論　文

文字の２次元出現密度分布に基づく文書画像の関連部分検索

正　員　黄瀬　浩一*　　非会員　辻野　雅章*,**
正　員　松本　啓之亮*

Retrieval of Relevant Parts of Document Images Based on 2D Density Distributions of Characters

Koichi Kise*, Member, Masaaki Tsujino*,**, Non-member, Keinosuke Matsumoto*, Member

This paper presents a new method of document image retrieval that is capable of spotting parts of document images relevant to users' queries. This enables us to improve effectiveness and usability of retrieval, since users are relieved from burdens of finding relevant parts in retrieved documents. The proposed method is based on the assumption that parts of document images which densely contain characters in queries are relevant to them. For the purpose of ranking relevant parts, two-dimensional density distributions of characters are calculated based on layout features such as locations of characters and distance to the nearest characters. Based on the experimental results of retrieving Japanese newspaper articles, it is shown that the proposed method is superior to a method without a function of retrieving the parts.

キーワード：文書画像検索, 文書検索, 出現密度分布, パッセージ検索
Keywords: document image retrieval, document retrieval, density distributions, passage retrieval

1. まえがき

文書画像とは，記号中心の情報表現である文書を，画像として保存した情報表現形態である。これまで文書画像は，認識処理によって紙文書を電子文書に変換するときに経由するだけの，いわば中間的な情報表現形態であった。しかし，近年，次のような理由によって最終的な情報表現形態としての価値が高まっている。

（1） 誤認識の問題が生じない。
（2） 閲覧の問題が生じない。

(1) は，既存文書を電子図書館に入力するときに顕在化した問題である。紙文書を認識処理により電子文書に変換すると，誤認識により可読性が低下してしまう。この問題を避けるために，国内外の多くの電子図書館では，紙文書を電子的に蓄積する形式として文書画像を採用している。また，(2) は，電子文書の規格に関する問題である。例えば，利用可能なフォントの違いによって，日本語環境で作成した電子文書が英語環境では閲覧できないことがある。一方，文書画像には原理的にこのような問題は生じない。大量の文書画像を扱う場合，以前は記憶容量が問題となることもあったが，最近では，高圧縮フォーマット[(1)]の開発や記憶装置の低廉化により解決されている。

以上のように，文書画像は紙文書や電子文書の保存形式として有効であるが，解決すべき問題点もある。最も大きな問題の一つは，柔軟な検索機能をいかに実現するかである。これまでに様々な検索手法が提案されているが[(2)]，中心的な手法は，Web 検索のように，検索質問をキーワードの集合として与え，関連文書を検索するものであろう。この範疇の研究事例としては，誤りを含む文字認識結果とキーワードをいかに柔軟にマッチングするかというものが多い[(3)~(5)]。理由は，キーワードとのマッチングができれば，通常の電子文書検索が適用できるためと考えられる。

しかしながら，この考え方には，少なくとも次の２つの問題がある。第一は，結果提示の単位に関する問題である。従来法では，検索結果として提示されるのは文書あるいはページであり，その中から検索質問に合致する箇所を探し出すことはユーザの負担となっている。対象文書が，新聞のように多段組の入り組んだレイアウトを持つマルチトピック文書の場合，これはかなりの負担となる。第二は，結果表示の問題である。通常，文書画像は，画面で表示可能な大きさにくらべてかなり大きい。したがって，検索結果を

* 大阪府立大学大学院工学研究科情報工学分野
〒 599-8531 大阪府堺市学園町１－１
Dept. of Computer and Systems Sciences, Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuencho, Sakai, Osaka 599-8531, Japan
** 現在, 松下電器産業（株）
Matsushita Electronic Industrial Co., Ltd.

表示するためには，文書画像を縮小するか，あるいはどこを表示するかを決定しなければならない。携帯情報端末のように画面が小さい場合，あるいは対象文書が新聞のように大きな画像である場合には，画像の縮小による対処では可読性を損なうことになる。

本論文では，上記の２つの問題点に対して，関連部分検索による対処法を提案する。関連部分検索とは，検索質問に関連する部分を，文書画像から検索する手法である。このような処理が可能であれば，関連部分を中心に選択的に表示することによって，上記２点の問題は自然に解決される。本研究では，電子文書検索の分野で検討されているパッセージ検索 [6] という考え方を文書画像検索に導入することによって，関連部分を検索する手法を検討する。特にここでは,「出現密度法」という考え方 [7][8] を用いる。これは，本研究の場合,「文書画像中で，検索質問に含まれる特徴的な文字が密集している部分は，検索質問に関連する可能性が高い」という考え方である。

以下では，まず **2.** において提案手法の詳細について述べる。次に，**3.** で日本語新聞画像を対象とした比較実験の結果について考察し，提案手法の有効性を検証する。最後に，**4.** で本研究の成果をまとめ，今後の課題を示す。

2. 出現密度分布に基づく関連部分検索

提案手法による処理は，索引付け (indexing) と検索処理 (retrieval) から構成される。索引付けは，データベース中の文書画像に対して，あらかじめ施される処理である。検索処理は，ユーザから検索質問 (query) を受けると起動され，検索質問に合致する文書画像をデータベースから選択する。その結果は，合致する部分を中心としてユーザに提示 (presentation) される。以下では，図 1 の処理例を参照しつつ，各部の処理について順に述べる。

〈2･1〉 索引付け　検索質問は単語 (キーワード) で表現されるので，文書画像のインデックスも単語単位で付与することが考えられる。ただし，日本語のように単語の境界が明確でない言語では，このために形態素解析が必要になる。しかし，文字認識誤りをはじめとする種々の誤りを含む対象に対して，形態素解析を施すことは簡単ではない。そこで提案手法では，文字を単位とした索引付けを考える。

索引付けの第一ステップは，文字の切り出しと認識である。図 1(b) は，図 1(a) から切り出された文字を黒の矩形で表したものである。各矩形には，文字認識の結果として得られる文字コードが対応付けられている。

次に，文書画像からレイアウトの特徴を抽出する。この処理の目的は以下のとおりである。我々人間が文書を読む場合を考えてみよう。いわゆるベタ書きの文書と比べて，美しくレイアウトされた文書は読みやすい。この理由は，構成要素（文字，文字列，ブロック）の論理的な関係 (読み順やタイトル-本文の対応など) がレイアウトから容易に知覚できることにある。このような知覚において，最も基本となる手がかりの一つは,「構成要素間の空白の大きさは互いの論理的な関係を反映している」ということであろう†。

そこで本手法では，これを基に，レイアウトの基本的な特徴を抽出する。具体的には，図 1(b) の白画素 (空白の画素) から最も近い文字領域 (同図の黒矩形) までの 4 近傍距離を距離変換により求め，距離分布として記録しておく。以下では，文書画像 p に対する距離分布を $K^{(p)}(x,y)$ と表す。図 1(b) に距離の数値例を，図 1(c) に距離分布を示す。図 1(c) では，画素が白いほど距離が大きいことを表す。

〈2･2〉 検　索　検索処理は，ユーザが検索質問を入力するたびに起動される。

最初の処理は，検索質問の処理である。検索質問として入力された文字列に対して形態素解析 [9] を施し，名詞相当語句 (名詞や未定義語など) をキーワードとして取り出す。一般には，この段階で複数のキーワードが得られる。次に，各キーワードから文字を取り出す。例えば，検索質問「衛星放送」からは,「衛星」,「放送」の 2 つのキーワードが得られ，各々はさらに文字に分解される。以下では，検索質問 q から得られたキーワードを $q_1,...,q_n$，キーワード q_i に含まれる文字を $q_{i1},...,q_{im}$ と表す。

次に，検索質問の出現密度分布を計算する。処理は以下の 4 ステップからなる。

Step 1　文字分布の作成

図 1(b) に示すような文字切り出し・認識の結果から，各文字 q_{ij} を捜し出す。図 1(d) は，文字「衛」に対する文字分布である。

以下では，文字 q_{ij} の位置を，その文字を囲む矩形の中心座標 (u,v) で表す。また，$P_{ij}^{(p)}=\{(u,v)\}$ により，文書画像 p における文字 q_{ij} の分布を表す。

Step 2　文字の出現密度分布の計算

窓関数を用いて文字分布を平滑化し，文字の出現密度分布を求める。図 1(e) に，図 1(d) から得られた文字の出現密度分布を示す。

計算方法は以下のとおりである。文書画像 p における文字 q_{ij} の出現密度分布 $D_{ij}^{(p)}$ は,

$$D_{ij}^{(p)}(x,y)=\sum_{(u,v)\in P_{ij}^{(p)}} W(x-u,y-v)\alpha^{(p)}(x,y,u,v) \quad (1)$$

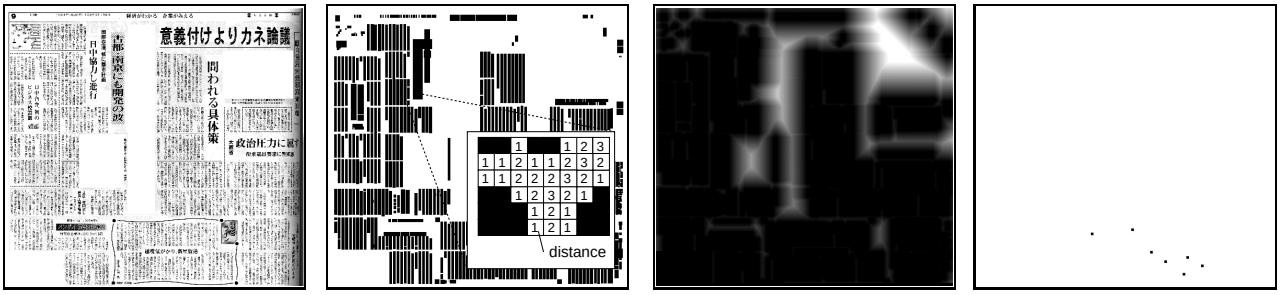
により与えられる。ここで，W は窓関数であり，$\alpha^{(p)}$ は距離分布 $K^{(p)}(x,y)$ から得られる重みである。

窓関数としては，図 2 に示す窓幅 M のピラミッド型関数を用いる。一方，$\alpha^{(p)}$ は，距離分布に基づいて，(u,v) にある文字の影響を減ずるための重みであり，次のように定められる。

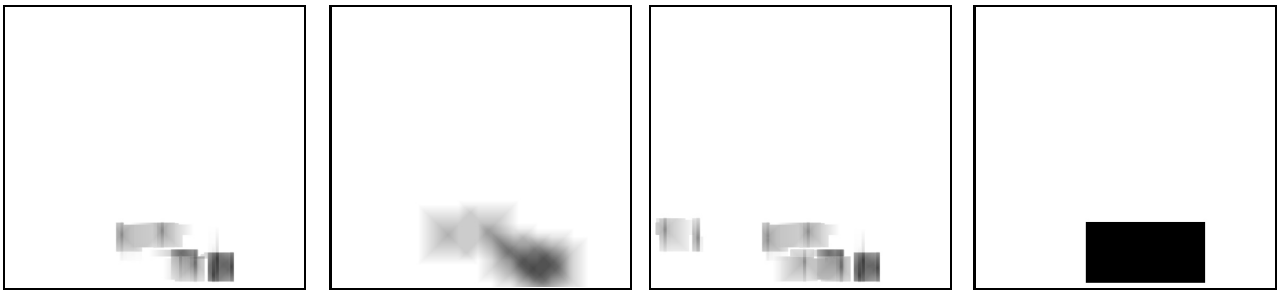
$$\alpha^{(p)}(x,y,u,v)=\begin{cases} T-K_{\max} & \text{if } K_{\max}<T, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

ここで，T はしきい値である。また，$K_{\max}$ は，領域 R

† 空白の大きさだけではなく，様々なレイアウトの特徴から構成要素間の論理的な関係を抽出する処理は「論理ラベリング (logical labeling)」と呼ばれている。すでに多数の手法が提案されているが，未解決な問題も多い。本手法では，この問題を避けるため，空白の大きさという単純な情報だけを用いている。



(a) document image (b) character regions (c) distance distribution (d) position of the first character “衛”



(e) density distribution of “衛” (f) density distribution of “衛” without (C) (g) density distribution of the query “衛星放送” (h) relevant article

図 1 処理例

Fig. 1. Examples of processing results.

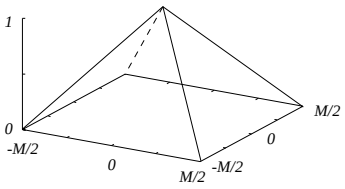


図 2 窓関数

Fig. 2. Window function.

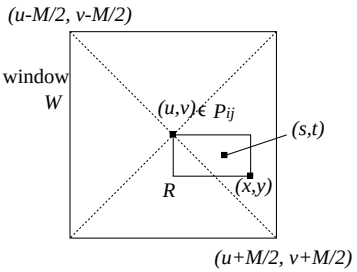


図 3 矩形領域 R

Fig. 3. Rectangular region R.

を図 3 に示すように 2 点 (u, v), (x, y) で定められる矩形領域としたとき，領域 R における距離分布の最大値

$$K_{\max} = \max_{(s,t)\in R} K^{(p)}(s,t) \cdots\cdots\cdots\cdots\cdots\cdots (3)$$

である。矩形 R 内に距離分布の値 $K^{(p)}(s,t)$ が大きい点が含まれていると，重み $\alpha^{(p)}$ の値は小さくなる。$K^{(p)}(s,t)$ の値がしきい値 T 以上であれば，$\alpha^{(p)}$ は 0 となり，点 (u, v) における文字の出現は点 (x, y) に影響を及ぼさない。

図 1(f) に，式 (1) で $\alpha^{(p)} \equiv 1$ としたときの文字の出現密度分布を示す。図 1(e) と比較すると，重み $\alpha^{(p)}$ は，出現密度分布を実際の記事の形状に近づける効果があるといえる。

Step 3　キーワードの出現密度分布の計算

文字 q_{ij} の出現密度分布 $D_{ij}^{(p)}$ をもとに，文書画像 p におけるキーワード q_i の出現密度分布 $D_i^{(p)}$ を，

$$D_i^{(p)}(x,y) = \min_j \beta_{ij} D_{ij}^{(p)}(x,y) \cdots\cdots\cdots\cdots (4)$$

のように求める。ここで，β_{ij} は q_{ij} の重みである。$D_i^{(p)}$ の値として最小値をとるのは，「キーワードが存在する領域は，キーワードを構成する文字をすべて含まなければならない。」という考えによる。

重み β_{ij} は，文字 q_{ij} がキーワード q_i の特定にどれほど重要なのかを表す度合である。文字 q_{ij} が多くの文書画像にまんべんなく分布するようであれば，この文字の出現は，キーワードの特定にはほとんど役立たない。一方，偏って存在するならば，より重要度が高い†。このような考えに基づく重み β_{ij} の定義を以下に示す。

$$\beta_{ij} = \sum_{p,x,y} \{\max_{p,x,y} D_{ij}^{(p)}(x,y) - D_{ij}^{(p)}(x,y)\} \cdots (5)$$

Step 4　検索質問の出現密度分布の計算

最後に，キーワード q_i の出現密度分布 $D_i^{(p)}$ から，文

† 情報検索の分野でよく用いられる「文書頻度の逆数」(IDF; inverse document frequency) に類似の考え方である。

書画像 p における検索質問 q 全体の出現密度分布 $D^{(p)}$ を求める。定義には様々なものが考えられるが，ここでは，以下のものを採用する。

$$D^{(p)}(x,y) = \max_i D_i^{(p)}(x,y). \cdots\cdots\cdots\cdots (6)$$

ここで，最大値をとる理由は,「キーワードのいずれかが含まれていれば，検索質問に関連する」と考えることである。

図 1(g) に，検索質問「衛星放送」の出現密度分布を示す。図 1(h) の正解領域とよく一致していることがわかる。

〈2･3〉 提　　示　検索処理により，データベース中の各文書画像に対して，検索質問の出現密度分布が求められる。ユーザに結果を提示する際には，次の提示処理を施す。

（1） 出現密度分布の最大値 $\max_{x,y} D^{(p)}(x,y)$ を文書画像 p のスコアとし，スコアの降順に文書画像をソートする。

（2） 最上位の文書画像から順に，ユーザに提示する。提示の際には，出現密度分布が最大値を取る部分を中心に，ユーザに表示する。

3. 実　　験

提案手法の有効性を検討するため，日本語新聞画像を対象とした検索実験を行った。

〈3･1〉 実験データ　検索実験を行うためには，文書画像，検索質問，ならびに個々の検索質問に関連する文書画像 (とその部分) のリストが必要となる。本実験では，電子文書検索システムを評価するために作成された「BMIR-J2」[10] に基づいて文書画像データセットを作成し，実験に用いた。

最初に，実験データ作成の元になった BMIR-J2 について簡単に紹介しておく。BMIR-J2 は，CD 毎日新聞 1994 年版から選出された，経済，工学分野の 5080 記事を対象とし，検索質問 50 件と各々の検索質問に対する正解記事を設定したテストコレクションである。正解には，A，B の 2 種類のランクが付与されている。A ランクは検索質問の内容を主題とする記事であり，B ランクは内容を少しでも記述している記事であることを表す。また，検索質問 50 件は，その処理に要求される機能別に分類されている。

本実験では，検索質問のうち最も基本的な機能，つまり「キーワードおよびキーワードのシソーラスによる展開語の存在確認」が要求されるものから 7 質問を選び，実験に用いた。選択した質問のリストを表 1 に示す。

各検索質問に対する正解記事としては，BMIR-J2 の A ランクのものから，検索質問ごとに 3 ないし 4 記事を選択した。次に，その記事が掲載されている新聞のページを毎日新聞 1994 年縮刷版から選び，解像度 800dpi で文書画像として取り込んだ。画像サイズは約 $6,000 \times 8,000$ 画素であり，本文文字の大きさはおおよそ 50×50 画素である。

以上のようにして得た文書画像から，CD 毎日新聞に含まれない部分 (広告や著作権交渉中の記事など) を削除し，検索対象の文書画像とした。図 1(a) の文書画像で一部欠けている部分があるのは，この理由による。実験に用いた文書画像の合計は 25 枚，これらの文書画像に含まれている記事の合計は 249 記事である。

表 1　検索質問
Table 1.　Queries.

ID	query	keywords	no. of relevant articles
1	任天堂またはセガ (Nintendo or SEGA)	任天堂，セガ	4
2	農薬 (agricultural chemical)	農薬	3
3	液晶 (liquid crystal)	液晶	3
4	減税 (tax reduction)	減税	4
5	衛星放送 (satellite broadcasting)	衛星，放送	3
6	賃貸住宅 (rental housing)	賃貸，住宅	4
7	核兵器 (nuclear weapon)	核兵器	4

最後に，索引付けについて述べる。検索質問に含まれる 22 種類の文字は，データベース中の文書画像にのべ 1514 文字含まれていた。文字切り出し･認識を施したところ，そのうち 74.1%が正しく認識された。認識率が低い理由は，スキャンした新聞が縮刷版であったためと考えられる。一方，検索質問の 22 種類の文字に誤って認識された他の文字は，合計 26 文字であった。

〈3･2〉 評価方法　実験では 2 種類の評価尺度を用いた。一つは再現率 (recall) と精度 (precision)，もう一つは平均精度 (mean average precision) である。

再現率と精度は，情報検索システムの評価でよく用いられる尺度である [11]。いま，X と Y を，それぞれ，ある検索質問に対して検索された文書の集合，その検索質問に対する関連文書の集合とする。このとき，再現率 R と精度 P は，$R = |X \cap Y|/|Y|$, $P = |X \cap Y|/|X|$ で定義される。再現率が高いほど検索洩れが少なく，また精度が高いほど誤りを含まず正確な検索といえる。

ところで，検索結果を比較する際には，再現率と精度のように複数の値による評価だけではなく，単一の値による評価も必要なことがある。このような目的を満たす尺度の一つに，平均精度†がある [11]。定義は以下のとおりである。一般に，文書検索の結果は，文書のランキングを表すリストにより表現される。いま，$r(i)$ により，リストの最上位から数えて i 番目の関連文書のランクを表すものとする。例えば，リストの上位から「関連文書，非関連文書，関連文書，...」と並んでいるとき，2 番目の関連文書 ($i = 2$) のランクは 3(上から 3 番目) である。i 番目の関連文書を検索した時点での精度は $i/r(i)$ なので，すべての関連文書に対する平均精度は，

$$\frac{1}{n}\sum_{i=1}^{n}\frac{i}{r(i)} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (7)$$

と表すことができる。ここで n は，現在の検索質問に対す

† 正確には，“mean average precision (non-interpolated) over all relevant documents” と呼ばれる尺度である。

表 2　パラメータの値
Table 2.　Values of parameters.

ID	threshold T	window width M
1	20	1150
2	20	1150
3	20	1150
4	20	1150
5	60	2400
6	90	1000
7	20	1150

る関連文書数である。すべての検索質問に対する平均精度 (以後，単に平均精度と呼ぶ) は，各検索質問に対する平均精度をさらに平均したものである。

最後に，評価の単位である「文書」について述べておく。再現率，精度，平均精度を求めるためには，何を「文書」という単位とみなして計算するかを定めなければならない。つまり，検索の結果としてランキングされるものが何なのかを定めておく必要がある。

本研究の場合には，文書画像を単位とする場合と記事を単位とする場合の 2 通りが考えられる。具体的には以下のとおりである。

単位が文書画像の場合，検索結果は 25 枚の文書画像のランキングとして表される。ある検索質問に対して，ある文書画像が関連文書であるとは，検索質問に関連した記事がその文書画像中に存在することとする。

一方，記事を単位とする場合，検索結果は 249 個の記事のランキングとして表される。提案手法では記事の切り出しを考慮していないので，ランキングの結果は次のように計算する。まず，あらかじめ文書画像中の記事の領域をすべて求めておく。次に，ある検索質問に対する記事のスコアを，その記事領域中の出現密度 (式 (6)) の最大値とする。最後に，スコアの降順に全記事をソートし，ランキングとする。これにより，出現密度の値が大きい部分が，どの程度，関連記事に対応しているのかを評価することができる。

〈3･3〉 パラメータの設定と実験方法　提案手法には，式 (2) のしきい値 T，図 2 の窓幅 M の 2 つのパラメータがある。本実験では，一つ抜き法により，パラメータを設定しつつ結果を求めた。例えば，検索質問 1 に対する再現率や精度を求める場合，まず，他のすべての検索質問 (2〜7) を用いて，パラメータ T, M の値を定め，次にその値を用いて，検索質問 1 に対する結果を得た。最終的には，そのようにして求めた検索質問ごとの結果を平均することにより，全体の結果とした。

一つを除く他の検索質問を用いて，パラメータの値を設定する方法は以下のとおりである。T を 10 ステップで 10〜100，M を 50 画素ステップで 100〜3000 画素まで変化させ，文書画像を単位とした平均精度が最大になる組合せを求める。これを，除いた検索質問に対するパラメータの値とする。各検索質問に対して求められたパラメータの値を表 2 に示す。

〈3･4〉 比較手法　比較手法として 2 手法を設け，提案手法と比較した。一方は，提案手法から関連部分検索の機能を取り除いた手法 (以下，MOD とする) であり，他方は，電子文書検索で最も基本的なベクトル空間法 (Vector Space Model; VSM) [11] である。

〈3･4･1〉 MOD　MOD は，提案手法において関連部分検索がどの程度有効に働いているかを検討するために設定した比較手法であり，提案手法で出現密度分布に基づいて検索結果を求める部分を，文字の出現個数を用いるように変更したものである。具体的には，式 (1), 式 (4) の $D_{ij}^{(p)}$，β_{ij} を，次のように変更した。

$$D_{ij}^{(p)} = \mathrm{cf}_{ij}^{(p)} \quad \cdots\cdots (8)$$

$$\beta_{ij} = \log \frac{N}{N_{ij}} \quad \cdots\cdots (9)$$

ここで，$\mathrm{cf}_{ij}^{(p)}$ は，文字 q_{ij} が文書画像 p に現れる頻度を表す。また，N はそれぞれ文書画像データベースに収められた文書画像の総数 (この実験では 25)，N_{ij} は文字 q_{ij} が出現する文書画像数である。

〈3･4･2〉 ベクトル空間法 (VSM)　実験に用いる BMIR-J2 は，そもそも電子文書検索を対象としたデータセットであるため，提案手法を電子文書検索と比較することが考えられる。ここでは，比較対象の手法としてベクトル空間法を採用する。

手法の概要は以下のとおりである。ベクトル空間法では，文書と検索質問を共にベクトルとして表す。ベクトルの次元数は，索引語の異なり数に対応する。索引語としては，対象文書に含まれる名詞相当語句 (名詞や未定義語) を用いる。各次元の値は，その次元に対応する索引語の重みを表す。ここでは，一般的に用いられる TF･IDF 重みを用いる。具体的には，ある文書 j における索引語 i の出現頻度を f_{ij}，全文書数を n，索引語 i を含む文書数を n_i とするとき，$\mathrm{TF} = \sqrt{f_{ij}}$, $\mathrm{IDF} = \log(n/n_i)$ とし，索引語の重みは両者の積とする。文書と検索質問の類似度は，文書ベクトルと検索質問ベクトルのなす角の余弦として定義される。

検索の対象とする電子文書としては，

- CD 毎日新聞に収められた電子文書
- 文字切り出し・認識の結果として得られた電子文書

の 2 通りを考える。後者は，提案手法で索引付けに用いた文字認識結果と同じものである。文字列や読み順 (reading order) の認定は，OCR の結果をそのまま用いた。以下では，前者と後者に対してベクトル空間法を適用したものを，それぞれ VSM(text)，VSM(OCR) と呼ぶ。

また，**〈3･2〉**で述べたように，検索単位の文書としては，提案手法と同様に,「文書画像」と「記事」の 2 通りが考えられる。ここで,「文書画像」を単位とした電子文書検索とは，文書画像に含まれる記事をすべて連結することにより，文書画像に対応する電子文書を新たに作成し，それに対して検索を行うことを意味する。本実験では両者について結果を求めた。

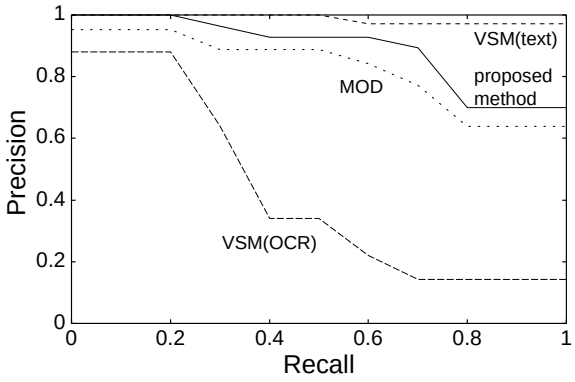


図 4　再現率－精度グラフ (文書画像単位)
Fig. 4.　Recall–precision graph for document images.

表 3　平均精度 (文書画像単位)
Table 3.　Mean average precision for document images.

method	mean ave. precision (%)
proposed method	88.2
MOD	80.9
VSM(text)	98.3
VSM(OCR)	39.4

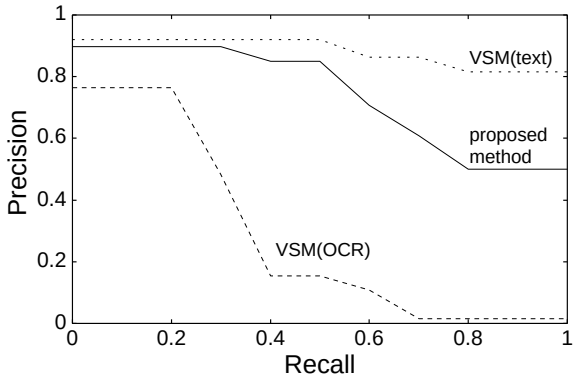


図 5　再現率－精度グラフ (記事単位)
Fig. 5.　Recall–precision graph for articles.

表 4　平均精度 (記事単位)
Table 4.　Mean average precision for articles.

method	mean ave. precision (%)
proposed method	72.9
VSM(text)	86.8
VSM(OCR)	27.7

〈3･5〉 実験結果

〈3･5･1〉 文書画像単位　文書画像単位の再現率－精度グラフを図 4 に示す。

まず，VSM(text)，VSM(OCR) と提案手法の結果を比較する。提案手法は，VSM(text) とは異なり，文字認識誤りを含むテキストを対象としているにもかかわらず，再現率が 0%から 70%程度までは大差ない結果を得ることができた。一方，文字認識誤りを含むテキストを対象とした VSM(OCR) は，ほぼ全域の再現率について，提案手法をかなり下回る結果となった。この理由としては，次のことが考えられる。提案手法は，検索質問を文字に分解し，その密集度合に基づいて文書画像をランキングしている。したがって，一部の文字に認識誤りがあっても，他の文字が正しく認識されていれば，それほど大きな問題を生じない。一方，VSM(OCR) は，索引語を基本とした手法であるため，索引語を構成する文字に一つでも認識誤りがあれば，その索引語としてはもはや取り出すことができない。VSM(text) の結果は，ある意味で，文書画像検索の上限値を表すと考えられるので，文字に分解して密集度合を計測するという提案手法の戦略は，低品質な文書画像を対象とした検索法としてうまく機能していると考えられる。

次に，MOD と提案手法を比較する。MOD は文字への分解は行うものの，文字の分布は考慮しない手法である。提案手法と MOD は共に VSM(text) に近く，VSM(OCR) を上回るものであるため，共に用いている「文字への分解」の方法が文字認識誤りへの対処には有効であると考えられる。ただし，提案手法は MOD を常に上回っていることから，密集度合を計測することは，検索精度の向上に貢献するものと考えられる。

各手法により得られた平均精度を表 3 に示す。これらの値から，提案手法は VSM(text) と MOD の中間的な精度を得る手法であり，VSM(OCR) を十分上回る精度を得ていることがわかる。

〈3･5･2〉 記事単位　次に，記事単位の結果について述べる。再現率－精度グラフを図 5 に示す。 MOD は文書画像単位の検索手法であるため，ここには現れていない。

文書画像単位の場合に比べて，提案手法と VSM(text) の差はより明確になっているものの，VSM(text)，提案手法，VSM(OCR) の関係は文書画像単位の場合と同様であり，この順に精度が優れている。再現率が 0%から 50%の間であれば，提案手法は，VSM(text) と比べてほぼ遜色のない結果を得た。

表 4 に記事単位の平均精度を示す。提案手法の平均精度は，VSM(text) には劣るものの，VSM(OCR) に比べれば十分優れているといえる。

〈3･6〉 考　察　検索質問に含まれる文字の認識率が 74.1%というかなり低品質な文書画像であるにもかかわらず，提案手法は，50%程度の再現率までであれば，電子文書検索手法の VSM(text) に迫る精度を得ることができた。しかしながら，50%を越える再現率の部分では，VSM(text) に及ぶことができなかった。これは，検索質問を構成するキーワードや文字があまり多く出現しない記事においては，少数の文字認識誤りが検索失敗に直接つながるためであると考えられる。再現率が高い部分においても，精度を VSM(text) に近づけるためには，文字認識誤りの傾向を考慮する手法 [3] [4] を導入する必要があろう。

また，検索質問を文字に分解して扱うという提案手法の処理は，文字認識誤りへの対処として効果を挙げた。ただ

し，検索質問を構成する文字が平仮名や頻出する漢字の場合，問題を生じることは十分考えられる。このような文字で構成される検索質問に対しても，有効な検索を実現するためには，文字の隣接関係など，つながりを考慮した処理が必要となるであろう。

最後に，出現密度の高い部分を提示するという提示方法について考える。文書画像データベースにおいて記事単位の検索を実現する最も直接的な手法は，次のようなものである。(1) 認識処理の結果として得られる記事領域を信頼し，文字認識結果とあわせて電子文書としての記事を同定する，(2) 同定した記事に対して電子文書検索を施し，結果を得る。しかしながら，このような方法では，認識処理で記事の同定に誤りが発生した場合，対処が非常に困難となる。一方，提案手法では，紙面で文字が近接することを利用して出現密度分布を求め，文書画像内で読むべき位置を指定する。紙面でのおおまかな位置を指摘されれば，そこから，例えば読み始めの位置を探すのは，ユーザにとってそれほど困難ではないであろう。これにより，上記のような問題を回避しつつ，関連部分検索を可能としている。

4. むすび

文書画像の柔軟な検索を実現するためには，検索質問に関連する部分が文書画像に含まれるかどうかを検査する必要がある。本論文では，このような観点に基づき，文書画像の関連部分検索を提案した。本手法は,「文書画像中で，検索質問に含まれる特徴的な文字が密集している部分は，検索質問に関連する可能性が高い」という考え方に基づくものであり，密集度合を 2 次元出現密度分布として数量化している点に特徴がある。日本語新聞画像を対象とした実験の結果，関連部分検索の機能を持たない手法では，文書画像検索の平均精度が 80.9%であったのに対して，提案手法は 88.2%であり，優位性が示された。

今後の課題は，より多くの文書画像を用いた実験に加え，精度をさらに向上させるために，文字認識誤りへの対処を導入することなどである。

謝　辞

本研究には，CD-毎日新聞 94 年版，テストコレクション BMIR-J2，ならびに形態素解析システム JUMAN を使わせて頂いた。関係各位に感謝する。本研究の一部は，科研費基盤研究 (C)(14580453) の補助による。

（平成 15 年 7 月 11 日受付，同 16 年 5 月 31 日再受付）

文　献

（1） L. Bottou, P. Haffner and Yann LeCun: “Efficient Conversion of Digital Documents to Multilayer Raster Formats”, Proc. ICDAR2001, pp.444–448, Seattle, USA (2001-9)

（2） D. Doermann: “The Indexing and Retrieval of Document Images: A Survey”, *Computer Vision and Image Processing*, **70**, 3, pp.287–298 (1998-6)

（3） K. Marukawa, H. Fujisawa, and Y. Shima: “Evaluation of Information Retrieval Methods with Output of Character Recognition Based on Characteristic of Recognition Error”, *T. IEICE Japan*, Vol.J79-D-II, No.5, pp.785–794 (1996-5) (in Japanese)
丸川勝美・藤澤浩道・嶋 好博: 「認識機能の出力あいまい性を許容した情報検索手法の一検討— 認識誤り特性に着目した検索手法の分析評価 —」, 信学論 (D-II), **J79-D-II**, 5, pp.785–794 (1996-5)

（4） M. Ohta, A. Takasu and J. Adachi: “Retrieval Methods for English-Text with Missrecognized OCR Characters”, Proc. ICDAR’97, Ulm, Germany, pp.957–961 (1997-8)

（5） T. Nakanishi, S. Omachi, and H. Aso: “High Precision Keyword Search System Adapted to Low Quality Document Images”, Tech. Rep. of IEICE, PRMU98-232 (1999-02) (in Japanese)
中西大雅・大町真一郎・阿曽 弘具: 「低品位文書画像に対応した高精度なキーワード検索システム」, 信学技報, PRMU98-232 (1999-2)

（6） J.P.Callan: “Passage-Level Evidence in Document Retrieval”, Proc. SIGIR ’94, pp.302-310 (1994)

（7） S. Kurohashi, N. Shiraki, and M. Nagao: “A Method for Detecting Important Descriptions of a Word Based on Its Density Distribution in Text”, T. IPSJ, Vol.38, No.4, pp.845–854 (1997-4) (in Japanese)
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」, 情報処理学会論文誌, **38**, 4, pp.845–854 (1997-4)

（8） K. Kise, M. Junker, A. Dengel and K. Matsumoto: “Experimental Evaluation of Passage-Based Document Retrieval”, Proc. of the 6th ICDAR, pp.592–596, 2001.

（9） http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html

（10） 木谷 強・小川泰嗣・石川 徹也・木本 晴夫・中渡瀬 秀一・芥子 育雄・豊浦 潤・福島 俊一・松井 くにお・上田 良寛・酒井 哲也・徳永 健伸・鶴岡 弘・安形　輝: 「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 (1998)

（11） R. Baeza-Yates and B.Ribeiro-Neto: Modern Information Retrieval, Addison-Wesley Pub. Co. (1999)

黄　瀬　浩　一

（正員） 1988 年 3 月大阪大学大学院工学研究科通信工学専攻博士前期課程修了。89 年同後期課程退学。90 年大阪府立大学工学部助手。現在，同大学大学院工学研究科助教授。00 年〜01 年大阪府在外研究員としてドイツ人工知能研究センター (DFKI) に滞在。博士（工学）。文書情報処理，情報検索の研究に従事。電子情報通信学会，情報処理学会，IEEE, ACM などの会員。

辻　野　雅　章

（非会員） 2002 年 3 月大阪府立大学大学院工学研究科電気情報系専攻博士前期課程修了。現在，松下電器産業 (株) に勤務。在学中，文書画像検索に関する研究に従事。

松　本　啓之亮

（正員） 1978 年 3 月京都大学大学院工学研究科精密工学専攻修了。同年 4 月三菱電機 (株) 入社。96 年大阪府立大学工学部情報工学科教授となり，現在，同大学大学院工学研究科教授。主に，知識情報処理やソフトウェアに関する研究に従事。工学博士。84 年電気学会論文賞受賞。情報処理学会，IEEE などの会員。