The Journal of GIS Association of Korea, Vol. 13, No. 3, pp. 239-252, November 2005

공간적 의사결정을 위한 공간 데이터 웨어하우스 설계 및 활용

박지만* · 황철수**

A Design and Practical Use of Spatial Data Warehouse for Spatial Decision Making

Ji-Man Park* · Chul-sue Hwang**

요 약

본 연구는 대용량의 공간자료에서 의미 있는 지식을 탐색 및 발견하고, 인터넷 환경에서 공간적 의사결정과정을 지원하기 위한 방안을 설계하였다. 그리고 설계를 기반으로 실험적 원형 시스템(Pilot Tested System)을 구현하여 의사결정자와 분석가 사이에서 특정주제에 적합한 다차원적 분석을 위해 양방향 통신이 가능하도록 공간 데이터 웨어하우스의 눈송이 스키마모델(snowflake schema model)을 활용하였다. 눈송이 스키마는 공간자료와 연계할 경우 기존 스타 스키마모델(star schema model)보다 확장성과 유연성이 뛰어난 분석중심의 설계방안이다. 또한 눈송이 스키마모델의 평가를 위해 백화점 거래사례를 통해 실험적 시스템을 구현하여 평가하였다. 이 시스템의 목적은 목표마케팅 지역설정이며, 의사결정자와 분석가의 양방향통신이 가능한 인터넷상 담시스템이다.

주요어 : 공간 데이터 웨어하우스, 시공간분석, 지식발견, 눈송이 스키마모델

ABSTRACT : The major reason that spatial data warehousing has attracted a great deal of attention in business GIS in recent years is due to the wide availability of huge amount of spatial data and the imminent need for turning such data into useful geographic information. Therefore, this research has been focused on designing and implementing the pilot tested system for spatial decision making. The purpose of the system is to predict targeted marketing

*경희대학교 대학원 박사과정(pjm754@khu.ac.kr)
**경희대학교 지리학과 조교수(hcs@khu.ac.kr)

area by discriminating the customers by using both transaction quantity and the number of customer using credit card in department store. Moreover, the pilot tested system of this research provides OLAP tools for interactive analysis of multidimensional data of geographically various granularities, which facilitate effective spatial data mining. Focused on the analysis methodology, the case study is aiming to use GIS and clustering for knowledge discovery. Especially, the importance of this study is in the use of snowflake schema model capabilities for GIS framework.

Keywords : spatial data warehouse, spatio-temporal analysis, decision making, snowflake schema model

1. 서 론

최근 인터넷 환경과 자료의 증가는 수치화된 지리정보의 폭발적인 증가를 가져왔다. 이러한 자료의 대량화에 따라 토지이용, 범죄, 의료, 마케팅 전략 등 공공부문과 민간부문에서는 매우 상세한 수준의 지리학적 해결책을 요구하고 있다.

의사결정을 위한 지리학적 지식발견 방법론은 공간 자료를 효과적으로 표현하고 질의할 수 있는 저장소가 필수적이다. 그러나 이와 같이 공간 자료를 표현하고 저장 및 검색이 가능한 공간 데이터베이스(spatial database)는 실세계의 여러 차원을 고려하지 않았고 정보 분석을 공간적(spatial)인 것에 한정하였다(Miller, 1999). 실세계를 고려한 저장소는 다차원적(multid- imensional) 정보 분석을 요구하고, 모델의 구현, 처리방식, 그리고 처리과정은 정보 분석의 주제마다 다르다. 따라서 지식발견을 위한 유연한 분석방법론은 기본적으로 대용량의 저장 공간과 함께 저장소내의 스키마 모델연구가 필요하게 되었다.

분석을 위한 지리학적 지식발견은 관심 있는 지역의 인구, 산업 특성과 공간적 특징을 고려하여 이루어지는데, 그 특징은 공간적 자기상관성(spatial autocorrelation)과 공간적 이질성(spatial heterogeneity)이라는 특성을 가지며, 시-공간적(spatio- temporal) 관련성은 더욱 복잡하게 연관되어 있다. 그러므로 각 변수에 따른 관계, 방향, 연결성은 각각의 차원에 따라 다르고 다양한 지리학적 차원에서 주요한 지식발견을 위해 시간, 공간적으로 추출해야 한다(Miller and Han, 2001). 이러한 맥락에서 기존 공간 데이터 웨어하우스의 성형스키마(star schema)는 지리학적 지식발견을 위한 유연한 방법론에 한계를 갖는다. 이러한 유연한 분석방법론을 위한 공간 데이터 웨어하우스 스키마 연구와 대비하여 지식발견에 관련한 연구는 매우 드물지만 몇몇 연구에서 주안점을 찾을 수 있었다.

Kamber(2001)는 전자회사의 고객관리를 위한 마케팅 전략에 눈송이 스키마(snowflake schema)를 고려하였다. 반면에,

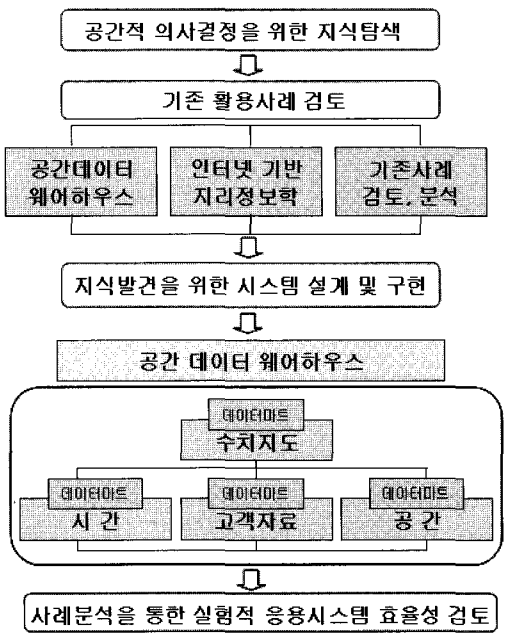
Murray와 Estivill-Castro(1998)는 대용량 공간저장소에서 지식발견을 위한 효율적인 방법론을 논의하였다. 또한 Miller와 Yuan(2003)은 공간저장소에서 지식발견을 위한 자료정제와 연관규칙을 기반으로 알고리듬연구를 진행하였고, Tung(2002)은 공간군집을 통한 지식발견 기법을 제안하였다.

이 연구는 다차원적인 공간 데이터 웨어하우스를 설계, 구현하는 것을 목적으로 하였다. 또한 사례분석을 통해 시공간적(spatio -temporal) 특징을 고려하여 실험적으로 구현된 공간 데이터 웨어하우스를 활용하였다. 그리고 고객의 인구속성, 구매 패턴 등을 분석하고, 서울시를 대상으로 목표 마케팅(target marketing) 지역을 예측하였다. 또한 이 예측된 결과로 서울시의 공간 및 비공간 자료들 사이에서 각기 다른 차원과 대비하여 데이터마이닝 기법 중 군집기법과 GIS의 공간분석 기법으로 시각화하였다.

2. 연구 내용 및 방법

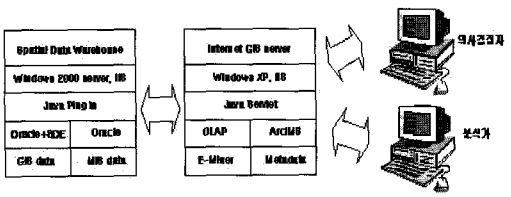
본 연구에서는 분석가의 의사결정을 지원하는 공간 데이터 웨어하우징 기법과 사례연구로서 공간 데이터 웨어하우스 눈송이 스키마 설계 및 구현, 데이터마이닝 기법 중 군집기법, GIS시각화 방법을 통해 얻은 지식을 기반으로 실제 서울시의 목표 마케팅 지역을 예측하기 위해 다음과 같은 절차에 따라 연구를 수행하였다.

첫째, 공간데이터의 가용성과 호환성을 가지고 공간 특성을 고려한 분석가의 공간분석을 위해 공간 데이터 웨어하우스의 각 단계별 프로세스를 설계하였다. 이를 통해 목표 마케팅 지역 예측을 위한 공간 데이터 웨어하우스를 구현하여 시스템의 개념적 기틀을 제공하였다. 둘째, 구현된 시스템을 통해 데이터 마이닝 기법 중 군집분석을 통해 의사결정시 요구되는 지식을 발견하도록 온라인 분석처리기법(OLAP : online analytical processing)을 이용하여 다차원적인 접근을 시도하였다. 이 연구는 주제에 따라 자료를 분류하고 실험적 시스템을 구현하는 단계와 효율적이고 합리적인 의사결정과정을 가능하게 하는 지식발견 단계로 구분할 수 있다.



[그림 1] 연구의 흐름도

전자의 단계에서는 특정 주제에 맞는 자료를 선별하여 각 차원에 따라 분류하고 정제하여 데이터마트를 설계하였다. 그 이후 개별적인 각각의 데이터마트는 하나의 공간 데이터 웨어하우스의 요소가 되며 하나의 시스템으로 구성된다. 각 데이터 마트에는 서울시 백화점 고객 자료를 공간적 차원으로 동별, 구별로 분류하고 시간적 차원으로는 월, 분기, 년도별로 구분하여 시계열 분석이 용이하고 분석에 있어 요약, 세분화가 가능하도록 하였다. 후자의 단계에서는 구현된 공간 데이터 웨어하우스에서 데이터 마이닝의 탐색적 자료 기법 중 하나인 군집분석의 k-평균 군집방법(k-means clustering) , 계층적 최근린 군집기법(nearest neighbor hierarchical clustering)을 활용하여 유사한 특성을 가진 몇몇의 집단으로 그룹화하여, 각 군집의 특성을 파악하고 지식과 패턴을 분석하여 서울시에서 목표 마케팅지역을 예측하는데 활용했다.



[그림 2] 실험적 시스템 흐름도

연구 대상으로는 1개월 시간간격의 2000, 2001년 백화점 고객데이터 170만건 중 카드로 물품을 구매한 고객 41,263명의 신상자료를 대상으로 하고 예측지역으로는 서울시로 국한하였다.

3. 실험적 시스템의 설계와 구현

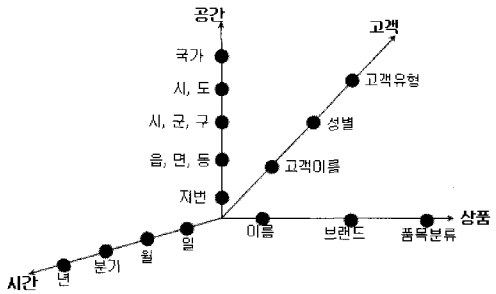
3.1 실험적 시스템 개념

공간 데이터 웨어하우스는 공간, 비공간 자료가 적용되고 점, 선, 면, grid-cell 등의 공간 객체를 저장, 색인하는 방법을 지원한다. 더불어 공간 객체(object)는 실세계의 지리학적 형상을 갖고 위상관계, 측정단위, 공간 해상도를 포함하며, 일반적인 데이터 웨어하우스와 유사하게 통합된, 주제 지향적이며, 시간성을 갖고, 비휘발성의 공간, 비공간 자료를 보유하는 저장소라고 정의한다. 공간 데이터 웨어하우스로 다차원 접근을 통한 공간 데이터는 기하학적인 사상을 갖고 측정을 하며 온라인 분석처리기법을 이용하여 다차원 분석을 수행한다. 다차원 모델링은 시간, 공간을 포함한 여러 차원을 사용하여 분석한다. 특히, 분석방법은 주제에 따라서 차원을 세분(drill-down)하거나, 요약(drill-up)할 수 있다.

[그림 3]은 중심으로부터 성형으로 라인이 뻗어 있다. 하나의 선은 각 차원을 뜻하며 선에 있는 점은 계층별 수준을 표현한다.

공간차원은 가장 세분화된 수준인 지번에서부터 국가수준으로 갈수록 일반화되며, 시간차원은 일, 월, 분기, 년별로 구분하여 계층구조에 따라 설계되었다. 각 수준을 압축하고 있는 점은 해당수준의 관련속성과 연결되어 있다. 그래서 특정 품목에 대한 고객들의 선호 품목, 선호 브랜드별 공간적 군집 패턴을 시계열적

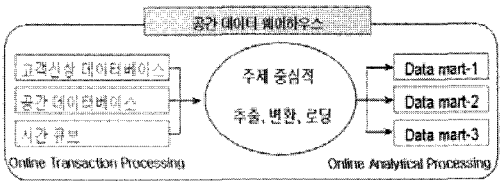
으로 분석할 수 있을 뿐 아니라 각 차원별로 세분화, 일반화하여 분석 목적에 따른 차원변경이 가능하다.



[그림 3] 다차원적 모델
(multi-dimensional model)

본 연구의 웹 기반 실험적 시스템은 공간 데이터 웨어하우스, 공간적 의사결정을 위한 지식 탐색 및 발견, 인터넷 지도 서비스로 구성되며 각 시스템의 구성요소는 상호 결합되어 있다.

공간 데이터 웨어하우스의 다차원 분석 패러다임 측정방법에는 첫째, 비기하학적 공간 차원(non-geometric spatial dimension)으로서 명목척도(nominal scale)에 해당한다(Guo, 2002). 이 차원은 행정경계 관리의 주기처럼 공간의 수치적인 성질이 요구되지 않으며 지도학의 공간 사상 구분에 해당한다. 둘째, 기하학-비기하학적 공간차원(geometric -to-non-geometric spatial dimension)으로 점, 선, 면의 공간 객체를 사용하여 사상으로 표현하고 그에 관련된 속성 값이 해당된다. 사상은 기하학적 X, Y좌표체계를 가지고 있으며 그 속성 값은 비 기하학적 성질을 가진다. 셋째, 기하학적 공간차원(fully geometric spatial dimension)이다. 이 타입은 원 자료가 기하학적 성질을 갖는다. 예를 들어 벡터타입의 면자료로 생성된 지번자료와 그 자료를 일반화시 거리는 같다. 공간의사결정을 위한 실험적 시스템에서 공간적 위치와 위상관계를 갖는 공간자료는 [그림 4]처럼 OLTP 방식의 운영시스템인 공간데이터베이스에 저장된다. OLAP 시스템에 필요한 자료는 공간객체, 시간, 공간큐브 등 OLTP 방식의 운영시스템으로부터 추출되어, 정제 및 변화과정을 거쳐 데이터마트에 저장된다.



[그림 4] 공간 데이터 웨어하우스 구성도

데이터마트를 구성함으로 관리자는 공간 데이터베이스의 원 자료를 보호하고 분석가에게 빠른 응답성능을 제공하며, 공간 자료에 거래자료, 고객신상자료가 통합된 양질의 자료를 제공한다. 또한 각 차원별 공간 자료집합체에 자료를 요약, 일반화된 형태로 논리적 계층수준을 두어 분할함으로써 분석가에게 공간, 시간별 세분화된 거래량과 상품품목을 분석하고 시-공간 차별적인 목표 마케팅 계획이 가능하였다.

3.2 요구사항분석

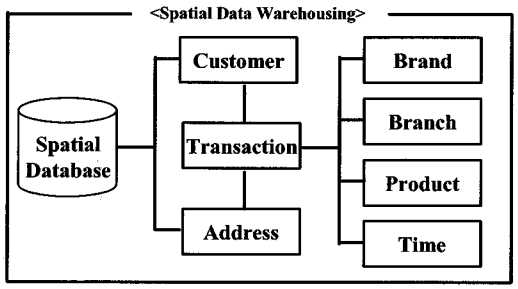
본 연구는 공간 데이터 웨어하우징을 활용한 사례로 시스템 설계 및 구축 계획을 수립하는데 있어 다음과 같은 요구

사항을 설정하고 연구를 진행하였다.

첫째, 본 연구에서 실험적 시스템의 구현은 공간 데이터 웨어하우징의 관리적, 분석적 측면을 고려한 활용을 목표로 한다. 둘째, 자료의 범위는 A백화점의 카드 고객 거래 170만 건 중 서울시 4개 지점에서 카드로 물품을 구매하고 2000년 5월부터 2001년 4월 기간에 거래한 서울시 거주 고객 41,263명의 신상 자료를 대상으로 했다. 셋째, 본 연구에서 고려한 실험적 시스템의 주 대상은 의사결정을 하는 분석가로 한정하고 시스템의 보안과 인터페이스의 설계는 이 연구에서 배제하였다. 넷째, 본 연구는 웹 환경을 통한 분석가의 공간적 의사결정을 지원한다. 그러므로 지리적 사상과 속성정보는 주제에 따라 다양한 표현법과 분류법이 요구된다. 즉, 공간분석의 수요에 맞도록 시스템은 반복되는 분류, 분석을 위해 초기 자료에는 손상이 없어야 하고, 정확하지 않는 자료와 시간에 따라 변형된 지리적 사상과 속성 변화는 수정, 편집할 수 있는 기능을 삽입하였다. 다섯째, 본 연구에서 사용하는 수치화된 공간 자료와 비공간 자료는 연구지역인 서울시의 실세계를 추상화(Abstraction), 단순화(Simplification) 한 것으로서, 자료의 내용과 형식, 용어와 개념을 정리한 메타데이터 문서를 제공한다. 이 문서는 미국 FGDC의 "디지털 지리 공간 자료에 대한 메타데이터 표준"에 따라 작성했다.

3.2.1 공간 데이터 웨어하우스 구성 방안

공간 데이터 웨어하우스의 벡터타입 점 자료는 고객의 위치를 표현하고, 도로와 철도는 선 자료, 행정동과 구별 행정구역은 면 자료로 구성하였다. 또한 래스터타입의 영상자료는 전 처리과정을 거쳐 공간데이터베이스에 저장하였다. 그리고 비공간 자료인 고객신상자료와 거래자료는 각각의 데이터베이스에서 추출 및 변화과정을 거쳐 정제되어 구성된다. 지도큐브는 공간 데이터 웨어하우스를 구성하고, 결합구조로 연결하였다. 분석가는 고객신상, 거래 데이터마트와 지도큐브에 구조화된 질의 언어를 통해 개별 접근이 가능하고, 공간 데이터 웨어하우스에 온라인 분석기법(OLAP)을 통해 다차원적 접근을 시도하였다.



[그림 5] 공간 데이터 웨어하우스의 지도큐브

3.2.2 지식탐색 및 발견을 위한 방안

실험적 시스템에서 목표 마케팅 지역 설정을 위한 지식은 카드고객이 어디에 밀집해 있으며 어느 정도를 구매하였는가? 또한 그들이 구매행위가 주로 발생하는 시간과 선호하는 브랜드 및 품목은 무엇인가? 라는 질문에 대한 해답일 것이다. 이 질의에 대한 해결책으로 공간과 시간, 선호하는 품목들에 대한 고객을 세

분화하고 지식을 탐색하여 목표 마케팅 지역을 설정하기 위한 시장과 고객을 세분화하고 거래가 빈번히 발생한 시간을 발견한다. 발견된 지식은 공간적 특징을 갖는 정보로서 특정지역에 마케팅 노력을 집중할 수 있다.

여기서 고객의 위치와 특정 차원에 따른 공간정보를 분석하여 수치지도로 목표지역을 시각화하였다. 그리고 의사결정자와 분석가 사이의 통신 수단을 만들어 상담할 수 있는 기능을 삽입했다. 지도서비스는 특정 차원에 따른 지식을 발견하고 분석가에게 의뢰하여 구매사건이 발생하는 각기 다른 차원들과 대비한 정보를 시각화한다. 여기에는 GIS의 공간분석, 데이터 마이닝의 군집분석이 활용되었다.

기존의 지식발견을 위한 공간통계기법은 지식 탐색을 위한 결과도출의 과정이나 결과 자체의 관점에서 데이터 마이닝과 비슷하다. 그러나 목표 마케팅지역을 설정하는데 공간통계기법이 거래행위의 각 시간 및 공간 등 각 차원별 추측을 하고 검토하는 것이라면, 데이터마이닝은 추측과 검정과정을 공간 데이터 웨어하우스에 접근하여 다차원 모델을 생성하고, 자동으로 공간자료와 비공간자료들을 상호 연계하여 지식을 탐색 발견하는 차이를 갖는다. 즉, 이 연구에서는 기존의 확정적 분석방법의 공간통계기법을 배제하고, 자동화된 마이닝 도구를 사용하여 공간 데이터 웨어하우스의 정제된 자료에서 카드구매에 따른 시공간적, 품목별 의미 있는 패턴이나, 상호연관성을 발견한다. 그리고 발견된 지식은 고객을 세분화하여 차별적인 고객관리 및 마케팅 전략수립을 가능하게 한다.

3.2.3 인터넷 기반 상담 시스템

웹 기반 상담 시스템은 의뢰자와 상담자간의 의사교환 및 일련의 과정을 웹 환경에서 도와주거나 대행하는 일이다. 즉, 직접대면을 통한 상담이 아닌 웹 기술을 기반으로 웹 환경에서 상담이 이루어진다. 본 연구에서 다루게 되는 실험적 시스템은 웹 환경에서 실시간으로 지리자료를 활용하여 지식을 발견하고 의사결정과정을 지원하는 시스템을 구현하였다. 최근 민간 기업에서 지리정보에 대한 관심이 높아지고 있는데, 그 이유는 고객에게 차별적인 서비스를 제공하기 위한 정보의 상당수가 지리정보이기 때문이다. 예를 들어 마케팅 전략가는 소매지점이 포괄하는 상권과 구매 고객들의 밀집지역을 공간적으로 세분화하기 원한다. 즉, 고객 개개인의 개성으로 자신에 맞는 특화된 상품과 서비스를 원하는데, 이러한 수요를 원하는 고객들은 특정 시간대에 지역적 집적의 형태로 공간상에 나타난다.

집적된 지역의 고객들을 확보하고 유지하는 일은 마케팅 전략에서 지리정보의 공유, 체계적인 관리, 다양한 활용이 중요시된다. 본 연구에서 구현한 인터넷을 활용한 실험적 시스템은 분석가와 의사결정자의 능동적인 대화가 가능하며, 분석에 필요한 많은 자료를 체계화하고, 자료관리 기능과 공간 분석을 포함한 여러 다차원적 분석을 수행한다.

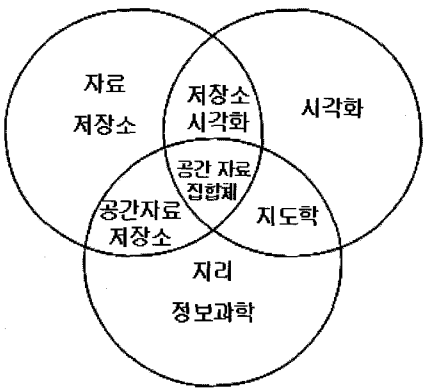
3.3 시스템 설계

공간 데이터 웨어하우스를 포함한 응용 시스템의 설계는 사용자 요구사항에서 도출된 개념을 통합하는 논리적 설계와 어떻게 논리적 설계를 특정 시스템에 연계하는지 결정하는 물리적 설계로 구분하였다.

3.3.1 논리적 설계

실험적 시스템은 데이터 웨어하우스, 지리정보학, 정보 시각화의 개념적 영역을 가진다[그림 6]. 각 영역은 결합관계를 이루고 있는데 데이터 웨어하우스와 지리정보과학의 개념적 결합으로 공간 데이터 웨어하우스, 지리정보과학과 정보 시각화 영역에서 지도학의 공간 자료 시각화, 데이터 웨어하우스와 정보 시각화 영역으로 데이터 웨어하우스의 스키마 모델을 시각화하였다.

마지막으로 이 세 가지 영역의 결합은 공간 자료 집합체로 개념적 모형을 설명하였다.

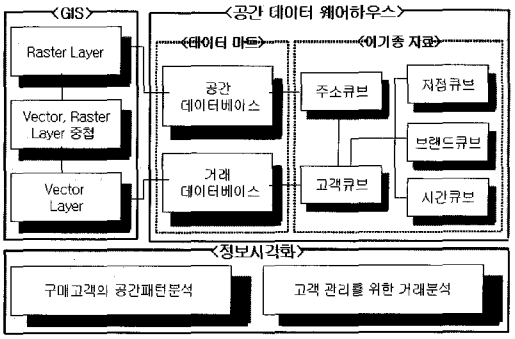


[그림 6] 실험적 시스템의 개념적 영역

첫째, 공간 데이터 웨어하우스의 영역에서 데이터 웨어하우스는 서울시통계자료가 담긴 이기종(heterogenous)의 데이터베이스, 지도큐브의 서울시 수치지도와 영상, 카드 거래 자료를 갖는 데이터 마트와 1:N의 형식으로 연결된다. 그리고 데이터 마트는 고객의 신상 자료가 있는 고객큐브와, 거래 시간 자료가 있는 시간큐브와 1:1로 연결된다. 그리고 지리적 사상의 시간적인 변동을 담기 위해 지도큐브와 연결된다.

둘째, 지리정보과학 영역에서 벡터 레이어는 공간 객체인 점, 선, 면과 1:N으로 연결되고, 각각의 객체가 담고 있는 사상은 고객위치, 주요도로와 철도, 행정도와 수계 등을 담고 있다. 이때 연구지역의 영상인 래스터 레이어는 벡터 레이어와 중첩하기 위해 지도큐브와 1:N으로 연결한다. 최종적으로 지리정보과학의 레이어는 데이터 웨어하우스와 N:M형식으로 연결된다.

셋째, 지리정보과학에서 수행하는 공간분석, 레이어 중첩, 가시화는 지도의 구성요소인 주기, 심볼과 적절한 지도학의 시각화 기법을 사용한다[그림 7].



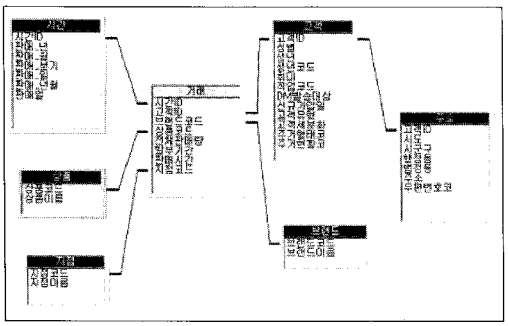
[그림 7] 실험적 시스템 논리적 체계

3.3.2 물리적 설계

실험적 시스템의 물리적 설계는 공간 자료 저장소와 비공간 자료인 백화점 자료로 구분하여 설계했다.

백화점 자료의 물리적 설계는 백화점 거래량을 사실 테이블로 데이터 마트를 설계하였다[그림 8]. 이 데이터 마트는 시간, 고객, 브랜드, 상품, 지점을 차원 테이블로 구성하였다. 각 차원의 시간ID, 고객ID, 브랜드 코드, 상품코드, 지점코드를 기본키(primary key)로 각 차원들과 다차원 테이블 조인을 수행하도록 구현했으며, 고객 차원은 고객ID 키를 사용하여 주소 차원과 조인이 가능하도록 눈송이 스키마모델(snow -flake schema model)로 설계했다. 그리고 주소차원의 고객ID는 공간 데이터베이스와 연결하여 지리정보과학에서 제공하는 공간 분석을 수행한다.

```
define cube trans_snowflake [time, customer, brand, produce, branch]:
        TotalAmount=sum(trans_total), installment_mon=count
define dimension time as (time_id, dat, week, month, quarter, year)
define dimension product as (product_cd, name, product_type)
define dimension branch as (branch_cd, name)
define dimension brand as (brand_cd, name)
define dimension customer as (customer_id, sex, birthday, hobby, job,
        dm, customer_type, resident_type,
    address (address, country, city, gu, dong))
```



[그림 8] snowflake schema model

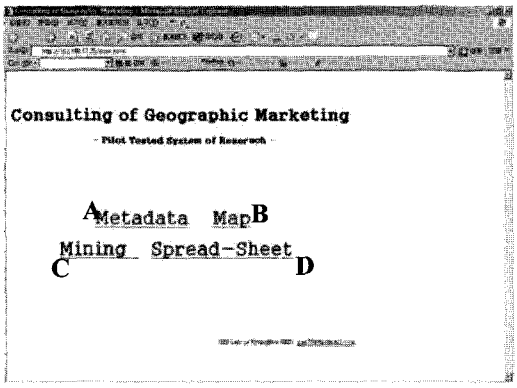
본 연구에서 사실 테이블의 고객ID는 고객 신상자료의 질의과정에서 웹 기반 지도서비스에서 공간객체인 점 사상으로 표현하기 위해 주소큐브와 연결했다. 이 지도큐브는 공간 데이터베이스의 공간객체들이 저장되었다.

이러한 구조는 고객 신상자료가 저장된 데이터마트와 공간 데이터베이스의 점 자료가 지도에서 사상으로 나타내기 위해 1:1구조 cartesian products로 순서쌍을 이루어야한다. 이 모델은 스타 스키마모델(star schema model)을 보완한 모델로 저장 공간을 최소화하고 유연성을 개선한 것이다. 그러나 성형 스키마모델보다 많은 결합(join)연산으로 인해 상대적인 느린 응답을 하는 단점을 가지고 있다.

3.4 인터넷 기반의 시스템 구현방안

본 연구에서 분석가는 구현한 실험적 시스템에 웹 환경에서 접근한다. 시스템의 H/W 및 S/W는 크게 세 가지로 분류된다. 공간 데이터 웨어하우스는 공간 자료와 비공간 자료를 저장, 관리하는 저장소로의 역할을 하며, 웹 환경 지도서버는 저장된 자료를 질의, 조회, 처리, 분석의 시각화기능을 수행한다. 그리고 서버에는 Servlet 엔진을 지원함으로써 마케팅 의사결정자의 클라이언트 접근, 시각화하는 채널을 제공한다. 마지막으로 다차원 분석을 위한 OLAP엔진과 마이닝 도구는 분석가 클라이언트의 소프트웨어에서 서버모듈로 연결하여 저장되며 의사결정자에게 다차원분석과 데이터 마이닝 결과

를 시각화하여 의사소통한다. 의사결정자의 요구에 따른 조회는 OLAP 엔진에서 구조화된 질의 언어로 변경되고 공간 데이터 웨어하우스에 전달된다. 전달된 요구에 의하여 분석가의 분석결과는 Arc-IMS의 서버모듈에서 Sevlet 엔진인 Jakarta-Tomcat에 전달되어 의사결정자에게 공간적 위치와 사상을 제공한다. 또한 경영정보시스템(MIS: management information system)의 거래 및 고객자료가 저장된 데이터 마트는 벡터 레이어의 점 사상 공간객체에 각 차원별 기본키(primary key)로 연결되어 요구에 따라 차원별 세분 및 요약되어 분석된다.



[그림 9] 실험적 시스템

[그림 9]는 웹 환경에서 구현된 실험적 시스템에 접근하기 위해 사용자 첫 화면을 구성하였다. 'A'에 해당하는 부분은 메타데이터와 연결되며 'B'는 분석된 자료를 공간 자료의 형태로 표현하기 위한 수치지도로 연결된다. 또한 'C'는 비공간 자료의 분석을 위한 데이터마이닝 결과 페이지로 연결되고 'D'는 비공간자료를 테이블 형식으로 제공한다.

4. 사례분석

4.1 전역적 공간분포패턴

데이터 웨어하우스에 저장된 고객 신상정보는 구, 행정동과 같은 공간단위를 기준으로 저장되어 있다. 이 고객의 위치와 관련된 속성정보를 활용하여 분석하기 위해 고객ID와 기본키(primary key)로 연결된 공간 데이터베이스에 저장된 공간 객체를 사용하였다.

전체 고객의 방향성은 북서에서 남동방향으로 밀집해 있고 신사동이 고객 분포의 중심지로 표현되었다. 전체 판매량에 최우수 1등급 고객의 분포형태는 장방향 축의 길이가 전체고객이 축보다 길고 단방향 축이 짧은 형태가 나타났다. 이러한 결과는 최우수 고객1등급의 고객이 동서 방향으로 더 길게 분포 했음을 알 수 있다. 또한 'tan'의 수치와 회전 값으로 남서 방향으로 치우쳐 분포하고 있다<표 1>, [그림 10].

<표 1> 고객세분화에 따른 표준편차 타원체

	고객전체	최우수고객 1등급
가중치 중심점	신사동	압구정 1동
tan	-4.89458	-6.26672
회전	11.547	9.06643
장반경 길이	7425.6m	7627.54m
단반경 길이	3889.7m	3250.82m

서울시 522개동에 고객수의 최대값은 1246명으로 잠원동이며 한명의 고객도 없

는 동은 삼선2동, 월곡4동, 동선2동, 가리봉2동, 강일동이다. 각 동에 대한 평균 고객수는 79명이고 표준편차는 121이다. 즉, 최대의 고객이 밀집한 지역은 잠원동, 압구정1동 등 주로 강남구와 서초구에 고객들이며 밀집해 있다.

지도 위에 있는 팝업창은 수치지도에 관련하여 기술한 설명이다. 이 기능은 지식을 발견하기 위한 의사결정자와 분석가 사이에서 양방향 통신을 수행한다. 고객의 구매량에 따른 공간적 분포의 특성을 밝히기 위해서 점 분포의 중심경향을 살펴보았다. 총 구매량에 가중치를 적용하여 생성한 표준거리(1)는 중심점을 중심으로 반경 8523.45m이다[그림 10].

표준거리 내에 고객들은 구매량에 따른 분포를 표현한다.

$$SD = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - x_{mc})^2 + \sum_{i=1}^{n} f_i (y_i - y_{mc})^2}{\sum_{i=1}^{n} f_i}}$$

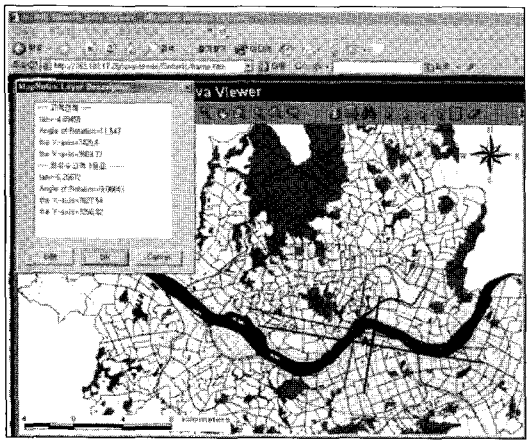
SD = 가중화된 표준거리 …(1)

f_i= 전체 구매량에 따른 가중치

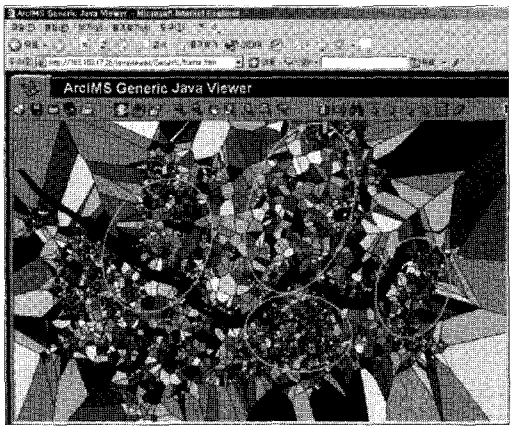
확률지역(probability area)과 표준타원체는(standard deviation ellipsoid) 연구지역내의 점 사상의 공간적 밀집을 효과적으로 시각화한다. 백화점 카드고객의 95%가 서울시의 동서방향으로 늘어진 분포패턴을 보였다.

4.2 공간적 세분화

[그림 11]은 고객의 전체 구매량에 따른 군집패턴의 결과로서 티센 다각형과 k-평균 군집분석을 행하였다. 그리고 티센 다각형으로부터 얻은 면 자료를 이용하여 전체 구매량 10%의 고객의 분포 패턴을 조사한 결과 불규칙한 형태로 흩어져 있었다. 군집의 계층을 파악하기 위해 최근린 계층군집을 수행하였다. 계층적 군집분석은 작은 군집부터 큰 군집으로 계층을 생성해간다. 이러한 원인은 군집분석이 감독학습(supervised learning)으로 절대적 기준 모델이 없기 때문이다. 그리고 계층적 군집을 이용하면 바람직한 군집의 수를 취할 수 있다.

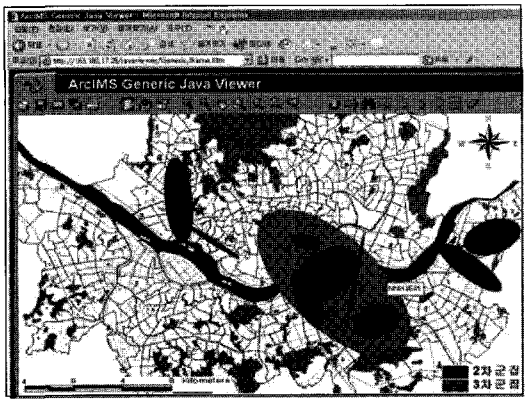


[그림 10] 고객의 공간적 분포패턴



[그림 11] 티센다각형과 군집

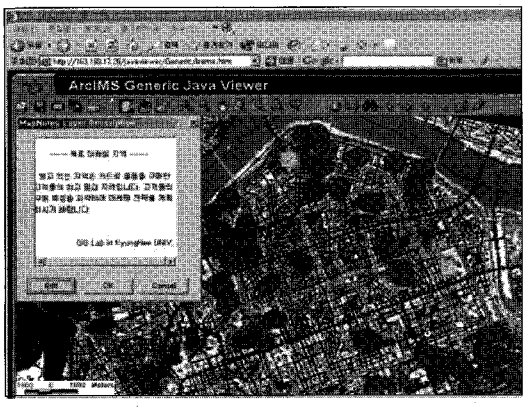
이 연구에서는 공간적 근접성과 구매량, 구매회수에 3번에 걸친 최 근린 계층 군집분석과 k-평균 군집결과 강남구와 서초구에 고밀도로 밀집해 있음을 확인 할 수 있었다[그림 12].



[그림 12] 계층적 군집분석

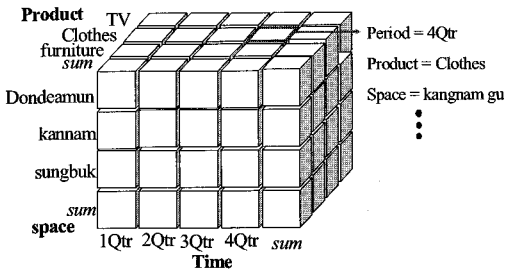
4.3 OLAP을 활용한 다차원적 분석

[그림 13]은 앞서 분석한 공간적 세분화의 결과로 우량고객이 밀집된 지역이다.



[그림 13] 목표마케팅지역

목표마케팅지역을 의사결정을 위한 분석가는 우수고객의 선호하는 구매시간, 브랜드, 그리고 상품 항목을 발견하여 마케팅에 활용한다. 공간적으로 강남, 서초구의 OLAP기법을 활용하여 공간차원의 동별, 시간차원의 2000년 월별, 상품과 거래량자료를 활용하여 고객을 세분화하였다([그림 14]).




```
select AL2.dong, count( AL2.dong), AL5.product_name, AL3.TotalAount
from customerAL1, addressAL2, transactionAL3, timeAL4, productAL5
where (AL1.customer_ID=AL3.customer_ID and AL5.product_CD=
      AL3.product_CD and AL4.time_ID=AL3.time_ID and AL2.
      customer_ID=AL1.customer_ID) and (AL2.gu='kangnam' or
      AL 2.gu='seocho' and AL1.sex=2 and AL4.TotalAmount_
      month='12' and AL1.customer_type='A' and AL1.customer_
      segmention=1)
group by AL2.dong, AL5.produduct_name, AL3.TotalAmount
order by 4 desc, 2
```

[그림 14] OLAP을 위한 백화점 자료큐브 모형

구매 패턴을 세분화하여 구매한 상품 정보를 발견하였는데, 주로 유명 디자이너 의류품목과 수입명품이 높은 판매실적을 기록했다. 이러한 초우량 1등급 고객은 고객 타입을 세분화한 분류에서 최상의 판매실적을 기록한 고객들이며 전체 판매량의 10%를 구매한 고객들이다. 또한 시간차원에서 분기별로는 4분기<표 2>, 차원을 세분화하여 일별 거래량은 1-10일 사이에 가장 많은 구매현상이 발생하고, 시간에 따른 거래량은 15시에서

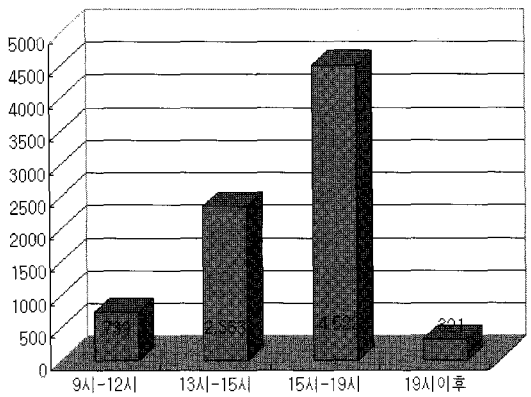
19시 사이에 가장 활발한 구매현상이 발생한다[그림 15]. 또한 각 차원별 세부적 정보를 요약, 세분화, 그리고 차원변경이 가능하여 공간적 의사결정을 위한 속성 정보에 대비하여 능동적이며 유연한 분석이 가능하였다<표 3>.

<표 2> 분기별 판매실적

분기	거래수	전체판매량(원)
1분기	3,514	401,895,620
2분기	4,098	404,664,670
3분기	3,400	396,928,830
4분기	3,963	533,390,540

<표 3> OLAP 수행의 결과

브랜드 명	상품종류	지점	전체판매량
Gucci	imports	압구정점	960,000
NIKE	sports	신촌점	915,000
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.



[그림 15] 시간대별 거래 고객 수

5. 결 론

이 연구는 인터넷 환경에서의 상호 운용성을 통해 대용량의 지리공간자료를 활용하여 효과적인 공간적 의사결정을 위한 지식발견에 초점을 맞추었다. 또한 실세계에서 발생하는 현상을 더욱 사실적으로 분석하고 표현하기위해 지리정보과학의 지리정보시스템과 고객관계관리의 요소를 삽입하여 공간 데이터 웨어하우스를 설계 및 구현하였다.

본 연구에서는 인터넷 환경에서 공간자료를 활용하기 위한 방법의 유형을 다음의 세 가지로 구분하고 그 특징과 한계를 분석하여 대안을 제시하였다. 첫째, 기존의 인터넷 기반 지리정보서비스는 의사결정을 위해 공간자료를 활용하여 분석하는 측면이 취약하다. 둘째, 공간자료와 해당속성자료를 제공하는 데이터베이스의 물리적 스키마설계는 실세계의 개념적 모형을 설명하기 불충분하다. 셋째, 최근 민간부문은 다양하고 급격하게 증가된 고객욕구를 지리자료에서 공간적 측면의 해결책이 요구되고 있다. 그러나 다양하고 세분화된 고객의 욕구를 수치화된 공간자료만으로 해결하기에는 한계를 갖고 있다.

연구에서는 민간기업의 활용성에 초점을 맞추어 실제 서울시 백화점 카드거래 자료를 활용하여 인터넷 환경에서 공간 데이터 웨어하우스의 다차원적 모델링을 구현하고, 기존 활용사례의 노출된 한계를 극복하기 위해 실험적 시스템의 설계 방향을 설정하였다. 첫째, 네트워크에서

전송되는 공간자료의 초기자료를 벡터타입의 공간객체로 표현 및 저장하며 관련 속성자료를 생성하여 연결 및 저장하였다. 둘째, 시공간적 차원과 더불어 특정 주제에 맞는 차원을 고려하기 위해 각 차원별 데이터베이스에서 추출 및 변화 등의 정제과정을 거쳐 데이터 마트를 구성하고, 눈송이 스키마모델을 사용하여 저장 공간을 최소화 및 유연성을 극대화 하였다. 셋째, 분석가는 OLAP기법을 통하여 공간 데이터 웨어하우스에 다차원적 접근을 시도하며, 각 차원별 계층적 수준을 두어 공간적 지식발견을 위해 의사결정자의 정의에 따른 차원변경, 세분화, 요약화가 가능하도록 구현하였다.

공간 데이터 웨어하우스는 특정한 목적을 갖고 설계되므로 실세계에서 발생한 관련된 차원들을 더욱 구체화하고, 시스템에 적용해야한다. 그리고 시공간적 특성을 명확하게 분석하여, 해당하는 측정방법과 의미 있는 지식을 발견할 수 있는 방법론에 기여하게 될 것이다.

참고문헌

Bedard, Y., 1999, "Visualization modelling of spatial databases towards spatial extensions and uml", Geomatica Tutorial

Guo, D., 2002, "Spatial Cluster Ordering and Encoding for High-Dimensional Geographic Knowledge Discovery", *GeoVISTA Center and Department of Geography, Pennsylvania State University*, pp.5-13.

Han, J., 1997, "Spatial Data Mining and Spatial Data Warehousing", Symposium on Spatial Databases(SSD97) Conference Tutorial.

Harms, S.L.D., and Tadesse, T., "Efficient Rule Discovery in a Geo-Spatial Decision Support System", NSF Digital Government Grant No. EIA-0091530

Han, J., 2003, "Data Mining in Spatial Database: A Multi-Disciplinary Promise", Database Systems Research Lab, URL:http://www.cs.uiuc.edu/~hanj

Kamber, M., and Han, J., 2001, "Data Mining and Techniiques", Morgan Kanfmann, pp.51-72

Miller, H. J. and Han, J., 1999, "Discovering Geographic Knowledge in Data Rich Environments", *Reporting of a Special Meeting held under the auspices of the Varenius Project, NCGIA*, Vol. 1pp.10-42

Murray, A.T., and Estivill-Castro, V., 1998, "Discovering association in spatial data an efficient method based approach", Proceedings of the Second Pacific-Asia Conference on Resea고 and Development in knowledge discovery and Data Mining, PAKDO-98 Berlin: Springer-Verlag, pp.110-111

Rawlings, J., and Kucera, H., 1997, Trials and tribulations of implementing a spatial data warehouse, *In Proceedings of the 11th Annual Symposium on Geographical Information System*, Vancouver, pp.510-513.

Tung, A., 2002, "Mining Spatial Datasets: A New Frontier for Data Mining", Technical Report in IBM Canada and Intelligent Database Systems Research Lab.

Yuan, M., 2003, "Geospatial Data Mining and Knowledge Discovery", *A Proposed Emergent Research Topic in GIScience.*