DOI: 10.3745/KIPSTD.2009.16-D.3.307

GIS-AMR 시스템에서 시공간 데이터마이닝 기법을 이용한 전력 소비 패턴의 분석 및 예측

박 진 형[†] · 이 헌 규[††] · 신 진 호[†††] · 류 근 호[††††]

요 약

이 논문에서는 자동 원격 검침(AMR) 시스템에서 수집되는 전력 사용량 데이터의 분석 결과를 실세계에 적용하기 위하여 시간과 공간의 변화에 따른 전력 소비 패턴의 주기성 탐사를 위한 시공간 데이터마이닝 기법을 제안하였다. 첫째, 고객의 전력 사용 목적에 따른 군집 분석을 위하여 분할 군집화 기법을 적용하였다. 둘째, 3차원 큐브 마이닝 기법을 적용하여 고객의 전력 소비 데이터가 갖는 시간 속성과 공간 속성에 대한 패턴을 탐색하였다. 셋째, 다양한 시간 도메인에서의 주기 패턴 발견을 위한 캘린더 패턴 마이닝 기법을 이용하여 탐사된 패턴들이 갖고 있는 시간 속성의 의미와 관계를 분석 및 예측하였다.

제안된 시공간 데이터마이닝 기법을 평가하기 위해 한국 전력 연구원에서 구축된 GIS-AMR 시스템에 의해 제공되는 고압 전력 소비 고객 3,256명의 2007년 1월부터 4월까지 총 266,426건의 데이터로부터 시간의 주기성 및 공간적 특성을 포함한 전력 소비 패턴을 분석하였다. 제안한 분석 기법을 통하여 특정 그룹에 속한 각각의 대표 프로파일이 시간과 공간상에서 갖는 주기성을 발견하였다.

키워드 : 전력 소비 패턴 분석, 3차원 큐브 마이닝, 시공간 데이터마이닝, 캘린더 패턴 마이닝

Analysis and Prediction of Power Consumption Pattern Using Spatiotemporal Data Mining Techniques in GIS-AMR System

Jin Hyoung Park[†] · Heon Gyu Lee[††] · Jin-Ho Shin[†††] · Keun Ho Ryu[††††]

ABSTRACT

In this paper, the spatiotemporal data mining methodology for detecting a cycle of power consumption pattern with the change of time and spatial was proposed, and applied to the power consumption data collected by GIS-AMR system with an aim to use its resulting knowledge in real world applications. First, partial clustering method was applied for cluster analysis concerned with the aim of customer's power consumption. Second, the patterns of customer's power consumption data which contain time and spatial attribute were detected by 3D cube mining method. Third, using the calendar pattern mining method for detection of cyclic patterns in the various time domains, the meanings and relationships of time attribute which is previously detected patterns were analyzed and predicted.

For the evaluation of the proposed spatiotemporal data mining, we analyzed and predicted the power consumption patterns included the cycle of time and spatial feature from total 266,426 data of 3,256 customers with high power consumption from Jan. 2007 to Apr. 2007 supported by the GIS-AMR system in KEPRI. As a result of applying the proposed analysis methodology, cyclic patterns of each representative profiles of a group is identified on time and location.

Keywords : Analyze Of Power Consumption Pattern, 3D Cube Mining, Spatiotemporal Data Mining, Calendar Pattern Mining

1. 서 론

지리 정보 시스템은 다양한 분야에 적용되어 실생활에 많은 도움을 주고 있다. 실제로 공간 정보 시스템과 센서 네트워크와 결합된 형태의 시스템을 통해 사용자는 실시간으로 공간 정보를 수집 및 분석할 수 있게 되었다. 현재 한전에서는 지리 정보 시스템(GIS)과 고압 전력 소비 고객의 원격 검침(AMR : Automatic Meter Reading) 시스템의 기술개발 추세를 반영하여 우리나라 배전계통에 적합한 부하분석시스템을 위해 "GIS-AMR 기반의 배전 고압계통 부하분석모델"을 정립하였다. 따라서 실시간 원격검침 정보와 인터넷 GIS 엔진인 SIAS(Small world Internet Application Server)를

※ 이 논문은 2009년도 정부(과학기술부)의 재원으로 한국과학재단(R01-2007-000-10926-0)과 2009년 교육과학기술부(지역거점연구단육성사업/충북BIT연구중심대학육성사업단)의 지원을 받아 수행된 연구임.
† 준 회 원 : 충북대학교 전자계산학과 석사과정
†† 정 회 원 : 한국전자통신연구원 우정물류기술부 연구원
††† 정 회 원 : 한국전력공사 전력연구원
†††† 종신회원 : 충북대학교 전기전자컴퓨터공학부 교수
논문접수 : 2008년 12월 10일
수 정 일 : 1차 2009년 3월 4일
심사완료 : 2009년 3월 4일

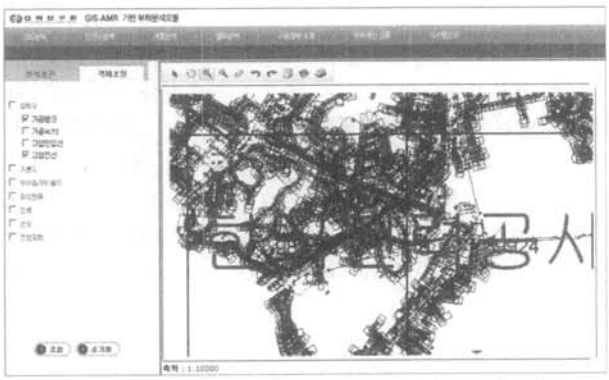
사용하여 (그림 1)과 같이 전국 고압 고객에 대한 실시간 부하현황에 대한 분석이 가능해졌다[1].

하지만 이러한 시스템 구조는 사용자에게 실시간으로 일시적인 데이터를 모니터링 하는 정도의 서비스만 제공할 뿐, 데이터가 갖는 의미있는 지식 제공에 대해서는 한계점을 갖고 있다. 따라서 실생활에 유용하게 적용 가능한 지식을 찾기 위한 많은 데이터 마이닝 기법들이 제안되어 왔다. 대표적인 기법으로는 특징 벡터 추출(Feature Selection)[2], 군집화(Clustering)[3], 회귀분석[4], 분류(Classification)[5, 15] 등이 있다. 하지만 기존 연구들은 대부분의 데이터가 시간과 공간 속성을 가졌음에도 불구하고 이를 고려하지 않고 있다. 이는 데이터 마이닝 기법 적용 결과가 아무리 정확하고 유용하더라도 그 지식을 적용할 시점과 지역을 표현하지 않으므로 모호한 지식이라 할 수 있다.

시공간 데이터마이닝은 이러한 기존 연구의 문제점을 해결하기 위해서 시간, 공간 및 시공간 특성을 포함하고 있는 방대한 양의 데이터 집합으로부터 이전에 알려지지 않았던 잠재적으로 유용한 시공간 지식을 탐사하기 위한 데이터 마이닝 연구 분야이다. 따라서 시공간 데이터마이닝을 통해 데이터들로부터 시점과 지역에 대한 다양한 형태의 지식을 탐사할 수 있다. 기존 연구에서 데이터가 갖는 시간과 공간 속성을 고려하지 않은 문제점을 해결하기 위하여 이 논문에서는 고압 고객의 전력 소비 대표 프로파일이 갖는 시간 및 공간적인 특성을 추출하고, 시간과 공간의 변화에 따른 대표 프로파일의 주기성 탐사를 위한 시공간 데이터마이닝 기법을 제안한다. 논문은 다음과 같은 내용으로 구성된다.

첫째, 기존의 고압 전력 소비 고객의 전력 소비 목적을 고려하기 위하여 계약종별 단위의 분할 군집화를 통해 군집분석을 실시한다. 둘째, 고압 고객의 전력 소비에 대한 대표 프로파일이 갖는 시간과 공간 속성을 고려한 패턴을 찾기 위하여 3차원 큐브 마이닝을 이용하고, 패턴 분석을 통해 각 속성간의 관계를 규명한다. 셋째, 이전 단계에서 찾아진 패턴으로부터 시간의 변화에 따라 적용 시점이 명확한 지식을 탐사하기 위하여 캘린더 패턴 마이닝 기법을 적용한다.

이 논문을 효율적으로 기술하기 위하여 다음과 같이 구성하였다. 2장에서는 기존 연구들을 정리하고 문제점을 분석한다. 3장에서는 실험에 사용된 데이터를 설명하고, 정확한 데이터마이닝 기법 적용을 위한 전처리 과정에 대해 기술한다. 4장에서는 3차원 데이터의 패턴 탐색을 위한 큐브마이너에 대해 설명한다. 5장에서는 시간 연관규칙 탐사의 문제 정의 및 알고리즘을 기술한다. 6장에서는 제안한 알고리즘을 구현하여 실험한 후에 성능 평가 및 결과를 분석한다. 마지막으로 7장에서는 이 논문에 대한 전체적인 결론과 향후연구로써 보완해야 될 문제점 및 향후연구 방향을 제시한다.



(그림 1) 부하분석모델 인터넷 GIS 스타일 적용[1]

2. 관련 연구

2.1 전력 소비 패턴 분석 및 예측

전력부하 예측은 발전 비용의 감소와 배전 설비 투자 결정, 전력 계통의 제어 등 배전 시스템의 효율적인 운용에 있어서 매우 중요하기 때문에 이를 위해 많은 데이터마이닝 기법들이 제안되었다. 대표적으로 신경망 분류기법(Multilayer Perceptron)[5]과 시계열 분석기법(ARIMA : Autoregressive Integrated Moving Average)[6], 온도 변화에 따른 전력 부하의 회귀 분석기법[4] 등이 있다. 신경망 분류기법은 은닉층의 수와 각 은닉층의 노드 수를 결정하는 적절한 방법이 존재하지 않는다는 단점을 갖고 있다[5]. 시계열 방법은 예측 모형을 설정하는데 있어서 자기 변수만을 고려하기 때문에 단순한 선형 모델을 형성함에 따라 좋은 예측 결과를 기대할 수 없다[6]. 그리고 온도 변화에 따른 회귀 분석기법은 온도에 민감한 지역 또는 계약 종별에 대해서 패턴들을 예측할 수 있지만, 패턴의 시간과 공간 속성에 대해서는 표현이 불가능하다[4]. 인공지능 계열의 데이터마이닝 기법 중에서 신경망 분류 기법이나 초평면을 이용한 분류기법(SVM)들은 비교적 좋은 성능을 보인다고 알려져 있다[16]. 하지만 시스템 운영상의 시간, 공간 복잡도로 인해 대용량 데이터 분석에 적합하지 않다. 따라서 데이터의 차원을 줄이기 위한 특징 벡터 추출 기법 기법들을 이용한 기법[2]들이 제안되었다. 하지만 특징 벡터 추출 기법이 정확한 특징을 추출하여도 원본 데이터의 상세한 특징이 제거된다는 한계점이 있다.

2.2 3차원 큐브 마이닝

빈발패턴이란 주어진 데이터 집합에서 사용자에 의한 임계값보다 빈발하게 발생되는 데이터 집합을 말한다. 2차원 데이터에서의 빈발패턴 탐색 알고리즘으로는 대표적으로 Apriori 기법[7]과 FP-Growth 기법[8]이 있다. 그리고 3차원 데이터일 경우에는 Reprisentative Slice Mining(RSM)[9] 기법과 CubeMining[9] 기법이 있다. 이 논문에서는 시간과 공간과 대표 프로파일간의 패턴을 찾기 위하여 3차원 데이터에서의 패턴 탐색 기법을 고려하였다.

(1) Representative Slice Mining

Representative Slice Mining(RSM)은 데이터를 0과 1로 표현되는 Slice 기반의 데이터 구조를 이용하여 2차원에서 고차원으로 빈발패턴을 탐색하는 기법이다[9]. RSM 기법은 최초 두 가지 속성(R/C)에 대한 Slice들을 작성하고, 각각의

〈표 1〉 이진 데이터 매트릭스 예제

	$H = h_1$					$H = h_2$						$H = h_3$					
R/C	c_1	c_2	c_3	c_4	c_5	R/C	c_1	c_2	c_3	c_4	c_5	R/C	c_1	c_2	c_3	c_4	c_5
r_1	1	1	1	0	1	r_1	1	1	1	1	1	r_1	1	1	1	0	0
r_2	1	1	1	0	0	r_2	0	1	1	1	0	r_2	1	1	1	0	0
r_3	1	1	1	1	1	r_3	1	1	1	1	0	r_3	1	1	1	1	0
r_4	0	0	1	0	1	r_4	1	1	1	0	1	r_4	1	1	0	1	1

〈표 2〉 RSM 예제 (minH = minR = minC = 2)

Height Set	Representative Slices	2D FCPs	3D FCCs
h2, h3	1 1 1 0 0 0 1 1 0 0 1 1 1 1 0 1 1 0 0 1	r1r3 : c1c2c3, 2 : 3 r1r3r4 : c1c2, 3 : 2 r1r2r3 : c2c3, 3 : 2	h2h3 : r1r3r4 : c1c2, 2 : 3 : 2
h1, h3	1 1 1 0 0 1 1 1 0 0 1 1 1 1 0 0 0 0 0 1	r1r2r3 : c1c2c3, 3 : 3	h1h3 : r1r2r3 : c1c2c3, 2 : 3 : 3
h1, h2	1 1 1 0 1 0 1 1 0 0 1 1 1 1 0 0 0 1 0 1	r1r4 : c3c5, 2 : 2 r1r3 : c1c2c3, 2 : 3 r1r2r3 : c2c3, 3 : 2	h1h2 : r1r4 : c3c5, 2 : 2 : 2
h1, h2, h3	1 1 1 0 0 0 1 1 0 0 1 1 1 1 0 0 0 0 0 1	r1r3 : c1c2c3, 2 : 3 r1r2r3 : c2c3, 3 : 2	h1h2h3 : r1r3 : c1c2c3, 3 : 2 : 3 h1h2h3 : r1r2r3 : c2c3, 3 : 3 : 2

슬라이드를 중첩해 나가면서 빈발한 패턴을 찾는다. 예를 들어 주어진 데이터집합이 <표 1>과 같을 때, RSM 기법은 <표 2>와 같은 빈발패턴 탐색 과정을 통해 3가지 속성에 대한 빈발 패턴을 찾는다. 하지만 슬라이드(H)의 개수가 n일 때, RSM은 슬라이드 중첩과정에서 $O(2^n)$ 만큼의 시간 복잡도를 갖기 때문에 대용량 데이터에는 부적합하다.

(2) CubeMiner

CubeMiner[9]는 RSM과 같은 데이터 구조를 이용하여 세 가지 속성을 모두 만족하지 않는 Cutter set을 이용하여 전체 데이터 집합에서 제거해나가는 방식으로 빈발한 큐브(FCC : Frequent Closed Cube)를 만든다. CubeMiner에 대한 자세한 설명은 4장에서 기술한다.

2.3 시간 연관규칙

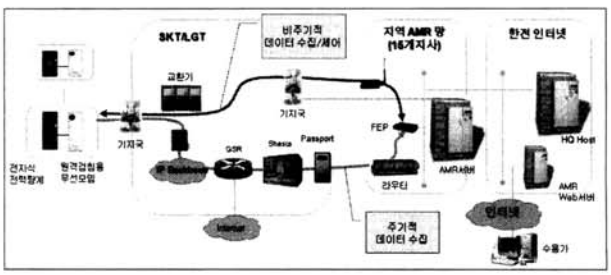
시간 연관규칙이란 기존 연관규칙에 시간 개념을 추가하여 시간 데이터로부터 시간 의미와 시간 관계를 가지는 유용한 지식을 탐사는 기법이다[10]. 시간 속성을 가지는 연관규칙에 대한 이전 연구들은 크게 주어진 시간 간격 동안 주기적으로 발생되는 현상, 즉 시간 간격에서의 완전한 주기성을 만족하는 연관규칙을 탐사하는 주기적 연관규칙 탐사[11]와 캘린더로 표현된 시간 패턴을 가지는 연관규칙을 탐사하는 캘린더 기반 연관규칙탐사[12]로 분류할 수 있다. 그러나 실제로 주어진 시간 간격 동안에 완전하게 유지되는 규칙은 존재하지 않으며 대부분 불완전한 주기를 이루고 있고, 주기적 연관규칙에서는 다양한 시간 단위를 표현하지 못하고 단 하나의 시간 단위만을 다룰 수 있다. 따라서 "매달 첫 번째 주 금요일"과 같이 실제 응용분야에서 적용되는 시간 표현은 불가능하다.

3. 데이터 전처리

3.1 원격 검침 데이터

한국 전력은 전기 사용 고객을 사용 전력에 따라 고압과 저압 고객으로 나눈다. 현재 국내 전체 전력 소비량의 약 70%를 차지하는 고압(100kw 이상) 고객(약 12만호)에 대하여 CDMA망을 이용한 자동 원격검침(AMR)을 실시하고 있다. 이를 통해 실시간 검침정보 조회가 가능하고, 수집된 정보는 분석 및 통계 등에 활용되고 있다. AMR 시스템은 (그림 2)와 같이 구성되어 주기적인 데이터 수집과 비주기적 데이터 수집 및 제어를 하고 있다.

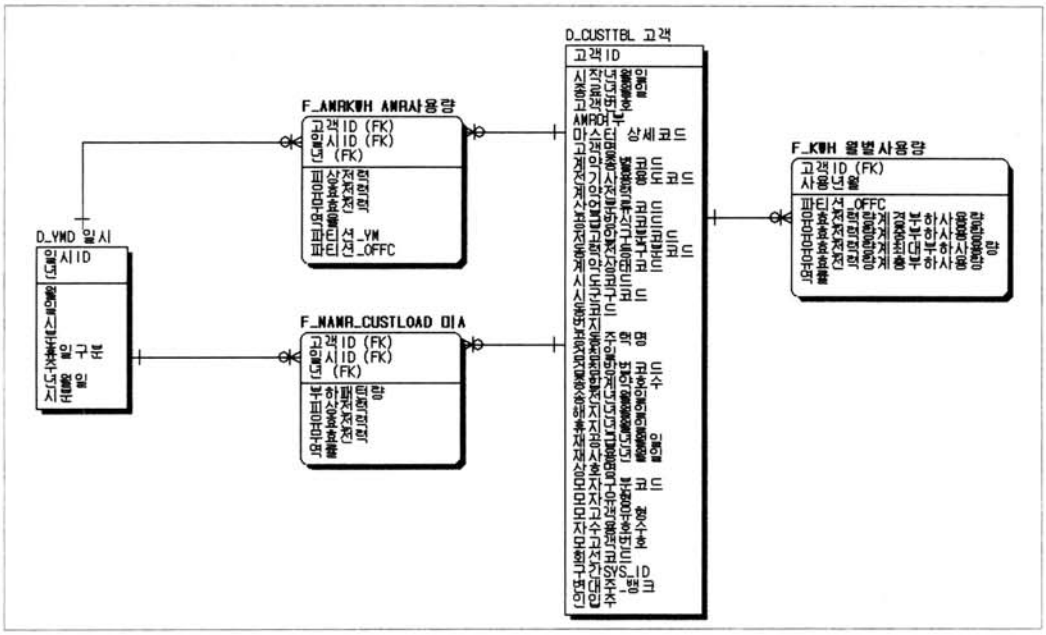
AMR 시스템에서 원격검침용 무선 모뎀은 각 고압 고객의 전력량계에 설치되어 15분간 전력 소비량을 저장하고 1시간 단위로 누적된 데이터를 기지국을 통해 AMR 서버로 전송하여 실시간 모니터링에 사용된다. 모니터링에 사용된 데이터는 부하 분석 모델을 위한 데이터 웨어하우스에 저장되어 분석 및 통계에 사용된다.



(그림 2) 자동 원격검침 시스템 구성도

3.2 데이터 추출 및 변환

3차원 큐브 마이닝을 위하여 수집된 데이터에서 시간, 공간 그리고 전력 소비에 대한 대표 프로파일을 추출해야 한다. 고압 고객에 대한 공간 정보를 추출하기 위해 (그림 3)의 AMR 부하 분석 대상 테이블 중 고객 테이블(D_CUSTTBL)에서 고객 ID, 고객의 지역코드, 계약종별 코드를 추출한다.



(그림 3) AMR 부하 분석 대상 테이블 (ERD)

이 논문에서는 일일 단위의 단기 부하 예측을 목표로 하기 때문에 15분 단위 전력 소비 데이터를 일일 사용량 패턴으로 변환해야 한다. 따라서 AMR 사용량 테이블(F_AMRKWH)에 저장된 15분 단위의 전력 소비 데이터를 다음 <식 1>과 같이 순차적으로 연결하여 1일 96개 데이터의 Vector형으로 표현하고, 패턴의 시간 정보를 추출하기 위하여 일시 테이블(D_YMD)의 년, 월, 일 속성을 이용한다.

$$V^{(b^{th})} = V_0^{(b^{th})}, \ldots, V_t^{(b^{th})}, \ldots, V_{2345}^{(b^{th})}$$

(b^{th} = 각 AMR 고객, t = 측정 시간, 0~2345) <식 1>

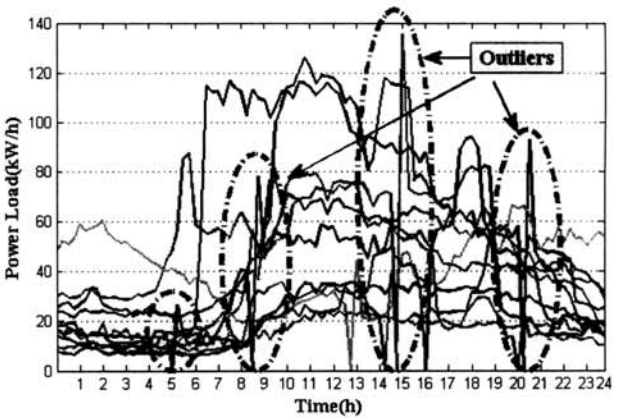
데이터 추출 과정을 통해 최종적으로 <표 3>과 같은 입력 데이터 집합을 형성한다.

〈표 3〉 AMR 부하 패턴 예측을 위한 질의 데이터 집합

속성명	데이터 타입	코드명
CUST_ID	nominal	고객 구분 코드
CNTR_KND_CD	nominal	계약 종별
SIDO_CD	nominal	시도 코드
SIGUNGU_CD	nominal	시군구 코드
DONGCD	nominal	동 코드
AVB_PWR	continuous	유효 전력 (일별 15분 간격)

3.3 데이터 전처리

AMR 시스템은 고압 고객의 전력 사용량을 측정하기 위해 전자식 전력량계를 이용하며, 무선으로 중앙 서버에 전송한다. 따라서 데이터 계측 및 전송 과정에서 누락 데이터 및 이상치 데이터를 포함하게 된다. 정확한 마이닝 작업을 위하여 누락 및 이상치 데이터를 제거해야 한다. 이상치 데이터를 처리하기 위한 정제 기법으로 SOMs 군집화 기법을



(그림 4) SOMs 군집화 정제 기법을 이용한 이상치 탐색

적용한다[3]. SOMs 군집화 정제 기법은 코호넨 네트워크 모델을 적용하여 (그림 4)와 같이 이상치 데이터를 찾았으며, 구성 매트릭스는 10 by 10(k=100)이다.

3.4 대표 프로파일 형성

고객정보 테이블(D_CUSTTBL)(그림 3)에서 계약 종별 속성은 유사한 전력 사용 목적을 갖는 고객들을 분류하기 위한 코드이다. 하지만, 하나의 계약 종별 집단 내의 고객들은 다양한 전력 소비 패턴을 갖기 때문에 계약종별 패턴만 가지고는 정확한 부하 집계를 할 수 없다. 기존 연구에서는 고객의 전력 사용 목적을 고려하지 하지 않았으며, 또한 대용량의 데이터를 군집화 함에 따른 시간 및 공간 복잡도를 고려하지 않는다는 점이 이전 연구의 한계이다[13]. 따라서 이 연구에서는 고객의 전력 사용 목적을 고려하고, 군집화의 효율성을 높이기 위해 고객의 계약 종별 단위로 분할하여 군집화를 실시한다. 계약 종별 단위의 군집 형성과 대표 부하 프로파일을 찾기 위한 SOMs 알고리즘[14]을 통해

〈표 4〉 3차원 큐브 마이닝을 위한 데이터 집합

속성명	데이터 타입	코드명
TIME	nominal	시간 ID
LOCAL	nominal	위치 코드 (시도, 시군구, 동코드)
PROFILE	nominal	대표 프로파일

<표 4>와 같이 시간과 공간, 그리고 계약종별 단위의 대표 프로파일에 대한 정보를 추출한다.

4. 큐브 마이너에 의한 3차원 빈발 패턴 탐색

4.1 문제 정의

큐브 마이닝은 3차원 데이터 집합에서 3가지 속성을 모두 만족하는 빈발패턴을 찾는 기법이다[9]. 이 기법은 전체 데이터 집합에서 cutter로 불리는 집합 Z를 이용하여 순환적으로 분리해 나가는 방법이다. 따라서 Z의 요소를 순차적으로 분리하여 cube의 값이 1 또는 사용자에 의해 지정된 임계값을 가질 때 까지 반복된다.

[정의 4.1] 행(column)의 집합 $C = \{c_1, c_2, ..., c_m\}$, 열(row)의 집합을 $R = \{r_1, r_2, ..., r_n\}$, 그리고 높이(height)의 집합을 $H = \{h_1, h_2, ..., h_l\}$라고 할 때, $k \in [1, l]$, $i \in [1, n]$, 그리고 $j \in [1, m]$이면, 3차원 데이터는 $n \times m \times l$의 이진 매트릭스 $O = H \times R \times C = \{O_{k,i,j}\}$로 표현된다. $O_{k,i,j}$는 높이가 k, 열이 i, 행이 j인 데이터를 가리킨다. 예를 들어, <표 1>에서 h_1과 r_4는 c_3와 c_5에 동시 포함하므로, $C(h_1 \times r_4) = \{c_3, c_5\}$으로 표현된다. 그리고 r_2와 c_1은 h_1과 h_3에 동시 포함되고, $H(r_1 \times c_4) = \{h_1, h_3\}$으로 표현된다.

[정의 4.2] 높이 지지 집합과 높이-지지도(Height Support Set and H-Support): 주어진 열의 집합 R′이 R에 포함되고, 주어진 행의 집합 C′가 C에 포함될 때, R′와 C′를 동시에 포함하는 높이의 최대빈발 집합을 높이 지지 집합 $H(R' \times C')$이라고 한다. 이 때, $H(R' \times C')$의 높이의 수는 높이-지지도(H-Support)라고 하며 $|H(R' \times C')|$로 표기한다. 예를 들어, <표 1>에서 $R' = \{r_1, r_2\}$이고, $C' = \{c_1, c_2, c_3\}$일 때, R′와 C′를 동시에 포함하면서 h_1과 h_3을 포함하는 더 큰 집합이 없기 때문에 높이에 대한 최대빈발 집합 $H(R' \times C') = \{h_1, h_3\}$이다.

[정의 4.3] 열 지지 집합과 열-지지도(Row Support Set and R-Support): 주어진 행의 집합 C′이 C에 포함되고, 주어진 높이의 집합 H′가 H에 포함될 때, C′와 H′를 동시에 포함하는 열의 최대빈발 집합을 열 지지 집합 $R(C' \times H')$이라고 한다. 이 때, $R(C' \times H')$의 열의 수는 열-지지도(R-Support)라고 하며 $|R(C' \times H')|$로 표기한다.

[정의 4.4] 행 지지 집합과 행-지지도(Column Support Set and C-Support): 주어진 열의 집합 R′이 R에 포함되고, 주어진 높이의 집합 H′가 H에 포함될 때, R′와 H′를 동시에 포함하는 행의 최대빈발 집합을 행 지지 집합 $C(R' \times H')$이라고 한다. 이 때, $C(R' \times H')$의 행의 수는 행-지지도(C-Support)라고 하며 $|C(R' \times H')|$로 표기한다.

[정의 4.5] Closed Cube: 3가지 집합 $R' \subseteq R$, $H' \subseteq H$, $C' \subseteq C$가 주어졌을 때, 세 가지 조건 (1) $R' = R(C' \times H')$; (2) $C' = C(R' \times H')$; (3) $H' = H(R' \times C')$;을 만족하는 큐브 $A = (H' \times R' \times C') \subseteq O$를 Closed Cube 라고 정의되며, $A = (H', R', C')$ 으로 표기된다.

[정의 4.6] Frequent Closed Cube(FCC): 만일 closed cube $A = (H' \times R' \times C') \subseteq O$의 각각 열-지지도 $|R(C' \times H')|$, 행-지지도 $|C(R' \times H')|$ 그리고, 높이-지지도 $|H(R' \times C')|$가 사용자에 의한 최소 지지도 minR, minC, minH를 만족할 때, Frequent Closed Cube(FCC)라고 정의한다.

4.2 큐브 마이너(CubeMiner)

큐브 마이너는 Frequent Closed Cube(FCC) 마이닝을 위한 새로운 알고리즘이다. 사용자 임계값을 기준으로 (H′, R′, C′)의 3개 속성을 동시에 포함하는 집합을 찾아낸다.

[정의 4.7] Cutter: 만일 $\forall t_k \in W$, $\forall l_i \in X$ 그리고 $\forall p_j \in Y$, $O_{t,l,p} = 0$일 경우 $(W, X, Y) \in Z$의 요소들을 Cutter라고 한다. 또한, W, X, Y 각각을 cutter(W, X, Y)의 왼쪽 원소, 가운데 원소, 오른쪽 원소라고 한다. T를 시간집합, L을 지역코드 그리고 P를 대표 프로파일 집합이라고 할 때, (T, L, P)의 서브집합 (T′, L′, P′)의 왼쪽 원소는 (T′\W, L′, P′), 가운데 원소는 (T′, L′\X, P′), 오른쪽 원소는 (T′, L′, P′\Y)로 정의된다. <표 5>는 <표 2>의 예제 데이터로부터 추출된 Cutter 집합이다.

만일 $W \cap T' \neq \emptyset$, $X \cap L' \neq \emptyset$ 그리고 $Y \cap P' \neq \emptyset$이면 Z집합에 있는 cutter(W, X, Y)는 cube(T′, L′, P′)을 분류하는데 사용할 수 있다. 반복적인 분류에 의해 모든 FCCs를 만들지만 몇몇 unclosed한 최대빈발 프로파일 패턴이 발생한다. 제거 단계에서 이러한 FCCs를 제거함으로서 완전한 FCCs를 가질 수 있다. 최대빈발 프로파일 패턴 중 unclosed한 대표 프로파일 패턴을 제거하는 단계는 다음과 같다.

(1) Left Track Checking: Cutter (W, X, Y)에 의해서 $O' = (T', L', P')$의 왼쪽에 위치한 자식 노드를 $L = (T'\backslash W, G', S')$이라고 하면 $W \cap TLo' \neq \emptyset$이면 L은 제거 할 수 있다.

(2) Middle Track Checking: Cutter (W, X, Y)에 의해

〈표 5〉 Cutter set Z 예제

W, X, Y
h_1, r_1, c_4
h_1, r_2, c_4c_5
$h_1, r_4, c_1c_2c_4$
h_2, r_2, c_1c_5
h_2, r_3, c_5
h_2, r_4, c_4
h_3, r_1, c_4c_5
h_3, r_2, c_4c_5
h_3, r_3, c_5
h_3, r_4, c_3

서 O′ = (T′, L′, P′)의 가운데 위치한 자식 노드를 L = (T′, G′\X, S′)이라고 하면 X ∩ TMo′ ≠ ∅이면 M은 제거 할 수 있다.

(3) Close Height Set Checking: O′의 가운데, 오른쪽 자식노드를 O″ = (T″, L″, P″) 그리고 Z를 전체 cutter 집합이고 하며 $l_x \in L''$, $P'' \cap P_y = \emptyset$에서 $\forall(\{t_w\}, \{l_x\}, P_y) \in Z$와 같이 $\exists tw \in (T \backslash T')$ (T는 O의 전체 time 집합)이면 O″은 time 집합에서 unclosed이며 제거 할 수 있다. 왼쪽 자식 노드가 조건에 전혀 만족하지 않으면 오직 가운데, 오른쪽 자식들은 이 검사가 필요하다.

(4) Close Row Set Checking: O′의 왼쪽, 오른쪽 자식 노드를 O″ = (T″, L″, P″) 그리고 Z를 전체 cutter 집합이고 하며 $l_x \in L''$, $P'' \cap P_y = \emptyset$에서 $\forall(\{t_w\}, \{l_x\}, P_y) \in Z$와 같이 $\exists tw \in (T \backslash T')$ (T는 O의 전체 time 집합)이면 O″은 time 집합에서 unclosed이며 제거 할 수 있다. 왼쪽 자식 노드가 조건에 전혀 만족하지 않으면 오직 가운데, 오른쪽 자식들은 이 검사가 필요하다.

3차원 전력 소비 패턴 데이터의 대표 프로파일 패턴을 발견하기 위해 큐브 마이닝 알고리즘에 적용한 결과는 (그림 5)와 같으며, 그림의 오른쪽 하단은 탐사된 패턴이다. (h, r, c)가 각각 높이, 열, 대표 프로파일일 경우, 각각에 대한 지지도를 (h, r, c : 8, 6, 4)로 설정하였을 때 탐사된 패턴이다.

(그림 5) 큐브마이너를 통한 3차원 데이터(Time-Local-Profile)의 패턴 탐사 결과

5. 시간 주기성 탐사를 위한 시간 데이터마이닝

이 장에서는 시간 데이터마이닝 중에서 캘린더 스키마에 대해 알아보고, 캘린더 스키마에 의해 정의되는 캘린더 패턴을 이용한 주기성 탐사 기법을 기술한다. 이 캘린터 스키마와 캘린더 패턴은 [11]의 정의를 사용하기 위하여 아래에 정리한다.

5.1 캘린더 스키마

캘린더 스키마는 달력의 개념 계층에 의해 결정되어지고 유효성 제약조건을 갖는 관계형 스키마이다.

[정의 5.1] 캘린더 스키마(*CS* : Calendar Schema): 캘린더 스키마는 달력 표현의 단위와 그 단위에서의 가능한 도메인의 집합으로 정의되며, 그 형태는 다음과 같다; $CS = (G_n:D_n, G_{n-1}:D_{n-1}, ..., G_1:D_1)$. $1 \le i \le n$에 대해, 속성 G_i는 년, 월, 일 등과 같은 달력 개념에서의 시간 단위이고, 각 D_i는 양의 정수의 유한 집합으로 속성 G_i의 도메인 값의 집합을 나타낸다.

캘린더 스키마가 $(G_n, G_{n-1}, ..., G_1)$이고 $1 \le i < n$일 때, 각 시간 단위 G_i는 유일하게 G_{i+1}에 포함된다. 이 논문의 실험에서는 '주'가 특정 '월'에 포함되도록 '주'의 도메인을 week:{1,2,3,4,5}으로 설정하고, 특정 '월'에 j 번째 주로 적용하였다. 또한 CS = (*year*:{1995~1999}, *month*:{1~12}, a *day*:{1~31}) 일 경우, {1995, 1, 20}은 유효한 시간 표현이 되나, {1996, 2, 31}은 유효한 조합이 되지 않는다.

5.2 캘린더 패턴

[정의 5.2] 캘린더 패턴(*CP* : Calendar Pattern): 캘린더 패턴은 주어진 스키마 $CS = (G_n:D_n, ..., G_1:D_1)$의 인스턴스이며, $CP = \{d_n, ..., d_1\}$으로 표현된다. 여기서 각 d_i는 D_i의 도메인 값이거나 문자 '*'이다. 만약 d_i가 '*'이라면, 그 의미는 도메인 D_i의 모든 값을 나타내고 "every"로 해석한다. 캘린더 패턴에 포함되어진 '*'의 개수에 따라 <표 6>과 같이 캘린더 패턴의 표현을 구분한다. i개의 '*'를 가지는 캘린더 패턴은 *i-star pattern*(CP_i)이라 부르고 *i-star pattern*들의 집합을 $\Phi(CP_i)$로 나타낸다. '*'를 전혀 포함하지 않는 캘린더 패턴에 대해서는 "기본시간단위"라고 부른다.

[정의 5.3] 최소발생도(minimum frequency): 시간 연관 규칙에서는 유용성 측면에서 최소발생도(minimum frequency)라는 새로운 임계값을 사용한다. 캘린더 패턴 $\Phi(CP)$에 속한 기본시간단위의 속성이 f % 이상을 만족한다면, 해당 속성은 '*'로 표기한다. 이때 f %를 최소발생도, *minFre*이라고 정의한다. 최소발생도의 적용 이유는 실제 응용에서 주어진 시간 간격 동안에 모두 성립되는 규칙은 거의 없기 때문에 유연성을 주기 위함이다.

[정의 5.4] 캘린더 패턴 사이의 포함관계, $CS = (G_n, G_{n-1}, ..., G_1)$, $CP = \{d_n, d_{n-1}, ..., d_1\}$, $CP' = \{d_n', d_{n-1}', ..., d_1'\}$, $1 \le i \le n$인 각 i에 대해, d_i = '*'이거나 $d_i = d_i'$일 때, CP는 CP'을 포함한다. 예를 들어 스키마가 (week, day, hour)일 경우, 정의 3.2에 대해 CP_1 = {1, *, 12}은 'day'에 대한 주기를 가지며, [정의 5.3]에 의해 또 다른 CP_0 = {2, 2, 12}을 포함한다. 따라서 기본시간단위 $CP_0 = \{d_n, ..., d_1\}$, $CP_0 \in \Phi$

〈표 6〉 캘린더 패턴(CP)의 표현 방식

'*'의 개수	캘린더 패턴	표현
0	0-star pattern	CP_0
1	1-star pattern	CP_1
...	...	...
i	i-star pattern	CP_i

(CP_0)이 주어지고 다른 i개의 '*'를 갖는 CP_i가 CP_0을 포함한다면, 그러한 CP_i들의 집합을 $\Phi(CP_i((CP_0))$로 나타낸다.

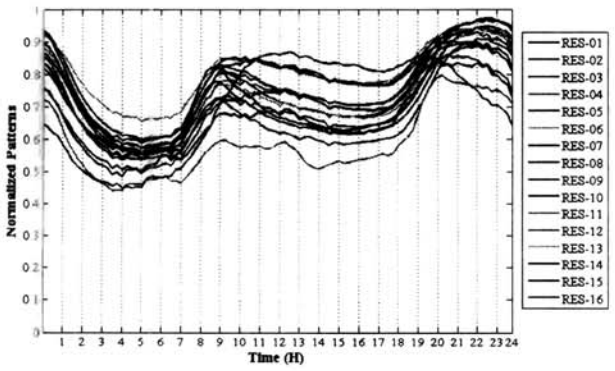
6. 실험 및 평가

본 연구에서 사용된 데이터는 한국 전력 연구원에서 구축한 배전 고압 계통 부하 분석 데이터 웨어하우스에서 AMR 시스템에 의하여 수집된 부하 데이터를 대상으로 한다. 실험에 사용된 데이터는 수도권역 고압 고객 3,256명의 2007년 1월부터 4월까지 총 266,426건의 데이터가 사용되었다. 이 장에서는 제안된 마이닝 알고리즘을 구현한 후에 여러 실험을 통해 제안한 방법론을 평가하고, 입력 파라미터들이 어떠한 영향을 미치는지를 분석해 보았다.

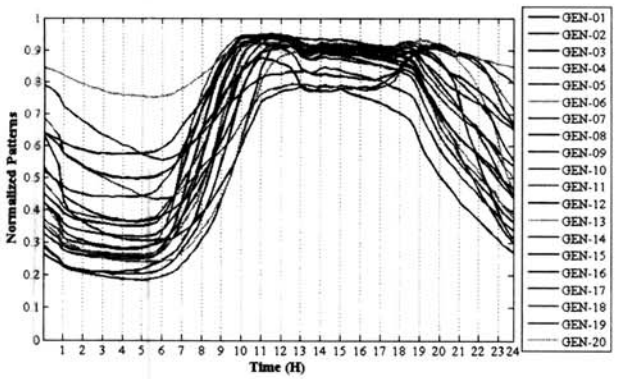
6.1 분할 군집 분석

고객의 군집 분석 단계에서는 기존에 사용되었던 군집 분석 기법의 단점을 보안하기 위해 고객의 전력 소비 특성을 고려한 계약종별 단위의 군집 분석을 실시하였다. (그림 6)~(그림 9)는 SOMs 군집 기법을 적용한 계약종별 대표 프로파일이다.

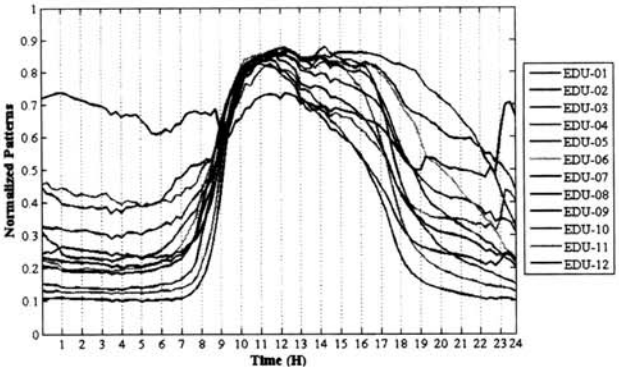
(그림 6)~(그림 9)와 같이 각각의 계약 종별 간에는 이질적인 전력 부하 패턴을 보이고 있고 같은 그룹 내에서는 유사한 패턴을 보이고 있다. 하지만 동일 그룹에서의 대표 패턴들은 특정 시간대에서 많은 차이를 보이고 있다. 이러한 분할 군집 방식을 통한 군집화 기법은 상대적인 군집수의 차이에 따른 패턴의 혼용을 방지하고, 군집형성에 따른 시간, 공간복잡도를 효율적으로 낮출 수 있었다.



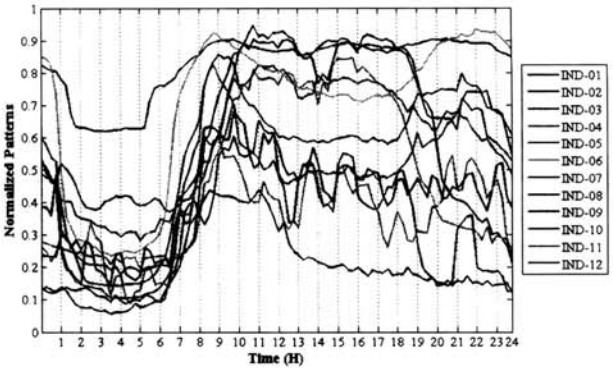
(그림 6) 주택용 대표 프로파일



(그림 7) 일반용 대표 프로파일



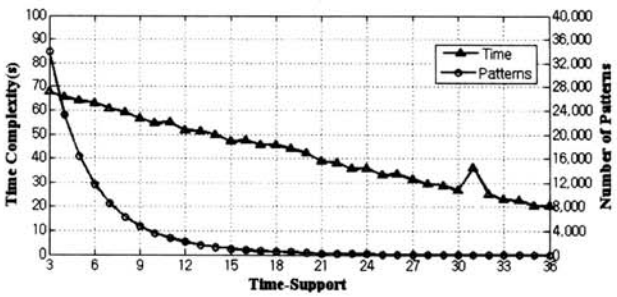
(그림 8) 교육용 대표 프로파일



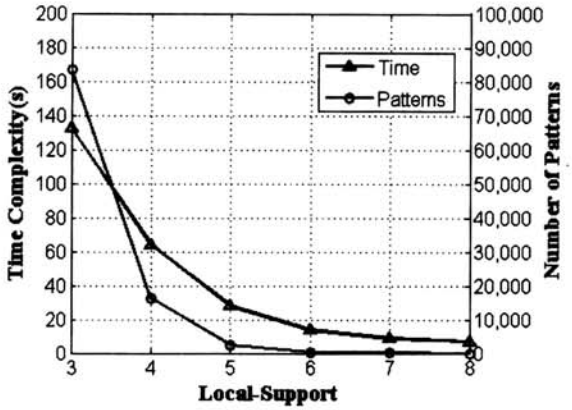
(그림 9) 산업용 대표 프로파일

6.2 3차원 큐브 마이닝

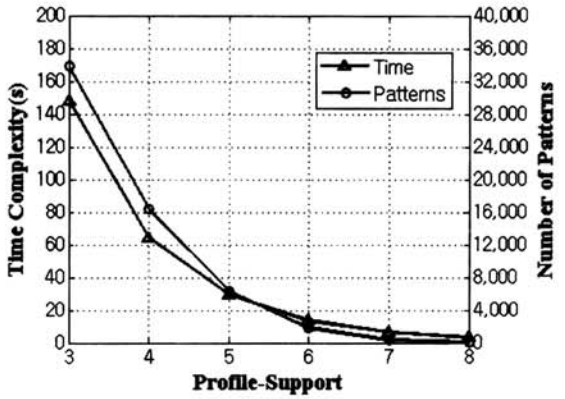
이 실험에서 큐브 마이닝 기법를 적용하여 시간 속성을 고려한 공간과 대표 프로파일간의 패턴 탐색 결과를 확인할 수 있다. 다음의 복잡도 그래프에서는 시간과 공간, 대표 프로파일의 지지도 변화에 따른 나머지 속성의 변화를 보여주고 있다. (그림 10)은 공간 및 대표 프로파일 지지도가 각각 4일 때, 시간 지지도의 변화에 따른 시간 복잡도와 생성된 패턴의 수의 추이를 보여주고 있다. (그림 11)는 공간 지지도의 변화에 따른 시간 복잡도와 생성된 패턴, (그림 12)는 대표 프로파일의 지지도의 변화에 따른 시간 복잡도와 생성된 패턴의 수의 추이를 보여주고 있다.



(그림 10) 시간 지지도의 변화에 따른 패턴수와 시간 복잡도 (local = 4, profile = 4)



(그림 11) 공간 지지도의 변화에 따른 패턴수와 시간 복잡도 (time=5, profile=4)



(그림 12) 프로파일 지지도의 변화에 따른 패턴수와 시간 복잡도 (time=5, local=4)

위의 (그림 11)과 (그림 12)에서 공간 지지도와 프로파일 지지도가 8 이상일 경우 더 이상의 패턴이 생성되지 않았다. 이를 통해 지역 속성이나 대표 프로파일은 패턴 탐색에 있어 주요 변인 요인이 아니라는 것을 알 수 있다. (그림 11)에서 시간과 대표 프로파일이 최소지지도일 때, 공간 지지도가 8 이상일 때 더 이상의 패턴을 찾을 수 없었다. 이는 공간 속성과 관련된 패턴이 7개 지역이내에서만 발생함을 의미한다. 다시 말해 특정 시간대에서 발생하는 대표 프로파일은 최대 7개의 지역패턴을 가진다는 것을 알 수 있다. (그림 12)에서도 역시 시간과 공간 속성이 최소지지도일 때 대표 프로파일의 지지도는 8을 넘을 수 없었다.

하지만 (그림 10)에서는 공간과 대표 프로파일 속성이 최소지지도일 때, 시간 지지도에 따라 다양한 패턴을 찾을 수 있었다. 특정 지역에서 발생되는 대표 프로파일은 시간 지지도가 37일 때 더 이상 패턴을 찾을 수 없었다. 이는 특정 지역에서의 대표 프로파일은 최대 36개의 기본 시간패턴을 가지며 많은 시간대에서 같은 패턴이 반복된다는 것을 알 수 있다.

세 가지 실험을 통해서 알 수 있듯이 대용량 데이터 기반의 패턴 탐색임에도 불구하고, 세 가지 속성에 대한 빈발 조건을 만족하는 패턴을 비교적 짧은 시간 내에 찾았다. 이를 통해 큐브 마이닝 기법은 대용량 데이터의 패턴 탐색에 효율적임을 할 수 있다.

6.3 캘린더 패턴 기반의 주기성 탐사

이전 실험을 통해 공간 속성과 대표 프로파일 속성이 정해진 가운데 시간 속성의 변화에 따라 많은 패턴이 생성됨을 확인하였다. 이 실험에서는 공간과 대표 프로파일 속성에 많은 영향을 미치는 시간 속성에 어떠한 의미와 관계가 있는지 분석해 보았다.

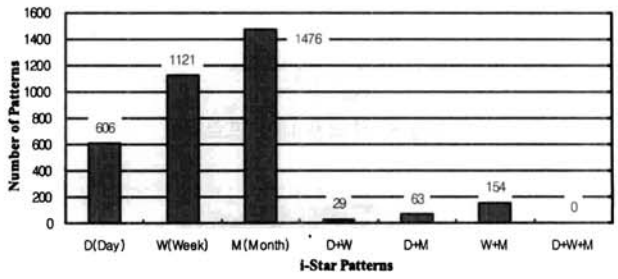
캘린더 패턴 갱신 알고리즘은 캘린더 스키마와 캘린더 패턴의 정의에 따라 구현하였다. <표 7>은 전력 소비 프로파일 데이터의 주기성을 탐사한 결과에 대한 해석이다. 예를 들어 지역 코드가 '363600'인 지역에서 대부분의 교육용 전력 사용자는 1월 매주 수요일에 대표 프로파일 코드가 'edu-0301'인 패턴을 형성한다. 또한, 일반용 전력 사용자는 매 달 마지막 주 금요일에 'gen-0101'인 패턴, 산업용 전력 사용자는 1월 매주 금요일에 'ind-0200', 주거용은 3월 3째주내내 'res-0303'인 패턴을 형성하였다.

(그림 13)은 최소지지도 (3, 4, 4)를 설정하였을 경우의 *i-star pattern*의 패턴수 이다.

캘린더 패턴 갱신을 통하여 대부분의 고객의 사용 패턴이 월 단위로 반복된다는 것을 확인 할 수 있다. 다시 말해서 대부분의 고압 전력 소비 고객들은 "매월 X번째주의 Y요일" 또는 "매월 매주 Y요일"과 같이 요일 단위의 주기성을 갖는 대표 프로파일이 많다는 것은 의미한다.

〈표 7〉 결과 해석

지역	계약종별	시간			대표 프로파일
		월	주	일	
363600	교육용	1	*	3	edu-0301
363600	일반용	*	4	5	gen-0101
363600	산업용	1	*	5	ind-0200
363600	주거용	3	3	*	res-0303
282403	일반용	*	3	6	gen-0403
282403	산업용	4	4	*	ind-0300
282403	주거용	1	3	*	res-0103
...	...	...	...	...	...



(그림 13) i-Star Patterns

7. 결 론

이 논문에서는 GIS-AMR 시스템에서 수집된 고압 전력 소비 고객의 데이터를 바탕으로 사용자에게 유용한 지식을 전달하기 위한 시공간 데이터마이닝 기법을 제안하였다. 첫째, 고객의 전력 소비 특성을 고려한 분할 군집화 기법을

적용하여 적절하고 효율적인 군집을 형성하였으며 실제 많은 고객들이 동일한 계약종별 내에서도 국부적인 유사성과 차이점을 확인할 수 있었다. 둘째, 3차원 큐브 마이닝 기법을 이용하여 시간과 공간, 대표 프로파일의 세 가지 속성에 대한 빈발패턴을 찾았다. 이 실험을 통해 우리는 공간과 대표 프로파일 간의 관계를 찾았으며, 시간 속성이 공간과 대표 프로파일 속성에 어떠한 영향을 미치는지를 규명하였다. 마지막으로, 캘린더 스키마와 캘린더 패턴식에 의한 주기성 탐사 기법을 통해 실제 특정 공간에서 발생되는 프로파일들이 갖는 일정한 주기성을 확인하였다. 이러한 주기성 탐색을 통해 실생활에서 데이터가 발생하기 이전에 어떠한 데이터가 생성되는지에 대한 예측 가능성을 보였다.

이 연구를 통해 우리는 GIS-AMR 시스템에 의해서 수집되는 데이터로부터 많은 지식을 찾았다. 이러한 시공간 데이터마이닝 기법을 통해 찾아낸 지식들은 실무자의 데이터 분석 및 의사 결정 지원 등에 많은 도움이 되며, 공간 정보 시스템에 적용하였을 경우 사용자에게 한 차원 높은 서비스를 제공할 것이다.

이 논문에서 보완해야 할 점은 탐색된 시간 주기성 검증을 위한 연구가 필요하다는 것이다. 또한 제안된 데이터마이닝 기법의 적용시점이 데이터 발생 시점을 기준으로 과거에 이뤄지기 때문에 과거의 지식이라는 범주를 크게 벗어날 수 없다. 따라서 향후 연구 방향은 입력되는 데이터를 바탕으로 3가지 기법이 스스로 진화해 나가는 통합 진화형 모델을 형성하는데 있다.

참 고 문 헌

[1] "부하분석모델 시공간데이터마이닝 기법적용 연구," 전력산업연구개발 보고서, 전력연구원, 2008.

[2] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," IEEE Transactions on Power Systems, Vol.18, No.1, pp.381-387, 2003.

[3] S. V. Verdu, "Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps," IEEE Transactions on Power Systems, Vol.21, No.4, pp.1672-1682, 2006.

[4] M. H. Pihao, J. H. Park, H. G. Lee, J. H. Shin, D. J. Chai, and K. H. Ryu, "Assessment of Temperature Sensitivity Analysis and Temperature Regression Model for Predicting Seasonal Bank Load Patterns," IEEE International Workshop on Semantic Computing and Applications, pp.81-84, 2008.

[5] A. A. El-Desouky, and M. M. Elkateb, "Hybrid adaptive techniques or electric-load forecast using ANN and ARIMA," IEE Proceedings Generation Transmission and Distribution, Vol.147, No.4, pp.213-217, 2000.

[6] D. C. Park, M. A. El-Sharkawi, R. J. Marks, L. E. Atlas, and M. J. Damborg, "Electric Load Forecasting Using an Artificial Neural Network," IEEE Transactions on Power Systems, Vol.6, pp.442-449, May, 1991.

[7] R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceedings of the 20nd international conference on Very Large Data Bases, pp.487-499, 1994.

[8] J. Han, J. Pei, and Y. Yiwen, "Mining Frequent Patterns Without Candidate Generation," Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp.1-12, 2000.

[9] L. Ji, K. L. Tan, and A. K. H. Tung, "Mining frequent closed cubes in 3D datasets," Proceedings of the 32nd international conference on Very Large Data Bases, pp.811-822, 2006.

[10] J. F. Roddick, and M. Spiliopoulou, "Temporal data mining: survey and issues," Research Report ACRC-99-007, University of South Australia, 1999.

[11] B. Ozden, S. Ramaswamy, and A. Silberschatz, "Cyclic association rules," 14th International Conference on Data Engineering, pp.412-421, 1998.

[12] Y. Li, P. Ning, X. S. Wang, and S. Jajodia, "Discovering calendar-based temporal association rules," Proceedings of the 8th International Symposium on Temporal Representation and Reasoning, pp.111-118, 2001.

[13] 박진형, 신진호, 박명호, 이헌규, 류근호, "전력 부하 패턴 분석을 위한 3차원 큐브 마이닝과 캘린더 패턴 기반 시간 데이터 마이닝," 제 29회 한국정보처리학회 춘계학술발표대회, Vol.15, No.1, pp.200-203, 2008.

[14] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," Microsoft Research Technical Report MSR-TR-98-14, 1998.

[15] M. H. Pihao, H. G. Lee, J. H. Park, and K. H. Ryu, "Application of Classification Methods for Forecasting Mid-Term Power Load Patterns," 4th International Conference on Intelligent Computing, Vol.15, pp.47-54, 2008.

[16] K. E. K. Saeed, M. H. Pihao, 이헌규, 신진호, 류근호, "전력배전 시스템에서의 취약 선로 분류를 위한 출현 패턴 마이닝," 제 30회 한국정보처리학회 추계학술발표대회, Vol.15, No.2, pp.325-327, 2008.

박 진 형

e-mail : neozean@cbnu.ac.kr

2008년 충북대학교 컴퓨터공학과(공학사)

2008년~현 재 충북대학교 전자계산학과 석사과정

관심분야 : 데이터베이스, 데이터마이닝 등

이 헌 규

e-mail : hg_lee@etri.re.kr
2002년 경기대학교 정보과학부(공학사)
2004년 충북대학교 전자계산학과(이학석사)
2004년~2006년 한국표준과학연구원 생활계측그룹 위촉연구원
2009년 충북대학교 전자계산학과(공학박사)
2009년~현 재 한국전자통신연구원 우정물류기술부 연구원
관심분야 : 데이터베이스, 데이터마이닝, 실시간 우편물류 운영기술 등

신 진 호

e-mail : jinho@kepco.co.kr
1996년 한밭대학교 전자계산학과(공학사)
2004년 충북대학교 전자계산학과(이학석사)
2004년 현재 충북대학교 전자계산학과 (박사과정 수료)
1996년~현 재 한국전력공사 전력연구원
관심분야 : 시공간 데이터마이닝, GIS, 모바일 등

류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr
1976년 숭실대학교 전산학과(이학사)
1980년 연세대학교 전산전공(공학석사)
1998년 연세대학교 전산전공(공학박사)
1976년~1986년 육군군수 지원사 전산실(ROTC 장교), 한국전자통신연구원(연구원), 한국방송통신대 전산학과(조교수) 근무
1989년~1991년 Univ. of Arizona Research Staff(TempIS 연구원, Temporal DB)
1986년~현 재 충북대학교 전기전자컴퓨터공학부 교수
관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 지식기반 정보검색 시스템, 유비쿼터스컴퓨팅 및 스트림 데이터 처리, 데이터마이닝, 데이터베이스 보안, 바이오인포매틱스