画像データに対するキーワードの話題境界を考慮した自動抽出手法の改良

岡田　真
大阪府立大学大学院　理学系研究科
okada@mi.s.osakafu-u.ac.jp

1. はじめに

近年，インターネット上には画像・動画といったマルチメディアデータが偏在するようになってきている．これらのマルチメディアデータを内容に基づいて検索したいという要求は古くからあり，さまざまな研究がおこなわれている．

以前，我々はウェブ上のマルチメディアデータの一つとして画像データを選び，それらの効率的検索のために付加するキーワードをその画像が含まれるウェブページ内から自動的に判別，抽出する手法について提案した[1]．この手法では画像データにキーワードを付与する際に，話題境界を判定して，同一話題の画像キーワードに画像を付与する必要があるが，話題境界の判定手法の精度面で十分ではなかった．そこで，今回は新たに中野らによる話題結束力による判定法[5]を用いることで話題境界の判定精度を向上させてより適切なキーワードが付与できるようにすることを試みた．

本稿では話題結束力による判定法をウェブ上の画像データに付加するキーワードを自動的に判別，抽出するシステムへ組み込んだ際の有効性について検証した．

2. システム概要

本研究ではウェブ上の画像に対してその画像を含んでいるファイルのテキストデータから以下のような手順で適切なキーワードを抽出する．

まず，ウェブ上のデータに対して，内部のテキストデータの形態素解析をおこなってキーワード候補を得る．次に，そのキーワード候補に基づいてデータ内の話題境界を見つける．これは，画像データとキーワード候補がどの話題に属するかを判定し，同じ話題内の画像データとキーワード候補群を関連が強いものとして結びつけるためである．この際，キーワード候補に方向を示す名詞(方向指示名詞)があった場合は，HTML タグなどのページ構成情報を用いて，名詞の指す方向における画像の有無を判定する処理をおこなう．また，抽出したキーワード候補について，各ファイル間での IDF(Inverse Document Frequency)値を用いて不要語を削除する．

このようにしてキーワード候補が得られる．次節では話題境界の判定手法について説明する．

3. 話題境界の判定

本稿では(1)スロット法[1]，(2)話題結束力による判定法[5]という２つの話題境界判定手法について説明する．

3.1　スロット法

スロット法では，まず，テキストデータを文単位に分割し，それぞれの文でキーワード候補の抽出をおこなう．次に，話題境界を探るために文を一定の個数ごとにまとめる．本研究では対象としたデータの性質から 3 文を一まとめとした．（以下，このまとまりを「窓」と呼び，一まとめとする文の数を「窓の幅」と呼ぶ．）

一窓ごとにどのようなキーワード候補が含まれ

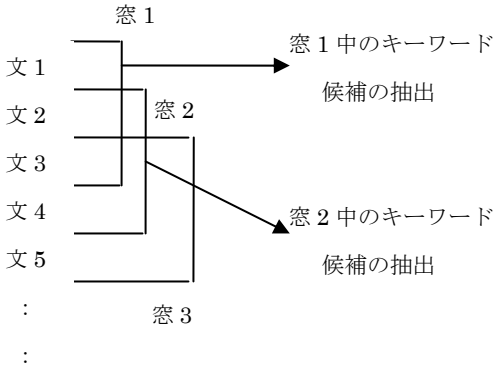

図 1．スロット法

るかを取得し，ある窓で得られたキーワード候補とその 1 回前の窓で得られたキーワード候補を比較する．この処理を先頭から順番に n 文ずつずらしながら文章の最後まで走査する．

この時，話題境界では隣接する窓に含まれるキーワード候補が変化すると予想し，キーワード候補の重なり具合が一定の値よりも下回ったところで話題が転換したと考える．今回はキーワード候補の重なりの割合が一定の割合を下回った時点で話題が転換したと判断した．

なお，ウェブページの最初の文はその長さにかかわらず単独で 1 番目の話題とした上で話題境界を判定して，その後 HTML タグなどを解析して話題境界の補正をおこなった．この処理については，最初の文は見出し文である場合が多く，本文部分と同列に扱うのは不適当だと判断したためである．

窓の幅，および窓のずらし幅を変えて評価実験をおこなった．実験の結果，スロット法においては窓の幅 3 文，窓のずらし幅 1 文，キーワード候補の重なりの割合の閾値 30%とした場合に最良の結果を得られた．

3.2 話題結束力による判定法[5]

本節の説明は参考文献[5]に準ずる．

ある語が複数の文で繰り返し現れるとき，この語を反復語と呼ぶ．反復語を含む二つの文の組み合わせを反復セット，ある反復セットにおいて，反復語が次に現れるまでに含まれる文の数を区間距離と呼ぶ．また，語の反復する区間距離が大きくなれば同一話題は維持されにくくなるので，区間距離に上限を設けることとする．この上限を文間距離限界と呼ぶ．反復語が繰り返し現れることによってそれらを含む文の間には同一の話題が存在し，そこには話題形成ポテンシャルが生じると考える．一対の反復セットで反復語がいくつ現れてもその反復セットでの話題形成ポテンシャルは一定の値を持つものとする．この時，全ての反復語で生じる話題形成ポテンシャルを合計したものが隣接する文間の結束力を示すものとして，これを話題結束力と呼ぶ．話題結束力の値が大きければそこには話題が存在し，値が下がったところで話題境界が生じる．

このようにして得られたすべての反復語の話題結束力の総和が対象文書の各文間の話題結束力となる．この話題結束力が大きく落ち込んだ箇所を話題境界と判定する．

また，文間限界距離，および話題境界判定のための閾値は参考文献[2]に従った．

4. 実験と考察

実験用のサンプルファイルとして，ウェブに存在する旅行記を用いた．それらのウェブデータはすべて同じ筆者に書かれたもので，各データで筆者の訪れた土地のポートレートとそれらに対する筆者のコメントが記されている．今回用いたデータの個数は全 13 ファイル，それらは 3 から 5 キロバイト程度の大きさで，各ファイルにつき 4 から 7 個の画像データを含んでいる．

すべての実験用サンプルファイルについて話題境界の正解データを作成し，それに対して各手法の再現率(recall), 適合率(precision)[3][4]を計算して比較した．再現率，適合率はそれぞれの次の式で求められる．

$$\text{再現率} = \frac{\text{システムが判定した話題境界の中で正解データと一致した数}}{\text{正解データの話題境界の数}}$$

$$\text{適合率} = \frac{\text{システムが判定した話題境界の中で正解データと一致した数}}{\text{システムが判定したの話題境界の数}}$$

また，再現率と適合率の両方を考慮して評価をおこなうために F 値[3][4]を計算した. F 値は次の式で求められる．

$$F\text{値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

F 値は再現率と適合率の両方を考慮した尺度で高い値を取るほどシステムとしての精度が高いと判断できる．

実験結果として表 1 に各手法の再現率，適合率を示す．

また，各サンプルのＦ値から各手法の平均値を求めた．スロット法では 0.55，話題結束力による

表 1　スロット法と話題結束力判定法の再現率と適合率

サンプル No.	スロット法		話題結束力判定法	
	再現率	適合率	再現率	適合率
1	0.40	1.00	0.60	0.75
2	0.40	0.67	0.80	1.00
3	0.67	1.00	0.60	0.60
4	0.67	0.80	0.33	0.50
5	0.50	1.00	0.50	0.67
6	0.40	1.00	0.40	0.40
7	0.33	0.33	0.67	1.00
8	0.33	0.50	1.00	1.00
9	0.60	1.00	0.60	0.50
10	0.17	0.33	0.50	0.75
11	0.40	1.00	1.00	1.00
12	0.40	1.00	0.60	0.50
13	0.33	0.67	0.33	1.00

手法では 0.66 となり，後者の方が有効であった．

次に，画像データに同一話題中のキーワード候補を関連付ける手法において，話題結束力による話題境界の判定手法を組み込んだ場合の有効性について実験をおこなう．話題境界を判定した場合に関連付けられるキーワード候補と判定しない場合のキーワード候補とを比較する．話題境界を判定する場合には，画像データを含む話題に含まれるキーワード候補すべて(話題キーワードと呼ぶ)を関連付け，判定しない場合には，画像データの前の段落，後の段落，前後の段落に含まれるキーワード候補すべて(それぞれ，前段落キーワード，後段落キーワード，前後段落キーワードと呼ぶ)の 3 通りのキーワード候補を文章の内容と画像の関係を考慮せずに関連付けるものとした．

関連付けたキーワードの比較のために，ある画像を含むサンプルファイル中のキーワード候補すべてについて，それらがその画像に対するキーワードとして適切かどうかをあらかじめ 4 段階(very good，good，bad，so bad)で評価しておき，

表 2　話題キーワードと各段落キーワードの点数の合計値

	話題キーワードの評価の方が良かった画像数	話題キーワードの評価の方が悪かった画像数
前段落キーワード	55	7
後段落キーワード	61	1
前後段落キーワード	52	10

次にその評価に合わせてキーワード候補に点数を与えた．各評価に対する点数はそれぞれ，very good を 8，good を 5，bad を 3，so bad を 0 とした．

この点数を用いて，前段落キーワード，後段落キーワード，前後段落キーワードを画像に関連付けた場合の点数の合計値を，話題キーワードを画像に関連付けた場合の点数の合計値と比較した．比較実験の結果を表 2 に示す．なお，使用した画像データは全部で 62 画像である．

実験の結果，話題キーワードの点数の合計値の方が段落キーワードの点数の合計値よりも高くなったのは 90.3(%)となった．話題境界を考慮して画像にキーワードを付与する手法に話題結束力を用いることには今回の実験では一定の評価を得られた．

段落キーワードの点数の合計値の方が話題キーワードの点数の合計値よりも高かった画像は 18 個あったが，その内 12 個に関しては評価 so bad のキーワードが話題境界を用いた場合よりも多く付けられていた．これは，キーワードを取得する範囲が大きく取られるなどの要因により，望ましいキーワードが多く付けられた結果点数としては良くなったが，ノイズとなるようなキーワードも同時に多く付けられたことを意味しており，本当に適切なキーワードを付与されたと言えない．

残りの画像では，本来使用しないはずの見出し行を含めてキーワード抽出および関連付けをおこ

なっていた．また，話題境界の判定ミスにより話題キーワードの適切な関連付けがおこなえなかったために，話題境界を用いた場合の方が合計値が低くなったものもある．これは話題境界判定手法の改良などで対処可能と考える．

5. まとめと今後の課題

本稿では，ウェブ上の画像データに付与するキーワードの自動抽出のための話題境界の推定手法であるスロット法と話題結束力による手法について，実験によりそれぞれの有効性を比較した．また話題結束力による手法を組み込んだ場合のキーワード抽出についても実験をおこなった例では有効性を確認した．今後の課題としては話題境界の推定精度の向上や構文解析などを用いた画像とキーワード間の関連性推定の改良などが考えられる．

謝辞

大阪府立大学人間社会学部人間科学科山口義久教授にはウェブページをテストデータとして使わせていただくことをご快諾いただいた．ここに謝意を表する．

参考文献

[1] 岡田 真, 浜田 浩史, 宝珍 輝尚: マルチメディアデータの効率的検索のためのキーワード自動抽出手法，情報処理学会研究報告，pp.73-78, 2005.

[2] M.L. Kherfi, D. Ziou, A. Bernardi: Image Retrieval from the World Wide Web: Issues, Techniques, and Systems, ACM Computing Surveys, Vol. 36, No. 1, pp. 35-67, 2004.

[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] 徳永 健信: 情報検索と言語処理, 東京大学出版会, 1999.

[5] 中野 滋徳, 足立 顕, 牧野 武則: 語の反復距離に基づく段落境界の認定，自然言語処理，Vol. 13, No.2, pp. 3-26, 2006.