2005－NL－170（8）
2005／11／21

WWW画像検索システムを用いた関連語の自動収集手法

竹安 真紀夫† 獅々堀 正幹† 柘植 覚† 北 研二‡

†徳島大学 工学部 知能情報工学科
‡徳島大学 高度情報化基盤センター

我々は，関連語を用いた検索質問拡張によるWWW画像検索システムの精度向上を目的に研究を進めている．本稿では，検索質問拡張に用いる関連語の自動収集手法を提案し，小規模な関連語収集実験による評価結果を述べる．提案手法は，既存のWWW画像検索システムから得られる検索結果画像の適合性を判断し，適合画像が含まれるWebページ中の単語 (関連語候補) を用いて検索質問の関連語を収集する．一般に，関連語候補には関連語に適した単語ばかりではなく，一般的な単語も多く含まれている．関連語に適した単語は関連語候補中の他の関連語候補と共起することが多いが，一般的な単語は関連語候補以外の単語とも多く共起する．提案手法では関連語候補間の共起の度合い (関連度) により各関連語候補の重要度を判定して，検索質問拡張に適した関連語を特定する．また，既存のWWW画像検索システムを用いれば，検索結果の適合性を文書を読まずに画像のみから判別可能であるため，ユーザの負担が少ないフィードバックが可能である．5種類の単語を用いて提案手法とRocchioの手法との比較実験をした結果，収集結果の上位25単語で約12%，上位50単語で約18%，提案手法がRocchioの手法を上回った．

Automatic Collection Method of Related Terms by using WWW Image Retrieval Systems

Makio Takeyasu† Masami Shishibori† Satoru Tsuge† Kenji Kita‡

†Department of Information Science & Intelligent Systems, Factory of Engineering, Tokushima University
‡Center for Advanced Information Technology, Tokushima University

We study about a WWW image retrieval system by using relevant terms. In this paper, we propose a automatic collection method of related terms for WWW image retrieval systems and described a evaluation result of the proposed method. In this proposed method, first, we select images which have relevance to a query from results retrived by WWW image retrieval system. Then, the query expantion is performed by using related terms in the Web page which is included the relevant images. The related term often appears with other related terms neverthless a lot of the general term appears excluding the related terms. Hence, we calculate a related-cost of the each related term using the co-occurence between related terms. Then, we select the related terms which are useful for query expantion and collected these terms for related terms. Additionally, in this method, we do not need to read the documents for related judgment because the images are used for this judgment. Therefore, we can reduce the user efforts. Experimental results showed that the proposed method could improve the accuracy of collecting the related terms compared to the Rocchio method.

1　はじめに

近年，インターネットの普及に伴い，誰もが情報を発信・受信できるようになり，WWW 空間には無数の Web サイトが存在するようになった．この急速な Web サイト数の増加は，ユーザの求める情報を素早く正確に収集することを困難にしている．この問題に対処するため，既存の WWW 検索システムの機能・精度を向上させるシステムの開発が求められている．

従来より，検索精度を改善する手法として適合性フィードバックが広く利用されている [1]．これは，検索結果の中からユーザが必要とする文書 (適合文書) とそうでない文書 (不適合文書) をシステムに教えることで検索精度を改善する手法である．適合性フィードバックの中でも代表的な手法として，Rocchio の手法 [1] が知られているが，適合文書の単語をすべて同等に扱っている点に問題がある．この問題を解決するために，適合文書において重要な単語を用いて精度向上を図る手法が提案されている [2][3][4]．しかし，これらの手法は文書間の類似性を計る類似文書検索に適用される手法であり，全文検索に基づく (単語を検索質問とする)WWW 検索システムに適用することは難しい．

一方，WWW 検索システムに対しては HTML タグを利用して検索精度を改善する手法が提案されている [5][6]．これらは，タグによる重みのみで出現単語とページ内容の関連を求めているため，必ずしもページ内容に即した単語を用いているとはいえない．また，WWW 検索システムを利用して関連語収集を行う手法 [7] も提案されているが，検索結果ページの適合性を判断しないため収集精度に問題がある．

本稿では，既存の WWW 画像検索システムから得られる検索結果画像の適合性を判断し，適合画像が含まれる Web ページ中の単語 (関連語候補) を用いて検索質問の関連語を収集する手法を提案する．関連語候補には，関連語に適した単語ばかりではなく，一般的な単語も多く含まれている．WWW 検索結果において，関連語に適した単語は限定されたページに存在し，一般的な単語は様々なジャンルのページに存在している．本研究では，関連語は他の関連語候補と共起する度合いが高く，一般的な単語は関連語候補以外の単語ともよく共起することに着目して関連語の収集を行う．本手法は，各関連語候補を入力として WWW 画像検索システムで再検索した結果に基づいて関連語候補間の共起の度合い (関連度) を求める．この関連度の値により関連語候補の差別化を図り，より関連語に適した単語を関連語として収集する．また，既存の WWW 画像検索システムを用いれば，検索結果の適合性を文書を読まずに，画像のみから判別可能であるため，ユーザの負担が少ないフィードバックが可能である．

2　従来の関連語収集手法

適合性フィードバックにより検索質問を拡張する代表的な手法として Rocchio の式 [1] 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り，式 (1) で表される．D_R は検索結果に含まれる適合文書の集合，D_N は不適合文書の集合を表し，$|D_R|$，$|D_N|$ はそれぞれの文書集合に含まれる文書数を表している．また，α,β,γ は 0 以上の定数であり，それぞれ検索質問，適合文書，不適合文書をどの程度重視するかを表している．

$$q' = \alpha q + \frac{\beta}{|D_R|}\sum_{d_i \in D_R} d_i - \frac{\gamma}{|D_N|}\sum_{d_j \in D_N} d_j \quad (1)$$

式 (1) は，適合文書と不適合文書の重みの調整を文書集合毎に正規化している．したがって，文書集合において平均的な単語の重みを用いて検索質問を拡張しているため，適合文書の中でも，より適合度の高い文書に含まれる重要な単語を再検索時に有効利用できていない点に問題がある．

適合度の高い文書に含まれる単語を有効利用する手法として，検索質問文と検索対象文書の類似度における各単語の影響を数値化した「単語寄与度」を用いた手法が提案されている [4]．単語寄与度を用いることで適合文書の特徴を表し，かつ元の検索質問文に含まれていない単語を抽出することが可能になり，抽出された単語を検索質問文に加えて検索質問拡張を行う手法である．これは，類似文書検索に適用される手法であり，検索質問には単語ではなく，文書 (複数の単語) を用いており，ユークリッド距離に代表される距離尺度を用いて類似検索している．そのため，全文検索に基づく (単語を検索質問とする)WWW 検索システムに適

用することは困難である．

一方，WWW 検索システムに対しては HTML タグを利用して検索精度を改善する手法が提案されている [5][6]．これらの手法は，主に WWW 画像検索などマルチメディアデータ検索システムに適用されている．既存のマルチメディアデータ検索システムは，データの周辺に出現する単語などによりデータとの関連の強さを重み付けすることによって，重みの高い順に検索結果を返している．この重みを HTML タグの構文によって割り振ることによって検索精度の改善を行っている．しかし，データの周辺に出現する単語が必ずしもデータと関連が深いとは限らず，HTML タグによる重み付けだけでは不十分である．また，WWW 検索システムを利用して関連語を収集する手法も提案されている [7]．これは，WWW 検索システムの結果から得られる単語の類似性に着目し，ある用語と関連度が高いと判断された単語群を関連語として出力する手法である．この手法は，検索結果の適合性を判断していないため，関連語候補にノイズとなる単語が含まれる可能性がある．したがって，収集精度は WWW 検索システムの検索結果に依存し，様々なジャンルのページが検索された場合には収集精度低下の原因になる．

3　検索質問拡張に基づく画像検索

3.1　本手法の概要

WWW 画像検索システムの検索精度の改善策として，検索質問拡張が考えられる．そこで，本稿で提案する関連語収集手法により検索質問拡張を行い，既存の WWW 画像検索システムにおいて検索を行う手法を提案する．

図 1 に本稿で提案する画像検索手法の流れを示し，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．

手順 1：正解画像の選択

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 (正解画像)$\mathrm{Image}_i(1 \leq i \leq n)$ を選択する．

手順 2：ページ内容の解析

Image_i にリンクする HTML ページを形態素解析し，出現単語 $\mathrm{w}_j(j \geq 1)$ と重み $\mathrm{Weight}(\mathrm{w}_j)$ を集計する．この出現単語 w_j を関連語候補とする．

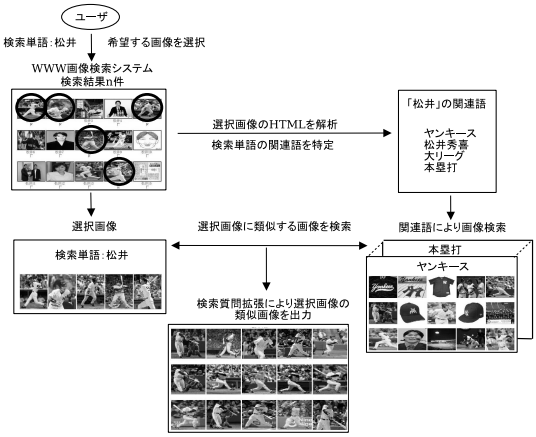

図 1: 本提案手法の概要

手順 3：検索質問の拡張

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 $\mathrm{Image}_k(1 \leq k \leq m)$ と正解画像 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」を正解画像としている．次に，手順 2 で正解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」等の関連語を収集する．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手順 1 で選択した正解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

図 1 の例において，ユーザは検索質問に「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．ユーザは「松井秀喜」の情報を知りたいにも関わらず，検索質問には「松井」と入力しているため，十分な結果を得ることができていない．つまり，検索質問が適切でないため，検索精度の改善には「松井秀喜」を特定できるような単語を検索質問に加える必要がある．ここでは,「松井」

の検索質問を補助する「ヤンキース」や「本塁打」などの関連語の収集方法について述べる．図 2 に関連語収集方法の概要を示し，以下の手順によって関連語を収集する．

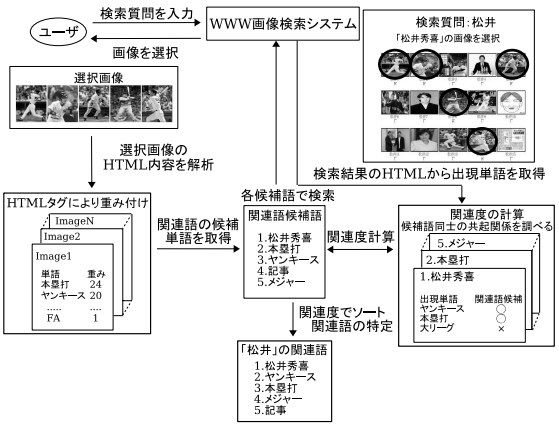

図 2: 関連語収集方法の概要

手順 1：関連語候補の重み付け

3.1 の手順 2 で得た関連語候補 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 [5].

手順 2：関連語候補が存在するページを検索

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 件の検索結果 URL を得る．

手順 3：関連語候補の関連度を計算

検索結果 URL 群に対応する HTML を形態素解析し，単語を得る．この単語群に関連語候補がどれだけ含まれているかを調べ，式 (2) により関連語を特定する．

$$\text{関連度} = \text{URL 群に存在する他の候補語数} \times \text{URL 群に他の候補語が存在する URL 数} \quad (2)$$

HTML 文書において，タイトルタグや見出タグには，少ない文字数でそのページの特徴を表す必要があり，検索質問と密接な関係がある単語が使用されている．また，画像を埋めこんでいるタグの alt 属性には，その画像と関係が強い単語が使われている．したがって，手順 1 において HTML タグを利用して出現単語の重み付けを行うことで，適合文書中の重要な単語を抽出することが可能になる．しかし，重み付けした単語の上位に現われる単語の中には，関連語に適した単語ばかりではなく，一般的な単語も含まれているため，それらの単語を差別化しなければならない．そこで，関連語に適した単語は WWW 検索において限定されたページに存在し，その中で関連語に適した単語同士は共起していると考え，手順 2 では関連語候補を入力として単語毎に検索結果の HTML ページを取得し，手順 3 で式 (2) により関連語の特定を行っている．例えば,「記事」のような一般的な単語を入力として WWW 検索を行うと様々なジャンルの Web ページがヒットするため「松井秀喜」と共起している可能性は低い．しかし,「ヤンキース」のような関連語となる単語は出現ページもある程度限定されるため「松井秀喜」と共起している可能性が高い．この関連語候補間の共起の度合いに着目して関連度の計算を行っている．

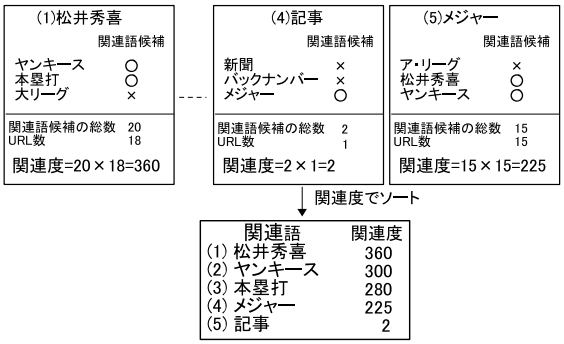

図 3: 関連語特定手順の例

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 件の検索結果 URL 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 URL に出現する関連語候補の総数と出現 URL 数から式 (2) により関連度を求める.「本塁打」等の関連語候補は n 件の URL 中に 20 回出現し，18 件の URL に含まれていたとすると「松井秀喜」の検索質問との関連度は 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から除去することができる．

4 評価実験

4.1 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，正解画像，正解数，正解画像のリンク先のページから得られた単語数を示す．正解画像は，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，正解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．

正解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	正解画像	正解数	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

4.2 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 4 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．また，表 2 には，表 1 中の検索質問を用いて，本関連語収集手法を適用して収集した関連語の上位 10 単語を示す．

図 4 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 (不正解画像) 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，単純に不正解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，式 (1) において $\gamma = 0$(不適合文書を使用しない) とした場合においても，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．

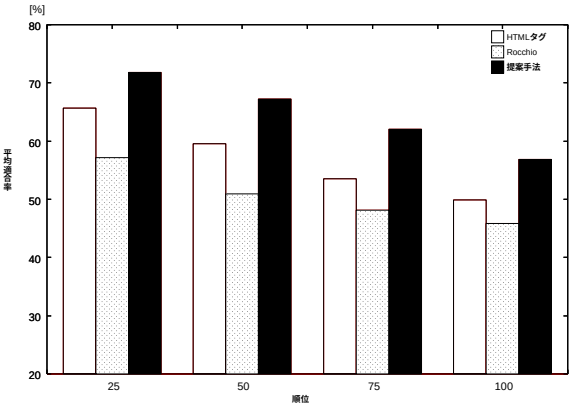

図 4: 検索質問 5 件の平均適合率の平均値

次に，本手法を適用しても高い精度が得られなかった検索質問の平均適合率を図 5，6 に示す．図 5 は検索質問に「小笠原」を入力し，サッカーの小笠原満男選手を正解画像として関連語を収集した結果である．HTML タグで重み付けした結果からも関連語候補にすでに多くのノイズ単語を含んでいることがわかる．これは，検索結果画像 20 件中，正解画像が 2 件であったため正解画像のリンク先のページから収集した単語が 123 単語と非常に少なく，最も良い精度が得られた検索質問「松井」では 7 件の画像を選択して，492 単語を収集し関連語を特定していることからも，少数の限定

表 2: 各検索質問を用いて本手法により特定した関連語の上位 10 件

順位	小笠原	小泉	中田 (blog 含)	中田 (blog 除)	松井	松坂
1	大友良行	小泉総理	中田英寿	中田英寿	ア・リーグ	中日
2	欧州組	イラク	セリエA	セリエA	松井秀	来季
3	日本代表	小泉内閣	サッカー	サッカー	安打	日本シリーズ
4	アルパイ	廃案	予選	日本代表	ノーヒット	松坂大輔
5	サッカー	予算	中田浩二	Jリーグ	秀喜	西武ライオンズ
6	Jリーグ	内閣	決勝	フィオレンティーナ	松井稼頭央	西武
7	スポーツ	参議院	スポーツ	決勝	二塁打	球界
8	中国戦	戦略	イチロー	スポーツ	松井秀喜	広島東洋カープ
9	ジーコ	中小	選手権	ニュース	満塁	大塚晶則
10	セリエA	地域	ニュース	野球	ヤンキース	野球

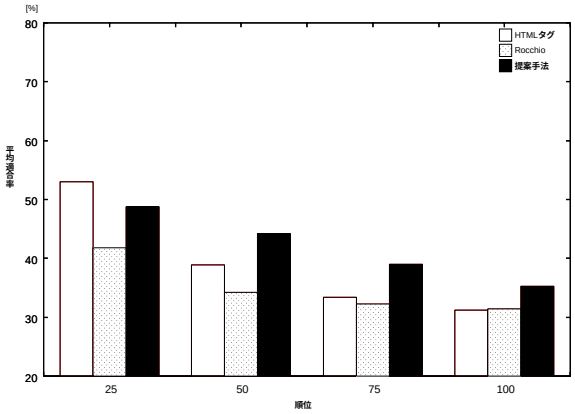

図 5: 検索質問「小笠原」の平均適合率

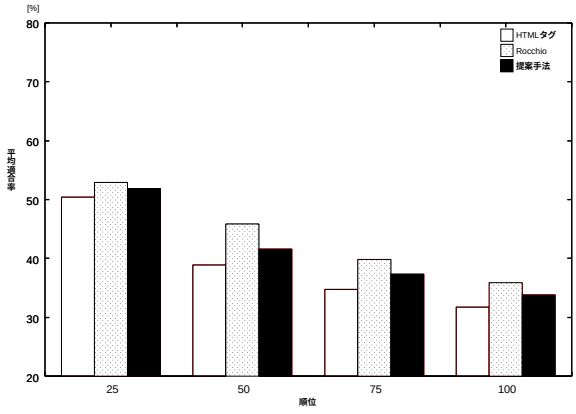

図 6: 検索質問「中田」の平均適合率

されたページから関連語を収集しようとしたことが原因であると考えられる．したがって，正解画像数，あるいは正解画像のリンク先のページから収集できる単語数など文書量が重要となり，正解画像数，単語数によりどの程度改善できるのか検討する必要がある．

図 6 は，検索質問に「中田」を入力し，サッカーの中田英寿選手を正解画像として関連語を収集した結果である．関連語候補にサッカーと関係ない単語が多く見られたため，正解画像とした 4 件のリンク先のページを確認した．その結果，2 件が blog(Weblog) であることがわかった．blog は，時事ニュースやある話題についてのコメントを掲載している形式が多く，一貫したテーマを扱っていないサイトが多く存在する．したがって，扱う内容も幅広く，それに伴い関連語候補の範囲が広くなり正解画像とは関係ない単語まで多く含まれたことが精度低下の原因であると考えられる．正解画像のリンク先のページが blog であるか否かを特定できれば，精度向上が期待できる．そこで，次に blog の影響を調べるため，検索質問「中田」の正解画像のリンク先のページ 4 件から 2 件の blog を除去して関連語を収集した．図 7 に結果を示す．

blog を除去し，本手法を適用した結果，約 25%精度が向上した．HTML タグで重み付けした結果だけをみても，多くのノイズ単語が blog から収集されていたことがわかる．また，2 件の blog を除去したことで収集できる関連語候補となる単語が少なくなり，上記の「小笠原」の結果のような影響が考えられたが，本手法により上位 25 件で約 75%の収集精度を得ることができた．これは，2 件のページから「小笠原」は 123 単語収集できたのに対し，

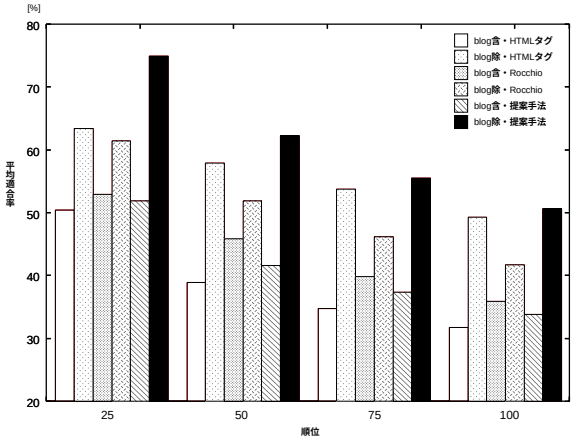

図 7: 「中田」の blog 除去時の平均適合率

「中田」では 222 単語を収集できたことによると考えられる．正解画像のページ数は同じにも関わらず,「中田」の結果が「小笠原」より良いのは，より多くの単語から関連語を特定したことによると思われる．したがって，関連語収集の精度向上には，(1)blog の除去，(2) 正解画像のリンク先のページから収集できる単語数が重要であるといえる．図 8 には，blog を除去した時の検索質問 5 件の平均適合率の平均値を示している．blog を除去していない図 4 に比べ，約 5%精度が向上した．

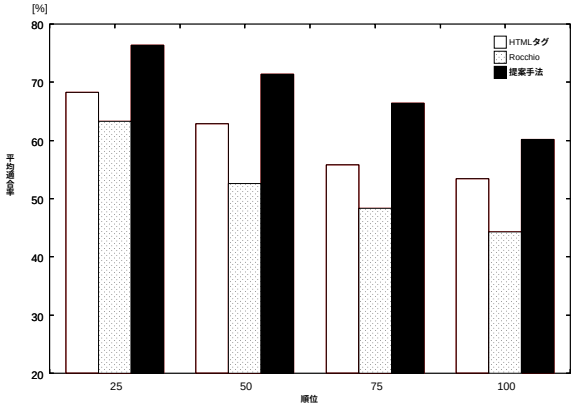

図 8: blog 除去時の検索質問 5 件の平均適合率の平均値

今回の実験では，正解とした画像のページが blog であるか否かの判断は，URL に “blog” の文字列を含んでいるか否かにより行った．これは Yahoo!などのポータルサイトが提供する blog の URL には，“blog” の文字列が含まれていることが多く，ある程度の blog ページを除去できると考えたからである．しかし，完全に除去できていないため精度向上にはページ内容からの blog 判別も行う必要がある．

5 まとめ

本稿では，既存の WWW 画像検索システムから得られる検索結果画像の適合性を判断し，適合画像が含まれる Web ページ中の単語 (関連語候補) を用いて検索質問の関連語を収集する手法を提案した．また，関連語の収集精度について評価を行い，本関連語収集手法の有効性と精度向上についての改善点を確認できた．今後は，blog 等，改善点を考慮しての精度向上を図り，本手法の有効性をさらに検討したい．

謝辞

本研究の一部は，科研費基盤研究 (B) 17300036, 科研費 基盤研究 (C) 17500644 を受けて行われた．

参考文献

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

[2] Salton, G.: Automatic Text Processing:the Transformation, Analysis, and Retrieval of Information by Computer, *Addison-Wesley* ,1988.

[3] 中島浩之, 木谷強, 岡田守: 検索語間における共起関係の特定によるレレバンスフィードバックの高精度化, 情報処理学会論文誌, Vol.40 No.3, pp.1236--1244, 1999.

[4] 帆足啓一郎, 松本一則, 井ノ上直己, 橋本和夫: 文書間の類似度における単語寄与度を利用した検索式拡張手法, 情報処理学会論文誌, Vol.40 No.8, pp.63--73, 1999.

[5] 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.

[6] 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.

[7] 小原恭介, 山田剛一, 絹川博之, 中川裕志: ウェブを利用した関連用語収集, FIT2004(第 3 回情報科学技術フォーラム), pp.183--184, 2004.