

Big Data, Bigger Opportunities And The Biggest Value!

2013 우수 프로젝트 사례집



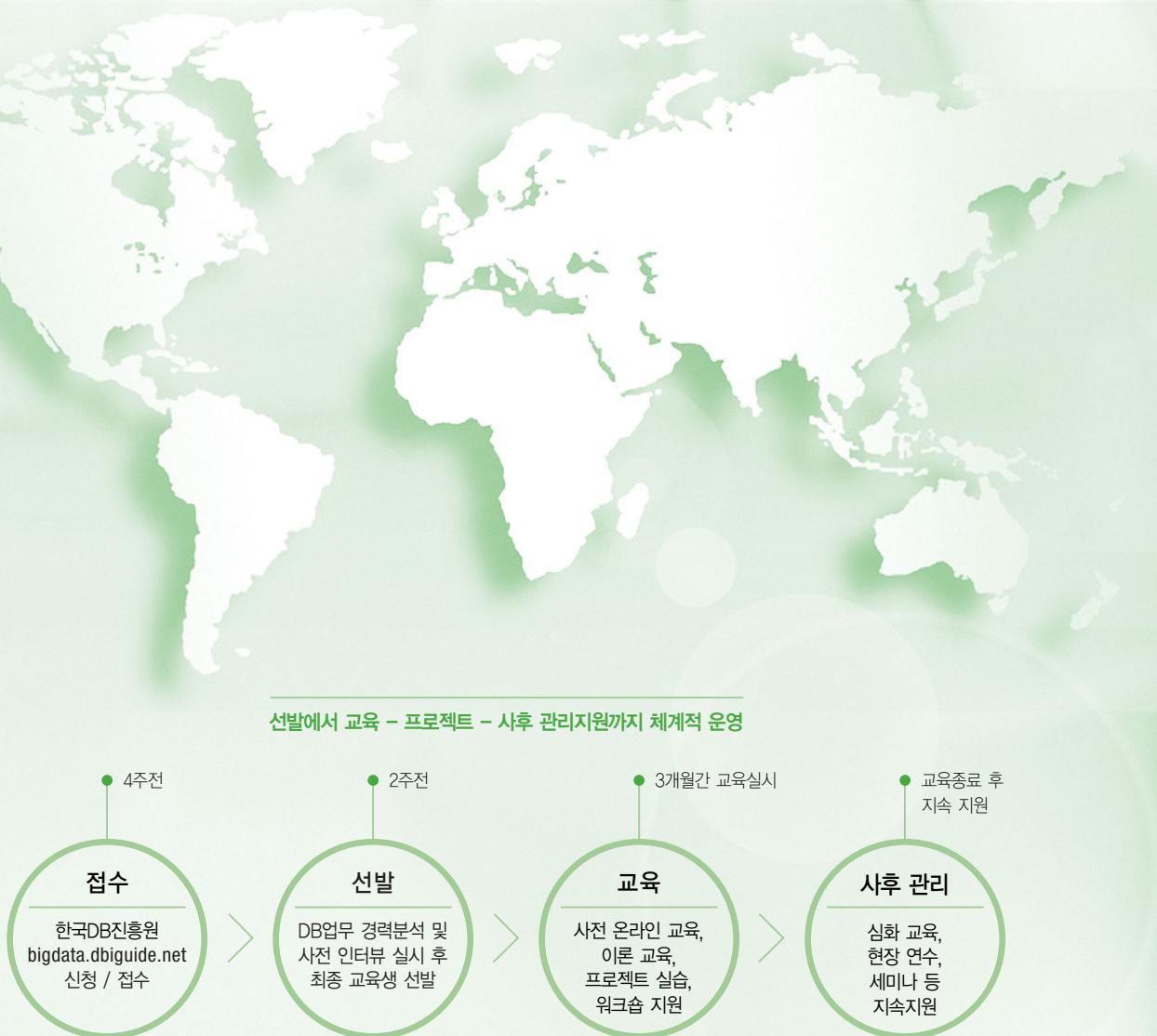
미래창조과학부
427-140 경기도 과천시 관문로 47, 4동
www.msip.go.kr

KODB 한국 데이터베이스 진흥원
KOREA DATABASE AGENCY
110-799 서울시 종로구 종로 51 종로타워 19층
TEL : 02-3708-5303
www.kodb.or.kr

*이 사례집은 미래창조과학부의 지원으로 제작되었습니다.

빅데이터 아카데미는

공공, 산업 등 사회 전반에서 활동중인 데이터 전문가들을 빅데이터 전문가로
전환 · 양성하여 글로벌 빅데이터 시장 선점에 기여하기 위한
빅데이터 전문가 양성 프로그램입니다.



2014년 빅데이터 아카데미 교육 안내

과정	내용	
사전 교육	(공통 역량) 빅데이터 전문가에 대한 요구역량 이해 빅데이터 환경 구축 방법(Hadoop)	빅데이터 분석 도구 사용법(R)
이론 교육		
	• (공통 역량) 빅데이터 이해 – 데이터 사이언스와 인문학적 통찰력 • (공통 역량) 빅데이터 분석 전략 – 분석 과제 정의, 마스터 플랜	
	• 빅데이터 아키텍처 • 하둡 분산파일 시스템 운영 • HDFS의 클러스터 운영 및 관리 • Map Reduce의 구조와 성능 • 분산 수집 시스템(FLUME) • 쿼리 분석 엔진(Hive/Pig) • 분석 시스템 용량/비용 계획 • 보안 관리	• 빅데이터 분석 및 활용 • 빅데이터 분석 환경(R) 사용법 • 시각화를 이용한 Insight 도출 • 통계 분석 • 데이터 마트 구축 • 데이터 마이닝 • Text Mining/Social Network Analysis • 시각화 디자인 및 구현
프로젝트 실습	• 추천(Micro Targeting) 시스템 구현 • 자사 빅데이터 프로젝트 POC	• 예측 모형(Prediction Model) 개발 • 자사 빅데이터 프로젝트 POC
워크숍	• 실시간 처리 기술(SQL on Hadoop) • 빅데이터 저장 기술(NoSQL)	• 빅데이터 분석 시각화(D3.js) • 빅데이터 고급 분석(Classification 등)
일정	5기 02월 17일 ~ 05월 09일	5기 03월 10일 ~ 05월 30일
	6기 03월 31일 ~ 06월 27일	6기 05월 05일 ~ 07월 25일
	7기 06월 02일 ~ 08월 08일	7기 06월 23일 ~ 09월 05일
	8기 08월 11일 ~ 10월 24일	8기 09월 08일 ~ 11월 21일

*2013년 실시한 빅데이터 직무분석 결과를 반영해 빅데이터 기획, 처리, 분석, 시각화, 운영관리를 포함한 빅데이터 아카데미 교육과정 개편 및 교육내용 업그레이드

Big Data, Bigger Opportunities And The Biggest Value!

2013 우수 프로젝트 사례집





Big Data, Bigger Opportunities and The Biggest Value!

미래창조과학부와 한국데이터베이스진흥원은 창조경제 실현의 핵심동력이 될 빅데이터 전문가를 양성하기 위해 2013년 6월 ‘빅데이터 아카데미’를 국내 최초로 개소하였습니다.

지난 한 해 ‘빅데이터 아카데미’는 ‘빅데이터 처리 기술 전문가’와 ‘빅데이터 분석 전문가’ 과정을 운영하여 금융·의료·제조·게임 등 다양한 산업분야에 종사중인 202명의 인력을 빅데이터 전문가로 육성했습니다.

빅데이터 아카데미를 수료한 연수생들이 연수기간 중 실시한 다양한 빅데이터 프로젝트 경험을 토대로 현업에 복귀하여 공공 및 민간 66개 기관에서 실시하는 빅데이터 프로젝트 70여 건에 참여하는 등 산업체 전반에서 큰 성과를 낸 것으로 나타났습니다.

이에 교육 연수 중 실시한 우수 프로젝트 사례를 자료로 배포하여 기술적·분석적 노하우를 공유하고, 신규 빅데이터 비즈니스가 발굴될 수 있도록 지원하기 위해 ‘빅데이터 아카데미 우수 프로젝트 사례집’을 발간하게 되었습니다.

본 사례집은 빅데이터 아카데미 교육 연수 프로그램의 파일럿 프로젝트들 중 우수사례로 선정된 4개에 대해 과제 발굴 단계부터 개발 과정과 프로젝트 수행 시 경험했던 문제점 진단까지 프로젝트 전반을 소개하고 있습니다.

모쪼록 본 사례집이 빅데이터 프로젝트를 기획하고 추진하는 기업과 전문가분들께 방향을 제시하고, 다양한 비즈니스 발굴과 아이디어 확보 도구로 활용되기를 기원합니다.

감사합니다.

2014년 4월
한국데이터베이스진흥원장

A handwritten signature in black ink, appearing to read '한국데이터베이스진흥원장' (President of the Korea Big Data Institute).

2013 우수 프로젝트 사례집

Contents

원장 인사말	04
분석 전문가 과정	
영화흥행 예측 분석 : '내일을 향해 쌔라'	06
상장폐지기업 예측 분석 : 데이터가 우리에게 말을 걸어오다	14
기술 전문가 과정	
쇼핑몰 상품 트렌드 분석 플랫폼 : 국내 최초로 쇼핑몰 실시간 분석에 도전하다	20
키워드 기반 트렌드 분석 플랫폼 : 데이터 분석 플랫폼의 공든 탑을 쌓다	28
한눈에 보는 빅데이터 아카데미	34
빅데이터 아카데미 설립배경	36
빅데이터 아카데미 교육안내	37



‘내일을 향해 쏴라’ 영화 흥행예측 분석



‘백 번 듣는 것보다 한 번 해보는 것이 낫다!’

분석 전문가 과정 집체교육을 받고 빅데이터 분석이 무엇인지를 알게 될 찰나, 수료 프로젝트가 기다리고 있었다. 공교롭게도 대전에서 참여한 수강생 5명이 한 팀이 되어 ‘영화 흥행예측’ 프로젝트를 수행했다. 비정형 데이터 분석을 통해 지금까지 어느 평론가나 영화전문가도 달성하지 못했던 높은 적중률에 도전한다. 대전영화팀의 당돌한 도전 결과는 어떻게 됐을까?

Challenges

영화흥행 예측을 프로젝트 과제로 정하다

빅데이터 아카데미 분석 전문가 1기 수강생 가운데 대전에서 올라온 사람이 5명이라는 것을 수료 프로젝트팀 구성할 때 알게 됐다. 집체교육 기간에 배운 것을 토대로 5명의 팀원이 모여 진행할 수료 프로젝트 과제를 놓고 고민했다.

팀원 모두 야구를 좋아한다는 공통점을 발견하고 ‘미국 메이저리그 야구팀의 성적과 구단의 나스닥 지수 연관성’을 분석 과제로 선정했다. 하지만 나스닥 어디에도 ‘양키스’나 ‘다저스’라는 단어를 찾을 수 없었다.

'실수를 통해 배운다'고 우리는 분석 과제 선정 시 무엇을 놓쳤는지를 생각해보았다. 주제도 흥미로워야 하지만, 비교·판단할 데이터가 필수적임을 알게 됐다. 요컨대 비교할 정형화된 데이터가 있어야 한다. 이로써 고민의 대상이 좁혀졌고 분명한 목적을 정해야 힘을 알게 됐다. 목적에는 명확한 대상이 포함되는데, 빅데이터를 활용한 분석이라는 목적에도 부합해야 했다.

이로써 흐릿했던 뭔가가 점점 윤곽을 드러내기 시작했다. '비교 대상이 될 만한 기초 데이터가 있으면서 소셜 네트워크 서비스(SNS)에서 많이 오가는, 사람들이 흥미로워할 주제가 무엇인가?'로 좁혀 생각하다 보니 영화가 어떻겠느냐는 의견이 나왔고 기초 조사를 통해 최종 주제로 선정했다.

Solution

정형 데이터로 영화진흥위원회의 데이터를 선택하다

데이터는 성격에 따라 영화의 상세 정보를 담고 있는 정형 데이터와 영화의 평점, 트위트, 각종 기사와 댓글 등을 포함한 비정형 데

이터로 양분했다. 앞서 분석할 데이터 수집이 가능한지부터 확인하고 주제를 잡았다

미국 영화로 좁혀본다면 boxofficemojo.com에서 충분한 정형 데이터를 얻을 수 있을 거 같았다. 한국 데이터도 역시 있었다. 한국 영화데이터베이스(KMDB)와 영화진흥위원회(KOFIC, www.kobis.or.kr/kobis/business/stat/theme/findAreaShareList.do)에도 원하는 정형 데이터가 있었다. 우리는 영화진흥위원회의 데이터를 사용하기로 했다. 엑셀 포맷으로 내려 받은 데이터를 다듬는 과정이 필요했다. 더불어 흥행 예측분석을 위해 필요로 했던 장르나 주연 배우 등 몇 가지 누락된 정보는 수작업으로 직접 수집했다. 이 과정을 통해 빅데이터 분석은 생각만큼 그렇게 환상적이지 않고 노력과 의지, 끈기가 중요함을 실감했다.

수집한 자료를 RDB에 넣고 데이터 무결성과 정합성을 검증해 보니 매우 빈약한 수준이었다. "분석 프로젝트에 들어가면 고객이 자료를 잘 준비해 놓았더라도 자료 정제에 프로젝트 기간의 반 이상을 소모한다"는 지도 교수님의 얘기가 바로 떠올랐다. 데이터 정리를 마치고 분석에 활용할 추가 변수를 생성해야 했다.

그림 1. 영화 흥행예측 분석팀의 빅데이터 분석 과정



배우, 감독, 배급사에 별도의 등급을 부여하여 변수로 활용하기 위해서다. 등급 부여는 예측을 하기 위해 매우 중요한 요소지만, 간단하게 처리하기로 했다. 더 정교한 모델은 실무 프로젝트에서 구현해야 하는 수준이므로 말이다. 작품당 평균 매출액을 기준으로 전체에서 몇 %에 속하는지에 따라 9개 등급으로 나눴다. 등급의 비율과 단계는 수능 등급과 동일하게 가져갔다.

비정형 데이터로 포털 서비스의 영화평을 수집하다

이제부터 본격적으로 빅데이터 분석을 위한 차례다. 책이나 세미나 자료는 중간과정이 압축돼 있기에 대전팀은 그 압축된 간격을 현실 세계에 맞게 풀어내야 했다.

집체교육 기간 중에 배운 것처럼 우선 오픈소스 분석도구인 R을 설치해 트위터에서 감상평 등 비정형 데이터를 수집했다. R에서 최근 트위터에 올라온 글을 검색하고 필요한 내용을 추출했다. 대부분의 트윗 메시지가 광고 데이터임을 확인했다. 더구나 트위터는 최근 2주 간의 데이터만 제공하기에 과거 데이터를 수집할 수 없다. 안타깝지만 이 때문에 트위터에서 데이터 수집을 포기했다. 대안으로 페이스북과 구글에서 영화관련 기사나 내용을 검색해 데이터를 수집하는 것도 생각해봤지만, 영화에 대한 개인의 감성이 잘 녹아 있고 데이터 양도 많은 포털이나 영화사 사이트의 감상평을

수집하는 것이 더 효율적이라고 판단했다.

수집해야 할 감상평은 많은데, 수작업으로 수집하면 프로젝트 기간 내내 감상평 수집만 하다 끝날 수 있었다. 상용 크롤링 프로그램을 구입해 감상평 데이터를 수집했는데, 수집은 잘 됐으나 속도가 느렸다. R에서도 웹사이트를 크롤링할 수 있음을 알고 영화관련 사이트에 들어가 감상평을 수집했다. 주말을 포함해 일주일 동안 '다음'과 '네이트'의 감상평을 수집했는데, 네이트에서만 2008년 이후의 영화에 대한 자료 84만여 건을 수집했다.

'시작이 반드시'

각각 떨어져 일하던 우리는 종간 점검을 위해 모였다. '시작이 반드시'라고 처음에는 우리가 과연 이 프로젝트를 완료할 수 있을까? 하는 걱정이 없지 않았는데 일단 부딪히고 볼 일이었다. 기초 데이터 정리가 절반이라고 했으니 프로젝트가 완료에 가까워졌다고 자신을 위로하며 다시 힘을 내기로 했다.

KONLP 패키지로 한글 자연어 처리를 하다

요건 정의와 데이터 가공까지 끝냈으니, 이제 이번 프로젝트의 하이라이트인 모델링만 남았다. 분석교육 때 실습도 해봤고, 그때 받은 감성분석사전도 있으니 어렵지 않게 진행될 거라고 생각했다.

그림 2. R용 크롤링 패키지를 이용해 '다음' 영화 감상평을 수집하는 함수

```
Source on Save | Run | Source
1  install.packages("XML")
2  install.packages("xtable")
3  install.packages("tm")
4  library(XML)
5  library(xtable)
6  library(tm)
7
8
9
10 #다음 영화 페이지에서 감상평 내용 수집하는 모듈
11 GetDaumMovieData <- function(num){
12   url = gsub(" ", "", paste("http://movie.daum.net/review/netizen_point/movieNetizenPoint.do?type=after&page=", num))
13   doc = htmlTreeParse(url, useInternalNodes = T, encoding="UTF-8")
14   xpathSApply(doc, "//div[@class='commentList']", xmlValue)
15   movie_nm <- xpathSApply(doc, "//span[@class='movieTitle fs11']", xmlValue) #영화명
16   score<-gsub("내 티즌별점","",xpathSApply(doc, "//span[@class='star_small1']", xmlValue)) #별점
17   reg_date<-xpathSApply(doc, "//span[@class='date']", xmlValue) #작성일
18   contents<-gsub("내 티즌별점|\r|\t|\n","",xpathSApply(doc, "//span[@class='comment_article']", xmlValue)) #내용
19   contents<-removeNumbers(contents) #숫자제거
20   contents<-removePunctuation(contents) #특수문자제거
21   contents<-stripWhitespace(contents) #공백제거
22   senti_data <-cbind(movie_nm,score,reg_date,contents) #데이터 바인드
23   return(senti_data)
24 }
25
```

감성 분석 시나리오

- 1단계: 감상평 데이터에서 명사 분리
- 2단계: 긍부정 단어 카운팅
- 3단계: 감상평 스코어링
- 4단계: 분석용 데이터셋트 작성
- 5단계: 알고리즘 트레이닝
- 6단계: 흥행 예측

먼저 어떻게 모델링할 것인지 시나리오를 만들어야 했다.

R의 extractNoun() 함수를 이용해 감상평 데이터에서 명사 분리를 시작하였다. 하지만 출발과 동시에 에러가 '터지기' 시작했다.

Warning message:

In preprocessing(sentence) :

It's not kind of right sentence :

'자연의경고가좀비라는존재로대체된것같네요!'

'글자 수가 0이거나 20자를 넘으면서 공백으로 분리된 단어의 수가 1보다 작거나 같으면' 이라는 경고 문구와 함께 종료되도록 지정돼 있어서 발생한 문제였다. 수집한 감상평 데이터에는 띄어쓰기 없이 20자가 넘는 게 많았다. 이에 이런 데이터가 들어오면 자동으로 일정한 크기로 잘라주는 모듈 하나를 추가했다.

이론과 현실의 차이로 시행착오를 겪다

두 번째 시나리오는 김경태 교수가 제공한 '한글감성분석사전'으로 수집한 감상평 데이터를 매치하는 것이다.

본격적인 코드 작성 전, 샘플링한 데이터에서 명사를 분리하고, 사전과 매칭하는 테스트를 해보았다. 여기서 또 다른 문제에 부딪혔다.

분석 대상 문장인 "정말 봐야됨 이런배우들이 떠야되는데ㅋ잼나게 봤어요"에서 감성을 나타내는 단어는 "봐야됨", "잼나게"다. 이때 extractNoun() 함수의 결과값은 "이런배우들이", "떠야되는덱잼나게"로 나왔다.

분석할 대상 데이터는 다양한 연령층, 다양한 교육수준을 가진 사람

들이 작성한 것이고, 인터넷에서 작성된 글이어서 띄어쓰기 등 문법이 무시되고, 은어 · 비속어 · 외국어 · 줄임말 등 변종어가 많다. 하지만 제공받은 감성분석사전에는 변종단어들이 포함돼 있지 않았다. 이 문제 때문에 우리는 책에 나와있는 '이론상 데이터'와 실무에서 다를 '현실 데이터'는 태생부터 달라서 분석 시 많은 시행착오를 겪을 수밖에 없음을 알게 되었다.

카운트된 긍부정 단어 수로 감상평을 스코어링하다

세 번째 시나리오는 감상평에 대해 분석을 위한 스코어를 부여하는 것이다. 이 스코어를 분석해 개봉 예정 영화의 흥행을 예측할 것이기 때문에 굉장히 중요한 과정 중 하나였다.

교육과정에서 받은 긍부정사전으로 감상평 데이터 스코어링만 하면 거의 끝날 것 같았다. 실제 스코어링하기 전에 제공받은 사전이 영화 감상평 데이터의 감성을 얼마나 정확히 맞추는지 확인했다. 뒤 페이지의 <그림 3>은 긍정 감상평 1000개, 부정 감상평 1000개를 ctree 알고리즘으로 분류한 결과다. 노드별로 pos(긍정) 또는 neg(부정) 쪽으로 편중돼 있어야 제대로 긍/부정이 분류된 것이다. 결과는 조건별로 긍/부정이 고루 분포된 모형이 나왔다.

앞에서 만든 분류 규칙을 토대로 감상평의 긍정/부정을 분류한 결과 63.25%의 정확도를 보였다. 한 가지로만 찍어도 확률이 50%인데 감성 분석의 정확도가 63%라면 분석의 의미가 없었다. 최소 70% 이상의 정확도는 나와야 했다. 정확도를 높이려면 어떻게 해야 할까?

감성분석의 정확도를 높여라

분야별로 쓰는 단어들이 다르고 단어의 의미도 달라질 수 있는데, 범용 사전으로 분석했기 때문에 정확도가 떨어졌다고 판단했다. 빅데이터 아카데미의 성공적인 수료를 위해 무조건 진격하기로 했다. 영화 감성 분석용 긍부정사전을 별도로 만들기로 한 것이다.

직접 만든 영화 감성 분석용 사전을 테스트해야 하는 순간 떨리는 가슴으로 지켜보았다. 이전과 동일한 방법으로 감성사전 테스트를 진행해 다음과 같은 결과를 얻었다.

<그림 4>의 분류 분포를 보면, 범용 감성사전과 달리 분포들이 위쪽이든 아래쪽이든 한쪽으로 몰려 있는 것을 알 수 있다. 이것은

해당 조건에서 감성 분석의 정확도가 높다는 것을 의미한다. 영화 감성분석 사전을 테스트한 결과 80.25%의 정확도가 나왔다. 이런 문제를 극복하고 분석의 정확도를 높이기 위해서는 반드시 해당 산업분야의 전용 사전을 구축하여 업무에 적용할 것을 추천한다.

정형 데이터와 맵핑해 통합 데이터세트를 작성하다

우리들이 만든 영화전용 긍부정 사전과 [senti_Score(), Cnt_Word()] 함수로 감상평 데이터의 긍/부정을 판단할 수 있었고, 이를 이용해 감성이 포함된 감상평 데이터 세트를 구축했다. 이제 할 일은 정형 데이터와 감상평 데이터를 통합해, 분석용 통합 데이터 세트를 만들기だ.

영화진흥위원회 데이터만을 이용해 Classification했다. 결과는 에러율을 31.97%로 약 68%의 정확도를 보였다. 감성 데이터만 이용해

Classification한 결과, 에러율은 45.26%로 정형 데이터의 경우보다 높았다.

영화 흥행에 영향을 주는 핵심 변수가 무엇인지와 감성 데이터가 흥행에 영향을 주는지 검증하기 위해 정형 데이터와 포털 사이트의 영화 감성 데이터(비정형 데이터)를 통합해 Classification을 했다. Variable importance 결과를 통해 영화 흥행에 가장 큰 영향을 미치는 변수는 감독 > 배우 > 감성 데이터 > 배급사 순인 것으로 확인됐다. 이제 영화 흥행 예측을 위해 아직 종영이 안된 2013년 6월 데이터를 이용해 흥행을 예측해보고, 모델을 검증할 차례다. 이전과 같은 방법으로 6월 개봉 영화의 정형 데이터와 감성 데이터를 수집·정제·통합해 분석용 데이터 세트를 만들고, 앞서 만들어진 모델에 넣어 predict를 진행했다. 결과는 다음과 같다. 정확도가 다소 떨어져서 약간 실망했지만, 우리는 놀라운 사실을 알 수 있었다.

분석 모델의 트레이닝 과정에서는 감성 데이터의 영향도가 미미했지만, 새로운 데이터의 예측에서는 영향을 크게 준다는 사실을 확인할 수 있었다. 더 눈길을 끌었던 점은 실제 예측에서 사용된 모델인 '영화진흥원 데이터 + D-15 감성 데이터'로 예측한 결과가 42%로 가장 높게 나온 것이다.

예측 결과가 의외였다. 분석 모델 개발 과정에서는 감성 데이터가 지금같이 크지는 않았는데, 아마도 6월에 개봉된 영화가 대부분 A class에 속해 있어서 그런 것 같았다. 한 집단에 집중 분포된 데이터를 분류하다 보니 에러율이 올라갔다. 더 많은 데이터를 수집하고, 학습을 시키면 예측의 정확도도 올라갈 수 있다는 의미였다.

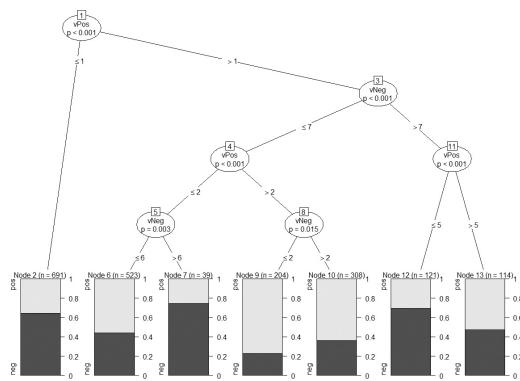


그림 3. 감상평을 ctree 알고리즘으로 분류한 결과 트리

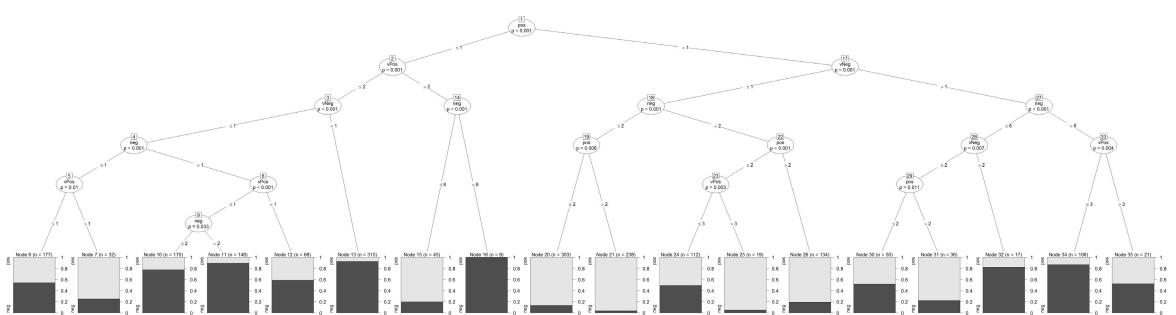


그림 4. 영화 감성 분석용 긍부정사전으로 분석한 결과

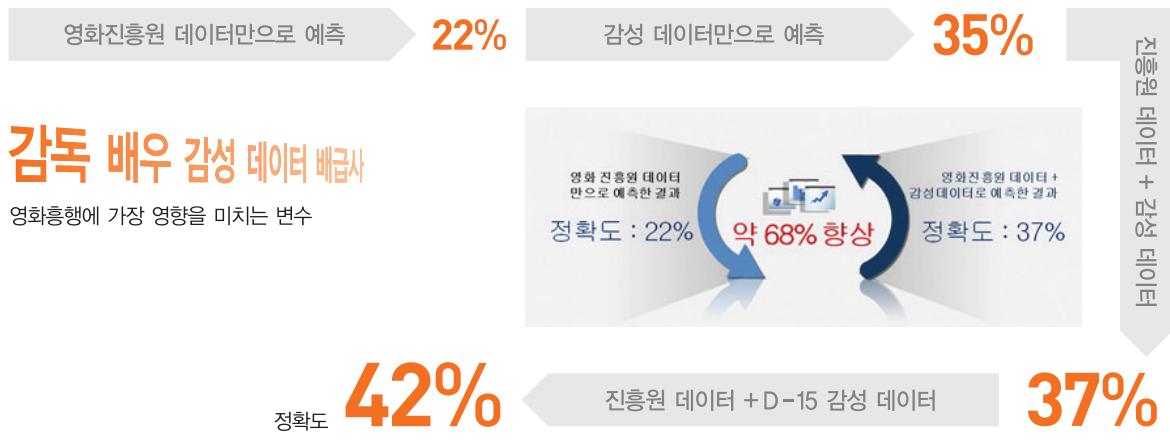


그림 5. 영화 흥행에 영향을 미치는 변수와 예측 정확도의 변화

영화 흥행예측 결과가 나오다

개발한 영화 흥행 예측 모델을 활용해 개봉예정인 영화 중에서 관객수를 예측하였다. 먼저 그 결과부터 확인해보자.

표. 빅데이터 감성분석을 반영한 예측과 실제 관객 수

영화명	예측	관객수(명)	적중여부
권법 쿵푸의 신	A (3만 이하)	1,652	적중
명탐정 코난	D (30만~100만)	389,873	적중
수평선상의 음모	E (100만 초과)	978,808	미적중

출처 : 영화진흥위원회 공식(연감)

통계 기준일 : 2013.09.09(권법 쿵푸의 신)/2013.10.06(명탐정 코난: 수평선상의 음모, 에픽: 숨속의 전설)

'권법 쿵푸의 신'과 '명탐정 코난 수평선상의 음모'는 정확히 예측 범위 안에 들어왔다. 100만 이상의 관객을 확보하면서 소위 대박을 칠 것이라 예측한 '에픽 숨속의 전설'은 아깝게 예측 범위를 벗어나고 말았다.

Conclusion

부족한 시간, 환경, 경험 속에서도 발생되는 문제점을 회피하지 않고 당당히 맞서 빅데이터 분석 프로젝트를 수행했고, 나름대로 성

공적인 결과를 얻었다.

대전영화팀이 진행했던 프로젝트를 요약해 보면 다음과 같다. 처음 계획대로 분석절차에 따라 진행하였으며, 각각의 에피소드를 진행하면서 발생한 문제들은 시행착오를 통해 해결했다. 감성분석을 통해 감상평을 스코어링했고, 이를 통해 개봉할 영화의 관객을 예측했다. 모두 적중하지 못했지만, 흥행 결과도 예측한 패턴대로 나왔다. 영화평에 대한 긍부정사전을 작성한 것은 매우 험난한 과정이었지만, 좀 더 보완해 나간다면 개봉 예정작에 대한 예측은 어떤 평론이나 영화전문가의 예측보다 높게 나올 것이라고 확신한다.

빅데이터라는 말을 처음 들었을 때, 데이터(정형 데이터)가 아닌 새로운 데이터(비정형 데이터)를 처리할 수 있다는 생각에 가슴이 뛰었다. 이후 기술적으로 빅데이터에 대한 정의가 많이 나왔지만 본질이 없는 특성에 대한 정의일 뿐이라는 생각을 많이 했다. 비정형 데이터도 결국엔 2진 데이터로 변환돼 저장되니까 말이다. 지금은 데이터를 처리·저장하는 것에 기술의 초점을 맞추고 있지만, 시간이 지나면 이 기술은 보편화될 것이고, 결국에 빅데이터는 데이터와 같은 말이 될 것이다. 물론 새로운 패러다임이 등장하면 빅데이터라는 용어를 또 사용할 수도 있다. 더불어 데이터가 있는 한 이를 분석하는 것은 언제나 그렇듯이 미래를 통찰하는 데 꼭 필요할 것이다. **BIG**



정근호 팀장(세림티에스지 기술연구소장)

“내 안에 숨은 놀라운 힘을 발견해보세요!”

빅데이터 아카데미 수료 후 어떤 변화가 있었나
2개월의 새로운 경험이 원가를 크게 바꿔 놓는다는 건… 대신 회사에서 관공서 대상의 대용량 데이터 분석 시스템 구축제안을 할 때면, 매우 구체적으로 접근할 수 있었습니다. 더불어 이미 진행중인 공공 빅데이터 프로젝트를 자세히 보면, ‘어느 단계까지 접근하는 사업이구나’ 하는 게 눈에 들어오더군요. 분석을 하지 않고 수집에 머무는 빅데이터 프로젝트들이 많다는 것도 알게 됐고요. 이런 안목으로 고객들과 접촉하다 보니 더 신뢰를 얻을 수 있었습니다.

데이터 분석 프로젝트를 해보고 느낀 점은

빅데이터에서 다루는 데이터도 결국은 2진 데이터이기 때문에 그동안 다뤄왔던 일반 데이터와 다를 게 없다는 생각을 하게 됐습니다. 그래서 요즘은 ‘빅데이터 분석’이라는 말 대신에 예전처럼 ‘데이터 분석’이라는 말을 쓰나 봅니다^^ ‘정형 데이터든 비정형 데이터든 분석할 데이터가 있다면, IT 전문가로서 경험을 동원하여 처리

할 수 있겠다’는 자신감을 갖게 된 것이 큰 소득이었습니다. 더불어 기회가 달지 않아 데이터 분석을 시도해보지 못한 분들에게는 ‘비슷한 입장에 있는 사람들이 저렇게 해내는 구나’ 하는 작은 본보기이자 희망이 된 것도 기분 좋은 일이고요.

영화 흥행예측 프로젝트의 규모가 매우 커다고 들었다

2주 동안의 집체교육으로 업무 공백이 컸기에 당초에는 수료 프로젝트 완료에 의미를 뒀는데, 하다 보니 일이 자꾸 커지더군요. 팀 프로젝트이므로 팀원들과 모여서 얘기를 나누다 보면 힘이 생겼고, 한번 해보자는 자신감이 들기 시작했습니다. 요즘 ‘기운이 사람을 좌우한다’는 생각을 많이 합니다. ‘저 사람이 마지 못해 참여한다’는 느낌이 들기 시작하면, 주변 사람들까지 열의가 떨어지면서 결과도 흐지부지해지기 쉽습니다. 팀원들 모두 즐겁고 의욕적으로 참가해서 여러 어려움을 잘 해결해 나갈 수 있었습니다. 그 과정에서 영화감성분석사전을 만드는 등 당초 계획보다 프로젝트가 커졌고요.

데이터 분석을 경험해 보지 못한 사람에게 해주고 싶은 얘기가 있다면
‘실행이 중요하다’고 생각합니다. IT 업계에서 몸담아온 분이라면 자신 안에 숨어 있는 충분한 능력이 있음을 믿고 일단 해보는 것입니다. 실행 과정에서 컴퓨터를 처음 배웠을 때의 두근거림도 경험할 수 있고, 내 안에 숨어 있는 놀라운 힘도 발견할 거라고 봅니다. 물론 당초 생각하지 못했던 어려움도 만나겠지만, 어려운 만큼 풀었을 때 느끼는 성취감도 클 거구요 ^^

모든 일을 쉽게 풀어나가는 능력이 있어 보인다

제가 자주하는 농담이 있어요. ‘밥을 많이 먹으면 어떻게 될까?’ 이렇게 질문하면 상황에 따라서 다르지만 여러 가지의 답이 나오죠. 전 이렇게 말합니다. ‘배불러’. 그러면 실소가 터져 나오죠. 좀 시간이 흐르고, 비슷한 질문을 또 합니다. ‘40대의 건강한 남자가 검은콩을 많이 먹으면 어떻게 될까?’ 좀 전의 질문과는 다르게 수식어가 붙으니까 또 여러 가지의 답이 나오니다. 전 또 이렇게 말합니다, ‘배불러.’ 제 말의 핵심은 수식어인 형용사와 부사를 빼고 나면 본질적인 문제(음식을 많이 먹는 것)만 남는데 그게 문제이고, 이에 대한 답(배부르다)은 간단하다는 거죠. 시스템도 마찬가지라고 봅니다. 소프트웨어, 하드웨어, 네트워크, DB, 그리고 최근의 이슈인 빅데이터는 모두 사람들을 사용하는 거예요. 본질적으로 사람을 편리하게 하고 도와주는 도구들일 뿐인 거죠.

수료 프로젝트에서 팀장의 역할이 매우 크다고 들었다

팀장은 아무도 지나간 흔적이 없는 눈길을 앞장서 가면서 뒷사람을 위해 발자국을 남겨주는 사람이 아닐까 합니다. 저의 작은 능력 가운데 하나가 ‘잘 할 수 있는 사람’을 더 잘하게 하는 촉매제 역할인 것 같습니다. 또 시작한 일은 즐겁게 함께 풀어나가자는 생각을 늘 갖고 있습니다. **BIG**

영화 흥행예측 분석



프로젝트 소개

소셜 데이터를 비롯해 포털 사이트의 영화 감상평을 수집·분석해 향후 개봉 예정작의 흥행을 예측하는 프로젝트다.
프로젝트 수행 결과, 한글의 특수성을 반영한 분야별 감성용어사전 등이 필수적임을 알게 됐다.

구분

분석 전문가 과정 | 예측 분석



진행

분석 전문가 과정 1기 영화 흥행예측 분석팀

• 정근훈 팀장	SI 업체 기술연구소장	컴퓨터공학 석사	경력 17년
• 김진도 팀원	DB 서비스 업체 개발자	문헌정보학 학사	경력 15년
• 배태승	입찰정보 서비스사 개발자	컴퓨터공학 학사	경력 13년
• 이동현	입찰정보 서비스사 개발자	멀티미디어학 학사	경력 07년
• 이상혁	SI 업체 개발자	컴퓨터공학 학사	경력 13년



지도

김경태



프로젝트 기간

2013년 06~07월



적용 도구

R, MySQL(RDBMS), Python, JAVA



수집 데이터

- 정형 데이터 : 한국영화진흥위원회 영화정보
- 비정형 데이터 : 네이버, 다음, 네이트의 영화평



산출물

- 영화 감성분석용 긍부정사전
- 영화 감성분석 모델링
- DBguide.net에 분석 프로젝트 진행경험 연재



교육참여 형태

자발적 참여 (3) / 회사 권유 (2)



빅데이터 아카데미 수강 후 달라진 점

- 수강 전 • 책이나 세미나로 접했지만 막연함 • 분명하지 못한 정보
 수강 후 • 빅데이터 분석 체험에 따른 자신감 • 오픈소스 분석도구를 알게 돼 바로 실무에 적용
 • 소속사는 데이터 분석 분야로 사업영역 확장

데이터가 우리에게 말을 걸어오다 상장폐지기업 예측 분석



로우 데이터를 확보할 수 있고, 팀원들의 장점을 최대한 활용할 수 있어야 한다는 기준으로 ‘상장폐지기업 예측 분석’을 프로젝트 주제로 진행했다. 눈으로 확인하기 전에 믿지 않으려는 참가자들의 태도는 ‘허술하게 접근해서는 데이터가 대답해 주지 않음’을 체험하면서부터 바뀌기 시작했다. 공교롭게 프로젝트 진행 과정에서 사회를 떠들썩하게했던 D사를 상장폐지 가능성이 높은 곳으로 예측한 것이 적중해 주목을 받았다. 상장폐지기업 예측 분석을 역으로 ‘대박주’ 예측 분석에 적용할 수 있을지 함께 알아보자.

Challenges

베테랑 회계사들, 데이터 분석과 거루다

2주간의 빠듯한 집체교육을 마치고, 긴장과 설렘 속에서 수료 프로젝트를 시작했다. 주제만큼은 어느 팀보다 빠르게 선정했다. 처음에는 팀원마다 자신이 소속된 기관 또는 기업 데이터를 분석하면 좋겠다는 의견을 내놓았다(‘빅데이터 아카데미’에 등록할 때, 이미 저마다 자신이 소속된 곳의 데이터 분석에 관심을 갖고 있지 않았나 싶다). 한 대기업의 품질 데이터, 공공기관의 데이터, 게임사의 데이터가 그것이다. 하지만 수료 프로젝트용 데이터 확보는 생각처럼 쉽게 풀리지 않았다. 한기원 팀장의 뛰어난 리더십 덕분에 외부 데이

터를 확보해 분석하기로 하고 ‘상장폐지기업 예측’을 최종 프로젝트 과제로 선정했다.

상장폐지기업 예측 분석팀(이하 상장폐지팀)의 한기원 팀장은 회계법인에서 다년간 근무하면서 데이터 분석을 통해 상장폐지 업체를 예측할 수 있다면, 여러 가지로 유용할 것이라고 호언했다. 회계사로 근무했던 안정국 팀원이 이 주제의 높은 효용성에 공감하면서 다른 팀원들도 적극 찬성하는 분위기가 형성됐다. 앞서 예비 분석 주제로 선정했던 대기업의 품질관리 데이터, 공공기관의 데이터, 게임업체의 데이터는 외부 유출이 차단돼 더 이상 분석과제를 놓고 고민할 필요가 없어졌다. 안정국 팀원은 소속사의 품질 데이터를 분석해 보면 기업 경쟁력 강화에 도움이 될 것이라고 판단, 분석 전문가인 김경태 지도 교수로부터 무료로 분석 컨설팅을 받을 좋은 기회를 놓치고 싶지 않았다. 하지만 보안문제 때문에 데이터 수급이 어려워지면서 접어야 했다.

아쉬움을 뒤로 하고, 2명의 회계사가 포함된 상장폐지팀은 이번 프로젝트가 비교적 쉽게 이뤄질 것이라고 호언하며 수료 프로젝트에 들어갔다. 상장폐지팀에는 10년 이상의 회계사가 2명이 포함됐기에 직감을 통한 예측과 분석을 통한 예측의 차이가 없을 것이라고 판단했다.

Solution

‘KISVALUE’를 정형 데이터로 선택하다

먼저 분석할 데이터를 확보했다. 한국신용평가정보에서 제공하는 ‘KISVALUE’를 정형 데이터의 확보처를 선정해 기업의 일반정보, 재무정보, 재무비율 데이터를 수집했다. 상장사의 데이터를 먼저 수집한 다음, 상장폐지사의 데이터도 별도로 수집했다. KISVALUE의 데이터 말고도 금융감독원에서 제공하는 DART에서 데이터를 가져올 수도 있으나, 회사별로 일일이 직접 수집해야 했기에 유료였지만 KISVALUE 데이터를 최종적으로 선정했다.

정형 데이터를 확보했으므로 이제 비정형 데이터를 추가해 예측분석의 정확도를 높여보고 싶었다. 비정형 데이터는 상장사의 공시정보가 적합할 것이라는 판단을 했다. DART의 상장사 공시를 두 명의 팀원이 담당해 JSON 파일로 획득했다. 이 데이터를 이용하기

위해 공시 제목과 공시 횟수를 수집하려 했으나, 공시의 종류가 너무 많고 동일한 제목의 공시가 어떤 것은 궁정적이고, 어느 것은 부정적이어서 유용하지 않을 것으로 판단해 일단 정형 데이터를 분석해본 후 성능이 나오지 않을 경우에 비정형 데이터를 적용하기로 했다.

처음 한 주간은 무척 진도가 빨랐다. 데이터를 모두 획득해 분석 전문가 과정 3기 어느 팀보다도 앞서 데이터 마트 생성을 완료했다. 프로젝트에서 90% 이상의 시간을 소모한다는 마트(MART) 구성이 1주도 지나지 않아 완성되자 팀원 모두 놀라워했다. 하지만 그것이 끝이 아니었다. 수집한 데이터를 보니 타깃 데이터가 너무 적었다.

3개년의 상장폐지사 데이터를 묶어 특성을 파악하다

지도 교수님과 상의한 결과, 타깃 데이터 건수가 5%는 넘어야 모델이 유의미하게 나올 것이라는 조언을 받았다. 고민 끝에 3년 간의 상장 폐지회사를 묶기로 했다. 하지만 3년간의 상장폐지사 데이터를 어떻게 묶는단 말인가?

분석 경험이 많지 않은 상장폐지팀은 3년 간의 데이터를 묶는 것과 헤더(Header) 정의를 쉽게 이해할 수 없었다. 어려움에 맞닥트리자 제대로 할 수 있을까 하는 두려움이 고개를 들기 시작했다. 결국 여려 차례에 걸쳐 데이터를 다시 가져오고 정렬하는 과정을 거듭하며 상장폐지사의 과거 3년 데이터 수집에 성공했다. 상장폐지 1년 전을 “Y-1”, 2년 전을 “Y-2”, 3년 전을 “Y-3”으로 지정해 데이터를 밀어 맞추는 작업을 했다. 정상 기업의 과거 3년 데이터와 합해 마트를 구성했다. 연도별로 구분된 헤더도 앞의 “Y-1”, “Y-2”, “Y-3” 형식으로 모두 바꾸며 TRAIN DATA를 완성했다. 테스트 데이터 생성은 학습효과 덕분에 더 수월하게 할 수 있었다. 이 과정에서 5명의 팀원 모두 데이터를 생성하며 DB 전문가의 노고를 몸소 경험했다.

이제부터 본격적인 실력을 발휘할 차례다. 회계 분야에서 다양한 경험을 가진 한기원 팀장의 진가가 드러나기 시작했다. 파생변수를 정의하는 날, 한 팀장은 고객과의 약속을 연기하고 늦게까지 남아 변수를 정의했다. 무리한 일정이었지만, 프로젝트 참가 팀원도 빨리 결과를 보고 싶어했기에 변수 정의를 강행했다.

드디어 모델링 단계에 진입했다. 생각과 달리 예측이 쉽지 않을 거 같다는 불안한 마음이 들기 시작했다. 직감이 적중할 거 같다는 생각을 버리지 못하던 팀원들은 데이터 정렬 과정에서 어려움을 겪으며 결과는 예측할 수 없다는 자세로 바뀌기 시작했다.

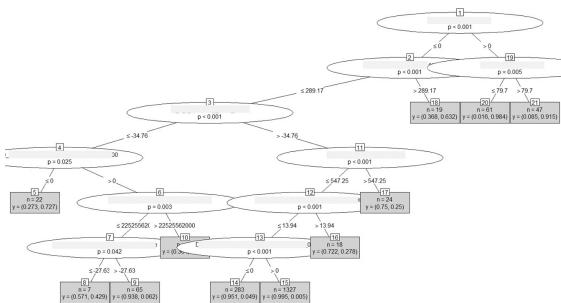
Target의 재정비: 프로젝트 모델링을 하다

최초 모델링 결과, 정확성은 90% 이상, Precision은 60% 정도, Detect rate 또한 60% 정도로 나왔다. 팀원들은 생각보다 낮다고 생각했으나, 지도 교수께서는 타깃이 5% 내외이므로 그 정도 수치면 상당한 의미를 갖는다고 조언했다. 이때부터 우려했던 분위기에 서 뭔가 될 거 같다는 분위기로 반전됐다. 수료 프로젝트가 시작된 지 2주를 막 지난 시점이었으므로 팀원들은 너무 일찍 끝났다는 생각에 들었다.

하지만 여기서도 또 다른 이슈가 기다리고 있었다. 예측 보고서 초안을 만드는 과정에서, 상장폐지사들 중에는 흡수합병된 업체들이 있음을 발견한 것이다. 이를 어떻게 처리할지 난감했다. 여러 흡수합병 사유가 있겠지만, 인수합병(M&A) 등이 많으므로 회사가 어려워 상장폐지됐다고 볼 수 없었다. 이를 제외해야 정확도를 높일 수 있었다.

결국 타깃(Target)을 재정비하고 모델링했더니 Precision과 Detect Rate가 다소 개선됐다. 여기까지는 어느 분석 프로젝트에

그림 1. party 모델. 주요 변수별 의사결정트리에 의해 나뉜 기업별 소속 집단(노드)과 집단 내 기업 개수(n), 정상/상폐 확률(y)



*예측한 상장폐지 기업명을 지움

서나 으레 겪는 일일 것이다. 이제부터 이 프로젝트의 진가를 드러내야 할 시점이다.

실제 상장폐지사를 예측해 분류분석의 진가를 확인하다

모델링 결과를 정리하던 중 어느 회사들이 상장폐지사로 예측됐는지 확인해 보자는 의견이 나왔다. 하나하나 분석 결과를 확인하면서 '그렇지, 이 회사들은 모두 이미 알고 있는 위태로운 곳들이야'라며 직감이 맞아떨어진 분석결과로 받아들였다. 공교롭게 그 즈음 D사에 대한 뉴스가 세상을 떠들썩하게 했다. 팀원들은 우리의 예측 모델이 D사를 예측했는지 찾아보았다. 리스트에 들어 있었다.

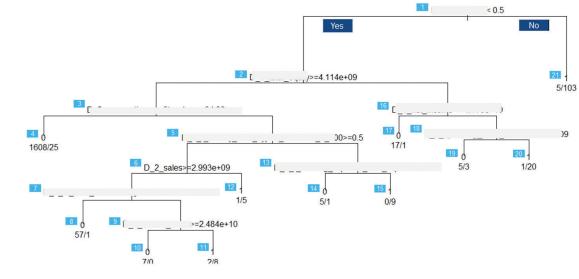
이때부터 상장폐지팀은 같은 기수의 다른 팀들로부터 주목 받는 팀으로 떠올랐다. 어떤 케이스로 예측됐는지, 어떤 변수들이 예측 변수로 쓰였는지 자세히 검증에 들어갔다.

〈그림 2〉은 한국증권거래소(www.krx.co.kr)의 '상장현황/상장폐지 현황'을 캡처한 것이다. 이 중 '에스와이코퍼레이션'과 '위다스'를 탐지해 정확한 예측 성능을 보여줬다.

분석 결과: 데이터가 우리에게 말을 걸어오다

상장폐지에 이르게 한 요인들은 다음과 같다. 감사의견, 유동성 지표, 부채비율, 영업이익, 몇몇 파생변수 등 모델의 예측 변수들이 이미 알고 있던 것들이었다. 여기서 어떤 의미를 더 찾을 수 있을

그림 2. rpart 모델. 중요 변수별 의사결정트리에 의해 나뉜 기업별 상폐 분류(0: 정상, 1: 상폐)와 정상/상폐 회사 수(정상 기업 수/상폐 기업 수)



*예측한 상장폐지 기업명을 지움

종목코드	기업명	폐지일	폐지사유
A008080	에스와이코퍼레이션	2013/09/25	신청에 의한 상장폐지
A049000	예당	2013/09/10	기업의 계속성 및 경영의 투명성 등을 종합적으로 고려하여 상장폐지기준에 해당
A056810	위다스	2013/09/10	의견거절(감사범위 제한)

그림 3. 한국증권거래소의 '상장현황/상장폐지현황' 화면

까 하는 생각이 들면서 한동한 들떴던 분위기가 가라앉았다. 하지만 프로젝트가 진행됨에 따라 상장폐지 기준에 없는 변수들도 있고, 우리가 아는 변수라 해도 1700여 개 상장사에 대해 일일이 이를 적용하면서 상장폐지 가능성을 판단하기란 어려울 것이라 생각을 하게 됐다. 무엇보다 전에는 그저 지표였던 숫자들이 의미를 갖고 우리에게 다가오기 시작했다. 데이터가 우리에게 원가를 말하고 있는 것을 알아 차리게 됐다. 아쉬운 점은 어렵게 수집했던 비정형 데이터를 사용하지 못했다는 것이다. 프로젝트 종료 시점까지 비정형 데이터를 정리했으나 워낙 경우의 수가 많고, 의미의 해석이 어려워 이 부분은 과제로 남겨 두기로 했다.

'대박주' 예측도 가능할까?

주식 시장에서 쉽게 사용할 수 있는 요인을 이용해 회사의 경영상태 예측이 가능하고, 이를 이용해 투자를 피해야 할 주식을 찾아냈 다. 그렇다면, 역으로 이 정보를 토대로 소위 대박주를 예측하는 것도 가능하지 않을까?

수료 프로젝트를 완료한 이후 이 모델을 더 발전시킬 가치가 있다 고 판단해 회계사로서 경험을 가진 안정국 팀원이 개인 프로젝트로 개선해나갔다. 당초 2개의 알고리즘을 7개로 확장하고, 3년간의 타깃 데이터를 이용하던 것에서 벗어나 1년 간의 타깃 데이터만으로 예측해 보았다. 이를 통해 상당한 효과를 보았다. 예측에만 사용하는 것보다 어려운 회사를 예측하는 것으로 사용할 계획으로 Detect Rate를 희생하고, 예측 회사 범위를 늘려 보았다. 이렇게 하니, 언론에 나오던 어려운 회사들이 탐지됐다. 경영상황이 어렵다고 언론에 보도되던 대기업 계열 건설사 3곳을 탐지했다. 아마 이 정보를 주식 투자에 이용했다면 적지 않은 투자자들이 손실을 피할 수 있지 않았을까?

기업 재무상태 파악 전문가라고 자처하던 한기원 팀장과 안정국 팀원은 이번 프로젝트를 통해 데이터가 무엇을 말하려는지에 대해 귀 기울일 필요가 있음을 알게 됐다. 그동안 현장에서 쌓은 경험이라는 변수 중심으로 상황을 파악하거나 직관을 토대로 결론을 내렸을 때의 위험함도 몸소 체험했던 계기가 됐다. 이는 회계사로 참여했던 두 명의 팀원에게는 재무정보 분석 실력을 한 단계 끌어올리는 계기가 됐다. 더불어 길기상, 강경민, 최동철 팀원은 데이터 분석의 힘과 재무정보에 대한 한층 높은 이해를 하는 계기가 됐다.

Conclusion

실무 경험과 데이터 분석의 결합을 기대하다

상장폐지팀은 수료 프로젝트 결과를 토대로 각자의 소속사에서 유사한 프로젝트를 진행하고자 하는 열의와 자신감을 갖게 되었다. 무엇보다도 실제로 데이터를 분석하면서 얻은 통찰력과 분석 전문가의 데이터 정결과 데이터 분석이 합쳐지면, 어떠한 힘을 얻게 될지도 실감했다.

특히 데이터 분석이 현업 전문가의 직감보다 예측 정확도가 높다는 것을 몸소 느끼면서 DB 전문가에서 분석 전문가로 거듭나야겠다는 생각을 하게 해줬다.

상장폐지 예측 모델링은 그 자체로서도 많은 분야에서 활용할 수 있을 것이다. 주식투자자에게 정보를 제공해 투자 손실을 줄이고, 해당 기업에 대해서는 어떠한 지표 때문에 예측됐는지를 활용해 위험을 극복할 방안을 모색하는 데 도움을 줄 수 있지 않을까 한다. 더 나아가 상장기업 정보뿐 아니라 비상장기업 정보까지 활용한다면, 금융기관의 기업 대출 부실화 예측, 건설사의 하청업체 부도 예측, 보증기금의 보증손실 예방 등에서 활용할 수 있을 것이다. **BIG**



안정국 분석 전문가

과정 3기 기장(The ECG 상무)

“경험이라는 틀에서 벗어나게 해준 데이터 분석”

빅데이터 아카데미 수강 후 달라진 점은

데이터 분석 방법을 배워야겠다는 마음으로 아카데미에 등록했지만, 수료 프로젝트를 진행하면서 당초 생각이 달라졌어요. ‘빅데이터가 현업 부서에 직접 도움을 줄 수 있겠구나’ 하는 생각이 들면서 빅데이터의 매력을 알게 됐죠. 일 배우러 갔다가 돈 버는 방법에 눈을 떴다고나 해야 할까요^^ 빅데이터 분석을 실체를 맛보고 나서 ‘제 경험의 테두리 안에서 제가 보고 싶은 대로 보려 했음’을 실감했습니다. 빅데이터 아카데미 교육이 제 생각의 프레임을 바꿔야 함을 알려줬습니다.

프로젝트에서 개발한 분석 기법을 실무에 적용할 수 있다고 보는가

예. 상장 폐지기업 예측 프로젝트 후 한 재무정보 서비스 업체로부터 이 모델을 추가 개발해줄 수 없느냐는 요청을 받았어요. 프로젝트 기간 중에 개발한 것을

토대로 추가 모델링을 했는데 예측의 정확도가 몰라보게 올라가는 것을 확인했습니다. 하지만 이 프로젝트는 최종적으로 완료되지 못했습니다. 상장폐지 가능성이 높은 기업 리스트가 외부에 공개됐을 때 그 파장이 적지 않고, 법적 문제로까지 연결될 가능성이 있었기 때문입니다. 이미 금융권에서는 내부적으로 비슷한 예측 시스템을 운영할 거라고 봅니다. 예측 결과를 외부로 공개하지 않은 채 말입니다.

분석 전문가 과정은 어떤 사람이 들으면 도움이 될 거라고 생각하나

생산 원가 분석, 마케팅, 경영기획 실무자들이 들으면 좋겠다는 생각을 했어요. 현업 담당자들은 일반적으로 현장 경험에서 얻은 자신의 직감이 빗나가지 않을 거라는 확신을 갖고 있는 경우가 적지 않아요. 이러한 사람들이 직접 데이터 분석을 해보면, 훨씬 객관적인 시각을 확보할 수 있을 거라고 봅니다. 현업 담당자들의 직감 기반의 예측과 데이터 분석을 통한 예측을 비교해보면 데이터 분석을 통한 예측이 훨씬 정확도가 높다는 것을 알게 될 겁니다.

어떤 일을 하고 있나

빅데이터 전문 업체인 ‘The ECG’에서 일하고 있습니다. 얼마 전까지 ‘그룹공통 재무시스템 컨설팅’ 업무를 했는데 ‘빅데이터 아카데미’가 인연이 되어 새출발을 했습니다. 대기업 계열 SI 업체에서 사회 생활을 시작했고, 공인회계사로서 회계법인에서 워크아웃 제도, 내부 관리회계 등 새로운 트렌드를 앞서 수용해 고객에게 컨설팅하는 일을 했습니다.

컨설팅을 했다면 비즈니스 인텔리전스도 잘 이해하고 있을 거 같다

어느 정도는 알고 있어요. 비즈니스 인텔리전스(BI)가 빅데이터 분석과 다른 점은 데이터 마이닝 없이 임의로 모델링했다는 점입니다. 경험을 토대로 ‘그럴 거야’ 하는 지표를 토대로 접근한 게 BI였습니다. ERP 같은 데이터를 임의의 지표를 토대로 분석·보고했던 것이 전통적인 BI라면, 여기서 마이닝을 통해 변수, 즉 객관적인 지표를 찾아서 분석하는 것이 빅데이터 분석입니다. 마이닝을 쉽게 설명하자면, 분석에 영향을 주는 변수를 찾아는 내는 과정입니다. 참고로 예전의 BI에서도 데이터 마이닝이 필요했지만, 그 필요가 매우 제한적이었습니다.

고객들에게 성공적인 빅데이터 분석 환경 구축 조건을 어떻게 소개하고 있나

기업에서 빅데이터는 분석 플랫폼 기술을 담당하는 IT 부서와 실제로 데이터 분석을 하는 실무부서에서 관리·활용하는 형태로 양분해 볼 수 있습니다. 빅데이터가 어떤 효과를 거둘 수 있는지를 자꾸 생각해야 성공적으로 구축·운영할 수 있다고 강조하고 있습니다. 어떤 측면에서 도움이 되고, 이에 따라 매출이 얼마나 늘어날지 효과를 염두에 두고, 예측에 필요한 데이터는 무엇이고, 어떻게 데이터를 수집해야 할지를 고민하는 과정이 따라야 합니다. **BIG**

상장폐지기업 예측 분석



프로젝트 소개

상장사의 재무정보를 이용해 상장폐지 가능성 예측한
분석 프로젝트. 추가 모델링을 통해 상장 폐지사 예측의 정확성 제고는
물론, 주가예측 등 활용범위를 넓여갈 수 있을 것으로 기대된다.

구분

분석 전문가 과정 | 분류 분석



진행

분석 전문가 과정 3기 상장폐지기업 예측 분석팀

• 한기원 팀장	회계법인 공인회계사	회계학 학사	경력 18년
• 안정국 팀원	데이터 분석업체 공인회계사	수학 석사	경력 15년
• 강경민	SI업체 수석 연구원	물리학 학사	경력 15년
• 길기상	게임 업체 설계 및 개발 SE	컴퓨터과학 학사	경력 13년
• 최동철	정부 산하기관 선임	공간정보공학 학사	경력 06년



지도

김경태



프로젝트 기간

2013년 09~10월



적용 도구

R, JAVA, MySQL, MS Excel



집집 데이터

- 정형 데이터 : KISVALUE의 재무 데이터
- 비정형 데이터 : DART 전자공시 데이터



산출물

- 3개년 재무정보를 이용한 상장폐지 예측 모델
- 2013년 예측 결과



교육참여 형태

자발적 참여 (3) / 회사 권유 (2)



빅데이터 아카데미 수강 후 달라진 점

- 수강 전 • 일상적인 업무의 일부로서 관리 포인트를 파악하려 함 • 데이터 사이언티스트와 분석은 다른 사람들의 일
 수강 후 • DB 전문가의 데이터 수집의 중요성 인식 • 현업 전문가의 경험과 IT 전문가의 협력을 통해 얻는 힘 체험
 • 기업 경쟁력 강화의 핵심 = 현업 부서의 노력 + IT 전문가의 지원 + 분석 전문가의 조언

국내 최초로 쇼핑몰 실시간 분석에 도전하다 쇼핑몰 상품 트렌드 분석 플랫폼



실시간 급상승 검색어 분석 및 상품별 트렌드 분석 시스템은 빅데이터라는 개념이 나오기 전부터 대형 쇼핑몰을 중심으로 이미 시도되고 있었다. 그러나 그 분석 데이터의 크기와 분석 주기 측면에서 빅데이터 기술과 비교해보면 차이가 크다. 빅데이터 이전의 분석 주기는 실시간이라는 말이 무색 할 정도로 20분 간격이나 된다. ‘쇼핑몰 트렌드 분석 플랫폼팀’이 선정한 과제는 10초 단위로 실시간 급상승 검색어를 분석해 SparkLine 차트로 시계열적으로 흐름을 보여주는 시스템이다. 일부 쇼핑몰에서나 도입한 실시간 트렌드 분석 플랫폼은 어떤 형태로 구축되는지 함께 알아보자.

Challenges

데이터 확보 가능여부를 기준으로 주제를 잡다

기술 전문가 과정의 집체교육 중에 수료 프로젝트 팀이 결성됐다. 팀원들은 첫 모임부터 프로젝트 주제로 어떤 것을 선정할지를 놓고 생각을 나눴다. 이를 후 각자 준비한 주제를 공유해 다수결로 최종 주제를 정하기로 했다. 다행스럽게도 ‘쇼핑몰팀’ 팀장이 직장에서 빅데이터 프로젝트를 앞둔 상태여서 빅데이터 요건을 수집하던 중이었다. 수료 프로젝트가 일종의 POC(Proof Of Concept)가 돼 업무에도 도움이 될 것이라는 기대를 하고 있었다.

쇼핑몰 팀장은 이미 하둡 에코시스템의 설치와 설정, 기본적인

MapReduce 코딩 등을 경험한 상태였다. 하지만 빅데이터 기술로 대용량 데이터를 실시간으로 빠르게 조회·시각화하는 부분은 경험이 없었다. 특히 예제 데이터 수준이 아닌 방대한 실제 데이터를 갖고 비즈니스 요건을 MapReduce 프로그램으로 배치(Batch)화해 2차 데이터를 생성하는 방법도 코딩해 보고 싶었다. 이렇게 두 가지 기술을 포함하는 요건으로 '실시간 급상승 검색어 분석 및 상품별 트렌드 분석 시스템'이라는 주제를 도출했다.

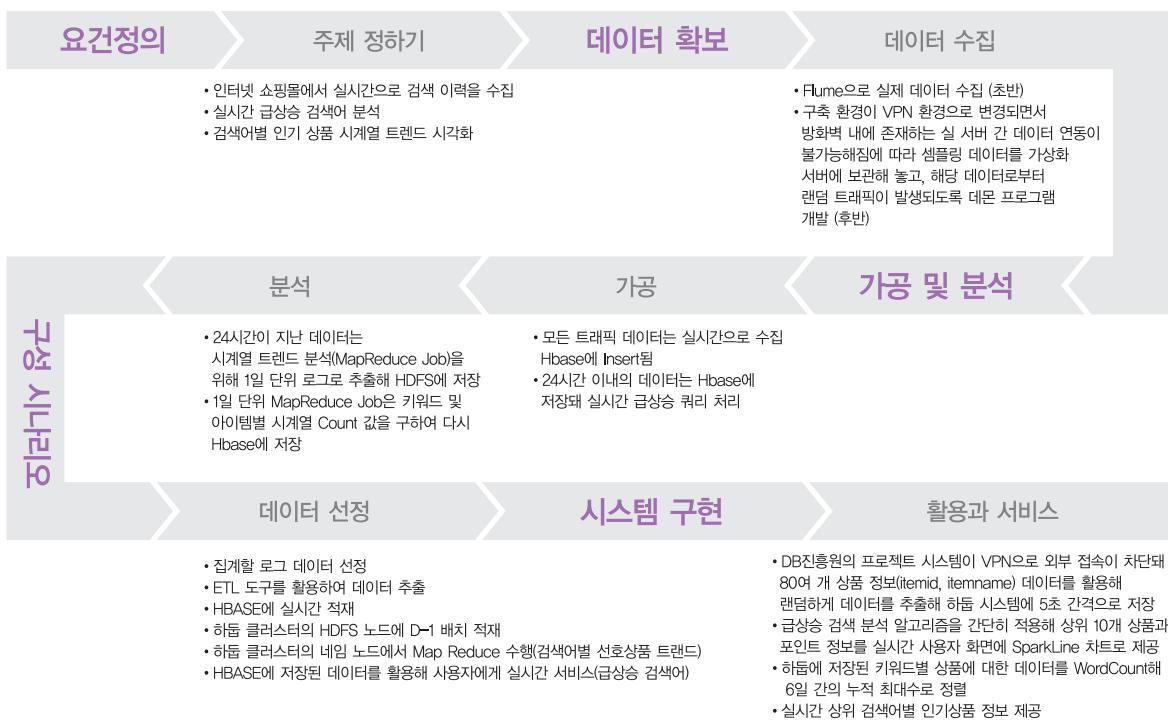
말로만 듣던 실시간 빅데이터 분석을 체험하다

쇼핑몰팀이 짧은 시간 안에 수료 프로젝트의 주제를 정할 수 있었던 배경은 단 한 가지였다. 3개 후보 주제 가운데 분석 데이터를 구할 수 있는 아이템은 팀장이 제안한 쇼핑몰 상품 트렌드 분석이 유일했기 때문이다. 외부 유출에 민감한 부분에 마스킹 처리한 회사의 데이터로 수료 프로젝트를 해보고 싶었던 팀장의 제안은 모든 팀원으로부터 흔쾌히 받아들여졌다.

쇼핑몰팀이 선정한 실시간 급상승 검색어 분석 및 상품별 트렌드 분석 시스템은 빅데이터라는 개념이 나오기 전에도 대형 쇼핑몰을 중심으로 이미 시도되고 있었다. 그러나 그 양과 분석 주기 측면에서 빅데이터를 활용했을 때와 큰 차이가 있다. 쇼핑몰팀 팀장이 소속된 회사의 인터넷 쇼핑몰만 보더라도, 이미 실시간 급상승 검색어 메뉴가 있었지만, 그 분석 주기는 실시간이라는 말이 무색할 정도로 20분 간격이나 된다. 국내 대표 포털 가운데 한 곳인 'N' 서비스 조차도 실시간 급상승 검색어를 수분 주기로 보여주고 있다. 모든 상품의 트렌드를 고객에게 제공하는 인터넷 쇼핑몰은 많지 않다. 매출 등 정형화한 데이터를 분석해 보여주는 서비스는 일부 쇼핑몰에서 제공하고 있지만, 검색어 기반의 비정형 데이터를 분석해 트렌드를, 정보계가 아닌 운영계 서비스로서 제공하는 곳은 구글 등 극히 소수의 외국 사이트에 국한돼 있다.

쇼핑몰팀이 선정한 주제는 10초 단위로 실시간 급상승 검색어를 분석해 SparkLine 차트로 시계열적으로 흐름을 보여주는 시스템

그림 1. 쇼핑몰 상품 트렌드 분석 플랫폼팀의 시스템 구축 과정



구축이었다. 검색어별 선호 상품의 시계열적인 트렌드를 빠르게 보여준다는 측면에서, 빅데이터 시스템의 다양한 장점을 극대화한 주제였다.

Solution

집체교육 기간 중에 프로젝트를 기획하다

쇼핑몰팀의 빅데이터 수료 프로젝트는 집체교육 기간 중에 시작됐다. 주제가 빨리 정해 졌고, 수업기간 중에 팀원들이 모여 스키마 디자인, UI(User Interface) 논의, 역할 분담, WBS 작성 등을 시작할 수 있었기 때문이다. 일반적인 폭포수 모델(Water Fall) 프로젝트에서는 분석 및 설계 단계가 프로젝트 업무의 2/5 정도를 차지한다. 쇼핑몰팀은 이 단계를 집체교육 동안에 마칠 수 있어서 구현 시간을 많이 아낄 수 있었다.

학창시절 아래 이렇게 공부에만 매진했던 적은 처음이었다. 특히 학창 시절에는 다양한 과목을 공부하지만, 이렇게 한 가지 주제를 2달에 걸쳐 공부한 것은 인생을 통하여 매우 드문 경험이었다. 그렇게 짧다면 짧은 8주 동안 빅데이터라는 한 가지 주제에 몰입해 집중으로 공부를 하니, 빠르게 많은 것을 받아 들일 수 있었다. 쇼핑몰팀은 매일 점심을 함께한 다음, 30분에 걸쳐 프로젝트 기획 회의를 했다.

첫 번째 난관은 예상했던 바대로 하둡 에코시스템 구축에서 닥쳤다. 시스템 구축은 정동민 팀원 몇이었는데, 목표했던 기능이 수업 시간에 배우지 않은 영역까지 포함하고 있는 데서 문제가 발생했다. 이에 따라 시스템 구축을 하는 데서 예상보다 시간이 더 걸렸다. 시스템 구축이 끝나야 다음 과정으로 넘어갈 수 있었으나, 하둡 에코시스템 구축에 2주가 넘는 시간이 소요됐다. 팀장은 조바심이 나기 시작했다.

하둡 에코시스템 구축에서 첫 번째 난관에 봉착하다

팀장은 만일에 대비해 정 팀원이 시스템 구성을 하던 것과 별도로 병렬로 다른 디렉터리에 또 다른 시스템을 구축하고 있었다. 문제는 여기서 발생했다. 이런 식의 병렬 구성은 한 번도 시도해 본 적이 없었다. 하둡이라는 HDFS 계정이 병렬 설정에 따라 추후에 문제를 일으킬 수 있음을 알기까지 또 다시 2주가 걸렸다.

생각보다 프로젝트 일정이 길어지자 팀장은 좀 더 빠르게 많은 에코 시스템을 한 번에 구축하고자 제안했다. 아파치 인큐베이터 프로젝트에 있는 Ambari(<http://incubator.apache.org/ambari/>) 모듈을 이용하기로 했다. Ambari를 이용하면, 기본적인 하둡 에코 시스템 구성을 한방에 해결할 수 있고, 덤으로 Nagios, Ganglia 같은 강력한 모니터링 툴까지 함께 설치할 수 있다(〈그림 1〉은 실제 가상 서버에 설정 완료된 모니터링 화면이다).

그림 1. 가상서버 환경에서 구현한 하둡 에코시스템 모니터링 화면

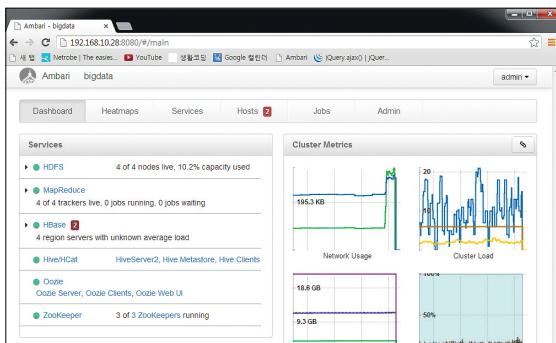


그림 2. 4대의 프로젝트용 VM 서버에 설치된 하둡 에코 시스템 클러스터
edu-hadoop-41 (3.7 GB, 1 cores)

NameNode Nagios Server Ganglia Collector Ambari Server Django Web Server

edu-hadoop-42 (3.7 GB, 1 cores)

SNameNode JobTracker Oozie Server ZooKeeper DataNode

edu-hadoop-43 (3.7 GB, 1 cores)

HiveServer2 Hive Metastore WebHCat Server ZooKeeper DataNode

edu-hadoop-44 (3.7 GB, 1 cores)

HBase Master ZooKeeper DataNode Flume-NG

하지만 Ambari를 사용하는 데서도 문제가 기다리고 있었다. 몇몇 계정들이 Reserved돼 있었는데, 원인을 찾아보니 수작업으로 해당 계정이나 그룹을 미리 만들어 사용할 경우, 미묘하고 찾기 힘든 문제를 야기했기 때문에 확인됐다. 이 때문에 쇼핑몰팀은 엉켜버린 시스템을 완전히 복구하고, 시스템 설정을 100% 마무리하는 데 추가로 2주가 더 소요됐다. 결과적으로 모든 세팅을 마무리 하는 데에 전체 6주 중에서 4주를 보내고 만 것이다.

〈그림 2〉를 설명하면 다음과 같다. 온라인 쇼핑몰에서 생성되는 데이터를 Flume으로 실시간 수집해 Hbase에 Insert한다. 24시간 이내의 데이터는 Hbase에 저장돼 실시간 급상승 쿼리를 처리하고, 24시간이 지난 데이터는 시계열 트렌드 분석(MapReduce Job)을 위해 1일 단위 로그로 추출해 HDFS에 저장한다. 1일 단위 MapReduce Job은 키워드 및 아이템별 시계열 카운트값을 구해 집계 요약 테이블로 Hbase에 2차 저장된다. 이후 집계할 로그 데이터를 선정해 배치(Batch) 스케줄러에 의해 Hbase에 주기적으로 적재된다. 하둡 클러스터의 HDFS 노드에 D-1 배치를 적재하고 하둡 클러스터상에서 검색어별 선호상품 트랜드 등을 MapReduce 한다. 최종 결과는 Hbase에 저장된 데이터를 Django 위의 Python Web Service가 Thrift를 통해 실시간 쿼리하고, Json 포맷으로 서비스해 jQuery의 SparkLine 실시간 차트로 사용자들에게 최종 서비스된다.

네트워크 보안이 두 번째 난관이 되다

주어진 6주 가운데 4주를 시스템 설정에 써버렸지만, 환경과 도구를 제대로 구비했기에 이제 달리기만 하면 된다.

빅데이터 프로젝트의 모든 시작은 수집 프로세스라 해도 과언이 아니다. 쇼핑몰팀은 4주차부터 더 집중력을 발휘했다. 팀원 모두 자신의 주말을 할애해 수집 프로세스를 완성했다. 우선은 팀장이 바로 접근 가능한 개발 서버 로그를 활용했다. Flume-NG를 이용해 Tail Log를 실시간으로 수집해 TCP 로그 수집서버에 실시간으로 저장하도록 했다. 수집서버는 성능도 좋고 가장 개발 생산성이 높은 파이썬을 이용했다. Flume의 TCP 통신 모드를 활용하면, 파이썬으로 만든 서버가 실시간으로 트래픽을 받는 시나리오를 그리 어렵지 않게 구현할 수 있을 것 같았다. 헤더 등을 통해 분기 로직을

넣을 수도 있고, Append에 HDFS보다 강점이 많은 HBase에 실시간 Insert하는 로직을 Thrift를 이용해 구현 시, Flume과 HDFS 조합의 단점인 실시간성과 분기성을 모두 해결할 수 있을 것 같았다.

문제는 수집 시스템 개발과 테스트를 프로젝트 시스템이 아닌, 팀장이 소속된 회사의 개발 서버에서 진행했던 데서 발생했다. 중간 발표 불과 2~3일 전에야 쇼핑몰팀은 보안상 VPN(Virtual Private Network)을 통해서만 VM 서버에 접속 가능하도록 정책이 바뀌었다는 공지를 들을 수 있었다.

회사의 개발서버와 DB진흥원의 프로젝트용 VM 서버 간에는 방화벽으로 차단되어 접속이 불가능했다. 일주일 전 까지만 해도, 회사 개발서버와 가상화 VM 서버 간에는 단방향 TCP 통신이 열려 있었다. 즉 VM에서 먼저 요청은 불가능하였지만, 개발 서버에서 TCP를 통해 외부로 먼저 Request를 보내고 Ack Response를 받는 것은 가능했다. 그렇게 우리의 첫 개발 작업도 난관에 부딪쳤고 해결 방법이 없어 보였다.

쇼핑몰팀은 다시 데이터 수집 메커니즘을 변경했다. 리얼 서버의 로그를 샘플링해 파일 리파지터리화하고, 해당 리파지터리를 랜덤 액세스해 랜덤하게 로그가 VM 안에서 발생되게 하는 Log Generator 제작에 들어갔다. 어쩔 수 없는 선택이었지만, 이로 인해 우리는 또 다시 시간을 허비하고 말았다.

협력과 기지를 통해 완료하다

프로젝트 완료까지 시간이 얼마 남지 않았다. 로그 제너레이터를 만들고 나니 비로서 MapReduce 코딩 및 HBase 코딩을 할 수 있는 단계에 진입했지만, 2~3일 만에 프로젝트 중간 보고를 해야 했다.

목요일 마지막 중간 보고 및 중간 점검을 위해 2주차 수요일 저녁에 팀원들이 철야 작업을 하기로 했다. 중간 점검일이 아니었다 하더라도, 최종 발표를 1주일 앞두고 있었기에 촉박한 상태였다. 완료 작업은 마지막 주말에 하더라도, 주중에 꼭 서버 프로그래밍은 마무리를 지어야 했다.

HBase를 통한 Insert, Search의 개발은 파이썬과 Django Web Framework를 이용했다. 그리고 Thrift를 이용해 파이썬에서

HBase에 직접 접근할 수 있도록, Thrift Wrapper를 제공하는 Happy Base 라이브러리를 이용했다.

마지막 중간 점검 시점까지 모든 CRUD(자료의 삽입-조회-수정-삭제)를 완성하기로 했다. 모든 CRUD를 REST Web Access 할 수 있도록 개발을 마치고, 모든 UI는 100% Ajax를 통해 개발하는 것이 목표였다. 프로젝트 중간 점검 시점에는 UI를 보여줄 수 없어서, 개발에 필요한 CRUD를 마무리해 해당 작동 여부를 브라우저를 통해 JSON 포맷으로 보여주는 것을 목표로 준비했다.

실시간 데이터 수집과 실시간 분석은 HBase(On HDFS) + HappyBase(with Thrift) + Python + Django의 조합으로 구현했다. 하지만 수개월 간의 상품 트렌드를 보여주기 위한 코딩은 일단위 MapReduce Batch Job으로 집계한 결과 파일만 저장되도록 구현했다. 이후 웹에서 시각화하는 경우에는 일 단위 집계를 끝낸 상태가 되기 때문에 빠른 액세스가 가능하며, 이 또한 매번 파일 IO가 일어나지 않도록 그 결과만을 야간에 HBase 안에 넣어 두도록 했다. 일 단위로 특정 검색어로 가장 많이 조회한 아이템을 집계해, 수개월 간의 트렌드를 보여주도록 하는 MapReduce 코딩

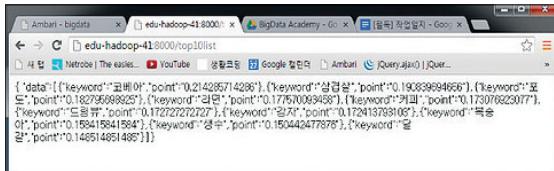


그림 3. 실시간 톱10 목록을 보여주는 Ajax 호출용 REST web API 호출 화면

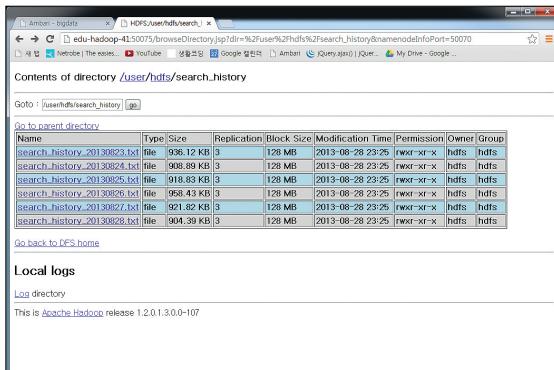


그림 4. MapReduce로 일단위 집계 요약한 결과

은 생각보다 간단했다. WordCount 코드를 이용해 검색어별 상품 ID를 카운트하면 됐다. 때문에 해당 작업은 익숙한 Java Word Count 프로그램을 변형해 개발했으며, 1일 단위로 특정 경로에 저장되도록 했다. 저장된 파일은 약속된 시간에 Python Batch에 의해 HBase에 그 결과만 누적되는 방식이다. 참고로 POC 프로젝트가 아닌 실제 프로젝트에서는 Python Batch가 아닌 Map-Reduce Batch 내에서 Reduce의 OutputFormater와 Hbase_OutputFormater를 사용하고, 이를 Oozie나 Azkaban 등 하둡 배치 집 전용 스케줄러로 관리할 것을 권장한다.

시각화 부분에 신경을 쓰다

마지막 주말과 주중에는 거의 매일 야간 작업을 했다. 다행스럽게도 모든 서버 프로그래밍은 완료 1주 전에 완성됐고, 마지막 UI 만을 남겨두고 있었다. UI는 간단한 듯 하지만 서버 프로그램과 달리, 투자 시간에 정비례해 결과가 드러나는 부분이다. 따라서 서버 프로그램은 '했다'와 '안 했다'가 명확하지만, UI는 잘했다 못했다가 주관적이라서 어느 정도에서 멈출지도 명확하지 않았다. 지금까지 수많은 난관을 뛰어넘으며 서버 프로그램까지 잘 해왔지만, 자칫 UI를 소홀히 하면 쇼핑몰팀의 작업 결과물이 초라하게 보일 수도 있었다.

쇼핑몰팀의 UI는 심플하면서도 다이나믹했다. 대전에서 원격으로 많은 작업을 한 김유성 팀원의 노고가 컸다. 김유성 팀원은 jQuery 의 Spark Line 차트를 이용해 실시간 수집되는 로그를 화면 깜빡

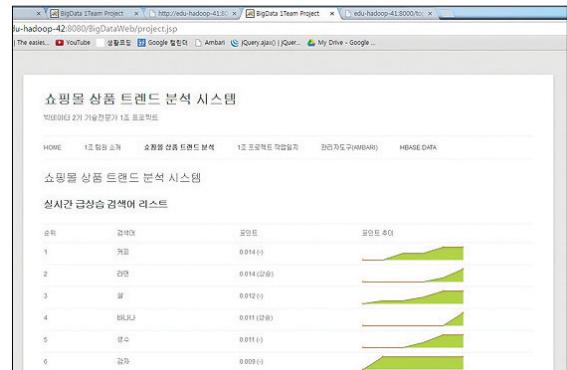


그림 5. 10초 단위로 페이지 리로딩 없이 가감을 보여주는 다이나믹 차트

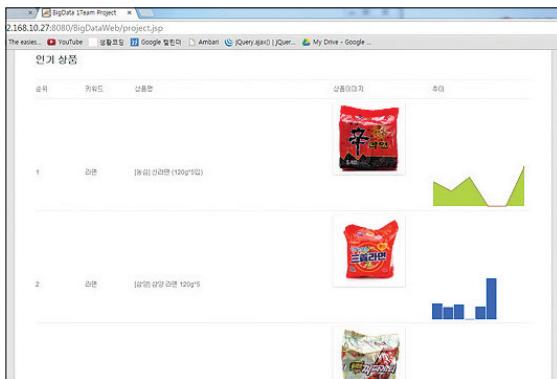


그림 6. 급상승 검색어 클릭 시 나타나는 장기간의 상품 선호도 트렌드



그림 7. 10초 단위로 보여주는 단기간의 검색어 선호 실시간 트렌드

임 없이 실시간 차트로 반영해 주었다. 키워드를 클릭하면, 긴 기간의 상품 인기 트렌드를 막대 그래프와 선 그래프로 심플하게 보여 줬다.

Conclusion

이론과 실제의 차이를 실감하다

무엇보다도 목표했던 바대로 마무리할 수 있어서 매우 기뻤다. 팀원들이 마음을 모아 맑은 업무를 잘 해주었고, 일부 팀원이 회사일로 공백이 생기면 나머지 팀원들이 해당 부분을 메워줘서 성공적으로 마무리할 수 있었다. 팀원 중 2명이 대전에서 일했기에 오프라인 모임을 자주 갖기 힘들었지만, ‘구글 드라이브’ 등 협업 도구를 이용하여 wiki 형태의 문서를 공유하며 작업했던 것이 많은 도움이 됐다.

쇼핑몰팀이 했던 프로젝트는 짧은 기간에 개념검증을 하는 성격이 강했다. 그러나 쇼핑몰팀 팀원들은 해당 프로젝트를 좀 더 발전시

켜서 실무에 적용 가능한 수준으로 올해 안에 보완 업그레이드 하기로 약속했다. 향후 보완할 부분은 다음과 같다.

1. HDFS 위에 NoSQL 레이어로 HBase를 뒀지만, 더 빠른 인메모리 클러스터 영역을 추가 레이어로 보충할 계획이다. 이를 위한 후보군은 MemChached와 Redis다.
2. 프로젝트에서는 배치 처리를 Cron으로 등록해 했다. 향후에는 Oozie로 좀 더 복잡한 스케줄 관리까지를 시도해볼 계획이다(2014년 3월 현재 팀장이 속한 빅데이터팀에서는 Oozie 스케줄러를 2~3개월 사용하다가 여러 가지 이유로 Azkaban 스케줄러로 교체해 현재 실제 운영환경에서 6개월째 사용 중이다).
3. Mahout과 R을 이용해 더 다양하고 정교한 분석 모델을 추가할 계획이다. 이로써 더 심층적인 아이템 트렌드 뷰 제공이 가능할 것으로 할 것으로 전망된다. 이렇게 되면 카테고리별, 개인화 그룹 성향별 개인화한 뷰도 제공할 수 있을 것이다.
4. 수료 프로젝트에서 수집 프로세스는 Flume NG를 이용했다. 좀 더 확장된 시스템에서는 Kafka, Camus, Avro 등을 적용할 계획이다. 또한 실시간 분석 레이어를 두어 Storm 등을 통한 실시간 질의도 도전해 볼 계획이다(2014년 3월 현재 팀장이 일하는 회사 빅데이터팀에서는 데이터 수집 및 데이터 이동 메커니즘을 Flume 대신 계획했던 kafka, camus, Avro를 사용하고 있다. Parquet File Format과 LZO, Snappy 압축을 함께 사용하고 있다).

무엇보다도 이론적으로 배웠던 부분을 프로젝트를 통해 실제 머릿 속에 정립하고, 실 프로젝트에서 나올 수 있는 많은 상황을 직접 접해볼 수 있어 많은 도움이 됐다. **BIG**



김훈동 팀장(신세계S.COM 빅데이터팀 과장)

“빅데이터 실무 프로젝트를 진행할 힘을 얻었습니다”

빅데이터 아카데미에서 얻은 도움은

교육 받은 시점이 공교롭게 회사에서 빅데이터 프로젝트에 착수를 앞둔 시점과 일치했기 때문에 교육과정 중에 가졌던 수료 프로젝트가 크게 도움이 됐어요. 빅데이터 아카데미에 등록하기 전에 관련 세미나와 책을 보면서 POC(개념 검증) 차원에서 PC 몇 대를 연결해 빅데이터 분석 플랫폼을 직접 구축해보기도 한 상태였고요. 저는 프로젝트 경험을 쌓기 위해 빅데이터 아카데미에 노크한 경우였습니다. 회사에서 외부 SI 업체에 의뢰하던 기존 IT 프로젝트와 달리, 빅데이터 시스템 구축 프로젝트는 내부 인력 중심으로 추진하기로 했기 때문입니다. 회사에서도 제가 빅데이터 구축 실무팀장을 맡고 있어서 처음 해보는 것이라 막연한 두려움 같은 게 있었어요. 하지만 빅데이터 아카데미에서 수료 프로젝트를 진행하면서 ‘할 수 있겠다’는 자

신감을 얻을 수 있었습니다. 무엇보다 저와 비슷한 일을 하는 많은 사람을 만나서 그들과 교류할 수 있었던 게 큰 도움이 됐습니다.

빅데이터 아카데미 수강 후 달라진 점이라면

데이터 분석에 더 흥미를 갖게 됐다는 점을 들 수 있습니다. 빅데이터는 스페셜리스트보다 제너럴리스트가 필요한 영역이라는 생각을 하게 됐어요. 프로그래밍 등 IT 지식에 수학, 통계학, 확률 등 전반적인 지식이 필요하더군요. 그래서 데이터 사이티스트가 되려면 빅데이터 아카데미에 개설된 기술 전문가 과정뿐 아니라 분석 전문가 과정도 함께 수강할 필요를 실감했습니다.

매우 열정적으로 수료 프로젝트에 참여했다고 들었다

팀장을 맡은 책임감 때문이지 싶습니다. 낮에는 회사에서 일하고 퇴근 이후와 주말을 주로 활용해 수료 프로젝트를 준비했습니다. 완료 3주를 앞두고는 아이들을 키즈 카페에 놀게 해 놓고 그 옆에서 준비하기도 했고요^^ 성공적으로 완료해야 한다는 생각으로 어려운 문제를 만날 때마다 ‘구글링’ 해가면서 그동안 모호하다고 여겼던 부분까지 이해할 수 있었습니다.

회사에서 하는 빅데이터 프로젝트를 소개해달라

신세계몰과 이마트를 상품을 믹스해 파는 ssg.com 사이트의 빅데이터 시스템 구축 사업입니다. 5년 정도의 로드맵 아래 총 6단계에 걸쳐 구축하는 장기 프로젝트입니다. 지난 1월 초에 1단계인 빅데이터 수집 인프라를 구축 완료해 오픈했습니다. 현재는 빅데이터 기술을 접목해 실시간 비정형 데이터를 분석해 상품이나 서비스를 추천하는 기능을 추가로 개발중입니다. 여기에 신세계그룹에서 운영하는 분스, 트레이더스, 신세계면세점 등의 온/오프라인 데이터를 통합해 빅데이터 기술을 접목해 추가적인 데이터 활용을 준비하고 있습니다.

빅데이터 프로젝트를 진행하면서 어려운 점이라면

소프트웨어는 보이지 않는 건물을 짓는 거라고들 하는데 빅데이터는 다른 어떤 IT 분야보다 보이지 않은 부분이 더 많은 영역이라는 생각합니다. 2000년대 초반을 전후에 유행했던 데이터 웨어하우스 방식으로 고가의 하드웨어와 상용 DBMS에서 4~5시간 걸려야 나올 분석 결과를 훨씬 저렴한 비용의 시스템에서 불과 몇십분 만에 도출했음에도 시각화에 신경을 쓰지 못하면 평가가 좋지 않더군요. 이 때 경험으로 데이터 분석 전문가들의 어려움을 이해할 수 있었고요. 데이터가 없으면 시각화가 불가능한데도 그 데이터보다 시각화에 더 관심을 두는 것을 보면서 약간 서운한 느낌마저 들었죠^^ 빅데이터팀에 시각화 전문가들을 참여시켜 현실을 수용하고 있습니다. 빅데이터 프로젝트를 할 때 자동화, 지능화, 머신러닝 등의 개념을 구현하더라도 시각화에 신경을 쓰지 않으면, 수고에 합당한 평가를 받기 어려울 수 있습니다. **BIG**

쇼핑몰 트렌드 실시간 분석 플랫폼



프로젝트 소개

Flume로 수집한 실시간 데이터를 HBase에 1차 적재하고, MapReduce로 HDFS에 하루 단위로 배치화해 2차 적재함으로써 긴 시계열의 대용량 트랜드 데이터와 트렌드 데이터를 동시에 실시간 분석 · 시각화한 시스템 구현

구분

기술 전문가 과정 | 쇼핑몰 분석 플랫폼



진행

기술 전문가 2기 쇼핑몰 실시간 빅데이터 분석 플랫폼 구축팀

• 김훈동 팀장	유통업체 IT부	컴퓨터공학 석사	경력 15년
• 김유성 팀원	연구소 전산실	산업경영학 석사	경력 15년
• 정동민	SI 업체 개발자	컴퓨터공학 학사	경력 12년
• 조용석	SI 업체 개발자	전자공학 학사	경력 15년
• 최승호	전문정보서비스기관	경영정보학 학사	경력 05년



지도

심탁길, 송주영



프로젝트 기간

2013년 7~8월



적용 도구

Hadoop(MapReduce), Flume, Hbase, Python, HappyBase, Thrift, Django, Zookeeper, Oozie, jQuery



산출물

- 쇼핑몰 실시간 분석 시스템 기획서 및 설계도
- 쇼핑몰 실시간 분석 모델 구현 엔진
- DBguide.net에 분석 프로젝트 진행경험 연재



교육참여 형태

자발적 참여 (4) / 회사 권유 (1)



빅데이터 아카데미 수강 후 달라진 점

- 수강 전 • 책이나 세미나에서 단편적으로 습득한 지식 보유 • 실무 경험이 없어 빅데이터에 대한 막연한 두려움
 수강 후 • 수료 프로젝트 과정에서 실제 프로젝트 중에 만날 문제점을 미리 파악 및 완료기간 단축 • 자신감과 필요한 지식 확보
 • 빅데이터 전문가 인적 네트워크 구축 및 커뮤니티를 통해 실무에서 만나는 각종 문제 해결

데이터 분석의 공든 탑을 쌓다 키워드 기반 트렌드 분석 플랫폼



하둡 에코시스템을 최대한 활용할 수 있는 주제로서
MapReduce에서 대용량 트위트 메시지 워드 카운트 프로
그래밍을 과제로 선정했다. HBase에 분석 · 저장된 키워드
데이터를 웹서버로 읽어와 실시간으로 그래프로 보여주기에
대한 레퍼런스가 부족해서 RESTful 방식으로 전달받은 파
라미터를 HBase와 연동해 조회된 데이터를 하나하나 확인
해가면서 구현해야 했다. NoSQL만 접해본 팀원들이 모여
구축한 키워드 기반 트렌드 분석 플랫폼은 어떤 모습일까?

Challenges

'빅데이터로 무엇을 할 것인가' 고민하다

교육을 받으며 알게 된 키워드 기반 트렌드 분석 플랫폼팀(이하 키
워드팀)의 팀원 6명이 모여 처음으로 할 수 있는 이야기는 그리 많
지 않았다. IT 분야에서 일하는 사람이라면 더 공감하겠지만, 회사
프로젝트에 참여중인 상태에서 6주 간의 수료 프로젝트까지 겹치는
상황이라면 참으로 난감할 수밖에 없다.

결국 키워드팀은 '빅데이터의 가장 본질적인 것과 자신의 업무에
지장을 초래하지 않는 수준에서 수행이 가능하고, 다른 프로젝트로
쉽게 전환할 수 있어야 한다'는 기준을 정해 놓고서야 주제를 찾을

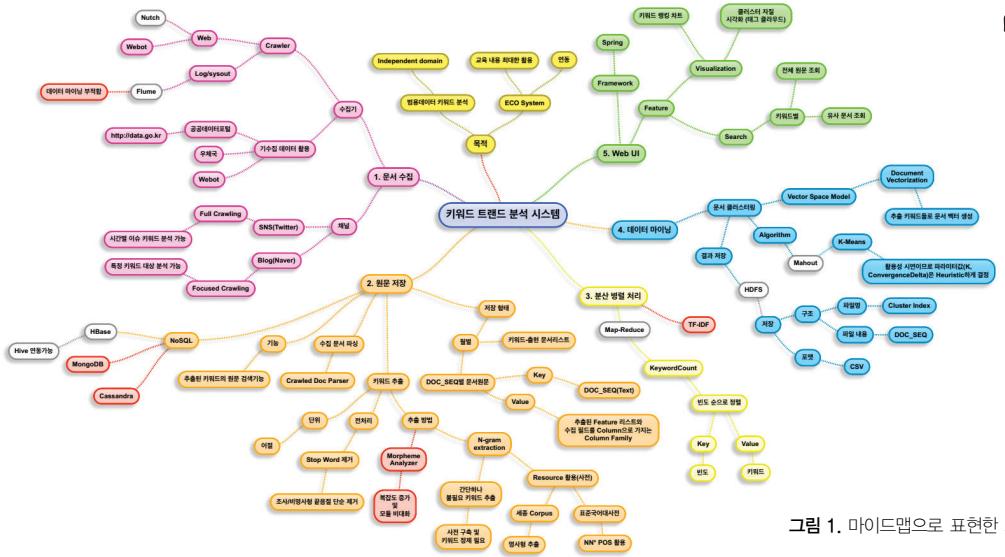


그림 1. 마인드맵으로 표현한 프로젝트 시스템의 구조

수 있었다.

집체교육 과정에서 배운 내용을 검증해보는 측면에서 기본에 가장 충실했고, 하둡(Hadoop) 에코시스템을 최대한 활용할 수 있는 주제로서 MapReduce 워드 카운트(word count) 프로그래밍을 생각해냈다. 배치(batch) 분석에 시간을 추가하면 키워드 분석으로 트렌드 파악 시스템을 만들 수 있을 것이라 생각했다. 처음에는 MapReduce와 HBase, mahout까지 모두 적용해 볼 계획이었다. 결과적으로 mahout 클러스터링 환경에서 하려 했던 데이터 마이닝은 제한된 시간 때문에 해보지 못한 채 MapReduce와 HBase에서 분석한 결과를 웹에서 시각화해 보여주는 것으로 마무리 지었다. 하지만 빅데이터에 대한 이해와 더 나아가 실무 프로젝트까지 진행할 수 있겠다는 자신감을 얻은 후회 없는 프로젝트였다.

여러 차례 회의를 거쳐 이용훈 팀원이 마인드맵으로 개발 시스템 구조를 그려내면서 구체화 됐다(그림 1 참고). 목적 설정에서 시작해 데이터 수집 > 데이터 저장 > 분산 병렬처리 > 데이터 마이닝 > 웹 UI 개발 단계로 개념검증 시스템을 구축하기로 했다.

Solution

개발 콘셉트를 정하자 속도가 불타

이전 기수들이 수료 프로젝트 기간이 짧아 어려움을 겪었다는 소식을 들었던 터라 키워드팀은 집체교육 기간 중에 프로젝트에 착수했

다. 팀원 중에 2명이 대구와 나주에서 왔기에 가능하면 집체교육 기간 중에 서로 호흡을 맞춰볼 필요가 있었다.

스키마 디자인, UI 설계, WBS(Work Breakdown Structure) 작성 등을 빠른 시간에 했다. 팀원 모두 회사일로 빠듯한 것은 마찬가지였으므로 점심시간과 쉬는 시간을 집중 활용했다. 개발 콘셉트를 잡고부터 속도를 더 낼 수 있었다. 처음에는 개발을 맡은 팀원들이 구현해온 시스템을 전 팀원이 검토하면서 피드백하는 형태로 진행했다.

플랫폼을 구축하고 데이터를 수집하다

프로젝트 수행을 위한 컴퓨팅 환경은 VM 기반의 4대의 클러스터로 구성됐다. 여기에 Hadoop, HBase, Zookeeper 등 하둡 에코 시스템과 자바 프레임워크 등을 구성했다. 키워드 분석 플랫폼이므로 정형 데이터가 아닌 비정형 데이터를 수집해 분석에 적용하기로 했다. 트위터 API를 얻기 위해 미리 확보해둔 트위터 ID로 5GB 상당의 한글 트위트 메시지를 수집해 분석 대상 데이터로 활용했다. 수집 데이터는 SCD(Structured Crawl Data) 파일 포맷으로 저장했다.

모든 트위트 메시지를 시간과 날짜로 구분해 가져왔다. MapReduce의 특성을 검증하는 과제이므로 실시간 처리가 아닌 배치(Batch) 처리가 적용됐다. 데이터 수집의 간격을 줄이면 실시간성에 가까워지므로 이 부분은 추후에 시도해볼 숙제로 남겨뒀다.

데이터에서 키워드를 뽑는 건 개발자들에게 그리 낯선 일은 아니다. MapReduce 기반의 워드 카운팅은 빅데이터 분야의 ‘Hello World!’라고 할 만큼 매우 기본적인 작업이다. 따라서 하둡이라는 말만 들어도 울렁증이 생기는, 아직 빅데이터의 실체를 경험하지 못한 사람들에게 충실향한 레퍼런스를 제공하겠다는 생각으로 하나씩 구현해 나갔다.

갖고 있던 생각을 바꿔야 할 때가 오다

개발을 담당한 이용훈 팀원과 이재호 팀원은 데이터 저장 목적으로 잠깐 NoSQL을 접해본 것 외에는 NoSQL 경험이 없는 상태였다. 교육 중에 실습용 데이터도 하둡 파일 시스템(HDFS)에서 갖다 사용한 터라 HBase에 적응하기까지 더 시간이 걸렸다.

시간대별로 수집한 데이터를 HBase에 어떻게 저장할 것인지와 MapReduce와 유기적으로 연동하는 문제, 향후 웹서비스까지 어떻게 수용할 것인지를 놓고 결정을 내려야 했다. 익숙한 RDB를 사용하지 않고 전 과정을 NoSQL로 하기로 한 이상 뒤로 물러설 수 없었다.

참고로 HBase를 비롯한 NoSQL은 칼럼을 유연하게 추가할 수 있는 것이 장점이다. 가변적 칼럼을 추가함으로써 HBase에서 시간 대별 키워드 저장 문제는 해결됐다.

MapReduce에서 워드 카운팅하기 전에 N-Gram Analyzer로 어절에서 명사를 추출하는 과정이 필요하다. 어절에서 명사를 추출하는 N-Gram Analyzer 프로그래밍은 대학원에서 자연어처리를 전공했던 이용훈 팀원이 개발을 맡았다. 갖고 있던 ‘개체사전’과 매핑하는 형태로 어절에서 명사를 추출했다. 이 명사들에서, 이재호 팀원이 MapReduce 키워드 통계 모듈로 시계열 키워드 정보와 빈도 정보를 추출해 Hbase에 저장함으로써 키워드 분석 플랫폼의 큰 틀은 개발이 끝났다. 개발을 진행했던 이용훈 팀원과 이재호 팀원은 오픈에스엔에스에서 함께 근무해서 짧은 수료 프로젝트 일정을 더 효율적으로 이용할 수 있었다.

제한된 일정 때문에 당초 하려 했던 마이닝은 하지 않고 시각화에 들어갔다. 수집한 원문에 대해 클러스터링 과정을 거치면 더 다양한 서비스를 개발할 수 있다. 예를 들어, 취미 · 종교 · 스포츠 등으로 카테고리를 분류해 놓으면, 무작위로 키워드를 추출했을 때 제

공할 수 없는 분야별 키워드 변화 등 훨씬 다양한 서비스를 제공할 수 있다.

이 과정에서 빅데이터 분석은 서비스 관점에서 접근하기보다 데이터 관점에서 접근해야 함을 실감했다. 예컨대 기존 RDB 환경에서 개발은 입력 결과에 대한 출력이라는 서비스 관점에서 접근하지만, NoSQL은 여기에 분석이라는 개념을 추가해야 하므로 관점의 변화가 필요했다.

더불어 하둡 애플리케이션을 구성하기까지는 생각보다 복잡한 과정을 거쳐야 했다. 그렇다고 시도 못할 정도로 복잡하지는 않았다. 어느 정도 현장 개발 경험을 가진 개발자라면 충분히 해쳐나갈 수 있을 정도다.

‘고통 없이는 얻는 것도 없다’

키워드팀이 기술 전문가 과정 4기의 수료 프로젝트에서 1위를 차지하는 데 시각화 부분이 크게 작용했다는 후일담을 들었다. N2M 소속의 황지선 팀원이 담당한 시각화는 HBase에 분석 · 저장된 데이터를 웹서버로 읽어와 실시간으로 그래프로 보여주는 작업이었다.

키워드 분석팀은 시작부터 시각화까지 완료하는 것을 목표로 HBase에서 데이터 설계 시 향후 웹서버와 연동을 고려했다. 이런 과정을 거쳤음에도 HBase에서 데이터를 읽어와 웹으로 시각화하는 부분은 쉽지 않았다. 일단 구글링을 해도 HBase와 웹을 연동한 시각화에 대한 자세한 레퍼런스를 찾을 수 없었고, 그나마 찾은 영문 자료도 개발에 적용할 수 없을 정도였다.

특히 DB 테이블 변경 및 데이터 조회를 할 때 어려움이 컸다. 별다른 방법이 없었으므로 WebUI에서 RESTful 방식으로 전달받은 파라미터를 HBase와 연동해 조회된 데이터를 하나하나 확인해가면서 처리해야 했다. 시간과 노력으로 해결한 셈이다. 시각화 레퍼런스가 풍부한 NoSQL을 선택했다면, 피해나갈 수 있었을지 모르지만, ‘고통 없이는 얻는 것도 없다’는 말처럼 이런 어려움이 HBase 라이브러리를 더 잘 이해하는 계기가 됐음에는 틀림 없다.

특히 HBase는 그동안 익숙한 RDB 형태의 쿼리가 아닌 것이 개발 담당자들을 난처하게 했다. select 문을 사용하는 기존 쿼리와 달리, 테이블명을 파라미터로 받아 테이블에 연결해 데이터를 조회하는 형태다. 참고로 Pig를 HBase와 연동하면 내부적으로 쿼리를

생성해 처리할 수 있지만, 이해를 목적으로 한 수료 프로젝트에서 단순히 결과만을 위해 자동화 도구를 이용할 필요까지 없다고 판단했다.

HBase에 분석·저장된 월별 키워드 순위 데이터를 불러와 Java Collection 라이브러리로 소트해 월별 키워드 순위를 메인 화면에 출력하고, 순위별로 출력된 키워드를 클릭하면 월별 키워드 카운트를 그래프로 보여주는 형태로 구현했다.

또한 검색 키워드 클릭 시 팝업 화면에 해당 키워드가 들어간 원문 및 태그 클라우드를 보여주도록 구현했다(그림 3). <그림 3> 좌측 상단의 ‘연월 선택’ UI 구현에도 많은 시간이 소요됐다. ‘Easy UI 프레임워크’를 사용한 것인데 달력 팝업을 축소하는 데도 적지 않은 시간과 노력이 들어갔다.

HBase의 웹 지원 성능은 프로젝트 경험을 기준으로 놓고 볼 때 그 리 권장할 정도로 나오지 않았다. 커넥션 타임이 기존 RDB에 비해 더 소요되는 느낌이었다. MongoDB나 Cassandra 같은 다른 NoSQL로 구현하면 어떤 결과를 보일지 추후에 확인해볼 계획이다.

Conclusion

아이디어를 받아들일 때 좋은 결과가 나온다

교육이 완료되고 프로젝트 평가 결과 1등을 차지했다는 이야기를 들었을 때 믿기지 않았다. 주변 팀에서 화려한 주제를 선정해 열정적으로 준비하고 있음을 잘 알고 있었기에 쟁쟁한 그 팀들을 제치

고 1등을 차지할 거라고는 생각하지 못했다.

당초 수료 프로젝트를 개시하기 전부터 팀원 간 원활한 소통을 강조했던 것이 적중하지 않았나 싶다. ‘성공 프로젝트와 실패 프로젝트는 참가자들 상호간 이해 정도의 수준이 좌우한다’는 말을 여러 차례 들었던 터라 빅데이터 수료 프로젝트에 적용했는데 역시나 통했다. 팀원 모두 회사 업무로 바빴지만, 빅데이터 기술 전문가 교육 과정에서 체험한 새로운 기술을 배울 때의 설렘과 함께 참여한 사람들의 태도 등이 큰 힘을 발휘하게 했다.

다른 팀의 프로젝트가 특정 목적을 염두에 두고 수행됐다면, 키워드 팀이 구축한 시스템은 특정 시스템이나 분석을 지원하는 플랫폼이라는 게 특징이다. 계획했던 플랫폼이 어느 정도 윤곽이 드러나면서 나중에 함께 상용화 해보자는 의견까지 나왔다. 향후 키워드 팀이 구축한 시스템을 보완해야 할 부분 몇 가지가 있다.

수료 프로젝트에서는 자동으로 시시각각 변하는 시계열 키워드 분석 기능까지 구현하지 않고 배치 작업으로 검증하는 선에서 만족했지만, 향후 스케줄러를 추가하면 자동화 서비스도 가능해진다.

수료 프로젝트를 진행하면서 내가 알고 있는 지식보다 수천 배 많은 아이디어와 생각이 있고, 자신이 그것을 받아들일 때 좋은 결과가 나온다는 것을 실감했다. 구성 팀원들이 얼마나 우수한지, 얼마나 열정이 있는지가 프로젝트의 성패를 좌우함을 다시 확인하게 해준 소중한 프로젝트였다. **BIG**

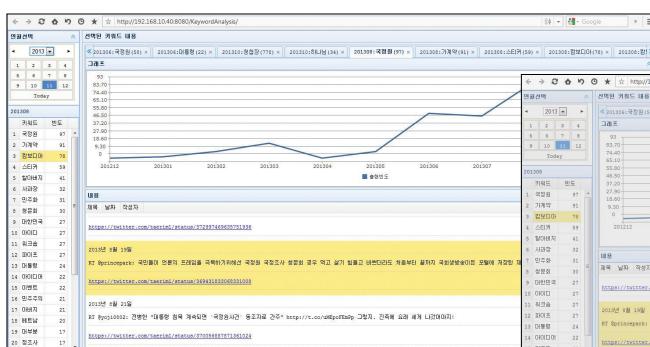
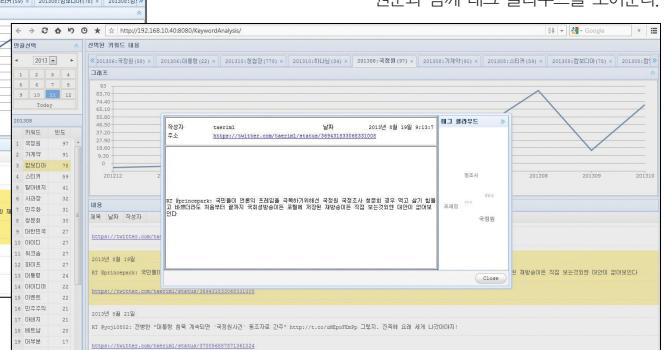


그림 2. 월 시각화 전체 화면

그림 3. 월별 키워드 조회 화면에서 특정 키워드를 선택하면 원문과 함께 태그 클라우드를 보여준다.





진현철 팀장(세종아이에스 부설연구소장)

“알게 되면 사랑하게 된다”

빅데이터 아카데미는 어떤 점에서 도움이 됐나

실무 지식뿐 아니라 빅데이터 관련 자료를 훨씬 쉽게 확보할 수 있게 됐다는 점입니다. 제가 일하는 회사에서 빅데이터 개발과제를 기획하기 위해 많은 자료를 구했지만, 체계화할 방법을 찾지 못해 어려움을 겪고 있었어요. 비슷한 목적으로 참여한 동기생들이 많아 서로 도움을 주고 받으면서 자료 체계화라는 제 바람도 이뤄졌습니다. ‘빅데이터를 제대로 공부해보겠다’는 마음이 통했는지 수료 프로젝트를 할 때도 서로 마음이 잘 맞았습니다.

지방에서 참여해서 어려움이 더 커질 거 같다

매일 열차로 출퇴근하면서 집체교육을 받았는데, 교육 장까지 왕복 4시간이 걸리더군요. 그날 배운 내용을 검토하고 더 공부해야 할 사항이나 질문할 내용을 메모하는 데 활용했기에 기차로 오가는 시간이 전혀 아

깝지 않습니다. 지방에서 출퇴근하면서 (빅데이터 아카데미에) 참석하는 분들이 많고, 디들 잘 적응하고 있기에 혹시라도 망설이고 있다면 일단 저질러 보면 원가 되지 않을까 합니다. 더불어 지방 도시에서도 빅데이터 아카데미 교육이 꼭 이뤄졌으면 하는 바람입니다.

빅데이터 아카데미 교육 수료 후 바뀐 점이라면

데이터에 대해 눈을 떴다고나 할까요? 개발을 담당했던 오픈에스엔에스의 이용훈·이재호 팀원, N2M의 황지선 팀원이 고생을 많이 했습니다. 수료 프로젝트 완료 후 서로 격려하는 자리를 가졌는데, 하나 같이 데이터 분석이 무엇인지를 알고 부터 데이터들을 더 관심을 갖고 보게 되었다고 하더군요. ‘뭐든지 일고 보면 좋지 않은 게 없다’는 말을 실감한 거지요. 개인정보의 소중함도 알게 됐고요. 이 생각이 일반 SI 프로젝트에서도 적용되더군요. ‘어떻게 설계하면 나중에 데이터 분석을 하는 데 도움이 되고, 활용도를 높일 수 있을까’를 염두에 두고 접근하고 있습니다. 이것이 변화라면 변화이고 가장 큰 소득이지 않나 싶습니다. 더불어 약간의 도움만 받으면, 실무 빅데이터 프로젝트도 직접 수행할 수 있겠다는 자신감도 얻었습니다. ‘호랑이를 잡으려면 호랑이굴로 들어가라’는 말처럼, 일단 한번 뛰어 들어가 볼 필요가 있습니다.

교육 과정 중에 인상적이었던 일은

수료 프로젝트를 함께했던 팀원들이 가장 기억에 남습니다. 목표를 잊지 않았기에 곳곳에서 불거진 문제들을 무사히 헤쳐나갈 수 있었습니다. 하다 보니 재미도 있었고요^^ 덕분에 당초 기대하지 못했던 기술 전문가 4기 수료 프로젝트 발표에서 1위라는 영광을 선물로 받았습니다. 실무는 팀원들이 대부분 했기에 저는 그저 무언가 해야 할지 방향만을 정해줬던 거 같습니다.

회사에 준비중인 빅데이터 사업을 소개한다면

세종아이에스는 철강산업 ERP 개발, 서버 공급과 운영을 전문으로 하는 기업이에요. 이 사업을 여러 해 동안 하다 보니 철강산업과 관련된 데이터를 많이 축적하게 됐어요. 자체 IDC센터를 운영하면서 ASP 형태로 ERP 서비스를 제공하고 있어서 ERP 데이터의 흐름을 한 눈에 볼 수 있습니다. ‘사물인터넷 개념 등을 수용해 데이터 분석 기반의 철강 수요 예측과 안전 관리 시스템 구축’을 목표로 준비 중입니다. **BIG**

키워드 기반 트렌드 분석 플랫폼



프로젝트 소개

하둡 에코시스템을 최대한 접해볼 수 있는 MapReduce 기반의 키워드 카운트 플랫폼을 주제로 선정했다. 트위터에서 수집한 비정형 데이터를 배치 처리해 얻은 키워드 데이터에 시간 개념을 추가해 시간 흐름에 따른 트렌드 변화를 파악할 수 있도록 구현했다.

구분

기술 전문가 과정 | 키워드 분석 플랫폼



진행

기술 전문가 과정 4기 키워드 분석 플랫폼팀

• 진현철 팀장	SI 업체 연구소장	컴퓨터공학 박사수료	경력 17년
• 유일선 팀원	SI 업체 개발자	전자정보통신공학 학사	경력 10년
• 이용훈	SI 업체 개발자	자연어처리 석사	경력 04년
• 이재호	SI 업체 개발자	컴퓨터공학 학사	경력 04년
• 황지선	SI 업체 개발자(N2M)	전산계산학 학사	경력 18년
• 황지선	공공기관 연구원	정보보호공학 석사	경력 10년



지도

심탁길



프로젝트 기간

2013년 11~12월



적용 도구

Hadoop, MapReduce, Hbase, Zookeeper, jQuery



산출물

- 키워드 분석 시스템 기획서 및 설계도
- 키워드 분석모델 구현 엔진
- HBase 기반의 웹 시각화 레퍼런스



교육참여 형태

자발적 참여 (4) / 회사 권유 (2)



빅데이터 아카데미 수강 후 달라진 점

수강 전 • 책이나 세미나에서 단편적으로 습득한 지식 보유 • 빅데이터에 대한 막연한 두려움

수강 후 • 자신감과 필요한 지식 파악 • 빅데이터 전문가 인적 네트워크 구축

- SI 프로젝트에서 분석을 고려한 데이터 설계 가능

빅데이터 아카데미에는 배움의 설렘과 성장의 기쁨이 있습니다



- 1 · 2 – 데이터 처리와 분석을 집중적으로 배우는 집체교육
- 3 · 4 – 실습을 통해 실무 지식을 습득 / 수료 프로젝트 발표
- 5 – 수료식
- 6 – 집체교육 중에 팀을 구성해 진행되는 수료 프로젝트
- 7 – 성공적인 수료 프로젝트를 위해 틈날 때마다 의견 나누고 확인하기
- 8 · 9 – 두근두근… 2개월 간의 교육의 결과를 평가 받는 수료 프로젝트 발표장





현재를 읽고 미래를 창조하라!

Big Data, Bigger Opportunities And The Biggest Value!



빅데이터 시장은 2015년에만 세계적으로 18조 2000억원, 한국은 약 300억원 규모로 형성 전망



빅데이터 전문가 수요 급증과 경력자 부족 등으로 전문인력 수급 난항, 사회적 비용 증가 예상



미국은 '빅데이터 R&D 이니셔티브'를 발표하고 빅데이터 인력양성 등에 집중 투자, 한국의 빅데이터 인력양성 체계는 시작 수준



국내에서도 '신직업 발굴 · 육성 추진방안' 발표, 빅데이터 전문가를 신직업으로 정의, 신규 일자리와 고부가가치 창출 도모



미래부 등 관계부처 합동으로 데이터 전문인력 양성과 일자리 연계를 골자로 한 '빅데이터 산업 발전전략' 발표



빅데이터 전문가 양성 체계 수립 및 실무 전문가 양성을 통해 국내 빅데이터 기술 경쟁력을 강화하고 글로벌 시장 선점의 토대 마련 필요



미래의 분석 전문가 여러분을 적극 환영합니다

■ 교육 대상

과정	내용
기술 전문가	<ul style="list-style-type: none">• 대상 : 개발자, DBA, SE 등 3년 이상 업무 경력자• 선발 우대조건<ul style="list-style-type: none">- 빅데이터 프로젝트 수행인력 또는 예정인력- 하둡(MapReduce, HDFS), NoSQL 및 캐싱기술 유경험자- 분산 파일시스템 또는 분산 데이터베이스 관리 유경험자
분석 전문가	<ul style="list-style-type: none">• 대상: CRM, 마케팅 및 기획 등 3년 이상 업무 경력자• 선발 우대조건<ul style="list-style-type: none">- 빅데이터 프로젝트 수행인력 또는 예정인력- SAS, SPSS, R 등 통계분석 툴, 데이터 마이닝 유경험자- SQL, OLAP, Query, Reporting 도구 유경험자

■ 참여 방법(연수생 선발 절차)

구분	내용
수강신청	http://bigdata.dbguide.net에서 원하는 교육과정을 신청
1차 선발	직무 및 업무경력 적합성 평가
2차 선발	프로젝트 경력 및 상세기술 온라인 질의 평가
최종 선발	2차 선발인원에 한하여 재직증명서 및 협약서 등 서류 제출



2013년 빅데이터 아카데미 연수생 현황

- 평균 연령 37.9세, 평균 경력 10.1년으로 산업분야에 대한 깊은 이해와 지식(Domain Knowledge)을 보유한 전문가를 대상으로 연수 진행
- 중소기업 재직자 위주(69.8%)로 연수생을 선발했으며, 개발 · 분석직무 등 현업에서 직접 데이터를 다루는 실무 전문가를 선발 · 교육

구 분	빅데이터 기술 전문가		빅데이터 분석 전문가		합 계
수료 인원	102명		100명		202명
연수생 연령	평균 37.8세		평균 38세		평균 37.9세
업무 경력	평균 10.6년		평균 9.6년		평균 10.1년
기업 분포	대기업	14.7%(15명)	대기업	30%(30명)	22.3%(45명)
	중소기업	76.5%(78명)	중소기업	63%(63명)	69.8%(141명)
	프리랜서	8.8%(9명)	프리랜서	7%(7명)	7.9%(16명)
직무 분포	개발 업무	57.8%(59명)	데이터 분석	47%(47명)	-
	DBA	32.4%(33명)	마케팅 기획	21%(21명)	-
	컨설팅	9.8%(10명)	컨설팅	32%(32명)	-

2013 우수 프로젝트 사례집

발행일 | 2014년 4월

발행처



미래창조과학부

427-140 경기도 괴천시 관문로 47, 4동
www.msip.go.kr



110-799 서울시 종로구 종로 51 종로타워 19층
www.kodb.or.kr

편집 · 디자인 | 글불크리에이티브
제작 | 친프로미디어