Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 28 - 29, 2006, pp. 195-197

기업 데이터의 한계와 외부데이터 보강 및 활용 사례

김은석[1] · 엄익현[2]

요 약

근래 기업은 영리 활동 중에 발생하는 대용량의 데이터를 축적하여 이윤창출에 도움이 되는 분야에 활용하고 있다. 기업에서 축적하고 있는 데이터는 주로 두 종류로 구분할 수 있는데, 이는 바로 트랜잭션(transaction) 데이터와 고객 속성(customer profile : 인구사회적 변수 + 고객신상변수) 데이터이다.

본 논고에서는 기업이 축적하고 있는 기업 데이터의 한계를 논하고, 외국의 사례를 통해 고객분석에 필요한 정보와 이를 확보하기 위한 방법을 살펴본다. 또한 기업에서 필요로 하는 외부데이터를 병합하기 위한 방법과 보강된 외부 데이터에 대한 활용 사례를 살펴보고자 한다.

트랜잭션 데이터는 기업 사정에 맞게 여러 가지 주제로 다양하게 활용되고 있으며, 기업에서는 다양한 통계 기법을 이용하여 마케팅 수행 또는 리스크 관리에 의미있는 분석 결과를 도출하고 있다. 트랜잭션 데이터는 일반적으로 규모가 방대(tera 수준)하며, 물리적/논리적으로 잘 정리(data warehousing)되어야 통계분석이 가능하나, 오늘날 시스템 성능 향상과 다양한 분석 기법의 적용으로 대용량 데이터를 어렵지 않게 처리할 수 있게 되었다. 다만, 트랜잭션 데이터를 분석하기 위해서는 사전에 반드시 데이터 정제가 필요하며, 이는 데이터 분석의 전 처리(pre-processing) 과정으로 고려하여야 한다.

고객 속성 데이터는 고객 분석 상에 중요한 의미를 가지나, 데이터의 갱신 및 필요정보의 부족으로 인해 심각한 한계를 지니고 있다.

외국의 경우(美, 佛), 고객분석에 필요한 고객 신상 정보 및 관련된 서비스를 제공하는 전문 데이터베이스 마케팅(DBM) 전문 기업이 활발하게 활동하고 있다. 1917년에 설립된 미국의 Donnelley Marketing이 대표적인 기업이라고 할 수 있다. 이러한 기업에서는 기업이 원하는 리스트(DB) 뿐만 아니라, 지속적인 데이터 수집을 바탕으로 각종 마케팅과 관련된 지수 및 데이터 보강 도구(system)를 개발, 판매하고 있으며, 필요 데이터의 보강은 주로 지역속성을 기반으로 제공된다. 즉, 특정 지역 내 거주자들의 인구사회학적 변수, 그들의 라이프스타일, 구매이력 정보 및 그 지역과 유사한 타 지역의 정보가 제공되고 있다.

[1]135-818 서울시 강남구 논현동 81-10 신창빌딩 4층, (주)지디에스케이, 대표이사/통계학박사.
E-mail : eskim@gdskorea.co.kr
[2]135-818 서울시 강남구 논현동 81-10 신창빌딩 4층, (주)지디에스케이, 이사/통계학박사.
E-mail : abodata@gdskorea.co.kr

외국과 같은 내용을 우리나라의 기업에게 적용하기 위해서는 여러 소스의 데이터를 통합할 수 있는 기술(data fusion techique)이 필요하다. 주민번호를 통해 데이터를 결합하면 쉬우나, 각 정보에는 주민번호를 포함하고 있지 않고, 있다하더라도 법적 문제가 있어 사용이 쉽지 않다는 것이 데이터 통합 기술이 필요한 이유라고 할 수 있다.

데이터 통합의 대상이 되는 데이터는 각기 다른 소스의 정보 (기업, 부동산, 정부)이며, 집계 단위 또한 다양하다. 대표적인 예로 행자부에서 매월 발표하는 "행정동별 성별/연령별 인구수"는 행정동 단위로 집계되며, "공동주택 평형/시세 데이터"는 호(戶)단위로 집계된다. 이러한 지역속성 데이터를 고객 속성 데이터와 병합하기 위해서는 key변수가 필요한데, 주소가 그 것이다.

그러나, 고객기재 주소는 우리나라 주소체계 문제(행정동, 법정동, 우편동의 혼용)와 기입의 다양성으로 인해 key변수 역할이 쉽지 않다. 기업 내부 고객 정보와 공동주택 DB나 센서스 DB를 주소에 의해 병합(Merge)하기 위해서는 기본적으로 주소가 단일 체계화되어야 하며, 이를 주소 표준화 기술이라고 한다. 물론 대용량의 고객주소를 표준화하기 위해서는 자동화된 시스템이 요구된다.

기업들은 주소 표준화를 통해 얻은 표준화된 주소를 key변수로 사용하여, 여러 지역속성 변수를 병합하고 이를 마케팅이나 리스크 관리에 활발히 활용하고 있다. 공동주택 정보를 이용하여 공동주택 거주자에 대한 거주공간의 평형, 시세 및 기타 변수에 대한 보강을 한다든지, 우리나라 센서스 정보를 이용하여 지역속성 정보 및 사회경제적 변수를 보강하는 것 등이 데이터 보강에 대한 대표적인 사례이다.

보강된 데이터를 활용하는 사례로는 가구 및 지역별 정보를 보강하여 산출된 데이터세트에 대해 데이터 마이닝 기법(신경망, 의사결정나무분석, 회귀분석)을 적용하여 가구별 소득을 산출하는 것을 들 수 있다, 이러한 추정 소득은 은행 또는 캐피탈사의 개인 대출이나 신용카드사의 한도 부여의 기준으로 활용되고 있다.

주요용어 : 기업 데이터, 트랜잭션 데이터, 고객속성 데이터, 데이터 통합 기술 주소표준화, 외부 데이터 보강.

Limitation of Business Data & Enrichment of External Data and Examples of Application

Eunseok Kim[1], *Ickhyun Um*[2]

Abstract

Recently, businesses are accumulating massive data produced from profit activity to apply in fields that helps creating profit. Data that is being accumulated by businesses can largely categorized in two types, which are Transaction data and Customer profile data.

This study will discuss the limitation of business data that is been accumulated by businesses and through analyzing examples of other countries, find information for customer analysis and methods to acquire it. Moreover, the study will find out how to merge external data required in businesses and examples of enriched external data application.

Keywords : Business Data, Transaction Data, Customer Profile, Enrichment, Merge/ Purge, Data Fusion.

[1]CEO/Ph.D., GDS Korea Inc., SinChang B/D 4F, 81-10, Nonhyun-dong, Kangnam-gu, Seoul 135-818, Korea. E-mail : eskim@gdskorea.co.kr

[2]Director/Ph.D., GDS Korea Inc., SinChang B/D 4F, 81-10, Nonhyun-dong, Kangnam-gu, Seoul 135-818, Korea. E-mail : abodata@gdskorea.co.kr