XML 기반 의료 문서의 압축전송

유의혁[O] 정종일 신동규 신동일
세종대학교
{solui[O], jijeong, shindk, dshin}@gce.sejong.ac.kr

Compress transmission of XML-based Clinical Document

Weehyuk Yu[O], Jongil Jeong, Dongkyoo Shin, Dongil Shin
Dept. of Computer Engineering, Sejong University

요 약

XML 기반의 CDA는 의료정보 데이터로써 환자의 개인정보, 과거 의료정보, 가족기록 및 검사기록 등 의료정보를 저장한다. 의료정보 데이터는 병원 시스템간에 교환 및 공유함으로써 양질의 의료서비스를 제공되고 데이터베이스에 저장되어 관리된다. 그러나 다양한 의료정보는 의료정보 문서 자체의 크기를 증가시키기 때문에 데이터베이스에 저장 시 공간증가와 저장시간 그리고 데이터의 전송 시 송,수신 시간을 증가한다. 따라서 의료정보 문서의 크기를 감소시켜 문서 처리시간을 단축시킴으로써 처리 효율성을 증가시킨다.[1]

1. 서 론

CDA(Clinical Document Architecture)는 XML 기반의 의료정보 데이터이다. CDA는 의료정보 데이터로써 환자의 개인정보, 과거 의료정보, 가족기록 및 검사기록 등 전체적인 정보를 갖는다. CDA의 이러한 특성은 XML 기반으로 구성되어 있기 때문에 환자의 데이터에서 원하는 데이터를 쉽게 찾을 수 있고 다른 환자 그룹에서 비슷한 환경 및 나이 대의 환자들의 치료에 참고 데이터가 된다[1][2]. 의료정보 통합 시스템은 그림 1과 같이 데이터베이스를 구축함으로써 각 의료기간 및 병원들에게 의료정보를 제공한다.

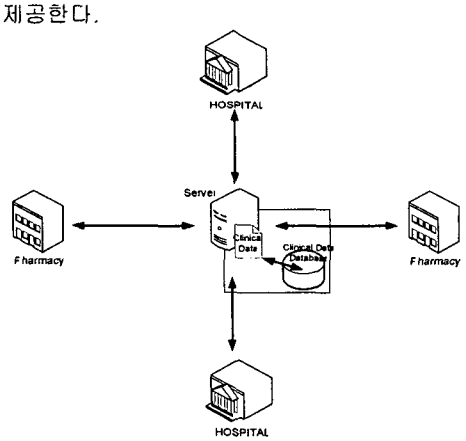

그림 1 CDA Database Middleware

CDA 문서의 다양한 컨텐츠는 환자의 다양한 의료정보를 표현하지만 문서 크기의 증가를 가져온다. CDA 문서의 크기 증가는 데이터의 전송 시 송,수신 증가와 데이터 베이스의 저장공간 증가를 갖기 때문에 이러한 문제점을 해결하기 위해서 기존의 데이터베이스에서는 압축기법을 이용하였지만 XML기반의 CDA 문서에서의 적용에는 어려움이 있다. 기존의 압축기법은 압축된 문서의 내용파악을 위해서는 디코딩 과정이 필요하고 문서의 부분적 열람이 불가능하다. 본 논문에서는 XML 저장기법 중 하나인 바이너리 압축 기법(Binary Compress)을 이용하여 CDA 문서의 적합한 압축기법을 적용 및 적합한 압축 기법을 알아본다.

2. 관련연구

기존 데이터베이스의 압축 기법과 비슷한 XML 저장기법 중 바이너리 압축 기법(Binary Compress)은 통신의 경우 데이터의 대역폭을 감소시키기 위해 사용된다. 이 기법은 XML 기반의 CDA 문서의 저장 및 압축에도 응용할 수 있다.

2.1 CDA

CDA는 HL7에서 문서 교환을 목적으로 정의하였고 텍스트, 이미지, 사운드 및 컨텐츠를 포함한다[1][3]. CDA는 HL7 Version 3 데이터 타입을 사용하고 계층적 구조를 갖고 문서의 구조는 크게 Header 와 Body 부분으로 나눠진다. 문서 버전이 올라감으로써 CDA는 환자의 정보를 구체적인 내용을 포함시킬 수 있다. 현재 CDA Release 2

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

을 기반의 CDA 문서의 저장을 기반으로 하는 연구가 진행 중이다[3]. CDA Release 2의 구조는 그림 2와 같다.

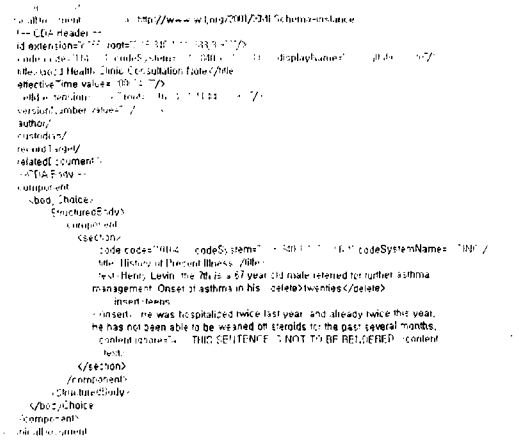

그림 2 CDA Release 2

CDA Body 의 내용은 환자의 세부적인 데이터를 가지고 있지만 그에 따른 문서의 크기는 증가한다. CDA Release 2 의 CDA Body는 Component Elemenet 단위로 구별 관리된다. StructureBody Element의 자식 노드인 Component Element는 Section Element 노드를 갖는다. Section Element는 환자 정보 및 검사 결과, 과거 의료 정보 등을 가지고 있지만 반복적인 노드 배열로 표현된다. 데이터의 반복성은 환자의 정보를 세부적으로 제공하지만 문서 자체의 크기를 증가시켜 문서의 처리 활용도를 감소시킨다. 그러므로 문서의 크기를 감소시켜 문서의 활용도를 증가시켜야 한다.

2.2 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 표1와 같다.

표 1 기본 압축기법

압축기법	설명
Differential encoding	인접한 값들의 차이를 인수 표현 저장
Offset encoding	기준 값으로부터 차이를 인수로 표현 저장
Dictionary encoding	문자열을 하나의 문자로 표현 저장
Lempel-Zip 77 or 78	사전적 매칭을 이용한 인코딩
Huffman encoding	자주 나오는 문자들을 짧은 비트 단위로 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. 그러나 기존의 데이터베이스의 압축기법은 XML 기반의 문서에 적용이 불가능하다. XML 기반의 압축기법이 필요하다.

2.3 XML 압축기법

XML 압축기법은 XML 저장기법 중의 하나이다. XML 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. 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 다음과 같다.

- 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

XML 압축기법의 적용은 CDA 문서 자체를 줄일 수 있지만 기존의 XML 압축 기법들은 데이터 송,수신의 시간 단축을 목적이기 때문에 압축 후 CDA문서의 형태를 알아보기 어렵다. Document 중심적 압축은 높은 압축효율성을 보여주지만 CDA 문서를 알아보기 위해서는 전체적인 디코딩 과정이 필요하다. CDA 문서는 정보의 교환을 목적으로 하지만 환자 기본적인 데이터의 열람이 가능해야 한다. CDA 문서적 특징을 살리기 위해서는 Element 중심적 압축이 필요하다. Element 중심적 압축은 문서 자체의 크기도 줄이고 중요한 데이터부분은 비압축을 하여 문서의 전체적인 내용 파악이 가능하다. 표 2와 같이 Element 중심적 압축은 Document 중심적 압축보다 압축효율성은 떨어지지만 원본 CDA 문서의 크기와는 큰 차이가 있는 것은 볼 수 있다.

표 2 압축 단위별 크기비교

	원본문서	전체압축	부분압축
문서크기	42MB	3.9MB	12MB

부분 압축된 CDA 문서는 데이터 송,수신 시 원본 문서보다 낮은 대역폭을 형성하기 때문에 전송 시간 및 데이터베이스 공간 감소의 효과를 갖는다. CDA 데이터베이스 미들웨어는 그림 3과 같이 처리된다.

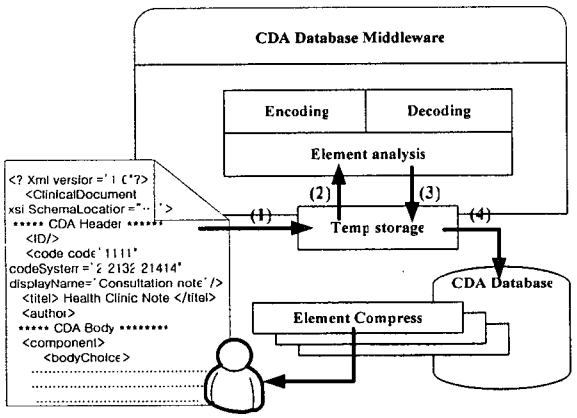

그림 3 CDA 데이터베이스 미들웨어

본 논문에서는 바이너리 압축 기법 중 하나인 XGrind 을 이용한 실험을 하였다. CDA의 Component Element의 자식 노드인 Section Element 단위로 압축을 하여 데이터베이스에 저장하였다. 데이터베이스에 저장 시간은 전체압축 시 저장 시간보다 시간이 더 소비되지만 전체 문서를 저장 시보다 시간의 감소를 가져온다. 데이터 베이스에 저장 시 각 Section Element 단위 별 데이터베이스 테이블에 저장 한 후 사용자가 특정 Element 에 대한 요청 시 임시 저장소에 각 Element에 대한 정보를 XPath를 이용하여 질의 후 제공한다.

전체문서 단위로 압축 시에는 사용자에게 특정 단위를 제공하기 위해서는 전체적인 디코딩 과정과 Element 단위별 압축 과정이 필수적으로 필요하지만 부분적 압축 기법을 이용 시에는 불필요한 디코딩 과정과 압축 과정이 생략으로 인해 전체적인 처리 시간의 증가를 가져온다.

문서 전체압축 및 부분압축에 따른 문서의 처리 시간은 차이가 나지만 사용자가 원하는 데이터의 전송용량은 같기 때문에 전송 시에 걸리는 시간은 동일하다. 문서를 압축하여 전송하기 때문에 원본문서를 전송 시보다 시간이 감소 및 대역폭의 감소를 가져온다.

4. 결론 및 향후과제

XML 압축 기법을 XML 기반의 CDA에 적용하여 저장공간 감소를 가져왔고, 대역폭 감소를 가져왔다 그러나 부분적 압축기법은 의료정보 문서의 Elemenet 단위 설정에 따라 압축효율이 바뀌기 때문에 Element Analysis에서의 정형화된 의료정보 문서의 처리로 비정형화 의료정보 문서인 경우에는 압축효율성이 정형화된 문서 압축효율성보다 낮다. 향후 Element Analysis에서의 정형화된 문서뿐만 아니라 비정형화된 문서에서의 처리를 하기 위해서는 Element Analysis에서의 인공지능 처리가 필요하다.

참고문헌

[1] P Marcheschi, A Mazzarisi, S Dalmiani, A Benassi, " HL7 Clinical Document Archecture to Share Cardiological Images and Structured Data in Next Generation Infrastructure" , Computers in Cardiology IEEE, page 617-620, 2004

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

[3] Kai U. Heitmann, Ralf Schweiger, Joachim Dudeck, " Discharge and referral data exchange using global standards-the SCIPHOX project in Germany" ,International Journal of Medical Informatics, page 195-203, 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"