고속도로 연계도로의 스피드 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 Speed 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 연계도로의 속도와 관계을 적용하고, 다양한 데이터 마이닝의 연산을 사용하여 결과를 도출한다.

1. 서론

근대사회에서 현대 사회로 발전됨에 따라, 많은 교통수단이 이용되었고, 그 중에서도 차를 이용한 수단이 발달되면서, 많은 사람들이 차를 소유하게 되었다. 차가 증가함에 따라, 교통은 혼잡하게 되고, 교통 체증은 더욱 심화된다. 특정 도로의 교통정보를 특정 주기로 데이터베이스에 구축하여 원시자료를 작성하고, 그 데이터를 이용해 가설을 설립하고, 가설에 대해 마이닝의 다양한 연산(클러스트링, 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 속성들간의 관계를 유추하기 위해 도로에 대한 교통 정보 데이터베이스를 구축하며, 가설을 설립하고, 데이터의 연관관계와 속성을 유추하여 속도에 영향을 주 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.

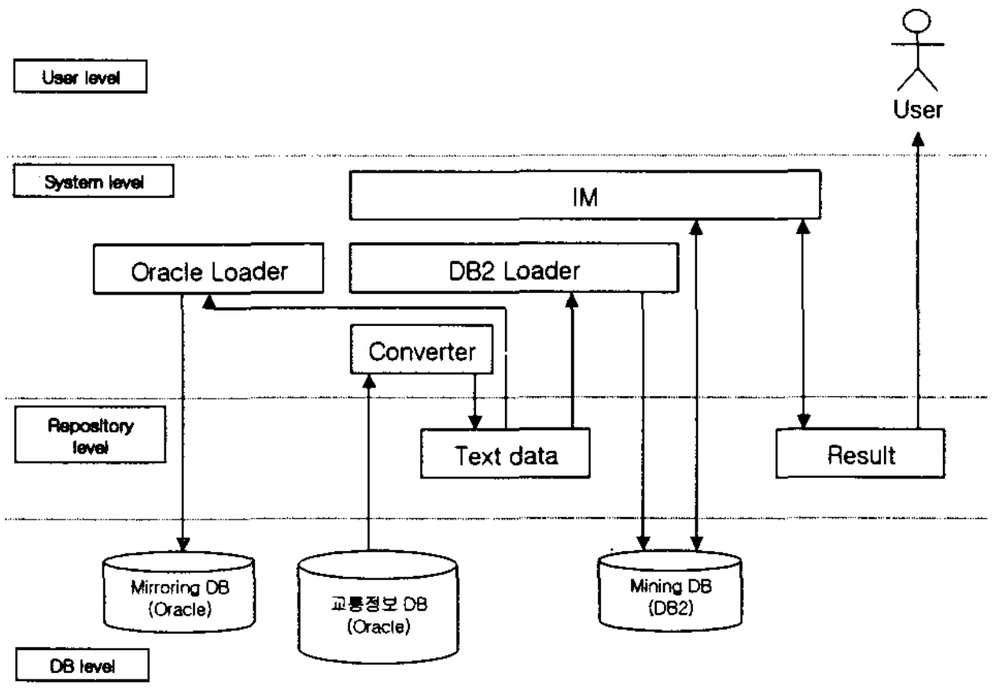
(2) 마이닝에 사용된 테이블 구조

- 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. 시스템 구조도

본 논문의 마이닝 시스템은 크게 위로부터 User level, System level, repository level, DB level로 이루어진다. 즉, 우리의 마이닝 시스템은 데이터베이스 레벨의 교통정보 DB로부터 사용자가 알기 쉬운 User level로 결과를 추출하는 시스템이다. 원시 자료로서 구축되어 있는 교통정보 DB는 매 5분마다 빈번하게 갱신되므로 무척 느리고, 사실상 다른 작업을 전혀 할 수 없는 상황이므로 우리는 같은 내용으로 다른 시스템에 데이터베이스를 이식하여야 한다. 따라서 이식할 시스템으로는 두 시스템이 필요하며 하나는 단순 DB작업을 할 수 있는 시스템이고, 다른 하나는 마이닝을 위한 DB작업을 할 수 있는 시스템이다.



4. 데이터 마이닝을 이용한 분석 및 결과

[가설] 톨게이트 인접 지역은 평균 속도가 40km/h 이하이다.

평균 속도가 40km/h 이하라는 검증을 위하여 우리는 visualization을 통하여 검증하고자 하였다. visualization으로 2가지를 보였다. 하나는 전일, 전 시간대의 평균 속력을 구하는 것이며, 다른 하나는 모든 날의 시간당 평균 속력을 구하는 것이다. 전체 평균 속도를 보이는 것은 텍스트로서 쉽게 이해가 가나, 시간당 평균 속도는 텍스트를 통한 visualization은 이해의 한계가 있으므로, 그래픽한 visualization을 동반하였다.

■ Visualization 1: 서울 톨게이트의 상·하행선 평균 속도를 보여준다.

- 방법: 1. tollgate라는 이름의 테이블을 생성한다.
2. 서울 톨게이트와 인접한 노드를 정보를 삽입한다.
3. 아래와 같은 질의를 통하여 톨게이트로 서울에서 진입하는 도로(KUT10070)와 서울로 진입하는 도로 (KDT10060)의 10일 간의 평균 속도를 추출한다.

```
SQL> select link_id, avg(speed) from tollgate
group by link_id;
```

- 결과 화면:

LINK_ID	AVG(SPEED)
KDT10060	80.428505
KUT10070	87.649533

- 결과 고찰

생각 외로 톨게이트와 인접한 도로 속도는 높았다. 기존에 톨게이트 주변이 막힐 것이라는 생각은 잘못된 생각임이 증명되었습니다.

■ Visualization 2: 서울 톨게이트의 시간 당 상·하행선 평균 속도를 보여준다.

- 방법:

1. 우리가 시험하고 있는 highway 테이블에 존재하는 ttime 컬럼은 시간과 분 정보를 동시에 가지고 있으므로 시간만으로 그룹화 시키기 위해서는 별도의 컬럼이 필요하다. 따라서 아래와 같은 질의문을 통하여 Visualization 1의 tollgate 테이블에 hour라는 컬럼은 추가시킨다.

ALTER TABLE tollgate ADD COLUMN (hour char(2));

2. 새로 추가한 컬럼에 시간만을 포함하는 정보를 삽입하기 위하여 아래와 같은 질의를 사용한다.

UPDATE tollgate SET hour = substr(ttime,1,2);

3. 해당 도로를 시간별로 그룹화 하여서 9일간 각 시간대의 속도를 그룹화 하여 평균을 구한다. 아래와 같은 질의를 사용한다.

SELECT LINK_ID, HOUR, AVG(speed) FROM TOLLGATE GROUP BY link_id, hour

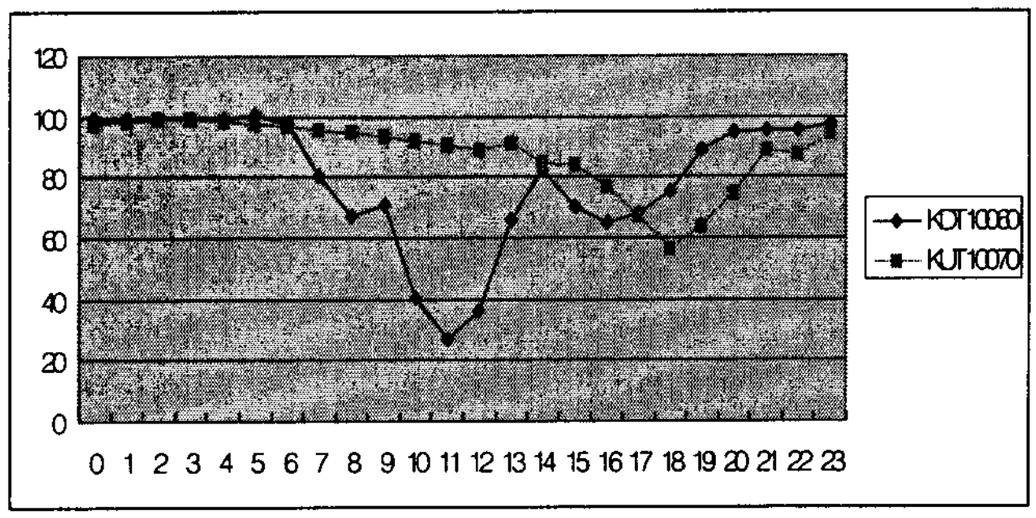
4. 그래픽한 결과를 보여주기 위하여 엑셀로 포팅(porting)한다.

5. 엑셀로부터 도표를 산출한다.

- 텍스트 결과 화면:

LINK_ID	hour	AVG(SPEED)			
KDT10060	00	98.802083	KUT10070	00	97.302083
KDT10060	01	99.354167	KUT10070	01	98.489583
KDT10060	02	99.020833	KUT10070	02	98.958333
KDT10060	03	98.808511	KUT10070	03	98.797872
KDT10060	04	99	KUT10070	04	98.072289
KDT10060	05	100.2619	KUT10070	05	97.380952
KDT10060	06	97.845238	KUT10070	06	96.630952
KDT10060	07	80.595238	KUT10070	07	95.488095
KDT10060	08	67.46988	KUT10070	08	94.626506
KDT10060	09	70.595238	KUT10070	09	93.083333
KDT10060	10	40.214286	KUT10070	10	91.964286
KDT10060	11	26.845238	KUT10070	11	90.690476
KDT10060	12	35.97619	KUT10070	12	88.595238
KDT10060	13	65.728395	KUT10070	13	90.765432
KDT10060	14	82.464286	KUT10070	14	84.892857
KDT10060	15	70.25	KUT10070	15	83.761905
KDT10060	16	64.776471	KUT10070	16	76.623529
KDT10060	17	67.705263	KUT10070	17	67.526316
KDT10060	18	74.916667	KUT10070	18	56.520833
KDT10060	19	88.715789	KUT10070	19	63.926316
KDT10060	20	94.791667	KUT10070	20	74.739583
KDT10060	21	95.083333	KUT10070	21	88.927083
KDT10060	22	95.677083	KUT10070	22	87.239583
KDT10060	23	97.541667	KUT10070	23	94.145833

- 그래픽 결과 화면



(X축:시간, Y축: 속도)

KDT10060 - 서울에서 부산 방향으로, 서울 톨게이트에 진입하는 도로

KUT10070 - 부산에서 서울 방향으로, 서울 톨게이트에 진입하는 도로

- 결과 고찰

시간대별 분포를 통해서도 알 수 있듯이 대부분의 시간에서 속력이 40km/h를 초과하였다. 단지 서울 톨게이트로 집입하는 하행선 도로(KDT10060)에서 10시-12시 사이의 속력이 40km/h 이하로 나타났다. 서울로 집입하는 상행선 도로(KUT10070)의 시간당 평균 속력은 항상 40km/h를 웃돌았다. 하행선이 시간에 대한 속도의 영향이 상행선에 비해 상대적으로 크게 나타나고 있는 것을 볼 수 있다.

특히, 상행선과 하행선에서의 시간당 속력의 그래프는 대부분의 작업이 시작되는 10시에서 12사이에 지체현상을 발생하는 것을 보였으며, 점심을 기점으로 다시 지체 현상을 나타냈으나, 하행선에서 진입하는 도로는 일과가 끝나는 시점에 지체 현상을 보였다. 결과 그래프는 영향받는 정도를 꺽은선 그래프로 나타낸 것이다.

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 생

성하여 준다. 데이터 마이닝을 수행하기 위해 세 가지 가설을 설정하였으며, 가설에 적합한 마이닝 연산을 적용하여 결과를 도출하였다.

첫 번째 마이닝의 가설은 '속도가 가장 빠른 시간대가 언제인가' 이며, 이를 위해 세 가지 방법(①양방향의 차선에 대한 속도를 클러스터링 한다. ② 한 구간의 자료를 뽑아서 클러스터링 작업으로 시간 분포의 특성을 분석한다. ③ 자료를 나누어서 속도에 대해 클러스터링 한다.)을 클러스터링 하여 속도를 평균속도, 저속도, 과속도로 군집시켜 특성을 도출한 결과 새벽 시간에 제일 빠른 속도가 결과로 제시되었다.

두 번째 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.

세 가지 가설에 의해 도로 교통에 대하 일반적인 편견이 실제 도로 상황에서는 적용이 되지 않음을 알게 되었고, 한 구간이 속도가 느리거나 막히면, 연관되어 있는 다른 구간들이 영향을 받는 일반적인 사실도 확인하게 되었다. 또한 평균 속도에 대해서는 오전, 오후에는 아주 비슷한 속도의 양상을 보였으며, 새벽 시간대에 속도가 빠른 것으로 나타났다. 향후계획으로 다양한 가설에 의해 다양한 연산을 적용하여 좀 더 많은 요소들을 도출하여 제시하는 것이 필요하다.

참고문헌

[1] "IBM DB2 Intelligent Miner for Data", IBM corp., 1999.

[2] I. Witten, E. Frank, "Data Mining", Morgan Kaufmann Publishers, 1999

[3] "Oracle Adminstration 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 Knowledg, 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.