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

桑原　昭裕† 　小山　聡†† 　角谷　和俊†† 　田中　克己††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 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,oyama}@dl.kuis.kyoto-u.ac.jp, ††{sumiya,ktanaka}@i.kyoto-u.ac.jp

あらまし　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．
キーワード　情報検索，情報統合，マルチメディア情報

Query Translation and Answer Integration Towards Multimedia Meta-Search

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

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

Abstract　The quantity and the kind of contents acquirable with WWW are increasing every year. When we want to acquire the information about a certain matter, since these information distributes to a lot of websites and it is expressed by various media, the function which searches these information effectively and unifies it is important. The conventional WWW meta-search engines are what unifies the retrieval by keyword Web pages based on the index information which each search engine has by performing duplicating removal, automatic classification, etc, and displays reference results. In this paper, unlike these, we propose the system which divides the question keyword group which the user inputted, performs reference processing to the search engines for various media, such as text search engines and picture search engines, and unifies these reference results automatically.
Key words　information retrieval, information integration, multimedia information

1. 緒　　論

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２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） 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.

- 検索結果では 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．また，検索を利用するにあたって，ユーザは自分の求める情報を取得するために，どんなキーワードをどんなサーチエンジンに入力したらよいのか分からない．

そこで本研究では，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う．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．

次に，検索結果として出てきた Web ページから検索キーワード群に関連している情報だけを抽出する．検索結果の Web ページ内に出現する単語の出現頻度に基づき，それがある閾値を超えたものを検索キーワードの関連語とし，関連語が出現する部分だけを Web ページから抽出する．抽出するものは画像，テキスト等などのマルチメディアオブジェクトである．このようにして Web ページ単位での検索結果をオブジェクト単位に置き換える．

最後に，抽出したオブジェクトを自動的に統合させ新たなコンテンツを生成させる．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．特に，ユーザがこの言葉はなんだろう，といったような時に使用することでその言葉がどんなものなのか効果的に理解することができるようになると考えられる．

以降，第２章で基本的事項と関連研究について，第３章でシステムの概要を，第４章で質問変換によるメタサーチについて，第５章で検索結果の統合について，第６章でプロトタイプとそれに基づく本研究の考察，第７章で結論を述べる．

2. 基本的事項と関連研究

2.1 基本的事項

2.1.1 メタサーチエンジン

メタサーチエンジンとは，自分自身でデータベースを持たず，ユーザからの検索要求を複数のロボット型検索サービスやディレクトリ型検索サービスに送り，各サーチエンジンの検索結果から Web ページの重複を除去したり，自動的に分類をするなどの，加工・編集や再び独自にランキングをつける等の操作をして，ユーザに検索結果を検索サービスである．

2.1.2 tf ・idf 法

Web ページを特徴付けるものの一つに，Web ページ内に出現する単語が挙げられる．Web ページ内に出現する単語のうち，その Web ページにおいて出現頻度が高い単語は，その Web ページを特徴付けるものとして考えることが出来る．これを単語の重み付けに使用する．ある文書 d 中に出現する単語 t の頻度を term frequency と呼び，これを tf(t,d) で表わす．この tf(t,d) を文書 d における単語の重み w_t^d と考えることができる．すなわち

$$w_t^d = tf(t,d) \tag{1}$$

とする．

しかし単語の頻度は単語の網羅性を高めるには貢献するが，単語の特定性には必ずしも役に立たない．各文書の頻度は考慮していても，文書集合内の他の文書の単語の分布については考慮していない．これを考慮するために特定性を表すための尺度 idf を定義する．

$$w_t^d = idf(t) = \frac{logN}{df(t)} + 1 \tag{2}$$

ここで N は検索対象となる文書集合中の全文書数，df(t) は単語 t が出現する文書数である．これを用いることによって，特定の少数の文書に出現する単語に大きい重みを与えることができる．

またこれらを単独で用いるよりも両方の性質を合わせ持つように，２つの尺度を組み合わせて単語の重みを計算することが考えられる．よって tf(t,d) と idf(t) の積

$$w_t^d = tf(t,d) * idf(t) \tag{3}$$

を重み付けに用いる．これを tf∗idf 重み付けと呼ぶ.

2.1.3 類 似 度

検索結果として出力された Web ページを特徴付けるために特徴ベクトルを生成する．特徴ベクトルを使い各 Web ページ同士の類似度を計算するにあたり，様々な類似度の尺度の中から，ベクトル間の余弦を用いた手法を用いる．

- 余弦

類似度として２ベクトル間の余弦の値を利用する方法である．各 Web ページの特徴ベクトルを $\vec{x}, \vec{y}$ とすると，以下のようになる．

$$sim(\vec{x}, \vec{y}) = \cos(\vec{x}, \vec{y})$$

図 1　検索サービス Naver

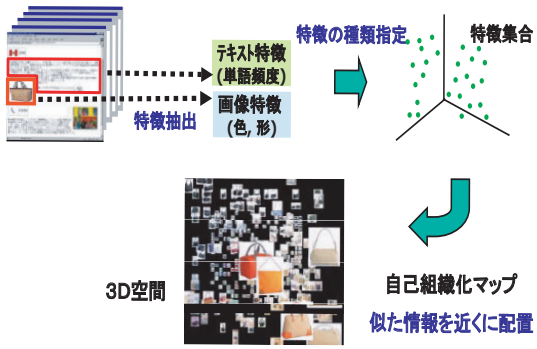

図 2　MIRADOR-Search

$$= \frac{\vec{x} \cdot \vec{y}}{\| \vec{x} \| \| \vec{y} \|}$$

$sim(\vec{x}, \vec{y})$ の値は 1 以下であり，これに近づくほど類似度が高くなる．ただし，この手法はベクトルの大きさ（ノルム）を考慮していないことに注意しなければいけない．

2.2　関 連 研 究

2.2.1　NAVER

検索サイトの NAVER の検索サービス [4] 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．

2.2.2　MIRADOR-Search

MIRADOR-Search [5] とは，富士通によって開発された，テキストによる意味的検索と画像による視覚的検索とを統合した「情報を眺めて選ぶ」クロスメディア検索機能を提供するサービスである．これは，ユーザが検索キーワードを入力，その情報を基に Web Crawler がインターネット上の Web ページを解析し，画像と説明テキストを自動収集する．そして検索結果として，収集してきた画像を画像の色や形などの画像特徴に基づいて，3 次元上にマッピングして表示するものである．しかしクロスメディア検索といってもその結果として出力されるのは画像のみであり，テキスト情報は画像の特徴量を出すためだけに使用される．画像に特化した検索だといえる点で，様々なメディアを扱うことを提案している本研究とは異なっている．

2.2.3　質問の階層構造を用いた検索手法

この手法 [6] は小山らによって提案された，検索エンジンに複数のキーワードを入力した場合における，検索質問の階層的構造化を用いた Web 検索手法である．同じキーワードからなる検索に対し，主題を表すキーワード，主題に関する内容を表すキーワードのようにキーワードごとに役割を区別することで検索精度を向上させるといったものである．また検索結果の表示方法として同じキーワードでも役割を変えて検索していった検索結果を比較して，大きな相違があった検索結果集合を並列に提示している．本研究ではこの手法を利用し，検索キーワードを，画像検索に使う質問，テキスト検索に使う質問のように変換させて利用していることで，検索キーワードに役割を持たせている．

3.　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,...,K_n(n≧2) からなる conjunctive query Q が与えられたとする．すなわち，Q = $K_1 \wedge K_2 \wedge \ldots \wedge K_n$ である．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．

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 $E_1, E_2, \ldots, E_m$(m≧2) に対して，Ans(Q) = Ans(Q, E_1)∪...∪Ans(Q, E_m) 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,...,E_m 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．

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，すなわち，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,...,E_m に対し

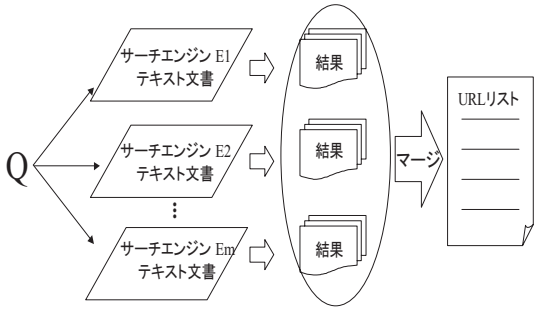

図 3　従来のメタサーチ

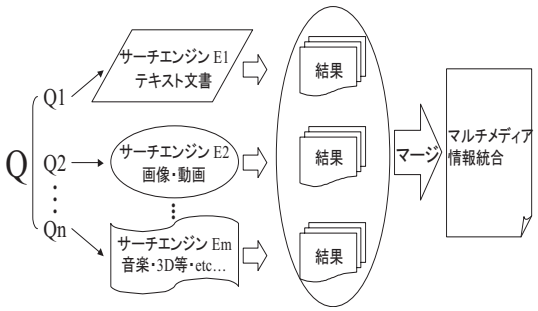

図 4　マルチメディア・メタサーチ

て，質問 Q を分解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．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集合として考える．

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の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．

4.　質問変換によるメタサーチ

マルチメディア・メタサーチのために質問を変換して，様々なメディアのサーチエンジンに対し質問をわたして検索を行う．

4.1　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,...,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 Google 画像サーチエンジン，E_2 として画像周辺の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

$\{K_1\}$，$\{K_2, \ldots, K_n\}$
$\{K_2\}$，$\{K_1, K_3 \ldots, K_n\}$
⋮
$\{K_1, K_2\}$，$\{K_3, \ldots, K_n\}$
$\{K_1, K_3\}$，$\{K_2, K_4 \ldots, K_n\}$
⋮
$\{K_1, \ldots, K_n\}$

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．　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５は従来の検索エンジンで検索キーワードが二つであった場合の収集してくるページ，図６は本手法で収集してくるページを表している．

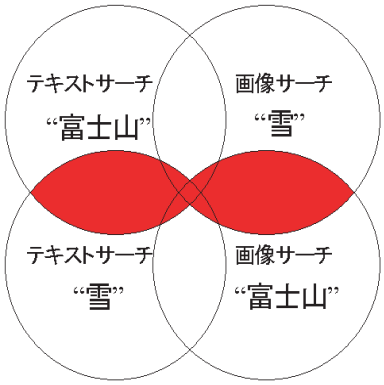

図 5　従来の収集ページ

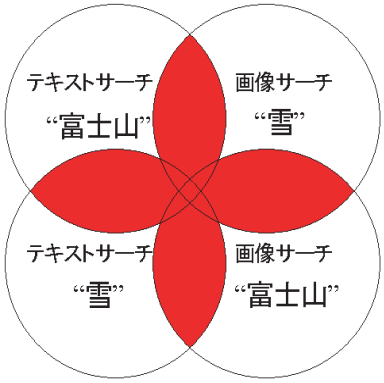

図 6　本手法での収集ページ

4.2 Web ページの収集

部分集合の各要素 $\{K_1\}$, $\{K_2\}$, ..., $\{K_1, K_2\}$, $\{K_1, K_3\}$, ..., $\{K_1, \ldots, K_n\}$ 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 $Ans(K_1, E_1), Ans(K_2, E_1), \ldots, Ans(K_1 \wedge K_2, E_1), \ldots, Ans(K_1 \wedge \ldots \wedge K_n, E_1)$ 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 $\{K_2, \ldots, K_n\}$ 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，$Ans(K_1, E_1) \cap Ans(K_2 \wedge \ldots \wedge K_n, E_2)$ 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

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

を得る．以下簡略化のため，画像検索に使ったキーワードが K_1 である解の集合を $Im(K_1)$ と表す．つまり，$Ans(K_1, E_1) \cap Ans(K_2 \wedge \ldots \wedge K_n, E_2)$ を $Im(K_1)$ として表し，$Ans(K_1 \wedge K_2, E_1) \cap Ans(K_3 \wedge \ldots \wedge K_n, E_2)$ を $Im(K_1 \wedge K_2)$ と表す．

ここで再び，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時の収集対象となるページの例を図 7 に示す．

4.3 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．

4.4 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

図 7　収集対象のページ

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．

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 Im(K_1) のクラスタにおけるある単語 t の頻度を tf(t,$Im(K_1)$) とする．各 Web ページの単語の頻度の計算結果を総計して，$Im(K_1), \ldots, Im(K_n), Im(K_1 \wedge K_2), Im(K_1 \wedge K_3,), \ldots, Im(K_1 \wedge \ldots \wedge K_n)$ のそれぞれにまとめる．また Ans(Q) 全体での単語 t の頻度を tf(t,Ans(Q)) とする．ここで $Im(K_1)$ のクラスタの Web ページ内の単語 t の重要度を以下のように定める．

$$
w_t^{Im(K_1)} = tf(t, Im(K_1)) * idf(t) \tag{4}
$$

$$
idf(t) = log \frac{N}{df(t)} + 1 \tag{5}
$$

N は Ans(Q) の Web ページ数，df(t) は単語 t が出現する Web ページの数である．$Im(K_1)$ 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．

次に文単位での重要度を求める．文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．つまり，ある文 p に対する重要度は (4) 式を使い

$$
w_p^{Im(K_1)} = \sum_{w \in p} w_t^{Im(K_1)} \tag{6}
$$

とする．これによって Web ページ内の文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文を Web ページから抽出する．$Im(K_1)$ の各 Web ページから抽出したものをオブジェクトと呼び，$Obj(K_1)$ とする．すなわち，$Obj(K_1)$ とは，Ans(K_1,E_1) 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 8 に抽出の様子を表した．

図 8　Web ページからの画像，テキストの抽出

5.　検索結果の統合

4 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 $Im(K_1),\ldots,Im(K_n),Im(K_1 \wedge K_2),Im(K_1 \wedge K_3,),\ldots,Im(K_1 \wedge \ldots \wedge K_n)$ 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．

以下にコンテンツの構成について述べる．現在の考察段階では二つのの表示方法を考えている．一つ目は網羅的な表示方法．二つ目はカテゴリー的な表示方法である．

5.1　網羅的な表示

網羅的な表示とは全体をまとめて表示する，つまり各クラスタをすべて表示するという方法である．各オブジェクトの大きさを小さくして，その分，表示する数を増やす．これにより全体を眺めながら比較して各クラスタを閲覧することができる．その分，当然膨大な量になってしまい，ユーザは閲覧しにくいという問題点がある．

5.2　カテゴリー的な表示

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 9 のようにユーザが画像検索に使われるキーワードを，テキスト検索に使うことによって，条件をゆるめて，次のクラスタを表示していくという表示方法である．ユーザの興味に合わせてクラスタを移動できるが，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 9　網羅的な表示方法

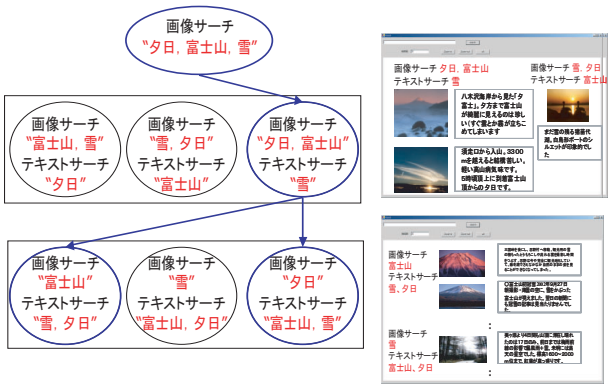

図 10　カテゴリー的な表示方法

6.　プロトタイプ

4 章，5 章で示した方法を基にプロトタイプを作成した．

まず，ユーザが複数のキーワード $K_1,K_2,\ldots,K_n$(n≧2) からなる And 検索を行うとする．すなわち，Q = $K_1 \wedge K_2 \wedge \ldots \wedge K_n$ である．次に，E_1 として Google 画像サーチエンジン，E_2 として Google 画像サーチで取得した画像の参照元の Web ページ内のテキストサーチを用いる．検索結果との Web ページ集合の和を解 Ans(Q) とする．この Ans(Q) を利用して，情報を統合させた新たなコンテンツを生成することによって，ユーザに質問 Q に対する検索結果を提示する．ここでコンテンツは html で書かれ，ユーザの要求に応じて動的にシステムが html ファイルを作成する．それを Web ブラウザが随時読み込むことによって，動的な統合を可能にした．

6.1　実 装 方 法

プロトタイプの実装に使用した環境は以下の通りである．

- Microsoft Visual Studio .Net/ C♯

プロトタイプの処理の流れは以下の通りである．

（1）　ユーザが入力した検索質問と検索数を読み取る．検索質問を空白を切れ目として単語ごとに切り分け，それをキーワード $K_1,K_2,\ldots,K_n$(n≧2) として保持する．

（2）　Google 画像検索エンジンに K_1 を入力しその結果の画像の URL と画像の参照元の URL を検索数分だけ収集してくる．これを $K_2,\ldots,K_n$ に関して繰り返し，すべての結果の URL を配列に格納する．

（3）　K_1 の結果として収集した画像の参照元の URL から得た html のソースコードからタグや Javascript 等を除いた，タイトル及び本文のテキスト文書を収集する．

（4）　キーワード $K_2,\ldots,K_n$ がすべてテキスト文書内に入っているかを調べ，入っているならテキスト文書を格納し，そうでないなら破棄する．

（5）　$K_2,\ldots,K_n,K_1 \wedge K_2,\ldots,K_1 \wedge \ldots \wedge K_n$ に関して繰り返し，すべての画像の参照元の URL について行う．

（6）　各テキスト文書に対して tf 値を調べるため茶筅 [3] を

図 11　実 装 画 面

用いて文を単語単位に分ける．4.4 節で用いた方法により単語の重要度を計算し Web ページにランク付けをする．

（7） ランクの高い Web ページから重要度が高い文を抽出し，Top layer として $Im(K_1), \ldots, Im(K_n)$ のオブジェクトを表示する．

（8） ユーザの興味によってコンテンツの機能を使用することによって，コンテンツを動的に変化させる．

6.2 考　　察

実装によって得られた本研究の考察を以下に述べる．

- テキスト，画像を抽出してくることによって，必要最小限の情報を効率よく手に入れることができたと考えられる．しかし，抽出してくる分 Web ページらしさが損なわれていると感じられる．
- 本研究では利用可能なサーチエンジンとして Google 画像サーチを最初に用い，その検索結果の Web ページ内に他の検索キーワードがあるかどうかによって，解集合 Ans(Q) を決めていた．しかし，これでは画像主体の検索となってしまいある単語の意味が知りたい，というようなテキスト主体のユーザの要求に応じることができない．
- 検索キーワードを分解することによって最初の検索結果が非常に多くヒットしてしまう．このため，ページの有用性を判断するために各 Web ページのテキスト文書を取得する際に検索結果すべての Web ページを収集してくるため各 Web ページへのアクセス時間が多くかかってしまい．検索単語や検索数を多く指定すればするほど，非常に効率が悪くなってしまう．
- 検索結果の Web ページが有用かどうか判断するのに，画像の参照元の Web ページ内のテキスト文書に検索キーワードがすべて入っているかどうかで判断している．検索キーワードが多数の場合，精度は高まるが，条件が厳しいため検索結果が出てこない場合がある．
- Web ページから関連部分を抜き出すときに単語の出現頻度に関する式でしか出していない．そのため，Web ページ内の他の画像等のテキスト情報以外の情報が無視されてしまう．

6.3 課　題　点

考察をふまえた上で改善していくべき課題点を以下に挙げた．

- システムの改良
  - 利用可能なサーチエンジンの増加

最初に質問を入力するサーチエンジンを増加させる．本研究では最初に使用するサーチエンジンとして Google 画像検索エンジンを使用しただけである．マルチメディア・メタサーチのため様々なメディアに対するサーチエンジンを使用しなくてはならない．かつ，それらの検索結果をうまく統合することが必要である．

  - 検索キーワードの修正

フィードバック制御を取り入れ，検索結果の情報，またそれに対するユーザの振る舞いによって，ユーザの質問をシステムが自動的に修正することでより最適な検索結果をユーザに提示できるようにしていく．

  - データベースの作成

メタサーチでは各サーチエンジンの処理の速さや，検索結果の Web ページのアクセスの状況によって，システムの処理が非常に遅くなってしまう．このため，様々なマルチメディアのデータをデータベースに蓄えることが重要であると考える．データベースを持つことによりメタサーチではなくなるが，さらなる速度の向上が期待される．

- データの改良
  - 抽出方法

本研究では Web ページからテキストを抽出する際にタグ情報は無視して，純粋なテキスト文しか抽出していない．しかし，タグやそれによる文章構造は重要なものである．よって，Web から文書だけを抜き出すだけでなく画像検索で取得した画像以外の画像や，またリンク等も考慮に入れて親ページやリンク先の情報も提示いくことが重要である．

  - 重要度の算出

単に単語の出現頻度だけでなく文脈を意識したり，文の位置情報なども考慮に入れて重要度を算出していくことが必要であると考えられる．

- システムの評価方法

本方式では，他の方式のような検索結果に URL リストが出力され，それらを訪れ Web ページを閲覧するという手間が省かれ，検索キーワードに関連する部分だけ簡単に閲覧することができる．このシステムの評価の尺度としては，再現率，適合率を用いることが考えられる．関連部分だけの統合という手法を用いているため，従来の検索システムにおけるユーザの有用な情報が得られるまでの振る舞い，つまり Web ページを閲覧した時間や，そこにたどり着くまでの手間を省くことができる．このためユーザの振る舞いを数値化したものも評価の尺度として取り入れるべきだと考える．

7. 結　　論

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くことが重要であると思われる．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きたい．

8. 謝　　辞

本研究の一部は，平成 14 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：14019048，代表：田中克己）および基盤研究 (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・呈示と放送コンテンツ生成」（課題番号：14208036，代表：田中克己）および平成 14 年度 NEC 共同研究「クロスメディア情報流通システムにおける情報メディアの活性化の研究」（代表：田中克己）による．ここに記して謝意を表します．

9. 参 考 文 献

文　　献

[1] 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)
[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/