PAPER

False Drop Analysis of Set Retrieval with Signature Files

Hiroyuki KITAGAWA† *and* **Yoshiharu ISHIKAWA**††, *Members*

SUMMARY Modern database systems have to support complex data objects, which appear in advanced data models such as object-oriented data models and nested relational data models. Set-valued objects are basic constructs to build complex structures in those models. Therefore, efficient processing of set-valued object retrieval (simply, set retrieval) is an important feature required of advanced database systems. Our previous work proposed a basic scheme to apply superimposed coded *signature files* to set retrieval and showed its potential advantages over the B-tree index based approach using a performance analysis model. Retrieval with signature files is always accompanied by mismatches called *false drops*, and proper control of the false drops is indispensable in the signature file design. This study intensively analyzes the false drops in set retrieval with signature files. First, schemes to use signature files are presented to process set retrieval involving "has-subset," "is-subset," "has-intersection," and "is-equal" predicates, and generic formulas estimating the false drops are derived. Then, three sets of concrete formulas are derived in three ways to estimate the false drops in the four types of set retrieval. Finally, their estimates are validated with computer simulations, and advantages and disadvantages of each set of the false drop estimation formulas are discussed. The analysis shows that proper choice of estimation formulas gives quite accurate estimates of the false drops in set retrieval with signature files.

key words: *access method, complex object, advanced data models, signature file, false drop, set retrieval*

1. Introduction

Requirements for the database systems to handle complex data objects have been increasing according to the recent expansion of the computer application domain. To meet the requirements, advanced data models such as nested relational data models [1], [9], [12], [16], [19] and object-oriented data models [3], [11], [23] are investigated. Those models manipulate complex data objects and, in particular, directly handle set-valued objects. Therefore, efficient set-valued object retrieval (simply, set retrieval) facilities are indispensable for database systems supporting those models. Although novel indexing schemes such as the nested index and the multi-index [2], [20] incorporating the nested structures in the complex data objects have been investigated, they are not designed to support set retrieval in general. We have proposed the use of superimposed coded *signature files* as efficient set retrieval facilities and showed their potential capabilities concentrating on the retrieval with the set inclusion operator ($\subseteq$) [10].

Signature files were originally designed for text retrieval [4]–[6], [8]. Although applications of signature files to non-text object retrieval, such as to the traditional record retrieval [7], [17], [18], to the Prolog clause retrieval [21], and to the object navigation in the object-oriented database [14], [22], were discussed, study focusing on their use in set retrieval has not been reported by other researchers. Retrieval with signature files is always accompanied by mismatches called *false drops*. The number of false drops has a direct effect on the performance [10]. Therefore, it is important to estimate the false drops and to properly control them in the signature file design.

In this paper, we first present schemes to use signature files to process four types of set retrieval based on the "has-subset," "is-subset," "has-intersection," and "is-equal" conditions, and derive generic formulas estimating the false drops. Then, we derive three sets of concrete formulas to compute the false drop probabilities. We evaluate the validity of each set of the formulas with computer simulations, and discuss their advantages and disadvantages with respect to their reliability and computation cost. Among the three sets of formulas, two are refinements of our previous work presented in [13]. The remaining one is based on the theoretical research by Murphree and Aktug on the signature generation by superimposed coding [15].

The paper is organized as follows: In Sect. 2, we give an overview of the set retrieval in our context and the set query processing with signature files. In Sect. 3, we derive a set of generic formulas and three sets of concrete formulas estimating the false drops in the set query processing. In Sect. 4, we evaluate validity of the formulas with computer simulations and discuss advantages and disadvantages of each set of the formulas. Sect. 5 is the summary and conclusion.

Manuscript received April 2, 1996.
Manuscript revised November 7, 1996.

†The author is with Institute of Information Sciences and Electronics, University of Tsukuba, Tsukuba-shi, 305 Japan. e-mail: kitagawa@is.tsukuba.ac.jp

††The author is with Graduate Institute of Information Science, Nara Institute of Science and Technology (NAIST), Ikoma-shi, 630-01 Japan. e-mail: ishikawa@is.aist-nara.ac.jp

MUSICIAN			
TID	MNO	MNAME	INSTRUMENTS
			INAME
T1	10	E. Jones	piano guitar banjo
T2	25	F. Kent	trumpet tuba saxophone flute

Fig. 1 Example nested relation.

2. Set Retrieval with Signature Files

2.1 Set Retrieval

One of the typical data models supporting set values and their retrieval is the nested relational model [1], [9], [12], [16], [19]. Figure 1 shows a sample nested relation representing musicians in an association. Attributes MNO and MNAME represent the musician number and name, respectively. Each tuple is tagged with the tuple identifier (TID). INSTRUMENTS is a *set attribute* and represents the set of instruments each musician can play. Tuple retrieval based on the INSTRUMENTS attribute value, for instance INSTRUMENTS $\supseteq$ {`piano`, `guitar`}, is an example of set retrieval investigated in this paper. In this case, each INSTRUMENTS attribute value is called a *target set*, and {`guitar`, `piano`} is called a *query set*.

Selection operation in many nested relational algebras considers selection conditions including (some of) the following set comparison operators [1], [9], [12], [16], [19]. Here, T and Q denote the target set and query set, respectively, and q denotes a simple value given in case of $T \ni q$.

1. $T \ni q$ (*has-element*): The target set has the simple value q as an element.
   Q1: Retrieve musicians who can play the piano.

2. $T \supseteq Q$ (*has-subset*): The target set has the query set as a subset.
   Q2: Retrieve musicians who can play both the piano and the guitar.

3. $T \subseteq Q$ (*is-subset*): The target set is a subset of the query set.
   Q3: Retrieve musicians who can only play some of the piano, guitar, bass, and violin.

4. $T \sqcap Q$[†] (*has-intersection*): The target set has intersection with the query set.
   Q4: Retrieve musicians who can play the tuba or the clarinet.

element		element signature
`piano`	→	0001000000000101
`guitar`	→	1100100000000000
`banjo`	→	0100001010000000
		⇓
target signature		1101101010000101

Fig. 2 Generation of a target signature.

1101101010000101	T1
1110110111000010	T2
⋮	⋮

Fig. 3 Signature file.

5. $T \equiv Q$ (*is-equal*): The target set is equal to the query set.
   Q5: Retrieve musicians who can play the piano, guitar, banjo, and nothing else.

As $T \ni q$ is a special case of $T \supseteq Q$, we consider four types of set retrieval conditions $T \supseteq Q$, $T \subseteq Q$, $T \sqcap Q$, $T \equiv Q$ in the remaining part of this paper.

2.2 Use of Signature Files

Signature files were originally designed for text retrieval [4]–[6], [8]. A *signature* is a bit pattern formed for each data object and stored in the signature file. A typical query processing with the signature file is as follows: When a query is given, a *query signature* is formed from the query value. Then, each signature in the signature file is examined over the query signature for potential match. If the signature satisfies a pre-defined condition implied by the query condition, the corresponding data object becomes a candidate that may satisfy the query. Such a data object is called a *drop*. The last step is the *false drop resolution*, and each drop is accessed and examined whether it actually satisfies the query condition. Drops that fail the test are called *false drops*, while the qualified data objects are called *actual drops*.

In set retrieval with signature files, a *target signature* is generated for each target set as shown in Fig. 2. First, each element in a target set is hashed to a binary bit pattern called an *element signature*. All element signatures have F bit length, and m bits are set to "1." Then, a target signature is obtained by bit-wise OR-ing (*superimposed coding*) element signatures of all the elements in the target set. Pairs of such a target signature and a TID of the tuple including the target set are stored in the signature file as shown in Fig. 3.

Queries $T \supseteq Q$, $T \subseteq Q$, $T \sqcap Q$, and $T \equiv Q$ are processed with the signature files in the following way:

[†]"$T \sqcap Q$" stands for "$T \cap Q \neq \emptyset$."

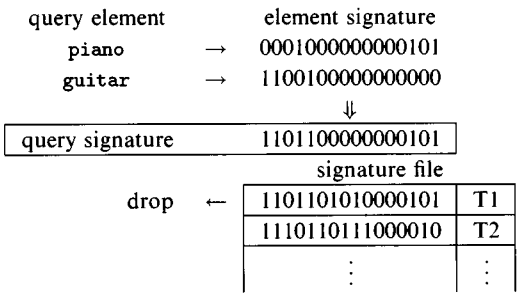

Fig. 4 Query processing of **Q2**.

1. A query signature is generated from the query set Q in the same way as the target signature.

2. The signature file is examined. Each target set becomes a drop if the following condition is satisfied†.

   $\boldsymbol{T \supseteq Q}$: *query signature* $\wedge$ *target signature* = *query signature*.

   $\boldsymbol{T \subseteq Q}$: *query signature* $\wedge$ *target signature* = *target signature*.

   $\boldsymbol{T \sqcap Q}$: *weight*(*query signature* $\wedge$ *target signature*) $\geq m$, where the function *weight*() returns the *weight*, namely the number of bits set to "1."

   $\boldsymbol{T \equiv Q}$: *query signature* = *target signature*.

3. The drops in step 2) are retrieved and checked whether they actually satisfy the query condition (false drop resolution).

Figure 4 illustrates steps 1) and 2) in query processing of the query **Q2**. The tuple T1 becomes a drop because its satisfies the above condition for $T \supseteq Q$.

The query $T \sqcap Q$ could be processed following the above procedure. However, it has been clarified in our previous work [13] that this processing scheme for $T \sqcap Q$ is sometimes undesirable because the number of false drops is rather large. An alternative processing scheme for $T \sqcap Q$ to resolve this problem is shown below:

1. An element signature is generated for each element in the query set Q.

2. The signature file is examined. Each target set becomes a drop if any element signature generated in step 1) satisfies the following condition.

   element signature $\wedge$ *target signature* = *element signature*.

3. The drops in step 2 are retrieved and checked whether they actually satisfy the query condition (false drop resolution).

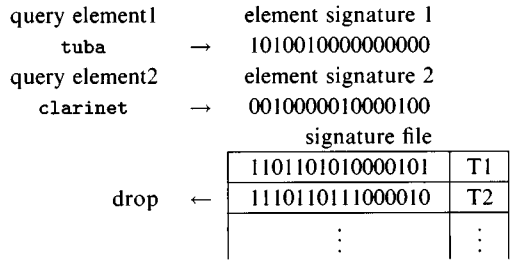

Fig. 5 Query processing of **Q4**.

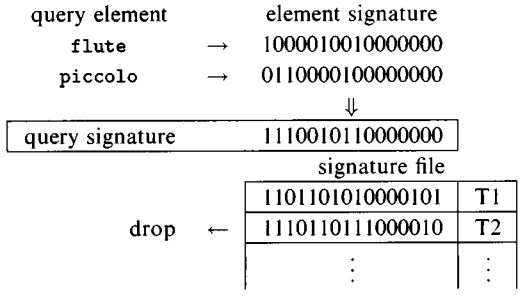

Fig. 6 False drop.

Figure 5 illustrates query processing of the query **Q4** under this scheme. In the following discussion, the query $T \sqcap Q$ processed under the first scheme is denoted by $T \sqcap_1 Q$, and that processed under the second scheme is denoted by $T \sqcap_2 Q$.

3. False Drop Analysis

Figure 6 shows a case where false drops occur. The query with the condition $T \supseteq$ {`flute, piccolo`} is processed in this example. Although the tuple T2 does not satisfy this condition, it becomes a drop. Thus, it is a false drop. The false drop is due to the collision of element signatures and the superimposed coding method. The frequency of false drop is usually measured in the *false drop probability* Fd as follows:

$$Fd = \frac{\text{false drops}}{N - \text{actual drops}},$$

where N is the total number of target sets [4]. In this section, we derive formulas estimating false drop probabilities for the four types of queries mentioned in Sect. 2. As discussed in [4], we can derive those formulas assuming the case of unsuccessful search, where we have no actual drops.

3.1 Basic Considerations

Table 1 shows symbols used in our analysis. We make

† '$\wedge$' stands for bit-wise AND operation.

Table 1 Symbols.

symbol	definition
F	Signature size in bits
m	Element signature weight
D_t	Cardinality of a target set T
D_q	Cardinality of a query set Q
N	Total number of target sets in the database
V	Cardinality of the set element domain

following assumptions:

1. The weight of an element signature is very small compared with the signature size ($m \ll F$).
2. The "1"'s are uniformly distributed in an element signature. Therefore, each bit position is set to "1" with the same probability.
3. There is no skew in element occurrences in the target sets. Namely, each possible set value of cardinality D_t has an equal likelihood of appearing as a target set.

From the assumption 2, each bit position in an element signature is set with the probability m/F. Therefore, the probability that a bit position b_t of a target signature is set to "1" is given by

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

Similarly, the probability that a bit position b_q of a query signature is set to "1" is given by

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

The following formula giving the false drop probability for $T \ni q$ was derived in [4]:

$$Fd_{\{T \ni q\}} = (p(b_t))^m \approx (1 - e^{-\frac{mD_t}{F}})^m. \tag{1}$$

We derive false drop probability formulas for $T \supseteq Q$, $T \subseteq Q$, $T \sqcap Q$, and $T \equiv Q$ taking these considerations as a starting basis.

3.2 Generic Formulas

First, we consider the case where all target sets have the same cardinality D_t. Let b_t^j $(1 \leq j \leq F)$ be the j-th bit position of the target signature and b_q^j $(1 \leq j \leq F)$ be j-th bit position of the query signature. For each i $(1 \leq i \leq F - m)$, the following equations hold:

$$\begin{aligned}
&\text{Prob}\left\{b_t^1 = 0 \wedge \cdots \wedge b_t^i = 0\right\} \\
&= \left(\frac{\binom{F-i}{m}}{\binom{F}{m}}\right)^{D_t} \\
&= \left(\frac{(F-m)(F-m-1)\cdots(F-m-i+1)}{F(F-1)\cdots(F-i+1)}\right)^{D_t} \\
&= \prod_{k=1}^{i}\left(1 - \frac{m}{F-k+1}\right)^{D_t}. \qquad (2)
\end{aligned}$$

If $\frac{m}{F-k+1} \ll 1$ is satisfied for $1 \leq k \leq i$,

$$\begin{aligned}
&\text{Prob}\left\{b_t^1 = 0 \wedge \cdots \wedge b_t^i = 0\right\} \\
&= \left(1 - \frac{m}{F}\right)^{D_t} \times \left(1 - \frac{m}{F-1}\right)^{D_t} \times \\
&\quad \cdots \times \left(1 - \frac{m}{F-i+1}\right)^{D_t} \\
&\approx \left(1 - \frac{1}{F}\right)^{mD_t} \times \left(1 - \frac{1}{F-1}\right)^{mD_t} \times \\
&\quad \cdots \times \left(1 - \frac{1}{F-i+1}\right)^{mD_t} \\
&= \left(1 - \frac{i}{F}\right)^{mD_t}. \qquad (3)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\text{Prob}\left\{b_q^1 = 0 \wedge \cdots \wedge b_q^i = 0\right\} \\
&= \prod_{k=1}^{i}\left(1 - \frac{m}{F-k+1}\right)^{D_q}. \qquad (4)
\end{aligned}$$

If $\frac{m}{F-k+1} \ll 1$ is satisfied for $1 \leq k \leq i$,

$$\text{Prob}\left\{b_q^1 = 0 \wedge \cdots \wedge b_q^i = 0\right\} \approx \left(1 - \frac{i}{F}\right)^{mD_q}. \tag{5}$$

In the following, we derive generic false drop probability formulas for $T \supseteq Q$, $T \subseteq Q$, $T \sqcap Q$, and $T \equiv Q$. The probability that the target signature weight is i is denoted by $p_t(i)$, and the probability that the query signature weight is i is denoted by $p_q(i)$.

(1) $T \supseteq Q$

A false drop occurs when the following condition holds for every bit position j $(1 \leq j \leq F)$:

$$b_t^j = 0 \quad \Rightarrow \quad b_q^j = 0.$$

If the target signature weight is i, the number of "0"'s in the target signature is $F - i$. Therefore, the probability $f_{\{T \supseteq Q\}}(i)$ that the target set becomes a false drop is derived from Eq. (4) as follows:

$$\begin{aligned}
f_{\{T \supseteq Q\}}(i) &= \text{Prob}\left\{b_q^1 = 0 \wedge \cdots \wedge b_q^{F-i} = 0\right\} \\
&= \prod_{k=1}^{F-i}\left(1 - \frac{m}{F-k+1}\right)^{D_q}. \qquad (6)
\end{aligned}$$

As distribution of the target signature weight is determined by $p_t(i)$, the false drop probability for $T \supseteq Q$ is given by

$$Fd_{\{T \supseteq Q\}} = \sum_{i=0}^{F} p_t(i) \prod_{k=1}^{F-i}\left(1 - \frac{m}{F-k+1}\right)^{D_q}. \tag{7}$$

If $\frac{m}{F-k+1} \ll 1$ holds for $1 \leqq k \leqq F-i$, we get from Eq. (5)

$$f_{\{T \supseteq Q\}}(i) \approx \left(1 - \frac{F-i}{F}\right)^{mD_q} = \left(\frac{i}{F}\right)^{mD_q} \tag{6'}$$

$$Fd_{\{T \supseteq Q\}} \approx \sum_{i=0}^{F} p_t(i) \left(\frac{i}{F}\right)^{mD_q}. \tag{7'}$$

The assumption $\frac{m}{F-k+1} \ll 1$ does not hold when $k \approx F-i$ and $i \approx m$ or $i < m$. Actually, however, $p_t(i) = 0$ for $i < m$, and we suppose that, for most probable D_t-values such as $D_t = 10$ and $D_t = 100$, $p_t(i)$ becomes very small for $i \approx m$. Therefore, if we can give a good estimate for $p_t(i)$, the formula (7') will also give good values. This point is validated in Sect. 4. In case $D_q = 1$, formulas (7) and (7') give the false drop probability for $T \ni q$.

(2) $T \subseteq Q$

A false drop occurs when the following condition holds for every bit position j $(1 \leqq j \leqq F)$:

$$b_q^j = 0 \quad \Rightarrow \quad b_t^j = 0.$$

If the query signature weight is i, the probability $f_{\{T \subseteq Q\}}(i)$ that the target set becomes a false drop is derived from Eq. (2) as follows:

$$\begin{aligned} f_{\{T \subseteq Q\}}(i) &= \text{Prob}\left\{b_t^1 = 0 \wedge \cdots \wedge b_t^{F-i} = 0\right\} \\ &= \prod_{k=1}^{F-i} \left(1 - \frac{m}{F-k+1}\right)^{D_t}. \end{aligned} \tag{8}$$

As distribution of the query signature weight is determined by $p_q(i)$, the false drop probability for $T \subseteq Q$ is given by

$$Fd_{\{T \subseteq Q\}} = \sum_{i=0}^{F} p_q(i) \prod_{k=1}^{F-i} \left(1 - \frac{m}{F-k+1}\right)^{D_t}. \tag{9}$$

If $\frac{m}{F-k+1} \ll 1$ holds for $1 \leqq k \leqq F-i$, we get from Eq. (3)

$$f_{\{T \subseteq Q\}}(i) \approx \left(1 - \frac{F-i}{F}\right)^{mD_t} = \left(\frac{i}{F}\right)^{mD_t} \tag{8'}$$

$$Fd_{\{T \subseteq Q\}} \approx \sum_{i=0}^{F} p_q(i) \left(\frac{i}{F}\right)^{mD_t}. \tag{9'}$$

A remark has been made with respect to the validity of the formula (7'). A similar remark applies to the formula (9').

(3) $T \sqcap Q$

1) $T \sqcap_1 Q$

A false drop occurs when the target signature weight and the query signature weight are at least m, and they have m or more bit intersection. Therefore, we get

$$\begin{aligned} &Fd_{\{T \sqcap_1 Q\}} \\ &= \sum_{i=m}^{F} p_t(i) \sum_{j=m}^{F} p_q(j) \sum_{k=\max(m, i+j-F)}^{\min(i,j)} \frac{\binom{i}{k}\binom{F-i}{j-k}}{\binom{F}{j}}. \end{aligned} \tag{10}$$

Here, $\sum_{k=\max(m, i+j-F)}^{\min(i,j)} \frac{\binom{i}{k}\binom{F-i}{j-k}}{\binom{F}{j}}$ is the probability that the target signature and the query signature have m or more bit intersection when their weights are $p_t(i)$ and $p_q(j)$, respectively.

2) $T \sqcap_2 Q$

The false drop probability for $T \sqcap_2 Q$ is simply expressed with $Fd_{\{T \ni q\}}$ as follows:

$$\begin{aligned} &Fd_{\{T \sqcap_2 Q\}} \\ &= \sum_{i=1}^{D_q} Fd_{\{T \ni q\}} \times (1 - Fd_{\{T \ni q\}})^{i-1} \\ &= 1 - (1 - Fd_{\{T \ni q\}})^{D_q}. \end{aligned} \tag{11}$$

(4) $T \equiv Q$

A false drop occurs when both the target signature weight and the query signature weight take the same value i, and the target signature is equal to the query signature. Therefore, the false drop probability for $T \equiv Q$ is given by

$$Fd_{\{T \equiv Q\}} = \sum_{i=0}^{F} p_t(i) p_q(i) \frac{1}{\binom{F}{i}}. \tag{12}$$

3.3 False Drop Probability Formulas

As shown in Sect. 3.2, probability distributions of the target and query signature weights denoted by $p_t(i)$ and $p_q(i)$, respectively, play an important role in estimating the false drops. In this subsection, we derive three sets of formulas by estimating $p_t(i)$ and $p_q(i)$ taking the following three different approaches.

3.3.1 Formulas F1

Here, we simply assume that the target and query signature weights are equal to their expected values $\overline{m}_t$ and $\overline{m}_q$, respectively, given as follows:

$$\overline{m}_t = F \times p(b_t) \approx F(1 - e^{-\frac{mD_t}{F}})$$

$$\overline{m}_q = F \times p(b_q) \approx F(1 - e^{-\frac{mD_q}{F}}).$$

Therefore,

$$p_t(i) = \begin{cases} 1 & \text{if } i = F(1 - e^{-\frac{mD_t}{F}}) \\ 0 & \text{otherwise} \end{cases}$$

and

$$p_q(i) = \begin{cases} 1 & \text{if } i = F(1-e^{-\frac{mD_q}{F}}) \\ 0 & \text{otherwise.} \end{cases}$$

The false drop probability formulas based on these $p_t(i)$ and $p_q(i)$ are as follows:

(1) $T \supseteq Q$

$$Fd_{\{T \supseteq Q\},\mathrm{F1}} = \prod_{k=1}^{Fe^{-\frac{mD_t}{F}}} \left(1 - \frac{m}{F-k+1}\right)^{D_q}. \quad (13)$$

$$Fd_{\{T \supseteq Q\},\mathrm{F1}} \approx (1-e^{-\frac{mD_t}{F}})^{mD_q}. \quad (13')$$

Note that Eq. (13') becomes Eq. (1) in case $D_q = 1$.

(2) $T \subseteq Q$

$$Fd_{\{T \subseteq Q\},\mathrm{F1}} = \prod_{k=1}^{Fe^{-\frac{mD_q}{F}}} \left(1 - \frac{m}{F-k+1}\right)^{D_t} \quad (14)$$

$$Fd_{\{T \subseteq Q\},\mathrm{F1}} \approx (1-e^{-\frac{mD_q}{F}})^{mD_t}. \quad (14')$$

(3) $T \sqcap Q$

1) $T \sqcap_1 Q$

$$Fd_{\{T \sqcap_1 Q\},\mathrm{F1}} = \sum_{k=\max(m,\, F(1-e^{-\frac{mD_t}{F}})+F(1-e^{-\frac{mD_q}{F}})-F)}^{\min(F(1-e^{-\frac{mD_t}{F}}),\, F(1-e^{-\frac{mD_q}{F}}))} \left\{ \frac{\binom{F(1-e^{-\frac{mD_t}{F}})}{k} \binom{F-F(1-e^{-\frac{mD_t}{F}})}{F(1-e^{-\frac{mD_q}{F}})-k}}{\binom{F}{F(1-e^{-\frac{mD_q}{F}})}} \right\}. \quad (15)$$

2) $T \sqcap_2 Q$

$$Fd_{\{T \sqcap_2 Q\},\mathrm{F1}} = 1 - \left(1 - \prod_{k=1}^{Fe^{-\frac{mD_t}{F}}} \left(1 - \frac{m}{F-k+1}\right)\right)^{D_q}. \quad (16)$$

$$Fd_{\{T \sqcap_2 Q\},\mathrm{F1}} \approx 1 - \left(1 - (1-e^{-\frac{mD_t}{F}})^m\right)^{D_q}. \quad (16')$$

(4) $T \equiv Q$

$$Fd_{\{T \equiv Q\},\mathrm{F1}} = \frac{1}{\binom{F}{\overline{m}}}, \quad (17)$$

where $\overline{m} = F(1-e^{-\frac{mD_t}{F}}) = F(1-e^{-\frac{mD_q}{F}})$. Note that Eq. (17) for $T \equiv Q$ is applicable only when $D_t = D_q$.

3.3.2 Formulas F2

In this approach, we assume that each bit position in the target signature and the query signature is set to "1" with probabilities $p(b_t)$ and $p(b_q)$ (given in Sect. 3.1), respectively, independently of other bit positions. Then, the distribution of the target and query signature weights follows the binomial distribution, and we get

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

$$p_q(i) = \binom{F}{i} p(b_q)^i (1-p(b_q))^{F-i}.$$

The false drop probability formulas based on these $p_t(i)$ and $p_q(i)$ are given below. $f_{\{T \supseteq Q\}}(i)$ is given by Eqs. (6) or (6'), and $f_{\{T \subseteq Q\}}(i)$ is given by Eqs. (8) or (8') as in F1.

(1) $T \supseteq Q$

$$Fd_{\{T \supseteq Q\},\mathrm{F2}} = \sum_{i=0}^{F} \binom{F}{i} (1-e^{-\frac{mD_t}{F}})^i e^{-\frac{mD_t}{F}(F-i)} f_{\{T \supseteq Q\}}(i). \quad (18)$$

(2) $T \subseteq Q$

$$Fd_{\{T \subseteq Q\},\mathrm{F2}} = \sum_{i=0}^{F} \binom{F}{i} (1-e^{-\frac{mD_q}{F}})^i e^{-\frac{mD_q}{F}(F-i)} f_{\{T \subseteq Q\}}(i). \quad (19)$$

(3) $T \sqcap Q$

1) $T \sqcap_1 Q$

$$Fd_{\{T \sqcap Q\},\mathrm{F2}} = \sum_{i=m}^{F} \binom{F}{i} p(b_t)^i (1-p(b_t))^{F-i} \left\{ \sum_{j=m}^{F} \binom{F}{j} p(b_q)^j (1-p(b_q))^{F-j} \left[\sum_{k=\max(m,\, i+j-F)}^{\min(i,j)} \frac{\binom{i}{k}\binom{F-i}{j-k}}{\binom{F}{j}} \right] \right\}. \quad (20)$$

2) $T \sqcap_2 Q$

$$Fd_{\{T \sqcap Q\},\mathrm{F2}} = 1 - \left[1 - \sum_{i=0}^{F} \left\{ \binom{F}{i} (1-e^{-\frac{mD_t}{F}})^i \right.\right.$$

$$e^{-\frac{mD_t}{F}(F-i)} f_{\{T \ni q\}}(i)\Big\}\Big]^{D_q}, \qquad (21)$$

where $f_{\{T \ni q\}}(i)$ is $f_{\{T \supseteq Q\}}(i)$ with $D_q = 1$.

(4) $T \equiv Q$

$$Fd_{\{T \equiv Q\},\,F2} = \sum_{i=0}^{F} \left\{ \binom{F}{i} \left(1 - e^{-\frac{mD_t}{F}}\right)^i \left(1 - e^{-\frac{mD_q}{F}}\right)^i e^{-\frac{m(D_t+D_q)}{F}(F-i)} \right\}. \qquad (22)$$

3.3.3 Formulas F3

In deriving the formulas F1 and F2, we have taken two different approaches to estimate the target and query signature weights. Murphree and Aktug derived a more strict mathematical formula giving the probability distribution of the set signature generated by superimposed coding [15]. Murphree and Aktug considered superimposed coding of D element signatures as a Markov process consisting of D stages. Let m_i $(1 \leqq i \leqq D)$ be the weight of the i-th element signature, Y_i be the weight of the set signature after the stage i, and W be the final signature weight. Then, $m_1 = Y_1 \leqq Y_2 \leqq \cdots \leqq Y_D = W$ holds. They derived the following probability distribution formula for the set signature weight by analyzing this Markov chain:

$$\text{Prob}\{W = w\} = \sum_{j=0}^{w-m_1} \left\{ \binom{F-m_1}{j} \binom{F-m_1-j}{w-m_1-j} (-1)^{w-m_1+j} \prod_{r=2}^{D} \frac{\binom{m_1+j}{m_r}}{\binom{F}{m_r}} \right\}, \qquad (23)$$

where $m_1 \leqq w \leqq \min(F, m_1 + \cdots + m_D)$. When we apply this formula to our context, we get the following formulas:

$$p_t(i) = \begin{cases} \displaystyle\sum_{j=0}^{i-m} \left[\binom{F-m}{j} \binom{F-m-j}{i-m-j} (-1)^{i-m+j} \left\{ \frac{\binom{m+j}{m}}{\binom{F}{m}} \right\}^{D_t-1} \right] & (\text{if } m \leqq i \leqq mD_t), \\ 0 \quad (\text{otherwise}), \end{cases}$$

$$p_q(i) = \begin{cases} \displaystyle\sum_{j=0}^{i-m} \left[\binom{F-m}{j} \binom{F-m-j}{i-m-j} (-1)^{i-m+j} \left\{ \frac{\binom{m+j}{m}}{\binom{F}{m}} \right\}^{D_q-1} \right] & (\text{if } m \leqq i \leqq mD_q), \\ 0 \quad (\text{otherwise}). \end{cases}$$

The false drop probability formulas based on these $p_t(i)$ and $p_q(i)$ are as follows:

(1) $T \supseteq Q$

$$Fd_{\{T \supseteq Q\},\,F3} = \sum_{i=m}^{\min(F, mD_t)} \sum_{j=0}^{i-m} \left[\binom{F-m}{j} \binom{F-m-j}{i-m-j} (-1)^{i-m+j} \left\{ \frac{\binom{m+j}{m}}{\binom{F}{m}} \right\}^{D_t-1} f_{\{T \supseteq Q\}}(i) \right]. \qquad (24)$$

(2) $T \subseteq Q$

$$Fd_{\{T \subseteq Q\},\,F3} = \sum_{i=m}^{\min(F, mD_q)} \sum_{j=0}^{i-m} \left[\binom{F-m}{j} \binom{F-m-j}{i-m-j} (-1)^{i-m+j} \left\{ \frac{\binom{m+j}{m}}{\binom{F}{m}} \right\}^{D_q-1} f_{\{T \subseteq Q\}}(i) \right]. \qquad (25)$$

(3) $T \sqcap Q$

1) $T \sqcap_1 Q$

$$Fd_{\{T \sqcap_1 Q\},\,F3} = \sum_{i=m}^{\min(F, mD_t)} p_t(i) \left\{ \sum_{j=m}^{\min(F, mD_q)} p_q(j) \sum_{k=\max(m, i+j-F)}^{\min(i, j)} \frac{\binom{i}{k}\binom{F-i}{j-k}}{\binom{F}{j}} \right\}. \qquad (26)$$

2) $T \sqcap_2 Q$

$$Fd_{\{T \sqcap_2 Q\},\,F3} = 1 - \left(1 - \sum_{i=m}^{\min(F, D_t)} \sum_{j=0}^{i-m} \left[\binom{F-m}{j} \binom{F-m-j}{i-m-j} (-1)^{i-m+j} \left\{ \frac{\binom{m+j}{m}}{\binom{F}{m}} \right\}^{D_t-1} f_{\{T \ni q\}}(i) \right] \right)^{D_q}. \qquad (27)$$

(4) $T \equiv Q$

$$
\begin{aligned}
&Fd_{\{T\equiv Q\},\mathrm{F3}} \\
&= \sum_{i=m}^{\min(F, mD_t, mD_q)} \left(\left[\sum_{j=0}^{i-m} \binom{F-m}{j} \right. \right. \\
&\quad \left. \binom{F-m-j}{i-m-j} (-1)^{i-m+j} \left\{ \frac{\binom{m+j}{m}}{\binom{F}{m}} \right\}^{D_t-1} \right] \\
&\quad \times \left[\sum_{j=0}^{i-m} \binom{F-m}{j} \binom{F-m-j}{i-m-j} \right. \\
&\quad \left. \left. (-1)^{i-m+j} \left\{ \frac{\binom{m+j}{m}}{\binom{F}{m}} \right\}^{D_q-1} \right] \frac{1}{\binom{F}{i}} \right).
\end{aligned} \tag{28}
$$

3.4 Varying Target Cardinality

Hitherto, we have focussed on the case where all target sets have the same cardinality D_t. It is not difficult to extend our study to the case where the cardinality of the target set varies. Let a *set element domain* be a set from which each target set element is taken, and its cardinality be V. If we do not consider the case that the target set is an empty set, the target set cardinality varies from 1 to V. Suppose the probability that the target set cardinality is D_t is given by the function $P(D_t)$ ($1 \leqq D_t \leqq V$). Then, we can derive the false drop probability with the following formula for each case of $T \supseteqq Q$, $T \subseteqq Q$, $T \sqcap Q$, and $T \equiv Q$:

$$
Fd_{\mathrm{VTC},c,f} = \sum_{D_t=1}^{V} P(D_t) Fd_{c,f}, \tag{29}
$$

where c and f are parameters indicating one of $\{T \supseteqq Q\}$, $\{T \subseteqq Q\}$, $\{T \sqcap_1 Q\}$, $\{T \sqcap_2 Q\}$, $\{T \equiv Q\}$, and one of F1, F2, F3, respectively. $Fd_{c,f}$ is the false drop probability formula derived in Sect. 3.3 for each combination. For example, the false drop probability for $T \supseteqq Q$ based on F3 (Eq. (24)) is given as follows:

$$
\begin{aligned}
&Fd_{\mathrm{VTC},\{T\supseteqq Q\},\mathrm{F3}} \\
&= \sum_{D_t=1}^{V} P(D_t) \sum_{i=m}^{\min(F, mD_t)} \\
&\quad \left[\sum_{j=0}^{i-m} \binom{F-m}{j} \binom{F-m-j}{i-m-j} \right. \\
&\quad \left. (-1)^{i-m+j} \left\{ \frac{\binom{m+j}{m}}{\binom{F}{m}} \right\}^{D_t-1} f_{\{T\supseteqq Q\}}(i) \right].
\end{aligned} \tag{30}
$$

4. Simulation Study

In this section, we evaluate validity of the formulas derived in Sect. 3 by simulations. A number of physical signature file organizations have been proposed [8]. In this study, we set parameter value $m = 2$ assuming the bit-sliced signature file organization, since our previous work [10] showed advantages of the bit-sliced signature files with small m-values in set retrieval. The set $\{0, \ldots, 9999\}$ is used as the set element domain, thus $V = 10000$. F, D_t, and D_q are variable parameters. First, for a given D_t, 10000 set values are generated, each of which contains randomly selected D_t distinct elements of the set element domain. Then, for a given value of F and $m = 2$, target signatures are created. Finally, for each query condition and a given value of D_q, ten query sets of cardinality D_q are generated using the same element domain, and the false drop probabilities are actually measured.

Simulation results with the target set cardinalities $D_t = 10$ and $D_t = 100$ are presented for queries $T \supseteqq Q$ (including $T \ni q$), $T \subseteqq Q$, $T \sqcap Q$ ($T \sqcap_1 Q$, $T \sqcap_2 Q$), and $T \equiv Q$. In the following discussion, we show simulation results and estimates by the formulas in Sect. 3. In figures, we tag simulation results with `sim` and estimates by the formulas F1, F2, and F3 with `F1`, `F2`, `F3`, respectively. In computing the estimates, we tried both the formulas without the approximation by Eqs. (6') and (8') (such as formulas (13), (14), and (16)) and those based on the approximation (such as formulas (13'), (14'), and (16')). Only in F3, we recognized a very small difference caused by rounding off $Fe^{-\frac{mD_t}{F}}$ in computing the former set of formulas. Except for this, we could find no recognizable difference between the two sets of formulas. In addition, the approximation generally contributes to the reduction of the computation cost. For these reasons, estimates by the formulas, if applicable, incorporating the approximation are presented below.

(1) $T \supseteqq Q$

Figure 7 shows the simulation results and estimates by Eqs. (13'), (18), and (24) for $D_t = 10$[†]. Figure 8 shows the case of $D_t = 100$. We can see that (a) there is almost no difference between false drop probabilities given by Eqs. (13'), (18), and (24), and that (b) the three equations give good estimates of the actual false drop probabilities.

(2) $T \subseteqq Q$

Figure 9 shows the simulation results and estimates by Eqs. (14'), (19), and (25) for $D_t = 10$. Figure 10 shows the case of $D_t = 100$. We can see that (a) the false drop probability given by Eq. (14') tends to be smaller than

[†] As aforementioned, Eq. (6') is used in Eqs. (18) and (24). Similar remarks apply to some of the remaining estimates.

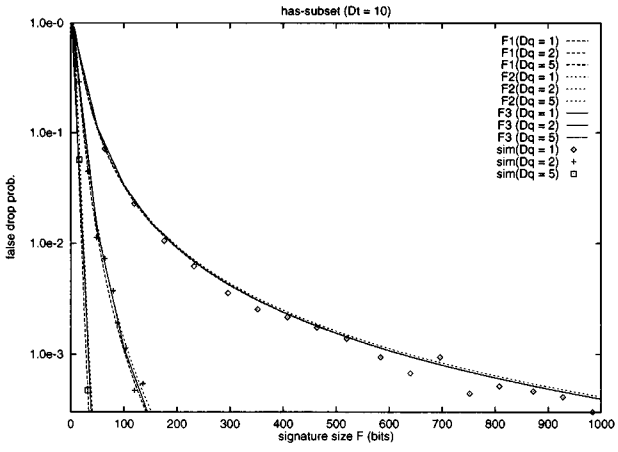

Fig. 7 $T \supseteq Q$.
($D_t = 10$, including $T \ni q$)
F1: Eq. (13'), F2: Eq. (18), F3: Eq. (24)

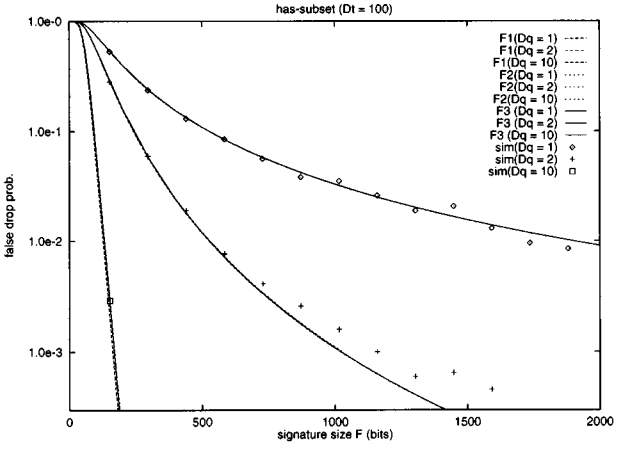

Fig. 8 $T \supseteq Q$.
($D_t = 100$, including $T \ni q$)
F1: Eq. (13'), F2: Eq. (18), F3: Eq. (24)

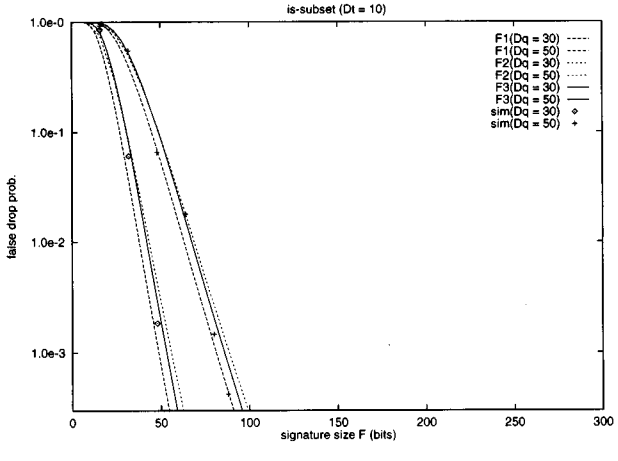

Fig. 9 $T \subseteq Q$ ($D_t = 10$).
F1: Eq. (14'), F2: Eq. (19), F3: Eq. (25)

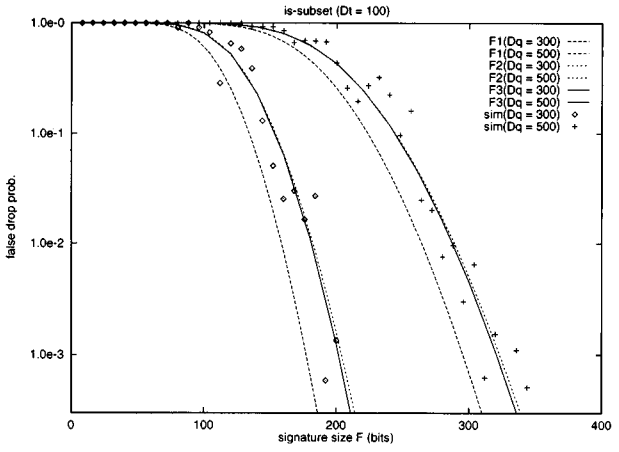

Fig. 10 $T \subseteq Q$ ($D_t = 100$).
F1: Eq. (14'), F2: Eq. (19), F3: Eq. (25)

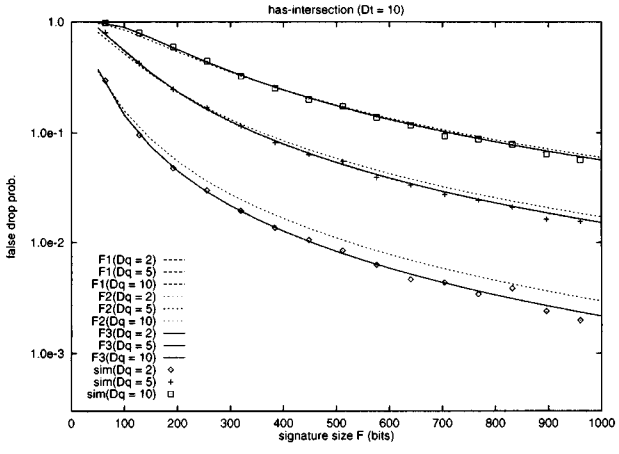

Fig. 11 $T \sqcap_1 Q$ ($D_t = 10$).
F1: Eq. (15), F2: Eq. (20), F3: Eq. (26)

those given by Eqs. (19) and (25), and that (b) the false drop probabilities given by Eqs. (19) and (25) coincide well with the simulation results. Therefore, we can see that Eqs. (19) and (25) give more correct estimates of the false drop probabilities.

(3) $T \sqcap Q$

1) $T \sqcap_1 Q$

Figure 11 shows the simulation results and estimates by Eqs. (15), (20), and (26) for $D_t = 10$. Figure 12 shows the case of $D_t = 100$. We can see that (a) Eqs. (15) and (26) give almost the same false drop probabilities coinciding well with the simulation results, and that (b) the false drop probability given by Eq. (20) does not fit the simulation results when D_q is very small. The latter phenomenon can be explained as follows. When D_q is very small, the query signature weight becomes mD_q in most cases. Therefore, the assumption for Eq. (20) that the distribution of the query signature weight follows the binomial distribution is not reasonable in such

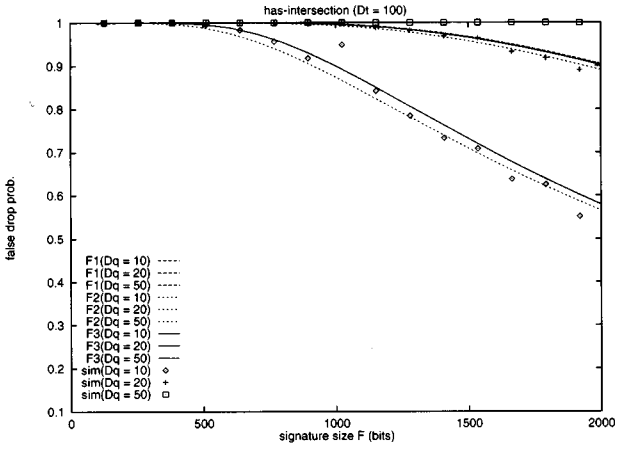

Fig. 12 $T \sqcap_1 Q$ ($D_t = 100$).
F1: Eq. (15), F2: Eq. (20), F3: Eq. (26)

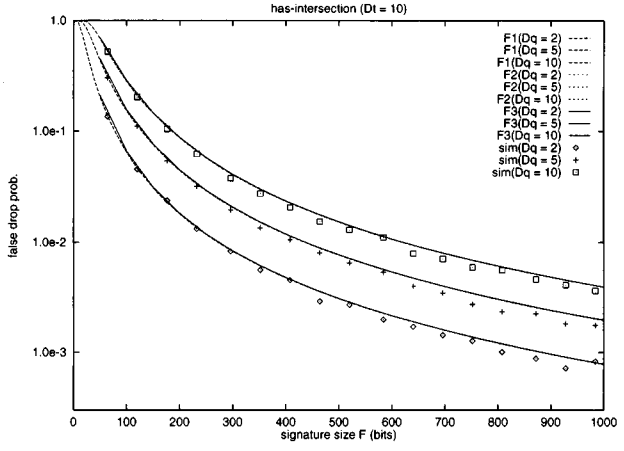

Fig. 13 $T \sqcap_2 Q$ ($D_t = 10$).
F1: Eq. (16'), F2: Eq. (21), F3: Eq. (27)

cases.

2) $T \sqcap_2 Q$

Figure 13 shows the simulation results and estimates by Eqs. (16'), (21), and (27) for $D_t = 10$. Figure 14 shows the case of $D_t = 100$. We can see that (a) there is no difference among false drop probabilities given by Eqs. (16'), (21), and (27), and that (b) they coincide well with the simulation results. We could have foreseen the results, since the false drop probability for $T \sqcap_2 Q$ is based on that for $T \ni q$, and each formula for $T \ni q$ in F1, F2, and F3 quite correctly estimates the false drop probability as shown above.

(4) $T \equiv Q$

Figure 15 shows the simulation results and estimates by Eqs. (17), (22), and (28) for $D_t = 10$. Note that Eq. (17) can be used only when $D_t = D_q$. Figure 16 shows the case of $D_t = 100$. We can see that (a) the false drop probability given by Eq. (17) does not fit the simulation results for large D_t-values, and that (b) Eqs. (22) and

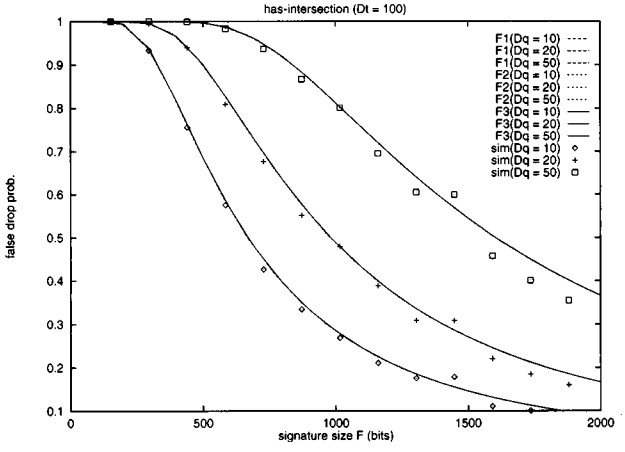

Fig. 14 $T \sqcap_2 Q$ ($D_t = 100$).
F1: Eq. (16'), F2: Eq. (21), F3: Eq. (27)

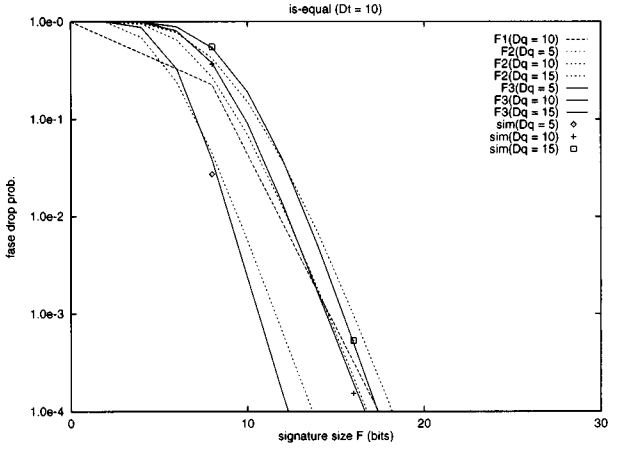

Fig. 15 $T \equiv Q$ ($D_t = 10$).
F1: Eq. (17), F2: Eq. (22), F3: Eq. (28)

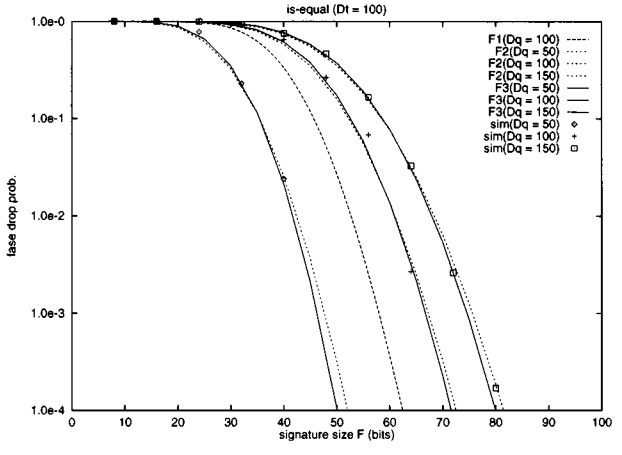

Fig. 16 $T \equiv Q$ ($D_t = 100$).
F1: Eq. (17), F2: Eq. (22), F3: Eq. (28)

(28) give good estimates in general but Eq. (28) seems to be slightly more reliable than Eq. (22).

5. Summary and Conclusion

In this paper, we have derived three sets of formulas F1, F2, and F3, estimating false drops in set retrieval with signature files based on the $T \supseteq Q$ (has-subset), $T \subseteq Q$ (is-subset), $T \sqcap Q$ (has-intersection), and $T \equiv Q$ (is-equal) conditions. For $T \sqcap Q$, two processing schemes $T \sqcap_1 Q$ and $T \sqcap_2 Q$ have been considered. For each condition, their estimates have been evaluated with computer simulations. In deriving the estimates, we have experimentally checked that the approximation by Eqs. (6') and (8') can be used without causing serious problems in situations under consideration. Observations in the simulation study can be summarized as follows:

$T \supseteq Q$: There is essentially no difference among the three formulas F1, F2, and F3, and each gives good estimates.

$T \subseteq Q$: F2 and F3 give good estimates. F1 tends to give smaller estimates.

$T \sqcap_1 Q$: F1 and F3 give good estimates. F2 does not fit the simulation results when D_q is very small.

$T \sqcap_2 Q$: The three formulas give good estimates.

$T \equiv Q$: F2 and F3 give good estimates, but F3 seems to be slightly more reliable. F1 is applicable only when $D_t = D_q$ and is not very reliable.

As for validity of the estimates, the formulas F3 have a more sound mathematical basis and generally give correct estimates. The formulas F2 also give good estimates for most of the cases. The formulas F1 fit the simulation results only for $T \supseteq Q$, $T \sqcap_1 Q$, and $T \sqcap_2 Q$. As far as validity is concerned, we can conclude that the formulas F3 are reliable for all cases, and that the formulas F2 are slightly inferior to the formulas F3. The formulas F1 are useful for rough estimation, but the formula for $T \equiv Q$ in F1 is only applicable to very limited cases.

As for the computation cost, the formulas F1 obviously have the smallest cost. The formulas F2 and F3 have their advantages and disadvantages. In F2, the distribution of the signature weight is assumed to follow the binomial distribution. Therefore, we need to get the sum for a very large space of index values to compute the false drop probability. For example, Eq. (18) involves the summation from $i = 0$ to F. On the other hand, the formulas F3 only require the summation for actually possible weight values. For example, Eq. (24) only needs the summation from $i = m$ to mD_t. For small m-values, this property contributes to reduction of the computation cost. A disadvantage of the formulas F3 is that their computation requires maintaining many significant digits. Since the probability distribution formula for the set signature weight (Eq. (23)) has a term $(-1)^{w-m_1+j}$, loss of significant digits often occurs. To prevent this, we have to maintain many significant digits. Thus, the superiority of F2 and F3 in terms of the computation cost heavily depends on the available mathematical computation tools.

Efficient set retrieval facilities are indispensable not only for advanced database systems but also for advanced information systems in general which handle various complex data objects. The signature file method is one of very promising approaches to efficient set handling. The analytical study presented in this paper provides a sound mathematical basis in design and development of advanced database systems based on the signature file method. In this study, we have assumed that there is no skew in occurrences of members in the set element domain. Analysis of the effect of such skew on the false drop probability is an important future research issue.

Acknowledgement

The authors would like to thank Mr. Yoshiaki Fukushima and Mr. Noriyasu Watanabe for their contribution to the study of signature files. They are also grateful to the anonymous referees whose comments and questions led to improvement of this paper. This work was supported in part by the Ministry of Education, Science, Sports and Culture, Japan under the Grant-in-Aid for Scientific Research on Priority Areas, No. 08244101 and No. 08244104.

References

[1] S. Abiteboul, P.C. Fisher, and H.-J. Schek, eds., "Nested Relations and Complex Objects in Databases," Lecture Notes in Computer Science, vol.361, Springer-Verlag, Berlin, 1989.

[2] E. Bertino and W. Kim, "Index techniques for queries on nested objects," IEEE Trans. Knowl. Data Eng., vol.1, no.2, pp.196–214, June 1989.

[3] R.G.G. Cattell, ed., "The Object Database Standard: ODMG-93," Morgan Kaufmann Publishers, San Francisco, 1996.

[4] C. Faloutsos and S. Christodoulakis, "Signature files: An access method for documents and its analytical performance evaluation," ACM Trans. Office Inform. Syst., vol.2, no.4, pp.267–288, Oct. 1989.

[5] C. Faloutsos, "Access methods for text," ACM Computing Surv., vol.17, no.1, pp.49–74, March 1985.

[6] C. Faloutsos and S. Christodoulakis, "Description and performance analysis of signature file methods for office filing," ACM Trans. Office Inform. Syst., vol.5, no.3, pp.237–257, July 1987.

[7] C. Faloutsos, "Signature files: An integrated access method for text and attributes, suitable for optical disk storage," BIT, vol.28, pp.736–754, 1988.

[8] C. Faloutsos, "Signature-based text retrieval methods: A survey," IEEE Database Eng., vol.13, no.1, 25–32, March 1990.

[9] P.C. Fischer and S.J. Thomas, "Operators for non-first-normal-form relations," Proc. IEEE COMPSAC 83, pp.464–475, 1983.
[10] Y. Ishikawa, H. Kitagawa, and N. Ohbo, "Evaluation of signature files as set access facilities in OODBs," Proc. ACM SIGMOD Conf., pp.247–256, May 1993.
[11] W. Kim and F.H. Lochovsky, eds., "Object-Oriented Concepts, Databases, and Applications," ACM Press, New York, 1989.
[12] H. Kitagawa and T.L. Kunii, "The Unnormalized Relational Data Model—For Office Form Processor Design," Springer-Verlag, Tokyo, 1989.
[13] H. Kitagawa, Y. Fukushima, Y. Ishikawa, and N. Ohbo, "Estimation of false drops in set-valued object retrieval with signature files," Proc. 4th Intl. Conf. on Foundations of Data Organization and Algorithms (FODO), Lecture Notes in Computer Science, vol.730, Springer-Verlag, Berlin, pp.146–163, 1993.
[14] W.-C. Lee and D.L. Lee, "Signature file methods for indexing object-oriented database systems," Proc. Intl. Computer Science Conf. (ICSC), pp.616–622, Hong Kong, 1992.
[15] E. Murphree and D. Aktug, "Derivation of probability distribution of the weight of the query signature," Department of Mathematics and Statistics, Miami University, Oxford, OH 45056, 1992.
[16] Z.M. Ozsoyoglu, ed., "Special issue on nested relations," IEEE Database Eng., vol.11, no.3, 1988.
[17] J.L. Pfaltz, W.J. Berman, and E.M. Cagley, "Partial-match retrieval using indexed descriptor files," Commun. ACM, vol.23, no.9, pp.522–528, Sept. 1980.
[18] R. Sacks-Davis, A. Kent, and K. Ramamohanarao, "Multikey access methods based on superimposed coding techniques," ACM Trans. Database Syst., vol.12, no.4, pp.655–696, Dec. 1987.
[19] H.-J. Schek and M.H. Scholl, "The relational model with relation-valued attributes," Inf. Syst., vol.11, no.2, pp.137–147, 1986.
[20] J. Stein and D. Maier, "Associative access support in GemStone," in On Object-Oriented Database Systems, K.R. Dittrich, U. Dayal, and A.P. Buchmann, eds., Springer-Verlag, Berlin, pp.323–339, 1991.
[21] K.F. Wong and M.H. Williams, "A superimposed codeword indexing schema for handling sets in Prolog databases," Proc. 2nd Intl. Symp. on Database Systems for Advanced Applications (DASFAA), pp.468–476, April 1991.
[22] H.-S. Yong, S. Lee, and H.-J. Kim, "Applying signatures for forward traversal query processing in object-oriented databases," Proc. 10th Intl. Conf. on Data Eng., pp.518–525, Feb. 1994.
[23] S. Zdonik and D. Maier, eds., "Readings in Object-Oriented Database Systems," Morgan Kaufmann Publishers, San Mateo, 1990.

Hiroyuki Kitagawa received the B.Sc. degree in physics and the M.Sc. and D.Sc. degrees in computer science, all from the University of Tokyo, in 1978, 1980, and 1987, respectively. He is an associate professor at Institute of Information Sciences and Electronics, University of Tsukuba, Japan. Before joining University of Tsukuba in 1988, he was a research staff member at Software Product Engineering Laboratory of NEC Corporation. His research interests include integration of heterogeneous information sources, object database systems, temporal database and version management, query processing, and text databases. Dr. Kitagawa is a member of ACM, IEEE Computer Society, IPSJ, and JSSST.

Yoshiharu Ishikawa received the B.S., M.E., Dr.Eng. degrees in information engineering from University of Tsukuba in 1989, 1991 and 1995, respectively. He is currently a research associate at Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan. His research interests include the design and implementation of advanced DBMSs and scientific databases. He is a member of the ACM, IEEE Computer Society, IPSJ, and JSSST.