□ 기술해설 □

문서구조화와 정보검색

충남대학교 맹성현*
한국전자통신연구원 주종철

1. 서 론

인터넷 및 인트라넷의 보편화로 정보 교환이 활발해지면서 필요한 정보를 손쉽게 취득하게 해 주는 정보검색 시스템의 역할이 과거 어느 때 보다도 중요하게 되었다. 일반적인 정보검색은 사용자의 질의에 기반하여 비정형 자료를 문서 단위로 찾아 주는 기능을 말하는데, 정형 데이타를 관리하면서 다양한 사용자 응용을 지원해 주는 전통적인 DBMS의 기능과는 구별된다. 과거에는 DIALOG와 같이 상용 데이타베이스를 대상으로 검색 엔진이 주로 사용되었으나, 근래에는 인터넷 자료만을 전문적으로 검색하는 시스템이 많이 보급되고 있으며(예: Inforseek, Lycos, Alta Vista, 심마니, 정보탐정 등), 기업이나 공공기관의 검색 요구사항(예: 국회 회의록, 특허 문헌, 사내 보고서 검색)을 만족하기 위해 문서의 내용과 구조에 기반한 검색기능에 대한 요구가 점점 증대되어 가고 있다.

기존의 검색기능은 주로 문서를 검색 단위로 간주하고 있을 뿐만 아니라 문서간에도 독립성을 가정하고 있어, 검색의 결과는 서로 독립적인 문서의 목록인 경우가 거의 대부분이고 사용자의 요구에 의해 전문(full document)을 제시해 준다. 그러나 대부분의 문서는 단순 텍스트외에 다양한 추가정보를 가지고 있는데 문서의 장, 절, 제목, 참고문헌 등을 명시하는 논리적(logical) 구조정보와 타 문서와의 연계성 구조정보, 그리고 문자형, 쪽 구분 등 제시(presentation)용 정보로 나눌 수 있다. 일반 정보검색의 경우 문서에 내재되어 있거나 명시되어 있는 이런 추가정보는 시스템의 검색과정에서는 보통 사용되지 않지만, 사용자가 전체 문서의 내용을 파악하거나 문서내의 필요한 부분을 찾아가는데 매우 중요한 역할을 한다.

문서간의 구조정보는 하이퍼텍스트(hypertext) 형태로서 명시적으로 표현될 수 있는데, 이는 문서를 선형적으로 보지 않고 작은 노드(node)라고 불리는 텍스트 조각이 링크(link)로 서로 연결되어 그래프를 형성하고 있는 형태로 봄으로써 문서 처리를 비선형적 혹은 비순차적으로 할 수 있게 한다는 개념이다. 하이퍼텍스트는 사용자가 브라우징(browsing)하고 링크를 따라 손쉽게 항해탐색 할 수 있는 문서 형태이지만, 텍스트의 양이 증가함에 따라 이러한 문서 처리 방식은 너무 많은 시간을 요구하므로 결국 검색 기능을 필요로 한다. 현재 웹 환경에서 보편화된 HTML(Hyper Text Markup Language) 문서도 하이퍼텍스트 형태로 정보검색의 대상이 되어 왔으나 대부분 웹 검색엔진의 경우 링크 정보를 무시하고 각 노드를 독립적인 문서로 간주하여 검색한다.

인터넷 정보의 대부분을 차지하는 HTML 문서는 링크 및 간단한 계층 구조를 지니고 있지만 ISO 국제표준인 SGML(Standard Generalized Markup Language) 문서, XML (eXtensible Markup Language) 문서, HyTime(Hypermedia Time-based Structuring Language)링크에 기반한 구조문서는 마크업으로 구분된 엘리멘트의 구조적 연관성을 잘 표현하고 있다. 예를 들어 학술 논문이 가지고 있는 제목, 저자, 초록, 장, 절, 문단, 참고문헌

*종신회원

등과 같은 텍스트 단위가 계층적으로 구성되어 장이 절을 포함하듯 하나의 텍스트 단위가 다른 것을 포함할 수 있다. 문서가 가지는 이러한 구조 정보는 문서 전체가 아닌 작은 단위의 검색을 가능하게 하여 사용자가 원하는 특정 영역에 바로 접근할 수 있게 해줌으로써 검색의 정확성을 증대시킨다.

일반적으로 문서는 묵시적이든 명시적이든 다양한 구조 정보를 가지고 있다. 이러한 구조 정보가 검색 관점에서는 대개 무시되어져 왔는데, SGML이 표준화되고 HTML문서가 웹에서 일반화되어 감에 따라 구조문서(structured document)의 검색 및 활용 대한 관심이 높아지고 있다. 문서구조화 기술에 의해 표현된 정보는 혼란도가 높은 데이타를 조직화/정규화/지식화 하는 효과를 가지고 있어 정보의 일관성과 재활용성을 높여 준다. 이는 이제까지의 문서정보처리 방식을 초월하여 정보획득 기회의 증대 및 정보 재사용을 통한 정보 유통 비용의 감소를 가능하게 하면서 정보 인프라로서의 새로운 파라다임을 제공한다. 즉 기존 정보 저장 방법의 비일관성, 자료의 중복 및 불일치, 정보접근의 한계 등 문제[26]로 발생하는 사용자 정보부하를 줄일 수 있는 가능성을 제공하는 것이다. 특히 XML이 웹에서의 응용에 표준으로 자리를 잡아가고 있는 시점에 CALS/EC 및 디지탈 도서관 등의 응용에서는 문서의 구조정보를 어떻게 활용하는가에 대한 문제가 필수적이라 할 수 있다.

본고의 2절에서는 구조문서 검색 기술의 배경이 되는 내용기반 검색 기술을 관련되는 부분을 중심으로 소개하며 3절에서는 문서 구조화의 동향에 관해 정리하고, 구조문서의 검색에 관련된 이슈 및 기법을 4절에서 기술한다.

2. 내용기반 문서 검색

구조문서 검색의 기반이 되는 정보검색에 관한 기술이 국내에 본격적으로 알려지기 시작한 지 몇 년 안되지만, 그 기술은 독립된 분야로서 약 30여년 동안 외국에서 꾸준히 발전되어 왔다. 정보검색은 근본적으로 "내용기반"검색을 의미하는데, DBMS의 검색 기능과 같이 특정 속성(attribute) 값을 만족시키는 자료를 얼마나 신속하게 찾느냐의 문제가 아니라, 자연언어로 구성된 문서의 내용과 명확하지 않은 사용자의 정보 요구가 얼마나 근접한가를 계산하여 그 값이 높은 순서로 문서를 검색하는 것이다. 구조문서 검색은 이러한 순수 내용과 구조 정보를 효과적으로 혼합하여 새로운 차원의 기능을 제공하는데 주안점을 둔다.

문서의 내용과 사용자 질의의 유사도를 계산하는 정합(matching)과정을 효율적으로 하기 위해서는 문서의 내용을 미리 분석하여 내부적인 표현으로 변환하는 것이 필요한데, 이를 색인(indexing)이라 한다. 색인 과정에서는 문서의 내용을 대표하는 용어를 선별하고 그 중요도를 계산하는데, 비교적 간단한 형태소 해석 수준에서의 자연어 처리와 통계 처리에 주로 의존한다. 정보검색의 역할이 적절한 문서와 그렇지 못한 문서를 구별하여 우선순위를 결정하는데 있으므로 적절한 용어 집합만으로 문서의 내용을 충분히 표현할 수 있다고 보는 것이 일반적인 견해이다. 구조검색을 위한 색인에서도 유사한 방법론을 사용할 수 있지만 구조정보를 어떻게 활용하는가 하는 것은 새로운 문제로서 검색기의 성능과 기능에 많은 영향을 줄 수 있다.

문서의 내용을 분석하여 표현하는 색인에 있어 고급 자연어처기술은 아직 큰 역할을 못하고 있는데, 그 이유로는 크게 두 가지를 들 수 있다. 첫째는 위에서 언급한대로 검색 기능 자체가 고급 자연어처리를 요구하지 않는다는 견해이고[24], 둘째는 자연어처리 기술이 현재 가지고 있는 한계와 연구방향이 검색 분야의 요구를 충분히 만족시키지 못하고 있다는 것이다. 결국 자연어처리는 주로 색인용어를 보다 정확히 검출해 내는데에 일조를 해왔는데, 근래에는 색인결과의 단위를 용어쌍, 구, 혹은 용어와 용어간의 관계성 등으로 확장시켜 검색의 신뢰도에 긍정적인 영향을 주는 방법론도 개발되었다[9, 15, 25]. 또한 단어의 의미적 모호성을 해소하는 경우 검색의 신뢰도 향상에 도움을 줄 수 있는데[10], 자동 모호성 해소 기술의 정확도가 그 실용성을 좌우한다고 할 수 있다[20]. 이러한 새로운 기술들이 주로 정확도

의 개선에 영향을 주므로 구조정보를 활용하는 데에 있어 상승효과를 제공할 것으로 기대된다.

내용기반 검색과 DBMS 프로세스와의 또 다른 차이점으로 유사도 계산 여부를 들 수 있다. 정보검색에서는 질의와 검색 대상 문서와의 유사도를 계산하여 문서의 검색 여부 혹은 랭크(rank)를 결정하게 되는데, 그 결과는 어떤 검색 모델을 사용하는가에 좌우된다. 가장 오래된 것으로 원칙적으로 문서랭킹을 계산할 수 없는 불리언(Boolean) 모델, 벡터공간(Vector Space) 모델, 이들을 통합하는 P-Norm 모델[19], 확률 모델, 추론망 모델[27], 논리기반 모델[28] 등을 축으로 하여 다양한 모델들이 개발되어 왔는데, 구조문서 검색에 적합한 모델의 개발이 필요하다.

내용기반 검색을 효율적으로 수행하기 위해서는 특정 용어를 가지고 있는 모든 문서를 빠르게 찾아내는 기능이 필요한데, 이를 위해서 역색인(inverted index) 구조를 사용하는 것이 일반화되어 있다. 문서를 색인단위로 하고 용어를 색인의 결과로 사용하는 경우 검색모델과는 거의 독립적으로 사용되는 이 하부구조는 크게 용어 사전에 해당하는 B^+ 트리, 문서식별자 목록을 가지고 있는 포스팅 파일(posting file), 문서 파일 등으로 구별되는데, 검색엔진 및 문서DB 혹은 색인의 특성에 따라 변형되기도 한다. 문서의 내용정보와 구조정보를 동시에 필요로 하는 구조검색에서는 이들이 효율적으로 사용될 수 있도록 설계된 새로운 하부구조가 필요하다.

3. 문서 구조화 동향

3.1 문서구조화 기술의 배경

조직내 80~90%의 문서정보가 데이타베이스 등의 조직화된 형태로 관리되지 않고 있다. 이러한 현실을 극복하고자 문서구조화 기술은 문서정보의 혼란도(entropy)를 낮추어 주는 단계로서 종이 상태로부터 디지탈화된 이미지로의 변환 단계와 OCR을 통한 문자 인식의 단계를 넘어서서 데이타에 마크업을 추가하여 조직화/정규화/지능화된 정보 시스템 개발을 유도하게 한다[6]. 문자나 비트 데이타로 구성된 기존의 순차적인 정보 생성 방법에서 인간의 사고 체계와 보다 유사한 단위와 링크 정보를 표현하여 비순차적 형태로의 활용성을 높여주는 기술로서 광의적으로 볼 때 하이퍼미디어 기술로 연구되어 오던 분야이다.

3.2 문서구조화 표준 SGML/XML/HTML 동향

문서구조화를 위한 표준적 정보표현 방법으로서 1986년 ISO 표준으로 제정된 SGML은 SGML 표준에 기반한 정보 시스템 구축 효과를 믿는 일부 사용자 그룹에서만 사용되어 왔다. SGML은 하이퍼미디어 정보의 표현(representation) 철학으로서 논리적 구조정보(제목, 장, 절, 인용문, 문단…)와 물리적 외양정보(스타일, 레이아웃…)를 분리하여 논리적 구조정보가 특정 응용 프로그램과 시스템에 종속되지 않도록 하는 것에 초점을 두고 있다. 반면에 SGML의 응용인 HTML은 사용자의 수용도를 높이기 위해 정보 제시(presentation)에 중점을 두었다. 웹브라우저들은 외양정보를 일정한 태그 집합에 대해 스타일 프로세싱 시멘틱스를 브라우저에 고정하여 가벼운 처리를 하고 인터넷 기술을 접목하여 세계 도처의 정보 서버를 연결하여 놀라운 호응을 얻게 되었다. 최근에 SGML의 웹환경 적용을 위하여 발표된 XML은 정보의 표현과 제시를 모두 중요시 함으로써 SGML의 기본 철학에 충실하면서도 수용도를 높일 뿐만 아니라 SGML보다 가벼운 처리를 위해 단순화된 스펙으로 산업계의 수용을 높이는 안이다. XML 패밀리 스펙인 XSL(eXtensible Stylesheet Language)과 XLL (eXtensible Linking Language)은 XML이 SGML에 대응되는 것과 같이 DSSSL (ISO10179 : Document Style Semantics and Specification Language) 과 HyTime(ISO 10744 : Hypermedia Time-based Structuring Language)의 링크 모듈에 대응하여 스펙화가 진행되고 있다.

XML은 HTML이 지니는 근본적인 한계를 극복하고 다음의 두 가지 측면에서 SGML의

본질적인 요소들이 웹환경에 적용되도록 타협된 안이다[13, 22].

첫째, 논리적 구조정보와 물리적 외양정보가 혼재된 기존의 HTML 문서와는 다르게 문서구조화의 기본적인 이념을 따르도록 외양정보를 분리하여 생성하는데 충실해진 점이다. 기존의 HTML 문서들도 최근 CSS(Cascading Style Sheet)로 외양정보를 분리하고 있지만 HTML 문서 코드를 살펴보면 근본적으로 논리적 구조정보와 물리적 외양정보가 혼돈스럽게 혼재되어 있으며 오히려 사용자의 수용도를 높이기 위해 외양 정보에 초점이 맞추어져 왔음을 알 수 있다.

둘째, 기존의 HTML문서에서는 태그 확장이 불가능한 반면에 XML에서는 사용자 태그 확장이 가능하도록 만들었다. 기존 HTML DTD(Document Type Definition)의 태그 집합은 하나의 DTD로 여러 문서 유형들을 포괄하기 위해 의미전달이 쉽지 않은 H1, H2 등과 같은 태그가 사용된다. XML에서는 SGML에서와 같이 여러 유형의 DTD를 정의할 수 있게 하고 이에 따른 유효한 문서를 생성하게 함에 따라 일부 적용 규칙이 제외되거나 단순화된 점 외에 SGML과 다르지 않다. 또한, DTD 생성에 따른 부담을 줄이고 기존 정보 시스템과의 접목을 높이기 위해 well-formed 문서라는 새로운 개념을 만들어 사용자가 자유롭게 태그를 확장할 수 있도록 하였다. 이는 다음 절에 설명하게 될 Filtered SGML의 접근방법과 거의 동일한 개념적 바탕을 지닌다고 볼 수 있다. 즉 구조화된 문서 객체간의 연관성에 관한 정보 처리가 시스템 측에서 거의 없는 정보 생성 또는 변환 방법이며 이에 대한 처리는 사용자의 부담으로 전이하는 방법이다.

이제까지의 SGML에 의한 문서처리 시스템 개발의 문제점은 논리적 구조정보에 대한 유효한 문서 생성 및 저장에 초점을 맞추고 물리적 외양정보에 대한 처리는 산업계의 현실적인 방법(FOSI/ CSS 등)을 채택하여 출시되어 왔었다는 점이다. 최근 관심이 높은 XSL 스펙은 웹브라우저의 시장 동향에 따른 상업적인 가치에 따라 가변적인 요소가 있으며 기존의 SGML 문서를 CSS 또는 업체 자체의 스타일 정의 규칙에 따라 상용화하여 온 흐름과 같아 개방성 부족의 위험은 상존하고 있다. 따라서, 문서구조화 시스템의 저장 포맷으로서 SGML(또는 유효한 XML)을 선택하는 것이 안전하며 전달은 XML이나 HTML 등으로 변환하여 사용자에게 전달하는 방법을 권유하는 근거를 제공한다. 즉 문서구조화 기술의 중심이 논리적 구조정보에 대한 유효한 정보의 생성 및 저장에 초점을 맞추는 것이 위험도가 낮다.

XML이 상업적인 측면에서 영향력을 높일 기술로 부각되면서 데이타베이스, 정보교환, 정보관리 관련 업체 및 기술 개발자의 관심을 모으고 있다. 태그가 고정된 HTML의 한계 극복이 가능하여 사용자에게 다양한 시멘틱스를 제공하는 각종 응용 분야 즉 데이타베이스 정보교환, 검색, 제시 등의 단계에서의 활용성이 높다. 그러나, 최근 국내외에서 일어나는 현상으로 XML의 부각이 문서구조화를 위한 새로운 움직임으로 인식되고 있지만 실은 SGML의 근본으로부터 다른 점이 거의 없다. 다만 복잡한 SGML 스펙을 단순화하여 파싱의 복잡도를 줄이고 웹환경 사용자 접근을 쉽게 하기 위한 제시 측면을 고려하고 있다. 따라서 문서구조화 기술의 모체로서 SGML/DSSSL/HyTime의 기본적인 이해가 매우 중요하며 SGML의 기본 철학의 유지와 상업화 또는 사용자 수용 측면의 양면성에 대한 기술적 저울질을 정확히 할 수 있는 능력 배양이 먼저 요구된다. 사용자 수용도를 높이기 위해 타협된 안인 XML/XSL/XLL의 동향에 민감하기보다는 본질적인 면에서 분석·생성·저장·검색·교환 전 문서 사이클을 지원하는 SGML 기반 정보 시스템을 이해하는 것이 중요하다.

3.3 문서구조화 접근 방법

지금까지 문서정보 전자화의 주류는 워드프로세서와 DTP(Desk Top Publishing) 관련 제품이었다. 그러나, 대부분의 제품들은 논리적 구조 정보와 물리적 외양정보가 혼재되어 특정 응용 제품에 종속됨으로써 정보의 재활용을 위해서는 특정 응용 프로그램간의 변환 프로그램에 의해서 가능하였다. 문제점으로서는 변환 프로그램의 수가 많고 정보의 손실이 많다는

점이다. 국제 표준적 방법에 의한 문서구조화는 특정 응용 프로그램과 시스템에 종속되지 않는 방식으로 정보의 활용성을 염두에 두고 설계된 정보의 생성으로 보다 적고 간단한 변환 프로그램으로 정보의 교환이 가능하다. 그 근거로는 SGML 출판 모델은 문서의 논리적 구조정보와 물리적 외양정보를 분리된 파일들로 생성하기 때문에 정보 변환시 논리적 구조정보와 물리적 외양정보가 혼돈스럽게 혼재되어 있는 경우보다 매우 간단해지게 된다.

따라서, SGML 출판 모델에서는 논리적인 구조정보에 대해서는 SGML에서 정의된 문법을 따르지만 물리적인 외양정보에 대해서는 DSSSL 표준으로 정의된다. SGML 출판 모델에 따른 첫번째 단계로 논리적 구조정보에 대한 문서구조화를 위해서는 Filtered SGML과 Native SGML의 두 가지 접근 방법이 있는데 이에 대한 충분한 인식과 기술적 검토를 통한 선택이 필요하다[7].

우선 교환에 초점을 둔 Filtered SGML은 모든 국제 표준이 공통적으로 가지는 장점으로서 기존에 이미 생성된 문서들을 표준적 형태로 만드는 자체에 의의가 높다. 주로 간단한 문서 구조에 적용되며 태그 맵핑 및 변환 검증 등의 비용으로 향후 운용 비용이 많이 들고 문서구조화가 지니는 모든 효과를 기대할 수 없는 점을 유념할 필요가 있다. 따라서, 새로이 구조화된 문서를 생성시 권장할 만한 방법이 되지 못한다. Native SGML 접근방법은 충분히 검토된 문서유형 DTD를 설계한 후 SGML 전용 편집기를 사용하여 DTD 문법에 유효한 정보를 생성하여 SGML이 추구하는 본질에 충실한 방법이다. 이는 문서구조화 기술의 모체가 되는 SGML이 지니는 본질이 정보 교환을 훨씬 능가하는 기대효과가 보장되는 정보 시스템의 인프라가 만들어지도록 하는데 필요한 기본적인 조건으로 볼 수 있다.

구조문서의 생성단계는 기존의 혼란도가 높은 문서정보의 조직화/정규화/지식화의 정도를 결정하는 단계로서 전 문서사이클에 걸쳐 특히 검색 및 관리 단계에서 정보 생성자가 예측하지 못하는 활용을 가능하게 하여 사용자의 정보 과부하를 줄여 주며 각종 응용 시스템의 질을 결정하게 된다. 이를 위해서는 문서유형의 설계가 여러 각도에서 활용성을 염두에 두고 진행되어야 한다는 점이 중요하다. 문서 사이클에 기반한 문서구조화의 모든 장점을 염두에 두고 설계된 DTD에 따른 구조화 정보의 생성에 따라 저장, 검색, 관리 등에 까지 확장된 정보 시스템에서 문서구조화의 진정한 효과가 기대될 수 있기 때문이다.

4. 구조문서 정보검색

정보검색의 관점에서 볼 때 구조문서는 다음과 같은 장점을 제공한다.

- 문서 의미정보의 명시화 : 문서의 논리적 구조는 문서의 내용을 파악하는데 많은 정보를 제공한다. 예를 들어, 학술논문의 경우 '서론'과 '결론'이 갖는 내용과 '관련연구' 혹은 '실험 및 결과'가 갖는 내용이 다르므로, 검색시 문서의 특성에 따라 특정 영역에만 초점을 맞추어 효율 및 검색 신뢰도를 향상시킬 수 있다.
- 정보접근점의 다양화 : 문서를 분리할 수 없는 하나의 단위로서 접근하는 것이 아니라 엘리먼트(element)를 노드로 갖는 계층적 트리 구조로 표현하므로 다양한 수준(level)에서의 엘리먼트를 독립적인 혹은 연관된 객체로서 검색 대상으로 삼아 질의를 표현할 수 있고 접근할 수 있다.
- 동적인 문서제시 기능 제공 : 문서를 하나의 균일한 문서로 보는 경우 검색 결과의 제시도 선형의 문서 리스트로 한정되는 반면에, 구조정보가 존재하는 경우 제시 대상이 임의의 엘리먼트가 될 수 있으므로 이들간의 계층적인 연관관계와 링크에 의한 연결관계 등을 동적으로 사용자에게 제시할 수 있다.
- 검색목적 정보 추가 가능 : 저자가 작성한 문서 원본에 추가하여 새로운 정보를 추가하고 사용하는 것이 용이하다. 예를 들어 문서에 관한 독자의 평이나 주석(annotation)을 포함하거나, 텍스트의 분석을 통해 얻어진 추가 정보를 구조문서에 포함시킬 수 있는 장치가 마련되어 있으므로 문서의 부가가치를 증대시키고 검색의 다양성을 제공할 수 있다.

정보검색 사용자 관점에서 볼 때, 문서의 구

조를 이용하여 직접 얻을 수 있는 혜택은 크게 두 가지로 요약될 수 있다. 첫째는 문서의 구조 정보를 활용하여 문서 검색의 신뢰도 및 효율을 높일 수 있는 방법이고, 둘째는 단순한 문서 단위의 검색을 초월하여 사용자로 하여금 문서 구조와 관련된 질의를 할 수 있도록 검색기에 새로운 기능을 부여하는 방법이다. 사용자의 새로운 정보요구를 만족시켜 주는 두번째 관점에서 볼 때 검색 단위가 문서의 특정 부분(예 : 절, 제목, 문단 등)일 수 있고 질의에도 구조적인 정보 및 엘리먼트 속성정보도 포함될 수 있다.

검색기술 자체와는 별도로 구조문서를 관리하고 검색을 지원해주기 위한 새로운 저장 구조에 대한 기술도 연구가 되고 있다. GMD[29]에서는 SGML문서를 저장하고 검색하기 위하여 객체지향 DBMS와 검색시스템을 통합하는 방법을 제시하고 내용기반 질의와 구조기반 질의를 객체지향 DBMS의 질의언어로 표현될 수 있도록 하였다. 미국 시라큐스 대학에서는 구조문서의 색인을 효율적으로 저장할 수 있는 기법을 개발하였고[12], 국내에서는 SGML문서의 다양한 검색을 지원하는 색인 저장 구조와 SGML문서를 관리하기 위한 O2 DBMS에 기반한 저장시스템이 개발되고 있고, SGML문서관리를 위한 저장기법에 대한 다른 유사한 연구도 수행되고 있다[2, 3, 4, 5].

다음은 정보검색 사용자 입장에서 구조화문서를 사용하는 접근방법 2가지를 중심으로 구조문서 정보검색 기술의 내용을 요약한다.

4.1 검색 신뢰도 향상을 위한 구조 정보의 사용

검색 신뢰도를 위한 구조 정보의 이용은 문서의 길이가 다양해 지면서 그 중요도가 증가하고 있다. 정보검색기술의 발전과정에서 볼 때 초기에는 검색 대상이 주로 논문의 초록이었으나 근래에 와서는 짧은 메모로부터 긴 보고서에 이르기까지 그 길이가 매우 다양하게 되었다. 긴 문서의 경우 사용자 질의를 만족시키는 부분이 특정 부분에 국한될 수 있으므로 단어의 전반적인 빈도 수에 의존하는 일반적인 검색기법으로는 한계가 있다. 긴 문서에 있어 문서의 주제가 변화하는 데에 따른 문제점을 극복하기 위해 문단검색(passage retrieval)이라는 방법이 개발되었는데, 이는 문서를 문단 구조에 따라 나누거나 일정한 길이의 인위적 문단으로 나누어 각각에 대한 적합성 계산을 한 후 적합한 부분만 전체 문서의 맥락과 함께 제시하거나, 각 적합성 계산 결과를 산술적으로 통합하여 전체 문서에 대한 최종 적합성을 계산하는 방법이다[30].

검색신뢰도를 향상시키기 위하여 구조정보의 특수 형태인 링크정보를 이용하는 방법도 연구되고 있는데, 링크는 일반 논문에서 볼 수 있는 상호참조에 의한 문서간의 링크일 수도 있고 HTML문서에서 볼 수 있는 단어와 문서간의 링크일 수도 있다. Savoy[21]는 문서간의 인용 관계를 링크로 간주하여 검색신뢰도를 향상시킬 수 있음을 보였고, 국내에서의 최근 연구[1]에서는 하이퍼링크의 방향성과 앵커 노드의 특성 등을 고려하여 문서 랭킹을 향상시키고 링크정보가 사용되지 않았을 때 질의어의 부재로 전혀 검색되지 않는 중요 문서를 검색할 수 있음을 보였다. 웹 문서가 대부분 HTML 문서로서 많은 링크 정보를 포함하고 있다는 면에서 볼 때, 이러한 연구의 결과는 웹에서의 응용가능성이 매우 높다고 할 수 있다.

4.2 구조 기반 질의 처리

구조 문서 검색은 기존의 검색 방법에 비해 문서로의 다양한 접근 경로를 제공한다. 사용자는 문서의 구조적 특성이나 속성 값을 사용하여 필요한 정보를 표현할 수 있고 질의를 구성하는데 있어 내용과 더불어 그 내용이 출현해야 되는 맥락(context) 정보를 명시적으로 포함시키므로써 보다 정확한 정보요구를 표현할 수도 있다. 예를 들어 단순히 "월드컵에 관한 신문 기사(문서)"라는 일반적인 검색질의 뿐만 아니라 "사회면에 게재된 월드컵에 관한 기사의 제목"과 같이 맥락(신문이라는 문서의 하위 엘리먼트인 사회면)정보가 제공되고 검색대상(제목)이 명시된 질의를 처리할 수 있는 검색환경을 말한다.

일반 문서와 비교하면 구조 문서는 SGML 문서의 경우와 같이 3가지 새로운 정보를 포함

하고 있다고 할 수 있다. 즉 엘리먼트간의 계층적 관계, 엘리먼트의 속성(attribute) 정보, 링크 정보가 텍스트와 함께 존재한다. 링크 정보는 속성 정보의 한 종류이긴 하나 하이퍼텍스트를 형성하는 독특한 기능을 가지고 있으므로 별도로 취급하는 경우가 많다. 결국 내용, 구조, 속성, 링크 정보가 혼합된 질의를 처리할 수 있는 정보검색 기법이 필요한데, 자연어로 표현된 질의의 예를 들면 다음과 같다.

"1998년 5월 15일 이후 사회면에 게재된 월드컵에 관한 신문기사를 인용한 논문의 제목"

여기서 "월드컵"은 내용 부분이고, "사회면" 및 "제목" 은 신문 문서의 구조적 정보이며, "1998년 5월 15일"은 속성 정보이다. 그리고 "신문기사를 인용한 논문"은 두 개의 문서 형태에 걸친 링크 정보를 표현하고 있다. 이러한 질의는 실제 시스템에서는 보다 형식화된 질의 형태로 변환되어 처리되어야 한다.

일반적인 검색엔진에서도 필드(field)를 명시하는 질의를 처리할 수 있는 기능을 갖추고 있는 경우도 있으나, 위에서 언급한 포괄적인 질의는 처리할 수 없다. 그러나 단순한 필드 검색 기능을 초월하면서 구조문서 검색기능을 만족시키기 위한 연구개발은 근래에와서 다수 진행되고 있다. SGML의 확장으로 나온 HyTime의 일부분으로 정의된 HyQ[17]은 내용부분만을 제외한 구조, 속성, 링크를 사용한 질의를 구성할 수 있게 해준다. [16]은 구조와 내용에 기반한 질의를 처리할 수 있는 강력한 표현력을 가지는 질의언어를 개발하였고, [8]은 위의 4가지 정보를 모두 사용할 수 있는 질의어를 정의하였으나 이를 처리하는 검색엔진은 아직 개발되지 않은 상태이다. [11]은 불확실성이론을 사용하여 구조정보만을 처리할 수 있는 모델을 제시하였으나 구현 및 실험은 아직 보고된 바가 없다.

국내에서는 추론 망에 기반한 새로운 모델이 개발되고 SGML문서의 색인기, 하부 저장 구조, 검색기가 구현되었다[14]. 이 연구에서는 주로 내용과 구조 정보에 초점을 맞추어 이들이 복잡하게 혼합된 모든 질의를 처리할 수 있는 새로운 기능을 제시한 것 이외에 문서의 SGML문서의 구조 정보를 적절히 이용할 경우 문서 단위 검색에 있어서의 신뢰도 향상에도 큰 영향을 줄 수 있음을 보였다. 유사한 연구로 누산기 개념[18]과 [12]의 구조문서 표현 기법을 활용하여 구조문서 검색의 효율성에 초점을 맞춘 연구가 있다[23].

위에서 사용된 것과 같은 질의를 수행하기 위해서는 색인 과정에서부터 검색 결과를 제시하는 과정까지 대부분의 과정이 기존 검색방법과는 다르게 진행되어야 한다. 우선 색인 과정에서는 각 SGML문서를 파스(parse)하여 DTD와 일관성 여부를 검사함과 동시에 문서 트리를 생성하여야 하고, 일반적인 검색에 사용되는 색인 과정을 거치되, 문서별 분석이 아닌 엘리먼트별 분석을 하여야 하므로 그 복잡도가 매우 높아진다. 구조문서 검색 모델에 따라 일반적인 색인과정에서 수행하지 않는 작업을 하기도 하는데, 예를 들면 상위 엘리먼트와 하위 엘리먼트와의 관계를 계산하여 구조정보에 내재되어 있는 의미정보를 사용하기도 한다[14].

색인결과는 기존 검색시스템에서 사용하는 색인구조와는 상당히 다른 구조로 저장되어야 한다. 이는 단지 색인이 엘리먼트 수준에서 수행되기 때문만이 아니라 문서의 트리구조가 저장되어 있어야 구조 관련 질의를 효율적으로 수행할 수 있기 때문이다. 예를 들어 "월드컵이라는 단어가 나오고 사진이 포함되어 있는 기사의 제목"이라는 질의에 답하기 위해서는 기사 엘리먼트와 제목 엘리먼트간의 연결관계가 저장되어 있어야 하고 사진 엘리먼트가 기사 엘리먼트에 포함되어 있다는 사실이 저장되어 있어야 한다.

검색과정도 일반 정보검색에서와 같이 각 엘리먼트를 독립된 객체로 볼 수 없을 뿐만 아니라 적합한(relevant) 하위 엘리먼트를 포함하는 상위 엘리먼트의 적합성 계산을 위한 모델이 정립되어 있어야 한다. 예를 들어 질의가 "월드컵이라는 내용을 가지고 있는 기사를 가지고 있는 사회면을 가지고 있는 신문"이라면 다수의 적합한 기사 엘리먼트를 가지고 있는 사회면 엘리먼트의 적합성 계산을 해야하며 이를 사용하여 신문 엘리먼트의 적합성 계산도 수행하여야 한다. 검색 후 결과를 제시하는 경우에도 검색된 엘리먼트가 문서 트리에서 가지

는 위치를 고려하여야 한다.

5. 결 론

구조화된 정보는 문서내 세부 요소들로 구성된 논리적인 객체와 각 객체간의 연관성을 정의하고 있고 외부 문서와의 연계도 가능하게 하므로 문서구조화는 방대한 정보의 조직화/정규화/지식화를 촉진시킨다. 따라서 평탄한 문서만을 대상으로 하던 정보검색 기술을 확장시켜 이러한 구조문서를 검색할 수 있는 기능을 제공하는 것은 디지탈도서관이나 CALS/EC분야를 비롯한 각종 응용 분야에 매우 중요하다. 뿐만 아니라, 구조문서는 일반 문서에 비해 문서 내용의 조직화 및 관련성에 대한 정보가 명시적으로 기술되어 있으므로 보다 상세한 검색 요구를 해결할 수 있는 가능성을 가지고 있다. 또한 구조문서에는 사용자의 주석이나 문서 가공을 거쳐 생산된 추가 정보들을 저장할 수 있는 여지가 있으므로 향후 새로운 형태의 정보 요구에도 부응할 수 있다. 따라서 구조문서 검색기능은 CALS/EC나 디지탈 도서관과 같은 새로운 응용분야에서 필수적인 기능으로 부각되고 있으며 정보검색 분야의 관점에서는 새로운 기술의 연구개발을 필요로 하는 하나의 분야가 되고 있다. 이미 일반 정보검색 기술은 상품화 관점에서 보면 성숙단계에 이르러 국내외적으로 치열한 경쟁이 이루어지고 있는데, 구조문서검색과 같은 새로운 개념의 검색기능에 투자하는 것이 기술개발이나 사업 측면에서 국내외적인 경쟁력을 갖출 수 있는 지름길이 될 수도 있을 것이다.

참고문헌

[1] 김동욱, 류준형, 주원균, 맹성현(1997). 링크정보를 이용한 검색 신뢰도의 향상. 한국정보과학회 춘계학술발표 논문집.

[2] 맹성현(1997). 구조화 정보검색 색인기 및 정보집약 알고리즘 개발. 한국전자통신연구원 부설 시스템공학연구소 위탁과제 최종보고서.

[3] 이원석 외 7인(1997). SGML문서 관리 시스템의 데이타베이스 관리기 설계 및 구현. 한국정보과학회 추계학술발표 논문집.

[4] 이용배, 손기락(1997). SGML문서의 저장을 위한 스키마변환기 및 자동삽입기의 설계 및 구현. 한국정보과학회 추계학술발표 논문집.

[5] 장재우(1997). 메타데이타 인덱스 구조 설계 및 구현. 한국전자통신연구원 부설 시스템공학연구소 위탁과제 최종보고서.

[6] Alschuler, L.(1995) "ABCD··· SGML : A User's Guide to Structured Information," International Thomson Computer Press.

[7] An ArborText White Paper "Native SGML vs. Filtered SGML", http://www.arbortext.com/natifilt.html.

[8] Arnold-Moore, T., Fuller, M., Lowe, B., Thom, J. and Wilkinson, R.(1995). The ELF Data Model and SGQL Query Language for Structured Document Databases. In Proc. of the Australian Database Conference, Adelaide, Australia.

[9] Evans, D., Ginther-Webster, K., Hard, M., Lefferts, R. and Monarch, I.(1991). Automatic Indexing Using Selective NLP and First Order Thesauri. In Proc. RIAO '91, Universitat Autonoma de Barcelona, Spain.

[10] Krovetz, R. and Croft, B.(1992). Lexical Ambiguity and Information Retrieval. ACM Transactions on Information Systems, 10 (2).

[11] Lalmas, M.(1997). Dempster-Shafer's Theory of Evidence Applied to Structured Documents : Modelling Uncertainty. In Proc. of the ACM SIGIR '97, Philadelphia, PA, USA.

[12] Lee, Y. K., Yoo, S. J., Yoon, K. and Berra, B.(1996). Index Structures for Structured Documents. In Proc. Digital Library '96.

[13] Light, R.(1997). "Presenting XML"

Sams.net Publishing.

[14] Myaeng, S. H., Jang, D.-H., Kim, M.-S., and Zhoo, Z.-C.(1998). A Flexible Model for Retrieval of Structured Documents. In Proc. of ACM SIGIR '98, Melbourne, Australia.

[15] Myaeng, S. H. and Liddy, E. D. Information Retrieval with Semantic Representation of Texts. In Proc. the 2nd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, USA.

[16] Navaro, G. and Baeza-Yates, R. (1995). A Language for Queries on Structure and Contents of Textual Databases. In Proc. of ACM SIGIR '95, Seattle, WA, USA.

[17] Newcomb, S., Kipp, N., and Newcomb, V.(1991). The "HyTime" Hypermedia/Time-Based Document Structuring Language. Communications of the ACM, 34 (11).

[18] Persin, M., Zobel, J., and Sacks-Davis, R.(1995). Filtered Document Retrieval with Frequency-Sorted Indexes." Journal of American Society for Information Science.

[19] Salton, G., Fox., E., and Wu, H.(1983). Extended Boolean Information Retrieval. Communications of the ACM, 26 (12).

[20] Sanderson, M.(1994). Word Sense Disambiguation and Information Retrieval. In Proc. of 17th ACM SIGIR Conference, Dublin, Ireland.

[21] Savoy, J.(1996). An Extended Vector-Processing Scheme for Searching Information in Hypertext Systems. Information Processing and Management, 32 (2).

[22] SGML/XML '97 Conference Proceedings, December 8-11, 1997, Sheraton Washington Hotel, Washington, D. C.

[23] Shin, D. W., Nam, S. J., Jang, H. L., Jin, H. L. and Zhoo, Z. C.(1997). Bottom-Up Query Evaluation of Structured Documents. In Proc. of NLPRS '97, Phuket, Thailand.

[24] Spark Jones, K.(1990). What exactly should we look to AI, and NLP especially, for? In Working Notes for AAAI Spring Symposium on Text Based Intelligent Systems, Stanford, CA, USA.

[25] Strzalkowski, T.(1995). Natural Language Information Retrieval. Information Processing and Management, 31 (3).

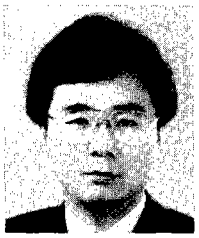
[26] Travid, Brian E. and Waldt, Dale C. (1995). The SGML Implementation Guide, Springer-Verlag.

[27] Turtle, H. and Croft, B.(1991). Evaluation of an Inference Network Based Retrieval Model. ACM Transactions on Information Systems, 9 (3).

[28] van Rijsbergen, C.(1986). A Non-Classical Logic for Information Retrieval. The Computer Journal, 29 (6).

[29] Volz, M., Aberer, K., and Bohm, K. (1996). Applying a Flexible OODBMS-IRS-Coupling to Structured Document Handling. In Proc. of the 12th International Conference on Data Engineering, New Orleans, USA.

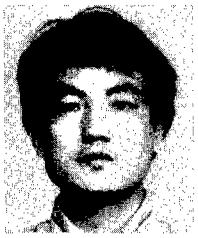
[30] Wilkinson, R (1994). Effective Retrieval of Structured Documents. In Proc. of ACM SIGIR '94, Dublin, Ireland.



맹 성 현

1983 미국 California 주립대 (B. S.)
1985 Southern Methodist University(M. S.)
1987 Southern Methodist University(Ph. D.)
1987～1988 미국 Temple University 교수
1988～1994 미국 Syracuse University 교수
1994～현재 충남대학교 컴퓨터과학과 교수

관심분야 : 정보검색, 자연어처리, 인간과 컴퓨터 상호작용, 디지탈도서관
E-mail:shmyaeng@cs.chungnam.ac.kr



주 종 철

1982 한양대학교 학사
1984 한양대학교 석사
1984～1985 Goldstar Honeywell Inc.
1985～1998 시스템공학연구소 정보검색연구실장
1991 미국 오하이오 주립대 석사/박사수료
1998～현재 ETRI 컴퓨터·소프트웨어기술연구소 문서정보팀장

관심분야 : SGML/XML, 정보검색 및 관리, Document-Based HCI, Mental Models, 인지시스템공학
E-mail:zczhoo@seri.re.kr

● '98 정보통신 하계워크샵 ●

- 일　　자 : 1998년 8월 20일(목)～21일(금)
- 장　　소 : 온양그랜드호텔
- 주　　최 : 정보통신연구회
- 문 의 처 : 충남대학교 컴퓨터공학과 최 훈 교수
  Tel. 042-821-6652
  E-mail : hchoi@comeng.chungnam.ac.kr