データベースシステム 92－5
（1993. 3. 22）

シグネチャファイルによる集合値検索の性能評価

福島慶明* 北川博之† 石川佳治‡ 大保信夫†

*筑波大学 理工学研究科
†筑波大学 電子情報工学系
‡筑波大学 工学研究科

データベース応用の多様化に伴い、オブジェクト指向モデルや入れ子型リレーショナルモデル等複雑なデータ構造を取り扱えるデータモデルが注目されている。これらのデータモデルを支援するデータベースシステムでは集合値検索の効率的処理が極めて重要である。我々はシグネチャファイルを利用した集合値検索の処理方式について研究を行なってきた。シグネチャファイルに基づく検索ではフォルスドロップと呼ばれるミスマッチが必然的に発生するため、シグネチャファイルの設計においてはフォルスドロップ確率の見積りが必要となる。

本論文では、subset、superset、intersection、equivalenceの4種類の基本的な集合値検索を対象として、幾つかの状況における確率論的なフォルスドロップ確率の見積り式を導出し、シミュレーションによりその妥当性を評価する。また、データ要素の出現頻度が一様でない場合等、確率論的な見積り式が導出できない、より複雑な場合におけるフォルスドロップ確率をシミュレーションにより解析する。

False Drop Estimation for Set-valued Object Retrieval by Signature Files

Yoshiaki Fukushima* Hiroyuki Kitagawa† Yoshiharu Ishikawa‡ Nobuo Ohbo†

*Master's Degree Program in Sciences and Engineering, University of Tsukuba
†Institute of Information Sciences and Electlonics, University of Tsukuba
‡Doctoral Degree Program in Engineering, University of Tsukuba
Tsukuba City, Ibaraki 305, Japan

Advaned database systems must support complex data structures treated in object-oriented models and nested relational models. In particular, efficient processing of set-valued object retrieval is indispensable for such systems. We have proposed the use of signature files for efficient set-valued object retrieval. Retrieval with signature files is always accompanied by mismatches called false drops, and it is very impotant in designing signature files to properly control the false drops.

In this paper, we derive probabilistic formulas estimating false drop probabilities in four types of set-valued object retrieval based on the subset, superset, intersection, and equivalence relationships. Then we evaluate their validity by computer simulations. Simulation study is also done to investigate false drop probabilities in more complex situations where probabilistic estimation is difficult.

1 序論

近年データベースの適用領域の拡大に伴い、複雑な対象をデータベース化する要求が高まっている。この要求に対応するために、入れ子型リレーショナルモデルやオブジェクト指向モデル等の研究がなされている。これらの新しいモデルでは階層構造や集合値を用いてより複雑な構造を持つデータを表現できるが、データ構造の複雑化に対応し検索処理の効率化を図る必要性がある。これまで階層構造を考慮した索引 [Bert 89] が提案されてきたが、集合値検索に対する効率的な手法は提案されてこなかった。我々はシグネチャファイル (signature file) を用いた集合値検索の手法を提案し、包含関係 ($\subseteq$) に関する集合値検索に対してその有効性を示した [Fuku 92, Ishi 93]。

シグネチャファイルは元来テキスト検索においてあるワードを含むテキストブロックを検索するために考案されたものであり [Falo 84, Chan 89]、その後レコード検索や Prolog 節検索をシグネチャで処理しようとする手法が提案されている [Sack 87, Wong 91]。しかし、シグネチャファイルによる各種の集合値検索に関する研究はこれまで十分行なわれてこなかった。シグネチャファイルを用いた検索ではフォルスドロップ (*false drop*) と呼ばれるミスマッチが存在するため、シグネチャファイル設計上フォルスドロップの見積りが極めて重要である。本研究ではシグネチャファイルを各種集合値検索に適用するが、その場合でもフォルスドロップの正しい見積りは効率的なシグネチャファイル設計のために欠くことができない。

本論文では、集合値属性に対する subset、superset、intersection、equivalence の 4 種類の検索を対象として、シグネチャファイルを用いた集合値検索のフォルスドロップ確率の確率論的な見積り式を導出する。またシミュレーションによりこれら見積り式の妥当性を検証する。更に、確率的な見積りが得られないような現実のより複雑な状況に対してもシミュレーションを行ないフォルスドロップ確率を分析する。

シグネチャファイルの物理的な構成法は種々提案されているが [Kent 88, Chan 89]、[Ishi 93] においてビットスライストシグネチャファイル (*bit-sliced signature file*) の集合値検索における有効性が示されている。そこで本論文中のシミュレーションでは、ビットスライストシグネチャファイルを想定した場合のパラメータ設定における結果を示す。

本論文は以下のように構成されている。2節では集合値検索のシグネチャファイルによる処理を示す。また 3節では 4 種類の集合値検索のフォルスドロップ確率の確率論的な見積り式を導出し、これを元に 4節においてビットスライストシグネチャファイルを想定したシミュレーションによりこの見積り式の妥当性の評価を行なう。更に確率的にはフォルスドロップ確率が求められないような状況に対してもシミュレーションを行ない、考察する。最後の 5節は結論である。

2 シグネチャファイルを用いた集合値検索

2.1 集合値検索

集合値を扱う代表的データモデルの一つとして、入れ子型リレーショナルモデルがある。入れ子型リレーションに対するデータ操作を記述するのに幾つかの入れ子型リレーショナル代数 [Kita 89] が提案されている。入れ子型リレーショナル代数の選択演算 (selection) では通常以下のような集合値に関する検索条件を指定することができる。以下検索の対象となる集合 (ターゲット集合 (*target set*)) を T で表し、検索条件として与える単純値を q、検索条件として与える集合 (問い合わせ集合 (*query set*)) を Q で表す。例としては図 1の入れ子型リレーション中の集合値属性 INAME に対する集合値検索の例を示す。

1. $q \in T$: ターゲット集合が、問い合わせで与えられた単純値を含む (*membership*)。
   Q1 : `pencil` を取り扱う部門を求める。
2. $Q \subseteq T$: ターゲット集合が、問い合わせ集合を含む (*subset*)。
   Q2 : `pen` と `pencil` のいずれのアイテムも取り扱う部門を求める。
3. $Q \supseteq T$: ターゲット集合が、問い合わせ集合に含まれる (*superset*)。
   Q3 : `pen`、`pencil`、`eraser`、`cutter` 以外のアイテムを取り扱っていない部門を求める。
4. $(Q \cap T) \neq \phi$: ターゲット集合が、問い合わせ集合と共通な要素を持つ (*intersection*)。
   Q4 : `eraser` と `stapler` のいずれかを取り扱う部門を求める。
5. $Q \equiv T$: ターゲット集合が、問い合わせ集合と等価である (*equivalence*)。
   Q5 : `pen`、`pencil`、`ink` だけを取り扱う部門を求める。

$q \in T$(membership) に関する検索は $Q \subseteq T$(subset) に関する検索の特別な場合であるので、以後後者に含めて考える。

2.2 シグネチャファイル

図 2に示すように、データベース中の検索対象となるターゲット集合に対する集合シグネチャ (ターゲットシグネチャ (*target signature*)) を生成する。まず、ターゲット集合中の各要素をハッシングして要素シグネチャ (*element signature*) と呼ぶ 2 進のビットパターンに符合化する。全ての要素シグネチャは長さ F ビットであり、F ビット中 m ビットが "1"にセットされる。以下あるシグネチャ中で "1"がセットされているビット数をシグネチャのウエイト (*weight*) と呼び、シグネチャのウエイト $\div F$ をシグネチャの密度 (*density*)

D:

{DEPT}			
TID	DNO	MNAME	{SALES_ITEM}
			INAME
T1	314	tanaka	pen pencil ink
T2	125	suzuki	notebook eraser clip cutter

図 1: 入れ子型リレーションの例

element		element signature		
pen	→	0001000000000101		
pencil	→	1100100000000000		
ink	→	0100001010000000		
		⇓		TID
set signature		1101101010000101	→	T1

図 2: 集合シグネチャの生成

と呼ぶ。集合の全ての要素の要素シグネチャのビット毎の論理和をとる (スーパーインポーズドコーディング) ことにより集合シグネチャを生成する。ターゲット集合に対する集合シグネチャと集合が属するタプルのタプルID を組にして格納したものがシグネチャファイルである (図 3)。

2.3 検索方法

2.1節で述べた各種の集合値検索は、以下の手順で処理可能である。

1) 問い合わせ集合に対する集合シグネチャ (問い合わせシグネチャ (*query signature*)) をターゲットシグネチャと同様の方法で生成する。

2) シグネチャファイルをスキャンし、ターゲットシグネチャが以下の条件を満たす時に該当するタプルが検索条件を満たす候補となる。

   $Q \subseteq T$
   問い合わせシグネチャ ∧ ターゲットシグネチャ
   = 問い合わせシグネチャ

   $Q \supseteq T$
   問い合わせシグネチャ ∧ ターゲットシグネチャ
   = 問い合わせシグネチャ

   $(Q \cap T) \neq \phi$
   $weight$(問い合わせシグネチャ∧ターゲットシグネチャ)
   $\geq m$
   ただし $weight()$ はシグネチャのウエイトを返す関数とする。

   $Q \equiv T$
   問い合わせシグネチャ = セットシグネチャ

3) 2) で候補として選択されたタプルが実際に問い合わせで与えられた条件を満足するか調べる (false drop resolution)。

図 4に **Q2** の問い合わせのシグネチャファイルによる処理を示す。

1101101010000101	T1
1110110111000010	T2
⋮	⋮

図 3: シグネチャファイル

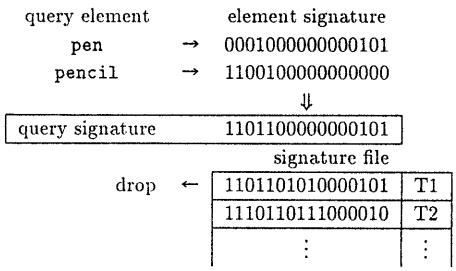

図 4: **Q2** の問い合わせの処理

3 集合値検索におけるフォルスドロップ確率

シグネチャファイル法ではフォルスドロップと呼ばれるミスマッチが存在する。これは異なる要素を同じ要素シグネチャにハッシングしたり、図 5に示すように要素シグネチャをスーパーインポーズドコーディングすることにより起こる。

フォルスドロップ確率を Fd で表し、M をターゲット集合の総数、M_a を実際に問い合わせ条件を満たすターゲット集合 (*actual drop*) の数、M_f をフォルスドロップとなるターゲット集合の数とする。この時、

$$Fd = \frac{M_f}{M - M_a} \tag{1}$$

によりフォルスドロップ確率が定義される [Falo 84]。

シグネチャファイルを用いた検索ではページアクセスの回数を抑えるために、フォルスドロップ確率を適正に制御することが必要となる。従ってフォルスドロップ確率を見積もることはシグネチャファイルの集合値検索に対する性能評価において非常に大きな意味を持つ。またフォルスド

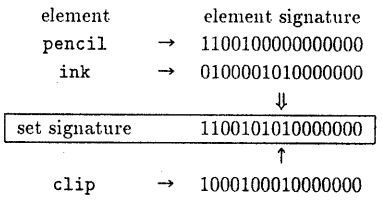

図 5: フォルスドロップの例

ロップ確率はシグネチャのビット長とも関連しデータ格納時のオーバーヘッドに直接影響する。

3.1 準備

以下の解析で用いる変数を表 1に示す。以下では次の仮定の下で、フォルスドロップ確率の見積り式を導出する。

1. $m \ll F$。
2. 要素シグネチャの各ビット位置は互いに独立であり、等しい確率で 1 がセットされる。

ターゲットシグネチャのあるビット b_t が"1"である確率を $p(b_t)$ とする。この時 $p(b_t)$ は、

$$p(b_t) = 1 - \left(1 - \frac{m}{F}\right)^{D_t} \simeq 1 - e^{-\frac{mD_t}{F}}$$

により与えられる [Falo 84]。同様に問い合わせシグネチャのあるビット b_q が"1"である確率を $p(b_q)$ とすると、

$$p(b_q) = 1 - \left(1 - \frac{m}{F}\right)^{D_q} \simeq 1 - e^{-\frac{mD_q}{F}}$$

で与えられる。F、m、D_t、D_q を一定とした時ターゲットシグネチャ及び問い合わせシグネチャのウェイトの期待値は、

ターゲットシグネチャ： $N_t = F \times p(b_t)$
問い合わせシグネチャ： $N_q = F \times p(b_q)$

となる。

$q \in T$(membership) の検索におけるフォルスドロップ確率は [Falo 84] において求められている。この場合 q は単純値なので、問い合わせシグネチャでは m ビットが 1 にセットされる。よって、

$$Fd_{\{q \in T\}} = (p(b_t))^m = \left(1 - e^{-\frac{mD_t}{F}}\right)^m \quad (2)$$

が得られる。この時フォルスドロップ確率は、

$$m = \frac{F \ln 2}{D_t} (= m_{opt})$$

の時に極小値をとる [Stia 60]。

表 1: 変数のリスト

記号	定義
F	シグネチャのサイズ (ビット長)
m	要素シグネチャのウェイト
D_t	ターゲット集合 T の要素数
D_q	問い合わせ集合 Q の要素数
M	データベース中のタプルの総数
V	ターゲット集合の要素となり得るデータの種類の総数

3.2 ターゲット集合の要素数が一定な場合

まず、各ターゲット集合の要素数 D_t が全てのターゲット集合に対して一定の場合について、4 種類の集合値検索におけるフォルスドロップ確率の見積り式を導出する。

まず以下で利用する近似式を示す。$b_t^j (1 \leq j \leq F)$ によりターゲットシグネチャの j 番目のビット位置を表し、$b_q^j (1 \leq j \leq F)$ により問い合わせシグネチャの j 番目のビット位置を表すものとする。$1 \leq i \leq F - m$ なる i に対して、

$$\begin{aligned} Prob(b_t^1 = 0 \wedge \cdots \wedge b_t^i = 0) \quad (3) \\ = \left(\frac{{}_{F-i}C_m}{{}_{F}C_m}\right)^{D_t} \\ = \left(\frac{(F-m)(F-m-1)\cdots(F-m-i+1)}{F(F-1)\cdots(F-i+1)}\right)^{D_t} \\ = \prod_{k=1}^{i}\left(1 - \frac{m}{F-k+1}\right)^{D_t} \quad (4) \end{aligned}$$

$1 \leq k \leq i$ に対して $\frac{m}{F-k+1} \ll 1$ が成立する場合、

$$\begin{aligned} Prob(b_t^1 = 0 \wedge \cdots \wedge b_t^i = 0) \\ \simeq \left(1 - \frac{1}{F}\right)^{mD_t}\left(1 - \frac{1}{F-1}\right)^{mD_t}\cdots\left(1 - \frac{1}{F-i+1}\right)^{mD_t} \\ = \left(1 - \frac{i}{F}\right)^{mD_t} \quad (5) \end{aligned}$$

同様に、

$$Prob(b_q^1 = 0 \wedge \cdots \wedge b_q^i = 0) = \prod_{k=1}^{i}\left(1 - \frac{m}{F-k+1}\right)^{D_q} \quad (6)$$

$1 \leq k \leq i$ に対して $\frac{m}{F-k+1} \ll 1$ が成立する場合、

$$Prob(b_q^1 = 0 \wedge \cdots \wedge b_q^i = 0) \simeq \left(1 - \frac{i}{F}\right)^{mD_q} \quad (7)$$

以上を踏まえた上でフォルスドロップ確率について考える。

3.2.1 見積り式の導出 (1)

まず全てのターゲットシグネチャのウェイトと、全ての問い合わせシグネチャのウェイトがそれぞれその期待値 $\bar{N}_t$ 及び $\bar{N}_q$ とほぼ等しいとみなせるものとした時の、フォルスドロップ確率の見積り式を導出する。

(1) $Q \subseteq T$

フォルスドロップが生じるのは $1 \leq j \leq F$ に対して、

$$b_t^j = 0 \text{ ならば } b_q^j = 0$$

が成立する場合である。ターゲットシグネチャでは $F - \bar{N}_t$ ビットが 0 であるのでフォルスドロップ確率は (6) 式より、

$$\begin{aligned} Fd_{\{Q \subseteq T\}} &= Prob(b_q^1 = 0 \wedge \cdots \wedge b_q^{F-\bar{N}_t} = 0) \\ &= \prod_{k=1}^{Fe^{-\frac{mD_t}{F}}} \left(1 - \frac{m}{F-k+1}\right)^{D_q} \end{aligned}$$

$1 \leq k \leq Fe^{-\frac{mD_t}{F}}$ に対して $\frac{m}{F-k+1} \ll 1$ が成立する場合には (7) 式より、

$$Fd_{\{Q \subseteq T\}} \simeq (1 - e^{-\frac{mD_t}{F}})^{mD_q} \quad (8)$$

となる。$Q \subseteq T$ の問い合わせにおいてもフォルスドロップ確率は $m = m_{opt}$ の時に極小となる。

(2) $Q \supseteq T$

フォルスドロップが生じるのは $1 \leq j \leq F$ に対して、

$$b_q^j = 0 \text{ ならば } b_t^j = 0$$

が成立する場合である。質問シグネチャでは $F - \bar{N}_q$ ビットが "0" であるのでフォルスドロップ確率は (4) 式より、

$$\begin{aligned} Fd_{\{Q \supseteq T\}} &= Prob(b_t^1 = 0 \wedge \cdots \wedge b_t^{F-\bar{N}_q} = 0) \\ &= \prod_{k=1}^{Fe^{-\frac{mD_q}{F}}} \left(1 - \frac{m}{F-k+1}\right)^{D_t} \end{aligned}$$

$1 \leq k \leq Fe^{-\frac{mD_t}{F}}$ に対して $\frac{m}{F-k+1} \ll 1$ が成立する場合には (5) 式より、

$$Fd_{\{Q \supseteq T\}} \simeq (1 - e^{-\frac{mD_q}{F}})^{mD_t} \quad (9)$$

となる。フォルスドロップ確率の極小値を与える m の値は、

$$m = \frac{Fln2}{D_q}$$

により与えられる。

(3) $(Q \cap T) \neq \phi$

フォルスドロップが生じるのは $1 \leq j \leq F$ に対して、

$$b_q^j = 1 \text{ かつ } b_t^j = 1$$

となるビット j が少なくとも m ビット存在する時である。言い換えると問い合わせシグネチャで "1" が立っているビット位置と同一の位置にターゲットシグネチャにおいて m ビット未満 "1" が立っている時以外にフォルスドロップとなる。いま問い合わせシグネチャで "1" が立っているビット位置と同一のターゲットシグネチャのビット位置に "1" が立つ確率がそれぞれ $p(b_t)$ で与えられるものとみなす。この時、問い合わせシグネチャとターゲットシグネチャの同一の位置に k ビット "1" が立つ確率は、

$${}_{N_q}C_k p(b_t)^k (1 - p(b_t))^{N_q - k}$$

により与えられるので、結局フォルスドロップ確率は、

$$Fd = 1 - \sum_{k=0}^{m-1} {}_{N_q}C_k p(b_t)^k (1 - p(b_t))^{N_q - k} \quad (10)$$

により与えられる。

(4) $Q \equiv T$

この場合 $D_t = D_q$ でありそれぞれ $\bar{N}_t = \bar{N}_q = N$ ビットが "1" にセットされているものとする。この時 F ビット中で N ビットが "1" であるシグネチャの組合せは、${}_FC_N$ により与えられる。よって二つのシグネチャが同一である確率、つまりフォルスドロップ確率は、

$$Fd_{\{Q \equiv T\}} = \frac{1}{{}_FC_N} \quad (11)$$

により与えられる。この場合 $N = F/2$ の時つまり $m = m_{opt}$ の時に ${}_FC_N$ の値が最大値となり、従ってフォルスドロップ確率が極小になる。

3.2.2 見積り式の導出 (2)

本節では、3.2.1節で導出した見積り式に、ターゲットシグネチャ、問い合わせシグネチャのウエイトの分布を考慮したフォルスドロップの見積り式を導出する。ターゲットシグネチャで各ビットが $p(b_t)$ の確率で "1" がセットされるものと仮定すると、ターゲットシグネチャのウエイトの分布は 2 項分布 $B(F, p(b_t))$ となり、あるターゲットシグネチャのウエイトが i である確率 $p_t(i)$ は、

$$p_t(i) = {}_FC_i p(b_t)^i (1 - p(b_t))^{F-i}$$

により近似できる。同様に問い合わせシグネチャのウエイトが i である確率 $p_q(i)$ は、

$$p_q(i) = {}_FC_i p(b_q)^i (1 - p(b_q))^{F-i}$$

により近似できる。

以下では、上記の近似によりシグネチャのウエイトの分布を考慮した見積り式を導出する。

(1) $Q \subseteq T$

ターゲットシグネチャのウエイトが i である時フォルスドロップ確率は (6) 式より、

$$\begin{aligned} f_t(i) &= Prob(b_q^1 = 0 \wedge \cdots \wedge b_q^{F-i} = 0) \\ &= \prod_{k=1}^{F-i} \left(1 - \frac{m}{F-k+1}\right)^{D_q} \end{aligned}$$

$1 \leq k \leq F - i$ に対して $\frac{m}{F-k+1} \ll 1$ が成立する場合には (7) 式より、

$$f_t(i) \simeq (1 - \frac{F-i}{F})^{mD_q} = \left(\frac{i}{F}\right)^{mD_q}$$

となる。ターゲットシグネチャのウエイトの分布は2項分布 $B(F,p(b_t))$ となるので、結局フォルスドロップ確率は、

$$Fd_{\{Q\subseteq T\}} = \sum_{i=1}^{F} p_t(i)f_t(i)$$
$$= \sum_{i=1}^{F} {}_FC_i(1-e^{-\frac{mD_t}{F}})^i e^{-\frac{mD_t}{F}(F-i)}\left(\frac{i}{F}\right)^{mD_q} \quad (12)$$

により与えられる。

(2) $Q \supseteq T$

問い合わせシグネチャのウエイトが i である時フォルスドロップ確率は(4)式より、

$$f_q(i) = Prob(b_i^1 = 0 \wedge \cdots \wedge b_i^{F-i} = 0)$$
$$= \prod_{k=1}^{F-i}\left(1-\frac{m}{F-k+1}\right)^{D_t}$$

$1 \le k \le F-i$ に対して $\frac{m}{F-k+1} \ll 1$ が成立する場合には(5)式より、

$$f_q(i) \simeq (1-\frac{F-i}{F})^{mD_t} = \left(\frac{i}{F}\right)^{mD_t}$$

となる。問い合わせシグネチャのウエイトの分布は2項分布 $B(F,p(b_q))$ となるので、結局フォルスドロップ確率は、

$$Fd_{\{Q\supseteq T\}} = \sum_{i=1}^{F} p_q(i)f_q(i)$$
$$= \sum_{i=1}^{F} {}_FC_i(1-e^{-\frac{mD_q}{F}})^i e^{-\frac{mD_q}{F}(F-i)}\left(\frac{i}{F}\right)^{mD_t} \quad (13)$$

により与えられる。

(3) $(Q \cap T) \neq \phi$

問い合わせシグネチャのウエイトが i である時フォルスドロップ確率は、

$$Fd = 1 - \sum_{k=0}^{m-1} {}_iC_k p(b_t)^k (1-p(b_t))^{i-k}$$

により与えられる。問い合わせシグネチャのウエイトの分布は2項分布 $B(F,p(b_q))$ となるので、結局フォルスドロップ確率は、

$$Fd_{\{(Q\cap T)\neq\phi\}} = \sum_{i=1}^{F} {}_FC_i p(b_q)^i (1-p(b_q))^{F-i}$$
$$\times\left(1-\sum_{k=0}^{m-1} {}_iC_k p(b_t)^k (1-p(b_t))^{i-k}\right) \quad (14)$$

により与えられる。

(4) $Q \equiv T$

フォルスドロップが生じるのはターゲットシグネチャのウエイトが i であり、かつ問い合わせシグネチャのウエイトが i である時にターゲットシグネチャと問い合わせシグネチャが同一のシグネチャである時である。従って $Q \equiv T$ の問い合わせにおけるフォルスドロップ確率は、

$$Fd_{\{Q\equiv T\}} = \sum_{i=1}^{F} p_t(i)p_q(i)\frac{1}{{}_FC_i}$$
$$= \sum_{i=1}^{F} {}_FC_i(1-e^{-\frac{mD_t}{F}})^i(1-e^{-\frac{mD_q}{F}})^i e^{-\frac{m(D_t+D_q)}{F}(F-i)} \quad (15)$$

により与えられる。

3.3 ターゲット集合の要素数が一定でない場合

本節では、ターゲット集合の要素数 D_t が一定ではなく、V 中の各要素が一定の確率で各ターゲット集合中に出現し、D_t の平均値は $\bar{D}_t$ により与えられる場合を考える [Iwai92]。この時各要素は $p = \bar{D}_t/V$ の一定な出現確率を持ち、ターゲット集合の要素数の分布は2項分布 $B(V,p)$ となる。

この場合、ターゲット集合の要素数の分布を考慮した、

$$Fd_b = \sum_{D_t=1}^{V} {}_VC_{D_t} p^{D_t}(1-p)^{V-D_t} Fd \quad (16)$$

により前記の4種類の集合値検索におけるフォルスドロップ確率を見積もることができると考えられる。ただし、Fd は3.2節で導出した4種類の集合値検索におけるそれぞれのフォルスドロップ確率の見積り式を表す。

4 シミュレーションによる検証

ここでは3節で述べたフォルスドロップ確率の見積り式の妥当性をシミュレーションにより検証すると共に、F の値の変化に対するフォルスドロップ確率の変化を調べる。

シミュレーションでは $M = 10000$、$V = 10000$ とし、また m の値はビットスライストシグネチャファイルを想定して $m = 2$ とする [Ishi 93]。$Q \subseteq T (q \in T$ を含む)、$Q \supseteq T$、$(Q \cap T) \neq \phi$ の3種類の問い合わせは、$D_t = 10$、$D_t = 100$ の二つの場合に対する結果を示す。また、$Q \equiv T$ の問い合わせは $D_t = 1$、3、5、10、100 に対する結果を示す。以下、図中においては、シミュレーション結果を sim により示し、3.2.1節の見積り式による見積り値を exp1 により示し、3.2.2節の見積り式による見積り値を exp2 により示す。

4.1 ターゲット集合の要素数が一定な場合

3.2節に対応して各ターゲット集合の要素数 D_t が一定の場合を考える。

(1) $Q \subseteq T$

シミュレーション結果と(8)式、(12)式による見積り値を、$D_t = 10$ の場合を図6に、$D_t = 100$ の場合を図7にそれぞれ示す。これより(8)式と(12)式による見積り値に

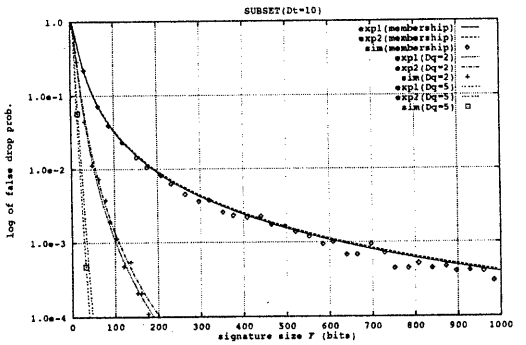

図 6: $Q \subseteq T$(含$q \in T$)($D_t = 10$)

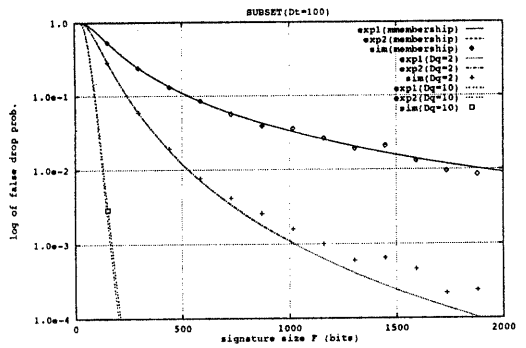

図 7: $Q \subseteq T$(含$q \in T$)($D_t = 100$)

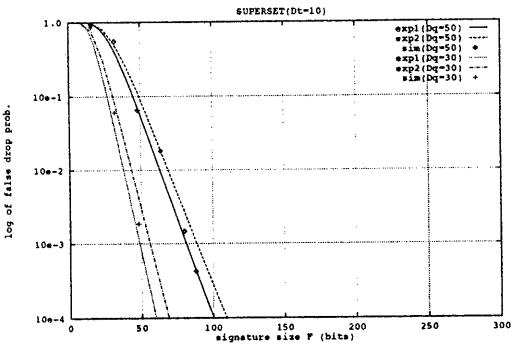

図 8: $Q \supseteq T$($D_t = 10$)

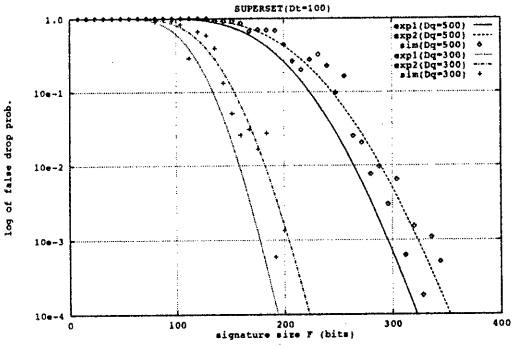

図 9: $Q \supseteq T$($D_t = 100$)

変化はほとんどなく、いずれの式によってもほぼ正確にフォルスドロップ確率を予測することができることが分かる。

これより、より単純な (8) 式を用いてフォルスドロップ確率を算出可能であると考えられる。

(2) $Q \supseteq T$

シミュレーション結果と (9) 式、(13) 式による見積り値を、$D_t = 10$ の場合を図 8に、$D_t = 100$ の場合を図 9にそれぞれ示す。(9) 式によるフォルスドロップ確率の見積り値は、シミュレーション結果と比較してフォルスドロップ確率を低く見積る傾向があることが分かる。一方、(13) 式によるフォルスドロップ確率の見積り値はほぼ正確にシミュレーション結果と一致する。

これより、より精度の高い (13) 式によりフォルスドロップ確率を算出するのが適当であると考えられる。

(3) $(Q \cap T) \neq \phi$

シミュレーション結果と (10) 式、(14) 式による見積り値を $D_t = 10$ の場合を図 10に、$D_t = 100$ の場合を図 11に示す。(10) 式の見積り式はシミュレーション結果とほぼ正確に一致する。一方 (14) 式は D_q が小さい時は誤差が大きい。

(4) $Q \equiv T$

シミュレーション結果と (11) 式、(15) 式による見積り値を $D_t = 1$、3、5 の場合を図 12に示す。また $D_t = D_q$ の制限を付けないより一般的なシミュレーション結果と (15) 式によるフォルスドロップ確率の見積り値を、$D_t = 10$ の場合を図 13に、$D_t = 100$ の場合を図 14に示す。図 12に示すように D_t の小さい時は (11) 式による見積りにより正確にフォルスドロップ確率を見積もることができる。D_t の値が大きくなるに従いシグネチャのウェイトの分散が大きくなってくるため、(11) 式による見積りは誤差が生じてくる。

D_t と D_q の値が小さくターゲットシグネチャ、問い合わせシグネチャのウェイトの分散がほとんどない場合は、(15) 式による見積りではフォルスドロップ確率を高めに見積もってしまう。図 14に示すように、$D_t = 100$ の場合は (15) 式によりフォルスドロップ確率を正確に見積もることができる。

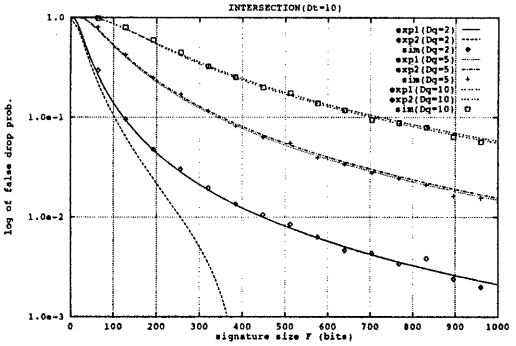

図 10: $(Q \cap T) \neq \phi (D_t = 10)$

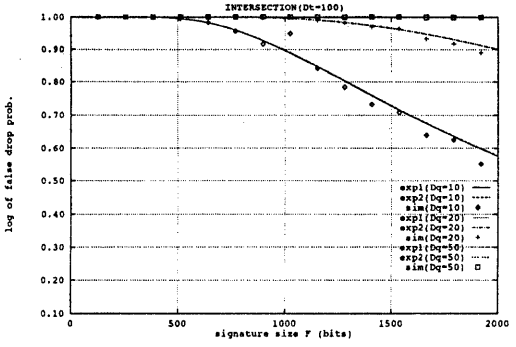

図 11: $(Q \cap T) \neq \phi (D_t = 100)$

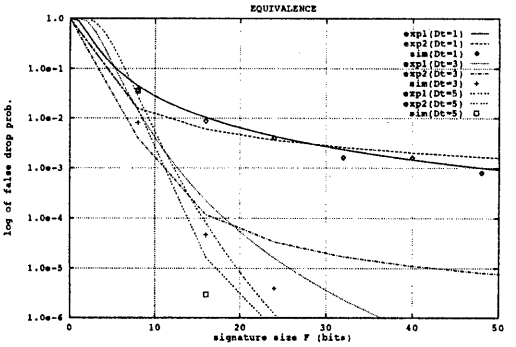

図 12: $Q \equiv T (D_t = D_q)$

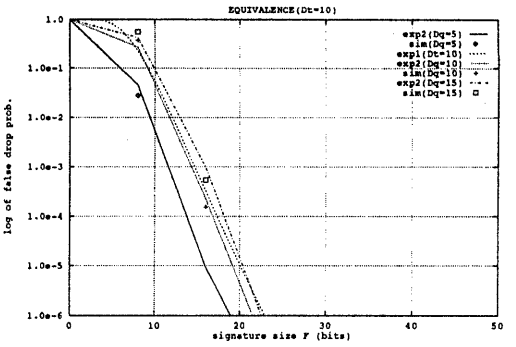

図 13: $Q \equiv T (D_t = 10)$

(5) $(Q \cap T) \neq \phi$ の処理方法の改善

図10、図11に示す通り、$(Q \cap T) \neq \phi$ の問い合わせではかなりフォルスドロップ確率が高くなってしまう。これはターゲットシグネチャと問い合わせシグネチャのいずれでも "1" が立っているビット位置がかなり多くなることによる。この問題を解決するためには、要素シグネチャをスーパーインポーズせずに、各要素シグネチャを問い合わせシグネチャとして各要素 q に対しての $q \in T$ の membership のマッチングを行なうことにより解決できる。この方式による Q4 の問い合わせのシグネチャファイルによる処理を図15に示す。

このような方式で $(Q \cap T) \neq \phi$ の問い合わせの処理を行なう場合、フォルスドロップ確率は、

$$\begin{aligned} Fd_{\{(Q \cap T) \neq \phi\}} &= \sum_{i=1}^{D_q} Fd_{\{q \in T\}} * (1 - Fd_{\{q \in T\}})^{i-1} \\ &= 1 - (1 - Fd_{\{q \in T\}})^{D_q} \end{aligned} \quad (17)$$

により与えられる。この場合、

$$\frac{\Delta Fd_{\{(Q \cap T) \neq \phi\}}}{\Delta Fd_{\{q \in T\}}} = D_q (1 - Fd_{\{q \in T\}})^{D_q - 1}$$

であるので、$0 < Fd_{\{q \in T\}} \leq 1$ より

$$\frac{\Delta Fd_{\{(Q \cap T) \neq \phi\}}}{\Delta Fd_{\{q \in T\}}} > 0$$

が成立する。よって $Fd_{\{q \in T\}}$ が極小の時、つまり $m = m_{opt}$ の時に $Fd_{\{(Q \cap T) \neq \phi\}}$ が極小となる。

この改良した方式で $(Q \cap T) \neq \phi$ の問い合わせを処理した場合についてシミュレーション結果と (17) 式による見積り値を、$D_t = 10$ の場合を図16に、$D_t = 100$ の場合を図17にそれぞれ示す。これより (17) 式による見積り値はシミュレーション結果に正確に一致することが分かる。以下、$(Q \cap T) \neq \phi$ の問い合わせはこの改良した方法により処理を行なうものとする。

4.2 ターゲット集合の要素数が一定でない場合

4.2.1 要素の出現頻度が一定である場合

本節では 3.3節に対応して、ターゲット集合の要素数が一定ではなく 2 項分布となる場合に対するシミュレーション結果を示し、要素数が一定な場合と比較する。図中では

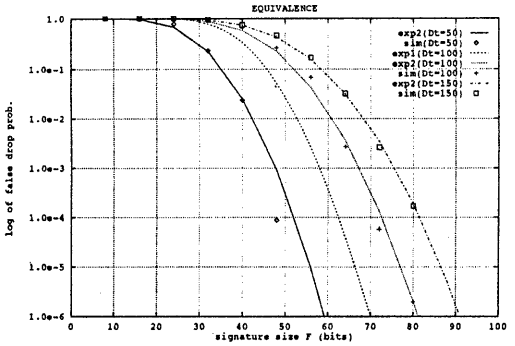

図 14: $Q \equiv T(D_t = 100)$

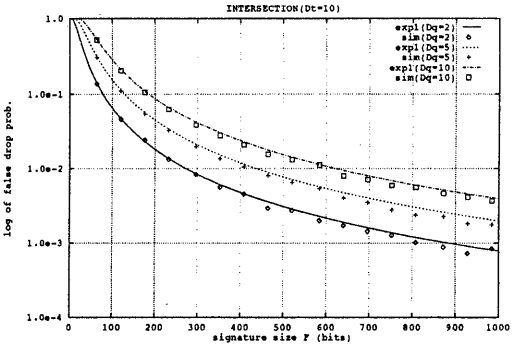

図 16: $(Q \cap T) \neq \phi(D_t = 10)$

query element1 → query signature 1
eraser → 1010010000000000
query element2 → query signature 2
stapler → 0010000010000100

signature file

	1101101010000101	T1
drop ←	1110110111000010	T2
	⋮	⋮

図 15: Q4 の問い合わせの処理

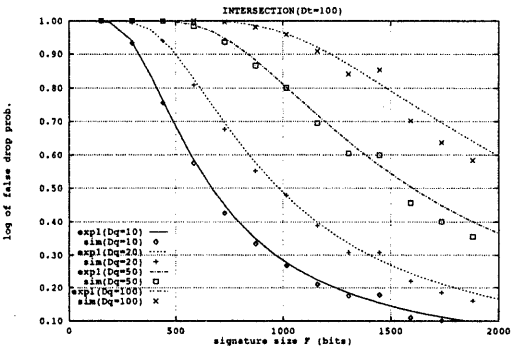

図 17: $(Q \cap T) \neq \phi(D_t = 100)$

3.3節で示したフォルスドロップ確率の見積り式による解析値を bd により示す。

(1) $Q \subseteq T$

シミュレーション結果と (16) 式において (8) 式を適用することにより得られる見積り値を、$\bar{D}_t = 10$ の場合を図 18に、$\bar{D}_t = 100$ の場合を図 19にそれぞれ示す。また、比較のため $D_t = \bar{D}_t$ とした (8) 式による見積り値も示す。(16) 式による見積り値はシミュレーション結果とほぼ正確に一致する。$Q \subseteq T$ の問い合わせでは D_t が大きくなるに従いフォルスドロップ確率が急激に大きくなるため、$D_t > \bar{D}_t$ となるターゲット集合のフォルスドロップ確率が影響し、(8) 式による見積りよりもシミュレーションにおけるフォルスドロップ確率は若干高くなる。しかし、D_t が一定な場合との大きなずれは見られない。

$\bar{D}_t = 10$ の場合は、$\bar{D}_t = 100$ の場合よりもターゲット集合の要素数の分散による影響が若干大きく現れている。これは、$\bar{D}_t = 10$ の場合は要素数が 1 から 23($\bar{D}_t$ の 10% から 230%) まで分散しているのに対して、$\bar{D}_t = 100$ の場合は 66 から 136($\bar{D}_t$ の 66% から 136%) まで分散しているためである。つまり要素数の分散が要素数の平均値と比較して大きい程フォルスドロップ確率が高くなることが分かる。

(2) $Q \supseteq T$

シミュレーション結果と (16) 式において (13) 式を適用することにより得られる見積り値を、$\bar{D}_t = 10$ の場合を図 20に、$\bar{D}_t = 100$ の場合を図 21にそれぞれ示す。また比較のため $D_t = \bar{D}_t$ とした (13) 式による見積り値も示す。(16) 式による見積り値はシミュレーション結果と一致することが分かる。$Q \supseteq T$ の問い合わせでは、D_t が小さくなるに従い同じ D_q の値に対するフォルスドロップ確率が急激に高くなるので、$D_t < \bar{D}_t$ となるターゲット集合値のフォルスドロップ確率が影響し、(13) 式の見積りよりもシミュレーションにおけるフォルスドロップ確率が高くなる。

$\bar{D}_t = 10$ の場合が $\bar{D}_t = 100$ の場合よりもターゲット集合の要素数の分布による影響が顕著であるが、これは $Q \subseteq T$ の場合と同様で、要素数の分散が要素数の平均値と比較して大きいほどフォルスドロップ確率が高くなる。

(3) $(Q \cap T) \neq \phi$

シミュレーション結果と (16) 式において (17) 式を適用することにより得られる見積り値を、$\bar{D}_t = 10$ の場合を

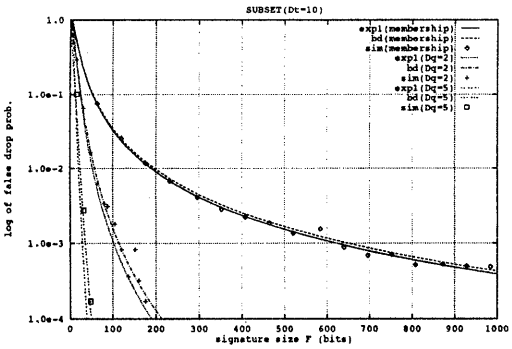

図 18: $Q \subseteq T$(含$q \in T$)($\bar{D}_t = 10$)

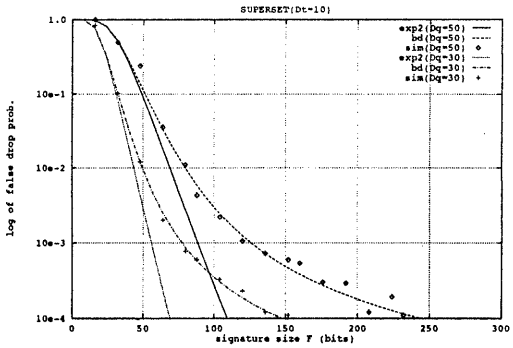

図 20: $Q \supseteq T(\bar{D}_t = 10)$

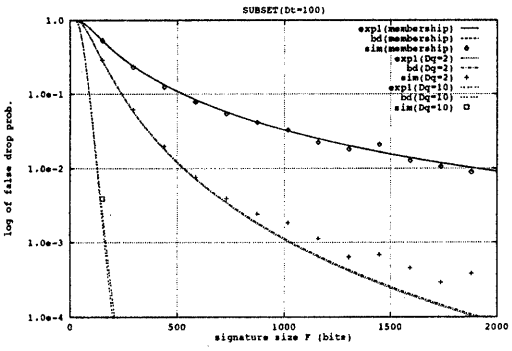

図 19: $Q \subseteq T$(含$q \in T$)($\bar{D}_t = 100$)

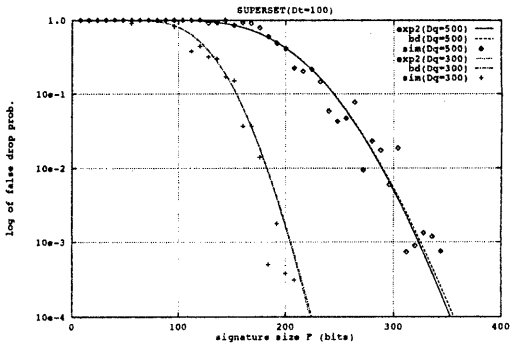

図 21: $Q \supseteq T(\bar{D}_t = 100)$

図22に、$\bar{D}_t = 100$ の場合を図23にそれぞれ示す。また比較のため $D_t = \bar{D}_t$ とした (17) 式による見積り値も示す。(16) 式による見積り値はシミュレーション結果と一致する。この場合も $Q \subseteq T$ と同じ傾向を示し、(17) 式を用いた場合でも見積りに大きな誤差はない。

(4) $Q \equiv T$

シミュレーション結果と (16) 式において (15) 式を適用することにより得られる見積り値を、$D_t = 10$ の場合を図24に、$D_t = 100$ の場合を図25にそれぞれ示す。また比較のため $D_t = \bar{D}_t$ とした (15) 式による見積り値も示す。(16) 式による見積り値はシミュレーション結果と一致する。また (15) 式を用いた場合でも大きな差異は存在しない。

4.2.2 要素の出現頻度が一定でない場合

ここでは V 中の要素の出現頻度が一定ではなく、一部のものが他より大きな出現頻度を持つものとする。ここでは以下の 3 つの場合を考慮する。

1. V 中の 100 個の要素が他の要素の出現頻度の 100 倍の出現頻度を持つ。
2. V 中の 1000 個の要素が他の要素の出現頻度の 100 倍の出現頻度を
3. V 中の 3000 個の要素が他の要素の出現頻度の 100 倍の出現頻度を持つ。

ただし、出現頻度の高い要素の数を h により示し p_1、p_2 によりそれぞれ出現頻度の高い要素の出現確率と低い要素の出現確率を示す。この時

$$\bar{D}_t = p_1 h + p_2(V - h)\ ,\ p_1 = 100 p_2$$

である。この式を用いて $\bar{D}_t = 10$ とすることによりデータを生成した結果、1 のケースでは $\bar{D}_t = 10$、2 のケースでは $\bar{D}_t = 11$、3 のケースでは $\bar{D}_t = 12$ となった。シミュレーション結果と $D_t = \bar{D}_t$ とした 3.2節で導出した見積り式による見積り値、D_t が 2 項分布をなすとして (16) 式に 3.2節の見積り式を適用した場合の見積り値を、1 の場合を図26に、2 の場合を図27に、3 の場合を図28にそれぞれ示す。

この結果、要素の出現頻度が一様でない場合においても、ターゲット集合の要素数が 2 項分布である場合と大差なく、$Q \supseteq T$ の問い合わせを除く三種類の問い合わせに対しては

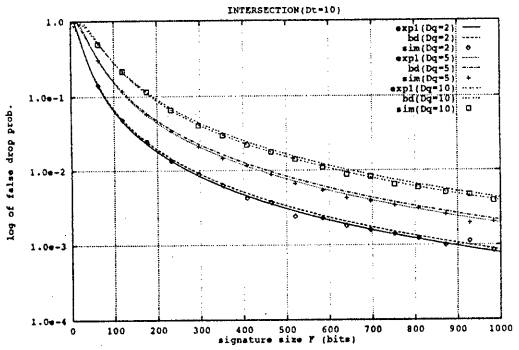

図 22: $(Q \cap T) \neq \phi(\bar{D}_t = 10)$

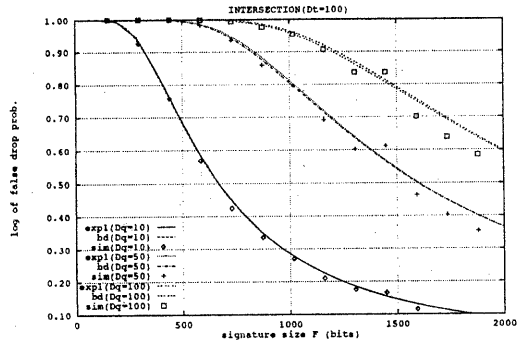

図 23: $(Q \cap T) \neq \phi(\bar{D}_t = 100)$

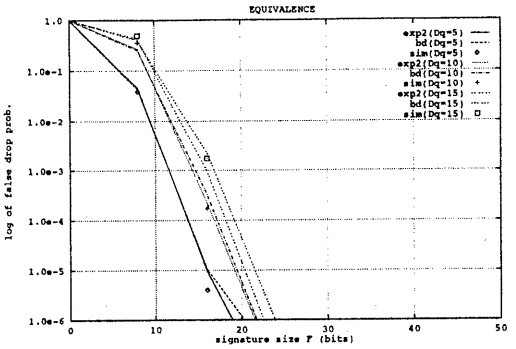

図 24: $Q \equiv T(\bar{D}_t = 10)$

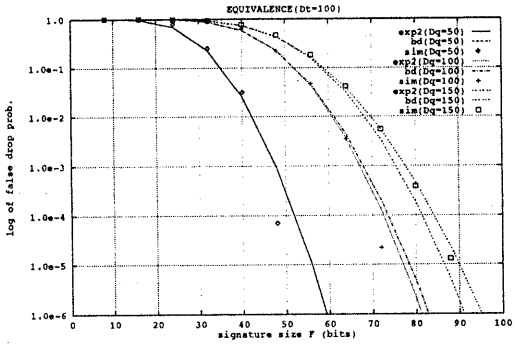

図 25: $Q \equiv T(\bar{D}_t = 100)$

$D_t = \bar{D}_t$ とした3.2節で導出した見積り式により、ほぼ正確にフォルスドロップ確率を見積もれることが分かる。

4.2.3 考察

$m = 2$ に m を固定した場合、ターゲット集合の要素数が一定でない4.2.1、4.2.2のいずれの場合でも、シミュレーション結果とターゲット集合の要素数が一定値 $D_t = \bar{D}_t$ とした見積り値に大きなずれがないことが以上の検討により確認できた。$Q \supseteq T$ の問い合わせの場合で $\bar{D}_t$ が小さい場合はずれが若干大きくなるが、$Q \supseteq T$ のフォルスドロップ確率が他の場合と比較して小さいため、この差はシグネチャファイルによる検索全体のコスト見積りにおいては大きな問題とはならないと考えられる。

5 結論

本論文では、シグネチャファイルを用いた4種類の集合値検索 $(Q \subseteq T,\ Q \supseteq T,\ (Q \cap T) \neq \phi,\ Q \equiv T)$ に対するフォルスドロップ確率の確率論的な見積り式を導出し、シミュレーションによりその妥当性を評価した。

その結果、これらの見積り式を用いることにより、ターゲット集合の要素数が一定な場合だけでなく、要素数の分散が2項分布となるような状況においても、シグネチャファイルを用いた集合値検索におけるフォルスドロップ確率をほぼ正確に予測することができることが分かった。更に要素の出現頻度が一定でない場合のような確率論的な見積り式を導出できない場合でも、これらの見積り式により十分フォルスドロップ確率の見積りが可能であることを示した。また、今回結果は示さなかったが、シーケンシャルシグネチャファイルを想定し $m = m_{opt}$ とした場合においてもシミュレーションを行い、本論文で導出した見積り式の妥当性を確認した。これらの結果に基づき、実際に集合値検索の支援を行なう際の適切なシグネチャファイルの設計や検索コストの見積りを行なうことができると考えられる。

謝辞

本研究を進めるに当り、貴重な御助言をして下さった藤原譲教授に深く感謝いたします。

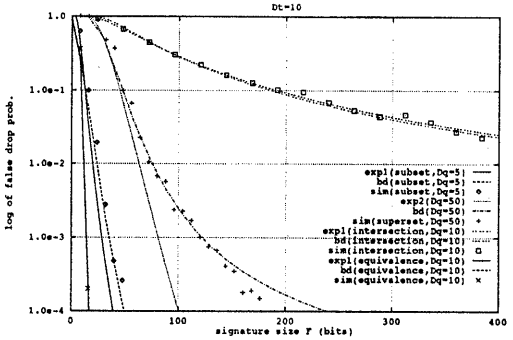

図 26: $h = 100, p_1 = 100p_2$

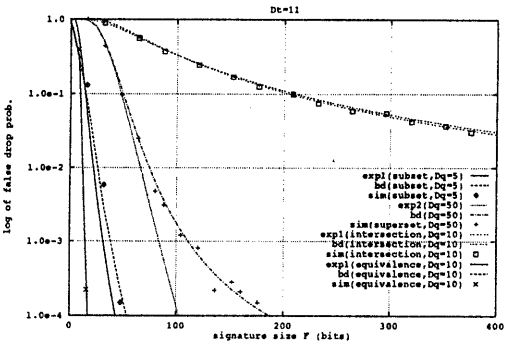

図 27: $h = 1000, p_1 = 100p_2$

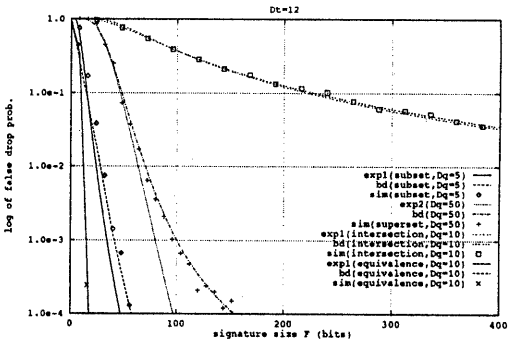

図 28: $h = 3000, p_1 = 100p_2$

参考文献

[Bert 89] E. Bertino and W. Kim, *"Index Techniques for Queries on Nested Objects,"* IEEE Trans. Knowledge and Data Engineering, Vol. 1, No. 2, June 1989, pp. 196-214.

[Chan 89] W. W. Chang and H. J. Scheck, *"A Signature Access Method for the Starburst Database System,"* Proc. 15th VLDB, 1989, pp. 145-153.

[Falo 84] C. Faloutsos and S. Christdoulakis, *"An Access Method for Documents and Its Analytical Performance Evaluation,"* ACM Trans. Office Information Systems, Vol. 2, No. 4, Oct 1984, pp. 267-288.

[Fuku 92] 福島, 石川, 干, 北川, 大保, "複合オブジェクトに対する索引機構の研究," 情報処理学会第４４回全国大会予稿集 (4), 1992, pp. 75-76.

[Ishi 93] Y. Ishikawa, H. Kitagawa and N. Ohbo, *"Evaluation of Signature Files as Set Access Facilities in OODBs,"* To appear in Proc. ACM SIGMOD 1993.

[Iwai92] 岩井原, 牧之内, "複合オブジェクトの集合演算のための索引構造," 情報処理学会第４５回全国大会予稿集 (4), 1992, pp. 131-132.

[Kent 88] A. Kent and R. Sachs-Davis and K. Ramamohanarao, *"A Superimposed Coding Schema Based on Multiple Block Descriptor Files for Indexing Very Large Data Bases,"* Proc. 14th VLDB, Aug 1988, pp. 351-359.

[Kita 89] H. Kitagawa and T. L. Kunii, *"The Unnormalized Relational Data Model - For Office Form Processor Design,"* Springer Verlag, 1989.

[Sack 87] R. Sacks-Davis and A. Kent, *"Multikey Access Methods Based on Superimposed Coding Techniques,"* ACM Trans. Database Systems, Vol12, No. 4, Dec 1987, pp. 655-698.

[Stia 60] S. Stiassny, *"Mathematical Analysis of Various Superimposed Coding Methods,"* American Documentation, Vol. 11, Feb 1960, pp. 155-169.

[Wong 91] K. F. Wong and M. H. Williams, *"A Superimposed Codeword Indexing Schema for Handling Sets in Prolog Databases,"* Proc. 2nd International Symposium on Database Systems for Advanced Applications, April 1991, pp. 468-476.