単語の 2 次元分布に基づく 擬似関連フィードバックを用いた文書画像検索

尹 沃棠† 黄瀬 浩一‡ 松本 啓之亮‡
大阪府立大学工学部情報工学科†
大阪府立大学大学院工学研究科情報工学分野‡
e-mail : {yin, kise, matsu}@ss.cs.osakafu-u.ac.jp

1 はじめに

印刷文書から効率よく情報を検索する手段として，文書画像データベースが注目されている．これまでも，様々な検索法が提案され [1]，利用されてきた．しかし，従来法の多くには,「単語と単語の頻度しか使わないため，新聞のように 1 ページに様々なトピックの記事が書かれた文書に対しては，記事の切り分けができず，処理精度が下がる」という問題がある．

黄瀬らは，この問題を解決するため，単語の位置情報を用いた文書画像検索法を提案している [2]．この手法では，文書画像に対して，検索要求に合致する部分が含まれるかどうかを計測するため，上記の問題を解決することができる．ただし，実験では日本語の新聞画像を対象としているだけであり，新聞以外の形式の文書に対する有効性は検討されていない．

そこで本研究では，英語の論文画像を対象として実験を行い，黄瀬らの手法の有効性を検討する．また，さらに精度を向上させるため，電子テキスト検索の擬似関連フィードバックと呼ばれる手法を文書画像検索に導入し，有効性を実験的に検討する．

2 2 次元出現密度分布に基づく文書画像検索

まず，提案手法の基礎となる手法 (2D-DD) を説明する．この手法は黄瀬らの手法を英語論文に適用するために変更したものである．処理の流れを図 1 に示す．

2.1 索引付け

文書画像に対して, 単語とその位置情報を取り出し，索引とする．このとき不要語 (stop word) の除去とステミング (stemming) を施す．

2.2 検索処理

検索質問からタームを取り出し，タームを検索単語に分解する．例えば，検索質問「color and multimedia documents」からは,「color」,「multimedia documents」の 2 つのタームが得られ，各々はさらに検索単語に分解される．

以下では，検索質問 q から得られたタームを $q_1, \cdots, q_n$，ターム q_i に含まれる検索単語を $q_{i1}, \cdots, q_{im}$ と表す．

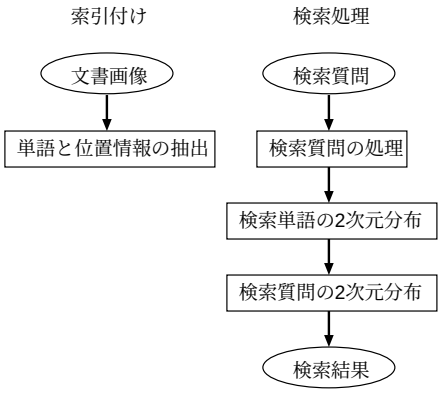


図 1: 2D-DD の処理の流れ

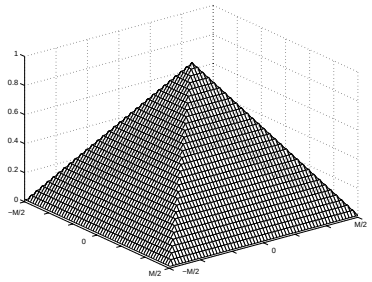


図 2: 窓関数

文書画像 p における検索単語 q_{ij} の 2 次元分布 $D_{ij}^{(p)}$ は，検索単語 q_{ij} の出現位置に窓関数 (窓幅 M)(図 2) の中心をあわせて置き，置かれた窓関数の値を画素ごとに加えて計算できる．

次に，文書画像 p におけるターム q_i の 2 次元分布 $D_i^{(p)}$ を，次のように求める．

$$D_i^{(p)}(x,y) = \prod_{j=1}^{m} \mathrm{Idf}_{q_{ij}} D_{ij}^{(p)}(x,y) \qquad (1)$$

ここで，$\mathrm{Idf}_{q_{ij}}$ は検索単語の重みであり，$\mathrm{Idf}_{q_{ij}} = \log(n/n_{q_{ij}})$ により与えられる．式中の n は全文書数，$n_{q_{ij}}$ は検索単語 q_{ij} を含む文書数である．

最後に，ターム q_i の 2 次元分布 $D_i^{(p)}$ から，文書画像 p における検索質問 q 全体の 2 次元分布 $D^{(p)}$ を次の式に基づい

Document Image Retrieval Using Pseudo-Relevance Feedback on Two-Dimensional Distributions of Terms
Yin Wuotang†, Koichi Kise‡ and Keinosuke Matsumoto‡
†College of Engineering, Osaka Prefecture University
‡Graduate School of Engineering, Osaka Prefecture University

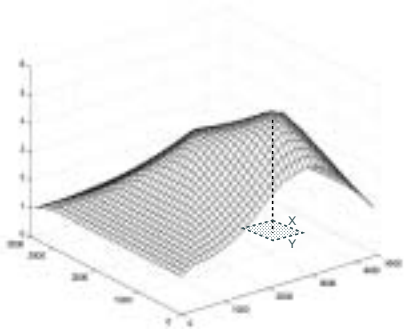


図 3: 検索質問の拡張

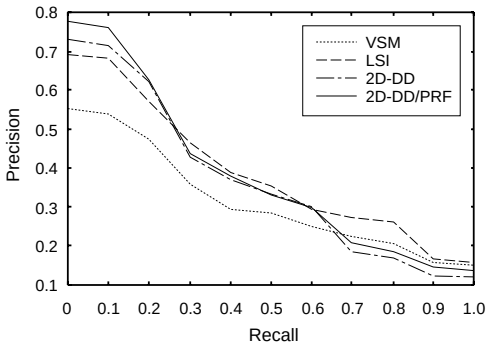


図 4: 実験結果

て求める．

$$D^{(p)}(x,y) = \sum_{i=1}^{n} D_i^{(p)}(x,y) \qquad (2)$$

検索質問の 2 次元分布の最大値 $\max_{x,y} D^{(p)}(x,y)$ 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．そして，最上位の文書画像から順に，ユーザに提示する．

3 擬似関連フィードバック

2D-DD はマルチトピックの文書を検索するのには適しているが，一般の文書のように必ずしも複数のトピックが含まれない文書に対しては，有効性が未知数である．特に，検索質問に含まれる単語が極端に少ない (数語) の場合には，検索に必要な情報を得ることができずに，精度が下がってしまうことが考えられる．

この問題を解決し，より検索精度を高めるため，本研究では，擬似関連フィードバック (Pseudo-Relevance Feedback) の導入を考える．以下では，この提案手法を 2D-DD/PRF と呼ぶ．具体的な処理は以下のようになる．

2D-DD の検索結果から，検索質問 q の 2 次元分布の最大値が大きい N 個の文書を選択する．そのような文書に対する 2 次元分布の例を図 3 に示す．次に，各文書において，検索単語が密集する箇所 $(X \times Y)$ から新しい単語 (拡張単語)$e_1, \cdots, e_k$ を取り出し，検索質問を拡張する．窓関数を用いて，各拡張単語の 2 次元分布 $E_i^{(p)}$ は同様に計算できる．拡張検索質問の 2 次元分布は次の式で求める．

$$D'^{(p)}(x,y) = D^{(p)}(x,y) + \alpha \sum_{i=1}^{k} \mathrm{Idf}_{e_i} E_i^{(p)}(x,y) \qquad (3)$$

4 実験

4.1 実験条件

提案手法の有効性を評価するため，英語論文の文書画像を対象として検索実験を行った．検索対象の文書画像は，ICDAR2001 に投稿された論文のうち，オーラルセッションのものの一部 (129 文書，673 ページ) である [4]．画像サイズは 3,500 × 4,500 画素である．検索質問は，論文集の目次にあるセッションのタイトルから設定した．検索質問数は１９であり，平均検索質問長は 2.37 単語である．検索タスクは，セッションタイトルからそのセッションに属する論文を検索するものとした．比較手法としては，ベクトル空間法 (Vector Space Model; VSM) [3] と潜在的意味インデキシング (Latent Semantic Indexing; LSI) [3] を用いた．なお，処理に用いたパラメータ値は，従来法を含めて予備実験により決定した．

4.2 実験結果

再現率 (Recall) と精度 (Precision) を用いて結果を評価した．ここで，S を検索された文書の集合，T を正解文書の集合とすると，再現率 R は，$R = |S \cap T|/|T|$，精度 P は，$P = |S \cap T|/|S|$ である．

実験の結果を図 4 に示す．まず，2D-DD と 2D-DD/PRF を比べると，再現率の低い部分と高い部分で，後者の精度が前者の精度を上回っていることがわかる．また，2D-DD/PRF と LSI，VSM を比べると，低い再現率の部分で，2D-DD/PRF の精度が高いことがわかる．以上より提案手法の有効性が明らかとなった．

5 まとめ

本稿では，文書画像から求めた単語の 2 次元分布を用いて，単語が密集する箇所から新しい単語を取り出し，検索質問を拡張することにより，文書画像データベースの検索精度が向上することを示した．

今後の課題としては，検索質問を拡張する方法を改良し，さらに高い検索精度を追求することがある．

参考文献

[1] Doermann,D.：“The Indexing and Retrieval of Document Images: A Survey”,*Computer Vision and Image Processing,* Vol.70，No.3，pp.287-298，1998.

[2] 黄瀬浩一，辻野雅章，松本啓之亮：“文字の 2 次元出現密度分布に基づく文書画像の部分検索”，信学技報 PRMU2002-16，2002.

[3] 北研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版，2002.

[4] http://icdar.djvuzone.org/