http://dx.doi.org/10.7236/JIIBC.2013.13.3.157
JIIBC 2013-3-21

공간 데이터마이닝 분석을 통한 데이터의 효과적인 활용

Effective Utilization of Data based on Analysis of Spatial Data Mining

김기범*, 안병구**

Kibum Kim, Beongku An

요 약 데이터마이닝은 데이터간의 상호 연관성과 다양한 패턴 분석을 통해서 우리가 알 수 없었던 새로운 발견을 할 수 있는 유용한 기술로서 현재 금융, 마케팅, 의료 등 다양한 분야에서 활용되고 있다[1]. 본 논문에서는 공간 데이터마이닝 분석을 통한 데이터의 효과적인 활용방법을 제안한다. 서울시에 거주하는 외국인들의 기본적인 데이터를 활용하고자 한다. 하지만, 이 데이터는 다른 분야의 데이터와 구별되는 특징이 있는데, 민감 정보로 분류된다는 것과 개인정보보호 등과 같은 법적인 문제가 있을 수 있다. 따라서 개인정보를 알 수 없는 기본적 통계적 데이터를 활용하고자 한다. 제안된 방법의 주요한 특징 및 기여도는 다음과 같다. 첫째, 큰 데이터를 여러 질의방법을 통해서 정보로서 이용할 수가 있으며, 정제를 통해서 클러스터링 할 수 있다. 둘째, 이러한 정보들을 새로운 패턴이나 앞으로의 의사결정에 이용할 수 있다. 질의 결과에서 얻은 새로운 정보를 사용자가 보고 판단하여 의사결정에 이용하고자 한다. 제안된 방법의 성능평가에서는 데이터들의 주제별 도식화를 통한 시각적 접근방법을 사용하고자 한다. 제안된 방법의 성능평가 결과는 데이터를 보다 가치 있게 활용하기 위해서 데이터마이닝 기술을 이용한 분석을 통해 우리가 알 수 없었던 새로운 패턴과 결과의 발견이 가능함을 보여준다.

Abstract Data mining is a useful technology that can support new discoveries based on the pattern analysis and a variety of linkages between data, and currently is utilized in various fields such as finance, marketing, medical. In this paper, we propose an effective utilization method of data based on analysis of spatial data mining. We make use of basic data of foreigners living in Seoul. However, the data has some features distinguished from other areas of data, classification as sensitive information and legal problem such as personal information protection. So, we use the basic statistical data that does not contain personal information. The main features and contributions of the proposed method are as follows. First, we can use Big Data as information through a variety of ways and can classify and cluster Big Data through refinement. Second. we can use these kinds of information for decision-making of future and new patterns. In the performance evaluation, we will use visual approach through graph of themes. The results of performance evaluation show that the analysis using data mining technology can support new discoveries of patterns and results.

Key Word : Data mining, Spatial data, Database, Big data

*준회원, 홍익대학교 컴퓨터정보통신공학과
**종신회원, 홍익대학교 컴퓨터정보통신공학과 (교신저자)
접수일자 2013년 4월 6일, 수정완료 2013년 5월 22일
게재확정일자 2013년 6월 14일

Received: 6 April 2013 / Revised: 22 May 2013 /
Accepted: 14 June 2013
**Corresponding Author: beongku@hongik.ac.kr
Dept. of Computer & Information Communications Engineering, Hongik University, Korea

I. 서 론

지난 수십 년간 여러 가지 형태로 저장되어 있는 데이터의 양은 기하급수적으로 증가되어 왔다. 그러나 이러한 데이터의 무제한적인 증가는 필요한 정보를 찾아내는데 불필요한 시간이 많이 소모되고 있다. 왜냐하면 대용량의 데이터로부터 의미 있는 지식을 찾아내고자 하는 목적에 반하여 오히려 데이터만 계속 누적되고 있기 때문이다. 이러한 상황에서 데이터마이닝 (data mining)[1-10]은 중요한 문제 중 하나이다. 데이터로부터 알려지지 않은 새로운 정보나 유용한 패턴과 상관관계를 추출하여 의사 결정에 이용하는 작업인 데이터마이닝은 이미 존재하는 데이터에서 새로운 것을 찾아낼 수 있다는 것에 가장 큰 매력이 있다. 여기서 우리는 서울시에 거주하는 외국인들의 기본적인 데이터(성별 및 거주하는 구)를 이용해서 패턴이나 상관관계를 알아보고자 한다. 그리고 분석결과를 시각적인 표현을 통해서, 데이터마이닝 이전에는 알 수 없었던 혹은 텍스트로 된 데이터만으로 알 수 없었던 중요한 발견을 기대할 수 있다는 것에 목적을 둔다.

본 논문은 다음처럼 구성되어 있다. Ⅱ장에서는 관련 연구를, Ⅲ장에서는 제안된 방법을, Ⅳ장에서는 데이터마이닝 결과를 토대로 성능측정을, Ⅴ장에서는 결론을 맺고자 한다.

II. 관련 연구

데이터 마이닝[1-10]은 대용량의 데이터로부터 알려지지 않은 새로운 정보나 유용한 패턴과 상관관계를 추출하여 의사 결정에 이용하는 작업이다. 이는 자료의 효율적 저장을 위한 기술(데이터 베이스, 압축, 통신)의 발달에 의해 기하급수적으로 늘어나는 방대한 데이터 양, 지식 정보화 사회에서의 새로운 지식의 습득에 대한 필요성, 컴퓨터 성능의 향상으로 인한 방대한 양의 데이터의 실시간 분석기능으로 인하여 가능하게 되었다.

데이터 마이닝[1-10]은 기업들이 보유한 기존의 경험적 지식을 재확인하는 역할을 수행함과 동시에 지금까지 인식하지 못했던 새로운 정보와 지식을 제공하여 경영의사결정에 도움을 준다. 특히 데이터마이닝에 의해 발견되는 기존의 관념을 깨는 지식은 기업 경쟁력 강화에 결정적인 역할을 한다. 그러나 데이터마이닝의 올바른 활용을 위해서는 이것의 한계를 명확히 이해하는 것이 중요하다. 흔히 데이터마이닝과 같은 용어로 사용되고 있는 KDD(지식발견 : Knowledge Discovery in Database)라는 용어로 인하여 데이터마이닝 도구들이 자동적으로 실행 가능한 지식을 발견 해주는 것으로 오해되고 있지만 데이터마이닝은 최종 사용자들이 데이터에 내제된 패턴을 찾아낼 수 있도록 도와 줄 뿐 발견된 패턴의 타당성이나 가치를 판단해 주지는 못한다. 데이터마이닝에 의해서 발견된 패턴들을 현실에 맞게 해석하고 타당성을 검증하여 실행 가능한 지식으로 지식화 하는 것은 최종 사용자의 책임이며, 이를 위해 사용자는 적용 도메인은 물론 분석 데이터의 속성 및 데이터마이닝 도구에 대한 지식을 보유하고 있어야 한다.

1. 데이터마이닝과 KDD

데이터마이닝은 그림 1과 같이 입력 데이터를 변환하여 유용한 정보를 도출하는 전체 과정인 데이터베이스 KDD(Knowledge Discovery in Database)의 핵심과정으로 이 과정은 데이터 전처리에서 데이터마이닝 결과의 후처리까지 일련의 변환 과정으로 구성된다[9].

그림 1. 데이터베이스에서 KDD과정
Fig. 1. KDD process in the database

입력 데이터는 일반파일, 스프레드시트, 관계 테이블 등의 다양한 형식으로 저장되며 중앙데이터 저장소나 다양한 시트에 걸쳐서 분산된다. 전처리의 목적은 입력 데이터를 변환한 후 분석에 적합한 형식으로 변환하는데 있다. 데이터 전처리에 필요한 과정으로는 다양한 소스로부터 데이터를 모으고, 잡음과 중복으로부터 데이터를 정제하고, 데이터마이닝 작업과 관련된 레코드의 특징들을 선택하는 과정이 있다. 데이터는 여러 방법으로 수집되고 저장되기 때문에 데이터 전처리는 전체 지식탐사과정에서 가장 노동 집약적이고 시간을 많이 소모하는 과정이다.

데이터마이닝의 결과는 의사결정 시스템으로 입력되어 활용되어야 하는데 비즈니스 응용에서 데이터마이닝 결과에서 도출된 정보는 영업 관리 도구와 통합되어 효과적인 마케팅 홍보에 적용되고 결과를 검증할 수 있어

야 한다. 이와 같은 통합은 후처리 과정이 타당성 있고 유용한 결과만을 의사결정 시스템으로 통합되도록 해야만 한다. 통계적 척도와 가설 검증 방법 또한 후처리에 적용되어 불필요한 데이터마이닝 결과를 제거하기도 한다[8].

또한 데이터마이닝, 지식마이닝, 데이터베이스로부터 지식발견 등의 연구도 활성화 되고 있다. 더욱이 검색 결과의 시각화나 브라우징이 가능하도록 데이터베이스 인터페이스로 표현하는 분야가 주목되고 있다. 여기에서는 기계학습에 관한 성과와 함께 데이터베이스 특유의 성질을 고려하여 발전시키고 있는 연구 분야가 많으며 각 속성, 속성치 간에 성립되는 규칙, 제약 그리고 규칙성이 전형적인 지식으로서 발견된다.

2. 데이터마이닝의 수행단계

데이터마이닝의 수행 단계는 그림 2와 같이 8단계로 구성된다. 각 단계는 여러 세부 작업으로 분할 가능하며 상이한 단계들 또는 작업 사이를 반복적으로 수행하는 것이 일반적이다[2-9].

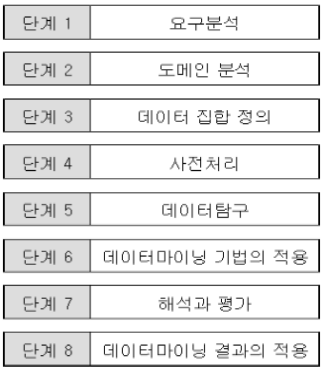

그림 2. 데이터마이닝의 수행단계
Fig. 2. Implementation of data mining

- **단계 1:** 요구 분석(requirements analysis): 대상 문제에 대한 명세화 또는 데이터마이닝의 목표에 대한 명확한 정의를 내리는 단계로 이 단계의 산출물은 이후 단계들의 준비와 실행에 관한 전략적 계획이다.
- **단계 2:** 도메인 분석(domain analysis):응용 도메인, 데이터, 환경적 특성에 관한 지식을 분석하여 초기 데이터마이닝 계획을 수립한다.
- **단계 3:** 데이터 집합 정의(definition of data sets):데이터마이닝의 대상이 될 데이터가 데이터베이스에 분산되어 있는 여러 개의 이질적인 데이터의 통합이 수행된다.
- **단계 4:** 사전 처리(preprocessing): 기법 적용 전에 필요한 모든 작업으로 데이터의 적재, 변환 및 클리닝 등이 포함된다. 클리닝은 데이터 내의 잡음제거, 잡음설명, 망실 데이터 필드에 대한 처리 전략 등에 관한 기본적인 작업을 수행한다.
- **단계 5:** 데이터 탐구(data exploration): 데이터마이닝에 적용할 데이터에 대한 통찰과 흥미 있는 데이터 또는 특성 부분집합을 파악한다. 단계 3에서 단계 5는 빈번하게 데이터 집합들과 이들에 관한 정보 및 중간 결과가 생성되므로 작업 수행 이력을 체계적으로 관리하는 것이 필요한다.
- **단계 6:** 데이터 마이닝 기법의 적용(application of data mining techniques): 다양하고 상이한 기능별 데이터마이닝 기법인 연관기억 (associative memories), 분류(classification), 군집화(clustering), 함수 근사와 예측(function approximation & prediction) 의 4가지로 선택적으로 적용된다.
- **단계 7:** 해석과 평가(interpretation and evalution):데이터마이닝의 결과는 사용자가 해석 가능한 용어 또는 의사결정에 이용할 수 있는 지식으로 표현되어야 하며, 단계 1에서 정의된 평가 기준에 의해서 평가된다.
- **단계 8:** 데이터마이닝 결과의 적용(deployment):성공적인 데이터마이닝 태스크의 결과는 의사결정 문제의 해결을 위해서 사용된다.

Ⅲ. 제안된 방법

본 연구에서 제안하고자 하는 방법은 텍스트데이터를 이용해서 공간데이터 마이닝을 실행하는 방법이다. 데이터를 시각화 즉, 공간화를 통한 분석을 통하여 새로운 정보를 발견하는 것이다. 그림 3은 공간 데이터 마이닝의 기본 개념을 보여주고 있다. 그림 4는 본 논문에서 제안한 공간 데이터 마이닝을 위한 PNU 코드화 방법을 보여주고 있으며, 그림 5는 제안된 PNU 코드화 방법을 구체적인 예를 들어서 설명하고 있다. 주소로 표현되어 있는 텍스트데이터를 코드화 시키는 방법인데, 그림 4와 같이 PNU코드화를 통해서 KLIS(한국토지정보시스템)에서 제공하는 지적도 및 주제도의 Primary Key와 그림 6의

개념으로 매칭을 시키고자 한다. 여기서 한가지의 통일된 양식이 아닌 무작위의 양식대로 저장되어진 텍스트 주소데이터를 그림 7처럼 코드화 시키는 과정은 상당한 시간이 소요될 수도 있다. 그림 7은 완성된 PNU 코드의 예를 보여주고 있다.

그림 3. 공간데이터 마이닝의 기본개념
Fig. 3. Basic concepts of spacial data mining

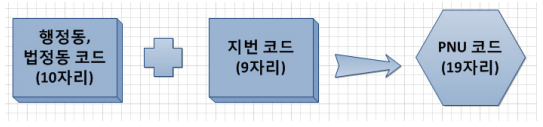

그림 4. 제안된 PNU 코드화 방법: 주소 데이터를 PNU 코드화
Fig. 4. Proposed PNU code method: PNU code into the address data

그림 5. 제안된 PNU 코드화 방법의 구체적인 설명
Fig. 5. The detail explanation of the proposed PNU code method

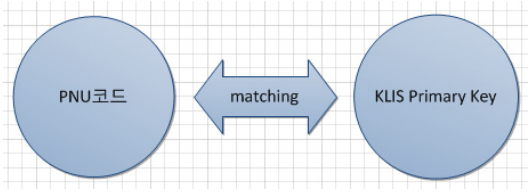

그림 6. PNU 코드와 KLIS Primary Key 값과 매칭
Fig. 6. KLIS Primary matching to PNU code

구분	성별	국적	국적별구분	국적별코드	체류지	동코드	지번코드	PNU
강남구	F	이란	중동	15	서울특별시 강남구 수서동 713번지	1168011500	107130000	1168011500107130000

그림 7. 완성된 PNU코드의 예
Fig. 7. PNU code example

PNU코드와 Primary Key의 매칭과정은 1:1 매칭이기 때문에 빅데이터의 경우에는 상당한 시간이 소요된다. 코드와 Primary Key의 매칭이 완료되면, 기본 지도(base map)에 공간화를 시작하고, 그림 8과 같이 공간화된 데이터를 토대로 사용자가 의사결정을 할 수 있게 된다[8,9].

그림 8. 공간화된 데이터의 예
Fig. 8. Example of data into space

Ⅳ. 성능평가

본 연구에서는 텍스트 데이터를 이용한 공간 데이터 마이닝을 제안하였다. 이를 위해서는 우선 텍스트 데이터를 가공하여야 한다. 즉, 텍스트 데이터에서 주소데이터를 코드화 하여 공간 마이닝에 이용하고자 한다. 이에 대한 본 성능평가에서는 제안된 방법에 대하여 ArcGIS를 이용한 환경에서 실험을 수행한다.

1. 성능평가 환경

성능 평가는 ArcGIS를 이용한 환경해서 시행하고자 한다.

그림 9. ESRI 사의 ArcGIS
Fig. 9. ESRI's ArcGIS

엑셀로 저장되어 있는 데이터를 dbf 파일화 하여 PNU 코드화를 통해 가공한다. 기본지도(base map)으로는 한국토지정보시스템의 서울시 기본 필지를 이용하였

다. 그림 10는 서울시의 기본지도를 보여주고 있다.

그림 10. 서울시의 기본지도
Fig. 10. Base map of Seoul

2. 성능평가 결과

ArcGIS에서 KLIS Primary Key와 가공화한 PNU코드를 매칭 시켰다. 1:1 매칭을 하여 기본지도(base map)에 약26만개의 데이터를 하나하나 공간화 하는 방식으로 상당한 시간이 소요되었다[9].

그림 11에서는 이탈리아의 거주 현황을 확인할 수 있는데, 여기에서 대부분의 이탈리아인이 중구, 용산구, 성동구, 서대문구 등지에 거주하는 것을 파악할 수 있다. 용산구에는 이탈리아의 대사관이 있는데 이와 근접한 중구와 성동구에도 많이 거주하는 이유를 알 수 있다. 서대문구는 다른 국가의 공간화 결과에서도 두드러지게 나타나는데 대학교가 많고, 외국인 유학생들이 많이 거주하기 때문에 외국인들에게 거주하기 편한 구역이라는 것을 파악할 수 있다.

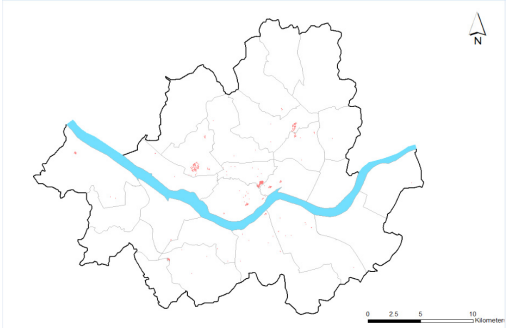

그림 11. 공간화된 이탈리아인의 거주현황
Fig. 11. Spatial the current state of Italian

그림 12와 그림 13의 공간화 결과를 보면, 앞서 데이터마이닝 시뮬레이션을 통해서 파악했던 결과와 유사하다는 것을 알 수 있다.

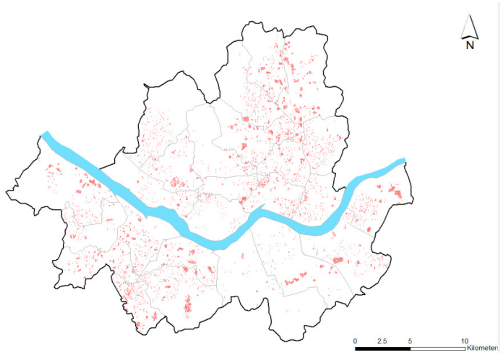

그림 12. 공간화된 베트남인의 현재 거주현황
Fig. 12. Spatial the current state of Vietnamese

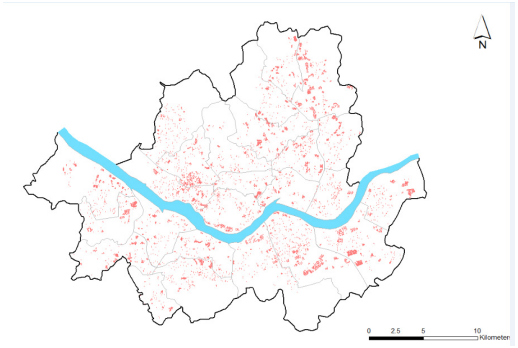

그림 13. 공간화된 일본인의 현재 거주현황
Fig. 13. Spatial the current state of Japanese

이 외에도 전체적인 국가를 한 공간에 표현하고(그림 14), 각 국가별로 분류를 해 보았다. 현재 성능평가에서 파악할 수 있는 특징 이외에도 보는 사용자에 따라 또 다른 특징이나 연관성, 패턴 등을 파악할 수 있을 거라 기대한다.

그림 14. 서울시의 전체 외국인 거주현황
Fig. 14. Current state of the entire foreign residents in Seoul

V. 결론

본 논문에서는 텍스트 데이터를 이용해서 사용자가 파악하기 쉽게 시각적인 표현을 한다는 계획대로 공간 데이터마이닝을 실행할 수 있었다. 중요한 것은 텍스트를 어떻게 공간화 시키냐는 것인데, 본 연구처럼 코드화를 통해서 KLSI의 키(key) 값과 매칭을 시키는 방법을 사용한다면 데이터를 시각적으로 이용할 수가 있다. 본 연구에서 부족했던 점은 좀 더 다양한 주제별로 묶어볼 수 있는 방법을 고안하지 못했다는 것과, 코드화와 매칭을 시켜서 공간화 하는 데에 상당한 시간이 소요 되었다는 것이다. 그리고 추가적으로 GPS 값과의 매칭에 대해서 고안해 본다면 더욱 유용하게 이용할 수 있을 것이라고 보인다. 보는 사용자에 따라 파악되는 특징이 다를 수도 있다는 것이 데이터마이닝의 흥미로운 점이다. 이번 결과 또한 본 연구에서 파악한 결과 이외에도 보는 사용자에 따라서 또 다른 정보들이 나타날 수 있기 때문에 여러 분야별 사용자의 의사결정에 도움을 줄 수 있다고 생각한다.

References

[01] http://cif.iis.u-tokyo.ac.jp/e-society

[02] http://www.kddi.com/variety/wireless_japan/pdf/

[03] http://www.nec.co.jp/rd/datamining/

[04] M. Ester et al., “Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support,” Data Mining and Knowledge Discovery, Vol. 4, pp. 193-216, 2000.

[05] M. Ester, H. Kriegel, and J. Sander, “Algorithms and Applications for Spatial Data Mining,” Geographic Data Mining and Knowledge discovery, 2001.

[06] J. Mennis and J. Liu, “Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change,” Transactions in GIS, Vol. 9, No. 1, pp. 5-17,2005.

[07] F. Verhein and S. Chawla, “Mining Spatio-Temporal Association Rules, Sources, Sinks, Stationary Regions and Thoroughfares in Object Mobility Databases,” In Proc. Int'l. Conf. on Database Systems for Advanced Applications, DASFAA, pp. 187-201, 2006.

[08] Duck-Ho Bae, Ji-Haeng Baek, Hyun-Kyo Oh, Ju-Won Song, “Design and Implementation of a Spatial Data Mining System,” Journal of Korea Spatial Information Society, vol.11, no.2, pp.119-132, June 2009..

[09] Gunhak Lee, “A Study on Spatial Patterns of Traffic Accidents using GIS and Spatial Data Mining Methods: A Case Study of Kangnam-gu, Seoul,” Journal of Korean Geographical Society, vol.39, no.3, pp. 457-472, 2004.

[10] Qin Ding, Qiang Ding, and William Perrizo, “PARM－An Efficient Algorithm to Mine Association Rules From Spatial Data,” IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS－PARTB: CYBERNETICS, vol. 38, no. 6, December 2008.

※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2012046780).

저자 소개

김 기 범(준회원)

• 2013년 : 홍익대학교 컴퓨터정보통신공학과 졸업(BS)

<주관심분야 : GIS Spatial Analysis, Data Mining, Database >

안 병 구(종신회원)

• 1988년 : 경북대학교 전자공학과 (BS)
• 1996년 : (미)Polytechnic University, Dept. of Computer and Electrical Eng.,USA (MS).
• 2002년 : (미)New Jersey Institute of Technology(NJIT), Dept. of Computer and Electrical Eng., USA. (Ph.D)
• 1989년 ~ 1994년 : 포항산업과학기술연구원(RIST), 선임연구원
• 2003년 ~ 현재 : 홍익대학교 컴퓨터정보통신공학과 교수
• 2012년 : 대한전자공학회 컴퓨터소사이어티 회장

<주관심분야 : Wireless Networks, Ad-hoc & Sensor Networks, Multicast Routing, QoS Routing, Cross-Layer Technology, Cooperative Communication, Network Coding, Bioinformatics, VLC, Network Security>