**데이터 마이닝 사례**

**<페이스북 대규모 감정 전이 가설>**

IME 201402308 이상인

1. 페이스북의 수익모델과 활성 사용자

21세기에 들어오면서 급격히 증가하는 인터넷 시장은 물리적인 면에서 컴퓨터를 넘어 모바일 기기까지 확대하고 있다. 이 시장에서는 자신의 제품이나 서비스를 제공 할 사이트가 필요한데 이 사이트를 유지하는데 비용이 들고 보통 트래픽 수 대비 돈을 받기 때문에 사람들이 많이 접속하는 사이트에서는 타 사이트보다 훨씬 더 많은 비용이 요구된다. 사이트 유지 비용을 벌기 위해서 사람들은 수많은 방법을 생각하였고 그 중 하나가 광고를 다는 것이다. 광고는 사이트에서 클릭이나 노출되는 수만큼 해당 광고를 올린 사이트 관리자에게 돈을 주며 사이트 관리자들은 대부분 여기에서 벌어들이는 수익으로 관리비 및 이익을 창출한다. 실제로 페이스북이 공개한 2014년 4분기 실적을 확인하면 전체 수익 중 95.5%, 약 56억 달러가 광고수입임을 확인할 수 있다. 이 광고수입을 늘리기 위해 사이트 관리자들은 사이트를 적극적으로 이용하는 활성 사용자들이 좀 더 많이 광고에 흥미를 느끼고 클릭하게 유도하도록 각종 방법을 생각하였는데 그 중 하나가 사용자 맞춤 광고 노출이다. 이 전략을 이용하기 위해서는 사용자, 특히 계속 활동을 하고 있는 활성 사용자들을 유지해야 했고, 이 활성사용자들이 많을수록 표적 광고의 정확성을 높일 수 있다. 결과적으로 활성 사용자가 늘수록 사업의 수익이 는다는 것을 안 페이스북은 활성 사용자들을 늘리고 유지하기 위해 여러 활동과 연구를 진행하였다.

과거에 광고의 주 타겟층을 특정하지 못했던 이유는 개개인의 충분한 데이터가 없었기 때문이었다. 하지만 시간이 갈수록 많은 사용자들이 오랜 시간 동안 인터넷을 이용하면서 개인들이 생성해놓는 데이터 양이 기하급수적으로 불어났다. 특히 페이스북의 경우 친구들이 많으면 많을수록 자신의 뉴스피드에 친구들이 좋아요나 공유를 한 게시물이 많이 나타나는데 현실적으로 모두 올릴 수가 없어 페이스북 개발자들은 개인 계정에 올라오는 뉴스피드를 그 개인이 관심 가지는 게시물 위주로 올라오도록 알고리즘을 제작하였다. 불어나는 데이터를 통해 페이스북은 효과적인 데이터 마이닝을 이용하여 뉴스피드 알고리즘을 개선하였고 다음의 본인이 소개할 뉴스피드 알고리즘을 뒷받침 하는 데이터 마이닝 결과물 중 하나인 ‘사회관계망을 통한 대규모 감정 전이’ 사례에 대해 소개하고자 한다.

1. 사례 소개

페이스북 데이터사이언스 팀의 애덤 크래이머를 포함한 3명의 데이터사이언티스트들은 페이스북 사용자들이 본인의 뉴스피드에 긍정적, 혹은 부정적 단어에 노출되어 있을 경우 어떠한 영향을 받는지 알아보고자 하였다. SNS 중 전세계 사람들이 대다수가 사용하는 페이스북을 대상으로 하였는데 페이스북의 모든 사용자들을 조사하기에는 모집단이 크므로 긍정, 부정적 단어에 노출되었던 경험이 있는 사용자 중 68만 9003명을 랜덤 추출하여 표본으로 설정하였다. 실험은 2012년 1월 11일부터 18일까지 일주일간 진행되었고 노출되는 뉴스피드가 생성된 날짜도 실험날짜와 동일하게 설정하였다. 긍정적인 단어의 빈도수, 부정적인 단어의 빈도수를 각각 독립변수로 설정하여 이에 따라 개개인 각각의 반응을 볼 수 있는 타임라인(본인만 쓸 수 있음)의 게시물들의 몇 가지 요소를 기준으로 두어 종속변수로 볼 것을 가정하였다.

1. 데이터베이스와 분산처리

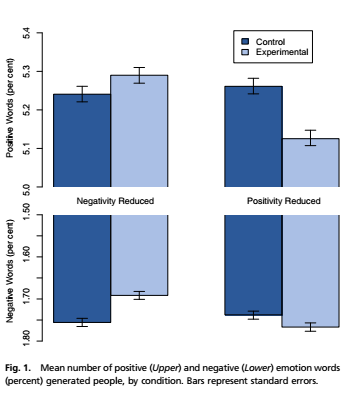
독립변수와 종속변수간의 연관성을 파악하기 전에 표본이 되는 데이터를 저장해야 하였다. 실험은 일주일간이었지만 약 69만명에 이르는 사람들 개인당 가지고 있는 일주일 간의 뉴스피드 글들은 매우 큰 원시데이터였다. 데이터 크기가 커서 한 컴퓨터에 저장할 수 없었고 저장이 가능 하더라도 분석과정에서 큰 데이터를 한번에 분석처리하기엔 시간과 비용 면에서 비효율적이기 때문에 분석가들은 자바 기반 오픈 소스 프레임워크인 하둡의 분산시스템인 HDFS(Hadoop Distributed File System)을 이용하였다. HDFS에 저장함으로써 대용량 데이터를 각각 128MB씩 분산된 서버에 저장하였고 이 저장된 데이터를 빠르게 처리할 수 있게 하였다. (하둡1.0까지는 64MB씩 분할하였으나 하둡2.0에서는 128MB까지 확장하였다.) 또한 긍정, 부정적인 단어가 없을 때 결측치(missing value)가 발생하고 실시간으로 데이터가 변화하지 않기 때문에 무결성이나 트랜젝션의 영향을 받지 않는 알맞은 저장소였다.

독립변수에 대해서 어떠한 단어가 긍정적인 단어인지, 어떠한 단어가 부정적인 단어인지 또 각각 빈도수가 얼마나 되는지 알아야 했다. 이를 위해 Linguistic Inquiry and Word Count software(LIWC2007) word counting system을 이용하였다. LIWC2007은 개인의 다양한 심리학적 변인을 탐지할 수 있어 감정 〮 정서적 과정, 사회적 과정, 인지적 과정 등을 연구하는데 쓰이는 시스템이었는데 본 분석에서는 감정 〮 정서적 과정을 위해 사용되었다. 하위 변인으로는 긍정적 정서, 긍정적 느낌, 낙천성 또는 활동성과 관련된 단어들을 ‘긍정’이라고 정의하였고 부정적 정서, 불안, 화, 슬픔과 관련된 단어들을 ‘부정’이라고 정의하였다. 빈도 수를 세기 위해서는 LIWC에서 맵리듀스(Mapreduce)를 사용하였다. Mapreduce는 Map 다음 Reduce 단계를 거치게 되는데 Map에서 각 단어들을 key값으로 변환시키고 reduce 클래스에서 같거나 성격이 유사한 (혹은 사용자 설정에 따라 같다고 정의한) key값들을 그룹화한 후 집계하였다. Mapreduce 자체가 여러 노드들을 대상으로 일괄 병렬 처리가 가능하기 때문에 HDFS과 같이 대용량 메타데이터를 파티션 별로 나누어 저장된 데이터들을 처리하는데 알맞은 방법이었다.

앞에서와 같이 데이터를 저장하고 분산 처리한 결과 총 300만개의 글들이 분석되었고 2억 2200만개의 단어가 검색되었으며 이중 긍정적인 단어가 4백만개(3.6%), 부정적인 단어가 180만개(1.6%)임을 알 수 있었다.

1. 관련성 분석

앞선 분산처리로 두 독립변수를 구하였다. 이 독립변수들로 인해 영향을 받는 종속변수들이 실제로 관련이 있는 것인지 알아보기 위해 관련성 분석을 진행하였다. 종속변수들의 기준은 활성 사용자들의 확실한 변화를 알 수 있는, 본인들의 타임라인을 몇 가지 기준으로 나타내었다. 독립변수와 종속변수 두 변수 사이의 연관성을 알아보는 것이었으므로 장바구니 분석을 이용하였다 .그 결과 두 가지 종속변수가 연관이 있는 것을 발견하였다.

첫 번째 변수는 본인의 타임라인의 긍정적 단어 빈도수의 변화이다. 뉴스피드의 긍정적 단어의 빈도수와 타임라인의 긍정적 단어의 빈도수를 분산형 그래프로 나타낸 결과 두 개의 군집을 확인할 수 있었는데 뉴스피드의 긍정적 단어가 많을수록 타임라인도 긍정적 단어가 많음을 확인할 수 있었다.

두 번째 변수는 타임라인의 부정적 단어 빈도수의 변화이다. 뉴스피드의 부정적 단어의 빈도수와 타임라인의 부정적 단어의 빈도수를 분산형 그래프로 나타낸 결과 두 개의 군집을 확인하였는데 뉴스피드의 부정적 단어 수가 낮으면 타임라인의 부정적 단어의 수도 낮음을 확인하였다.

긍정독립변수-긍정종속변수, 부정독립변수-부정종속변수 간의 관계가 있는가 하면 긍정독립변수-부정종속변수, 부정독립변수-긍정종속변수의 관계도 확인할 수 있다.

뉴스피드의 긍정적 단어와 타임라인의 부정적 단어를 그래프로 나타내면 앞에서와는 반대인 군집을 확인할 수 있다. 즉, 뉴스피드 긍정적 단어가 많으면 타임라인의 부정적 단어가 적음을 확인하였고 뉴스피드의 부정적 단어가 많으면 타임라인의 긍정적 단어가 많음을 알 수 있었다. 뉴스피드의 부정적 단어와 타임라인의 긍정적 단어와의 관계에서도 동일한 효과를 볼 수 있었다.

통제집단과 실험집단과의 관계에서도 확인할 수 있듯이 뉴스피드의 긍정적, 부정적 단어가 뉴스피드의 긍정적, 부정적 단어 사용에 영향을 줌을 알 수 있었다.

1. 윤리적인 문제

윤리적인 문제란 실험에서 단순히 생명의 생사를 가르는 여부만을 의미하지 않는다. 여기에서 말하는 윤리적인 문제는 사람들의 감정을 조작하여 데이터를 수집한 부분에서 발생한다. 이 분석은 데이터 수집을 제외한 나머지에서는 어떠한 윤리적인 문제도 발견되지 않았다. 하지만 전반적인 진행 단계에서 처음에 시작하는 데이터 수집에서 문제가 발생하면 나머지도 그 문제를 안고 간다는 것은 자명한 사실이다.

이 분석 당시 분석가들은 보다 정확한 데이터를 수집하기 위해 실험 대상이 된 약 69만명의 사람들의 뉴스피드 알고리즘을 임의로 수정하였다. 긍정, 부정과 관련된 단어 전체 중 10%~90%에 해당하는 만큼 무작위로 삭제되는 알고리즘을 추가하였다. 그 결과 15만 5천명의 사람들의 뉴스피드에는 긍정적인 단어가 포함된 글들이 많이 올라왔었고 다른 15만 5천명의 사람들의 뉴스피드에는 부정적인 단어가 포함된 글들이 대다수가 되어 긍정적인 콘텐츠를 읽은 사용자들은 긍정적 콘텐츠를 올리고, 부정적 글을 읽은 사용자들은 부정적인 글을 자신의 타임라인에 올릴 가능성이 커진다는 사실이 드러났다.

페이스북측에서는 “서비스 개선을 위해 사용자의 데이터가 사용될 수 있다”라는 약관에 사용자들이 동의를 하여 문제가 없다고 설명하였지만 데이터가 사용될 수 있다는 거지 조작할 수 있다는 것은 아니다. 이에 한 미국 대기업의 리서치 연구원은 “페이스북은 이용자들이 행복해지길 원한다.”라고 하였다. 만약 이 연구결과가 반영되어 페이스북측에서 자신들의 수익을 위해 행복하고 긍정적인 글들만 올라가게 알고리즘을 만든다면 우린 어디까지 이 것을 알아챌 수 있을까라고 심각하게 생각해봐야 한다.

이러한 논란이 심각하다고 뒷받침하는 다른 근거는 이 데이터 수집 시기 자체(2012년 1월 11일 – 2012년 1월 18일)에도 있다. 1월 13일에는 지중해에서는 크루즈선이 침몰하여 약 30명이 사망하였으며 15일에는 유타 주지사가 후보에서 사퇴하였고 18일에는 수백 개의 인터넷 사이트가 SOPA와 PIPA를 반대하는 블랫아웃 운동이 있었다. 페이스북은 내가 좋아요 하는 게시물 위주로 계속 업데이트가 되므로 위의 사례들이 잘 나오지 않고 긍정적인 콘텐츠만 나오면 사회에 참여하고 관심을 가지게 되는 기회가 차단되는 것이다. 우리가 행복한 콘텐츠만 본다고 해서 우리의 주관으로 행복하게 되는 건지 다시 한번 생각해 볼 필요가 있다.

또한 신문 같은 경우에는 안보면 그만이고 구독을 끊으면 그만이다. 하지만 페이스북 같은 경우에는 요즘 SNS계에서 1인자 자리를 공고히 하고 있듯이 경쟁사가 거의 전무 하다시피 하므로 이러한 윤리적인 문제에 조심하고 또 조심해야 할 필요성이 있다.

1. 결론 및 요약

페이스북 같은 여러 SNS에서는 자신의 수익모델의 대부분이 광고비용이므로 SNS계의 주 목적 중 하나가 광고비용을 늘리는 것이라고 해도 과언이 아니다. 보통 누르거나 노출시켜야 효과가 있는 광고이므로 페이스북에서는 사용자 맞춤 표적 광고를 만들기를 희망하였다. 과거에는 사용자가 다른 물건, 사람과 형성하는 관계에 대해서 부족했을지 몰라도 지금과 같이 데이터의 홍수를 겪고 있는 이 시대에서는 충분히 사용자 맞춤 광고를 만들 수 있다. 이에 페이스북 측은 이 광고를 보고 눌러줄 활성 사용자들의 이탈이 일어나면 표적 광고의 정확성이 떨어져 결과적으로 수익이 떨어지므로 이를 방지해야 한다고 생각하였다. 그래서 뉴스피드 알고리즘의 긍정, 부정 단어를 검색하여 무엇이 어떠한 관계가 있는지 밝히고자 하였다. 먼저 HDFS에 원시데이터를 나누어 저장하고 맵리듀스를 이용하여 분산되어 있는 데이터를 한번에 병렬 처리하여 효율적으로 데이터를 뽑아내어 실제 독립변수의 빈도 수를 확인하였다. 다음에 데이터 마이닝의 군집 분석을 이용하여 독립변수와 종속변수 사이를 분석해본 결과

첫째, 뉴스피드에 좋은 글이 올라오면 사용자도 좋은 글을 올리고 그 반대면 사용자도 부정적인 글을 올린다.

둘째, 뉴스피드에 긍정적 단어가 포함된 글이 올라오면 타임라인에 부정적인 단어가 포함된 글이 줄어들고 그 반대도 똑같이 성립한다.

위 두가지를 알 수 있었다. 또한 이번 실험의 윤리적인 문제가 드러나면서 우리들 스스로도 ‘파블로프의 개’가 아닌 SNS데이터 자체를 조심스럽고 비판적으로 바라봐야 할 필요성을 알 수 있었고 앞으로 분석을 할 경우에도 이렇게 개인에게 얼마나 개입할 수 있는지 생각해 볼 수 있는 교훈을 얻을 수 있었다.