質問緩和法によるクロスメディア・メタサーチ

桑原　昭裕† 　角谷　和俊† 　田中　克己†

† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †kuwabara@dl.kuis.kyoto-u.ac.jp, ††{sumiya,ktanaka}@i.kyoto-u.ac.jp

あらまし　現在の Web 空間には膨大な量のコンテンツが分散し，様々なメディアのコンテンツが存在している．よって，ユーザが有効な情報を検索する際は，より効果的な情報を大量のコンテンツの中から選択し，様々な情報を統合する機能が重要である．そこで，我々はこれを解決するために，テキスト検索エンジンや画像検索エンジンなどの多様なメディア向けの検索エンジンを利用するクロスメディア・メタサーチを提案し，このシステムを実現するために検索質問の緩和法を提案した．しかし，現在のままの質問緩和法ではユーザが複数のキーワードを入力すると，非常に効率が悪い．本論文では，これを解決するために質問緩和法の拡張を提案し，また質問緩和法の効率化について考察する．

キーワード　情報検索，情報統合，マルチメディア情報，メタサーチ

Cross-Media Meta-Search by Query Relaxation

Akihiro KUWABARA†, Kazutosi SUMIYA†, and Katsumi TANAKA†

† Graduate School of Informatics, Kyoto University　Yosidahonmati, Sakyou-ku, Kyoto,606–8501 Japan
E-mail: †kuwabara@dl.kuis.kyoto-u.ac.jp, ††{sumiya,ktanaka}@i.kyoto-u.ac.jp

Abstract　Recently, the huge of contents are scattered in several sites and the contents have varied media types in Web space. When users retrieve valid information, it is important to choice more efficient information from many contents, and integrate and arrange various information. We have proposed *Cross-Media Meta-Search*, which uses search engines for various media, such as text search engines and picture search engines, to solve this problem and query relaxation to realize this system. However, when users input more keywords in the system, efficiency is getting worse. In this paper, we propose an expansion of query relaxation and an efficient method of query relaxation.

Key words　information retrieval, information integration, multimedia information ,meta-search

1. はじめに

インターネット技術の進歩に伴って，Web ページの数は劇的に増大してきた．またブロードバンドやデジタルカメラ等の普及により，画像，動画などのマルチメディアコンテンツも非常に増加してきている．このようなことから，ユーザが自分によって有益な情報をサーチエンジンを用いて探すことが非常に重要になってきている．

情報を効果的に検索し統合する手段として，メタサーチエンジンが挙げられる．しかし，メタサーチエンジンに共通する点として３つが挙げられる．

（1） メタサーチで利用する各々のサーチエンジンは同一メディア.

既存のメタサーチではほぼテキストサーチエンジンしか利用していない．これにより，Web ページ内のテキスト文書しか考慮に入れていないため，現在の多様なメディアを有する Web ページ上では十分な検索ができないと考えられる．

（2） どのサーチエンジンに対しても同一の検索質問が実行される．

多数の検索キーワードがあった場合や，検索キーワードごとに関連が全くない場合などは検索結果が思うようにでてこない．このようなことを解消するためには，与えられたキーワードをそのまま利用するのではなく，なんらかの形に変換させる必要があると考えられる．

（3） 統合した検索結果として Web ページへのリンクが示される．

ほとんどの検索システムでは Web ページへのリンクが検索結果として表示されている．そのため，ユーザが検索結果の Web ページを閲覧する際，有益な情報だと判断できる内容がかかれている Web ページを発見するまで，検索結果の一つ一つの Web ページを閲覧するという動作を繰り返さなければならないために非常に労力がかかる．また，一つの Web ページ内には様々な

内容が記述されているために有益な情報だけを効率よく収集することができない．

このようなことから，従来の検索システムではユーザにとって有益な情報を得ることは容易とはいえない．ユーザにとって重要なことは，一つのサーチエンジンで検索キーワードを入力しただけで，テキスト，画像，動画などの様々な情報が得られることである．またユーザにとって有益な情報だけを閲覧しやすい状態で表示することである．

そこで我々はこれまでにクロスメディアメタサーチという手法を提案してきた．この手法では，様々な情報を得るために，既存の様々なメディアに対するサーチエンジンを使用することで，情報量の増加と，情報の種類の多様化を図っている．ここで各サーチエンジンを効率よく利用するために，検索質問を各サーチエンジンに適した形に変換する必要があると考えられる．また，クロスメディアメタサーチでは検索結果として Web ページへのリンクを提示するのではなく，各 Web ページから検索キーワードに関連している部分を抽出し，それらのコンテンツを統合させる．これによって，ユーザの検索キーワードに関する情報が分かりやすく記述されているような Web ページのコンテンツを新たに生成する．

本論文ではクロスメディアメタサーチのための検索質問の緩和方法を提案し，その効率化について述べる．これにより，ユーザの入力したキーワードの数にかかわらず，より効率的に情報を収集することが可能であることを示す．以降，第２章で関連研究，第３章でクロスメディア・メタサーチの概要，第４章で検索質問の緩和について，第５章で質問緩和法の効率化について，第６章でまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 NAVER

検索サイトの NAVER の検索サービス [5] は Web 上にある HTML ページを始め，動画，イメージ，サウンド，文書などを同時に検索し，検索語別にユーザーの検索意図を予想して検索結果を提供する．統合検索では，Web ページ，動画，イメージ，サウンドの検索を同時に行い，検索結果を一画面にまとめて表示するという，機能を持っている．検索質問は従来のサーチエンジンのような入力方式であり，また統合といっても，画像は

図 1　検索サービス Naver

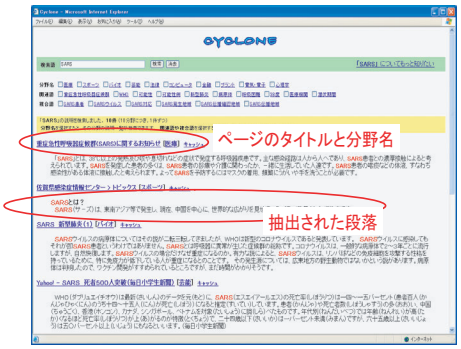

図 2　ウェブ事典検索システム CYCLONE

画像の領域，テキスト検索の出力結果としての URL のリストを表示する領域は分割されている．様々なメディアを扱ってはいるが，この点で本研究とは異なっている．

2.2 Cyclone

Cyclone [6] はウェブを使用した百科事典のような検索システムである．検索キーワードとなる単語を入力すると，Web ページから抽出されたその単語の意味，分野，および関連する単語が，検索結果として表示される．検索結果として Web ページを表示するのではなく，必要な部分だけを提示するという点が本研究と関連している．しかし，このシステムでは１つの単語を入力したときでないと適切な解を得ることができない．つまり，「国語辞典」的な検索結果となっていると考えられる．

3. クロスメディアメタサーチ

複数のキーワードからなるクエリー Q が与えられたとする．従来のメタサーチでは，この複数のキーワードからなるクエリー Q をそのままいくつかの検索エンジンに利用している．またこの時，利用する検索エンジンはほとんどがテキスト検索の検索エンジンである．そしてその後に検索結果として様々なサーチエンジンの結果から重複などを除去して解の Web ページへのリンクを示している．

クロスメディア・メタサーチでは，利用可能なサーチエンジンは異種のものを許している，すなわち，例えば，通常の Google [7], AltaVista [8], Google 画像サーチエンジン [9], また音楽サーチエンジンなどのように，タイプの異なるサーチエンジンの混在を許している点が特徴的である．さらに，与えられた質問 Q，および，タイプの異なるサーチエンジンに対して，質問 Q を変換して各サーチエンジンに送り，その結果を統合しようというものである．この変換する方法として，検索質問の緩和という方法を使用している．

また，検索結果として従来のような Web ページへのリンクではなく，解の Web ページから検索キーワードに関する情報だけ抽出してそれを統合して，検索結果としてユーザに提示するものである．このような方式を用いることにより，従来のように URL を示す検索結果とは異なり，Web ページの内容を抽出して直接表示することによって Web ページへのリンクを辿り閲覧する作業をなくし，検索キーワードに関連している情報だけをまとめて見ることができる．　このようなシステムを利

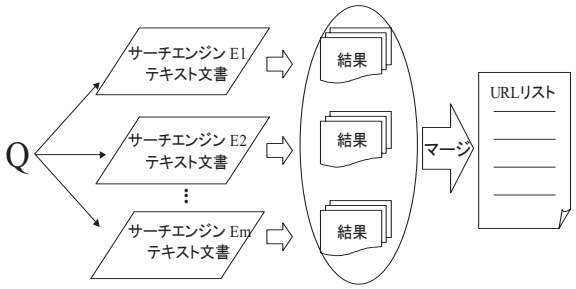

図 3　従来のメタサーチ

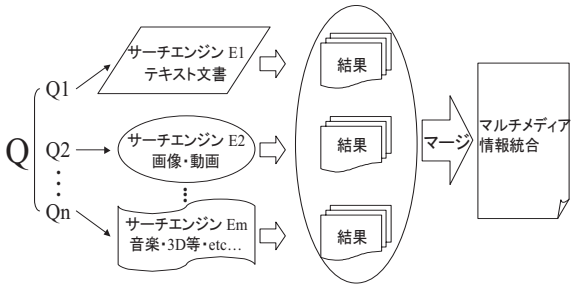

図 4　クロスメディア・メタサーチ

用することによって，ユーザは検索キーワードを入力するだけで，その検索キーワードについての様々な情報を簡単に閲覧できることができる．

4.　検索質問の緩和

4.1　検索キーワードの役割変換

ユーザが複数のキーワード K_1,K_2,...,K_n(n≧2) からなる conjunctive query Q を入力したとする．すなわち，Q = $K_1 \wedge K_2 \wedge \ldots \wedge K_n$ である．また種々の利用可能なサーチエンジンを，E_1,E_2,...,E_m　(m≧2) とする．質問 Q に対する解である Web ページ集合を，Ans(Q) とする．また，質問 Q をサーチエンジン E_i（1≦i≦m）に対して行って得られる解集合を Ans(Q, E_i) と表すものとする．但し，解集合は，Web ページ，画像，音楽などのファイル集合である．

本論文では，E_1 としてテキスト検索，E_2 として Google 画像サーチエンジンを用いる．まず，質問 Q をサーチエンジンの数に合わせて，部分集合に分割する．すなわち，ここではサーチエンジンは２つであるので

$\phi, \{K_1, \ldots, K_n\}$
$\{K_1\}, \{K_2, \ldots, K_n\}$
$\{K_2\}, \{K_1, K_3 \ldots, K_n\}$
$\vdots$
$\{K_1, K_2\}, \{K_3, \ldots, K_n\}$
$\{K_1, K_3\}, \{K_2, K_4 \ldots, K_n\}$
$\vdots$
$\{K_1, \ldots, K_n\}, \phi$

という部分集合に分解する．ここで，部分集合の前者の要素をテキスト検索への，後者の要素を画像検索への入力とする．

このような役割を割り当てる理由は次のとおりである．テキスト検索では，検索キーワードが１つのページ内に書かれていればヒットするが，画像検索の場合は，ファイル名や，画像へのアンカーテキストに検索キーワードが含まれているものがヒットする．ここでヒットするとは検索結果として出力されるということである．よって，テキスト検索よりも検索キーワードに対するヒット率が低く，検索の条件としては厳しいものになっている．そこで，従来では検索キーワードをすべてＡＮＤ検索で画像検索にかけていたものを，部分集合に分けていくつかのキーワードをテキスト検索のキーワードとして使用するものである．これは，画像検索するよりもテキスト検索したほうがヒットしやすいことをふまえて，検索質問を緩和していくものであると考えられる．

4.2　質問緩和法における解集合

部分集合の各要素 $\{K_1, \ldots, K_n\}$, ..., $\{K_1, K_2\}$, $\{K_1, K_3\}$, ..., $\{K_1\}$, $\{K_2\}$, ...をそれぞれ E_2 である Google 画像検索にかける．これによって $Ans(K_1 \wedge \ldots \wedge K_n, E_2)$,..., $Ans(K_1 \wedge K_2, E_2)$,..., $Ans(K_1, E_2)$, $Ans(K_2, E_2)$, ...を得ることができる．これは各要素を Google 画像検索にかけた解集合である．解集合は，画像検索の画像とその画像の参照元の Web ページへの URL によって構成される．

次に，検索結果として出力された画像の参照元の Web ページを収集する．$Ans(K_1, E_2)$ の Web ページに対しては，まだ使用していない部分集合の要素 $\{K_2, \ldots, K_n\}$ が画像の参照元の Web ページにすべて含まれているかを調べる．すべて含まれている場合はこの Web ページを解として収集する．この操作をすべての部分集合に対して行う．ここで解として収集した Web ページは，K_1 で画像検索をし，$K_2, \ldots, K_n$ でテキスト検索をし，両方の検索結果として出力されたページだけを収集することと変わりは無いはずである．つまり，$Ans(K_2 \wedge \ldots \wedge K_n, E_1) \cap Ans(K_1, E_2)$ である．これをすべての部分集合に対し繰り返し行うことによって，Ans(Q) として

$$
\begin{aligned}
Ans(Q) \;=\; & Ans(K_1 \wedge \ldots \wedge K_n, E_2) \\
& \cup \; (Ans(K_1, E_1) \\
& \qquad \cap Ans(K_2 \wedge \ldots \wedge K_n, E_2)) \\
& \cup \; (Ans(K_2, E_1) \\
& \qquad \cap Ans(K_1 \wedge K_3 \wedge \ldots \wedge K_n, E_2)) \\
& \cup \; \ldots \\
& \cup \; (Ans(K_1 \wedge K_2, E_1) \\
& \qquad \cap Ans(K_3 \wedge \ldots \wedge K_n, E_2)) \\
& \cup \; (Ans(K_1 \wedge K_3, E_1) \\
& \qquad \cap Ans(K_2 \wedge K_4 \wedge \ldots \wedge K_n, E_2)) \\
& \cup \; \ldots \\
& \cup \; Ans(K_1 \wedge \ldots \wedge K_n, E_1)
\end{aligned}
$$

を得る．

ここで質問の緩和度という尺度を定義する．緩和度とはいくつのキーワードを画像検索からテキスト検索に緩和させたかを表すものである．検索キーワードが３つの場合を例に挙げると，検索キーワードの３つを And 検索で画像検索にかけた場合は緩

和度０とし，検索キーワードの２つを And 検索で画像検索にかけて残りの１つのキーワードをテキスト検索に使用した場合は緩和度１とするものである．実際に簡易実験によりこの質問緩和法を用いることによって，

4.3 検索質問の緩和の拡張

しかし，上記のような総当り的な方法では，ユーザが大量のキーワードを入力した場合，非常に効率が悪くなってしまう．よって，今回の論文では検索質問 Q が多量のキーワードで構成されている場合を想定し，質問の緩和を利用した効率化手法について考察する．多量のキーワードで検索質問が構成される場合とは，ユーザが自分の欲しい情報を絞って探すために，多くのキーワードを検索エンジンに入力した場合，また，ユーザが Web ページを閲覧している時に，もっと理解したい文章をドラッグ等によって指定した場合などが考えられる．以下に多量のキーワード（N 個のキーワード）に対する検索質問の緩和アプローチを列挙する．

（1） 部分質問の組のラティス構造をどこから実行するか．現在のシステムでは緩和度を０から N まで各部分集合をそれぞれすべて検索エンジンに入力して Web ページを収集してきているため非常に効率が悪い．よってどの緩和度から検索を実行し，その次にどの緩和度にいくかということを決めて，検索を実行することが重要である．

（2） テキストサーチエンジンへの部分質問に，inTitle, inText の概念を導入する．部分集合内の各単語の役割として，その単語は inTitle か inText で使われているかを役割分担させる．これによって，検索キーワードの中でどの単語をメインにして考えていくかを決めることができる．しかし，部分質問をさらに役割によって分割するわけなので，部分質問の組の数が増えることになり，効率は悪くなる．

（3） 質問Qから，不要なキーワードをフィルタリングして除去する．検索質問があまりに多量のキーワードで構成されている場合，その検索質問に対する解の Web ページは見つけることができない．そこで，ユーザの質問の各単語に重要度を設定することによって，重要でない単語を除去し重要な単語だけを抽出し，その単語を検索キーワードとして利用する．こうして新たな質問 Q ’を生成することによって，検索の効率をよくしていく．

（4） メディア毎のサーチエンジンの特性を考慮して，部分質問の処理順序を決める．例えば，画像検索ではキーワードが多いと検索結果がでなかったり，検索しやすい単語があったりする．このように各サーチエンジンにはそれぞれ特性がある．このようなことを考慮に入れて各サーチエンジンを効率よく利用する方法を考える．

5. 質問緩和法の最適化

システムの拡張としては，4.3 で述べたことが考えられるが，本論文では，その中から多数のキーワードが入力されたときの質問を部分質問に分けた時のどのラティス構造から検索質問を実行するのかという点に焦点を当てて論じていく．多数のキーワード（N 個）が入力された時で多種のサーチエンジン（M 個）を利用する場合を考える．今回は簡略化のため，サーチエンジンの数を M=2 として論じる．この時の部分質問のラティス構造は図４のようになる．サーチエンジンが２個の時，N 個のキーワードがあるとする．そのとき部分質問は図４のようなラティス構造になる．図中の各部分集合の前者の要素はテキスト検索に，後者の要素は画像検索に使用するキーワードとする．つまりは上から緩和度０，緩和度１となっていくものである．

図のラティス構造を見れば明らかだが，サーチエンジンが２個でもキーワードがN個であると，部分質問の数は非常に膨大な量となってしまう．このようなことから，クロスメディアメタサーチで効率よく解を収集するためには，総当り的にすべての部分集合において解ページを収集するのではなく閾値を用いて枝刈りする必要性，またどの緩和度の部分集合から検索を実行するのかを考える必要性がある．

まずは枝刈りについて考える．キーワード K_1 と K_2 をサーチエンジン E_1 に入れた解，つまり $Ans(K_1 \wedge K_2, E_1)$ よりも，それにキーワードを付け加えた $Ans(K_1 \wedge K_2 \wedge K_3, E_1)$ は条件が厳しくなるので，解の数は $Ans(K_1 \wedge K_2, E_1)$ 以下になるはずである．よって $Ans(K_1 \wedge K_2, E_1)$ が解を持たないとき，$Ans(K_1 \wedge K_2 \wedge K_3, E_1)$ は解を持たない．つまり個別のサーチエンジンで見ると，テキスト検索については図の下にいけばいくほど解は減少していき，画像検索は図の上にいけばいくほど解が減少してく．これを考慮に入れると，ある部分質問 S の前者の要素，つまりテキスト検索のキーワード群 S_{text} が解を持たない時，その部分質問 S は解を持たない．また S_{text} をテキスト検索のキーワード群の一部に含む部分集合も解を持たない．画像検索においても同様である．このようにして解を持たない部分集合を枝刈りすることができる．

また次にどの緩和度の部分集合から検索を実行するのかを考える．これについては現在のところ，緩和度のちょうど真ん中から実行し，枝刈りを行いながら，上下に移動する．上下の移動方向をどのように決めるかはヒット件数が多い方向に実行していくことを考えている．しかし，これが最適かどうかはわからないので，実験をして検証していく予定である．

6. 検索結果の表示方法

現在の所，プロトタイプでは図６のような表示方法をしている．検索結果を Web ページ単位でなく，Web ページから抽出した段落及び画像としているのである．このように自動的に統合して新しいコンテンツを生成し，これを検索結果としてユーザに提示する．従来のメタサーチエンジンとは違い，検索結果は URL リストの表示ではなく，検索結果の画像，テキスト文書が新しいコンテンツとしてまとまって表示されることが特徴である．

6.1 Web ページからの検索

多数のキーワードを，入力可能なことから、Web ページへの応用が考えられる．Web ページを閲覧しながらその背後で本論文のようなシステムを動かし，ユーザが単語や文章でわからないものがあったならば，その部分を選択することによって，吹

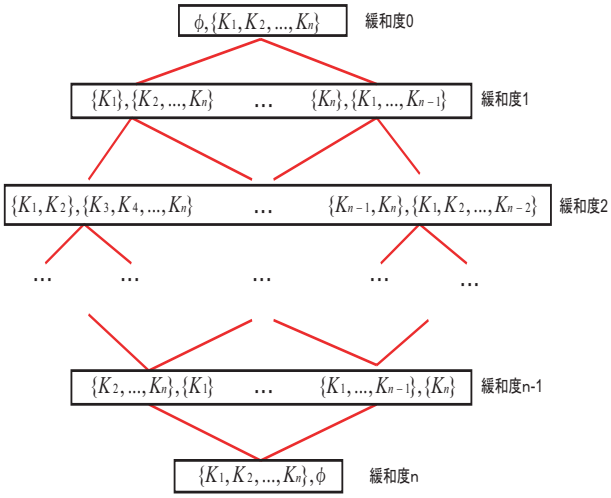

図 5　ラティス構造

図 6　プロトタイプ

き出しのようなものや別 Window によって，わからない情報について統合した結果が表示されるシステムを考えていきたい．こうすることによって，Web ブラウジングを楽しみながらいちいち検索サイトに移動することなく，順次分からない言葉の意味を知ることができ，快適なブラウジングができると考えられる．国語辞典的な利用が考えられる．図 7 にイメージ図を示す．

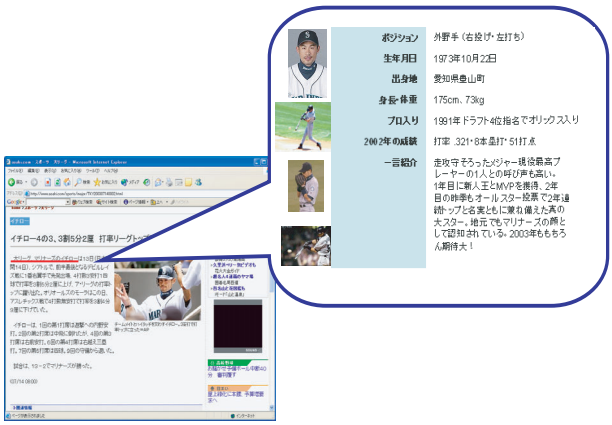

図 7　Web ページからの検索

7.　まとめと今後の課題

本論文では，以前我々が提案したクロスメディアメタサーチを実現する上で使用している質問緩和法の拡張及びその効率化に焦点をおいて考察した．

今後の課題としては以下のようなことが挙げられる．今回，検索キーワードは各サーチエンジンごとに重複を許していない（テキスト検索，画像検索の両方で同じキーワードを利用するのを許していない）ので考える必要性があると思われる．4.3 で述べた拡張方法を組み合わせることによって，よりクロスメディアサーチを効率化していく必要がある．また，今回は検索結果を表示するインターフェースについて触れていないが，インターフェースを考えなくてはならない．また，インタフェース的な面だけではなく，検索結果で Web ページから抽出した文章を単に羅列するのではなく，検索キーワードに応じてユーザの意図をできるだけ反映させるような意味のある配列にしていかなくてはならない．

8.　謝　　辞

本研究の一部は，平成 15 年度科研費特定領域研究 (2)「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：15017249，代表：田中克己）および 21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」による．ここに記して謝意を表すものとします．

9.　参 考 文 献

文　　献

[1] 桑原 昭裕，小山 聡，角谷 和俊，田中 克己：マルチメディア・メタサーチのための質問変換と検索結果の統合，DBSJ Letters Vol.2，No.1
[2] M.C. Schraefel, Yuxiang Zhu, David Modjeska, Daniel Wigdor, Shengdong Zhao : Hunter Gatherer: Interaction Support for the Creation and Management of Within-Web-Page Collections, WWW2002, pp.130–131(2002)
[3] Corin R. Anderson, Eric Horvitz : Web Montage: A Dynamic Personalized Start Page, WWW2002, pp.468–469(2002).
[4] 奈良先端科学技術大学松本研究室茶筅ホームページ : http://chasen.aist-nara.ac.jp/index.html.ja
[5] NAVER Japan : http://www.naver.co.jp/
[6] Cyclone : http://cyclone.slis.tsukuba.ac.jp/
[7] Google : http://www.google.co.jp/
[8] Altavista : http://altavista.com/
[9] Google image : http://images.google.co.jp/