文書分類を利用した画像検索結果のフィルタリング

正木裕一
茨城大学大学院
理工学研究科

新納浩幸
茨城大学工学部
システム工学科

佐々木稔
茨城大学工学部
情報工学科

1　はじめに

現在、一般的な通信回線の高速化もあり、ネットワークを介して、画像などの容量の大きなマルチメディア情報を自由にやりとりすることができるようになった。それにともない、Web 上に無数に存在する画像データから自分の所望するデータを検索する画像検索技術が要求されている。

現在、Google で提供されている画像検索は、画像周辺にあるテキストを利用している。もし画像の周辺のテキストに検索クエリが含まれていれば、その画像を検索クエリにマッチした画像としている。ランキングの方法は不明である。このような技術は、クロスメディア検索と呼ばれ、様々な検索エンジンで行われている。しかし、その検索結果には、画像自体を解析しているわけではないので、要求にあわない画像が検索される場合も多い。

相良らによって HTML テキストの重要文を用いた画像ラベリング手法 [1] や、画像検索技術として、画像の類似性から画像を検索する類似画像検索も研究されているが、現在広く普及しているパーソナルコンピュータの性能、通信回線速度を考えると実現するのは今のところ難しいと思われる。そこで、本研究では検索エンジンによる検索結果をフィルタリングして、要求にあわない画像を取り除いたり、検索結果のランキングを変更したりする研究を行った。具体的には、検索された画像を含む HTML 文書のテキスト部分が入力されたキーワードと関連性が高いかどうかに注目する。今回は特に人名による画像検索に特化して実験を行った。

2　画像検索手法

現在利用されている画像検索は図 1 のように検索ウインドウにキーワードを入力すると、キーワードに一致する画像データがサムネイル形式で一覧表示され、希望の画像をクリックすることにより、画像を含むページ全体が表示される。一覧表示された各サムネイル画像の下には、その画像ファイルの名称、画像を含むサイトの URL、画像サイズなどの付加情報も表示される。

これらの検索された画像は各検索サイトのインデックスに含まれる画像の中から次のようなアルゴリズムに沿って検索されている。各検索サイトに共通した主なアルゴリズムは、画像のキャプション、ページ内の画像の周辺テキスト、画像の alt 属性など画像の周辺情報内に検索ウィンドウに入力されたキーワードを含む場合にキーワードに関連する画像とし、検索された画像を各検索サイトの独自のアルゴリズムにより順位付けし表示している。

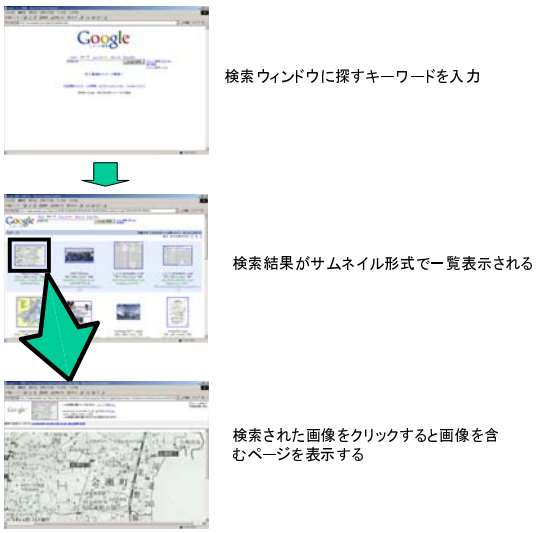

図 1: 画像検索の例 (Google)

3　フィルタリング手法

3.1　Naive Bayes 法

ここでは、検索結果のフィルタリングに Naive Bays 法 [2] を用いた。Naive Bayes 法は全ての未知量を確率変数として扱い、未知パラメータも確率分布として推定を行う確率論に基礎をおく推定方法を用いた分類法であり、スパムメールの自動フィルタリングなどの

文書分類において広く利用されている。その特徴としては、最終的な仮定成立の確率計算に事前知識を利用し、仮定が成立するかどうかを真か偽ではなく、確率で明示的に表現できる点にある。

ある事例 x を素性のリストとして、 $x = (f_1, f_2, \cdots, f_n,)$ とし、 x の分類先のクラスの集合を $C = \{c_1, c_2, \cdots, c_m,\}$ としたとき、分類問題は $P(c \mid x)$ の分布を推定することで解決できる。実際に、 x のクラス c_x は以下の式で求まる。

$$c_x = \arg\max_{c \in C} P(c \mid x) \tag{1}$$

また、ベイズの定理より、

$$P(c \mid x) = \frac{P(c)P(x \mid c)}{P(x)} \tag{2}$$

となる。よって、式 (1)、式 (2) より、以下の式が成立する。

$$c_x = \arg\max_{c \in C} P(c)P(x \mid c) \tag{3}$$

しかし、$P(c)$ は比較的容易に推定できるが、$P(x \mid c)$ の推定は現実的に難しいため、Naive Bayes のモデルでは以下の仮定を導入する。

$$P(x \mid c) = \prod_{i=1}^{n} P(f_i \mid c) \tag{4}$$

3.2 実験データ

今回使用する実験データには、Google によって「イチロー」、「小泉純一郎」、「渡辺謙」について画像検索を行った結果の、それぞれ上位 20 件の画像元ページの HTML 文書をデータとして使用した。これらの人物名を実験データの対象とした理由として、スポーツ、政治、芸能の各分野で昨年大いに活躍し、様々なメディアに多く登場した各分野を代表する人物であるということで実験データの対象人名とした。それぞれのデータは画像検索結果の上位 20 件の画像とその画像を含む HTML 文書である。作成した実験データの例を図 2 に示す。なお、形態素解析には京都大学の JUMAN を使用した。

図 2: 実験データの例

4 実験

実際に実験データを入力する前に、まず Naive Bayes による学習を行う。学習される分類器の入力は HTML 文書であり、識別クラスは「イチロー」「小泉純一郎」「渡辺謙」「その他」の 4 つである。訓練データは「イチロー」「小泉純一郎」「渡辺謙」をキーに Google で検索を行い、その検索結果の HTML 文書をそれらのクラスの訓練データとした。「その他」のクラスの訓練データとしては国際、経済、社会の文書を適当に用意した。これらの訓練データを用いて、Naive Bayes による学習を行い分類器を作成した。

次に Google で「イチロー」をキーに画像検索した結果の上位 20 件の画像と HTML 文書を取り出し、その HTML 文書を先の学習できた分類器により、その文書が「イチロー」「小泉純一郎」「渡辺謙」「その他」のどのクラスに属するかのスコアーを求めた。その合計スコアーの「イチロー」に対するスコアーの比率をその文書の「イチロー」に対する関連度とした。合計スコアーを Sum 、そのテキストのクラスのスコアー $Score$ から、次式により各テキストファイルごとの各クラスに属する比率 $Ratio$ を求める。算出式は次式のようになる。

$$Ratio = \frac{Score}{Sum} \tag{5}$$

この関連度により、 Google の検索結果のランキングを変更した。「小泉純一郎」「渡辺謙」に対しても同様の実験を行った。この実験の流れを図 3 に示す。

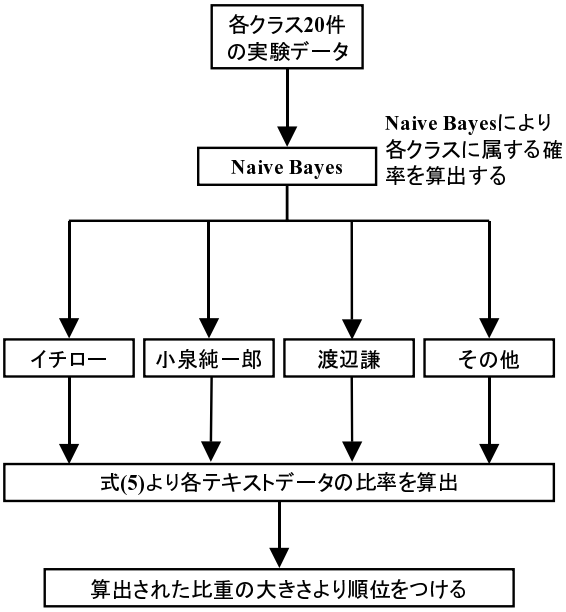

図 3: 実験のフローチャート

画像検索結果について、算出された各クラスの順位と比率を表 1 に示す。また、各クラスのフィルタリングを行った結果の上位、下位各 3 件の比較する。「イチロー」の結果を図 4 に、「小泉純一郎」の結果を図 5 に、「渡辺謙」結果を図 6 に示す。

表 1: 実験結果 (%)

順位	「イチロー」	「小泉純一郎」	「渡辺謙」
1	70.8	82.0	75.2
2	70.8	82.0	71.1
3	69.7	81.6	70.4
4	65.5	76.8	68.7
5	56.6	76.3	66.5
6	55.6	73.6	66.3
7	55.6	69.6	64.3
8	54.3	65.6	61.7
9	53.9	64.6	61.0
10	53.2	63.4	59.4
11	49.9	62.3	57.1
12	41.1	61.9	51.6
13	40.3	61.8	49.8
14	38.3	58.0	48.9
15	22.9	55.0	無
16	13.6	35.0	無
17	無	無	無
18	無	無	無
19	無	無	無
20	無	無	無

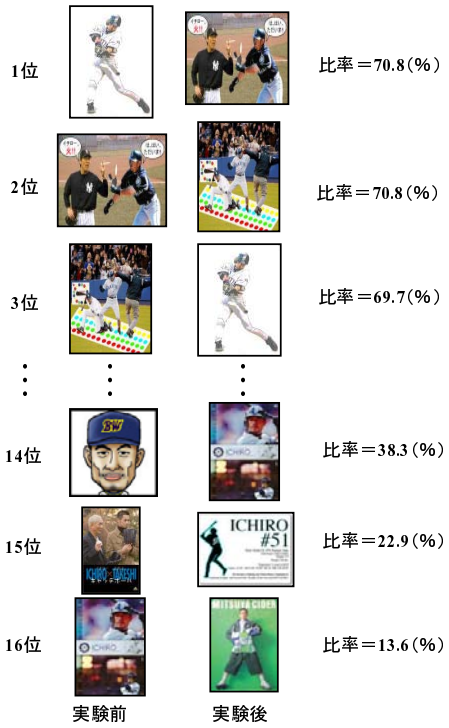

図 4: 「イチロー」の上位, 下位各 3 件の画像の比較

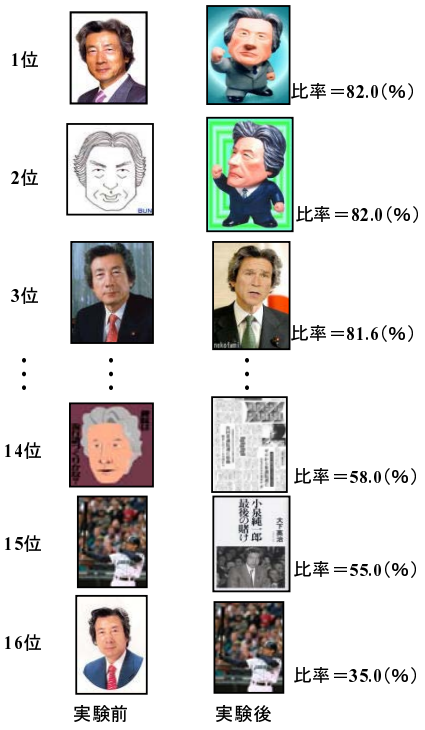

図 5: 「小泉純一郎」の上位、下位各 3 件の画像の比較

図 6: 「渡辺謙」の上位、下位各 3 件の画像の比較

5　考察

今回の Naive Bayes を利用したフィルタリング実験の結果、画像を含むページのクラスの全クラスからみた比率を見ることによって、検索された画像の新しい順位付けを行った。

「小泉純一郎」の Google イメージ検索結果に対して、フィルタリング実験を行った結果では、上位 3 件では 1、2 件目が小泉純一郎の人形の画像になり小泉純一郎に関連性ある画像となった。これらは画像元は同じページであり、これらの画像の比率が、高くなった理由として考えられることは、画像元のページにおいて、文章がキーワードである「小泉純一郎」のみであったためと考えられる。また、下位では、新聞の画像など人物像では無い結果となり、比較的良好な結果を得ることができた。

「イチロー」の結果については、実験前後で大きく順位が変動する画像は少ない結果となった。これには、画像検索結果の件数が関係しているのではないかと考えられる。他の検索結果件数はそれぞれ、「小泉純一郎」が約 1200 件、「渡辺謙」が約 700 件に対し、「イチロー」の結果は約 8400 件と約 8 倍程度の検索結果となった。今回は 8000 件中の上位 20 件のみを実験データとして使用したため、このような結果となったと思われる。

「渡辺謙」については、渡辺謙自身の画像が最上位の結果となったが、2 位、3 位の画像が同姓同名の人物の画像が入ってしまう結果となってしまった。、この原因としては、キーワードである「渡辺謙」が考えられる。「渡辺謙」は芸能という分野の代表として使用したが、他 2 分野と比べこの分野特有と呼べるような単語が少ないため、算出された各テキストデータの比率にばらつきが少なく、今回の様に同姓同名の人物が上位になる結果となったのではないかと考えられる。このような結果から、Naive Bayes 法によって比較的良好なフィルタリングを行うことできるが、検索サイトでの検索結果件数や、検索する人名によってはそれらが原因となり、良い結果が得られない場合があることがわかった。

6　まとめ及び今後の課題

本研究では、キーワードを含む文章を文書分類を用いることにより、誤った検索結果を除外するフィルタ作成に関する研究を行った。今回の研究では、同姓同名の人物の画像など、要求していない画像を排除し、より精度の高い順位付けを行う手法を確立することを目標に研究を行った。実験の結果、比較的良好な結果を得られた例もあったが、一部順位を下げる事のできない画像が未だ存在する、実験前後で類似した順位付けとなってしまう、という例があった。これらのことから、本研究の目的とする所である、より高精度な順位付けの手法を確立することができたとは言い難い。

そのため今後の課題として、今回の Naive Bayes 法を用いたフィルタリングを基礎とし、それに加え新たな判断規準を設ける必要がある。現在、その 1 つとして考えているものは、画像元のページ中の本文、画像の alt 属性、画像のキャプションなど検索キーワードを含む文章がそのページのどの部位にあるかにより、そのページに重み付けを行ない、Naive Bayes 法による結果を併せて総合的に順位付けを行う方法で現在研究中である。また、今回は上位 20 件のみの結果であったが、今後 Google イメージ検索の検索結果全てにおいて実験することも考えている。

参考文献

[1] 相良直樹, 砂山渡, 谷内田正彦： “HTML テキストの重要度を用いた画像ラベリング手法”, 人工知能学会全国大会第 17 回 (2003).

[2] Tom M.Mitchell,Thomas Michell ： “Machine Learning(Mcgraw-Hill Series in Computer Science)”, (1997).