SGML/XML 정보검색 시스템의 구성과 구현 방법론 사례연구 : STEER-SGML/XML

박영찬, 김문석, 김남일, 주종철
전자통신연구원, 컴퓨터·소프트웨어 연구소 자연어처리연구부
대전시 유성구 가정동 161, 우 :305-350
{parkyc,mskim,nikim,zczhoo}@etri.re.kr

Constructing and Implementing SGML/XMLInformation Retrieval Systems with a Case Study : STEER-SGML/XML

Young C. Park, Mun-Seok Kim, Namil Kim, Zong-Cheol Zhoo
Dept. Natural Language Processing , ETRI-Computer·Software Laboratory

요 약

SGML/XML은 임의 형태 문서, 임의 응용에 대해 일반화 마크업을 정의하기 위한 방법을 기술하는 메타언어이다. 즉 문서의 작성시에 고려되는 문서의 논리적 정보를 표현 가능하다. 이러한 논리적 구분을 이용하여 정보사용자에게 좀 더 정확한 검색을 제공할 수 있다. SGML/XML을 이용하여 표현된 계층적 논리정보를 이용하여 다양한 문서 접근점을 제공할 수 있으며, 문서의 재사용 및 동적인 문서제시를 가능케 한다. 본 논문에서는 SGML/XML 정보검색의 장점과 이러한 시스템을 구현하기 위한 구현 단계 및 구성요소를 알아보고자 한다. 아울러 구현사례로 STEER-SGML/XML 검색 시스템을 알아본다.

1 서론

기존의 정보검색 방법론들은 대부분 하나의 문서 전체를 색인과 검색의 단위로 삼는다. 문서에 표현되어 있는 논리적 구조 정보 (예를 들어 논문인 경우 문서제목, 저자, 요약, 내용, 참고문헌에 대한 구별)들은 모두 무시되어 문서는 일련의 문자열로 간주되어 색인되게 된다. 따라서 정보사용자는 문서에 접근하기 위해 색인어로 추출된 단어만을 사용가능하다. 문서 생성시에 주의깊게 고려된 여러가지 다양한 정보가 모두 손실되게 된다. 이러한 문제는 기존의 문서 표현 형식이 문서의 논리적 구조와 물리적 정보를 혼합하는 방식이기 때문이다. 예를 들어 논문을 워드프로세서로 작성하게 되면, 문서작성자는 제목부분, 요약부분을 문서의 외적인 물리적 정보 (스타일정보) 등으로 표시하여 준다. 그러나 문서를 색인하는 측면에서 이러한 정보는 서로 구분가능 하지 않으므로 이를 모두 무시하게 된다. 결과적으로 사용자는 문서에 색인어만을 통해서 접근할 수 밖에 없게 된다 [2].

SGML (Standard Generalized Markup Language) / XML (eXtensible Markup Language)은 이러한 문서의 물리적 정보를 제거하고 논리적 정보를 마크업을 통해 구분 짓는다 [4, 5, 10]. 이러한 마크업의 사용은 문서 내부에 내제된 논리적 정보를 구분지어 표현함으로 문서의 검색에서 다양한 접근점을 제공해주고, 문서의 재사용, 교환을 극대화 할 수 있다. 본 논문에서는 SGML/XML등의 구조문서가 제공하는 검색의 효용성을 알아보고, 구조검색 시스템 구축을 위한 네 가지의 접근방법에 대해 기술한다. 또한 사례연구로 STEER-SGML//XML (STructured Entity & Element Retrieval system) 검색 시스템에 대해 알아보고자 한다 [2].

2 SGML/XML과 정보검색

2.1 SGML/XML 문서와 논리적 구조정보

전자화된 문서가 점차로 증대함에 따라 단순히 문서의 초록만을 검색하고자 하는 요구에서 문서의 전문을 검색하고자 하는 요구가 증대되었다. 따라서 검색의 대상인 문서의 길이가 상당히 길어지게 되었다. 문서의 길이가 길어지면서 문서단위로 검색을 제공해도 사용자는 문서에서 자신이 필요로 하는 정보를 얻기위해 문서내 검색을 다시 수행해야 하는 불편이 생긴다. 예를 들어 검색시스템이 책 한권을 검색해 주어도 필요한 장, 절에 대한 정보를 얻기위해 다시 책 한권내에서 검색을 수행해야 한다. 이러한 불필요한 정보검색을 줄이기 위해 연구되는 것이 구절검색(passage retrieval)과 구조검색(structured document retrieval)이다 [7, 8, 9].

구절검색은 문서를 n개의 부분으로 자동 혹은 반자동으로 구분하고, 사용자의 검색요구에 대한 문서전체가 아닌 구분된 n개의 부분 중 일부를 보여준다 [7]. 이러한 방법은 사용자에게 좀 더 정확한 부분을 검색하여 줄 수 있으나, 문서 저작자가 의도한 문서의 논리적 구분과 검색 시스템이 임의로 구분한 n개의 논리적 구분과 서로 일치하지 않게 되며, 사용자는 문서의 저작자가 의도한 문서의 구조를 통해 검색을 수행할 수 없다.

구조검색은 SGML/XML등의 마크업 언어를 통해 문서 저작시점에서부터 저작자에 의해 부여된 논리적 구분을 검색에 사용한다 [8, 9]. 이러한 검색은 문서를 n개의 부분으로 구분지어 검색할 수 있음은 물론 저작자가 의도한 논리적 구분이 그대로 검색에도 사용되어 질 수 있다. 또한 논리적 구조의 계층적 특성에 따라 다양한 검색을 수행할 수 있다. 그러나 이러한 구조 정보를 문서저작단계에부터 고려해 주어야 하는 단점이 있다.

다음은 전자통신연구원 자연어처리부에서 작성한 정보과학회를 위한 SGML문서 예이다 [1].

```
<!DOCTYPE 논문 SYSTEM [
<!ENTITY fig01 SYSTEM "9609a_07f01.bmp"
NDATA bmp>
<!ENTITY fig02 SYSTEM "9609a_07f02.bmp"
NDATA bmp>
<!ENTITY fig03 SYSTEM "9609a_07f03.bmp"
NDATA bmp> ]>
<논문> <전방> <논문지정보>
<잡지명 언어="한글">정보과학회논문지(A)
</잡지명><ISSN>1226-2315</ISSN>
<권>제 23 권</권><호>제 9 호</호> <출판일>
1996. 9</출판일></논문지정보><제목그룹>
<제목 언어="한글">메쉬 멀티프로세서 분할을
위한 효율적인 이차원 Packing 알고리즘</제목>
<대체제목>An Efficient Two-Dimensional Packing
Algorithm for Partitioning Mesh ultiprocessors
</대체제목></제목그룹>
<저자그룹><저자 rid="aua"><이름 언어="한글">
황인재</이름><대체이름>Injae Hwang</대체이름>
<회원구분>정 회 원</회원구분></저자>
<소속><소속기관>충북대학교</소속기관>
<소속부서>컴퓨터교육과</소속부서>
<직책>교수</직책></소속></저자그룹> .....
```

그림1. SGML 예제 문서 (정보과학회 논문)

그림 1과 같은 SGML/XML문서는 문서의 스타일 정보는 표현되어 있지 않으며 문서의 내용과 논리적 구조를 구분지어 표현하고 있다. 그림 1의 문서에 대한 논리적 구조는 다음과 같다.

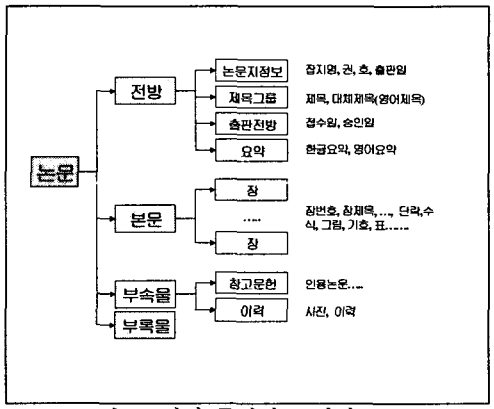

그림 2: 예제 문서의 논리적 구조

2.2 SGML/XML 구조 검색의 장점

SGML/XML 문서는 계층적 구조와 이에 따른 내용으로 구분되므로 다음과 같은 검색측면에서의 장점을 갖을 수 있다. 그림 1의 SGML예제 문서와 그림 2의 논리적 구조를 통해 그 장점에 대해 기술한다.

➢ 문서접근 (document access point)의 다양성

기존의 단순 색인어만으로 문서에 접근하는 기능 외에 문서의 계층적 구조정보와 내용정보를 혼합하여, 다양한 각도로 문서를 검색할 수 있다. 예를 들어 특정한 키워드가 제목에 포함되거나, 저자 소속, 요약등의 문서의 특정 부분을 지칭하여 검색의 범위를 한정할 수 있으며, 6개의 장으로 구성된 문서, 세명이상의 저자가 있는 문서등 문서의 구조 정보를 이용하여 검색이 가능하다. 이러한 검색은 사용자가 생각하는 모든 조합이 가능하다.

➢ 동적인 문서 제시 (dynamic presentation) 가능

SGML/XML은 문서의 일부만을 골라내어 자유롭게 조합하여 사용자에게 제시할 수 있다. 즉 문서전체를 보낼 필요없이 문서의 요약, 저자부분, 참고문헌만을 독립적으로 분리하여 볼 수 있으며, 스타일정보를 외적으로 부가하여 여러형태로 보여줄 수 있다. 따라서 불필요한 네트워크의 과부하와 시스템의 과부하를 감소시킬 수 있다. 예를 들어 1장만 분리하여 보여줄 수 있으며 사용자가 원하는 형태로 변경가능하다. 또한 문서내에 포함된 그림만을 조합한 새로운 형태로 제시가 가능하여 제시된 문서의 재사용 또한 보장된다.

➢ 검색정보의 일관된 관리

일반적으로 문서에서 특정 부분을 자동 인식(예를 들어 제목부분)하면 오인식이 항상 개입되게 된다. 따라서 제목이 아닌 엉뚱한 부분이 제목으로 인식되어 검색될 수 있다. 그러나 SGML/XML은 구조에 대한 유효성(validation)을 마친 상태이므로 항상 일관된 정보를 바탕으로 검색에 사용될 수 있다.

➢ 다양한 부가 정보 제공

문서의 여러부분이 계층적으로 구분되어 있으므로 검색외의 여러 부가 정보를 얻을 수 있다. 예를 들어 가중치계산을 위한 용어 분포정보, 문서 요약, 분류를 위한 용어분포정보등을 부가적으로 얻어 활용 할 수 있다. 또한 문서의 브라우징을 위한 다양한 정보를 제공한다. 따라서 참고문헌에 있는 논문을 이용하여, 본 논문을 참고하고 있는 논문에 대한 브라우징 및 본 논문의 저자가 쓴 다른 논문 검색등을 가능케 해준다.

3 SGML/XML 검색 시스템

3.1 SGML/XML 검색 시스템 분류

본 논문에서는 구조문서를 검색하는 검색 시스템을 구현하기 위한 방법을 4가지로 분류하여 설명하고자 한다. 분류기준은 SGML//XML이 표현하는 구조정보를 검색에서 얼마만큼 손실없이 제공하는지에 대한 척도로 분류하였다. 다음은 4가지의 구조검색시스템의 분류이다.

Phase 1 : 단순 검색 방법 (Simple Approach)
Phase 2 : 독립 필드 검색 방법
(Independent Field Approach)
Phase 3 : 통합 필드 검색 방법
(Unified Field Approach)
Phase 4 : 구조 엘리먼트 검색 방법
(Structured Element Approach)

각 분류에 대한 설명은 다음과 같다.

Phase 1: 단순 검색 방법 - 이 방법은 SGML/XML의 마크업을 모두 무시하고 하나의 문서를 하나의 검색단위로 하여 기존의 검색 시스템에 그대로 적용하는 방법이다. 이러한 시스템은 기존의 검색시스템을 그대로 이용할 수 있으나 SGML/XML문서가 갖는 모든 정보를 손실하게 되어 구조검색이 갖는 장점을 전혀 제공할 수 없다.

Phase 2 : 독립 필드 검색 방법 – 이 방법은 SGML/XML 문서를 SGML 파서 혹은 XML 파서를 사용하지 않고 특정 부분을 n개의 독립된 텍스트 문서로 분리하는 방법이다. SGML/XML 문서를 구절검색을 이용하여 검색하는 시스템이라고 할 수 있다. 이러한 시스템은 기존의 검색 시스템을 약간 수정하여 이용가능 하다. 그러나 각각의 분리된 독립된 텍스트 문서들은 개개의 문서로 취급되어 부분/전체의 통합 기능이 없으며 문서 접근점 또한 구절 검색 수준으로 제약된다. 그리고 원래의 SGML/XML의 정보는 손실되므로 동적인 문서제시가 불가능 하다.

Phase 3 : 통합 필드 검색 방법 – 이 방법은 SGML 파서 혹은 XML 파서를 이용하여 문서를 n개의 논리적 부분으로 나누며 각 분리된 부분은 전체로 통합되어 하나의 문서전체를 이룬다. 이러한 방법은 Phase 2보다 많은 검색접근점을 제공할 수 있으며 분리된 부분 단위로 동적인 제시가 가능하며 부분과 전체의 제한된 통합기능을 제공한다. 그러나 새로운 역화일 구조의 구성이 필요하여 기존의 검색시스템으로 구현시 색인저장 구조의 낭비가 생기며 SGML/XML 파서가 필요하다.

Phase 4 : 구조 엘리먼트 검색 방법 – 이 방법은 SGML/XML의 파싱된 결과를 모두 분리하여 저장하며 서로의 연관관계를 모두 구분하는 방법이다. 즉 SGML/XML 파싱트리 자체를 색인구조화 하여 사용한다. 이러한 시스템은 SGML/XML이 주는 모든 장점을 사용할 수 있으나, 상당한 색인저장 공간이 요구되며 검색측면에서 빠른 속도를 얻기 위해 여러 기술적 어려움이 존재한다.

3.2 SGML/XML 검색을 위한 검색 시스템의 요구사항

앞에서는 SGML/XML 검색이 주는 장점과 각 구현 단계에 대해 알아보았다. 그러나 이러한 검색기능 외에 SGML/XML 검색 시스템이 가져야 할 요구 사항은 다음과 같다.

문서 검색 기능 - Phase 3 또는 Phase 4

파싱 - SGML/XML validation 파싱 기능

Packer/Unpacker 기능 – SGML/XML문서가 포함하는 그림등의 외부 엔터티를 네트워크상에 분리된 사용자에게 전달하기 위해 문서와 외부 엔터디등 필요한 요소를 하나로 묶어 전송하여 사용자측 클라이언트에서 재구성하는 기능

동적 제시를 위한 SGML/XML 포맷팅 기능 - SGML/XML 문서는 자체에 문서의 스타일 정보를 갖고 있지 않다. 사용자에게 문서를 제시하기 위해서는 DSSSL (Document Style Semantics and Specification Language), XSL (eXtensible Style Language)등을 이용하여 포맷팅 하는 기능

SGML/XML repository 기능 - SGML/XML 문서를 저장하고 꺼내오는 기능, 문서 단위뿐만이 아닌 문서의 일부를 가져오는 기능도 포함되어야 한다.

3.3 개발 사례 : STEER-SGML/XML

본 논문에서는 SGML/XML을 위한 검색 시스템으로 STEER-SGML/XML의 구현 사례를 다루고자 한다. 본 시스템은 전자통신연구원 컴퓨터.소프트웨어 연구소 자연어처리부에서 개발한 정보검색시스템으로 SGML/XML 검색 시스템이다 [2]. STEER는 Phase 3에 해당하는 통합 필드 검색 방식을 지원한다. STEER-SGML/XML 시스템의 구성 요소는 다음과 같다.

SGML/XML 파서
SGML/XML meta fileter
SGML-XML converter
DSSSL을 이용한 XML-to-RTF 실시간 변환

STEER는 meta-fileter방식을 이용하여 Phase 3의 검색기능을 제공하는데 *meta-fileter*는 SGML/XML 문서의 각 엘리먼트를 하나의 필드로 분리하며 분리된 필드는 모두 합쳐서 전체 문서를 구성하게 된다. 또한 각각에 구분된 필드의 내용은 XML로 저장되며, 개별적인 저장과 꺼내오기가 가능한 간단한 SGML/XML repository 기능을 지원한다. 그림 3은 STEER-SGML/XML의 구성 개념도이다.

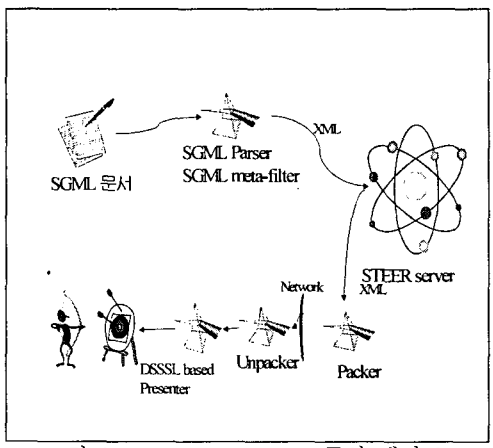

그림 3. STEER-SGML/XML 구성 개념도

STEER를 이용한 검색 예제는 다음과 같다. 메타필터를 이용하여 검색 필드를 생성하여 이를 검색 화면에 적용하였으며, 실시간 포맷팅 기능을 지원한다. 그림 4는 예제 질의로 "알고리즘"이란 검색어가 문서내 임의의 위치에 존재하고 영문요약에 "retrieval"이라는 검색어가 있는 문서를 검색하고자 하는 화면이다. 그림 5는 그림 4의 예를 통해 검색된 검색 결과보기 화면이다. 그림 6과 7은 검색 결과를 부분/전체를 DSSSL [6]을 이용하여 실시간 포맷팅하여 RTF(Rich Text Format)로 변환하여 사용자에게 동적으로 제시하는 화면이다. RTF viewer는 MS-word를 사용하였다.

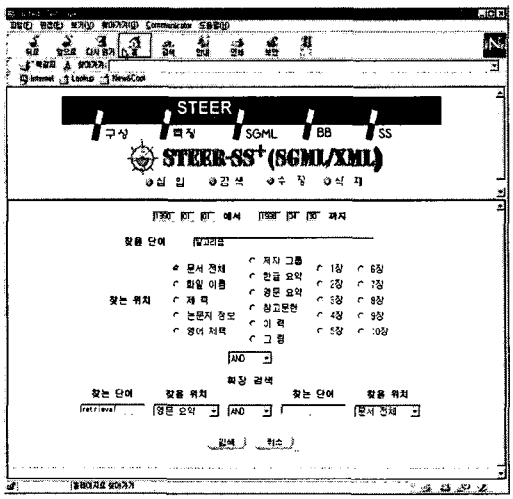

그림 4: STEER-SGML/XML 검색 화면

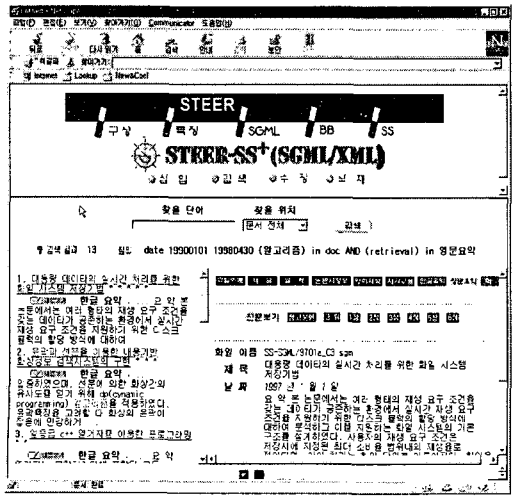

그림 5. STEER-SGML/XML에서의 검색결과화면

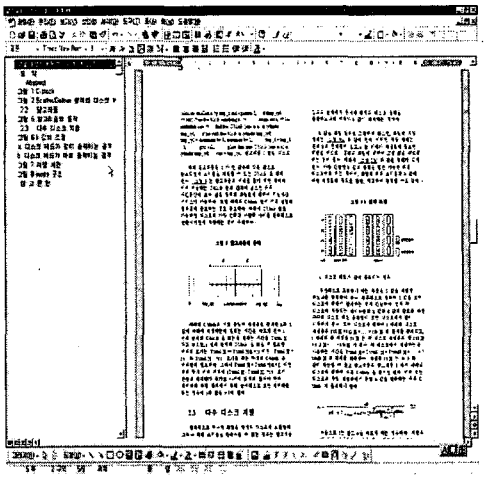

그림 6 : STEER-SGML/XML의
동적 문서 제시 화면 (two column)

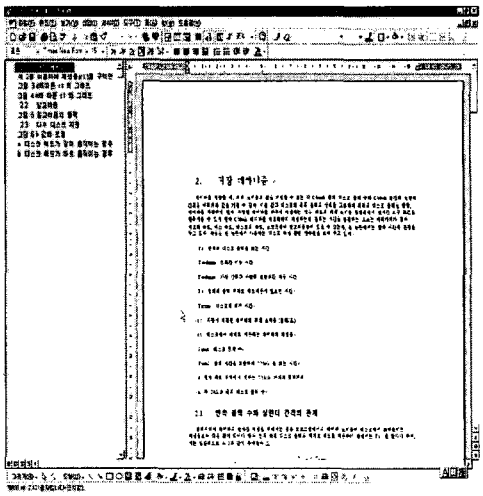

그림 7 : STEER-SGML/XML의
동적 문서 제시 화면 (one column)

4. 결론

SGML/XML정보검색 시스템은 SGML/XML 문서내에 표현되어 있는 계층적 논리적 구분을 이용하여 사용자에게 다양한 문서 접근점 및 각 논리구분의 의미에 따라 다양한 각도로 문서에 대한 접근 점을 달리 해석 할 수 있다. 또한 문서의 스타일정보가 분리되어 있으므로 동적인 문서제시가 가능하여 불필요한 문서전송 없이 원하는 부분을 원하는 형태로 사용자측에 전달할 수 있다. SGML/XML

자체가 문서의 전송 및 재사용을 목적으로 하므로 검색된 문서를 재차 가공없이 재사용이 가능하다. 이러한 SGML/XML 정보검색 시스템을 구현하는 단계로는 SGML/XML에 표현된 정보를 얼마만큼 손실없이 유지하느냐에 따라 Phase 1에서 4로 나누어 각각의 장단점을 알아보았다.

SGML/XML정보검색 시스템의 구현 사례로 전자통신연구원 자연어처리부에서 개발된 STEER-SGML/XML 시스템을 알아보았다. SGML/XML이 표현하는 모든 구조정보를 검색에 이용하기 위해서는 효율적인 Phase 4 수준의 검색시스템이 개발되어야 하며, 구조적 질의에 대한 연구 및 구조정보를 손쉽게 질의 할 수 있는 질의 인터페이스 등이 연구되어야 한다.

참고문헌

[1] 최종연구보고서 SGML 테스트 데이터 구축, 전자통신연구원, 1997

[2] 우리말 정보처리 S/W 기술 개발에 관한 연구, 정보통신부, 전자통신연구원, 1997, 1998

[3] Brian E. Travis, Dale C. Waldt, "The SGML Implementation Guide" Springer, 1995

[4] Charles F. Goldfarb,, "The SGML Handbook", Clarendon Press, Oxford, 1990

[5] ISO 8879, Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML), Includes Amendment 1, 1988

[6] ISO/IEC 10179:1996(E), Information Technology – Processing Language- Document Style Semantics and Specification Language (DSSSL)

[7] Massimo Melucci, Passage Retrieval : A Probabilistic Technique, Information Processing & Management, Vol34, No. 1, pp. 43-68, 1998

[8] M. Lalmas, "Dempster-Shafer's Theory evidence applied to structured documents: modeling uncertainty", Proc. Of ACM SIGIR-97, Philadelphia, pp.110-118, 1997

[9] R. Wilkinson, "Effective retrieval of Structured Documents", Proc. Of ACM SIGIR'94, pp 311-317, Dublin City, 1994

[10] XML (eXtansible Markup Language), XSL(eXtensible Style Language) , available via http://www.w3c.org/