K-Means 알고리즘과 베이시안 네트워크를 활용한 경로추천시스템

정용규, 류종범, 최병렬
한국공학대학교 컴퓨터공학과

Route Recommendation System Using K-Means and Bayesian Network

YongGyu Jung, JongBeom Ryu, ByoongYoul Choi
Tech University of Korea Computer Engineering

요 약

본 논문에서는 기존 길 안내 어플리케이션이 사용자의 시간적인 비용과 정보의 한계로 발생한 역 선택 문제를 발생한다는 문제의식을 바탕으로 K-Means 알고리즘과 베이시안 네트워크(Bayesian Network)를 활용한 어플리케이션을 개발하였다. 본 어플리케이션은 기존 어플리케이션의 길 안내 시스템을 바탕으로 사용자화(User customization)에 초점을 두었다. 크롤링을 통한 데이터 정보 추출 및 가공, K-means 알고리즘을 바탕으로 군집화(Clustering), 베이시안 네트워크(Bayesian Network) 활용하여 사용자의 니즈(needs)에 맞는 경유지를 추천하고 사용자의 사용이 누적될수록 사용자에게 보다 적합한 경유지를 추천할 수 있도록 하였다.

1. 서론

기존 길 안내 어플리케이션은 출발지와 목적지의 다양한 경로를 제공하는데 초점이 되어있다. 이는 편리한 기능을 제공하는 것은 분명하지만, 사용자의 목적지가 다양한 경우나 사용자가 중간에 다른 목적지를 필요로 하는 경우에 효율적이지 못한 측면이 있다. 가령 사용자가 여행을 목적으로 특정 지역을 가정할 때, 사용자가 특정 목적지 외에 방문할 의사가 있는 경우에는 사용자가 사전에 가는 경로 도중에 다른 경유지를 찾아야 하는 시간적인 비용 문제와 정보의 한계로 인한 역 선택 문제가 발생한다. 사용자가 중간에 다른 목적지를 필요로 하는 경우에도 비슷한 문제가 발생한다. 이동 중에 사용자는 시간적인 자원이 제한되기 때문에 비슷한 상황에 직면할 확률이 높다.

길 안내 시스템이 효율적인 동선으로 사용자의 시간적 비용을 줄여준다는 점에 초점을 둘 때, 기존의 길 안내 시스템이 목적을 다하는 것처럼 보이지만 사용자화(User customization) 측면에서 불리한 점이 있어 효율적이지 못한 측면이 존재한다. 사용자에게 정해진 출발지에서 목적지까지 정확하고 효율적인 경로를 안내하는 것은 분명하지만, 사용자의 니즈(needs)에 따라 이것이 과연 개개인에게 효율적이고 합리적인 동선인지 불분명하다. 이를 해결하기 위해 본 논문에서는 K-Means 알고리즘과 베이시안 네트워크(Bayesian Network)를 통해 사용자화를 실현함으로써 시간적인 비용 문제와 정보의 한계로 인한 역 선택 문제를 해결한 경로 추천 시스템을 제안하고자 한다.

2. 관련 연구

2.1 RFM(Recency Frequency Monetary)

RFM은 시장 분석 기법의 하나로 고객의 미래 구매 행위를 예측할 때, 과거의 구매내용을 기반으로 예측하는 기법이다. [1] RFM은 최근성(Recency), 빈도성(Frequency), 총구매액(Monentary)을 각각 의미하는데, RFM에 대해 각각 가중치 부과가 가능하다.

$$RFM\ Total = (R\ weight) \times R + (F\ weight) \times F + (M\ weight) \times M$$

본 논문에서는 장소에 대해 RFM 기법을 유사 적용하여 사용자 데이터를 활용하여 최근성, 빈도성, 평점을 기준으로 산출된 RFM 점수를 활용한다. 산출된 RFM 점수는 사용자의 사용횟수가 누적될수록 사용자의 경향성을 수치화할 수 있는 장점이 있다. 이는 경유지 간의 선호를 구분할 수 있는 비교정보를 제공하고, 사용자의 변화하는 성향을 따라갈 수 있다.

2.2 K-Means Algorithm

K-Means 기법은 객체의 특성을 나타내는 속을 기반으로 클러스터링을 통해 K개의 군집으로 나누는 것이다. [1] 보다 자세히 말하자면, 각 객체를 가장 가까운 중심점에 할당하는 과정을 반복하여 K개의 군집으로 나누는 것이다. 이 과정에서 같은 군집 내의 객체와는 유사도는 증가하고 다른 군집 내의 객체와는 유사도가 감소한다. [2] K-Means은 유사도 기반으로 객체를 집합으로 만든다는 측면에서 객체들을 레이블을 해주고 객체의 적합성을 비교해주는 장점이 있다.

본 논문에서는 K-Means 알고리즘이 가진 이런 장점을 이용해 각 객체 즉 경유지를 향용사 테그, 날씨, 온도, 분류, 등을 기반으로 클러스터링을 하고 사용자에게 가장 적합한 경유지를 추천하는데 활용하고자 한다.

2.3 Bayesian Network

베이시안 네트워크는 사전에 일어난 일을 기반으로 사후의 확률을 추론하는 방법이다. [1] 이는 사람의 선택의 성향과 경향을 반영한다는 측면에서 인간의 의사결정 과정과 유사한 측면이 있다. 베이시안 네트워크는 속성들이 서로 독립적인 가정이 전제되는데 이런 가정 덕분에 알고리즘이 단순하고 견고해지는 장점이 있다. 이런 특징 때문에 베이시안 네트워크는 대규모 데이터베이스를 빠르고 정확하게 처리할 수 있다. 특징을 표현할 수 있는 n차원의 T 벡터와 m개의 선택지를 Di 클래스라 할 때 베이시안 네트워크의 식은 다음과 같다. [2]

$$P(Di|T) = \{P(T|Di) \times P(Di)\}/P(T) \ (1<=i<=m)$$

베이시안 독립 가정 하에,

$$P(T|Di) = P(Di)\prod P(T|Di) \ (i\text{은 }1\text{부터 }n\text{까지})$$

위 식에서 P(T)의 값은 항상 일정하므로 P(T|Di) x P(Di)의 값이 비교 기준이 된다. [2] 이 값이 높을수록 그 선택지를 선택할 확률이 높다고 볼 수 있고, 가장 높은 값은 최종선택일 가능성이 높다. 데이터를 가장 높은 값을 기준으로 내림차순으로 정렬하고 위에 있는 선택지를 여러 개 추천한다면 사용자의 실제 선택이 이 추천 안에 있을 확률은 더욱 높아진다.

3. 세부 설계 및 구현

3.1 개발환경

본 논문의 어플리케이션은 윈도우 10 환경에서 개발하였다. Python Crawling을 이용하여 경유지 정보를 크롤링을 하였으며, Maria DB를 활용하여 데이터베이스를, N Cloud를 이용하여 서버를 구축하였다.

3.2 시스템 구성

이 어플리케이션의 기본적인 동작방식은 출발지와 목적지를 입력하면, 자체 개발한 API를 통해 경유지를 추천한다, 데이터베이스는 유저 정보 및 크롤링 정보를 저장하고 있는데 API는 이 데이터를 활용해 출발지와 목적지 사이의 경유지를 클러스터링하여 사용자에게 적합한 경유지를 추천한다. 사용자가 실제 선택한 경유지 정보는 데이터베이스에 저장되어 차후 추천에 활용된다.

그림 1. 시스템 구성도

3.3 핵심 알고리즘

3.3.1 형용사 태깅 알고리즘

형용사 태그는 사용자의 주관적인 성향을 경로 추천에 반영하기 위해 사용하였다. 이는 사용자의 태그와 유사한 태그를 갖고 있는 장소가 보다 사용자에게 적합할 것이라는 전제와 사용자가 간 장소들의 형용사 태그를 조사하면 사용자의 성향을 파악할 수 있다는 전제를 기반으로 한다. 장소에 대한 형용사 태깅은 다음의 과정을 거친다.

단계 1 : 유사한 형용사를 그룹화하는 데이터 셋을 데이터베이스화
단계 2 : 데이터 크롤링한 댓글에서 형용사를 추출
단계 3 : 유사한 형용사를 통합하여 데이터 카운팅
단계 4 : 빈도가 높은 형용사 4개를 장소에 대한 태그로 지정
단계 5 : 4개의 형용사 비율을 1을 기준으로 비율 계산 및 저장

위 단계를 바탕으로 장소에 대한 형용사 태깅을 완료하면 사용자의 방문 기록을 기반으로 사용자의 형용사 태그를 추적할 수 있다.
사용자의 형용사 태그를 추적하는 과정은 다음과 같다.

단계 1 : 사용자가 형용사 태그를 선택했다면 이를 기반으로 추천하고 선택하지 않았다면, ***귀납 오류**를 해결하기 위해 경유지 100개의 방문 데이터가 쌓이기 전까지 추천할 때 형용사의 태그를 고려하지 않고 추천한다.
단계 2 : 방문한 장소에 대한 ***RFM 점수**를 구한다.
단계 3 : 장소의 RFM 점수와 형용사 태그의 비율을 곱한 후 더한다.
단계 4 : 가장 높은 형용사 태그 4개를 사용자의 태그로 지정한다.

***귀납 오류 :** *초기의 값을 추천에 반영하면 잘못된 값이 대푯값이 되어 추천 결과의 부정확성을 야기할 수 있다.*

***RFM 점수 :** *RFM 점수 산정은 다음과 같다. 산출된 각 RFM 점수가 0인 경우, R은 최저점인 0.1 F는 1 / 최고 높은 방문 횟수, M은 1 / 평점으로 한다.*

R	최근 20회중 방문 횟수 / 최고 높은 방문 횟수
F	방문 횟수 / 최고 높은 방문 횟수
M	평점 / 총 평점

이 과정을 통해 나온 형용사 태그는 사용자 경유지 추천에 활용하게 되고 사용자가 경유지를 방문하면 위 단계를 반복한다. 여기서 RFM 점수는 최근성, 빈도, 평점을 반영하여 사용자의 주관을 보다 정확하게 반영할 수 있다.

3.3.2 장소 속성을 활용한 K-Means 알고리즘

장소의 날씨, 온도, 사용자 형용사 테그 와 평점을 활용하여 클러스터링한다. 날씨 온도는 날씨 API를 활용해 크롤링한 데이터 상의 방문한 날의 평균기온과 날씨를 추출하여 저장한 데이터를 이용하여 활용한다.

이를 통해 실제 사용자의 방문에 영향을 미치는 유동적인 온도나 날씨 정보를 추천에 활용하고자 한다. 사용자 형용사 태그를 활용하여 사용자의 니즈를 반영하고 평점을 통한 클러스터링을 통해 사용자들의 평점을 기반으로 장소의 적합성을 평가하여 추천하게 한다.

3.3.3 사용자 속성을 활용한 K-Means 알고리즘

사용자의 나이, 형용사 태그,, 성별과 방문지를 활용하여 클러스터링 하였다. 유사한 나이, 형용사 태그,, 성별인 사람은 동일한 선택을 할 것이라는 전제를 기반으로 하였다. 나이는 10살 단위로, 형용사 태그는 3.3.1 형용사 태깅 알고리즘을 바탕으로 대표되는 형용사로, 성별은 남, 녀를 기반으로 클러스터링 하였다. 이를 통해, 유사한 사용자 그룹 내에 다양한 장소 표본을 확보하고 경유지 내의 많은 데이터를 필터링할 수 있다.

3.3.4 베이시안 네트워크를 활용한 추천 시스템

장소 속성을 활용한 K-means로 방문 데이터 내에서 클러스터링을 한다. 사용자 속성을 활용한 K-Means 알고리즘을 기반으로 사용자와 비슷한 그룹의 방문 데이터를 추출한다. 사용자 방문 데이터를 기반으로 방문지 선호도를 계산할 때, 사전 확률 계산을 통해 가중치 비율을 적용한다. 가중치가 적용된 사전 선호도 확률이 높은 장소 중에서 장소가 결정되면 최종 경유지 결정을 위한 사후 선호도 확률을 계산한다. 이를 바탕으로 경유지를 추천하게 되고 확률이 높은 순서대로 가볼만한 곳에 추천하게 된다.

3.4 알고리즘 적용 및 어플리케이션 구현

3.4.1 네이버 맵 API

```
@Component
public class NaverMap {
    private String ncloudUrl = "https://naveropenapi.apigw.ntruss.com/map-direction/v1/driving";
    private String clientID = "z61zcmt5wp";
    private String clientSecretKey = "O89vy2cYe04XasNZssFfJspMZqfUzW7qaZ9MwYch";

    public String sendNaverMap(ABox param) throws IOException {
        String result = "";
        HttpURLConnection con = null;
        try {
            // http client 생성

            ncloudUrl += "?start=" + param.getString("start");
            ncloudUrl += "&goal=" + param.getString("goal");
            if (param.containsKey("waypoints")) {
                ncloudUrl += "&waypoints=" + param.getString("waypoints");
            }
            if (param.containsKey("option")) {
                ncloudUrl += "&option=" + param.getString("option");
            }
            ncloudUrl = ncloudUrl.replaceAll(" ", "");
            // ncloudUrl = URLEncoder.encode(ncloudUrl, "UTF-8");

            URL url = new URL(ncloudUrl);
            con = (HttpURLConnection) url.openConnection();

            con.setRequestMethod("GET");
            con.setRequestProperty("Content-Type", "application/json");
            con.setRequestProperty("X-NCP-APIGW-API-KEY-ID", clientID);
            con.setRequestProperty("X-NCP-APIGW-API-KEY", clientSecretKey);
            int responseCode = con.getResponseCode();
            BufferedReader br;
            if (responseCode == 200) {
                br = new BufferedReader(new InputStreamReader(con.getInputStream()));
            } else {
                br = new BufferedReader(new InputStreamReader(con.getErrorStream()));
            }
            String inputLine;
            StringBuffer response = new StringBuffer();
            while ((inputLine = br.readLine()) != null) {
                response.append(inputLine);
            }
            result = response.toString();
            br.close();
        } catch (Exception e) {
            System.out.println(e);
        }

        return result;
    }
}
```

그림 2. 네이버 맵 API

네이버 맵 API를 활용해 네비게이션 기능을 활용하였다. 출발지와 목적지를 입력하면 네이버 맵 API를 통해 3가지 루트를 추천한다.

3.4.2 경로 내의 1차 필터링

경로를 선택하면경로 내의 일정 벡터 거리 안의 장소들을 장소 속성을 활용한 K-means 알고리즘을 통해 1차적으로 필터링을 한다. 이를 통해 현재 날씨, 온도, 장소 평점과 형용사 태그에 적합한 장소들을 필터링한다.

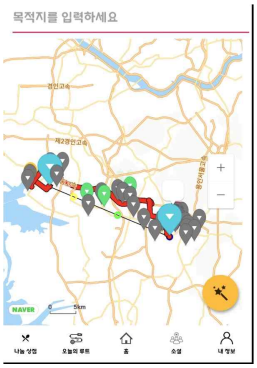

그림 3. 경로 내의 1차 필터링

3.4.2 경로 내의 2차 필터링

1차적으로 필터링된 장소를 기준으로 사용자 속성을 활용한 K-Means 알고리즘을 통해 유사한 속성의 그룹의 방문지로 필터링한다. 만약 필터링된 장소를 기준으로 유사한 속성의 그룹의 방문지가 없는 경우 1차적 필터링 데이터를 2차 필터링 데이터로 간주한다.

3.4.2 베이시안 네트워크를 활용한 경로 추천

2차적으로 필터링 된 경유지를 사용자 방문 데이터를 기반으로 방문지 선호도를 계산한다. 처음에는 표본이 없으므로 2차 필터링 내에서 사용된 사용자 속성을 활용한 K-Means 알고리즘을 통해 유사한 속성의 그룹의 방문지를 표본으로 이용한다. 이후 사용자의 방문 데이터가 쌓이면 이를 바탕으로 추천하게 된다. 베이시안 네트워크를 통해 추천된 경유지 3곳을 경로에 기본적으로 적용하고 맵에 대체 경유지나 추가 경유를 띄어 사용자가 자유롭게 추가 삭제할 수 있게 하였다.

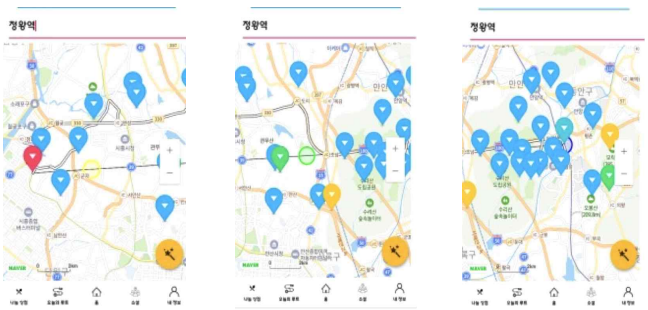

그림 7. 구현 결과 및 테스트

3.5 구현결과

본 논문에서 제안한 경로 추천 시스템은 사용자의 성향을 반영하여 사용자가 경유지 정보를 찾는데 시간을 할애하지 않고 적합한 경유지를 찾을 수 있는 정도로 잘 구현되었다. 네비게이션 기능 및 경유지 계산은 기존 네비게이션 속도와 비슷한 정도의 성능을 내었다. 사용자의 니즈를 만족시키기 위해 베이시안 네트워크나 K-Means를 통해 산출된 경유지를 List 형식으로 각 경유지마다 테마를 부여하고 적합도 순으로 나열하도록 하였다.

하지만, 경유지 필터를 할 때, 일정 벡터를 벗어나는 범위에 경유지 누락 문제가 발생할 가능성이 높다. 이 경우 사용자의 효율성과 효과성이 충돌하는 딜레마가 발생한다.

4. 결론 및 향후 연구과제

현재 경유지를 자동으로 추천하며 경로를 효율적이고 효과적으로 안내하는 어플리케이션은 없다. K-Means와 베이시안 네트워크를 활용한 경로 안내 시스템은 사용자가 사용해야 하는 복수의 어플리케이션을 통일시켜 사용자의 편의성을 높일 것이다.

1차 필터링 단계에서 경유지 필터 벡터 범위를 어떻게 할 것인지에 따라 네비게이션의 방향성을 크게 달라진다. 가령 단순한 경로를 지나면서 좋은 경유지를 방문하는 것과 여행 목적으로 지역을 꼼꼼하게 방문하는 것과 같은 다양한 방향성을 가질 수 있다. 향후 연구에서 이런 벡터의 범위를 산정하는 데에 있어서 다양한 변수를 고려하여 머신러닝을 통해 효율적인 벡터의 범위를 찾고 사용자의 패턴과 설정에 맞게 추천해주는 기능을 추가해보고자 한다.

참고문헌

[1] 네이버 지식백과,“https://terms.naver.com/”

[2] 추천시스템을 활용한 k-means 기법과 베이시안 네트워크를 이용한 가중치 선호도 군집 방법 - 박희범, 조영성. 고형화 “https://koreascience.kr/article/JAKO201334446973156.pdf”