

고속도로 연계도로 정보 데이터베이스를 이용한 교통체증 마이닝에 관한 연구

이기성*

*호원대학교 컴퓨터학부

e-mail:ygslee@sunny.howon.ac.kr

A Study of Traffic Mining used High expressway Connection Road Information Database

Gi-Sung Lee*

*Dept of Computer Science, Howon University

요 약

교통 체증이나 도로의 속도를 이전의 통계를 이용하여 예측할 수 있다면 상당히 도움이 될 것이다. 본 논문은 다양한 종류의 도로 중 고속도로의 속도에 영향을 주는 요소를 분석하여 상호 영향을 주는 요소를 고찰한다. 이를 수행하기 위해 고속 도로 교통에 대한 데이터베이스를 구축하며, 도로 교통 데이터베이스에 연계도로와 관계를 적용하고, 다양한 데이터 마이닝의 연산을 사용하여 결과를 도출한다.

1. 서론

현재 교통 체증으로 인해 차의 평균 시속이 해마다 줄고 있다. 또한, 우천이나 눈이 내려 도로의 사정이 좋지 않을 때는 차의 평균 시속이 현저히 감소된다. 이러한 도로의 상황을 단순한 통계나 관측을 이용하여 일반 운전자에게 숙지할 수 있는 정보로 가공하기는 많은 어려움이 있다. 하지만, 특정 도로의 교통정보를 특정 주기로 데이터베이스에 구축하여 원시자료를 작성하고, 그 데이터를 이용해 가설을 설립하고, 가설에 대해 마이닝의 다양한 연산(클러스터링, 연관화 등등)을 적용하면 데이터의 연관관계나 분포, 밀접성들의 결과를 쉽게 도출하여 자동차의 속도에 영향을 받는 속성들을 유추하여 분석할 수 있다. 본 논문은 이러한 도로에 대한 속성들 간의 관계를 유추하기 위해 도로에 대한 교통 정보 데이터베이스를 구축하며, 가설을 설립하고, 데이터의 연관관계와 속성을 유추하여 속도에 영향을 주는 요소들을 도출한다. 또한 방대한 데이터의 자료로

인한 오차율을 막기 위해 많은 도로 중 고속도로에 대한 교통 정보 데이터를 이용한다.

2. 데이터베이스 구조

마이닝에 사용할 데이터베이스는 현재 도로교통망 정보서비스에서 사용하고 있는 데이터베이스로서 DBMS로는 오라클 8(Oracle 8)을 사용한다. 구축된 데이터베이스 시스템의 특성은 다음과 같다.

(1) 원시 Database 개요

- 특정 기업의 고속도로 정보 서비스를 위한 데이터베이스를 사용.
- 데이터베이스에는 월요일부터 일요일까지의 일주일 분량의 정보가 저장.
- 5분 단위로 새로운 정보가 추가
- 전국 20개의 고속도로 중 경부고속도로(상/하행)만 추출하여 데이터베이스를 새롭게 구성.
- 경부고속도로는 56개의 구간으로 분리되어 있고,

이중 29개의 구간을 중점적으로 사용.

- 각 구간은 인터체인지(IC), 분기점(JC), 톨게이트(TG), 휴게소를 기준으로 구분.
- DBMS는 Oracle8를 사용.

(2) 마이닝에 사용된 테이블 구조

- ddate : 날짜
 - ttime : 시간
 - link_id : 구간 id
 - from_node : 시작지점 id
 - to_node : 도착지점 id
 - congestion_grade : 일반도로 상태 id
 - speed : 일반도로 속도 (km/h)
 - travel_time : 일반도로 소요시간 (초)
 - bus_congestion_grade : 버스전용도로 상태 id
 - bus_speed : 버스전용도로 속도 (km/h)
 - bus_travel_time : 버스전용도로 소요시간 (초)
 - suspension : 차단통제 정보 id
 - announcement : 공지사항 id
 - weather : 도로기상 id
 - queue_length : 지체길이 (m)
- * id로 표기되는 것은 별도의 table이 존재하기 때문에 비교하여 확인.

(3) 전처리

전처리는 원시 자료를 데이터베이스에 로드해서, 데이터베이스에서 작업을 수행하기 위한 형태로 변경시켜 주는 작업을 수행한다. 전처리 작업 모듈은 프로그램 언어 중 하나인 java를 이용하여 구축하였다. 전처리 작업을 마친 후의 데이터는 다음과 같다.

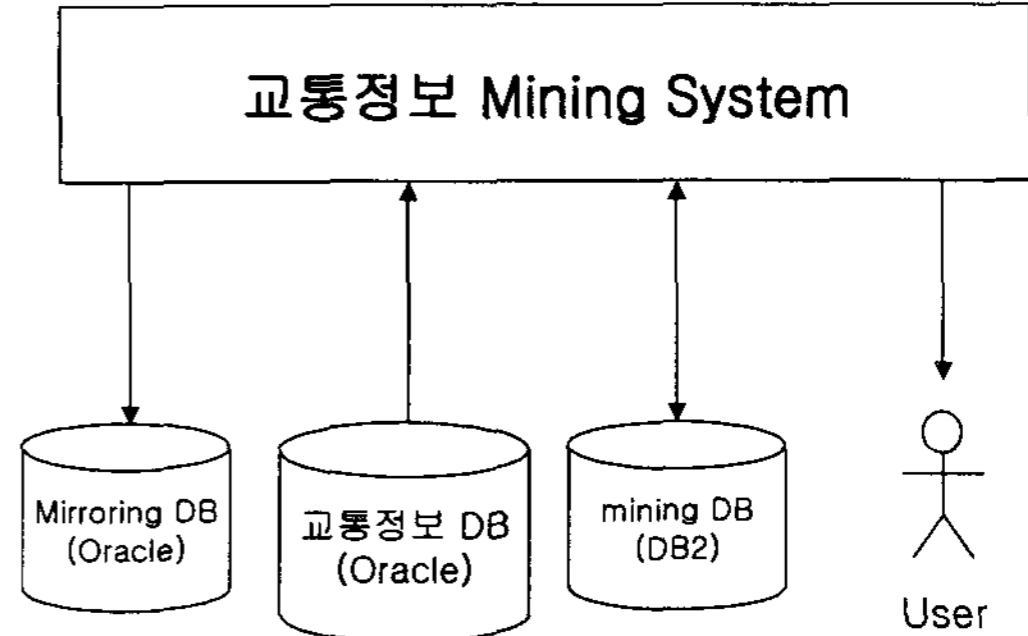
```

"20000508","1655","KDT10010","KKN10001","KN10002","1",100,116,"0",
0,"0000","G102","E109",0
"20000508","1655","KDT10020","KKN10002","KN10003","2",68,126,"0",
0,"0000","G103","E109",0
"20000508","1655","KDT10030","KKN10003","KN10004","1",90,101,"0",
0,"0000","G103","E109",0
"20000508","1655","KDT10040","KKN10004","PN10005","1",94,310,"0",
0,"0000","G102","E109",0
"20000508","1655","KDT10050","KPN10005","KN10006","1",100,41,"0",
0,"0000","G102","E109",0
"20000508","1655","KDT10060","KKN10006","KN10007","1",99,131,"0",
0,"0000","G102","E109",0
"20000508","1655","KDT10070","KKN10007","KN10008","1",95,169,"0",
0,"0000","G102","E109",0
"20000508","1655","KDT10080","KKN10008","YN10009","1",95,144,"0",
0,"0000","G102","E109",0
    
```

3. 시스템 구조도

본 논문의 마이닝 작업은 오라클로 구축된 데이터베이스로부터 작업하기 용이한 별도의 시스템에 같

은 데이터를 미러링(mirroring)하여 작업하며, 또한 마이닝에 사용할 도구인 인텔리전트 마이너(IM)을 사용하기 위하여 DB2 DBMS에서 이식하여 작업한다. 개념도의 대략은 아래 그림과 같다.



4. 데이터 마이닝을 이용한 분석 및 결과

가설 '가'지역의 속도가 A가 되면 '나'지역도 속도가 A가 되는가?

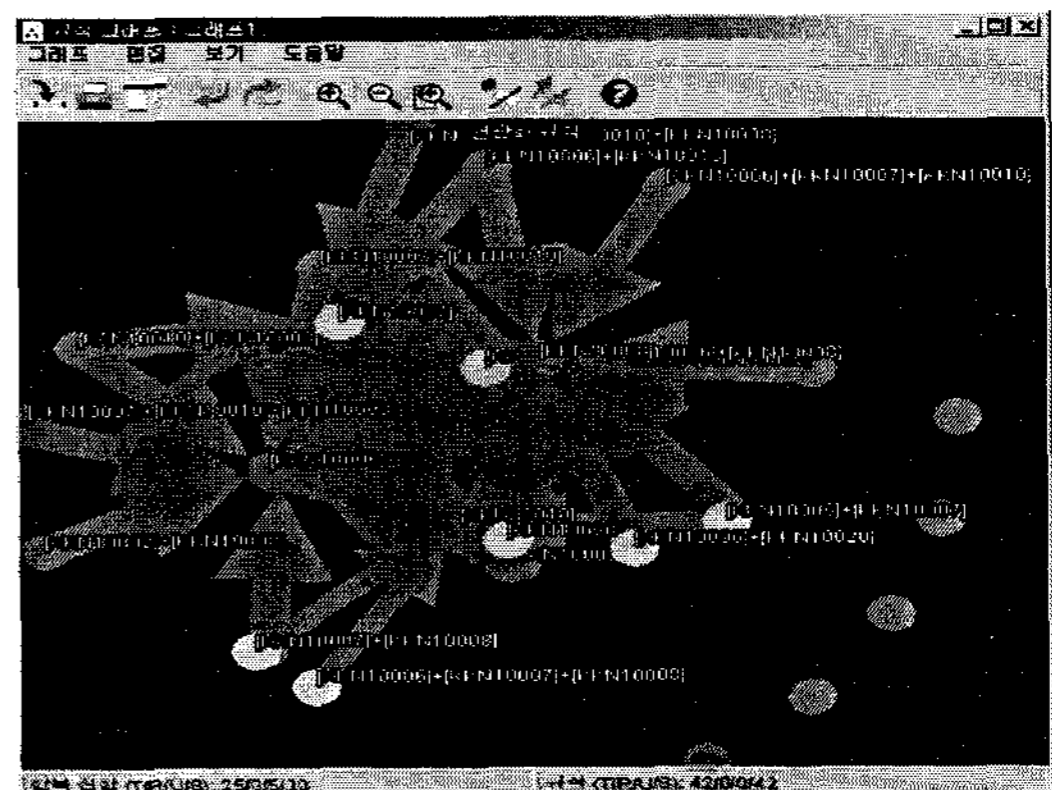
고속도로에서 운전을 하다보면 어느 구간에서 속도가 저속도가 나오면 저속도가 똑같이 나는 구간이 존재 할 것이다라는 생각에서 위 가설을 만들었다. 가설을 검증하는 방법은 아래의 방법으로 실행한다.

- 1) 하행선에 대한 구간에 대하여 연관화를 실행
- 2) 상행선에 대한 구간에 대하여 연관화를 실행

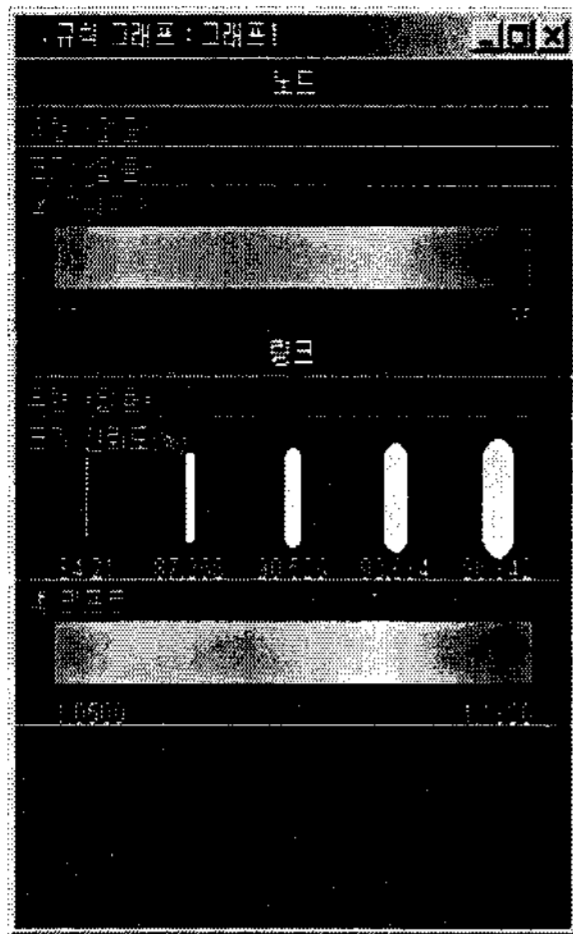
■ 연관화 1 : 하행선에 대하여 연관화를 실행한다.

- 방법: 1. 속도가 연속된 값으로 되어 있어서 연관성을 찾는데 많은 트랜잭션이 생기게 되므로 속도에 대해서 이산화 작업을 수행한다.
- 2. 속도는 0부터 200까지 5차이로 한다.
- 3. 추가 자료 후, 이산화 된 속도에 구간을 연관화한다.

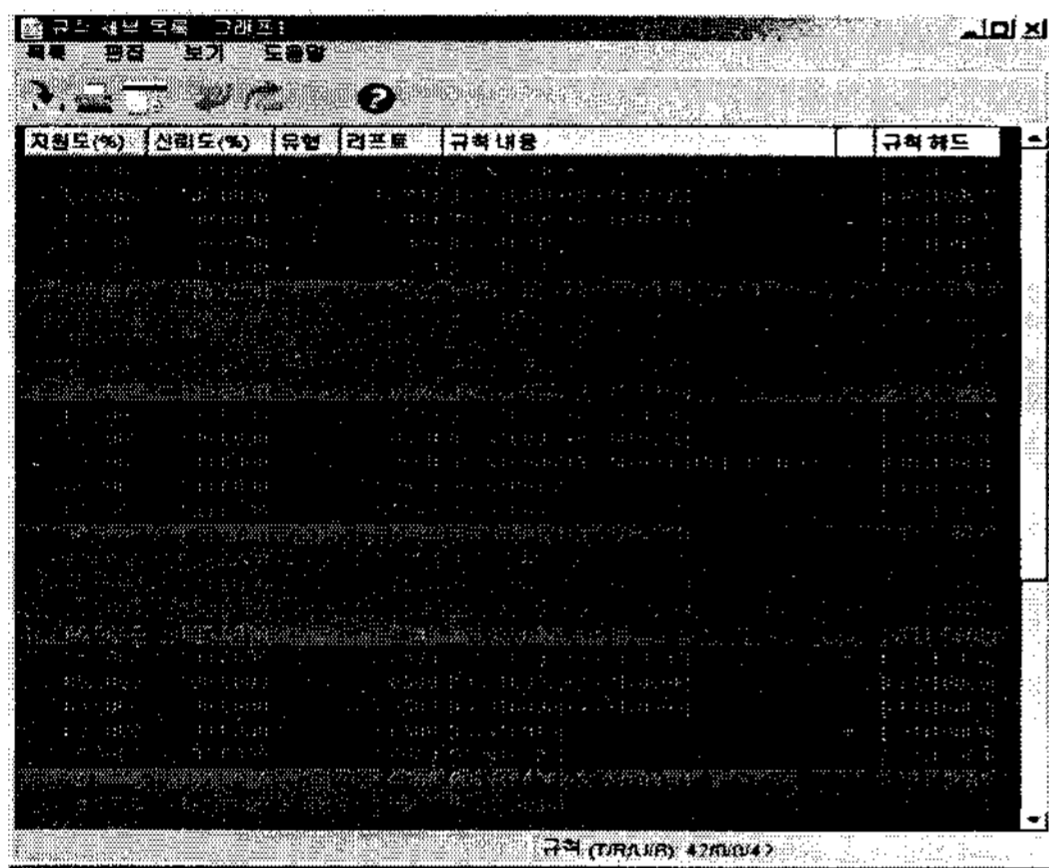
-결과 화면



- 자료1 -



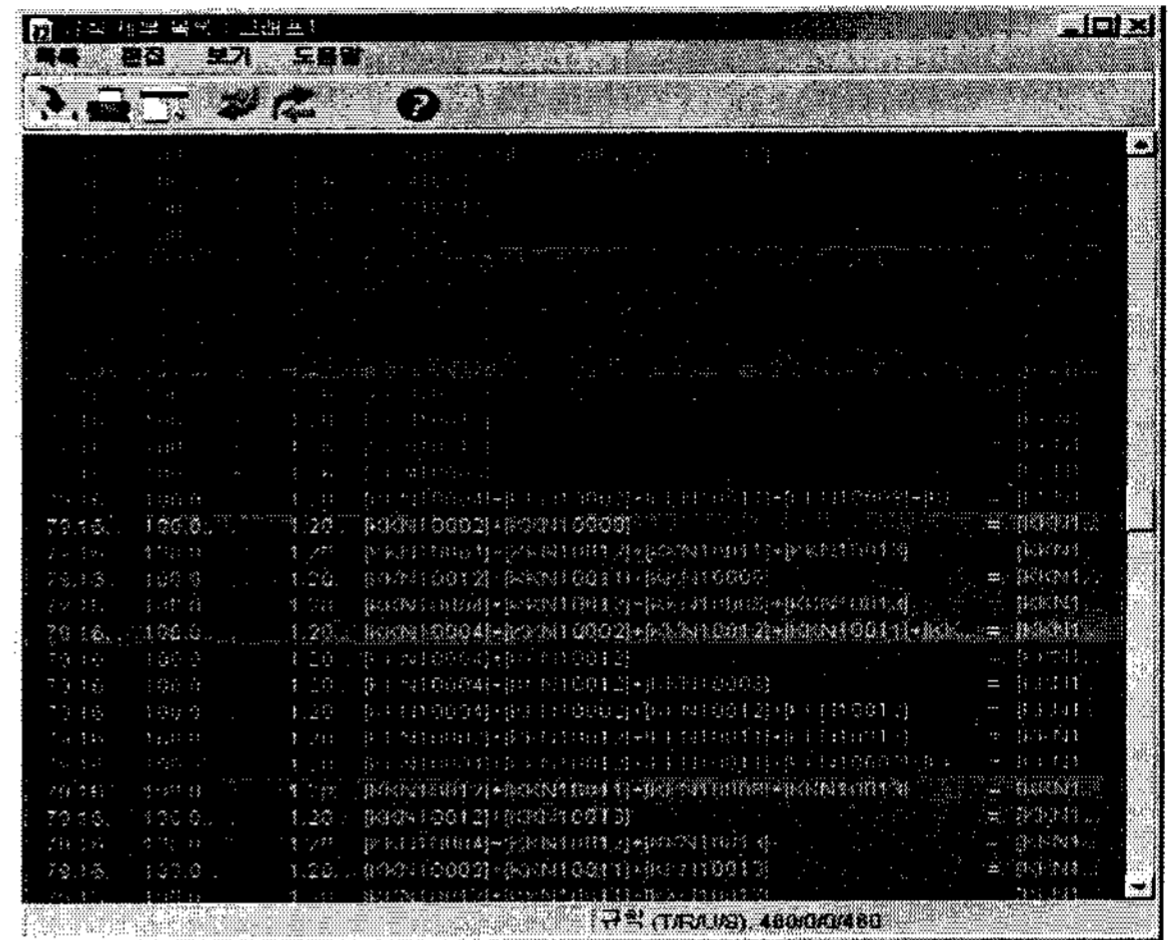
- 자료2 -



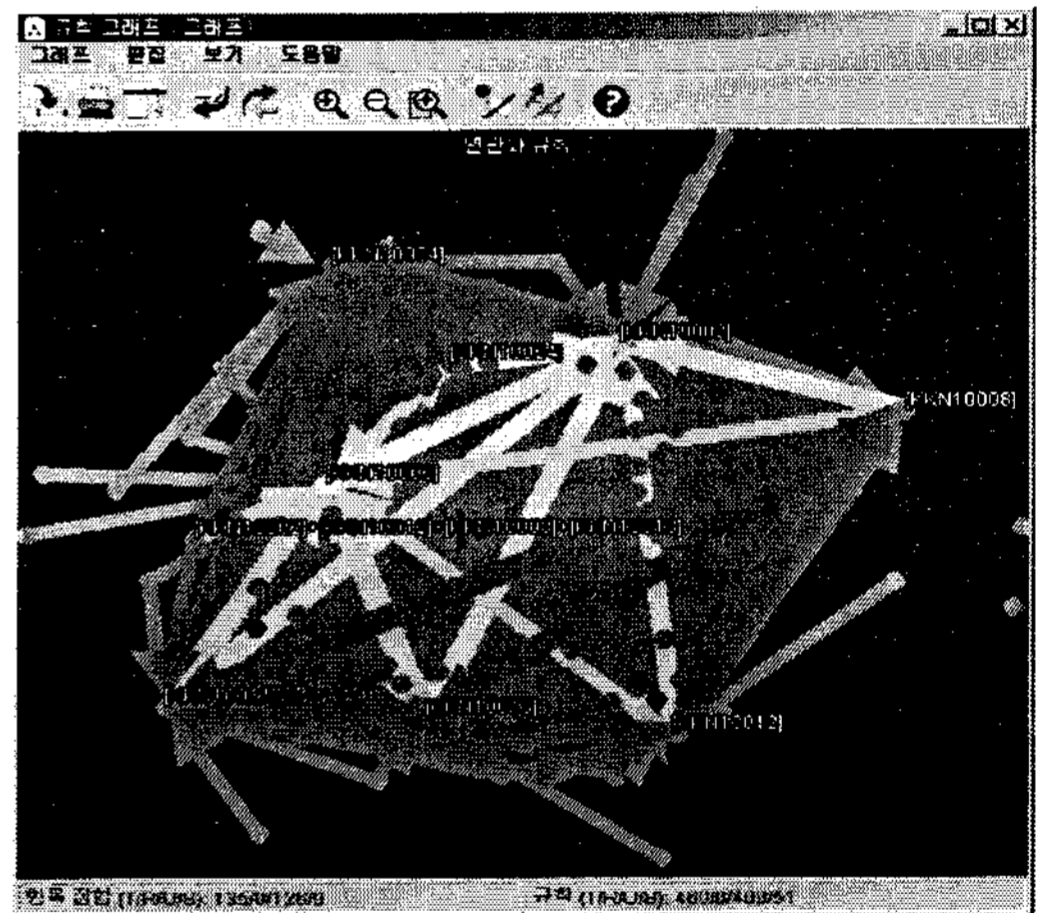
알 수 있다.

■ 연관화 2: 상행선에 대하여 연관화를 실행한다

- 방법: 1. 연관화1의 과정과 같다.
- 결과 화면



- 자료1 -



- 자료2-

- 자료3 -

(KKN10006 : 판교IC, KKN10010 : 수원IC, KKN10008 : 죽전휴게소, KKN10007 : 서울TG)

- 결과고찰

위 자료3에 의하면

판교IC, 수원IC, 죽전휴게소의 속도가 같으면 서울TG도 일정속도가 된다.

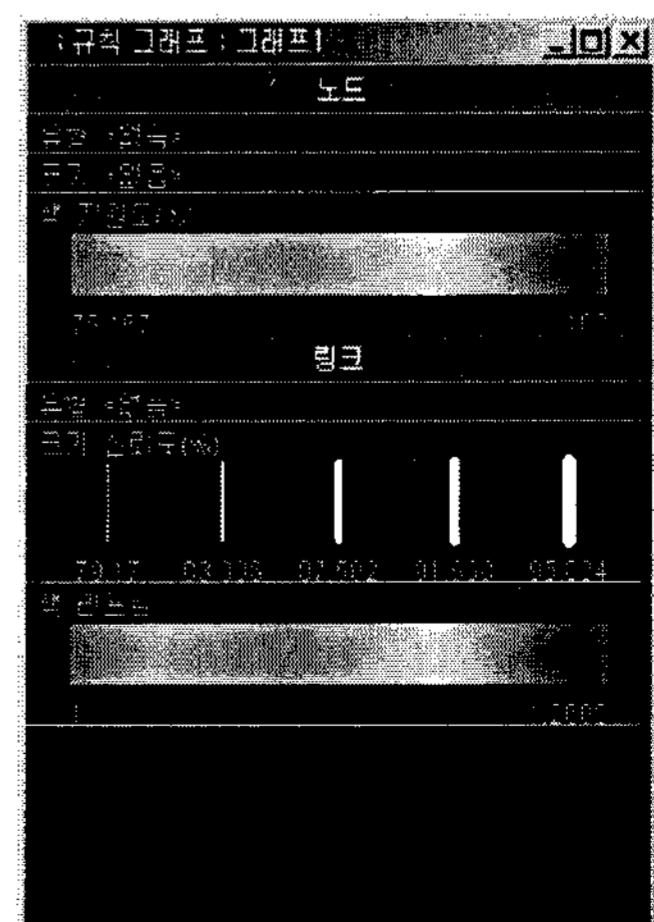
수원IC, 죽전휴게소의 속도가 같으면 서울TG도 일정속도가 된다.

판교IC, 수원IC의 속도가 같으면 서울TG도 일정속도가 된다.

수원IC의 속도로 서울TG의 속도를 알 수 있다.

따라서 위 4개의 지역은 서로 많은 연관관계를 가지고 있는 것을 알 수 있다. 서울TG의 속도를 빠르게 하려면 위 3곳의 속도를 빠르게 하면 될 것이라고 본다.

결론적으로 하행선에서는 연관화가 존재한다는 것을



- 자료3 -

(KKN10011 : 기흥휴게소, KKN10012 : 기흥IC,
KKN10013 : 오산IC, KKN10008 : 죽전휴게소)

- 결과 고찰

위 자료1과 자료2를 보면 기흥휴게소의 속도와 죽전휴게소의 속도가 관계가 있다.
죽전휴게소의 속도와 오산IC의 속도가 관계가 있다.
따라서 상행선에서도 속도에 대하여 연관성이 있는 구간이 있다.

5. 결론

본 논문은 교통 정보 데이터베이스에 데이터 마이닝을 적용해서 고속도로의 속도에 영향을 주는 요소를 도출하고 있으며 정보 데이터베이스 스키마, 데이터 인스턴스, 시스템 구조도를 내포하고 있다. 교통 정보 데이터베이스는 도로의 상태, 날씨, 구간, 기상등의 정보를 포함하고 있으며, 시스템 구조는 이종간의 시스템간의 작업을 처리하기 위해 전처리 과정을 수행할 수 있는 컨버터를 구축하였으며, 컨버터는 데이터 마이닝을 하기 위한 기본 자료를 생성하여 준다. 데이터 마이닝을 수행하기 위해 세 가지 가설을 설정하였으며, 가설에 적합한 마이닝 연산을 적용하여 결과를 도출하였다.

마이닝의 가설은 '두 지역간의 속도가 연관이 있는가'이며, 이를 위해 두 가지 방법(① 하행선에 대한 구간에 대하여 연관화를 수행한다. ②상행선에 대한 구간에 대하여 연관화를 수행한다.)을 적용하였고, 상행선과 하행선 모두 상호간의 구간이 속도에 밀접하게 영향을 주는 지역이 많이 있음을 발견하였다. 마지막 마이닝의 가설은 '톨게이트 인접 지역은 평균 속도가 40km/h 이하이다.'이며, 이를 위해 두 가지 방법(①전날 전 시간대의 평균 속력을 구하는 방법 ②모든 날의 시간 당 평균 속력을 구하는 방법)에 의하여 톨게이트 인접 지역의 속도는 40km/h이하가 아님을 도출하였다.

참고문헌

[1] "IBM DB2 Intelligent Miner for Data", IBM corp., 1999.
[2] I. Witten, E. Frank, "Data Mining", Morgan Kaufmann Publishers, 1999

[3] "Oracle Administration Handbook", Oracle press., 198.
[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview. In Advances in Knowledge Discovery and Data Mining", pp. 1-34. AAAI Press, Menlo Park, CA, 1996.
[5] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh, "Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals", Data Mining and Knowledge, 1997.
[6] M. Chen, J. Han, and P. Yu, "Data mining: An overview from database perspective", IEEE Transactions on Knowledge and Data Eng., 8(6):866--883, December 1996.
[7] M. Holsheimer, M. Kersten, H. Mannila, and H. Toivonen, "A perspective on databases and data mining", In 1st Intl. Conf. Knowledge Discovery and Data Mining, Aug. 1995.