형태소 분석의 가장 큰 문제는 형태소 분석의 결과가 너무 많다는 것이며, 이를 형태소 과잉분석이라고 한다. 예를 들면, 어절 "하나가"에 대한 형태소 분석 결과 수는 132개를 얻었다. 형태소 과잉분석의 원인은 여러 가지가 있을 수 있으나, 가장 중요한 이유 중 하나는 형태소 배열규칙(morphotactics)의 제약조건이 부족하기 때문이다. 형태소 배열규칙의 제약조건을 강화하기 위해서 본 논문에서는 어절범주 정보를 사용한다. 즉, 형태소를 분석하기 전에, 한국어 어절에 대한 어절범주를 결정하여, 형태소 분석기의 탐색공간을 줄이고자 한다. 본 논문에서 어절에 대한 어절범주를 결정하는 것을 어절분류(eojeol classification)라고 하며, 본 논문에서는 사례기반 방법을 이용해서 어절을 분류한다. 사례기반 방법은 기존의 사례들을 학습하고, 학습된 사례들을 이용하여 새로운 사례를 분류하는 방법이다. 이러한 사례들을 벡터로 구성되어야 하며, 이를 자질벡터(feature vector)라고 한다. 본 논문에서는 오토마타 및 사전을 이용하여 자질벡터를 구성한다.

본 시스템의 성능을 평가하기 위하여 두 종류의 말뭉치를 사용하였다. 평가 방법으로서 정확도를 측정하였다. 또한 실험에 사용된 두 종류의 말뭉치는 학습에 필요한 사례를 충분히 만들지 못하여 교차 검증 방법을 사용하였다. 본 시스템은 22개의 자질을 사용하였을 경우, 각각 평균 97%와 평균 96.5%를 보였으며, 두 종류의 말뭉치를 합쳤을 경우, 평균 95.9%의 성능으로서 1%정도의 성능 차이를 보였다. 이는 두 종류의 말뭉치의 장르가 다르기 때문이라고 생각된다. 또한 최적 자질을 선정하기 위한 실험에서 16개의 자질을 선택하여 시스템의 성능을 평가했을 경우, 평균 0.2%정도의 성능 향상을 보였다. 또한 본 시스템을 형태소 분석기에 적용해 보았을 경우, 어절범주를 사용하지 않은 분석결과보다 평균 35% 정도의 축소율을 보였다.

30. 객체관계형 데이터베이스에 기반한 XML 문서 저장 및 검색 시스템의 설계 및 구현

컴퓨터공학과 곽 용 원
지도교수 박 휴 찬

인터넷의 급속한 발전에 가장 큰 원동력이었던 웹에서 컨텐츠 표현 및 정보전달의 중요한 수단으로 HTML(HyperText Markup Language)이 사용되었다. HTML은 특별한 데이터 타입이나 제한 없이 사용자가 쉽고 간단하게 사용할 수 있다는 장점을 가지고 있다. 하지만 너무나 간단하고 기존에 정의되어진 태그 이외에는 사용할 수 없고 문서의 표현 양식을 기반으로 만들어졌기 때문에 문서의 구조를 표현하는 데도 어려움을 가지고 있다.

XML이 소개되기 전까지, 표현 능력이 뛰어나고 문서 구조를 기반으로 방대하고 복잡한 기능을 제공하는 SGML(Standard Generalized Markup Language)이 사용되었지만 문법이 너무 복잡하고 사용하기에 어려운 단점을 가지고 있었다. 그리하여 W3C(World Wide Web Consortium)은 HTML의 단점을 보완하고 SGML의 복잡성을 제거한 XML(eXtensible Markup Language)을 웹 문서의 표준으로 지정하였다.

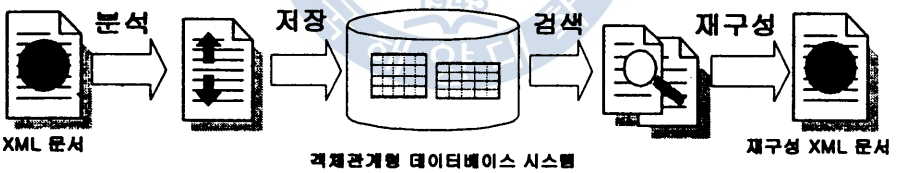
XML은 사용자 스스로 확장이 가능하며, 이기종 간의 정보 교환이 용이하며, 구조적 문서를

표현하는데 유용하다. 이런 장점으로 최근 전자 도서관, 전자 상거래, EC/EDI(Electronic Commercial/ Electronic Data Interchange)를 포함한 다양한 분야에서 활용되고 있다. 이에 따라 XML 문서를 보다 효과적으로 관리하기 위한 연구들이 진행중이다. 특히 데이터베이스를 이용한 저장 방법과 검색 방법에 관한 연구가 활발히 진행되고 있다.

전통적인 관계형 데이터베이스에 XML 문서를 저장할 경우 검증된 우수한 성능을 쉽게 사용할 수 있다. 하지만 구조적 관계를 효과적으로 표현하지 못하고, 집합 값(set-value)을 지원하지 못하는 제약 등을 가지고 있다. 객체지향형 데이터베이스는 객체의 특성을 활용하여 객체들간의 관계를 효과적으로 저장할 수 있지만 문서의 임의 위치 접근과 검색 시 비효율적인 면을 가지고 있다.

이런 점을 고려하여 본 논문에서는 객체관계형 데이터베이스에 기반한 XML 문서 저장 및 검색 시스템을 설계 및 구현하였다. 본 논문에서는 XML 문서 정보 및 구조 정보를 DOM(Document Object Model) API(Application Programming Interface)을 사용하여 추출하였다. 저장 스키마는 DTD(Document Type Definition)의 구조 정보에 영향을 받지 않도록 DTD 독립적이며, 문서 내용을 엘리먼트(element)별로 나누어 저장하는 분할 저장 방식으로 설계하였다. 이렇게 XML 문서의 각 특성을 고려하여 저장 스키마는 문서 테이블, 구조 테이블, 엘리먼트 테이블, 속성(attribute) 테이블, 내용 테이블로 구성하였다. 데이터베이스에 저장된 값들에 대한 다양한 방식의 검색을 위하여 구조 기반 검색, 내용 기반 검색, 혼합 검색과 속성 검색이 가능하도록 하였고, 검색된 결과 값을 재구성하여 XML 문서로 사용자에게 보여지도록 하였다.

그림은 본 논문에서 XML 문서를 분석하여 데이터베이스에 저장하고 검색되어지는 과정을 나타내고 있다.



본 논문에서는 사용자들이 쉽게 XML 문서를 저장 및 검색할 수 있도록 사용자 중심의 인터페이스를 제공하고 있다. 특히 관계형 데이터베이스와 객체지향형 데이터베이스의 장점을 취하여 장시 집합 값을 저장할 수 있으며 관계형 데이터베이스에서 표현하기 힘든 XML 문서의 구조 정보를 효과적으로 표현할 수 있다. 뿐만 아니라 효과적인 저장 스키마 설계로 인하여 문서 특성을 유지하면서 정보들을 저장할 수 있으며, 다양한 검색 질의 및 계층적 질의를 빠르게 처리할 수 있다.