데이터 마이닝을 이용한 빅데이터 분석 시스템 적용 방안

전중호*, 박석천**, 김종현***
*가천대학교 일반대학원 모바일소프트웨어학과
**가천대학교 컴퓨터공학과 정교수(교신저자)
***위세아이텍 대표이사
lover030237@naver.com

Applied Method of Analysis System Using Data Mining for Big Data

Jung-Ho Jeon*, Seok-Cheon Park**, Jung-Hyun Kim***
*Dept. of Mobile Software, Gachon University
**Dept. of Computer Engineering, Gachon University
***Representative Director, WISEITECH co., ltd

요 약

스마트폰의 배급과 SNS의 성장으로 최근 데이터양은 급증하고 있다. IDC에 따르면 지난 10년간 생성된 데이터 보다 최근 2년 사이에 생성된 데이터양이 많은 걸로 나타났고 앞으로 점점 늘어날 것으로 예상된다. 이러한 대규모의 데이터인 빅데이터가 사회적 이슈가 되고 있고 이를 활용하려는 시도가 끊임없이 일어나고 있다. 본 논문에서는 빅데이터 상의 데이터 마이닝을 통하여 고객의 패턴을 분석하고 이용자에게 신뢰성 있는 데이터를 제공 할 수 있는 방안을 제시한다.

Ⅰ. 서론

최근 스마트폰과 트위터, 페이스북등 SNS의 확산 등으로 데이터양이 기하급수적으로 증가함에 따라 “빅데이터”가 사회적 이슈가 되고 있다.

미국의 IT분야 리서치 및 자문회사인 가트너는 (Gartner)는 빅데이터 이슈에 대해 2012년에 이어서 2013년의 10대 전략 기술로 선정하고 있을 정도로 IT 산업의 핵심 키워드로 대두되고 있다.

빅데이터는 과거 데이터에 비해 규모가 크고 형태가 다양하여 기존의 방법으로 수집, 저장, 분석이 어려운 방대한 크기의 데이터를 의미한다.

기업에서는 이러한 방대한 크기의 데이터인 빅데이터 분석을 통해서 고객에게 필요한 서비스를 제공함으로서 이익을 창출하고자 하는 시도가 계속 되고 있다.

본 논문에서는 빅데이터의 데이터마이닝을 통해 고객에게 맞춤 정보를 제공함으로서 고객에게 신뢰성 있는 정보와 편의를 제공 할 수 있는 빅데이터 분석 시스템 적용방안을 제시하고자 한다.

Ⅱ. 빅데이터의 개요

2.1 빅데이터의 정의

빅데이터는 통상적으로 일반적인 데이터베이스, 소프트웨어로 관리가 어려운 대용량 데이터를 의미한다. 최근에는 대용량 데이터를 수집, 저장, 분석하기 위한 도구, 플랫폼, 분석기법등을 포괄하는 용어로 변화하고 있다.

기존의 관계형 데이터와 비교하여 양, 속도, 다양성 및 복잡성에서 그 차이를 볼 수 있다.

데이터에는 정형화된 데이터와 비정형화된 데이터가 있는데 최근에 논의되고 있는 빅데이터는 정형화된 데이터든 아니든 상관없이 엄청난 양의 데이터를 말한다. 빅데이터에 대한 정의는 다양하지만 기업적인 측면에서의 빅데이터를 기업의 효과적인 전략 도출에 필요한 상세하고 높은 빈도로 생성되는 다양한 종류의 데이터로 정의할 수 있다[1].

다음 [표2.1]은 각 기관에서 내린 빅데이터의 정의이다.

[표2.1]각 기관의 빅데이터 정의

기관	정의
맥킨지(2011)	기존 방식의 저장, 관리, 분석 할 수 있는 범위를 초과하는 규모의 데이터[7]
IDC(2011)	다양한 종류의 빅데이터로 부터 낮은 비용으로 가치를 추출하고 데이터 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처로 정의[8]
가트너(2011)	빅데이터는 21세기의 원유로 다양한 종류의 데이터가 기업이 감당할 수 없을 정도로 빠르게 생성되는 현상으로 정의

2.2 빅데이터의 구성요소

빅데이터는 일반적으로 데이터규모(Volume),데이터 속도(Velocity), 데이터 다양성(Variety)등의 3가지요소에 1V(Value)나 1C(Complexity)가 추가되어 설명한다.

데이터의 규모(Volume)은 데이터의 발생량으로 페타바이트, 제타바이트 이상의 정보를 기준으로 물리적 크기뿐만 아니라 데이터를 처리하는데 어려움이 없는지 여부를 의미한다.[2]

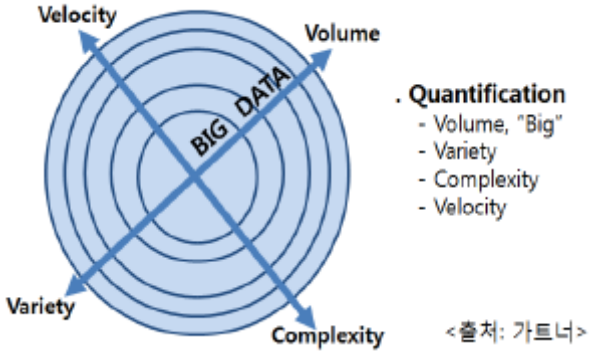
데이터의 속도(Velocity)는 데이터를 처리하는 속도로 배치 분석만을 의미하는 것이 아니라 필요에 따라 수많은 사용자 요청을 실시간으로 처리한 후 처리결과를 반환해주는 기능에 대한 시간적인 의미이다.

데이터의 다양성(Variety)은 기존의 데이터베이스에서 관리하는 구조적인 데이터 사용된 수치 및 텍스트뿐만 아니라 사진, 동영상 등 기존의 구조화된 데이터가 아닌 다양한 형태의 데이터를 의미한다.

가트너는 앞에 언급한 3V에 복잡성(Complexity)을 추가하여 4개의 축으로 빅데이터를 설명하였다.

데이터의 복잡성(Complexity)은 데이터 구조, 도메인 규칙, 저장 타입 등 데이터의 발생, 처리, 정제 등의 과정에 포함된 모든 요소가 복잡해지는 것을 의미한다[1].

아래 [그림2.1]은 가트너가 제시한 빅데이터의 구성요소를 나타낸다.



[그림2.1] 빅데이터의 구성요소

2.3 빅데이터의 국내 활용사례

국내 기업에서도 대기업을 중심으로 빅데이터를 분석하여 고객에게 필요한 정보를 제공하는 서비스들이 늘어가고 있다.

SK텔레콤의 네비게이션 서비스인 티맵은 프로그램을 기계에 내장하는 일반 내비게이션과 달리 SK텔레콤 서버에 접속하여 고성능 컴퓨터가 계산한 길안내 결과를 수신하며 전국 도로의 교통정보를 5분 단위로 실시간 수집 분석하고 있다.

신세계백화점에서는 2011년 인천 명품관을 오픈하면서 빅데이터를 활용하여 "타깃 고객 마케팅"을 실시하였다. 데이터 마이닝 기술 이용하여 특정 고객을 집중 공략 한 결과 4만명의 손님이 매장을 방문하였고 4만명 손님 중 2만명 이상이 실제 인천 점에서 명품을 구매하였다. .

소셜 미디어 분석 서비스 기업인 다음소프트는 '와인오퍼'를 개발하여 와인 가격, 포도품종, 어울리는 음식등 와인과 관련된 정보를 제공하고 있으며 사용자 관심사 및 상황에 맞게 와인을 추천하고 다양한 와인 랭킹을 제공하고 있다.

국내 기업뿐만 아니라 정부에서도 빅데이터를 활용한 스마트 정부구현방안을 마련하여 2012년 8월 마스터플랜을 기반으로 빅데이터 활용가능성 검증을 위해 파일럿을 구축하여 시범 사업을 추진 중에 있다.

이외에도 국민 권익위원회, 한국석유공사 국립 보건 연구회등 여러 분야에서 빅데이터를 활용하여 서비스를 제공하고 있다[2].

Ⅲ. 빅데이터의 분석기법

앞에서 언급한 것과 같이 빅데이터란 방대한양의 데이터를 의미한다. 이런 방대한양의 데이터에서 필요한 정보를 추출하여 사용하려면 데이터 마이닝 기술을 사용하여야 한다.

데이터 마이닝이란 데이터에서 숨겨진 패턴과 관계 등을 파악하여 의사결정이나 미래를 전망 할 수 있는 유용한 정보를 추출한다는 의미로 사용되고 있다. 최근 소셜 미디어등 비정형 데이터의 증가로 인해 분석기법들 중 텍스트 마이닝 ,오피니언 마이닝, 소셜 네트워크분석, 군집분석등이 주목 받고 있다[5].

3.1 텍스트 마이닝(Text Mining)

텍스트 마이닝은 비/반정형 텍스트 데이터에서 자연어처리(Natural Language Processing) 기술을 기반으로 유용한 정보를 추출,가공 하는 것을 목적으로 하는 기술이다.

텍스트 마이닝 기술을 통해 텍스트 뭉치에서 의미있는 정보들을 추추해내고 다른 정보와의 연계성을 파악하여, 텍스트가 가진 카테고리를 찾아내거나 단순한 정보 검색 그 이상의 결과를 얻어 낼 수 있다.

3.2 오피니언 마이닝(Opinion Mining)

오피니언마이닝은 소셜 미디어와 웹사이트 등에 나타난 여론과 의견을 분석하여 유용한 정보로 재가공하는 기술이다.

오피니언마이닝을 통해 정형/비정형 텍스트의 긍정, 부정, 중립 등의 선호도를 판별 할 수있다[6].

3.3 소셜 마이닝(Social Mining)

소셜 마이닝 혹은 소셜 네트워크 분석은(Social Network Analysis)은 수학의 그래프 이론을 기반으로 하고 있으며 소셜 네트워크 연결구조 및 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 측정하여, 소셜 네트워크상에서 입소문 중심이나 허브 역할을 하는 사용자를 찾는데 활용된다.

소셜 미디어에 올라오는 글과 사용자를 분석하여 소비자의 흐름이나 패턴 등을 분석하고, 판매나 홍보에 적용할 뿐만 아니라 사회의 흐름과 트랜드, 여론 변화 추이를 읽어내는 소셜 미디어 시대의 새로운 마이닝 기법이다[4].

3.4 군집분석(Cluster Analysis)

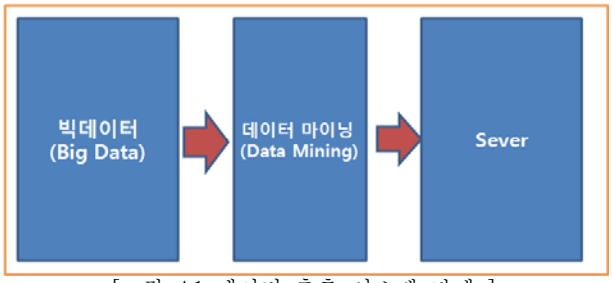
군집분석은 비슷한 특성이 있는 개체를 합쳐가면서 최종적으로 유사 특성의 군을 발굴 하는데 사용된다.

군집분석을 이용하면 페이스북, 트위터같은 소셜 네크워크 상에서 다양한 관심사나 취미에 따른 사용자군을 분류 할 수 있다[6].

Ⅳ. 빅데이터의 분석 시스템 적용방안

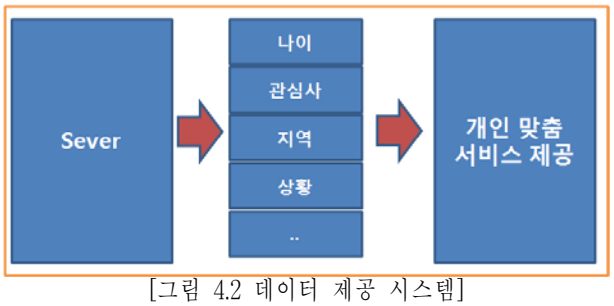
기존의 많은 기업이나 국가에서 빅데이터를 이용하여 개인에게 맞는 서비스들을 실행하고 있지만 간혹 상황에 맞지 않는 서비스를 제공하여 서비스를 이용하는 이용자에게 신뢰성을 잃는 경우들이 있다.

이점을 착안하여 본 논문에서는 [그림 4.1] 과 같이 빅데이터에 데이터마이닝 기법을 통하여 고객들의 나이 , 지역 , 관심사 등 카테고리 별로 서버에 저장 후 서버에 있는 데이터를 기반으로 이용자에게 신뢰성 있는 데이터를 제공 하고자 한다.



[그림 4.1 데이터 추출 시스템 단계]

여러 마이닝 기법을 통해서 나이, 지역 , 관심사등 카테고리별로 구분되어 있는 데이터를 서버에 미리 저장하고 [그림 4.2] 와 같이 이용자의 상황 , 나이 , 관심사 , 지역 등을 고려하여 이용자에 맞는 서비스를 제공 하도록 한다.



[그림 4.2 데이터 제공 시스템]

이러한 서비스 구성은 서비스를 제공해주는 입장에서도 다양한 이용자들의 성향을 먼저 분석하기 때문에 성향에 맞게 구매 전략을 세울 수 있으며 이용자에 경우 필요한 정보들만 얻을 수 있기 때문에 서비스에 신뢰감을 얻을 수 있을 것으로 판단된다.

Ⅳ. 결 론

스마트폰의 배급과 SNS의 성장으로 최근 데이터양은 급증하고 있다. IDC에 따르면 지난 10년간 생성된 데이터 보다 최근 2년 사이에 생성된 데이터양이 많은 걸로 나타났고 앞으로 점점 늘어날 것으로 예상된다.

다양한 분야에서 이러한 대규모의 데이터를 활용하려는 노력이 끊임없이 일어나고 있다.

본 논문에서는 빅데이터 환경에서 데이터마이닝 처리한 데이터를 서버에 저장한 후 이용자의 상황에 맞는 서비스를 제공하는 모델을 제안하였다.

서비스를 제공해주는 입장에서는 데이터 마이닝을 통해 이용자들의 경향 등을 분석하고 이용자는 신뢰성 있는 데이터를 제공받을 수 있을 것으로 예상된다.

향후 스마트 폰의 GPS 기능과 연계하여 상황에 맞는 서비스를 재공 하고자 한다. 또한 빅데이터는 다양한 경로로부터 데이터를 제공 받기 때문에 데이터 마이닝을 함에 있어서 개인정보유출 등의 보안에 관련된 부분을 보강 할 예정이다.

빅데이터를 단순히 수집 저장하는 목적이 아니라 효과적인 분석을 통해 이용자의 패턴을 찾아 낼 수 있다면 빅데이터는 기업이나 국가 간의 큰 경쟁력이 될 수 있을 것이다.

사사의 글

본 연구는 2013년도 지식경제부의 SW 전문 인력양성사업의 재원으로 정보통신산업진흥원의 고용계약형 SW석사과정 지원사업(HB301-13-1003)으로부터 지원받아 수행되었습니다.

참고문헌

[1]김지숙,“빅데이터 활용과 분석기술 고찰”,고려대학교,2013

[2]박준규,“빅데이터를 위한 분석기술 활용방안”,세종대학교,2013

[3]배동민,박현수,오기환, “빅데이터 연구 동향과 시사점”, 정보통신정책연구원,2013

[4]하연편집부 ,“빅데이터와 DBMS의 시장전망”,하연 ,2012

[5]정지선,“성공적인 빅데이터 활용을 위한 3대요소”,한국정보화 진흥원,2012

[6]김정숙,“빅데이터 활용과 관련기술 고찰”, 한국콘텐츠학회,2012

[7]ames Manyika & Michael Chui, “Big data: The next frontier for innovation, competition, and productivity”, McKinsey Global Institute, 2011

[8] John Gantz & David Reinsel,“ Extracting Value from Chaos ”, IDC IVIEW June, 2011년

[10]http://ko.wikipedia.org/wiki/빅데이터