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

Query Translation and Answer Integration Towards Multimedia Meta-Search

桑原 昭裕♥ 小山 聡♦
角谷 和俊♦ 田中 克己♦

Akihiro KUWABARA Satoshi OYAMA
Kazutoshi SUMIYA Katsumi TANAKA

WWWによって取得可能なコンテンツはその量や種類が年々増大しており，ある事象に関する情報を取得したい場合に，これらの情報が大量のWebサイトに分散し，かつ，種々のメディアによって表現されているため，これらの情報を効果的に検索し統合する機能が重要である．従来のWWWのメタサーチエンジンは，各々のサーチエンジンが有するインデックス情報をもとにキーワード検索したWebページを，重複の除去・自動分類などを行うことによって統合して検索結果を表示するものであった．本論文では，これらとは異なり，ユーザが入力した質問キーワード群から，これを部分質問への自動分割・変換を行い，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に対して検索処理を行い，これらの検索結果を自動的に統合する方式を提案する．

The quantity and the kind of contents acquirable from WWW are increasing for these years. When we want to acquire the information about a certain matter, since these information distributes over a lot of web sites and it is expressed by various media, the function which searches these information effectively is important. The conventional WWW meta-search engines basically integrate the retrieval results each of which is obtained by applying the same user query to each Web search engine. Many meta-search engines have functions for duplication, removal, automatic classification, and displaying retrieval results. In this paper we propose a system which automatically divides and transforms a given keyword query, applies transformed sub queries to the search engines for various media, such as text search engines and picture search engines, and integrates these retrieval results automatically.

1. はじめに

インターネット環境がますます普及してきたためWebページの数は増大する一方であり，またブロードバンドやデジタルカメラ等の普及によりWebページのコンテンツは画像等が多く取り入れられて，ますます多種多様になってきている．このように，Web空間には様々な情報が氾濫しているため，ユーザが有益な情報だけを収集してくることは非常に困難になってきている．

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．従来のメタサーチエンジンでは，ユーザが検索キーワード群を入力し検索を実行すると，各サーチエンジンに入力されたキーワード群を渡し，各々のサーチエンジンがキーワード検索しWebページを収集してくる．そしてメタサーチエンジンは，各々のサーチエンジンが収集したWebページの重複を除去したり，自動的に分類等の操作を行い，検索結果を出力する．ここで，どのメタサーチエンジンにも共通する点として３つのことが挙げられる．

(1) メタサーチで利用する各々のサーチエンジンは同一タイプである点.
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点.
ユーザが入力した検索質問をそのまま使っていたのでは，キーワードが多ければ検索結果がでてこない場合もあるし，また少なければ余分な情報まで収集してきてしまう．よって，適切な検索質問を適切なサーチエンジンに使用することが重要である．

(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点.
検索結果がWebページへのリンクで表示されているため，ユーザが検索結果のWebページを閲覧する時に，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操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.

このように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．またシステムが関連のない情報を取り除き，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．

そこで本論文では，ユーザが入力した検索キーワード群を分割して，テキスト検索エンジンや画像検索エンジンなど，多様なメディア向けの既存の検索エンジンに対して検索処理を行い，これらの検索結果を自動的に統合する方式を提案する．このようなシステムを利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と考えられる．いわばWebを利用した百科辞典的なものである．

2. 質問変換と検索結果の統合の概要

複数のキーワード $k_1, k_2, \ldots, kn$ $(n \geqq 2)$ からなる conjunctive query Q が与えられたとする．すなわち，$Q = k_1$

♥ 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 京都大学大学院情報学研究科
{ tanaka, sumiya, oyama }@i.kyoto-u.ac.jp

∧k_2∧...∧kn である．種々の利用可能なサーチエンジンをE_1,E_2, ... ,Em (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Ei（1≦i≦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．

従来のWebのメタサーチエンジンの最も基本的なものは，質問Q，サーチエンジン E_1,E_2, ... , Em (m≧2) に対して，Ans(Q) = Ans(Q, E_1)∪...∪Ans(Q, Em) として，Ans(Q) に対して自動分類などを行ってユーザに提示するものである．また，E_1,E_2, ... ,Em は，すべて同一のタイプのサーチエンジンであり，多くの場合それらはテキストサーチである．

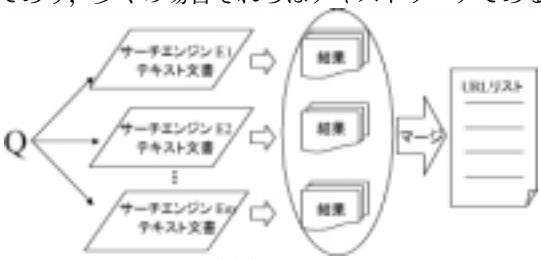

図 1 従来のメタサーチ
Fig.1 Conventional Meta-Search

本論文で提案するマルチメディア・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．例えば，E_1 は通常のGoogle [7]，E_2 はAltaVista [8]，E_3 はGoogle画像サーチエンジン[9]，E_4 は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，提案する方式では，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 E_1,E_2, ... ,Em 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．これは，どのキーワードにどの検索エンジンを割り当てたものが，ユーザの要求に最も合うかは分からないので，それを補助するために検索キーワードの分割パターンを多数生成するものである．従来の検索エンジンの検索結果の提示方法のほとんどは，Webページのリンクである．よって，ユーザはリンク先のWebページを訪れて，有用な情報が得られるWebページまでこれを繰り返さなくてはならない．また一つのWebページには様々な内容が書かれているため有用な情報が書かれているかどうかを判断するにも労力がかかる．本論文ではこれを解決するために，検索結果のWebページから検索キーワードに関するところだけを抽出する．解として得られたAns(Q) に対して，単語の出現頻度を用い検索キーワードとの関連性を考える．これによって解をWebページ集合とするのではなくて，抽出したオブジェクトの集合として考える．

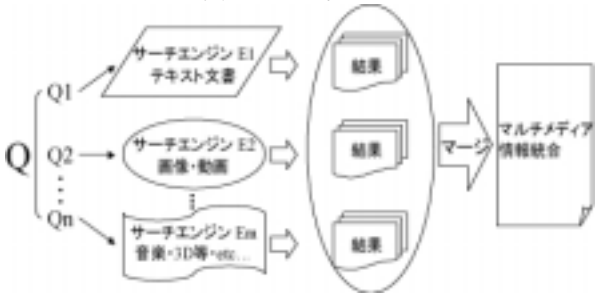

図 3 マルチメディア・メタサーチ
Fig.2 Multimedia Meta-Search

本研究では，以降，利用可能なサーチエンジンとしてE_1 をGoogle画像サーチエンジン， E_2 を通常のテキスト検索として考えていく．以下にシステムの流れを示す．

(1) まず，質問Q を利用するサーチエンジンの数と同数要素をもつ部分集合すべてに分割する．これらの部分集合に分解した質問Q に対して，画像サーチに使用する要素，テキストサーチに使用する要素というように役割を設定する．

(2) 画像サーチに使用するキーワードをGoogle 画像サーチエンジンに入力し，その結果として出力されたWeb ページの中にテキストサーチで使用するキーワードが含まれているならば，そのWeb ページを有用なものとして収集する．

(3) Webページから検索キーワードに関連した部分だけ抽出する．

(4) 最後にこれらを統合させ，WebページへのURLではなくて，画像やテキスト文書等のWebページの内容が書かれている新しい検索結果のコンテンツを生成し，提示する．統合したコンテンツにはユーザとのインタラクションによって検索結果の表示方法を動的に変化させることができる機能を持たせる．

3. 質問変換によるメタサーチ

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, ... ,kn)(n≧2) からなるconjunctive query Q を入力したとする．すなわち，Q = k_1∧k_2∧...∧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, ... ,Em (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Ei（1≦i≦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Google 画像サーチエンジン，E_2 としてテキストサーチを用いる．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, ... , kn}
{k_1} , {k_2, ... , kn}
{k_2} , {k_1, k_3, ... , kn}
...
{k_1, k_2} , {k_3, ... , kn}
{k_1, k_3} , {k_2, k_4, ... , kn}
...
{k_1, ... , kn} , {φ}

という部分集合に分解する．これは，どのキーワードにどの検索エンジンを割り当てたものが，ユーザの要求に最も合うかは分からないので，それを補助するために検索キーワードの割当パターンを多数生成するものである．ここで部分集合の要素の前者に対しては画像検索，後者に対しては通常のテキスト検索にかけるという役割を持たせる．

3.2 Web ページの収集

部分集合の各要素{k_1},{k_2}, ... ,{k_1, k_2},{k_1, k_3},... , {k_1, ... ,kn} をそれぞれE_1であるGoogle画像検索にかける．これによってAns(k_1, E_1),Ans(k_2, E_1), ... , Ans(k_1∧k_2, E_1), ... ,Ans(k_1∧...∧kn, E_1) 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．ここで，各画像は部分集合の要素に対しての画像検索だけの結果であるので，有益な情報をフィルタリングするためにページ内のテキスト文書に注目する．Ans(k_1, E_1)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, ... , 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, ... , 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_1, E_1) ∩ Ans(k_2 ∧... ∧ kn, 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

Ans(Q) = (Ans(k_1∧...∧kn, E_2))
∪ (Ans(k_1, E_1) ∩ Ans(k_2 ∧...∧ kn, E_2))
∪ (Ans(k_2, E_1) ∩ Ans(k_1∧k_3∧...∧kn, E_2))
∪ ...
∪ (Ans(k_1∧k_2, E_1) ∩ Ans(k_3∧...∧kn, E_2))
∪ (Ans(k_1∧k_3, E_1) ∩ Ans(k_2∧k_4∧...∧kn, E_2))
∪ ...
∪ (Ans(k_1∧...∧kn, E_1))

を得る．以下簡略化のため，画像検索に使ったキーワードが K_1 である解の集合Pを(k_1) と表す．つまり，Ans(k_1, E_1)∩ Ans(k_2 ∧...∧ kn, E_2) をP(k_1) として表し，Ans(k_1∧k_2, E_1) ∩ Ans(k_3 ∧...∧ kn, E_2) をP(k_1∧k_2) と表す．

ここで例として，ユーザの入力した検索キーワードが「富士山」，「雪」という二つであった場合の収集してくるページを考える．図３はその時のベン図である．図の赤い部分が収集してくるWebページと対応している．このようにして従来では収集してないWebページまで網羅的に収集してくる．

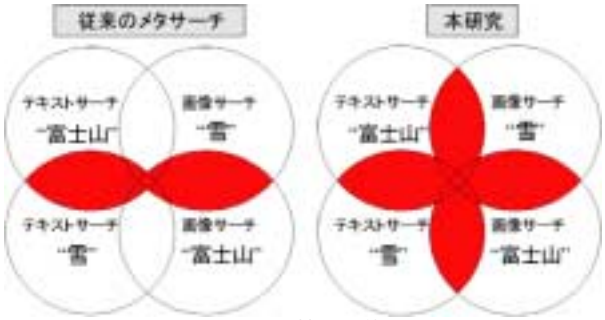

図 3 質問変換
Fig.3 Query Translation

実際に「富士山」「雪」「夕日」という３つのキーワードに対し質問変換を行い，Webページを収集してきた結果を示す．

画像サーチのキーワード	テキストサーチのキーワード	ヒットしたページ数
富士山 ∧ 夕日 ∧ 雪		０件
夕日 ∧ 雪	富士山	４件
富士山 ∧ 雪	夕日	８件
富士山 ∧ 夕日	雪	11件

表１ 実際の収集結果
Table.1 Collection result

3.3 特徴ベクトルの生成

Ans(Q) として得られた各Web ページに対し，各Web ページを特徴付けるために，Web ページ内の各単語の出現頻度に基づく特徴ベクトルを作成する．特徴ベクトルの要素としては，各Webページ内に出現する各単語の出現頻度であるtf値を用いる．

3.4 Web ページからの関連文書の抽出

一つのWebページ内には検索キーワードと関係のない話題も含まれている．関係のない話題を除去することによって効率よく有益な情報を取得することができると考える．これに基づいて，本研究では，検索キーワードに関連する部分だけをWebページ内から抽出する．検索キーワードへの関連性を考えるにあたって，Webページ内のテキスト中の単語の頻度を利用する．単語の頻度に基づき単語の重要度を計算し，文が含む単語の重要度に基づいて文の重要度を計算するという手法を用いる．計算式は以下の通りである．

$$w_S^{P(K)} = \sum_{t \in S} w_t^{P(K)}$$

$$w_t^{P(K)} = tf\ (t, P(K))$$

まずAns(Q)の各Webページに対して，Webページ内のテキスト中に出現する単語の頻度を計算する．各単語の種類としては「名詞」「形容詞」「動詞」「未知語」を利用する．また解析には茶筅[3] を利用した．ここであるクラスタP(K) のWebページ内の単語t の重要度を単語の頻度を利用してtf(t,P(K))とする．P(K) での単語の頻度を用いることにより，同じ単語であってものどのクラスタに属しているかによって重要度の値が異なってくる．これによって各クラスタごとの重要度を際立たせることができると考える．これを各クラスタに対して求める．

次にある段落S での重要度を求める．段落の重要度には文中に含まれる各語の重要度の合計を使用する．これによってWebページ内の段落に対して重要度を計算することができる．このように求めた重要度によりある閾値を超えた段落を有益なものとしてWebページから抽出する．

図５に抽出の様子を表した．抽出してくる内容は，画像検索で得られた各画像と，その各画像の参照元のWebページから上で求めた重要度を基にして検索キーワードに関連している部分を抽出してきた文書である．

図 5 Web ページからの抽出
Fig.5 Extraction from a Web page

4. 検索結果の統合

３章によって，検索キーワードは部分集合に分けられ，各部分集合の検索結果はP(φ), P(k_1), ... , P(kn), P(k_1∧k_2), P(k_1∧k_3), ... , P(k_1∧...∧kn) として各クラスタに入っている．ここで，各クラスタの中のWebページから抽出したオブジェクトの集合を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，統合結果は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．検索結果を統合したプロトタイプシステムを図６に示す．

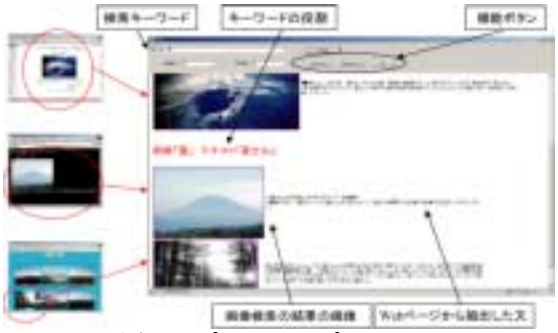

図 6　プロトタイプシステム
Fig.6　Prototype System

以下にコンテンツの構成について述べる．現在の考察段階カテゴリー的な表示方法を考えている．

カテゴリー的な表示とは，ユーザの興味に合わせて各クラスタを表示していくことである．有効な画像が得られなかった場合に，図７のように画像検索に使われるキーワードの内の一つのキーワードを画像検索でなく，テキスト検索に使ったクラスタを表示する．画像検索よりテキスト検索は検索結果がヒットしやすいので，検索質問の条件を少しゆることで有益な情報を得るわけである．また，クラスタを移動していく時には同じキーワード数で最も特徴の違う検索結果を表示させる．これにより，キーワードの役割を変えることによっては違う検索結果の側面があることをユーザに示すことができる．この表示方法ではユーザの興味に合わせて検索結果の表示を変化させることできるのが利点であるが，その分，システムとのインタラクションが必要となってしまう．

図 7　カテゴリー的な表示のイメージ図
Fig.7　Image figure of a category display

5. まとめと今後の課題

本研究では，ユーザの情報検索を支援するために，ユーザが入力した質問を変換することにより既存の様々なメディアに対するサーチエンジンを利用し，検索キーワードに関連したマルチメディアオブジェクトをWeb空間上から抽出し，それらを統合し，検索結果として新たなコンテンツを作成するマルチメディア・メタサーチを提案した．

しかし，マルチメディア・メタサーチと言っても，まだテキスト文書と画像のみの検索でしか考えていないので，動画や音声，Web空間上にあるあらゆるメディアに対して検索の対象としていかなければならない．また，収集してきたマルチメディア情報をどのように統合し，どのようにユーザに見せたらユーザの情報検索をより支援できるのかを模索していくことが重要である．

今後の課題としては，システムの評価が挙げられる．従来の検索システムにおける再現率，適合率ではなく，ユーザがどれだけ有効な情報をどれだけ効率的に検索することができたかということをあるかを評価できる尺度を考え，その尺度をもとにしてシステムを評価していきたいと考えている．

[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)および平成14～16年度科研費基盤研究(Ａ)(２)「モバイル環境におけるコンテンツのマ ルチモーダル検索・提示と放送コンテンツ生成」(14年度課題番号：14208036，代表：田中克己)，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」(代表：上林弥彦）による．ここに記して謝意を表します．

[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] NAVER Japan : http://www.naver.co.jp/
[5] 富士通 : MIRADOR-Search, http://www.labs.fujitsu.com/News/1999/Dec/9-2.html
6] 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
[7] Google : http://www.google.co.jp/
[8] Altavista : http://altavista.com/
[9] 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

小山 聡　Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助手．2002 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．情報検索，データマイニングなどの研究に従事．電子情報通信学会，人工知能学会，ACM，AAAI 各会員．

角谷 和俊　Kazutoshi SUMIYA
京都大学大学院情報学研究科社会情報学専攻助教授．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．