XML 기반의 통합 임상정보를 효율적으로 저장하기 위한 XML 압축 기법에 대한 연구

유의혁, 정종일, 이태헌, 신동규, 신동일
세종대학교 컴퓨터공학과
e-mail : {solui, jijeong, leeth, shindk, dshin}@gce.sejong.ac.kr

A Study on XML Compress method for efficient integration and storing of XML-based Clinical Information

Weehyuk Yu, Jongil Jeong, Taeheon Lee, Dongkyoo Shin, Dongil Shin
Dept. of Computer Engineering, Sejong University

요 약

임상정보 문서는 환자 진료기록뿐만 아니라 처방전, 개인적 유전자정보를 가지고 있다. 이러한 임상정보 문서는 병원 시스템들간에 교환 및 공유함으로써 양질의 의료서비스를 제공할 수 있다. 이와 관련하여 임상정보의 통합을 위한 기존의 연구들은 각각 HL7 메시지를 XML 문서로 변환하고 XML 기반의 CDA 를 관계형 데이터베이스에 저장하는 연구가 진행되었다. 그러나 관계형 데이터베이스는 문서의 데이터 별 테이블 단위로 생성, 저장된다. 그러나 HL7 과 CDA 는 문서 중심의 XML 문서이기 때문에 관계형 데이터베이스에 저장 시 문서 별 많은 변이가 존재하여 테이블 증가를 갖는다. 따라서 비정규적인 구조에 적합한 데이터베이스를 선택하기 위해 XML 전용 데이터베이스와 관계형 데이터베이스 비교하고 효율적 저장을 위해 압축기법을 제시한다. 압축기법을 적용한 임상정보 데이터베이스는 대용량 임상정보 문서의 크기를 압축함으로써 문서의 크기를 줄임으로써 데이터베이스의 효율적 저장을 향상시킨다.[1]

1. 서론

최근 환자에게 양질의 의료서비스를 제공하기 위해 각 병원 시스템에서 개별적으로 관리되고 있는 의료 및 진료 데이터를 공유하고 통합하려는 노력이 진행되고 있다. 이질적인 병원 시스템들간에 데이터를 교환하고 공유하기 위해서는 특정 시스템에 독립적으로 운용될 수 있는 연구가 필요하다. 이와 관련하여 임상정보의 통합을 위한 기존의 연구들 [1][2] 에서는 각각 HL7 (Health Level Seven) 메시지를 XML 문서로 변환하고 XML 기반의 CDA (Clinical Document Architecture)를 관계형 데이터베이스에 저장하는 연구가 진행되었다. 그러나 HL7 과 CDA 는 사람이 읽을 목적으로 작성되는 문서 중심의 XML 문서 (Document-centric Document) 이기 때문에 정규적인 구조를 갖는 관계형 데이터베이스에 저장하는 것은 너무 많은 변이가 존재해서 테이블 수가 많아지고 널(null)값을 갖는 컬럼이 너무 많아져서 비효율적일 수 있다.

XML 전용 데이터베이스는 비정규적인 구조를 갖는 문서를 저장하는데 적합하다. 그러나 비정규적인 구조는 부가적인 데이터를 많이 포함할 수 있기 때문에 XML 의 기본 저장단위인 문서의 크기가 커질 수 있다. 따라서 문서 중심의 XML 문서 같은 비정규적인 구조를 갖는 문서를 효율적으로 저장하기 위한 연구가 필요하다.

본 논문에서는 관계형 데이터베이스와 XML 전용 데이터베이스의 저장공간 활용도를 비교하고 효율적

[1] 본 연구는 보건복지부 보건의료기술진흥사업의 지원에 의하여 이루어진 것임. (0412-MI01-0416-0002)

저장을 위해 압축기법을 소개한다 [3].

2. 관련연구

2.1 관계형 데이터베이스와 XML 전용데이터베이스의 저장공간 활용도 비교

관계형 데이터베이스와 XML 전용 데이터베이스의 비정규화된 문서의 저장 시 공간활용도는 각 데이터베이스에 데이터를 저장하는 단위를 통해 알 수 있다. 관계형 데이터베이스는 문서의 데이터 단위로 테이블을 생성 후 저장하는 반면에 XML 전용 데이터베이스는 XML 문서 자체를 저장한다.

비정형화된 문서를 관계형 데이터베이스에 저장하는 경우 문서에 포함된 각 구성요소를 저장하기 위한 테이블이 생성되어야 하기 때문에 그 결과 다수의 테이블 생성으로 인해 저장공간의 활용도가 낮아지게 된다. 반면 XML 전용 데이터베이스는 XML 문서 자체가 저장단위이기 때문에 비정규화된 문서에 포함된 각 구성요소를 위한 별도의 저장공간이 필요하지 않다. 따라서 비정규화된 문서를 저장하는 경우 XML 전용 데이터베이스가 관계형 데이터베이스에 비해 저장공간의 활용도가 높다.

2.2 데이터베이스 압축기법

데이터베이스의 저장되는 데이터 양이 증가함으로써 효율적 저장에 대한 연구가 진행되었다 [4]. 그 중 압축기법은 저장 전 데이터를 압축하여 데이터 크기를 줄어들게 하는 기법이다 [4]. 압축기법은 테이블에 저장되는 데이터 단위로 압축하는 기법과 전체 데이터 단위로 압축하는 기법으로 나눌 수 있다. 압축기법은 표 1 와 같다.

<표 1> 표준 압축 기법 [4]

압축기법	설명
Differential encoding	인접한 값들의 차이를 저장
Offset encoding	기준 값으로부터 차이를 저장
Dictionary encoding	문자열을 하나의 숫자로 대응
Lempel-Zip 77 or 78	적응적인 사전적 인코딩
Huffman encoding	자주 나오는 문자들을 더 짧은 비트로 표현

Differential Encoding 과 Offset Encoding 은 데이터 값이 숫자일 경우 사용하고 Dictionary encoding, Lempel-Zip 그리고 Huffman encoding 은 데이터 값이 문자열일 경우 압축기법으로 사용한다.

2.3 XML 파일을 파일 시스템에 저장하는 방식

XML 문서를 데이터베이스에 저장하는 것 외에 파일 시스템에 저장하는 방식이 있다. 파일 시스템에 저장하는 방식은 텍스트 파일 방식, 역 리스트 그리고 XML 압축 파일이 있다 [3].

- 텍스트 파일 형식은 XML 문서를 텍스트 형식으로 저장하여 관리된다. 텍스트 파일은 사용의 용이성과 독립적인 처리가 가능하다는 장점을 가지고 있다. 질의 처리시 직접적인 처리를 제공하지 않고 파서를 이용하는 단점이 있다.
- 역 리스트 형식은 텍스트 파일 형식의 질의 처리 단점을 해결한다. XML 문서의 형식을 역 리스트(inverted list)를 이용하여 질의 처리를 제공한다. 역 리스트 방식은 XML 요소 별 관리가 가능하다. 그러나 XML 문서를 수정 시 역 리스트도 수정해야 하는 단점이 있다.
- 압축기법은 파일 시스템의 공간 활용도를 높이기 위해 사용된다. XML 규칙성을 이용한 압축을 통해 파일 처리의 효율성을 높인다. 압축된 파일은 텍스트 기반의 파일보다 문서 파싱의 횟수를 줄일 수 있지만 역 리스트 기법과 동일하게 XML 문서를 수정 시 수정해야 하는 단점이 있다.

2.4 XML 압축기법

XML 문서는 각 요소의 시작 태그로 시작하여 해당 태그에 해당되는 값을 포함시킨 후 태그를 닫는 규칙성을 가지고 있다. 이러한 규칙성과 XML 문서의 원소 별 빈 공간은 문서의 크기를 증가시킨다. 이러한 규칙성을 이용하여 문서의 크기를 감소시키는 기법이 XML 압축기법이다. XML 압축기법으로는 XMill [5], ICT XPress [6], XGrind [7] 이 있다. 각 기법 별 제공하는 기능은 표 2 와 같다.

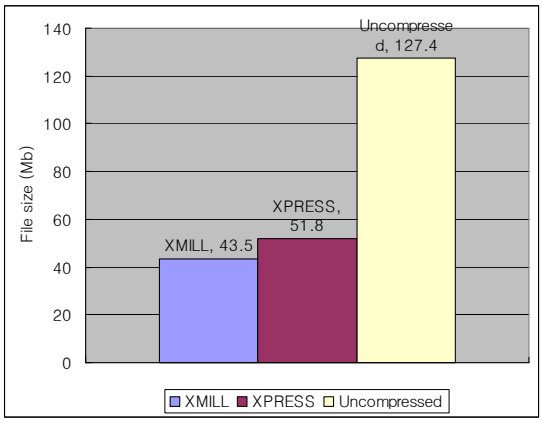
<표 2> 압축기법 별 제공하는 기능 [8]

	XMill	XGrind	ICT XPress
압축된 후 질의		○	○
사용 편리성	○		○
data & structure 분리	○	○	○
DTD 이용		○	

- XMill 은 XML 최대 압축률을 제공한다 [8]. 태그는 Dictionary 압축을 통해 압축이 되고 데이터 값은 사용자가 선택한 압축으로 된다. 사용자에게 사용 편리성을 제공하지만 압축된 데이터의 질의처리가 불가능하다.
- XGrind 는 미리 정해진 압축기법인 Dictionary 와 Huffman 를 이용하여 XML 데이터를 압축한다. 다른 기법과 다르게 XGrind 는 DTD 을 이용한 압축을 사용한다. 그러나 다른 기법들과 다르게 사용자에게 사용 편리성을 제공하지 못하지만 압축된 데이터의 질의처리를 제공한다.
- ICT XPress 는 XML 데이터의 값들의 타입을 추출하여 이에 맞는 기법으로 압축한다. 사용자에게 사용 편리성을 제공하고 압축된 데이터의 질의처리를 제공 한다.

3. 본론

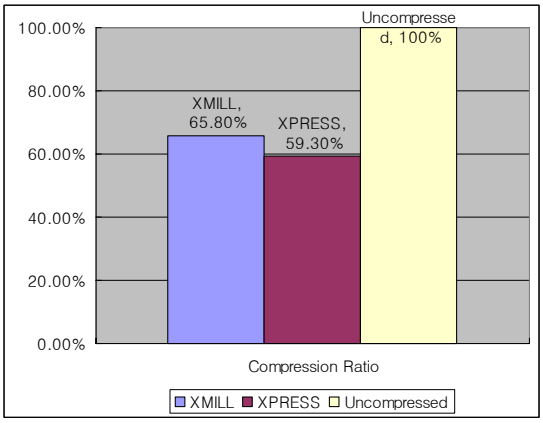
데이터베이스의 저장되는 임상정보 문서를 압축하면 효율적 저장을 제공할 수 있다. 그림 1 은 임상정보 문서 파일을 압축 전후를 용량을 비교한 그림이다.



(그림 1) 압축 후 용량 비교

문서 크기가 127.4MB 인 문서를 XMill 과 XPress 를 이용하여 압축을 하였다. XMill 을 이용하여 압축 시 파일의 크기는 43.5MB 가 되었고 XPress 를 이용한 압축 시에서는 파일의 크기가 51.8MB 되었다. 위와 같이 압축된 파일의 크기는 XPress 보다 XMill 을 이용할 경우 문서의 크기가 더 줄어든 것을 알 수 있다.

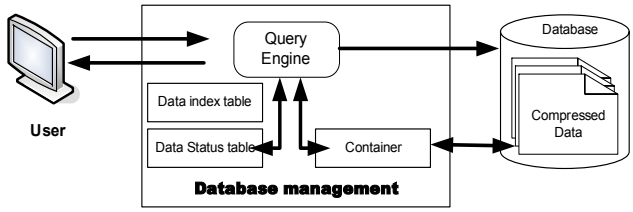
그림 2 은 각 압축 기법의 문서 압축률을 나타낸다. 문서압축률은 [1 – {압축된 문서의 크기/압축 전 문서의 크기)]으로 계산한다.



(그림 2) 압축률 비교

문서 압축률은 XMill 이 XPress 보다 효율적이다. 그러나 XMill 은 압축된 데이터의 질의처리가 불가능한 단점을 가지고 있다. 이런 단점은 데이터베이스의 질의처리 시 문제로 발생된다. 데이터베이스 내 저장된 데이터의 질의처리가 불가능할 경우, 데이터베이스는 압축된 데이터의 정보파악이 불가능하다. 압축된 데이터에 대한 정보를 인덱스 테이블을 이용하여 관리를 해도 테이블 정보로 인한 데이터베이스 용량증가 및 정보파악의 한계점이 있다. 따라서 데이터베이스에 저장될 압축기법은 압축 된 데이터의 질의처리가 가능해야 한다.

임상정보 데이터베이스는 양질의 의료서비스를 제공하기 위해서는 사용자에게 문서 전체뿐만 아니라 문서의 특정 요소 질의처리, 문서의 수정기능도 제공되어야 한다. 그러나 압축기법은 문서 수정 시에는 압축해제 과정이 필요하다. 이와 같은 문제점을 해결하기 위해서 그림 3 과 같은 데이터베이스 관리 시스템이 필요하다.



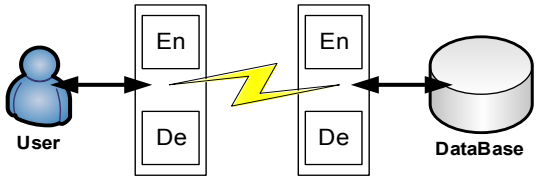
(그림 3) 데이터베이스 관리 시스템

데이터베이스 관리 시스템은 사용자로부터 요청 받은 문서나 문서의 특정 요소 데이터를 관리하는 역할을 한다. 임상정보 문서 자체를 요청할 경우 데이터베이스 관리 시스템 내 문서의 정보가 들어있는 데이터 인덱스 테이블을 참조하여 사용자에게 전송하고 문서의 특정 요소 데이터를 요청 시에는 데이터베이스 관리 시스템 내 질의 엔진을 이용한다. 질의 엔진은 데이터베이스에 XQuery [9] 기반으로 사용자에게 질의처리를 제공한다.

압축된 문서의 수정 시 데이터베이스 내 컨테이너와 데이터 상태 테이블을 이용하여 처리한다. 사용자로부터 원하는 문서의 수정을 요청 받으면 해당 문서를 데이터베이스 관리 시스템 내 컨테이너로 복사 후 압축 해제를 한다. 그리고 데이터 상태 테이블은 해당 문서의 버전 별 관리를 담당하고 문서의 수정된 부분을 표시한다. 컨테이너에서 수정 작업이 완료되면 문서를 재 압축을 하여 데이터베이스 내 원본문서 위치에 복사한다. 복사가 완료되면 데이터 상태 테이블에 문서의 버전과 수정된 부분을 표시한다.

압축된 문서의 데이터베이스 구현효과는 데이터베이스의 효율적 저장을 향상시킬 뿐만 아니라 데이터의 전송 시에도 대역폭이 감소효과를 가져온다.

대용량 문서의 전송 시 문서의 크기가 작아질수록 대역폭이 감소된다. 대역폭의 감소는 사용자에게 고속처리를 제공하고 데이터베이스 서버 측에게 시간적, 자원적 이득을 제공한다. 데이터베이스 측에서 전송된 압축문서는 사용자 측에서 압축해제 과정 후 사용자가 읽을 수 있는 문서형태를 갖는다. 반대로 사용자 측에서 데이터베이스로 전송 시, 사용자로부터 전송된 문서는 데이터베이스 관리 시스템에서 데이터 인덱스 테이블에 등록 후 데이터베이스에 저장된다.



(그림 4) 문서의 전송 시 압축 및 해제

그림 4 와 같은 처리를 하기 위해서는 사용자와 데이터베이스 간에 동일한 압축 기법을 사용하고 양쪽 모두 압축과 압축해제 기능이 제공되어야 한다.

4. 결론 및 향후 연구

최근에, 양질의 의료서비스를 위해 이질적인 병원 시스템간에 문서들을 중앙 데이터베이스 통합저장에

관한 연구가 진행되었다. 그러나 임상정보 문서인 HL7 과 CDA 는 문서 중심의 XML 문서이기 때문에 관계형 데이터베이스에 저장 시 비효율성이 제시되었다. 또한 임상정보 문서는 비정규적인 구조를 가지고 있어서 부가적인 데이터를 갖는다. 이러한 부가적인 데이터는 문서의 크기를 증가시킴으로써 데이터베이스 크기의 증가를 가져온다. 그러므로 데이터통합뿐만 아니라 데이터베이스의 효율적 저장기법이 필요하다. 본 논문에서는 XML 전용 데이터베이스를 사용하여 데이터베이스의 공간 활용도를 높이고 효율적 저장을 위해 XML 압축기법을 제시하였다.

압축기법을 적용한 XML 전용 데이터베이스는 관계형 데이터베이스보다 공간 효율성이 뛰어나고 효율적 저장을 제공할 수 있다.

또한 압축된 문서는 사용자에게 문서를 전달 시 대역폭의 감소를 가져와 사용자에게 고속처리와 데이터베이스 서버 측에 시간적 자원적 이득을 제공한다. 그러나 압축된 데이터를 저장하는 데이터베이스 저장방식은 질의 처리의 향상 및 처리 시간 단축은 가져오지 못한다. 향후 데이터베이스의 질의 처리 향상에 관한 연구가 필요하다.

참고문헌

[1] 이민경, 정재헌, 전종훈, 유수영, 김보영, 최진욱, “The LEX System : HL7 을 사용하는 전자의무기록의 효율적인 교환과 공유를 위한 XML 기반 통합의료 환경의 구축”, 정보처리학회,2002

[2] 박수진, 김일곤, 조훈, 곽연식, “CDA 문서를 관계형 데이터 베이스에 저장 관리하기 위한 시스템”, 한국정보과학회 가을 Vol.30. No.2,2003

[3] 민준기, 박명제, 안재용, 정진완, “다양한 저장소에서의 효율적인 XML 저장기법에 대한 연구”,정보과학회논문지:데이터베이스, 제 19 권, 제 1 호, page 1-14,2003

[4] 조형주, 정진완, “다차원 색인 구조를 위한 효율적인 압축 방법”, 정보과학회논문지:데이터베이스, 제 30 권, 제 5 호, page 429-437,2003

[5]XMill,http://www.oledo.com/harmut/XMill/XMill.html

[6] Intelligent Compression Technologies. XML-Xpress. http://www.ictcompress.com.

[7] P.Tolani and J.R. Haritsa, “XGrind: A Query-friendly XML Compressor”, In Proceedings of 18th IEEE International Conference on Data Engineering, page 225-234, 2004.

[8] Smitha S.Nair, “XML Compression Techniques:A survey, University of Iowa”

[9] XQuery,http://www.w3.org/xquery/