7.7 분석 파일럿 실행 5단계

머하웃과 스파크ML을 이용한 머신러닝

머하웃과 스파크ML 같은 머신러닝 기술은 복잡도가 높은 비즈니스 로직을 자동으로 생성 및 관리하거나, 대규모 단순 반복 작업에서 패턴들을 찾아 효율화하는 데 사용 된다. 이때 자동으로 만들어진 프로그램을 모델이라고 하며, 모델은 대규모 데이터에서 과거의 패턴을 찾아 정의하는 학습 과정을 통해 만들어진다. 학습이 완료된 모델에 현재의 데이터를 입력해서 앞으로 발생할 일들을 예측하면서 신속한 의사결정을 내리도록 지원한다. 이번 장에서는 지금까지 수집, 적재, 처리한 스마트카의 데이터셋을 가지고 세 가지 마이닝 기법인 추천, 분류, 군집 기능을 머하웃(추천)과 스파크ML(분류, 군집)을 이용해 좀 더 활용성 있는 분석을 진행한다. 참고로 스파크ML의 작업 환경으로 제플린을 사용한다.

7.7 분석 파일럿 실행 5단계–머하웃 추천

스마트카 차량용품 추천

머하웃 추천 – 스마트카 차량용품 추천 _ ⌨ 실습

데이터 마이닝의 추천에 사용될 데이터셋은 “스마트카 차량용품 구매 이력” 정보로서 하이브의 Managed 영역에 있는 Managed_SmartCar_Item_BuyList_Info 테이블에 약 10만 건의 데이터가 적재돼 있다.

	car_number	sex	age	marriage	region	job	car_capacity	car_year	car_model	item	score	biz_month
0	V0067	남	19	기혼	광주	자영업	2000	2011	E	Item-022	2	201606
1	U0084	여	61	기혼	경남	회사원	3000	2012	D	Item-004	3	201606
2	U0006	여	35	미혼	강원	공무원	1200	2005	H	Item-009	2	201606
3	X0052	여	40	미혼	광주	학생	3000	2004	F	Item-006	5	201606
4	V0039	여	66	미혼	경기	자영업	2500	2015	G	Item-028	5	201606
5	C0024	여	51	미혼	울산	회사원	2000	2012	C	Item-023	2	201606
6	C0017	남	56	미혼	경북	학생	3000	2001	H	Item-023	5	201606
7	D0067	남	55	미혼	전북	학생	1200	2008	B	Item-010	3	201606
8	K0096	남	32	미혼	경기	자영업	1200	2001	E	Item-005	1	201606
9	Q0028	남	59	미혼	전북	자영업	1700	2005	F	Item-010	2	201606

7.7 분석 파일럿 실행 5단계–머하웃 추천

스마트카 차량용품 추천

구매일	구매자	상품명	상품코드	평가점수
9월1일	사용자 A	카시트	Item – 001	5
9월5일	사용자 A	충전기	Item – 002	4
9월6일	사용자 A	거치대	Item – 003	3
9월7일	사용자 B	카시트	Item – 001	5
9월7일	사용자 B	거치대	Item – 003	3
⋮				

유사도 측정

유사도 측정

7.7 분석 파일럿 실행 5단계–머하웃 추천

스마트카 차량용품 추천

01. 먼저 “스마트카 용품 구매 이력” 데이터를 머하웃의 추천기에서 사용 가능한 형식으로 재구성한 파일을 만들어야 한다. 휴의 Hive Editor에서 다음 QL을 실행한다.

```
1
2 insert overwrite local directory '/home/pilot-pjt/mahout-data/recommendation/input'
3 ROW FORMAT DELIMITED
4 FIELDS TERMINATED BY ','
5 select hash(car_number), hash(item), score from managed_smartcar_item_buylist_info
```

7.7 분석 파일럿 실행 5단계-머하웃 추천

스마트카 차량용품 추천

02. 추천기의 입력 데이터로 사용될 파일이 정상적으로 만들어졌는지 확인한다. Server02에 SSH로 접속하고 다음 명령을 실행한다.

```
$ more /home/pilot-pjt/mahout-data/recommendation/input/*
```

```
[root@server02 input]# more /home/pilot-pjt/mahout-data/recommendation/input/*
80900631,1240943830,2
79977169,1240943770,3
79976923,1240943775,2
82747637,1240943772,5
80900540,1240943836,5
63353605,1240943831,2
63353577,1240943831,5
64277253,1240943797,3
```

7.7 분석 파일럿 실행 5단계-머하웃 추천

스마트카 차량용품 추천

03. 이제 앞서 생성한 "000000_0" 파일을 머하웃 추천기의 입력 데이터로 사용하기 위해 HDFS에 /pilot-pjt/mahout/recommendation/input/ 경로를 생성하고 "000000_0" 파일을 저장한다.

```
$ hdfs dfs -mkdir -p /pilot-pjt/mahout/recommendation/input
$ hdfs dfs -put /home/pilot-pjt/mahout-data/recommendation/input/*  /pilot-pjt/mahout/
recommendation/input/item_buylist.txt
```

7.7 분석 파일럿 실행 5단계-머하웃 추천

스마트카 차량용품 추천

04. 이제 머하웃의 추천 분석기를 실행한다. 머하웃의 주요 명령은 C://예제소스/bigdata2nd-master/CH07/Mahout/에 있으니 활용하도록 한다.

```
$ mahout recommenditembased -i /pilot-pjt/mahout/recommendation/input/item_buylist.txt -o
/pilot-pjt/mahout/recommendation/output/ -s SIMILARITY_COOCCURRENCE -n 3
```

사용된 매개변수와 옵션은 다음과 같다.

- i: 추천 분석에 사용할 입력 데이터
- o: 추천 분석 결과가 출력될 경로
- s: 추천을 위한 유사도 알고리즘
- n: 추천할 아이템 개수

```
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=437274
        File Output Format Counters
                Bytes Written=179641
16/12/19 22:20:12 INFO driver.MahoutDriver: Program took 27852 ms (Minutes: 0.4642)
```

7.7 분석 파일럿 실행 5단계–머하웃 추천

스마트카 차량용품 추천

05. 분석 결과가 저장된 HDFS의 /pilot–pjt/mahout/recommendation/output/에 있는 파일을 휴의 파일 브라우저로 열어서 확인한다. 해당 경로에는 두 개의 파일인 "part–r–00000", "part–r–00001"이 생성돼 있을 것이다. 이 가운데 하나를 열어서 추천 결과를 확인한다.

7.7 분석 파일럿 실행 5단계–머하웃 추천

스마트카 차량용품 추천

추천받은 차량번호	첫 번째 – 추천상품 상품ID: 추천값	두 번째 – 추천상품 상품ID: 추천값	세 번째 – 추천상품 상품ID: 추천값
61506498	[1240943774:4.381886	1240943859:4.3754354	1240943769:4.371212]
61506500	[1240943770:4.4761906	1240943798:4.47266	1240943775:4.4653916]
61506502	[1240943767:4.5166845	1240943805:4.510496	1240943771:4.5015078]
61506504	[1240943767:4.7208376	1240943768:4.7191577	1240943833:4.7176113]
61506506	[1240943837:4.4397306	1240943769:4.438384	1240943802:4.435185]
61506528	[1240943836:3.7480755	1240943767:3.7338078	1240943806:3.7289279]
61506530	[1240943833:3.8017108	1240943772:3.7908888	1240943805:3.7882187]

7.7 분석 파일럿 실행 5단계-머하웃 추천

스마트카 차량용품 추천

06. 추천 분석을 재실행할 때는 기존 결과 파일을 삭제한 후 재실행해야 한다. 아래 명령은 관련 삭제 명령이니 참고하기 바란다. 이후 분류, 군집 분석에서도 같은 명령을 중복을 실행할 때 이미 존재하는 파일(경로)이라는 에러가 발생할 수 있다. 그때는 해당 경로의 파일을 삭제한 후 재실행한다.

```
$ hdfs dfs -rm -R -skipTrash /pilot-pjt/mahout/recommendation/output
$ hdfs dfs -rm -R -skipTrash /user/root/temp
```

7.7 분석 파일럿 실행 5단계-머하웃 추천

스마트카 차량용품 추천

실습