# 인문전산학 활용을 위한 데이터마이닝기법

곽호형*, 방혜자**
*서울산업대학교 산업대학원 컴퓨터공학과 석사과정
**서울산업대학교 컴퓨터공학과 교수
e-mail:kwakx@shinbiro.com

# Data Mining Technology for Application in Humanistic Computing

Ho-Hyung Kwak*, Hye-Ja Bang**
Dept of Computer Engineering, Seoul National University of Technology
*Graduate school of Industry and Engineering, **Professor

요 약

데이터마이닝은 대량의 실제 데이터로부터 이전에 잘 알려지지는 않았지만 묵시적이고 잠재적으로 유용한 정보를 추출하는 작업으로, 본 논문은 최근 인문학 정보 자료가 전산화되고 있는 가운데 대량의 정보와 특정 체계를 갖춘 '조선왕조실록' 전산자료를 분석하고 기존의 단순한 정보 검색이 아닌 데이터마이닝 기법을 적용한 상세하고 예측가능 한 정보자료 추출법을 제시한다. 먼저 텍스트화 되어 있는 컨텐츠를 형태소분석기법을 사용하여 색인어를 추출하고 집계를 낸다. 질의어와 유관한 색인어의 군집정도와 출현시점을 분석하는데, 사용된 마이닝 기법은 연관규칙분석과 클러스터링 분석기법이다. 최종 결과치는 기존의 인문학연구 결과물과 비교하여 그 정확도를 분석해 보인다.

## 1. 서론

최근 정보기술의 발달과 인터넷의 사용증가로 인해 데이터베이스에 저장되는 데이터의 양이 기하급수적으로 늘고 있다. 이런 데이터와 데이터베이스의 급속한 증가는 데이터를 유용한 정보로 변환할 수 있는 새로운 기술과 도구의 필요를 반영하며 결과적으로 데이터마이닝이라는 분야가 중요한 연구대상으로 자리 잡게 되었다.

아울러 인문학 분야에서도 다량의 정보가 데이터화 되어서 관련 연구자들에게 많은 편의를 제공하고 있지만 자료이용성의 한계, 아직은 많지 않은 자료, 체계화 되어있지 않은 데이터 등의 이유로 그 사용범위는 한정되어 있었다.

본 논문의 핵심이 되는 '조선왕조실록'은 인문학연구에 있어 국학연구의 핵심이 되는 자료이며 이미 1997년에 전산화 되어 관련 연구자들에게 많은 편리함 제공하고 있는데 단순한 질의어에 대한 정보 조회나 자료 소장의 용도 외에는 이용의 범위가 한정되어왔었다. 그러나 이용성이 떨어질 만큼 데이터의 규모가 적거나, 데이터의 구조가 체계적이지 못한것도 아니었다.

본 논문에서는 '조선왕조실록' 전산데이터를 이용하여 데이터마이닝 기법에 맞게 데이터구조를 리모델링 해보고 다양한 데이터마이닝 기법을 제시해 보고자 한다.

## 2. 인문전산학의 개념과 특징

인문전산학의 개념은 인문학을 연구함에 있어 관련된 정보 자료의 전산화가 그 근간이 되며 이를 활용하는 방법론에 대한 통칭이라고 하겠다. 인문학은 인간의 가치와 정신적인 면을 강조하여 인간만이 지니는 표현능력을 인식하기 위한 연구 방법을 취하는 문화 과학이자 정신과학으로 불리는 학문이다[5].

인문 전산학의 효시는 1949년 부사(R. Busa)가 IBM 사장에게 토마스 아퀴나스의 전작을 전산화해 줄 것을 부탁한 때로 잡는다. 당시의 컴퓨터는 진공관을 쓰고 있을 때므로 인문전산학은 컴퓨터공학과 함께 출발했다고 해도 과언이 아니다.

그 이후 언어학이나 고고학 통계 자료처리 등의 영역에서 골고루 사용되었는데 제한적이나마 데이터

베이스화되어 다양한 구조와 형태로 이용되었다.

전산화되기 위한 인문학 정보의 특징은 아래와 같다.

① 인문학 정보는 표현을 위한 행위나 동작 또는 작품 그 자체가 중요한 정보적 가치를 지니기 때문에 정보자료의 정형을 유지하기 보다는 영상을 통해 복제를 하거나 제3의 방법으로 자료화 하는 경우가 많다.

② 고전자료나 오래된 문서 자료는 그 자체가 학술적 가치를 지니는 연구실적인 동시에 정보내용을 담고 있는 자료일수 있는데 자료화를 통해 자료의 가치와 생명을 잃게 되는 경우도 있을 수 있다.

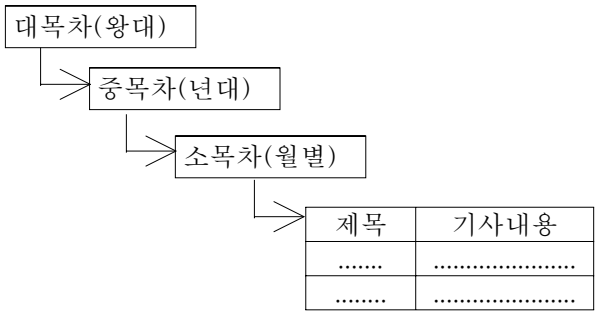
③ 이런 특성으로 자연과학이나 사회과학의 경우와 같이 서지 자료를 통해 요약이나 통계 등 2차 자료화 등의 일반화가 학문분야의 관심이 되기보다. 원저나 실물의 추적에 더 관심을 갖기 때문에 서지의 활용이 학문의 과정에서 크게 중요시 되지 않는다

④ 하지만 인문학에서의 연구 성과가 사회과학이나 자연과학적인 형태로 나타나고 유관 학문과의 연계가 빈번하므로 정보자료의 2차 가공의 필요성은 절실하다.

## 3. 국역 '조선왕조실록'의 전산화와 활용

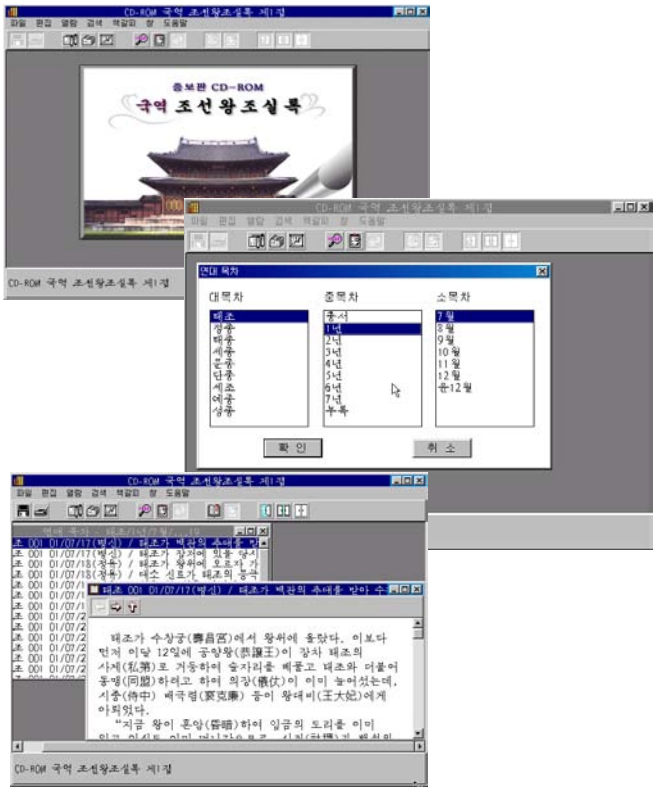
'조선왕조실록'은 조선시대 태조에서 철종까지 25대 472년에 이르는 역사를 왕대, 년도, 일자별로 기록해 놓은 역사서이다.(4의 73 주석)

1997년에 전산화 되어 CD롬으로 출시되었는데 그 데이터의 구조를 살펴보면 아래와 같다.




(그림1) '조선왕조실록'의 구성

왕대별로 시간순서에 맞게 구성되어있는데 이는 데이터를 목차 순서대로 활용하고 이외에도 색인어 검색을 통한 특정 기사로 접근도 가능하다. 다수의 이용자는 단순한 열람의 목적이 아닌 특정 색인어를 이용하여 검색하고 검색 결과를 열람하는 형태로 이




(그림2) '조선왕조실록'이용화면

용하고 자료의 양이 많아서 매우 비효율적이다.

## 4. 데이터마이닝의 적용

### 4.1 데이터의 구성

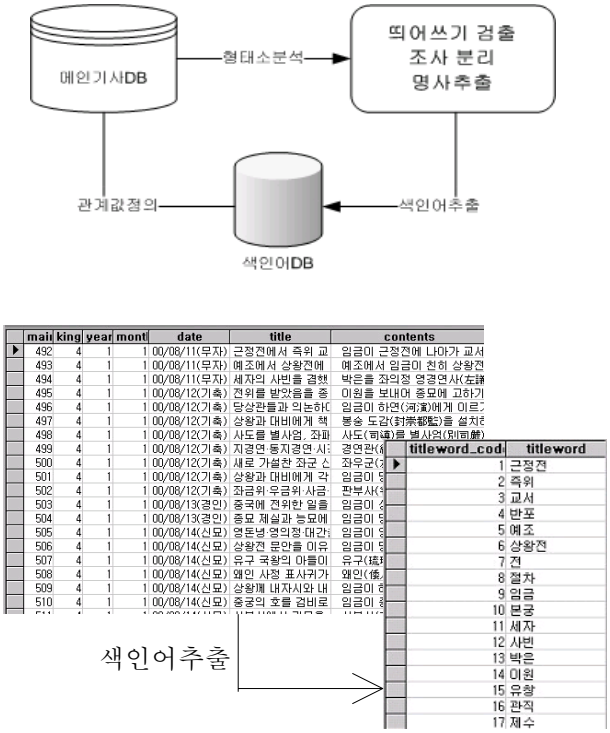
데이터마이닝 기법을 사용함에 있어 가장 중요한 요소중 하나는 정보의 양이다. 우선적으로 다량의 정보가 있어야 정보의 특성과 연관 규칙을 파악할 수 있다. '조선왕조실록'이 보유하고 있는 데이터 양은 다음과 같다[6].

<표1>'조선왕조실록'의 데이터 통계

| 데이터구분 | 수치 |
|---|---|
| 재위년수 | 493 |
| 기사총수 | 362161 |
| 사론총수 | 5792 |
| 원주총수 | 38269 |
| 기사평균 | 734.6 |
| 사론평균 | 11.7 |
| 원주평균 | 77.6 |

여기서 주로 사용되어질 데이터는 기사인데 한 기사는 서술적인 자료 형태로 유효 색인어의 개수는 평균 30개 내외였다. 색인어의 추출은 형태소 분석을 통해 이루어졌다. 따라서 기사내의 색인어의 개수는 약 10864830개가 되는 것이다. 이 정도의 양

이면 충분히 데이터마이닝을 구현하여 연관규칙 기법과 클러스터링 기법등을 적용할 수 있을 것이다.




(그림3) 유효색인어 추출과정과 결과

### 4.2 연관규칙기법

연관규칙이란 어떤 사건이 일어나면 다른 사건이 일어나는 트랜잭션내의 항목간의 연관성을 의미한다[1,2]. 가령 어떤 통계치를 통해 빵과 버터를 구매하는 사람은 우유를 구매한다 라는 연관 규칙을 알 수 있다고 했을때 아래와 같이 표기한다.

[빵],[버터] ->[우유](support 12.5%, confidence 90%)

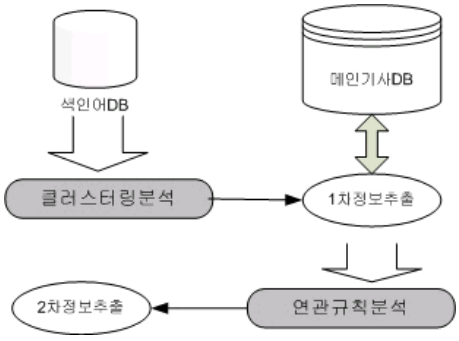
- support(지지도) : 지지도는 생성된 연관규칙이 전체 아이템속에서 차지하는 비율을 말한다. 즉 데이터베이스에 속한 전체 트렌젝션의 개수 중, 그 연관규칙이 차지하는 트랜젝션의 비율을 의미한다.
- confidence(신뢰도) : 신뢰도는 연관규칙의 강도를 의미하며 전체부를 만족하는 트랜젝션이 결론부까지를 만족하는 비율을 말한다[3].

### 4.3. 클러스터링 기법

클러스터링이란 물리적 혹은 추상적 객체를 비슷한 객체군으로 그룹화 하는 과정이다. 이 때 유사성에 따라 함께 모여진 객체의 셋을 클러스터라 한다. 클러스터링 작업은 먼저 필수 객체들이 셋으로 모여지고 이로부터 일련의 규칙이 유도된다. 클러스터링 기법은 어떤 목적 변수를 예측하기 보다는 속성이 비슷한 정보들을 묶어서 몇 개의 의미있는 군집으로 나타내는 것을 목적으로 한다[4].

## 5. 구현

'조선왕조실록'에서의 데이터마이닝 정보 탐사과정은 다음과 같다.




(그림4) 정보탐색의 과정

먼저 비교대상이 되는 관련 인문학 논문 한편을 선정하여 그 논문의 핵심 주제어를 질의어로 삼고 논문의 결과와 데이터마이닝의 결과를 비교하여본다.

선정논문 : 한국 정치제도의 변화상 ; 조선초기 경차관과 외관 (김순남, 한국사학보,18권,2004년)
논문의 결론 : 조선초기 경차관은 중앙에서 지방으로 파견된 관리로서 지방의 백성들을 기근으로부터 진휼한다던지 위급한 상황을 해결하기 위해 파견되었는데, 정규 지방관리인 관찰사 또는 수령과 상호 보완적 이면서도 감시와 견제를 함께 하여 효율적인 지방통치를 위해 만든 관직이다.

위 논문에서 알 수 있는 핵심 주제어는 '경차관'과 '외관' 이다. 두 단어를 질의어로 사용하여 클러스터링 기법을 적용한 결과 다음과 같은 결과가 나왔다.

<표2> 클러스터링 분석 결과

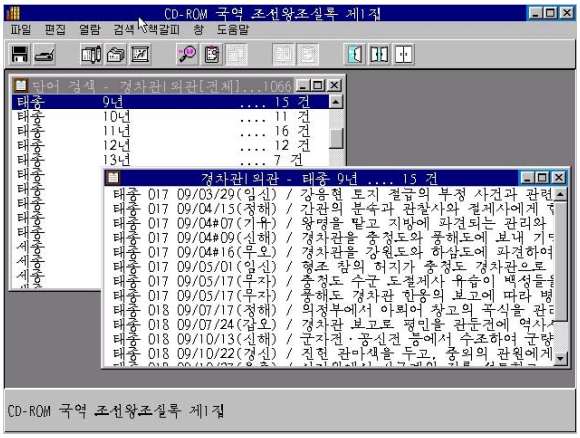
| 기사횟수 | 색인어 |
|---|---|
| 987 | 파견 |
| 785 | 삼남 |
| 554 | 평안도 |
| 512 | 진휼 |
| 503 | 충청도 |
| 489 | 수령 |
| ... | ..... |

1차로 질의어에 대한 색인어의 군집도를 알 수 있는데 가장 빈번하게 기사에 나타난 순서대로 나열하면 질의어에 대한 기본 성향을 알 수 있다. 2차 작업으로 빈출 색인어에 대한 기사내의 연관규칙을 조사하면 다음과 같다. (1:기사내포함,0:미포함)

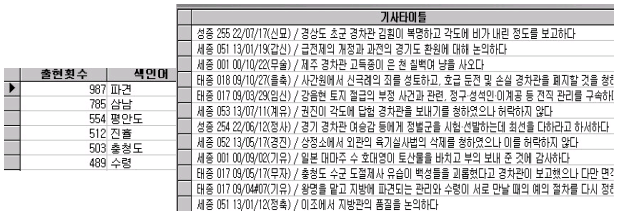
<표3> 연관규칙분석을 위한 색인어 기사분포데이터

| 번호 | 색인어 | 기사번호 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 745 | 789 | 587 | 265 | 428 | 456 | 874 | 354 | 785 | 854 |
| 1001 | 파견 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| 1152 | 삼남 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 1230 | 평안도 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 1355 | 진휼 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |

특정색인어의 기사 출현 빈도와 기사간의 연관규칙의 신뢰도가 높은 기사의 순서대로 기사의 리스트를 나열하게 되면 좀더 유효하고 정확성이 있는 정보가 우선적으로 보여질 수 있게 된다.




(그림5) 기존 '조선왕조실록' 정보검색 결과




(그림6) 데이터마이닝 적용 검색 결과

## 6. 결론

본 논문에서는 데이터마이닝 기법을 이용하여 '조선왕조실록' 데이터를 효율적으로 활용하는 방법을 구현하였다. 기존의 CD롬 형태의 프로그램에서는 이용자가 제시하는 단순한 질의에 대한 결과치의 리스트만 나열하는 형태로 구현되는 데이터를 데이터마이닝의 연관 규칙과 클러스터링 기법을 활용하여 사용자가 원하는 정보를 보다 상세하고 효율적 이용할 수 있게 하였다. 사용자가 입력한 질의와 관련한 색인어 테이블을 별도로 구성하여 메인DB에서의 출현여부 그리고 해당 기사리스트에서의 유효 색인어 출현 빈도를 체크하여 중요도 위주로 다시 정보값을 뿌려 주어 이용자가 원하는 정보를 재생산 하였다.

본 논문에서는 제외되었지만 메인 데이터베이스를 활용하기 위하여 형태소 분석기법을 사용하였다. 유효색인어의 자동검출을 위해 반드시 필요한 기법이고 향후 이용자의 서술형질의에 대한 유효단어 검출에도 사용할 수 있을 것이다.

현재 다수의 국학자료가 전산화 되고 있지만 그 활용적인 측면에서는 단순한 정보검색 외에는 특별한 형태로의 사용은 아직 이루어지지 않고 있는 현실이다. 인문학정보 활용이라는 측면에서 볼 때 데이터마이닝 기법은 아주 효율적인 정보탐사방법이다. 앞으로 단순한 질의에 대한 기법적용이 아닌 웹을 활용한 실시간 데이터 공유로 이용자의 정보활용 성향이나 이용빈도를 데이터화 하여 보다 효율적인 정보 활용이 가능하게끔 하는 것도 좋을 이용법이 될 것이다.

## 참고문헌

[1] U. Fayyad and R. Uthurusamy, eds, "Data Mining," Special Issue, Comm of the ACM, 39(11), Nov, 1996.

[2] J. Han, "Data Mining Techniques," Proc. of 1996 ACM SIGMOD Intl Conf. on Management of Data, 1996.

[3] 최민희, 오형재, 홍의경, "데이터마이닝기법을 이용한 효율적인 웹 검색엔진의 설계 및 구현," 서울시립대학교 정보기술연구소 논문집, 제2집 pp. 7~15, July, 2000.

[4] 이용범, "데이터마이닝의 농업적인 활용," 한국농업기계학회 바이오시스템공학, 29권, 1호, pp. 79~96, 2004.

[5] 배순자, "인문과학 정보자료의 전산화와 정보봉사," 전주대학교 인문과학종합연구소 인문과학연구, Vol.3, No.0, pp. 239~252, 1997.

[6] 이남희, "전산화를 통해서 본 조선왕조실록," 서지학회 서지학연구, 13권, pp. 73~101, 1997.