社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－ＦＩ－ 80 (11)
2005－ＮＬ－169 (11)
2005／9／30

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

岡田 真, 浜田浩史, 宝珍輝尚

大阪府立大学 大学院 理学系研究科 情報数理科学専攻

本研究ではマルチメディアデータの効率的検索のために, ウェブ上のデータからマルチメディアデータをよく説明するキーワードを自動的に抽出する手法について提案する. 提案手法の手順は, まずウェブページを形態素解析し, 名詞と形容詞をキーワード候補として抽出する. 次にページ内の話題転換部を判定し, 同話題内の画像とキーワード候補を関連づける. そして2種類のIDF値を用いて不要語の削除と, キーワードの選出をおこなう. 提案手法の有効性を実験により評価したところ, 「良い」「まあまあ良い」と判定されたキーワードは, 名詞については, 文書キーワードで約 92(%), 話題キーワードで約60(%)であり, 形容詞については, 文書キーワードで約 59(%), 話題キーワードで約68(%)だった. 話題については「適切」が 79(%), 「不適切」が 21(%)となった.

A Method of Automatic Keyword Extraction for an Efficient Retrieval of Multimedia Data

Makoto Okada, Hiroshi Hamada, Teruhisa Hochin
Department of Mathematics and Information Science, Graduate School of Science,
Osaka Prefecture University

In this paper, we proposed a method of automatic keyword extraction for a retrieval of multimedia data. In order to put a set of keywords to image data, firstly, all of nouns and adjectives are extracted from web pages by using morphological analysis. Next, change points of topics are estimated in the pages, and these words are related to the image data having the same topic. Finally, keywords and "stop words" are chosen from the words by using 2 types of Inverse Document Frequency (IDF). We confirmed effectiveness of the proposed method through experiment.

1. 始めに

近年, インターネット上に画像・動画・音といったマルチメディアデータが遍在するようになってきている. これらの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, 様々な研究がおこなわれている.

筆者らは, マルチメディアデータを効率良く利用することを目的として, マルチメディアデータを印象や感性に基づいて相互に直接的に検索する機構を開発してきた[1][2][3]. しかし, ウェブ上のマルチメディアデータを検索対象にしようとすると, マルチメディアデータの特徴量のみからの検索では雑音が多くなりすぎて使用に耐えないという問題がある. これは, 例えば, ある音楽に合った画像を求めるという検索でも, 風景がほしい場合や人物写真がほしい場合があり, 単にその音楽に合った画像を結果とすれば良いわけではないからである. また, ウェブ上のメディアデータにはそれらへの説明や印象が付加され

ている場合がある. これらの記述からキーワードを抽出し積極的に使用することで検索精度の向上が図れる可能性もある. そこで用いられるキーワードとしては, 従来よく用いられている名詞に形容詞を加えて用いることが有効であると考えられる. なぜなら, 形容詞は人間の感性を表現することが多い品詞であり, そのメディアデータに対しての印象の情報を持っている可能性が高いからである.

そこで本稿では, 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(名詞と形容詞)を自動的に判別および抽出する手法について提案する. まずキーワード抽出のために必要となる手順について説明し, それらの手法を用いて実際にキーワードを抽出した実験について述べ, その実験結果に考察を加える. 最後にまとめと今後の課題について述べる.

2. キーワード抽出手法

今回はウェブ上のマルチメディアデータとして画像データを対象とし, それ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適切なキーワードを抽出して付加するために, 以下のような手順を取る.

ウェブ上のデータに対して, その中に含まれる画像データに関する情報を取得し, その後に内部のテキストデータの形態素解析をおこなって, 頻度を元にキーワード候補を得る. 次に各データ内の話題の転換部を得る. これは, 画像データとキーワード候補群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係があるものとして結びつけるためである. さらに, 抽出したキーワード候補について, 各データ相互での IDF(Inverse Document Frequency)と各データ内部の話題間での IDF という2種類の IDF 値を得て, 不要語・削除語の選別とキーワードの重要性の判定をおこなう.

以下, 各手法について説明する.

2.1 キーワード候補の抽出

ウェブ上のテキストデータから HTML タグなどの不必要な情報を取り除いて, 残った文書の内容に関わる記述部分を形態素解析に掛ける. 形態素解析エンジンとしては茶筌(ChaSen)[1]を用いる. そうして得られた解析結果である形態素の中から名詞と形容詞をキーワード候補として選び出し, ファイル毎にそれらの頻度を集計する. この二つの品詞に絞った理由は, 名詞は一般的にキーワードとして抽出されるためであり, 形容詞は人間の感性を表すために使われることが多く, 本研究の前提条件である画像データに対する感性検索をおこなう場合に重要となると考えたからである.

そして, その頻度の集計結果を頻度の降順でソートして, 出現頻度の多いものをキーワード候補として選び出す. ただし, 今回は形容詞は全て候補として選出した. その理由は以下の通りである. まず, 形容詞は感性を表す表現と考えられ, 本研究の目的として名詞よりも重要な意味を持つこと. そして形容詞自体の出現頻度が名詞に比べるとはるかに低く, 単純な頻度で比較すると候補からほとんどの場合に除外されてしまうのでそれを避ける必要があることである.

また, 今回はHTMLタグの持つ性質を使ったキーワード抽出もおこなっている[4]. 具体的には <TITLE>タグや ALT 属性に記述されている文章もキーワード候補として抽出する. <TITLE>タグは文章自体のタイトルを記述するのに用いられるので, そこから得られた候補はファイル全体のキーワードとして扱う. ALT 属性は画像を表示する<IMG>タグなどで画像の内容を説明するのに使用される属性であり, その記述は画像データに関係が深いキーワード候補として扱う.

2.2 話題の転換部の判定

前述の方法で抽出されたキーワード候補を適切な画像データに付与するために, それらの候補がど

[1] http://chasen.naist.jp/hiki/ChaSen/

の画像データを説明するものなのかを判別する必要がある．そのために対象ファイルのテキストデータがどの部分で話題が転換しているかを調べて，同じ話題内から得られる画像データとキーワード候補群とを関連があるものとして結びつけることにする．

まず，テキストデータをパラグラフ単位に分割し，それぞれのパラグラフでキーワード候補の抽出をおこなう．次に，話題の転換部を探るために，パラグラフを一定の個数毎にまとめる．本研究では対象としたデータの性質よりから 3 パラグラフをひとまとめとした．ひとまとめ毎にどのようなキーワード候補が含まれるかを取得し，先頭から順番に一パラグラフずつずらしながら最後まで走査する．この時，話題の転換部では抽出されるキーワード候補が変化すると予想し，ある走査で得られたキーワード候補とその一回前の走査でのキーワード候補を比較し，そ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，そこに話題の転換部としての情報を埋め込むことにする．今回はキーワード候補の重なりの割合が 30(%)を下回った時点で話題が転換したと判断した．

2.3 IDF による不要語とキーワードとの判定

先に，キーワード候補の判定基準としてファイル内での出現頻度を用いると説明したが，それに加えて IDF (Inverse Document Frequency)[5]も用いることにする．これを用いることにより，IDF 値が低ければそのキーワード候補がさまざまなデータに表れる一般的な単語として判断できるので不要語の抽出などが可能となる．また値が高ければ，その候補が特定のファイルや話題にのみ現れるものと判定して，そのファイルの内容をよく表すキーワードや，ファイル内のある話題に関連の強いキーワードとして抽出することができるようになる．

本研究では，2 種類の IDF 値を用いる．それは 1) 全体 IDF (*IDFg*), 2) 話題 IDF (*IDFt*)である．前者は一般的な IDF 値であり，次式で求められる．

$$IDFg = \log\left(\frac{nf}{nf_i}\right)$$

ここで，nf は対象となる総ファイル数，nf_i はあるキーワード候補を含む文書数である．また，後者は次式で求められる．

$$IDFt = \log\left(\frac{nt}{nt_i}\right)$$

ここで，nt はあるファイル内での話題の総数，nt_i はあるキーワード候補を含む話題数である．

まず，各ファイルの内容をよく表すキーワード(ファイルキーワード)と不要語を全体 IDF 値を利用して得る．すべてのファイルからキーワード候補を抽出し，それらの全体 IDF を求める．その値と各候補の頻度 (*TF*)を乗じた数値 *TF * IDFg* を各ファイルにおけるキーワード候補の重みとする．ここで，すべてのファイルに出現する単語はその *IDFg* の性質から重みが 0 になるので，キーワードとしては不適切と判断してファイル全体の不要語リストに加える．また，各ファイルにおいて *TF * IDFg* 値の上位のものをファイルキーワードとする．今回は各ファイルにつき上位 5 個とし，同位のキーワード候補は切り捨てた．

次に，話題 IDF を使って各話題に固有のキーワードを求める．まず，各ファイル内のキーワード候補について各話題で話題 IDF を求め，先ほどと同様に *TF *IDFt* 値を得る．この値が高いと，そのキーワード候補は特定の話題で頻出し，さらにその話題にしか出現しない話題固有のキーワードと判定できる．このようにして各話題に対し *TF *IDFt* 値の高い候補を上位から話題キーワードとして選出する．今回はこちらも上位 5 個としたが，同位のキーワード候補はすべて選出した．ファイルキーワードと話題キーワードの例を図 1 に示す．(a)はキーワード付加対象の画像，(b)は全体キーワード，(c)は話題キーワードであり，(b)と(c)の数字は上位から何番目かを表す．

(a) 対象画像

名詞 1. フィレンツェ, 2. ダビデ, 3. ヴェッキオ, 4. ミケランジェロ,
形容詞 1. 高い, 2. 古い, 3. よい

(b) 全体キーワード

名詞 1. 聖堂, 街, 複製, フィレンツェ, 2. 写真, ブロンズ, 中腹, 好き, 本物, 下, 場所, 一つ, 煙, 宮殿, 政, 庁舎, 鐘楼, 印象, 全体, ジョット, ドーム, 中心, ダビデ,
形容詞 1. 高い

(c) 話題キーワード

図 1 キーワードの例

Fig.1. An example of extracted keywords.

3. 実験

3.1 実験方法

本稿で提案した手法の有効性を, 実際にウェブ上に存在するデータをテストデータとして用いて確認する. 今回はテストデータとしてウェブ上に存在する各地の旅行記を用いることにした. それらのウェブデータで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有名な風景や美術品などのポートレートとそれらに対する筆者のコメントが記述されている. 例を図 2 に示す. 今回用いたデータの個数は全 13 ファイル, 3 から 5(KB)程度の大きさで, 各ファイルにつき 4 から 7 個のポートレート画像データを含んでいる.

実験の内容としては, まず前述のテストデータそれぞれに対し, 2 章で述べたキーワード候補抽出, キーワードと話題の転換部の判定, 2 種類の IDF 値を用いたキーワードの重み付けと不要語削除などの各処理をおこなう. その結果, 各ファイル毎にファイルに含まれる画像データ, 話題の転換部, そのファイルに含まれる全体キーワードと話題キーワードが得られる. 抽出された画像データにはどの話題の範囲に含まれるかにより 5 個以上の話題キーワードが付加される.

そのようにして得られた結果を男子大学生 1 名が見て, 結果として適切かどうかを判定した. 判定対象は抽出した 2 種類のキーワード(文書キーワード, 話題キーワード)と話題の転換部であり, それらが適切かどうかを以下に述べる基準に沿って判定をおこなった. 判定基準は, キーワードに関しては 4 段階(良い, まあまあ良い, あまり良くない, 悪い)を, 話題に関しては 2 段階(適切, 不適切)を設けた.

3.2 実験結果

キーワードの判定結果を表 1 に示す. 「良い」「まあまあ良い」と判定されたキーワードは, 名詞につい

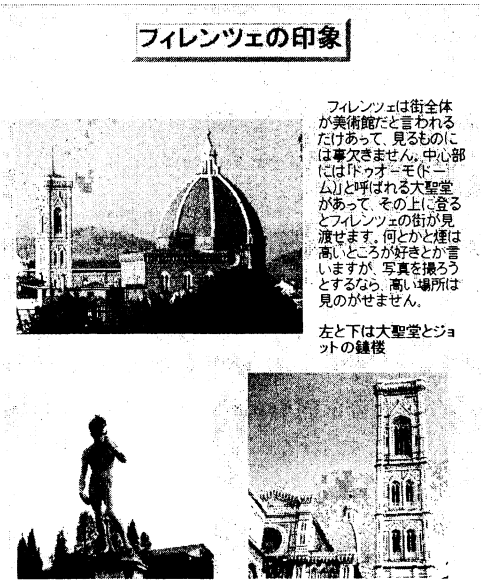

図 2 ウェブページの例

Fig.2. An example of a web page.

ては，文書キーワードで約 91.5(%)，話題キーワードで約60.1(%)であり，形容詞については，文書キーワードで約 59.2(%)，話題キーワードで約 68.3(%)だった．

話題については話題の総数42個に対して「適切」が 33 個(78.57(%))，「不適切」が 9 個(21.43(%))となった．考察については次章で述べる．

4. 考察

以下，実験結果について考察し，現時点で判明している問題点とその解決案について述べる．

まず，キーワードの判定結果について述べる．文書キーワードについては，名詞に関しては「良い」と「まあまあ良い」が約 90(%)となり，満足のいく結果であると言える．その反面，形容詞の評価は「良い」と「まあまあ良い」で約 60(%)であり，あまり満足がいくものではない．このように形容詞のキーワードの評価が低くなった原因としては，画像との適切な対応付けが困難であることが挙げられる．形容詞は元々文書中の出現数が少ないため，本手法のように出現頻度を元にしてキーワード候補を決定する手法ではキーワードとしての判定や画像データへの関連づけが難しい傾向があると考えられる．精度の向上のためには，文脈解析などにより，抽出された形容詞のキーワード候補がどの画像をどのように評価をしているのかをより正確に判定する必要がある．

話題キーワードに関しては，名詞，形容詞の両方で 60(%)台にとどまった．その原因として，同位のキーワード候補が多くなりノイズが増えたためであると考えられる．各話題の文章は多くて数パラグラフと短いため，キーワード候補として同じ単語が繰り返し出現することは少ない．そのため話題キーワード候補の出現頻度は 1 から 2 回程度の場合が多くなり，その結果同位の候補数が増加する．今回は話題キーワードの抽出方針として最低 5 個，ただし同位の候補があればすべて抽出するとしたため，話題に関わりのない候補も多く選出されやすくなり，その結果それらがノイズとして精度を引き下げたと考えられる．今回，我々は話題固有のキーワードを得るために話題IDFを用いたが，精度向上のためにそれに加えて新たなキーワード評価指標を用いる必要があるといえる．

次に，話題の転換部についての評価は「適切」が約 78(%)でありほぼ満足だといえるが，まだ改善しなくてはいけない点が多い．現在すでに分かっている問題点を以下に述べる．今回は転換部の判定時にテキストデータを 3 パラグラフ単位で扱い，先頭から 1 パラグラフずつずらしながら走査していき，各パラグラフから抽出されたキーワード候補の重なりが一つ前のデータと比べて 30(%)を下回った場合，現在見ているパラグラフ群の先頭で話題が転換したと考えていた．だが，今回のテストデータでは画像のすぐ下に表題を入れてある場合が多く，今回の手法では表題の直前に話題の転換部を挿入せざるを得ず，そのため評価が低くなる場合が見受けられた．このことから，話題の転換部の挿入位置の決定方法を改善する必要があると考えられる．例えば，現在見ている 3 つのパラグラフの第一パラグラフからキーワー

表 1 実験結果

Table 1 Experimental results

文書集合全体の平均	良い	まあまあ良い	あまり良くない	悪い
文書キーワード(平均)				
名詞	47.44%	44.10%	6.54%	1.92%
形容詞	15.48%	43.79%	32.14%	8.59%
話題キーワード(平均)				
名詞	18.97%	41.13%	31.03%	8.87%
形容詞	14.29%	53.97%	26.98%	4.76%

ド候補を抽出し，それらが前回のパラグラフ群からのキーワード候補と現在のパラグラフ群とのキーワード候補とのどちらにより重なりが多いかにより話題の転換部の挿入位置を第一パラグラフの前にするか後にするかを決定するという手法などが考えられる．

また，HTMLの解釈の問題もある．今回のテストデータでもよく見られた特徴として，画像データを表示する際に，整形のためにテーブルタグなどを使って表示する場合がよくある．この場合，表中に画像とその説明文が混在することが多くなるため，表中の画像と説明文の対応を取らなければならなくなる．そのため，HTMLタグの解釈をおこなう解析エンジンや，文書中の，特に方向を指示する語句(左，右，上，下など)を含む文に関して，それらがどこを示しているのかを解釈する解析手法を組み込む必要があると考えられる．

その他，キーワード候補抽出時の未知語の問題などもある．今回，未知語は適宜辞書に追加したが，扱うデータが大きくなる場合も考えて，自動的に処理する手法の使用についても考慮の必要があるだろう．

5. 終わりに

本稿では，マルチメディアデータである画像データの効率的検索のために，ウェブ上のテキストデータからキーワード候補を抽出し，話題の転換部や2種類のIDF値を利用して画像に付与するキーワードを判定する手法について提案し，実験によりその有効性を確認した．今後は4章で考察した点について改良をおこなっていきたいと考えている．

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頂くことをご快諾頂きました．ここに謝意を表します．なお，本研究は，一部，文部科学省科学研究費補助金(課題番号：16500067)による．

＜参考文献＞

1) 宝珍 輝尚，都司 達夫：感性に基づくマルチメディアデータの相互アクセス法，情報処理学会論文誌，43，SIG 2(TOD 13)，69-79，2002.

2) 宝珍 輝尚，都司 達夫：画像の特徴量からの感性の主因子の因子得点の推定，第63回情報処理学会全国大会講演論文集，3-229 - 3-230，2001.

3) 宝珍 輝尚，熊切 健夫，井田 俊博，都司 達夫，樋口 健：感性マルチメディア検索における擬逆行列を用いた一個人適応法，感性工学研究論文集，4，1，27-30，2004.

4) 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，March，pp. 35-67，2004.

5) 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.