빈발 패턴 탐사 기법을 이용한 반구조적 데이터로부터의 공통구조 추출

이영언°, 문봉희
숙명여자대학교 컴퓨터과학과

Extracting Common Structure of Semistructured data Using mining frequent patterns

Young-Eon Lee° Bong-Hee Moon
Dept. of Computer Science, Sookmyung Women's University

요 약

인터넷의 발달로 웹에는 엄청난 데이터가 존재하나, 불규칙적인 구조를 이루고 있는 반구조적 데이터가 대부분이다. 이러한 반구조적 데이터는 데이터들간의 어떤 정확하게 정해진 구조를 갖고 있진 않지만 불완전하고 불규칙한 구조 정보를 포함하고 있는 것으로, 데이터들 간의 관계를 규명할 수 있는 공통 구조 정보를 추출하여 효과적으로 구조화시킴으로써 정보로서의 가치를 높일 필요성이 대두되게 되었다. 또, 데이터 처리 과정에서 기존의 잘 정의된 구조를 가진 데이터베이스의 장점을 수용하기 위해서는 반구조적 데이터 집합의 불완전한 구조 정보로 부터 공통 구조를 추출하는 것이 요구된다.
본 연구에서는 후보 항목 집합의 생성이 없는 빈발 패턴 탐사 기법을 사용하여 반구조적 데이터 집합으로부터 공통구조를 추출하고자 한다.

1. 서론

반구조적 데이터(Semistructured data)란 데이터들간의 어떤 정확하게 정해진 구조를 갖고 있진 않지만 불완전하고 불규칙한 구조 정보를 포함하고 있는 데이터를 말한다 [1]. 즉, 각 데이터마다 어떤 내재된 구조가 존재하는 데이터를 반구조적 데이터라고 한다. 이러한 반구조적 데이터들의 예로는 BibTex, HTML, XML등의 문서들을 들 수 있으며 이들은 일반 문서와 달리 태그에 의한 구조 정보를 포함하고 있으나, 각 데이터에 대한 구조만을 표현하고 있다. 따라서, 반구조적 데이터는 본질적인 특성상 기존의 관계형 모델이나 객체 지향형 모델로 모델링하는 것은 어려우며, 대부분의 경우에 있어서 관련 데이터 집합의 공통적인 구조를 미리 결정하는 것은 불가능하다. 이러한 이유로 지금까지 반구조적 데이터에 관한 연구는 반구조적 데이터에 대한 데이터 모델을 제시하고, 각 모델에 대한 질의에 관한 방향으로 많이 이루어져 왔다. 또한, 웹에서의 HTML 문서들과 같은 유용한 반구조적 데이터들간의 의미적, 구조적 관계를 설정하고 불완전한 구조 정보를 이용하여 커다란 반구조적 정보 공간에서 데이터들 사이의 관계를 규명할 수 있는 공통 구조 정보를 추출하여 효과적으로 구조화하려고 시킴으로써 정보로써의 가치를 높일 필요성이 대두되게 되었다.

이러한 구조 정보는 사용자에게 데이터 집합에 대한 일관된 뷰를 제공하여 데이터들 간의 관계를 명확히 해주고, 사용자가 필요한 정보를 찾기 위한 질의의 기반이 된다. 따라서 반구조적 데이터에서 구조 정보를 추출함으로써 유용한 정보를 효과적으로 얻을 수 있으며, 반구조적 데이터의 데이터베이스 구조화를 가능하게 함으로서 데이터베이스 관리 시스템의 여러 장점을 수용할 수 있다. 또, 여러 데이터 소스 통합을 용이하게 하며 문서의 유효성 검사, 효과적인 데이터 처리, 개선된 저장소 그리고 사용자를 위한 안내로써 매우 중요한 방법으로 이용된다.

본 논문에서는 빈발 패턴 탐사 기법을 사용하여 반구조적 데이터 집합에서 공통 구조를 추출하는 방법을 제안하고자 한다. 즉, 빈발 패턴 탐사 기법을 사용하여 사용자 정의의 최소 지지도를 만족하는 빈발 트리 패턴을 찾아내고자 한다.

2. 관련연구

[3]은 반구조적 데이터의 스키마를 간결(concise)하고 정확하게(accurate) 요약하는 DataGuide를 소개하고 있다. DataGuide는 데이터베이스 구조를 볼 수 있고, 질의를 생성하고 통계 정보와 샘플 값을 저장할 수 있는 동적 스키마를 제공한다.

반구조적 데이터는 고정된 스키마가 없으므로 사용자에게 정보의 전체구조와 내용에 관한 요약이 필요하다. [4]에서는 반구조적 데이터의 각 구조가 서로 다르므로 완전하고 정확한 기술을 기대할 수는 없지만 데이터 집합을 적절히 분류하여 타입 계층를 추출하는 방법론을 제시하고 있다.

[5]에서는 데이터로그 프로그램의 최대 고정점을 이용하여 타입을 추출하는 방법을 제시하고 있다. 근사 타입 추출방법을 보이는데 단순히 최대 고정점을 이용하여 타입을 추출하게 되면, 많

은 수의 타입이 생성되며 경우에 따라서는 타입이 실제 데이터와 비슷한 양 만큼이 생성되는 경우도 있기 때문에 이것을 해결하는 방법으로 클러스터링(clustering) 방법을 이용하여 타입을 줄이고 있다.

위의 방법들은 반구조적 데이터를 처리하는 것에 비중을 둔 스키마 추출 방법이다. 다시 말해, 스키마를 반구조적 데이터의 저장과 질의 처리에 이용하는 측면에서 스키마 추출이 목적이다. 반구조적 데이터를 나타내기 위해 객체를 노드로 나타내고 객체들간의 관계를 간선 위의 레이블로 표현하는 그래프 모델에서 타입을 찾는 것으로, 같은 종류의 타입으로 분류되는 객체에서 나오거나 또는 들어가는 같은 종류의 간선들을 갖는 객체들을 그룹핑하는 것이다.

반면에 [6]에서는 스키마를 표현하는 방법인 DTD를 정확하게 추출하는데 중점을 두고, 특히 XML문서에 존재하는 구성 요소들을 정규 표현식으로 정확히 표현하는데 목적이 있다.

[7]는 마이닝 기법을 이용하여 반구조적 데이터의 구조 추출 방법을 제시하고 있다. 사용자가 정의한 최소 지지도를 넘는 최대한의 빈번한 트리 구조를 찾는 것을 목적으로 하고 있다.

3. 공통 구조의 추출

3.1 추출 모델

반구조적 데이터는 라벨 간선 그래프로 표현되는 OEM(Object Exchange Model)을 기본 모델로 하여 주로 표현되고 있다. 그러나 실제적으로는 반구조적 데이터의 한 형태인 XML[2]에 대한 관심이 급증하면서 이에 대한 연구가 활발히 진행 중이므로 본 논문에서는 XML 문서에 대한 공통구조를 추출하고자 한다.

XML 문서는 중첩된 태그 엘리먼트들로 구성되어있으며, 모든 태그 엘리먼트들은 여러 개의 속성과 그 값을 가질 수 있고 여러 개의 하위 엘리먼트를 가질 수 있는 특징이 있으며, XML문서는 어떤 특정 데이터 모델을 가지고 정의된 것은 아니지만, 대신에 응용프로그램에서 XML문서에 대한 처리를 위해 표준적으로 정의된 DOM(Document Object Model)이 있다.

DOM은 XML문서가 응용프로그램상에서 용이하게 사용될 수 있도록 트리 형태를 취하며 XML데이터에 대한 처리의 시작점 역할을 한다.

OEM에서 라벨은 시작점과 다른 객체와의 관계를 나타내는 포인터 역할을 하지만 XML DOM에서 각각의 엘리먼트는 식별 태그를 가지고 있다.

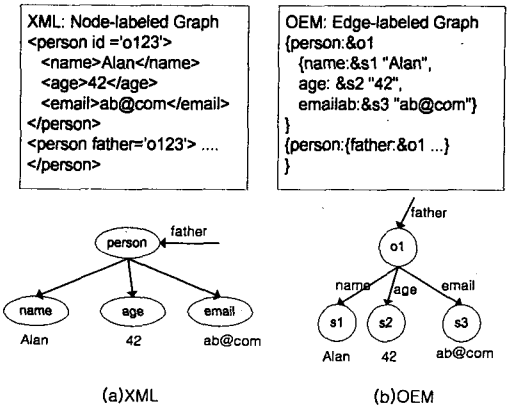

그림1. XML 과 OEM

그림1에서 (a)와 (b)는 XML과 OEM모델을 따르는 반구조적 데이터를 각각 트리 형태로 나타낸 것이다. 가장 큰 차이는 라벨이 XML 트리에서는 노드에 붙어 있고 OEM에서는 간선에 붙어 있다는 것이다. (a)와 (b)에 나타난 노드 라벨 그래프와 간선 라벨 그래프는 구현상의 차이일 뿐이다.

3.2 공통 구조 추출

빈발 항목 집합을 찾는 알고리즘은 후보(candidates)라 불리는 빈발 가능성이 있는 항목 집합들의 생성을 포함하는데, 이러한 알고리즘의 목적은 후보 항목들의 수와 빈도를 줄이는데 있다. 만약 이 부분을 고려하지 않는다면 항목들의 크기는 기하급수적으로 증가하게 되어 알고리즘의 성능을 급격히 저하시키는 결과가 된다.

따라서 본 논문에서는 [8]에서 제안한 후보항목 생성없이 빈발항목 패턴을 찾는 알고리즘을 반구조적 데이터의 공통구조를 추출하는데 사용가능 하도록 수정하여 적용하고자 한다.

먼저 공통구조를 발견하는 작업에서 각 XML 문서로부터 객체를 수집하는 작업이 필요하다. 즉, 만약 사용자가 영화 객체에 흥미를 갖고 이에 대한 공통구조를 추출하고자 한다면 영화 객체들을 표현할 수 있는 노드경로 표현법이 필요하며 이를 수집해야 하고, 공통구조를 추출하는 단계는 다음과 같다.

단계1 XML 문서로부터 공통 구조를 추출하는 방법에 있어 본 논문에서의 가장 중요한 특성은 XML DOM 트리로부터 [7]에서 제시한 것과 비슷하게 노드 경로를 표현하는 방법이다. 깊이 우선 순위(depth-first-order)방법으로 노드의 경로를 표현하고 이 경로 표현(path-expression) 들이 구조를 나타내는 트리 표현식(tree-expression)을 구성한다.

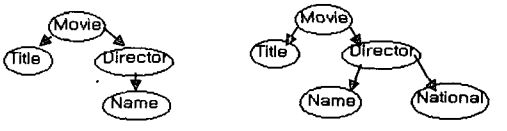

그림 2. 영화객체의 트리 표현의 예

예를 들어, 그림2와 같은 XML DOM 트리 구조가 있다면 노드 경로를 표현하는 pi는 깊이 우선 순위로 추출되어 다음과 같이 표현 되어진다. 여기서 ㅗ는 널(nil) 구조임을 표시한다.

p1 = [Movie, Title, ㅗ]
p2 = [Movie, Director, Name, ㅗ]
p3 = [Movie, Director, National, ㅗ]

그리고, (T1) 와 (T2) 의 트리 표현은 다음과 같이 나타낼 수 있다

a={Movie:{Title : ㅗ}, Movie : {Director :{Name :ㅗ}}
b={Movie:{Title : ㅗ}, Movie:{Director : {Name : ㅗ,
Nationality : ㅗ}}}

위의 두 트리 표현 T1은 2-시퀀스 (p1, p2) 로 이루어졌으며, T2는 3-시퀀스 (p1, p2, p3)로 이루어져 있다.

이러한 방법으로 각 XML 문서에 대한 DOM 트리로부터 1-트리 표현 즉, 모든 1-시퀀스 pi를 추출한다.

단계2 추출된 1-시퀀스 즉, 경로 표현에 대해 공통구조를 추출하기 위한 첫번째 작업으로 사용자가 정의한 지지도를 만족하는 빈발 경로 표현 항목을 계산하여 뽑아낸다. 각 XML 문서 단위별로 추출된 1-시퀀스는 데이터베이스에 저장되어 스캔 작업을 거쳐 1-시퀀스의 빈도수가 계산된다.

단계3 계산되어진 최소지지도만을 만족하는 빈발 경로 표현 항목들을 (빈발 경로 표현 항목i : 지지도)의 형식으로 내림차순의

리스트를 구축한다.
각 문서에 대해서 리스트에 따른 각 빈발 경로 표현 항목들을 선택하고 내림차순으로 정렬한다.
트리의 루트는 "널(null)" 값을 갖고, 루트의 자식들은 (항목 이름, 총계, 노드링크)로 이루어진 항목 전위 부트리(ite-m prefix subtree) 의 집합과 (항목 이름, 노드링크 헤더)로 이루어진 빈발항목 헤더 테이블로 구성되어진 빈발 패턴 트리(FP-tree:Frequent pattern tree)를 구축한다.

Input : XML 문서(Doc) 1-path-expression DB 와
최소 지지도
Output : frequent pattern tree (FP_Tree)

Method
1. Create Order List L *//최소지지도를 만족하는 항목을 내림차순으로 정렬한 리스트*
2. Construction FP-Tree
   1) Create the root of an FP_Tree
   2) Order List L 에 따라 각 XML 문서에 대해 Pattern list P추출
   3) While (Doc_i != null)
      While (P != null)
        If T의 child N.node_name = p_i
                                // p_i : P의 item
          child N.node_name 의 count++
        else
          create new node N
          new node N count = 1
          parent node 와 link

<FP-tree 구축 알고리즘>

단계4 FP-tree로부터 조건부 패턴(Conditional Pattern) 을 발견한다. FP-tree의 빈발 헤더 테이블에서 시작하여 각 빈발 경로 표현 항목에 대해 FP-tree의 링크를 추적한다. 즉, 조건부 패턴을 구축하기 위해 빈발 항목의 모든 선행 노드의 경로를 추적한다.
단계5 각 패턴 기반에 대해 발견된 조건부 패턴에서 각 항목에 대한 총계를 계산하여 조건부 항목의 빈발 항목들에 대해 조건부 FP-tree를 구축하여 빈발 트리 패턴을 추출해 낸다.
그림3은 위의 다섯 단계를 영화 객체에 적용하여 추출된 빈발 트리 패턴을 트리 구조로 보인 것이다.

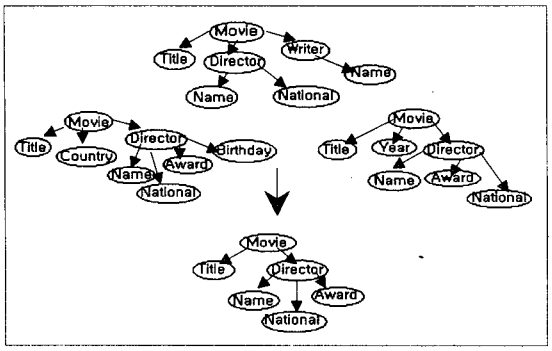

그림 3. 공통 구조 추출의 예

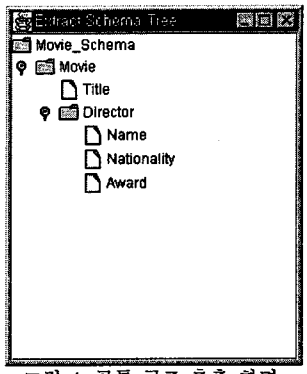

그림 4. 공통 구조 추출 화면

4. 결론

본 연구에서는 반구조적 데이터의 한 인스턴스인 XML 데이터 집합에 대해 후보항목 생성없이 빈발 패턴을 탐사하는 기법을 적용하여 공통 구조를 추출해 보았다. 지식 탐사 기법을 반구조적 데이터 처리 과정에 적용해 본 것은 향후 연구로도 좋을 것으로 기대된다.
웹은 하나의 커다란 데이터 창고와 같다. 그러나, 웹 데이터는 공통된 스키마를 가질 수 없고, 그 구조가 불규칙하므로 본 연구에서와 같은 공통 구조를 추출하는 기법의 데이터 처리과정을 거친다면 웹에서 얻어진 정보를 잘 정의된 구조를 가진 데이터베이스의 장점도 수용할 수 있을 것이다.

5. 참고 문헌

[1] Serge Abithboul, "Querying Semistructured Data", Proceedings of the '97 ICDT, Delphi. Greece, 1997
[2] Dan Suciu, "Semistructured Data and XML", Proceedings of International Conference on Foundations of Data Organization , 1998
[3] R. Goldman, J. Widom. "DataGuide : Enabling Query Formulation and Optimization In Semistructured Databases", VLDB, pp. 436-445, 1997
[4] S. Nestorov, S. Abiteboul, R. Motwani, "Inferring Structure in Semistructured Data", Proceeding of the Worshop on Management of Semistructured Data, Tucson, Arizona, 1997
[5] S. Nestorov, S. Abiteboul, R. Motwani, "Extracting Schema from Semistructured Data", SIGMOD, pp. 295-306, 1998
[6] M. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri, K. Shim. "XTRACT: A System for Extracting Document Type Descriptors from XML Documents", SIGMOD, pp. 165-176, 2000
[7] K. Wang, H.Q. Liu, "Discovering association of structure from semistructured objects", To appear in IEEE Transactions on Knowledge and Data Engineering, 1999
[8] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'00), Dallas, TX, May 2000