図表と説明テキストの対応付けを利用した重要図表抽出

4V－2

野久 仁志　黄瀬 浩一　松本 啓之亮

大阪府立大学 工学部

1 はじめに

多くの文書ではテキストとともに図や表が用いられる．図表は複雑な構造をわかりやすく表現できるので，人間の直観的な理解を助ける．そのため文書の概略を把握する時には，図表だけを拾い読みすることがよく行われる．したがって文書の主題と密接に関連した図表をあらかじめ特定しておけば，より効率良く文書の概略を把握することができる．

本稿では，電子文書においてテキストと併用される図表の中から，文書の主題に関連する重要な図表を抽出する手法について述べる．本手法の特徴は，図表とその説明テキストとの関係を対応付けで数値化し，説明テキストの重要度を用いて図表の重要度を推定する点にある．

2 図表の特徴

図表は視覚を通じて複雑な構造や関係を表現できるため，幅広い分野の文書において利用されている．主な利用例として，システムの構成を示す図や実験結果をまとめた表などがある．これらの図表は，テキストによる表現では冗長で難解になる説明を簡潔に表すので，直観的な理解をする上で有用である．

しかし図表による表現には，おおまかな原則はあるが厳密な規則はない．したがって，図表が単独で曖昧性なく情報を伝えることは難しい．そのため通常は，自然言語を併用して図表を説明する．図表を説明している本文テキストを説明テキストと呼び，本文テキスト以外で図表に付随している自然言語文をキャプションと呼ぶ．これらは図表とともに相補的に情報を表現している．

このような特徴を持つ図表を計算機で直接扱うことは非常に困難である．図表は明確な書式がないパターン情報であるので，計算機が図表中で使われている図形を解析し，図表の意味や文書中での重要度までを判断することは現実的でない．そのため何らかの近似表現を用いて図表の重要度を推定する必要がある．

3 重要図表抽出

重要図表抽出手順を図1に示す．図表を直接解析して重要度を求めることは困難であるので，本手法では図表に対する説明テキストを利用する．図表と説明テキストとの関連を調べる対応付けの後に，対応付けられた説明テキストの文書中での重要度をもとに図表の重要度を推定する．テキストの重要度も，複雑な自然言語処理を避け数値計算から近似的に求める．

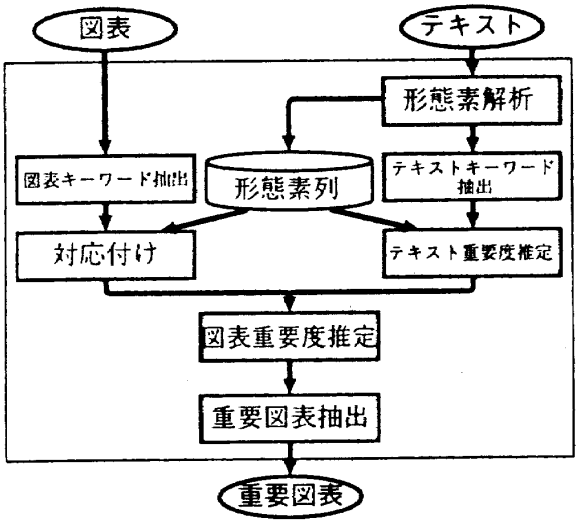

図1 重要図表抽出手順

3.1 図表と説明テキストの対応付け

ある語に関して自然言語で説明する場合には，必然的にその語を繰り返し用いる[1]．同様に本文テキストによって図表を説明する場合にも，図表中で使われている語もしくはキャプション中の語を繰り返し用いる傾向がある．この特徴を利用してテキストでの図表キーワードの出現密度計算[2]から，図表と説明テキストを対応付ける．手順を以下に示す．

1. 図表キーワード抽出… 図表中の語とキャプションを形態素解析し，自立語を図表キーワードとする．
2. テキスト中のキーワード検出… テキストを形態素解析し，得られた語の列（テキスト）中の各図表キーワードの出現位置を調べる．
3. 出現密度計算… 図表キーワード周辺の語に，ハニング窓関数で出現の影響を与え[3]，それらをを加算することにより各位置の出現密度を求める．

対応付けの結果，図2のような出現密度分布を得る．図2の横軸は語の位置，縦軸は出現密度である．ハニング窓関数により，図表キーワードが密集する部分の出現密度が高くなる．図表キーワードが密集する部分は，図表の説明テキストである可能性が高いので，出現密度の大小は図表とテキストとの関連の大小を示すと考える．

3.2 テキスト重要度の推定

テキストの重要度を厳密に判断するには，複雑な自然言語処理が必要となる．しかしこのような処理には多くの知識と時間を要するため，本手法ではキーワードの出現密度からテキストの重要度を近似的に推定する．

Selection of Important Figures and Tables Using Links to Their Expository Text

Hitoshi Nohisa, Koichi Kise and Keinosuke Matsumoto
College of Engineering, Osaka Prefecture University

テキストは形態素解析し，語の列として扱う．文書の主題と密接に関連する語をテキストキーワードとし，出現密度を求める．前述のようにある語を説明する場合には，その語を頻繁に使う傾向がある．したがって，文書の主題に関する部分とテキストキーワードの出現密度が高い部分は一致する可能性が高い．本手法では，テキストキーワードの出現密度をテキスト重要度と定義する．テキストキーワードの候補として次の語が考えられる．

- 高い頻度で本文テキスト中に出現する自立語
- 文書のタイトル，章や節のタイトル中の自立語

それぞれの候補に含まれる語は大部分が共通しているので，本手法では抽出の容易なタイトル中の自立語[a]をテキストキーワードとする．

本文テキスト中のテキストキーワードの出現密度から図3のようなテキスト重要度グラフを得る．

3.3 対応付けを利用した図表重要度推定

本手法では図1に示すように，図表の対応付け結果 (図2) とテキスト重要度推定結果 (図3) から図表重要度を推定し，重要図表を抽出する．

図表重要度推定 対応付け結果 (図2) を出現密度の最大値で正規化する．正規化により図表と説明テキストとの関連を0から1の間の数値で表現できる[b]．一方，テキスト重要度グラフ (図3) は，本文テキスト内部の重要なテキストの分布を数値で表している．そのためこれらの数値の積は，本文テキスト全体と図表との関連を表していると考えられる．これを図表重要度と定義し，図4実線のような図表重要度グラフを得る．

重要図表抽出 図表重要度グラフ (図4: 実線) は，テキスト重要度グラフ (図4: 破線) を変形・縮小した形状になる．したがって，テキスト重要度グラフに対する図表重要度グラフの面積比を求め，面積比の大きい順に図表を並べると図表に順位をつけることができる．面積比が最大となる図表は，他の図表に比べて文書の主題と類似した情報を表しているので，重要図表として抽出する．

4 処理例

本手法の処理例を示す．サンプルには情報処理学会学会誌の電子商取引に関する解説記事[4]を用いた．サンプルは，図を4個と表を1個含み，本文テキストは167文から構成されている．形態素解析には，JUMAN 3.5[5]を用いた．また対応付けとテキスト重要度推定に用いるハニング窓関数の窓幅は600語とした．

図2〜図4は，文献[4]の"図3: 画像半開示と透かしの組合せ"に対する各グラフである．この図の図表重要度の面積比は31.0%で全図表中の第5位となり，重要図表として抽出されなかった．文献[4]の重要図表として抽出されたのは，"図1:EC アーキテクチャ"で図表重要度の面積比は74.8%であった．この図は文書の主題を図示しており，人間の重要度の判定とも一致している．

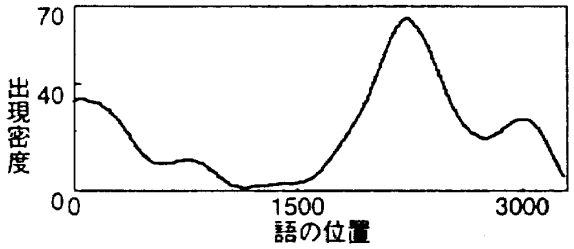

図2 図表キーワードの出現密度分布

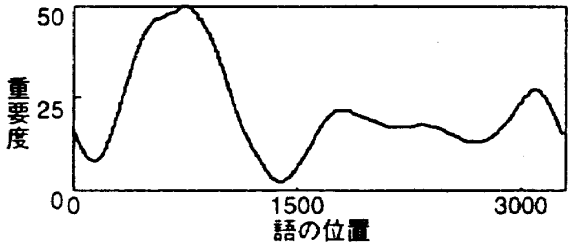

図3 テキスト重要度グラフ

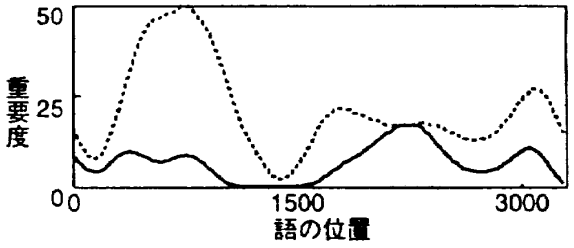

図4 図表重要度グラフ

5 おわりに

本稿では，電子文書で用いられる図表の中から重要図表を抽出する方法を提案した．本手法の特徴は，図表と説明テキストの関連を対応付けにより出現密度分布で表し，説明テキストの文書中での重要度から図表の重要度を推定する点である．

今後の課題としては，図表抽出精度向上のためのキーワード選択法および図表重要度の推定法の改良，ならびに重要図表と重要テキストから構成される図表付き抄録作成への応用があげられる．

参考文献

[1] 新谷 研，角田 達彦，大石 巧，長尾 眞: "単語の共起頻度と出現位置による新聞の関連記事の検索手法"情報処理学会論文誌，Vol.38, No.4, pp.855-862(1997).

[2] 水野 浩之，黄瀬 浩一，松本 啓之亮: "単語の出現密度分布を用いた図表と説明テキストの対応付け"情報処理学会第57回全国大会 4V-1(1998).

[3] 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).

[4] 安原 隆一: "EC の技術動向: デジタルコンテンツ作成流通技術"情報処理学会学会誌，Vol.38, No.9, pp.785-791(1998).

[5] 黒橋 禎夫，長尾 真: "日本語形態素解析システム JUMAN version 3.5"京都大学工学部大学院工学研究科 (1998).

[a] 形式名詞，副詞的名詞，ひらがなだけから構成される動詞は除く．

[b] 起伏が激しいグラフと平坦なグラフで最大値が等しい場合には，最大値での正規化によって後者の方が大きな面積となる．