FP-tree를 이용한 효율적인 수강신청 로드맵 제시 기법

박영욱*, 이승철*, 김응모*
*성균관대학교 정보통신 공학부
e-mail : {sewhi, eddie, umkim}@ece.skku.ac.kr

Roadmap of an application for attending the lecture by FP-tree

YoungWook Park*, SeungCheol Lee*, Ung Mo Kim*
*Dept of Computer Engineering, Sungkyunkwan University
e-mail : {sewhi, eddie, umkim}@ece.skku.ac.kr

요 약

데이터베이스 시스템 사용이 거의 모든 분야에 걸쳐서 필수적인 요소가 되어가고 있다. 따라서 데이터베이스 내에 축적된 정보들의 양과 더불어 정보와 정보 사이의 연관성이 중요한 관심사로 대두되고 있다. 이를 충족하기 위한 구체적인 방안으로 데이터마이닝 기법이 개발되고 발전해나가고 있다.

현재 수강신청시 많은 학생들이 필수 로드맵이라는 단편적인 정보를 가지고 수업 시간표를 구성함으로써 개개인이 많은 시간을 허비하고 있다. 이에 본 논문에서는 관련성 있는 정보 추출에 용이한 FP-Growth 마이닝 기법을 이용하여 수강신청시 도움이 되는 수강신천 로드맵 기법을 제안한다.

1. 서론

컴퓨터 시스템의 급속한 발전과 데이터베이스 시스템 사용의 보편화는 데이터베이스에 저장되는 데이터의 양적 증가를 초래하게 되었다. 아울러 의사 결정자는 단지 방대한 데이터에 묻혀 있는 가치 있는 지식들을 추출해주는 도구들이 없다는 이유로, 데이터베이스에 저장된 풍부한 정보를 바탕으로 의사결정을 하는 것이 아니라 그의 직관에 근거하여 중요한 의사 결정을 하곤 한다[1]. 이에 데어터마이닝 기법이 제시되고 점차 중요성이 강조되어 왔으며 현재는 관련 연구들이 활발히 진행되고 있다. 따라서 본 논문은 대학의 수강신청 시스템에 연관규칙 마이닝을 적용함으로써 실제로 학생들에게 도움이 되는 유용한 로드맵 자료를 제시하려 한다. 현 수강신청 대상자를 위한 로드맵은 대상자가 필수적으로 수강해야하는 수업에 대해서 학기별로 나열하는 방식에 그치고 있다. 이는 실제로 수강신청을 하는 대상자에게 기본적인 정보제공을 할뿐 실제로 특정 시간에 어떠한 수업을 중복되지 않게 배치할 수 있는가에 대한 구체적인 문제에 대해서는 해결책을 제시하지 않는다. 이에 구체적인 대안을 제시하기 위해 본 논문은 우선 특정 전공의 특정 학년이 실제로 그 해에 어떻게 수강신청을 하였는지에 대한 데이터베이스를 구축한다. 이후 연관규칙 마이닝 기법을 이용하여 매년 실제로 수강신청을 해야 하는 대상자들에게 필수 로드맵(학생들이 필수적으로 수강해야 하는 과목들을 나열한 기존 로드맵)을 만족시킬 수 있는 현실적인 자료를 제공을 제안한다. 이를 통하여 학생들은 보다 적은 시행착오를 거쳐서 본인의 수강 신청을 할 수 있게 된다. 본 논문의 구성은 다음과 같다. 2장에서는 수강신청 로드맵 제시를 위해 사용되는 연관규칙 마이닝 기법과 그중 실제로 적용된 후보 생성이 없는 마이닝 기법인 FP-Growth 기법에 대해 소개하고, 3장에서는 본 논문에서 제안하는 수강신청 로드맵 제시를 위한 마이닝 기법에 대해 설명하고 4장에서는 전체적인 고찰 및 발전된 연구를 위한 향후 연구과제를 제시함으로써 결론을 맺는다.

2. 관련 연구

이 절에서는 관련 연구로서 유비쿼터스 환경에서 보다 중요성이 강조 되고 있는 연관규칙 마이닝의 개요와 후보 항목을 생성하지 않는 효율성이 뛰어난 FP-Growth (Frequent Pattern Growth)기법에 대해 설명한다.

2.1 연관 규칙 마이닝

연관 규칙 마이닝[2,3]은 주어진 데이터 집합에서 흥미로운 관련성을 찾아낸다. 즉 특정 데이터베이스 상에서 연관성있는 항목들을 찾아내는 방법 중 하나이다.

정의 1.

$I = \{i_1, i_2, \ldots, i_n\}$을 m개의 항목을 가지는 집합이라 하자. 트랜잭션 T는 $T \subseteq I$인 항목들의 집합이고 D는 트랜잭션 T들의 데이터베이스라고 하자. X, $Y \subset I$ 이고 $X \cap Y = \varnothing$을 만족 할때 X, Y를 아이템셋이라 부르고, 연관 규칙은 $X \Rightarrow Y$의 형식으로 표현한다.

연관 규칙의 두가지 중요한 측정요소로서 **지지도**와 **신뢰도**가 있고 다음과 같이 정의 할 수 있다.

정의 2. 연관 규칙의 **지지도**는 $X \cup Y$의 트랜잭션수를 D (데이터 베이스의 모든 트랜잭션의 수)로 나누었을때의 백분율로 나타난다.

그러므로 만약 연관규칙에서 5%의 지지도를 가진다고 한다면 그것은 총 데이터베이스의 트랜잭션의 5%가 $X \cup Y$의 항목집합을 포함하고 있다는 것을 나타낸다. 지지도는 연관규칙의 통계적인 의미이다.

예를 들면 연관 규칙에서의 5% 지지도의 의미는 전자제품 대리점에서 전체구입고객(D)의 5%가 TV(X)와 DVD(Y) 플레이어를 동시에 구매한다는 의미가 된다.

대리점 사장은 이러한 연관 규칙의 지지도를 근거로 특정 제품들을 비슷한 위치에 진열하거나 할인 판매를 하는 식으로 유용하게 이용할 수 있을 것이다.

정의 3. 신뢰도는 연관규칙 $X \Rightarrow Y$에서 $X \cup Y$의 트랜잭션의 수를 X를 포함하는 트랜잭션의 수로 나누었을때의 백분율로 나타난다.

그러므로 85%의 신뢰도를 가지는 연관규칙이 있다는 것은 X를 포함하는 항목집합 중 85%가 Y역시 포함하고 있다는 것을 의미한다. 연관규칙에서의 신뢰도는 X와 Y사이의 상호 연관성을 나타낸다고 할 수 있다.

즉 85% 신뢰도는 TV를 구입한 고객의 85%가 DVD플레이어도 구매한다는 것을 의미한다.

연관규칙에서 최소 지지도와 최소 신뢰도를 동시에 만족하는 경우 강한 규칙이라 정의하므로 신뢰도는 강한 규칙을 판별하는 측정치가 되기도 한다.

일반적으로 데이터베이스에서 연관 규칙 마이닝은 사용자나 전문가가 정해놓은 최소 지지도 임계값과 최소 신뢰도 임계값을 만족하는 모든 규칙을 찾는 것이다.

2.2 FP-Growth(Frequent Pattern Growth: 빈발 패턴 증가)

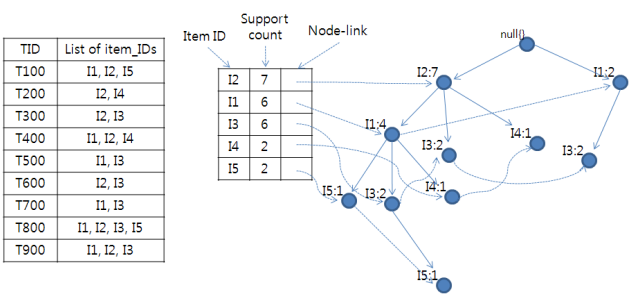
FP-Growth기법[4,5]은 연관 규칙 마이닝의 알고리즘 중 하나로 후보생성 없이 분할-정복 기법을 사용하여 빈발항목을 찾아낸다. FP-Growth는 깊이 우선 알고리즘으로 FP-Tree라는 자료구조를 사용한다. 이후 FP-Tree에 압축된 데이터베이스를 하나의 빈발 항목에 대하여 연관된 조건 데이터 베이스의 집합으로 분할하고 분할된 각각의 데이터베이스에 대헌 개별적 마이닝을 통해 빈발 항목집합을 얻어낼수 있다. 이를 이용한 마이닝 절차는 다음과 같다.

FP-Growth기법은 모든 빈발 항목집합을 얻어내기 위해 2번의 데이터베이스 스캔만을 필요로 한다.

첫 번째 데이터베이스 스캔으로 모든 1항목집합을 찾아낸다. 이 항목집합들 중 최소 지지도를 만족하는 것들만 헤더테이블에 내림차순으로 정렬한다.

이후, "null"로 표시된 트리의 루트를 생성하고, 두 번째로 데이터베이스를 스캔한다. 이때 헤더테이블에 내림차순으로 정렬된 각 트랜잭션마다 가지가 생성된다. 만약 스캔한 트랜잭션의 항목집합의 접두부가 이미 존재할 경우에는 기 생성된 가지의 노드부분과 지지도 카운터를 공유한다. 카운터는 루트로부터 노드까지의 경로를 나타내는 항목집합을 포함하는 트랜잭션의 숫자를 저장한다. 카운터는 매 데이터베이스 스캔시 트랜잭션이 새로운 가지를 추가시킬 때마다 갱신된다. 그림 1의 (a)는 예시 데이터이고 그림 1의 (b)는 예시 데이터의 FP-Tree이다.

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3



(a)예시데이터 (b)FP-tree

그림1

3. 수강신청 로드맵 제시 마이닝 기법

이 절에서는 첫째로 새롭게 제안된 수강 신청 로드맵 제시 마이닝 기법이 적용된 시스템의 전반적인 흐름에 대해 설명하고 둘째로 예시 데이터를 실제로 FP-Growth 기법의 FP-tree를 이용하여 최종적으로 사용자가 얻게 되는 패턴 정보에 대해 설명한다.

3.1 제안된 수강신청 시스템

수강신청 서버는 사용자가 수강신청시 사용자에게 사전에 마이닝 되어 패턴 평가 모듈에 의해 데이터화된 빈발한 연관정보를 제공 해준다. 이후 수강 신청을 마친 학생의 수강신청 데이터는 정제 및 분할되어 데이터 웨어하우스에 추가/갱신되며 이는 이전 데이터들과 통합되어 새롭게 마이닝 되고 패턴 평가 모듈에 의해 새로운 빈발 패턴 데이터를 생성 한다. 즉, 매 수강신청 시즌 마다 새롭게 업데이트 되면서 수강 과목 변경사항 및 필수 항목들을 유연하게 패턴 데이터에 반영한다. 그림2는 제안된 시스템의 전체적인 흐름도이다.

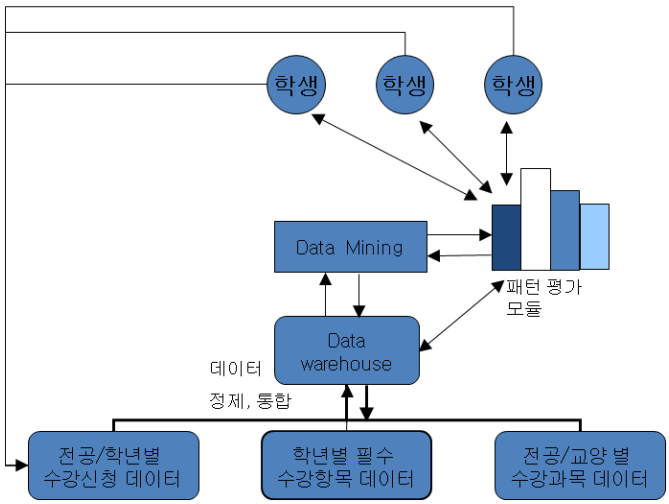


그림2. 시스템 흐름도

3.2 FP-tree를 이용한 제안된 시스템의 마이닝 및 패턴 평가

표1. 수강신청 데이터

STID	Grade	List of course_ids
100	1	전공1, 교양1, 교양2
	2	전공2, 전공3, 교양3
	3	전공4, 전공5, 전공6, 전공7
	4	교양4, 교양5, 전공8
200	1	전공1, 교양1, 전공2
	2	전공3, 교양2
	3	전공4, 전공5, 전공7, 교양3
	4	교양4, 교양5
300	1	전공1, 교양4
	2	전공2, 전공3
	3	전공4, 전공5, 교양2
	4	교양1, 교양3, 교양5

다음과 같은 전제조건을 만족할 때 시스템은 마이닝 과정을 수행한다. Ⅰ)사용자에게 수강 패턴 정보를 제공하기 위해 우선 데이터 웨어하우스에 특정학과 학생들의 최소 1학기분의 수강신청 데이터가 미리 저장되어 있어야 한다. Ⅱ)이에 앞서 저장될 데이터는 학생들이 졸업 이수 요건을 만족하였음을 판별하여 선택적으로 저장되어야 한다, Ⅲ)마지막으로 데이터 웨어하우스에 저장되어 있는 데이터는 마이닝 되기전 표1과 같은 형태로 정제되어 있어야 한다. 표1은 실제 마이닝에 사용되는 데이터 웨어하우스에 저장된 수강신청 데이터이다. 이를 이용하여 사용자의 수강 신청에 도움을 주기위한 유용한 정보를 본 논문 2.2절에 설명한 FP-Growth 기법을 응용하여 그림3과 같이 각 학생의 학년별 FP-tree를 생성한다.(각 트리의 최소 지지도는 2라고 한다.)

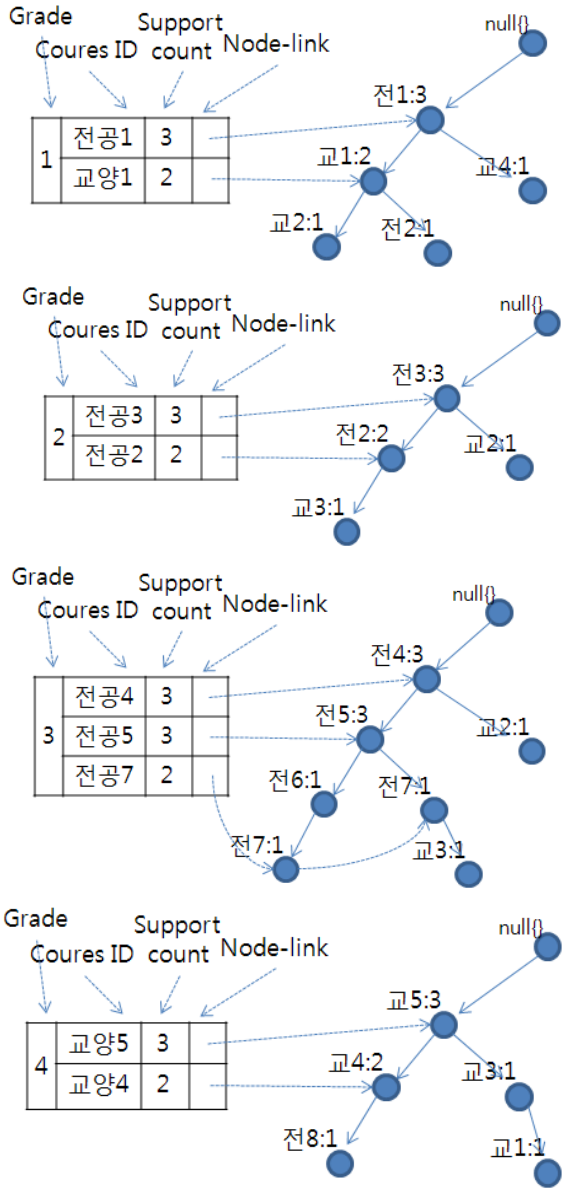


그림3. FP-tree

이후 생성된 트리의 빈발 패턴을 분석해보면 그림4와 같은 패턴 정보를 얻을 수 있다. 이는 각 트리별로 최소지지도를 만족하는 가장 빈발하게 연관된 수강신청 패턴이다. 사용자는 그림4의 수강신청 패턴 정보를 바탕으로 학년별 수강 신청을 유연하게 수행 할 수 있다.

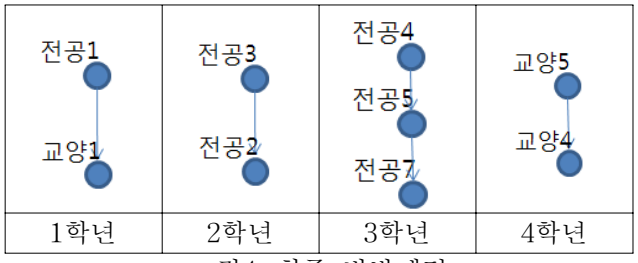


그림4. 최종 빈발패턴

4. 결론 및 향후 연구과제

최근 정보 산업 분야에서 데이터 마이닝이 주목을 받고 있다. 이는 데이터베이스 분야의 발전과 보급화에 따른 자연스러운 발전 과정이라 할 수 있다. 데이터베이스 분야에서의 초기의 관심사는 데이터 수집 및 데이터베이스 구축 기법의 발전 이었다면, 이후 질의를 통한 신속한 정보 추출 등이 차례로 관심의 대상이 되었고, 현재는 데이터를 분석하고 연관성있는 혹은 의미를 가지는 패턴을 찾아내어 기업 전략, 과학/의학 분야에서의 응용 등에 이용하는 데이터 마이닝 기법으로 그 관심분야가 진화 한 것이라 볼 수 있다. 이러한 흐름에 발맞추어 본 논문은 연관규칙을 이용한 수강신청 로드맵 제시 기법을 제안하였다. 기존의 FP-tree에서의 기본적인 데이터 베이스에 특정 항목(본 논문에서는 학년 플래그)를 추가함으로써 구조적 변화를 취했다. 이를 이용하여 현재는 제공 되지 않고 있는 수강 신청 로드맵 데이터를 학생들에게 제공 할 수 있는 발판을 제시하였다. 본 논문의 향후 연구과제로는 객관성을 고려할 수 있는 패턴 평가 모듈의 개발과 사용자의 시행착오를 줄이면서 성능을 보장하는 새로운 마이닝 기법의 연구가 진행되어야 할 것이다.

5. 감사의 말(Acknowledgment)

본 연구는 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스 컴퓨터 및 네트워크 원천기반기술 개발사업의 지원에 의한 것임

참고문헌

[1] J. Han, M. Kamber "Data Mining : Concepts and Techniques" Acadamic Press 2000

[2] R. Agrawal, T. Imielinski and A. Swami "Mining association rules between sets of items in large database" In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data(SOGMOD'93), page 207-216, Washington, DC, May 1993

[3] M. H. Dunham, Y. Xiao, L. Greenwald, Z. Hossain " ASurvey of association rules" Dallas, Texas

[4] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. ACM-SIGMOD Int'l Conf. Mangement of Data, pp 1-12, May 2000

[5] J. Han, J. Pei, Y. Yin and R. Mao, "Mining Frequent Patterns without Candidate Generation : A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, vol. 8, no. 1, pp. 53-87, 2004