통계적 추론을 이용한 전문가 Belief 기반의 Web Usage 패턴 검증

고세진[0] 안계순 정준 이필규
인하대학교 전자계산공학과
{suksj, kyesun, jjeong, pkrhee}@im.inha.ac.kr

Web Usage Patterns Validation Based on Expert Belief Using Statistical Reasoning

Se-Jin Ko[0] Kye-Sun Ahn Jun Jeong Phill-Kyu Lee
Dept. of Computer Science and Engineering, Inha University

요 약

발견된 Web Usage 패턴들은 분석하는 전문가에게는 불필요하고 흥미롭지 못해 의사결정에 도움이 못되는 경우가 많다. 따라서 발견된 패턴에 대한 도메인 전문가의 사전 Belief에 기반한 패턴 검증 과정이 필요하다. 발견된 패턴의 유용성 여부는 패턴의 Unexpectedness를 측정함으로써 결정할 수 있다. 본 논문에서는 패턴의 Unexpectedness를 전문가의 Belief에 기반하여 검증하기 위한 새로운 방법론 제안한다. 발견된 패턴과 전문가 Belief를 매칭 알고리즘을 이용하여 패턴을 4가지(완전일치, 조건부 일치, 결과부 일치, 완전 불일치)로 분류하는 1차 검증과 1차 검증 결과의 4가지 분류데이터를 통계적 추론 방법인 Dempster-Shafer에 적용한 2차 검증으로 나뉜다. 1차 검증 과정은 패턴의 분류 용이성을 부여하나 패턴의 Unexpectedness에 대한 신뢰성을 제공하지 못한다. 이 문제점을 2차 검증 과정을 통해 해결한다.

1. 서 론

Web Usage Mining[4]은 일반적인 Data Mining 방법을 웹 도메인에 적용하여 Web Usage Patterns 을 찾아내고 이를 바탕으로 개인화된 정보를 사용자에게 전달하는 기술이다. 그러나 사용자의 패턴을 발견하는 Data Mining의 연관규칙(Association Rule) 알고리즘 즉, Apriori[2]는 많은 수의 패턴을 발견할 뿐만 아니라 의사결정에 참여한 도메인 전문가에게 이미 알려진 패턴이거나 불필요한 패턴, 즉 흥미롭지 못한 패턴을 발견한다[2,3,5,6]. 따라서 발견된 패턴에 대한 도메인 전문가의 사전 Belief에 기반한 패턴 검증 과정이 요구된다.

발견된 패턴의 유용성 여부는 Interestingness[2,3,5]를 이용하여 판단되어 진다. 패턴의 Interestingness는 객관적 척도(Objective Measures)와 주관적 척도(Subjective Measures)로 나누어지며 전자는 패턴발견 과정에서 사용되는 데이터와 그 구조에 의해 측정되어진다. 대표적인 객관적인 척도(Objective Measures)는 패턴의 지지도(support), 신뢰도(confidence) [2,3,5] 등이며 반면 후자는 전문가의 사전 지식 또는 Belief를 패턴 평가에 이용한다.

본 연구에서는 패턴 검증의 척도로 전문가 Belief를 이용하는 주관적 척도(Subjective Measures)[2,3,5]로 하며 이는 다시 두가지 요소로 나뉘어 진다.

- Unexpectedness : 패턴이 전문가에게 놀란만한 사실임을 전해주는 정도를 나타냄
- Actionability : 발견되어진 패턴이 전문가로 하여금 특정 행위를 하여 이익을 얻게 하는 정도

Actionability는 개념적인 요소로서 실제로 사용되기는 어렵다. 그러나 두 요소는 상호 배타적이 아니며 일반적으로 Unexpectedness 패턴은 Actionability하며 그 역도 성립한다[2]. 따라서 Unexpectedness를 패턴 검증의 주관적 척도(Subjective Measures)로 인식한다. 본 논문에서는 패턴의 Unexpectedness를 전문가의 Belief에 기반하여 검증하기 위한 새로운 방법론 제안한다.

2. 관련 연구

현재 진행 중인 대부분의 연구는 패턴의 interesting-ness 식별을 위해 주관적인 척도인 Unexpectedness 측정에 맞춰지고 있다. [2]에서는 사용자가 제시한 Belief의 degree을 정하는데 Bayes rule을 이용하고 있다. 즉 Belief의 패턴 a에 대한 degree는 조건부 확률 p(a|E)으로 나타내어지며 E는 주어진 Belief에 대한 증거이다. 이 방법은 해당 Belief에 대한 증거 공간이 단일하게 주어지므로 다양한 증거가 필요할 경우에는 적용하기 어렵다. [5]는 패턴 및 전문가 Belief의 형태가 A->B형태에서 작성되며 발견된 패턴과 각 Belief의 condition part와 일치하는 아이템의 빈도수를 계산하여 4개의 공간 중 하나로 분류시키는 방법을 제안하였다. 이 방법은 Belief degree 계산 과정이 단순하며 명시적이다.

3. Web Usage 패턴 검증

패턴 검증은 웹로그 데이터 전처리 과정을 거쳐 최종적인 사용자 트랜잭션 파일을 입력 대상으로 한다[4]. 전문가가 자신의 Belief를 직접 입력하여 패턴과 Belief의 매칭을 수행하여 패턴을 분류하는 1차 검증과 1차 검증 결과를 Dempster-Shafer에 적용하는 2차 검증으로 패턴 검증 과정을 나눈다.

3.1 1차 검증

발견된 패턴 및 전문가 Belief는 다음과 같이 표현된다.

패턴 : $A_1, A_2, \ldots, An \rightarrow B_1, B_2, \ldots, B_m$

Belief : $S_1, S_2, \ldots, Sn \rightarrow V_1, V_2, \ldots, V_m$

<A,B,S,V는 웹페이지>

각 패턴과 Belief의 우측부분을 consequent part, 좌측부분을 condition part라고 한다. 1차 검증에서는 패턴과 Belief의 condition, consequent part를 매칭하여 패턴을 Confirming, Unexpected consequent, Unexpected condition, Both-side unmatched pattern 4가지[5]로 분류한다.

3.1.1 매칭 알고리즘

패턴과 Belief의 condition part의 웹페이지 매칭 비율을 L, 패턴과 Belief의 consequent part의 웹페이지 매칭 비율를 R이라고 정의한다. i번째 패턴의 L_i과 R_i 값은 아래의 알고리즘(1)을 이용하여 구한다[5].

(단, L : condition, R은 consequent, M은 매칭수, N은 총 페이지 수)

$$\text{if } \frac{LM_{ij}}{LN_i} > \frac{RM_{ij}}{RN_i} \text{ then}$$

$$L_{ij} = \min\left(\frac{LM_{ij}}{LN_i}, \frac{SM_{ij}}{SN_j}\right);$$

$$R_{ij} = \frac{RM_{ij}}{RN_i};$$

$$\text{else} \quad R_{ij} = \min\left(\frac{RM_{ij}}{RN_i}, \frac{SM_{ij}}{SN_j}\right);$$

$$L_{ij} = \frac{LM_{ij}}{LN_i};$$

<알고리즘 (1) : i번째 패턴과 j번째 Belief의 L_i, R_i>

3.1.2 분류 매칭 degree

알고리즘(1)에서 구한 L_i, R_i값을 이용하여 4가지 분류에 속하는 매칭 degree를 알고리즘(2)를 이용해 구한다[5].

$$confm_{ij} = L_{ij} * R_{ij};$$

$$unexpConseq_{ij} = \begin{cases} 0 & L_{ij} - R_{ij} \le 0 \\ L_{ij} - R_{ij} & L_{ij} - R_{ij} > 0 \end{cases};$$

$$unexpCond_{ij} = \begin{cases} 0 & R_{ij} - L_{ij} \le 0 \\ R_{ij} - L_{ij} & R_{ij} - L_{ij} > 0 \end{cases};$$

$$bsUnexp_{ij} = 1 - max(confm_{ij}, unexpConseq_{ij}, unexpCond_{ij});$$

<알고리즘 (2) : i번째 패턴의 분류 매칭 degree>

3.2 통계적 추론을 이용한 2차 검증

1차 검증에서 구해진 매칭 degree는 패턴과 Belief와의 웹페이지 빈도수를 바탕으로 하기 때문에 간단하게 패턴을 4가지로 분류하는 데 충분하지만 패턴의 interesting-ness 값을 신뢰하기는 부족하다. 따라서 신뢰성을 확보를 위한 추가 검증이 필요하다.

3.2.1 Dempster-Shafer theory

Dempster-Shafer는 상호 배타적인 특성을 지니는 속성으로 이루어진 가설 공간, 즉 Frame of Discernment 라는 집합을 정의한다. 예를 들면 의사가 환자의 병명을 결정하기 위해 설정해 놓은 가능한 모든 병명으로 이루어진 집합에 해당한다. Dempster-Shafer에 따르면 가설 공간 집합 Θ의 부분집합 2^{Θ}에 대한 확률을 기본확률(a basic probability)라고 하며 이는 Belief에 대한 검증 자료가 되는 증거(evidence) 집합들의 확률이다. 이 확률값을 이용하여 Dempster-Shafer는 주어진 증거(evidence)가 Belief를 만족시키는 구간 [Belief, Plausibility]에 놓이는 지를 측정하게 한다[2].

1차 검증 결과에서는 각 패턴마다 4가지 분류의 확률값을 가진다. 이는 Dempster-Shafer의 가설 공간 Θ의 부분집합 2^{Θ}에 대한 기본확률 값을 나타내며 가설 공간

집합은 {confm , unexpConseq , unexpCond, bsUnexp} 이며 구간 측정 알고리즘은 아래와 같다.

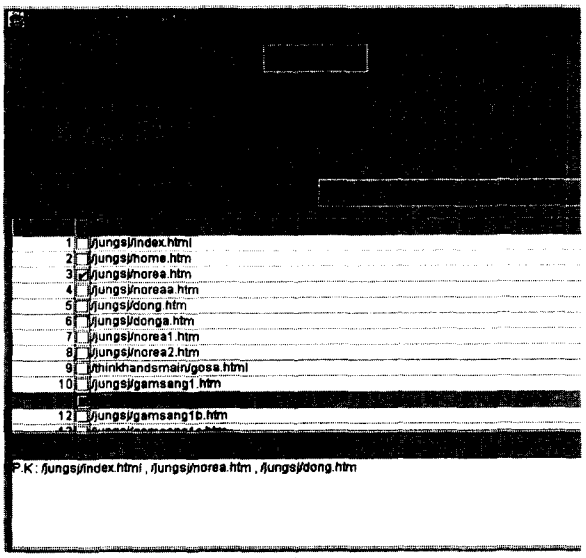
Let E ={confm,unexpConseq,unexpCond,bsUnexp}
m(confm) = confm * 0.25
m(unexpConseq) = unexpConseq * 0.25
m(unexpCond) = unexpCond * 0.25
m(bsUnexp) = bsUnexp * 0.25

Belief(E) = $\sum_{E'} m(E')$ **(E' are all subsets of E)**

Plausibility(E) = 1 - Belief(E^c)

4. 실험

웹로그파일을 전처리 하여 250개의 파일로 이루어진 208개의 트랜잭션을 얻었으며 지지도 0.1, 신뢰도 0.9로 하여 Apriori 알고리즘 이용하여 50개의 패턴을 찾아내었다. 다음 전문가의 Belief를 입력 받아 각 패턴마다 confm, unexpConseq, unexpCond, bsUnexp 값을 계산하고 2차 검증 모듈의 입력 데이터로 이용하였다. <그림1>은 전문가로부터 Belief를 입력받는 화면이다.



<그림1>

<표1>은 1차 검증의 unexpConseq 패턴에 대한 2차 검증 결과를 보여주고 있다. 5개의 patterns에서 1차 검증의 결과가 2차 검증에서 3,5번 pattern이 accept 되지 않음을 보이고 있다.

N	pattern	unexpConseq	1차 accept	B	P	2차 accept
1	home,index->dong	0.75	yes	0.5	1	yes
2	/jungsi/index -> /jungsi/minsok	0.65	yes	0.5	0.85	yes
3	/jungsi/minsok -> study, dona	0.5	yes	0.65	0.85	no
4	home, dong -> donga, index	0.5	yes	0.5	0.85	yes
5	home, dong ,index-> study, donga	0.6	yes	0.45	0.55	no

<표1>

5. 결론

발견된 Web Usage 패턴들은 분석하는 전문가에게는 불필요하고 흥미롭지 못해 의사결정에 도움이 못되는 경우가 많다. 이 문제점을 해결하기 위해 전문가의 사전 Belief에 기반한 패턴 검증 방법을 제안하였다. 검증 방법은 1,2차로 나누어 지며 특히 2차 검증은 통계적 추론 방법인 Dempster-Shafer theory를 이용하여 검증의 효율성을 보였다.

6. 참고문헌

[1] R. Agrawal and R. Srikant. "Fast algorithm for mining association rules." In Proc. of the 20th VLDB Conference, 1994.

[2] Silberschatz, A., and Tuzhilin, A. "what make patterns interesting in knowledge discovery systems." IEEE Trans. On Know. And Data Eng.,1996

[3] Liu,B., and Hsu,W. "post-analysis of learned rules." AAA-96, 1996

[4] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." In *SIGKDD Explorations*, 2000.

[5] Bing Liu, Wynne Hsu, Shu Chen and Yiming Ma. "Analyzing the Subjective Interestingness of Association rules." IEEE Intelligent Systems, 2000

[6] Adomavicius, G., and Tuzhilin, A. "Expert-Driven Validation of Rule-Based User Models in Personalization Applications." International Journal on Data Mining and Knowledge Discovery. January 2001