

데이터 마이닝과 필터링을 이용한 내비게이션에서의 정보제공

정순원*, 이은주*, 김응모*
*성균관대학교 컴퓨터공학과
e-mail:gogo1260@naver.com

Offering Information in Car Navigation using Data Mining Techniques and Filtering

Soon-won Jung*, Eun-ju Lee*, Ung-mo Kim*
*School of Information and Communication Engineering,
Sungkyunkwan University

요 약

내비게이션 시장의 확대에 따라 주요 서비스인 길안내 외에 다양한 콘텐츠 제공 기술개발로 경쟁력을 갖춰 나갈 필요성이 대두되었다. 이러한 흐름에 발맞추어 운전자의 특성, 관심사를 고려, 운전자가 선호할 만한 서비스 정보를 제공하여 내비게이션의 경쟁력을 갖출 수 있는 방법을 제안한다. 본 논문에서는 데이터 마이닝, 필터링과 추천방법을 통하여 기존의 내비게이션이 경로를 탐색할 때 운전자의 기호와는 상관없는 정보를 제공한 것과 다르게 운전자가 선호할 만한 서비스 정보를 효율적으로 도출하는 방법을 제안 한다. 또한 내비게이션이 제공하는 불필요한 정보를 제함으로써 빠르고 효율적인 데이터관리를 할 수 있도록 한다.

1. 서 론

차량용 내비게이션(car navigation)이란 현재 위치에서 목적지까지의 거리 및 교통 상황을 고려하여 선택한 최적의 경로를 따라 안내를 하는 도로 및 교통정보 시스템으로 위치파악기술과 양방향 이동통신망을 제공하는 단말을 통해 각종 정보를 실시간으로 주고받을 수 있도록 하는 서비스다[1]. 지난해 국내 카 내비게이션 시장은 약 6000억원(150만 개)에 달하는 것으로 나타났다. 올해는 30~40% 성장한 약 1조원(200만 개)으로 전망 된다. 특히 신규업체의 잇따른 진출에 기존업체들은 기술력, 콘텐츠, 유통망 확대 등으로 맞서고 있다. 더불어 내비게이션의 주요 서비스인 길안내 외에 DMB뿐만 아니라 다양한 콘텐츠제공을 위해 많은 노력을 기울이고 있다. 업계가 포화상태에 이르고 있는 만큼 고객만족을 위한 기술개발로 경쟁력을 갖춰 나가는 기업이 성공할 것이다. 기존 제품들 중 몇몇 제품들은 인터넷 사이트, 출판업체 등과 제휴를 맺어 전국 곳곳의 다양한 맛집 정보를 제공한다. 본 논문은 운전자의 필요와 욕구에 부합하는 서비스를 제공하여 내비게이션의 경쟁력을 갖출 수 있는 방법을 제안 한다.

본 논문에서는 운전자의 데이터베이스 로그파일을 분석하여 불필요한 정보를 배제하고 내비게이션의 정해진 경로를 따라 운전자의 특성, 관심사를 고려한 POI(Point Of Interest : 사용자가 내비게이션 단말에서 지리정보와

함께 볼 수 있는 시설물, 지역 등에 대한 정보를 의미하며, land mark나 way point와 같은 용어가 사용되기도 한다.)정보 제공을 목표로 한다. 예를 들어 운전자가 오후 12시에 내비게이션으로 시청을 검색할 때 내비게이션은 시청으로 가는 경로를 운전자에게 안내한다. 또한 추가적인 정보로 점심을 할 수 있는 식당, 그리고 식사 후 쉬어 갈수 있는 커피 전문점 등을 운전자에게 알려준다.

논문구성은 2장의 관련연구, 3장에서 데이터마이닝을 이용한 내비게이션 정보제공 방법에 대해 설명한다. 4장 결론에서는 데이터마이닝 기법을 내비게이션에 적용 할 때 나타나는 효과, 향후 연구에 대해 논한다.

2. 관련 연구

2.1 데이터 마이닝

데이터마이닝이란 대량의 데이터로부터 유용한 정보를 추출하는 것으로 흥미 있는 지식(interesting knowledge), 고려할 만한(non-trivial), 잠재적인(implicit), 과거에는 알지 못했던 유용한(potentially useful)정보를 규칙이나 패턴 형태로 추출하여 이해하기 쉬운 형태로 변환, 의사 결정 과정에 적용하는 과정이다[2,3,6]. 통계학, 데이터베이스 기술, 인공지능, 기계학습(machine learning), 그리고 마케팅

등 매우 다양한 학문이 유용한 정보 추출의 목적 달성을 위해 관련되어 있다. 활용 분야에는 고객관리(CRM), 기업 의사 결정(decision making), 인터넷에서 방문자 이용 패턴을 위한 웹 마이닝(webmining), 생물정보학(bioinformatics)에서의 데이터분석 등의 다양한 응용분야에 적용된다[4].

2.2 정보 제공 방법

2.2.1 연관 규칙(association Rule)

연관 규칙은 한 번에 구매한 각각의 상품에 해당하는 항목들의 빈도수와 동시 발생 확률을 이용하여 항목간의 관계를 찾아 연관성을 규칙으로 표현하는 기법이다. 연관 규칙($X \rightarrow Y$)에서 지지도(Support(s))는 전체 트랜잭션 중 X와 Y를 동시에 포함하는 비율을 의미한다. <표 1>은 동시에 구매한 Item목록을 보여주고 있다. 지지도는 식 (1)을 이용하여 지지도를 계산할 수 있다. 이때 $\sigma(\text{Milk, Diaper, Beer})$ 는 milk, diaper, beer를 포함하는 트랜잭션의 수이고 $|T|$ 는 데이터베이스 내 전체 트랜잭션의 수이다. 연관 규칙에서의 신뢰도(Confidence(c))는 X를 포함하는 트랜잭션 안에서 Y안에 있는 아이템들이 얼마나 나타나는가에 대한 비율을 나타내며 이는 식 (2)와 같다[2,8].

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

<표 1> 동시 구매한 Item 목록

$$\text{Support}(s) = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4 \text{ 식 (1)}$$

$$\text{Confidence}(c) = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67 \text{ 식 (2)}$$

2.2.2 인구 통계학 추천(Demographic-based)

인구통계학적 추천시스템은 인구 통계학적 정보(사용자의 성별, 나이, 직업 등)에 의한 사용자 유형별 특징 분석을 통해 상품을 추천한다[5].

2.2.3 내용기반 추천(contents-based)

내용기반 추천 시스템은 사용자가 어떤 상품에 대해 평가한 특징 정보와 다른 상품에 포함된 텍스트의 특징 정보의 유사도를 이용하여 필터링 한다[6].

2.2.4 협업 필터링 (collaborative Filtering)

협업 필터링은 기호정보(taste information)를 많은 사

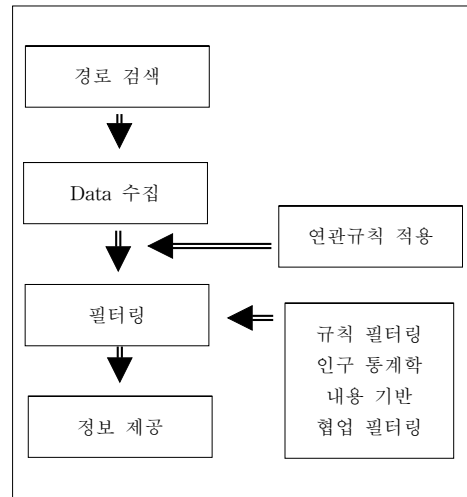
용자들로부터 얻어 관심사들을 자동적으로 예측하게 해주는 방법이다. 이 접근법의 근본적인 가정은 사용자들의 과거의 경향이 미래에서도 그대로 유지 될 것이라는 전제에 있다. 예를 들어, 음악에 관한 협업 필터링 혹은 추천시스템(recommendation system)은 사용자들의 기호(좋음, 싫음)에 대한 부분적인 목록(partial list)을 이용하여 그 사용자의 음악에 대한 기호를 예측하게 된다. 이 시스템은 일부 사용자 정보에 국한된 것이 아니라 많은 사용자들로부터 수집한 정보를 사용한다는 것이 특징이다. 이것은 단순히 투표 한 수를 기반으로 각 아이템의 관심사에 대한 평균적인 평가로 처리하는 방법과 차별화 된 것이다. 즉 고객들의 선호도와 관심 표현을 바탕으로 선호도, 관심에서 비슷한 패턴을 가진 고객들을 식별해 내는 기법이다 [2,4,9].

2.2.5 규칙기반 필터링(Rules-based Filtering)

규칙기반 필터링은 몇 가지 질문들을 한 이후에 운전가 답한 질문들을 바탕으로 적합한 contents을 전달하는 것이다. 규칙기반 필터링에서 제공하는 질문은 사용자들이 자신의 선택을 통해 내용을 스스로 구성하기 위한 목적이 아니라 질문을 통해 사용자들을 구분하고, 또한 개개인을 구별하기 위한 목적으로 사용된다[2,4].

3. 정보제공 연구 모형

본 논문에서 제안하는 정보제공 연구모형은 (그림 1)과 같다.



(그림 1) 시스템 구조도

(그림 1)에서 경로 검색 후 위치정보 등 데이터를 수집하여 연관규칙을 적용한다. 다음 단계에서는 수집된 데이터에 규칙필터링, 인구통계학 추천, 내용기반추천, 협업 필터링을 통하여 필요한 정보를 추출한다. 위의 단계를 거쳐 운전자에게 정보를 제공한다.

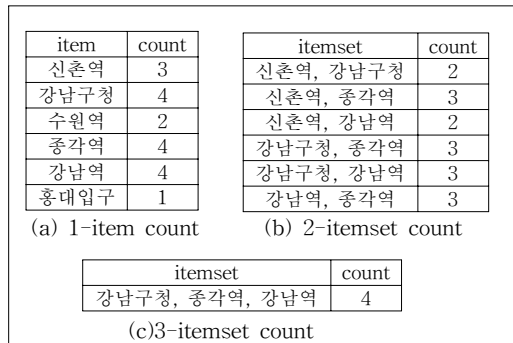
3.1 연관규칙(association rule)적용

예를 들어 운전자가 강남역이란 경로를 검색 한다. 시스템에서는 지난 5일 동안 운전자가 단말기를 통해 검색한 경로인 <표 2>의 수집된 데이터를 이용하여 연관 규칙을 적용 한다.

Date	경로 검색목록
08.9.1	강남구청, 강남역
08.9.2	강남구청, 종각역, 신촌역, 홍대입구
08.9.3	강남역, 종각역, 신촌역, 수원역
08.9.4	강남구청, 강남역, 종각역, 신촌역
08.9.5	강남구청, 강남역, 종각역, 수원역

<표 2> 날짜별 경로 검색목록

연관 규칙에는 두 단계가 있다. 첫 번째로 최소 지지도(minimum support)가 3이라고 할 때 이것을 만족하는 모든 항목 집합(Itemsets)을 찾는다. 두 번째 단계는 빈발항목집합(frequent itemset)을 추출하여 이로부터 연관규칙을 생성한다. 빈발항목집합을 생성함으로써 내비게이션에서 정보를 제공할 때 제한된 메모리를 효율적으로 사용하고 정보전송시간을 줄일 수 있다. <표 2>에 대한 빈발항목집합을 생성 과정은 (그림 2)처럼 우선순위 알고리즘(Apriori Algorithm)을 이용한다. (그림 2)의 (a)와 (b)에서 최소지지도를 만족하지 않는 데이터는 삭제하여 (c)를 생성한다[2].

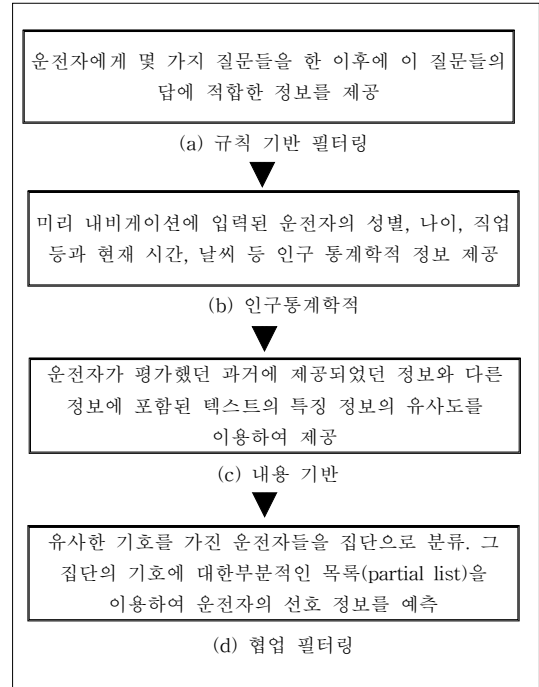


(그림 2) 우선순위를 이용한 빈발항목집합 생성

3.2 필터링을 통한 정보 제공

(그림 2)의 생성된 빈발항목 (c)을 바탕으로 필터링을 적용한다. 필터링 적용은 (그림 3)과 같다. 예를 들면 (그림 3)의 (a) 규칙기반 필터링은 운전자가 주거래 은행 정보에 ABC은행을 입력/체크했을 경우 다른 은행이 아닌 ABC은행과 해당 은행의 ATM기기의 위치정보 등을 제공 받는다. (b) 인구 통계학적 추천은 27살의 은행원인 남성 운전자가 비오는 날 저녁시간에 내비게이션에서 검색을

통하여 인구통계학적 정보에 의해 해당 음식집정보를 제공 받는다. (c) 내용기반 추천 방법은 운전자의 충분한 과거 프로파일 정보를 축적해야 하며, 이 프로파일과 비교하여 높은 점수를 얻은 정보를 제공한다. 예를 들어 햄버거를 좋아하는 운전자는 비슷한 종류의 패스트푸드점의 위치를 제공 받을 수 있다.



(그림 3) 필터링

(d) 협업 필터링의 첫 번째 단계는 운전자의 선호도 데이터를 가지고 운전자와 정보간의 행렬을 구성하는 것이다. 다음으로 운전자간의 유사도를 계산한다. 식 (3)은 운전자 i, j간의 상관계수 r_{ij} 을 구하는 식이다. S_{ik} 는 운전자 i의 정보 k에 대한 평가 점수이고 \bar{s}_i 는 운전자 i의 평가 점수 평균이다.

$$r_{ij} = \frac{Cov(i,j)}{\delta_i \delta_j} = \frac{\sum_k (S_{ik} - \bar{S}_i)(S_{jk} - \bar{S}_j)}{\sqrt{\sum_k (S_{ik} - \bar{S}_i)^2} \sqrt{\sum_k (S_{jk} - \bar{S}_j)^2}} \quad \text{식 (3)}$$

식 (4)은 사용자와 정보간의 행렬을 운전자별로 벡터로 표현하여, 두 운전자의 유사도를 벡터간의 코사인 값으로 계산하는 식이다. 이 경우에 사용자 a, b를 m차원의 벡터로 표현하게 된다.

$$\text{similarity}(\vec{a}, \vec{b}) = \cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|} \quad \text{식 (4)}$$

식 (3)이나 식 (4)을 통해 계산된 사용자들 사이의 유사도를 통해 유사 운전자를 선정한다. 협업 필터링의 마지막 단계는 상품에 대한 고객의 선호도 점수를 예측하는 것이다. 이는 식 (5)로 나타낼 수 있다. 식 (5)은 고객 i 의 정보 k 에 대한 선호도 점수인 P_{ik} 을 예측하는 식으로, $Rater(k)$ 는 상품 k 를 평가한 고객의 집합을 의미한다[10].

$$P_{ik} = \bar{S} + \frac{\sum_{l \in Rater(k)} (S_{lk} - \bar{S}_l) r_{il}}{\sum_{l \in Rater(k)} |r_{il}|} \quad \text{식 (5)}$$

이러한 필터링을 통해 운전자는 내비게이션에서 선호하지 않는 정보를 제외한 기호에 맞는 POI(커피숍, 서점, 맛 집, 극장, PC방, 관공서, 은행 등)정보를 제공 받을 수 있다.

4. 결론

현재 데이터베이스 분야에서 각광받고 있는 데이터마이닝을 이용하여 본 논문에서는 내비게이션의 정보제공에 데이터 마이닝과 필터링을 적용하였다. 기존의 내비게이션은 경로를 탐색할 때 운전자의 기호와는 상관없는 정보를 제공한 것과 다르게 본 논문은 운전자가 선호할 만한 서비스 정보를 효율적으로 도출하는 방법을 제안 하였다. 따라서 길 안내뿐만 아니라 운전자에게 알맞은 데이터를 제공할 수 있는 효과를 가져 온다. 향후 다양한 환경에 적합한 시스템으로 발전하기 위해서는 지금 보다 좀 더 많은 연구가 필요 할 것이다.

감사의 글

본 연구는 지식경제부 및 정보통신 연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었으며(ИТА-2008-C1090-0801-0028), 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 지식경제부의 유비쿼터스 컴퓨팅 및 네트워크 원천 기반기술 개발사업의 08B3-B1-10M 과제로 지원된 것임.

참 고 문 헌

- [1] 조성익 외 5명 "Trends and Perspectives of the next-generation navigation technology" 2007.6 전자통신 동향분석 제 22권 제 3호
- [2] J. Han, M. Kamber "Data Mining : concepts and techniques" Academic Press 2000
- [3] 장남식, 홍성완, 장재호, "데이터마이닝", 서울:대청미디어, 2002

- [4] 이기욱, 정창규 "데이터 마이닝 기법을 이용한 추천 시스템 구현" 2006. 3. 한국컴퓨터정보학회 제 11권 제 1호
- [5] B.kruwich, "Lifestyle Finder : Intelligent User Profiling Using Large-Scale Demographic Data", Artificial magazine, vol. 18, no. 2, pp.37-45, 1997
- [6] R.Burke, "Hybrid Recommender Systems : Survey and Experiments", User Modeling and User-Adapted Interaction, vol. 12, No. 4, pp.331-370 2002
- [7] 김정자, 이도현, "데이터마이닝 기술 및 연구동향" 1998. 9 한국정보과학회 정보과학회지, 제16권, 제9호
- [8] R. Agrawal and R. Strikant, "Fast Algorithm for Mining Association Rules" 1944, Proc of 20th Int'l Conf on VLDB, pp487-499
- [9] <http://ko.wikipedia.org> From Wikipedia, the free encyclopedia
- [10] Sarwar, B., Karypis, J. Konstan and J. Riedl. 2001 "Item-Based Collaborative Filtering Recommendation Algorithms" Proceedings of WWW10, Hong Kong, 285-295