Journal of The Korea Society of Computer and Information
Vol. 19, No. 8, August 2014

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.8.073

금융 상품 추천에 관련된 빅 데이터 활용을 위한 개발 방법

김 석 수*

A study on development method for practical use of Big Data related to recommendation to financial item

Seok-Soo Kim *

요 약

본 연구에서는 활용 기술로 데이터 저장 레이어, 데이터 처리 레이어, 데이터 분석 레이어, 시각화 레이어 등의 빅 데이터 기술을 활용한 개발 방법을 제안한다, 각 단계에서 저장, 처리, 분석된 데이터는 시각화를 통하여 볼 수 있게 하였다. Hadoop을 통하여 데이터를 처리한 후 처리된 데이터를 Mahout으로 실행하여 분석 결과를 시각화 하였다. 이 과정을 통해서 금융 상품에 가입된 고객의 여러 특성을 파악하였고, 각 고객에 따른 금융 상품의 추천을 적시에 수행할 수 있었다. 본 연구에서는 빅 데이터의 배경 및 문제점을 소개하고, 빅 데이터가 새로운 비즈니스 기회를 어떻게 창출하는지 금융상품 추천 사례를 중심으로 개발 방법과 사례 연구를 논의한다.

▸ Keywords : 빅 데이터, 하둡, 머하웃, 거버넌스, 군집화, 추천

Abstract

This study proposed development method for practical use techniques compromise data storage layer, data processing layer, data analysis layer, visualization layer. Data of storage, process, analysis of each phase can see visualization. After data process through Hadoop, the result visualize from Mahout. According to this course, we can capture several features of customer, we can choose recommendation of financial item on time. This study introduce background and problem of big data and discuss development method and case study that how to create big data has new business opportunity through financial item recommendation case.

▸ Keywords : Big Data, Hadoop, Mahout, Governance, Clustering, Recommendation

• 제1저자 : 김석수
• 투고일 : 2014. 7. 22, 심사일 : 2014. 7. 27, 게재확정일 : 2014. 8. 11.
* 가천대학교 컴퓨터공학과 (Dept. of Computer Science, Gachon University)

I. 서 론

유2012년 이후 현대사회는 정보화 사회에서 빅 데이터 서비스 사회로 점차 이동하고 있다[1]. 우리는 하루에도 수많은 데이터가 만들어지고 있는 시대에 살고 있다. 60초 안에 1억 6천 8백만 개의 E-mail이 발송되고, 51만개의 페이스북 주석이 달리며, 1만 3천개의 아이폰 어플이 다운로드 되는 등 소셜네트웍을 통한 데이터가 기하급수적으로 생성되고 있다. 정보 시스템의 고도화, 모바일, 클라우드, 소셜네트웍에서 생성되는 데이터의 양이 제타바이트 시대에 돌입하고 있다[2]. 더불어 정형 데이터보다는 비정형데이터의 폭증으로 전 세계 데이터 량은 매년 40%씩 증가하고 있다. 이렇게 증가하는 데이터를 분류해 보면 정형데이터가 15%, 비정형 데이터가 85%로 나누어진다. 정형데이터는 기존 데이터베이스 관리시스템인 Oracle, DB2, SQL 등으로 관리되어 왔으며 분석도구 역시 잘 알려진 데이터 마이닝 분석 도구를 통하여 활용되어 왔다. 기업이나 조직에서 분석에 필요한 데이터가 기존 데이터 기술로는 관리할 수 없는 데이터로 새로운 분석 도구를 필요로 한다. 특히 빅 데이터 중에서도 비구조적인 데이터로서 첫째, 로그파일 같은 반 구조 데이터, 둘째, 데이터 집합이 하나 이상의 구조를 갖는 데이터, 셋째, 하나의 데이터 세트가 어떤 구조도 갖지 않는 데이터로 구분할 수 있다[2]. 그리고 사회적인 분류로 보면 오픈데이터와 소셜데이터가 있다. 오픈데이터는 기상정보, 교통정보 등 정부, 공사 또는 지방자치단체 등이 보유하고 있는 공통데이터베이스로서 일반에 공개된 데이터이며, 소셜데이터는 소셜미디어 즉 페이스북이나 트위터 등에 개인 또는 법인이 올린 데이터로서 그 분석의 가치가 매우 중요시되고 있다. 또한 빅 데이터는 매년 60%정도 증가될 것으로 예상되고 있으며, 축적된 데이터 활용 능력이 중요한 과제로 대두되고 있다. 2020년까지 33 제타바이트까지 폭증할 것으로 예상되고 있다[3]. 오늘날의 기업환경에서 데이터에 기초한 과학적이고도 합리적인 의사결정은 이를 사용하는 개인이나 조직의 경쟁력에 결정적인 영향력을 미친다. 전통적인 데이터관리에서, 데이터베이스, 데이터마이닝, 고객관계관리 등 데이터 자원의 관리가 기술적 진보 및 경영환경의 변화에 따라 발전되어 가고 있다. 최근에 빅 데이터라는 개념은 클라우드, 소셜네트웍, 보안, 모바일 등과 함께 정보통신서비스 시장의 주요한 테마로 대두되고 있다. 특히, 기업은 빠르게 변하하는 고객 수요와 고객의 정보 지능에 대응하여 기존에는 분석되지 않았던 새로운 유형의 데이터를 정보자원으로 레버리지하여 고객의 이탈을 최소화시킨다거나, 매출 증대와 보다 향상된 품질의 제품을 제공하는 등 확장된 가치 흐름을 창출하고 있다[4].

본 연구에서는 2장에서 빅 데이터 이론 및 개발 사례를 소개하고, 3장에서는 금융 상품 추천을 위한 개발 모델을 제안한다. 4장에서는 빅 데이터가 새로운 비즈니스 기회를 어떻게 창출하는지 금융상품 추천 사례를 중심으로 개발 방법과 사례 연구를 기술한다.

II. 빅 데이터 이론 및 개발 사례

1. 빅 데이터 특징

빅 데이터의 특징은 다섯 가지이다. 첫째, Volume으로 기존 DB보다는 규모가 훨씬 크고 일정 기준으로 구분하지 않는다. 둘째, Velocity로 배치, 리얼타임, 스트림형태, 실시간 분석과 반응을 필요로 한다. 셋째, Variety로 구조적데이터와 비구조적데이터를 포함한다. 다양한 구조의 데이터를 서로 연관해서 분석할 수 있어야한다. 넷째, Complexity로 위의 3가지 특성에 따라 보관, 운영, 활용하는 것이 매우 복잡하다. 마지막, Value로서 기존 구조적데이터는 거래를 안전하게 처리하기 위한 목적이지만 이는 경쟁력 및 운영효율성에 직접 큰 영향을 줄 수 있다[5]. 빅 데이터의 가치는 크게 두 가지로 나누어 볼 수 있다. 첫 번째로, Agility를 들 수 있는데 이벤트 감지, 데이터 확보, 분석수행 의사결정, 행동착수라는 일련의 행동 과정으로 빅 데이터 분석을 빠르게 수행하여 경쟁우위 요소를 가질 수 있게 한다.

둘째로 Relevance로 VOC 감성분석, 위치정보 연계, 웹로그 분석을 통한 고객구매심리파악, 장바구니분석을 통한 구매포기 요인 파악, 센서 행동 분석 기반으로 고객 상황 인지를 통하여 연관성에 기초한 가치 있는 제안을 가능케 한다. 따라서 높은 고객만족, 재 구매 및 반복구매, 고객 이탈방지와 같은 가치 있는 행위를 유발하게 할 수 있다[6].

2. 분석의 중요성

리서치기관의 조사에 따르면 경영성과가 높은 기업이 낮은 기업보다 분석을 더 많이 활용하고 있다는 통계가 있으며 활용분야에서도 재무관리에서부터 영업, 마케팅, 고객관리는 물론이고 인력관리에까지 활용을 넓히고 있다. 특히 운영효율성, 전략수립, 고객서비스에서도 높은 활용도를 보이고 있다. 분석활용은 업종에 따라 다르며, 새로운 분석을 발견하기 위한 노력이 많아지고 있다. 실시간분석으로는 은행의 신용위험

및 시장 위험 분석, 은행의 부정사용 및 자금세탁 탐지, 금융 및 통신회사의 이벤트마케팅, 유통업종의 마크다운 최적화, 공공분야의 보상 및 과제 부정청구 등으로 들 수 있다. 배치성 분석으로는 항공회사의 예방정비, 소셜미디어 감성분석, 제조업체의 수요예측, 전자의료 기록 관리의 질병분석, 전통적 데이터웨어하우징, 마이닝 테스트, 비디오감시 분석 등이 있다.

3. 기술과 오픈소스

기존 데이터와는 다른 특성을 가진 빅 데이터는 그 분석의 기술 역시 다른 점을 가지고 있다. 이러한 특성을 해결하기 위해서도 기계가 스스로 학습할 수 있는 역량 및 표현기술, 통계분석 시스템, 자연어처리기술, 데이터를 어떻게 수집할 것인지 등에 대한 복잡한 기술이 존재한다. 이를 나열해 보면 NoSQL, Text Mining, IT(Search), NLP, Machine Learning, Semantics, Visualization, Statistics(R) 등이 있다. 그리고 이제 막 시장에서 나타나고 있는 기술로서 많은 부분에서 오픈소스 생태계를 이루고 있다. 분야로 보면, NoSQL, Cache, RPC, Script Language, Work flow, Queue, Machine Learning, Matrix file system, Search Engine 등에 많은 오픈 소스들이 경쟁력으로 나타나고 있다[7].

4. 빅 데이터 개발 사례

아직도 초기 기술로서 그다지 많은 기업이 개발하고 있지는 않지만 대기업 또는 인터넷 서비스 기업을 중심으로 시험 개발이 되고 있다. 그 내용으로는 제조업체의 수요예측 분석, 이벤트 마케팅 분석, 비정형 로그데이터 분석을 통한 서비스 품질 분석, 생산 설비 장애 분석 시스템 등이 있다.

III. 금융 상품 추천을 위한 개발 모델

본 연구에서 제안된 개발 모델은 활용 기술과 분석 프레임 워크로 구성된다.

1. 활용 기술

첫 번째 단계는 mahout 기반 추천엔진 잡의 실행에 필요한 데이터들을 Hadoop의 HDFS로 로딩 하는 것이다[7,8]. 서비스에 따라 다르겠지만 많은 경우 관계형 데이터베이스내의 테이블들이나 아파치와 같은 웹서버의 액세스 로그들이 그 대상이 된다. 이를 로딩 하는 방식은 간단하게는 hadoop fs -copyFromLocal처럼 HDFS 커맨드라인 명령을 이용하는 것일 수도 있고, 관계형 데이터베이스의 경우에는 Sqoop과 같은 오픈소스 패키지를 이용하는 것일 수도 있고, 액세스 로그의 경우에는 Flume등의 오픈소스 패키지일 수도 있다[9]. 전 단계에서 업로드 된 데이터들은 사실 그대로 사용할 수 있는 경우는 거의 없다. 어느 정도 가공을 통해서 잡음처리를 해야 하고, mahout에서 필요한 형태의 포맷으로 변환해주어야 하는데 이를 데이터 전처리라고 한다. 여기서는 전처리 또한 Hadoop상에서 맵리듀스 프로그램을 작성하여 해결한다. 추천엔진의 실행이 끝나고 나면 사용자별로 혹은 상품별로 추천되는 컨텐츠가 HDFS에 저장되게 된다. 이것을 읽어 들여서 보통 후처리를 하게 된다[10]. 마지막 단계는 최종 데이터를 웹서비스가 접근할 수 있는 어떤 레이어로 업로드 하는 것이다. 이는 관계형 데이터베이스가 될 수도 있고, 데이터의 크기가 너무 크다면 NoSQL이 될 수도 있다. ElasticSearch와 같은 검색엔진이 될 수도 있다.

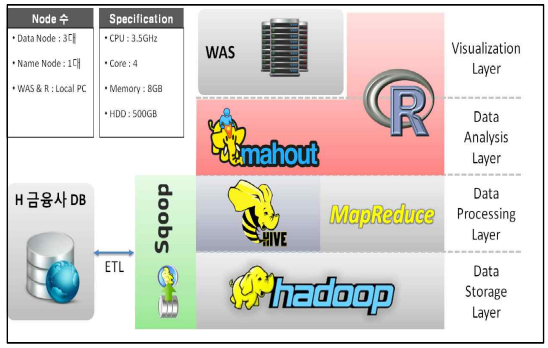

그림 1. 활용 기술
Fig. 1. Practical Use Technique

Hadoop 1.x에 해당하는 다른 배포판 들로는 클라우데라의 CDH3와 호튼웍스의 HDP1이 있다. 이것들은 대부분 아파치 Hadoop과 호환이 되기는 하지만 앞서 Single point of failure 문제를 해결하거나 Hadoop 관련 다른 패키지들을 다 같이 패키징 해주고, 설치나 모니터링 등과 관련해서 별도의 툴들을 제공해준다. 물론 서포트나 트레이닝도 제공해주는데 그것은 유료이다. 앞서 언급한 Hadoop 1.0의 약점을 보완하기 시작한 Hadoop 버전이 0.23이다. 이것을 Hadoop 2.0이라고 부른다. 특히 맵리듀스의 경우 맵리듀스 2.0이라고 불리기도 하고 YARN (Yet Another Resource Negotiator)라고 불리기도 한다. 이는 아직 개발 중인 버전으로 아직 널리 사용되지는 않는다. 클라우데라의 경우 CDH4가 이에 해당하고 호튼웍스의 경우에는 HDP2가 이에

해당한다. CDH4의 경우 CDH3에 있는 Hadoop도 같이 포함하고 있다. 그림 1은 활용 기술에 대한 내용이다.

2. 분석 프레임워크

SAS의 SEMMA 분석 절차 방법론에 따라서 분석 프레임워크를 구성하였다. Sampling에서는 분석 데이터 생성, 통계적 추출, 조건 추출, 모델 평가를 위한 데이터 준비를 한다. Exploration에서는 분석 데이터 탐색으로 기초 통계, 그래픽 탐색과 변수 유의성 및 상관분석을 수행 한다. 또한 데이터 오류 검색, 데이터 현황을 통한 비즈니스 이상 현상과 변화 탐색을 수행한다. Modification에서는 분석 데이터 수정 및 변환으로 수량화, 표준화, 변환, 그룹화가 수행되고, 데이터 특성을 극대화 할 수 있는 방안 모색과 최적의 모델링을 위한 다양한 형태의 파생변수를 생성하고 선택한다. Modeling에서는 모델 구축, 데이터 패턴 발견, 비즈니스 문제 해결을 위한 모델 및 알고리즘을 적용한다. Assessment에서는 모델 평가 및 검증, 모델 간 비교, 추가 분석 수행 여부를 결정한다.

2.1 샘플링

H 증권회사의 데이터베이스로부터 추출한 2011년~2013년까지 데이터를 활용하였다. 그림 2는 샘플링에 대한 내용이다.

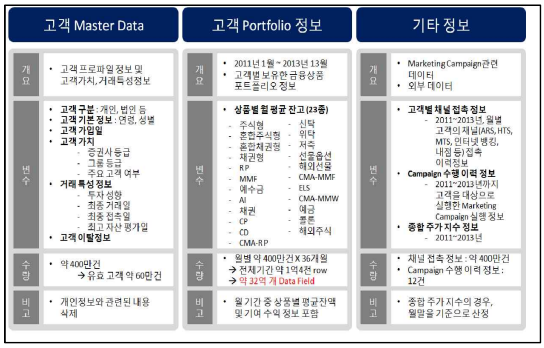

그림 2. 샘플링
Fig. 2. Sampling

고객 Master Data는 고객 프로파일 정보 및 고객가치, 거래특성정보를 기술한다. 고객 Portfolio 정보는 2011년 1월부터 2013년 13월까지 고객별 보유한 금융상품 포트폴리오 정보를 기술한다. 기타 정보에는 Marketing Campaign 관련 데이터와 외부 데이터를 기술한다.

2.2 탐색

ANOVA 분석 및 기본적인 데이터 시각화를 통해 데이터의 속성 파악하였다. 그림 3은 탐색에 대한 내용이다. 모든 변수에 대해서 서로 간의 영향도를 알기 위해 상관성 분석을 시행하였다. 예상했던 것보다 서로간의 상관성이 높은 변수들이 너무 많았다. 상관성이 높은 변수가 많다는 것은 분석적인 접근으로 보았을 때, 긍정적인 측면과 부정적인 측면 모두를 가지고 있다. 긍정적인 측면으로는 각각 변수 간에 서로 간에 어떠한 영향을 주고받는지에 대한 개략적인 정보를 얻을 수 있고, 때로는 여기서 새로운 발견을 할 수 있다. 하지만 부정적인 측면을 보면, 변수들끼리 서로 같이 움직인다는 의미로, 분석의 입력변수로써 적합하지 않다는 것으로 해석할 수도 있다. 상관성 분석을 통해 동일하게 움직이는, 혹은 전혀 반대로 움직이는 변수들의 특성을 알 수 있었고, 고객의 성향을 분명히 나누어주는 요소를 확인할 수 있었다. 예를 들면, 주식, 채권에 투자하는 성향과 선물 옵션, 증권 위탁에 투자하는 성향은 서로 간에 정 반대의 특성을 가진다. 반면에, 주식과 채권끼리는 매우 유사한 고객의 투자 성향을 가진다는 것을 알 수 있었고, 의외로 선물 옵션과 증권 위탁끼리 또한 유사한 투자 성향을 가지는 것을 알 수 있었다.

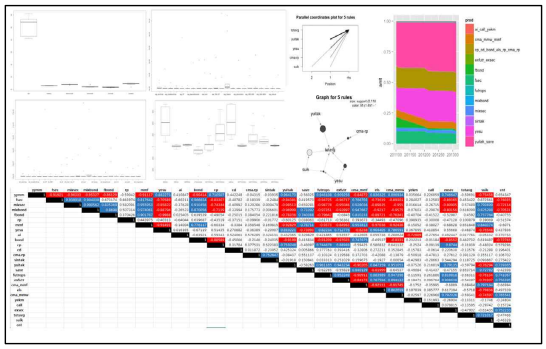

그림 3. 탐색
Fig. 3. Exploration

2.3 변형

분석할 데이터를 선택하고, 분석에 용이하도록 파생변수를 생성하였다. 그림 4는 변형의 내용이다.

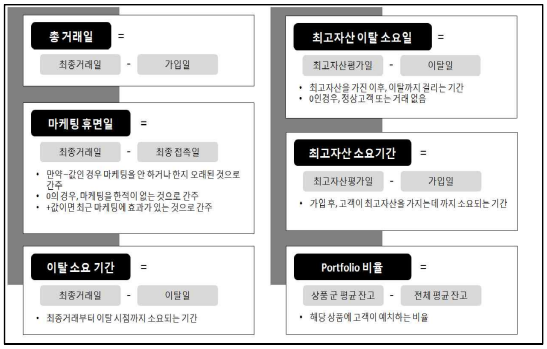

그림 4. 변형
Fig. 4. Modification

여러 종류의 파생변수를 생성하였지만, 대표적으로는 총 거래일, 마케팅 휴면일, 이탈 소요 기간, 최고자산 이탈 소요일, 최고자산 소요기간, Portfolio 비율 등과 같이, 기존에 가지고 있는 데이터의 변수들을 비즈니스 적으로 의미가 있으면서 결과 해석이 용이하도록 일정한 수식에 의해 생성하여 분석하였다. 이러한 파생변수들 또한 기초적인 분포분석 및 상관성 분석을 하였다.

2.4 모델링

기초 분석된 결과와 데이터들을 토대로 다양한 분석 모델을 적용하였다. 그림 5는 모델링의 내용이다. K-Means 알고리즘을 사용할 경우, K의 값에 따라 나눠지는 군집의 개수와 특성에 영향을 많이 미치게 된다. 따라서 K의 값을 점차 올리면서 Iteration 작업을 반복하여 K에 따른 Sum of squares 값을 확인하였다. Sum of squares가 나타내는 것은 군집끼리의 Centroid가 서로 얼마나 떨어져 있는지를 보여주는 값이라고 볼 수 있다. 군집의 개수를 늘렸을 때 Centroid가 일정 이상 가까워지거나, 혹은 갑자기 가까워지는 변곡점이 있을 경우, 해당 군집의 개수를 활용하는 것이 가장 합리적이라 판단하였다. 고객별 맞춤 상품을 추천하기 위해 Recommendation(Collaborative Filtering) 알고리즘을 사용하였다.

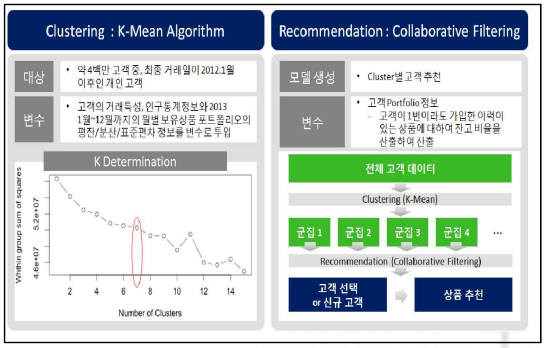

그림 5. 모델링
Fig. 5. Modeling

2.5 평가

도출된 Cluster간의 유의미하다고 판단되는 변수에 대하여 해당 Cluster와 대립되는 Cluster간의 귀무가설을 토대로 T-검정을 시행하였다. Clustering을 통해 도출된 군집들은, 그 자체로는 크게 의미가 없었다. 단순히 나눠놓은 것에 불과했기 때문에, 이를 비즈니스 적으로 재정의 하고 각각의 군집의 특성을 Profiling하는 것이 중요하였다. 결국 '핵심 고객군', '관리 사각지대 고객군', '장려 고객군'의 크게 3가지 축으로 재 정의하였고, 해당 군집의 대표적 특성이라 보이는 점들을 해당 Cluster와 다른 Cluster간의 귀무가설을 토대로 T-검정을 시행하였다. Two Sample T-Test를 통해 P-Value가 0.05미만으로 나온다면 해당 귀무가설을 채택하여 분석결과가 의미가 있었음을 검증하였다. 그림 6은 평가의 내용이다.

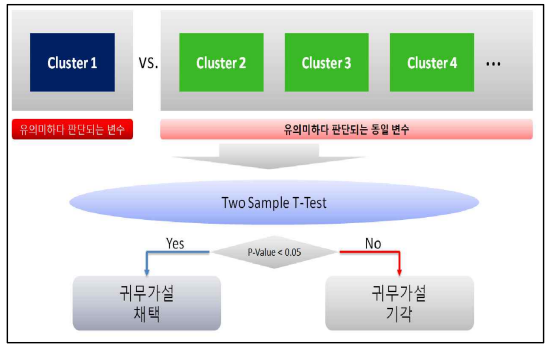

그림 6. 평가
Fig. 6. Assessment

IV. 사례 연구

1. 분석과제 개요

고객 포트폴리오, 인구통계학적 데이터, 거래특성 등의 데이터를 이용해서 고객 Clustering, Classifying분석 기법 등을 통하여 유사성 높은 흥미롭거나, 숨겨진 고객그룹을 발견한다. 그 내용은 상품판매 가능성 있는 고객 군 탐색, 핵심 관리대상 고객 군 탐색, 이탈 고객 군 패턴 분석을 통한 이탈

예측 모델 및 이탈 방지 방안 마련 등이다. 군집과 분류 분석 결과에 따라 도출된 관심 고객 군에 대하여 Contents 기반필터링 및 협업필터링 등 머신러닝 기법을 통하여 상품 Recommendation을 시행 한다.

2. 분석 방법

군집화 분석으로 객체를 잘 나타내는 특성을 정의해서 벡터화 하고, 객체들의 가까운 정도(거리측정법)를 표현할 측정치를 정의한다. 그 후 사용할 군집화 알고리즘을 적용 한다. 모든 내용은 시각화를 통하여 보여준다.

2.1 고객 군집 분석

약 4백만 고객 중 최종거래일이 2012.1월 이후인 고객을 대상으로 K-means Clustering 기법을 적용 한다. 변수는 고객의 거래특성, 인구통계정보와 2013년 1월~12월까지의 월별 보유상품 포트폴리오의 평잔, 분산, 표준편차 정보를 변수로 설정한다. 파생변수를 가공 하여 총 거래일수(최종거래일자 - 가입일자), 최종 마케팅일수(최종거래일자 - 고객직접접촉일자)로 적용한다. 그림 7은 고객 군집 분석의 공통 사항을 보여준다.

그림 7. 고객 군집 분석
Fig. 7. Customer Cluster Analysis

그림 8은 Cluster별 직접 접촉 경과일 평균과 Cluster별 최종 거래 경과일 평균을 보여준다.

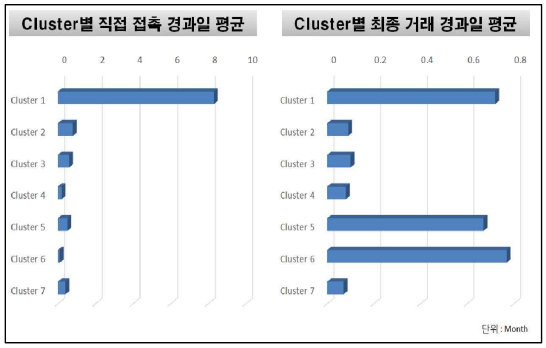

그림 8. Cluster별 평균
Fig. 8. Cluster Average

고객 군별 잔고 비중을 살펴볼 때, 2, 3번 군집이 전체의 93%를 차지하고, 수익 비중 또한 전체의 97%를 차지하므로 2, 3번 군집을 '핵심 고객 군'으로 분류된다. 그림 9는 cluster별 고객 잔고 비중과 고객 수익 비중을 보여준다.

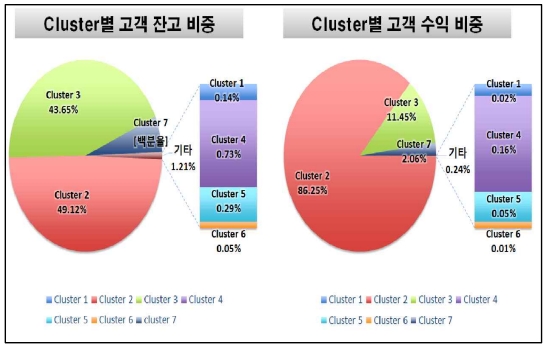

그림 9. Cluster별 비중
Fig. 9. Cluster Specific Gravity

그림 10은 핵심 고객 군을 보여준다.

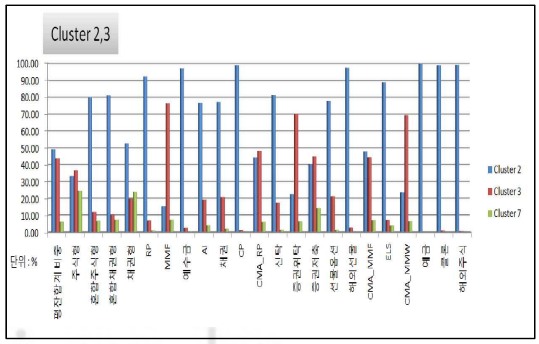

그림 10. 핵심고객군
Fig. 10. Core Customer Cluster

2번 군집의 경우, 상품 유형에 대하여 전반적으로 고르게 분산 투자를 하고 있으며 비중이 매우 높아서 전문투자 자문 서비스를 받고 있거나 특별 관리를 받는 고객일 가능성이 높다. 3번 군집의 경우, 증권 위탁에 대한 투자비율이 상대적으로 매우 높고, 안전 자산에 대한 선호도가 높아서 분산 투자를 유도할 수 있다. 그리고, 계속해서 핵심 고객의 지위를 유지할 수 있도록 VIP 관리 서비스를 발굴 및 제공해야 한다.

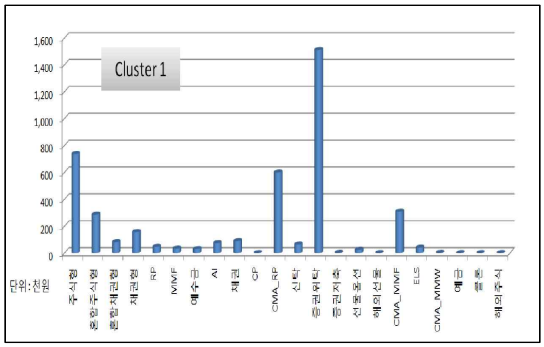

그림 11. 관리 사각지대
Fig. 11. Management Not Concern

그림 11은 관리 사각지대를 보여준다. 평균 잔고 비중이 0.14%, 수익 비중 0.02%로 미비하고, 평균 거래기간이 17년 이상 된 고객이면서, 최종 거래가 7개월 정도로 아직 유효한 고객 군으로 주식형 및 증권위탁에 대한 비중이 높아서 다소 공격적인 투자 성향을 가진 것으로 판단되어 주식 수익률 대회 등의 이벤트 마케팅으로 당사의 HTS나 MTS등의 채널로 유도할 수 있는 마케팅이 요구된다.

그림 12는 장려 고객 군을 보여준다. 고객 수 비중이 약 5%대로, 각종 고객 등급은 중간 이상의 일반 고객이지만 대부분이 관리대상으로 등록되어있는 우수 고객 군이다.

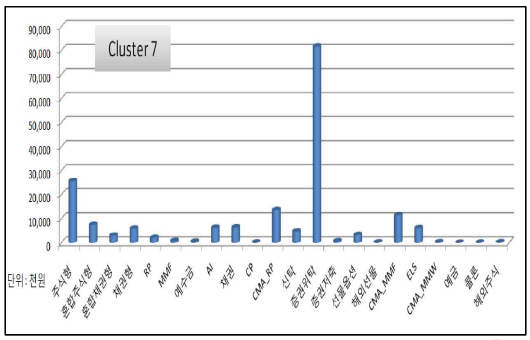

그림 12. 장려 고객군
Fig. 12. Promotion Customer Cluster

평균 거래 기간이 15년으로 길고, 최종 거래 경과일수가 1달 이내여서 대체로 거래가 활발한 고객 군이다. 전체 잔고 비중이 6% 정도로, 핵심 고객 군 이외에 가장 견실한 고객 군이다. 안정 또는 위험 중립형의 고객으로 종합자산관리 마케팅 등을 통해 포트폴리오 다양화를 유도할 수 있는 마케팅 전략이 필요하다.

2.2 금융 상품 추천 분석

Mahout으로 실행하기 위한 설정 값과 계수는 그림 13과 같다. 유사 집단을 50으로 하고, useriD가 10000039인 고객에게 5개의 상품을 추천한다.

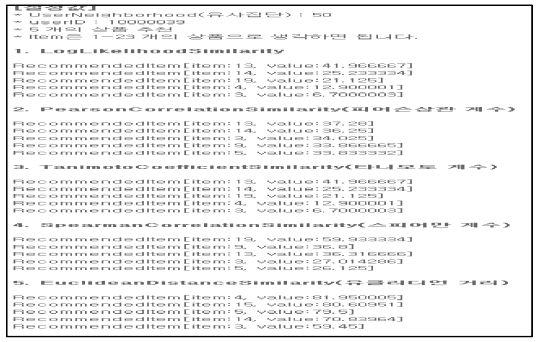

그림 13. 설정 값과 계수
Fig. 13. Setting Value and Coefficient

그림 14와 같이 고객 portfolio별 상품추천을 위해서 원하는 항목의 내용을 입력한다.

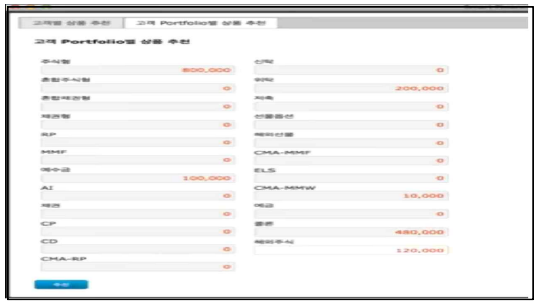

그림 14. 고객 상품 추천 입력 항목
Fig. 14. Recommendation Input Item

동일 고객 군 portfolio 유형과 고객번호 10000039님의 portfolio 내용이 그림 15처럼 나타난다.

그림 15. 고객 상품 추천 내용
Fig. 15. Customer Item Recommendation

이 고객은 관리 사각지대에 속하는 고객이다. Log Likelihood 기반 추천, Pearson 기반 추천, Tanimoto 기반 추천, Spearn 기반 추천, Euclidean Distance 기반 추천으로 해당 고객에게 상품을 추천해 준다.

본 연구에서는 빅 데이터를 이용하여 증권사의 고객의 특성을 파악하고 고객 군을 세분화하여 고객 군별 Profile을 바탕으로 최적의 마케팅 방향을 설정할 수 있었다. 예상하기로는 경제활동이 활발한 30대 후반에서 40대의 높은 분포를 가질 것으로 생각되었으나, 분석결과, 50대 중 후반의 고객이 가장 많은 것을 확인 할 수 있었다.

거래특성별 Clustering을 통해 도출된 7개의 군집을 비즈니스 적으로 이해할 수 있도록 다시 4개의 군집으로 분류하였다. 0.25%의 소수 고객이 증권사 전체 수익 비중의 대부분을 차지하기 때문에, 해당 고객들로 인해 기업의 영업이익률이 좌지우지될 수 있다. 고객 군별 Profiling 결과, 해당 증권사는 일반 소규모 고객보다는 자산이 많은 우수한 고객에 더욱 중점을 두는 영업을 하고 있는 것을 확인 할 수 있었다. 고객별 상품 추천의 경우, 증권사에 손님이 방문하여서 고객을 인식하는 순간부터 직원으로 하여금 적절한 상품을 제시하는 영업 전략을 사용할 수 있다. 뿐만 아니라, HTS, MTS나, 인터넷을 통하여 접속한 고객을 목표로 하여 추천된 상품을 노출시키고 광고함으로써 효과를 극대화 할 수 있다.

V. 결 론

본 연구에서는 활용 기술로 데이터 저장 레이어, 데이터 처리 레이어, 데이터 분석 레이어, 시각화 레이어가 사용되었다. 각 단계에서 저장, 처리, 분석된 데이터는 시각화를 통하여 볼 수 있게 하였다. Hadoop을 통하여 데이터를 처리한 후 처리된 데이터를 Mahout으로 실행하여 분석 결과를 시각화하였다. 이 과정을 통해서 금융 상품에 가입된 고객의 여러 특성을 파악하였고, 각 고객에 따른 금융 상품의 추천을 적시에 수행할 수 있었다.

본 연구에서 제안된 개발 방법은 빅 데이터의 활용에 대한 실제적인 내용을 보여줌으로서 빅 데이터의 개발 사례에 좋은 본 보기가 될 것이다. 이 내용은 금융, 제조, 공공, 서비스 등 여러 분야에 빅 데이터를 활용한 고객 관리를 위하여 활용될 수 있다. 이를 위해서 몇 가지 시사점을 제시한다. 먼저, 도메인 지식의 적극적인 활용 및 분석이 필요하다. 두 번째는 빅 데이터로 하여금 업무 프로세스의 개선점을 포착하는데 활용 할 수 있다. 더불어 이러한 빅 데이터 전쟁에서 승리하기 위해서는 먼저, 내부 엔지니어링 조직을 갖추는 것이 중요하다. 이는 많은 기업들이 IT관련 부분을 비용적인 측면에서 모두 아웃소싱하고 있는 실정이지만 빅 데이터는 한 번에 프로젝트가 마무리 될 수 있는 것이 아니기 때문에 내부에 엔지니어링 조직을 갖추는 것이 중요하다. 이는 조직내부의 IT 거버넌스와 밀접한 관련이 있기 때문에 CTO나 CIO 수준에서 지속적으로 빅 데이터를 구축하고 활용 하여야 한다. 빅 데이터는 단순히 많은 데이터를 분석하는 것뿐만 아니라 나아가 조직내부에서 빅 데이터로부터의 시사점을 수용하고 내재화할 수 있는 적응 능력이 있어야 한다. 이를 위해서는 빅 데이터 아키텍처가 구성된 빅 데이터 거버넌스가 앞으로 연구되어야 한다.

참고문헌

[1] Seong-Hee Lee, "Understand of Big Data - Value and Induction Strategy," Journal of Korea Information and Technology, Vol. 10, No. 1, pp.63-68, July, 2012.

[2] Sunil Soares, "Big Data Governance An Emerging Imperative," MC Press, pp.9-28, 2012.

[3] Special online collection: Dealing with Big Data. http://www.sciencemag.org/site/special/data/, 2011.

[4] Big Data. http://www.nature.com/news/specials/bigdata/index.html, 2008.

[5] Michael Minelli, Michele Chambers, Ambiga Dhiraj, "Big Data , Big Analytics," John Wiley & Sons, Inc. , pp.4-15, 2013.

〔6〕 Min Chen, Shiwen Mao, Yin Zhang, Victor C.M. Leung, "Big Data related Technologies, Challenges and Future Prospects," Springer, 2014.
〔7〕 Sean Owen, Robin Anil, Ted Dunning, Ellen Friedman, Mahout in Action," Manning Publications, 2012.
〔8〕 Kee-Yong Han, "Do it Hadoop with Big Data," Easy Publications, 2013.
〔9〕 KoDB, "The Guide for Advanced Data Analytics Professional 2014 Edition," Korea Database Agency, 2014.
〔10〕 M Grobelnik. Big Data tutorial. http://videolectures.net/eswc2012grobelnikbigdata/, 2012.
〔11〕 Seok-soo Kim, Hwa-sil Lee, "A Model of implementation Data Architecture for Enterprise Architecture," Journal of Korea Society of Computer and Information, Vol. 16, No. 9, pp.175-183, Aug. 2011.
〔12〕 Seok-soo Kim, "The Analysis of Data Governance model for Business and IT Alignment," Journal of Korea Society of Computer and Information, Vol. 18, No. 7, pp.69-78, July, 2013.

저 자 소 개

김 석 수
1982: 숭실대학교 전자계산학과 공학사
1987: 숭실대학교 전자계산학과 공학석사
1998: 숭실대학교 전자계산학과 공학박사
1989~현재: 가천대학교 컴퓨터공학과 교수
관심분야: DB, EA, 정보공학, DG, Big data
E-mail: sskim@gachon.ac.kr