Vol. 40 No. 12 情報処理学会論文誌 Dec. 1999

テクニカルノート

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

水野浩之† 黄瀬浩一† 松本啓之亮†

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，その一例として文書中の各図表に対する説明箇所の範囲を特定する手法について述べる．本手法の特徴は，以下のとおりである．(1) 図表から自立語を抽出してキーワードとし，それらのテキスト中での分散を調べ，偏り具合に応じて重み付けをする．(2) 図表と説明箇所の対応付けにキーワードの出現密度を用いる．10文書，78図表を用いた実験の結果，適合率50.3%，再現率72.1%を得た．

Linking Figures and Tables to Their Expository Texts Using Word Density Distributions and Their Biases

HIROYUKI MIZUNO,† KOICHI KISE† and KEINOSUKE MATSUMOTO†

Passage retrieval is an important subfield of document retrieval. As an example of passage retrieval, we propose a method of extracting texts which explain a figure/table. The characteristics of the method are as follows: (1) Each figure (table) is represented as a set of keywords, i.e., content words in it, each of which is weighted according to the bias of keyword density distribution in a text. (2) Each figure (table) is correlated with its expository texts by thresholding the density distribution of its keywords. From the experiments with 10 documents, we obtained 50.3% precision and 72.1% recall.

1. まえがき

近年，文書の電子化にともない文書検索の研究がさかんになっている．従来の文書検索は，検索対象として主に文書単位を扱ってきた．しかし，利便性をさらに高めるため，検索結果として文書中の部分を提示できるようなものが求められている．ここでは，そのようなものを部分テキスト検索（passage retrieval）と呼ぶ．

本稿では，部分テキスト検索の対象として，図表とそれを説明するテキスト（説明テキスト）に着目し，黒橋らによって提案された語の出現密度分布[1]を利用して，文書中の各図表に対する説明テキストの範囲を自動的に特定する手法を提案する[2]．出現密度分布とは，文書における語の出現の疎密を数量化したものである．本手法の特徴は，図表をその中で使われている単語とキャプションの単語を用いて表現し，それらのテキスト中での分布の偏りから求めた偏出度を語の重み付けに利用している点にある．

† 大阪府立大学工学部情報工学科
Department of Computer and Systems Sciences, College of Engineering, Osaka Prefecture University

2. 図表と説明テキストの対応付け

2.1 単語の出現密度分布と偏出度

図表は，情報表現の中心である本体部分，図表番号，キャプションから構成される．図表には厳密な書式がないので，単独で情報を曖昧性なく表現するのは難しく，テキストから説明を受けることが多い．ここでは，そのようなテキストを図表の説明テキストと呼ぶ．

テキスト中で図表を説明する場合，図表番号を用いて明示的に引用することが多いため，説明テキストは図表の引用箇所の周辺に存在しているといえる．また，説明テキストは図表が引用されている場所の周辺だけでなく，テキスト全体に点在していることもある．いずれの場合でも，説明テキストには図表を説明するために図表中の語を繰り返し用いるという傾向がある．

本手法は，「テキストにおいて図表中の語を使用する

頻度が高い部分はその図表の説明テキストである可能性が高い」という考えに基づいて，図表と説明テキストの対応付けを行う．語が密集している度合いを出現密度という値で表し，テキストにおける出現密度分布を調べることによって説明テキストを特定する．ただし，テキスト全体にわたって出現するような一般的な語は，説明テキストを特定する際にノイズとなる．このような語の影響を抑えるために，語の出現の偏りに応じて重み付けをする．ここでは，偏りの度合いを偏出度という値で表す．

2.2 処理のながれ

以下に具体的な処理手順について説明する．

(1) 図表からのキーワードの抽出

図表中で使われている語とキャプションの語を人手で抽出し，形態素解析（juman3.5）[3]によって単語に分割し，その中の自立語をキーワードとして抽出する．

(2) テキストにおけるキーワードの検出

テキストも同様に形態素解析を用いて単語に分け，単語 $a(l)\,(0 \le l < L)$ の列として表現する．ここで L は，テキスト中の全単語数であり，l は単語の出現位置を表す．次に，キーワードのすべての出現位置を調べる．位置 l にキーワード k が出現する場合 $m_k(l) = 1$，出現しない場合 $m_k(l) = 0$ とする．

(3) 偏出度に応じたキーワードの重み付け

図表の説明テキストを同定するためには，全体にわたって出現するキーワードよりも局所的に出現するようなキーワードの方が重要である．ここでは，各キーワードに出現の偏りが大きいほど高い値をとる偏出度を用いて重み付けをする．偏出度の算出には，文献 1) で導入されているハニング窓関数で計算された出現密度を使う．ハニング窓関数 $h(i)$ は，窓幅（重みを与える範囲）を W とすると以下の式で表される．

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

キーワード k に対する出現密度 $d_k(l)$ は以下の式で計算する．

$$d_k(l) = \sum_{i=-\frac{W}{2}}^{\frac{W}{2}} h(i) \cdot m_k(l-i) \quad (2)$$

（$i < 0$ または $i \ge L$ では $m_k(i) = 0$ とする）

この $d_k(l)$ を用いてキーワード k の偏出度 $b(k)$ を以下の式で求める．$\max_l d_k(l)$ はキーワード k に対する最大出現密度である．

$$b(k) = \begin{cases} 0 & (\forall l\ d_k(l) = 0) \\ \frac{1}{L}\sum_l \left(1 - \frac{d_k(l)}{\max_l d_k(l)}\right) & (\text{otherwise}) \end{cases} \quad (3)$$

偏出度 $b(k)$ は $d_k(l)$ が一定の場合に 0，1 カ所に集中する場合に最大値をとり，一般にはキーワードの分布が偏っているほど高い値を示す．$b(k)$ を用いてキーワード k に対し，以下のように重み $w(k)$ を与える．

$$w(k) = b(k)^n \quad (4)$$

ここで，$n\,(n \ge 0)$ は偏出度による重みの程度を調整するパラメータであり，n が大きいほど偏りをより重要視することを意味する．この n を重み付け次数と呼ぶ．なお，$n = 0$ のときは $b(k) = 0$ ならば $w(k) = 0$，それ以外ならば $w(k) = 1$ を表すものとする．これは重み付けに偏出度を用いないことを意味する．

(4) 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，ある図表に現れるすべてのキーワードの出現密度を計算する．位置 l に対するキーワード全体の出現密度 $d(l)$ はキーワード k に対する重み $w(k)$ と出現密度 $d_k(l)$ を用いて次式で計算される．

$$d(l) = \sum_k w(k) \cdot d_k(l) \quad (5)$$

(5) 説明テキストの抽出

出現密度 $d(l)$ をその最大値 $\max_l d(l)$ で正規化したものを $\hat{d}(l)$ とする．**図 1** のように閾値 T を設定し，$\hat{d}(l) \ge T$ を満たす位置 l の単語を考え，それを含

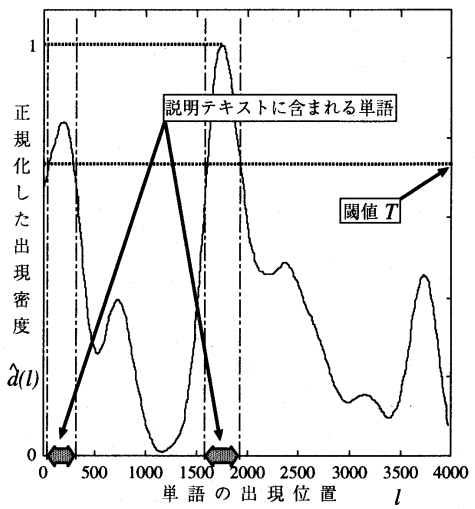

図 1 キーワードの出現密度を用いた説明テキストの抽出
Fig. 1 Selection of expository texts using density distribution of keywords.

む文を図表に対する説明テキストとして対応付ける．

3. 実　　験

3.1 実験条件

学習用データセットを用いたパラメータ決定実験（実験1），評価用データセットを用いた評価実験（実験2）の2種類を行った．各データセットは解説記事が5つ，教科書形式のものから3章分，論文形式のものが2つの計10サンプルで成り立っている．各データセットの概要を**表1**に示す．

各サンプルに対する正解データ，すなわち，サンプル中の各図表に対応する文は，本手法を知らない4人の大学生に設定してもらった．このとき，本文には手を加えずそのまま提示するとともに，図表番号の周辺以外でも関連すると思われる箇所は正解に含め，複数箇所を正解としてもよいとした．また，各図表の正解データ設定を異なる2人が担当するようにし，各人が正解として認めた部分の論理和を正解データとした．

対応付けの結果は，正解データに対する適合率 P と再現率 R を用いて，両者の和 $P+R$ を評価値として評価した．適合率 P と再現率 R は，図表に対する正解データの文に含まれる単語の数を N_C，対応付けられた文に含まれる単語の数を N_L，対応付けられた文の中で正解データと一致する文の単語の数を N_E とすると $P=N_E/N_L$，$R=N_E/N_C$ となる．ここで，単語の数で評価する理由は，文長を考慮するためである．

実験1では，重み付け次数 n を0から5.0まで0.5刻み，窓幅 W を単語単位で400から800まで50刻み，閾値 T を0.50から0.90まで0.05刻みで動かしたとき，評価値が最大となる n, W, T の組を求めた．

実験2では，実験1で設定したパラメータ n, W, T を用いて評価用サンプルに対する実験を行った．

表1　実験で用いたデータセット
Table 1　Data sets used in the experiments.

	学習用	評価用
サンプル数	10	10
内容	エレクトロニックコマース，情報処理，ソフトウェア，マーケティング戦略，ハードウェア	情報検索，生物，地学，マーケティング戦略，インテリジェント交通システム，ソフトウェア
各サンプルの平均単語数	5272	4842
図表の数	図81，表18	図67，表11
各図表の平均キーワード数	25.0	24.5
各図表の明示的引用の平均回数	1.17	1.30
各図表の正解データの平均文数	25.7	20.9

3.2 実験結果と考察

実験1の結果，最大の評価値を与えるパラメータは $n=4.0$，$W=550$，$T=0.60$ となり，このとき適合率55.3%，再現率74.5%を得た．実験1において，$n=4.0$ で窓幅 W と閾値 T を変化させたときの適合率と再現率の変化を**図2**に示す．また，比較対象として偏出度を用いない場合（$n=0$）と，文単位の出現密度を用いた場合，すなわち各文に対して M（＝その文中でキーワードとマッチした単語数/その文の全単語数）を求め $M \geq T_{base}$ を満たすものを出力とする場合の適合率と再現率の変化も示す．このグラフから本手法は，文単位の出現密度を大幅に改善したものであることが分かる．また，偏出度による改善効果も明らかである．窓幅と閾値については，窓幅を広げ，閾値を下げることによって対応付けられる範囲が広がり，再現率重視の対応付けとなる．逆に適合率を重視する場合は，窓幅を狭め，閾値を上げればよい．

学習サンプルの全正解データのうち文が連続するものを1ブロックとして考えると，全部で197ブロックあった．この中で，図表を明示的に引用していないものは85ブロックであり，最大評価値を得たパラメータを用いて対応付けを行ったところ，ブロック中の1文でも対応付けることができたものは31ブロックであった．このことから，本手法は図表を明示的に引用

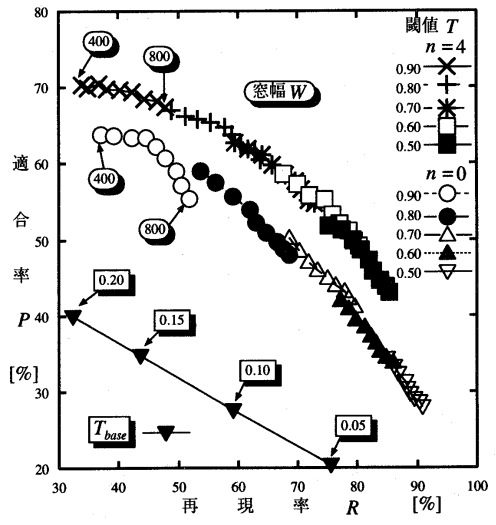

図2　実験1における適合率と再現率の変化（$n=0, n=4$，文単位の出現密度）
Fig. 2　Recall and precision in the experiment 1 ($n=0, n=4$, and with the density distribution for each sentence).

キーワード

S/W, HTML, サーバ, 変換, 電子, 商品, 商店, 作成, 決済, モールサーバ, バーチャルショップ, データベース, ツール, コンテンツ, アプリ, Web, DB, CGI

太字は出力，点線内は正解

通常，送信したい原データ（電文）からメッセージダイジェストと呼ばれるコンパクトで一定サイズのデータを生成するときにこのハッシュ関数を用いる．**この共通鍵暗号方式，公開鍵暗号方式，ハッシュ関数を組み合わせて使うことにより，EC で必須要件の秘匿，相手認証，非改ざん証明などが可能となった．3 種の方式をどのように組み合わせてこれらを実現するかの原理については本特集「5.EC の技術動向：セキュリティ技術」に譲る．**

3.3 電子モール構築技術

電子モールは EC の中でバーチャルショップを開設するプラットフォームである．これはデータベース，Web サーバ，決済処理ソフトウェアを組み合わせて構築される一種のオンライントランザクションサーバである．Web サーバに掲載する HTML コンテンツの作成にはデザイン的要素が強く，専門のデザイナがコンテンツ作成ツールを用いて作成することが多い．ほかに，データベース内のデータから HTML データを自動生成する変換ソフトウェアもあり，デザイナが作成するコンテンツと自動生成されるコンテンツを組み合わせて美麗で機能的なバーチャルショップが構築できる．コンテンツの作成技術に関しては本特集「6.EC の技術動向：デジタルコンテンツ作成流通技術」に譲る．電子モールの構築は個別のアプリケーション開発的色彩が強く，一般論的には論じにくいが，一例として図-1 に示すような構成がとれる．消費者側の端末からは Web サーバにアクセスして商品情報を閲覧し，消費者側の決済ソフトウェアと呼応して動く商店用決済ソフトウェアが決済サーバと連携して電子決済を行う．

3.4 電子決済技術

現在提案されている電子決済技術は現実の経済活動で利用されている決済方式のいずれかをモデルとしている．これらに共通する特徴をひと口でいうと「電子署名つきの金銭価値情報」ということができる．主な電子決済方式として電子現金決済，電子プリペイド型決済，電子個人小切手決済，デビットカード決済，カード決済などがある．

図 3 説明テキストの出力例
Fig. 3 An example of processing results.

していないような説明テキストでもある程度は対応付けが可能であるといえる．

実験 1 で得られたパラメータを用いて実験 2 を行った結果，適合率 50.3%，再現率 72.1%となった．学習用と評価用のサンプルが異なる文書からとられているにもかかわらず，大差のない結果となった．このことから，本手法による対応付けは文書の記述内容にそれほど左右されないといえる．

次に，対応付けられたテキストを詳しく見ていくことにする．具体的な出力例を**図 3** に示す．これは，情報処理 38 巻 9 号 p.774 の図に対して説明テキストを対応付けた結果である．太字部分が出力結果で，点線で囲まれた部分が正解データである．この例では，正解データをカバーする出力が得られているが両端に余分な文が入っている．

実際に対応付けられたすべてのテキストを見ると，ほぼ図表に何らかの関連があるテキストが対応付けられていた．しかし，適合率，再現率ともに 100%という結果はほとんどなかった．この理由は，図 3 の例のように説明テキストの境界部分でいくらか過不足を生じるためである．これに対処するためには，検索対象文書ごとに動的に閾値を変更する必要があろう．

対応付けに失敗しているものは，主に図表中で使われているすべての語が文書全体でよく使われていたり，図表の内容が例を示すもので文書内で説明がほとんどされていないものであった．前者の図表については，その文書の内容に対してのまとめが書かれている部分がいくつか対応付けられていた．また，後者のような図表は，読者に示すことで理解させるものだと考えられるので，説明テキストの必要性は薄い．この場合，図表にテキストを対応付けないための処理が必要である．

4. むすび

本稿では，部分テキスト検索の一例として，文書中の各図表に対して説明テキストの範囲を自動的に特定する手法を提案した．今後の課題としては，より汎用性や精度を向上させるために，大量サンプルに対する実験や対象に合わせた自動パラメータ設定があげられる．

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，情報処理学会論文誌，Vol.38，No.4，pp.845–854 (1997).
2) 水野浩之，黄瀬浩一，松本啓之亮：出現密度分布を用いた図表と説明テキストの対応付け，第 57 回情報処理学会全国大会論文集，Vol.4V-1 (1998).
3) 黒橋禎夫，長尾 眞：日本語形態素解析システム JUMAN version3.5，京都大学工学部大学院工学研究科 (1998).

(平成 11 年 5 月 17 日受付)
(平成 11 年 9 月 2 日採録)