고속도로 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 Information Database

Gi-Sung Lee*
*Dept of Computer Science, Howon University

요 약

차가 증가함에 따라, 교통은 혼잡하게 되고, 교통 체증은 더욱 심화된다. 만약에, 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 교통 체증의 시간대의 가설을 적용하고, 다양한 데이터 마이닝의 연산을 사용하여 결과를 도출한다. .....

1. 서론

현재 우리나라는 많은 교통 체증을 겪고 있다. 또한, 우천시나 눈이 내려 도로의 사정이 좋지 않을 때는 차의 평균 시속이 현저히 감소된다. 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.

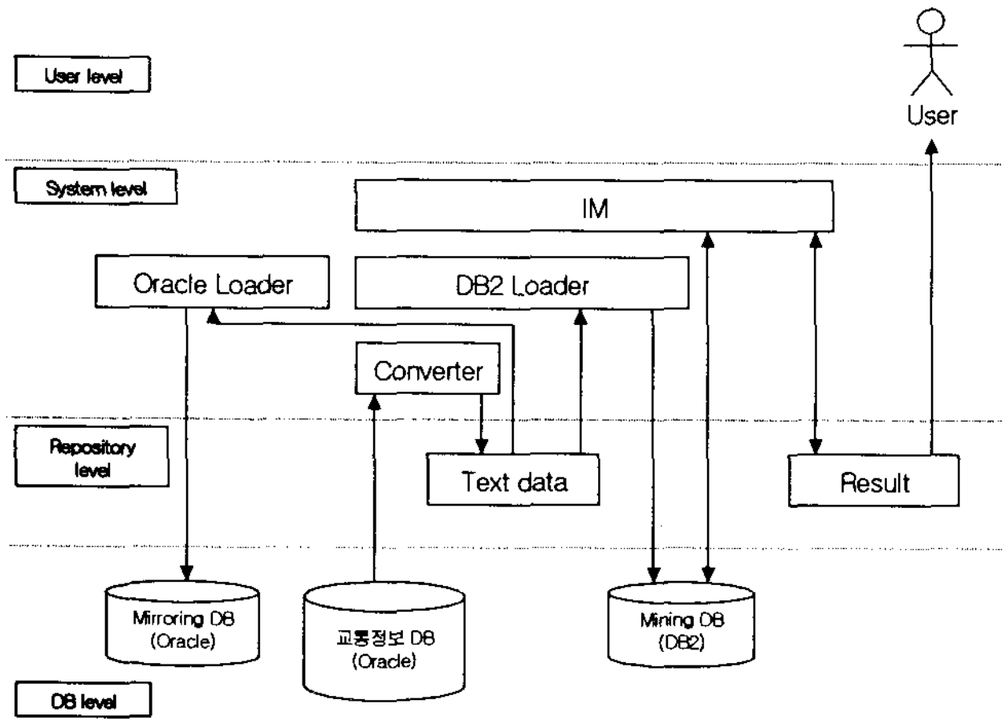
(2) 마이닝에 사용된 테이블 구조

- 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.



자료 예제는 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
```

4. 데이터 마이닝을 이용한 분석 및 결과

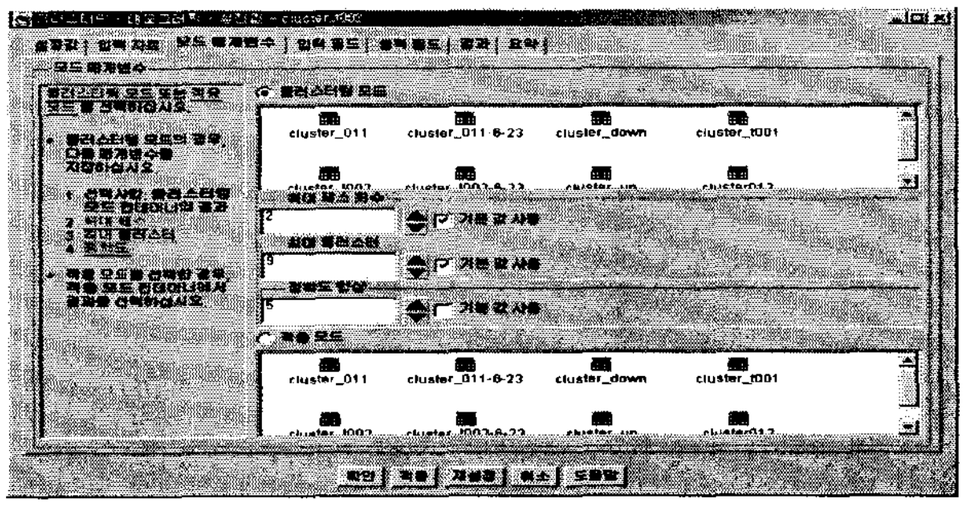
[가설] 속도가 가장 빠른 시간대는 언제인가?

고속도로에서 운전을 하다보면 어느 시간대에서는 차량 통행량이 많아 평균 속도가 늦은 경우가 있다. 가설을 검증하는 방법은 아래의 방법을 이용하여 클러스터링을 수행한다.

1) 상행선, 하행선의 구분없이 양방향의 차선에 대해 속도에 대해 자료들을 클러스터링한다.
2) 하행선의 한구간의 자료를 뽑아서 클러스터링 작업으로 시간 분포의 특성을 보았다.
3) 자료를 두 개로 나누어서 상행선, 하행선 부분으로 나누어서 속도에 대해 클러스터링.

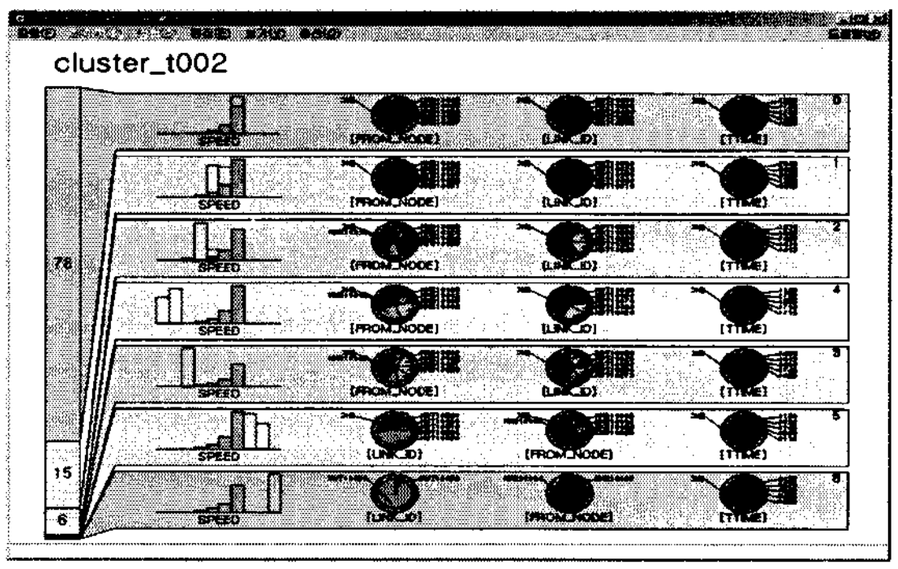
■ clustering 1: 상행선, 하행선의 구분 없이 양방향의 차선에 대한 속도에 대한 클러스터링을 수행한다.

- 방법: 1. roadhist라는 table을 생성한다.
  2. 전처리과정을 거친 자료를 테이블에 올린다.
  3. 자료들 중에서 하루의 자료의 양이 많이 차이가 나는 것을 제거한다.
  4. 아래의 방법으로 클러스터링을 과정을 수행.



많은 필드 중에서 클러스터링을 작업을 위한 부분은 speed필드를 이용하였다. 보조자료는 LINK_ID와 시간대인 TTIME을 넣었으며, 정확한 위치를 알기 위해 FROM_NODE를 넣었다.

- 결과 화면:



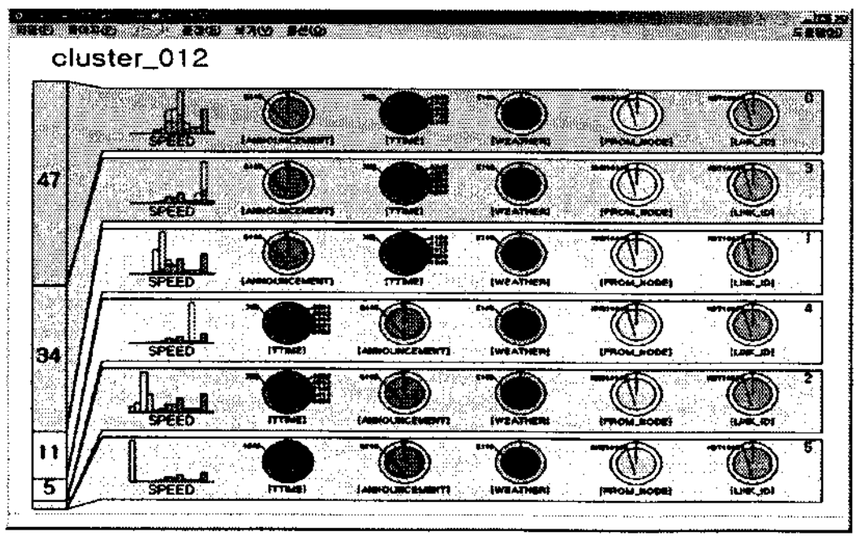
- 결과 고찰

고속도로 상태의 도로는 거의 모든 자료가 독립적인 성격을 갖기 때문에 서로의 연관성을 찾는 것은 힘들다는 것을 위의 자료에서 알 수가 있다. 위 자료에서는 전반적으로 평균속도를 가지는 집합이 많은 부분을 차지하다는 것을 알 수 있다. 그리고 다음으로 잡히는 집합은 저속도 집합이 많이 잡히고 마지막으로 작은 집합은 과속을 하는 집합이다.
가장 많이 지체가 되는 구간은 하행선 부분이며 일상적으로 정체가 자주되는 서초와 반포IC, 판교JC 부분이다. 상행선 부분은 죽전휴게소와 판교IC부분이다. 여기서 판교 부분이 많이 나왔다는 것은 위 시간대에서 판교 부분은 언제나 속도가 느리다는 것이다.

■ clustering 2: 하행선의 한 구간의 자료를 뽑아서 시간 분포의 특성을 보았다.

- 방법: 1. 위 clustering 1의 방법과 같다. 단지 하나의 구간을 오라클에서 뽑아서 따로 자료를 만들었다. 구간은 하행선인 서초IC에서 양재IC 사이의 구간이다.

- 결과 화면:



- 결과 고찰

위의 자료를 보면 시간과 속도와의 관계를 가지는 것은 아주 극단적으로 속도가 빠르거나 느린 경우라는 것을 알 수가 있다.
정속도를 이루는 부분과 약간 속도가 느린 부분은 시간과 관계가 없다. 그러나 정체를 나타내는 부분이 나오는 부분부터는 시간과의 관계가 생기는 것을 알 수가 있다.

먼저 아주 빠른 속도를 나타내는 집합을 보면 시간은 새벽에 속도가 아주 빠르게 나타나는 것을 알 수 있다. 기흥부근에는 새벽에 아주 빠른 속도로 지나 갈 수 있다는 것을 알 수 있다.
가장 저속도를 이루는 집합에서 시간대는 거의 오전 10에서 11시경과 저녁 7시, 8시대임을 알 수가 있다.

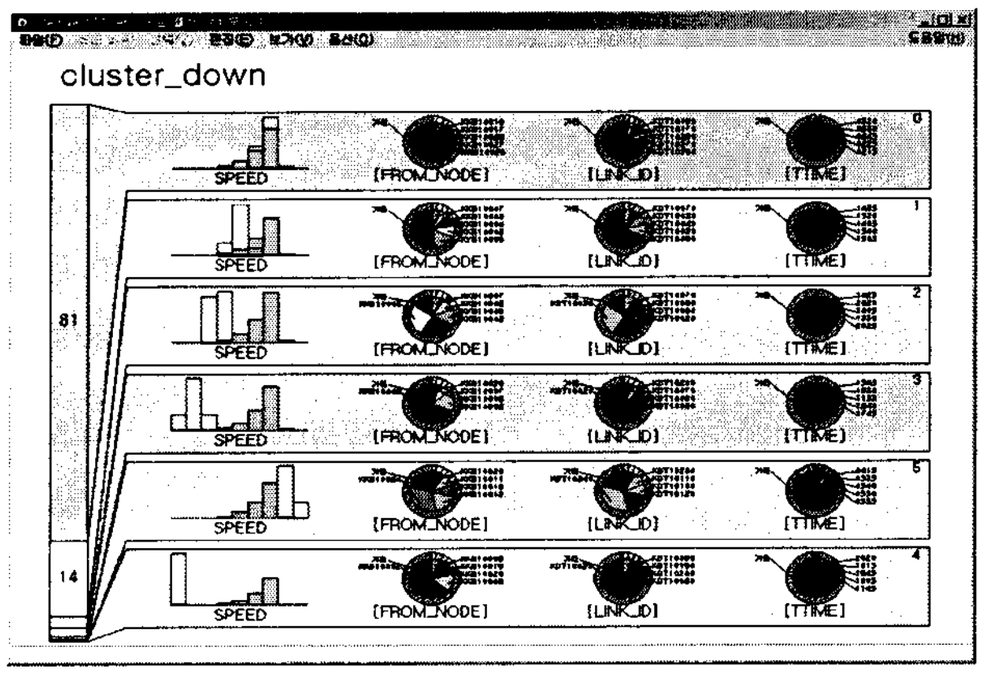
- 결과 고찰

전체적인 자료는 클러스터링1의 자료와 거의 같다. 여기서 다른 점은 저속를 이루는 집합들 사이에 정속도를 이루는 구간이 있다는 것이다. (클러스터 4) 속도를 이루는 클러스터1이 알려주는 것은 오전에 9시경부터 속도가 느려지기 시작한다는 것을 볼 수 있다. 이외의 날씨와 공지사항의 필드는 속도에는 아무런 관계가 없다는 것을 알 수 있다.

■ clustering 3: 하행선 부분으로 나누어서 속도에 대해 클러스터링.

- 방법: 1. 위 clustering 1의 방법과 같다. 자료는 전체 자료는 많은 시간을 소비하는 관계로 서울에서 대전사이의 경부고속도로구간을 추출하여 실험하였다.

- 결과 화면:



- 결과 고찰

위의 자료를 보면 시간과 속도와의 관계를 가지는 것은 아주 극단적으로 속도가 빠르거나 느린 경우라는 것을 알 수가 있다.
정속도를 이루는 부분과 약간 속도가 느린 부분은 시간과 관계가 없다. 그러나 정체를 나타내는 부분이 나오는 부분부터는 시간과의 관계가 생기는 것을 알 수가 있다.

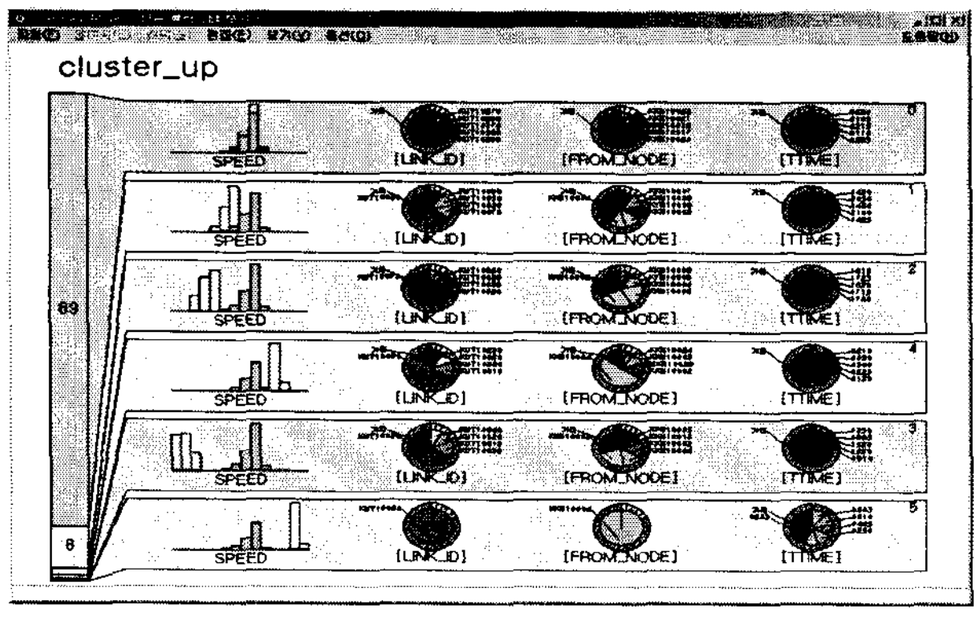
먼저 아주 빠른 속도를 나타내는 집합을 보면 시간은 새벽에 속도가 아주 빠르게 나타나는 것을 알 수 있다. 기흥부근에는 새벽에 아주 빠른 속도로 지나 갈 수 있다는 것을 알 수 있다.
가장 저속도를 이루는 집합에서 시간대는 거의 오전 10에서 11시경과 저녁 7시, 8시대임을 알 수가 있다.

■ clustering 4: 상행선 부분으로 나누어서 속도에 대해 클러스터링.

- 방법: 1. 위 clustering 1의 방법과 같다. 자료는

전체 자료는 많은 시간을 소비하는 관계로 서울에서 대전사이의 경부고속도로구간을 추출하였다.

- 결과 화면:



- 결과 고찰

위의 자료는 보면 하행선의 경우와 같은 결과가 나오고 있다. 점점 정규속도에서 속도가 느려지기 시작하면서 시간과의 관계가 나오고 있다. 그리고 하행선과 다른 점은 속도가 정규속도보다 아주 빠른 경우에 시간과의 관계가 다른 집합에 비해 아주 크다는 것을 알 수 있다.

속도가 느려지는 시간대는 오후 3시부터 6시대임을 알 수 있다.

속도가 아주 빠른 시간대가 새벽임을 알 수 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 '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.

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.