불확실한 시간 간격을 위한 퍼지 최소 간격 분할 기법

허 문 행[†] · 이 광 규[††] · 이 준 욱[†††] · 류 근 호[††††] · 김 홍 기[†††††]

요 약

시간 데이터베이스에서 데이터의 이력 관리를 목적으로 확장된 시간 차원은 조인 연산에 대한 복잡도와 연산 비용의 증가를 초래한다. 이와 같은 문제를 해결하기 위해 조인 연산의 대상이 되는 시간 범위를 일정 정도의 시간 간격으로 분할한 후 분할된 세그먼트 단위의 시간 데이터들을 조인하는 방법이 제안되었다. 하지만 기존의 방법들은 분할점 근방에서 분할에 적용된 시간 단위 등의 문제로 인해 발생하는 시간 경계의 모호성 문제를 해결할 수 없다. 이러한 문제를 해결하기 위해 이 논문에서는 분할선 근방의 경계가 불확실한 시간 간격을 고려한 분할 방법으로 퍼지 이론의 가능성 분포를 도입한 퍼지 최소 간격 분할(Fuzzy Minimum Interval Partition : FMIP)기법을 제안하고 이를 기존 기법들과 비교 평가함으로써 FMIP의 타당성을 검증한다.

Fuzzy Minimum Interval Partition for Uncertain Time Interval

Moon Haeng Huh[†] · kwang kyu Lee[††] · Jun Wook Lee[†††]
Keun Ho Ryu[††††] · Hong Gi Kim[†††††]

ABSTRACT

In temporal database, extended time dimension for history management brings about complexity of join operation and increased cost. To solve this problem, a method that joins the divided segment time data after partition the time range into fixed time interval is introduced. But existing methods can't solve the ambiguity problem of time border that caused by temporal granularity in the partition point. In this paper, We suggested Fuzzy Minimum Interval Partition (FMIP) method that introduced the possibility distribution of fuzzy theory considered uncertainty time interval border in the partition line.

키워드 : 퍼지 최소 간격 분할(Fuzzy Minimum Interval Partition : FMIP), 간격 분할(interval partition), 퍼지 이론(fuzzy theory), 불확실성(uncertainty), 가능성 분포(possibility distribution)

1. 서 론

인공지능과 데이터베이스 분야는 의사결정, 의료진료, 범죄판정 등 많은 응용 분야에 있어서 불확실(uncertainty)하거나 모호한(ambiguity)정보를 다루고, 지식 기반 및 데이터베이스 시스템은 불확실한 데이터를 저장하고 처리하는 기능을 제공하였다. 최근에 인터넷의 확산 및 각종 멀티미디어 데이터의 증가와 함께 새롭고 복잡한 데이터 유형들 즉, 공간, 시간, 오디오와 비디오 데이터 등이 보편화되면서 이들 새로운 유형의 데이터들을 데이터베이스를 통해 관리하기 위한 데이터 모델이 소개되었다[1, 2]. 이와 같은 응용 분야에서 데이터의 이력 관리 또는 객체의 상대적 시점을 관리하는 데이터의 시간 차원은 매우 중요한 요소로서 IXSQL[3], TQUEL[4] 및 TSQL2[5]등과 같은 연구에서는 시간 데이터베이스를 위한 데이터 모델, 연산 및 질의어 표준에 대한 다양한 연구를 제시하였다.

실세계 정보에 대하여 릴레이션들은 실세계의 불확실한 시간 간격 정보를 모델링할 수 있다. 또한 릴레이션들은 서로 다른 시간정보 측정 방법에 따라 서로 다른 시간단위(granularity)를 갖는 타임스탬프로써 표현된다. 예를 들어 특정 지역을 움직이는 이동 사용자와 네트워크 상태 정보의 경우 사용자 이동 정보는 불확실한 시간 간격정보를 표현할 필요가 있으며 두 릴레이션의 경우 타임스탬프 단위가 불일치하고 있어 두 릴레이션에 대하여 시간 데이터베이스 질의 처리는 불확실성 또는 애매성 문제를 해결해야 한다. 실세계 정보의 불확실성은 매우 빈번하기 때문에 이를 표현할 수 있어야 하며 이를 표현한 릴레이션 등에 관한 최적화 분할은 시간 데이터 베이스 연산에 있어 매우 중요하다.

시간 데이터베이스에서 시간 간격(time interval)은 질의처리 시 여러 문제점을 제기하는데 그 중 하나는 데이터의 부분 검색이나 조인 연산 등에 적용하기 위해 수행되는 시간 간격 분할(time interval partition)[6-14]이다. 시간 간격 분할시 여러 단편(fragment)들이 생성되고 어떤 특정한 시

※ 이 연구는 한국 과학재단 특정기초연구(R01-1999-00243) 지원으로 수행되었음.
† 정 회 원 : 한국소프트웨어진흥원 단장
†† 정 회 원 : 신흥대학 컴퓨터정보계열 교수
††† 준 회 원 : 충북대학교 대학원 컴퓨터과학과
†††† 종신회원 : 충북대학교 전기전자 및 컴퓨터공학부 교수
††††† 정 회 원 : 충북대학교 전기전자 및 컴퓨터공학부 교수
논문접수 : 2002년 2월 18일, 심사완료 : 2002년 5월 10일

간 도메인에서 이들 데이터 객체들은 서로 다른 단편을 교차(intersect)한다. 시간 조인을 수행하기 이전에 시간 간격을 분할하는 가장 큰 문제점은 분리(disjoint)가 아닌 데이터 단편을 생성하는 것이다. 이는 질의 처리시 상당한 연산 부하를 준다. 그러므로 시간 데이터베이스에서 시간 간격 분할은 신중하고 타당성 있는 논리 전개가 필요하다. 지금까지 많은 연구가들에 의해 간격 분할 문제를 다루는 알고리즘 설계[6, 7, 9, 10, 12, 14]와 인덱스 구조[13, 15-18]가 소개되었지만 가장 최적의 분할점을 구할 수 있는 방법은 제시된 적이 없다. 특히 간격 분할 과정에서 분할에 적용되는 시간 단위(time granularity) 문제로 인해 필연적으로 발생하는 분할선 근방의 모호성 문제를 고려한 최적 분할 기법에 대한 연구는 미흡한 실정이다. 따라서 이와 같은 문제를 해결하기 위해 이 논문에서는 분할선 근방에서 발생되는 불확실한 시간 간격의 분할선을 퍼지 이론의 가능성 분포를 도입하여 시간 속성인 타임스탬프의 시작시점이 모호한 상황에서도 간격을 분할할 수 있는 퍼지 최소 간격 분할(Fuzzy Minimum Interval Partition : FMIP)기법을 제안하고 기존 간격 분할 기법들과의 비교 평가를 통해 FMIP의 타당성을 검증한다.

이를 위한 논문의 구성은 다음과 같다. 2장에서는 시간 데이터베이스의 시간 변수의 불확실성과 관련된 문제 및 시간 조인을 위한 시간 간격 분할 기법들에 대한 기존 연구들을 고찰한다. 3장에서는 시간 간격의 불확실성을 퍼지 이론을 이용하여 모델링하기 위한 사항들을 기술하고 4장에서 3장의 이론적 배경을 토대로 MIP 알고리즘을 개선한 FMIP 분할 기법을 제안한다. 또한 5장에서는 4장에서 제안된 FMIP 분할 기법에 대한 실험 및 평가를 수행하고 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

시간이라는 개념은 항상 시간단위를 가지고 있으며, 시간의 연속성이라는 것은 타임스탬프에 의해 결정되므로 실질적으로는 시간 단위에 의한 시간 간격을 갖게된다. 그러므로 당연히 시간 단위에는 불확실성이 동반된다고 할 수 있다[19]. 또한 정의된 시간 단위보다 더 작은 단위의 시간 단위가 요구될 경우 불확실성이 발생한다. 따라서 현실세계에서 불확실성이 발생하게 되는 원인은 시간 단위에서의 애매 모호성과 시간 정보 자체의 불확실성 등에 있으며 이러한 원인들로 인해 발생하는 시간 데이터베이스에서의 시간 변수로 인한 불확실성과 시간 조인과 같은 연산 시 불확실성의 해결은 정확하고 신뢰성 있는 정보를 사용자들에게 제공하기 위해 반드시 필요하다. 불확실성 처리의 중요성은 [29]의 연구에서 많은 예를 찾을 수 있다.

불확실성을 갖는 시간 조인의 해결은 간격 분할시 다양한 범위를 정의해 줄 수 있는 퍼지 이론의 가능성 논리가 제기될 수 있다[20, 21]. 자데의 가능성 이론은 애매하고 불확실한 시간 지식(temporal knowledge)을 모델링하는 구조로 사용되어 왔으며, 잘 모르는 시점(ill-known time point), 퍼지 경계를 갖는 시간 간격(fuzzy boundary time interval), 퍼지 지속시간(fuzzy duration), 사건들의 불확실한 관계 등은 가능성 분포(possibility distribution)의 의미에서 정확하거나 확실하지 않은 유용한 정보들을 표현할 수 있다[21]. 또한, [22]에서 시간은 주관적으로 인지 가능하며 "after"(직후), "before"(직전)를 표현하는 순서화된 관계 값을 소개하기도 하였다. [21]에서 시간 지식에 관한 퍼지화는 정보의 일부 문맥에서 사용되는 자연어 표현이 가능한 퍼지 시간 간격(fuzzy temporal interval)으로 after와 before등을 퍼지 언어항으로 모델링하여 불확실한 시점을 처리하였으며, [23]은 [24]에 의해 처음으로 제안된 시간 간격의 퍼지 집합을 사건(event)의 지식 표현으로 모델링하였다. 또한 [25]는 믿음 함수(belief funtion)를 이용하여 시간축에 애매함과 불확실성을 표현하였다. 한편 [26]은 데이터의 불확실성과 시간을 결합한 데이터베이스 모델을 제안하였다. [26]에서 제시한 모델은 퍼지집합에 기반한 가능성이론(possibility theory)과 언어변수를(linguistic variable) 사용하여 불확실하고 애매한 자료를 처리하기 위한 정량자(quantitative), 정성자(qualitative), 다중값 논리(multivalued logic)등 세 개의 도메인을 모델링 하였으나 구체적인 실험결과를 제시하지 못하였다.

시간 조인의 간격 분할에 관한 기존의 연구로 제시된 균등 분할(uniform partition)[16], 언더플로우(underflow)분할[18]은 알고리즘이 단순하고 실행 시간이 빠르다는 장점이 있지만 불확실한 시간 간격의 분할점을 결정 할 수 없고, 또한 단편내 많은 중복이 발생하여 조인 연산시 전체적인 성능이 저하되는 것이 문제점으로 지적되었다. 간격의 중복을 최소로하는 분할점을 결정하는 최소 간격 분할(Minimum Interval Partition : MIP)기법[14]은 단편내 중복을 최소로하는 분할점을 결정할 수 있으나 MIP 역시 분할선 근방의 모호한 값은 최소 분할선으로 결정 할 수 없다는 것이 단점으로 제기되었다. 이 밖에도 [27, 28]연구는 분할에 기반한 시간 조인으로 시간인덱스의 사용의 중요성을 언급하였지만 퍼지 시간 분할에 대하여는 제시되지 못하고 있다.

이 논문에서는 이와 같은 문제를 해결하기 위해 퍼지 이론의 가능성 분포를 도입하여 일정값 이상을 만족하는 임계값(threshold)을 최소 분할선으로 결정하는 퍼지 최소 간격 분할(Fuzzy Minimum Interval Partition : FMIP)기법을 제안한다. 제안된 기법은 기존의 시간 간격 분할 기법으로 사용된 MIP 알고리즘을 기반으로 시간경계가 모호한 경우의 분할설 결정을 위해 적용될 수 있는 기법으로, 종료시점 및 종료시점 근방에서 최적의 간격 분할선이 결정됨을 보여준다. 제시된 기법은 불확실성을 갖는 시간 데이터의 최적 분할을 통하여 애매성을 갖는 기존 시간 조인의 문제를 해결할 수 있으며 또한 성능과 신뢰도 향상에 있어 시간

조인 연산 시 필수적인 기법이다.

3. 불확실성을 표현한 퍼지 언어항

자데의 가능성 이론은 불확실한 시점, 퍼지 경계를 갖는 시간 간격, 퍼지 지속시간, 사건 발생의 정확한 시간 등 애매하고 불확실한 시간 지식을 모델링하는 구조로 사용되었다[20]. 불확실한 시작시점 a가 정확히 시간 t를 가질 가능성의 수치적 표현은 $\pi_a(t)$로 표현하며, $\pi_a(t)=0$ 일 때 a는 분명히 t와 다른 것을 의미한다. $\pi_a(t)$가 점점 커질때 a는 t와 같아질 가능성이 커지며, $\pi_a(t)=\pi_a(t')=1$이 되는 t와 t'이 존재할 수 도 있다. 실제로, '시점 a가 t일 가능성은 1이다' 라고 하는 것은 'a가 t와 확실히 같다'라고 하는 것보다 약한 표현이다[20]. 왜냐하면, a가 취할 수 있는 가능성 1인 또 다른 t'의 존재성을 배제할 수 없기 때문이다. π_a는 모호한 시점 a값의 퍼지 집합 A를 정의할 수 있는데, 그 소속함수 μ_A는 모든 $t\in$T에 대해서

$$\mu_A(t)=\pi_a(t) \tag{1}$$

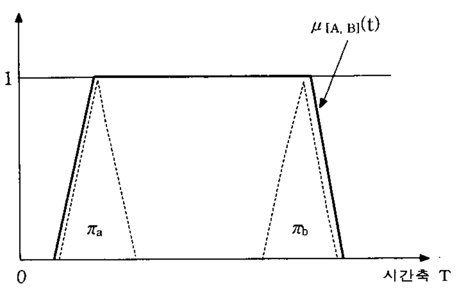
로 정의하고, 가능성 분포 π_a로서 시점의 언어적 표현으로 정의된다. 또한 시점이 불확실한 경우의 직후와 직전은 퍼지 시간 간격의 교집합으로 다음과 같이 정의한다[21].

【정의 1】 퍼지 시간 간격(fuzzy time interval)

불확실한 시점 a와 b에 대한 가능성 분포를 각각 π_a와 π_b라 하고, 퍼지집합을 각각 A와 B라 하면 a 직후와 b 직전의 시간점들의 퍼지집합 $[A, B]$는 다음과 같은 소속함수로 정의된다.

$$\forall t\in \mathrm{T},\ \mu_{[A,B]}(t) = \sup_{s \le t \le s'} \min(\pi_a(s),\ \pi_b(s')) \tag{2}$$

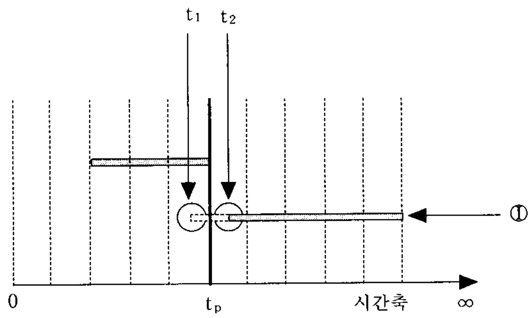
즉, $[A, B]$는 $[A, +\infty)$과 $(-\infty, B]$의 퍼지 교집합(fuzzy intersection)에 의해 얻어지며, 이 논문에서는 시작시점이 불확실하게 표현된 퍼지 시간 간격의 분할점을 결정하는 방법을 (그림 1)과 같이 α-절단과 가능성 분포를 이용하여 불확실한 상황에서 간격 분할 문제를 해결한다.



(그림 1) 퍼지 시간 간격

4. 시작시점이 불확실한 FMIP

불확실한 퍼지 시간 간격을 포함하는 간격 분할에서 시작시점의 모호한 상태는 분할선 근방에서 직전에 시작하거나 직후에 시작하느냐에 따라 분할선 결정이 달라질 수 있다. 시작시점이 모호하여 분할선 결정이 변경되면 조인 연산 비교 횟수가 달라져 전반적인 조인 성능에 영향을 미친다[14]. 하지만 분할선 근방에 가장 가깝게 근접한 경계점을 최적의 분할선으로 결정하기 위해서는 퍼지 이론의 가능성 분포로 접근 할 수 있다. 즉, 시작시점이 분할선 근방에서 직전에 시작한다고 하면 MIP와 동일하게 임의의 단편내 간격수 X를 초과하지 않으므로, α-절단을 이용하여 임계값 이상을 만족하는 값을 간격의 분할선으로 결정할 수 있다. (그림 2)에서 간격 ①의 시작시점을 a라 하면 a가 모호하여 분할선 t_p 이후, 즉 $a(=t_2) > t_p$ 인 경우는 단편내 X를 초과하지 않으므로 분할선은 $t_p \le$ 분할선 $\le t_2$의 범위내에서 결정되지만, $a(=t_1) < t_p$ 인 경우는 간격수가 X를 초과하여 이 경우는 α-절단을 이용하여 임의의 임계값을 만족하는 간격을 최소 분할선으로 결정할 수 있다.



(그림 2) 분할선 t_p 근방에서 간격 ①의 시작시점이 t_1 또는 t_2인지가 불확실한 경우

간격 ①의 시작시점을 식 (3)으로 정의된 가능성 분포로 표현했을 때, (그림 3)과 같은 사다리꼴 퍼지수로 주어진다면,

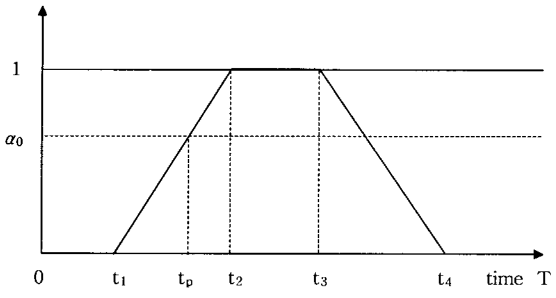
$\alpha > \alpha_0$인 α-절단을 취했을 경우 t_p가 분할점이 될 수 있고,

$\alpha < \alpha_0$인 α-절단을 취했을 경우는 t_p가 분할점이 될 수 없다.

즉, 임계값을 만족하는 α_0를 기준으로 간격의 분할점을 결정할 수 있다.

다음과 같은 가능성 분포로 결정된 분할점은 시간 간격을 최적으로 분할하여 단편과 중복의 합을 최소로하는 MIP 알고리즘[14]에 적용하여 최소 분할점을 결정할 수 있다. MIP 알고리즘은 간격의 중복을 최소로하는 분할점을 종료시점에서 결정하지만 분할선 근방이 모호한 경우에는 분할선을 결정할 수 없다. 따라서 이와 같이 애매하고 모호한경우는 FMIP의 가능성 분포를 이용한 분할선 결정 과정을 통해 임의의 임계값 이상을 만족하는 종료시점 근방의 분

$$\pi_a(t) = \begin{cases} 0, & t < t_1 \ \text{또는}\ t > t_4 \\ \dfrac{t-t_1}{t_2-t_1}, & t_1 \le t < t_2 \\ 1, & t_2 \le t \le t_3 \\ \dfrac{t-t_4}{t_3-t_4}, & t_3 < t \le t_4 \end{cases} \tag{3}$$



(그림 3) 간격의 불확실한 시작시점 a의 가능성 분포 π a(t)

할선을 최소 분할선으로 결정할 수 있다. 이렇게 최소 분할선을 결정하여 전체 데이터에 대한 간격 분할 수행은 MIP 알고리즘을 이에 맞게 수정함으로써 수행할 수 있다. 다음의 (그림 4)는 FMIP 알고리즘을 간략히 나타낸 것이다. (그림 4)에서 변수 c[i]는 단편내 중복의 누적수를 관리하는 변수이며 fwd[i]는 최소 분할점을 저장하기 위한 변수이다. 또한 min_bnd는 분할대상 단편내 최소 분할점을 결정하는데 사용되는 변수이며 min_j는 최소중복 발생위치를 나타내는 변수이고, End_Vector는 분할에 참여하는 시간 간격들의 종료점을 갖고 있는 변수이다. 또한 각각의 단편을 나타내는 $F_k(i, j)$는 정지점 j에서 간격 종료점 i까지 단편내에 존재하는 간격수를 의미한다. MIP알고리즘과 달리 FMIP에서는 불확실한 시작시점을 갖는 간격들에 대하여 α-절단을 이용하여 임의의 임계값 threshold를 만족하는 간격을 최소 분할선을 결정하기 위하여 alpha-partition()함수를 이용하며 계산된 퍼지간격 fuzzy_interval 등이 중복수 계산 등에 포함된다. 따라서 FMIP 알고리즘은 우선 min_bnd와 c[i] 변수를 이용해 최소 분할선의 후보를 선별한 후 이들 최소 분할선 후보들 중에서 최소 분할점을 결정하여 fwd[i]에 할당함으로써 알고리즘을 수행한다.

```
c[0] = 0 ;
break_ndx = 0 ;
fwd_cnt = 1 ;
for ( i = 1 ; i <= n ; i++ ){  /* 간격 종료점들의 수 만큼 순환 */
    min_j = NULL ;
    min_bnd = MAX_Cardinalty ;
    for ( j = break_ndx ; j < i ; j ++ ) {
                        /* 정지점부터 간격 종료점까지 순환 */
        qi = End_Vector[i] ;
        qj = End_Vector[j] ;
        fuzzy_interval = alpha_partition(qi, threshold) ;
                        /* a-절단시의 퍼지 간격수 */
        if (Fk (i, j) + fuzzy_interval) ≤ X ) {
            tmp_bnd = OR(qj) + c[j] + fuzzy_interval ;
                        /* 퍼지 간격을 고려한 중복수 계산 */
            if ( tmp_bnd < min_bnd ){
                        /* 계산된 중복수의 최소 중복 여부 조사 */
                min_bnd = tmp_bnd ;
                        /* 계산된 중복수를 최소 중복으로 설정 */
                min_j = j ;     /* 최소중복 위치 설정 */
            }
        }
    }
    c[i] = min_bnd
    if ( min_j != NULL ) {     /* 최소 중복 위치가 존재하면 */
        fwd[fwd_cnt++] = End_Vector[min_j] ;
                        /* 종료점을 최소중복 위치로 설정 */
        break_ndx = min_j ;   /* 다음 중복검사 시작위치를 최소
                                 중복 발생위치로*/
    }
}
```

(그림 4) FMIP 알고리즘

5. FMIP 실험

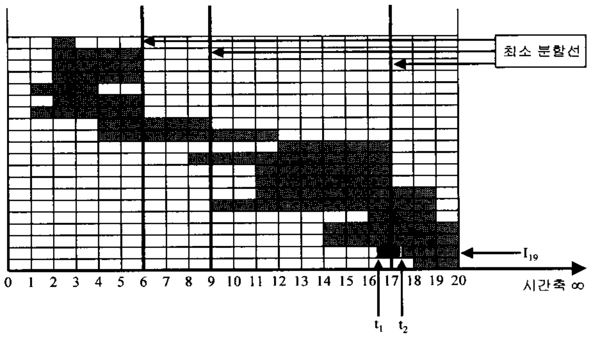
5.1 시나리오 환경 및 실험 방법

이 논문에서 제안한 기법의 평가를 위해서 알고리즘은 C 언어를 사용하여 구현하였으며 성능평가를 위해서 (그림 5)의 시나리오는 가능성 분포를 이용하여 표현하였다. 그리고 성능평가를 위해 사용된 시나리오 데이터는 임의의 간격을 20개로, 모든 간격의 경우에도 동일하게 적용되지만 임의의 간격 I_{19}의 시작시점이 t_1 또는 t_2인지가 불확실하다고 가정하였다. 성능 평가를 위해 기존 분할 기법으로 사용된 균등분할, 언더플로우 분할을 선정하였다. 균등 분할은 5개의 크로논으로 분할하였으며, 언더플로우와 MIP는 단편내 최대 간격수 X를 10개로 제한하였다. 또한, 조인 연산시 투플의 비교 횟수는 병렬 조인 연산의 대칭 분할[12]에 적용하였고 분할선이 명확하게 종료시점에서 교차된다면 MIP와 동일하게 종료시점을 최소 분할선 후보로 선택하지만 경계점이 모호한 경우는 FMIP의 분할 기법을 적용하여 α-절단값은 0.8이상을 만족하는 임계값으로 결정한다.

이 논문에서 제안한 FMIP기법은 퍼지 시간간격을 고려한 차별적인 분할기법으로써 실험의 비교 대상은 제한적으로 균등분할 및 언더플로우 분할기법에 대한 분할의 최적성에 맞추었다. 평가는 시나리오와 같이 경계점이 모호한 경우에도 최적 분할이 가능한 지를 보여주는 것과 또한 시간 조인 성공률 비교 방법을 채택하였다. 이러한 근거는 단편내의 타임스탬프 교차를 통한 조인 실패율이 분할의 최적성을 보여준다는 것[11]을 기반으로 한다.

따라서 평가에서는 FMIP 기법이 시간 조인과 같은 연산에 적용 시에 어떠한 성능 향상요인을 제공할 수 있는지를 평가 제시하였다.

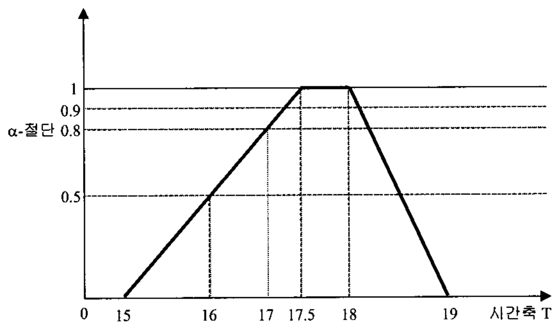
간격 I_{19}의 모호한 시작시점 a의 가능성 분포 $\pi_a(t)$는 식 (3)으로부터 식 (4)로 표현하고, (그림 6)에서 α-절단값 0.8 이상을 만족하는 값 즉, I_{19}의 시작시점이 t_2이면 분할점이



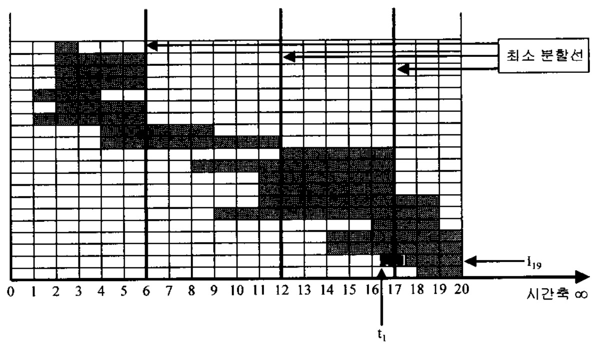
(그림 5) 분할선 17 근방에서 간격 I_{19}의 시작시점이 t_1 또는 t_2 인지가 불확실한 경우

17로 선택되어 [14]에서 제안된 MIP와 동일하게 최소 분할점은 {6, 9, 17}이 선택된다. 하지만 시작시점이 t_1인 경우 α-절단값을 0.8미만으로 취하면, 단편내 최대 간격수 10을 초과하므로 최소 분할점은 MIP 알고리즘에 의하여 17이전에서 종료시점에 가장 가까운 분할점 12가 결정된다. 따라서 분할점은 (그림 7)과 같이 {6, 12, 17}로 변경된다. 그 결과 조인 비교 횟수가 달라지게 된다.

$$\pi_a(t) = \begin{cases} 0, & t < 15 \text{ or } t > 19 \\ \frac{2}{5}t - 6, & 15 \le t < 17.5 \\ 1 & 17.5 \le t \le 18 \\ -t + 19 & 18 < t \le 19 \end{cases} \tag{4}$$



(그림 6) 간격의 불확실한 시작시점 a의 가능성 분포 $\pi_a(t)$



(그림 7) 분할선 17 근방에서 간격 I_{19}의 시작시점이 t_1인 경우

5.2 FMIP 분할 기법의 평가

현실세계는 여러 가지 측면의 관점들이 존재하며, 현실세계 자체가 완전하게 규정되어 있지 않기 때문에 불확실한 정보들이 존재한다. 그러므로 이들 불확실한 데이터를 기존에 사용된 분할 기법으로는 처리가 불가능하다. 따라서 간격의 시작시점이 모호한 경우는 분할점 결정을 할 수 없다는 문제점이 제기된다. 제안된 FMIP 분할 기법은 퍼지 언어항인 직후와 직전을 도입하여, 퍼지 시간 간격으로 확장된 간격의 시작시점이 모호한 경우에도 간격의 최소 분할점을 결정할 수 있었다. 5.1절의 실험에서 FMIP는 모호한 경계점 결정 방법을 제외한 나머지 부분은 모두 MIP와 동일하므로 간격 I_{19}의 근방에 따라 나타나는 결과 역시 MIP와 동일한 결과를 산출한다. 반면 균등 분할 및 언더플로우 분할은 간격 I_{19}의 근방에서는 분할선 결정에 따른 영향을 받지 않는다. 시작시점이 명확한 경우는 MIP 알고리즘에 의하여 종료시점을 최소 분할점으로 결정하지만, 시작시점이 모호한 경우는 FMIP 분할 기법을 적용하여 임계값 이상을 만족하는 분할선 근방의 경계점도 최소 분할점으로 선택할 수 있었다. <표 1>에서는 기존에 제안된 간격 분할 기법과 FMIP 분할기법의 성능을 비교하였으며, <표 2>에서 FMIP 분할 기법으로 결정된 최소 분할점을 토대로 시간 조인을 수행했을 경우의 성능을 비교하였다.

〈표 1〉 명확한 시간 간격 분할과 FMIP 성능 비교

간격 분할 \ 구 분		단편합	중복합	최소 분할점
균등 분할		38	18	×
언더플로우		32	16	×
FMIP	시작시점 = t_1	34	14	종료시점, 종료시점 근방
	시작시점 = t_2	30	10	종료시점, 종료시점 근방

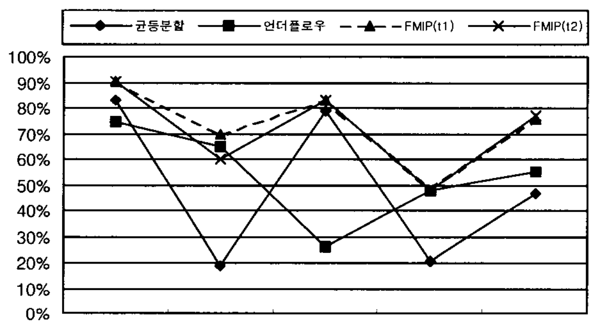
〈표 2〉 FMIP의 시간 조인 성능 비교

간격 분할		투플 비교 횟수	조인 성공	조인 중복	조인 실패	조인 성공률(%)	중복 (%)	실패 (%)
균등 분할		373	173	130	70	46.3	34.8	18.7
언더플로우		313	173	85	55	55.2	27.1	17.5
FMIP	t_1	254	192	66	26	65.3	22.4	8.8
	t_2	234	180	33	21	76.9	14.1	8.9

또한 각 단편내 조인 성공율과 실패율 및 중복율은 (그림 8), (그림 9), (그림 10)에서 분석 제시한다. 여기서 각 단편들 사이에 나타나는 실험 결과의 변화 현상은 실험에 사용된 데이터의 특성과 분할 기법들이 갖고 있는 분할 전략에 따라 나타나는 현상이다. 균등 분할의 경우 분할될 단편내에 포함되는 시간 간격들의 최적화에 대한 고려 없이 항상 일정한 크기의 시간 간격만을 기준으로 분할을 수행한다. 따라서 분할된 단편내에 포함되는 시간 간격의 수에 대한 변동폭이 매우 심하게 나타난다. 언더플로우 분할 기법은 균등 분할이 갖고 있는 문제점인 단편내 시간 간격들을 일정하게 유지할 수 있도록 개선된 방법을 제안하고 있지만 단편들 사이의 중복성에 대한 문제를 고려하지 못하

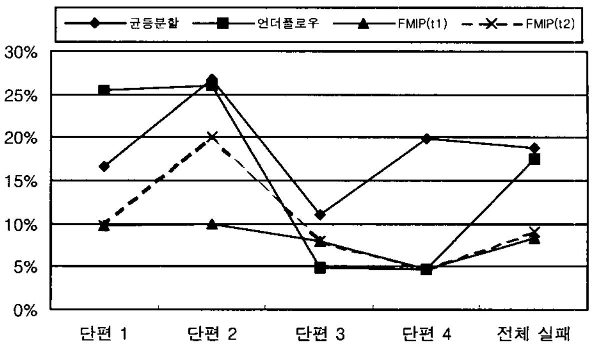
고 있다. 따라서 언더플로우 방법은 단편내에 중복이 과다하게 나타나는 문제가 있다. 이와 같은 현상은 결국 분할된 단편들의 최적화가 이루어지지 않았음을 의미하며 (그림 8), (그림 9) 및 (그림 10)에서 실험 결과가 심하게 변화하는 원인이 된다. 반면 FMIP의 경우 분할된 단편들에 대해 각각의 단편들이 일정한 수에 수렴하는 시간 간격을 포함할 수 있도록 제안되었으며 언더플로우 분할 기법에서 고려하지 못하는 중복 문제를 해결함으로써 분할된 단편의 최적화가 향상된 기법이다. 따라서 (그림 8), (그림 9) 및 (그림 10)의 실험 결과의 변동폭이 상대적으로 적게 나타나고 있다.

조인 성공율은 분할된 간격들 간의 타임스탬프가 교차되는 비율을 의미하며 다음의 (그림 8)은 분할된 각 단편에서 발생한 조인 성공율을 비교 분석한 것으로서 FMIP가 균등분할 및 언더플로우 분할보다 투플 비교 회수 및 조인 성공률이 우수한 것으로 나타났다. 이는 FMIP에 의해 분할된 단편을 이용한 시간 조인 수행 시 최적의 시간 조인을 가능케 한다. 특히 FMIP의 경우 (t_1)과 (t_2)의 두 경우를 나누어 평가하고 있는데 이는 (그림 5)의 시나리오에서 간격 I_{19}가 갖는 시작시점의 모호한 경계값의 성능을 평가하기 위한 것이다. 반면 균등 분할이나 언더플로우 분할의 경우 경계가 모호한 상태를 처리 할 수 없으므로 모호한 경계값을 만날 경우 불확실한 경계값에 최적화된 분할점을 결정하지 못한다. 이는 결국 전체적인 조인 성능을 저하시키는 주원인이 된다. 하지만 (그림 8)의 결과에서 확인할 수 있는 바와 같이 모호한 경계를 포함하는 경우에도 FMIP는 항상 균등 분할이나 언더플로우 분할보다 분할에 기반한 시간 조인에 적용 시 우수한 성능 향상을 위한 최적 분할점을 제공하고 있다.



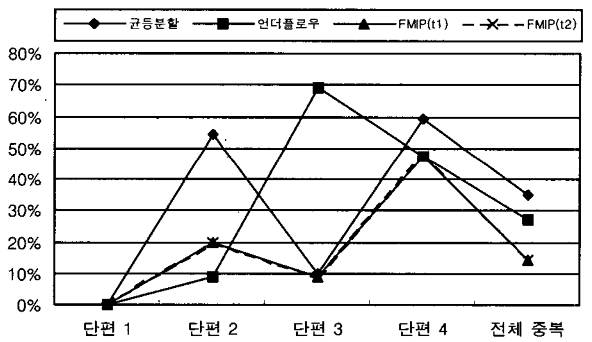
(그림 8) 단편의 조인 성공률

다음의 (그림 9)는 각 단편내에서 발생하는 조인 실패율을 비교한 것이다. 조인 실패율은 분할된 간격이 조인 상대인 간격과 조인을 수행하였으나 타임스탬프가 교차되지 않는 비율을 의미하며 조인 실패율이 적을수록 분할은 최적화가 된다[11]. (그림 9)에서 알 수 있는 바와 같이 조인 실패율 역시 FMIP가 균등 분할이나 언더플로우 분할보다 실패율이 적어 성능이 우수한 분할 기법임을 알 수 있다.



(그림 9) 단편의 조인 실패율

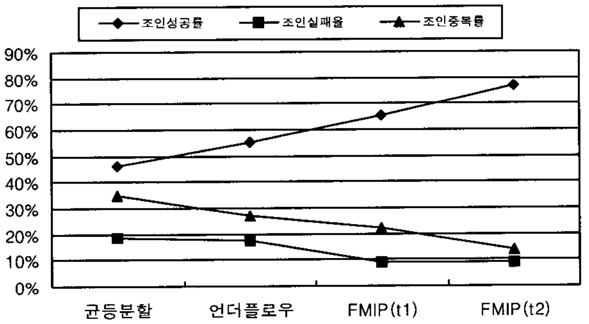
조인 중복율은 간격들이 단편내에 중복되어 조인 연산시 여러번 조인에 참여하는 비율을 의미한다. 따라서 조인 중복이 높다는 것은 불필요한 조인 수행을 초래하여 연산 효율성이 저하되며 전체적으로 조인 비용을 증가시키는 원인이 된다. 따라서 조인 중복율이 낮은 분할 기법이 보다 우수한 분할 기법이라 할 수 있다. 다음의 (그림 10)은 균등분할, 언더플로우 분할 및 FMIP가 각 단편내에서 발생되는 조인 중복율을 나타낸 것으로 단편들내에서 중복율 변동폭이나 실제적인 중복율 수치 모두 FMIP가 균등 분할 및 언더플로우 분할보다 더 낮게 나타나고 있다.



(그림 10) 단편의 조인 중복율

이상의 분석 과정을 통해 각 단편내에서 나타나는 조인 성공율, 조인 실패율, 조인 중복율에 대해 균등 분할, 언더플로우 분할 그리고 FMIP 사이의 수행 상태를 비교 분석하였다. 분석의 결과 모든 경우에도 FMIP가 균등 분할 및 언더플로우 분할보다 시간 조인 연산에 적용 시 성능향상에 기여할 수 있음을 확인하였으며, 특히 각 단편들 사이의 변동폭이 균등 분할 또는 언더플로우 분할보다 작아 기존 분할 기법보다 최적의 분할을 수행하였음을 알 수 있었다. 아울러 FMIP의 경우 경계점이 모호한 상황을 해결할 수 있는 적절한 분할 수단을 포함하므로 기존의 분할 기법에서 해결할 수 없었던 문제들을 처리할 수 있는 특성을 갖고 있다.

한편 앞의 (그림 8), (그림 9) 및 (그림 10)의 분석은 각 분할 기법들에 의해 수행한 분할의 결과로 나타나는 각 단편 내의 특성을 비교 분석한 것으로 이들 분할 기법들이 개별 단편내의 성능 특성을 분석한 것이라면 다음의 (그림 11)은 각 단편의 성능 특성을 종합한 비교 분석이다. (그림 11)에서 볼 수 있는 바와 같이 전체 단편에서의 성능 특성 역시 FMIP가 우수함을 확인할 수 있다.



(그림 11) 각 분할 방법간 성능 비교

이 장에서는 균등 분할, 언더플로우 분할 및 FMIP 사이의 성능 특성을 비교 분석하였다. 이 비교 분석과정에서 확인할 수 있는 바와 같이 균등 분할 및 언더플로우 분할은 MIP기법이나 FMIP에 비해 전체적으로 성능이 저하됨을 확인할 수 있었다. 하지만 간격의 시작시점이 모호한 경우에 MIP 기법으로는 이에 대한 처리 방법이 불가능하므로 조인의 연산 비용을 줄이는 분할을 수행하지 못한다. 반면 FMIP의 경우 간격의 시작시점이 모호한 경우에도 신뢰성 있는 분할을 결정하므로서 전체적인 간격 분할을 최적화시켜 조인 성능 향상 요인을 제공할 수 있었다. 따라서 불확실한 시간 간격을 포함하는 시간 간격의 조인 연산시 FMIP를 적용하여 간격 분할을 수행할 경우 최적화된 분할 기법을 통해 전반적인 시스템의 수행 성능을 향상시킬 수 있다.

6. 결 론

시간 데이타베이스는 관계형 데이타베이스와는 달리 타임스탬프 속성 값으로 인하여 조인 연산시 간격 분할을 위한 최적의 전략과 효율적인 시간 인덱스의 사용을 필요로 한다. 특히 간격분할에 기반한 시간 조인시 최적의 분할점 결정은 시간 인덱스와 함께 조인 성능 향상이 중요한 요소이다. 이 논문에서는 동적인 간격분할 기법인 MIP기법을 퍼지 간격 데이터에 대하여 확장하여 시작시점이 모호한 상태에서도 가능성 분포를 이용함으로써 최적의 분할점이 결정할 수 있는 FMIP기법을 제안하였다. 기존 간격의 시작시점이 퍼지 시간 간격으로 확장되어 시작시점이 애매한 경우 α-절단을 이용함으로써 간격 분할점을 결정할 수 있음을 증명하였다. 제안된 FMIP 분할 기법은 MIP 알고리즘을 확장한 분할 기법으로써, FMIP의 실험을 통해 여러 분할 기법과의 연구 논점을 비교 평가하여 시간 조인 측면에서 수행 성능의 효율성을 제공할 수 있음을 입증하였다. 제안된 FMIP 분할 기법은 조인 연산 단계 이전에 간격 분할의 최소 분할점을 결정하는 이론적인 배경이 되며, 아울러 기존에 사용된 분할 기법으로는 처리가 불가능한 불확실한 간격의 시작시점에 간격 분할을 접근 시도하였다는 것에 의미를 부여할 수 있다. 향후 연구에서는 FMIP를 기반으로 시간 인덱스와 동기화 또는 비동기화된 연산 수행 시 시간 조인의 성능향상을 위한 전략을 제공할 필요가 있다.

참 고 문 헌

[1] C. Breiteneder, S. Gibbs, and D. Tsichritzis, "Modelling of Audio/Video Data," In Proc.of the Int. Conf. on the Entity-Relationship Approach, Karlsruhe, Germany, pp.322-339, Oct., 1992.

[2] E. Ardizzone, M. L. Cascia, "Automatic Video Database Indexing and Retrieval," Multimedia Tools and Applications, Vol.4, No.1, pp.29-56, Jan., 1997.

[3] N. Lorentzos and Y. Mitsopoulos, "SQL Extension for Interval Data," Technical Rep. 105, Informatics Lab., Agricultural University of Athens, 1994.

[4] R. Snodgrass, "The temporal query language TQuel," ACM Transactions on Database System, Vol.12, No.2, pp.247-298, 1987.

[5] R. Snodgrass, "The TSQL2 Temporal Query Language," Kluwer, Sept., 1995.

[6] T. Leung and R. Muntz, "Temporal Query Processing and Optimization in Multiprocessor Database Machines," In Proc.18th Int. Conf. onVLDB, Vancounver, Canada, pp. 383-394, Aug., 1992.

[7] D. DeWitt and Gerber, "Multiprocessor hash-based join algorithms," In Proc. of the Int. Conf. on Very Large Data Base(VLDB), Stockholm, Sweden, pp.151-164, 1985.

[8] C. Ravishankar, "Spatial Hash-Joins," In Proceedings ACM SIGMOD Conference on Management of Data, Montreal, Canada, pp.247-258, 1996.

[9] X. Jeffrey, Yu and Kian-Lee Tan, "Scheduling Issues in Partitioned Temporal join," Joint Computer Science Technical Report Series, May, 1995.

[10] H. Gunadhi and A. Segev, "Query Processing Algorithms for Temporal Intersection Joins," In Proc. of the IEEE Int. Conf. on Data Engineering, pp.336-344, 1991.

[11] H. Lu, B.-C. Ooi, and K.-L. Tan, "On Spatially Partitioned Temporal Join," In Proc. of the Int. Conf. on Very Large Data Bases(VLDB), Santiago de Chile, pp.63-113, Sept., 1994.

[12] M. Soo, R. Snodgrass, and C. Jensen, "Efficient Evaluation of the Valid-Time Natural Join," In Proc. of the 10th Int. Conf. on Data Engineering, Houston, 1994.

[13] H. Zhou, "Two-stage m-way graph partitioning," Parallel Computing, Vol.19, No.12, pp.1359-1373, Dec., 1993.

[14] 이광규, 신예호, 류근호, 김홍기, "시간 데이터베이스에서 시간 간격 분할 알고리즘의 구현 및 평가", 한국정보과학회논문

지(C), 제8권 제1호, 2002.
[15] 김동호, 이인홍, 류근호, "주기억장치에서 시간지원 데이타베이스의 집계함수 설계 및 구현", 한국정보과학회논문지, 제21권 제8호, pp.1405-1415, 1994.
[16] P. Mishra and M. Eich, "Join Processing in Relational Databases," ACM Computing Surveys, pp.63-113, Mar., 1992.
[17] R. Elmasri, G. Wuu, and V. Kouramajian, "The Time Index and the Monotonic B+-tree," In Tansel and et al. [10], chapter 18, pp.433-456.
[18] H. Gunadhi and A. Segev, "Efficient Indexing Methods for Temporal Relations," IEEE Transactions on Knowledge and Data Engineering, Vol.5, No.3, pp.496-509, June, 1993.
[19] C. E. Dyreson and R. T. Snodgrass, "Temporal Indeterminancy," TR 91-30, The Univ. of Arizona, Dec., 1991.
[20] "Fuzzy sets as a basis for a theory of possibility," Fuzzy Sets System., Vol.1, pp.3-28, 1978.
[21] D. Dubois D. and H. Prade, "Processing Fuzzy Temporal Knowledge," IEEE Trans. Systems, Man, and Cybernetica, Vol.19, No.4, pp.729-744, July/August, 1989.
[22] M. Nowakowska, "A new theory of time : generation of time from fuzzy temporal relations," BUSEFAL(L.S.I., Univ. P. Sabatier, Toulouse, France), No.6, pp.87-94, 1981.
[23] S. Dutta, "An event-based fuzzy temporal logic," in Proc. 18th IEEE Int., Symp. Multiple-Valued Logic, Palma de Mallorca, Spain, pp.64-71, 1988.
[24] "Fuzzy Information and fuzzy time," in Proc. IFAC Symp. Fuzzy Information, Knowledge Representation and Decision Analysis, Marseille, France, pp.159-162, 1983.
[25] T. C Fall, "Evidential reasoning with temporal aspects," In Proc. of the AAAI Nat. Conf. Artificial Intelligence, Philadelphia, PA, pp.891-895, 1986.
[26] W. Kurutach, J. Franklin, "On Temporal-fuzziness in Temporal Fuzzy Databases," DEXA, pp.154-165, 1993.
[27] D. Son and R. Elmasri, "Efficient Temporal Join Processing Using Time Index," In Proc. of Conf. on Scientific and Statistical Database Management, pp.252-261, 1996.
[28] D. Zhang, V. J. Tsotras, B. Seeger, "Efficient Temporal Join Processing using Indices," In Proc. of ICDE'02, San Jose, CA, Feb.-March, 2002.
[29] D. Pfoster and N. Tryfona, "Capturing Fuzziness and Uncertainty of Spatiotemporal Object Objects," Time Center Technical Report, TR-59, 2001.

허 문 행

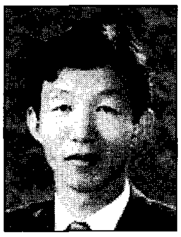
e-mail : moonh@software.or.kr
1979년 숭실대학교 전산학과(학사)
1989년 연세대학교 공학대학원 전산전공 (석사)
1980년~2000년 한국통신 책임연구원
2001년~현재 한국소프트웨어진흥원 단장
관심분야 : 멀티미디어검색, 정보통신

이 광 규

e-mail : kklee@shc.ac.kr
1985년 동국대학교 수학과(학사)
1991년 동국대학교 대학원 응용수학(이학 석사)
1995년~1998년 충북대학교 대학원 전자계산학과 박사과정 수료
1996년~현재 신흥대학 컴퓨터정보계열 조교수
관심분야 : 시간 데이터베이스, 퍼지 이론, 정보통신

이 준 욱

e-mail : junux@dblab.chungbuk.ac.kr
1997년 충북대 컴퓨터과학과 졸업
1999년 충북대 대학원 전자계산학과(이학 석사)
1998년~1999년 한국전자통신 연구원 위촉 연구원 근무
2002년 현재 충북대학교 대학원 컴퓨터과학과 박사과정
관심분야 : 시간 데이터베이스, 시공간 데이터베이스, CRM, 시간 데이터 마이닝, 시공간 데이터 마이닝



류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr
1976년 숭실대 전산과 졸업
1980년 연세대학교 공학대학원 전산전공 (공학석사)
1988년 연세대학교 대학원 전산전공(공학 박사)
1976년~1986년 육군군수지원사전산실(ROTC장교), 한국전자통신 연구소(연구원), 한국방송통신대, 전산학과(조교수) 근무
1989년~1991년 Univ. of Arizona 연구원(TempIS Project)
1986년~현재 충북대학교 전기전자 및 컴퓨터공학부 교수
관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 객체 및 지식베이스 시스템, 지식기반 정보검색시스템, 데이터 마이닝, 데이터베이스 보안 및 Bio-Informatics

김 흥 기

e-mail : hgkim@cbucc.chungbuk.ac.kr
1961년 연세대학교 수학과(학사)
1975년 연세대학교 교육대학원 응용수학 교육학과(교육학 석사)
1985년 중앙대학교 대학원 응용수학(이학 박사)
1980년~현재 충북대학교 전기전자 컴퓨터 공학부 교수
관심분야 : 퍼지 이론, 정보통신