WWW 画像検索システムにおける再検索型検索質問拡張に基づくフィードバック検索

Feedback Retrieval based on re-Retrieval Query Expansion on WWW Image Retrieval Systems

竹安 真紀夫 † **獅々堀 正幹** † **柘植 覚** † **北 研二** ‡

Makio Takeyasu Masami Shishibori Satoru Tsuge Kenji Kita

1. はじめに

現在，WWW 上には膨大な量の画像データが存在する．この膨大な量の画像データの中から一度の検索でユーザが所望する画像を取得することは困難である．そこで従来より，ユーザからのフィードバック情報を利用した検索システムが提案されている．一般に，WWW 画像検索システムにおけるフィードバック検索は，フィードバック情報を用いて検索結果をソーティングするものが主流であり，検索質問自体にフィードバック情報を反映させていなかった．

本稿では，WWW 画像検索において，フィードバック情報を用いて検索質問を拡張し，拡張された検索質問に対してフィードバック検索を行う手法を提案する．本手法は，ユーザが選択した画像にリンクする HTML ページ中の言語情報より，検索質問の関連語を収集する．この関連語を用いて再検索し，ユーザが選択した画像と関連語で検索した画像を比較し，内容的に類似した画像をユーザに出力する．

2. 従来のフィードバックモデル

画像検索システムにおける従来のフィードバックモデルは，ユーザの意図に適合する画像を選択し，選択した画像と検索結果の画像を比較して，類似している画像を上位に入れ換えるものであった．しかし，検索画像数が多くなるにつれ，ノイズとなる画像が増加し検索精度の限界があった．これに対して，予め人手で登録した (検索単語の) 関連語での検索結果をフィードバック情報に則して順位付けし，より多くの正解画像を検索する手法 [1] や，再検索の際に画像を含む HTML 文書内の頻出単語を補助キーワードに用いて多くの画像を収集する手法も提案されている [2]．しかし，用いる関連語がフィードバック情報に必ずしも即した単語であるとは限らず，また，関連語特定の自動化も必要であった．そこで，本稿では，フィードバック情報から自動的に関連語を収集する手法を用いる．

文書検索の分野では，ユーザのフィードバック情報により検索質問を拡張する代表的な手法として Rocchio の式 [3] が知られている．これは適合文書に含まれる単語の重みを大きくし，不適合文書に含まれる単語の重みを小さくするように検索質問の修正を行う手法である．WWW 画像検索システムにおいても，ユーザが選択した画像にリンクするページを適合文書，不選択の画像にリンクするページを不適合文書として検索質問を拡張することも考えられる．しかし，Google 等の WWW 画像検索システムは，画像近傍の言語情報から検索結果を表示しているため，画像内容とページの内容が異なる場合も見受けられる．このため，不選択画像であるにも関わらず，ページの内容が選択画像ページの内容と類似していたために，関連語となるべき単語を見落としてしまう可能性が高い．そこで，本稿では選択画像にリンクするページ内の言語情報のみを用いて検索単語を拡張し，既存の WWW 画像検索システムにおいてフィードバック検索を行う手法を提案する．

3. 検索質問拡張に基づくフィードバック画像検索

3.1 本手法の概要

図 1 に本稿で提案するフィードバック検索手法を示し，手順を説明する．なお，手順 2 で示す関連語候補の重み付け，および手順 3 で示す検索質問の拡張方法については 3.2 で詳しく述べる．

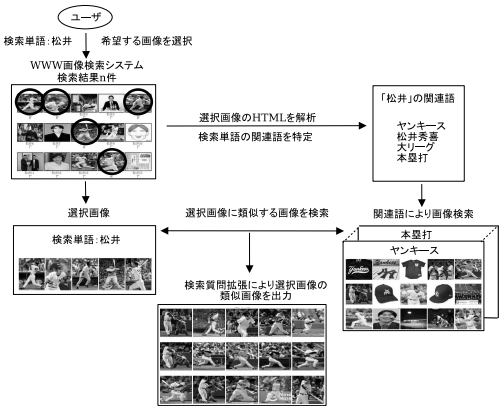

図 1: 本提案手法の概要

手順 1: 画像の選択

検索単語を WWW 画像検索システムに入力し，上位 n 件からユーザの希望する画像 $Image_i (1 \leq i \leq n)$ を選択する．

†徳島大学工学部

‡徳島大学高度情報化基盤センター

手順 2: ページ内容の解析

$Image_i$ にリンクする HTML ページの出現単語 $w_j(j \geq 1)$ と重み $Weight(w_j)$ を集計する．この w_j を関連語候補とする．

手順 3: 検索単語の拡張

関連語候補 w_j から多義性のある一般的な単語を除去し関連語を特定する．

手順 4: 類似画像の特定

手順 3 で特定した関連語を WWW 画像検索システムに入力し，上位 m 件の検索結果 $Image_k(1 \leq k \leq m)$ と選択画像 $Image_i$ との類似度を計算する．

図 1 では，手順 1 において，検索単語に「松井」を入力し,「松井秀喜が野球している画像」を選択している．次に，手順 2 で選択画像のページ内に含まれる単語を取得し，手順 3 において,「ヤンキース」等の関連語を特定する．最後に，これらの関連語を WWW 画像検索システムに入力した検索結果と選択画像の類似画像検索を行い，結果を出力する．

3.2 検索質問の関連語収集

検索単語の関連語を収集し，関連語を元の検索単語に加えることで検索質問の拡張を行う．2. で述べたように画像近傍の言語情報により画像検索を行う WWW 画像検索システムにおいては，画像とリンク先のページ内容が必ず一致するとは限らない．したがって，画像のみを見てリンク先のページが適合文書か不適合文書であるかの判断することは困難である．例えば，図 1 では「松井」に関する画像を検索し,「松井秀喜が野球をしている画像」を選択している．他の野球選手の画像も検索結果に含まれているが，これは不選択画像である．しかし，不選択画像となっている野球選手のページ中には，「松井」の関連語が多く含まれているため，このページを不適合文書とすると適合文書中に含まれる関連語の重要度を低下させてしまう．そこで，選択画像の言語情報のみを用いて以下の手順によって関連語を収集する．また，図 2 に関連語収集手法の概要を示す．

手順 1: 関連語候補の重み付け

関連語候補 w_j の周辺の HTML タグを利用して単語 w_j の重み付けを行う [4].

手順 2: 関連語候補が存在するページを再検索

上位の w_j を WWW 画像検索システムに入力し，単語毎に上位 n 件の再検索結果ページを得る．

手順 3: 関連語候補の関連度を計算

再検索結果の各ページ内に存在する関連語候補の

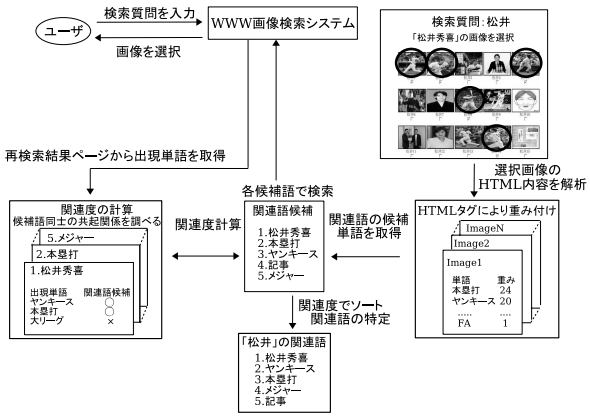

図 2: 関連語収集方法の概要

頻度を求め，式 (1) により関連度を計算する．

$$\text{関連度} = \text{再検索結果内に存在する他の関連語数} \times \text{再検索結果内に他の関連語が存在するページ数} \quad (1)$$

式 (1) の関連度計算において，局所的に頻出する単語の影響を抑え，より多くの再検索結果ページに出現する単語を重視するためにページ数をかけている．

図 3 に上記の手順に従い，関連語の特定を行った例を示す．いま，5 つの関連語候補があるとする．まず，関連語候補「松井秀喜」を WWW 画像検索システムに入力し，上位 n 件の再検索結果ページに出現する単語を取得する．このとき「ヤンキース」,「本塁打」,「大リーグ」の単語が得られたとする．次に，この単語中に他の 4 つの関連語候補が含まれているかを調べると「ヤンキース」と「本塁打」が含まれていることがわかる．最後に，n 件の再検索結果ページに出現する関連語候補の総数と出現ページ数から式 (1) により関連度を求める.「本塁打」等の関連語候補は n 件の再検索結果ページ中に 20 回出現し，18 件のページに含まれていたとすると「松井秀喜」の検索単語との関連度は 360 となる．また,「記事」を検索質問として検索されるページには，他の関連語候補がほとんど含まれていないため，関連度は低くなっている．このように本手法を適用すると「記事」のような一般的な単語を関連語候補から除去することができる．

3.3 HSI 色情報による類似画像検索

3.2 の方法に従い収集した関連語により検索質問を拡張し，再検索することでより多くの画像を収集することが可能になる．しかし，検索質問拡張だけでは画像内容に基づいた検索を行っていないため，ユーザの希望する画像とは異なる場合がある．そこで，本稿では初期の

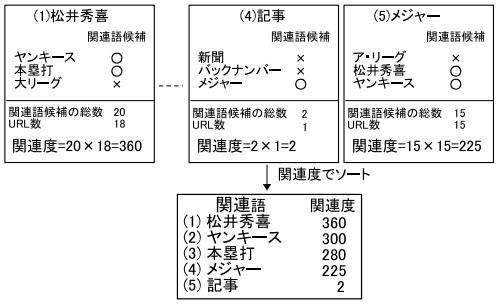

図 3: 関連語特定手順の例

検索結果から選択した画像と検索質問拡張により再検索した画像 (収集画像) の類似画像検索を行うことにより，多くの収集画像から選択画像に類似する画像を上位に出力する．図 4 に類似画像検索の概要を示す.

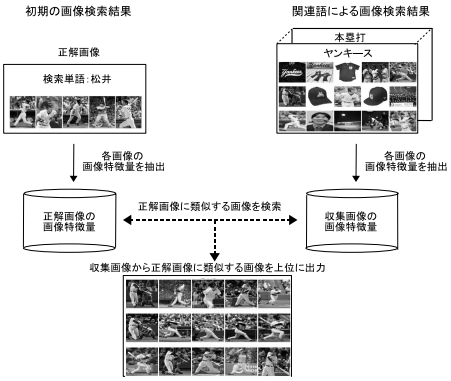

図 4: 類似画像検索の概要

今回，類似画像検索には画像特徴量として HSI 色情報 432 次元の特徴ベクトルを用いた [1]．具体的には，画像全体を 3×3 分割し，各分割領域の H(色相)，S(彩度)，I(明度) の各情報を 16 次元の特徴ベクトルで表す．すなわち，各分割領域につき 48 次元の HSI 色情報を抽出していることになる．このようにして，画像全体で $3 \times 3 \times 48 = 432$ 次元の特徴ベクトルを作成する．なお，類似度の尺度には選択画像と収集画像との特徴ベクトル間のユークリッド距離を用いる．

4. 評価実験

4.1 関連語収集の評価

4.1.1 評価条件

本稿で提案した関連語収集手法の有効性を確かめるために検索質問の関連語を収集して評価を行った．表 1 に評価に用いた検索質問，選択画像，選択数，選択画像のリンク先のページから得られた単語数を示す．選択画像は，各検索単語を入力したときに選択した画像の内容を示している．また，選択数は検索結果画像の上位 20 件から選択した画像の数を表している．

選択画像のリンク先のページから得られた単語のうち HTML タグにより重み付けした単語の上位 100 単語を関連語候補として，本手法と Rocchio の手法により関連語の収集を行った．また，関連語であるか否かの判断は人手により行い，精度評価には平均適合率を用いた．

表 1: 実験データ

検索質問	選択画像	選択数	単語数
小笠原	小笠原満男	2	123
小泉	小泉純一郎	9	435
中田	中田英寿	4	304
松井	松井秀喜	7	492
松坂	松坂大輔	5	345

4.1.2 実験結果

各手法により特定した関連語のうち上位 25，50，75，100 位までの関連語を対象にして平均適合率を求めた．図 5 に，表 1 に示す検索質問毎に平均適合率を求め，さらにすべての検索質問の平均適合率の平均値を求めた結果を示す.

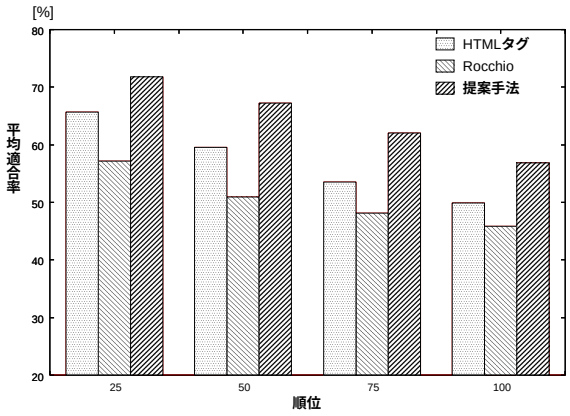

図 5: 関連語収集における平均適合率の平均値

図 5 より，本手法は，HTML タグで重み付けした結果よりもよい結果を得ていることがわかる．一方，Rocchio の手法を用いた場合，上位 25 単語でも平均適合率の平均値が約 60%であり，HTML タグで重み付けした結果より悪くなっている．これは，検索結果 n 件中，選択していない画像のページ内に適切な関連語候補を含んでいたためである．画像のみからページ内容を判断することは困難であり，単純に不選択画像のページを不適合文書と見なすことはできないことがわかった．また，適合文書には関連語に適した単語だけではなく，ノイズとな

る単語も存在している．Rocchio の手法では，適合文書に存在する単語の重みを大きくするため，適合文書にしか存在しないノイズ単語が重要な単語となってしまうことも精度低下の原因の一つである．したがって，不適合文書を使用しない場合においても，適合文書のノイズ単語が精度向上の足枷となるため，WWW 画像検索システムを用いての関連語収集には適用できないといえる．

4.2 関連語を用いたフィード バック画像検索の評価

4.2.1 評価条件

本手法により収集した関連語を WWW 画像検索システムに適用したときの有効性を確かめるために，関連語により検索質問を拡張し画像検索を行った．表 1 に示す検索単語に 4.1 で収集した関連語の上位 10 単語を加えて新たに検索質問を生成し，Google イメージ検索に入力して画像を収集する．さらに収集した画像と選択画像との類似画像検索を行った結果の上位 100 件の画像に対して評価する．また，本手法との比較として HTML タグで重み付けした単語の上位 10 単語を用いて同様の評価を行った．適合画像であるか否かの判断は人手により行い，精度評価には平均適合率を用いた．なお，入力する検索質問は “検索単語 AND(関連語 1 OR 関連語 2 OR $\cdots$ OR 関連語 N)” の形式とした．

4.2.2 実験結果

検索単語に加える単語を増やしながら画像検索を行ったときの検索結果位 100 件の平均適合率と収集できた適合画像数を調べた．図 6 に結果を示す．なお，平均適合率，適合画像数は各検索単語から求まった値の平均値を表している．

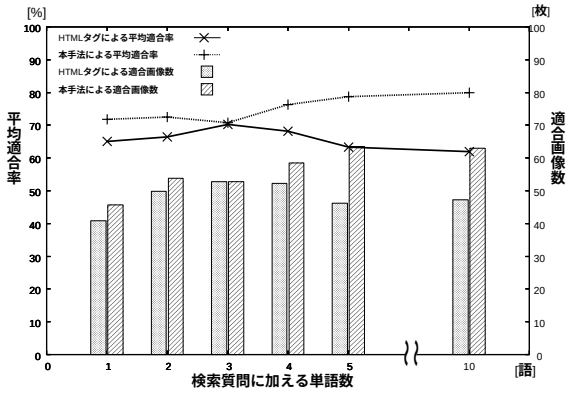

図 6: 画像検索における平均適合率と収集画像数

図 6 より，本手法は HTML タグで重み付けした単語で検索した結果よりも平均適合率，適合画像数ともに良い結果を得ることができている．これは，本手法で収集した関連語を用いることで検索単語を補助する単語としてふさわしい単語を選択できているためである．出現単語を重み付けしただけでは，関連語にふさわしい単語だけでなく,「コラム」のような一般的な単語も上位に出現してしまう場合があるため，それらの単語を用いての検索は精度低下の原因となる．本手法はこのような一般的な単語を除去することが可能であり，フィードバック画像検索に適用したときの有効性が確認できた．

次に，検索質問に加える単語数が検索精度に与える影響について考える．図中より検索質問に加える単語を増やすごとに平均適合率も適合画像数も増加していることがわかる．複数の関連語を用いることでより多くの画像を精度良く収集できることがわかった．しかし，上位 3 単語では本手法と HTML タグによる結果がほぼ同じ結果になっている．これは，HTML タグで重み付けした単語の上位 3 単語に検索単語との関連が強い単語が存在していたためである．このような場合もあるが，より多くの画像を収集したい場合には複数の関連語を用いる必要があり HTML タグによる重み付けだけでは限界がある．このことは上位 10 単語で本手法との差が大きくなっていることからも確認できる．今回は上位 10 単語までの関連語で評価を行ったが，上位 N 単語までの関連語を用いればよいのか検討する必要がある．

5. まとめ

本稿では，WWW 画像検索において，フィードバック情報を用いて検索質問の関連語を収集し，関連語を用いて画像検索を行う手法を提案した．また，関連語の収集精度と関連語を用いた画像検索の精度について評価を行い，本手法の有効性が確認できた．今後は，さらに用いる関連語数を変化させての実験と画像検索に用いる画像特徴量について検討を行い，検索精度の向上を図りたい．

謝辞

本研究の一部は，科学研究費補助金基盤研究 (B)(17300036)，科学研究費補助金基盤研究 (C)(17500644) を受けて行われた．

参考文献

[1] 獅々堀正幹，小泉大地，柘植覚，北研二: 画像知識データベースを用いた WWW 画像検索システムの開発, 電子情報通信学会論文誌, VOL.J87-D-I NO.2，pp.154--163，2004.

[2] Kenji Yanai：Image Collector II: A System for Gathering More Than One Thousand Images from the Web for One Keyword,*In Proc. of IEEE International Conference on Multimedia and Expo,* volume I, pp.785--788, 2003.

[3] Rocchio, J. J.: Relevance feedback in information retrieval, The SMART Retrieval System-Experiments in Automatic Document Processing, Salton, G.(Ed), Prentice Hall, pp.313--323, 1971.

[4] 杉尾敏康，竹野浩，藤本典幸，萩原兼一: WWW に対するマルチメディアデータ検索エンジンの HTML 構文を活かしたスコア付け手法の提案, 第 13 回データ工学ワークショップ (DEWS2002)，2002.