가중치가 부여된 FP-tree를 이용한 여행지 추출 기법

김민주*, 이은주*, 김응모*

*성균관대학교 정보통신공학부

soyara@skku.ece.ac.kr

Mining Technique of Tour Destination by weighted FP-tree

MinJu Kim*, EunJu Lee*, Eung-Mo Kim*
*School of Information and Communication Engineering, Sungkyunkwan University“

요 약

최근 컴퓨터와 통신의 기술이 빠르게 발달함에 따라 사회 각 부분은 그동안 경험하지 못했던 정보화라는 새로운 변화를 겪었다. 그 결과 정보화 수준이 점점 고도화 될수록 더욱 다양하고 방대한 데이터가 생성되어 데이터베이스를 이루게 되었다. 방대한 데이터에서 유용한 정보를 얻는 데이터마이닝 기법이 중요한 문제로 대두되었다. 데이터마이닝 기법은 점점 더 많은 분야에서 합리적인 선택을 위해 필수적으로 사용된다. 본 논문은 마이닝 기법을 적용하여 방대한 데이터베이스가 최적의 여행 경로 선택을 제공한다. 본 논문은 빈발 패턴 증가 기법에 가중치를 두어 여행자가 여행지를 선별하기 좋은 환경을 제공한다. 미래 산업 중 가장 중요한 산업 중 하나인 관광 산업은 계속적으로 성장하고 있으며 논문에서 제시하는 데이터 마이닝 기법으로 더 큰 발전을 기대한다.

1. 서론

최근 컴퓨터와 통신의 기술이 빠르게 발달함에 따라 사회 각 부분은 그동안 경험하지 못했던 정보화라는 새로운 변화를 겪었다. 특히, 정보화는 기업에서 새로운 경쟁력 확보의 수단으로 평가되어 각 조직들은 앞 다투어 사내의 정보화를 추진했다. 그 결과 정보화 수준이 점점 고도화 될수록 더욱 다양하고 방대한 데이터가 생성되어 통제하기에 어려움이 있다. 즉, 방대한 데이터는 축적 되었지만 실제로 가치를 지니는 의미 있는 데이터를 추출하는 과정이 복잡해졌다. 따라서 앞서 언급한 문제점을 해결하기 위해 특정 기준에 따라 정보를 구분, 분류하여 저장하는 데이터베이스의 개념이 등장하였다. 하지만 단순한 정보의 집적으로는 방대한 데이터들을 통제하고 기업의 의사결정에 활용할 수 없었다. 이러한 단점을 극복하기 위해 등장한 개념 중 하나가 데이터 마이닝(DM: Data Mining)이다[1]. 이러한 마이닝 기법은 최근 광범위한 분야에 적용되고 있다. 일례로 흔히 주변에서 볼 수 있는 대형 할인마트에서는 소비자들의 구매 정보를 축적, 분류, 분석하여 판매 품목을 어떻게 배치하는 것이 효율적인지 판단하기 위해 마이닝 기법을 적용된다[2].

본 논문에서는 최근 큰 미래 성장 가능성을 지닌 여행 산업과 앞서 언급한 마이닝 기법을 적용하여 선호도가 높고 효율적인 여행 경로를 제시하는 서비스를 제안한다. 현재까지의 여행 경로는 단순히 여행 책자나 인터넷 정보 등을 통해 얻어진 정보로 결정되는 것이 대부분이었다. 이와 같은 방법을 통해 얻어진 정보는 개인별 취향이 많이 반영된 것이므로 그 정보를 이용하는 사용자의 특성과 맞지 않을 수 있으므로 효율적인 여행 경로를 결정하지 못할 가능성이 높다. 그러나 마이닝 기법을 통해 축적된 여행 경로 정보를 통해 여행 경로를 선택할 경우 특정 패턴을 검색 할 수 있으므로, 검증된 여행 경로를 선택할 수 있게 된다. 제안된 기법의 아이디어는 다음과 같다. 여행객이 특정 지역을 방문하였을 때 그 이후 어떠한 여행지로 이동하는 지에 대한 정보를 축적한다. 이와 같은 과정을 통해 축적된 정보를 바탕으로 패턴을 추출하여 이후 여행객이 특정 여행지를 방문 할 경우 과거 어떠한 여행 경로가 이용되었는지에 대한 정보를 제공함으로써 여행객은 보다 쉽고 효율적인 여행 경로를 선택하게 한다.

본 논문은 다음과 같이 구성된다. 2장에서는 기본적인 마이닝 기법을 설명한다. 3장에서는 본 논문에서 제시하는 마이닝 기법을 이용한 여행 경로 설정 기법에 대해 서술한다. 4장에서는 전체적인 고찰과 발전된 연구를 위한 및 향후 연구 과제를 제시하면서 결론을 맺는다.

2. 배경지식

여행 경로 추출 기법에 사용 될 수 있는 마이닝 기법에는 여러 가지 방법이 있으나 본 절에서는 가장 잘 알려진 연관 규칙 마이닝과 빈발 패턴 증가 기법에 대해 설명한다.

2.1 연관 규칙 마이닝 기법

연관 규칙 마이닝[3,4] 기법은 주어진 데이터 집합에서 서로간의 연관성이 있는 항목을 찾아내는 방법 중 하나이다.

$I=\{i_1,i_2 \ldots i_m\}$ 인 I는 M 개의 항목을 가지는 집합이라 하자. 트랜잭션 T는 $T \subset I$인 항목들의 집합이고, D는 트랜잭션 T들의 데이터베이스라고 하자. 이 중 X, $Y \subset I$이고 $X \cap Y = \varnothing$ 을 만족 할 때 X, Y를 Item Set 이라고 부르고, X와 Y의 연관 규칙은 $X \Rightarrow Y$으로 나타낸다. 이러한 연관 규칙 중 중요한 개념으로서 지지도와, 신뢰도가 있다.

연관 규칙의 지지도는 $X \cup Y$ T를 D로 나눈 백분율 값이다. 그러므로 만약 연관규칙에서 8%의 지지도를 가진다고 한다면 그것은 총 데이터베이스의 트랜잭션의 8%가 $X \cup Y$의 항목집합을 포함하고 있다는 것을 나타낸다. 예를 들면 연관 규칙에서의 8% 지지도의 의미는 대형 할인점에서 전체구입고객(D)의 8%가 맥주와 기저귀를 동시에 구매한다는 의미가 된다. 할인매장 마케팅 담당자는 이러한 지지도를 근거로 특정 제품들을 비슷한 위치에 진열하거나 카탈로그의 쿠폰 북에 이용하는 등, 고도의 마케팅 전략을 펼칠 수 있다.

다음으로 신뢰도는 연관규칙 $X \Rightarrow Y$에서 $X \cup Y$의 전체 T의 수를 X를 포함하는 T의 수로 나눈 백분율 값이다. 그러므로 80%의 신뢰도를 가지는 연관규칙이 있다는 것은 X를 포함하는 항목집합 중 80%가 Y역시 포함하고 있다는 것을 의미한다. 연관규칙에서의 지지도가 통계자료를 의미하는 반면 신뢰도는 X와 Y사이의 상호 연관성을 나타낸다. 즉 80% 신뢰도는 맥주를 구입한 고객의 80%가 기저귀도 구매한다는 것을 의미한다.

일반적으로 데이터베이스에서 연관 규칙 마이닝은 사용자나 전문가가 정해놓은 최소 지지도 임계값과 최소 신뢰도 임계값을 만족하는 모든 규칙을 찾는 것이다.

2.2 빈발 패턴 증가 기법

2.2.1 FP-tree기본 개념

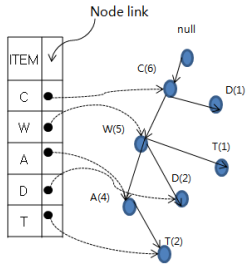
FP-tree기법[5,6]은 연관 규칙 마이닝의 알고리즘 중 하나로 후보생성 없이 분할-정복 기법을 사용하여 빈발항목을 찾아낸다. FP-tree 알고리즘은 첫 번째 단계에서 빈발 항목의 Count를 알아낸다. 두 번째 단계에서 빈발 항목들을 저장하기 위해 카운터 순으로 정렬한다. 세 번째 단계는 FP-tree 구조에 저장하는 단계로 처음 Root는 null 값을 주고, 각각의 Transaction들을 FP-tree로 저장한다. 마지막 단계는 헤더 테이블과 트리 노드를 링크시키는 것이다. FP-Growth는 모든 빈발 항목집합을 얻어내기 위해 2번의 데이터베이스 스캔을 요구하므로 빠르며 다음은 (그림1)의 (a), (b), (c)는 예시데이터와 FP-Tree 이다.

TID	ITEM
100	ACTW
200	CDW
300	ACTW
400	ACDW
500	ACDTW
600	CDR

(a) 데이터

ITEM	지지도
C	6
W	5
A	4
D	4
T	4

(b)지지도



(c) FP-tree

(그림 1) FP-tree 알고리즘 예시데이터

2.2.2 가중치를 둔 빈발 패턴 증가 기법

항목에 가중치를 두어 계산하는 방법이다[7]. 아이템 항목들에 일정 가중치를 지정하여, 가중치와 지지도 값을 이용하면 기존의 지지도와는 다른 결과를 가진다. <표1>은 아이템 항목과, 선호도가 있는 항목을 나타낸 표이다.

<표1> 데이터와, 빈발 항목들

TID	Set of items	Frequent Item List
100	a, c, d, f, m, r	c, d, f, m, r
200	a, c, d, f , i, m	c, d, f, m, r
300	b, c, f, m, p	c, f, m, p
400	b, d, f, m, p, r	d, f, m, p, r

다음으로 각 아이템의 가중치를 부여한다. 이 가중치는 지지도 값과 곱하여 새로운 지지도 값을 만들어 낸다.

<표2> 항목 당 지지도 및 가중치 설정

Item(min_sup = 2)	a	b	c	d	f	i	m	p	r
Support	2	2	3	3	4	1	4	2	2
Weight(0.2<wr<0.7)	0.5	0.3	0.6	0.4	0.7	0.3	0.5	0.2	0.7
Weight(0.8<wr<1.3)	1.1	1.0	0.9	1.0	0.7	0.9	1.2	0.8	1.3

가중치 값과 지지도 값을 곱하여 만들어낸 새로운 지지도 값을 이용하여 항목들의 순서를 새로 만든다.

<표3> 가중치가 부여된 빈발항목들

TID	WFI(0.2<wr<0.7)	WFI(0.8<wr<1.3)
100	d, f, m	c, d, f, m, r
200	d, f, m	c, d, f, m
300	f, m	c, f, p, m
400	d, f, m	d, f, m, p, r

<표3>은 <표2>의 가중치 값과 지지도 값을 곱하여 새로운 지지도 값을 구한 뒤, 이를 비교하여 WFI(Weight Frequent Item: 가중치 빈발 항목)생성 한다.

3. 데이터마이닝을 이용한 여행 경로 추출

3.1 여행 데이터 시나리오

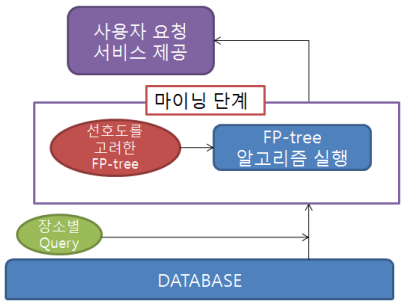
여행자는 장소별로 자신이 원하는 여행지가 다를 것이다. 현재 여행 계획을 세울 때 얻을 수 있는 정보는 인터넷과 여행 책자가 전부이지만 이 정보들은 여행자의 특성에 맞는 정보를 담고 있지 않고 분류되어 있기 때문에 여행자는 모든 정보를 확인하고, 다시 원하는 정보를 선택하여 자신이 원하는 여행 루트를 계획하여야 한다. 최적화된 여행을 계획 한다면 많은 정보를 얻기 위한 상당한 노력이 소모된다. 예를 들어 여행자 A는 독일의 뮌헨과 함부르크에 가고 싶고 미술에 관심이 있는 사람이라고 했을 때, A는 뮌헨과 함부르크에 있는 미술관을 모두 검색해야 하고, 자신이 만족하는 미술관인지 확인한 후 해당 미술관을 관람 할 지 결정해야 한다.

이런 상황에서 방대하게 축적되어 있는 데이터베이스에서 여행자에게 적합한 데이터를 추출하여 제공한다면 보다 더 편하게 여행을 계획할 수 있다. A는 데이터베이스에서 뮌헨, 함부르크와 가장 관심 있는 미술관을 쿼리로 넘긴다. 그 후 마이닝 기법을 이용하여 그 도시와 미술관을 갔다 온 많은 여행자들이 선택했던 다른 여행지를 확인 할 수 있다. 이는 A가 관광지들을 찾는 수고를 덜어주고, 최적화된 여행지를 선택하게 하여 계획을 세울 때 큰 도움이 된다.

3.2 여행지 추출 시스템 구성

데이터베이스에는 현재까지의 여행객들이 다녀온 도시와 관광지에 대한 데이터가 상당히 축적되어 있다고 하자. 다음으로 여행을 떠나려는 여행객은 데이터베이스에서 여행하고 싶은 장소를 입력한다.

선별된 데이터를 이용하여 기호를 부여하고 FP-tree 알고리즘을 이용하여 이전의 여행객들이 가장 많이 간 여행지를 검색할 수 있다. 검색된 데이터를 이용하여 여행자는 여행 경로를 보다 쉽게 계획 할 수 있다. 제안하는 시스템은 다음과 같다.



(그림 2) 전체적 시스템 블락도

(그림 2)는 세부적인 데이터 마이닝 모듈 블락도를 나타내고 있다. 총 3단계로 구성되며 첫 단계로 현재까지의 여행객이 다녀온 정보 정보가 통합 데이터베이스에 저장된다. 데이터 저장 시 추가로 여행자는 자신이 여행하였던 곳의 도시명과, 선호도를 따지기 위해 추천 Best와 Worst를 입력한다. 둘째 마이닝 단계로, 장소에 따라 분류된 데이터를 가지고 FP-tree 알고리즘과 기존의 선호도 정보를 이용하여 마이닝을 한다. 셋째, 추출된 데이터를 사용자에게 제공한다.

3.3 선호 가중치가 적용된 여행 경로 패턴 추출 기법

본 논문에서는 <표4>를 기반으로 선호도가 추가된 유용한 행동 패턴을 추출한다.

<표4> 지역 정보를 담은 통합 데이터베이스

User ID	여행지	Best	Worst
100	미술관B, 박물관A, 미술관C, 미술관D	미(D)	*
200	미술관C, 박물관A, 미술관D, 박물관D	미(D)	*
300	미술관B, 박물관D, 미술관C	*	박(D)
400	미술관B, 미술관C, 박물관B, 미술관F	미(C)	미(F)
500	미술관G, 박물관A, 미술관C	미(C)	*
600	미술관C, 박물관E, 미술관G, 미술관D	미(D)	*
700	미술관B, 박물관A	박(A)	*
800	미술관B, 박물관A, 미술관E	*	미(B)
900	미술관B, 박물관D, 미술관F, 미술관F	*	박(D)

<표4>는 여행지역의 정보를 담은 데이터베이스로 여행자들이 방문하였던 여행지와 여행한 곳 중에서 가장 선호했던 지역과 가장 선호하지 않은 지역에 대한 데이터를 담고 있다. 이 데이터를 FP-tree알고리즘에 적용한다. 먼저 tree로 저장하기 전에 빈발항목들을 알아내기 위한 카운트를 확인한다. 최소지지도를 3으로 하고, 기존의 지지도를 만족하는 아이템에 여행자들이 선택한 선호 데이터와 선호하지 않은 지역의 데이터에 대해 가중치를 부여하여 지지도를 새로 구한다. 각각의 아이템에 대한 지지도를 나타낸 것이 <표5>이다.

<표5> 최소 지지도 3을 만족하는 관광지

여행지	지지도	선호 지지도
미술관 C	6	6.4
미술관 B	6	5.8
박물관 A	5	52
박물관 D	3	2.6
미술관 D	3	3.6

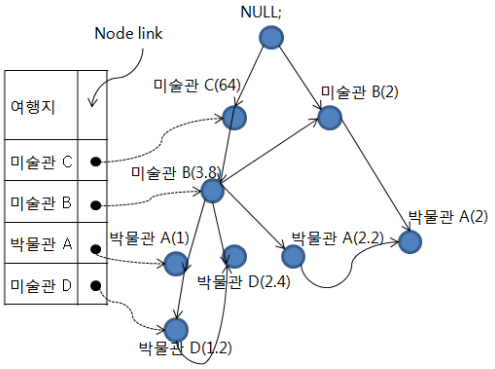
(a) 기존의 지지도와 비교

여행지	선호 지지도
박물관 A	1+1+1+1+1×1.2=5.2
박물관 D	1+1×0.8+1×0.8=2.6.
미술관 B	1+1+1+1+1×0.8+1=5.8
미술관 C	1+1+1+1×1.2+1×1.2+1=6.4
미술관 D	1×1.2+1×1.2+1×0.8=3.6

(b) 가중치 지지도

<표5>는 Best와 Worst에 대해 각각 1.2, 0.8의 가중치를 두어 기존의 지지도에 해당 가중치를 곱하였다. 이 값은 지지도 값을 구하기 위해 설정한 임의의 값이다. Best는 1보다 큰 값으로, Worst는 1보다 작은 값으로 설정한다. 가중치를 부여하지 않았을 때, 박물관 D의 기존 지지도는 3으로 빈발 항목에 속한다. 반면에 가중치를 적용하면 지지도가 2.6으로 되어 요구하는 최소 지지도 3을 만족하지 못해 빈발 항목에서 제외됨을 보인다.

다음의 (그림 3)은 <표2>의 결과를 이용하여 만든 FP-tree이다.



(그림 3) FP-tree

위의 자료를 이용하여 여행자는 선별된 미술관 C, B D와 박물관 A를 선택 할 수 있음을 볼 수 있다.
본 논문에서는 데이터마이닝을 이용하여 여행지를 추천하는 구조만 보였다. 경로 선택의 도움을 주기 위하여 여행지를 Item Set으로 정의하였다. 더 나아가 선호 여행지 간의 거리를 구해 주는 기법까지 제안 한다면 여행자가 여행 경로를 선택 하는 데 큰 도움이 될 것이다.

4. 결론 및 향후 연구

본 논문에서는 마이닝 기법을 이용하여 최적의 여행 경로를 결정 할 수 있도록 하였다. 앞으로 마이닝 기법이 발전함에 따라 다양한 질의를 이용할 수 있음이 보여 질 것이다. 또한 이 질의들은 보다 빠르게 응답을 받을 수 있을 것이다. 본 논문에서 사용한 FP-Tree 알고리즘이 가진 장점으로 완전성과 밀접성이라는 Completeness와 Compactness가 있다. 그러나 원본 DB보다 데이터의 양이 커지지 않는 장점에도 불구하고, 메모리 처리의 문제를 가지고 있다[8]. 향후 연구로는 데이터 마이닝 기법이 FP-tree의 문제점을 보안하는 계속적으로 연구가 필요하며 관광 산업 뿐만 아니라, 좀 더 다양한 곳에서의 마이닝 적용이 필요하다.

감사의 말

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

참고문헌

[1] J. Han, M. Kamber "Data Mining : Concepts and Techniques" Acadamin Press 2000
[2] Jiawei Han, Jian Pei, and Yiwen Yin, "Mining frequent patterns without candidate generation", Proceedings of 2000 ACM SIGMOD Int. Conf. Management of Data(SIGMOD'00), Dallas, TX, pp. 1-12
[3] Rakesh Agrawal , Ramakrishnan Srikant "Fast Algorithms for Mining Association Rules"
[4] M. H. Dunham, Y. Xiao, L. Greenwald, Z. Hossain "A Survey of association rules" Dallas, Texas
[5] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. ACM-SIGMOD Int'l Conf. Mangement of Data, pp1-12, May 2000
[6] 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
[7] Unil Yum and John J. Leggett, WFIM: Weighted Frequent Itemset Mining with a weight range and a minimum weight
[8] YoungWook Park, SeungCheol Lee, Ung Mo Kim, "Roadmap of an application for attending the lecture by FP-tree", Dept of Computer Engineering, Sungkyunkwan University