論文番号:C2-1

単語の出現頻度を用いた ドキュメントデータベースのメタデータ自動生成方式

河本穣 † 図子泰三 ‡ 清木康 ††

† 慶應義塾大学 総合政策学部

‡ 慶應義塾大学 大学院政策・メディア研究科

†† 慶應義塾大学 環境情報学部

概要

本稿では，ドキュメントデータベースを対象とし，単語の出現頻度を反映するアルゴリズム (tf×idf[4, 5]) を適用したメタデータ自動生成方式を提案する．本方式によって生成されるメタデータ群を，ベクトル計算により意味的相関を計量するドキュメントデータベース検索システムに適用する方法について示す．意味的な計量が可能なデータベース検索システムでは，情報検索における同義性，類似性の判定に，ドキュメントの内容を端的に説明した適切な数のメタデータが必要である．このようなメタデータを人間が生成するオーバーヘッドは非常に大きく，大規模なドキュメントデータベースを対象とした場合には大きな障害となる．また，人間により記述されるメタデータ間では，一貫性を欠きやすいという欠点がある．メタデータ間での一貫性の欠如は，実際にシステムを利用する場合において検索結果を偏ったものとする原因となる．ドキュメント群を対象としたメタデータ群は，一貫性のある客観的な基準で，自動的な処理によって生成される必要がある．本稿では，ドキュメント内での単語の出現頻度を用いたアルゴリズム (tf×idf) を用いて出現単語の重要度を反映したメタデータ自動生成方式について示す．また，生成されたメタデータ群を対象とした意味的ドキュメントデータベース検索に関する実験を行った結果を示し，提案するメタデータ自動生成方式の有効性を検証する．実験では，サンプルとして，比較的小規模なドキュメントデータベースを対象とし，それらのドキュメント群について，人間が一貫性を欠くことなく記述したメタデータ群を用いる方法との比較において，提案方式は，その方法と同程度の妥当性をもって自動的にメタデータ生成を実現可能であることを示す．この実験により，メタデータ自動生成方式の有効性を確認し，より大規模なドキュメントデータベース群への適用可能性を示す．

keywords: ドキュメントデータベース，メタデータ，自動生成，出現頻度, tf×idf

1 はじめに

現在，電子的なメディアで利用可能なドキュメントデータベースは多数存在し，また，ドキュメント数は増加を続けており，それらのデータ群は，知識・情報の源として重要な存在となっている．これらの大規模なドキュメントデータベースを対象とし，利用者の要求に適合した知識，情報を提供可能な性能の高い検索手法が必要とされている．

ドキュメントデータベースを対象とした検索を行う際，パターンマッチングによる方法を用いる場合には，言語の持つ類似性や多義性を扱うことが困難である．類似性，多義性を持つことばは，意味が文脈に応じて確定する性質がある．パターンマッチングを用いた情報検索は，文脈に応じた検索処理を行えないという点で柔軟性を欠いている．パターンマッチングでは扱うことが困難な類義性や多義性は，単語の表現でなく，ドキュメントの持つ意味に着目して文脈に応じた意味的な解釈を行うことで扱うことが可能となる．

ドキュメントの持つ意味に着目した検索手法の例として，意味の数学モデル [2, 3] や LSI[6](Latent Semantic Indexing) などが挙げられる．こうした検索手法はドキュメント間，あるいは検索質問とドキュメント間の同義性や類似性を，ドキュメントの持つ意味的な内容を端的に示すメタデータによって

判定することで実現されている．

意味的な内容を反映した検索を行うためにはドキュメントの内容を端的に示すメタデータをドキュメント毎に付与する必要がある．意味的検索のための各ドキュメントデータのメタデータを自動抽出するためには，ドキュメント中に使われている重要な単語群を必要十分な数だけ抽出することが重要となる．

また，人間がドキュメントからメタデータを生成するオーバーヘッドは非常に大きく，大規模なドキュメントデータベースを対象とした場合には大きな障害となる．また，人間により記述されるメタデータ間では，一貫性を欠きやすいという欠点がある．メタデータ間での一貫性の欠如は，実際にシステムを利用する場合において検索結果を偏ったものとする原因となる．ドキュメント群を対象としたメタデータ群は，一貫性のある客観的な基準で，自動的な処理によって生成される必要がある．

本稿では，ドキュメント内での単語の出現頻度を用いたアルゴリズム (tf×idf)[4, 5] を用いて出現単語の重要度を反映したメタデータ自動生成方式について示す．本方式は，ドキュメント群を対象として，意味的検索のための必要十分な数の重要な単語をメタデータとして自動抽出する方式である．そして，それらの方法に従って生成されるメタデータ群を対象とした意味的ドキュメントデータベース検索に関する実験を行った結果を示し，提案するメタデータ自動生成方式の有効性を検証する．

2　メタデータ生成方式の概要

本節では tf×idf の値を反映させたドキュメントのメタデータ生成方式の概要について述べる．

tf×idf の値は，以下のような特徴をもっている．

- 一文書内で，より出現頻度の高い単語についてより大きな値を示す．
- 文書群全体で，より出現文書数の大きい単語についてより小さな値を示す．

tf×idf の値をメタデータ生成に反映させることにより，ドキュメントにおいて重要となるメタデータ群を選択的に用いることが可能となる．これにより，問い合わせにおいて各ドキュメントの各メタデータ (単語) の重要度を反映した問い合わせ結果を求めることが可能となる．

N 個のドキュメントで構成されるドキュメントデータベースにおけるドキュメント d の持つ単語 t の重み tf×idf の値である $TFIDF_{d,t}$ は以下に示す方法で求める．

$$TFIDF_{d,t} = FREQ_{d,t} \cdot (1 + \log \frac{N}{DFREQ_t})$$

$FREQ_{d,t}$: 単語 t のドキュメント d における出現頻度
$DFREQ_t$: 単語 t の対象ドキュメント群での出現文書数

対象とするドキュメント群を自動的な処理で形態素に区切る．区切られた単語の中から検索システムにおいて検索対象のメタデータとして意味が定義されている単語のみを抜き出し，その他の語は破棄する．得られた単語列に対して tf×idf のアルゴリズムを適用し，ドキュメント d のメタデータを求める際の基準として，単語 t のドキュメント d における tf×idf の値 $TFIDF_{d,t}$ をそれぞれ求める．

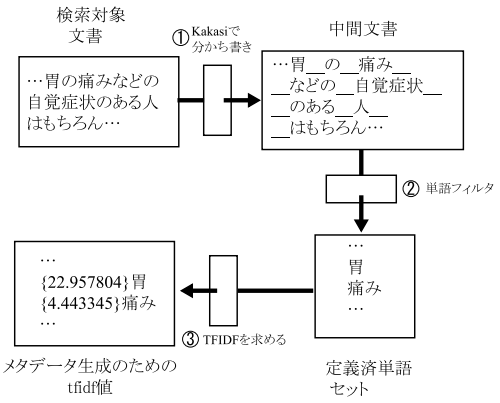

図 1: メタデータ生成手順

上に述べた手順により求められた tf×idf の値を用いて，次に挙げる方法により，複数のメタデータ集合を生成した．

方法 a　出現する全ての定義済み単語．重みは持たない

方法 w　全ての定義済み単語．tf×idf の値をメタデータの重みとして持つ

方法 c　tf×idf のドキュメント毎の中央値を閾値とし，それを越えた単語群．重みは持たない

方法 m　tf×idf のドキュメント毎の平均値を閾値とし，それを越えた単語群．重みは持たない

方法 t_n　tf×idf の値 n を閾値とし，それを越えた単語群．重みは持たない

方法 u_m　tf×idf の値の大きい順に，上位 m 件までの単語群．重みは持たない

以上の手順により，複数のメタデータ集合を生成する．

3　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 [2, 3] に述べられている．

3.1　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（f_1,f_2,$\cdots$,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 2）．このとき，M は，列ごとに 2 ノルムで正規化されている．

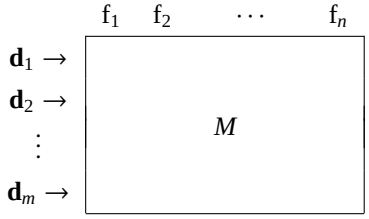

図 2: データ行列 M によるメタデータの表現

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．

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

$0 \le \nu \le n.$

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．

3.2　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．

$$\mathbf{p} = \bigoplus_{i=1}^{t} \mathbf{o}_i := (\mathrm{sign}(o_{\ell_1 1}) \max_{1 \le i \le t} |o_{i1}|,$$

$$\mathrm{sign}(o_{\ell_2 2}) \max_{1\le i\le t} |o_{i2}|,$$
$$\cdots, \mathrm{sign}(o_{\ell_n n}) \max_{1\le i\le t} |o_{in}|).$$

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．

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

3.3 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. $P_{\lambda_i} : \mathcal{MDS} \to span(\mathbf{q}_i)$.

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．

3.4 意味解釈オペレータ 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．
   $$\mathbf{G}^+(s_\ell) := \frac{\left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right)}{\left\| \left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \right\|_\infty}.$$
   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 $\Lambda_{\varepsilon_s} := \{\, i \,|\, |(\mathbf{G}^+(s_\ell))_i| > \varepsilon_s\}$ とする．

3.5 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，3.4 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．

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

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

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

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

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．

4 実験

本節ではメタデータ生成の具体的な方法について述べる．続いて，実験方式により得られたメタデータと人手により生成されたメタデータを用いて，それぞれについて意味の数学モデルを用いて検索を行った結果を相互に比較する．また，実験方式により得られたメタデータと人手により生成されたメタデータを用いてパターンマッチングによる部分一致検索での検索を行った結果を相互に比較する．

4.1 実験環境

読売新聞に連載記事として掲載された 95 件の記事を本方式に適用して実験を行った．

Perl で作成したプログラム内で KAKASI[1] の機能を用いて対象とするドキュメントを形態素解析し，本方式を適用する意味的連想検索の医療ドキュメントに関する意味のベクトル空間において医療分野に関連する単語として選んだ 1048 単語と一致する単語を抽出し，本方式によるメタデータの生成を行った．

意味的連想検索の実験においては，医療分野を説明する 316 単語を特徴語群とし，部位，症状，病名を表す 1048 単語を空間生成用メタデータとして構築した 270 次元の正規直交空間上に，検索対象とするドキュメント群を，生成したメタデータを用いて写像し，意味的連想検索を行った．

パターンマッチングによる部分一致検索の実験においては，検索語と同じパターンを含むメタデータを有する文書を検索結果として出力する検索プログラムを Perl で実装して実験を行った．

4.2 メタデータの生成

読売新聞に掲載された医療に関連する連載記事 95 件を対象としてメタデータ生成を行う．

対象とする新聞記事データを形態素解析ツールの KAKASI[1] を用いて形態素解析[*1]を行う．形態素解析によって得られた単語列のうち，意味の数学モデルの医療ドキュメント空間 [7] 内で意味が定義さ

[*1] 意味のまとまりを重視して単語の区切りをつける分かち書きオプション (-w) を用いる

れたもののみを抜き出す．抜き出された単語群の，各ドキュメント内での出現頻度を用いて tf×idf を算出する．

2節に示したメタデータ生成方法 a, w, c, m, t_n, u_m 各々にしたがって，求められた tf×idf の値を利用した，これらのドキュメントを対象としたメタデータを生成した．

人手により付与されたメタデータの一部を表 1に示し，実験方式によって得られたメタデータの一部を表 2,3,4に示す．

4.3 検索による実験

実験方式の有効性について，それぞれのメタデータ群の性質を，実際に意味的連想検索とパターンマッチングの二つの方法で検索を行った際の再現率および適合率を計算することによって検証する．

4.3.1 実験 1:意味的連想検索

実験方式によって得られたメタデータと，人手によって付与されたメタデータを，意味の数学モデルを用いた検索にそれぞれ適用し，比較する．

意味の数学モデルの医療ドキュメント空間に配置する検索対象のメタデータとして，実験方式で生成されたメタデータ群をそれぞれ適用し，「がん，腫瘍 (図中では文脈 1)」,「心臓，心室，心筋 (文脈 2)」,「肺，肺炎 (文脈 3)」の 3 種類の異なる文脈，下で意味的連想検索を行う．検索結果の上位 20 件と，人手によって生成されたメタデータで検索を行った場合の検索結果上位 20 件を比較し，20 件のうち，両者の検索結果が一致している件数をそれぞれのメタデータ群での検索性能とみなし，それぞれ比較する．

結果は表 5, 図 3に示すとおりである．

4.3.2 実験 2:パターンマッチングによる検索

生成されたメタデータ群を用いて，パターンマッチングによる検索を行う．複数種類の検索語を用いて検索を行い，人手によって生成されたメタデータで検索を行った場合の検索結果と，自動生成されたメタデータでの検索結果の双方に含まれるドキュメントを正解とし，それぞれ再現率，適合率として以下の割合を求めた．

ドキュメント	メタデータ	
がんのＱＯＬ 01	がん	肺がん
	肺	リンパ節
がんのＱＯＬ 02	がん	肺がん
	肺	腰椎
	しびれ	ぎっくり腰
胃がん 01	胃	胃がん
	がん	食道
	痛み	異物感
	ポリープ	早期がん
	消化器	
胃がん 02	胃	胃がん
	がん	胃かいよう
	吐血	ポリープ
	粘膜	消化器
⋮		⋮

表 1: 人手によって生成されたメタデータ [*y*]

ドキュメント	メタデータ		
がんのＱＯＬ 01	肺	頭	元気
	がん	興奮	声
がんのＱＯＬ 02	がん	骨	肺
	腰	声	元気
	頭	痛み	髪
	気分	顔	目
胃がん 01	胃	食道	痛み
	がん	苦痛	ポリープ
胃がん 02	痛み	出血	胃
	粘膜	ポリープ	上皮
	胃壁	筋肉	細胞
⋮		⋮	

表 2: 出現する全ての定義済み単語をメタデータとしたもの [*a*]

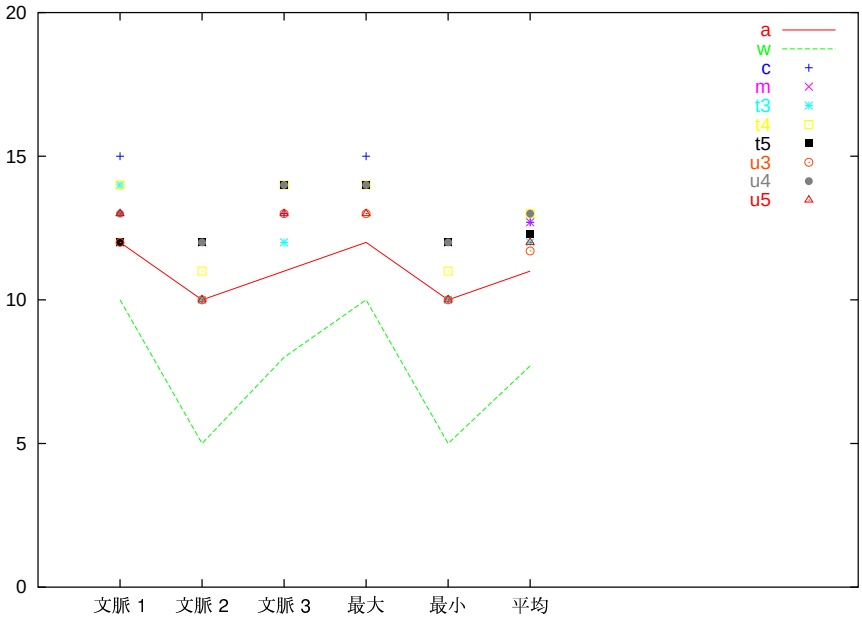

図 3: 意味的連想検索の結果:比較実験との一致件数

ドキュメント	メタデータ	
がんのＱＯＬ 01	{15.1} 肺	{4.1} 興奮
	{3.1} 元気	{3.1} 声
	{2.6} 頭	{2.0} がん
がんのＱＯＬ 02	{9.1} 顔	{7.1} 骨
	{7.2} 腰	{5.0} 肺
	{4.1} 髪	{3.1} 気分
	{3.1} 元気	{3.1} 声
	{2.7} 目	{2.6} 頭
	{2.2} 痛み	{2.0} がん
胃がん 01	{22.1} 胃	{9.7} 苦痛
	{4.4} 痛み	{4.1} ポリープ
	{3.9} 食道	{2.0} がん
胃がん 02	{33.3} ポリープ	{16.1} 粘膜
	{13.4} 出血	{12.6} 上皮
	{11.1} 胃	{6.1} 痛み
	{4.1} 胃壁	{3.1} 筋肉
	{2.1} 細胞	
⋮	⋮	

表 3: tf×idf の重みをつけたもの [w]

$$\text{再現率}：\frac{\text{出力された正解のドキュメント数}}{\text{正解ドキュメント数}}$$

$$\text{適合率}：\frac{\text{出力された正解のドキュメント数}}{\text{検索結果として出力されたドキュメント数}}$$

結果は表 6, 図 4,5,6,7,8,10, 9,11のとおりである.

4.4 考察

実験 1 では，実験方式により生成されたメタデータを用いて意味的連想検索を行い，検索の結果を人手により生成されたメタデータで検索を行った場合の結果と比較した．意味の定義されている単語をそのまま抜き出した場合 (方法 [a]) と，tf×idf の値をそのままメタデータの重みとした場合 (方法 [w]) との比較において，tf×idf を，単語をメタデータとして選択する際の閾値として用いる方法 [t_{3-5}]，あるいは並び替えによる選択の基準として用いる方法 [c,m,u_{3-5}] の方が，すべての文脈においてより高い性能を持つことがわかった．

実験 2 では，実験方式により生成されたメタデータを用いてパターンマッチングによる検索を行った．この検索の結果を，人手により生成されたメタデータで検索を行った場合の結果と比較した．パ

ドキュメント	メタデータ
がんのＱＯＬ 01	肺　興奮　元気
	声　頭
がんのＱＯＬ 02	顔　骨　腰
	肺　髪
胃がん 01	胃　苦痛　痛み
	ポリープ　食道
胃がん 02	ポリープ　粘膜　出血
	上皮　胃
⋮	⋮

表 4: tf×idf の値順の上位 5 件までを選んだもの [u_5]

	文脈 1	文脈 2	文脈 3
方法 a	12	10	11
方法 w	10	5	8
方法 c	15	10	13
方法 m	12	12	14
方法 t_3	14	10	12
方法 t_4	14	11	14
方法 t_5	12	12	14
方法 u_3	12	10	13
方法 u_4	13	12	14
方法 u_5	13	10	13
方法 y	20	20	20

	最多	最少	平均
方法 a	12	10	11.0
方法 w	10	5	7.7
方法 c	15	10	12.7
方法 m	14	12	12.7
方法 t_3	14	10	12.0
方法 t_4	14	11	13.0
方法 t_5	14	12	12.3
方法 u_3	13	10	11.7
方法 u_4	14	12	13.0
方法 u_5	13	10	12.0
方法 y	20	20	20.0

表 5: 意味的連想結果の結果:比較実験との一致件数

	適合率	再現率	適合率	再現率
	熱		高血圧	
a	0.46	1.00	0.44	0.80
w	0.46	1.00	0.44	0.80
c	0.57	0.89	0.25	0.20
m	0.89	0.89	1.00	0.20
t_3	0.46	1.00	0.44	0.80
t_4	0.64	1.00	0.50	0.20
t_5	0.73	0.89	0.50	0.20
u_3	0.80	0.67	0.33	0.20
u_4	0.86	0.86	0.25	0.20
u_5	0.88	0.88	0.25	0.20
	ストレス		肥満	
a	0.47	0.82	0.50	0.71
w	0.47	0.82	0.50	0.71
c	0.58	0.64	0.60	0.43
m	1.00	0.36	0.50	0.29
t_3	0.58	0.64	0.50	0.71
t_4	0.58	0.64	0.60	0.43
t_5	0.58	0.64	0.60	0.43
u_3	1.00	0.18	0.80	0.57
u_4	1.00	0.36	0.80	0.57
u_5	0.88	0.64	0.80	0.57
	疲労		糖尿病	
a	0.47	1.00	0.58	1.00
w	0.47	1.00	0.58	1.00
c	0.55	0.75	0.94	0.89
m	0.50	0.71	1.00	0.78
t_3	0.58	0.88	0.89	0.89
t_4	0.58	0.88	0.89	0.89
t_5	0.58	0.88	1.00	0.78
u_3	0.50	0.57	1.00	0.83
u_4	0.55	0.86	0.94	0.83
u_5	0.55	0.86	0.94	0.83

表 6: パターンマッチングによる検索結果

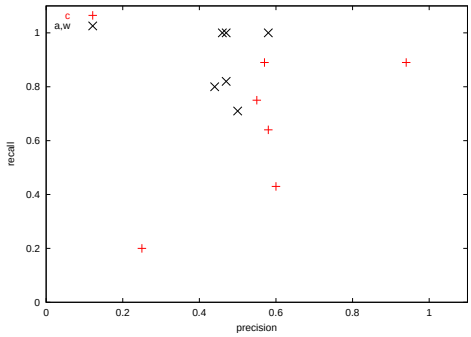

図 4: パターンマッチングによる検索結果 (c)

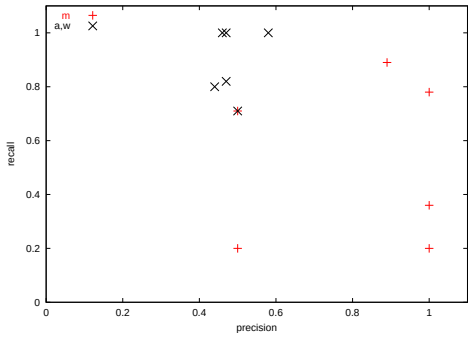

図 5: パターンマッチングによる検索結果 (m)

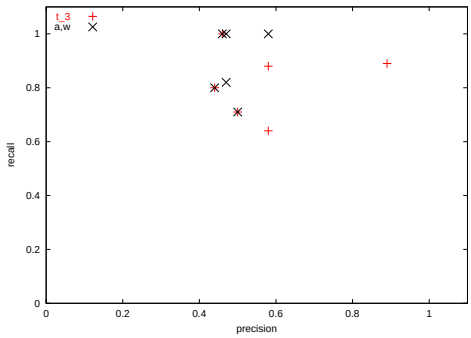

図 6: パターンマッチングによる検索結果 (t_3)

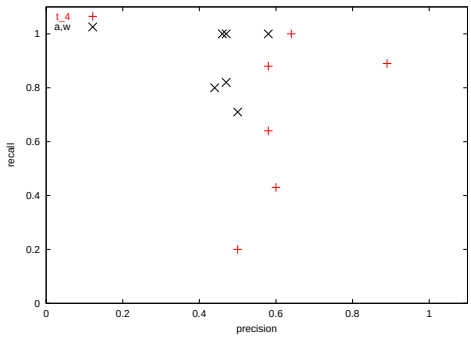

図 7: パターンマッチングによる検索結果 (t_4)

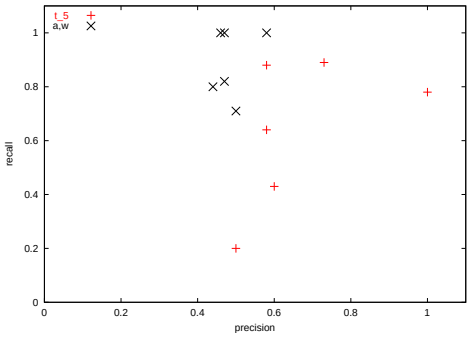

図 8: パターンマッチングによる検索結果 (t_5)

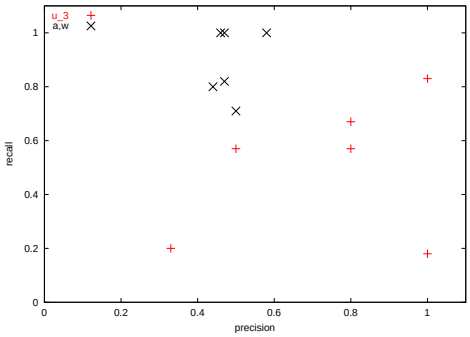

図 9: パターンマッチングによる検索結果 (u_3)

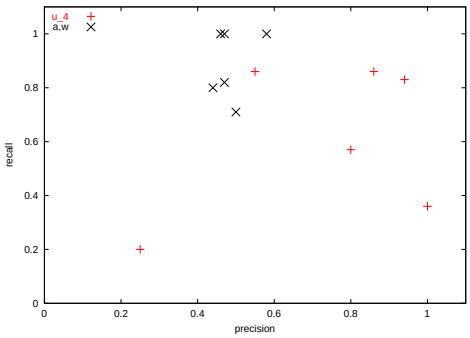

図 10: パターンマッチングによる検索結果 (u_4)

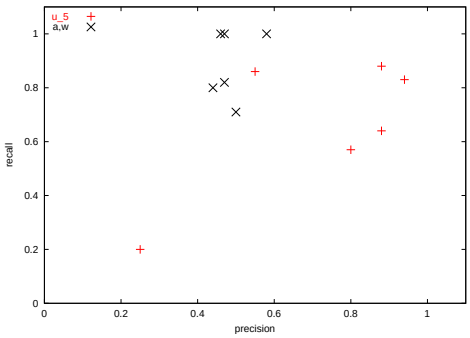

図 11: パターンマッチングによる検索結果 (u_5)

ターンマッチングによる検索に用いる場合では，意味の定義されている単語を全て抜き出したもの (方法 [a]) は他の方式により生成されるメタデータを包含する集合であるため，再現率はいずれの検索語でも最高値になっている．また，方法 [w] によって得られたメタデータは方法 [a] と同じ単語の集合に重みが付与されているものであり，パターンマッチングの検索に用いる限り，両者の検索結果は全く同じものとなっている．tf×idf を，単語をメタデータとして選択する際の閾値として用いる方法 [t_{3-5}] や，あるいは並び替えによる選択の基準として用いる方式 [c,m,u_{3-5}] では，方法 [a] や方法 [w] との比較において，検索語によっては，適切な絞込みを行うことができ再現率の減少を軽度に抑え，かつ，適合率の増加にはたらいている場合もある．しかし一方で，入力となる検索語によっては適合率の増加にはたらかない場合も存在し，これらの方法が一般的に有用であるとまでは言えない．検索結果の再現率を犠牲にしてでも検索結果のノイズを減らしたいという限定された場面においては有効に機能すると言える．

以上より，方法 [a, w] との比較において，本論文で提案したメタデータ生成方式，方法 [c,m,t_m,u_n] が特に意味的連想検索のメタデータとして用いる場合に有効であることを示した．

5 結論

本稿では，単語の出現頻度を用いたアルゴリズムにより，ドキュメントからメタデータを自動的に抽出する方式について述べた．本方式によってドキュメントから自動的にメタデータを生成し，意味的相関を計量する検索システムでのメタデータとして利用可能であることを確認した．また，実験により，生成されたメタデータが良好な検索結果をもたらすことが確認できた．

現在，電子的なメディア上に存在するドキュメントは莫大な量が蓄積されており，これらの豊富な知的資源を利用するためには，意味的相関を計量可能であるなどの高度な機能を持ち，かつ，大量のドキュメントに対して自動的に，客観的な統一された基準で処理を行うことのできる手法が有効である．本方式により意味的相関の計量に用いるメタデータを自動的な処理によって，同一の客観的な基準で生成することが可能であると確認できた．

参考文献

[1] KAKASI project: http://kakasi.namazu.org/, 1999-2001.

[2] 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.

[3] 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. (also published as:
Multimedia Data Management – using metadata to integrate and apply digital media –, McGrawHill, A. Sheth and W. Klas(editors), Chapter 7, 1998.)

[4] Salton, G. and Buckley, C.: “Term-weighting approaches in automatic text retrieval,” Information Processing and Management, 24, pp.513-523, 1988d.

[5] Salton, G. and Buckley, C.: “Improving retrieval performance by relevance feedback,” Journal of the American Society for Information Science, 41(4), pp.288-297, 1990.

[6] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman: “Indexing by Latent Semantic Analysis,” Journal of the American Society of Information Science, (1990)

[7] 吉田 尚史, 図子 泰三, 清木 康, 北川 高嗣: “ドキュメントデータ群を対象とした文脈依存動的クラスタリングおよび意味的データマイニング方式”, 情報処理学会論文誌：データベース, Vol. 41, No. SIG 1 (TOD5), pp. 127-139,2000.