The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 20, No. 1, pp.19-25, Feb. 29, 2020. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2020.20.1.19
JIIBC 2020-1-3

Fuzzy Utility를 활용한 연관규칙 마이닝 시스템을 위한 알고리즘의 구현에 관한 연구

A Study on the Implementation of an optimized Algorithm for association rule mining system using Fuzzy Utility

박인규*, 최규석**

In-Kyu Park*, Gyoo-Seok Choi**

요 약 빈발 패턴 마이닝에서 각 패턴이 가지는 불확실한 정보로 인하여 정보의 손실을 수반하기 마련이다. 또한 실제적인 환경에서는 패턴들의 중요도가 시간에 따라서 변하기 때문에 이러한 요구에 부합하기 위하여 퍼지논리를 적용하고 패턴이 가지는 중요도의 동적특성을 고려하여야 한다. 본 논문에서는 웹 로그 데이터베이스에서 퍼지 유틸리티 기반 웹 페이지 집합 마이닝을 통해 웹 로그 데이터베이스에서 빈발 웹 페이지 집합의 추출을 위한 퍼지 유틸리티 마이닝 기법을 제안한다. 여기서 퍼지 집합의 하향 폐쇄 특성은 최소 퍼지 유틸리티 임계 값(MFUTV) 및 사용자 정의 백분위 수(UDP)에 의해 넓은 공간을 제거하기 위해 적용된다. 여러 실험을 통하여 제안하는 기법은 매우 효과이며 확장성이 좋은 것임을 보인다.

Abstract In frequent pattern mining, the uncertainty of each item is accompanied by a loss of information. AAlso, in real environment, the importance of patterns changes with time, so fuzzy logic must be applied to meet these requirements and the dynamic characteristics of the importance of patterns should be considered. In this paper, we propose a fuzzy utility mining technique for extracting frequent web page sets from web log databases through fuzzy utility-based web page set mining. Here, the downward closure characteristic of the fuzzy set is applied to remove a large space by the minimum fuzzy utility threshold (MFUT)and the user-defined percentile(UDP). Extensive performance analyses show that our algorithm is very efficient and scalable for Fuzzy Utility Mining using dynamic weights.

Key Words : Frequent pattern, Quantitative value, Fuzzy data mining, Fuzzy FP-Tree, Dynamic weight

Ⅰ. 서 론

웹 사용 마이닝은 웹 마이닝 범주 중 하나로서 방문자가 웹 사이트를 서핑 할 때 웹 서버의 로그 파일에 기록된 사용자 활동에서 시간에 따라서 변하는 유용한 패턴을 마케팅 전략에 활용되고 있다. 예를 들어, 웹상의 데이터와 지식은 부정확하고 불완전하며 불확실한 데이터로 구성되어 진다. 퍼지 개념은 종종 이러한 데이터를 처리하는 데 사용되므로, 퍼지 및 언어 지식을 밝히기 위한 여러 가지 퍼지 웹 마이닝 기법이 존재한다[1].

이러한 거래 데이터베이스로부터 유틸리티 기반의 최상위 유틸리티 아이템 집합 마이닝에서는 선명한 경계

*정회원, 중부대학교 게임소프트웨어공학과
**종신회원, 청운대학교 컴퓨터공학과(교신저자)
접수일자: 2019년 12월 16일, 수정완료: 2020년 1월 16일
게재확정일자: 2020년 2월 7일

Received: 16 December, 2019 / Revised: 16 January, 2020 / Accepted: 7 February, 2020
*Corresponding Author: ikpark@joongbu.ac.kr
Dept. of Computer Science, Chungwoon University, Korea

문제로 어려움을 갖게 되기 때문에, 퍼지 이론을 적용하여 얻어진 높은 정확도를 기반으로 임계값을 이용하여 빈발 패턴이 추출된다[2].

실제적으로 이진 논리를 기반으로 정량 데이터베이스를 처리하기는 한계가 있다. 따라서 빈발패턴(Frequent Pattern) 집합을 추출하기 위해 많은 전략이 개발되었다. 그러나 기존 전략은 많은 패턴후보를 생성하고 트랜잭션에서 퍼지 유틸리티를 계산하는 데 많은 시간을 소비해야했다. 이러한 이유로 양적 데이터베이스에서 퍼지 규칙을 효율적으로 찾는 것이 매우 중요한 문제로 대두된다. 빈발패턴을 추출하기 위해 거의 모든 기존 알고리즘이 먼저 후보 패턴 집합을 생성 한 다음 각 후보의 정확한 유틸리티를 계산하여 빈발패턴을 식별하고, 웹 로그 데이터베이스에서 퍼지 유틸리티 웹 페이지를 추출하는 기법을 제안한다. 이 기법에서는 집합의 각 멤버에 적절한 값을 할당하기 위해 새로운 퍼지 집합 멤버쉽 기능이 정의되었다. 이 방법은 후보 집합을 생성하지 않지만 퍼지 기반의 빈발 웹 페이지 집합만 저장하며 사용자는 최소 퍼지 유틸리티 값의 백분율 값을 설정할 수 있다. 또한 웹 데이터베이스에서 하나의 스캔만 필요하기 때문에 전체적으로 계산 시간을 단축시켜 마이닝 속도를 향상시킨다. 제안된 기법은 실행 시간에서 IHUP (Incremental High Utility Pattern), UP-Growth, HUI-Miner (High Utility Itemset Miner) 및 FHM (Faster High-Utility Itemset Mining)의 기존 알고리즘보다 양호하다는 것을 보이고자 한다.

II. 연구 배경

1. 빈발 패턴 마이닝(Frequent Pattern Mining)

집합 I는 $\{i_1, i_2, \cdots i_m\}$로 구성되는 패턴들이고, 집합D는 $\{T_1, T_2, \cdots T_n\}$를 트랜잭션으로 구성되는 데이터베이스라 하자. 트랜잭션 $T_i \in D$는 I의 부분 집합으로 구성되어진다. 이 때 데이터베이스에서 패턴 $X = \{x_1, x_2, \cdots x_n\}$를 포함하고 있는 트랜잭션의 개수를 패턴의 빈도수라고 한다. 트랜잭션 데이터베이스에 나타난 여러 패턴에서 빈도수가 주어진 임계값보다 크거나 같은 패턴을 빈발 패턴이라고 한다. 만일 어떤 패턴 p가 빈발하지 않은 패턴이면 p를 포함하는 모든 집합은 빈발하지 않게 된다. 따라서 FP-트리를 이용한 FP-Growth 알고리즘은 이러한 문제를 극복하여 많은 성능 개선을 보여왔다[3].

2. 퍼지 유틸리티 빈발패턴 마이닝(Fuzzy Utility Frequent Pattern Mining)

QARM (Quantative Association Rule Mining)은 데이터베이스 내부에 숨겨져 있을 수 있는 상관 관계 및 패턴을 추출한다. 기존의 빈발 패턴 마이닝 알고리즘은 트랜잭션에서 빈발 패턴의 빈도와 패턴 수량만 고려한다. 그러나 실질적인 비즈니스 환경에서 상품구매와 가격 패턴은 시변(Time Variant)적이다. 또한 웹분석 환경에서도 웹 페이지의 중요도는 다르게 설정된다. 따라서 가중치나 중요도는 실제적인 응용에서 중요한 역할을 한다. 이로 인해 트랜잭션 데이터베이스에서 유틸리티를 기반으로 하는 마이닝이 생겼다. 상한(Upper) 모델을 사용하여 데이터베이스의 최대 가중치(이익) 값을 각 트랜잭션의 가중치 상한으로 채택한 새로운 하향 폐쇄 특성을 제안했다. IHUP[4]기법은 다중 데이터베이스 스캔을 피함으로써 마이닝의 성능을 향상 시켰고, UP-Growth은 UP-Tree에서 높은 유틸리티 패턴 집합을 효율적으로 얻을 수 있지만 두 번의 스캔으로 시간적인 문제가 있다[5]. MUGrowth 트리 기반 알고리즘은 다수의 후보를 줄임으로써 높은 유틸리티 패턴집합의 마이닝 효율을 가진다[6]. HUI_Miner는 후보를 생성하지 않고 마이닝하여 실행 시간을 줄였지만 패턴 검색비용이 증가한다[7]. 또한 항목에 대한 유틸리티를 동시에 적용하여 필터링을 통한 유틸리티 기반 패턴집합 마이닝기법으로 FHM이 제안되었다[8]. 그러나 이러한 방법들은 선명한 경계 문제로 인해서 메모리 공간이 많아지고 계산 시간이 길어져서 사실상 큰 데이터베이스에서는 처리하기 어려운 단점이 존재한다. 이 문제를 해결하기 위해 퍼지 집합의 언어적 변수를 이용하여 표현되는 양의 선명한 경계를 부드럽게 처리하여 정보의 손실을 막을 수 있다. 따라서 선명한 경계 문제와 물리적 측정의 부정확성을 보다 잘 처리하기 위하여 퍼지 유틸리티 마이닝(fuzzy utility mining)을 활용하고자 한다.

III. 제안된 방법

1. 퍼지 분할

초기에 각 속성의 퍼지 분할을 위하여 퍼지 c-means 클러스터링에 의한 클러스터(Cluster)의 중심점을 구하여 퍼지집합을 각각 구성한다. 퍼지 c-means 알고리즘

은 다음과 같이 정의된다.

$$\sum_{i=1}^{N}\sum_{j=1}^{c}\mu_{ij}{}^{m}\|x_i-c_j\|^2 \quad (1)$$

$$\mu_{ij}=1/\sum_{k=1}^{c}\left(\frac{\|x_i-c_j\|}{\|x_i-c_k\|}\right)^{\frac{2}{m-1}}$$

$$c_j=\sum_{i=1}^{N}\mu_{ij}^{m}x_i/\sum_{i=1}^{N}\mu_{ij}^{m}$$

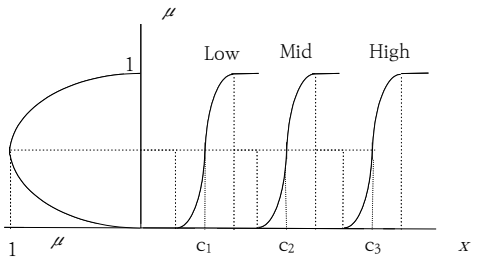

그림 1. 퍼지 분할
Fig. 1. Fuzzy Partition

$$F(x)=-x\times\ln(x)-(1-x)\times\ln(1-x) \quad (2)$$

표 1. 각 속성의 정의된 외부 유틸리티
Table 1. Predefined external utility of each attribute

Attr.		a_1	a_2	a_3	a_4	a_5
EU	t_1, t_2, t_3	3	7	5	2	5
	t_4, t_5, t_6	6	2	5	3	5
	t_7, t_8, t_9	6	4	7	4	6

표 2. 속성의 내부 유틸리티를 갖는 트랜잭션
Table 2. Transactions with internal utility of attributes

Tid / Attr.	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
a_1	5	8	0	5	7	0	5	3	0
a_2	0	2	3	3	0	2	2	0	4
a_3	10	3	9	10	9	8	5	10	0
a_4	2	0	0	0	3	3	0	2	2
a_5	9	0	0	3	0	0	0	2	0

그림 1에서 c_1, c_2 와 c_3는 퍼지 클러스터의 중심점으로 각각의 퍼지 분할로 언어적 변수(Low, Med, High))를 나타낸다. 그런 다음에 이 변수에 대하여 샤논함수(Shannon Function)를 적용하여 각 속성의 값의 식(2)의 정보 엔트로피(Information Entropy)를 구하여 최종적인 퍼지 집합(F)을 구성한다.

2. 정의

마이닝에서 사용되는 외부 유틸리티 값은 표 1처럼 1에서 10까지로 항목의 가중치로 가정하고. 데이터베이스의 트랜잭션은 내부 유틸리티로서 표 2와 같이 항목이 가지는 빈도수로 구성된다. 예를 들어 i 번째 항목의 유틸리티 값 (a), 즉 $U(a_i)$는 내부 유틸리티(IU)와 외부 유틸리티(EU) 값을 이용하여 다음과 같이 정의된다. 유틸리티 값을 모두 계산하고 정규화해 준다. 그리고 그 정규화한 값을 지정된 임계값을 기준으로 그 이하의 값을 가지는 항목들을 필터링한다. ur은 0과 1 사이로 지정한 0.6으로 가정한다.

$$U(a_i)=ur\times IU(a_i)+(1-ur)\times EU(a_i) \quad (3)$$

T_i에서 j 번째 웹 페이지 a_j의 값 V_{ij}의 퍼지 집합 F_{ij}[9]는 다음과 같이 웹 페이지 a_z에 대해 주어진 멤버쉽 함수에 의해 다음과 같이 나타낼 수 있다.

$$F_{ij}=\sum_{k=1}^{n}f_{ijk}/R_{jk} \quad (4)$$

여기서 n은 웹 페이지 a_j의 영역 수이고, R_{jn}은 a_j의 n 번째 퍼지 영역이고, f_{ij} n은 V_{ij}의 퍼지 멤버쉽 값을 나타낸다. 정량적 트랜잭션 T_i에서 패턴 a_j의 k 번째 퍼지 영역의 퍼지 유틸리티 FU_{ijk}는 외부 유틸리티이다 [10]. a_j의 $EU(a_j)$는 k 번째 퍼지 영역 R_{ijk}에서 V_{ij}의 정량적 값 V_{ij} 및 퍼지 멤버쉽 값 F_{ijk}를 곱한 것으로 다음과 같다.

$$FU_{ijk}=F_{ijk}\times V_{ij}\times EU(a_j) \quad (5)$$

T_d에서 웹 페이지 a_i의 최대 퍼지 유틸리티 (Maximum Fuzzy Utility) MFU_{di}은 다음과 같이 정의된다.

$$MFU_{di}=\max\{\sum_{k=1}^{n}FU_{dik}\} \quad (6)$$

여기서 FU_{di}은 T_d에서 웹 페이지 a_i의 n 번째 퍼지 영역 (R_{in})의 퍼지 유틸리티 값이다. DB에서 웹 트랜잭션 (T_d)의 최대 퍼지 트랜잭션 유틸리티 (MFTU)[11]는 T_d에 있는 모든 웹 페이지의 최대 퍼지 유틸리티 값을 더한 것으로 다음과 같다.

$$MFTU(T_d)=\sum_{a_1\in T_d}MFU(a_{id}) \quad (7)$$

MFU(a_{id})는 d 번째 트랜잭션 T_d ∈DB에서 i 번째 웹 페이지의 최대 퍼지 유틸리티이다. T_d ∈DB에서 웹 페이지 집합 X의 퍼지 트랜잭션 가중 이용률 (Fuzzy Transaction Weight) FWU[12]는 다음과 같이 정의된다.

$$FWU(X) = \sum_{X \subseteq T_d \in D} MFTU(T_d) \quad (8)$$

최소 퍼지 유틸리티 임계 값(Minimum Fuzzy Utility Threshold) FUT_{min}[25]는 사용자에 의해 고정 된 값이며 총 최대 퍼지 트랜잭션 유틸리티에 의존하며 다음과 같다[13].

$$FUT_{\min}(X) = ur \times \sum_{T_d \in D} MFTU(T_d) \quad (9)$$

3. 마이닝 과정

웹 로그 데이터에서 퍼지 유틸리티 기반 웹 페이지 집합 마이닝 알고리즘은 그림 2와 같다. 첫 번째 단계는 웹 서버에서 웹 로그 데이터 세트를 수집하여 웹 로그 트랜잭션 데이터베이스(DB)를 형성하는 것이다. 표 1과 같이 MFUTV, EU의 사용자 정의 값을 갖는 두 개의 매개 변수가 각각 마이닝 프로세스에 제공된다. 두 번째 단계는 DB에 속하는 트랜잭션 T_d를 스캔하고, 표 2와 같이 해당 웹 페이지의 내부 유틸리티를 사용하여 웹 트랜잭션 테이블을 만든다. 최종적으로 표 3과 같이 그림 1의 퍼지 집합을 이용하여 웹 트랜잭션을 퍼지 분할하여 퍼지집합 (Low, Med, High)을 구성한다.

표 3. 각 속성의 퍼지 영역에 대한 퍼지 유틸리티
Table 3. Fuzzy utilities of fuzzy region of each attribute

Attr. \ Tid	t1	t2	t3	t4	t5	t6	t7	t8	t9
a1(A).L	0.271	0	0	0.271	0	0	0.271	0.687	0
a1(A).M	0.185	0.555	0	0.185	0.691	0	0.185	0	0
a1(A).H	0	0.403	0	0	0.093	0	0	0	0
a2(B).L	0	0.479	0.687	0.687	0	0.479	0.479	0	0.576
a2(B).M	0	0	0	0	0	0	0	0	0
a2(B).H	0	0	0	0	0	0	0	0	0
a3(C).L	0	0.687	0	0	0	0	0.271	0	0
a3(C).M	0.004	0	0.245	0.004	0.245	0.555	0.185	0.004	0
a3(C).H	0.630	0	0.659	0.630	0.659	0.403	0	0.630	0
a4(D).L	0.479	0	0	0	0.687	0.687	0	0.479	0.479
a4(D).M	0	0	0	0	0	0	0	0	0
a4(D).H	0	0	0	0	0	0	0	0	0
a5(E).L	0	0	0	0.687	0	0	0	0.479	0
a5(E).M	0.245	0	0	0	0	0	0	0	0
a5(E).H	0.659	0	0	0	0	0	0	0	0

표 4. 최대 트랜잭션 퍼지 유틸리티
Table 4. Maximum transaction fuzzy utility

Attr. \ Tid	t	t2	t3	t4	t5	t6	t7	t8	t9	FTWU
a1(A).L	0.974	0	0	1.949	0	0	1.949	2.96	0	55.38
a1(A).M	0.666	3.199	0	1.332	6.967	0	1.332	0	0	62.37
a1(A).H	0	2.323	0	0	0.942	0	0	0	0	22.85
a2(B).L	0	1.609	3.461	0.989	0	0.460	0.920	0	2.21	46.7
a2(B).M	0	0	0	0	0	0	0	0	0	0
a2(B).H	0	0	0	0	0	0	0	0	0	0
a3(C).L	0	2.472	0	0	0	0	2.273	0	0	0
a3(C).M	0.052	0	2.644	0.052	2.644	5.332	1.554	0.07	0	88.81
a3(C).H	7.565	0	7.119	7.565	7.119	3.871	0	10.5	0	83.67
a4(D).L	5.747	0	0	0	1.483	1.483	0	0.92	0.92	63.24
a4(D).M	0	0	0	0	0	0	0	0	0	0
a4(D).H	0	0	0	0	0	0	0	0	0	0
a5(E).L	0	0	0	2.472	0	0	0	1.37	0	28.83
a5(E).M	2.644	0	0	0	0	0	0	0	0	21.45
a5(E).H	7.119	0	0	0	0	0	0	0	0	21.45
MFTU	21.41	7.28	10.58	12.98	15.57	7.28	5.14	15.86	3.13	

표 5. FTWU를 가지는 2-웹 페이지 집합
Table 5. 2-webpage ses with FTWU

No.	2-itemsets	FTWU	No.	2-itemsets	FTWU
1	a1.L, a1.M	39.167	12	a1.M, a4.L	52.473
2	a1.L, a2.L	18.122	13	a2.L, a3.L	*12.422*
3	a1.L, a3.M	55.025	14	a2.L, a3.M	35.982
4	a1.L, a3.H	65.453	15	a2.L, a3.H	30.84
5	a1.L, a4.L	36.903	16	a2.L, a4.L	*10.412*
6	a1.L, a5.L	28.838	17	a3.M, a3.H	83.313
7	a1.M, a1.H	22.85	18	a3.M, a4.L	59.753
8	a1.M, a2.L	25.402	19	a3.M, a5.L	28.838
9	a1.M, a3.L	12.422	20	a3.H, a4.L	59.753
10	a1.M, a3.M	54.737	21	a3.H, a5.L	28.838
11	a1.M, a3.H	65.453			

세 번째 단계는 웹 로그 데이터베이스에서 IU 및 할당된 EU 값을 사용하여 T_d에 속하는 모든 웹 페이지(a_i) 에 대하여 MFTU를 계산하고, MFTU를 이용하여 FTWU 값을 계산한다. 표 5에서 T_1 트랜잭션의 최대 퍼지 거래 유틸리티 (MFTU)는 다음과 같이 계산된다. MFTU(T_1) = max(a1.L, a1.M,) + max(a3.M, a3.H) + a4.L + max(a5.M, a5.H) = max(0.974, 0.666) + max (0.052, 7.565) + 5.747 + max (2.644, 7.119) = 21.405. 비슷한 방법으로 T_2, T_3, T_4, T_5, and T_6의 MFTU를 계산한다.

네번째 단계는 FTWU(a_i)의 값이 MFUTV (μ)의 값보다 작은 경우에 해당하는 빈발패턴은 제거한다.

2-항목의 빈발패턴의 (a1.L, a3.H)의 FTWU는 (a1.L, a3.H) = T_1+T_4+T_5+T_8 = 21.405 + 12.98 + 15.57 + 15.858 = 65.453. 마찬가지로, 2-항목의 패턴의 다른 조합에 대한 FTWU 값을 표 5과 같이 계산할 수 있다. 또한 표 5에서 9, 13과 16번의 집합인 {(a1.M, a3.L =

12.422), (a2.L, a3.L = 12.422), (a2.L, a4.L = 10.412)} 최소 유틸리티 기준 (μ = 15)을 충족하지 않으므로 삭제되어 진다.

다섯 번째 단계에서는 같은 방식으로 3,4,5-항목에 대해서도 각각의 빈발패턴을 생성하고 최소 퍼지 임계 값 (μ = 15)의 임계값의 필터링을 통하여 퍼지 유틸리티 기반의 빈발패턴을 추출한다.

```
Algorithm
Input: Web transactioin database, center, band, clusters(c1,c2,c3)
Output: the complete set of frequent patterns
Do clustering of attributes in transactions
Calculate information entropy according to the clusters
Read transaction Td, where Td∈DB
Fuzzify web transaction into fuzzy regions(Low, Med, High)
Calculate MFTU for each Td∈DB
Calculate FTWU of every webpage (ai)∈Td
If FTWU(ai) ≥ MFUT(μ)
   Calculate FU of frequent 1-itemset
   Generate {frequent 2-itemsets s.t. MFUT(μ)}
   Prune {all sets with v | v 〈 min fuzzy threshold}
   Generate {frequent n-itemsets | FTWU ≥ MFUT}
   Return fuzzy utility webpage sets
else
   Discard webpage(ai)
```

그림 2. 제안된 알고리즘
Fig. 2. The proposed algorithm

Ⅳ. 실험 및 결과

알고리즘의 성능을 알아보기 위하여 실행시간을 측정하였다. 제안된 방법의 검증을 위하여 기존의 IHUP(UP-Growth) 및 HUIMiner(FHM 알고리즘)와 비교하였다. 일관성을 위하여 최소 퍼지 유틸리티 임계값 (MFUT)을 15로 유지하면서 실험하였다[14,15,16].

그림 3은 제안된 기법이 기존의 알고리즘인 IHUP, UP-Growth, HUI-Miner 및 FHM 보다 실행시간이 짧다는 것을 알 수 있다. IHUP 알고리즘은 과도한 후보 항목의 발생으로 인하여 계산 시간이 가장 높다. UP-Growth는 하나의 트리와 두 번의 스캔을 통하여 시간을 비교적 단축시켰다. HUI-Miner는 후보 생성을 피하여 시간을 크게 단축하지만 많은 시간이 소요되는 조인(join)작업을 수행한다. FHM은 후보 생성 및 유틸리티 계산 수를 줄여서 실행 시간을 단축하였다. 제안된 기법은 이러한 전체 알고리즘 중에서 가장 짧은 시간에 빈발 항목을 추출하였다. 그 이유는 후보 항목 집합을 생성하지 않고 오히려 퍼지 항목만 저장하고 바람직하지 않은 항목 집합을 제거하기 때문이다.

그림 4는 서로 다른 최대 퍼지 유틸리티 임계값인 MFUT에서 기존의 방법인 IHUP, FHM, HUI-Miner, UP-Growth 및 제안된 기법과의 소요 된 시간을 비교한 것이다. 웹 트랜잭션 수만 건에 대해 MFUT를 백분율로 변경하여 실험을 수행하였다. MFUT가 증가함에 따라 많은 수의 낮은 퍼지 유틸리티 웹 페이지 제거로 인해 실행 시간이 실질적으로 감소함을 보여준다. 처음에서 제안된 방법의 실행 시간이 비교적 높은 이유는 MFUT의 비율이 증가할수록 많은 수의 웹 트랜잭션이 처리되기 때문이다. MFUT 4 % 이후부터 퍼지 낮은 유틸리티를 가지는 웹 피이지는 제거되므로 실행 시간이 더 빠르다. 따라서 제안된 방법의 실행 시간이 양호하여 퍼지 유틸리티 기반의 웹 페이지 집합 마이닝의 효율성이 높아졌다고 볼 수 있다.

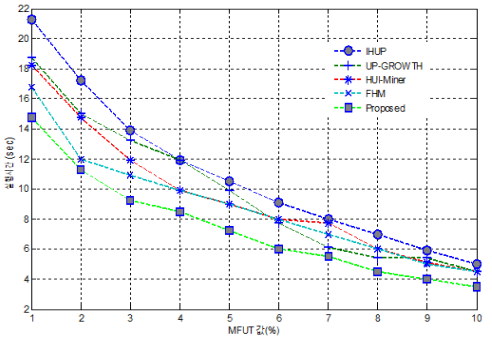

그림 3. 실행시간 대 웹 트랜잭션 수
Fig. 3. Running time vs number of web transactions

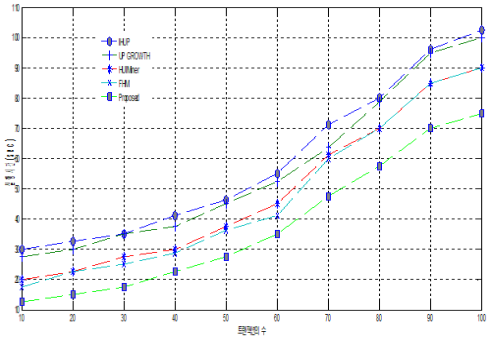

그림 4. 실행시간 대 최소 임계값
Fig. 4. Running time vs min_threshhold

그림 5은 서로 다른 MFUT 백분율 값에서 생성된 웹 페이지 집합의 여러 가지 빈발 패턴을 나타낸다. 고정 된 웹 트랜잭션 수를 유지하면서 여러 MFUT 값에서 생성되는 웹 페이지의 빈발 패턴 수를 관찰하기 위해 수행되었다. IHUP와 UP-Growth는 가장 많은 수의 웹 페이지

집합 패턴을 생성한다. 다른 방법들은 적은 수의 빈발 웹 페이지 집합을 생성하지만 높은 유틸리티 웹 페이지 집합이다. 그림 6은 웹 트랜잭션 수에 대해 생성 된 웹 페이지 집합의 여러 가지 패턴을 보여준다. 제안된 방법은 적은 수의 빈발웹 페이지 집합을 생성하지만 퍼지 유틸리티가 높다. IHUP은 웹 페이지의 빈발집합을 많이 생성하지만 웹 페이지 빈발집합이 높은 유용성을 보장하지는 않는다. 결국 제안된 방법은 높은 유틸리티를 가지는 웹 페이지 빈발집합을 추출하는 측면에서 기존의 알고리즘보다 양호하다.

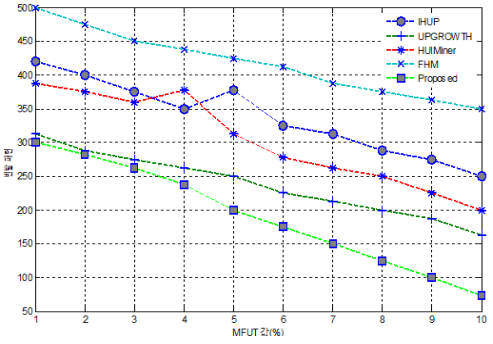

그림 5. 빈발 웹 페이지 집합 패턴 대 최소 임계값
Fig. 5. Frequent webpage sets vs minimum threshhold

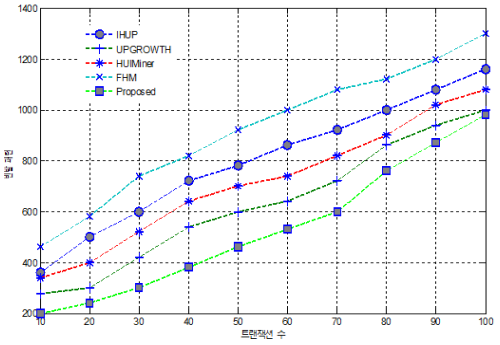

그림 6. 빈발 웹 페이지 집합 대 트랜잭션 수
Fig. 6. Frequent webpage sets vs number of transactions

V. 결 론

본 논문에서는 웹 로그 데이터베이스에서 퍼지 유틸리티를 기반으로 웹 페이지의 빈발집합을 추출하는 웹 페이지를 마이닝 기법을 제안하였다. FCM과 엔트로피 개념을 사용하여 속성들의 초기분할을 수행하고, 새로운 퍼지 유틸리티 기능을 도입하고 퍼지 집합의 하향 폐쇄 특성이 최소 퍼지 유틸리티 임계 값 (MFUT) 및 사용자 정의 백분위 수 (UDP) 접근 방식으로 넓은 공간을 제거하는 데 적용되었다. 실험 결과는 제안한 방법이 실행 시간 및 웹 로그 데이터베이스에서 높은 유틸리티 웹 페이지 집합 추출 측면에서 기존의 알고리즘보다 양호함을 보였다. 기존의 알고리즘보다 약 20~ 35 % 더 효율적이다. 따라서 이러한 접근 방식(제한된 방식)을 통해 얻은 결론은 인공지능형 검색엔진 내부에 지식 베이스를 추가하여 지능형 검색에 활용될 수 있을 것으로 사료된다.

References

[1] M. Zdravko and T. L. Daniel, “Data Mining the Web, Uncovering Patterns In Web Content, Structure and Usage”, John Wiley & sons Inc., New Jersey, USA, pp. 115-132, 2007.
https://doi.org/10.18637/jss.v025.b01

[2] C. W. Lin and T. P. Hong, “A survey of fuzzy web mining”, Data Mining Knowledge Discovery, Vol.3, No. 13, pp. 190-199, 2013
https://doi.org/10.1002/widm.1091

[3] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases", In: Proc. of ACM Sigmod Record, ACM. Vol. 22, pp. 207-216, 1993.
https://doi.org/10.1145/170036.170072

[4] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo, “Fast discovery of association rules”, In: Proc. of Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, pp. 307-328, 1996.
https://doi.org/10.1109/fskd.2008.332

[5] U. Yun and J.J Leggett, "Wfim: Weighted frequent itemset mining with a weight range and a minimum weight", In: Proc. of International Conf. on Data Mining, SIAM, pp. 636-640, 2005.
https://doi.org/10.1137/1.9781611972757.76

[6] C.F. Ahmed, S.K. Tanbeer, B. S. Jeong, and Y. K. Lee, "Efficient tree structures for high utility pattern mining in incremental databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 12, pp. 1708-1721, 2009.
https://doi.org/10.1109/tkde.2009.46

[7] S. Vincent, C. W. Tseng, B. E. Wu, Shie, and P. S. Yu, “UP-Growth: An Efficient Algorithm for High Utility Itemset Mining”, In: Proc. of ACM-Knowledge Data Discovery, Washington, DC, USA, pp. 253-262, 2010.
https://doi.org/10.1145/1835804.1835839

[8] U. Yun, H. Ryang, and K.H. Ryu, “High utility itemset mining with techniques for reducing over-estimated

utilities and pruning candidates", Expert Systems with Applications, Vol. 41 No. 8, pp.3861-3878, 2014.
https://doi.org/10.1016/j.eswa.2013.11.038

[9] C. M. Wang, S. H. Chen, and Y. F. Huang, "A fuzzy approach for mining high utility quantitative itemsets", In: Proc. of Fuzzy Systems, FUZZ-IEEE International Conf. on IEEE, pp. 1909-1913, 2009.
https://doi.org/10.1109/fuzzy.2009.5277408

[10] K. K. Mohbey, "High fuzzy utility based frequent patterns mining approach for mobile web services sequences", International Journal of Engineering (IJE), TRANSACTIONS B: Applications", Vol. 30, No. 2, pp. 182-191 , 2017.
https://doi.org/10.5829/idosi.ije.2017.30.02b.04

[11] L.A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning-1", Information Science, Vol.8, pp.199-249, 1975.
https://doi.org/10.1016/0020-0255(75)90017-1

[12] S. G. Matthews, M. A. Gongora, A. A. Hopgood, and S. Ahmadi, "Web usage mining with evolutionary extraction of temporal fuzzy association rules", Knowledge Based System, Vol. 54 pp. 66-72, 2013.
https://doi.org/10.1016/j.knosys.2013.09.003

[13] V. Kumar, R. S. Thakur "High Fuzzy Utility Based Sets Mining from Weblog Database", Int. Journal of Intelligent Engineering & System, Vol. 16, pp. 191-200, 2017.
https://doi.org/10.22266/ijies2018.0228.20

[14] Y. S. Im, E. Y. Kang, "MPEG-2 Video Watermarking in Quantized DCT Domain," The Journal of The Institute of Internet, Broadcasting and Communication(JIIBC), Vol. 11, No. 1, pp. 81-86, 2011.
https://doi.org/10.1109/tip.2006.873476

[15] I. Jeon, S. Kang, H. Yang, "Development of Security Quality Evaluate Basis and Measurement of Intrusion Prevention System," Journal of the Korea Academia-Industrial cooperation Society (JKAIS), Vol. 11, No. 1, pp. 81-86, 2010.
https://doi.org/10.5762/kais.2010.11.4.1449

[16] J. S. Oh, B, S. Lee, "A Study for Lifespan Prediction of Expansion by Temperature Status,", The Journal of KISTI, Vol. 19, No. 10, pp. 424-429637, 2018.
http://dx.doi.org/10.5762/KAIS.2018.19.10.424

저 자 소 개

박 인 규(정회원)

- 제10권 5호 참조
- 현 중부대학교 게임S/W학과 교수
- 관심분야 : 데이터 마이닝, 지능시스템

최 규 석(종신회원)

- 제9권 6호 참조
- 1991 ~ 1995년 : (주)SK텔레콤 중앙연구원 책임연구원
- 현 청운대학교 컴퓨터공학과 교수
- 관심분야 : 인공지능, ITS, 이동컴퓨팅

※ 본 논문은 2019학년도 청운대학교 교내학술연구조성비에 의하여 지원되었음.