集合間の相違を明確にする要素辞書

村田 美友紀† 掛下 哲郎††

集合検索はDBにおける重要なトピックであり，キーワードを用いた文献検索や仕様に基づいたソフトウエア検索等の分野に応用できる．本論文では集合検索を効果的に行うために要素辞書の概念を定義し，要素辞書の構成アルゴリズムを提案する．要素辞書を用いることで，以下の特徴が得られる．(1) 検索対象集合を特定する検索条件が常に構成できる．(2) 任意の集合間の相違に関する質問に答えられる．(3) 要素辞書を最適化することで索引のサイズを最小化できる．互いに異なる n 個の集合に対して要素辞書のサイズは $\log n$ 以上である．最小の要素辞書構成問題はNP完全になるが，与えられた n 個の集合に対してサイズが高々 $n-1$ の要素辞書が多項式時間で構成できる．さらに，検索対象集合の追加削除に伴って要素辞書を多項式時間で再構成するアルゴリズムを提案する．

Element Dictionary to Specify Difference among Sets

MIYUKI MURATA† and TETSURO KAKESHITA††

Set retrieval is an important topic in databases and has many application domains such as document retrieval using keywords and software retrieval based on their specification. In this paper, we define the notion of element dictionary to effectively retrieve sets and propose construction algorithms for the dictionary. The following advantages can be obtained by using the element dictionary. (1) A retrieval condition can always be constructed to identify the target set. (2) Difference among sets can always be specified. (3) Index size can be minimized by optimizing the dictionary. The size of an element dictionary is at least $\log n$ for n distinct sets. Although the construction of the minimum element dictionary is NP complete, a dictionary of at most $n-1$ elements can be constructed in polynomial time. Furthermore we propose polynomial time algorithms to reconstruct the element dictionary when a target set is added to or deleted from the database.

1. はじめに

データベース（DB）中の様々な情報をモデル化する際に集合は重要な役割を果たす．たとえば，文献検索においては検索対象の文献をキーワード集合で特徴づけている[1)~3)]．クエリーはそれによって検索されるDB中のデータ集合によって仕様を記述できる．同様のことがビューや一貫性制約に対しても言える．プログラムの仕様は，一般的に入出力データと副作用の組の集合で定義される．同様に，クラスもインスタンスの集合により特徴づけられる．

近年，知識発見技術（KDD）やデータマイニングが関心を集めている[4),5)]．これによって生成された知識は相関ルール，決定木，クラスタリング等のさまざまな形式を取るが，いずれも特定のデータ集合を抽象化した二次情報と考えることができる．従って，各種の知識も集合によって特徴づけられる．

DBシステムの複雑化に伴い，上述したクエリー，ビュー，一貫性制約，プログラム，知識等の再利用が重要性を増している．これを支援するためには仕様に基づいたクエリー等の検索が重要である．クエリー等の仕様は集合を用いて表現できるため、集合検索は仕様に基づいたクエリー等の検索に対応する．集合を対象とする検索方式として，我々はサンプルを用いた検索機構[6),7)]を提案している．この手法では集合の要素から構成されるサンプルを用いて目的の集合を検索する．

本論文ではサンプルを用いた集合検索を効果的に行うために，要素辞書の概念を導入する．要素辞書には集合の要素が登録されている☆．また，検索対象とな

† 八代工業高等専門学校情報電子工学科
Department of Information and Electronics Engineering, Yatsushiro National College of Technology
†† 佐賀大学理工学部知能情報システム学科
Department of Information Science, Faculty of Science and Engineering, Saga University

☆ 必ずしもすべての要素が登録されているとは限らない．

る任意の2つの集合に対して，要素辞書中には両者を区別できる要素が登録されている．要素辞書を用いて集合検索を行うことにより以下の利点が得られる．(1) 任意の検索対象集合を特定するサンプルが常に構成できるため，検索がより厳密に行える．(2) 集合間の相違に関する質問に常に答えられる．(3) 要素辞書を最適化することにより，集合に付加されるキーワードの全体集合が小さくなるため，シグネチャファイル[8),9)]や転置ファイル[1)]等の索引のサイズを小さくできる．

要素辞書は，文献検索におけるキーワード辞書構成問題にも活用できる．要素辞書をキーワード辞書として用いることで，任意の文献を識別することができる．また，要素辞書を最小化することにより，シソーラスや索引のサイズを小さくできる．

論文や特許等のようにオリジナリティを重視する文献を検索する際には，候補となる文献間の相違を明確にすることが重要である．また，ソフトウエアの再利用を行う際にも，細かい相違を明確にしなければ最適なものを選択することはできない．本論文で提案する要素辞書はこのために有用である．さらに，集合を検索する過程で，集合間の違いを表す要素を利用者に提示できるため，検索作業を支援できる．

本論文は以下のように構成されている．2節では要素辞書を定義する．また，互いに異なる n 個の集合に対して要素辞書のサイズ（要素数）が $\log n$ 以上であり，サイズ $n-1$ 以下の要素辞書が存在することを示す．3節では，最小の要素辞書構成問題のNP完全性を示す．4節では，サイズが高々 $n-1$ となる要素辞書の構成アルゴリズムを示し，集合の追加削除に伴う辞書の再構成アルゴリズムを提案する．5節では要素辞書の適用例として，サンプルを用いた集合検索とキーワード辞書構成問題について述べる．

2. 要素辞書

本節では，集合間を明確に区別する要素の集合を要素辞書として定義し，その性質について考察する．

定義1 要素集合 U の部分集合を要素とする集合 $G=\{g_1,\cdots,g_n\}$ を考える． G の要素辞書 D_G は U の部分集合であり，任意の $g,g'\in G(g\neq g')$ に対して $g\cap D_G\neq g'\cap D_G$ を満足する． □

以下に要素辞書の例を示す．

例1 図1の集合 $G=\{$学生, 卒研生, 院生, 研究室, 教官$\}$ を考える．G に対して定義される要素辞書は，$\{$山田, 斉藤, 井上$\}$ である．集合 g と $g\cap D_G$ を表1に示す． □

G の要素辞書 D_G は一般に複数存在する．G の要素

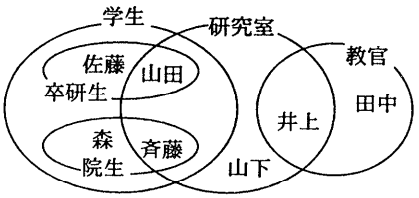

図1 集合の例

g	$g\cap D_G$
学生	{山田, 斉藤}
卒研生	{山田}
院生	{斉藤}
研究室	{山田, 斉藤, 井上}
教官	{井上}

表1 要素辞書の例

g, g' について，$diff(g,g')=(g-g')\cup(g'-g)$ を定義する．

補題1 G の任意の2つの要素 $g,g'(g\neq g')$ について，$D_G\cap diff(g,g')\neq\phi$ が成立する． □

[証明] G の要素 g,g' について，$D_G\cap diff(g,g')=\phi$ と仮定する．仮定より，$D_G\cap((g-g')\cup(g'-g))=\phi$ である．よって，$D_G\cap(g-g')=\phi$, $D_G\cap(g'-g)=\phi$ となるため，$g\cap D_G=g'\cap D_G$ である．これは要素辞書の定義に反する． □

補題1より，要素辞書を用いれば G に属する任意の2つの集合に対してその違いを表す要素が求められることが分かる．よって，この要素を検索することにより，集合間の相違に関する質問に答えることができる．また，補題1は逆も成立する．すなわち集合 $\{e|e\in diff(g,g'),g,g'\in G,g\neq g'\}$ は G の要素辞書である．ただしその要素数は $O(n^2)$ となる．

定理1 任意の集合 G に対して，G の要素辞書 D_G は少なくとも $\log_2 n$ 個の要素を含む． □

[証明] D_G が G の要素辞書であるとき，G の各要素 g について $D_G\cap g$ は互いに異なる．$D_G\cap g$ は D_G の部分集合である．D_G の部分集合は高々 $2^{|D_G|}$ 個であることから，定理1が成立する． □

定理1で示した要素数の下限を満足する要素辞書の例を以下に示す．

例2 要素集合 U のべき集合 2^U を考える．$G=2^U$ の任意の要素 $g,g'(\subseteq U)$ について，$g\cap U\neq g'\cap U$ である．よって，U は集合 2^U の要素辞書である．このとき，$|U|=\log_2|2^U|$ が成立する． □

定理2 任意の集合 G に対して，要素数が $n-1$ 以下の要素辞書 D_G が存在する． □

[証明] $n=1$ のとき $D_G=\phi$ より定理は成立する．

g	$g \cap E$
g_1	$\{e_1\}$
g_2	$\{e_2\}$
$\vdots$	$\vdots$
g_{n-1}	$\{e_{n-1}\}$
g_n	ϕ

表2 要素辞書の要素数が $n-1$ となる例

$n=k$のとき，$|D_G| \leq k-1$の要素辞書が存在すると仮定する．ここで，$G' = G \cup \{g_{k+1}\}$ を考える．D_GはGのすべての要素を唯一に識別できるため，Gの要素のうちg_{k+1}と区別できない要素g_tは高々1個である．g_{k+1}とg_tを区別するためには，要素$e \in diff(g_{k+1}, g_t)$をD_Gに追加すればよい．このため，G'の要素辞書$D_{G'}$の要素数はk以下となる． □

定理2は，要素辞書の要素数の上限が$n-1$であることを示す．以下に要素数が$n-1$の要素辞書が必要な集合の例を示す．

例3 集合Gにおいて．任意の集合g, g' $(g \neq g')$について，$g \cap g' = \phi$とする．Gに対して，$n-2$個の要素から構成される要素辞書D_Gが存在すると仮定する．このとき，$D_G \cap g_i = \phi$となる集合がG中に少なくとも2個存在する．これは定義に矛盾する．

Gに対して$e_1 \in g_1, \cdots, e_{n-1} \in g_{n-1}$からなる要素集合$E$を考える．このとき$|E| = n-1$である．**表2**に示すように，各$g$について$E \cap g$は互いに異なる．従って，$E$は$G$の要素辞書である． □

3. 最小の要素辞書構成問題

本節では，最小の要素辞書構成問題を定式化し，そのNP完全性を証明する．

最小要素辞書構成問題

インスタンス： 要素集合Uの部分集合を要素とする集合G，正の整数$K < |U|$．

質問： Gに対して要素数がK以下の要素辞書D_Gが構成できるか． □

最小要素辞書構成問題がクラスNPに属することは容易に示せる．すなわち，互いに異なる$g, g' \in G$に対して，与えられたD_Gが要素$e \in diff(g, g')$を含むことを検査すればよい．このために，任意の$e \in U, g \in G$についてeがgに属することが多項式時間で検査できることを仮定する．

節点被覆問題は，各節点の次数が4以下の平面グラフに限定した場合でもNP完全になることが知られている[10]．本節ではこの問題を最小要素辞書構成問題に帰着することでNP完全性を示す．

節点被覆問題

インスタンス： グラフ$\hat{G} = (\hat{V}, \hat{E})$，正の整数$\hat{K} < |\hat{V}|$．ただし$\hat{G}$の各節点の次数は4を超えない．

質問： $\hat{G}$について，要素数が$\hat{K}$以下の節点集合$\hat{V}' \subset \hat{V}$が存在するか．ここで，$\hat{V}'$は各枝$(u, v) \in \hat{E}$についてvまたはuの少なくとも一方を含む．

□

節点被覆問題のインスタンス$\hat{G}$と$\hat{K}$に対して，以下の手順で最小要素辞書構成問題のインスタンスU, G, Kを構成する．

$$U = \{v_1, v_2, v_3, v_4 | v \in \hat{V}\}$$
$$K = 3|\hat{V}| + \hat{K}$$

各$v \in \hat{V}$に対応して以下の集合を定義する．

$$g_v^1 = \{v_2, v_3\}$$
$$g_v^2 = \{v_1, v_3\}$$
$$g_v^3 = \{v_1, v_2\}$$
$$g_v^4 = \{v_1, v_2, v_3\}$$

各$e = (u, v) \in \hat{E}$に対応して以下の集合を定義する．

$$g_e = g_v^k \cup \{u_4, v_4\}$$

ここで，$k \in \{1, 2, 3, 4\}$はv（またはu）を始点とする各枝について互いに異なるように選択する．$\hat{G}$の各節点の次数は4以下であることより，これは常に可能である．以上に基づいて集合Gを以下のように定義する．

$$G = \{g_v^1, g_v^2, g_v^3, g_v^4 | v \in \hat{V}\} \cup \{g_e | e \in \hat{E}\}$$

U, G, Kが多項式時間で構成可能なことは容易に示せる．Gについて，以下の補題が成り立つ．

補題2 グラフ$\hat{G} = (\hat{V}, \hat{E})$がサイズ$\hat{K}$以下の節点被覆を持つならば，集合$G$はサイズ$K$以下の要素辞書$D_G$を持つ． □

[証明] 仮定を満たす$\hat{G}$の節点被覆を$\hat{V}'$とし，以下の要素集合$S \subseteq U$を考える．

$$S = \{v_1, v_2, v_3 | v \in \hat{V}\} \cup \{v_4 | v \in \hat{V}'\}$$

明らかに$|S| = 3|\hat{V}| + \hat{K} = K$である．以下，互いに異なる任意の$g, g' \in G$について，$diff(g, g')$と$S$が共通要素を持つことを示す．$g \neq g'$より$diff(g, g') \neq \phi$である．

$g = g_v^i, g' = g_v^j$ $(i \neq j)$の場合，$diff(g, g') \subseteq \{v_1, v_2, v_3\} \subseteq S$である．

$g = g_v^i, g' = g_u^j$ $(u \neq v)$の場合，$diff(g, g') \subseteq \{v_1, v_2, v_3, u_1, u_2, u_3\} \subseteq S$である．

$g = g_e, g' = g_v^i$の場合，$e = (u, v)$とすると$\{u_4, v_4\} \subseteq diff(g, g')$となる．ここで，$\hat{V}'$は節点被覆のため$u$または$v$が$\hat{V}'$に属する．従って，$u_4$または

v_4はSに属する．

最後に$g = g_e, g' = g_{e'}$ $(e \neq e')$の場合を考え，$e = (u, v), e' = (u', v')$とする．$v = v'$ならば定義より$g \cap \{v_1, v_2, v_3\} = g_v^k$, $g' \cap \{v_1, v_2, v_3\} = g_v^{k'}$ $(k \neq k')$となる．従って，$diff(g_v^k, g_v^{k'}) \subseteq diff(g, g')$となるが，$diff(g_v^k, g_v^{k'}) \subseteq \{v_1, v_2, v_3\} \subseteq S$より$S \cap diff(g, g') \neq \phi$である．$v \neq v'$ならば，$g \cap \{v_1, v_2, v_3\} = g_v^k$, $g' \cap \{v_1, v_2, v_3\} = \phi$より$S \cap diff(g, g') \neq \phi$となる．

□

補題3 集合Gの要素辞書D_Gについて，$\{v_1, v_2, v_3 | v \in V\} \subset D_G$ □

[証明] 任意の$v \in \hat{V}$と$i \in \{1, 2, 3\}$に対して定義より，$diff(g_v^i, g_v^4) = \{v_i\}$が成立する．よって補題1より，$D_G \cap diff(g_v^i, g_v^4) \neq \phi$である．従って，$v_i \in D_G$．

□

補題4 集合Gの要素辞書D_Gについて，$(u, v) \in \hat{E}$ならば，$\{u_4, v_4\} \cap D_G \neq \phi$ □

[証明] $e = (u, v)$に対応する集合g_eに対して$g_v^k \in G$が常に存在し，$diff(g_e, g_v^k) = \{u_4, v_4\}$が成立する．よって補題1より$\{u_4, v_4\} \cap D_G \neq \phi$． □

補題5 集合GがサイズK以下の要素辞書D_Gを持つならば，グラフ$\hat{G} = (\hat{V}, \hat{E})$はサイズ$\hat{K}$以下の節点被覆を持つ． □

[証明] 補題3より$\{v_1, v_2, v_3 | v \in \hat{V}\} \subset D_G$が成立する．ここで$\hat{V}' = \{v | v_4 \in D_G\}$を考えると，$\hat{V}' = D_G - \{v_1, v_2, v_3 | v \in \hat{V}\}$より$|\hat{V}'| = \hat{K}$である．補題4より$\hat{V}'$は$\hat{G}$の節点被覆である． □

以上より所望の定理が証明された．

定理3 最小要素辞書構成問題はNP完全である．

□

4. 要素辞書の再構成アルゴリズム

本節では，任意の集合Gに対して要素数が高々$n-1$の要素辞書を構成するために要素辞書木を導入する．要素辞書木のデータ量は$O(n)$である．また，集合の追加，削除が行なわれた場合に，要素辞書および要素辞書木を多項式時間で更新するためのアルゴリズムを提案する．

4.1 要素辞書木

集合Gに対して要素辞書木T_Gを以下のように定義する．

定義2 集合Gに対する要素辞書木T_Gは以下の条件を満足する木である．

(1) 各中間節点vは要素$e \in U$に対応しており，2個の子節点v_+, v_-を持つ．枝(v, v_+)のラベルはeであり，枝(v, v_-)はラベルを持たない．

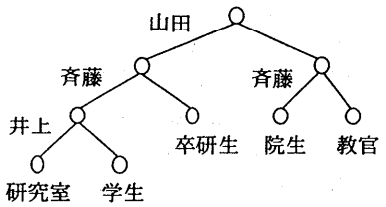

図2 要素辞書木の例

(2) 各葉節点vは$g \in G$と一対一に対応している．根節点からvに至る中間節点に対応する要素の集合をE_vとすると，根節点からvに至る経路上のラベル集合は$g \cap E_v$である．

□

g	E_v	$g \cap E_v$
研究室	{ 山田, 斉藤, 井上 }	{ 山田, 斉藤, 井上 }
学生	{ 山田, 斉藤, 井上 }	{ 山田, 斉藤 }
卒研生	{ 山田, 斉藤 }	{ 山田 }
院生	{ 山田, 斉藤 }	{ 斉藤 }
教官	{ 山田, 斉藤 }	ϕ

表3 要素辞書木の葉節点に対する集合

要素辞書木は2分木のため必ずしも平衡しない．図1の集合Gに対して構成される要素辞書木を**図2**に示す．また，図2における各$g \in G$, E_v, $g \cap E_v$を**表3**に示す．要素辞書木の定義より，以下の補題が成り立つ．

補題6 要素辞書木T_Gのデータ量は$O(n)$である．

□

[証明] 定義2の条件1よりT_Gにおける葉以外の節点数は葉節点数-1である．条件2より葉節点数はnである．従ってT_Gの節点数は$2n-1$である．またT_Gは連結した木であることより，枝数は$2n-2$である．以上より補題が成り立つ． □

補題7 要素辞書木T_Gの中間節点に対応する要素の集合はGの要素辞書D_Gであり，$|D_G| \leq n-1$である． □

[証明] 定義2の条件2より，T_Gを用いると任意の2つの集合$g, g' \in G$について，g, g'に共通する先祖の中で最も下位にある中間節点に対応するeはどちらか一方の集合にのみ含まれる．よって，$e \in D_G \cap diff(g, g') \neq \phi$である．また，$T_G$の中間節点数は$n-1$であることから$|D_G| \leq n-1$である． □

4.2 集合の追加に伴う要素辞書の再構成

要素辞書が定義されている既存の集合Gに新たに要素gを追加する際に，要素辞書中の要素ではgが他の集合と区別できない場合がある．このような場合には適切な要素$e \in U$を求め，要素辞書木を再構成する必要がある．本節では，集合を追加した場合の要素辞

書再構成アルゴリズムを提案する．ここで，T_Gの中間節点に対応する要素集合をD_Gとする．

要素辞書再構成アルゴリズム（集合追加時）

入力：要素辞書木T_G，要素辞書D_G，集合g

出力：要素辞書木T'_G，要素辞書D'_G

(1) 追加する集合gに対して，$g \cap E_k = g_k \cap E_k$を満足する集合$g_k$を検索する．ここで，$E_k$は$T_G$の根節点から$g_k$に対応する葉節点に至る中間節点に対応する要素の集合である．

(2) $e \in diff(g, g_k)$を検索する．

(3) $D'_G \leftarrow D_G \cup \{e\}$．

(4) g_kに対応する葉節点v_kを中間節点とし，eを対応づける．

(5) v_kの子節点v_+, v_-を作成する．枝(v_k, v_+)のラベルはeとする．$e \in g$ $(e \notin g)$ならば，v_+とv_-をそれぞれgおよびg_k(g_kおよびg)と対応づける． □

ステップ1において，g_kを求めるアルゴリズムを以下に示す．

集合検索アルゴリズム

入力：要素辞書木T_G，集合g

出力：$g \cap E_k = g_k \cap E_k$を満足する集合$g_k \in G$

(1) 節点vをT_Gの根節点とする．

(2) vが葉節点ならばvに対応する集合をg_kとする．

(3) vに対応する要素eについて，$e \in g$ならば$v \leftarrow v_+$，そうでなければ$v \leftarrow v_-$とする．

(4) 2へ戻る． □

上記のアルゴリズムで検索されたg_kに対して$g \cap E_k = g_k \cap E_k$が成立することは自明である．この条件を満足する$g_k$がただ1個しか存在しないことは以下の補題で示される．

補題8 要素辞書木T_G中で$g \cap E_k = g_k \cap E_k$を満足する集合$g_k$はただ1個しか存在しない． □

[証明] 2つの集合$g_k, g_{k'}(g_k \neq g_{k'})$について，$g \cap E_k = g_k \cap E_k$かつ$g \cap E_{k'} = g_{k'} \cap E_{k'}$が成立すると仮定する．$g_k \neq g_{k'}$より，$v_k$と$v_{k'}$の共通祖先の中で最も下位の中間節点に対応する要素$e \in D_G$が存在して，$e \in g_k$かつ$e \notin g_{k'}$となる（$e \notin g_k$かつ$e \in g_{k'}$の場合も以下の証明は同様）．このとき，$e \in g_k \cap E_k$かつ$e \notin g_{k'} \cap E_{k'}$である．ここで，$e \in g$の場合を考えると，$e \in g \cap E_k$かつ$e \in g \cap E_{k'}$となり，矛盾が導かれる．$e \notin g$の場合も同様である． □

再構成アルゴリズムのステップ2において，eを求める方法としては次の4つが考えられる．

(1) eをD_G中で検索する．

(2) eをDB中で検索する．

(3) g, g'を与えて，一方にのみ含まれる要素を自動抽出する．

(4) g, g'を利用者に与えて，eを作成させる．

ここで，方法(1)が適用できた場合，要素辞書には新たな要素を追加する必要がない．方法(3)の例としては文献検索におけるキーワード自動抽出アルゴリズム[2),3)]を利用する方法等が考えられる．

ステップ3では要素辞書の更新，ステップ4，5では要素辞書木の再構成をそれぞれ行う．

補題9 要素辞書木T_Gに対して，再構成アルゴリズムを適用して再構成された木T'_Gは定義2を満足する． □

[証明] ステップ5では，新たに中間節点となったv_kに対して要素eを，v_kから出る枝の一方に対してラベルeを対応づけている．従って条件1を明らかに満足する．根節点からv_kに至る経路上の中間節点に対応する要素の集合をE_kとすると，$E_k \cap g = E_k \cap g_k$である．$v_k$に対応する要素$e$について，$e \in diff(g, g_k)$より，$(E_k \cup \{e\}) \cap g \neq (E_k \cup \{e\}) \cap g_k$である．このとき，$e \in g$ $(e \notin g)$ならばv_{k+}とv_{k-}にはそれぞれgとg_k (g_kとg)が対応する．従って，T'_Gは条件2を満足する． □

定理4 任意の$e \in U, g \in G$に対して，$e \in g$が多項式時間で判定できるならば，再構成アルゴリズムを用いたT_GおよびD_Gの再構成（ステップ2を除く）は多項式時間で行える． □

[証明] 定義2より，T_Gの中間節点数は$n-1$であるため，T_Gの高さはnを超えない．これと仮定により，ステップ1は多項式時間で実行できる．ステップ3，4，5は定数時間で行なえる．故に定理が成立する． □

ステップ2において，方法(1)，(2)は仮定より多項式時間で実行できる．これらの方法が失敗したときには，最悪の場合，方法(4)を実行しなければならない．しかし，1個の集合を追加した場合，利用者が作成すべき要素数は高々1個である．

4.3 集合の削除に伴う要素辞書の再構成

集合Gから要素gを削除する際には，D_Gの要素数を$n-1$以下に保つために，集合の識別に不要な要素を削除し，要素辞書木を再構成する必要がある．本節では，集合を削除した場合の要素辞書再構成アルゴリズムを提案する．

例4 集合$G = \{g_1, g_2, g_3\}$に対して**図3**の要素辞書木T_Gを定義する．Gよりg_1を削除した場合，要素e_1はg_1とg_2を識別するためにのみ必要なため削除できる．要素e_2はg_2とg_3の差を表現するために必要なため削除できない． □

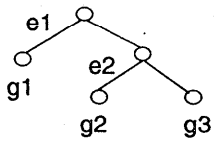

図3 要素辞書からの要素の削除

要素辞書再構成アルゴリズム（集合削除時）

入力： 要素辞書木 T_G，要素辞書 D_G，集合 g

出力： 要素辞書木 T'_G，要素辞書 D'_G

(1) 削除する集合 gに対応する葉節点を vとし，その親節点を v_pとする．

(2) v_pに対応する要素 eと同一の要素が対応する節点がT_G中に存在しないならば $D'_G \leftarrow D_G - \{e\}$．

(3) v以外の v_pの子節点を v'とする．

(4) 節点 v_p, vおよび枝 (v_p, v), (v_p, v') を削除する．

(5) v_pが T_Gの根節点ならば v'を新たな根節点とする．

(6) v_pが根節点でなければ，v_pの親節点 v_{gp}について枝 (v_{gp}, v_p) を (v_{gp}, v') に変更する． □

ステップ1で求めた vの親節点 v_pに対応する要素 eが削除の候補となる．ステップ2では，eが D_Gから削除可能であることを確認した後，削除を実行する．ステップ3から6では要素辞書木の再構成を行なう．ステップ6では枝 (v_{gp}, v_p) のラベルは変更しない．

補題10 要素辞書木 T_Gに対して，再構成アルゴリズムを適用して再構成された木 T'_Gは定義2を満足する． □

[証明] v_pを除く中間節点と枝のラベルは変更されないため，明らかに条件1を満足する．v_pを先祖とする各葉節点 v_iに対応する集合を g_i とする．T_Gおよび T'_Gの根節点から v_iに至る中間節点に対応する要素の集合をそれぞれ E_i，E'_iとする．また，T_Gおよび T'_Gの根節点から v_iに至る経路上のラベル集合をそれぞれ L_i，L'_iとする．ここで，$E'_i = E_i - \{e\}$ より，$g_i \cap E'_i = (g_i \cap E_i) - \{e\} = L_i - \{e\}$ となる．$e \in g_i$ のとき，枝 (v_p, v') が削除されるため，$L'_i = L_i - \{e\}$ である．一方 $e \notin g$のとき，L_iは e を含まないため，$L_i - \{e\} = L'_i$ である．よって，$L'_i = g_i \cap E'_i$である．また，v_pを先祖としない葉節点と T_G の根節点との間の経路は変更されないため，T'_Gは条件2を満足する． □

補題7，9，10より，要素辞書 D_Gの要素数は $n-1$ 以下であることが保証される．

定理5 再構成アルゴリズムを用いた T_Gおよび D_Gの再構成は，$O(n)$ 時間で行える． □

[証明] ステップ1は定数時間で行なえる．中間節点数が $n-1$ であることより，ステップ2は $O(n)$ 時間で行なえる．ステップ3から6までは G の要素数に関わらず定数時間で行なえる． □

5. 要素辞書の適用例

5.1 サンプルを用いた集合検索

本節では，サンプルを用いた集合検索の定義を行ない，これに要素辞書を適用する．

サンプル Sは要素の集合であり，各要素は正例または負例のいずれかが明示されている．サンプル Sが集合 gを検索するとき，S中のすべての正例 e に対し $e \in g$，かつ S中のすべての負例$\overline{e}$に対し $\overline{e} \notin g$ である．$G = \{g_1, g_2, \cdots, g_n\}$ について，$Set(G, S) = \{g \in G | S$は gを検索する $\}$ を定義する．以下にサンプルを用いた集合検索の例を示す．

例5 図1の集合 $G =$ { 学生, 卒研生, 院生, 研究室, 教官 } に対して，サンプル $S_1 =$ { 山田 } を用いて検索を行なうと，$Set(G, S_1) =$ { 学生, 卒研 生, 研究室 } となる．G に対して，S_1に‘斉藤’を追加したサンプル $S_2 =$ { 山田，斉藤 } を用いて検索すると，$Set(G, S_2) =$ { 学生, 研究室 } となる．さらに負例‘$\overline{\text{森}}$’を追加したサンプル $S_3 =$ { 山田，斉藤，$\overline{\text{森}}$ } を用いて検索すると，$Set(G, S_3) =$ { 研究室 } となり，集合が唯一に決定する． □

上記の方式に要素辞書を適用する．要素辞書を D_Gとすると定義より，任意の集合 g, g' $(g \neq g')$ について，$D_G \cap g \neq D_G \cap g'$である．よって，実体 $e \in D_G \cap g$ を正例，$e \in D_G - g$を負例とするサンプル Sを定義すると，$Set(G, S) = \{g\}$ である．従って，要素辞書を用いると任意の集合 gについて，gを特定するサンプルを常に作成できる．また補題1より，与えられた任意の集合 g, g'間の相違に関する質問に常に答えられる．

要素辞書木の任意の節点 vについて以下に定義するサンプル S_vを定義する．S_vは根節点から vに至る経路上の中間節点に対応する要素の集合であり，枝 (v_i, v_j) がラベルを持つならば v_iは正例，ラベルを持たないならば v_iは負例とする．このとき，補題11が成立する．

補題11 節点 vと vに対して定義されたサンプル S_vに対して，$Set(G, S_v) = \{v$を先祖とする葉節点に対応する集合 $\}$ が成立する． □

[証明] まず，S_vは vを先祖する葉節点 u に対応する集合 gを検索することを示す．根節点から u に至る経路上の中間節点に対応する要素の集合を E_u，根節点から v, u に至る経路上のラベル集合をそれぞれ L_v, L_u とする． $S_v \cap L_u$は L_uの要素中で根節点から vに至る経路上の節点集合に対応するため，$S_v \cap L_u = L_v$

が成立する．よって，$L_v \subseteq L_u$である．要素辞書木の定義より，$g \cap E_u = L_u$であることから，$e \in L_u$に対して$e \in g$である．同様に，$e \in L_v$に対しても$e \in g$となる．また，$g \cap E_u = L_u$より$e \in E_u - L_u$に対して$e \notin g$であることが分かる．$S_v \subseteq E_u$より，$S_v - L_u \subseteq E_u - L_u$が成り立つ．$S_v \cap L_u = L_v$より$S_v - L_u = S_v - L_v$である．従って，$e \in S_v - L_v$に対して$e \notin g$となる．以上より，$S_v$は$g$を検索する．

次に，S_vはvを先祖としない葉節点uに対応する集合gを検索しないことを示す．uとvの最も近い共通先祖をwとする．仮定より$w \neq v$である．wがS_vの正例のとき，根節点からuに至る経路上にはラベルwの枝が存在しない．従って，$w \notin g \cap E_u$かつ$w \in E_u$より$w \notin g$である．よって，S_vはgを検索しない．wがS_vの負例の時もS_vはgを検索しないことが同様に示せる． □

目的の集合を検索するためのサンプルの作成は利用者が行なう．集合数をnとすると，要素辞書を用いることで，高々$n-1$個の要素の中から目的の集合を検索するサンプルが常に作成できる．

5.2 キーワード辞書構成問題

文献検索におけるキーワード辞書には，文献を登録した際に抽出されたキーワードが登録される．キーワード抽出を行うためには，手作業の他に自然言語処理や重要度等を用いた自動化手法も知られている[2),3)]．しかし，これらの手法では，文献間の差を表現するキーワードが常に抽出されるとは限らない．従って，検索対象の文献を唯一に識別できない場合がある．また，抽出されたすべてのキーワードを登録するため，文献数の増加に伴いキーワード辞書のサイズが増大する．

利用者による検索効果を高めるためにシソーラスが提案されている．シソーラスは文献に付加されるキーワードと利用者が指定する検索語 (検索式中に出現する語) の間の関連を保持する．このためシソーラスのサイズは検索語数とキーワード数の積に比例する．また，文献検索の高速化のためにシグネチャファイルや転置ファイルなどの索引が提案されている．シグネチャファイルは，文献とそれに対して定義されるシグネチャの対応を保持する．文献のシグネチャは，文献に付加されたキーワード毎に定義されたシグネチャの論理和である．転置ファイルは，キーワードとそれを付加する文献との対応を保持する．これらの索引のサイズは文献数とキーワード数の積に比例する☆．以上の考察により，キーワード辞書のサイズが増大すると，シソーラスや索引のサイズも大きくなる．

要素辞書をキーワード辞書として用いることで，これらの問題を解決できる．抽出されたキーワードをキーワード辞書に登録する際に，要素辞書再構成アルゴリズムを用いると，高々1個のキーワードを追加するだけで常に任意の文献を唯一に識別するキーワード辞書が構築できる．文献数nに対して構成されたキーワード辞書のサイズは高々$n-1$であるため，シソーラスや索引のサイズも小さくできる．

☆ 文献のシグネチャサイズは，フォルスドロップ確率を一定に保った場合，キーワード数に比例する．

6. おわりに

本論文では，集合間の相違を明確にするための要素辞書を定義し，その性質を調べた．最小の要素辞書構成問題はNP完全になるが，n個の集合Gに対して，要素数が$n-1$以下の要素辞書が多項式時間で構成できる．

今後の課題としては，要素辞書を用いた集合検索支援機構の構築が挙げられる．また，文献検索以外でのアプリケーションにおいて，$e \in diff(g, g')$を自動抽出するための研究を行なう予定である．これが可能になれば，集合の追加に伴う要素辞書の再構成を自動化できる．

謝辞 本研究の一部は文部省科学研究費重点領域研究（課題番号 08244105）の援助を受けている．

参考文献

1) 伊藤哲郎: 情報検索, 昭晃堂 ソフトウエア講座 19 (1986).
2) 大本晴夫: 日本語新聞記事からのキーワード自動抽出と重要度評価, 電子情報通信学会論文誌, Vol. J74-D-I, No. 8, pp. 556–566 (1991).
3) 原正巳, 中島浩之, 木谷強: テキストのフォーマットと単語の範囲内重要度を利用したキーワード検出, 情報処理学会論文誌, Vol. 38, No. 2, pp. 299–309 (1997).
4) Fayyad, U. and R. Uthurusamy, E.: Data mining and knowledge discovery in databases, *Comm. ACM*, Vol. 39, No. 11, pp. 24–68 (1996).
5) Chen, M.-S. et al.: Data mining: An overview from a database perspective, *IEEE Trans. Knowledge and Data Eng.*, Vol. 8, No. 6, pp. 866–883 (1996).
6) 村田美友紀, 掛下哲郎: データベース化されたビューに対するサンプル を用いた検索, 電子情報通信学会 DEWS'97 論文集 (1997).
7) 村田美友紀, 掛下哲郎: サンプルを用いた論理式検索機構の評価, 情報処理学会 DBS 研報 113–3 (1997).

8) Faloutsos, C. and Chistodoulakis, S.: Signature files: an access method for documents and its analytical performance evaluation, *ACM Trans. Database Syst.*, Vol. 2, No. 4, pp. 267–288 (1984).
9) Ishikawa, Y. et al.: Evaluation of signature files as set access facilities in OODBs, *Proc. ACM SIGMOD*, pp. 247–256 (1993).
10) Garey, M. R. and Johnson, D. S.: *Computers and Intractability: A Guide to the Theory of NP Completeness*, Freeman (1979).

(平成10年 9 月 20 日受付)
(平成10年12月 27 日採録)

(担当編集委員 細野 公男)

村田美友紀（正会員）

昭和 46 年生．平成 6 年佐賀大学情報科学科卒．平成 8 年同博士前期課程修了．修士（理学）．同年 八代工業高等専門学校 情報電子工学科助手．非手続的データベース操作言語の研究に従事．

掛下 哲郎（正会員）

昭和 37 年生．昭和 59 年 九州大学情報工学科卒．平成元年 同博士後期課程修了．工学博士．同年 佐賀大学理工学部講師を経て，現在，助教授．データベースおよびソフトウェア工学の研究に従事．本学会の他，電子情報通信学会，ACM 等会員．