データベースシステム 125−31
（２００１．７．１８）

情報通信分野を対象とした意味的連想検索機構による WWW 検索エンジンの実現

大橋 英博† 清木 康†

本稿では情報通信分野を対象とした意味的連想検索機構による WWW 検索エンジンの実現方法を提案する．本方式の特徴は，用語辞典を元にしたメタデータ空間の半自動生成及び，Web 上から取得した HTML ファイル群を元にした検索対象メタデータの自動生成を可能とする点にある．本方式により，意味的連想検索機構を WWW 検索エンジンに適用する際に必要となるメタデータ空間及び検索対象メタデータの効率的な作成が可能となる．実際の HTML ファイルを用いた実験結果を示し，提案方式の実現可能性および有効性を確認する．

An Implementation Method of WWW Search Engine by A Semantic Associative Search Mechanism in Information and Communication Fields

Hidehiro Ohashi† and Yasushi Kiyoki†

In this paper we present a method for realizing WWW search engine by a semantic associative search mechanism in information and communication fields. The main feature of the method is to create a metadata space semi-automatically based on dictionaries and to extract metadata of retrieval candidates automatically from HTML files collected on the Web. By using this method, we can efficiently create a metadata space and extract metadata of retrieval candidates which are necessary for applying the semantic associative search system to WWW search engine. We clarify feasibility and effectiveness of the method by showing several experimental results using actual HTML files.

1. はじめに

現在の Web 検索エンジンではパターンマッチングによる検索方式が主流である．この方式では，利用者自身は自ら求める Web ページを絞り込める検索キーワードを知っていることを必要とされる．しかし，現実には利用者が，適切な検索キーワードを知っているケースは少ない．検索結果を十分に絞り込めない抽象的な検索キーワードを指定した場合，利用者は大量の検索結果を得ることになる．その場合には，さらに検索キーワードを指定して絞り込んでいく必要がある．逆に，具体的過ぎる検索キーワードを指定した場合は得られる検索結果は少なくなり，利用者が求める Web ページがその中に含まれる可能性は少なくなる．このようにパターンマッチングによる検索方式で適切な結果を得るためには，利用者は適切な検索キーワードを入力しなければならない．検索エンジンにおいて，意味的解釈を伴う検索が実現されれば，検索者へ高度な情報獲得の機会を提供することが可能となる．本論文では情報通信分野を対象とした意味的連想検索機構 [1)~4)] による WWW 検索エンジンの実現方法を提案する．本方式の特徴は用語辞典を元にしたメタデータ空間の半自動生成及び，Web 上から取得した HTML ファイル群を元にした検索対象メタデータの自動生成を可能とする点にある．本方式により，意味的連想検索機構 [1)~4)] を WWW 検索エンジンに適用する際に必要となるメタデータ空間及び検索対象メタデータの効率的な作成が可能となる．実際の HTML ファイルを用いた実験結果を示し，提案方式の実現可能性および有効性を確認する．本方式では Web 検索エンジンによる検索に意味の数学モデルを適用することにより，利用者が必ずしも適切な検索キーワードを知らない場合でも，それに意味的に近いキーワードを知っていれば，そのキーワードに対して意味的に近い Web ページを獲得可能とする方式を提案する．

† 慶應義塾大学 環境情報学部
Faculty of Environmental Information, Keio University

2. 意味的連想検索方式

ここでは，意味的連想検索方式ついて概説する．詳細は，文献 [1)~4)] に述べられている．

2.1 メタデータ空間 $\mathcal{MDS}$ の設定

初めに，m 個の基本データについて各々n 個の特徴（$\mathrm{f}_1,\mathrm{f}_2,\cdots,\mathrm{f}_n$）を列挙した特徴付ベクトル $\mathbf{d}_i(i =$

$1, \cdots, m$) が与えられているものとし，そのベクトルを並べて構成する $m \times n$ 行列を M とおく（図 1）．このとき，M は，列ごとに 2 ノルムで正規化されている．

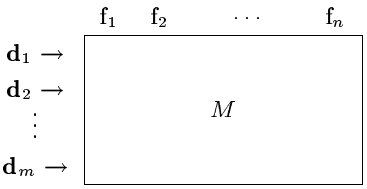

図 1　データ行列 M によるメタデータの表現

(1)　データ行列 M の相関行列 M^TM を計算する．

(2)　M^TM を固有値分解する．

$$M^TM = Q \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_\nu & & \\ & & & 0 & \\ & & & & \ddots 0 \end{pmatrix} Q^T,$$

$0 \leq \nu \leq n$.

ここで行列 Q は，

$$Q = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n)$$

である．この $\mathbf{q}_i (i = 1, \cdots, n)$ は，相関行列の正規化された固有ベクトル（以下，“意味素”）である．相関行列の対称性から，この固有値は全て実数であり，その固有ベクトルは互いに直交している．

(3)　メタデータ空間 $\mathcal{MDS}$ を以下で定義する．
非ゼロ固有値に対応する固有ベクトル（以下，“意味素” と呼ぶ）によって形成される正規直交空間をメタデータ空間 $\mathcal{MDS}$ と定義する．この空間の次元 ν は，データ行列のランクに一致する．この空間は，ν 次元ユークリッド空間となる．

$$\mathcal{MDS} := span(\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_\nu).$$

$\{\mathbf{q}_1, \cdots, \mathbf{q}_\nu\}$ は $\mathcal{MDS}$ の正規直交基底である．

2.2　メディアデータのメディアデータベクトルの作成方式

ここでは，メディアデータを表現するメディアデータベクトルを形成する方法を示す．

(1)　**Step-1:** メディアデータの特徴づけ
t 個の印象語（あるいは，t 個のオブジェクト）$\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_t$ から成るメディアデータ P を次のように特徴づける．

$$P = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_t\}.$$

ここで，各印象語 $\mathbf{o}_i$ は，データ行列の特徴と同一の特徴を用いて表現される特徴付ベクトルである．

$$\mathbf{o}_i = (o_{i1}, o_{i2}, \cdots, o_{in})$$

(2)　**Step-2:** メディアデータ P のベクトル表現
メディアデータ P を構成する t 個の印象語 $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t$ が，それぞれ n 次元ベクトルで定義されている．オブジェクト $\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_t$ の和演算子 $\oplus$ を次のように定義し，メディアデータのメディアデータベクトル $\mathbf{p}$ を形成する．

$$\mathbf{p} = \bigoplus_{i=1}^{t} \mathbf{o}_i := (\mathrm{sign}(o_{\ell_1 1}) \max_{1 \leq i \leq t} |o_{i1}|,$$
$$\mathrm{sign}(o_{\ell_2 2}) \max_{1 \leq i \leq t} |o_{i2}|,$$
$$\cdots, \mathrm{sign}(o_{\ell_n n}) \max_{1 \leq i \leq t} |o_{in}|).$$

この和演算子 $\bigoplus_{i=1}^{t}$ は，t 個のベクトルから各基底に対して絶対値最大の成分を選ぶ演算子である．
ここで sign(a) は，“a” の符号（正，負）を表す．また，$l_k (k = 1, \ldots, t)$ は，特徴が最大となる印象語を示す指標であり，次のように定義する．

$$\max_{1 \leq i \leq t} |o_{ik}| = |o_{\ell_k k}|.$$

2.3　意味射影集合 Π_ν の設定

メタデータ空間 $\mathcal{MDS}$ から固有部分空間（以下，意味空間）への射影（以下，“意味射影”）の集合 Π_ν を考える．P_{λ_i} を次の様に定義する．

$P_{\lambda_i} := \lambda_i$ に対応する固有空間への射影
$i.e.\ P_{\lambda_i} : \mathcal{MDS} \rightarrow span(\mathbf{q}_i)$.

意味射影の集合 Π_ν を次のように定義する．

$$\Pi_\nu := \{\ 0\ , P_{\lambda_1}, P_{\lambda_2}, \cdots, P_{\lambda_\nu},$$
$$P_{\lambda_1} + P_{\lambda_2}, P_{\lambda_1} + P_{\lambda_3}, \cdots, P_{\lambda_{\nu-1}} + P_{\lambda_\nu},$$
$$\vdots$$
$$P_{\lambda_1} + P_{\lambda_2} + \cdots + P_{\lambda_\nu}\}.$$

i 次元の意味空間は，$\frac{\nu(\nu-1)\cdots(\nu-i+1)}{i!}, (i = 1, 2, \cdots, \nu)$ 個存在するので，射影の総数は，2^ν となる．つまり，このモデルは，2^ν 通りの意味の様相の表現能力をもつ．

2.4　意味解釈オペレータ $\mathbf{S}_p$ の構成

検索者の印象やメディアデータの内容を与える文脈を表す ℓ 個の検索語列

$$s_\ell = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_\ell)$$

と，しきい値 $\varepsilon_s (0 < \varepsilon_s < 1)$ が与えられたとき，それに応じた，意味射影 $P_{\varepsilon_s}(s_\ell)$ を構成するオペレータ（以下，“意味解釈オペレータ”）S_p を構成する．T_ℓ を長さ ℓ の検索語列の集合とすると，S_p は，次のように定義される．

$S_p : T_\ell \longmapsto \Pi_\nu$
ここで，$T_\ell \ni s_\ell, \Pi_\nu \ni P_{\varepsilon_s}(s_\ell)$.

また，$\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_\ell\}$ の各要素は，特徴付ベクトルであり，データ行列 M の特徴と同一の特徴を用いて表される．

オペレータ S_p は以下の計算を行う．

(1) $\mathbf{u}_i(i=1,2,\cdots,\ell)$ をフーリエ展開する．
検索語列 s_ℓ を構成する ℓ 個の検索語を各々メタデータ空間 $\mathcal{MDS}$ へ写像する．
この写像では，ℓ 個の単語を各々メタデータ空間 $\mathcal{MDS}$ 内でフーリエ展開し，フーリエ係数を求める．これは，各検索語と各意味素の相関を求めることに相当する．
$\mathbf{u}_i$ と $\mathbf{q}_j$ の内積 u_{ij} は次のようになる．

$$u_{ij} := (\mathbf{u}_i, \mathbf{q}_j), \quad j = 1,2,\cdots,\nu.$$

ベクトル $\widehat{\mathbf{u}}_i \in \mathcal{MDS}$ を次のように定める．

$$\widehat{\mathbf{u}}_i := (u_{i1}, u_{i2}, \cdots, u_{i\nu}).$$

これは，単語 $\mathbf{u}_i$ をメタデータ空間 $\mathcal{MDS}$ に写像したものである．

(2) 検索語列 s_ℓ の意味重心 $\mathbf{G}^+(s_\ell)$ を求める．
まず，各意味素ごとに，フーリエ係数の総和を求める．これは，検索語列 s_ℓ と各意味素との相関を求めることに相当する．このベクトルは，ν 個の意味素があるため，ν 次元ベクトルとなる．このベクトルを，無限大ノルムによって正規化したベクトルを，以下，検索語列 s_ℓ の意味重心 $G^+(s_\ell)$ と呼ぶ．

$$\mathbf{G}^+(s_\ell) := \frac{\left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right)}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \|_\infty}.$$

ここで，$\|\cdot\|_\infty$ は無限大ノルムを示す．

(3) 意味射影 $P_{\varepsilon_s}(s_\ell)$ を決定する．
検索語列 s_ℓ の意味重心を構成する各要素において，しきい値 ε_s を越える要素に対応する意味素を，メディアデータのメタデータを射影する意味空間の構成に用いる．意味射影 $P_{\varepsilon_s}(s_\ell)$ を次のように決定する．

$$P_{\varepsilon_s}(s_\ell) := \sum_{i \in \Lambda_{\varepsilon_s}} P_{\lambda_i} \quad \in \Pi_\nu.$$

ただし $\Lambda_{\varepsilon_s} := \{\ i \mid |\left(\mathbf{G}^+(s_\ell)\right)_i| > \varepsilon_s\}$ とする．

2.5 意味空間における相関の定量化

文脈（文脈を表す検索語列）を対象として，2.4 節で示したオペレータ S_p を用いて選択された意味空間（部分空間）上で，その文脈に対応したメディアデータを選び出す意味的連想検索方式を示す．

メタデータ空間に写像されたメディアデータ群に対応する各ベクトル（メディアデータベクトル）について，選択された意味空間（部分空間）上におけるノルムを求め，文脈に相関の強いメディアデータの検索を行う．意味空間におけるメディアデータベクトルのノルムの大きさをその文脈とメディアデータとの相関の強さとする．

文脈 s_ℓ が与えられた場合のメディアデータ $\mathbf{x}$ のノルム $\rho(\mathbf{x}; s_\ell)$ を次のように定める．

$$\rho(\mathbf{x}; s_\ell) = \frac{\sqrt{\sum_{j \in \Lambda_{\varepsilon_s} \cap \mathcal{S}} \{c_j(s_\ell) x_j\}^2}}{\|\mathbf{x}\|_2},$$

$$\mathcal{S} = \{i | sign(c_i(s_\ell)) = sign(x_i)\},$$

$$c_j(s_\ell) := \frac{\sum_{i=1}^{\ell} u_{ij}}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \|_\infty},$$

$j \in \Lambda_{\varepsilon_s}$.

ここで，意味空間を構成する意味素（固有ベクトル）群において，文脈に関係しているのは，正と負のどちらか一方である．そこで，意味空間を構成する意味素の符合を考慮するため，意味空間を構成する意味素の符合と正負が逆の成分についてはノルムの計算において無視している．

また，メディアデータを特徴づける特徴の数が多いと，どのような意味空間が選ばれても，意味空間におけるメディアデータのノルムが大きくなる傾向がある．そのため，本来，文脈との相関が強いと考えられるメディアデータベクトルのノルムよりも，特徴の数が多いメディアデータベクトルのノルムが大きくなってしまい，適切な抽出が行われないことがある．そのため，メタデータ空間でのメディアデータベクトルを 2 ノルムで正規化している．

3. メタデータ空間および検索対象メタデータの自動生成方式

本節では意味的連想検索機構を用いた WWW 検索エンジンの実現のためのメタデータ空間および検索対象メタデータの自動生成方式を示す．

3.1 本方式の概要

本方式の概要を図 2 に示す．まず，メタデータ空間の半自動生成方式について説明する．メタデータ空間は用語辞典を元に作成する．用語辞典の各用語を見出し語とし，それを説明する特徴語を用語を説明する説明文から抽出する．特徴語とは見出し語を説明するための単語群である．特徴語の抽出には形態素解析処理により行う．次に検索対象メタデータの自動生成方式について説明する．検索対象メタデータは HTML ファイル等のドキュメントを元に生成する．ドキュメントに対して形態素解析を行い，既に作成されたメタデータ空間を参照して検索対象メタデータを生成する．

本方式を用いた WWW 検索エンジンの実現の手順を以下に示す．まず，(1) 一般の用語辞典から形態素解析処理を用いてメタデータ空間を作成する．次に，(2) 独自のロボットプログラムをによって Web 上から HTML ファイルを収集する．(3) 収集した HTML ファイルと先にメタデータ空間を使用して検索対象メタデータを作成する．(4) 作成したメタデータ空間と

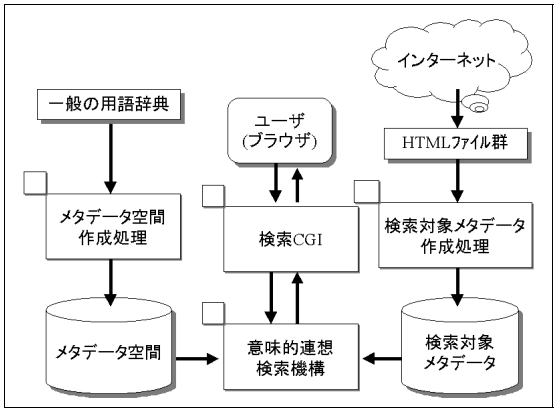

図 2　本方式の概要

検索対象メタデータに対して意味的連想検索機構を適用して Web ブラウザから意味的連想検索を実行する．

3.2　メタデータ空間の半自動生成方法

意味的連想検索機構に使用するメタデータ空間の半自動生成方法について述べる．

3.2.1　用語辞典を元にしたテキストデータの作成

まず，一般の用語辞典の中の用語と説明文の組をテキストファイルの形式で保存する．今回は情報通信分野を対象としたメタデータ空間を作成するために，例として文献 [5] を参照する．図 3 に辞典を元にしたテキストデータを示す．

No.	用語	説明文
1	アイコン	データ・ファイルやソフトウェアの個々の機能を示す小さな図形。GUI環境で用いられる。
2	アイテニアム	米Intel 初の64ビットCPUとして2000年に出荷されるCPU。
3	あいまいさ	あいまいな値を使う論理演算に関する理論。1965年に米カリフォルニア大学のL. A. ザデー教授が提唱した「ファジー集合論」に端を発している。
4	アウトソーシング	ユーザ企業が情報システムの構築や運用を社内のシステム部門から外部の専門業者に委託する。
5	アウトライン・フォント	文字の形状を輪郭線で表現したフォント。輪郭を構成する要所要所の点を座標データとして持ち，各点をスプライン曲線やベジェ曲線で補間しながら結ぶことで文字の形を表現する。
6	アクセス回線	通信事業者の（基幹）ネットワークとユーザーを結ぶ回線。通信サービスの提供に不可欠な回線。通常は最寄の局を結ぶ回線を指す。

図 3　辞典を元にしたテキストデータ

3.2.2　用語の説明文に対する形態素解析処理

次に用語の説明文に対して形態素解析処理を行う．形態素解析には Breakfast [8] を使用した．図 4 に形態素解析処理を行った結果を示す．

No.	用語	説明文
1	アイコン	データ ・ ファイル ソフトウェア 個々 機能 図形 GUI 環境
2	アイテニアム	米 Intel 64 ビット CPU 2000 年 出荷 CPU
3	あいまいさ	値 論理 演算 理論 1965 年 米 カリフォルニア 大学 A ザデー 教授 提唱 ファジー 集合
4	アウトソーシング	ユーザ 企業 情報システム 構築 運用 社内 システム 部門 外部 専門 業者 委託
5	アウトライン・フォント	文字 形状 輪郭 線 表現 フォント 輪郭 構成 要所 要所 点 座標データ 点 スプライン 曲線 ベジェ曲線 補間 こと 文字 形 表現
6	アクセス回線	通信 事業 基幹 ネットワーク ユーザー 回線 通信 サービス 提供 回線 通常 最寄 局 回線

図 4　形態素解析を行った結果

この処理により説明文を形態素単位に分割し，名詞相当の単語を抽出する．基本的にこれらの単語が用語を説明する特徴語となる．これらの単語をそのまま特徴語とすると，用語の数より特徴語の数が多くなる．特徴語を用語辞典中の重要度順にソートし，上位から用語数分だけ取り出し，これを特徴語とする．重要度は出現頻度や TF・IDF [6),7)] 等により決定する．この処理を図 5 に示す．これにより，最終的に作成されるメタデータ空間は用語数分の行と列を持つ行列となる．

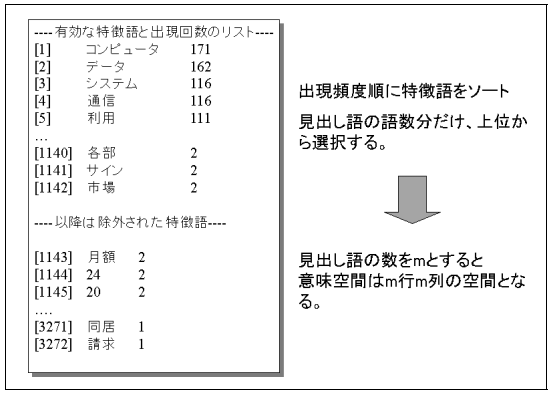

図 5　特徴語決定のための処理

3.2.3　分野によるメタデータ空間の分割

上記のプロセスにより，情報・通信分野の用語を説明するメタデータ空間を作成した．この空間を情報分野と通信分野の２つのメタデータ空間に分割する方法を示す．分割によって得られた空間は 4.3 節で使用する．前項で作成したメタデータ空間の各用語に対して情報分野に関する用語か，通信分野に関する用語かを示す識別子を設定する．この識別子を元に２つのメタデータ空間に分割する (図 6)．これらの空間を使用するにより，情報分野・通信分野それぞれに対する意味的検索が可能となる．

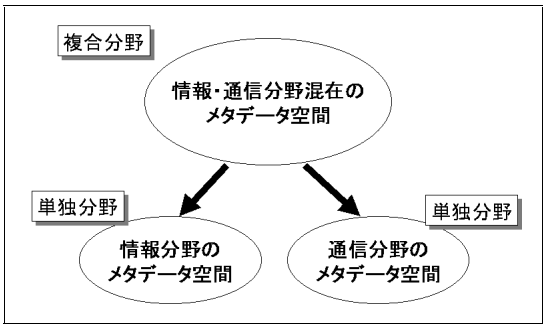

図 6　空間の分割

3.3　検索対象メタデータの自動生成方法

意味的連想検索機構で使用する検索対象メタデータの自動生成方法について説明する．

3.3.1 HTML データに対する文字コード変換及びテキスト抽出処理

検索対象メタデータは独自のロボットプログラムで収集した HTML データを対象として作成する．はじめに，収集した HTML データを統一した文字コードに変換する．文字コード変換ツールである nkf32 を使用して，全ての HTML データの文字コードをシフト JIS コードに変換する．次に，HTML データからテキスト部分のみを抽出する．抽出対象タグは，TITLE タグと BODY タグとし，これらのタグで囲まれたデータを抽出する．その中にタグが含まれる場合は取り除く．これらのプロセスを図 7 に示す．全ての HTML ファイルに対してこれらのプロセスを実行し，テキストデータを抽出する．

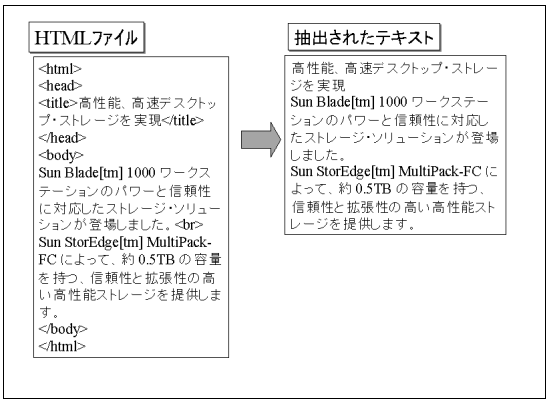

図 7　HTML からのテキストの抽出

3.3.2 テキストデータに対する形態素解析処理

次に，抽出したテキストデータに対して形態素解析処理を実行する．形態素解析処理は 3.2.2 節で使用した Breakfast[8] を使用する．形態素解析を行った結果を図 8 に示す．

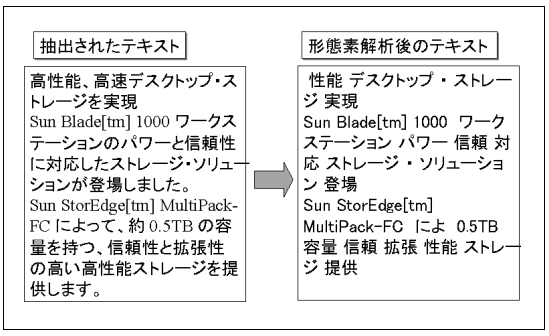

図 8　形態素解析処理の結果

3.3.3 URL に対する見出し語割り当て処理

形態素解析処理が行われたテキストデータと 3.2 節で作成した３つのメタデータ空間，つまり情報･通信分野，情報分野，通信分野を使い，それぞれのメタデータ空間に対応した検索対象メタデータを作成する．

表 1　メタデータ空間と検索対象メタデータ

メタデータ空間	検索対象メタデータ
情報・通信分野	情報・通信分野のメタデータ空間を元に作成した検索対象メタデータ
情報分野	情報分野のメタデータ空間を元に作成した検索対象メタデータ
通信分野	通信分野のメタデータ空間を元に作成した検索対象メタデータ

検索対象メタデータはテキストデータの URL に対してメタデータ空間の見出し語を割り当てることで作成する．テキストデータに対する見出し語の割り当ては以下の２つのルールにしたがって行う．

ルール１：分割された単語群に見出し語が存在した場合，その HTML データの URL に存在した見出し語を割り当てる

ルール２：分割された単語群の中に見出し語が存在しない場合，見出し語を定義する特徴語を一定の割合（ここでの一定の割合を閾値と呼ぶ）以上含んでいた場合，その見出し語を URL に割り当てる

前項の段階で HTML ファイルから抽出されたテキストデータは単語群に分割されている．この単語群の中にメタデータ空間で定義されている見出し語が含まれていた場合はルール１に該当し，当該テキストデータの URL にその見出し語が割り当てる．単語群の中に見出し語が含まれていない場合，ルール２が該当しているかどうかを調べる．見出し語は一語以上の特徴によって説明されている．テキストデータの単語群の中に，見出し語を説明している特徴語を閾値以上含んでいる場合はルール２に該当し，その場合も当該テキストデータの URL に見出し語を割り当てる．ここでの一定の割合を閾値と呼ぶ．閾値は検索対象メタデータ作成プログラムにパラメータとして渡される．図 9 にルール２に該当する例を示す．この２つのルールにより検索対象メタデータを作成する．

4. 実験と考察

本方式の有効性を検証するための実験方法を示す．また，その実験結果に関する考察を述べる．

4.1 実験環境と方法

実験環境とその方法について説明する．

4.1.1 対 象 分 野

本論文の対象分野としては Web 上になんらかの情報が公開されている情報・通信分野とした．情報通信分野は他の分野と比べて比較的 Web 上の情報量が多く，また専門用語が多いため本方式による WWW 検

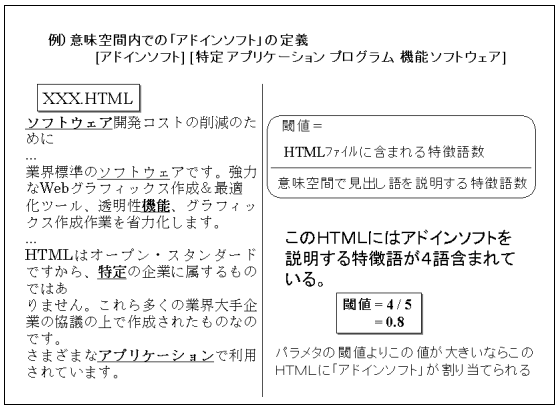

図 9　検索対象メタデータの作成

索エンジンがより効果的であると考えられる．

4.1.2　対象 HTML データの収集

WWW 検索エンジンの検索対象となる HTML データは，独自のロボットプログラムを使用し，jp ドメイン配下の HTML データを収集した．収集した HTML データは，約 5 万件である．

4.1.3　実験環境

実験は次のような環境で行う．まず，メタデータ空間の作成には文献[5]の中の 1000 語を対象として３つのメタデータ空間を作成した．

表 2　メタデータ空間一覧

No.	分野	説明
1	情報・通信分野	情報通信の両分野を対象とした空間
2	情報分野	情報分野のみを対象とした空間
3	通信分野	通信分野のみを対象とした空間

検索対象メタデータはメタデータ空間に対応して自動作成されるため，３つのメタデータ空間のそれぞれに対応した検索対象データを自動作成した．また情報・通信分野の検索対象メタデータについては実験１で手作業で作成した場合と自動で作成した場合との比較を行うために，収集した 5 万件の HTML ファイルのうち 20 件を選択して，手作業で URL に対して見出し語を割り当てることにより，検索対象メタデータを作成した．

表 3　検索対象メタデータ一覧

No.	分野	説明
1	情報・通信分野	自動生成された情報通信の両分野を対象とした検索対象メタデータ
2	情報・通信分野	情報通信の両分野を対象とした空間
3	情報分野	情報分野のみを対象とした空間
4	通信分野	通信分野のみを対象とした空間

4.1.4　実験方法

実験１は，検索対象メタデータを手作業で作成した場合と，本研究で提案している方式で生成した場合で適合率・再現率の変化を調べる．適合率と再現率は以下の計算式で算出する．

表 4　適合率・再現率の計算式

適合率 ＝ (検索結果として得られた中の正解数)/(正解数)
再現率 ＝ (検索結果として得られた中の正解数)/(検索結果数)

これらの値は 0.0〜1.0 までの値をとり，1 に近いほど良い検索結果が得られていることを示す．検索結果の正解の集合である正解セットはロボットプログラムにより収集した HTML ファイルから 10 件選んで作成した．　実験２では情報・通信の２つの分野が混在している空間を使用した場合と，その空間を情報分野と通信分野の２つの分野に分割した場合で再現率と適合率がどのように変化するかを調べた．正解セットは情報分野と通信分野それぞれに 10 件の HTML ファイルを選んで作成した．

4.2　実験結果

4.2.1　実験１：手作成した検索対象メタデータを使用した場合と，自動生成した検索対象メタデータを使用した場合の比較

ここでは実験１の結果について述べる．実験１の結果を以下の図 10 に示す．

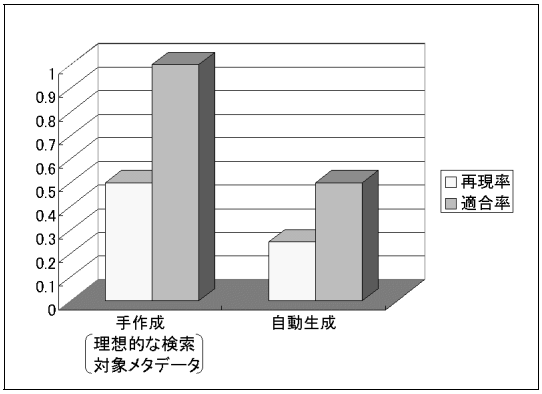

図 10　実験結果 1

この結果から検索対象メタデータを自動生成した場合，手作成した理想的な検索対象メタデータを使用した場合に比べて適合率・再現率ともに低下していることが分かる．

4.3　実験２：情報・通信を１つの空間とした場合と，情報と通信をそれぞれ別の空間に別けた場合の比較

ここでは実験２の結果について述べる．以下の図ではメタデータ空間を情報と通信の２つの分割した場合の適合率と再現率の変化を示している．

図 11 より情報分野のみの空間を使用した方が情報通信混合の空間を使用した場合に比べ再現率，適合率ともに向上している．同様に図 12 より通信分野につ

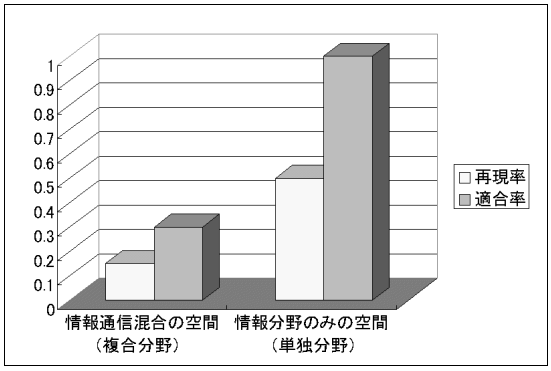

図 11　実験結果 2-1

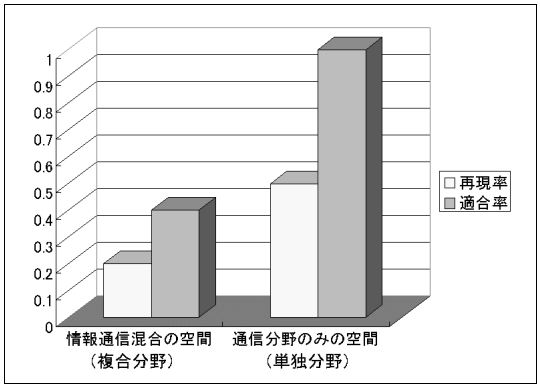

図 12　実験結果 2-2

いても，通信分野のみの空間を使用した方が情報通信混合の空間を使用した場合に比べ，再現率，適合率ともに向上していることがわかる．これらの結果より，メタデータ空間は複数の分野の情報を含んでいる場合より，より限定された分野の情報だけを対象とする場合の方が適合率や再現率が向上することが分かることがわかる．

4.4 考　　察

4.4.1 実験１の考察　（検索対象メタデータを自動生成したことによる検索精度の変化について）

実験１の結果から得られた考察について述べる．実験 1 の結果から，本方式で自動生成した検索対象メタデータを使用し場合，人手で作成した検索対象メタデータを使用する場合に比べて，適合率や再現率は低下する．本方式では HTML ファイルのテキスト部分に形態素解析処理を施して単語単位に分割し，それらの単語群とメタデータ空間内の見出し語や特徴語と比較してメタデータを付与しているが，HTML ファイルにおける各単語の意味が全て同列に扱われているため，単語間の重要度の違いが反映されていない．そのために人手により作成した場合と比べて適合率や再現率が低下しているものと思われる．しかし，WWW 検索エンジンで意味的連想検索機構を使用する場合は，大量の HTML ファイルをもとにして検索対象メタデータを作成する必要があり，効率的に検索対象メタデータの作成が行えるような仕組みが必要である．本方式は完全に自動で検索対象メタデータを作成するものであり，WWW 検索エンジンに意味的連想検索機構を適用する場合に有効な方式であると考えられる．

4.4.2 実験２の考察　（メタデータ空間を分割したことによる検索精度の向上について）

実験２の結果から得られた考察について述べる．実験２の結果から，メタデータ空間は複数の分野の情報が混在している場合に比べて，より限定された分野の情報を対象とした場合の方が適合率や再現率を向上させることが出来ると考えられる．今回使用した用語辞典は情報と通信の２分野をカバーする辞典であるため，そのままメタデータ空間を作成すると２つの分野の情報を持つ空間になる．このような場合は各用語がどちらの分野を説明する用語かを人手により判断し，それぞれの分野に振り分けて一分野一空間になるように空間を作成する．こうすることで２つの分野が混在したメタデータ空間を使用する場合と比べて高い再現率や適合率を得ることが出来るようになる．本論文ではメタデータ空間を半自動で生成する方式を提案したが，半自動生成したメタデータ空間でも一分野一空間として空間を生成することで十分な検索精度が得られる．

5. 結　　論

実験の結果より検索対象メタデータを自動生成した場合，人手で作成した場合と比べて検索精度が若干落ちるが，大量のデータに対して効率良くメタデータを作成することが出来る．また２つの分野を１つの空間として作るより，より限定した分野を対象として空間を作ることで検索精度を向上させることが出来る．以上のことから大量の検索対象メタデータを作成する必要がある場合は自動生成を行うことが適切であるといえる．またメタデータ空間はある程度人手が介在することになっても，一分野一空間として作成することが望ましい．１つの分野に限った空間であれば本論文で提案した形態素解析によるメタデータ空間の半自動生成方式でも十分な検索精度が得られる．本研究によって意味的解釈を伴う検索が実現でき，検索者へ高度な情報獲得の機会を提供することが可能となったと考えられる．また，メタデータ空間作成の半自動化および検索対象メタデータ作成の自動化を実現した．これにより意味的連想検索機構を実際の WWW 検索エンジンに適用することが容易になったと考えられる．

参 考 文 献

1) Kitagawa, T. and Kiyoki, Y.: The mathematical model of meaning and its application to multidatabase systems, Proceedings of

3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp.130-135, April 1993.

2) Kiyoki, Y., Kitagawa, T. and Hayama, T.: A metadatabase system for semantic image search by a mathematical model of meaning, ACM SIGMOD Record, vol. 23, no. 4, pp.34-41, 1994.

3) Kiyoki, Y.,Kitagawa, T. and Hitomi, Y.: A fundamental framework for realizing semantic interoperability in a multidatabase environment, Journal of Integrated Computer-Aided Engineering, Vol.2, No.1, pp.3-20, John Wiley & Sons, Jan. 1995.

4) 清木 康, 金子 昌史, 北川 高嗣: 意味の数学モデルによる画像データベース探索方式とその学習機構, 電子情報通信学会論文誌,D-II,Vol.J79-D-II,No.4,pp.509-519, 1996.

5) 日経 BP 社: 情報 · 通信新語辞典　2001 年版

6) Salton, G. and Buckley, C. : "Term-weighting approaches in automatic text retrieval." Information Processing and Management, 24, pp.513-523, 1998d

7) Salton and Buckley, 1990 Salton, G. and Buckley, C. : "Improving retrieval performance by relevance feedback." Journal of the American Society for Information Science, 41(4), pp.228-297, 1990.

8) 形態素解析プログラム　Breakfast　株式会社富士通研究所　颯々野学氏