한국컴퓨터정보학회 하계학술대회 논문집 제26권 제2호 (2018. 7)

지능형 지식서비스를 위한 의미기반 XML 마이닝 시스템 연구

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

Development of Semantic-Based XML Mining for Intelligent Knowledge Services

Juryon Paik*, Jinyeong KimO
O*Dept. of Digital Information & Statistics, Pyeongtaek University

● 요 약 ●

XML을 대상으로 하는 연구가 최근 5~6년 사이에 꾸준한 증가를 보이며 이루어지고 있지만 대다수의 연구들은 XML을 구성하고 있는 엘리먼트 자체에 대한 통계적인 모델을 기반으로 이루어졌다. 이는 XML의 고유 속성인 트리 구조에서의 텍스트, 문장, 문장 구성 성분이 가지고 있는 의미(semantics)가 명시적으로 분석, 표현되어 사용되기 보다는 통계적인 방법으로만 데이터의 발생을 계산하여 사용자가 요구한 질의에 대한 결과, 즉 해당하는 정보 및 지식을 제공하는 형식이다. 지능형 지식서비스 제공을 위한 환경에 부합하기 위한 정보 추출은, 텍스트 및 문장의 구성 요소를 분석하여 문서의 내용을 단순한 단어 집합보다는 풍부한 의미를 내포하는 형식으로 표현함으로써 보다 정교한 지식과 정보의 추출이 수행될 수 있도록 하여야 한다. 본 연구는 범람하는 XML 데이터로부터 사용자 요구의 의미까지 파악하여 정확하고 다양한 지식을 추출할 수 있는 방법을 연구하고자 한다. 레코드 구조가 아닌 트리 구조 데이터로부터 의미 추출이 가능한 효율적인 마이닝 기법을 진일보시킴으로써 다양한 사용자 중심의 서비스 제공을 최종 목적으로 한다.

키워드: 시맨틱 마이닝(semantic mining), 트리구조데이터(tree-structured data), 지능형 지식서비스(intelligent knowledge service)

I. Introduction

정보 검색이나 유용 지식 추출 방법이나 기술들은 레코드 구조의 데이터 위주로 이루어져왔으며, 이는 기하급수적으로 증가하며 웹 또는 온라인 데이터들을 이루고 있는 트리 구조 데이터인 XML에 적용하기에는 시간과 비용 측면에 있어서 비효율적일 뿐만 아니라 데이터가 제공하는 정보의 손실 또한 초래한다.

지능형 지식서비스 제공을 위한 환경에 부합하기 위한 정보 추출은, 텍스트 및 문장의 구성 요소를 분석하여 문서의 내용을 단순한 단어 집합보다는 풍부한 의미를 내포하는 형식으로 표현함으로써 보다 정교한 지식과 정보의 추출이 수행될 수 있도록 하여야 한다. 그러나 기존 XML 데이터에서 정보를 추출하는 방법들은 한계점을 갖고 있기 때문에, 사용자 만족도를 최대화할 수 있는 순수 트리 구조 데이터의 의미 추출과 이를 기반으로 한 응용 분야들의 발전을 위해서는 범람하고 있는 XML 기반 데이터들에 대한 효율적이고 효과적인 의미(semantics)를 고려한 유용하고 적절한 지식 정보 추출에 대한 새로운 방법론의 필요성이 대두된다.

II. Preliminaries

1. Related works

지식 공유 및 유용 정보 추출과 사용을 위한 프로젝트들은 주로 대형 병원과 대학 캠퍼스들을 중심으로, 환자 또는 수업관련 정보를 DB화한 후, 관련 의사나 학생들이 모바일 디바이스들을 사용하여 개개의 관심에 따른 정보를 찾거나 적절한 행위를 할 수 있도록 솔루션을 제공해주는 시스템을 구축하였다. 솔트룩스사는 지식의 축적을 넘어 지식의 공유를 주요 목표로 하여, 컴퓨터가 <사람처럼> 정보와 지식을 이해하고 처리할 수 있도록 하며 사용자가 원하는 지식을 <정확>하게 <맞춤형>으로 제공받을 수 있도록 할 뿐만 아니라, 기 등록 된 지식들의 상관관계성 분석 등을 통해 새로운 <지식발굴>을 추구한다. 솔트룩스사는 웹 데이터들의 지식 표현, 저장, 관리와 공유에 중점을 두어 연구를 진행하지만 주 데이터는 메타 데이터 위주의 개발에 중점을 두고 있다.

KT는 유무선 통신 뿐만 아니라 데이터 통합, IT 컨설팅 업무에 인공지능 검색 연계 연구에 앞서 시맨틱 웹의 추론 엔진 개발 연구를 선행하였다. 이는 지능형 지식 검색 서비스를 제공함은 물론 다양한

스마트 모바일 단말기에서 사용 가능하도록 연동형 엔진도 지원하기 위함이다.

슬로베니아 류블리아나 대학의 AI 연구실은 미국 베이어 대학과 공동으로 1차 전립선 절제 수술 환자들의 재발 가능성 예측을 위한 시스템 구현을 위해, 센서로부터 얻어진 환자의 상황 데이터들로부터 필요한 정보를 추출하기 위한 마이닝 기술과 추론 엔진을 연계하여 의사의 직접적인 진찰이나 치료 없이 위급한 상황에 대한 환자의 결정을 모바일 단말기를 통해 지원하도록 하였다.

2. Background concept

레코드나 객체형을 갖는 정형화된 기존 데이터 모델에 비해서 XML 기반 데이터는 좀 더 복잡한 구조로 모델링이 이루어지며 가장 일반적인 형태는 그래프 구조이다. 그러나 그래프 구조는 알고리즘적인 복잡도 측면에 비춰서 이론적으로 매우 바람직하지 않은 특성이 존재하는데 바로 어떤 그래프와 다른 그래프의 서브 그래프가 서로 동일한가 동일하지 않은가를 결정하는 효율적인 알고리즘이 부재하다는 것이다 [1]. 그래프로 XML 기반 데이터를 모델링 할 경우 발생하는 또 다른 심각한 문제점은 대다수의 반정형 데이터마이닝 알고리즘에서 적용하는 서브 그래프의 체계적인 열거를 위한 효율적인 알고리즘 역시 부재하다는 것이다. 이를 위해 특별한 그래프인 트리를 사용하여 XML 기반의 모든 데이터들은 모델링되는데, 단일 루트에 사이클이 존재하지 않는 특징이 있는 트리는 알고리즘적인 측면에서 수용할 수 있는 범위의 복잡도를 제공한다 [2, 3, 4]. 그러나, 트리 구조는 직접적으로 해당 데이터들로부터 마이닝을 수행하기는 불가능하기 때문에, 계층적인 트리 특성을 반영하여 원래의 XML 기반 데이터가 지닌 내용을 변형시키지 않으면서 마이닝을 수행할 수 있는 지식표현이 요구된다 [5, 6].

지식표현은 일반적으로 동일 노드나 동일 트리에 대해서는 같은 표현이 생성되어야 하며, 원본 트리 구조보다는 메모리나 디스크의 공간 사용이 적으며 좀 더 컴팩트하게 표현되도록 모델링한다. 그러나 이 때 문장에서의 단어, 단어와 단어의 관계, 문장과 문장의 관계 그리고 문장 구성 성분이 갖고 있는 의미의 명시적 표현이 반드시 필요한데 트리 구조는 기존 지식표현을 보완하거나 전혀 다른 형태의 표현법이 요구된다. [그림 1]은 시맨틱 정보를 고려하지 않고 XML 데이터로부터 지식을 추출할 경우 발생하게 되는 치명적인 오류를

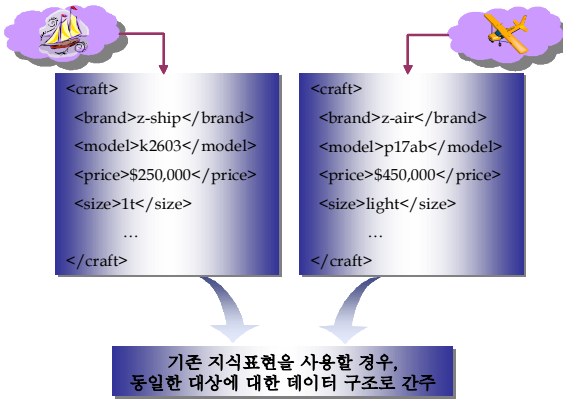

Fig. 1. 시맨틱 처리에 따른 전혀 다른 지식정보 표현

보인다.

데이터 마이닝과 의사결정시스템의 통합 환경을 위한 방법은 해당 시스템의 적용 목적, 데이터 특색, 시스템의 적용 환경 등 여러 인자에 의해 모델링된다. [그림 2]는 대표적인 6 가지 연계 모델링을 보인다. DS는 의사결정 시스템(Decision-making System)을 의미하며 DM은 데이터마이닝(Data Mining)을 나타낸다. 본 연구에서 적용할 연계 모델은 순차 적용 혹은 병렬적용이다.

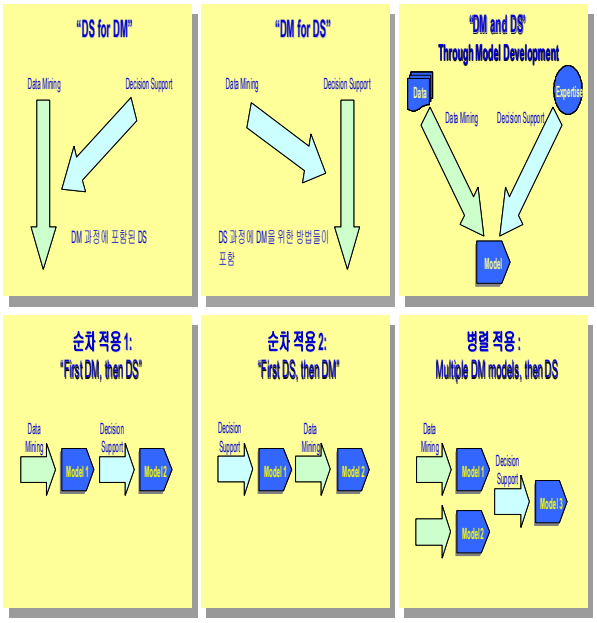

Fig. 2. 대표적인 6 가지 연계 모델링

III. The Proposed Scheme

[그림 3]은 본 연구의 최종 목표가 되는 시스템 구조도로 앞서 출판 된 상황인지기반 미들웨어를 근간으로 한다 [7].

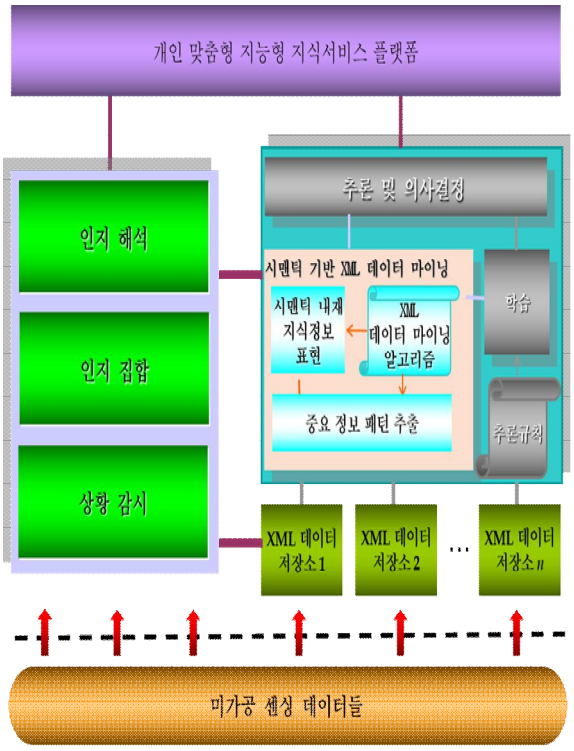

Fig. 3. 지능형 지식정보서비스를 위한 XML 기반 마이닝시스템

주어진 시스템 아키텍처에 근거하여 지능형 지식서비스의 상황인식 정확성의 증대를 위해서는 우선, 데이터베이스 내에 존재하지 않는 데이터에 대해서도 적절한 응답을 하도록 하는 추론 엔진이 필수적이며 해당 추론 엔진과 밀접한 연관을 갖는, 지식을 저장하고 관리하는 지식 데이터베이스에 대한 분석이 요구된다. 이를 위해 다양한 원천으로부터 데이터를 획득하여 사용자의 결정에 필요한 정보 처리를 할 수 있도록 설계되어야 할뿐만 아니라, 사용자의 의사결정이 이루어지는 동안에 발생 가능한 환경의 변화를 반영할 수 있도록 유연하게 설계되어야 한다.

XML 기반 데이터들의 시맨틱 마이닝을 위해서 제안한 시스템은 우선, 레코드 기반 데이터보다 복잡한 구조인 트리 기반 데이터를 대상으로 하는 분석적인 데이터 획득할 수 있어야 하며, 이를 위해 트리 데이터의 계층 간 특징을 반영한 데이터 분석이 필연적이다. 또한, 트리의 각 노드에 위치하는 구체적인 데이터의 의미까지 고려한 마이닝 기법을 통한 향상된 결과 도출이 수행되어야 한다.

제안하는 시스템은 사용자 요구의 의미까지 파악하여 정확하고 다양한 지식을 추출할 수 있는 장점을 가지며, 이를 이용하여 지식서비스 플랫폼의 데이터 분석 능력을 향상시킬 수 있다. 또한 서비스 디스커버리 및 기존 애플리케이션의 활용성을 높일 수 있다.

IV. Conclusions

지능형 지식서비스 환경에서의 효율적이면서 트리구조 데이터의 의미 추출이 가능한 마이닝 기법을 진일보시킴으로써 다양한 사용자 중심의 서비스는 물론 여러 가지 응용 시스템의 기반을 닦을 수 있다. 지식서비스를 위한 데이터 관리와 유지 기술의 발전과 여러 가지 응용 서비스의 활성화는 물론, 정보화 사회에서의 국가 또는 기업의 경쟁력으로 이어질 것이다. 본 연구의 최종 결과물은 지능형 지식서비스 플랫폼에서 다양한 응용 분야의 핵심 데이터 관리 기술 연구로서 추후 연구 결과의 기술이 산업체로 이전될 경우, 각 업체들은 국내외에서 인정받는 기술력의 토대를 마련할 수 있다.

ACKNOWLEDGMENT

이 논문은 2018년도 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1A2B1007015).

REFERENCES

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, 2nd ed., "Introduction to Algorithms", The MIT Press, 2001.

[2] A. Termier, M-C. Rousset, M. Sebag, "TreeFinder : a First step towards XML data mining", Proceedings of IEEE International Conference on Data Mining (ICDM'02), 2002.

[3] S. Abiteboul, P. Buneman, D. Suciu, 1st ed., "Data on the Web", Morgan Kaufmann, 2000.

[4] Y. Chi, S. Nijssen, R. R. Muntz, J. N. Kok, "Frequent subtree mining - An overview", Fundamenta Informaticae 66(1-2), 2005.

[5] Y. Chi, Y. Xia, Y. Yang., R. R. Muntz, "Mining closed and maximal frequent subtrees from databases of labelded rooted trees", IEEE Trans. Knowkedge and Data Engineering 17(3), 2005.

[6] J. Paik, J. Nam, D. Won, and U. M. Kim, "Fast extractoin of maximal frequent subtrees using bits repreerntation", J. of Information Science and Engineering 25(2), 2009.

[7] J. Paik and U. M. Kim, “An XML context data mining

mathod for improved decision making on ubiquitous middleware platforms", Telecommunications review 17(2), 2007.