빅데이터 품질관리 방법 및 활용 방안

안하철*, 박석천**, 김종현***
*가천대학교 일반대학원 모바일소프트웨어학과
**가천대학교 컴퓨터공학과 정교수(교신저자)
***위세아이텍 대표이사
e-mail : ahc84@nate.com

Quality Management and Utilization Method for Big Data

Ha-Chul An*, Seok-Cheon Park**, Jung-Hyun Kim***
*Dept of Mobile Software, Gachon University
**Dept of Computer Engineering, Gachon University(Corresponding Author)
***Representative Director, WISEITECH co., ltd

요 약

최근 IT와 미디어의 발전으로 인하여 데이터의 양이 많아졌다. 기존에 컴퓨터를 이용하여 인터넷을 이용했던 것을 넘어 스마트 폰의 보급으로 인한 모바일 시장이 급격이 성장함으로써 데이터의 양은 급격히 증가하고 있는 추세이다. 이와 같이, 엄청난 데이터들을 저장하고 관리하며 분석 할 수 있는 기술이 필요하게 되면서 등장한 것이 빅데이터이다. 빅데이터는 다양한 정보가 결합하고 있는데 이 데이터의 가치를 누가 먼저 효율적으로 추출해 내는 것이 기업의 성패를 가늠할 만큼 중요하게 되었다. 따라서 본 논문에서는 빅데이터를 효율적으로 품질 관리함으로써 정보의 가치를 높이고 신뢰성 있는 데이터로 만들어 활용할 수 있는 방법에 대해 연구한다.

Ⅰ. 서 론

미래의 정보기술 가운데 빅데이터(Big Data)는 최고의 이슈로 대두 되고 있다. 스마트 폰의 보급이 보편화 되면서 누구나 스마트폰을 소유하는 시대가 열었고, 이와 더불어 모바일 시장이 크게 성장하게 되었다. 또한 많은 사용자들이 이용함에 따라 더 많은 양의 콘텐츠를 제공을 해야 하기 때문에 그만큼 데이터는 점점 증가 할 수밖에 없는 상황이다.

최근에 소셜네트워크서비스(Social Networks Services), 트위터(Twitter), 카카오스토리(KakaoStory) 등과 같은 새로운 미디어가 등장함에 따라 더 많은 데이터를 생산하게 되었다. 그리고 이곳에 저장된 데이터는 각 사용자의 취미, 성격, 가치를 알 수 있는 중요한 정보로 인식되면서 기업 마케팅에 활용되고 있다.

이와 같이, 빅데이터는 많은 분야에서 활용할 수 있는 중요한 데이터 원석이라고 불리고 있다. 누가 이 데이터를 어떻게 활용하는가에 따라 다이아몬드가 될 수 있는 중요한 가치가 될 수도 있는 것이다[8].

하지만, 수많은 데이터가 있는 만큼 무분별한 데이터들이 존재한다. 잘못된 정보를 가진 데이터를 마케팅에 활용하였다가 기업은 자금만 소모하는 상황이 발생하고 그 손실로 인하여 빅데이터의 데이터 신뢰성은 떨어질 수밖에 없다. 따라서 쓸모없는 정보, 가치가 없는 정보를 중요한 정보를 가진 데이터로 가공하기 위한 빅데이터 품질관리가 필요하다.

본 논문에서는 부문별하고 정리가 되지 않은 정보의 가치를 높이고 신뢰성 있는 데이터를 만들기 위한 빅데이터 품질관리 및 활용 방안을 제안한다.

Ⅱ. 관 련 연 구

2.1 빅데이터 출현배경

빅데이터의 출현 배경은 스마트폰의 확산을 들 수 있으며, 스마트폰 출현으로 가장 가치 있는 개인정보 및 위치정보가 양산되고 있다. 규모로 보면 크지는 않지만 경제적 가치는 최고인 개인정보와 결합되는 빅데이터에 최고의 부가가치를 창출하고 있다[1].

IT혁명에는 인터넷이 세계 경제의 변화를 촉진하고 있으며 그 중 빅데이터는 스마트폰의 핵심자원으로서 세계 경제 변화를 이끌 제 4의 경영자원으로 부상하고 있는 상황이다.

스마트폰 혁명과 소셜 네트워크 효과로 디지털 공간의 데이터 빅뱅이 발생하여 정보에서는 데이터 지식 확보 및 활용 방안을 요구하고, 이에 데이터 활용 시 예산 절감 및 대내외 변화에 대한 신속한 대처, 삶의 질과 정부 신뢰도

향상이 가능해 졌다고 한다.

대표적인 사례로는 그림 1에서 보는 바와 같이 오바마 정부의 필박스(Pillbox) 통한 의료개혁이다. 필박스는 약 검색 서비스로써, 빅데이터를 통하여 수집된 정보의 통계치를 분석하여 연간 5,000만 달러 비용을 절감할 수 있었으며, 독일은 연방 노동기관에서 빅데이터 활용 맞춤형 고용으로 인하여 3년간 백업 유로 비용을 절감했다.

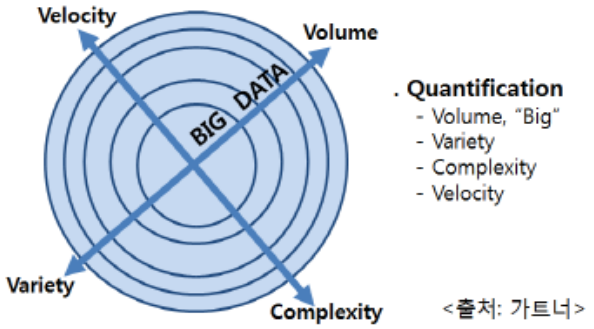
또한 기업에서는 고객 데이터 추적행위 및 수집행위가 증가하고 있는 추세이고, 소셜네트워크서비스의 급격한 확산과 비정형 데이터의 폭증과 정보 수집 가능한 센서 저변의 확대로 매달 수백억 개의 콘텐츠가 페이스북에서 공유되고 있다[1].

(그림 1) 빅데이터 사례(오바마 정부의 필박스)

2.2 빅데이터 정의

빅데이터는 '현재 시스템으로 처리 가능한 범위를 넘어서는 데이터'로 정의된다[7]. 또한, 빅데이터는 페타(Peta: 1015), 엑타(Exa: 1018), 제타(Zeta: 1021)바이트 등 기존의 데이터 단위를 넘어서는 엄청난 양(Volume), 데이터의 생성과 흐름이 매우 빠르게 진행되는 속도(Velocity), 사진, 동영상 등 기존의 구조화된 데이터가 아닌 다양한(Variety) 형태의 정보 등 3가지 속성을 가진다[5].

이처럼, 3가지 속성을 가진 데이터가 '빅데이터'라는 게 대다수 전문가들의 공통된 의견이고 그림 2와 같이 가트너는 3V에 복잡성을 추가해 3V+C로 정의하기도 한다[2].



(그림 2) 빅데이터의 정의

2.3 데이터 품질관리 정의

데이터 품질(Data Quality)이란 조직의 목적 달성을 위해 관리되는 데이터가 조직 구성원, 고객 등 데이터 이용자의 만족을 충족시킬 수 있는 수준을 의미한다.

데이터 품질관리(Data Quality Management)란 조직이 운영하는 정보시스템과 데이터베이스를 활용하는 이용자의 기대를 만족시키기 위해 지속적으로 수행하는 데이터 관리 활동을 의미한다[3].

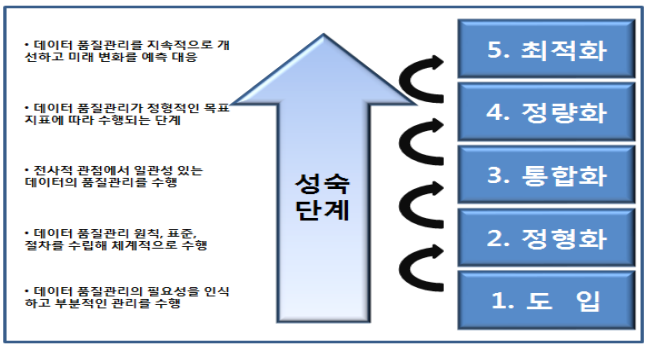
최근 기업의 업무가 정보화되면서 업무별 정보시스템 간에 심각한 데이터 중복성과 불일치성의 문제가 대두되고 있다.

예를 들어, 정보시스템 이용자의 의사결정을 효과적으로 지원하기 위해 전사 차원에서 데이터를 통합한 데이터웨어하우스(DW, Data Warehouse)를 운영하는 조직이 증가하고 있으나, 분산된 정보시스템을 통합하고 운영하는 과정에서 정보시스템에 적재된 오류 데이터가 적절히 통제되지 못하고 있다[3].

이러한 오류 데이터로 인한 잘못된 의사결정으로 피해가 발생하고 있으며 저품질 데이터를 이용한 고객의 불만도 증가하고 있다[3].

2.4 데이터 품질관리 현황

국내 한국데이터베이스진흥원에서는 매년 공공기관 데이터 품질관리 성숙수준을 온라인 설문조사로 진행하여 발표하고 있다[6]. 그림 3에서 보는 것처럼 '도입-정형화-통합화-정량화-최적화'의 1~5 레벨로 측정하고 있으며 2011에 측정된 품질관리 성숙 수준은 1.1 레벨로 조사되었다. 1.1 레벨의 '도입 단계'는 데이터 품질의 문제점과 필요성에 대해 인지하고 부분적인 데이터 품질활동을 시행하는 단계이다[4].



(그림 3) 데이터 품질관리 성숙모형
(한국데이터베이스진흥원, 2011.8)

III. 빅데이터 품질관리 방법 및 활용방안

3.1 데이터 품질관리 수행 절차

품질관리의 수행 절차는 다음과 같다. 먼저, 데이터 품질기준과 관련된 자료를 조사. 특히, 품질평가 기준, 유사

프로젝트 등을 중점적으로 조사하여 조직 내부에 적합한 품질기준의 후보를 수집한다.

그 후에 조직 내에서 활용할 데이터 품질기준은 실제 품질 진단 절차를 고려하여 선정하고, 세부 데이터 품질기준을 도출한다. 마지막으로, 업무담당자, 품질전담조직, IT담당자 등과의 협의를 거쳐 최종적으로 데이터 품질기준을 확정한다.

그림 4는 한국데이터베이스진흥원에서 제안하는 품질진단 프로젝트 절차이다. 각 단계별로 계획, 정의, 측정, 분석, 개선 등 5가지를 실행함으로써 데이터 품질관리의 수행절차를 진행한다.

(그림 4) 데이터 품질진단 절차

3.2 데이터 품질기준

표 1 은 데이터 품질기준을 유형에 따라 분류한 것으로 완전성, 유일성, 유효성, 일관성, 정확성 등 5가지로 분류할 수 있다.

<표 1> 데이터 품질기준

품질기준	상세 데이터 품질기준
완전성	컬럼 값은 누락이 없어야 한다.
유일성	컬럼 값은 유일해야 하며 중복이 되면 안 된다.
유효성	컬럼 값은 정해진 데이터 유효범위 및 도메인을 충족해야 한다.
일관성	데이터가 지켜야 할 구조, 값, 형태가 일관 되게 정의되고 서로 일치해야 한다.
정확성	실세계에 존재하는 객체의 표현 값이 정확히 반영이 되어야 한다.

기업의 환경 또는 여건에 따라서 완전성, 유일성, 유효성, 일관성, 정확성 등 범용 데이터 품질기준은 부족할 수 있기 때문에 범용 데이터 품질기준을 확장하여 상세 품질기준을 정의 할 수 있다.

상세 데이터 품질기준은 계량화되어 측정 가능한 것만을 상세 데이터 품질기준으로 도출해야 한다. 이를 위해, 품질기준의 진단방법과 결과물들을 사전에 조사하여 명확한 측정이 가능한 상세 데이터가 품질기준을 선정해야 한다.

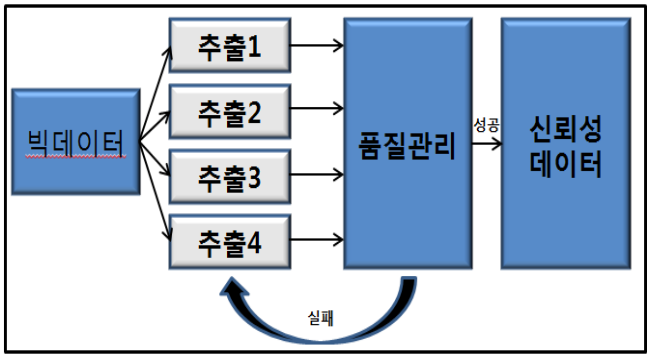
3.3 빅데이터 품질관리 방법

그림 5에서 보는 바와 같이 빅데이터는 일반적으로 가공되지 않은 상태로써 무분별한 데이터들이 있으며, 많은 정보를 가지고 있는 상태이다. 그 다음 단계는 추출단계로써 빅데이터에서 원하는 자료 부분을 추출 및 분석하는 곳이다.

본 논문에서 제안하는 것은 추출 다음 단계인 품질관리 시스템 부분이다. 이 시스템은 그 전 단계에서 추출한 데이터를 받아서 품질관리를 실시를 하며 기본적으로 한국데이터베이스진흥원에서 제공하는 품질관리 방법에 맞추어 정확한 데이터를 만드는 과정을 수행한다. 정확한 데이터로 합격을 하게 되면 성공으로써 신뢰성 있는 데이터가 생성되는 것이다.

하지만, 품질관리 시스템에서 제공하는 품질관리 방법에 통과하지 못하면 실패로써 다시 추출하는 곳으로 데이터를 되돌려 준다.

이처럼, 단계별 과정을 거친 후, 빅데이터 품질관리에서 완성된 데이터는 기존에서 사용하고 있는 하둡(Hadoop), 트위터의 스톰(Storm)과 같은 빅데이터 처리 기술에 전달하여 데이터를 처리한다.



(그림 5) 빅데이터 품질관리 시스템 구성

3.4 빅데이터 품질관리 활용방안

그 동안 빅데이터를 사용한 것 중에 가장 많이 활용한 분야는 카드 분야라고 볼 수 있다. 사람들이 사용하는 카드 내역을 분석한 후, 개인 한사람마다 사용 내역을 기준으로 쇼핑행사, 쿠폰행사 등 다양한 행사에 대한 쿠폰 및 적립금 등을 사용하는데 빅데이터를 활용해 왔다.

예를 들면, 커피를 좋아하는 사람에게는 카드 결제 시 추가 할인이 된다는 내용의 문자를 받으면 그 카드로 결재하는 것처럼 빅데이터를 사용한 마케팅 방법은 점점 변화되어 가고 있다.

또한, 이미 데이터 품질관리의 중요성을 알고 공공기관(서울시청, 산림청, 대한주택보증, 국민연금), 금융기관 등에서 데이터 품질관리를 실행하고 있다.

특히, 국민연금이나 금융기관 같이 자금과 관련된 곳은 데이터 품질관리를 하지 않으면 금전적 손실, 마케팅 대상 감소 등에 피해를 초래 할 수 있기 때문에 데이터 품질관리를 중요하게 여기고 있다.

이처럼, 많은 분야에서 활용하고 있는 빅데이터의 정보가

신뢰성 없고 정확하지 않으면 마케팅 한 기업은 잘못된 데이터로 인하여 시간 및 자금 등 손해를 볼 수밖에 없다. 따라서, 본 논문에서 제안한 빅데이터 품질관리 수행절차 및 방법을 활용하면 효율적이고 정확한 데이터와 오류를 줄일 수 있다.

Ⅳ. 결 론

최근 스마트 폰이 보편화되어 모바일 시장의 규모가 광범위해졌으며 이와 함께 많은 콘텐츠가 생성 및 발전함에 따라 데이터의 양이 증가했다. 즉, 빅데이터 환경이 구축되었으며 그 만큼 빅데이터는 중요하고 활용할 기대치가 높아졌다.

그러나, 빅데이터의 데이터 품질이 정확하지 않고 오류가 있으면 데이터를 이용하는 사람은 잘못된 정보로 판단을 하여 피해를 보기 때문에 빅데이터에 대한 신뢰도는 떨어질 수밖에 없다.

따라서, 본 논문에서는 빅데이터에서 데이터를 추출하는 과정에서 데이터 품질관리를 함으로써, 불필요한 데이터, 오류가 있는 데이터, 중복된 데이터를 제거 하여 데이터의 가치를 높이고 신뢰성 있는 데이터를 생산 및 활용하도록 제안했다.

향후, 빅데이터 품질관리 시스템을 구현하고 테스트를 진행하여 생성된 결과는 바탕으로 제안 시스템을 수정 보완할 예정이다.

Ⅴ. 사사의 글

본 연구는 2013년도 지식 경제부의 SW전문인력양성사업의 재원으로 정보통신산업진흥원의 고용계약형 SW석사과정 지원사업(HB301-13-1003)으로부터 지원받아 수행되었습니다.

참고문헌

[1] 최 성, 우성구 “빅데이터 정의, 활용 및 동향”, 2012
[2] 김정숙 ″빅 데이터 활용과 관련기술 고찰″, 2012
[3] 한국데이터베이진흥센터 “데이터 품질 가이드 라인”, 2011
[4] 한드림 “데이터 품질이 공공기관의 서비스 품질에 미치는 영향에 대한 사례연구”, 2011
[5] 김병곤 “빅데이터 기반 기술을 활용한 분산 처리 및 실시간 처리 방안“, 2012
[6] 문성은 “메타데이터와 연계한 데이터품질관리의 경제적 효과 분석 및 사례 연구”, 2013
[7] 정성우 “빅데이터 개요와 관련 기술, 그리고 오라클의 지원 전략”, 2012
[8] 이미영 “빅데이터 분석을 위한 빅데이터 처리 기술 동향”, 2012