

# CMF-based Priority Processing Method for Multi-dimensional Data Skyline Query Processing in Sensor Networks

Jin-Whan Kim<sup>†</sup> · Kwang-Mo Lee<sup>\*\*</sup>

## ABSTRACT

It has been studied to support data having multiple properties, called Skyline Query. The skyline query is not exploring data having all properties but only meaningful data, when we retrieve informations in large data base. The skyline query can be used to provide some information about various environments and situations in sensor network. However, the legacy skyline query has a problem that increases the number of comparisons as the number of sensors are increasing in multi-dimensional data. Also important values are often omitted. Therefore, we propose a new method to reduce the complexity of comparison where the large number of sensors are placed. To reduce the complexity, we transfer a CMF(Category Based Member Function) which can identify preference of specific data when interest query from sync-node is transferred to sub-node. To show the validity of our method, we analyzed the performance by simulations. As a result, it showed that the time complexity was reduced when we retrieved information in multiple sensing data and omitted values are detected by great dominance Skyline.

**Keywords :** Skyline Query, Multi-dimensional Data, Preference Data, Sensor Network

## 1. 서 론

센서 네트워크는 기지국과 센서 노드들로 구성된다. 기지국은 사용자 질의를 입력 받아 이를 모든 센서 노드에 전달하고 센서 노드들은 수집한 데이터에서 질의 조건에 만족하는 데이터들을 판별하고 무선 통신 방식을 이용하여 기지국으로 전달하는 방식으로 상호 협력적인 네트워크 구성을 통하여 다양한 환경 및 상황 정보를 수집하여 사용자에게 제공한다[1,2]. 이와 같은 센서 네트워크를 구성하는 센서의 하드웨어적 성능이 향상됨에 따라 하나의 장치에서 센싱할 수 있는 데이터의 종류가 다양화되었고 센서에서 측정되는 데이터의 차원은 증가 된다. 따라서 센서 네트워크 환경에서 상황 인지 및 분석을 위해 다중 속성을 갖는 데이터에 대한 처리 기법에 대한 연구가 다양하게 진행되어 왔으며 다차원의 측정값으로 부터 의사 결정을 내려야 할 경우에 스카이라인 질의를 유용하게 활용할 수 있다고 보고되었다[4-6].

스카이라인 질의는 대량의 다차원 데이터 집합에서 모든 값들의 탐색하지 않고 각 속성에 대해 의미 있는 값을 탐색하는 것으로써, 모든 속성이 모든 값들을 지배하거나 적어도 한 가지 이상의 속성이 다른 값을 지배하는 포인트들의 집합을 구한다[3]. 이러한 스카이라인 질의는 센서 네트워크 환경에서의 환경 감시 응용에서도 활용될 수 있다. 예를 들면 한 지역의 온도와 습도 정보를 감지한다고 하면, 온도가

높을수록 그리고 습도가 낮을수록 화재 발생 확률이 높기 때문에 온도와 습도의 속성 값에 대한 최댓값 또는 최솟값인 스카이라인이 유용하게 활용될 수 있다[1,4].

기존의 스카이라인 탐색 방식은 중앙 집중 방식과 분산 환경 방식으로 나눌 수 있다. 중앙 집중 방식의 스카이라인 탐색을 다차원 센서 네트워크에 적용할 경우 스카이라인이 선출되기 전 모든 포인트들과 비교 검색을 수행해야 하므로 선형적으로 증가하는 계산상의 어려움이 발생한다. 그리고 분산 방식에서는 In-Network 질의 처리 및 필터링 기법을 적용하여 데이터 전송량을 감소시켰으나 센서의 연속적인 질의 형식에서는 필터의 선정 및 배포, 비교검색에서 비교횟수가 증가한다. 또한 필터링의 적용이 크지 않을 경우 차원의 증가로 발생할 수 있는 비교횟수의 증가에 따른 센서의 과부하와 지연은 고려되지 않았다. 또한 스카이라인 처리에서 지배력이 큰 포인트에 의해 상대적으로 낮은 값은 제외 된다. 이때 단일 값으로도 의미 있는 포인트도 함께 제외 되어 특정 속성이 우수한 값을 검색하지 못한다.

따라서 본 논문에서는 싱크 노드로 부터 관심질의(interest query)를 노드로 전송할 때 질의의 주요 속성들의 선호도(preference)를 판별할 수 있는 카테고리 기반 소속함수(이하 CMF : Category Based Member Function)를 함께 전송하여 스카이라인 선출 시 차원 및 센서의 수 증가에 따라 발생하는 비교 계산의 복잡성을 감소시키고 의미 있는 값을 효율적으로 검출하기 위해 우선순위 데이터처리 기법을 제안한다.

본 논문의 구성은 2장에서 기존의 스카이라인 질의 처리 기법과 연구 동기를 기술하고 3장에서 제안된 스카이라인 질의 처리 기법을 설명한다. 4장에서는 실험을 통하여 제안된 기법을 평가하며 5장에서 결론을 기술하였다.

<sup>†</sup> 준 회 원 : 한림대학교 컴퓨터공학과 박사과정  
<sup>\*\*</sup> 중 신 회 원 : 한림대학교 정보전자공과대학 컴퓨터공학과 교수  
논문접수 : 2012년 5월 19일  
수 정 일 : 1차 2012년 7월 9일, 2차 2012년 7월 27일, 3차 2012년 8월 7일  
심사완료 : 2012년 8월 8일  
\* Corresponding Author : Kwang-Mo Lee(kmlee@hallym.ac.kr)

## 2. 관련 연구

스카이라인 질의는 다중 속성을 갖는 데이터 집합에서 다른 부분집합에 지배(dominate) 당하지 않는 부분집합을 찾는 것이다. 스카이라인을 검출하는 가장 기본적인 방법은 각각의 값들의 지배관계를 정의하는 것이다. 데이터 내의 각 값  $r$ 을 다른 값들과 비교하여  $r$ 을 지배하는 값이 없다면  $r$ 을 스카이라인으로, 그렇지 않다면  $r$ 을 지배하는  $r'$ 가 스카이라인으로 선출된다.

예를 들어 Fig. 1과 같이 “질의 : 숙박비용이 저렴하고 해변에서 가까운 호텔을 검색하라.” 라는 스카이라인 질의를 수행한다면  $x$  축은 해변에서의 거리  $y$ 축은 숙박비용으로 좌표를 표현할 수 있다.

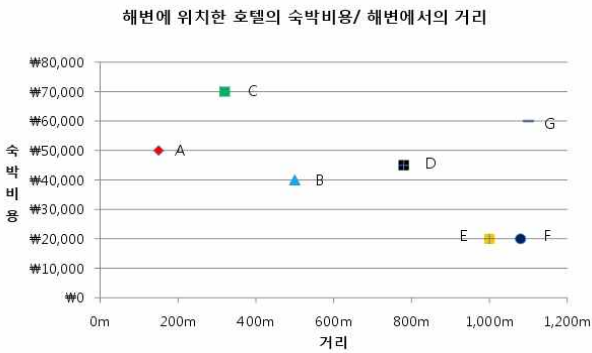


Fig. 1. Skyline of hotels located on the beach

호텔 A는 C, G에 비해 숙박비용이 낮고 해변으로부터 거리가 가깝다. 이때 A가 C, G를 지배 한다고 한다. B의 값은 D, G보다 숙박비용이 낮고 해변으로부터 거리가 가깝다. 그

리고 E와 F는 숙박비용은 같으나 F가 해변에서 더 멀다. 따라서 “질의”의 결과는 다른 어떤 값들로부터 지배되지 않는 A, B, E가 선호하는 데이터로 결정된다. 이와 같이 스카이라인 질의는 산재되어 있는 정보 중 선호하는 정보만을 제공하여 의사결정에 도움을 준다.

Fig. 2는 센서 네트워크에서 수집되는 데이터를 기준으로 데이터의 수집과 비교를 나타내었다. 센서 네트워크에서의 스카이라인 질의는 Fig. 1과 같이 숙박비용과 해변에서 호텔까지의 거리를 고려한 호텔을 찾는 일차원적인 스카이라인 질의와 달리 시간에 종속적인 데이터의 연속성 및 다차원 데이터라는 차이점을 갖는다. 즉, 센서 데이터는 시간에 따라 연속적으로 발생된다. 따라서 검색하고자 하는 차원(속성), 노드의 수와 데이터의 연속성에 의해 많은 양의 데이터가 발생하게 되고 스카이라인 선출시 더욱 많은 처리가 필요하다.

대표적인 중앙 집중 방식은 Block Nest Loop(BNL)[7], Divide and Conquer(D&C)[7], Nearest Neighbor Search (NN)[3], Branchand Bounded Search(BBS)[8] 등이 있고 분산 방식은 웹 환경의 스카이라인 알고리즘[9], 스트림 데이터처리를 위한 알고리즘[10]과 센서 네트워크 환경에서의 MFTAC기법(Min score Filtering Tuple applied Aggregation method based on Cluster)[4] 등이 있다.

중앙 집중 방식의 스카이라인 탐색 알고리즘 중 BNL (Block Nest Loop)은 메모리상에 스카이라인 후보 값들의 리스트를 저장하고 그 외의 값들과 모두 비교한다[7]. 처리 방식을 나열하면 (i) 값  $v$ 가 리스트 값에 의해 지배당한다면  $v$ 는 스카이라인에서 배제된다. (ii)  $v$ 가 리스트 값을 지배한다면  $v$ 를 리스트에 삽입하고  $v$ 에 의해 지배되는 값은 삭제한다. (iii) 만약  $v$ 가 리스트 내의 값에 지배하지도, 당하지도 않는다면  $v$ 는 스카이라인이 될 수 있어 리스트에 삽입된

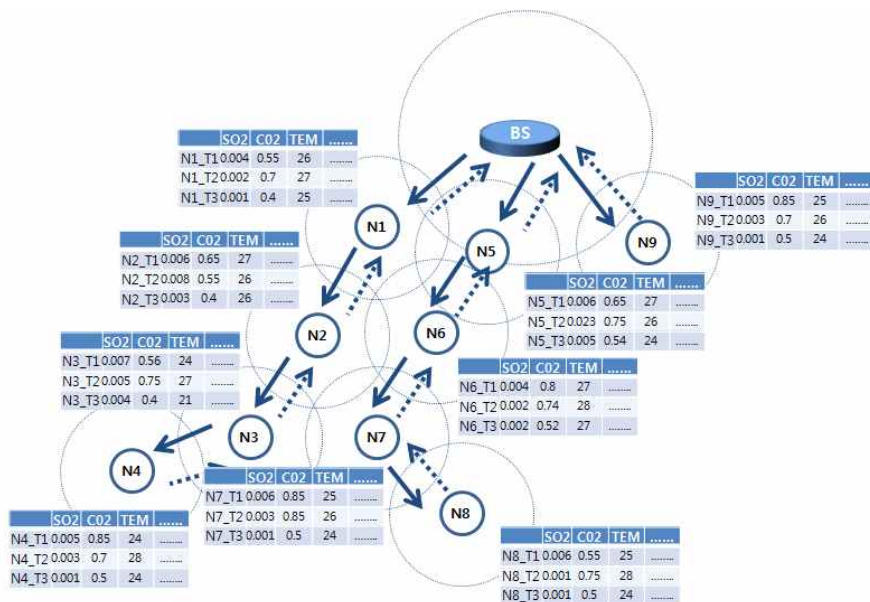


Fig. 2. Continuity of data in sensor networks

다. 그러나 BNL은 다수의 값을 비교할 경우 제한적인 리스트의 크기로 메모리 오버플로우가 발생할 수 있으며, 모든 값들을 일대일 비교를 하여야 함으로 계산상의 어려움이 발생한다. 따라서 다차원의 데이터 집합과 제약적인 센서 네트워크 환경에서는 적합하지 않다.

D&C(Divide and Conquer) 알고리즘은 전체의 데이터 집합을 메모리에 알맞게 다수의 영역으로 나누고 각 영역별 스카이라인을 계산한다[7]. 비교적 크기가 작은 데이터 집합에는 용이하나 파티션 분할이 수행될 때, 전체 데이터 집합을 적어도 한번 이상 로드 되어야 하기 때문에 많은 양의 스트리밍 데이터 환경에서는 적합하지 않다.

Wolf-Tilo Balke et al.[9]와 같은 웹 기반의 분산 방식에서는 웹 정보 시스템을 대상으로 하고 있기 때문에 무선 센서 네트워크에는 부적합하며, Xuemin Lin et al.[10] 또한 센서 네트워크 제한적인 환경의 특성을 고려하지 않았다.

Zhiyong Huang et al.[11]은 MANET(Mobile Adhoc Network)환경의 경량화된 장치를 위한 분산 스카이라인 질의 처리 기법이다. 질의 결과에서 불필요한 데이터를 제거하기 위해 필터링 기법을 활용한다. 그러나 필터의 생성을 위한 정보를 상위 노드에서 결정하기 때문에 하위 노드들의 정보는 반영하지 못한다. [11]의 방법 보다 센서 네트워크 환경에서의 MFTAC[4]는 효과적인 필터링을 제공한다. 하위 센서로부터 초기 데이터를 수집한 싱크 노드는 MFT(Min-Socre Filtering Tuple)를 결정하여 하위 센서로 전송한다. 하위 센서들은 수신 받은 MFT를 이용하여 데이터를 필터링하여 싱크 노드로 전송한다. 따라서 불필요한 데이터 전송을 줄일 수 있다. 또한 클러스터링을 통한 MFT 관리로 모든 센서에서 필터 관리를 하지 않아 효율적이다. 그러나 클러스터링에 의해 필터의 생성과 배포가 줄어들고, 연속적인 스카이라인 질의처리를 하기 위해서는 매 주기마다 MFT를 계산하여 클러스터 내의 센서들로 배포하여야 하며, 상대적으로 컴퓨팅 파워가 낮은 센서 네트워크 환경에서 빈번한 클러스터내의 정보 갱신은 클러스터 헤더와 같은 센서의 수명을 단축시켜 센서 네트워크 전체에 효율성을 감소시킬 수 있다. 또한 차원과 센서의 수가 증가하면 MFT필터 계산 및 배포의 영향으로 비교횟수의 증가를 가져온다.

따라서 기존의 스카이라인 선출에서는 데이터 필터링을 통한 전송 데이터를 줄이는 목적의 연구가 진행 되었으나 차원과 센서의 수 증가에 따른 비교횟수의 증가는 고려되지 않았다. 따라서 센서 네트워크 환경에서 다차원 데이터를 처리할 때 비교횟수를 감소하기 위한 방법을 제안한다.

### 3. CMF를 기반으로 한 스카이라인 질의 처리 기법

센서 네트워크에서 기지국은 관심 질의를 센서 노드에게 전달하고 센서 노드들은 질의에 부합하는 데이터를 기지국으로 전달하여 전송된 모든 데이터는 기지국에서 관리된다. 센서에서 전달되는 데이터는 연속적이며 불필요한 다수의 데이터가 포함되어 있다. 따라서 다수의 데이터에서 관심

값을 선출하기 위해 스카이라인 질의를 사용할 수 있다.

그러나 스카이라인 질의에서는 다수의 노드와 다차원의 데이터를 가질 경우 스카이라인 선출을 위해 많은 계산 복잡도를 갖는다. 따라서 이와 같은 다차원 데이터에서 발생하는 비교 계산의 복잡성을 감소시키기 위해 차원에 따른 모든 센서 데이터의 값들을 비교 평가하는 것이 아닌 기지국에서 관심 질의를 전송할 때 CMF를 전송하여 계산된 함수의 소속 값으로 관심 포인트를 선출하는 기법을 제안한다. 그리고 지배력이 큰 스카이라인 포인트로부터 제외되는 의미 있는 속성 값을 처리하기 위해 우선순위 값 처리기법을 제안하여 선출되는 스카이라인의 신뢰성을 향상시킨다.

#### 3.1 CMF의 정의

카테고리 기반 소속 함수(CMF : Category Based Member Function)는 센서로부터 측정되는 값을 소속 값 0부터 1까지의 카테고리(범주)로 나타낸다. 이와 같이 “0부터 1까지” 카테고리(범주)를 소속 값으로 반환하는 함수를 CMF라 정의한다.

센서 네트워크의 기지국에서는 측정되는 값에 대한 종류와 범위를 알고 있음을 가정한다. 따라서 기지국은 관심 질의에 관한 CMF를 미리 정의하여 활용할 수 있다.

Table 1. Type of sensor / range of values (ex.)

Sensor 종류	측정범위
Temperature Sensor(온도)	-20℃ ~ 50℃
Humidity Sensor(습도)	0 ~ 100 RH
Wind Speed Sensor(풍속)	0 ~ 150 knot
Wind Direction Sensor(풍향)	16 Bearing
Rainfall Sensor(강우)	Under 5cm/hr
Solar Radiation Sensor(일사)	100mV/Wm <sup>-2</sup>
Barometric Sensor(기압)	900mbar ~ 1146mbar

질의에 종류에 따라 하위 센서로 전달되는 CMF는 선택된다. 예를 들어 1차원 적인 질의로 “기온이 높은 곳은?”이라 질의하면 CMF는 (-20℃ ≤ x ≤ 50℃)범위의 정의역을 갖고 (0 ≤ F(x) ≤ 1)인 치역을 갖게 된다. 따라서 기온의 CMF는  $F(x) = \frac{1}{70}x + \frac{2}{7}$  로 정의 할 수 있다. 그리고 여러 가지 차원을 이용하는 질의로는 “화재 위험이 높은 지역은?” 이면 온도, 습도, 풍속의 값이 이용될 수 있다. 이 질의를 CMF로 나타내면 Table 2와 같다.

CMF는 센서의 데이터 측정 범위 및 데이터의 모형에 따라 선택적으로 사용할 수 있으며 센서의 측정값을 정의역으로 하여 치역은 0부터 1까지의 실수에 바운드시켜 사용한다.

#### 3.2 스카이라인 처리

센서 네트워크를 구성하는 다수의 센서와 데이터 종류의 다양화로 데이터 차원은 증가 한다. 이와 같은 센서 네트워크

Table 2. Range of CMF values

질의	센서	함수 그래프	범위	CMF
“화재 위험이 높은 지역은?”	온도		$0 \leq F(x) \leq 1$	$T(x) = \frac{1}{70}x + \frac{2}{7}$
	-20 ~ 50 (°C)			
	습도		$1 - (0 \leq F(x) \leq 1)$	$H(x) = -\frac{1}{100}x + 1$
	0 ~ 100 (RH)			
	풍속		$0 \leq F(x) \leq 1$	$W(x) = \frac{1}{150}x$
	0 ~ 150 (knot)			

Table 3. Example of sensor data tuples

S_NUM	온도	기압	습도	풍속	풍향	조도	일사	월사	강우	CO2	.....
S_01	0.07	1011.5	63	0.9	NE	25	0.00	0.00	0.92	0.0025	.....
S_02	0.22	1010.7	66	0.9	NE	28	0.00	0.00	1.02	0.0026	.....
S_03	0.36	1009.9	70	0.9	NE	30	0.00	0.00	1.12	0.0028	.....
S_04	0.50	1009.2	73	0.9	E	15	0.00	0.00	1.23	0.0029	.....
S_05	0.64	1008.4	77	0.8	E	22	0.00	0.00	1.33	0.0031	.....
S_06	0.78	1007.6	80	0.8	ENE	35	0.00	0.00	1.43	0.0032	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

크에서는 관심 데이터를 수집하기 위해 싱크 노드는 하위 센서에게 질의를 전송한다. 이때 하위 센서로 전달되는 질의에 관심 포인트의 정보를 판별할 수 있는 CMF를 함께 설정하여 전달하게 된다.

그리고 센서에서 발생하는 데이터는 다양한 접근성을 갖는다. 즉 질의에 따라 필요한 데이터가 변경되거나 삭제될 수 있다. 예를 들어 한국태양에너지학회의 표준기상데이터에서는 건구온도(섭씨), 기압(hPa), 운량(%), 적설량(cm), 강수량(mm), 풍향, 상대습도(%), 수평면일사(MJ/m2), 운형, 풍속(m/s), 습구온도(°F), 절대습도(kg/kg), 공기밀도(kg/m2), 엔탈

피(Kcal/kg), 직달일사(mJ/m2) 등으로 되어있다[12]. 질의를 “화재위험이 높은 곳?”으로 데이터를 검색하려 한다면 측정된 데이터 중 화재에 영향을 줄 수 있는 값들만을 활용하는 것이 전체 데이터를 비교 검색하는 것 보다 효율적이며 또한 차원의 증가에 따른 비교횟수의 증가를 줄일 수 있다.

Table 3과 같이 각 센서별 튜플이 정의 된다면, 센서들의 측정값 중 “질의 : 화재위험이 높은 곳?”의 질의 처리를 통하여 관심 포인트를 검출할 수 있다. 이때 기존의 스카이라인 처리에서는 센서의 튜플들을 모두 비교하여 지배관계를 정의하여 관심 포인트를 추출한다. 그러나 “질의”와 같이 화

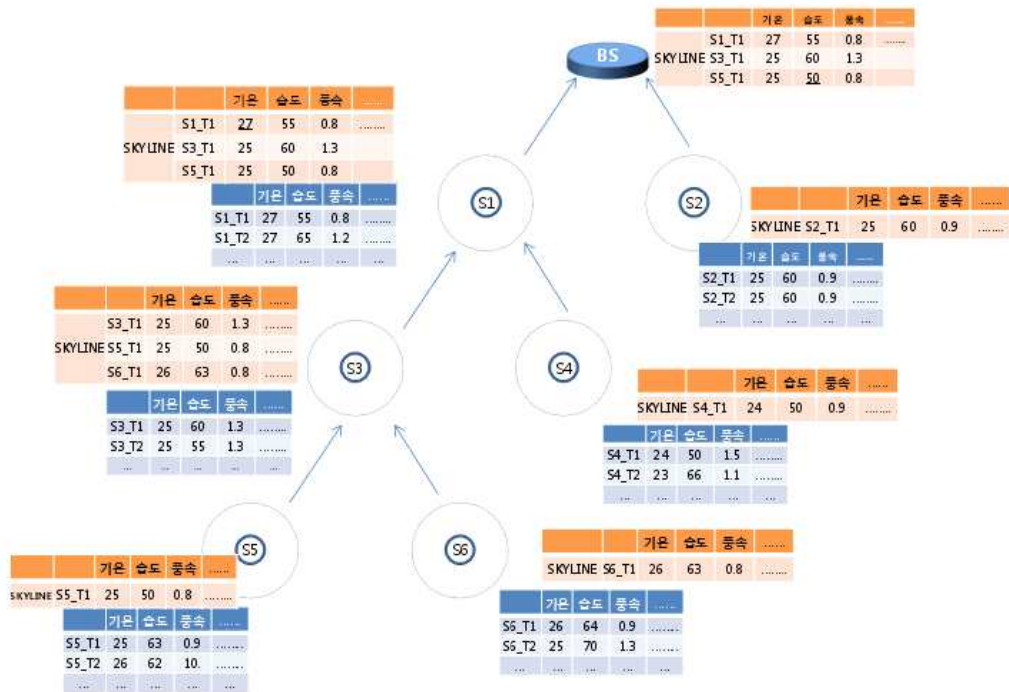


Fig. 3. Update of Skyline Decision list

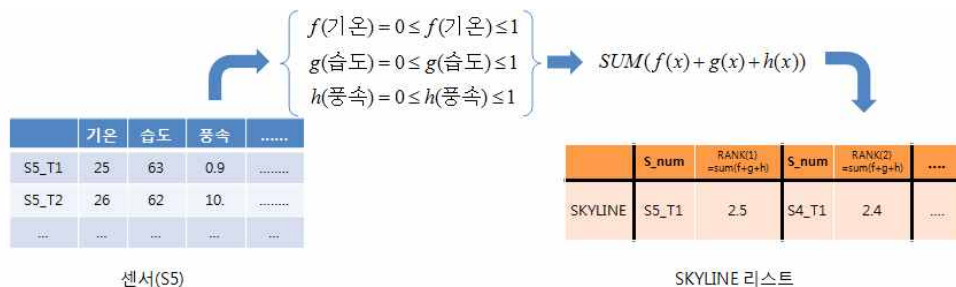


Fig. 4. Member Values of CMF

재 위험에 영향을 주는 값은 온도, 기압, 습도, 풍속이다(화재 기장 : 기온/ 습도/ 풍속/ 기압배치의 값이 특히 화재의 원인이 된다[13]). 따라서 “질의”에서는 센서의 모든 센서 값 중 영향을 주는 특정 값들이 사용되며, 모든 센서의 튜플 모두를 비교 검색하는 것보다 특정 튜플만 비교 검색하여 비교횟수가 감소한다.

Fig. 3과 같이 센서의 측정 데이터에서 스카이라인을 처리할 경우 하위 센서에서부터 데이터를 비교 검색하여 스카이라인 결정 리스트에 정보를 저장하여 상위 센서로 전달한다. 상위 센서에서는 스카이라인 결정 리스트 값을 비교하여 스카이라인을 갱신하여 싱크 노드에서는 찾고자 하는 값에 가장 부합하는 센서 정보를 얻을 수 있다. 그러나 이와 같은 방법에서는 스카이라인 리스트를 모든 센서 데이터와 비교 검색하기 때문에 비교횟수가 증가한다. 따라서 Fig. 4와 같이 센서의 튜플별 기온/습도/풍속의 속성 값을 모든 센서의 값들과 비교하지 않고 센서에서 계산된 CMF 소속 값의 합을

SKYLINE 리스트에 추가하여 전승함으로 스카이라인 선출에서 발생하는 튜플별 속성 값의 비교 횟수를 감소시켰다.

CMF 기반 스카이라인 처리에서는 센서의 데이터를 최저 0부터 최대 1까지의 사이 값으로 변환하고 그 값들의 합을 기준으로 스카이라인을 선정한다. CMF의 소속 값이 클수록 (작을수록 : 측정값에 따라 역수를 취할 경우) 값의 중요도는 커진다. Fig. 4와 같이 “화재 위험이 높은 곳?”의 질의에 대한 처리라면 기온이 높을수록, 습도는 낮을수록, 풍속은 빠를수록 화재 위험이 높아진다. 이때 습도에 대한 CMF의 처리는 의미론적 역수(습도가 낮을수록 1에 가깝게)로 처리한다. 만약 반대의 질의로 “화재 위험이 낮은 곳?”으로 질의한다면 화재에 영향을 주는 기온, 습도, 풍속에 관한 CMF를 앞선 질의와 반대로 재정의 하여 적용할 수 있다.

이와 같이 스카이라인의 모든 튜플을 비교 검색하지 않고 CMF의 소속 값을 이용하기 때문에 모든 튜플들을 비교 검색하는 것 보다 비교횟수가 감소한다.

센서	풍속(cm)	온도(°C)
s_01	50	25
s_09	100	18
s_15	280	8
s_07	60	25
s_08	80	20
s_02	12	28
s_05	30	22
s_04	15	24
s_13	210	10
s_14	150	12
s_17	250	16
s_19	220	14
s_12	200	15
s_20	275	5
s_16	27	10
s_06	50	23
s_03	15	20
s_11	150	20
s_18	210	20
s_10	150	16

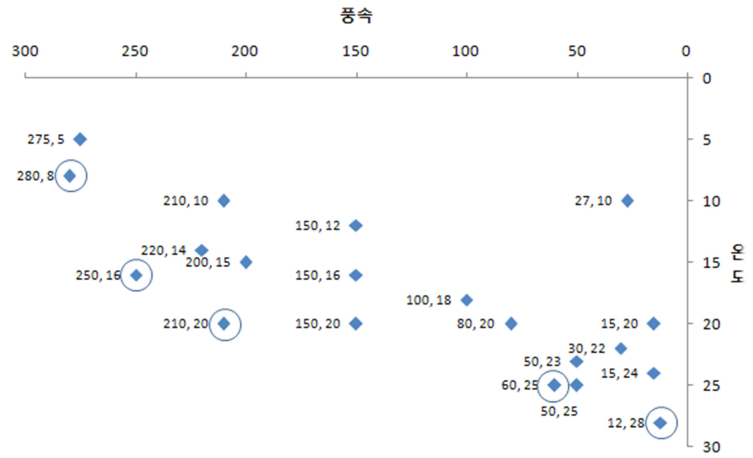


Fig. 5. Skyline detection of existing method

센서	풍속(cm)	온도(°C)	소속값(합)	순위
s_01	50	25	1.00	11
s_09	100	18	0.93	13
s_15	280	8	1.20	3
s_07	60	25	1.03	8
s_08	80	20	0.93	13
s_02	12	28	0.97	12
s_05	30	22	0.83	18
s_04	15	24	0.85	17
s_13	210	10	1.03	9
s_14	150	12	0.90	16
s_17	250	16	1.37	1
s_19	220	14	1.20	3
s_12	200	15	1.17	5
s_20	275	5	1.08	7
s_16	27	10	0.42	20
s_06	50	23	0.93	13
s_03	15	20	0.72	19
s_11	150	20	1.17	5
s_18	210	20	1.37	1
s_10	150	16	1.03	9

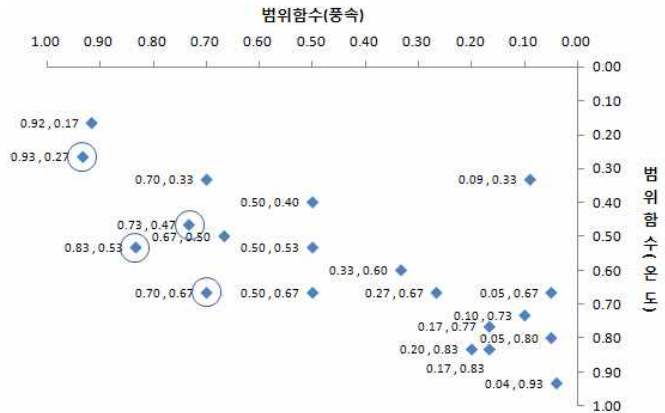


Fig. 6. Skyline detection of CMF

3.3 우선순위 데이터 처리

CMF를 이용하여 스카이라인을 처리할 경우 소속 값의 합에 따라 스카이라인이 결정되어 진다. 그러나 계산된 소속 값의 합의 크기는 관심 있는 속성을 모두 만족시킬 경우에만 높은 값을 얻을 수 있다. 따라서 검색하고자 하는 스카이라인 중 특정 하나의 속성이 우수한 값은 검색하지 못한다. 예를 들어 Fig. 5와 같이 온도와 풍속의 센서 데이터의 경우 “화재위험이 높은 곳?”의 질의에 대한 처리에서는 소속 값의 합이 모두 큰 경우에 스카이라인으로 선택되어진다.

기존의 방법으로 검출되어진 스카이라인(s\_15, s\_07, s\_02, s\_17, s\_18)과 CMF 스카이라인(s\_15, s\_17, s\_19, s\_18)을 비

교하여 보면 Fig. 6에서는 센서 s\_07, s\_02가 제외되고 s\_19가 추가되어있다. 이것은 센서 s\_07, s\_02의 데이터가 특정 속성에서만 높은 값을 갖기 때문에 소속 값의 합으로 검출한 스카이라인에서 제외된 것이다.

또한 Fig. 5의 기존의 스카이라인 검출 방식에서 센서 s\_19의 경우 데이터의 속성이 중요한 의미를 포함하나 s\_17의 지배 여부에 따라 제외된 값이다. 또한 센서 s\_20은 센서 s\_15에 의해 스카이라인으로 선출되지 않았지만 풍속이 매우 빠르다. 그리고 s\_01은 s\_07에 의해 지배되지만 여전히 화재 위험이 높은 데이터에 속한다. 이와 같이 의사 결정을 함에 있어 데이터의 단일 값으로도 중요한 의미를 가질 수 있다. 따라서 센서 데이터의 스카이라인 검색 시 의미 있는



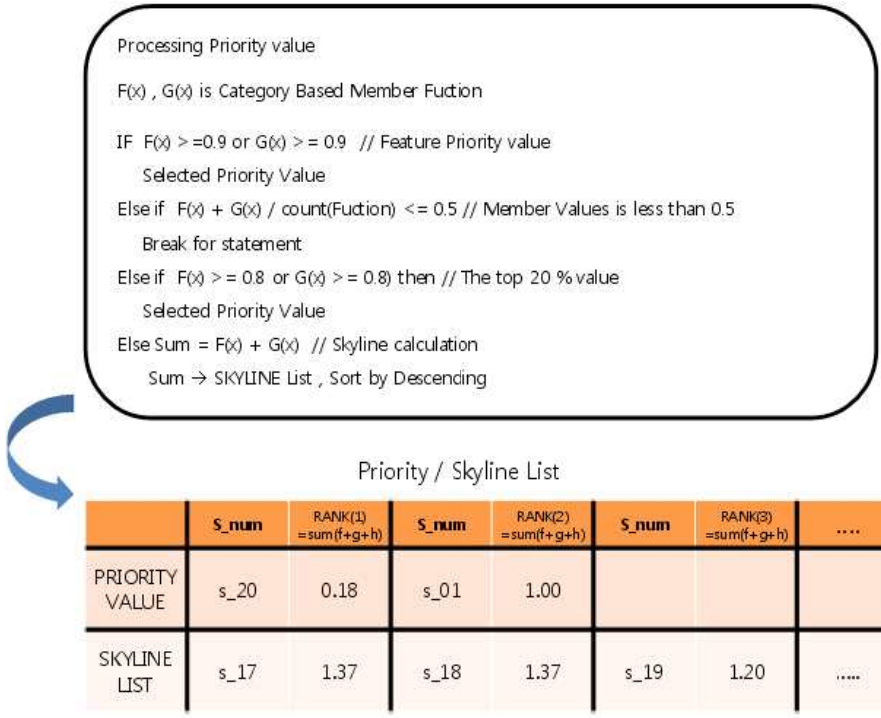


Fig. 7. Priority processing / Skyline list

Table 4. Skyline of CMF-based priority processing method

기존 스카이라인 처리			CMF 기반의 우선순위 데이터 처리 기법				CMF 기반의 스카이라인 처리		
센서	풍속(cm)	온도(℃)	센서	소속값의 평균	풍속 (소속값)	온도 (소속값)	센서	소속값(합)	순위
s_01	50	25	s_01	0.50	0.17	0.83			
s_09	100	18	s_09	0.47	0.33	0.60	s_09	0.93	8
s_15	280	8	s_15	0.60	0.93	0.27			
s_07	60	25	s_07	0.52	0.20	0.83			
s_08	80	20	s_08	0.47	0.27	0.67	s_08	0.93	8
s_02	12	28	s_02	0.49	0.04	0.93			
s_05	30	22	s_05	0.42	0.10	0.73	s_05	0.83	13
s_04	15	24	s_04	0.43	0.05	0.80	s_04	0.85	12
s_13	210	10	s_13	0.52	0.70	0.33	s_13	1.03	5
s_14	150	12	s_14	0.45	0.50	0.40	s_14	0.90	11
s_17	250	16	s_17	0.68	0.83	0.53			
s_19	220	14	s_19	0.60	0.73	0.47	s_19	1.20	2
s_12	200	15	s_12	0.58	0.67	0.50	s_12	1.17	3
s_20	275	5	s_20	0.54	0.92	0.17			
s_16	27	10	s_16	0.21	0.09	0.33	s_16	0.42	15
s_06	50	23	s_06	0.47	0.17	0.77	s_06	0.93	8
s_03	15	20	s_03	0.36	0.05	0.67	s_03	0.72	14
s_11	150	20	s_11	0.58	0.50	0.67	s_11	1.17	3
s_18	210	20	s_18	0.68	0.70	0.67	s_18	1.37	1
s_10	150	16	s_10	0.52	0.50	0.53	s_10	1.03	5
s_15, s_07, s_02, s_17, s_18			s_01, s_15, s_07, s_02, s_17, s_20				s_18, s_19, s_12, s_11, s_10, s_13		

값을 보다 효율적으로 찾기 위해 단일 값으로 중요한 의미를 가질 수 있는 속성 값을 우선 처리 하고 CMF의 소속 값을 기준으로 스카이라인을 선출한다.

싱크 노드의 질의에 의해 하위 센서로부터 수집된 데이터를 싱크 노드로 전달할 때 Fig. 7과 같이 CMF 소속 값을 스카이라인 결정 큐에 높은 순으로 정렬하여 전송한다. 이때 우선순위 값의 정의는 특성 우선순위 값으로 소속값이 0.9 보다 큰 포인트를 우선 선출하고 각 소속 값이 평균 0.5 이상이고 단일 속성의 값 상위 20%에 속하는 값을 우선순위 값으로 처리한다.

우선순위 데이터를 처리한 후 소속 값 합이 상위 값을 이용하여 관심 포인트를 추가 한다. 소속 값의 평균 0.5 이상이며 값의 합이 큰 순으로 추가된다. 값이 클수록 주요 포인트로 선출할 수 있다.

Table 4는 기존의 스카이라인 방식과 우선순위 데이터 처리 기법을 적용한 CMF 기반의 스카이라인을 나타내었다. 기존의 스카이라인 처리를 통하여 검출되는 포인트는 s\_15, s\_07, s\_02, s\_17, s\_18 이다. CMF 기반의 우선순위 데이터 처리를 통하여 검출되는 포인트는 단일 소속 값이 0.9 이상인 특성 우선순위 값 s\_15, s\_02, s\_20과 우선순위 값 s\_01, s\_07이며 우선처리 후 CMF의 소속 값의 합의 순위에 의해 추가로 검출되는 포인트는 s\_19, s\_12, s\_11, s\_18 이다.

#### 4. 실험 및 성능 평가

본 논문에서는 센서 네트워크 환경에서 센서의 수와 차원의 증가로 스카이라인 처리 시 증가하는 비교횟수를 감소시키는 기법과 다수의 차원의 데이터에서 효율적인 관심 값을 검출해 내기 위한 CMF 스카이라인 처리 방법을 제안하였다.

다차원 데이터 처리의 성능 평가를 위해서 센서의 수, 차

원의 수, 선출 스카이라인의 수에 따른 비교횟수를 BNL 방식과 MFTAC 기법과 비교하였으며, BNL과 MFTAC에서의 스카이라인과 CMF 스카이라인을 비교 평가 하였다. 센서의 수는 100부터 500단위로 4500개 까지 이용하였으며, 센서의 값은 0 ~ 1 사이의 랜덤 값으로, 차원은 10차원부터 50차원 까지 생성된 데이터를 이용하여 비교횟수를 측정한다. 비교횟수는 라운드별 스카이라인 선출에 따른 응답 시간을 기준으로 하였다. MFTAC의 경우 클러스터 정책은 반영되지 않았으며, MFT 필터는 라운드별 1회 적용되었다. CMF의 경우 합수 소속 값에 대한 계산은 1 라운드 당 1회 진행하였다.

노드의 수와 차원의 수가 증가함에 따라 각 스카이라인 값의 선출시간은 상대적으로 증가하며 BNL 방식과 MFTAC 방식은 노드와 차원의 증가에 따라 스카이라인 처리 시간이 유사하게 진행되었으며, 제안된 CMF에서는 차원의 모든 값을 비교 검색을 하지 않기 때문에 Fig. 8과 같이 응답시간이 가장 빨랐다. 10 Dim 이하의 경우에는 응답시간이 크게 차이를 나타내지 않고 있다. Fig. 9에서는 시간의 범위를 줄여 내용을 확인한다.

효율적인 관심 데이터 처리를 위해 제안된 우선순위 데이터 처리 기법을 평가하기 위하여 불규칙적인 데이터 범위를 갖는 경우와 정규분포 데이터 범위를 갖는 두 가지 데이터 유형을 이용하였다.

Fig. 10은 우선순위 데이터 처리 기법을 성능 평가하기 위해 사용된 센서의 [라운드 1] 랜덤 값 분포이다. 라운드 1에서는 200개의 센서와 3차원으로 0 ~ 10 까지 데이터 범위를 갖는다. 각 차원별 센서 값의 분포는 불규칙적인 데이터로 구성되어 있다. Fig. 10에서와 같이 불규칙적인 값의 범위를 갖는 데이터를 Table 5와 같이 4개의 라운드로 구성하여 센서의 수를 달리 하여 우선순위 데이터 처리 기법을 통하여 검출되는 포인트들을 기존의 스카이라인 처리를 통하여 검출되는 포인트의 수를 비교한다.

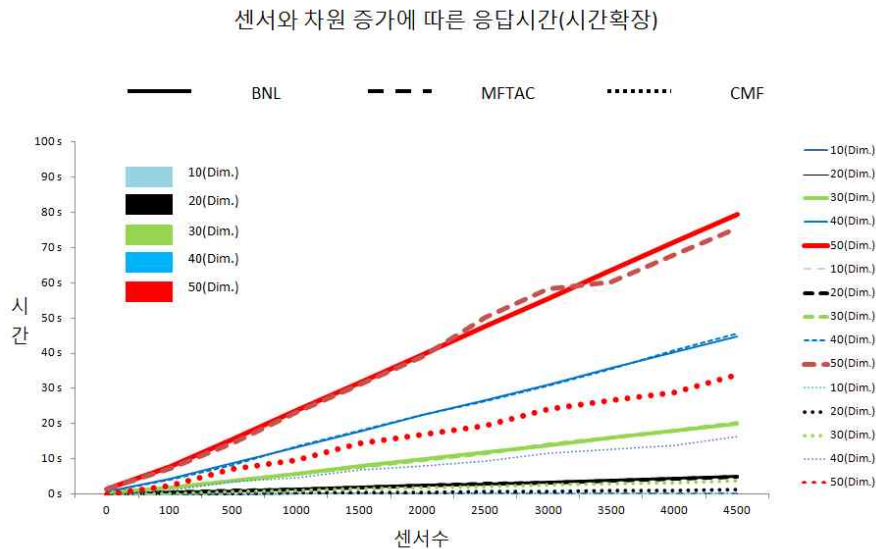


Fig. 8. Response time BNL/ MFTAC/ CMF



센서와 차원 증가에 따른 응답시간

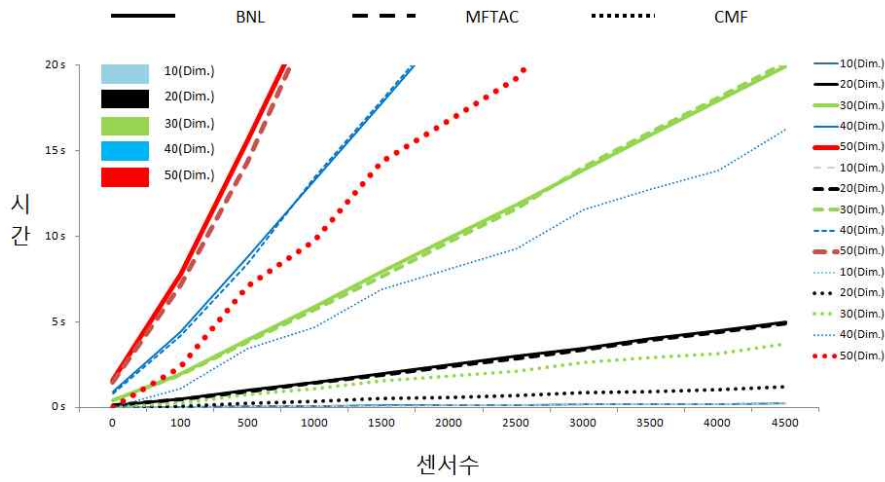


Fig. 9. Response time BNL/ MFTAC/ CMF

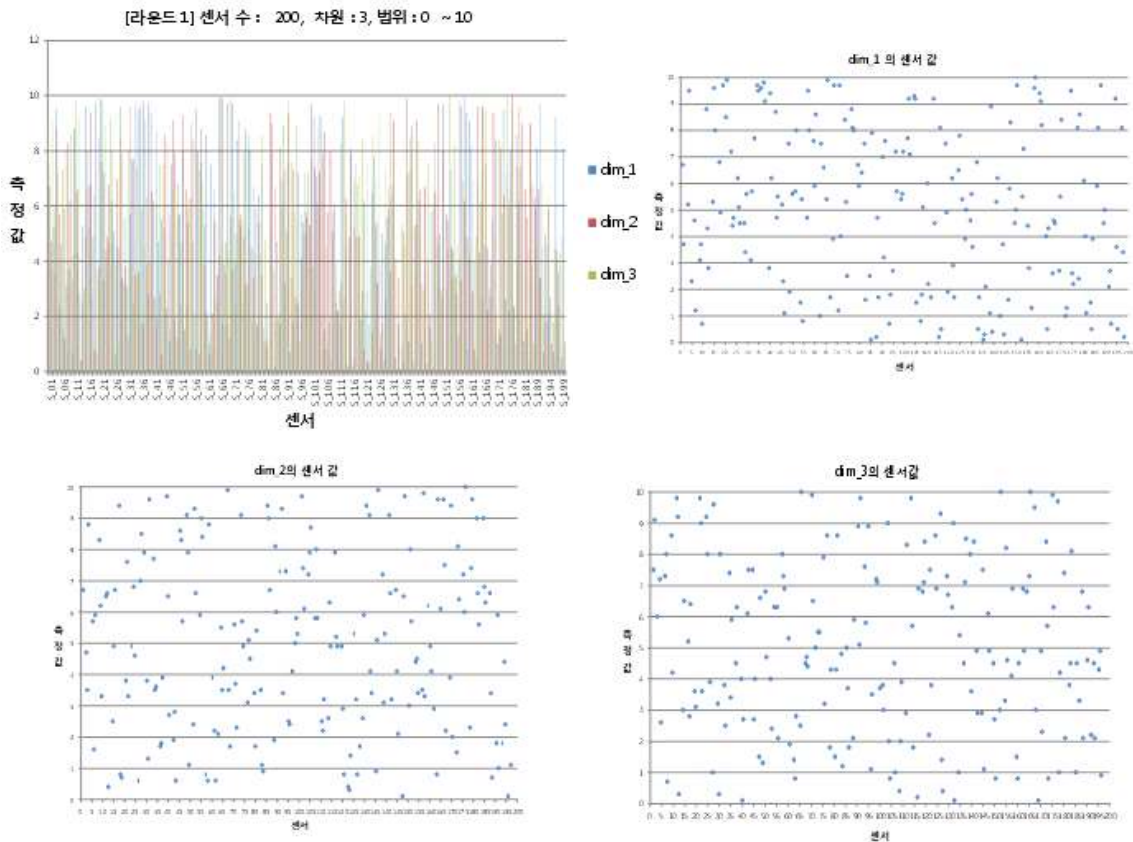


Fig. 10. Random range of values

Table 5. Round information of sensors

라운드	센서 수	차원	범위
R_1	200	3	0~10
R_2	150		
R_3	100		
R_4	50		

Fig. 11은 센서의 데이터 범위가 정규분포 범위를 갖는 데이터 유형이다. 정규분포 유형의 데이터에서도 Table 5와 같이 4개의 라운드를 적용하여 스카이라인 수를 비교하였다.

아래 Fig. 12, 13는 불규칙적인 데이터 범위를 갖는 경우와 정규분포 데이터 범위를 갖는 두 가지 데이터 유형을

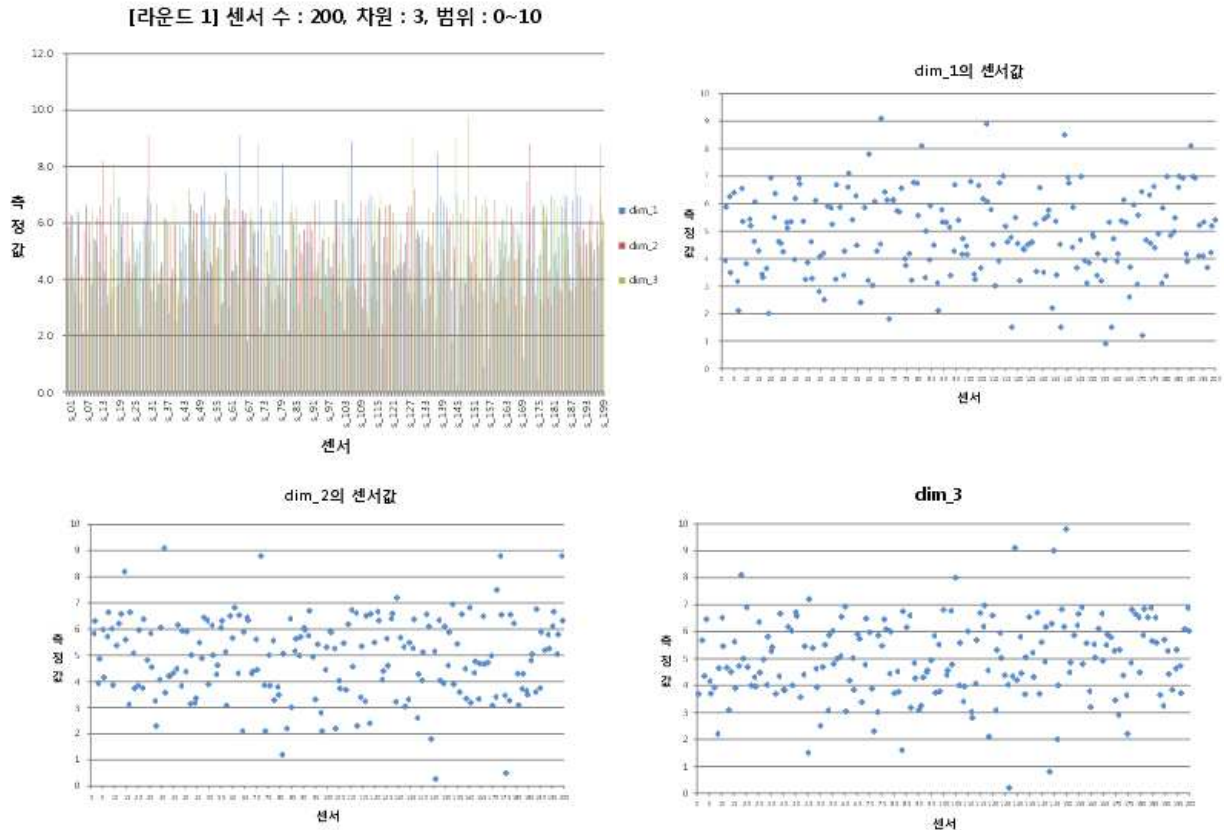


Fig. 11. Normal distribution range of values

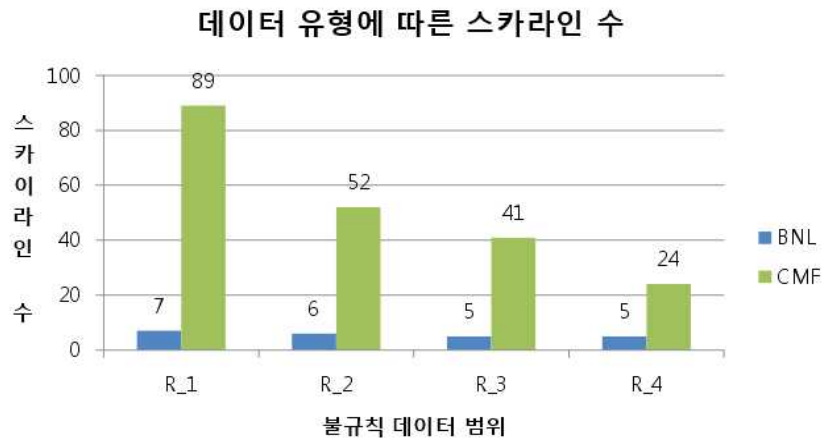


Fig. 12. Skyline of random data type

Table 5와 같이 4개의 라운드로 BNL, CMF 기법으로 스카이라인 포인트를 선출하였을 때 선출되는 스카이라인의 수를 비교하였다. CMF에서 우선순위 데이터 처리 후 소속 값의 합의 순위에 의해 추가로 검출되는 포인트 수는 사용자 정의로 전체 센서 수의 10%로 제한하였다.

Fig. 12는 불규칙적인 랜덤 데이터를 이용하였을 경우의 스카이라인 선출 결과로 CMF의 스카이라인수가 BNL 보다

다수 검색되었다. 이는 우선순위 데이터 처리 기법을 이용하여 중요도 높은 속성의 값을 먼저 반영하여 나타난 결과이다. 이와 비교하여 Fig. 13은 정규 분포 범위의 데이터를 이용하였을 경우로 불규칙 범위의 데이터를 이용하였을 때 보다 BNL의 스카이라인 수의 차이가 줄었다. 이와 같은 차이는 정규 분포 범위의 데이터 유형에서는 우선순위 데이터 처리 기법을 통하여 추가되는 포인트가 줄었기 때문이다.

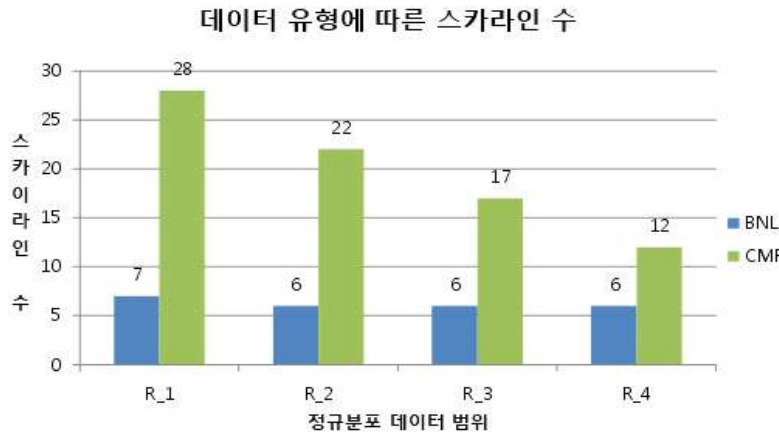


Fig. 13. Skyline of normal distribution data type

하지만 CMF 기반의 우선처리 기법의 결과는 기존의 스카이라인 처리 기법 보다 많은 수의 관심 포인트를 포함한다. 이는 각각의 단일 속성으로 의미 있는 데이터를 검출하고 차원의 속성 중 우수한 값을 가지고 있는 포인트를 전체 센서의 수의 10% 만큼 유지하였기 때문이다.

### 5. 결론 및 향후 연구

본 논문에서는 센서 네트워크 환경에서 관심 포인트 값을 찾기 위해 사용되는 스카이라인 질의 처리를 이용할 경우 발생하는 문제에 관하여 분석하고, 이를 해결하기 위해 CMF의 소속 값을 이용하는 기법을 제안하였다. CMF를 이용하여 스카이라인을 검색할 경우 센서의 수와 차원의 증가로 급격히 증가하는 비교횟수를 줄일 수 있었으며 우선순위 데이터 처리를 통해 지배력이 큰 스카이라인 포인트로부터 제외되는 의미 있는 속성의 값도 검출 할 수 있었다. 여기서 사용되는 우선순위 데이터 처리 기법은 정규분포를 따르는 데이터 집합에서는 유용하나 밀집도 높은 균등분포 데이터에서는 상대적으로 많은 관심 포인트 값을 갖는다. 따라서 향후 연구에서는 데이터 분포에 따른 가변적인 우선순위 데이터 처리 방법에 대하여 진행할 계획이다.

### 참 고 문 헌

[1] Alan Mainwaring, David Culler, Joseph Polastre, Robert Szewczyk, and John Anderson, "Wireless sensor networks for habitat monitoring," Proc. the 1st ACM international workshop on Wireless sensor networks and applications, pp.88-97, 2002.

[2] Ning Xu, Ning Xu, Krishna Kant Chintalapudi, Deepak Ganesan, Alan Broad, Ramesh Govindan, and Deborah Estrin, "A wireless sensor network for structural monitoring," Proc.

the 2nd international conference on Embedded networked sensor systems, pp.13-25, 2004.

[3] Donald Kossmann, Frank Ramsak, and Fen Rost, "Shooting stars in the sky: an online algorithm for skyline queries," Proc. VLDB, pp.275-286, 2002.

[4] Yoon KWON, Jae-Ho CHOI, Yon-Dohn CHUNG, and SangKeun LEE, "In-Network Processing for Skyline Queries in Sensor Networks," IEICE TRANS. COMMUN., Vol.E90-B, No.12, pp.3452-3459, 2007.

[5] Hekang Chen, Shuigeng Zhou, and Jihong Guan, "Towards Energy-Efficient Skyline Monitoring in Wireless Sensor Networks," Proc. European Workshop on Wireless Sensor Networks, 2007.

[6] Junchang Xin, Guoren Wang, Lei Chen, Xiaoyi Zhang, and Zhenhua Wang, "Continuously Maintaining Sliding Window Skylines in a Sensor Network," Proc. DASFAA, 2007.

[7] Stephan Börzsönyi, Donald Kossmann, and Konrad Stocker, "The Skyline Operator," Proc. the 17th International Conference on Data Engineering, pp.421-430, 2001.

[8] Dimitris Papadias, Yufei Tao, Greg Fu, and Bernhard Seeger "An optimal and progressive algorithm for skyline queries," Proc. SIGMOD Conf. pp.467-478, 2003.

[9] Wolf-Tilo Balke, Ulrich Güntzer, and Jason Xin Zheng, "Efficient Distributed Skylining for Web Information Systems," Proc. EDBT Conf. pp.256-273, 2004.

[10] Xuemin Lin, Yidong Yuan, Wei Wang, and Hongjun Lu, "Stabbing the sky: efficient skyline computation over sliding windows," Proc. ICDE, pp.502-513, 2005.

[11] Zhiyong Huang, Christian S. Jensen, Hua Lu, and Beng Chin Ooi, "Skyline Queries Against Mobile Lightweight Devices in MANETs", Proc. IDCDE, pp.66, 2006.

[12] 한국태양에너지학회, 대한민국표준기상데이터 86년-05년 기준, 2010.

[13] 한국화재보험협회, "기상과 화재와의 관계" 화재안전점검 13호, pp.66-70, 1980.



**김진환**

e-mail : kim@hallym.ac.kr  
2003년 한림대학교 컴퓨터공학과,  
수학과(학사)  
2005년 한림대학교 컴퓨터공학과(석사)  
2012년 한림대학교 컴퓨터공학과  
박사과정

2005년~2006년 한림대학교 교무연구처 연구원  
관심분야: 센서 네트워크, 대용량 데이터 처리, 분산처리,  
클라우드 컴퓨팅 등



**이광모**

e-mail : kmlee@hallym.ac.kr  
1975년 서울대학교 응용수학과(학사)  
1984년 서울대학교 계산통계학과(석사)  
1992년 서울대학교 계산통계학과(박사)  
1985년~현 재 한림대학교 정보전자공과  
대학 컴퓨터공학과 교수

관심분야: 병렬처리, 분산처리, 클러스터 컴퓨팅 등

## 센서 네트워크에서 다차원 데이터 스카이라인 질의 처리를 위한 CMF 기반의 우선처리 기법

김진환<sup>\*</sup> · 이광모<sup>\*\*</sup>

### 요 약

데이터베이스 분야에서 다수의 속성을 갖는 데이터의 효율적인 의사 결정을 지원하는 스카이라인 질의에 관한 연구가 활발히 진행되고 있다. 스카이라인 질의란 대량의 데이터에서 필요한 관심 정보를 검색할 때 모든 속성의 데이터를 탐색하지 않고 속성 내에 의미 있는 데이터만 탐색하는 것이다. 이와 같은 스카이라인 질의는 센서 네트워크에서 다양한 환경 및 상황정보를 수집하여 사용자에게 제공하기 위해 유용하게 활용할 수 있다. 그러나 기존의 스카이라인 선출 방식은 다차원 데이터에서 스카이라인 선출시 센서의 수와 차원이 증가함에 따라 비교 계산 횟수가 급격히 증가하며 또한 지배력이 큰 값에 의해 단일 속성으로도 의미 있는 값이 제외될 수 있다. 따라서 본 논문에서는 싱크 노드로부터 관심(interest) 질의를 하위 노드로 전송할 때 전체 데이터 중 일부 데이터들의 선호도(preference)를 관별할 수 있는 카테고리 기반 소속 함수(CMF : Category Based Member Function)를 함께 전송하여 스카이라인 선출시 차원의 증가로 발생할 수 있는 비교 계산의 복잡성을 감소시키고 선호도 높은 우선순위 데이터를 처리하는 기법을 제안한다. 제안된 기법의 우수성을 보이기 위해 시뮬레이션을 통한 성능평가를 수행하였으며 그 결과 다차원의 센서 데이터 집합에서 데이터 검출 시 카테고리 기반 소속 함수를 기반으로 한 처리기법에서 시간 복잡도가 감소함을 보였으며 지배력이 큰 스카이라인으로부터 제외된 의미 있는 속성 값을 검출할 수 있었다.

**키워드 :** 스카이라인 질의, 다차원 데이터, 선호 데이터, 센서 네트워크