아파트 경매를 위한 웹 기반의 지능형 의사결정지원 시스템 구현

나 민 영† · 이 현 호††

요 약

아파트 경매는 일반시민들이 주택장만을 위해 많이 이용하는 제도이다. 본 논문에서는 경매 의사결정 지원을 위하여 OLAP(Online Analytical Processing) 기법과 데이터마이닝 기법을 적용한 웹 기반의 지능형 의사결정지원 시스템의 구현을 다루었다. 구현된 시스템은 실제 경매 데이터로 구성된 경매 데이터베이스 상에서 작동하며 그 핵심은 데이터웨어하우스 기반의 OLAP 지식추출기와 데이터마이닝 기법을 적용한 데이터마이너로 구성된다. OLAP 지식추출기는 OLAP 기법을 경매 데이터베이스에 적용하여 요구되는 지식을 추출하고 그 결과를 시각화하였다. OLAP 기법은 사실(fact), 차원(dimension), 계층구조(hierarchies) 등을 사용하며 롤업(roll-up), 드릴다운(drill-down), 슬라이싱(slicing), 다이싱(dicing), 피보팅(pivoting) 등을 통한 데이터 분석결과(지식)을 제공한다. 경매 데이터마이너는 데이터마이닝 기법 중 분류기법을 경매 데이터베이스에 적용시켜 낙찰가를 예측하는 것으로 lazy model-based classification 알고리즘을 기반으로 하였으며 결정필드 추출, 동적 도메인 정보 구성, 필드가중함수 등의 개념을 적용하여 경매 데이터의 특성을 반영하였다.

Implementation of a Web-Based Intelligent Decision Support System for Apartment Auction

Min-Young Ra† · Hyun-Ho Lee††

ABSTRACT

Apartment auction is a system that is used for the citizens to get a house. This paper deals with the implementation of a web-based intelligent decision support system using OLAP technique and data mining technique for auction decision support. The implemented decision support system is working on a real auction database and is mainly composed of OLAP Knowledge Extractor based on data warehouse and Auction Data Miner based on data mining methodology. OLAP Knowledge Extractor extracts required knowledge and visualizes it from auction database. The OLAP technique uses fact, dimension, and hierarchies to provide the result of data analysis by means of roll-up, drill-down, slicing, dicing, and pivoting. Auction Data Miner predicts a successful bid price by means of applying classification to auction database. The Miner is based on the lazy model-based classification algorithm and applies the concepts such as decision fields, dynamic domain information, and field weighted function to this algorithm to reflect the characteristics of auction database.

※ 본 논문은 정보통신부 1998년도 초고속정보통신 응용기술개발사업 및 1996년도 한국과학재단 특정연구(No.96-0101-08-01-3) 연구비 지원에 의한 것임.
† 정 회 원 : 육군사관학교 전산학과 교수
†† 정 회 원 : 삼성 SDS 연구원
논문접수 : 1999년 4월 2일, 심사완료 : 1999년 10월 25일

1. 서 론

의사결정지원시스템은 의사결정자에게 필요한 고급 정보를 적시에 제공하고 다양한 분석기능을 제공하기 위한 시스템으로 단순한 자료처리보다는 실제로 의사결정을 도와줄 수 있는 시스템이어야 한다. 최근 들어 의사결정시스템이 주목받게된 동기는 기업환경에서의 불확실성 증대로 인해, 경영은 운영과 경영통제보다는 계획과 전략분석에 초점을 맞추게 되었고 또한 컴퓨터 기술의 발달로 인하여 새로운 형태의 정보기술 활용분야로 인식되었기 때문이다. 본 논문에서 다루는 의사결정지원시스템은 기존의 시스템과는 달리 지식을 이용하여 사용자의 의사결정을 도와주는 지능형 시스템이다.

본 논문에서는 이러한 지능형 의사결정지원을 위하여 데이터 웨어하우스/마이닝 기법을 적용한다. 데이터 웨어하우스(Data Warehouse)는 중복을 허용함으로써 정보를 빠르고 쉽게 검색하고 분석하여 의사결정에 직접적으로 도움을 줄 수 있는 전문화된 시스템으로 최근 들어 많은 연구가 진행되고 있는 분야이다[4, 8]. OLAP(On Line Analytical Processing)는 이러한 데이터웨어하우스를 기반으로 중요한 정보처리를 빠르고 다차원적으로 해주는 온라인 분석처리이다. OLAP는 OLTP(On Line Transaction Processing) 시스템의 단점을 보완하기 위해 대두되었다[9, 11]. 왜냐하면 우리가 보통 응용체계에서 사용하는 전형적인 OLTP 시스템은 매일 삽입과 갱신이 빈번하게 일어나는 시스템의 운영을 자동화하기 위해 설계된 것으로 데이터를 빠르고 안전하게 데이터베이스로 저장하는데는 매우 좋으나 의미 있는 분석 결과를 얻는데는 효과적이지 못하기 때문이다. 데이터웨어하우스에 관해서는 최근들어 질의모델, 뷰관리기법, 인덱싱, 시제품구현 등의 분야에서 많은 연구가 이루어지고 있다. 국외에서는 Stanford 대학의 WHIPS 프로젝트[5]가 대표적인 웨어하우징 프로젝트인데 이 프로젝트에서는 웨어하우스 전반에 관한 연구를 수행하고 있다. 국내에서도 대학과 기업을 중심으로 활발한 연구활동이 진행되고 있다[16, 17].

데이터 마이닝(Data Mining)은 대규모 데이터베이스에 존재하는 감추어진 지식을 찾아내는 작업으로써 최근 들어 데이터베이스 분야에서 중요한 응용으로 관심을 끌고 있는 분야이다[2, 3]. 즉 현실세계에서 데이터베이스가 발달하여 수많은 데이터가 쌓여가고 있으므로 이로부터 감추어진 정보를 캐내어 응용하고자하는 요구가 발생하기에 이른 것이다. 마이닝에 관한 연구에는 연관규칙, 분류규칙, 클러스터링 등 여러 분야별로 활발한 연구가 수행중에 있다[2, 6]. 이 중 분류는 가장 관심이 많은 분야중 하나이다.

경매는 일반 시민들이 주택 장만을 위해 많이 이용하는 제도이다[14, 15]. 그러나, 그 절차가 복잡하고 분석된 정보를 구하기가 어려워 쉽게 접근하기가 곤란하다. 경매정보 서비스에 관해서는 국내에서 여러 컨설팅 및 정보기술 관련 회사들이 서비스를 제공하고 있으나 데이터웨어하우스나 데이터 마이닝 기술을 이용하는 것이 아니라 단순히 경매 사실 위주로만 서비스하고 있는 실정이다. 효과적인 경매 정보 분석을 위해서는 데이터웨어하우스 기반의 OLAP 분석 및 데이터 마이닝 기술 적용이 필요하다.

본 논문에서는 OLAP 기법과 데이터마이닝 기법을 적용하여 경매 데이터베이스로부터 다양한 관점에서 경매 데이터 분석결과를 추출하고 경매 낙찰가를 예측하여 의사결정을 지원하는 경매 의사결정지원 시스템을 구현하고, 실제 데이터를 이용하여 구축된 경매 데이터베이스로부터 그 결과를 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 사용하는 경매 데이터베이스의 구성을 다루고, 3장에서는 제안된 시스템 구조를 설명한다. 4장에서는 시스템 구현 및 결과에 관하여 설명하고 5장에서 결론을 맺는다.

2. 경매 데이터베이스의 구성

2.1 법원 경매 데이터베이스

경매는 넓은 의미로 볼 때 성업공사 등에서 실시되는 경매와 법원에 의해 실시되는 경매가 있으나[15] 본 논문에서 다루는 경매는 법원경매를 대상으로 하고 여러 경매물건 중 서울지역 아파트만을 대상으로 하였다. 본 논문에서의 경매 데이터베이스는 실거래 데이터로 이루어진 데이터베이스로서 1998년 7월 1일부터 1999년 1월 8일까지 서울지역 5개법원 즉, 서울민사지방법원, 동부지원, 북부지원, 남부지원, 서부지원에서 경매 결과 낙찰된 아파트 1303건으로서 구성되었다.

2.2 구성 필드

경매 데이터베이스를 이루는 애트리뷰트(필드)로서는 경매와 관련된 모든 정보가 필드로 구성될 수 있으

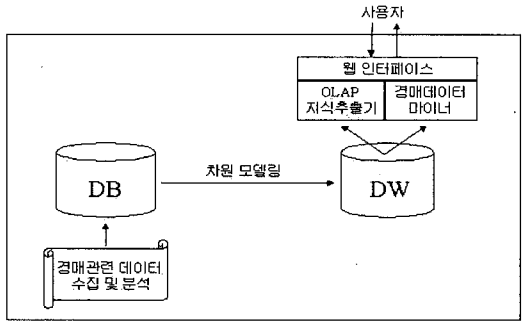
나 본 논문에서 다루는 데이터베이스 필드는 이 중 OLAP 분석 및 마이닝에 필수적인 다음의 필드로 구성된다.

- 최저경매가 : 새로운 경매에 부쳐지는 가격으로 감정가로부터 출발해 한번 유찰될 때마다 20%씩 감액된다.
- 낙찰가 : 경매아파트가 낙찰된 가격
- 낙찰일 : 경매아파트가 입찰결과 하자없이 낙찰된 날짜
- 소재지 : 등기부상에 기재되는 부동산의 주소
- 유찰횟수 : 경매부동산의 유찰된 내력
- 면 적 : 경매 아파트의 면적으로 여기서는 평형으로 처리
- 감정가 : 법원이 경매 전 감정평가사를 통해 실시한 부동산의 시세 가격

3. 시스템 구조

구현된 시스템은 (그림 1)과 같은 구조를 갖는다. 수집된 경매데이터는 데이터베이스로 저장되고 이는 차원모델링을 거쳐 데이터웨어하우스 구조를 갖게 된다. 이렇게 저장된 데이터에 대해 사용자는 웹 인터페이스를 통해 OLAP 지식추출기와 경매 데이터마이너를 이용하여 의사결정에 필요한 원하는 정보를 얻게 된다.

(그림 1)에서 보는 바와 같이 구현된 시스템에서 핵심이 되는 부분은 OLAP 지식추출기와 경매 데이터마이너이다. OLAP 지식추출기(OLAP Knowledge Extractor)는 법원경매에 대한 시장동향, 정책수립 등 거시적인 관점에서의 지식을 제공하기 위한 시스템으로 OLAP 분석모듈, 질의변환 모듈, 시각화 모듈 등으로 구성된다. 이 시스템을 이용하면 소재지 간의 낙찰가 관계, 최저경매가, 유찰횟수, 낙찰가, 평형 간의 관계, 유찰횟수가 낙찰가에 미치는 영향, 향후 경매 낙찰가액 증가/감소 전망 등을 분석할 수 있다. 이때 분석 대상이 되는 측정 애트리뷰트로는 낙찰가를, 측정 애트리뷰트 값을 식별해 주는 차원 애트리뷰트로는 소재지, 크기(평형), 낙찰시기를 설정하였다. 경매 데이터마이너(Auction Data Miner)는 경매에 응하는 사용자 개개인의 경매의사결정을 직접적으로 도와주기 위해 분류(classification)기법을 이용하여 예상 낙찰가를 추출하고 낙찰 여부를 예측할 수 있도록 도와준다.



(그림 1) 시스템 개요

4. 시스템 구현

본 시스템은 웹을 이용한 경매분석 및 예측이 가능토록 구현하였다. 이 장에서는 구현된 시스템을 크게 OLAP 지식추출기와 경매 데이터마이너를 중심으로 구현 세부내용 및 그 결과를 기술한다.

4.1 OLAP 지식추출기

4.1.1 차원 모델링

고전적 데이터 모델링인 ER 모델링은 트랜잭션 처리 성능이 우수한 운영 데이터베이스를 구축하기 위해 수행되어 왔다. 그러나 데이터웨어하우스에서는 질의를 중심으로 사용되기 때문에 ER 모델로서는 부적합하고 새로운 모델링 기법이 필요하게 되었다. 이것이 바로 차원 모델링으로서 이는 사용자 관점에서 스키마를 작성하며 업무상 중요한 수치 데이터와 이들을 설명하는 설명 데이터를 다른 테이블로 분리하여 설계한다[10, 11]. 엔티티, 관계성, 기능적 분해, 및 상태 전이 분석 등에 의존하는 OLTP 시스템을 설계하는 기법과 비교해 볼 때 이 기법은 사실(fact), 차원(dimension), 계층구조(hierarchies) 등을 사용한다.

본 논문에서 다루는 경매 데이터베이스는 소재지, 낙찰시기, 크기의 세 차원의 데이터웨어하우스의 개념을 갖는 구조로 재구성된다. 데이터웨어하우스란 원래 다양한 정보 소스로부터 데이터를 미리 뽑아서 별도의 물리적 저장소에 통합해 놓은 것을 의미하는데 본 논문에서의 데이터웨어하우스는 정보 소스가 하나이고 설명 데이터 또한 다른 테이블로 분리할 만큼 다양하지도 않아서 경매 DB에 경매 차원정보가 첨가된 것을 가리킨다. 이를 위하여 각 차원을 나타내는 애트리뷰트는 더 낮은 레벨이 표현될 수 있도록 분리된다. 예

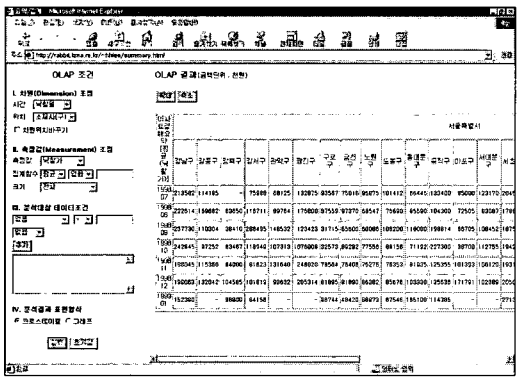
를 들어 낙찰시기는 일, 월, 년으로 구분되어 각각 하나의 필드로 유지된다. 각 차원은 다음과 같은 애트리뷰트 계층을 가지고 있다.

- 소재지 차원 : 동 → 구 → 시
- 크기 차원 : 평형 → 평형분류
- 낙찰시기 차원 : 일 → 월 → 년

여기서, '평형분류'는 '소형(24평이하)', '중형(40평이하)' 또는 '대형(40평초과)'을 값으로 갖는다. 이러한 애트리뷰트의 계층은 데이터의 집단화(aggregation)를 암시한다. 즉 '동'은 모여서 '구'가 되고, '구'는 모여서 '시'가 되는 것이다. 이러한 계층은 다차원 분석에 효과적이다.

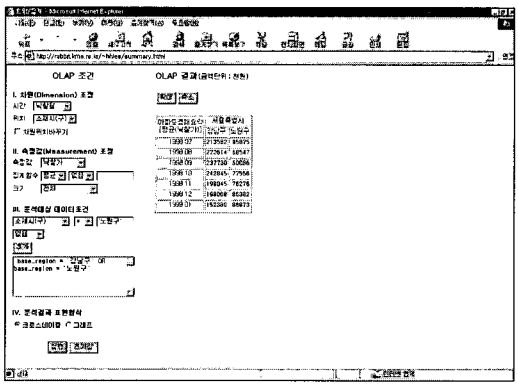
4.1.2 OLAP 지식의 추출

OLAP 분석 즉 지식을 추출하는데 많이 사용되는 기법에는 롤업(roll-up), 드릴다운(drill-down), 슬라이싱(slicing), 다이싱(dicing), 피보팅(pivoting) 등이 있다 [12, 17]. 본 논문에서 구현된 지식추출기에서는 이러한 기법을 모두 구현하여 경매 데이터에 대한 분석지식을 웹을 이용하여 얻게 해 준다. 이때 분석기법 적용에 따른 데이터의 집단화로서는 합계, 평균, 개수를 사용하였다. (그림 2)는 낙찰월, 소재구의 관점에서 낙찰가의 평균을 분석한 결과를 보이고 있다.



(그림 2) OLAP 분석결과

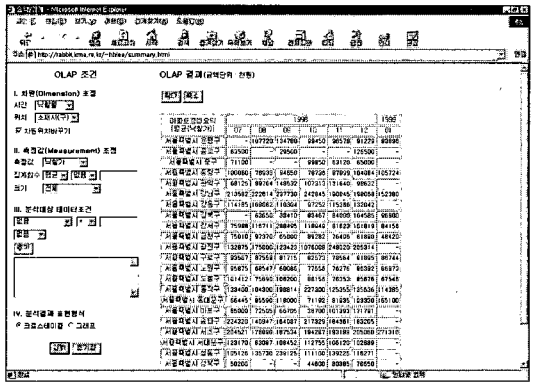
롤업은 차원의 일반화(generalization)를 의미하는 것으로 집단화 정도를 높이는 기법이다. 반면, 드릴다운은 차원의 구체화(specialization)을 의미하는 것으로 집단화 정도를 낮추는 기법이다. 슬라이싱은 차원의 부분집합을 취하는 기법이다. 본 시스템은 낙찰시기, 소재지, 크기의 3가지 차원을 사용하는데 슬라이싱으로 이들 3가지 중 한 두가지 차원만으로 분석결과를 나타낼 수 있다. (그림 2)는 크기차원을 제외한 낙찰시기와 소재지 차원만으로 구성된 점에서 슬라이싱이 적용되었음을 알 수 있다.



(그림 3) (그림 2)에 대한 다이싱 결과

다이싱은 차원이나 분석대상 데이터에 제한 조건을 둠으로써 다차원 데이터의 부분집합을 취하는 기법이다. (그림 3)은 (그림 2)에서 다이싱을 통하여 소재구가 '강남구' 또는 '노원구'인 데이터만을 대상으로 분석한 결과를 나타낸다. 이와 같은 분석을 통하여 '강남구', '노원구' 지역의 경매 낙찰가(평균)에 대한 추세 지식을 얻을 수가 있다.

피보팅은 결과 뷰(view)의 단계에서 적용되는 것으로 차원의 위치를 바꾸어 결과를 나타내는 기법이다. (그림 4)는 (그림 2)를 피보팅하여 나타낸 분석결과이다.



(그림 4) (그림 2)에 대한 피보팅 결과

4.1.3 질의변환

본 연구에서의 차원 모델링은 논리적으로는 다차원 데이터베이스의 기본이 되는 데이터 큐브(cube)에 기반을 두고 있지만 물리적으로는 하나의 flat 릴레이션으로 구현된다. 그러므로, OLAP 분석결과는 여러 가지 차원 관점에서의 OLAP 조건들을 SQL 형태로 변환하여 flat 릴레이션을 대상으로 질의를 수행한 결과를 재구성한 것이다.

OLAP 조건의 SQL 변환에는 다음과 같은 원칙을 적용한다[17].

1) 계층구조를 가지고 있는 차원에 관한 조건을 그룹핑 대수에 적용한다. 즉, 차원조건으로 제시된 차원값과 그것을 기준으로 상위계층에 위치한 차원값들 모두로 "Group By"절을 구성하고 "Select"절에 포함시킨다.

2) 분석대상이 되는 측정 애트리뷰트에 집계함수를 적용하여 "Select"절에 넣는다.

3) 차원 모델링의 물리적 구현 형태인 flat 릴레이션명을 "From"절에 넣는다.

4) 다이싱(dicing) 조건을 "where 절"에 넣는다.

5) 측정 애트리뷰트의 집계함수 값에 대한 제한조건이 있으면 "Having"절에 넣는다.

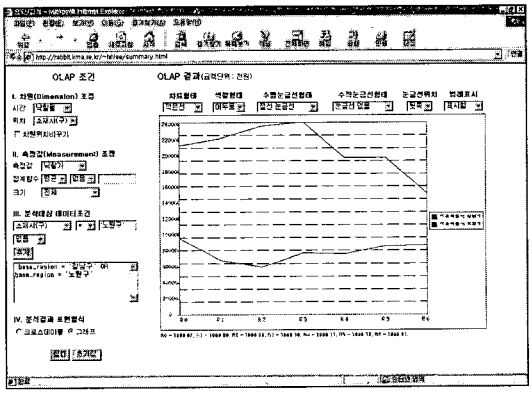
6) 크로스 테이블이나 챠트 형태로 질의 수행 결과를 재구성하는 것을 돕기 위해 "Group By"절과 같은 내용으로 "Order By"절을 구성한다. □

이상의 원칙을 적용하면 본 논문에서 제시한 차원 및 계층구조를 기반으로 한 어떠한 OLAP 조건도 SQL 질의로 변환이 가능하다. 예를 들어, (그림 3)의 OLAP 조건을 SQL 질의로 변환하면 다음과 같다.

SELECT 낙찰년도, 낙찰월, 소재시, 소재구, AVG(낙찰가)
FROM OLAP를 위한 flat relation
WHERE 소재구 = "강남구" or 소재구 = "노원구"
GROUP BY 낙찰년도, 낙찰월, 소재시, 소재구
ORDER BY 낙찰년도, 낙찰월, 소재시, 소재구

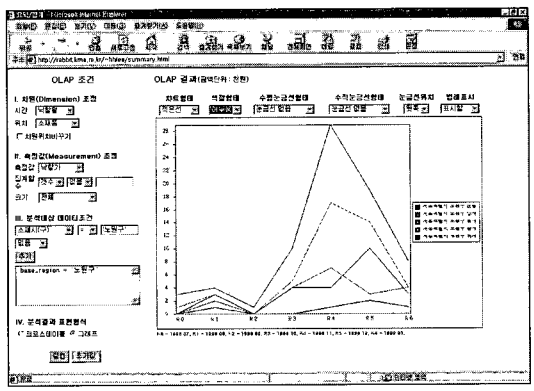
4.1.4 분석결과의 시각화(Visualization)

OLAP 분석결과는 4.1.2절의 예와 같이 일반적으로 크로스 테이블(cross table)의 형태로 나타낸다. 그러나, 크로스 테이블로는 전체적인 데이터의 흐름을 한눈에 파악하기 어렵다. 시각화는 주로 2차원 또는 3차원의 챠트(막대, 꺽은선, 파이 등) 형태로 분석결과를 보여주는 것으로 데이터의 전체적인 성향 파악 등 지식추출에 도움을 준다. (그림 5)는 (그림 3)의 '강남구'와 '노원구'의 낙찰가(평균) 추세 분석결과를 시각화한 것으로 시기에 따른 낙찰가 동향과 두 지역 간의 비교가 한 눈에 파악된다.



(그림 5) OLAP 분석 결과의 시각화 (1)

(그림 6)은 1998년 하반기부터 '노원구'에서 일어난 아파트 경매 건수의 분석결과를 시각화하여 보여주고 있다. (그림 6)의 꺽은선 챠트로부터 '노원구'(특히, '노원구 상계동')에서는 1998년말에 경매가 활발하게 이루어졌음을 쉽게 파악할 수 있다.



(그림 6) OLAP 분석결과의 시각화 (2)

4.2 경매 데이터마이너

4.2.1 결정필드의 추출

경매 데이터베이스는 2.1절에서 밝힌 바와 같은 필

드들로 구성되어 있다. 본 시스템의 목적은 이러한 필드 정보를 근거로 예상 낙찰가를 예측하는 것이다. 모든 필드가 직·간접적으로 낙찰가에 영향을 미치는 요소가 될 수 있으나 동일 비중으로 영향을 미치지는 않는다. 그러므로, 예측하고자 하는 필드 즉, 예측필드(predicted field)인 낙찰가에 크게 영향을 미치는 몇가지 필드 즉, 결정필드(decision field *or* predicting field)를 추출하여 분류 알고리즘을 적용한다.

결정필드를 추출하기 위해서는 필드 간의 관계성(relationship)을 이용해야 하는데 관계형 데이터베이스에서의 함수적 종속성(functional dependency)을 바로 적용할 수는 없다. 왜냐하면, 함수적 종속성은 어떤 릴레이션 R에서 X, Y를 애트리뷰트의 셋이라 할 때, R에 있는 어떤 두 투플 t1, t2 에 대하여 t1[X] = t2[X] 이면 반드시 t1[Y] = t2[Y] 이어야 한다는 것인데 반해, 본 논문에서의 경매 데이터베이스에서는 어떠한 필드집단도 예측필드인 낙찰가를 정확하게 확정짓지 못하기 때문이다. 그러나, 예를 들어 "서울시 강남구에 소재한 30평형 아파트의 낙찰가는 1억 3천만원대에서 형성된다"는 식으로 낙찰가격대의 형성에 중요한 영향을 미치는 필드집단은 분명히 존재하므로 이러한 필드집단 즉, 결정필드를 추출하기 위해서 다음과 같이 의존성 관계를 정의하였다.

[정의 1]

릴레이션 R의 애트리뷰트 집합 {X}가 애트리뷰트 Y 값에 영향을 미칠 때, 'Y는 {X}에 의존적이다'라 하고 '{X} → Y'로 표현한다. 이를 의존성 관계(dependency relationship)라 정의한다. □

결정필드 추출방법은 휴리스틱 도메인 지식(heuristic domain knowledge)에 기반하여 낙찰가에 관계된 몇개의 의존성 관계들을 선정한 다음, 통계적인 접근방법으로 의존성 관계 각각의 낙찰가 의존도(dependency degree)를 구하여 낙찰가 의존도가 가장 높은 필드집단을 결정필드로 삼는 것으로 의존도를 정의하기 위한 방법은 다음과 같다.

1) 트레이닝 셋인 릴레이션 R에서 결정필드인 애트리뷰트 집합 {X}의 튜플 인스턴스 값이 같은 것끼리 하나의 군으로 묶는다.

2) 군에 속한 {X}의 튜플 인스턴스 각각에 대한 예측필드 Y 값들의 표준편차를 구한다.

3) 각 군들의 표준편차 값의 평균을 구하여 의존성 관계 '{X} → Y'에 대한 의존도로 삼는다.

위의 접근방법은 트레이닝 셋 내에서 결정필드의 값이 같은 튜플들의 예측필드 값이 편차가 적게 결정될수록 해당 결정필드가 예측필드에 더 큰 영향을 미친다는 아이디어를 기본으로 한 것이다. 이에 근거하여 의존성 관계 '{X} → Y'에 대한 의존도는 다음과 같이 정의하였다.

[정의 2]

릴레이션 R의 애트리뷰트 집합 {X}와 애트리뷰트 Y 사이에 의존성 관계 {X} → Y가 성립할 때, {X} → Y의 의존도를 $T_Y(\{X\})$라 하면,

$$T_Y(\{X\}) = \frac{1}{n} \sum_{i=1}^{n} STDEV(\pi_Y(\sigma_{\{X\}=\{x_i\}}(R)))$$

(*STDEV() : 표준편차*)

(여기서 x_i는 {X}에서 똑같은 값을 갖는 튜플 인스턴스를 가리키고 n은 이러한 x_i의 갯수임)

로 정의한다. 이 때, $T_Y(\{X\})$ 값이 작을수록 의존도가 높다. □

(예 1)

다음과 같은 릴레이션에서 $\{X_1, X_2\}$ → Y 의 의존도와 $\{X_1, X_2, X_3\}$ → Y를 구해 보자.

X_1	X_2	X_3	Y
a	α	y	10
b	β	z	30
a	α	y	30
a	β	y	10
b	β	y	20

a) 의존성 관계 $\{X_1, X_2\}$ → Y의 의존도

정의에 따라서, $\{x_i\}$는 1번째와 3번째 튜플의 인스턴스인 {a, α}와 2번째와 5번째 튜플의 인스턴스인 {b, β}이다. 그리고, n = 2이다.

$STDEV(\pi_Y(\sigma_{\{X\}=\{a,\alpha\}}(R))) = STDEV(10, 30) = 14.1$

$STDEV(\pi_Y(\sigma_{\{X\}=\{b,\beta\}}(R))) = STDEV(10, 30) = 7.1$

그러므로, $\{X_1, X_2\}$ → Y의 의존도는 이들의 평균값인 10.6이다.

b) 의존성 관계 $\{X_1, X_2, X_3\}$ → Y의 의존도

정의에 따라서, $\{x_i\}$는 1번째와 3번째 튜플의 인스턴스인 {a, α, y}이고, n = 1이다.

$STDEV(\pi_Y(\sigma_{\{X\}=\{a,\alpha,y\}}(R))) = STDEV(10, 30) = 14.1$

n=1이므로 $\{X_1, X_2, X_3\} \rightarrow Y$의 의존도는 14.1이다. □

본 연구에서 경매 휴리스틱 도메인 지식에 기반하여 추출된 후보 의존성 관계는 다음과 같다. 여기서 (4.1), (4.2), (4.3)은 휴리스틱 도메인 지식에서 추출된 의존성 관계이고 (4.4), (4.5), (4.6)은 이들 의존성 관계로부터 유도된 것이다. 물론 의존성 관계의 폐포(closure)에 속하는 다른 의존성 관계도 있을 수 있으나 본 논문에서는 낙찰가에 직간접적으로 크게 영향을 미치는 의존성 관계만 다루었다.

- 소재지, 면적 → 감정가 (4.1)
- 감정가, 유찰횟수 → 최저경매가 (4.2)
- 최저경매가 → 낙찰가 (4.3)
- 소재지, 면적, 유찰횟수 → 최저경매가 (4.4)
- 소재지, 면적, 유찰횟수 → 낙찰가 (4.5)
- 소재지, 면적, 감정가, 유찰횟수 → 낙찰가 (4.6)

본 연구에서 사용한 경매 데이터베이스 트레이닝 셋을 대상으로 (4.1)~(4.6)에서 제시한 의존성 관계 각각에 대한 의존도 T_Y값을 계산하면 다음과 같다.

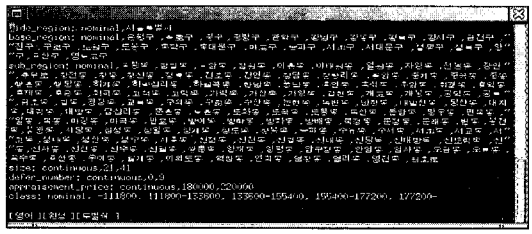
〈표 1〉 의존성 관계의 의존도 (단위:천원)

의존성 관계	T_Y
(4.1)	15703
(4.2)	3375
(4.3)	15151
(4.4)	7252
(4.5)	9235
(4.6)	5633

<표 1>에서 낙찰가 의존성에 직접 관련된 (4.3), (4.5), (4.6)을 비교하면 (4.6)의 의존도가 가장 높게 나타났음을 알 수 있다. <표 1>의 결과를 토대로 경매 마이닝 시스템에서 낙찰가에 대한 결정필드는 {소재지, 면적, 감정가, 유찰횟수}로 결정하였다. 이상에서 알 수 있는 바와 같이, 예측필드 {낙찰가}는 최저경매가보다는 소재지, 면적, 감정가 등에 더 크게 의존하고 있음과, 영향을 미치는 휴리스틱 도메인 지식이 복합되면 더욱 의존도가 높아짐을 알 수 있다.

4.2.2 동적 도메인 정보구성

이 절에서 설명할 도메인 정보란 어떤 시점에서의 데이터베이스가 실제 저장하고 있는 분류에 필요한 결정필드와 예측필드 각각에 대한 필드값들의 집합(또는 범위)을 의미하며 다음 절에서 설명될 lazy model-based classification 알고리즘 구현의 필수적인 요소이다. 실제 이 알고리즘이 제시된 환경[7]에서 사용된 데이터는 인공지능 분야에서 주로 사용하는 소규모이고 정적인 데이터이나 본 논문에서 사용하는 데이터는 실세계의 운영 데이터베이스로서 대규모이고 동적인 데이터이다. 그러므로, 결정필드와 분류의 목적이 되는 예측필드의 도메인 정보도 동적으로 구성해야 한다. (그림 7)은 본 연구에서 사용한 동적 도메인 정보의 예이다.



(그림 7) 동적 도메인 정보 구성

결정필드 {소재지, 면적, 감정가, 유찰횟수} 중에서 소재지는 다시 광역구역(wide_region : 도/광역시단위), 기초구역(base_region : 구단위), 서브구역(sub_region : 동단위)으로 나누었는데 이는 아파트 시세가 시/구/동에 따라 차이가 많다는 휴리스틱 도메인 지식을 반영한 것으로 분류 결과도 시단위, 구단위, 동단위 등으로 다양하게 얻을 수 있는 장점이 있다.

결정필드의 도메인 정보는 하나의 테스트 인스턴스에 대해 분류 알고리즘을 적용할 때마다 트레이닝 셋에 대한 다음과 같은 질의를 통해 동적으로 구성됨을 원칙으로 한다.

SELECT DISTINCT(wide_region), DISTINCT(base_region), DISTINCT(sub_region), MIN(size), MAX(size), MIN(appraisement_price), MAX(appraisement_price), MIN(defer_number), MAX(defer_number)
FROM table name for classification

그러나, 본 논문의 분류 알고리즘은 테스트 인스턴

스의 필드값에 대한 예측필드의 값을 추정하는 것이므로 면적, 감정가, 유찰횟수와 같이 연속성을 갖는 수치필드의 도메인은 테스트 인스턴스의 해당 필드값을 중심으로 제한하는 것이 효율적일 수 있다.

예측필드의 도메인 정보도 필요하다. 본 논문의 분류 알고리즘에서 사용되는 예측필드의 도메인은 불연속적 클래스 데이터이어야 하나[7], 실제 예측필드인 {낙찰가}의 도메인은 수치 데이터로 연속적이다. 따라서, 본 논문에서는 다음의 방법으로 {낙찰가}를 클래스화 한다.

1) 트레이닝 셋에서 테스트 셋(테스트 인스턴스)에 해당하는 낙찰가의 최소값과 최대값을 질의를 통해 구한다.

2) 전체 낙찰가 구간은 [최소값 - ξ, 최대값 + ξ]로 정한다. 이 때, ξ는 4.2.1절의 의존성 관계 (4.6)의 $2 * T_Y$값으로 한다. 이것은 아파트 시세가 10000천원 정도의 등락폭이 상존하고 있음을 반영한 것이다.

3) 구간 수를 사용자 입력으로 받아 전체 낙찰가 구간을 구간 수로 나누어 클래스를 정한다. □

이상에서 살펴본 바와 같이, 결정필드는 트레이닝 셋에 따라, 예측필드는 테스트 셋과 트레이닝 셋에 따라 동적으로 구성됨을 알 수 있다.

4.2.3 분류 알고리즘

분류는 트레이닝 셋이라 하는 과거에 알고 있는 데이터베이스 정보로부터 새로운 데이터베이스 튜플을 분류해낼 수 있는 분류 규칙을 생성해 내는 기법이다[1, 2]. 분류의 뿌리는 기계학습 분야에서 많이 쓰인 고전적 트리분류기인 ID3이다. 이 외에도 IC[1], SLIQ[6], PUBLIC[13] 등이 성능향상을 위한 트리분류기로 제안되었다. 분류의 응용은 크레딧 승인, 마켓팅, 의료진단, 상점 위치결정 등 여러 분야에 쓰일 수 있다. 본 논문에서 다루는 경매는 그 결과가 낙찰여부로 나타나므로 분류를 적용할 수 있는 좋은 예이다. 즉 낙찰된 데이터를 분석하여 어떤 경우에 낙찰이 되는가를 규칙으로 나타내고 이 분류규칙을 이용해 분석을 원하는 대상에 대하여 적절한 신뢰도를 갖는 예상낙찰가를 제공함으로써 사용자로 하여금 최적의 입찰가를 결정할 수 있게 해준다.

본 연구에서 개발된 시스템에 사용된 분류 알고리즘은 공개 알고리즘인 lazy model-based classification 알고리즘[7]을 변형하여 구현하였다. Lazy model-based classification 알고리즘은 테스트 셋에서 하나의 인스턴스 이벤트의 필드값을 바탕으로 하향탐색(top-down search)방식을 이용하여 최적의 규칙(rule)을 찾는 것으로 규칙은 IF-THEN 형식으로 표현된다. 탐색과정은 다음과 같이 요약된다.

1) 인스턴스 이벤트를 만족하는 가장 일반적인 씨앗 규칙(seed rule)을 정하여 탐색을 시작한다.

2) 인스턴스 이벤트를 만족하는 한단계 더 구체화된 규칙들을 찾아 후보규칙(candidate rule)을 정한다.

3) 후보규칙 각각에 평가함수(evaluation function)을 적용하여 결과값이 가장 높은 규칙을 지역최적규칙(local optimal rule)으로 정한다. 본 알고리즘의 평가함수는 어미규칙의 클래스 분포와 자식규칙의 클래스 분포 간의 유클리드 거리(Euclidean distance)를 이용한 것이다.

4) 지역최적규칙과 어미최적규칙(parent optimal rule) 간의 유클리드 거리가 0이 아니면(즉, 지역최적규칙의 평가함수값이 0가 아니면) 2) ~ 4)과정을 반복하고 그렇지 않거나, 더 이상 구체화된 규칙을 찾을 수 없으면 5)번으로 간다.

5) 현 시점의 지역최적규칙을 최적의 규칙으로 삼는다. □

(예 2)

Lazy model-based classification 알고리즘의 적용과정에 대한 이해를 돕기 위한 예제로 다음과 같은 트레이닝 셋과 테스트 셋을 가정해 보자.

- 트레이닝 셋 : 결정필드 {X_1, X_2, X_3}와 예측필드 {X_4}로 구성된 100개의 튜플을 가 진 테이블
- 테스트 셋 : 결정필드 X_1, X_2, X_3이 각각 x_1, x_2, x_3 값을 가진 인스턴스 튜플
- 예측필드의 도메인(클래스) : {c_1, c_2}

알고리즘은 먼저 다음과 같은 씨앗규칙(seed rule) r_0를 찾는다.

$$r_0 : (X_{\{1,2,3\}} = \text{ANY}) \rightarrow X_4 \in [30, 70]$$

r_0에서 [30, 70]은 클래스 분포를 나타내는 것으로 트레이닝 셋의 예측필드의 값이 c_1인 튜플이 30개, c_2인 튜플이 70개라는 뜻이다.

다음으로 알고리즘은 테스트 셋의 인스턴스 튜플 값을 반영하기 위해 r_0보다 한단계 더 구체화된 규칙 r_1, r_2, r_3를 찾는다.

$$r_1 : (X_1 = x_1) \rightarrow X_4 \in [16, 64]$$
$$r_2 : (X_2 = x_2) \rightarrow X_4 \in [14, 6]$$
$$r_3 : (X_3 = x_3) \rightarrow X_4 \in [30, 70]$$

후보규칙 r_1, r_2, r_3 각각에 평가함수를 적용하여 결과값이 가장 좋은 규칙을 지역최적규칙으로 삼는다. 예를 들어, r_2가 지역최적규칙으로 결정되었다면 r_2에 대한 한단계 더 구체화된 규칙을 찾는다. 다음과 같은 r_4, r_5가 구체화된 규칙이라면

$$r_4 : (X_2 = x_2) \wedge (X_1 = x_1) \rightarrow X_4 \in [0, 0]$$
$$r_5 : (X_2 = x_2) \wedge (X_3 = x_3) \rightarrow X_4 \in [14, 6]$$

r_4와 r_5에 평가함수를 적용하는데 r_4는 매치(match)되는 튜플이 없고 r_5의 클래스 분포는 r_2와 같으므로 r_4와 r_5 둘 다 r_2와의 유클리드 거리가 0이다. 그러므로, r_2를 최적의 규칙으로 삼는다. □

본 연구에서는 위의 알고리즘을 적용하는데 있어서 필드가중함수를 제안하였다. 이는 결정필드 {광역구역, 기초구역, 서브구역, 면적, 감정가, 유찰횟수}가 예측필드 {낙찰가}에 동등한 영향을 미치지는 않는다는 휴리스틱 도메인 지식을 반영한 것이다. 필드가중함수를 다음과 같이 정의한다.

[정의 3]

릴레이션 R의 결정필드를 $\{X_1, ..., X_i, ..., X_n\}$라 하고 예측필드를 Y라 하고 의존성 관계 $\{X\} \rightarrow Y$의 의존도를 T_Y라 하자. 그러면, X_i에 대한 필드가중함수 field_weight(X_i)는 다음과 같이 정의된다.

field_weight(X_i) =

$$\frac{T_Y(X_1,...,X_{i-1},X_{i+1},...,X_n) - T_Y(X_1,...,X_{i-1},X_i,X_{i+1},...,X_n)}{\sum_{j=0}^{n}\left(T_Y(X_1,...,X_{j-1},X_{j+1},...,X_n) - T_Y(X_1,...,X_{j-1},X_j,X_{j+1},...,X_n) \right)}$$

□

일반적으로 X_i가 포함되지 않은 의존성 관계의 의존도와 X_i가 포함된 의존성 관계의 의존도와의 차이가 클수록 X_i가 예측필드 Y 값에 미치는 영향이 크다 즉, X_i에 대한 Y의 의존도가 높다라고 말할 수 있다. (정의 3)은 X_i의 포함 유무에 따른 의존도의 차이를 field_weight(X_i)에 반영한 것이다. 그러므로, 필드가중함수는 Y의 의존도를 높이는 X_i(즉, field_weight(X_i)의 값이 큰 X_i)를 알고리즘 적용과정에서 다른 결정필드들에 비해 살아남는데 유리한 조건을 주게 하여 알고리즘이 찾는 최적의 규칙에 포함될 확률을 높이는데 목적이 있다.

본 연구에서 사용하는 결정필드 중에서 {광역구역, 기초구역, 서브구역}은 '소재지' 필드를 3개의 필드로 나눈 것으로 의미적인 계층구조가 성립한다. 즉, '기초구역'은 '광역구역'을 의미적으로 내포하고 있으므로 '기초구역'에 대한 필드가중함수 값을 구할 때, X_i를 {광역구역, 기초구역}으로 잡는 것이 타당하다. 마찬가지로, '서브구역'에 대한 필드가중함수 값을 구할 때, X_i는 {광역구역, 기초구역, 서브구역}으로 잡는다.

본 연구에서 사용한 트레이닝 셋에서 계산한 결정필드 {광역구역, 기초구역, 서브구역, 면적, 감정가, 유찰횟수} 각각에 대한 필드가중함수 값은 <표 2>와 같다.

〈표 2〉 결정필드 각각에 대한 필드가중함수 값

결정필드	필드가중함수 값
광역구역	0
기초구역	0.09
서브구역	0.24
면적	0.14
감정가	0.25
유찰횟수	0.27

<표 2>에서 '광역구역'의 필드가중함수 값이 0인 것은 트레이닝 셋의 데이터가 '서울특별시'로 한정되어 있기 때문이다. 그리고, '서브구역' 필드가 '면적' 필드보다 필드가중함수 값이 크게 나타나는 것에서, 서브구역(동단위)에 따른 아파트 시세의 차이가 크다는 것과 같은 서브구역 내에는 대체로 면적이 비슷한 아파트들이 모여있다는 것을 알 수 있다. 또한, '감정가'와 '유찰횟수'가 낙찰가 형성에 비중있게 작용하고 있음을 알 수 있다.

필드가중함수 field_weight(X_i)는 결정필드 X_i가 구체화된 후보규칙의 평가함수값에 곱해져 최종 평가함수값으로 사용된다.

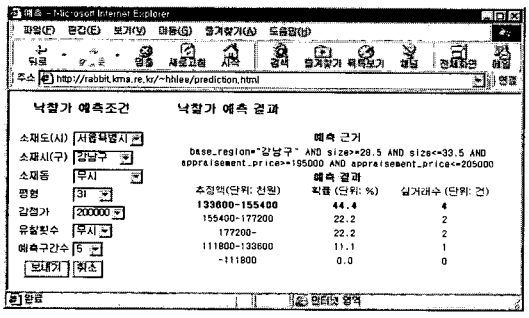
(예 3)

어미규칙이 r인 규칙 r'의 평가함수값을 F(r', r)이라

하고 r'는 r로부터 결정필드 Xi가 구체화된 규칙이라고 하면 r'의 최종 평가함수값은 field_weight(Xi) * F(r', r)이 된다. (예 1)에서 r_4의 최종 평가함수값은 field_weight(X_1) * F(r_4, r_2)이고 r_5의 최종 평가함수값은 field_weight(X_3) * F(r_5, r_2)가 된다.

4.2.4 분류 알고리즘 적용 결과

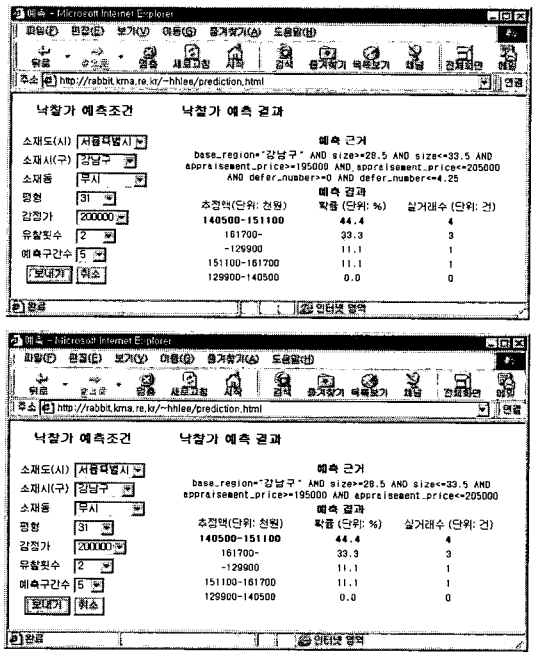
이 절에서는 지금까지 설명된 알고리즘을 이용하여 경매 데이터마이너가 예측필드 {낙찰가}를 예측한 결과를 보인다. (그림 8)은 사용자 질의구성 및 그 결과를 보여준다. 이 예에서 사용된 사용자 질의는 "서울특별시 강남구 소재하고 감정가가 200000천원인 31평형 아파트의 낙찰가를 5개 구간으로 예측하라"이다. (그림 8)의 화면에서 왼쪽이 이 질의 구성을 보여주고 질의 결과는 화면의 오른편에 디스플레이된다. 디스플레이된 질의결과의 의미는 기초구역(base_region : 구)이 '강남구'이고 면적(size : 평형)이 '28.5평 - 33.5평'이고 감정가(appraisement_price : 천원)가 '195000천원 - 205000천원'이면, 낙찰가가 '133600천원 - 155400천원'일 확률이 44.4%, '155400천원 - 177200천원'일 확률이 22.2%, '177200천원 이상'일 확률이 22.2%, '111800천원 - 133600천원'일 확률이 11.1%이며 트레이닝 셋에서 이러한 질의조건 및 결과로 이루어진 거래수는 각각 4개, 2개, 2개, 1개라는 뜻이다.



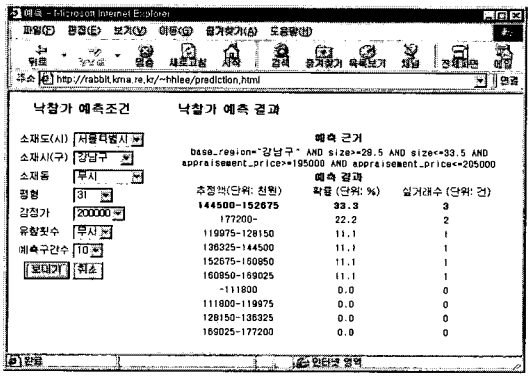
(그림 8) 사용자 질의 및 질의 결과

(그림 9)는 (그림 8)의 사용자 질의에 '유찰횟수 = 2'인 조건을 덧붙여 사용자 질의를 구성한 결과로, 이는 앞에서 설명한 필드가중함수 적용 유무를 비교한 것이다. (그림 9)에서 보는 바와 같이, 시스템에서 제시된 '예측근거'를 비교해 보면, 필드가중함수를 적용한 경우가 적용하지 않은 경우에 비해 'defer_number >= 0 and defer_number <= 4.25'라는 조건이 추가되어 있다. 즉, 필드가중함수를 적용하면 유찰횟수(defer_number)의 범위가 지정됨으로써 좀 더 정확하고 유용한 예측 근거를 제시할 수 있음을 알 수 있다.

(그림 10)은 4.2.2절에서 설명한 동적 도메인정보 구성을 보여주는 예로 (그림 8)에서 낙찰가 예측구간을 10개로 늘려 예측한 결과를 보여준다. 이와 같이 낙찰가 예측구간을 사용자가 임의로 정함으로써 다양한 형태의 예측이 가능하게 된다.



(그림 9) 필드가중함수 적용 비교 : 유(上) · 무(下)



(그림 10) 동적 도메인 정보 구성의 예