

한이음 공모전 2017

개 발 보 고 서

2017. 9. 2

프로젝트명	국문	모두의 통계청
	영문	Every Statistics Office
작 품 명	Alba Sherpa	
신 청 자	한신대학교 / 김석준	

요 약 본

팀 정보

팀 명	자비스			
팀 원	이 름	소 속	부서/학과	직위/학년
멘 토	백송이	우리에프아이에스	기반인프라부	과장
지도교수	김성기	한신대학교	컴퓨터공학부	
멘티 1(팀장)	김석준	한신대학교	수학과	4학년
멘티 2	장근호	한신대학교	수학과	4학년
멘티 3	김예지	한신대학교	수리금융학과	4학년
멘티 4				
멘티 5				



작품 정보		
프로젝트명	국문	모두의 통계청
	영문	Every Statistics Office
작품명	Alba Sherpa(알바 셸파)	
작품 소개	Alba Sherpa란 셸파(Sherpa)가 등산가들의 짐을 덜어주는 도우미라는 뜻이 있다. 즉 아르바이트의 짐을 덜어주겠다는 뜻으로 아르바이트 데이터를 데이터 마이닝 기법으로 수집, 분석, 시각화하고 사용자들이 한 눈에 볼 수 있는 정보를 제공하여 자신의 환경에 맞는 아르바이트를 선택할 수 있도록 도움을 주는 프로그램이다.	
작품 구성도		
작품의 개발배경 및 필요성	모든 데이터를 평균값으로만 분석하여 보여주거나 선형적(직선)으로만 분석해야 할까? 모든 데이터를 직선으로 분석하는 것이 아닌 데이터를 분석할 때 그 데이터의 분포를 미리 추측하고 그 데이터에 가장 적합한 함수모형을 선정해야 한다 보았고 분석해볼 종목을 아르바이트로 선정하여 그에 알맞은 함수모형을 생각해보며 적용시켰다.	
작품의 특징점	분석기법으로 원래 존재하는지는 모르겠지만 '극값 데이터 마이닝'이란 기법을 개발하였고 단순히 '선형회귀 모형'으로만 분석하는 것이 아닌 2차함수 모형, 단조 증가하는 3차 함수 모형으로 분석을 실시해보며 가장 적합한 모형으로 시각화를 시켜준다. 또한 위의 분석기법들을 체험해 볼 수 있는 분석기가 존재하며 데이터 크롤링 기법을 이용하여 데이터를 대량으로 수집하였다.	
작품 기능	<ul style="list-style-type: none"> * 알바 데이터 회귀 분석 기능 : 두 개의 속성을 선정하여 두 속성간의 변화를 크롤링을 통한 데이터들과 사용자가 입력한 데이터를 이용하여 분석한다. * 댓글을 이용한 정보 공유: 댓글창을 구현하여 종목에 따른 사용자들 간의 다양한 의견을 자유롭게 소통이 가능하며 분석 결과로는 얻을 수 없는 정보들을 얻게 된다. * 알바 종목 랭킹 : Top N으로 자신이 우선순위를 두고 싶은 속성을 누르면 그 속성으로 정렬이 된다. 예로 시급이 높은 순 등으로 정렬할 수 있다. * 데이터 분석기 : 분석기법을 체험해 볼 수 있는 분석기가 존재한다. 자신의 매월 수익을 입력하여 다음 달에 얼마의 수익이 있을지 예측이 가능하다. 또한 추가로 극값 데이터 마이닝 기법을 제공한다. 	
작품의 기대효과 및 활용분야	선형회귀가 아닌 함수모형 회귀분석으로 더욱 다양한 분석이 가능해지며, 아르바이트에 대한 여러 정보들을 제공해주어 자신에게 알맞은 아르바이트를 선택할 수 있게 도와준다. 구직사이트 광고효과 및 통계, 빅데이터 분야 교육적 효과도 가져온다.	

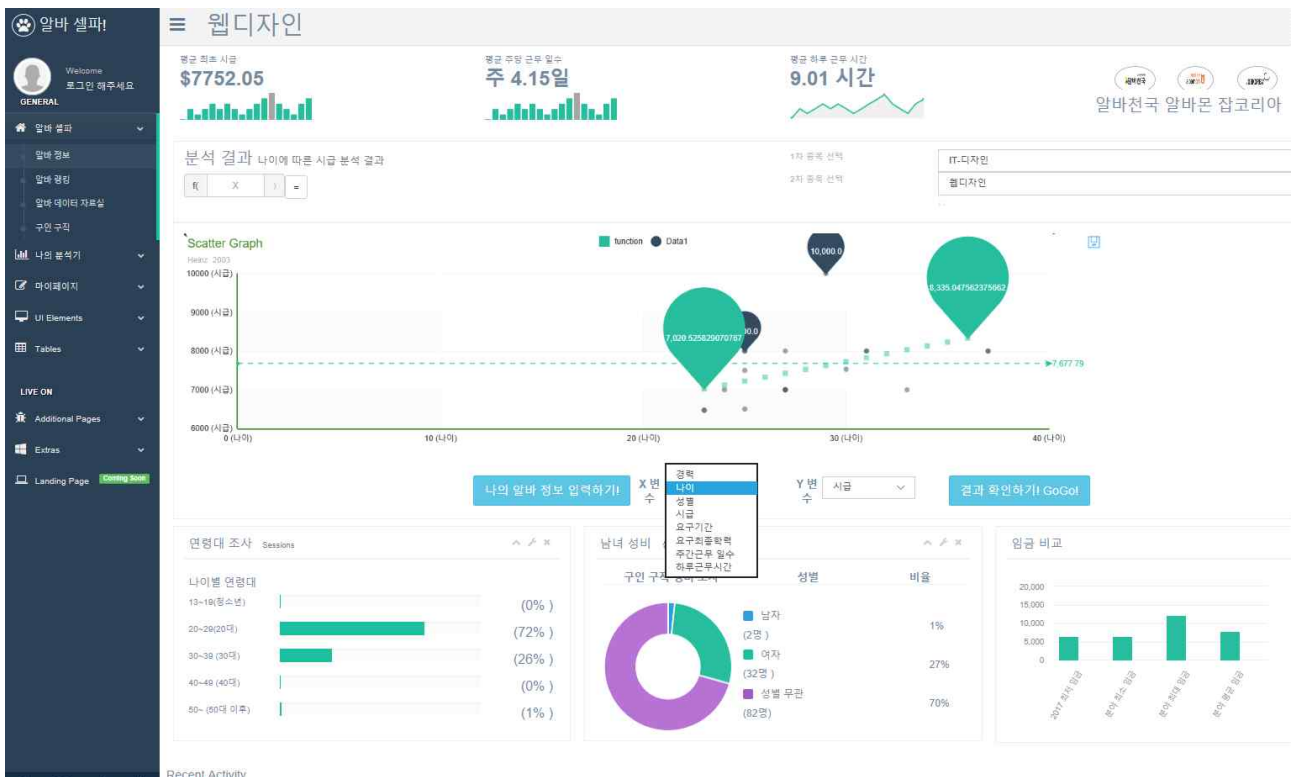
본 문

I. 작품 개요

※ 평가항목 : 기획력 (필요성, 차별성)

1. 작품 소개

- Alba Sherpa (아르바이트 도우미)
- 여러 가지 알바 분석 자료를 토대로 자신에게 맞는 아르바이트를 선택할 수 있도록 도움을 주는 프로젝트이다.
- 데이터 마이닝(크롤링)과 사용자의 입력을 이용하여 데이터를 수집함.
- 회귀분석을 이용해 다양한 변수에 대해 상관관계를 파악할 수가 있음.
- 분야별로의 댓글을 통해 사용자들 간의 정보공유 및 의사 결정에 도움을 줌
- 개발한 알고리즘을 통해 임의의 자신의 데이터를 넣어보며 분석해 볼 수 있는 분석기 존재



< 그림 1. Alba Sherpa >

2. 작품의 개발 배경 및 필요성

- 청년들의 금전적 문제
 - 최근 대부분의 대학생들의 경우 아르바이트를 하지 않으면 부모님에게 손을 벌리지 않는 이상 대학교 생활을 해나가기 어려운 상태이다. 어떤 경우는 학비를 벌기 위하여 대학생생활을 아르바이트로 시간을 보내게 되는 경우도 있어 대학생생활에 지장이 가여 필요성이 제기된다.
- 객관적인 정보의 필요성
 - 대부분의 대학교 신입생들은 어떤 알바가 자신의 성격과 환경에 맞을지 생각 안하고 아르바이트를 하게 된다.
 - 웹상에 퍼져있는 소량의 정보들은 가공되어 있지 않아 특정한 근무지의 영향을 많이 받게 된다. 따라서 객관성이 부족하여 신뢰도가 떨어진다.
- 아르바이트별 고충 및 노하우
 - 어떤 업무 던 간에 사회 초년생들이 처음 접하는 일은 다소 두렵게 느껴지거나 익숙하지 않을 수 있다. 예로 학원 아르바이트의 경우 환경을 접해보면서 자신의 재능을 이용 및 개발할 수 있는 반면 ‘처음’ 접하는 입장에서는 매우 부담스럽게 느낄 수 있기 때문이다. 따라서 생각보다 할 만하다 던지 노하우를 조언을 해줄 수 있는 사람이 존재하지 않다면 접해보기 어려운 경우가 많다.
- 아르바이트의 환경근무에 따른 다양한 상관관계 도출
 - 상관관계란 ‘하루에 일하는 시간이 많을수록 시급을 더욱 많이 받을까?’와 같은 개념이며 통계학과, 빅데이터 전문가가 아니더라도 결과를 통해서 정보를 알 수 있다. 예시) 실제로 나의 경험 상 ‘학원 강사’의 경우는 ‘경력’에 따라 ‘시급’이 급격하게 오르기도 함. 이런 결과를 보고 아르바이트를 선택할 때 자신의 재능에 맞춰 고른다면 매우 도움이 됨.
- 개인 분석기를 통한 자신의 데이터 분석
 - 웹상에서 제공하는 데이터 분석기가 존재한다. 이를 활용하면 자신이 추후에 얼마의 시급을 받을지 등을 추측 해 볼 수 있고 다양한 분석방식을 통해 그에 맞게 의사결정을 할 수도 있게 된다.
- 데이터 분포에 가장 적합한 함수 선정의 필요성 제시
 - 모든 데이터를 평균값으로만 분석하거나 선형적으로만 분석해야 할까? 모든 데이터를 직선으로 분석하는 것이 아닌 데이터를 분석할 때 그 데이터의 분포를 미리 추측하고 그 데이터에 가장 적합한 함수모형을 선정해야 한다는 의견.

3. 작품의 특징 및 장점

- 웹 크롤링

- 데이터들을 대량으로 수집하기 위해 알바 천국 (www.alba.co.kr/) 사이트에 구인 구직중인 정보들을 웹 크롤링 방법을 실시하여 데이터를 추출 하였고 실제로 최근의 정보를 수집함으로써 신뢰도를 높인다. 100,000개가량의 데이터가 수집되었다.

- 다양한 분석 기법

- 분석 기법 중 Linear Regression 문제를 모티브로 분석 알고리즘을 설계하였고 Linear Regression을 변형하여 2차 다항 함수 모형, 3차 다항 함수 중 ‘단조 증가’ 하는 모형으로 변형하여 데이터들의 분포에 가장 적합한 모형을 자동으로 선정 되게끔 설계하였다. 장점으로는 기존에 다른 어플리케이션들은 단순히 key값에 따른 value값의 ‘평균값’ 으로 만 분석했다면 여기서는 추가로 기계학습을 통한 분석이 이루어진다.

- 분석 모형의 선정 (단순 회귀분석 , 비선형 회귀분석)

- 어떤 두 개의 변수간의 상관관계를 분석하게 될 때 함수의 모형을 선정하여 분석해야 하는데 예를 들면 ‘경력에 따른 시급분석’ 의 경우 경력에 따라 시급이 줄어들고 있는 구간이 생긴다면 신뢰도가 매우 떨어지게 된다. (경력에 따라 시급이 계속적으로 증가 또는 그대로(단조 증가)유지되어야 한다.) 따라서 각 변수간의 관계를 추측하고 함수모형을 선정하고 있어 신뢰성을 높인다.

- 데이터 분석기 및 극값 데이터 마이닝 설계

- 위에서 설명하는 분석기법과 함께 극값을 이용한 데이터 분석 기법을 추가로 설계하였고 자신이 임의의 데이터를 입력하여 분석해볼 수 있는 분석기를 제공한다.

- 익명 댓글을 통한 사용자들 간의 정보공유

- 유사 제품으로는 (<https://www.wanted.co.kr/salary>) 원티드의 연봉추정 방식이 있는데 이 제품은 ‘댓글기능’ 이 제공되고 있지 않다. 정보를 필요로 웹서핑을 하는 경우 그 내용의 댓글을 보면서 신뢰를 갖거나 글이 잘못되었다는 것을 판단하는 경우가 많고 통계적 기법이 아닌 자연어로 작성된 정보는 매우 유용하게 쓰일 수 있다.

- Caching 테이블을 이용한 내부 처리

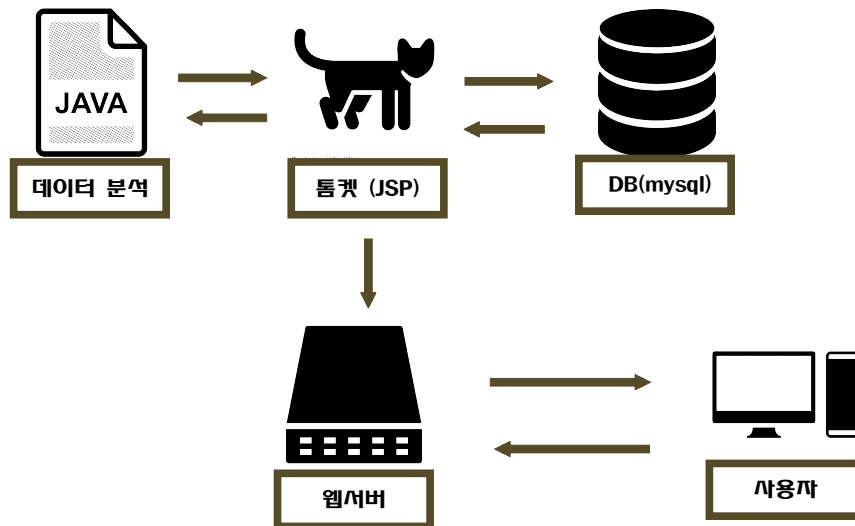
- 대량의 데이터를 사용자들이 이용할 때마다 매번 분석하게 되면 매우 비효율적이므로 caching(영속화 방식)을 이용하여 서버 측에서 주기적으로 분석기를 실행하고 그 결과를 caching 테이블에 담으면 그 ‘분석 결과’ 만 사용자들에게 시각화 하는 방식을 이용하여 빠른 분석 결과를 제공한다. 분석기로는 HeartbeatMessage클래스(Java Application과의 통신)를 이용해 Thread 로 사용자의 데이터를 받아내 체크하여 분석하고 결과를 저장하여 사용자가 들어올 때 테이블을 Reduce하여 하나의 함수 클래스로 변형하여 사용자들에게 보여준다. 이는 최초 1회만 실시하여 서버과부화의 부담을 매우 덜어준다.

II. 작품 내용

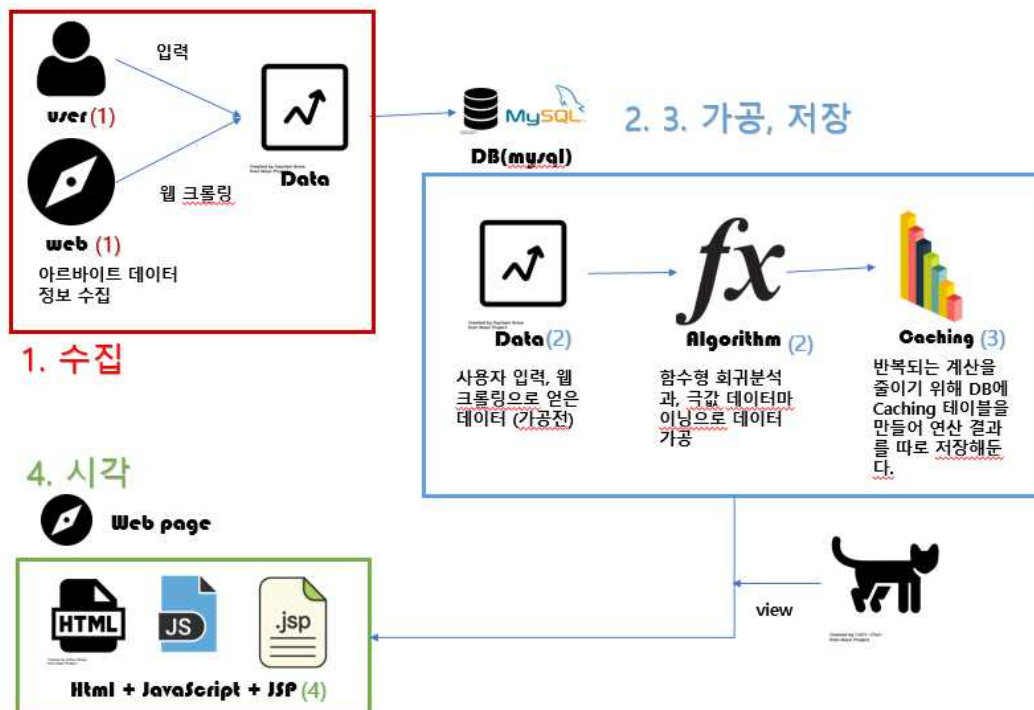
※ 평가항목 : 기술력 (기능구체성, 난이도, 완성도)

1. 작품 구성도

○ 구성도



< 그림 2. 시스템 구성도 >



< 그림 3. 시스템 구조>


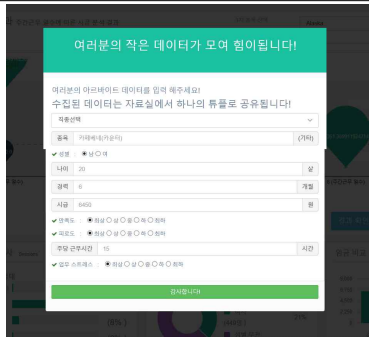
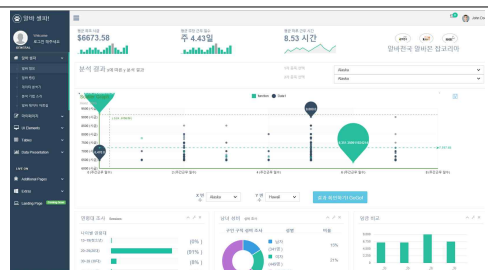
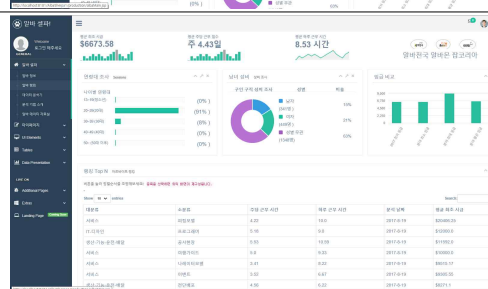
2. 작품 기능


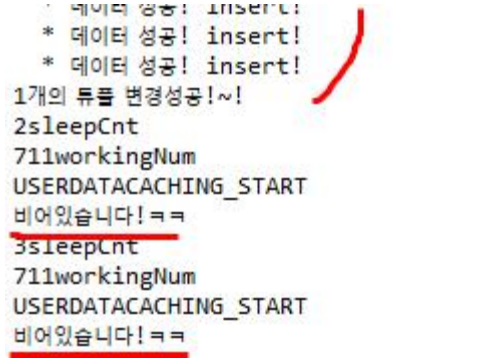
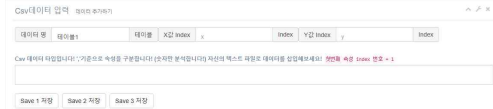

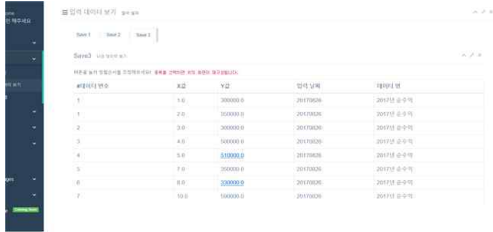
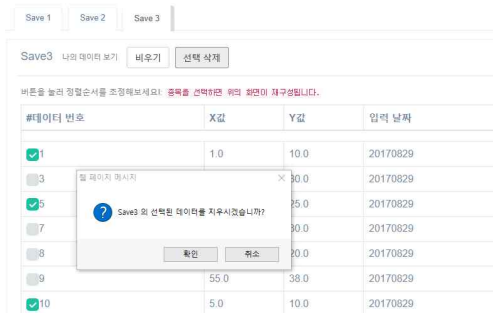
2-1. 전체 기능 목록

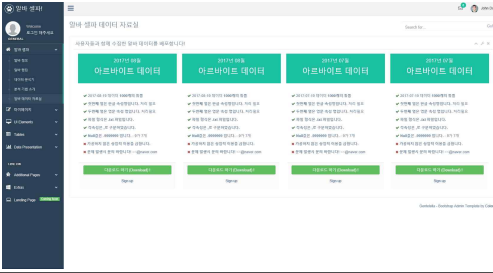
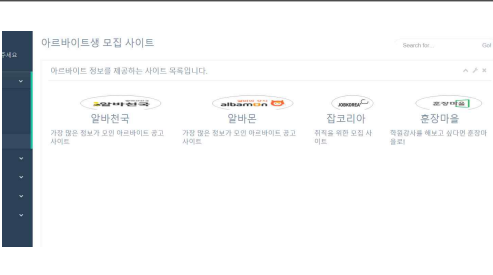
구분		기능	설명	현재진척도(%)
S/W	회원 기능	회원가입	회원가입할 수 있는 기능을 제공함	100%
		로그인	아이디/비밀번호를 이용하여 로그인 할 수 있는 기능을 제공함.(시저암호 도입)	100%
		비밀번호 찾기	비밀번호를 모를 때 회원정보를 이용하여 비밀번호를 변경할 수 있는 기능을 제공함.	100%
		프로필	로그인 후 자신의 회원정보를 수정할 수 있는 기능을 제공함.	100%
	1)수집	자신의 알바 데이터를 입력하는 기능	알바 데이터를 수집하기 위해 입력 폼을 만들어 사용자들이 자신의 데이터를 입력 할 수 있도록 구축한다.	100%
		데이터 웹 크롤링	알바 천국 구인구직 페이지에서 데이터를 추출해내는 추출기 개발	100%
	2)가공 및 분석	회귀분석 데이터 분석	점과 선을 이용하여 분석결과 함수를 시각화 한다.	100%
		평균값을 이용한 데이터 분석	평균 값 ($E[X]$ 기댓값)을 통해 굳이 데이터 마이닝을 할 필요가 없을 경우 막대 그래프로 시각화 한다.	100%
		자동으로 데이터 크롤링 및 caching 테이블 세팅	sleep메소드로 주기적으로 데이터를 크롤링하고 caching테이블을 자동으로 세팅한다.	100%
	3) 시각화	회귀분석 데이터 분석 시각화 (점과 선 그래프)	점과 선을 이용하여 분석결과 함수를 시각화 한다.	100%
		평균값을 이용한 데이터 분석 시각화 (막대 그래프)	평균 값 ($E[X]$ 기댓값)을 통해 굳이 데이터 마이닝을 할 필요가 없을 경우 막대 그래프로 시각화 한다.	100%
		리스트를 이용한 Top N 분석 및 시각화	핵심적인 사항인 꿀 알바와 극한 알바를 찾기 위해 자신이 우선순위를 두고 가중치를 두어 정렬하여 Top N을 보여준다.	100%
		값 추출	분석된 함수에 x값을 대입하여 y값을 추출하여 볼 수가 있다.	100%
		댓글 기능	사용자들끼리 자연어로 정보공유 한다.	100%
		알바 데이터의 변수를 선정하는 기능	경력에 따른 시급뿐만이 아닌 다양한 변수에 관해 관계를 추출해 낸다.	100%
	1) 데이터 분석기 삽입	자신의 분석할 데이터 삽입 (회원전용)	데이터 분석기에서 돌려봤던 데이터를 저장하고 다시 써 볼 수 있도록 회원가입 자들에게 제공한다.	100%
		텍스트로 ,를 기준으로 데이터를 삽입하는 기능	빅데이터 또는 통계학과 학생들이 자주 사용하는 ,로 구분되는 텍스트형식으로 개인 분석기에 데이터를 삽입 가능하다.	100%

	1) 분석기 데이터 수정 및 삭제	자신의 데이터 테이블을 볼 수 있는 테이블 뷰	전에 입력했던 데이터를 전부 볼 수 있는 페이지가 존재한다.	100%
		자신의 데이터를 비우거나 데이터를 선택하여 부분삭제할 수 있는 기능	자신의 데이터를 볼 수 있는 페이지에 가서 전체비우거나 선택으로 삭제할 수 있다. 이때 부분 삭제하였을 경우 남은 데이터로 분석을 다시 하여 결과를 다시 저장한다.	100%
	2) 분석기 분석	Thread를 이용한 HeartBeatMessage 클래스를 이용하여 개인 데이터 분석시 서버 과부하를 막기 위한 방식 설계 (Java Application과의 통신)	사용자들의 데이터 분석기를 구현할 때 새로고침 할 때마다 분석기를 실행하는 것이 아니라 데이터를 보내면 최초 1회 분석을 시작하여 그 결과를 테이블에 저장하고 사용자들이 들어왔을 때 테이블을 Reduce하여 하나의 함수 결과 클래스로 만들어낸다.	100%
		위의 함수 형 회귀분석과 극값을 이용한 데이터 분석기법 개발 및 구현	분석 기법을 응용하여 극값의 패턴을 파악하고 다음 극값이 언제 일어날지 찾아내는 분석기법을 개발한다.	100%
	3) 분석기 결과 시각화	회귀분석,극값데이터 마이닝 분석결과 시각화	점과 선을 이용하여 분석결과 함수를 시각화 한다.	100%
		분석기 값 추출	분석된 함수에 x값을 대입하여 y값을 추출하여 볼 수가 있다.	100%
	기타	알바셀파 및 분석 기법 소개글 페이지	프로그램에서 사용하는 분석기법들을 설명해주는 페이지이다.	100%
		알바 구직 사이트 링크 페이지	알바천국, 알바몬 ,훈장마을 등의 여러종류의 사이트를 링크를 통해 접근할 수 있게 된다.	100%
		수집된 아르바이트 데이터를 공유하는 자료실	크롤링과 사용자들에게 얻어진 알바데이터 정보를 txt파일로 ';' 속성 구분을 두어 사용자들에게 1개월 간격으로 저장하여 배포하게 된다.	100%

2-2. S/W 주요 기능

기능	설명	작품실물사진
데이터 크롤링을 이용한 데이터 수집	알바천국 웹사이트의 정형화된 정보를 자바프로그램으로 크롤링을 실시한다. 현재 1000000개 가량의 데이터가 수집되어있다.	
데이터 입력폼을 통한 데이터 수집	사용자들에게 데이터를 수집하게 되는 입력폼이 존재한다. 이때 어느정도 기각 범위를 설정하여 잘못된 데이터라고 판단된다면 기각 테이블로 따로 저장된다.	
주기적으로 크롤링을 실시와 caching테이블을 자동으로 세팅	자동으로 크롤링을 실시할수 있도록 Thread.sleep메소드를 사용하여 서버 컴퓨터를 틀어놓으면 크롤링 및 caching테이블을 세팅해준다. caching테이블은 데이터 양이 많으므로 매번 사용자들이 들어올 때마다 분석하는 것이 아닌 분석결과를 caching테이블을 이용하여 알고리즘 계산을 하고 결과를 저장하여 테이블을 함수로 바꾸어 결과만 출력해주는 방식을 사용한다.	<pre>HeartbeatMessage heartbeat= new HeartbeatMessage(workingNum); //쓰로드 heartbeatMessage 1초마다 사용자의 데이터를 계산해야 하는지 물어본 Thread heartbeatMessage = new Thread(heartbeat); heartbeatMessage.start(); while(true){ try { Thread.sleep(1000); } catch (InterruptedException e) { // TODO Auto-generated catch block e.printStackTrace(); } if(workingNum[0]==HeartbeatMessage.CRAWLER_START_NUMBER){ System.out.println("CRAWLER_START"); mainServer.crawlerServer.dataCrawler(startNum,getData); startNum = getData; sleepCnt++; workingNum[0] = 0; } if(workingNum[1]==HeartbeatMessage.CACHING_START_NUMBER){ System.out.println("CACHING_START"); mainServer.cachingServer.cachingStart(lowestTimeMoney); sleepCnt=0; workingNum[1] = 0; } if(workingNum[2]==HeartbeatMessage.USERDATACACHING_START_NUMBER){ System.out.println("USERDATACACHING_START"); mainServer.userDataCachingServer.userDataCachingStart(); workingNum[2] = 0; } }</pre>
알바데이터를 이용한 회귀, 평균 분석결과 시각화	다양한 변수에 대해 다양한 함수 형태로 분석하고, 사용자의 정보를 입력받아 분석결과와 비교해준다. 다양한 변수를 선택하여 시각화되어 두 변수간의 상관관계 또한 도출해낼 수 있다.	
원하는 항목으로 정렬해 시각화하는 TOP N 리스트 뷰	아르바이트별로 우선순위를 두고 정렬하여 자신이 원하는 순으로 알바정보를 볼 수 있다. 예를 들면 시급 순으로 정렬하게 되면 시급이 높은 순으로 정렬이 가능하다.	

<p>사용자들이 체험해볼수 있는 데이터 분석기</p>	<p>LinearRegression 문제를 토대로 데이터 마이닝 기법을 학술적으로 분석하여 선형 함수모형, 2차함수모형, 3차함수 모형으로 각각 분석해보는 '함수형 데이터마이 닝'과 '극값을 이용한 데이터마이닝 기법'을 설계 하였다. 자신이 데이터를 직접 입력하여 분석 해 볼 수가 있다.</p>	
<p>HeartbeatMessage 클래스를 이용한(Thread) 서버 과부하 처리</p>	<p>웹페이지에서 작동하는 분석기를 설 계할 때 서버 과부하 문제를 막기 위 해 사용자가 데이터를 보내면 자바 어플리케이션인 HeartbeatMessage 클래스가 1초 간격으로 체크하며 대 기하다가 사용자에게 데이터가 들어 오면 최초 1회 알고리즘 계산을 하고 결과를 테이블에 저장하여 테이블 Reduce를 통해 함수로 바꾸어 결과 만 출력해주는 방식을 사용한다.</p>	
<p>사용자의 데이터를 텍스트로 구분자()로 데이터를 삽입해주는 기능</p>	<p>통계학과 학생들 또는 빅데이터를 공부하는 학생은 보통 ,로 구분하는 파일을 가지고 있다. 따라서 속성을 ,로 구분하여 자신의 테이블에 삽입할수 있는 기능을 갖추었다.</p>	
<p>값 추출</p>	<p>테이블에 저장된 결과를 Reduce하고 분석 결과 함수로 변형하여 $f(x)$의 x값에 숫자를 대입하면 자바스크립트로 값을 추출할 수 있다.</p>	
<p>나의 데이터 삽입 및 삽입된 데이터 시각화</p>	<p>최대 3개의 저장테이블을 사용할 수 있으며 그것을 테이블로 시각화한다. 예시로 자신의 '매달 순이익' 같은 데이터를 삽입하여 다음 달에는 얼마의 수익을 받을지 추측 가능하다</p>	
<p>나의 데이터 비우기 및 선택하여 삭제하기</p>	<p>자신의 데이터를 전체 비우거나 선택하여 삭제할 수 있다. 이때 선택 삭제하였을 경우 남은 데이터로 분석 을 다시 하여 결과를 다시 저장한다.</p>	

<p>사용자들에게 수집된 데이터 파일 공유</p>	<p>크롤링과 사용자들의 입력으로 모여 진 데이터를 txt파일로 저장하고 ,로 구분자로 두어 데이터를 사용해야하 는 사람들에게 제공한다. 사용 예로는 통계학과 학생들이 있다.</p>	
<p>아르바이트 구인구직사이트 소개 및 링크연결</p>	<p>아르바이트 구인구직 사이트를 소개하고 각 사이트별로 특징이 작성되어 있다. 예로 '학원강사' 아르바이트를 해보려면 훈장마을사이트를 통해야 잘 구할 수 있다.</p>	

3. 주요 적용 기술

수집	데이터 크롤링	웹 상에 퍼져있는 데이터를 일정한 기준을 두고 텍스트 형태로 추출하여 모으는 것이다. 여기서는 ‘알바천국’의 데이터를 가져와 우리의 데이터베이스에 가공하여 담았다.
저장	MySQL	표준 데이터베이스 질의언어인 SQL을 사용하는 관계형 데이터베이스 관리 시스템이다. 매우 빠르고 유연하며 사용하기 쉬운 특징이 있다.
분석	JSP	데이터 분석은 JSP와 Java를 이용하여 구현하였다. Java는 객체지향프로그래밍 언어로서 C/C++에 비해 간략하고 쉬우며 네트워크 기능의 구현이 용이하기 때문에 인터넷 환경에서 가장 활발히 사용되는 프로그래밍 언어이다. JSP는 자바서버페이지로 자바로 구현된 기술로서 특히 서블릿이라는 서버 프로그래밍기술에 기반한 웹 프로그래밍 언어이다. 자바서버 페이지는 실행시에는 자바서블릿으로 변환된 후 실행되므로 서블릿과 거의 유사하지만 HTML 표준에 따라 작성되기 때문에 웹디자인을 하기 편하다.
	Java	
시각화	HTML	HTML은 웹페이지의 큰 뼈대를 만들고, JSP와 JavaScript는 웹페이지의 동작을 만든다. 이를 이용하여 그래프나 웹페이지 디자인을 한다. HTML은 웹 문서를 만들기 위해서 사용하는 기본적인 프로그래밍 언어의 한 종류이며 태그라는 명령어를 사용하여 HTML을 작성한다. HTML에서는 문서가 별도의 코드를 인식하여 완벽한 하이퍼텍스트를 만들 뿐만 아니라 단어 또는 단문을 인터넷의 다른 장소나 파일로 연결시킬 수 있다. 자바스크립트는 객체 기반의 스크립트 프로그래밍 언어이다. 웹 브라우저 내에서 주로 사용하며, 다른 응용 프로그램의 내장 객체에도 접근할 수 있는 기능을 가지고 있다.
	JSP	
	JavaScript	
기반환경	Apache Tomcat	아파치 톰캣은 웹 서버와 연동하여 실행할 수 있는 자바 환경을 제공하여 자바 서버 페이지(JSP)와 자바 서블릿이 실행할 수 있는 환경을 제공한다. 서블릿이나 JSP를 실행하기 위한 서블릿 컨테이너를 제공하며, 상용 웹 애플리케이션 서버에서도 서블릿 컨테이너로 사용하는 경우가 많다.

○ 데이터 크롤링

현재 웹상에 있는 데이터들은 무궁무진하다고 할 수 있다. 이를 활용하기 위해서 데이터 크롤링방식을 통해 웹상에 퍼져있는 데이터를 추출해 낼 수가 있는데 여기서는 ‘알바천국’ (www.alba.co.kr)의 데이터를 가져와 데이터 베이스에 가공하여 담았다. Java의 URL클래스로 웹사이트의 소스를 가져와서 문자열을 잘라내고 속성을 숫자로 변환하기 위해 (회귀분석시 값들이 상수이어야 함) 숫자에 매핑시켰다. 또한 금액 같은 경우도 제공되어 있는 자료가 ‘시급’으로 작성되어 있을 수도 있고 ‘월급’ 또는 ‘주급’ 등으로 작성되어 있을 수 있어 이를 해결하기 위해 월급은 주당 근무 일수와 하루 근무시간 등을 나누어 모두 시급으로 저장되도록 수정하였고 주급 또한 이런 형태로 변형하였다.

또한 Null값이라는 것을 데이터베이스에 삽입 할 때 추후 오류가 날 가능성이 있으므로 Null값으로 넣지 않고 나올 수 없을 것으로 추정되는 값인 -9999999로 Null값을 대체 하였다.

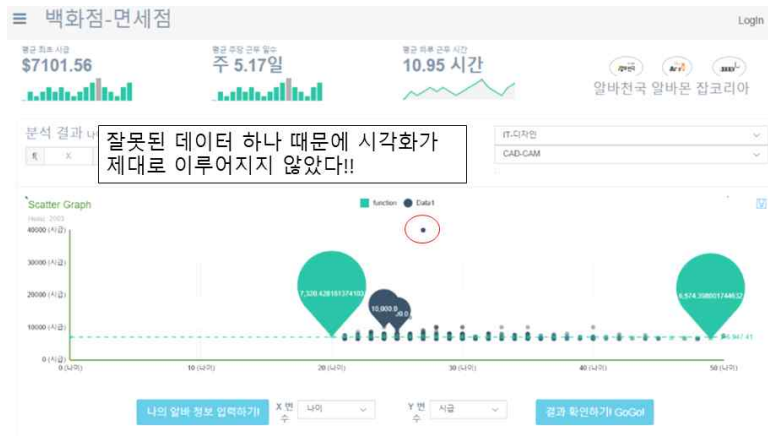
The screenshot displays a Java application window titled 'AlbaHeavenCrawler'. The interface is divided into several sections:

- Left Sidebar:** Contains a list of job categories (e.g., '경력(연차무관)', '성별무관', '연령', '학력') and a '모집내용' (Recruitment Content) section with details like '기타매장 > 청소', '편의점 > 카운터, 진열, 재고관리', '고용형태: 아르바이트', and '모집인원: 3명'.
- Main Window:** Shows a table of job postings with columns for job ID, title, location, and other details. A message '10000 rows in set (0.06 sec)' is visible, indicating the number of rows retrieved from the database.
- Bottom Console:** Displays the execution of SQL queries and the resulting data. The queries show the insertion of job data into a table, with values for job ID, title, location, and other attributes. The console output shows the results of these queries, including job details like '경력: 0, 성별: -1, 나이: 31, 최종학력: 0, 대분류: 임시·보습학원, 소분류: 수학, 근무 형태: 계약'.

< 그림 4. 데이터 크롤링 구현 >

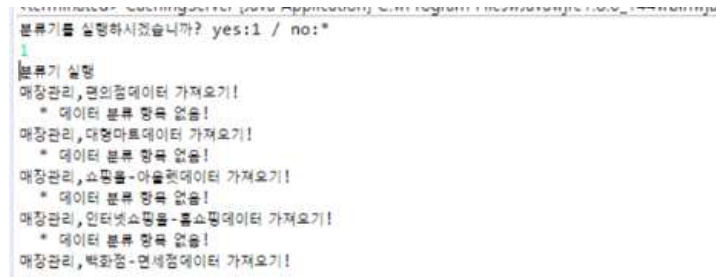
현재는 대략 1000000개 가량의 데이터를 수집하였다.

○ 데이터의 분류



< 그림 5. 데이터 분류 전 >

데이터가 분류되지 않으면 위의 그림과 같이 시각화 또는 분석에 문제가 생길 수 있다. 예상되는 분류할 데이터들은 구인구직 공고를 올리는 사장님이 사이트에 올릴 때 “하루 일급”으로 적어야 하는 돈을 “시급”으로 잘못 체크하여 위와 같은 현상이 일어날 것이라고 보는데 이를 처리하기 위해 데이터 분류 작업을 취한다.



< 그림 6. 데이터 분류기 >

따라서 분류장치를 두어 아르바이트 별로 level을 정하여 데이터의 범위 ex) 카페 아르바이트 (level 1) = 시급 ~10000원


대형마트 아르바이트 (level 2) = 시급 ~15000원

이런식으로 level을 두고 범위를 정하여 아래와 같이 데이터를 걸러내는 작업을 하였다.



< 그림 7. 데이터 분류 후 >

○ 회귀 모형 알고리즘 목록 표

설명\분석 모형	선형 회귀분석	2차 함수모형 분석	단조 증가하는 3차 함수 모형	함수모형 회귀분석 (제네릭 버전)
웹 분석 페이지				
분석 방식	일반적인 선형회귀분석을 기계학습을 통해 구현 할 수 있다.	선형회귀를 변형하여 2차 함수모형으로 기계학습을 시켜 분석한다.	선형회귀를 변형하여 단조 증가하는 3차 함수 모형으로 기계학습을 시켜 분석한다.	데이터들을 선형 , 2차 모형 , 단조 증가 3차 모형들로 각각 분석해 보고 그 중에서 데이터 분포에 가장 적합하다고 판정되는 것으로 출력한다.
분석의 특징	두 데이터 값(x,y) 간의 선형적 상관관계를 파악한다 (예시) 경력에 따라 시급이 오를까? =>분석시 함수 $f(x)$ 로 도출 => $f(x) = 30x + 270$ 존재하지 않는 데이터도 함수로 추측가능 => 3개월 차 월급 $f(3) = 30 \cdot 3 + 270 = 90 + 270 = 360$ (만)	어떤 데이터가 증가하다(+) 감소(-) 또는 감소하다(-) 증가(+) 하는 데이터라면 2차 함수모형이 적합하다 $f(x) = w(x - x_1)^2 + y_1$ 존재하지 않는 데이터도 함수로 추측가능	어떤 데이터가 '정체 구간'을 표현하기에 적합하다. (예시) 경력에 따라 수익이 오르지 않는 구간도 생길 수 있으며 계속적으로 증가하는 함수이므로 분석 시 감소하는 구간이 있으면 안 되는 데이터에 적합하다.	앞의 데이터들을 각각 데이터의 분포에 알맞은 정도(cost)로 기준을 두고 가장 분포에 알맞은 함수를 채택하여 보여준다.

○ Linear Regression 문제

Linear Regression 문제를 말한다. 통계학적으로는 ‘선형 회귀 분석’의 모델을 컴퓨터가 학습할 수 있도록 설계하여 상수 값을 갖는 두 개 이상의 변수를 갖는 데이터가 주어지면 그 변수간의 상관관계를 학습시켜 가장 적합한 하나의 직선을 찾아내는 알고리즘이다. 여기서 가장 적합한 하나의 직선이 되는 기준은 cost라는 함수와 데이터간의 차이의 제곱의 평균 (분산을 생각하면 된다.)이 가장 작을 때의 함수가 가장 적합하다는 기준이 되는데

예를 들면 직장인들의 월급을 분석해보는다면 경력에 따른 월급으로 분석해볼 때 경력이 오르면 월급은 오를 것 이다! 즉 경력과 월급은 변수간의 상관관계가 존재한다는 것인데 이 관계를 직선(선형적인 관계)으로 표현한다면 $y=30x+270$ 이라는 함수가 나

오게 된다 해보자 이 함수가 주어지게 되면

일반적으로 경력 1년차는 300만원 , 2년차는 330만원 ... 이런식으로 도중에 없는 데이터가 존재하여도 어느 정도 추측을 할 수가 있다는 것이다. 더 나아가 미래에는 내가 얼마정도의 월급을 받을 지도 대략 추측할 수 있게 되며 이러한 결과는 사람의 의사결정에 도움을 주기도 한다.

○ 2차 함수모형

위에서 ‘선형적’이란 단어의 뜻은 그래프 관점에서 보면 ‘직선적인 관계’로 볼 수 있다.

그러나 모든 데이터의 분포를 선형적으로만 분석하게 된다면 문제가 있다고 보았다.

따라서 비선형적인 모델로 분석하기 위해 정의역이 주어진 전사함수를 선정하게 되었고 가장 일반적인 함수들을 채택하게 되었다. $f(x) = w(x-a)^2 + b$

○ 3차 단조 증가 함수 모형

시간 복잡도에 관한 설명은 위의 내용과 비슷하며

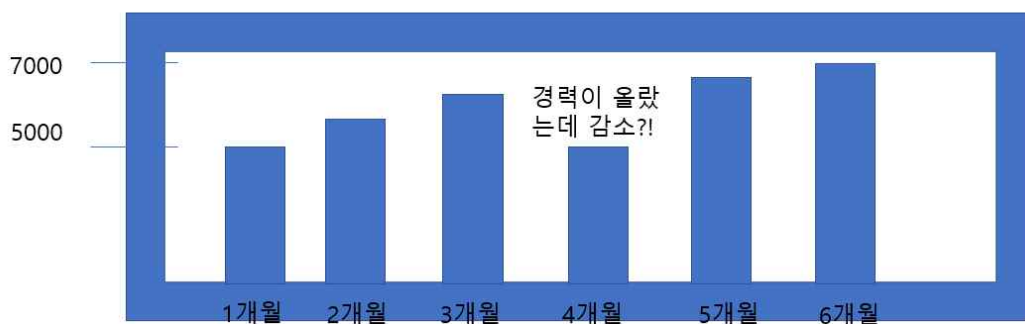
여기서 3차 함수를 $f(x)$ = 단조 증가하는 3차 함수 $\Rightarrow w(x-a)^3 + b$

로 선정한 이유는 이 함수만의 특징이 있기 때문이다

먼저 앞서 설명하면 경력에 따른 월급분석을 할 때 경력이 오르고 있는데 월급이 떨어지게 되는 구간이 있다는 것은 일반적으로 분석이 잘못 됐다고 생각하여 신뢰도가 떨어지게 될 것이다.

일반적으로 경력에 따른 시급 분석이라고 하면
이때 분석해야할 모형은 반드시 “증가”만 하는 “함수 모형”을 선택해야 한다!

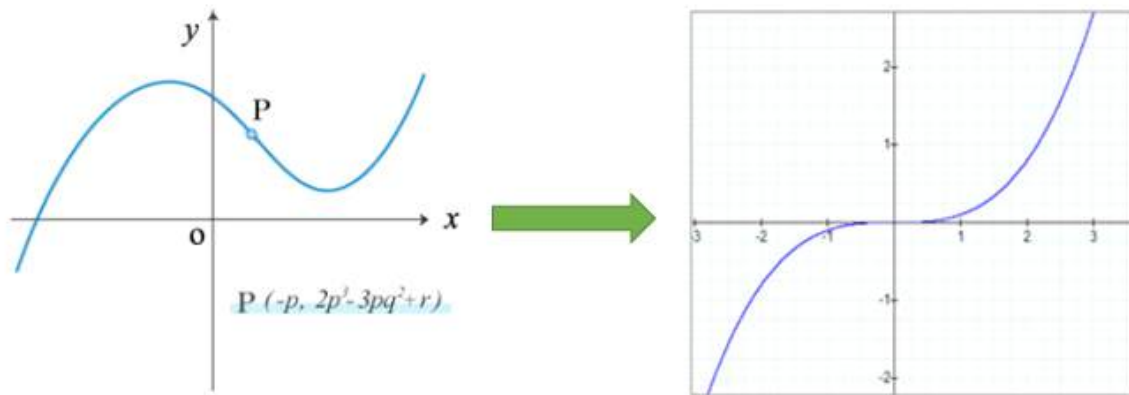
예를 들어보자 !



< 그림 8. 분석결과가 단조증가 해야 하는 예시 >

따라서 원래 3차함수모형은

$$f(x_i) = ax^3 + bx^2 + cx + d \quad \text{로 본다면} \quad f(x_i) = w(x-a)^3 + b$$



< 그림 9. 일반 3차 함수에서 단조 증가하는 3차함수로의 변형 >

으로 변형하여 다음과 같이 수식을 변형하였다.

즉 이 함수모형은 어떤 증가하다가 ‘정체되는 구간’을 표현하기 좋은 모형이며 시간복잡도 또한 수정할 weight값이 적어 실시간 응답에 적합한 함수모형이다.

○ 선형, 2차, 단조 증가 3차 모형 알고리즘

위의 알고리즘을 구현하기 위해서는 크게 다음과 같은 순서로 진행된다.

1) 데이터를 1, 2차 단조 증가 3차 함수 모형이라고 각각 가정하고 그에 따라 추측되는 중심 좌표

(알고리즘 상 centerPoint)를 찾아낸다.

2) 그 좌표를 함수식의 (a, b) 값에 대입하여 weight값을 수정해 나간다.

(gradient descent 알고리즘)

3) function Regression일 경우 각 모형들을 분석했을 때 가장 cost(비용)값이 작은 함수 모형을 채택하여 보여준다.

먼저 1)의 내용은 각 함수식을 표준형으로 추상해야한다는 것에 중점을 둔다.

일반형	$f(x) = w_1x^2 + w_2x + w_3$	수정해야할 weight값이 많다. (3개)
표준형	$f(x) = w(x-a)^2 + b$	꼭지점 좌표를 대략 찾아내 수정할 weight값을 1개로 줄인다.

$f(x) = w(x-a)^2 + b$ 가 그 중 선정된 2차 함수식이며 알고리즘은 다음과 같이 작동한다.

1) 데이터를 2차 함수 모형이라고 가정하고 그에 따라 추측되는 중심 좌표

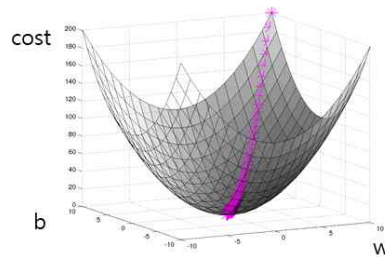
(알고리즘 상 centerPoint)를 찾아낸다.

2) 그 좌표를 설정한 뒤 weight(w)값을 수정해 나간다.

일반형으로 분석하게 되면 물론 결과 함수는 더욱 정확하겠지만 이러한 알고리즘 방식으로 작동하면 시간 복잡도나 그래프를 추상하기 매우 간단해 진다.

예를 들면 $y=wx+b$ 의 형태의 단순한 모형도 따지고 보면 조정해야할 변수가 2개이다 그런데 gradient descent 알고리즘을 사용하기에 적합한지 확인하고자 하면 전체적인 그

래프를 그려보게 되는데 조정해야할 변수가 2개이고 cost값을 y축에 세워보면 결국 3차원의 곡면이 최소 cost값으로 수렴할 수 있도록 그래프가 이루어 져야한다.



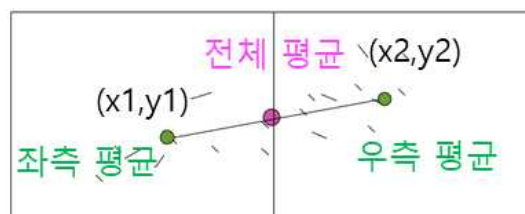
<그림 10.조정할 weight 값을 1개로 줄이는 이유 ($y=wx+b$ 의 예시)>

3차원 까지는 무난하다 하자 그런데 2차 함수 모형이나 단조 증가하는 3차 함수 모형에서는 조정해야할 변수가 3개가 되는데 이를 gradient descent알고리즘을 사용하기 적합한지 확인하려면 4차원을 생각하게 되므로 가능하다 하더라도 추상하기 어려움이 있다.

따라서 여기서는 어떤 centerPoint(a,b) 좌표를 두어 모형별로 “어떤 좌표를 대략 지날 것인가?” 를 통계적 방식으로 먼저 추측하고 그 좌표를 상수 값으로 대입하여 하나의 weight만을 수정하겠다는 방식으로 구성되었다.

따라서 표준형으로 함수식을 설계하고 그에 따른 중심 좌표(a,b)를 미리 구하는 것이 매우 중요한데 이는 다음과 같다.

1-1) Linear Regression(선형회귀) 의 centerPoint(a,b)



< 그림 11. Linear Regression의 경우 centerPoint와 초기 w값 선정 방법에 대한 예시 >

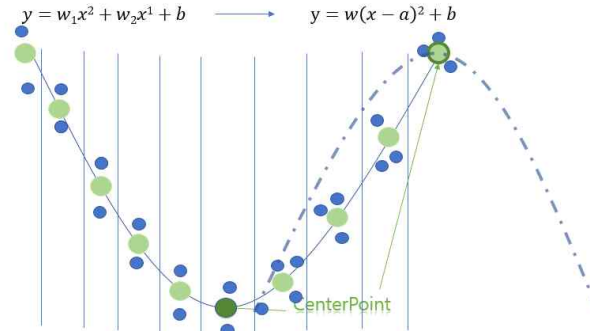
그림에서 centerPoint는 빨간색 좌표를 뜻한다.

즉 데이터들을 선형으로 분석한다는 것은 데이터의 분포가 선형적으로 분포 되어있다고 가정하고 시작할 수 있으므로 “선형적으로 분포되어 있을 때 구하고자 하는 회귀 직선은 어떤 좌표를 지나게 될까?” 의 대답은 전체 데이터들의 x,y값의 평균 값을 반드시 는 아니어도

centerPoint(a,b)로 선정하기에는 무리가 없다고 생각하였다.

또한 위에서 centerPoint를 기준으로 좌측 평균,우측 평균도 추가로 구하였는데 이것의 효과는 초기의 weight값을 대략적으로 추상하여 기울기(w)의 초기값을 적절하게 선정하기 위함이다. ($w = \frac{y_2 - y_1}{x_2 - x_1}$)

1-2) 2차 함수 모형의 centerPoint(a,b)



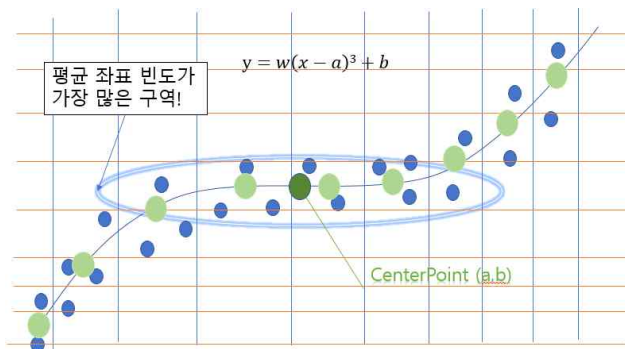
< 그림 12. 2차 함수 모형의 centerPoint >

“2차 함수 모형으로 분포되어 있을 때 구하고자 하는 회귀 포물선은 어떤 꼭지점 좌표 위치를 지나게 될까?”

라고 출발하여 적절한 centerPoint좌표를 찾아낸다.

1. 데이터들의 최소 x값, 최대 x값을 구한 뒤 그 범위를 10등분하고 등분된 부분의 평균 포인트들을 구한다.
2. 양수일 경우 : 10개의 평균 Point들의 최소 Point
음수일 경우 : 10개의 평균 Point들의 최대 Point
3. 구해진 2개의 좌표로 gradient descent 알고리즘을 각각 실행하여 둘 중 cost값이 낮은 centerPoint와 weight를 가져온다.

1-3) 단조 증가하는 3차 함수 모형 centerPoint(a,b)



< 그림 13. 3차 함수 모형의 centerPoint >

“단조 증가하는 3차 함수 모형으로 분포되어 있을 때 구하고자 하는 회귀선은 어떤 좌표를 중심으로 지나게 될까?”

라고 출발하여 적절한 centerPoint좌표를 찾아낸다.

1. X (Key)의 최소 값과 최대 값으로 x축 10등분
2. 10등분된 각 범위별 평균 좌표를 찾아냄
3. Y (Value)의 최소값과 최대값으로 y축도 10등분
4. 등분된 구역중 2에서 얻어낸 평균 좌표의 빈도가 가장 많은 부분을 찾는다
5. 그 구역의 평균좌표가 CenterPoint 이다.

이렇게 채택된 centerPoint로 시작하여 gradient descent 알고리즘을 하게 된다.

2) 함수 모형에 따른 gradient descent 알고리즘

〈표 함수모형식과 gradient descent 함수식〉

	모형 함수 식	gradient descent
1차 함수 모형	$f(x) = w(x-a)^1 + b$	$w = w - \frac{r}{n} \sum_{i=1}^n (w(x_i - a)^1 + b - y_i)(x_i - a)^1$
2차 함수 모형	$f(x) = w(x-a)^2 + b$	$w = w - \frac{r}{n} \sum_{i=1}^n (w(x_i - a)^2 + b - y_i)(x_i - a)^2$
단조 증가 3차 함수 모형	$f(x) = w(x-a)^3 + b$	$w = w - \frac{r}{n} \sum_{i=1}^n (w(x_i - a)^3 + b - y_i)(x_i - a)^3$

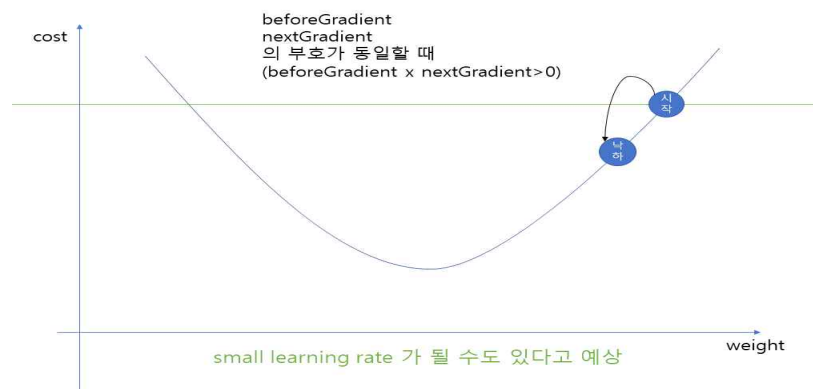
함수식이 변형됨에 따라 gradient descent 알고리즘의 함수식 또한 cost 함수를 w 에 대해 편미분 시켜 변형된다.

여기서 중요한 부분은 learning rate라고 불리는 함수식에서의 r 값인데 이 값이 너무 높으면 overShooting현상 또는 small rate현상이 일어나기 때문에 적절한 학습계수를 선정하는 것이 중요하다.

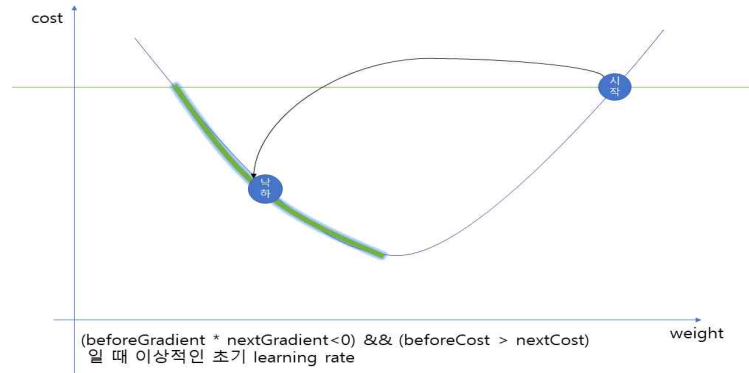
필자는 다음과 같은 방식으로 학습계수를 조절하였다.



< 그림 14. overShooting 예상 범위 >



< 그림 15. small learning rate 예상 범위 >



< 그림 16. 이상적인 초기 learning rate 선정 범위 >

여기서 하고자하는 learning rate 조절 방식은 “초기값”을 잘 선정하여 이상적인 learning rate값을 찾아낸 뒤 overShooting 또는 small rate 현상이 일어날 것으로 보이면 learning rate값을 조금씩 늘리거나 줄여나가 학습시키자는 방식이다.

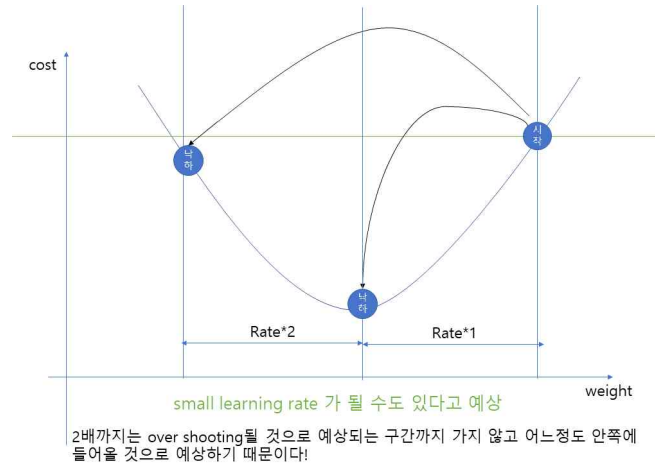
```
rate = toInitLearningRate(w,centerPoint,0.01,degree);
// learningRate값 초기값 (초기화)
for (int i = 0; i < randomCount; i++) {
```

```
.... //생략
if(beforeCost<afterCost) {
//OverShooting 예상됨 초기화 된 rate값을 줄인다.
    rate*=0.1;
}else if(beforeGradient*afterGradient>0) {
// learning rate 올려도 됨. (*꼭 2배 이하)
    rate*=2;
}
if(beforeCost>afterCost) {
    resultW = w;
    beforeCost = afterCost;
}
if(afterGradient==0) {
break;
}
}
```

```
return resultW;
```

*rate값을 늘릴 때는(2배 이하로 늘려야 한다.)

$\text{beforeGradient} * \text{afterGradient} > 0$ 의 의미는 참고로 기울기의 부호가 서로 같다는 것을 표현하기 위해 표현하였다. 이때 rate 값을 1배보다는 크고 2배 이하로 조절하게 되는데 이유는 다음과 같다.



< 그림 17. rate 값을 늘릴시 1~2배 이하로 선정해야 하는 이유 >

2배까지는 overShooting이 될 것으로 예상되는 구간까지 가지 않고 어느정도 안쪽에 들어올 것으로 예상하기 때문에 안정적이라 생각하였다.

○ Function Regression 알고리즘

여기서 말하는 Function Regression 알고리즘은 Linear Regression이

한번의 분산값 (cost)값을 구하기 위해 $\frac{\sum_{i=1}^n (f(x_i) - y_i)^2}{n}$ 의 모형을 따른다.

이를 위의 함수식을 미분하여 gradient descent 알고리즘으로 weight값을 추출한다. Linear Regression은 위의 식에서 $f(x_i)$ 가 $f(x_i)=wx+b$ 라고 보면 된다. 그렇다면 여기 있는 $f(x_i)$ 를 수정하자는 생각인데 알바 셀파에서는 함수를 3가지로 정하였다.

	함수식	gradient descent	cost값
LinearRegression	$f(x) = w(x-a)^1 + b$	$w = w - \frac{r}{n} \sum_{i=1}^n (w(x_i - a)^1 + b - y_i)(x_i - a)^1$	100
2차함수 모형	$f(x) = w(x-a)^2 + b$	$w = w - \frac{r}{n} \sum_{i=1}^n (w(x_i - a)^2 + b - y_i)(x_i - a)^2$	3000
단조 증가 3차 함수 모형	$f(x) = w(x-a)^3 + b$	$w = w - \frac{r}{n} \sum_{i=1}^n (w(x_i - a)^3 + b - y_i)(x_i - a)^3$	80
...	추가되는 함수모형들..
function Regression	단조 증가 3차 함수 모형이 cost값이 가장 작음 = 단조 증가 3차 출력		80

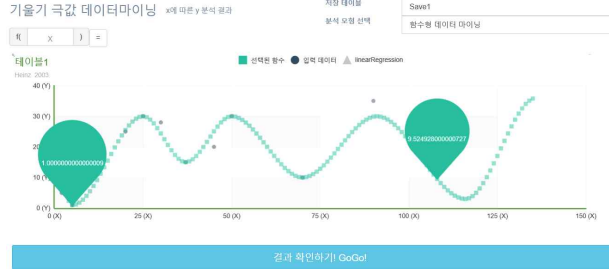

다음 함수들을 다음과 같이 변형하여 각 함수 모형에 따른 비용 cost값이 존재하는데 결과적으로 cost값이 높으면 안 좋다고 보면 된다.

(cost 값은 함수에 대해 데이터들의 떨어져 있는 정도를 나타낸다.)

따라서 여러 함수모형을 대입하여 그 중 최종적으로 가장 적합한 함수 모형을 출력하자는 알고리즘이다.

○ 극값 데이터 마이닝 알고리즘

데이터간의 “극값” 이란 매우 중요한 정보로 사용이 가능하다 생각하여
극값들의 주기를 분석해 다음에 일어날 극값이 언제 일어날지 분석하는 기법이다

설명\분석 모형	극값 데이터 마이닝 알고리즘	
	‘기울기’를 이용한 극값 데이터 마이닝	‘선형 회귀’를 이용한 극값 데이터 마이닝
웹 분석 페이지		
분석 방식	<p>각 key값에 따른 value값들의 평균을 내어 그 평균 값들중 ‘극 값’이라고 생각되는 평균값만 뽑아내 가장 (*최근 바로 전 구간)의 기울기와 과 거의 기울기를 통하여 유사한 구간을 찾아내고 다음번 째 극값이 언제 일어날지를 추측하여 추가 한 뒤 3차 함수의 미적분을 통하여 연결한 모형</p>	<p>각 key값에 따른 value값들의 평균을 내어 그 평균 값들중 ‘극 값’이라고 생각되는 평균 값만 뽑아내 가장 (*최근 바로 전 극값)과 선형회귀분석과 의 차이와 과거의 극값과 선형회귀분석과의 차이를 비교하여 유사한 위치를 찾아내고 다음번 째 극값이 언제 일어날지를 추측하여 추가한 뒤 3차 함수의 미적분을 통하여 연결한 모형</p>
분석의 특징	<p>기울기는 작용과 반작용의 느낌이다. 예를 들면 주식으로 연상하면 ‘빠르게 오르면 빠르게 떨어질 것’ 이다! 또는 ‘약하게 밀면 약하게 튕겨 나온다’던지 최근의 기울기를 보고 과거에 있던 패턴을 가져와 뒤에다 연결하게 되는 방식으로 분석된다.</p>	<p>선형회귀 모형은 분석 할 데이터의 분포가 ‘이미 선형적으로 움직인다.’라는 것을 알고 있을 때 사용하는 분석방식이며 ‘데이터의 분포가 선형적이라는 정보’는 매우 실용적인 정보가 된다. 따라서 선형적인 데이터라는 것을 이용하여 선형적 분석 결과 함수와 얼마나 극 값이 떨어져 있는 지를 기준으로 하여 추측하고 과거의 패턴을 가져와 뒤에 연결하는 방식으 로 분석된다.</p>

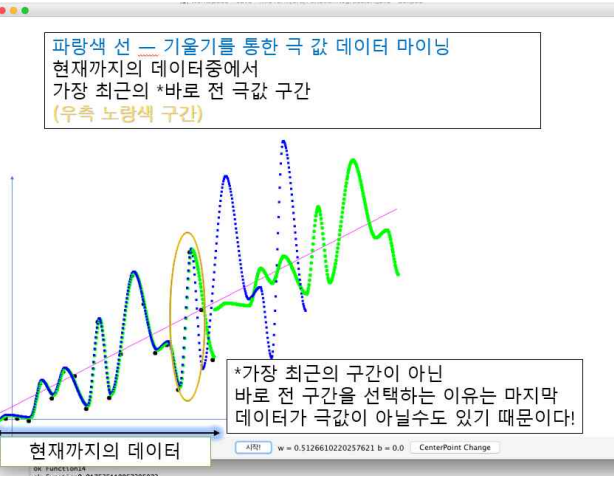
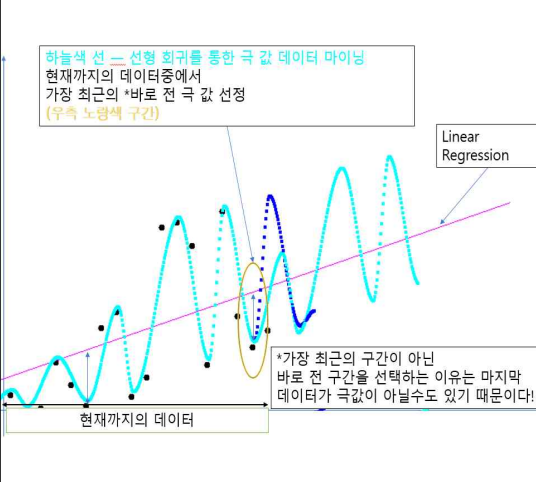
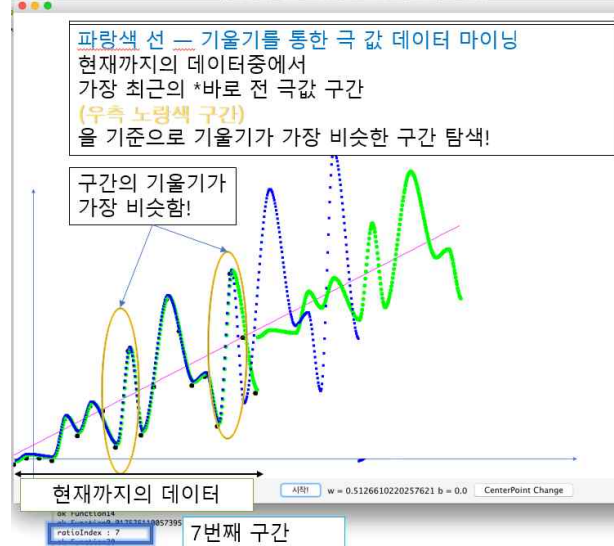
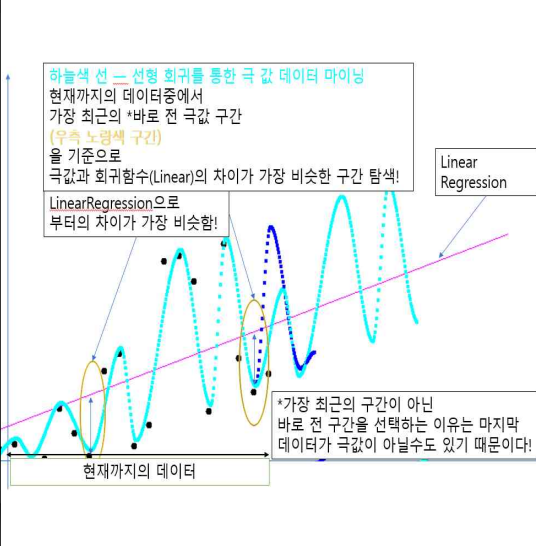
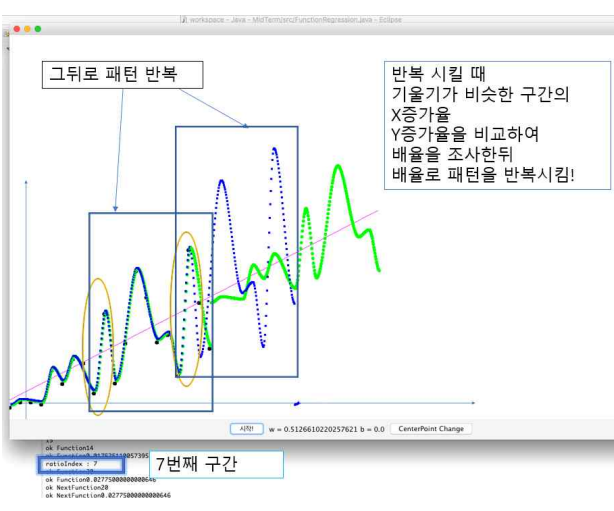
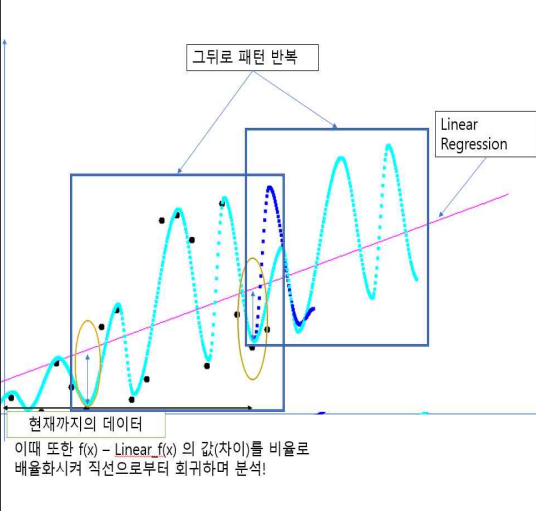
이러한 극 값을 분석하는 방식이 존재 하는지는 모르나 다음과 같이 방식을
설계하였고 알고리즘으로 개발하였다.

(*가장 최근 ‘바로 전’ 극값 구간) 이 되는 이유는 마지막 데이터가 극값으로 보일
진 몰라도

다음 번째의 데이터가 더 큰 값(또는 더 작은 값) 이 나온다면 마지막 값은 극값이
아니게 된다. 따라서 정확하지 않은 두 극값의 사이의 특징(기울기 등)을 기준으로
사용하게 되면 알맞지 않으므로

그 바로 전의 극값 구간을 채택하여 기준(pivot)으로 삼아 과거와 대조해 본다는
것이다.

한마디로 가장 최근의 극값구간을 잡아내 특징을 선정하여 과거와 대조한다는 것이다.

설 명 \ 분석 모 형	기울기를 통한 극값 데이터 마이닝	선형 회귀분석을 통한 극값 데이터 마이닝
1단계 -극값의 기준 (pivot) 선정	<p>파랑색 선 — 기울기를 통한 극 값 데이터 마이닝 현재까지의 데이터중에서 가장 최근의 *바로 전 극값 구간 (우측 노랑색 구간)</p>  <p>*가장 최근의 구간이 아닌 바로 전 구간을 선택하는 이유는 마지막 데이터가 극값이 아닐수도 있기 때문이다!</p> <p>현재까지의 데이터</p> <p>Linear Regression</p>	<p>하늘색 선 — 선형 회귀를 통한 극 값 데이터 마이닝 현재까지의 데이터중에서 가장 최근의 *바로 전 극 값 선정 (우측 노랑색 구간)</p>  <p>*가장 최근의 구간이 아닌 바로 전 구간을 선택하는 이유는 마지막 데이터가 극값이 아닐수도 있기 때문이다!</p> <p>현재까지의 데이터</p> <p>Linear Regression</p>
2단계 - 유사 기준 탐색	<p>파랑색 선 — 기울기를 통한 극 값 데이터 마이닝 현재까지의 데이터중에서 가장 최근의 *바로 전 극값 구간 (우측 노랑색 구간) 을 기준으로 기울기가 가장 비슷한 구간 탐색!</p>  <p>구간의 기울기가 가장 비슷함!</p> <p>현재까지의 데이터</p> <p>7번째 구간</p> <p>Linear Regression</p>	<p>하늘색 선 — 선형 회귀를 통한 극 값 데이터 마이닝 현재까지의 데이터중에서 가장 최근의 *바로 전 극값 구간 (우측 노랑색 구간) 을 기준으로 극값과 회귀함수(Linear)의 차이가 가장 비슷한 구간 탐색!</p>  <p>LinearRegression으로 부터의 차이가 가장 비슷함!</p> <p>*가장 최근의 구간이 아닌 바로 전 구간을 선택하는 이유는 마지막 데이터가 극값이 아닐수도 있기 때문이다!</p> <p>현재까지의 데이터</p> <p>Linear Regression</p>
3단계 -다음 함수로의 연장	<p>그뒤로 패턴 반복</p>  <p>반복 시킬 때 기울기가 비슷한 구간의 X증가율 Y증가율을 비교하여 배율을 조사한뒤 배율로 패턴을 반복시킴!</p> <p>현재까지의 데이터</p> <p>7번째 구간</p> <p>Linear Regression</p>	<p>그뒤로 패턴 반복</p>  <p>이때 또한 $f(x) - \text{Linear}_f(x)$의 값(차이)을 비율로 배율화시켜 직선으로부터 회귀하며 분석!</p> <p>현재까지의 데이터</p> <p>Linear Regression</p>

또한 여기서 각 구간들을 3차 함수 형태로 연결하였다.

알고리즘 작동 방식은 다음과 같다.

각 계급별로 그 평균값을 뽑아내 그 평균값이 증가하다가 감소하는 구간이 생기면 그 값을 극 값이라 생각하고 이를 이용하여 3차 함수의 미적분을 통해 $f'(x) = (x-a)(x-b) \Rightarrow$ 적분 $\Rightarrow f(x) = \int f'(x)dx + C \Rightarrow$ 극값 좌표 2개를 대입 및 연립 $\Rightarrow f(x)$ 하는 방식으로

극값 좌표 2개가 들어왔을 때 그 좌표를 극값으로 갖는 3차함수를 추출해내고 각 구간별로 반복한다는 것이다.

따라서 하나의 구간별로 3차 다항 함수를 얻어낼 수 있는데

이때 극 값들 중 마지막에 있는 극값을 빼고 가장 마지막 구간을 채택해 특징을 선정하고 그 특징을 모두 측정한다 그 다음 전에 있던 구간들 중 가장 특징이 유사한 구간을 채택한 뒤 그 뒤를 유사한 구간의 뒤로 이어 나간다는 이론이다.

※ 왜 3차함수인가?

각 구간을

1. 직선으로 연결하여 분석할 수도 있고
2. 극값을 극값으로 갖는 ‘다항 함수’로 만들 수 있으며
3. 다음과 같이 각 구간 별로 3차 함수를 이용하여 분석할 수도 있다.

그런데 왜 3차함수를 적용한 이유는 다음과 같다

- 1) 직선은 기울기가 일정하여 미분이 불가능한 구간이 생긴다.

미분이 불가능 한 구간은 순간이동 같은 느낌을 준다 생각하여 현실과는 맞지 않다고 생각하였다.

- 미분 불가 존재함

- 한번의 $f(x)$ 를 구하기 위한 시간복잡도 $O(n)$

- 2) 시간복잡도가 극값이 n 개이며 그 극값을 극값으로 갖는 다항 함수의 경우

구해보면 $2n-1$ 차 함수 \rightarrow 대략적인 연산 횟수 $= \frac{(2n-1)(2n)}{2}$

-한번의 $f(x)$ 를 구하기 위한 시간복잡도 $O(n^2)$

- 시간복잡도 $O(n^2)$

- 3) 3차 함수인 경우는 미분도 가능하며 시간복잡도 또한

- 적합하다!

상수 시간 $(3+2+1) \times (\text{각 구간의 개수}) = O(n)$ 이 된다.

따라서 미분이 가능해야 하며 시간 복잡도는 ‘빅 데이터’라는 관점에서 보면 매우 크게 작용할 것이므로

따라서 3차함수가 가장 적합하다!

○ ‘기울기’를 통한 극값 데이터 마이닝 알고리즘

사용 예시를 들면

- 주식의 전환 점 추측(상승세에서 하락세로 변환 시점),
- 나의 수익 최고 점 추측 등등..

으로 볼 수 있다.

위에서 설명한 방식 중 마지막의 특징을 여기서는 ‘기울기’로 채택하였다는 것이다.

(*최근 바로 전 구간)의 기울기와 과거의 기울기를 통하여 유사한 구간을 찾아내고

다음번 께 극값이 언제 일어날지를 추측하여 추가한 뒤

3차 함수의 미적분을 통하여 연결한 모형이다.

그 다음 그 뒤의 모형을 평행이동 시키고 기울기를 대조했던 두 개의 구간끼리

비교하여 확대 또는 축소시킨다.

따라서 다음의 발상은 가장 최근의 극값 간의 기울기를 구하여 과거의 극값간의 기울기들과 비교해보고 가장 비슷한 과거를 그 뒤로 이어나가는 분석 기법이다.

기울기를 선정한 이유는

-썩게 치면 썩게 튕겨 나간다거나

-주식이 빠르게 올라가면 빠르게 떨어진단지

과거에 있던 기울기를 비교하여 분석한다는 것은 변화율에 따른 다음 패턴을 그려내 다음에 일어날 극값을 찾아내는 기법이다.

극값을 실제로 적용해보면 주식의 전환점(상승세에서 하락세로 변하는 시점) 등이나 최대 이익 과 최소이익을 날짜별로 나열하다보면 언제 다시 최대이익이 일어날지 나름대로 분석 해볼 수 있을 것이라 생각한다.

○ ‘선형 회귀분석(Linear Regression)’을 통한 극값 데이터 마이닝 알고리즘

다음의 발상은 극값은 찾고 싶은데 데이터가 이미 선형적으로 움직인 다는 것을 알고 있을 때 사용하는 기법이다.

이 기법은 선형모형 LinearRegression 이 어느정도 (평균선) 과 비슷하다는 생각에 평균선과의 차이를 비교하여 가장 최근의 극값이 평균선보다 얼마큼 떨어져 있는지 분석하고 과거와 비교하여 그 다음 극값 들을 찾아내는 기법이다.

방식은 위의 기울기를 통한 분석방식과 유사하게 마지막 극값을 빼고 그전의 극값과 마지막 극값의 차이를 비교하고 가장 유사한 차이를 갖는 위치부터

그 뒤를 그려나가는 데 그 차이의 닮음비를 이용하여 확대,축소시킨다.

추가한다면 Function Regression(함수형 회귀 분석)을 통한 극값 데이터 마이닝

알고리즘을 생각해 볼수 있는 데 위에 설계 식은 ‘선형’ 으로 회귀하는 폼으로 다루었다면 여기서는 ‘함수형’ 으로 교체해보고 설계할 수도 있다.
즉 선형함수 대신에 자신이 선정한 함수로 교체하여 회귀 시킨다는 이야기 이다.

- ‘다 변수(특징)’ 에서의 극값 데이터 마이닝 알고리즘
- 이 내용은 구현하지 않고 추상만 하였다.
위의 내용은 각각 “하나의 특징” 으로만 유사구간을 판정하여 뒤를 이어나간다.
그런데 하나의 변수보다 다양한 특징들이 모두 유사한 구간을 판정하여 구현할 수 있는 데 예를 들면 gap(차이)을 정규화 시킨 뒤 분산 값(cost)으로 대조시키는 방법이다.

이러한 극값을 이용한 분석방식의 메인은 ‘미래를 예측한다는 개념’ 으로 접근하여 알바 셀파의 데이터마이닝 기법하고는 맞지가 않다. (데이터 분포와의 적합성)
따라서 분석기에만 존재하고 알바 셀파의 아르바이트 분석에는 포함되지 않는다.
※위에서 설명한 극값 데이터 마이닝 기법은 개인 분석기에서만 존재한다.
※알바 데이터를 분석할 때는 함수형 데이터 마이닝 기법만을 사용한다.

4. 작품 개발 환경

구분		상세내용
S/W 개발환경	OS	window, Mac OS
	개발환경(IDE)	
	개발도구	Eclipse, MySQL , Atom , SQLGate
	개발언어	Java, mysql, html, java Script , JQuery
	기타사항	알바천국 데이터 수집
프로젝트 관리환경	형상관리	
	이슈관리	
	의사소통관리	카카오톡
	기타사항	토즈카페 모임

III. 프로젝트 수행 내용

※ 평가항목 : 수행능력 (문제해결능력, 수행충실성)

1. 멘티(참여학생) 업무분장

번호	이름	대학	학과	학년	역할	담당업무
1	김석준	한신대학교	수학과	4학년	팀장	멘토와 진행내역 공유 및 역할 배분 분석 알고리즘,Caching 설계 및 구현 요구사항정의 및 시스템 분석/설계 웹페이지 개발 및 DB 구현 관련 산출물 작성
2	장근호	한신대학교	수학과	4학년	팀원	요구사항정의 및 시스템 분석/설계 웹페이지 개발 및 DB 구현 관련 산출물 작성
3	김예지	한신대학교	수리금융학과	4학년	팀원	요구사항정의 및 시스템 분석/설계 웹페이지 개발 및 DB 구현 관련 산출물 작성

2. 프로젝트 수행일정

프로젝트 기간 (한이음 사이트 기준)		2017.04.03. ~ 2017.11.30.											
구분	추진내용	프로젝트 기간											
		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
계획	어떤식으로 구성할것인지 추진				0								
분석	알고리즘 및 데이터의 분포와 관계가 있는 함수 선택 및 종목별 기각시킬 데이터 기준 설정				0	0							
설계	화면의 view단과 sql선택, 실질적인 내부 구성 도면					0	0						
개발	html, 디자인 구성,크롤링 개발						0	0					
	분석기 적용 및 서버 구축, 기능 구현							0	0				
	크롤링 Caching 자동화 및 서버 과부하 처리								0				
	기능구현								0	0			
테스트	버그 오류 수정 및 편의성을 위한 수정									0	0		
종료												0	

3. 프로젝트 추진 과정에서의 문제점 및 해결방안

3-1. 프로젝트 관리 측면

○ 계획 단계의 주제 선정

- 처음 계획 시 데이터의 분야 상관없이 모든 데이터를 분석할 수 있는 분석기를 주제로 선정함.
- 하지만 데이터 분야가 너무 광범위 하여 프로젝트를 진행하는데 한계 발생.
- 그래서 분야를 아르바이트로 한정하여 주제 결정

○ 회의 및 협업을 위한 시간과 장소

- 한이음에서 토즈 스터디카페 장소를 제공해줌.
- 학교 세미나실도 함께 이용하여 회의 및 성과공유.

○ 파일 공유 시 관리문제

- 서로 맡은 부분의 파일을 공유하면서 예전 파일이랑 섞여서 문제발생.
- 파일을 저장할 때 날짜를 덧붙이는 등 파일 이름을 변경해서 저장

3-2. 작품 개발 측면

○ 자바 jre1.7, jre1.8 호환 문제, 톰캣 오류

- jre1.7로 구현하던 프로젝트를 jre1.8로 바꿈.

- jre 버전을 바꾸는 과정과 프로젝트 실행 시 톱캣에서 오류가 자주 일어남.
- 인터넷 검색 및 팀원 간의 의견공유를 통해 오류수정.
- 서로 다른 운영체제로 인한 인코딩 문제
 - 보통은 window 운영체제를 사용하지만 Mac 운영체제를 사용하는 팀원도 있음
 - 따라서 파일을 공유할 때 파일에 한글이 다 깨져있는 경우도 발생함.
 - 페러럴즈를 이용한 Window 사용 및 UTF-8로 통일
- 데이터 수집 문제
 - 아르바이트에 대한 여러 데이터를 직접 입력하기로 계획
 - 직접 입력은 신뢰도가 떨어져 데이터 크롤링으로 수집방법을 변경
 - 100만개의 데이터 수집
- 서버 과부화 문제
 - 아르바이트 데이터 및 사용자 분석기 개발시 매번 새롭게 분석하기에는 시간 및 서버 과부화 문제 우려
 - Caching테이블과 HeartBeatMessage를 이용하여 최초 1회 분석 및 결과 저장
- 부트스트랩을 이용한 시각화
 - Bootstrap과 JQuery를 이용한 반응형 웹(Responsive Web)을 사용
 - 그래프를 보기 편하도록 구현

4. 프로젝트를 통해 배우거나 느낀 점

- 김석준 : 저희가 모두 복수 전공자 팀 이어서 주변에 물어볼 곳이 없어 정보가 많이 부족했는데 한이음 프로젝트에서 멘토님을 만나 데이터 크롤링과 같은 기법이나 라이브러리 같은 것들이 존재 한다는 것을 알게 되고 찾아보고 구현할 수가 있었습니다.

또한 프로젝트를 하기 위해 Tensorflow와 LinearRegression 문제등의 기법을 스스로 찾아보며 공부 해볼 수 있는 기회가 되었고 독창적으로 분석기법을 설계해보기도 했으며 분석기 개발 시 서버 측에서 과부하가 걸리지 않을 까? 라는 문제를 보안하기 위해 caching테이블과 HeartbeatMessage클래스를 생각하여 서버 과부하 문제를 덜어주는 구조를 짜보았고 실제로 구현시켜보니 매우 즐겁고 흥미로웠습니다.

이번 프로젝트를 진행하면서 개발자가 얼마나 가치가 있는지 더욱 느낄 수 있었고 꼭 배우지 않았어도 스스로 찾아보고 구현 할 수 있다는 생각이 들었습니다. 앞으로 프로그램을 개발하게 되면 이러한 자신감으로 좋은 프로젝트를 개발하는 사람이 되겠습니다.

- 김예지 : 이 프로젝트를 시작할 때 느낌은 기대 반 걱정 반 이었습니다. 그 동안 배웠던 것을 종합하여 우리만의 프로그램을 만든다는 기대도 있었지만 아직 배운지 얼마 안되서 프로젝트를 하는 게 어렵진 않을까하는 걱정이 있었습니다. 하지만 혼자가 아니라 팀으로 같이 하니 걱정했던 것 보다는 수월하게 진행되었습니다.

저희 팀은 프로젝트의 주된 기능이 분석기능이다 보니 통계분야의 공부도 필요했습니다. 그래서 회귀분석에 대해 자료를 찾아보았고 데이터 마이닝에 대한 개념도 공부하게 되었습니다. 학교에서 배우지 못한 개념들을 스스로 공부하는 거라 어려울 때도 있었지만 팀원끼리 서로 설명해주면서 흥미롭게 공부할 수 있었습니다. 또한 서로 의사소통 하는 것도 정말 중요하다고 느꼈습니다. 혼자할 때 어려운 것도 수월하게 했고 진행속도도 빨라졌습니다.

이번 프로젝트를 하면서 학교에서 배운 기초적인 프로그래밍 개념을 활용하는 법을 배웠고 학교 수업 때 작은 팀 프로젝트만 하다가 좀 더 제대로 된 프로젝트를 하고나니 뿌듯했습니다. 함께한 팀원들과 도와주신 멘토님께도 고마움을 많이 느꼈습니다. 또한 진행중에 프로그래밍 실력이 조금씩 느는 것을 느꼈고 앞으로 다른 프로젝트를 하게 되면 이를 계기로 부족한 부분을 보완하여 열심히 하겠습니다.

- 장근호 : 프로그래밍 랭귀지와 컴퓨터를 배우고 큰 프로젝트를 처음 하게 되었습니다

다. 프로젝트 시작 초기에 주제선정에서 창의적이고 독창적인 주제를 선정하는 것에 어려움이 있었습니다. 좋은 결과물을 내기 위해 여러 가지 주제를 생각해 서 주제를 선정하였고, 이후에도 구체적인 주제로 구상하기 위해 노력했습니다.

그리하여 선정주제를 다듬어가며 계획을 준비했고, 소프트웨어 공학에서 배운 개발 방법을 보가며 프로젝트 계획, 분석, 설계 등의 단계를 진행하였습니다. 그동안 배우기만 해왔던 공학설계 방법을 실제 프로젝트에 개발에 적용하게 되면서 이론으로 뿐만 아니라 실제로 프로젝트 기획에 있어 여러 가지 요소를 빠지지 않고 착실하게 수행해야한다는 것을 많이 깨닫게 되었습니다.

단순히 머릿속에서 추상적으로 생각되는 내용을 구현 하는 게 아니라 체계적인 팀 협업과 의사소통 속에서 계획해나가고 내가 맡은 일을 해내서 팀의 업무가 원활히 진행될 수 있도록 노력하였습니다.

이번 프로젝트를 진행하게 되면서 기획과 분석이 잘된다면, 구현과정에서 겪게 되는 어려움이 없게 되고, 결국 기획과 분석이 프로젝트 진행에 핵심적인 역할을 쥐고 있다는 걸 생각하게 되었습니다. 이번 경험을 통해 앞으로 팀 프로젝트를 하게 된다면 팀원 간의 협동을 통해 좋은 시너지를 발휘 할 수 있습니다. 노력하면 할 수 있다는 자신감으로 늘 열심히 할 수 있는 동기가 되었습니다.

IV. 작품의 기대효과 및 활용분야

※ 평가항목 : 기획력 (활용가능성)

1. 작품의 기대효과

- 대표적으로 아르바이트에 대한 여러 정보를 제공하여 자신에게 알맞은 아르바이트를 결정할 수 있도록 도움을 준다.
- 아르바이트에 대한 상관관계를 보여주며 가장 자신의 생활환경에 맞는 아르바이트를 찾을 수 있도록 도와준다.
- 아르바이트 사장의 경우 아르바이트생들의 시급을 얼마정도 주어야 할지 가늠할 수 있게 되며 근무지 환경의 고충이 소통이 되어 개선이 가능하다.
- 작은 데이터 하나하나를 모아 적절하게 분류하여 큰 데이터를 추출하여 알바 통계 분석 자료가 필요한 사용자에게 제공해 준다.
예시로 통계학과 또는 빅 데이터 공부중인 학생들이 사용할 자료가 만들어진다는 것.
- 빅 데이터나 분석 알고리즘을 몰라도 자신이 데이터를 삽입하여 분석해 볼 수 있다.
- 설계한 극값 데이터마이닝 기법이 논문으로 작성되어진다.
- 구인 구직 사이트에 광고효과를 가져온다.

2. 작품의 활용분야

- 아르바이트 구직자들에게 아르바이트 정보를 제공
- 아르바이트생을 구하는 기업들의 시급 금액 측정 정보 제공
- 아르바이트생들의 정보공유, 의사소통 및 정보의 축적
- 임의의 정보를 통계적 기법을 사용하여 자신의 정보를 분석 및 미래 예측
- 통계학과 또는 빅데이터 학습자료로 활용
- 분석 기법 설계로 논문 작성

V. 개발산출물

※ 평가항목 : 평가 전반에 참고

○ 시스템의 기능



< 그림18. 시스템 기능 >

시스템의 기능은 크게 7가지로 구성된다. 아르바이트 데이터 수집, 회귀분석을 통한 정보, Top N을 보여주는 랭킹정보 리스트, 구인구직 사이트 정보, 데이터 분석기, 알바 데이터 자료실, 분석 기법을 소개해주는 분석 방식 정보이다.

아르바이트 데이터 수집은 중요한 것이 ‘신뢰도’ 인데 실제로 알바 천국 (www.alba.co.kr/) 사이트에 구인 구직중인 정보들을 웹 크롤링 방법을 실시하여 최근의 정보를 수집함으로 써 신뢰도를 높였다. 1000000개 가량의 데이터가 수집되었다. 또한 사용자들에게 아르바이트 정보를 입력하면 그 데이터가 오류가 없는 데이터인지 분류를 하여 신뢰도를 높인다.

아르바이트 회귀분석을 통한 정보 제공은 다양한 변수에 대해 선택해보며 두 변수간의 상관관계를 느껴볼 수 있도록 구성하였다.

예를 들면 나이에 따른 시급이라 하면 나이가 들면 들수록 시급이 오를 것 인가? 라는 결과가 분석결과로도 선형적으로 증가하고 있었다.

이처럼 변수를 선택하여 시각화 가능하도록 변수선택을 할 수가 있으며 추가로 댓글을 통하여 자연어로 작성되어진 분야별 아르바이트 정보 등을 형식 없이 공유할 수 있도록 구현하였다.

아르바이트 랭킹정보는 가장 최근에 분석된 아르바이트의 평균적인 정보를 테이블로 시각화 하여 자신이 원하는 우선순위를 선택하여 정렬할 수 있도록 구성되었다. 자신이 원하는 것이 ‘시급’ 이라면 시급을 선택하여 시급이 높은 순으로 정렬할 수가 있다.

구인구직 사이트 정보는 구인구직사이트는 다양한데 특정한 아르바이트는 어떤 사이트에 특화되어 있는 경우가 있다.

예를 들면 학원강사의 경우 ‘알바천국’ 등의 사이트에서 찾아봐도 자신의 지역에서 구하기 힘든 경우가 많다. 그러나 ‘훈장마을’ 이라는 사이트에서 찾게 되면 나의 경우 올린 다음날 하루만에 전화가 7통 왔던 기억이 있다. 즉 이러한 정보도 찾아보지 않으면 알기 힘들다는 것이다.

또한 사이트 별로 사이트 이미지를 누르게 되면 해당 페이지로 이동가능 하도록 링크를 설정해놓았다.

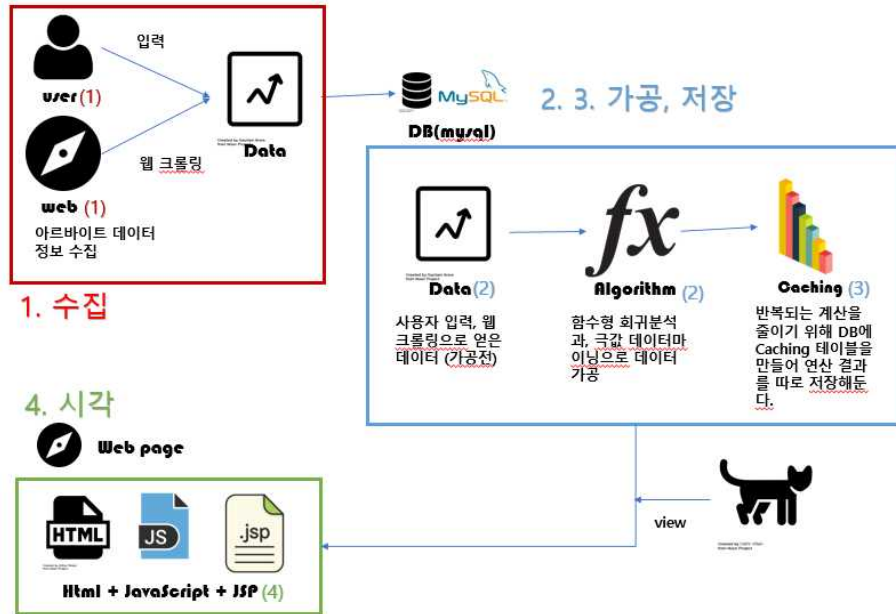
데이터 분석기는 사용자들이 분석해보고 싶은 데이터를 빅데이터 분석방식이나 통계학과 학생이 아니더라도 분석해 볼 수 있는 시스템이다. 분석방식은 선형, 2차, 단조 증가 3차, 함수형 등 회귀분석방식으로 가능하고 원래 존재하는지는 모르겠지만 독자적으로 개발한 ‘극값 데이터 마이닝’ 기법을 이용하여 ‘기울기’, ‘선형적’ 이라는 특징들을 적용시켜 중간에 존재하지 않는 데이터나 미래를 추측할 수 있게 된다.

분석기의 활용방안으로는 나의 매달 수익 등을 삽입시키면서 저장해 놓으면 분석되어 다음달에는 내가 얼마의 수익이 있을지 예측이 가능하다는 것이다. 이는 다양한 방도로 활용이 가능하다.

알바데이터 자료실은 통계학과 학생들이나 빅 데이터를 공부하고 있는 학생들은 반드시 대량의 데이터가 필요하게 된다. 하지만 실제로 데이터를 구하기란 사람들이 공유하지 않아 구하기가 쉽지 않다. 따라서 알바셀파에서 수집된 데이터를 사용자들에게 익명 데이터 자료로 제공함으로써 교육적 자료로 활용할 수 있게 된다.

분석기법 소개는 사람들이 어떤 ‘관계성’ 에 대해 모르고 회귀분석 자료를 보게 된다면 뭔지도 모르고 분포만 보게 될 가능성이 높다. 따라서 간단하게 결과를 이해할 수 있는 소개 페이지를 구성하고 더 나아가 분석방식의 깊은 내용까지 소개를 한다. 이 또한 교육적 자료로 활용이 가능하다.

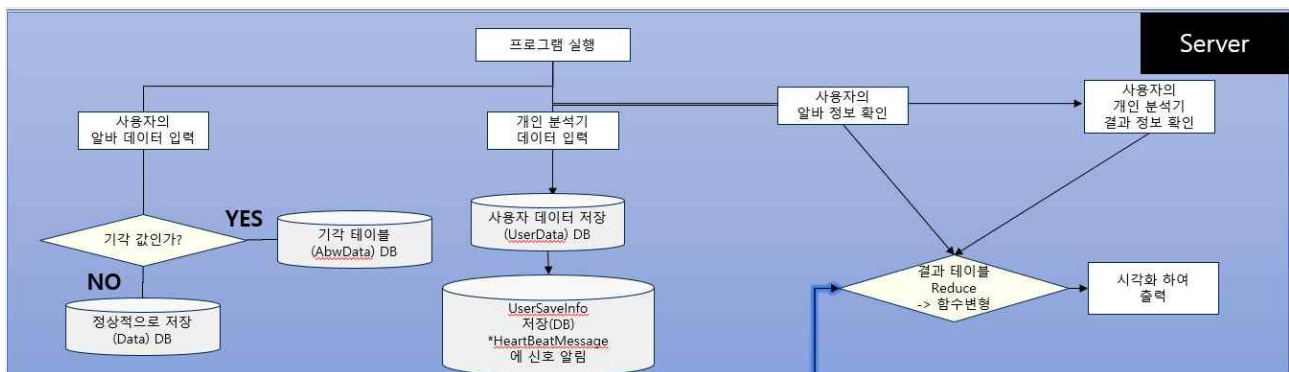
○ 시스템 구성도



<그림19. 시스템 구성도>

웹 크롤링은 Java의 URL클래스를 이용하여 크롤링을 실시하였고 가공, 결과 저장방식으로는 Caching테이블과 HeartBeatMessage라는 Thread 클래스가 중요하게 작용하는데 Caching테이블은 수집된 알바 데이터의 양이 1000000개라 하면 그것을 사용자들이 ‘새로고침’ 할 때마다 알고리즘을 돌려 분석하고 시각화 하게 되면 서버의 과부하가 일어날 것이다.

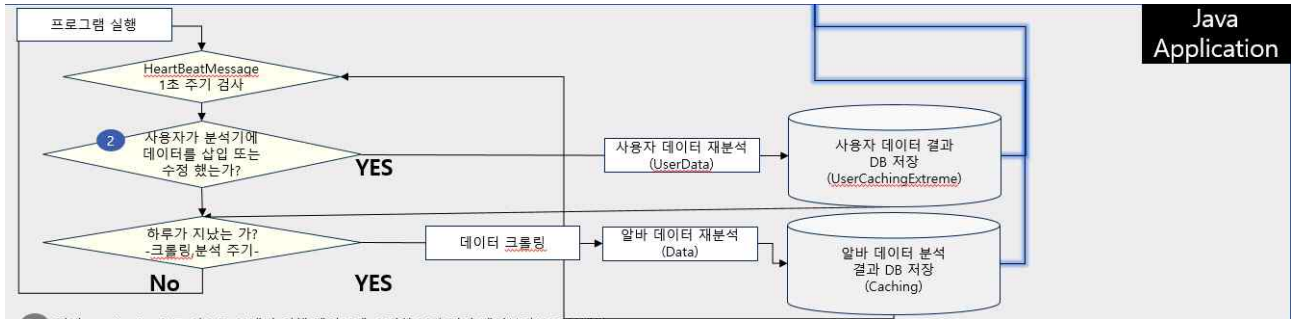
따라서 여기서는 과부하를 방지하기 위해 주기적으로 분석하고 그 결과를 저장하여 Caching이라는 테이블에 저장하게 된다.



<그림 20. 사용자 서버 흐름도>

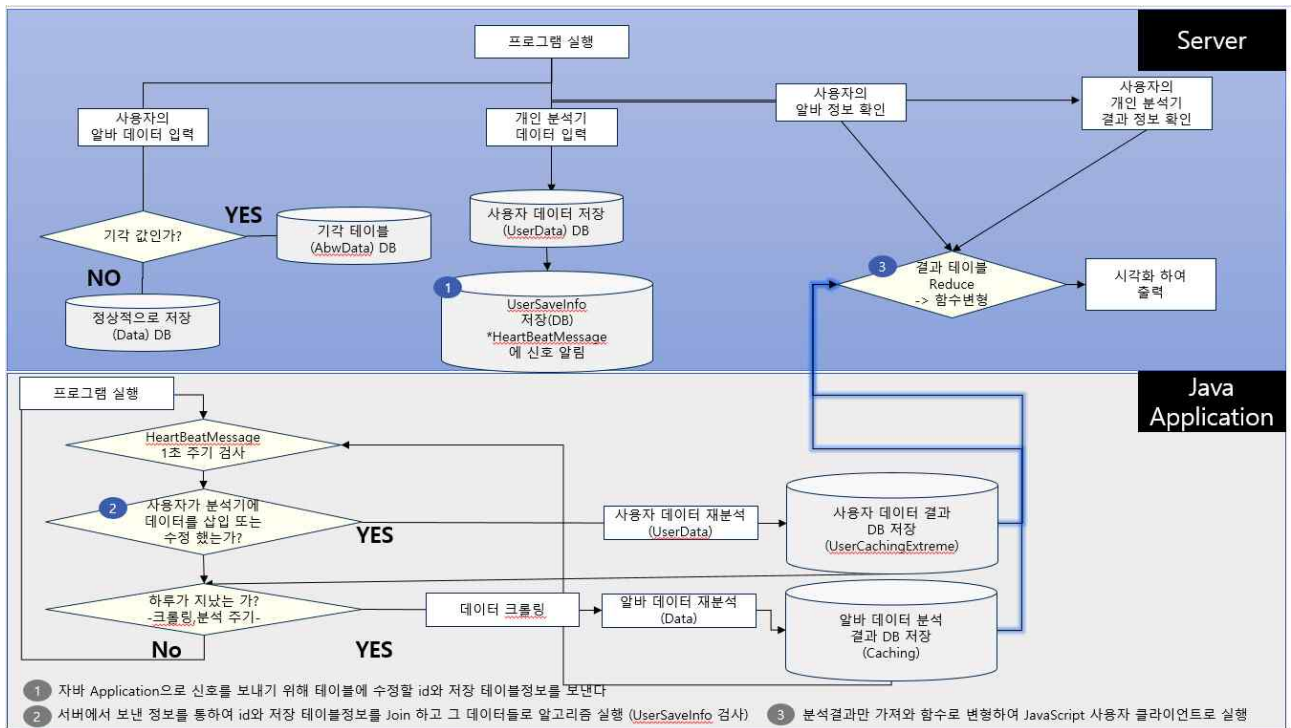
다음과 같이 JSP의 Tomcat 서버에서는 사용자들이 입력하는 데이터를 저장하고 결과를 가져오거나 요청만 할뿐 알고리즘 계산을 하지 않도록 구성하였고 또한 사용자 분석기에서도 적용이 되는데 사용자들이 데이터를 입력하거나 수정하게 되면 동시에 분석을 실행하라고 요청한다. (Java Application과의 통신)

이렇게 수정된 테이블은 HeartBeatMessage에서 체크하여 잡아내고 자바 어플리케이션이 알고리즘을 실행하여 그 결과 함수들을 저장하게 된다.



<그림 21. JavaApplication (HeartBeatMessage) 요청 순서도>

하둡 내용 중 HeartBeatMessage가 빅데이터 HDFS의 네임노드와 데이터 노드간의 통신방식인 HeartBeat가 떠올라 클래스 명 또한 HeartBeatMessage 로 정하게 되었다. 이는 1초주기 간격으로 현재 내가 해야 하는 일을 알려주는 역할을 하며 비교적 간단한 연산만 하므로 Thread를 이용하여 구현하였다. 이렇게 만약 요청이 들어오면 자바 어플리케이션이 알고리즘을 실행하여 분석 결과를 데이터베이스에 담는 다.



<그림 22. 시스템 순서도>

결론적으로 JavaApplication과 Tomcat을 통한 서버가 서로 통신하기 위해 HeartBeatMessage가 실행되고 이로 인한 효과는 분석기와 아르바이트 데이터 분석이 새로고침 할때마다 알고리즘을 통해 분석 되는 것이 아닌 모두 최초 1회 실행 되어 결과만 보여주게 되고 사용자에게 빠르게 응답할 수 있게 되는 장점을 가져온다.

FunctionName	degree	w	centerPointX	centerPointY	cost	extremePointX1	extremePointY1	extremePointX2	extremePointY2	coef1	coef2	coef3
선형모형	1	0.536568506102937	11	20.7777777777778	9706.30987037318	0	0	0	0	-9999999	-9999999	-99999
2차함수모형	2	0.0299412297044734	1	3	11869.8861084474	0	0	0	0	-9999999	-9999999	-99999
중간 3차함수 모형	3	-0.00136743643848599	11	20.7777777777778	9985.7050962446	0	0	0	0	-9999999	-9999999	-99999
기울기 극값 데이터마이닝	-9999999	-9999999	-9999999	-9999999	-9999999	1	3	3	3	0	3	3
기울기 극값 데이터마이닝	-9999999	-9999999	-9999999	-9999999	-9999999	3	3	11	111	0	33	33
기울기 극값 데이터마이닝	-9999999	-9999999	-9999999	-9999999	-9999999	11	111	20	18	0	220	-15
기울기 극값 데이터마이닝	-9999999	-9999999	-9999999	-9999999	-9999999	20	18	22	33	0	440	-
기울기 극값 데이터마이닝	-9999999	-9999999	-9999999	-9999999	-9999999	22	33	24	20	0	528	-
기울기 극값 데이터마이닝	-9999999	-9999999	-9999999	-9999999	-9999999	24	20	31	22	0	744	-27
기울기 극값 데이터마이닝	-9999999	-9999999	-9999999	-9999999	-9999999	31	22	35	18	0	1085	-
선형 극값 데이터마이닝	1	0.536568506102937	11	20.7777777777778	9706.30987037318	1	3	3	3	0	3	3
선형 극값 데이터마이닝	1	0.536568506102937	11	20.7777777777778	9706.30987037318	3	3	11	111	0	33	33
선형 극값 데이터마이닝	1	0.536568506102937	11	20.7777777777778	9706.30987037318	11	111	20	18	0	220	-15
선형 극값 데이터마이닝	1	0.536568506102937	11	20.7777777777778	9706.30987037318	20	18	22	33	0	440	-
선형 극값 데이터마이닝	1	0.536568506102937	11	20.7777777777778	9706.30987037318	22	33	20.9816433377811	33	0	461.596153431183	-21.490821668899
선형 극값 데이터마이닝	1	0.536568506102937	11	20.7777777777778	9706.30987037318	20.9816433377811	33	16.9082166889053	-21.9912597598223	0	354.762172044529	-18.944930011334
선형 극값 데이터마이닝	1	0.536568506102937	11	20.7777777777778	9706.30987037318	16.9082166889053	-21.9912597598223	12.3256117089201	25.362325033358	0	208.404113597731	-14.61691419891
선형 극값 데이터마이닝	1	0.536568506102937	11	20.7777777777778	9706.30987037318	12.3256117089201	25.362325033358	11.3072550467012	17.724650066716	0	139.368835199367	-11.816433377811
선형 극값 데이터마이닝	1	0.536568506102937	11	20.7777777777778	9706.30987037318	11.3072550467012	17.724650066716	7.73382836787548	31.9816433377811	0	81.7947476587878	-4.270541727263

<그림 23. 사용자 분석기용 Caching테이블>

이런 형태로 분석한 결과 함수에 대해 저장이 되며 특히 ‘극값 데이터 마이닝’ 부분은 다항식의 계수와 좌표들이 여러 개 존재하기 때문에 이를 Reduce하여 하나의 함수 클래스로 변형시켜 시각화 하게 된다.



<그림 24. 테이블 Reduce를 통한 함수 결과 시각화>

결과적으로 시각화로는 Bootstrap을 이용하여 다양한 방식으로 통계자료를 시각화 하였고 JSP,HTML,CSS등으로 페이지를 개발하고 JavaScript,JQuery로 동적 페이지를 완성시켰다.JDBC를 통하여 MySql과 통신하였고 웹서버는 Tomcat을 이용하였다.

추가로 위 내용을 토대로 논문 2부 제출 및 외부 컨퍼런스 발표를 하였다.

-KIPS 일어날 사건의 사전 예측 분석을 위한 극값 데이터 마이닝 알고리즘

-KIPS 함수모형 회귀 분석 및 알고리즘

데이터야 놀자 10.13일 발표 : 김석준

함수모형 회귀분석 및 알고리즘

김석준*, 장근호**, 김예기*
*한신대학교 수학과 수리금융학과
e-mail : raeloc7607@naver.com

Function Regression algorithm

Seok Jun Kim*, Geon Ho Jang**, Ye Ji Kim*
*Dept of Mathematics, Han-Shin University
**Dept of Finance Mathematics, Han-Shin University

요 약

Linear Regression 문제를 토대로 변형하여 선형회귀분석, 2차함수모형 회귀분석, '단조 증가(감소)' 3차 함수모형 회귀분석과 그에 따라 변형되는 gradient descent 알고리즘을 기술한다.

1. 서론

데이터를 분석하여 보여줄 때 모든 데이터를 '선형적 회귀분석'으로만 분석하여 보여주어야 할까? 라는 생각으로 시작하여 LinearRegression 문제를 기반으로 선형, 2차 함수모형 회귀분석, '단조 증가(감소)' 3차 함수모형 회귀분석 알고리즘을 구현 하였다.

회귀분석은 선형일뿐 아닐지도 시작하지만 함수모형을 선형일지 따라 그 모형의 특징을 살릴 수 있기 때문이다. 예를 들면 2차 함수모형은 감소하다가 증가하는 특징, '단조 증가(감소)'하는 3차 함수모형은 증가하다가 감소하게 되는 특징을 살릴 수 있다는 것이 이 논문의 논제이다.

2. Linear Regression

<표 1> 선형회귀

선형 회귀분석
두 데이터 값(x,y)간의 선형적 상관관계를 파악한다. 예시) 전력에 따라 시금치 오름가? => 분석식 함수 f(x)로 도출 => 분석식 함수 f(x)로 도출 => 분석식 함수 f(x)로 도출
일반적인 선형회귀분석을 기계학습을 통해 구현 할 수 있다.

Linear Regression 문제를 말한다.
통계학적으로는 '선형 회귀분석'의 모델을 컴퓨터가 학습 할 수 있도록 설계하여 함수 값을 갖는 두 개 이상의 변수를 갖는 데이터가 주어지면 그 변수간의 상관관계를 학

습시켜 가장 적합한 하나의 직선을 찾아내는 문제이다.
여기서 가장 적합한 하나의 직선이 되는 기준은 cost라는 비용으로 함수와 데이터 간의 차이(편차)의 제곱의 평균 (분산과 비슷하다) 생각하면 된다.)가 가장 작을 때의 함수가 가장 적합하다는 기준이 되며 gradient descent 알고리즘을 통해 적당 함수를 찾아낸다.

3. 2차 함수모형 회귀분석

<표 2> 2차 함수모형

2차 함수모형 분석
어떤 데이터가 증가하다(+) 감소(-)또는 감소하다(-) 증가(+)하는 데이터라면 2차 함수모형이 적합하다
선형회귀를 변형하여 2차 함수모형으로 기계학습을 시켜 분석한다.
2차 함수모형 알고리즘 위에서 '선형적'이란 단어의 뜻은 그래프 관점에서 보면 '곡선적'인 관계로 볼 수 있다. 그러나 모든 데이터의 분포를 선형적으로만 분석하게 된다면 문제가 있다고 보았다. 따라서 비선형적인 모델로 분석하기 위해 경의력이 주어진 전사함수를 선정하게 되었고 가장 일반적인 함수들을 채택하게 되었다.

일어날 사건의 사전 예측 분석을 위한 극값

데이터 마이닝 알고리즘

김석준*, 장근호**

*한신대학교 수학과
e-mail : raeloc7607@naver.com

Extreme Data Mining Algorithm for Pre-Predictive

Analysis of Occurrences

Seok Jun Kim*, Geon Huh**

*Dept of Mathematics, Han-Shin University

요 약

데이터간의 '극값'이란 매우 중요한 정보로 사용이 가능하다. 생각하여 최근의 극값 특징을 기준으로 과거의 극값들과의 그 특징이 유사한 극값을 찾고 다음에 일어날 극값이 언제 일어날지 분석하는 기법이다.

1. 서론

데이터간의 '극값'이란 매우 중요한 정보로 사용이 가능하다. 예를 상승하다가 하락하게 되는 전황점으로 볼 수도 있게 되는데 이러한 전황점의 중요 사항으로는 주식의 상승에서 하락으로 전환되는 전환점 또는 나의 수익이 언제까지 상승하다가 떨어지게 될지, 등으로 적용될 것이라 보고 이 알고리즘이 실제로 데이터에 적용된다면 그에 따라 알맞게 대비할 수가 있다.

이해를 돕기 위한 예로는 주식으로 치면 현재 주식을 매수,매도 해야하는 지를 파악한다는 것인데 아직 해당 알고리즘으로 실제 데이터를 실험해보지는 못하였다. 따라서 여기서는 극값들을 분석해 다음에 일어날 극값(전환점)이 언제 일어날지 분석하는 것이 주제이다.

2. 극값 데이터 마이닝 (기반 아이디어)

먼저 데이터를 2차원 평면에서 분포시켜 볼 때 극값이란 상승하다 하락 또는 하락하다 상승하는 어떤 전환점 같은 의미로 생각할 수 있다. 순서는 다음과 같다.

- 1) key값으로 오동차순 정렬한다,
- 2) 데이터들의 key값 별로 +군집화하여 따른 value값들을 모두 평균을 낸 (key_avgValue) 형태로 만든다,
- 3) 만들어진 평균값들을 차례로 스캔하며 avgValue값이 증가하다가 감소하게 될 때 또는 감소하다 증가하게 될 때를 확인해 좌표들을 '극 값'이라 정한다,
- 4) 이렇게 모은 극 값 중 가장 최근에 떨어진 극값을 두고 그 극값의 특징을 선정할 뒤 (그 특징 값을 기준 pivot)으로 삼는다,
- 5) 나머지 극값들 또한 앞에서 선정 한 특징 값을 각각 구해 기준(pivot)의 특징과 가장 유사한 위치를 찾아내고 그

위치의 뒤를 미러로 보고 마지막에 이어 붙인다.
5) 각 극값들을 2개의 경우 3차함수의 미적분론 통해 연 결하여 하나의 함수 f(x)로 만들어낸다.

먼저 이 알고리즘의 핵심은 가장 최근에 일어나는 특징(현상)을 선택해 과거에 그 특징(현상)과 가장 유사했던 경험들 토대로 미래를 분석한다는 것이 주제이다.

2차에서 +군집화과 곡선한 이유는 만약 key값이 너무 많거나 경우가 아닌 실수 값을 경우를 생각하여야 한다.

이때는 각 key값들을 범위를 정하여 계급별로 모아 그 계급 값(계급의 중앙 값)을 key값으로 모은 데이터들의 전체 value값들의 평균을 value값으로 처리하여 구 할하여야 한다. 여기서는 그 특징(현상)을 '기울기'와 '선형 회귀분석'을 통해 특징(현상)을 선정하여 구현하였다.

3. '기울기'를 이용한 극값 데이터마이닝

<표 1> '기울기'를 이용한 극값 데이터 마이닝

'기울기'를 이용한 극값 데이터 마이닝
각 key값에 따른 value값들의 평균을 대어 그 평균 값들을 '극 값'이라고 생각되는 평균값만 뽑아 내 가장 (<=> 바로 전 구간)의 기울기와 과거의 기울기를 통하여 유사한 구간을 찾아내고 다음번 세 극값이 언제 일어날지를 추적하여 추가한 뒤 3차 함수의 미적분론을 통하여 연결한 모형
'빠르게 오르면 빠르게 떨어질 것' 또는 '약하게 밀면 약하게 펴겨 나간다' 등지 최근의 기울기를 보고 과거에 있던 패턴을 가늠해 뒤에 다 연결하게 되는 방식으로 분석된다.

<그림 25. KIPS 학술대회 논문>

(함수 모형 회귀)

< 그림 26. KIPS 학술대회 논문 >

(극값 데이터 마이닝)