XML 데이터의 유사내용 검색을 위한 Bootstrap Mining

이한수°, 박종현, 강지훈
충남대학교 컴퓨터과학과
hansu@cnu.ac.kr°, {jhpark, jhkang }@cs.cnu.ac.kr

Bootstrap Mining for Searching Similar Content of XML Data

Han-Su Lee°, Jong-Hyun Park, & Ji-Hoon Kang
Dept. of Computer Science, Chungnam National University

요 약

인터넷 상의 정보교환을 위한 국제표준인 XML은 여러 분야의 응용에 사용되며 응용의 특성에 따라 다양한 형태의 구조로 정의되어 사용된다. 이러한 XML은 응용에 따라 의미적으로 유사한 정보라 하더라도 서로 다른 구조정보를 가질 수 있으며 때로는 스키마(DTD)가 없는 XML문서 형태로 존재하기도 한다. 그 결과 특정 영역(동일 스키마 따르는)의 응용들 사이의 통합은 용이해졌으나 서로 다른 영역 또는 영역에서 소외된 응용과의 통합은 여전히 문제로 남아있다.

본 연구에서는 대부분의 XML문서는 구조정보에 의미를 내포하고 있다는 특성을 고려하여 문서의 구조정보만을 이용하여 서로 다른 영역의 정보들 사이의 유사성을 판단하고 이를 이용하여 의미적으로 유사한 정보를 찾는다. 또한 XML 문서의 특성을 고려하여 보다 정확한 유사정보를 찾기 위하여 처리의 단위를 정의하고 이를 기반으로 프로토타입 시스템을 구현하였다.

1. 서론

XML[1]은 인터넷 상의 정보교환을 위한 표준으로 이 기종 컴퓨터 간의 정보를 교환 할 수 있을 뿐만 아니라 정보의 의미를 표현할 수 있는 구조화된 언어이므로 다양한 분야에서 사용 중에 있으며 또한 사용하고자 시도 중에 있다. 이러한 XML문서는 다양한 분야에서 많은 양의 정보가 존재하며, 그 정보의 의미를 표현하는 구조정보(스키마)[2] 또한 다양하다.

일반적으로 동일한 목적을 갖는 응용들간에는 동일 구조의 XML 문서를 정의하고 정보를 공유하고자 한다. 그러나 동일한 목적을 가진 모든 응용이 동일 구조를 따르고 있지 못하는 것이 현실이며, 때로는 XML문서의 사용이 일반화 되면서 구조정보 없는 형태의 문서들도 존재한다. 이러한 현실은 정보의 검색 측면에서 검색의 범위를 제한하거나 검색의 범위를 확장하기 위해서는 모든 응용에서 사용하는 XML 문서의 구조정보를 참조해야만 가능하다는 문제점을 야기시킨다. 그러므로 구조정보의 참조 없이 문서들 사이의 구조정보를 판단하고 이를 기반으로 내용정보를 추출할 필요성이 대두 되었다.

본 논문에서는 동일 목적을 갖지만 다양한 구조정보를 가지고 있는 XML문서들을 대상으로 최소한의 입력정보를 이용하여 구조정보에 대한 지식이 없이도 문서 자체로부터 구조정보(XPath)[3]를 유추하고, 유추한 구조정보를 이용하여 유사 내용정보를 추출하는 방법을 제안한다. 또한 이를 보다 정확하게 수행하기 위하여 우리는 유용한 검색의 처리의 단위를 제안한다.

우리의 검색 방법은 입력정보에 따라 단일 질의 검색(Single Keyword Search)과 복합 질의 검색(Mixed Keyword Search)을 제공하므로 사용자에게 보다 다양한 검색을 가능하도록 하며, 향후 XML 기반의 정보검색을 위한 연구를 위한 자료로 참조될 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 기술하고 있으며, 3장에서는 본 논문에서 설계하고 구현한 XML기반의 Bootstrap Mining 엔진의 구조와 여기에서 사용하고 있는 방법에 대해서 설명하고 있다. 마지막으로 4장에서는 본 논문의 결론 및 향후 연구에 대해 기술한다.

2. 관련연구

인터넷의 빠른 보급과 함께 많은 정보가 공유되고 있으며 그 정보의 가장 일반적인 형태는 아직까지 HTML다. 이러한 HTML은 다양하고 방대한 정보를 내

* 이 연구는 BK21 충남대학교 정보통신 인력 양성 사업단과 소프트웨어 연구 센터의 지원을 받았음.

포하고 있다. 그러므로 HTML문서에 내포된 방대한 정보를 검색하기 위한 많은 노력들이 진행되고 있으며, 그 노력의 한 분야로 HTML문서에서 구조정보만을 이용하여 유사한 의미를 갖는 내용정보를 추출하는 것이다[4,5]. 즉 HTML의 태그 구조 중 의미 있는 정보를 표현할 때 주로 <List>,<Table>태그 등의 제한된 태그를 사용하여 표현된다는 논거 아래 관심의 대상을 일정한 태그정보에 제한 하는 것이다. 예컨대 의미 있는 태그를 <Table>로 제한하고 테이블의 행 과 열 정보를 이용하여 정보의 유사성을 결정하는 것이다. 그러나 이 방법은 정보추출의 단위를 일정한 태그들로 제한함으로 정보의 누수가 발생 할 수 있고, 또한 HTML의 특성상 고정된 구조정보로 인하여 정보의 유사성을 판단하기 위해서는 모든 테이블의 행과 열의 정보를 모두 참조 하여 불필요한 작업을 수행한다는 단점이 있다.

XML은 HTML과 함께 인터넷상에 정보를 표현하기 위해서 사용되는 언어이며 현재 매우 활발히 사용 중에 있다. 그러므로 많은 연구에서는 앞서 언급한 HTML과 함께 XML문서에서 구조정보 만으로 유사한 문서 혹은 내용을 찾기 위한 연구가 진행 중에 있다. 그 중 대표적인 연구는 [6,7,8]로, 반 구조화된 문서의 구조를 정규화하고 정규화된 구조의 패턴을 추출하여 유사정보를 추출하는 방법이다. 즉 정규화된 구조정보에 일정한 패턴이 반복 될 경우 그 패턴은 유용한 정보를 담고 있다고 가정하고 이 패턴을 효과적으로 검색 할 수 있는 방법들을 제안하고 있다.

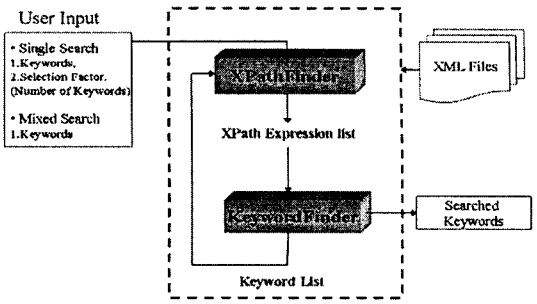
본 논문에서는 이러한 앞선 방법들의 목적과는 달리 XML문서에 대한 내용정보의 검색을 목적으로 한다. 그러므로 XML 문서의 구조적 패턴이 아닌 XML문서가 가지고 있는 구조정보로부터 구조적인 유사성을 판단하고 이를 이용하여 유사한 내용정보를 추출한다. 또한 우리는 앞선 방법들과 달리 유사 내용검색을 보다 정확하게 검색하기 위하여 처리단위를 정의한다.

3. Bootstrap Mining Engine

3.1 Bootstrap Mining Engine의 구조

본 논문에서 제안하는 Bootstrap Mining 엔진은 XML문서에 대한 검색 엔진으로서, [그림 1]과 같이 추출 할 정보에 대한 위치정보(XPath)를 탐색하는 XpathFinder 모듈과 찾아진 위치정보를 기준으로 구조적인 유사정보를 추출하는 Keyword Finder 모듈로 구성 된다.

Mining 엔진은 사용자로부터 검색하고자 하는 키워드들을 입력으로 받아 XML 문서에서 해당 키워드의 구조정보를 검색한다. 이렇게 검색된 구조정보는 다시 검색어가 되어 XML 문서에서 해당 구조정보를 만족하는 키워드를 검색한다. 이 두 과정은 현재 추가적인 정보를 검색하지 못할 경우까지 반복해서 수행하게 된다.

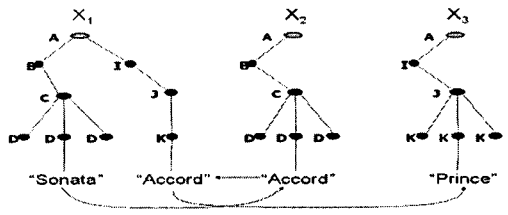


[그림 1] BootStrap Mining 엔진의 구조

우리의 Mining 엔진은 이러한 구조를 중심으로 사용자의 요청에 따라 단순정보구조(word단위)를 질의하는 단순질의검색과 복합정보구조(Subtree 단위)를 질의하는 복합질의검색을 제공한다.

3.2 단순질의검색 (Single Keyword Search)

단순질의검색은 같은 목적을 가진 응용들에 존재하는 다른 구조의 XML문서들로부터 주어진 키워드와 유사한 의미를 갖는 유사 키워드들을 추출하는 검색이다 단순질의검색의 입력은 동일한 의미를 갖는 "키워드들"과 키워드들 중 최소한 몇 개의 키워드를 만족할 경우 동일 의미를 갖는다고 결정할 수 있도록 하는 "최소검색을 위한 수"로 구성된다.



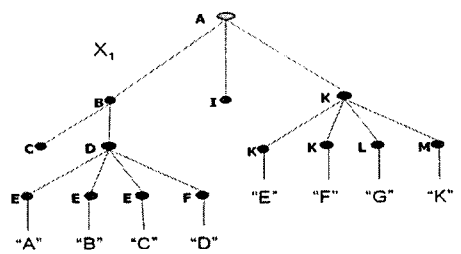
[2] 단순질의검색 방법

[그림 2]와 같이 문서 X_1 , X_2 , X_3 이 서로 동일한 응용의 분야(자동차판매)에서 사용하는 서로 다른 구조의 XML문서들이 존재할 경우, 사용자는 임의의 자동자명(예:"Sonata")을 입력 키워드정보로 입력하고 (이때 최소검색을 위한 수는 1) 입력한 키워드 정보와 유사한 의미를 갖는 정보를 XML문서들로부터 추출하고자 한다.

[그림 2]에서 초기 키워드인 "Sonata"는 XpathFinder모듈의 입력으로 제공되어, 구조정보(/A/B/C/D)를 찾도록 하며 이렇게 찾아진 구조정보는 다시 KeywordFinder 모듈의 입력으로 제공되어 동일문서 혹은 다른 XML문서에서 동일한 위치에 있는 정보들을 유사키워드로 추출한다. 즉 X_2 문서의 /A/B/C/D에 존재하는 "Accord"가 새로 찾아진 유사정보이며, 찾아진 유사정보는 다시 다음 반복을 위한 입력정보에 포함 된다. 그 결과 두 번째 반복 에서는

동일한 구조정보에 존재하는 "Prince" 라는 정보를 찾을 수 있다.

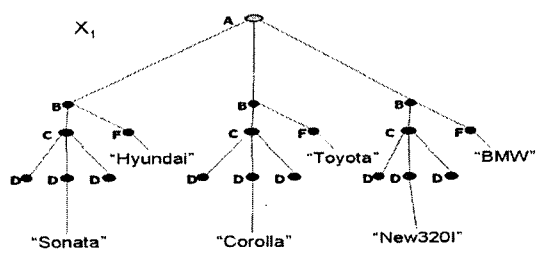
구조정보를 추출할 시, 입력정보인 각 키워드는 동일한 의미를 갖는 서브트리의 내부에 포함되어야 한다. 본 논문에서는 이러한 서브트리를 Bootstrap Mining을 위한 처리단위(Processing Unit)로 정의하고 이를 찾기 위한 방법을 제안한다. 즉, 유사한 키워드들은 처리단위 내부에서 동일한 위치에 있는 정보들이다. 처리단위를 찾기 위한 방법은, 주어진 입력키워드 정보를 기반으로 해당 텍스트 노드를 검색하고, 텍스트 노드의 부모 노드를 찾아 가면서 그 부모 노드가 반복된다면 그 노드의 부모 노드를 처리단위의 최상위 노드로 정의한다. [그림3]은 D와 K는 각각 키워드가 "A","B","C" 또는 "E","F"인 경우 처리단위의 최상위 노드를 나타낸다.



[그림 3] 처리단위(Processing Unit)

3.3 복합질의검색 (Mixed Keyword Search)

복합질의검색은 단순질의검색과 동일한 검색처리 절차를 따른다. 그러나 복합검색의 입력은 "연관된 의미를 갖는 키워드들"과 각 키워드들의 "가중치" 이다.



[그림 4] 복합질의 검색방법

[그림4]는 자동차회사와 자동차회사별 자동차명을 보여주는 XML문서의 한 예로 초기 사용자 키워드들은 "Hyundai"와"Sonata"이다. 예제에서는 단일의 XML 문서를 기반으로 보여주고 있으나 이들의 처리는 다수의 문서라 하더라도 달라질 것은 없다. 복합질의처리에서 역시 단순질의처리에서와 동일하게 처리단위를 가장 먼저 검색한다. 그러나 다른 점은 단순질의처리에서는 텍스트 노드로부터 처리단위를 검색했으나, 복합질의처리의 경우 복합 질의에 해당하는 텍스트 노드들의 조상 노드들이 최초로 만나는 노드를 시작 노드로 처리단위를 결정한다. 그러므로 위 예제의 경우 A 노드가 처리단위의 최상위 노드이다. 이렇게 처리단위가 결정되면 단순검색질의 처리와 동일하게 처리단위 내부에서 동일한 경로를 갖는 키워드들을 검색하여 유사 키워드로 결정한다. 그러나 복합질의처리에서는 추가적으로 키워드들간의 경로 역시 동일해야지만 유사한 의미의 키워드로 검색된다.

4. 결론

본 논문에서는 동일한 의미의 내용정보를 표현하지만 서로 다른 구조를 갖는 XML 문서들 간에 유사정보를 추출하기 위하여, XML 문서에서 유사한 구조정보를 추출하여 이를 기반으로 유사한 의미의 내용정보를 추출하는 방법을 제안하고 있다. 또한 우리는 검색의 처리 단위를 정의하여 보다 정확한 유사정보를 추출하기 위한 방법을 제안하고 있다.

우리의 연구는 향후 XML 데이터의 정보 검색의 측면에서 효율적으로 사용될 수 있을 뿐만 아니라 간단한 알고리즘의 추가만으로 HTML 데이터의 검색을 위해서도 유용하게 사용될 수 있다. 이러한 특징은 현재 웹 상에 혼재되어 있는 XML데이터와 HTML데이터를 동시에 검색할 수 있다는 장점을 제공한다.

5. 참고 문헌

[1] W3C, Extensible Markup Language (XML) Version 1.0, Recommendation, Feb. 1998,

[2] W3C, XML Schema Part 0: Primer, Recommendation, May 2001

[3] W3C, XML Path Language(XPath), Recommendation, November 1999

[4] S. Hirokawa, E. Itoh, & T. Miyahara, "Semi-Automatic Construction of Metadata from a Series of Web Documents", Proc. Australian Conference on Artificial Intelligence 2003, 2003.

[5] Y. Yamada, N. Craswell, T. Nakatoh, & S. Hirokawa ", Testbed for information extraction from deep web" WWW (Alternate Track Paper & Poster) 2004.

[6] M.Noguchi, & S.Hirokawa, "A Prototype of Search Engine for Tables on the Web", Proc. ISEE2003, 2003.

[7] J.-W. Lee, K. Lee, & W. Kim, "Preparations for Semantics-Based XML Mining", Proc. ICDM 2001.

[8] J. Myllymarki, "Effective Web Data Extraction with Standard XML Technologies", Computer Networks, Volume 39, Number 5, pp. 689-696, 2001.

[9] C.-H. Chang, S.-C. Lui, &Y.-C. Wu "Applying Pattern Mining to Web Information Extraction" Proc. PAKDD 2001, 2001.