1) 침해지표 기반 구축된 데이터 셋에 대한 연관성 분석

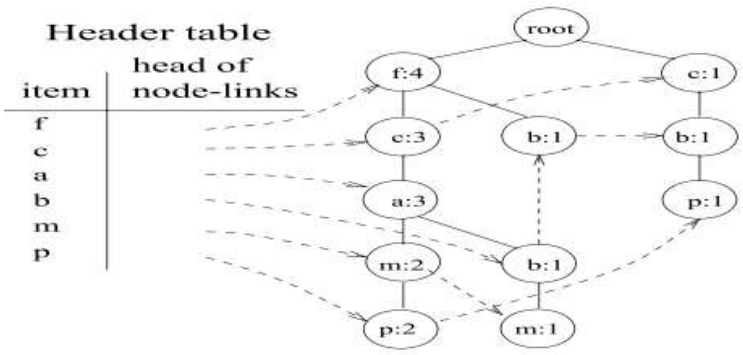
본 단계에서는 앞서 구축한 데이터 셋들을 기반으로 침해사고 분류 및 연도를 중점으로 연관성 분석을 수행한다. 연관성 분석에서는 FP-growth 알고리즘을 적용하였으며, 해당 알고리즘을 통해 각 데이터 셋 혹은 두 가지 이상의 데이터 셋을 결합하여 다양한 규칙들을 생성한다.

FP-growth 알고리즘이란 연관규칙 알고리즘의 하나로, 처음 제안되었던 apriori 알고리즘의 문제를 해결하기 위해 제안되었다. Apriori 알고리즘은 아이템 셋 후보를 만들고 이들을 하나씩 검사하는 방법으로 트랜잭션의 수가 많을 때는 해당 트랜잭션을 모두 읽기 때문에 시간이 오래걸린다는 단점이 존재한다. 따라서 FP-growth 알고리즘은 해당 문제를 해결하기 위해 트리와 노드링크를 기본으로 조건을 만족하는 아이템들을 연결해 아이템 셋을 생성한다. 즉, 아이템의 조건을 확인하고 조건을 만족한다면, 다음 아이템과 연결하고, 만족하지 않는다면 중지시켜서 결과적으로 후보를 만들지 않으면서 조건 확인이 가능하다는 것이다. 이는 결과적으로 단 2번의 DB스캔만 이루어지면 되기 때문에 Apriori 보다 빠르다.

해당 알고리즘은 다음과 같은 순서로 진행된다. ① 트랜잭션 리스트를 스캔해 트랜잭션이 포함하고 있는 아이템마다 support를 계산한다. ② 구해진 support를 기준으로 minsup 이상의 아이템들만 골라낸다. ③ 골라낸 아이템들을 support 내림차순으로 정렬해 [그림 38]과 같이 테이블에 정렬한다. ④ 트랜잭션 리스트를 다시 스캔하면서, 트랜잭션에 포함된 아이템 중 테이블에 있는 아이템만 골라내 트리에 삽입한다. ⑤ 트랜잭션마다 루트부터 삽입하며, 새로운 아이템이 나타나면, 포인터로 연결해 노드를 가리키고 동일한 아이템이 나올 경우, 노드를 합치고 support에 1을 추가한다. 이렇게 5가지 과정을 반복해 트리를 생성한다.

TID	Items Bought	(Ordered) Frequent Items
100	f, a, c, d, g, i, m, p	f, c, a, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m, n	f, c, a, m, p

[그림 38] Support에 따른 테이블 정렬



[그림 39] FP-Growth 내 tree 생성 과정

본 과제에서는 FP-growth를 적용하여 수행된 침해사고 기반 악성코드를 통해 추출된 아티팩트를 활용한 정규화된 데이터 셋이 연관성 분석, 기계학습 및 인공지능 등과 같은 알고리즘에 사용될 수 있는 지를 확인 할 것이다.