한국컴퓨터정보학회 하계학술대회 논문집 제27권 제2호 (2019. 7)

가중치 부여 부정 트리 패턴 추출

백주련*, 김진영O
평택대학교, 데이터정보학과O
평택대학교, 데이터정보학과*
e-mail: jrpaik@ptu.ac.kr*, wlsdud1517@naver.comO

Weighted Negative Tree Pattern Discovery

Juryon Paik*, Jinyoung KimO
Dept. of Digital Information & Statistics, Pyeongtaek UniversityO
Dept. of Digital Information & Statistics, Pyeongtaek University*

● 요 약 ●

사물인터넷(IoT)은 지금의 우리가 살고 일하는 모든 방식을 변화시키고 있다. IoT를 통해 데이터를 생성하고 저장하고 연결된 장치와 상호작용하여 비즈니스는 물론 우리의 일상 생활을 개선하고 있는 것이다. 무수히 많은 센서들이 연결된 세상은 센서들에 의해 그 어느 때보다 거대한 양의 데이터들을 생산하고 있다. JSON, XML 같은 트리 구조의 데이터 타입은 대량 데이터 저장 · 전송 · 교환 등에 주요하게 사용되는데 이는 트리 구조가 이형 데이터 간의 유연한 정보 전송과 교환을 가능하게 하기 때문이다. 반면에, 효용성 높은 정보나 감추어져 있는 정보들을 트리 구조의 대량 데이터들로부터 추출하는 것은 일반 데이터 구조에 비해 훨씬 어려우며 더 난해한 문제들을 발생시킨다. 본 논문에서는 트리 구조의 대량 스트리밍 데이터로부터 가중치가 부여된 주요한 부정 패턴들을 추출하기 위한 방법을 공식화한다.

키워드: 연관규칙(association rules), 부정패턴(negative pattern), 가중치 임계값(weighted thresholds)

I. Introduction

세상의 모든 사물들이 네트워크로 연결되고, 지능화되며, 이를 활용하여 사용자에게 유익하고 편리한 다양한 서비스 제공을 목적으로 하는 사물인터넷(Internet of Things)은 학계뿐만이 아니라 거의 모든 분야에서 관심이 증폭되고 있다. 2015년 맥킨지 보고서[1]에 의하면 사물인터넷이 갖는 잠재적 경제 효과는 2025년까지 매년 3조 9천억 달러부터 11조 1천억 달러까지로 전망하고 있다.

최신 기술들로 무장한 사물인터넷 환경은 근간이 되는 센서 기술로 인해 거대한 양의 스트리밍 데이터들을 생산한다. 과거 웹 데이터 범람 시 웹 데이터 저장·전송·교환의 표준으로 사용되었던 XML처럼 작금의 스트리밍 센서 데이터들의 저장·전송·교환을 위해 XML 보다 가볍지만 이형 데이터 처리에 용이한 JSON이 대표적으로 사용된다. XML과 JSON의 공통점은 데이터 구조로 모두 트리 구조를 갖는다. 트리 구조는 데이터 종류에 상관없이 저장과 전송 그리고 교환에 있어서 유연하기 때문에 이형 데이터 처리에 주로 사용된다. 그러나, 사용의 용이성은 분석에 어려움을 초래한다. 평면적인 구조의 데이터보다 복잡한 트리 구조는 특유의 계층 구조로 인해 유용하고 숨겨진 정보를 추출하는 데이터마이닝 분야에서는 어려운 대상으로 여겨진다. 게다가 연속적으로 발생하는 스트리밍 데이터라면 그 어려움과 복잡함은 배가 된다. 본 논문은 대량의 트리 구조 스트리밍 데이터로부터 드러나지 않지만 발생 가치에 따라 가중치를 갖는 부정 패턴을 추출하고자 한다.

II. Preliminaries

1. Tree-structured data

사물인터넷에서 가장 대중적인 데이터 인코딩은 JSON의 사용이다. XML처럼 JSON 또한 표현이 유연하고 이형 데이터 간에 교환이 용이할 뿐만 아니라 가볍기 때문에 센서로부터 발생하는 데이터의 저장과 전송에 주로 이용된다. Fig. 1에 보인 JSON 코드는 Fig.2에 나타낸 것처럼 트리 구조로 모든 데이터를 표현하기 때문에 트리형 데이터가 갖는 장점과 단점을 갖게 된다. 즉, 이형 간의 자유로운 전송 교환이라는 장점 대신 가치 있는 정보 추출에 있어서 어렵다는 단점이다.

```
{
   "firstName": "John",
   "lastName": "Smith",
   "age": 25,
   "address": {
      "streetAddress": "21 2nd Street",
      "city": "New York",
      "state": "NY",
      "postalCode": "10021-3100"
   },
   "gender": {
      "type": "male"
    }
   "spouse": null
   "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    },
    {
      "type": "mobile",
      "number": "123 456-7890"
    }
   ]
}
```


Fig. 1. JSON Code

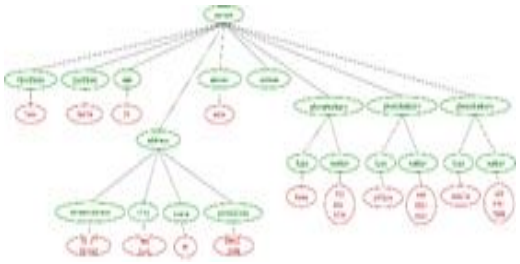

Fig. 2. JSON Code의 트리 구조화

2. Negative association rules

대표적인 연관규칙 추출은 의미 있는 긍정패턴들을 얼마나 정확하게 발견하는가 하는 것이다. 긍정패턴은 발생하는 데이터들 간의 관계를 분석 후 특정 임계값을 상회하는 패턴의 경우 자주 발생하며 서로 연관이 있다고 판별하는 것이다. 부정패턴은 긍정패턴과는 달리 발생하지 않은 데이터들 간의 연관을 분석한다. 그런데 발생하지 않은 데이터들 간의 연관을 분석하기 위해서는 발생한 데이터들의 역을 사용하기 때문에 그 과정이 긍정 연관규칙보다 더 복잡하고 어렵다. 긍·부정 연관규칙 모두 기본적으로 두 개의 주요 임계값인 support와 confidence를 적용한다.

2014년 Paik et al.에 의해 발표된 논문[2]은 트리 구조 스트리밍 데이터들에 대한 support와 confidence 값을 정의한다. 사용된 스트리밍 트리 데이터 세트는 다음과 같이 정의되는데 블록 단위로 대량의 스트리밍 데이터들이 전송되며 각 블록 TB_i는 타임스태프 정보도 같이 전송된다.

$$|S| = \sum_{i=1}^{L}|TB_i| = |TB_1| + \cdots + |TB_L|$$
$$= \left|\sum_{j=1}^{k_1} T_j\right| + \left|\sum_{j=1}^{k_2} T_j\right| + \cdots + \left|\sum_{j=1}^{k_n} T_j\right|$$
$$= \left|\sum_{i=1}^{n}\sum_{j=1}^{k_i} T_{ij}\right|.$$

데이터 세트 S로부터 특정 트리 X에 대한 support는 해당 트리 구조의 발생 빈도로 다음과 같이 정의된다. 전체 데이터 크기에 비례해서 특정 서브트리가 어느 정도의 빈도로 발생했는가를 계산한다.

$$sup(X) = \frac{|\{T_{k_1}|(X \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| + \frac{|\{T_{k_2}|(X \subseteq T_{k_2}) \wedge (T_{k_2} \in TB_2)\}| + \cdots}{|S|} + \frac{\left|\{T_{k_L}|(X \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\}\right|}{|S|}}{|S|}$$

위의 공식에 기반을 두어 발생하지 않은 트리 Y, ¬Y라 표기, 와의 support 값은 다음과 같다.

$$sup(X \Rightarrow \neg Y) =$$
$$\frac{|\{T_{k_1}|(X \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| + \cdots + \left|\{T_{k_L}|(X \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\}\right| -}{|S|}$$
$$\left(\frac{|\{T_{k_1}|(X \subseteq T_{k_1}) \wedge (Y \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| + \cdots}{|S|}\right.$$
$$\left. + \frac{\left|\{T_{k_L}|(X \subseteq T_{k_L}) \wedge (Y \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\}\right|}{|S|}\right)$$

support(X ⇒ ¬Y)가 의미하는 것은 특정 트리 X가 발생하면 특정 트리 Y는 발생하지 않는다는 대표적인 부정 연관규칙을 위한 임계값으로 주어진 임계값보다 크다면 의미 있는 부정 연관규칙이라고 할 수 있다. 이와 더불어 또 다른 임계값 confidence는 두 개의 특정 서브트리 X와 Y의 발생이 어느 정도의 상관을 갖고 발생하는 정도를 측정한다.

$$conf(X \Rightarrow \neg Y) = \frac{sup(X \Rightarrow \neg Y)}{sup(X)}$$
$$= 1 - \frac{\|\{T_{k_1}|(X \subseteq T_{k_1}) \wedge (Y \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| + \cdots + |\{T_{k_L}|(X \subseteq T_{k_L}) \wedge (Y \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\}|}{|\{T_{k_1}|(X \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| + \cdots + |\{T_{k_L}|(X \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\}|}$$

III. The Proposed Scheme

블록 단위로 전송되는 스트리밍 데이터들에 대한 일반적인 연관규칙은 전체 데이터 세트에 대한 발생 빈도를 근간으로 한다. 즉, 동일 블록 내에서 발생한 특정 패턴의 경우 몇 번을 발생하더라도 해당 가중치는 1이다. 블록 크기가 작다면 가중치는 의미가 없겠지만 일반적으로 블록들의 크기 역시 대량의 사이즈다. 따라서 블록별로 동일한 패턴이 발생했다 할지라도 내포하고 있는 정보의 의미가

다른 것이다. 본 논문에서는 이를 해결하기 위해 가중치를 부여하기 위해 블록별로 support와 confidence 값을 계산하고자 한다. 전체 데이터 세트 S가 아닌 특정 블록의 대상으로 한 support와 confidence 공식은 다음과 같다.

$$bsup(X \Rightarrow \neg Y, TB_i) =$$

$$\frac{\left|\left\{\left\langle T_{k_i}\middle|(X \subseteq T_{k_i})\right\rangle\right\}\right| - \left|\left\{\left\langle T_{k_i}\middle|(X \subseteq T_{k_i}) \wedge (Y \subseteq T_{k_i})\right\rangle\right\}\right|}{|TB_i|}$$

$$bconf(X \Rightarrow \neg Y, TB_i) =$$

$$1 - \frac{\left|\left\{\left\langle T_{k_i}\middle|(X \subseteq T_{k_i}) \wedge (Y \subseteq T_{k_i})\right\rangle\right\}\right|}{\left|\left\{\left\langle T_{k_i}\middle|(X \subseteq T_{k_i})\right\rangle\right\}\right|}$$

같은 부정 패턴에 대해 블록별로 임계값을 차별적으로 부여할 수 있으며 이를 기반으로 어느 블록에서는 의미 있는 패턴이 어느 블록에서는 무의미한지를 판별할 수 있으며 판별 정도에 따라 블록별로 가중치에 따른 의미 부여가 가능하다.

Fig. 3은 두 개의 블록으로 구성된 스트리밍 트리 데이터 세트 S이다. 각 블록은 4개의 트리 데이터들로 구성된다. 공식의 간단화를 위해 블록 사이즈를 동일하게 가정한다. 전체 데이터 사이즈 |S|는 8개의 트리로 구성된다. Fig.4는 임의의 트리 패턴 X와 부정 패턴 ¬Y에 대한 가능한 연관규칙 X ⇒ ¬Y를 보인다.

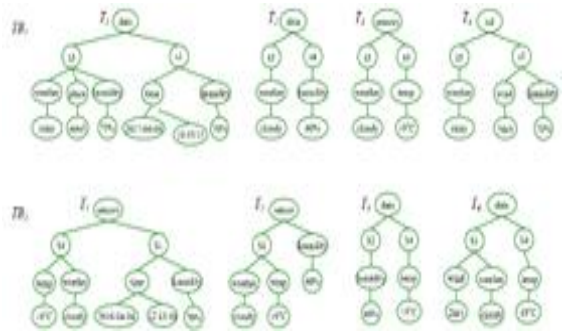

Fig. 3. 두 개의 블록으로 구성된 스트리밍 데이터세트 S

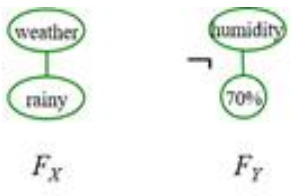

Fig. 4. 임의의 부정패턴 X ⇒ ¬Y

두 임계치 support와 confidence를 S 전체에 적용한 결과와 블록별로 적용한 결과는 다음과 같다. 동일한 패턴일지라도 전체에 대한 값과 TB_1 그리고 TB_2 에 따라 그 값이 상이함을 알 수 있다. 만약 support 임계값을 0.3으로 정했다면 패턴 X는 전체 데이터 세트 S에 대해서는 기준 값을 만족하지 못하기 때문에 제거되지만 각 블록에서는 기준 값보다 크기 때문에 추출됨을 알 수 있다. 즉, 추출될 부정 패턴의 효용을 스트리밍 데이터 전체로 할 것인지 아니면 특정 블록별로 할 것인지에 따라 추출되는 부정 패턴의 범위가 달라진다.

$$sup(F_X) = \frac{2}{8} = 0.25$$

$$sup(F_Y) = 1 - \frac{3}{8} = 0.625$$

$$bsup(F_X, TB_1) = \frac{2}{4} = 0.5$$

$$bsup(F_X, TB_2) = \frac{0}{4} = 0$$

$$bsup(F_Y, TB_1) = 1 - \frac{2}{4} = 0.5$$

$$bsup(F_Y, TB_2) = 1 - \frac{1}{4} = 0.75$$

연관규칙 추출에 있어서 기본이 되는 support와 confidence 임계값 외에 추출되는 부정패턴 제거단계의 정확성을 높이기 위해 두 개의 또 다른 임계값, interestingness와 correlation coefficient 값을 추가하며 해당 임계값들에 대해서도 가중치 부여를 한다. Fig. 5에 제시된 알고리즘은 총 4개의 임계값들에 대한 가중치 부여 유무를 포함한다.

INP: **S** OUTP: *WNT or N-WNT*

1. IF (*weight*)
2. FOR EACH block $TB_i \in \mathbf{S}\ (1 \le i \le k)$
3. FOR $j \leftarrow 1$ to n
4. IF $freq(X_{ij}, \mathbf{S}) \ge |\mathbf{S}| \times \delta$
5. THEN $FT = FT + \{X_{ij}\}$;
6. FOR itemset $X \subset FT, Y \subset FT, X \cap Y = 0$
7. IF $sup\ (X \Rightarrow Y) \le ms$
8. or $conf\ (X \Rightarrow Y) \le mc$
9. THEN
10. IF $interest(X, \neg Y) < mi$
11. THEN
12. IF $\phi_{(X, \neg Y)} \le -0.3$ or $\phi_{(X, \neg Y)} \ge +0.3$
13. THEN $WNT \leftarrow WNT + \{X \Rightarrow \neg Y\}$;
14. ELSE THEN
15. $WNT \leftarrow WNT + \{X \Rightarrow \neg Y\}$;
16. ELSE THEN
17. $WNT \leftarrow WNT + \{X \Rightarrow \neg Y\}$;
18. ELSE THEN
19. FOR SOME block $TB_i \in \mathbf{S}\ (1 \le i \le k)$
20. FOR $j \leftarrow 1$ to n
21. IF $freq(X_{ij}, TB_i) \ge |TB_i| \times \delta$
22. THEN $FT = FT + \{X_{ij}\}$;
23. FOR itemset $X \subset FT, Y \subset FT, X \cap Y = 0$
24. IF $bsup\ (X \Rightarrow Y) \le ms$
25. or $bconf\ (X \Rightarrow Y) \le mc$
26. THEN
27. IF $binterest(X, \neg Y) < mi$
28. THEN
29. IF $b\phi_{(X, \neg Y)} \le -0.3$ or $b\phi_{(X, \neg Y)} \ge +0.3$
30. THEN $N\text{-}WNT \leftarrow N\text{-}WNT$
31. $+ \{X \Rightarrow \neg Y\}$;
32. ELSE THEN
33. $N\text{-}WNT \leftarrow N\text{-}WNT$
34. $+ \{X \Rightarrow \neg Y\}$;
35. ELSE THEN
36. $N\text{-}WNT \leftarrow N\text{-}WNT + \{X \Rightarrow \neg Y\}$;
37. RERURN *NTS or N-WNT*
38. END

Fig. 5. 4개 임계값 가중치부여 부정패턴 추출 알고리즘

IV. Conclusions

대량의 스트리밍 트리 구조 데이터로부터 블록별 가중치 부여에 의한 부정패턴 연관규칙 추출은 특정 데이터 구간에 대해서 발견할

수 있는 정보의 차별성을 부여할 수 있다. 일괄적으로 전체 데이터 세트를 대상으로 했을 경우 제거될 가능성이 있는 정보가 특정 구간에서는 매우 유용할 수 있기 때문이다. 차후 본 연구에서는 실제적인 데이터를 대상으로 한 실험을 통해 정확한 부정패턴의 추출을 증명하고자 한다.

ACKNOWLEDGEMENT

이 논문은 2019년도 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1A2B1007015).

REFERENCES

[1] J. Manyika and M. Chui, (2015). By 2025, Internet of things applications could have $11 trillion impact, *Available:http://www.mckinsey.com/insights/mgi/in the news/by 2025 internet of things applications could have 11 trillion impact.*

[2] J. Paik, J. Nam, U. Kim, and D. Won, (2014). Association rule extraction from xml stream data for wireless *sensor* networks, Sensors, 14, 12937-12957.