質問緩和法による クロスメディア・メタサーチ

Cross-Media Meta-Search by Query Relaxation

桑原 昭裕♥ 田中 浩也♦
角谷 和俊♣ 田中 克己♠

Akihiro KUWABARA Hiroya TANAKA
Kazutoshi SUMIYA Katsumi TANAKA

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．

Recently, huge contents in a variety of media types are distributed in several Web sites. For acquiring valuable information from those Web sites, it is important how to retrieve pertinent information effectively from many sites, and how to integrate those information. In our previous work, we proposed the notion of the *cross-media meta-search*, which uses search engines for various media types, such as text search engines and image search engines, to solve this problem, and a way of query relaxation to improve the recall ratio without decreasing the precision ratio. When users input more keywords in the system, the efficiency of our previously-proposed method is getting worse. In this paper, we propose an extension of the query relaxation method and a more efficient method for processing the query relaxation.

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) サーチエンジンに対し，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) 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，２章でクロスメディア・メタサーチの概要について，３章で質問緩和法について，４章で質問緩和法の拡張について，５章でまとめと今後の課題について述べる．

2. クロスメディア・メタサーチの概要

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

♥ 学生会員 京都大学大学院情報学研究科修士課程
kuwabara@dl.kuis.kyoto-u.ac.jp
♦ 正会員 東京大学生産技術研究所
tanaka@csis.u-tokyo.ac.jp
♣ 正会員 兵庫県立大学環境人間学部環境人間学科
sumiya@shse.u-hyogo.ac.jp
♠ 正会員 京都大学大学院情報学研究科
ktanaka@i.kyoto-u.ac.jp

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[8]，AltaVista[9]，Google画像サーチエンジン[10]，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ページへのリンクを辿り閲覧する作業をなくし，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．

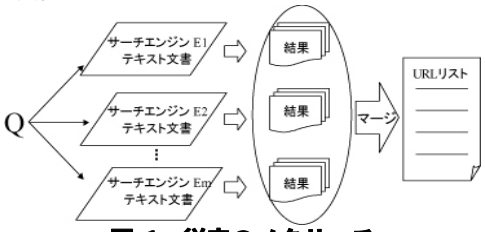

図 1　従来のメタサーチ

Fig.1　Conventional Meta-Search

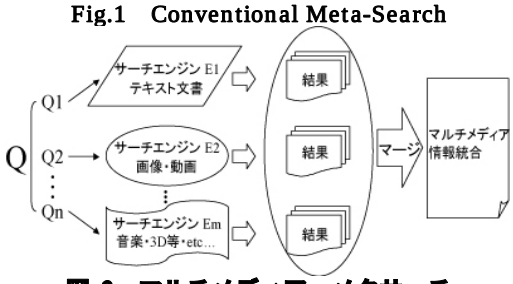

図 2　マルチメディア・メタサーチ

Fig.2　Multimedia Meta-Search

3. 検索質問の緩和

3.1　検索キーワードの変換

ユーザが複数のキーワード $k_1, k_2, \ldots, kn$ $(n \geqq 2)$ からなる conjunctive query Q を入力したとする．すなわち，$Q = k_1 \wedge k_2 \wedge \ldots \wedge kn$ である．また種々の利用可能なサーチエンジンを，$E_1, E_2, \ldots, Em$ $(m \geqq 2)$ 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 Ei（$1 \leqq i \leqq m$）に対して行って得られる解集合をAns(Q, Ei) 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

ϕ ，$\{k_1, \ldots, kn\}$
$\{k_1\}$ ，$\{k_2, \ldots, kn\}$
$\{k_2\}$ ，$\{k_1, k_3, \ldots, kn\}$
...
$\{k_1, k_2\}$ ，$\{k_3, \ldots, kn\}$
$\{k_1, k_3\}$ ，$\{k_2, k_4, \ldots, kn\}$
...
$\{k_1, \ldots, kn\}$ ，ϕ

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．

3.2　質問緩和法における解集合

部分集合の各要素 $\{k_1, \ldots, kn\}$ ，... ，$\{k_1, k_2\}$ ，$\{k_1, k_3\}$ ，... ，$\{k_1\}$ ，$\{k_2\}$ ，... 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$Ans(k_1 \wedge \ldots \wedge kn, E_2)$ ... ，$Ans(k_1 \wedge k_2, E_2)$，... ，$Ans(k_1, E_2)$,$Ans(k_2, E_2)$，... 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 $\{k_2, \ldots, kn\}$ 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，$\{k_2, \ldots, kn\}$ 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，$Ans(k_2 \wedge \ldots \wedge kn, E_1) \cap Ans(k_1, E_2)$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で，Ans(Q) として

$$
\begin{aligned}
Ans(Q) = {} & (Ans(k_1 \wedge \ldots \wedge kn, E_2)) \\
& \cup (Ans(k_1, E_1) \cap Ans(k_2 \wedge \ldots \wedge kn, E_2)) \\
& \cup (Ans(k_2, E_1) \cap Ans(k_1 \wedge k_3 \wedge \ldots \wedge kn, E_2)) \\
& \cup \ldots \\
& \cup (Ans(k_1 \wedge k_2, E_1) \cap Ans(k_3 \wedge \ldots \wedge kn, E_2)) \\
& \cup (Ans(k_1 \wedge k_3, E_1) \cap Ans(k_2 \wedge k_4 \wedge \ldots \wedge kn, E_2)) \\
& \cup \ldots \\
& \cup (Ans(k_1 \wedge \ldots \wedge kn, E_1))
\end{aligned}
$$

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．図３に具体的なキーワードを入れた場合の検索質問の例をあげる．

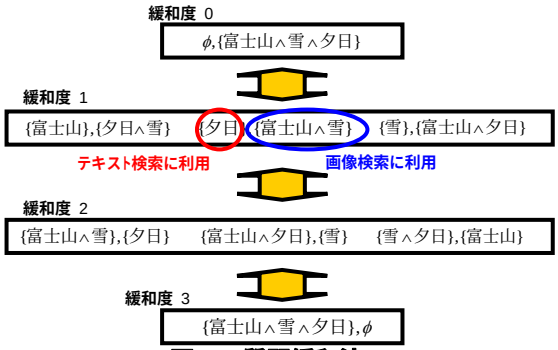

図 3 質問緩和法

Fig.3 Query Relaxation Method

3.3 質問緩和法の実験・考察

実際に簡易実験によりこの質問緩和法が有効であるかどうか検証した．図４に実験結果の一部分を示す．ヒット件数は画像検索のキーワードを入れた時の画像検索結果の中からテキスト検索のキーワードが含まれたページの数を示している．有効ページとは，ヒットしたページの中で検索キーワードの内容を適切に記述しているページであると人が判断したものである．また実験の各検索キーワードごとの再現率と適合率についてのグラフを図５に示す．適合率，再現率とも緩和度ごとに表示している．ここで，適合率とは各緩和度ごとの結果に対する，平均の適合率である．また，再現率とは，Web上での解集合全体は分からないため，各実験ごとの有効なページの総和を解集合全体とするものとする．よって緩和度２の時の再現率は１００％であり，それと相対的に見た評価であることに注意を要する．

検索キーワード
「京都」∧「紅葉」∧「高台寺」

緩和度	画像検索	テキスト検索	ヒット数	有効ページ
0	京都、紅葉、高台寺		6件	6件
1	京都、紅葉	高台寺	38件	13件
	京都、高台寺	紅葉	24件	23件
	紅葉、高台寺	京都	10件	8件
2	京都	紅葉、高台寺	7件	1件
	紅葉	京都、高台寺	6件	2件
	高台寺	京都、紅葉	54件	11件

図 4 実験結果の表

Fig.4 Experimental result

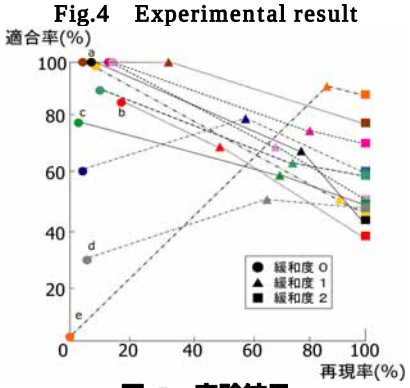

図 5 実験結果

Fig.5 Experiment of Cross-media Meta-Search

実験に対する考察を行う．まず，結果の再現率に注目すると緩和度０から緩和度１にすることによって非常に再現率が上がることが分かる．これは検索キーワードをテキスト検索に利用することが，検索の条件を緩めて網羅的に収集していることの証明になっている．

またグラフより，適合率－再現率グラフには，おおまかに二つのパターンの形が存在することが読み取れる．一つ目は(a)，(b)，(c)のパターンで，緩和度を増やしていくことによって，再現率が上がり適合率が下がっていく一般的な形である．二つ目には(d)，(e)のパターンで，緩和度０の時よりも，緩和度１の時の方が適合率，再現率ともに高くなっている場合である．特に(e)の場合，従来のような緩和度０だけを検索しても結果は得られないが，緩和することによって検索結果を得ることができるのである．これは質問を緩和したことによる有効性が顕著に示されるパターンだと考える．特に緩和度０の結果が少ない時にこのパターンが現れることが多い．実験より質問緩和法を利用することで適合率を適度に保ちながら，再現率を向上させることができると考えられる．

4 質問緩和法の拡張

4.1 質問緩和法の問題点と拡張案

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，in-Title，in-Text 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in-Titleかin-Text 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．

4.2 質問緩和法の最適化

システムの拡張としては，4.1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．そしてN個のキーワードが入力されたとする．その時，部分質問は図6のようなラティス構造になる．図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に枝刈りを行う．

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実験をして検証していく予定である．

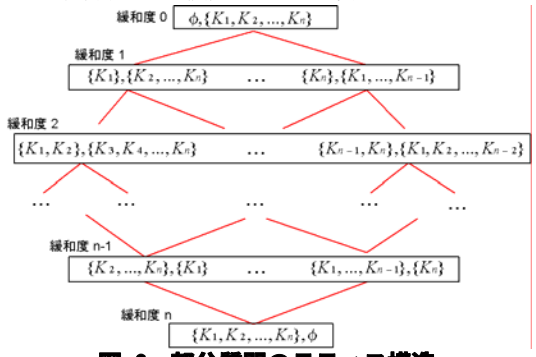

図 6　部分質問のラティス構造

Fig.6　Lattice structure of subqueries

5. 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．4.1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，インターフェイス，および検索結果の画面においてどのように文章を関連付けて必要な情報だけをユーザに提示していけばよいかを考えていかなくてはならない．

[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．

[文献]

[1] M.C. Schraefel, Yuxiang Zhu, David Modjeska, DanielWigdor, Shengdong Zhao : Hunter Gatherer: Interaction Support for the Creation and Management of Within-Web-Page Collections, WWW2002, pp.130–131(2002)
[2] Corin R. Anderson, Eric Horvitz : Web Montage: A Dynamic Personalized Start Page, WWW2002, pp.468–469(2002).
[3] 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
[4] 小山聡，田中克己 : 質問の階層的構造化を用いたWeb検索手法の提案，DBSJ Letters Vol.1, No.1
[5] 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
[6] NAVER Japan : http://www.naver.co.jp/
[7] Cyclone : http://cyclone.slis.tsukuba.ac.jp/
[8] Google : http://www.google.co.jp/
[9] Altavista : http://altavista.com/
[10] Google image : http://images.google.co.jp/

桑原 昭裕　Akihiro KUWABARA
京都大学大学院情報学研究科修士課程在学中．2003 年京都大学工学部情報学科卒業．日本データベース学会学生会員

田中 浩也　Hiroya TANAKA
東京大学生産技術研究所助手. 2003年東京大学大学院工学系研究科博士後期課程修了、工学博士．情報処理学会、日本建築学会、バーチャルリアリティー学会、認知科学会各会員.

角谷 和俊　Kazutoshi SUMIYA
兵庫県立大学環境人間学部環境人間学科教授．1998 年神戸大学大学院自然科学研究科博士後期課程修了，工学博士．マルチメディアデータベース，データ放送の研究開発に従事．IEEE Computer Society，ACM，映像情報メディア学会，情報処理学会，日本データベース学会等各会員.

田中 克己　Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院博士前期課程修了，工学博士．主にデータベース，マルチメディアコンテンツの処理の研究に従事. IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員.