画面イメージから取り出したテキストに対する情報検索の課題

熊谷　摩美子　阿部 洋丈　岡部 正幸　梅村 恭司　（豊橋技術科学大学）

研究の背景と目的

コンピュータの操作画面を画面イメージとして一定時間間隔で蓄積し、画面イメージをOCR処理してOCRテキスト（**画像OCRコーパス**）を得るシステムを実装した。画像OCRコーパスには特徴として、一定間隔で画面イメージを取得するため、全体または一部の似通った画面イメージの取得の繰り返しが生じ、OCRテキストに同じ文字列の繰り返しが見られる。通常の検索テキストとは文字列分布が異なると考える。よって、文字列の分布に関して、通常の検索テキストと比較を行い、通常の検索における文字列の重み付けについて従来の方法でよいのか考察する。

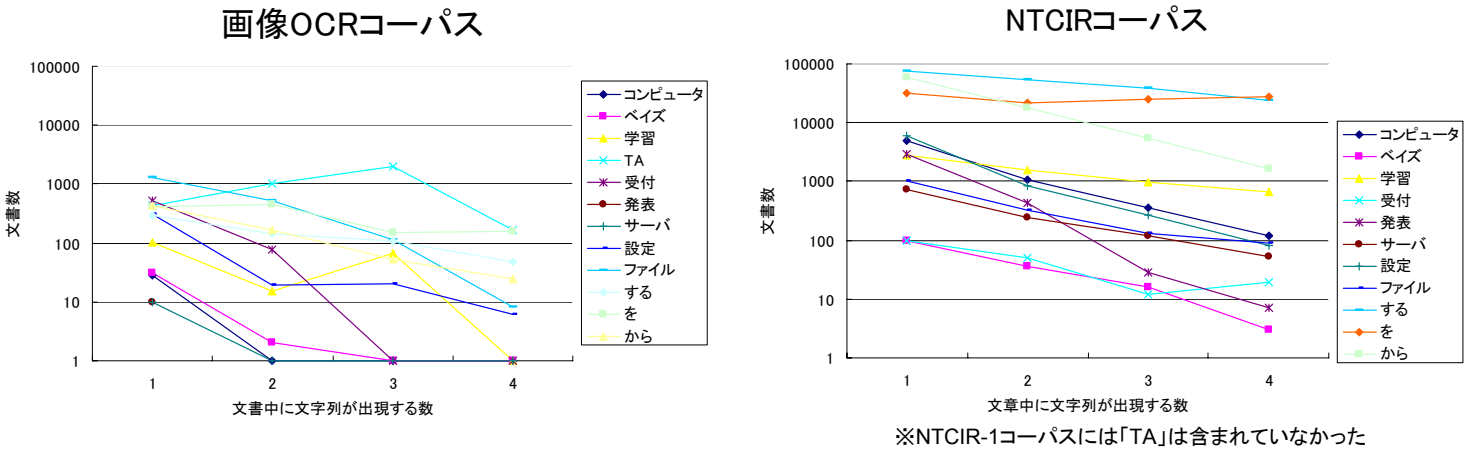
実装したシステム

実装したシステムは、コンピュータで見た情報の検索の実現を目的としている。基本コンセプトを図1に示す。コンセプトとしては、見つけたい情報に関するキーワードを入れると、検索結果として見つけたい情報を含む画面イメージを表示する。

図1．システムの基本コンセプト

画像OCRコーパスの特徴

NTCIR-1コーパス（332918件の文書）という通常のコーパスとの比較により、画像OCRコーパス（4456件の文書）の特徴について述べる。画像OCRコーパスは、実装したシステムを実際に自分のコンピュータで実行して得た文書集合である。実装したシステムで検索キーワードとなりそうな文字列（検索文字列）と通常の検索におけるストップワードが文書中にn回現れる文書の数を求める（n=1,2,3,4）。各文字列の分布を図2に示す。



※NTCIR-1コーパスには「TA」は含まれていなかった

図2.NTCIR-1コーパスと画像OCRコーパス文字列の分布

検索文字列の分布は、NTCIR-1コーパスの場合、単調減少する分布となっている。画像OCRコーパスの場合、明らかに単調減少しない分布がみられる。主として複数のOCRテキストに同じ文字列が繰り返し出現するため、**検索文字列の分布が通常コーパスであるNTCIR-1コーパスとは異なるという特徴**が見られた。

検索の問題と課題

情報検索においては、単語の分布と適切な重み付けには関係があることが知られている[1]。画像OCRコーパスでは、通常のコーパスとは異なり、検索文字列の分布が単調減少でないため、検索文字列に従来の重み付けの局所的重み付けが適応できず、期待した結果が得られないと考える。画像OCRコーパスおいて、検索における従来の文字列の重み付けでは問題がある。

画像OCRコーパスのための新しい情報検索アルゴリズムに関する検索の重み付けを考える必要があるが、具体的な重み付けを考え、それを評価することは現状できておらず、今後の課題である。

[1]情報検索とアルゴリズム：北研二，津田和彦，獅々堀正幹.共立出版(2002年).ISBN4-320-12036-1.