2004－DBS－134　(II)　(58)
2004／7／14

意味的連想検索におけるメタデータ表現のための ベクトル自動生成方式

高　松　耕　太†　　清　木　　康　††

本稿では，意味的連想検索におけるメタデータ表現のためのベクトルを特徴語が自動的に抽出可能な辞書より自動生成する方式を示す．本方式では，次のステップにより，対象とする辞書からベクトル群を自動生成する．1) 辞書によって定められた見出し語の定義を特定の単語群 (特徴語群) として抽出する．2) 辞書より基本データと説明語群を抽出する．3) 説明語群に表記の揺れを取り除くフィルタ群を適用する．4) 辞書によって定められた特徴語群と説明語群から一致するものを基本データの定義とする．5) どの特徴語にも一致しなかった説明語に対しては，特徴語に変換するフィルタを適用し，基本データの定義とする．これらのステップより，基本データの定義を特徴語によって説明している辞書からベクトル群を自動生成することが可能になる．本稿では本方式を意味的連想検索に適用し，本方式の実現可能性を示す．

An automatic vectors creation method for metadata expression of semantic associative search

In this paper, we present an automatic creation method for metadata vectors according to feature words specified in an authorized dictionary. By our method, we can obtain a large number of metadata vectors for semantic search systems. Our method realizes automatic creation for metadata vectors as the following five steps: 1) extracting feature words from the target dictionary, 2) extracting basic words and explaining words from the dictionary, 3) applying noize reduction filters that normalize variants of expression, 4) creating word definitions according to feature words and explaining words, 5) converting explaining words, which have no corresponding feature words, by applying special filters. We clarify effectiveness of our method by applying to the semantic associative search method.

1. はじめに

現在，多種多様なメディアデータ群が広域ネットワークを介して情報システム上において存在するが，それらを対象にして意味的に相関の強いメディアデータ群の獲得を可能とする意味的検索機構を実現する必要がある．より質の高い検索を目的として，文献[1)~4)]では言葉と言葉の相関量を計量する意味の数学モデルが提案されている．意味の数学モデルを用いた意味的連想検索方式[1)~4)]では，検索語ベクトルと検索対象データをメタデータ空間と呼ばれる直交空間上に写像する．この方式では，指定した検索語ベクトルに対応する動的な文脈に応じて，意味的に近い情報の検索を実現している．意味的連想検索を行うためにはあらかじめメタデータ空間，検索対象メタデータ，検索語メタデータを生成しなければならない．本稿では，見出し語の定義を特定の単語群で説明している辞書により，意味的連想検索のためのベクトル群を自動生成する方式について示す．これにより，大規模データを対象として，意味的連想検索を適用可能な応用分野を広げることが可能となる．本稿では，文献[4)]で示されている方式を応用し，大規模データ群を対象とした方式を実現する．本方式では次の５ステップにより，対象とする辞書からベクトル群を自動生成する．辞書によって定められた特徴語群を抽出する．辞書より基本データ（見出し語から抽出）と説明語群（その見出し語を説明している語群）を抽出し，説明語群に活用や語尾の変化など，表記の揺れを取り除くフィルタ群を適用する．辞書によって定められた特徴語群と説明語群から一致するものを基本データの定義とする．どの特徴語にも一致しなかった説明語に対しては，特徴語に変換するフィルタを適用し，基本データの定義に追加する．これにより，基本データの定義を特徴語によって説明している

† 慶応義塾大学政策・メディア研究科
†† 慶應義塾大学環境情報学部

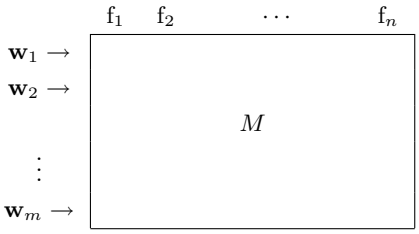

図 1 データ行列によるメタデータ表現

辞書からベクトル群を自動生成することが可能となる．

関連研究として文献[5]，に示された研究がある．文献[5] は基本データ群と説明語群より，特徴語群抽出したメタデータ空間を半自動的に生成する方式である．この方式[5] は多様な形態の辞書を利用可能とする方式であるが，大規模データについて，メタデータ付与することには不向きである．この方式との比較において，本方式は，大規模なデータに対してメタデータの自動生成が可能である．

2. 意味的連想検索におけるメタデータ表現のためのベクトル自動生成方式

本節では，特徴語が自動的に抽出可能な辞書を対象とした意味的検索空間のためのベクトル自動生成方式を示す．本方式は文献[4] にて提案された意味的連想検索方式に適用するための，大規模データを対象としたベクトルの自動生成方式を実現する．

意味的連想検索では，初めに，m 個の基本データについて各々n 個の特徴 $(f_1, f_2, ..., f_n)$ を列挙した特徴語つきベクトル $w_i(i = 1,...,m)$ 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)． 「**特徴語 (feature)**」とは，図 1 の特徴 $(f_1, f_2, ..., f_n)$ にあたり，「**基本データ**」とは，$(w_1, w_2, ..., w_m)$ を指す．「**説明語**」とは，辞書の見出し語の定義に使われている単語群に相当する．「**基本データセット**」とは基本データに対して特徴語群から，基本データと関係ある単語群（＝説明語群）を基本データの定義として表現したものであり，表 1 に示すように構成となっている． 本方式は，辞書の見出し語を基本データとし，見出し語を説明している説明語群を各種フィルターに適用することにより，特徴語群を抽出する方式である．本方式の主要なプロセスを以下にまとめる（図 2）．本方式では，対象とする辞書へこれらの **Step-1**〜**Step-5** のステップによって，辞書より基本データセットを結果として得る．この基本データセットを対象として，各基本データについて，０，１，および－１の３値により表現し，ベクトルを自動生成する．

表 1 基本単語の説明の例

基本データ	説明語群
excite	cause2 lose -calm1 have2 strong feelings expectation happiness
grief	great1 sorrow1 feelings suffering death loved person

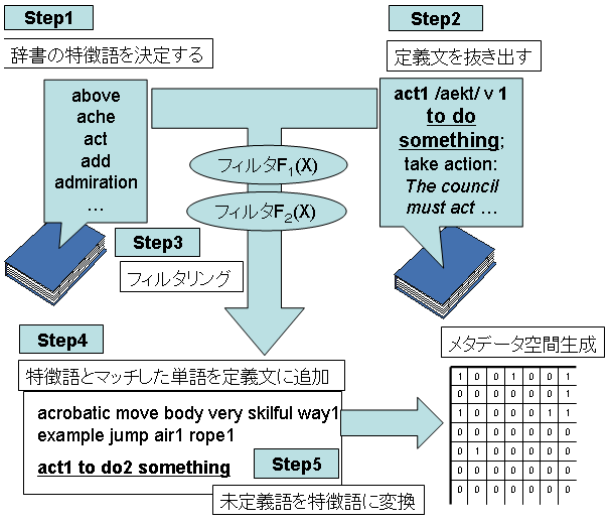

図 2 本方式概観

Step-1: 特徴語群の抽出

辞書の定義に使われている特徴語群 $f_1, f_2, ..., f_i$ を抽出する．本方式において用いる特徴語群は，利用する辞書が説明語彙として指定している単語群をそのまま特徴語群として抽出する．

Step-2: 基本データと説明語群の抽出

辞書の情報を参照し，基本データ w_j と基本データを説明する説明語群 $d_{j1}, d_{j2}, ..., d_{jk}$ を抽出する．

Step-3: 説明語の表記の統一

活用や語尾の変化など，説明語の表記の揺れを解消する表記変換フィルタ群 $F_1(X), F_2(X), ..., F_l(X)$ を設定する．表記変換フィルタ $F_l(X)$ は文字列を入力にとり，変換後の文字列を出力する．

$$F_l(string) \longrightarrow string'$$

Step-2 において抽出された説明語群 $d_{j1}, d_{j2}, ..., d_{jk}$ について，表記変換フィルタ群 $F_1(X), F_2(X), ..., F_l(X)$ からフィルタを任意の組み合わせで使用し，変換された説明語群 $d'_{j1}, d'_{j2}, ..., d'_{jk}$ を生成する．

Step-4: 特徴語群と説明語の比較

Step-3 において変換された説明語群 $d'_{j1}, d'_{j2}, ..., d'_{jk}$ のそれぞれについて，特徴語群 $f_1, f_2, ..., f_i$ との

比較を行う．特徴語群と一致した説明語 d'_{jk} は基本データセット W_j の定義の単語群に追加する．どの特徴語とも一致しなかった説明語群を未定義語群とよび，未定義語群 $u_{j1}, u_{j2}, ..., u_{jk}$ とする．

Step-5: 特徴語群に一致しなかった説明語群に対する変換

説明語を特徴語群に変換する特徴語変換フィルタ群 $Z_1(X), Z_2(X), .., Z_r(X)$ を設定する．特徴語変換フィルタ群 $Z_r(X)$ は説明語を入力にとり，特徴語群を出力する．

$Z_r(string) \longrightarrow feature_1, feature_2, ..., feature_n$

Step-4 によって生成された未定義語群 $u_{j1}, u_{j2}, ..., u_{jk}$ について，特徴語変換フィルタ群 $Z_r(X)$ から一つのフィルタを選んで使用し，出力された特徴語群を基本データセット W_j の定義の単語群に追加する．

3. 実 現 方 式

ここでは，本方式において述べたプロセスの具体的な実現方式について示す．ここでは例として，ロングマン現代英英辞典[7] を参照し，本方式を適用する．

Step-1: 特徴語群の抽出

辞書の定義に使われている特徴語群 $f_1, f_2, ..., f_i$ を抽出する．ロングマン現代英英辞典では見出し語の定義に使われる単語として 2148 語が指定されている．ここでは，この 2148 語をそのまま特徴語として採用する．

Step-2: 基本データと説明語群の抽出

辞書の情報を参照し，基本データ w_j と基本データを説明する説明語群 $d_{j1}, d_{j2}, ..., d_{jk}$ を抽出した．ロングマン現代英英辞典の見出し語から 47849 語を抽出する．

Step-3: 説明語の表記の統一

表記変換フィルタ $F_l(X)$ として以下のフィルタを実現する．

(1) $F_1(X)$ **Stemming フィルタ**

本フィルタは，入力された説明語に対して Stemming（活用を取り除き出力）を行うフィルタである．説明文中の活用を取り除くことにより，説明語を特徴語に一致させる．

(2) $F_2(X)$ **接頭辞除去フィルタ**

本フィルタは，入力された説明語から接頭辞を取り除き，出力するフィルタである．対象となる説明語において否定の意味を持つ接頭辞である un,im,dis がついていた場合，単語にマイナスをつけて出力する（-1 となる）．フィルタ $F_1(X)$ では取り除くことのできない接頭辞を取り除くことにより，説明語と特徴語を一致させる場合に使用する．

(3) $F_3(X)$ **品詞除去フィルタ**

本フィルタは，入力された説明語が文法的に意味を持たない品詞に該当する場合はその説明語を排除し，それ以外の場合は説明語をそのまま出力するフィルタである．文法的に使われてる品詞としては接続詞，冠詞，前置詞，代名詞などが挙げられる．このフィルタは単語の定義に必要ではない品詞を排除する場合に使用する．

(4) $F_4(X)$ **基本データ除去フィルタ**

本フィルタは，入力された説明語が基本データ自身と一致するときにはその説明語を排除し，それ以外の場合は説明語をそのまま出力するフィルタである．基本データの説明語にその基本データ自身を使う例は，基本データの定義として使うのではなく，基本データをシンボルとして文の中で説明するために使っている．この説明語を排除する場合に使用する．

(5) $F_5(X)$ **Stop-word フィルタ**

本フィルタは，入力された説明語がメタデータ空間生成者によって登録されている単語のリスト (ストップワードリスト) に一致する場合はその説明語を排除し，それ以外の場合は説明語をそのまま出力するフィルタである．このフィルタはメタデータ空間生成者が特に空間生成で必要でないと判断した単語を排除する場合に使用する．**Step-2** において抽出された説明語群 $d_{j1}, d_{j2}, ..., d_{jk}$ に対し，上記の表記変換フィルタ群 $F_1(X)$,$F_2(X)$, ...,$F_5(X)$ からフィルタを任意の組み合わせで使用し，変換された説明語群 $d'_{j1}, d'_{j2}, ..., d'_{jk}$ を生成した．

Step-4: 特徴語群と説明語の比較

Step-3 において変換された説明語群 $d'_{j1}, d'_{j2}, ..., d'_{jk}$ のそれぞれについて，特徴語群 $f_1, f_2, ..., f_{2148}$ との比較を行う．特徴語群と一致した説明語 d'_{jk} は基本データセット W_j の定義の単語群に追加する．どの特徴語とも一致しなかった説明語群を未定義語群 $u_{j1}, u_{j2}, ..., u_{jk}$ とする．

Step-5: 特徴語群に一致しなかった説明語群に対する変換

特徴語変換フィルタ $Z_r(X)$ として次のフィルタを適用する．

(1) $Z_1(X)$ **特徴語除去フィルタ**

入力された未定義語 u_{jk} を削除する．

(2) $Z_2(X)$ **定義追加フィルタ**
入力された未定義語 u_{jk} に対し，一致する基本データ w_s を探し，その定義に使われている特徴語 $f_{s1}, f_{s2}, ..., f_{sk}$ を出力する．
(3) $Z_3(X)$ **重要語追加フィルタ**
入力された未定義語 u_{jk} に対し，一致する基本データ w_s を探し，その定義に使われている特徴語 $f_{s1}, f_{s2}, ..., f_{sk}$ のうち辞書における出現頻度が n 回以下の特徴語群 $f_{st}, ..$ のみを出力する．
Step-4 によって生成された未定義語群 $u_{j1}, u_{j2}, ..., u_{jk}$ に対し，特徴語変換フィルタ群 $Z_r(X)$ からひとつを選んで使用し，出力された特徴語群を基本データセット W_j の定義の単語群に追加した．

4. 実　　験

4.1 実験概要

本方式の有効性を検証するために行った実験の結果を示す．ロングマン現代英英辞典[7] を参照し，本方式を適用した結果，47849 語の基本データ群と 2148 語種類の特徴語群が抽出された．抽出された基本データ群と特徴語群から，基本データセットを生成し，メタデータ空間を生成した．

4.2 実験:自動生成方式に関する実験

第 2 節の **Step-3** および，第 3 節の **Step-3** における $F_l(X)$ の性能を評価する．生成されたすべての基本データセット群のうち，未定義語を含まない基本データセット群の割合を P と定義する．本実験では，自動生成に使用する表記変換フィルタを組み合わせることにより，未定義語を含まない基本データセット群 P の数が最も多くなるようなフィルタの組み合わせを検証した．実験結果を表 2 に示す．

4.3 考察:自動生成方式に関する実験

この実験結果から表記変換フィルタ $F_l(X)$ を組み合わせることで未定義語を含まない基本データセット群 P_j の数が高くなることがわかる．$F_1(X)$(Stemming フィルタ) によって，未定義語全体の約 4 割の 25000 語の未定義語を特徴語に変換することができた．自動生成において活用を除去するフィルタ $F_1(X)$ が効果があることがわかる．同様に品詞による除去や接頭辞による除去など文法的な変化を取り除くフィルタ $F_2(X), F_3(X)$ を組み合わせて使用することにより，未定義語を約 4 割まで減らすことができた．未定義語を含まない基本データセットの割合については，フィルタ $F_1(X)$〜$F_3(X)$ 適用前の 31%から，適用後は 60%へと変換することができた．文法的なルールだけではなく，辞書の表記のルールに注目した $F_4(X)$(基本データ除去フィルタ) についても，適用することで未定義語を減らし，未定義語を含まない基本データセットを増やすことができた．メタデータ空間生成者によって定義されたルールに着目した $F_5(X)$(ストップワードフィルタ) によっても，未定義語を減らすことができた．これらの結果は辞書からのベクトル自動生成方式の実現可能性を示している．

4.4 実験:特徴語変換フィルタ $Z_r(X)$ の性能に関する実験

本実験では，自動生成中に出現する未定義語群に対する特徴語変換フィルタ $Z_r(X)$(**2 節の Step-5**) の正確性について検証する．それぞれのフィルタを用いて，基本データの未定義語を特徴語群に変換した．表 3 は基本データセット happy と基本データセット pleasure の説明語群である．基本データセット happy の説明語群のうち，pleasure のみが，未定義語になっている．表 4 は未定義語 pleasure を含む基本データ happy について，フィルタ $Z_1(X)$ から $Z_3(X)$ をそれぞれ適用して作られた基本データセットである．

2 節の **Step-5** および，3 節の **Step-5** における $Z_r(X)$ の正確さを評価する．ここでは本方式により，

表 2 表記変換フィルタ $F_l(X)$ による実験結果

フィルタ	種類	のべ出現数	未定義語を含まない基本データセット	P
適用前	7000	61543	14950	31%
$F_1(X)$	5095	35165	23851	50%
$F_{1\sim2}(X)$	5020	28549	27155	57%
$F_{1\sim3}(X)$	4801	25757	28754	60%
$F_{1\sim4}(X)$	2796	21964	31000	65%
$F_{1\sim5}(X)$	2742	21672	31130	65%
全フィルタ適用結果	-4258	-39871	+16180	+34%

表 3 基本データセット happy と基本データセット pleasure の定義

基本データ	説明語群
happy	have feeling example so1 best have happen1 (pleasure)
pleasure	enjoy(175) feeling(1375) satisfaction(13) get(695) experience(191) enjoy(175)

括弧内の数字は，辞書内における特徴語の出現数である．

表 4 変換語の基本データセット happy の定義

基本データ	説明語群
特徴語除去フィルタ $Z_1(X)$	have feeling example so1 best have happen1
定義追加フィルタ $Z_2(X)$	have feeling example so1 best have happen1 enjoy feeling satisfaction get experience enjoy
特徴語追加フィルタ $Z_3(X)$	have feeling example so1 best have happen1 satisfaction enjoy

フィルタ $Z_r(X)$ を適用したベクトルを生成する．検索語のベクトルとして適用し，2945 語各々について手動で生成した 2945 個のベクトルを対象として意味的連想検索方式による検索[1)~4)] を行い，それぞれの検索結果を比較した．表 5, 表 6, 表 7 はそれぞれ，フィルタ $Z_1(X)$,$Z_2(X)$,$Z_3(X)$ によって生成された，未定義語 pleasure を含む基本データ happy を検索語とし，手動で作られた 2945 語のベクトルデータを対象として意味的連想検索[1)~4)] を行った結果である．

4.5 考察:特徴語変換フィルタ $Z_r(X)$ の性能に関する実験

表 6 は pleasure の定義を直接に特徴語に変換したフィルタを適用し検索を行った場合の検索結果である．表 5 と比較して表 6 は pleasure や enjoy など，意味的に近い語が上位に検索されているが，nurse1 や chairperson など関連性の低い単語も上位に検索されている．

表 7 は pleasure の定義のうち，出現頻度の低い単語のみを特徴語に変換したフィルタを適用した場合の検索結果である．表 5 と表 7 を比較すると，delight2 や pleasure といった単語がフィルタ $Z_3(X)$ によって追加された単語の存在により，上位に検索されている．辞書の限定された説明語の中において，出現頻度が低い単語は見出し語と意味的に強い関係があると考えられる．pleasure の説明文はは本来は happy の説明文ではなく，そのまま happy に追加された場合，happy とは直接関係のない nurse1 や chairperson などが pleasure との関連において上位に現れてしまう．happy と直接関係の深い satisfaction，enjoy を補うことにより，pleasure を用いる代わりに pleasure と意味的に近い特徴付けを行うことが可能となる．これにより，pleasure の定義に偏らず happy の未定義語の情報を補うことができる．

本実験では，重要語追加フィルタ $Z_3(X)$ が未定義語を特徴語を変換するフィルタとして適切であると考えられる．これらの結果は，本方式における複数の特徴語変化フィルタ $Zr(X)$ の適用結果を比較，検討可能な環境を実現できたことを示している．

5. 結　　論

本稿では意味的連想検索方式の意味空間を基本データの定義を特徴語によって説明している辞書から自動生成する方式について述べた．本方式によって，特徴語の定まった辞書を用いることで大規模データを対象とした意味的連想検索が実現可能となる．今後の展望として，特徴語群を生成する方式[5)] と本方式を組み合わせることにより，辞書の形式の如何に関わらず、メタデータ空間を自動生成する方式の実現を目標とする．

謝辞　本研究に多大なご助言をいただいた慶應義塾大学院政策・メディア研究科，吉田尚史氏，倉林修一氏，河本 穣氏，佐々木史織氏に感謝いたします．

表 5 特徴語除去フィルタ $Z_1(X)$ によって生成された happy による検索結果

	相関量	ワード名	ワードのメタデータ
1	0.325539	excite	cause2 lose -calm1 have2 strong feelings expectation happiness
2	0.287712	excited	full1 strong feelings expectation happiness -calm1
3	0.286175	kind2	caring2 happiness feelings others
4	0.285992	now3	something has happened
5	0.285992	suppose2	happen1
6	0.283928	like1	regard2 pleasure fondness have3 good1 feelings enjoy
7	0.279477	worship1	strong religious feelings love1 respect1 admiration when1 shown God god
8	0.277864	heaven	place1 God god supposed1 live1 place1 complete1 happiness souls good1 people1 go1 after1 death
9	0.276885	grief	great1 sorrow1 feelings suffering death loved person
10	0.276787	hope1	wish1 expect want1 something happen1 have2 some1 confidence happen1

各特徴語の語尾に付与されている数字は多義語を識別するための識別子である.

表 6 定義追加フィルタ $Z_2(X)$ によって生成された happy による検索結果

	相関量	ワード名	ワードのメタデータ
1	0.323492	nurse1	person typically woman trained take1 care1 sick1 hurt1 old people1 directed doctor1 hospital
2	0.313992	pleasure	state1 feeling1 happiness satisfaction resulting experience1 one2 enjoys
3	0.312968	enjoy	get pleasure thing experiences1 like1
4	0.307492	like1	regard2 pleasure fondness have3 good1 feelings enjoy
5	0.307388	care1	process1 looking after1 giving attention1 someone needs2 sick1 old person
6	0.291720	get	receive experience2
7	0.284713	chairperson	person charge2 meeting who directs work1 committee organization
8	0.283389	nurse2	take1 care1 like2 nurse1
9	0.282047	suppose2	happen1
10	0.282047	now3	something has happened

各特徴語の語尾に付与されている数字は多義語を識別するための識別子である.

表 7 重要語追加フィルタ $Z_3(X)$ によって生成された happy による検索結果

	相関量	ワード名	ワードのメタデータ
1	0.342297	like1	regard2 pleasure fondness have3 good1 feelings enjoy
2	0.319121	worship1	strong religious feelings love1 respect1 admiration when1 shown God god
3	0.317451	excite	cause2 lose -calm1 have2 strong feelings expectation happiness
4	0.310620	grief	great1 sorrow1 feelings suffering death loved person
5	0.307217	delight2	cause2 someone great1 satisfaction enjoyment joy1
6	0.306380	suppose2	happen1
7	0.306380	now3	something has happened
8	0.297206	pleased	happy satisfied
9	0.296964	pleasure	state1 feeling1 happiness satisfaction resulting experience1 one2 enjoys
10	0.296150	polite	having2 showing1 good1 manners sensitivity other people feelings correct1 social1 behaviour

各特徴語の語尾に付与されている数字は多義語を識別するための識別子である.

参 考 文 献

1) T. Kitagawa and Y. Kiyoki, “A 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) 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,Vol.j79-D-2 No.4,pp.509-519,April.1996.

3) Kiyoki, Y., Kitagawa, T. and Hitomi, Y.,“A fundamental framework for realizing semantic interoperability in a multidatabase environment ”, International Journal of Integrated Computer-Aided Engineering, Vol.2, No.1(Special Issue on Multidatabase and Interoperable Systems), pp.3-20, John Wiley & Sons, Jan. 1995.

4) 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.

5) 河本 穣, 清木 康, 吉田尚史, 藤島清太郎, 相磯貞和, “医療分野ドキュメント群を対象とした意味的連想検索空間の実現方式””, 日本データベース学会 Letters, Vol.1, No. 2, pp.12-15, March 2003

6) 中西 崇文, 岸本 貞弥, 櫻井 鉄也, 北川 高嗣 :“複数の書籍の索引部を用いたメタデータ空間拡張統合方式”, 第 15 回データ工学ワークショップ (DEWS2004), 電子情報通信学会, (2004).

7) “Longman Dictionary of Contemporary English”, Longman(1987)