w_j を求める. w_j は, 索引語 t_j がテキスト全体に渡って出現すると小さくなるように設定される. t_j を含む部分テキストの数を df_j, 全部分テキスト数を N_p とすると w_j は以下の式で表される.

$$w_j = \log\left(\frac{N_p}{df_j - \min_i df_i + 1}\right) \quad (4)$$

そして, P_l から部分テキストベクトル

$$\boldsymbol{M}(l) = (m_1(l), m_2(l), \cdots, m_n(l)) \quad (5)$$

を設定する. ここで, $m_j(l)$ は $b_j(l)$ を

$$b_j(l) = \begin{cases} 1, & (t_j = a(l)) \\ 0, & (\text{otherwise}) \end{cases} \quad (6)$$

$(l \leq 0$ または $l > L$ では $b_j(l) = 0$ とする$)$

とすると以下の式で表される.

$$m_j(l) = w_j \sum_{i=-\frac{W}{2}}^{\frac{W}{2}} f(i) \cdot b_j(l-i) \quad (7)$$

部分テキストベクトルの計算例を図 2を用いて説明する. 『情報』, 『検索』, 『関係』 をそれぞれ索引語 t_1, t_2, t_3 とすると,

$$\boldsymbol{M}_5 = (\{f_H(-2) + f_H(0)\} \cdot w_1, \\ f_H(1) \cdot w_2,\ f_H(3) \cdot w_3,\ 0, \cdots, 0)$$

となる.

(4) 部分テキストの抽出

$\boldsymbol{Q}$ と $\boldsymbol{M}(l)$ の類似度

$$s(l) = \frac{\boldsymbol{M}(l) \cdot \boldsymbol{Q}}{\|\boldsymbol{M}(l)\| \cdot \|\boldsymbol{Q}\|} \quad (8)$$

を求め, 類似度 $s(l)$ の最大値を $\max_k s(k)$ とする. 次に, 閾値 Th_1 を設定し, 図 3のように

$$s(l) \geq Th_1 \cdot \max_k s(k) \quad (9)$$

となる $s(l)$ を持つものを求める. そして, 条件を満たす P_l を考え, P_l と重なりを持つ文を全て選択し, 検索質問に該当する部分テキストとして抽出する.

3.2 出現密度法

出現密度法は，検索者から与えられた検索質問に対し，「テキストにおいて検索質問の語を使用する頻度が高い部分は，その質問に関連するテキストである可能性が高い」という考えに基づくものである．この手法では，検索質問として入力される語の分布の疎密を窓関数を用いて出現密度という値で表す．その後，出現密度が高い部分を，該当する部分テキストとして特定する [8, 9]．

以下に処理のながれを示す．

(1) 検索質問からのキーワードの抽出

ベクトル空間法と同様にしてキーワードを得る．

(2) テキストにおけるキーワードの検出

キーワードの全ての出現位置を調べ，$b_j(l)$ を求める．

(3) テキストにおけるキーワードの出現密度の計算

テキストに対して，ある検索質問に現われる各キーワードの出現密度を窓関数を用いて計算する．位置 l における出現密度 $d(l)$ は以下の式で計算する．

$$d(l) = \sum_{k \in K} w_k \sum_{i=-\frac{W}{2}}^{\frac{W}{2}} f(i) \cdot b_k(l-i) \qquad (10)$$

$(l \leq 0$ または $l > L$ では $b(l) = 0$ とする$)$

ここで w_k は，キーワード k に与える重みであり，式 (4) と同様に求められる．

(4) 部分テキストの抽出

出現密度 $d(l)$ の最大値を $\max_k d(k)$ とする．次に，閾値 Th_2 を設定し，ベクトル空間法のときと同様に

$$d(l) \geq Th_2 \cdot \max_k d(k) \qquad (11)$$

を満たす単語 $a(l)$ を考え，それを含む文を選択し，検索質問に該当する部分テキストとして抽出する．

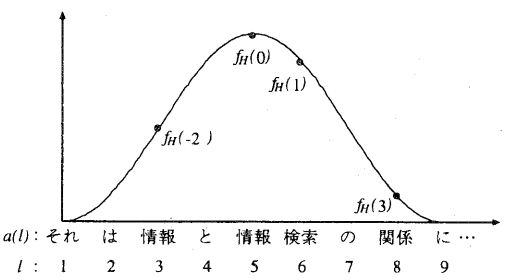

図 2: 部分テキストベクトルの計算例

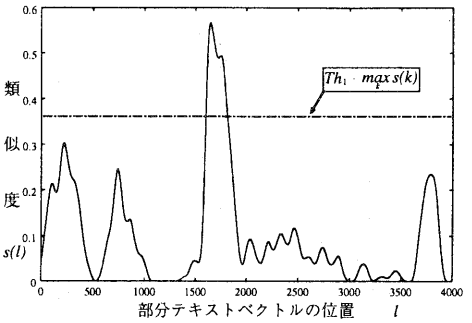

図 3: ベクトル空間法による部分テキストの抽出

3.3 計算式による比較

ベクトル空間法において，計算された部分テキストベクトル $\boldsymbol{M}(l)$ は，部分テキスト P_l に与えられるものである．ここで，$\boldsymbol{M}(l)$ を単語 $a(l)$ に与えることを考える．この場合の $\boldsymbol{M}(l)$ を単語ベクトルと呼ぶことにする．$\boldsymbol{M}(l)$ を単語ベクトルとして与えると，ベクトル空間法と出現密度法の類似性が見えてくる．このことを計算式による比較で明らかにする．

ベクトル空間法による部分テキスト抽出の条件式は式 (9) より

$$\frac{\boldsymbol{Q} \cdot \boldsymbol{M}(l)}{\|\boldsymbol{Q}\| \cdot \|\boldsymbol{M}(l)\|} \geq Th_1 \cdot \max_k \frac{\boldsymbol{Q} \cdot \boldsymbol{M}(k)}{\|\boldsymbol{Q}\| \cdot \|\boldsymbol{M}(k)\|} \qquad (12)$$

とかける．ここで q_j は 0 と 1 で表されたベクトルであるため

$$\boldsymbol{Q} \cdot \boldsymbol{M}(l) = \sum_{k \in K} m_k(l) = d(l) \qquad (13)$$

となる．ゆえに，(12) 式は

$$\begin{aligned} d(l) &\geq Th_1 \cdot \|\boldsymbol{Q}\| \cdot \|\boldsymbol{M}(l)\| \cdot \max_k \frac{\boldsymbol{Q} \cdot \boldsymbol{M}(k)}{\|\boldsymbol{Q}\| \cdot \|\boldsymbol{M}(k)\|} \\ &= Th_1 \cdot \|\boldsymbol{M}(l)\| \cdot \max_k \frac{d(k)}{\|\boldsymbol{M}(k)\|} \end{aligned} \qquad (14)$$

となる．

出現密度法の部分テキスト抽出の条件式 (11) と比較すると閾値を除く差異は，

$$\frac{\|\boldsymbol{M}(l)\|}{\|\boldsymbol{M}(l^*)\|} \qquad l^* = \arg\max_k s(k) \qquad (15)$$

となる．この値は，位置 l によって変化する．

このことから，窓関数を用いた部分テキスト検索においてベクトル空間法と出現密度法は，単語ベクトルを考えると判断基準が多少異なるだけとなる．

表 2: 比較手法

基本とする手法	窓関数	窓を動かす幅	ベクトルを与える単位	手法の略称
ベクトル空間法 (V)	矩形窓 (R)	単語単位 (T)	部分単位 (P)	VRTP
	ハニング窓 (H)	単語単位 (T)	部分単位 (P)	VHTP
	矩形窓 (R)	単語単位 (T)	単語単位 (T)	VRTT
	ハニング窓 (H)	単語単位 (T)	単語単位 (T)	VHTT
	矩形窓 (R)	窓幅単位 (W)	部分単位 (P)	VRWP
出現密度法 (D)	矩形窓 (R)	単語単位 (T)	-	DRT
	ハニング窓 (H)	単語単位 (T)	-	DHT

4 比較実験

ベクトル空間法と出現密度法の特長, 違いを調べるために比較実験を行った. また, これまで説明した手法に加えて, ベクトル空間法に基づくもので窓を 1 単語ずつではなく, 窓幅単位で動かして部分テキストを決定する手法を加えた. これは, 部分テキストベクトルを得る際に, 窓の重なりを許さないものにあたる. 比較対象は 7 手法であり, まとめたものを表 2に示す.

評価用のデータは, BMIR-J2[10] と文献 [8] で用いた図表データを対象とした.

ここで, 部分テキスト検索の実験方法としては, 2 種類が考えられる. 出力された部分テキストをそのまま評価する方法と, 出力された部分テキストを利用して文書検索を行い, 文書単位で評価する方法である. 今回の実験は, 前者の方法で行った.

4.1 BMIR-J2 を用いた実験

4.1.1 実験条件

BMIR-J2 では, CD 毎日新聞 1994 年版から選出された経済・工学分野の 5080 記事を検索対象とし, 検索質問 60 件とそれらに対応する正解データが付けられている. 正解データは, 検索質問との関連度合から A ランク, B ランクの 2 段階の評価が与えられている. A ランクのものは, 検索質問の内容を主題とするもので, B ランクのものは, 検索質問の内容を少しでも記述しているものである. 今回の実験では, A ランク, B ランク両方を正解とした. また, 基本機能[†]のみを必要とする 14 件の検索質問を用い, 検索質問に付けられた補足説明文から関連していると思われる語を加えた. 検索質問を表 3に示す. この表の中で『』で示されたものが元々の検索質問で, それ以外が補足説明文から抽出したものである. また, () で囲まれた語は, 文書内で頻出し, 部分テキストの特定に悪影響があると考え除外した.

実験方法としては, 部分テキスト検索に対応させるため, 5080 記事全てを連結して一つの文書とみなし, 各検索質問に対して検索を行った.

各手法による出力結果は再現率 R と適合率 P を用いて評価した. R と P は正解データに含まれる単語数を A, 出力された部分テキストの単語数を B, 出力された部分テキストの中で正解データと一致した単語数を C とすると $R = C/A$, $P = C/B$ で表される. ここで, 単語単位で評価するのは文長を考慮するためである.

評価値には, 再現率と適合率の積を用いた. 各検索質問に対する結果から得られた評価値を平均したものを用いて各手法を比較した.

パラメータとしては, 閾値 Th_1, Th_2 を 0.05 から 0.90 まで 0.05 刻みで動かし, 窓幅 W は 50～500 まで単語単位で 50 刻みで動かした.

4.1.2 実験結果

実験結果として表 4に各手法に対して最大の評価値を与えるパラメータとそのときの評価値を示す. 最も良い評価値を得たのは VHTP, DHT であったが, 他の手法もほぼ同等の結果を得た. また, 全体的に最適な閾値が 0.10～0.15 とかなり低い値になった.

最も良い評価値を得た VHTP と DHT, 最も悪い評価値を得た VRWP に対し, 閾値を最適にして窓幅を変化させたときの再現率, 適合率の動きを図 4に示す. このグラフから, 各手法間においてほぼ同じように再現率, 適合率が変化していることがわかる. ある適度の窓幅までは, 窓幅を広げることで再現率が上がり, 適合率が下がる. しかし, 窓幅を広げすぎると再現率が上がらずに適合率のみが下がることになる.

[†]キーワードの存在確認, キーワードのシソーラスによる展開語の存在確認

表 3: 検索質問

質問番号	検索質問
01	『菓子, (メーカー)』, 江崎, グリコ
02	『(国内), 航空, (大手)』, 全日空, エアシステム
03	『任天堂, セガ』, ニンテンドー, エンタープライゼス
04	『農薬』, 除草剤, 殺虫剤
05	『飲料, (品)』, 清涼, アルコール
06	『液晶』, LCD
07	『ビデオデッキ』, VTR, ビデオ
08	『携帯, 電話, パーソナルハンディホン』, PHS, テレビ電話
09	『減税』
10	『衛星, 放送』, BS, CS
11	『賃貸, 住宅』, 社宅, 別荘
12	『核兵器』, 核, 査察, 軍縮
13	『ソフトウェア』, ソフト, OS, ゲームソフト
14	『教育, (産業)』, 塾, けいこ, 私立

4.2 図表データを用いた実験

4.2.1 実験条件

図表データは, 文献 [8] で用いたもので, 図表付きの文書を対象に, 図表に対する説明文を出力することで検索精度を評価するものである. 対象文書は, 20 文書であり, 学習用データセット (10 文書), 評価用データセット (10 文書) で成り立っている. 正解データは, 本手法を知らない大学生 4 人によって作成された. このとき, 図表を引用して説明している

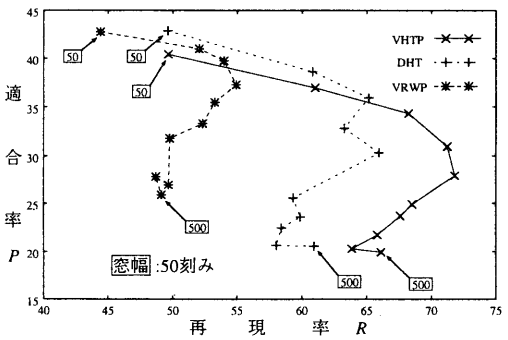

図 4: BMIR-J2 に対する実験で窓幅を変化させたときの再現率, 適合率の動き

表 4: BMIR-J2 に対する実験結果

手法	閾値	窓幅	再現率 (%)	適合率 (%)	評価値
VRTP	0.15	100	64.14	35.96	0.230
VHTP	0.15	150	68.22	34.33	0.234
VRTT	0.10	150	59.43	38.22	0.227
VHTT	0.10	300	59.96	38.29	0.229
VRWP	0.10	200	64.29	33.47	0.215
DRT	0.10	150	69.03	32.23	0.222
DHT	0.15	100	60.86	38.59	0.234

箇所以外でも図表と関連があると考えられる箇所は正解とした.

実験方法としては, 学習用サンプル 10 文書 (図 81, 表 18) でパラメータを決定し, 評価用サンプル 10 文書 (図 67, 表 11) を用いて評価する.

閾値は BMIR-J2 に対する実験と同様に動かし, 窓幅は単語単位で 50～700 まで 50 刻みで動かした.

4.2.2 実験結果

実験結果を表 5に示す. 両データセットにおいて平均して良い精度を示したのは, VRTT, VHTT, DRT であった. また, 両データセット間の評価値の差が小さいのは, VRTP, VHTP であった. 最適な閾値は, 0.65～0.75 と BMIR-J2 のときと比べて高い値を示した. 最適な窓幅は, 手法ごとにばらつきがみられ, 各手法とも BMIR-J2 のときと比べて大きめの幅となった.

学習用データセットにおいて最も良い評価値を得た VHTT と最も悪い評価値を得た VRWP に対し, 閾値を最適にして窓幅を変化させたときの再現率, 適合率の動きを図 5に示す. BMIR-J2 のときと同様に各手法においてほぼ同じように再現率, 適合率が変化していていることがわかった. 窓幅を広げることで再現率が上がり, 適合率が下がるが, BMIR-J2 のときよりも変化が緩やかであった.

4.3 考察

4.3.1 窓関数の違い

窓関数の違いからどのような影響があるかを見るために, 図 6に図表データから矩形窓とハニング窓を用いて得られた出現密度分布の例を示す. このグラフは, 出現密度 $d(l)$ を最大出現密度で正規化した値 $\hat{d}(l)$ を用いている. ハニング窓を用いることにより矩形窓のグラフで平坦になってしまう所が,

表 5: 図表データに対する実験結果

手法	閾値	窓幅	学習用データセット			評価用データセット		
			再現率 (%)	適合率 (%)	評価値	再現率 (%)	適合率 (%)	評価値
VRTP	0.75	150	68.42	53.45	0.365	68.77	50.35	0.346
VHTP	0.75	200	69.32	50.35	0.356	71.45	48.13	0.343
VRTT	0.65	350	74.55	51.43	0.383	73.05	46.92	0.342
VHTT	0.75	650	71.02	56.09	0.398	69.03	50.40	0.347
VRWP	0.70	400	74.63	45.30	0.338	72.99	39.01	0.284
DRT	0.75	200	71.02	53.51	0.380	71.85	48.70	0.349
DHT	0.70	200	70.95	52.36	0.371	68.59	46.59	0.319

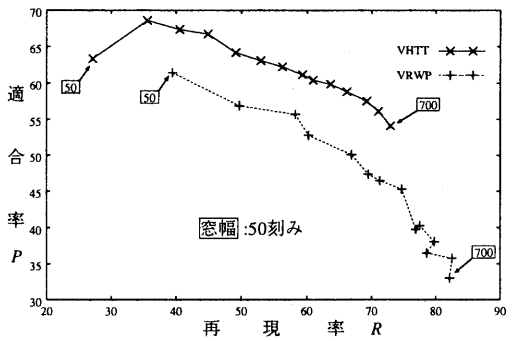

図 5: 図表データに対する実験で窓幅を変化させたときの再現率, 適合率の動き

山なりに盛り上がる. また, 細かく上下している部分を滑らかな変化にする. このらことから, ハニング窓を使うことによって検索質問の語の密集度合をよりはっきりと得ることができる. 同様のことが類似度の方にも当てはまる.

4.3.2 各手法の違い

2つの実験結果において, 各手法の評価値を比較すると, 窓の重なりを許さない手法以外はほぼ類似の結果が得られた. 計算式による比較から, 単語ベクトルを使ったベクトル空間法と出現密度法は, ほとんど差異がないという結果が得られたが, 実験結果からも裏付けがなされたといえる.

4.3.3 検索対象の違い

全体的にみて BMIR-J2 よりも図表データに対する検索精度の方が高かった. この理由として, 検索質問と検索対象文書の特徴が挙げられる.

BMIR-J2 では, 非常に短い検索質問を使っているため, 1つの語の出現が大きく効いてくる. また, 記事をつなぎ合わせているため, 複数の内容が混在する. これらのことから, 検索質問で使われている語が少しでも出現すると, 該当しない部分であっても抽出されてしまうことになる. このような状態は, BMIR-J2 においてある検索質問から得られた類似度のグラフをみても明らかである. 類似度のグラフを図 7に示す. 図 7のような類似度分布から部分テキストを抽出する場合, ある程度の再現率を得るためには閾値を低く設定しなければならない. その結果, キーワードの出現した場所の周辺を検索結果とするキーワード検索に近いものになる. ここで, キーワード検索との差を明らかにするため, 窓を窓幅単位で動かす手法において閾値を 0 にした場合と比較する. これは, 窓で区切られた部分テキスト内にキーワードが出現したものを全て抽出するものである. この手法の結果は, 窓幅 200 で再現率 64.78%, 適合率 32.08%, 評価値 0.207 となった. 単純にキーワードの出現する部分テキストを取り出す手法と比較しても, それほど改善されているとはいえない.

一方, 図表データでは, 検索質問が図表に含まれる語を用いているため, 多くの単語から構成される. また, 図表は検索対象となる文書から得られたものであるため, 図表と文書の間には関連がある. これらのことから, 図表データに対する検索は, 全体的に検索質問と関連のある文書から, 特にあることがらについて書いてある部分を抽出するものになる. 図表データから得られた類似度分布は図 3のグラフのようになる. このグラフから図表と文書は全体的に関連があり, 特に図表との関連が強い部分だけを抽出するために閾値が高くなることがわかる. また, 先程と同様に, 窓を窓幅単位で動かす手法において閾値を 0 にした場合の評価をすると窓幅 50 で再現率 89.62%, 適合率 14.86%, 評価値 0.133 となる. 各手法によって得られた値は, この値と比較して大幅に改善されていることがわかる.

窓関数を用いた部分テキスト検索は, 検索質問

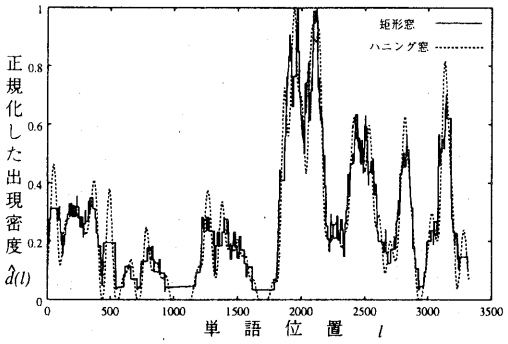

図 6: 矩形窓とハニング窓のグラフの違い (出現密度法)

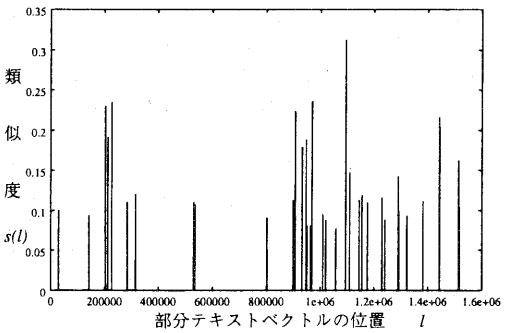

図 7: BMIR-J2 における類似度分布の例 (ベクトル空間法, ハニング窓)

の語の有無で部分テキストが特定できるような対象にはあまり効果がないことがわかった. また, 語の有無ではなく密集度合が重要になる対象には有効であることがわかった.

5 おわりに

本稿では, 窓関数を用いた部分テキスト検索の中からベクトル空間法と出現密度法に基づく手法を説明し, 両者を比較した. 両手法間に理論的, 実験的にそれほど差異がないことがわかった. また, 実験結果から検索質問の語の密集度合が重要になるような対象に対して, 窓関数を用いた部分テキストが有効であることがわかった. 今後の課題としては, 検索質問をシソーラスや共起関係を用いて拡張することと索引語ベクトルの次元削減を行って検索をすることがある.

謝辞

本研究では, 情報検索システム評価用テストコレクション BMIR-J2, CD 毎日新聞 1994 年版ならびに日本語形態素解析システム JUMAN を使わせて頂きました. 関係各位に感謝致します.

参考文献

[1] 黒橋禎夫, 白木伸征, 長尾眞: 出現密度分布を用いた語の重要説明箇所の特定, 情処学論, Vol.38, No.4, pp.845-853, 1997.

[2] Salton, G., Singhal, A., Buckley, C. and Mitra, M.: Automatic Text Decomposition Using Text Segments and Text Themes. In *Proceedings of ACM Hypertext 96*, pp.53-65, 1996.

[3] Callan, J. P.: Passage-Level Evidence in Document Retrieval. In *Proceedings of ACM SIGIR 94*, pp.302-310, 1994.

[4] Hearst, M. A. and Plaunt, C.: Subtopic Structuring for Full-Length Document Access. In *Proceedings of ACM Hypertext 93*, 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. 情報処理学会研究会資料 FI-51, pp.39-46, 1998.

[6] Kaszkiel, M. and Zobel, J.: Passage Retrieval Revisited. In *Proceedings of ACM SIGIR 97*, pp.178-185, 1997.

[7] 黒橋禎夫, 長尾眞: 日本語形態素解析システム JUMAN version3.6, 京都大学大学院工学研究科, 1999.

[8] 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, 情処学論, Vol.40, No.12, pp.4400-4403, 1999.

[9] Kise, K., Mizuno, H., Yamaguchi, M. and Matsumoto, K.: On the Use of Density Distribution of Keywords for Automated Generation of Hypertext Links from Arbitrary Parts of Documents. In *Proceedings of ICDAR 99*, pp.301-304, 1999.

[10] 木谷強ほか: 日本語情報検索システム評価用テストコレクション BMIR-J2, 情処研報 DBS-114-4, pp.15-22, 1998.