연관규칙기반 Pattern Miner의 설계 및 구현

김 지현† , 성 유진† , 박 종수† , 지 원철‡
† 성신여자대학교 전산학과 ({jhkim, yjsung, jpark}@cs.sungshin.ac.kr)
‡ 홍익대학교 산업공학과 (jhee@wow.hongik.ac.kr)

요약

방대한 양의 데이터들 속에 존재하는 일관된 흐름이나 경향을 파악해 내는 데이터 마이닝에 대한 관심이 확산되고 있다. 특히 항목들 상호간의 연관성을 나타내는 연관 규칙과 시간 개념이 포함되어 항목들 사이의 순서를 찾아내는 순차 패턴의 탐사는 데이터 마이닝에서 중요한 역할을 하고 있다. 본 논문에서는 트랜잭션 데이터베이스에서 연관 규칙과 순차 패턴을 탐사하는 시스템의 설계 및 구현에 관하여 기술한다. 연관 규칙을 위해 Aproiri, DHP를, 순차패턴을 위해 AprioriAll등 기존에 연구된 대표적인 알고리즘들을 사용하였고, Windows NT상에서 Visual C++과 JAVA언어로 구현하였다. 편리한 사용자 환경 구축을 위해, 데이터의 입력 형식으로 텍스트 타입과 MDB (Microsoft Access)형태를 모두 처리할 수 있게 하였고, 출력형식은 스프레드시트이다. 입력 데이터로 실험 데이터와 통계청의 DB 이용 로그 데이터에 대하여 본 시스템 을 수행하였다.

1. 서론

데이터베이스 양이 방대해짐에 따라 많은 사람들의 관심은 DB내에 저장되어 있는 데이터에서 사용자가 사전에 예측할 수는 없지만 미리 예정된 일정한 패턴의 정보를 찾아내는 일에 집중되어졌다[1, 4, 6]. 이들 지식탐사에 관련한 규칙들은 RDBMS에서 기존의 질의 언어(SQL)로는 얻을 수 없는 결과로서 이를 위한 탐사시스템의 필요성이 대두되었다.

본 시스템은 연관 규칙(Association Rule)[2, 5, 7]이나 순차 패턴(Sequence Pattern)[3, 8]을 요구하는 질의를 처리하기 위하여 개발되었으며, 이러한 유형의 지식들은 경영 진단이나 의사결정에 매우 유용하게 이용될 수 있다. 연관 규칙 및 순차 패턴은 간략하게 다음과 같이 설명할 수 있다. 연관 규칙 탐색은 주어진 특정 사건, 거래 또는 DB의 한 레코드 내에서 동시에 발생되는 항목들(items)들을 분석해내는 것이다. 순차 패턴 탐사는 한 트랜잭션 안에서 발생하는 항목들간의 연관 규칙에 시간의 변이를 추가한 것이다. 즉 연관 규칙은 트랜잭션 안에서 어떤 항목을 함께 사는가 하는 문제로 트랜잭션 내의 문제인 반면 순차 패턴을 발견하는 것은 트랜잭션 상호간의 문제이다.

본 논문에서는 최근 관심이 고조되고 있는 DB 마케팅 분야의 장바구니 분석(Basket Analysis)에 유용한 연관 규칙과 순차패턴 탐색을 기존에 소개된 알고리즘을 이용하여 구현하였다.

본 논문의 구성은 다음과 같다. 제2장에서 연관 규칙과 순차 패턴의 정의를 살펴보고, 제3장에서는 구현된 시스템의 설계에 대하여, 제4장에서 실세계의 데이터인 통계청 데이터의 특징과 분석에 대하여, 제5장에서 결과 분석에 대하여, 그리고 6장에서 결론 및 향후과제로 맺는다.

2. 연관 규칙과 순차 패턴의 정의

2.1 연관 규칙 탐사 (Association Rule Mining)

연관규칙은 빈발 항목집합을 찾은 후 이를 기반으로 연관규칙을 생성해 내는 작업으로 나눌 수 있다. 항목들(예를 들면, 소매점에서 판매된 물품 항목들)의 집합 $I=\{i_1, i_2, \ldots, i_m\}$이 주어지면, 트랜잭션 T는 I의 부분집합으로 정의된다($T \subseteq I$). 집합과 같이 트랜잭션들은 중복된 항목을 허용하지 않는다. 트랜잭션과 다른 모든 항목집합들 내에 있는 항목들은 정렬된 것으로 가정한다. 데이터베이스 D는 n개의 트랜잭션들의 집합이고 각 트랜잭션은 고유한 트랜잭션 번호(TID)가 부여된다고 하자. 만일 트랜잭션 T가 X의 모든 항목들을 포함하면 ($X \subseteq T$), T가 집합 X(물론, $X \subseteq I$)를 "지지한다(support")고 한다. supp(X)은 X를 지지하는 D에 있는 모든 트랜잭션들의 개수를 의미한다. 만일 주어진 최소 지지도 min_supp에 대하여 supp(X) ≥ min_supp이라면, 집합 X는 빈발하다. 이 경우에 항목들의 집합 X를 일반적으로 빈발항목집합(large itemset 또는 frequent itemset)이라고 한다.

이상의 빈발항목의 모든 공집합이 아닌 부분집합들에 대하여 연관 규칙을 찾는다. 연관 규칙(association rule)은 R: $X \rightarrow Y$ 형식의 함축이고, 이때 X와 Y는 서로 같은 원소를 갖지 않는 항목집합이다: $X, Y \subseteq I$이고 $X \cap Y = \emptyset$이다. 연관 규칙은 만일 한 트랜잭션이 X를 지지한다면, 어떤 확률에 의해 Y도 지지할 것이라는 예측으로 이해될 수 있다. 이런 확률을 이 규칙의 신뢰도(conf(R)로 표시)라 한다. R의 신뢰도는 X를 지지하는 T에 대하여 Y또한 지지할 조건부 확률로 정의된다. 지지도와 신뢰도의 형식적인 표현은 다음과 같다.

$$\text{supp(R)} = \frac{supp(X \cup Y)}{\text{전체트랜잭션개수}} .$$

$$\text{conf(R)} = \frac{supp(X \cup Y)}{supp(X)} .$$

Apriori 알고리즘은 전 단계에서 빈발항목집합이었던 항목만을 대상으로 후보 항목집합을 만든다. 이 단계를 통하여 후보 항목집합의 수를 줄일 수 있다. 그리고 생성된 후보 항목집합에 대하여 데이터베이스를 읽으면서 지지도를 계산하고 빈발항목집합을 결정한다. DHP (Direct Hashing and Pruning) 알고리즘은 해쉉 기법을 이용하여 처음 단계의 후보 항목집합을 기존의 알고리즘의 방법보다 훨씬 작게 만듦으로써, 초기단계에서의 병목현상을 해결한다. 또한 매 반복 시행에서 불필요한 항목들과 트랜잭션들을 삭제하여 데이터베이스를 재구성함으로써 전체적인 성능을 개선한다.

2.2 순차패턴탐사 (Sequential Patterns Mining)

트랜잭션 데이터베이스를 D라 하고 각 트랜잭션은 고객ID, 트랜잭션 시간, 그 트랜잭션에서 구매된 항목들로 구성되었다고 하자. 그리고 같은 고객들에 대해서 같은 시간에 두개 이상의 트랜잭션은 존재하지 않는다고 가정한다. 또한 항목의 수량은 고려하지 않는다.

항목집합 $i = \{i_1, i_2, \ldots, i_m\}$에서 i_j는 하나의 항목을, 시퀀스 $s = \langle s_1, s_2, \ldots, s_n\rangle$에서 s_j는 항목집합을 나타낸다. 만일 존재하는 $i_1 < i_2 < \ldots < i_n$이고, $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, ..., $a_n \subseteq b_{i_n}$이면, 시퀀스 $\langle a_1, a_2, \ldots, a_n\rangle$은 다른 시퀀스 $\langle b_1, b_2, \ldots, b_m\rangle$에 속한다고 한다. 아래에서 "$\lesssim$"을 속해 있다는 표시로 사용한다. 예를 들어, <(3)(4 5)(8)> $\lesssim$ <(7)(3 8)(9)(4 5 6)(8)>은 (3) $\subseteq$ (3 8)이고 (4 5) $\subseteq$ (4 5 6)이고 (8) $\subseteq$ (8)이기 때문에 성립한다. 그러나 <(3)(5)>는 <(3 5)>에 속하지 않는다. 전자의 (3), (5)는 3을 구매한 후에 5를 샀다는 것이고 후자의 (3 5)는 같이 샀다는 의미이기 때문이다. 이러한 시퀀스의 집합에서 다른 시퀀스에 포함되지 않은 시퀀스를 최대 시퀀스(maximal sequence)라고 한다.

각 고객들의 트랜잭션을 시간 순서로 볼 수 있는데 이것을 소비자 순차집합(customer sequence)이라고 한다. 소비자 순차집합의 형태는 <항목집합 (T_1) 항목집합 (T_2) ... 항목집합 (T_n)>의 시퀀스이다. 한 시퀀스가 특정 고객에 대한 소비자 순차집합에 속해 있다면 그 고객은 이 시퀀스를 지지한다고 말한다. 시퀀스에 대한 지지도는 그 시퀀스를 지지하는 전체 고객들의 수이다. 주어진 고객에 대한 트랜잭션 데이터베이스 D에서 순차 패턴 탐사는 사용자가 정의한 최소 지지도를 만족하는 모든 시퀀스들 사이에서 최대 시퀀스를 찾는 것이다.

AprioriAll은 연관규칙 탐사 알고리즘인 Apriori의 응용으로서 먼저 모든 빈발 1-sequence를 찾고 이를 기반으로 기존의 데이터베이스를 새로운 데이터베이스로 변환한다. 그리고 해쉬트리를 이용하여 후보 sequence를 생성하고 변환된 데이터베이스로부터 고객 sequence를 읽으면서 지지도를 계산한다. 이 과정은 빈발 sequence를 찾아나가면서 반복된다. 모든 빈발 sequence가 찾아지면 최대 시퀀스를 찾는다.

3. Pattern Miner 시스템의 설계

Patttern miner 시스템은 크게 전처리 과정(Preprocessing), 패턴 탐색(Pattern mining), 후처리 과정(Postprocessing)으로 나눌 수 있다. 시스템의 구조도는 그림 3.1과 같다.

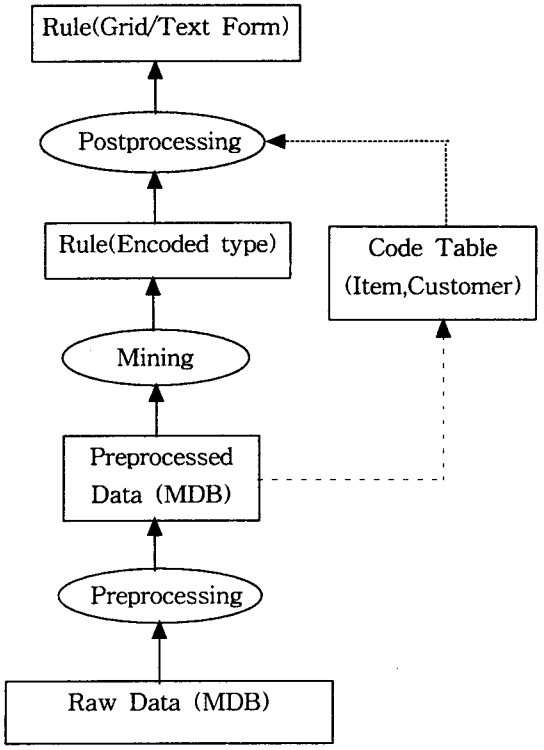


그림 3.1: 시스템의 전반적인 구조도

3.1 Preprocessing

가공되지 않은 원래 데이터의 전처리 과정으로서 DB 마이닝의 가장 첫 번째 단계이자 가장 시간을 요하는 단계이다. 불필요한 데이터 및 필드를 소거함으로써 데이터의 크기를 줄인다. 또한 데이터 수집 과정에서 발생한 noise를 제거하거나 적절한 값으로 대치한다. 또한 트랜잭션에 존재하는 중복된 항목들은 하나의 항목만을 남기고 제거한다. 데이터가 단지 시간의 순서에 의해 축적된 경우 적당한 시간 윈도우(Time window)로 잘라서 하나의 트랜잭션으로 간주하여 처리한다. 따라서 사용자의 요구에 맞추어 트랜잭션 크기가 결정된다. 마지막으로 데이터베이스를 구성하고 있는 항목들을 mining단계에서 처리할 수 있도록 적당한 정수형으로 인코딩한다. 이때 구성된 매핑 테이블은 파일 형태로 저장하여 후처리 단계에서 활용한다. 4.2는 통계청 데이터에 대한 전처리 과정에 대하여 설명한다.

3.2 Pattern Mining

Pattern mining은 마이닝 시스템의 핵심으로서 각종 규칙들을 찾아낸다. 전처리된 데이터베이스는 이 과정의 입력이 된다. 연관 규칙(association rule), 순차 패턴(sequential pattern), 일반화된 연관규칙(generalized association rule), 일반화된 순차패턴 (generalized sequential pattern), 클러스터링(clustering), 유사도 탐색(similarity searching)등의 규칙들이 있다. 본 시스템에서 초점을 둔 것은 연관 규칙과 순차 패턴 탐색으로 Apriori, DHP, AprioriAll등의 기존에 연구되어진 알고리즘을 이용하였다. 현재 데이터의 입력형태는 MDB와 텍스트 타입의 데이터가 모두 가능하며, 특히 MDB의 핸들링을 위해서 Jet engine을 이용한 DAO클래스 (MS Visual C++5.0 제공)를 사용한다. 마이닝 후

얻어진 결과는 텍스트 형태로 출력되며 이는 후처리과정에서 사용자 요구에 따라 적합한 형태로 시각화하거나 텍스트 파일 형태로 저장한다. 규칙에 따라서 적절한 measurement값(support, confidence, conviction, interest)을 함께 출력하여 사용자가 의사결정에 참고할 수 있게 하였다.

3.3 Postprocessing

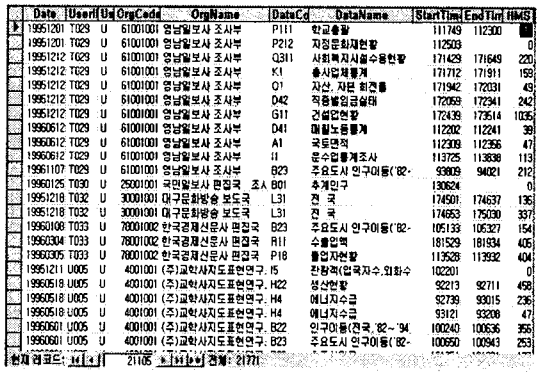
이 시스템은 mining후 얻어진 결과를 두 가지 형태로 출력한다. 한가지는 Grid 형태이며 다른 하나는 텍스트 파일 형태로 저장한다. Grid형태는 MS Visual C++에서 제공하는 ActiveX 컨트롤로서 Formula1이라는 프로그램을 이용하였다. 이 과정에서는 전처리 단계에서 생성된 코드 테이블과 pattern mining 단계에서 얻어진 규칙을 입력으로 사용한다. 4.4는 본 시스템의 후처리 과정을 설명한다.

4. 시스템의 구현

실험 데이터[2]를 생성하여 구현된 시스템을 점검하였고, 실제 데이터로는 통계청의 DB 이용 로그를 사용하였다. 다음은 실제 데이터를 적용한 경우의 데이터의 특징 및 패턴 탐사의 과정을 사용자 인터페이스와 함께 보여준다.

4.1 데이터의 특징

통계청의 DB 이용 로그 데이터는 1995년 12월 1일부터 1996년 11월 30일(1개년도), 1996년 12월 1일부터 1997년 11월 30일(2개년도)까지 각 기관에서 이용한 통계청 자료에 대한 로그 데이터이다. 단지 시간의 순서에 의해서 누적된 데이터이므로 전처리 과정에서 적당한 시간 간격으로 전체 데이터를 트랜잭션화할 수 있도록 하였다. 예를 들어, 시간 간격(Time Window)을 7일로 주었을 때, 전체 트랜잭션 수는 21,771개이고 이용자수는 297명이며, 이용 자료는 325건이다. 이들은 MDB형식으로 저장된다. 그림 4.1은 전처리과정 전의 데이터베이스를 보여준다.



Date	UserID	UG	OrgCode	OrgName	DataCd	DataName	StartTime	EndTime	HMS
19951201	T029	U	61001001	영남일보사 조사부	P111	학교총람	111749	112300	
19951201	T029	U	61001001	영남일보사 조사부	P212	지정문화재현황	112503		0
19951212	T029	U	61001001	영남일보사 조사부	Q311	사회복지시설수용현황	171429	171649	220
19951212	T029	U	61001001	영남일보사 조사부	K1	총사업체통계	171712	171911	159
19951212	T029	U	61001001	영남일보사 조사부	O1	자산, 자본 최종률	171942	172031	49
19951212	T029	U	61001001	영남일보사 조사부	D42	적정불입금실태	172059	172341	242
19951212	T029	U	61001001	영남일보사 조사부	G11	건설업현황	172439	173514	1035
19960512	T029	U	61001001	영남일보사 조사부	D41	매칠노동통계	112202	112241	39
19960512	T029	U	61001001	영남일보사 조사부	A1	국토면적	112309	112356	47
19960512	T029	U	61001001	영남일보사 조사부	I1	운수업통계조사	113725	113838	113
19961107	T029	U	61001001	영남일보사 조사부	B23	주요도시 인구이동('82-	93809	94021	212
19960125	T030	U	25001001	국민일보사 편집국 조사	B01	추계인구	130624		0
19951218	T032	U	30001001	대구문화방송 보도국	L31	전 국	174501	174637	136
19951218	T032	U	30001001	대구문화방송 보도국	L31	전 국	174653	175030	337
19960108	T033	U	78001002	한국경제신문사 편집국	B23	주요도시 인구이동('82-	105133	105327	154
19960304	T033	U	78001002	한국경제신문사 편집국	R11	수출입액	181529	181934	405
19960305	T033	U	78001002	한국경제신문사 편집국	P18	들업자현황	113528	113932	404
19951211	U005	U	4001001	(주)교학사지도표현연구.	I5	관광객(입국자수,외화수	102201		0
19960518	U005	U	4001001	(주)교학사지도표현연구.	H22	생산현황	92213	92711	458
19960518	U005	U	4001001	(주)교학사지도표현연구.	H4	에너지수급	92739	93015	236
19960518	U005	U	4001001	(주)교학사지도표현연구.	H4	에너지수급	93121	93208	47
19960601	U005	U	4001001	(주)교학사지도표현연구.	B22	인구이동(전국,'82~'94	100240	100636	356
19960601	U005	U	4001001	(주)교학사지도표현연구.	B23	주요도시 인구이동('82-	100650	100943	253

그림 4.1: Original database

(* 각 필드별 설명.
- Date: 이용 일자
- UserID: 이용자 ID
- UserGroup: 이용 그룹
- OrgCode: 그룹 내 기관 코드
- OrgName: 기관 명
- DataCode: 이용 자료 코드
- DataName: 이용 자료 명
- StartTime: 이용 시작 시간(시/분/초)
- EndTime: 이용 종료 시간(시/분/초)
- HMS: 이용 시간)

4.2 전처리 과정

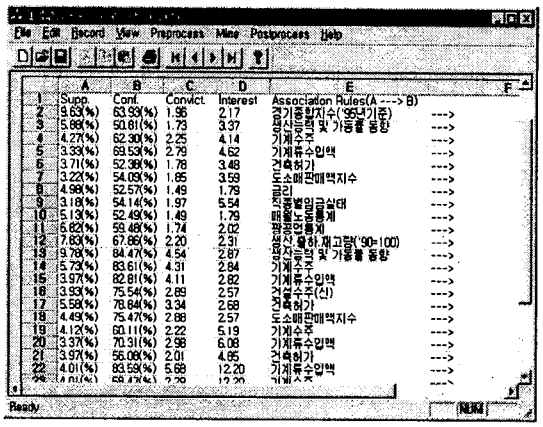
전처리 과정에서는 데이터에 순서를 매긴 형태의 가공된 데이터베이스와 함께 각 항목에 대한 매핑 테이블을 생성한다. 이 테이블은 후처리 과정에서 규칙을 구성하는 항목들을 실제의 항목들로 변환 시에 이용된다. 본 데이터의 특징상 고객과 이용 자료 항목에 관한 두개의 매핑테이블이 존재하며 이들은 데이터의 사전적인 순서에 의해 정렬하고, 정수 1부터 차례대로 코드값를 부여하여 1:1 로 대응시킨다. 고객에 관한 매핑테이블은 UserGroup와 OrgCode를 오름차순으로 정렬하고, 1부터 차례대로 코드 값를 부여하여 CustomerID와 1:1 대응시킨 결과이다. 이용 자료에 대한 매핑 테이블은 DataCode를 오름차순으로 정렬하고, 차례로 1부터 번호를 부여하여 ItemID와 1:1 대응시킨 결과이다.

4.3 패턴 탐사

전처리과정을 거친 데이터는 본격적으로 패턴 마이닝의 입력으로 사용되며 세 가지 알고리즘을 선택할 수 있다. Apriori, DHP는 연관규칙 탐사를 위한 알고리즘이며 입력 파라미터로 최소 지지도와 최소 신뢰도를, AprioriAll은 순차패턴을 위한 알고리즘으로 최소 지지도만을 필요로 한다.

4.4 후처리 과정

생성된 규칙들을 실제의 항목들로 해독하는 단계로서 전처리 과정에서 생성된 매핑 테이블을 참조하여 이루어지며 결과는 두 가지 형태, 즉 Grid 형식과 텍스트 파일 형식 중 사용자가 원하는 형태로 생성한다. 그림 4.2는 최소지지도 30%, 최소 신뢰도 50%일 때 추출된 연관 규칙들을 Grid형식으로 보여준다. 그림 4.3은 최소 지지도를 12%로 주었을 때의 추출된 순차패턴을 Grid형식으로 보여주고 있다.



	A	B	C	D	E	F
	Supp.	Conf.	Convict.	Interest	Association Rules(A ---> B)	
1	9.63(%)	63.99(%)	1.96	2.17	경기종합지수('95년기준)	--->
2	5.66(%)	50.81(%)	1.73	3.37	음식점등록 및 가동률 동향	--->
3	4.27(%)	62.30(%)	2.25	4.14	기계수주	--->
4	3.33(%)	69.53(%)	2.79	4.62	기계류수입액	--->
5	3.71(%)	52.39(%)	1.78	3.48	건축허가	--->
6	3.22(%)	54.09(%)	1.85	3.59	도소매판매액지수	--->
7	4.98(%)	52.57(%)	1.49	1.79	금리	--->
8	3.18(%)	54.14(%)	1.97	5.54	직종별임금실태	--->
9	5.13(%)	52.49(%)	1.49	1.79	매월노동통계	--->
10	6.82(%)	59.48(%)	1.74	2.02	원공인통계	--->
11	7.83(%)	67.66(%)	2.20	2.31	생산 물하,재고량('90=100)	--->
12	3.78(%)	64.47(%)	4.54	2.87	음식점등록 및 가동률 동향	--->
13	5.73(%)	83.61(%)	4.31	2.84	기계수주	--->
14	3.97(%)	82.81(%)	4.11	2.82	기계류수입액	--->
15	3.93(%)	75.54(%)	2.89	2.57	건설수주(신)	--->
16	5.58(%)	78.64(%)	3.34	2.68	건축허가	--->
17	4.49(%)	75.47(%)	2.88	2.57	도소매판매액지수	--->
18	4.12(%)	60.11(%)	2.22	5.19	기계수주	--->
19	3.37(%)	70.31(%)	2.98	6.08	기계류수입액	--->
20	3.97(%)	56.06(%)	2.01	4.85	건축허가	--->
21	4.01(%)	83.59(%)	5.68	12.20	기계류수입액	--->
22	4.01(%)	68.27(%)	2.79	17.20	기계수주	--->

그림4.2: 후처리과정후 추출된 연관규칙(Grid형식)

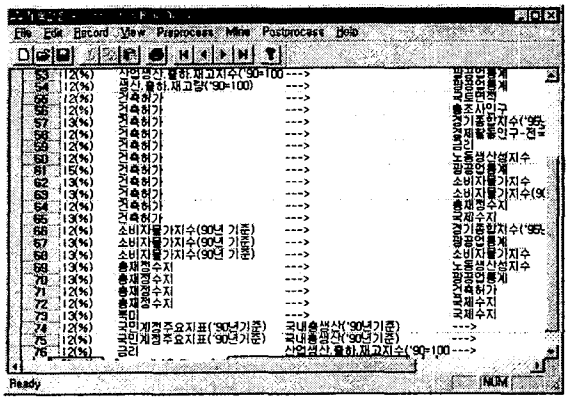


그림 4.3: 후처리 과정후 추출된 순차패턴(Grid형식)

5. 결과 패턴 분석

표 5.1은 최소지지도 2%, 최소 신뢰도 50%일 때의 연관 규칙 탐사의 결과 예이다. 이 가운데 "생산능력 및 가동률동향" --> "산업생산 .출하. 재고지수('90=100')" 라는 규칙은 9.67%의 지지도와 83.77%의 신뢰도, 4.35의 Conviction, 2.85의 Interest값을 가진다. Conviction은 규칙을 구성하는 전제부와 결과부 항목집합들 사이의 의존성을 측정키 위한 것으로 1보다 커질수록 관계성이 크다. Interest는 규칙을 구성하는 전제부와 결과부 항목집합들이 얼마나 함께 나타나는가를 측정하는 것으로 1보다 클수록 높은 것이다. 표 5.2는 최소 지지도를 12%로 주었을 때의 순차패턴이다. 예로써 "생산자물가지수"--->"산업생산.출하.재고지수('90=10')"은 전체 고객 중 13%가 "생산자물가 지수"를 참조한 후 다음 트랜잭션에서 "산업생산.출하.재고지수('90=100')"을 참조하였음을 의미한다.

표 5.1: 연관 규칙의 예

연관 규칙	Supp(%)	Conf(%)	Con.	Inter.
"국민계정주요지표('90년기준)" ---> "국내총생산('90년기준)"	2.96	50.32	1.87	7.25
"기계수주" ---> "경기종합지수('95년기준)"	4.27	62.30	2.25	4.14
"생산능력 및 가동률 동향" --->"산업생산.출하.재고지수('90=100')"	9.67	83.77	4.35	2.85
"산업생산.출하.재고지수('90=100')"생산자물가지수"---> "소비자물가지수"	2.85	57.14	1.97	3.71

표 5.2: 순차 패턴의 예

순차 패턴	지지도
"생산자물가지수"---> "경기종합지수('95년기준)"	12(%)
"생산자물가지수" ---> "산업생산.출하.재고지수('90=100')"	13(%)
"생산자물가지수" ---> "소비자물가지수(90년 기준)"	13(%)

6. 결론 및 향후 과제

본 논문에서는 연관규칙과 순차패턴을 탐색하기 위해 널리 알려진 알고리즘인 Apriori와 DHP 그리고 AprioriAll을 Visual C++ 및 JAVA를 이용하여 구현하였다. 실험데이터와 실세계의 데이터인 통계청 DB 로그가 입력 데이터베이스로 쓰여졌다. 편리한 사용자 인터페이스를 위해 Grid와 텍스트 형식으로 결과를 보여주고 있다.

향후 각 패턴의 효율적인 탐사를 위해 보다 개선된 알고리즘을 추가적으로 구현한다. 또한 일반화된 연관규칙(generalized association rule), 일반화된 순차패턴(generalized association rule), 일반화된 순차패턴(generalized sequential pattern)등 taxonomy를 적용한 패턴의 추출 등으로 확장해 나갈 것이다.

참고 문헌

[1] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", *Proceedings of ACM SIGMOD*, pp. 207-216, 1993.

[2] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules", *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 478-499, 1994.

[3] R. Agrawal and R. Srikant, "Mining Sequential Patterns", *Proceedings of the 11th IEEE ICDE*, pp. 3-14, 1995.

[4] A. Berson and S.J. Smith, *Data Warehousing, Data Mining, and OLAP*, McGraw-Hill, NewYork, 1997.

[5] S. Brin, Rajeev Motwani, Jeffrey D. Ullman and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", *Proceedings of ACM SIGMOD*, pp. 255-264, 19937

[6] M.S. Chen, J. Han, and P.S. Yu, "Data Mining" An Overview from a Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 886-883, Dec. 1996.

[7] J. S. Park, M. S. Chen and P. S. Yu, "An Effective Hash-Based Algorithm for Mining Association Rules", *Proceedings of ACM SIGMOD*, pp. 175-186, 1995.

[8] R. Srikant and R. Agrawal, "Mining Generalized Association Rules",*In Proceedings of the 21st VLDB Conference*, Zurich, Swizerland, 1995.