FP-growth 마이닝을 이용한 효율적인 여행경로 수립 기법

유기범*, 조경수, 김응모
*성균관대학교 정보통신공학부
e-mail : idprayforu@gmail.com

A Technique for Making Efficient Travel Routes using the Mining Method of Frequent Patterns-growth

Kibeom Yoo*, Kyungsoo Cho, Ung-Mo Kim
*School of Information and Communication Engineering,
Sungkyunkwan University

요 약

컴퓨터의 활용이 다양해 지면서 예전과 다르게 다양한 이유로 많은 사람들이 여행을 하고 나서 여행에 대한 정보 블로그나 웹 상에 저장하고 공개한다. 이렇게 웹 상에 많은 양의 여행 관련 데이터가 존재함에도 불구하고 데이터들이 산발적으로 존재하고 체계적으로 데이터 베이스화 되어 있지 않아서 여전히 정보를 검색하고 여행 일정을 세우는 데에 많은 시간과 노력이 필요하다. 따라서 본 논문은 FP-tree 기반의 빈발 패턴 증가 기법을 이용한 여행 계획 수립 기법을 제안한다. 제안되는 기법에서 데이터들은 FP-tree 방식으로 저장되어 검색에 필요한 시간과 노력을 극적으로 줄이고, FP-growth 마이닝 기법을 이용해 효과적인 여행 경로를 선택할 수 있게 도와준다.

1. 서론

최근 업무의 중요성 만큼이나 휴식의 중요성이 부각되면서 그에 따른 여행의 수요도 꾸준히 증가하는 추세를 보이고 있다. 이러한 추세는 삶의 질을 중요시하는 인식과 연관되어 끊이지 않고 이어질 것으로 보인다. 여기서 주목되는 특징은 여행을 하는 수요의 연령대가 폭 넓어 지고 있다는 것이다. 여행의 주 소비계층이라 여겨졌던 30, 40 대의 수요가 증가하는 것 뿐만 아니라 학습과 경험을 목적으로 하는 10, 20 대의 수요도 크게 증가하고 있고, 관광을 즐기는 50, 60 대의 수요도 많아지면서 이제 여행은 단순한 휴식이 아니라 삶의 질 향상을 위한 기본적인 생활 요소로 사회적 인식이 바뀌고 있는 것이다. 또 다른 특징에는 여행을 함에 있어서 국내와 해외를 구분하지 않는 성향이 두드러져 보인다는 점이다. 사회적으로 지구촌이라는 말이 이미 널리 사용되는 만큼, 여행을 할 때에도 국내와 해외 여행을 구분하지 않는 글로벌 마인드적인 인식이 널리 퍼져있는 이유 뿐만 아니라, 국내와 다른 외국 환경에 대한 호기심 또한 이러한 특징에 일조를 하고 있다.

현재 여행 일정을 수립하는 것은 많은 시간과 노력이 필요하다. 여행 일정과 전체적인 경로를 수립할 때 가장 중요하게 여겨지는 것이 동선(動線)이다. 누구나 여행을 할 때에는 자연스러운 동선에 따라서 여행지를 결정하고 경로를 수립하는데 현재 여행 서적이나 웹 데이터들은 효율적인 여행 일정을 세우는 데에 부족함이 많기 때문에 많은 시간과 자료가 필요한 것이다. 일단 여행 서적 자료를 보면 많은 서적에서 동선에 따른 여행 경로를 보여주고는 있지만, 저자 개인의 성향 만을 반영한 경로라는 약점을 가지고 있기 때문에 경로 선택의 다양성을 보여주지 못하는 단점이 있다. 따라서 다양한 여행지 선택의 기회를 놓치게 되는 결과를 낳을 수 있다. 그리고 많은 웹 블로그 데이터는 여행 장소에 대한 정보만 알려줄 뿐 일정이나 경로가 고려되지 않았기 때문에 모든 자료를 찾아본 후에 동선에 따라 여행 경로를 수립해야 한다. 이 경우, 모든 여행지에 대한 정보를 따로 찾는 수고와 번거로움이 있을 뿐만 아니라 해당 지역에 대한 충분한 지식의 부재로 효율적인 여행 경로를 세우는데 어려움을 겪게 되고 때에 따라서는 여행의 효율성이 떨어지는 경로를 산출해 낼 수 있다는 단점이 있다.

이와 같은 신중한 분석 끝에, 다양한 의견을 반영하지 않고 산발적으로 존재하는 데이터와 함께 현재 효율적인 여행 일정 수립을 어렵게 하는데에는 시스템적인 문제가 있다고 판단하였다. 따라서 만약 이 문제들만 해결된다면 여행 일정을 세우는 데에 큰 효율의 향상을 보여줄 것이라 기대된다.

이러한 문제점을 해결하기 위해서 다음 두가지 관점을 가지고 접근해 보았다.

첫째, FP-tree 를 이용하여 자료들을 데이터 베이스화 한다. 다양한 관점의 자료가 쌓여갈 수록 각 노드들의 빈발 빈도(frequency count)가 결정되고 더 많은 빈발 빈도를 가진 패턴이 다른 경우보다 잘 선택될 수 있도록 노출이 된다. 즉, 각 여행지마다 방문 횟수가 기록되면서 그 횟수에 따라서 자주 발생하는 여행 경로를 가지고 있게 된다. 이에 대한 장점은 다수의 가장 많이 발생하는 경로 뿐만 아니라 소수가 이용하는 경로까지 보여주면서 기존의 문제점을 극복하여 경로 선택의 폭을 넓힌다는 것이다.

둘째, FP-tree 기반의 FP-growth 마이닝 기법을 이용하여 여행 경로를 수립하는데 도움을 준다. 예를 들어, 같은 경유지를 거치더라도 두 가지의 여행 경로가 존재할 때 각각의 장단점을 비교하면서 좀 더 효율적인 계획을 수립하는데 도움을 준다.

본 논문은 다음과 같이 구성된다. 2 장에서는 데이터 마이닝과 빈발 패턴 성장 기법을 소개한다. 3 장에서는 빈발 패턴 성장 기법을 이용한 효율적인 여행 일정 수립 기법을 제안한다. 마지막으로 4 장에서는 전체적인 고찰 및 더욱 발전된 기법을 위한 향후 연구 과제를 제시한다.

2. 관련연구

이 절에서는 여행일정 생성 기법에 필요한 데이터 마이닝에 대해 설명하고, 여러 마이닝 기법 중에서도 본 생성 기법에 사용될 수 있는 빈발 패턴 증가 기법을 소개한다.

2.1 데이터 마이닝 (Data mining)

간단하게 말해서, 기술적으로 데이터 마이닝[1,2]은 어마어마하게 큰 데이터에서 정보를 추출(extracting)하거나 채굴(mining)하는 것을 의미한다. 즉, 데이터 마이닝은 서로 다른 시각(perspectives) 이나 종합된 의견(summarizing) 을 분석하여 유용한 정보로 변환하는 과정을 말한다. 여기서 정보는 이익을 증대시키거나, 비용을 절감하거나, 혹은 두가지 모두에 이용될 수 있다. 데이터 마이닝은 다음과 같은 순차적인 단계로 수행된다.

1. 데이터 세척(Data cleaning) : 이 단계에서는 데이터속의 잡음(noise)과 모순된 데이터를 제거한다.
2. 데이터 통합(Data integration) : 다수의 데이터가 합쳐지는 단계이다.
3. 데이터 선택(Data selection) : 데이터 베이스로부터 분석작업에 연관된 데이터가 추출된다.
4. 데이터 변환(Data transformation) : 데이터가 종합(Summary) 과정을 거쳐 마이닝에 적합한 유형으로 변환된다.
5. 데이터 마이닝(Data mining) : 데이터 패턴을 추출해 내기 위해 지능적인 기법들이 적용되는 꼭 필요한 과정이다. 여기서 말하는 지능적인 기법들에는 예를 들어 연관 규칙 마이닝 기법, 그리고 빈발 패턴 증가 기법등 여러 기법들이 존재한다.
6. 패턴 평가(Pattern evaluation) : 데이터에서 보여지는 패턴을 구별한다.
7. 정보 표현(Knowledge presentation) : 사용자에게 채굴된 정보를 보여주는 단계.

요즈음 데이터 마이닝은 소매, 경제, 통신, 그리고 마케팅 조직 등 강한 소비자 집중(consumer focus) 성향을 가진 기업에서 사용되어지고 있다. 이 기술은 위와 같은 기업들을 가격이나 상품위치, 직원의 능력 등의 내부적인 요소들의 상관 관계를 결정할 수 있게 도와주고 또한 경제적인 지표(economic indicators), 경쟁, 그리고 소비자 통계 등의 외부적인 요소 간의 상관 관계를 결정할 수 있게도 해준다.

2.2 FP-Growth (Frequent Pattern Growth : 빈발 패턴 성장)

빈발 패턴 성장 마이닝 기법[2,3]은 빈발 패턴에 대한 중요한 정보만을 압축적으로 저장하게 고안된 FP-tree 구조를 기반으로 마이닝을 한다. 큰 데이터 베이스가 아주 밀집된 구조로 압축되었기 때문에 반복되는 데이터 스캔의 큰 비용을 효과적으로 절감할 뿐만 아니라 조각 빈발 패턴 성장에 의한 빈발 패턴의 완벽한 마이닝까지 제공한다.

FP-tree 는 다음과 같은 절차로 생성된다

.

i. 트랜젝션 데이터베이스 DB 를 한번 스캔한다. 그 중 빈발 항목들과 그에 대한 절대적인 빈발 사건(support)를 수집한다. 그 후에 빈발 항복의 리스트인 L 로서 빈발 항목 F 를 내림차순 정렬한다.

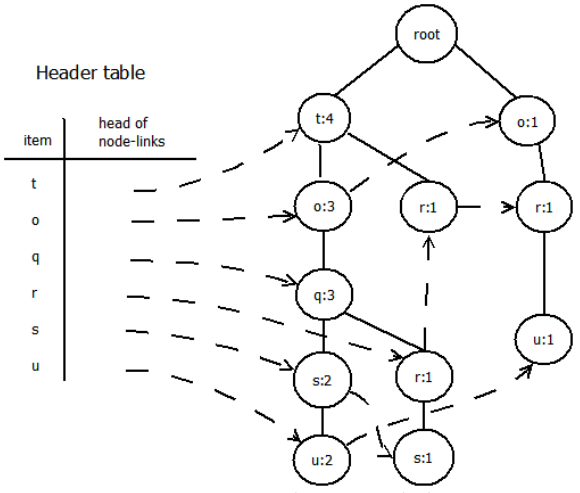
ii. 이름이 "*null*" 인 FP-tree T 를 만들고, DB 의 각 트랜젝션 *Trans* 에 다음 과정을 수행한다. 빈발 항목 리스트 L 의 순서에 따라 하나의 트랜젝션인 *Trans* 에 있는 빈발 항목을 선택하고 정렬한다. 정렬된 빈발 아이템을 $[p|P]$라고 할 때, 여기서 p 는 첫번째 원소이며 P 는 나머지 리스트 들이다. 그 후에 insert_tree($[p|P]$,T)를 호출하게 된다. 이 함수는 만약 T 의 자식 노드 중 p 와 같은 이름을 가진 노드 N 이 있다면 N 의 빈도를 1 증가시킨 후 부모인 T 에 링크된다. 이 과정은 P 가 비어있지 않다면 반복적으로 수행되게 된다.

위와 같은 과정을 통해서 모든 트랜젝션에 대한 FP-tree 가 완성되면 빈발 패턴 성장 마이닝을 할 수 있게 된다. 다음은 위 알고리즘을 이용한 데이터 베이스 트랜젝션의 예시와 그에 따른 FP-tree 를 보여준다.

<표 1> 트랜젝션 데이터 베이스의 예시

TID	Items Bought	(Ordered) Frequent Items
100	t, o, q, d, g, i, s, u	t, o, q, s, u
200	q, r, o, t, l, s, y	t, o, q, r, s

300	r, t, h, j, y	t, r
400	r, o, k, s, u	o, r, u
500	q, t, o, e, l, u, s, n	t, o, q, s, u



(그림 1) 트리 구조 예시

트리 탐색을 용이하게 하기 위해서 아이템 헤더 테이블은 트리 내의 사건을 가리키고 있다. 같은 이름을 가진 아이템은 노드 링크를 통해서 연속적으로 연결되어 있다. 따라서 모든 트랜젝션의 스캔 후에는 트리의 노드들이 (그림 1)에서 보는 것처럼 모두 연결되어 있는 것을 볼 수 있다.

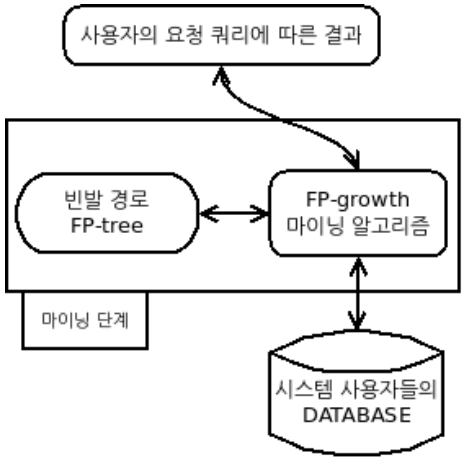
3. FP-growth 마이닝을 이용한 여행일정 수립

이 절에서는 여행 경로 수립 기법의 FP-tree 를 이용한 데이터 베이스 구축 과정을 소개하고, 그 데이터 베이스를 기반으로 빈발 패턴 성장 기법을 이용하였을 때 얻어낼 수 있는 효과적인 여행 경로 수립 절차를 설명한다.

3.1 여행 데이터의 FP-tree 구축 단계

데이터 베이스에 현재까지 많은 양의 여행 경로 데이터가 저장되어 있다고 가정했을 때, 데이터 베이스가 FP-tree 에 저장된 구조는 (그림 1)과 같은 모습을 보여준다.

(그림 1)에서 볼 수 있듯이 FP-tree 는 특정 도시에 대한 다양한 여행 경로를 담고 있다. 손실없이 저장된 데이터는 노드별 횟수(count)에 의존하여 사장 많이 이용되는 여행 경로를 보여준다. 이 정보를 이용하여 사용자는 객관적이고 효율성이 높다고 믿을 수 있는 여행 경로를 선택할 수 있게 된다. 다음 (그림 2)는 전체적인 시스템의 흐름을 보여준다.



(그림 2) 전체적인 시스템

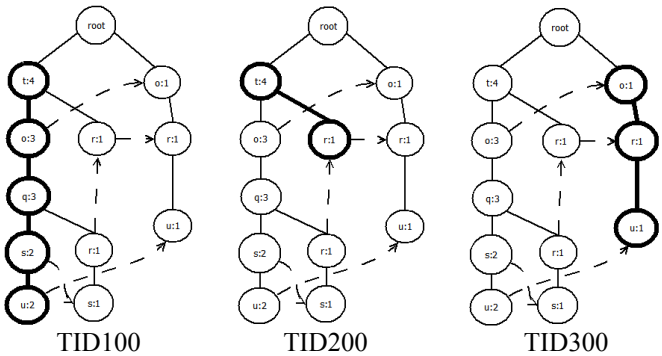
이 시스템은 사용자가 특정 쿼리를 요청했을 때, FP-tree 에 저장된 데이터를 FP-growth 마이닝 기법을 통해서 사용자의 쿼리에 맞는 데이터를 추출하여 사용자에게 제공해준다. 그 후, 사용자는 시스템이 제공한 데이터를 그대로 만족하여 여행 경로를 수립하거나, 조금 더 자신에 맞는 경로로 수정하여 경로를 확정하게 된다. 그럼 시스템은 다시 사용자가 결정한 경로를 FP-tree 알고리즘으로 분석하여 빈발 패턴을 찾아낸 후 에 기존의 데이터들과 통합되어 데이터 베이스는 최근 쿼리에 대한 정보를 담아 갱신되어 저장된다. 그리고 반복된 결과에 따른 새로운 빈발 패턴들은 다음 사용자의 쿼리 요청이 있을 때 갱신되어 제공된다.

3.2 FP-tree 기반의 여행 경로 마이닝

현재 저장되어 있는 데이터 베이스가 (그림 1)이라고 할때, 다음 <표 2>에서 제시하는 예시 데이터를 기반으로 하여 이 시스템이 어떻게 효율적인 여행 경로를 선택할 수 있도록 도와주는지 마이닝 단계를 설명한다.

<표 2> 여행 경로를 담은 데이터

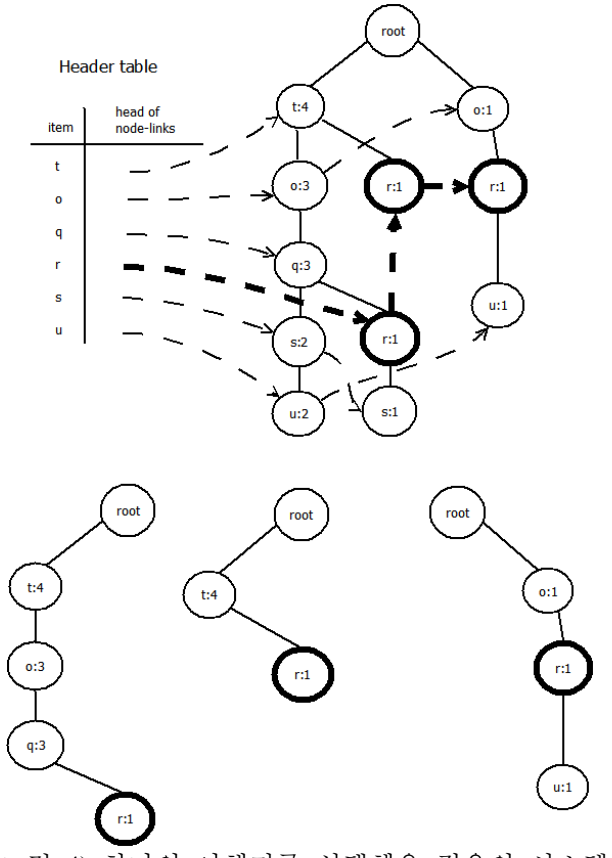
TID	여행지	정렬된 빈발 여행경로
100	t, o, d, q, g, i, s, u	t, o, q, s, u
200	t, r, q, d, s, u	t, r
300	O, q, s, l, r, u	o, r, u



(그림 3) 데이터에 따른 FP-tree 내의 경로 추출

다음 <표 2>와 같은 트랜젝션 데이터가 있다고 가정하자. 시스템은 각 사용자의 요청 쿼리에 따라서 FP-tree 알고리즘을 이용하여 빈발 패턴을 추출해 내고 사용자별 여행 경로를 추천해 준다. 예를 들어, Id 가 200 인 사용자가 t, r, q, d, s, u 를 경유하는 여행 경로를 원할 때, FP-tree 기반의 마이닝 과정을 거쳐서 많은 사람들이 이용했던 t, r 여행 경로를 사용자 200 에게 보여주게 된다.

또 다른 경우를 경우를 예로 들어 시스템의 동작 과정을 설명하고자 한다. 만약 시스템을 사용하는 ID 가 400 인 사용자가 다수의 여행지를 입력하지 않고, 하나의 여행지 만을 지정하여 그 여행지를 경유하는 가장 많이 이용되고 효율적인 여행 경로를 알고 싶을 때, 이 기법은 FP-tree 의 아이템 헤더링크를 이용하여, 해당 여행지가 포함된 여행 경로를 가장 빈번하게 이용된 경로 순으로 내림차순 정렬하여 사용자에게 보여준다.



(그림 4) 하나의 여행지를 선택했을 경우의 시스템

예를 들어, 한 사용자가 여행지 r 를 경유하는 효율적인 여행 경로를 알고 싶을 때, FP-tree 는 r 아이템의 헤더 테이블을 이용하여 r 를 경유하는 다음 세가지의 여행 경로를 사용자에게 결과로 보여주고 선택할 수 있게 도와준다.

위와 같이 많은 양의 데이터를 이용하여 빈발 패턴을 찾아내고 이 패턴을 이용하여 사용자는 많이 이용된 객관적이고 효율적이라고 믿을 수 있는 여행 경로를 얻을 수 있게 된다.

4. 결론 및 향후 연구과제

본 논문에서는 데이터 마이닝 기법 중에서도 FP-tree 를 이용한 빈발 패턴 성장 기법을 이용하여 효과적인 여행 일정 세우기를 할 수 있도록 했다.

여행 일정 수립에 이 기법을 적용하는 것에는 몇 가지 장점들이 존재한다. 첫째, 자료를 데이터 베이스화 하는 것에 있어서 경량화된 FP-tree 를 이용함으로써 여행지 검색에 걸리는 많은 시간을 줄여준다. 둘째, 데이터를 FP-tree 로 구성하는 과정에서 어떠한 손실도 일어나지 않기 때문에 빈도가 높은 여행 경로뿐만 아니라 발생했던 가능한 여행 경로를 보여주기 때문에 여행 경로 선택의 다양성을 보장받을 수 있다. 마지막으로, 이 기법은 하루의 여행 일정을 세우는 데 뿐만 아니라 나아가 여행의 전체적인 일정을 수립하는 데에 큰 도움을 줄 수 있다. 장기 해외 여행을 예로 들어, 첫 단계로 여행 할 나라를 결정하는 데에 저장된 데이터에 따라 경로를 결정할 수 있고, 둘째 단계로 각 나라의 도시 이동에 대한 일정과 마지막으로 각 도시에서 장소에 대한 경로를 수립할 수 있다. 이는 본 논문이 제시한 기법이 단순한 여행 경로 수립에 뿐만 아니라 여행 플래너의 기능을 수행 할 수 있는 가능성을 보여주고 있다.

더 나아가 이보다 발전된 기법은 지도 서비스와 결합된 형태로 제시되어야 할 것이다. 현재 웹 상의 지도 서비스는 단순히 지형을 표시하는 과거의 기능을 뛰어넘어 대중교통 서비스, 이동 수단에 따른 네비게이션 서비스, 각 장소에 대한 DB 를 저장하고 보여주는 서비스 등 다양한 기능을 보여주고 있다. 여기에 여행 경로 기법이 결합된다면 여행자는 각 여행지를 지도를 통해 한 눈에 볼 수 있을 뿐만 아니라, 그에 따라서 더욱 효율적인 경로를 수립할 수 있게 된다. 또 각 여행지에 대한 객관적인 정보 뿐만 아니라 다른 여행자들의 의견 또한 별다른 검색의 노력 없이 볼 수 있어서, 여행 일정을 수립하는 데에 더 많은 효율 향상을 보일 것으로 기대한다.

감사의 글

이 논문은 2009 년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0075771).

참고문헌

[1] J. Han, M. Kamber "Data Mining : Concepts and Techniques" Multiscience Press 2006

[2] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc.ACM-SIGMOD Intl Conf. Management of Data, pp 1-12, May 2000

[3] J. Han, J. Pei, Y. Yin and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, vol. 8, no.1, pp. 53-87, 2004