지방자치단체 통합 데이터 활용성 분석 : 데이터마이닝 기법을 활용하여*

Analysis of the Utilization of Integrated Data on Local Governments: Using Data Mining Techniques

임 광 혁
배재대학교

Kwang Hyuk Im
Paichai University

요약

자체별 또는 연도별과 같은 공통 요소를 통한 교차 분석이 가능하도록 통합 DB를 설계함으로써 다면적인 연구 분석을 위한 프레임워크를 구축하는 데 중점을 두어야 한다.
본 연구는 지자체 통합 데이터베이스에 구축되고 있는 지자체 통합 데이터의 활용성 검증을 위하여, 데이터마이닝 기법을 적용하여 지자체 통합 데이터를 분석하고자 한다. 데이터마이닝 기법을 적용하기 위해서 데이터의 추출, 적재, 변환의 과정을 수행하고, 이렇게 준비된 데이터를 실제 데이터마이닝 기법을 이용하여 분석하고 결과를 검증하는 과정을 수행하면서 지자체 통합 데이터의 활용성을 검증하고자 한다.

I. 서론

지방 균형발전을 위한 지자체 현황 분석을 위해서는 종합적인 자료의 활용이 필수적이지만, 경제, 사회, 문화(축제) 전반에 걸친 지자체들의 다양한 예산집행 활동 및 실태에 대해서는 각 예산별로 중앙정부 담당 부처가 다른 이유로 통합적인 관리가 이루어지지 않고 있다. 지자체에 대한 체계적이고 다각적인 분석과 연구를 지원하기 위해서는 이렇게 분산되어 있는 데이터를 종합적으로 제공할 수 있는 지자체 통합 DB구축이 필요하다. 이러한 필요성을 바탕으로 지자체 통합 DB 구축이 진행되고 있다.

구축되고 있는 지자체 통합DB의 활용성을 검증할 필요가 있으며 통계 기반 분석과 아울러 데이터마이닝 기반 분석도 병행하여 활용성을 검증할 수 있다. 본 연구는 데이터마이닝 기법을 이용하여 활용성 분석을 수행하고자 한다.

Ⅱ. 지자체 통합DB

1. 통합 코드 체계

다양한 데이터를 통합하기 위해서는 통합 코드 체계를 구축하여야 한다. 통합 코드 체계에는 대분류, 소분류, 소분류 코드를 정의하고 각 데이터는 이 통합 코드 체계에 맞게 데이터 구조를 정의되어야 한다.

통합 코드 체계와 아울러 지자체 통합 DB에서는 지역과 데이터 조사 연도가 매우 중요하므로 각 데이터는 통합 코드와 아울러 지역코드와 기준연도를 저장하여야 한다.

통합 코드 체계와 지역코드, 기준연도는 대부분의 데이터 테이블에 포함되어야 한다. 데이터 테이블의 예시는 〈표 1〉과 같다.

표 1. 데이터 테이블(예시)

필드명	필드설명
대분류코드	FK, PK
중분류코드	FK, PK
소분류코드	FK, PK
지역코드	FK, PK
기준연도	PK
실데이터	INT, Float, etc

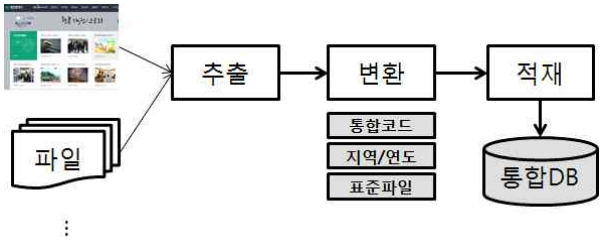
* PK : Primary Key, FK : Foreign Key

2. 통합 DB 구축

지자체 통합DB의 데이터의 원천이 중앙정부에 집적되어 있는 경우도 있지만, 대부분 각 부처별로 분산되어 있고 지방자치단체 자체예산으로 시행된 경우 각 지방자치단체 해당 부서나 대행집행 기관에 데이터가 축적되어 있다.

* 이 논문 또는 저서는 2015년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015S1A5B4A01036638).

이렇게 분산되어 있는 데이터는 온라인 및 오프라인 채널을 통하여 수집하고 표준 서식에 맞춰 데이터를 변환한 후 데이터베이스에 적재하게 된다. 모든 항목을 수집할 수는 없기 때문에 지자체 효율성 분석과 관련된 항목을 우선적으로 수집하여 다양한 관점에서 지자체 효율성 분석을 수행하고 그 결과를 바탕으로 지방균형발전의 방향을 제시할 수 있다[1-5].



▶▶ 그림 1. 데이터 추출, 변환, 적재 과정

실데이터 추출, 변환 및 적재를 통하여 데이터 테이블에 저장된 실제 데이터는 〈표 2〉와 같다. 대분류, 중분류, 소분류는 생략하여 나타내었다.

표 2. 연도별, 기초단체별 교통사고 관련 통계데이터

기준연도	지역코드	교통사고 발생건	사망자수	부상자수
2015	11110	1,414	8	1,933
2015	11140	1,338	8	1,803
2015	11170	1,516	18	2,110
2015	26110	337	6	456
2015	26170	528	12	700
		...		
2015	26200	394	6	524

Ⅲ. 데이터 활용성 분석

1. 통계 기반 분석

엑셀을 통한 그래프 분석을 포함하여 연관분석, 회귀분석, 구조방정식을 통한 분석, DEA분석 등 다양한 통계기법을 기반으로 데이터를 분석 및 활용할 수 있다. 통합DB의 각 항목 중에서 지자체 경영효율성 분석에 영향을 미칠 수 있는 항목들을 선택하여 다양한 통계 기법을 통하여 분석할 수 있다. 통계 데이터는 현황 파악에 용이하고 연관 항목 추출 및 영향력 분석을 통하여 직접적인 영향을 분석할 수 있지만, 비슷한 성향을 나타내는 지자체를 그룹핑하거나, 숨겨진 유용한 지식을 발견하기에는 한계가 있다.

2. 데이터마이닝 기반 분석

데이터마이닝은 대용량 데이터를 분석하여 숨겨진 지식을 찾는 기법으로 문제유형은 크게 Clustering, Classification, Affinity Grouping, Estimation(Prediction)으로 나눌 수 있다. Clustering 은 유사한 데이터를 가지고 있는 개체끼리 그룹화하는 유형이고, Classification 은 그룹화 되어 있는 데이터를 구분하는 규칙이나, 그룹 내의 특성을 정의할 수 있는 유형이다. Affinity Grouping 은 서로 연관되어 있는 데이터를 찾는 유형으로 시간을 고려할 수도 있고 시간을 고려하지 않을 수도 있다. Estimation(Prediction)은 과거 데이터를 분석하여 규칙을 찾아내고 이를 이용하여 미래를 예측하는 유형이다.

지자체 통합DB에 저장된 지자체 데이터를 분석하여 비슷한 성향을 나타내는 지자체끼리 그룹화 할 수 있고, 그룹화 되어 있는 지자체의 특성을 추출할 수 있다. 경제, 사회, 문화별로 연관이 있는 지자체를 추출하거나 다양한 통계데이터의 향후 변화를 예측할 수도 있다.

지금까지 수행해 왔던 통계 기반의 분석에 데이터마이닝 기반 분석을 추가한다면 훨씬 다양한 분석 결과 데이터를 얻을 수 있으며, 이를 통하여 다양한 정책 결정을 수행할 수 있을 것이다.

■ 참 고 문 헌 ■

[1] 이혁주, 박희봉, “도시행정서비스의 생산특성과 비효율 분석”, 한국행정학보, 제30권, 제4호, pp.121-137, 1997
[2] 문춘걸, “자료포락분석법 및 그 변형기법을 통한 공공부문의 생산성 측정: 한국 중소도시의 생산성 분석”, 한국조세연구원 재정포럼, pp.67-69, 1998
[3] 김재홍, “도농통합 행정구역 개편이 지방정부의 효율성 변화에 미친 영향 연구”, 한국정책학회보, 제9권, 제2호, pp.48-66, 2000
[4] 이시원, 민병익, “시 · 군 통합에 따른 행정구역 개편의 효율성 분석 : DEA분석 기법을 적용한 통합 전후의 비교를 중심으로”, 한국사회와행정연구, 제12권, 제3호, pp.79-101, 2001
[5] 고동원, “지방자치단체 사회복지시설 운영의 효율성 분석”, 한국콘텐츠학회논문지, 제11권, 제9호, pp.238-245, 2011