확장된 질의 처리를 위해 경로간 의미적 유사도를 고려한 XML 문서 순위화 기법

(A Ranking Technique of XML Documents using Path Similarity for Expanded Query Processing)

김 현 주 † 박 소 미 †† 박 석 †††
(Hyun Joo Kim) (Somi Park) (Seog Park)

요 약 정보기술의 표준으로 사용되고 있는 XML환경에서 방대한 양의 데이터에 대한 사용자의 질의를 효율적이고 정확하게 처리하기 위한 연구가 이슈화되고, 특히 웹 환경에서의 XML문서들은 용어적, 구조적인 측면에서 다양한 형태로 존재하고 있다. 이러한 특성을 갖는 XML 문서들을 대상으로 사용자가 특정한 정보를 얻고자 한다면, 사용자의 질의가 가진 용어 및 구조적 특성과 정확히 일치하지 않는 문서의 정보에 대해서 추가적인 기법이 필요하다. 본 논문은 이와 같은 경우에도 동일한 용어 및 구조를 사용하던 환경에서와 마찬가지로 최상위 순위로 정보를 검색할 수 있는 기법을 제시한다. 또한 정확히 일치하지 않는 문서의 경우에 대해서도 사용자 질의 측과의 경로간 의미적 유사성을 측정하여 사용자 질의와 의미적으로 유사한 경로를 가진 순으로 문서들을 순위화하여 제공한다. 제안된 기법은 실험을 통하여 기존의 기법보다 세밀하고 정확한 검색 결과를 도출함을 보인다.

키워드 : 정보검색 시스템, XML 문서 검색, 검색 순위화, 웹 데이터베이스

Abstract XML is broadly using for data storing and processing. XML is specified its structural characteristic and user can query with XPath when information from data document is needed. XPath query can process when the tern and structure of document and query is matched with each other. However, nowadays there are lots of data documents which are made by using different terminology and structure therefore user can not know the exact idea of target data. In fact, there are many possibilities that target data document has information which user is find or a similar ones. Accordingly user query should be processed when their term usage or structural characteristic is slightly different with data document. In order to do that we suggest a XML document ranking method based on path similarity. The method can measure a semantic similarity between user query and data document using three steps which are position, node and relaxation factors.

Key words : Information Retrieval System, XML Document Searching, Ranking System, Web Database

· 본 연구는 한국과학재단 세계수준의 연구중심대학(WCU) 육성사업(R33-2008-000-10110-0) 지원으로 수행되었음

† 학생회원 : 삼성전자 정보통신총괄 연구원
lastemperor99@gmail.com
†† 학생회원 : 서강대학교 컴퓨터공학과
thathal@sogang.ac.kr
††† 종신회원 : 서강대학교 컴퓨터공학과 교수
spark@sogang.ac.kr

논문접수 : 2009년 7월 27일
심사완료 : 2009년 11월 29일

정보과학회논문지: 데이타베이스 제37권 제2호(2010.4)

1. 서 론

정보기술의 표준으로 사용되고 있는 XML환경에서 사용자 질의 처리에 대한 연구는 꾸준히 진행되어 왔으나 최근의 Web환경에서의 XML문서들은 서로 각기 다른 종류의 정보를 지니고 있다. 특히 문서 작성자의 의도에 따라 서로 다른 구조와 용어를 이용해 작성되므로 이에 대한 정보가 부족한 사용자는 고전적인 질의 처리 방식으로는 처리되지 못하는 형태의 질의를 작성할 가능성이 높아졌다. 바람직한 정보 제공을 위해서는 사용자가 문서 측 정보에 대해 무지(無知)할 경우 발생하는 "정확한 정보를 찾아 제공"의 어려움을 해결하는 것 이

외에 ① 문서 내에 사용자의 질의와 정확히 일치하는 정보 자체가 없을 경우와 ② 사용자 질의가 자신이 원하는 바를 명확히 표현하고 있지 못할 경우에 대한 고려가 필요하다.

기존의 데이터베이스 관리시스템(Database Management System; DBMS)에서는 사용자가 원하는 바를 질의로 정확히 표현하지 못하는 경우에 대해 결과를 도출하지 못하였다. 반면 정보검색시스템(Information Retrieval System; IRS)에서는 수집된 자료를 분석 및 가공하여 축적해놓고 사용자의 질의를 분석하여 색인화된 문서로부터 원하는 정보를 검색해준다. 이렇게 검색된 정보들을 순위화(Ranking)하여 사용자가 원하는 정보에 대하여 더욱 쉽게 접근할 수 있도록 용이성을 높여준다.

XML 문서 환경의 고전적인 질의처리 방식 중 가장 큰 문제점인 용어와 구조에 대한 절대적인 비교를 극복하기 위하여 질의의 재작성과 처리에 관한 기법들[1-3]이 제안되었다. XML문서와 사용자 질의간의 상이한 용어 및 구조 형태를 각각 극복할 수 있는 기법들은 제시되었으나 동시에 효율적으로 극복할 수 있는 기법은 아직까지 미흡하다.

본 연구에서는 사용자 질의를 처리하는 기준을 확장시켜 문서와 정확히 일치하는 구조 및 용어를 사용하지 않더라도 가급적 사용자가 얻고자 하는 정보를 유사하게 제공 할 수 있는 문서에 대해서 사용자 질의 측과의 경로간 의미적 유사성에 기반한 순위를 결정한다. 그리고 사용자가 다수의 문서 중 자신이 작성한 질의와 정확하게 구조적 용어적으로 일치하는 정보를 가지고 있는 문서를 최상위(Top-k) 결과로 제공 받게 되고, 이후의 문서들에 대해서는 질의와 의미적으로 유사한 정보를 가진 순으로 순위화 된 문서에 대한 정보를 제공 함으로써 확장된 질의 처리를 가능케 한다.

2. 관련연구

2.1 정보 검색과 순위화

정보검색(Information Retrieval; IR)은 앞서 수집된 정보 또는 자료의 내용을 분석한 뒤 적절히 가공하여 축적해 놓은 정보파일로부터 사용자의 요구에 적합한 결과를 탐색하여 찾아내는데 의의가 있다. 이러한 정보검색시스템과 기존의 데이터베이스 관리시스템 사이의 유사점과 차이점을 간단하게 분석하면 표 1과 같다. 정보검색은 질의(Query) 형태로 표현되는데 사용자의 질의어를 분석하여 질의의 단어와 구조 등을 변경함으로써 문서를 분석한다. 이에 관련된 기술을 질의어 확장(Query Expansion)이라 한다. 검색된 문서는 색인화 과정을 거쳐 속성집합 형태로 표현된다.

질의어와 문서가 분석되고 나면 이를 비교해서 관련

표 1 정보검색 시스템(IRS)과 데이터베이스 관리시스템(DBMS)의 비교

	차이점			유사점	
	다루는 개체	주요 문제	검색 특성		
정보검색 시스템 (Information Retrieval System)	텍스트 형식의 문서	문서의 내용을 분석하여 효율적으로 핵심어를 추출하는것	확률적	대량의 문서를 대상으로 하는 정보 시스템	문서의 첨가 변경 제기 등을 처리하는 데이터 구조와 알고리즘이 요구됨
데이터베이스 관리시스템 (Database Mnagement System)	레코드 (구조적 형태의 정보)	분석된 문서에 대해 정보를 효율적으로 저장하는 것	결정적		

성(Relevance)이 높은 문서 순으로 정렬해야 하는데 이를 위해 각 문서에 점수를 할당한다. 이때 사용되는 수식이 검색 모델이며 정렬된 문서는 순위화되었다고 할 수 있다. 정보검색의 중요측정기준은 재현률(Recall)과 정확률 혹은 정확도(Precision)이다. 재현률은 검색과 관계되는 문서 전체 중에 몇 개를 찾아내느냐를 보는 것이며, 정확성은 검색결과 중에서 상위 몇 위까지 중 질의와 관계되는 문서가 몇 개인가를 보는 것이다. 이것은 결과의 "정확도"를 측정 하는 자료로서 검색 엔진의 "순위화"가 얼마나 잘 되어 있는가 측정하는 자료라고도 볼 수 있다. 따라서 문서 모음의 크기가 커질수록 더 높은 정확률을 가진 검색엔진이 필요하게 된다. 기본적으로 문서를 순위화하는 까닭은 사용자가 원하는 정보를 가급적 정확하거나 유사한 형태로 제공해 주기 위함이다. 만일 다수의 데이터 중 사용자가 원하는 바와 정확히 일치하는 정보가 있다면 그것을 제공해 주면 되겠지만 전체 데이터에 비해 정확한 정보의 양이 너무 적거나, 존재하지 않는 경우, 또 정확하지 않지만 그와 유사한 정보로써 사용자에게 도움이 될 수 있는 가능성이 있는 정보를 제공하기 위해서는 사용자의 요구에 따라 문서들이 바람직한 순으로 고려되고 분류되어야 한다.

2.2 XML 환경

XML[6]은 W3C XML 1.0권고안에 언급되어 있는 문법을 잘 지켜서 작성된 문서를 적격문서(Well-Formed XML Document)라 하고 여기에 XML로 개발된 특정 마크업 언어에 맞게 작성된 문서를 유효한 문서(Valid XML)라 한다. XML은 어떤 구조로 만들고 어떤 엘리먼트(Element)를 집어 넣을 수 있는지를 정의하는 DTD (Document Type Definition)를 이용해 XML문서의 구조와 사용하는 엘리먼트를 작성자가 직접 정의함으로써 새로운 마크업 언어를 개발 할 수 있다. 일반적으로 특정 마크업 언어로 작성된 문서, 즉 특정 DTD를 기준으로 만들어진 문서를 그 마크업 언어에 대해 유효(Valid)

한 문서라고 부른다. 본 연구에서는 XML 문서의 정보 내용에 초점을 두지 않고 그 정보를 분류하는 기준인 노드 엘리먼트에 초점을 맞추고 있다. 또한 XML문서에서 사용하는 태그 노드 엘리먼트의 종류와 그 상하 관계가 명확해야 하므로 본 논문에서는 특정 DTD기반의 유효한 문서만을 고려 대상으로 한다.

XML구조는 계층적(Hierarchical)인 구조를 가지고 있으며 트리(Tree) 형태로 표현 될 수 있기 때문에, 사용자가 XML문서로부터 필요한 정보를 얻기 위해서는 XPath 질의를 작성해 사용할 수 있다. XPath는 XML의 계층구조로부터 데이터를 쉽게 참조할 수 있도록 설계되어 있다. XML문서 상에서 원하는 정보를 쉽게 찾기 위해서는 문서의 각 노드에 인덱스가 할당되어 있어야 하는데 이러한 방법을 레이블링(Labeling) 기법이라 한다.

XML레이블링 기법은 데이터의 계층적인 구조 정보를 유지함으로써 XML데이터의 저장과 질의 처리에 상당히 효과적인 성능을 보이고 있다. XML문서의 구조에 대한 질의를 가능하게 하기 위해 구조적 특성을 고려하여 순서정보를 유지하기 위한 대표적인 레이블링 기법에는 그림 1에서 보여주고 있는 듀이 레이블링 기법(Dewey Order Labeling)[4]이 있다. 본 연구의 핵심은 문서의 레이블링이 아니므로 정적인 데이터를 대상으로 노드 간의 부모/자식, 조상/자손, 형제 관계의 파악이 용이하여 경로를 도출해 내기 쉬운 듀이 레이블링 기법을 채택하여 사용한다.

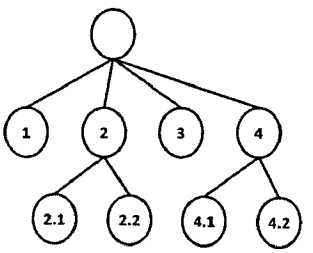


그림 1 듀이 레이블링 기법[4]

2.3 XML 문서 검색 기법

최근의 웹 환경에서 XML문서는 다양한 내용을 표현하고 저장하기 위한 수단으로 사용되므로 각 엘리먼트를 표현하는 용어가 문서 작성자에 의해 다양하게 정의될 수 있으며, 문서 내 정보의 계층적 관계를 나타내는 구조 또한 다양할 수 밖에 없다. 사용자는 원하는 정보를 얻기 위해 XML문서 측에 XPath[7] 질의를 작성하여 보내게 되는데 웹 상의 XML문서가 서로 다른 용어적, 구조적 특성을 가지고 있어서 효율적인 질의처리를 하는데 문제로 작용한다. 고전적인 데이터베이스(Database)의 개념으로 수행하는 질의 처리(Query Process)의 경우는 사용자 질의가 문서 측에서 처리할 수 있는 형태이기 위해서 사용자 질의를 구성하는 모든 엘리먼트 노드들이 문서 측에 동일한 용어로 표현되어 존재하여야 하며 그 나열 순서 또한 질의 측과 문서 측이 동일해야만 해당 질의는 처리가 가능하다고 판단된다. 말하자면 사용자의 질의가 처리 될 수 있는가 없는가에 대한 고려는 크게 ① 엘리먼트 노드의 유무(有無)에 대한 용어적 비교(Ontology)와 ②엘리먼트 노드의 나열에 대한 구조적 비교(Structure, Ontology)로 이루어 지는데, 이 중 한 가지라도 문서 측과 다를 경우 해당 질의는 거절되게 된다.

고전적인 질의 처리 방식에서의 사용자 질의를 처리하는 기준이 가지는 가장 큰 문제점은 용어와 구조에 대한 단순하고 절대적인 비교로 인하여 자원의 낭비가 발생한다는 점이다. 이러한 질의 처리 방식이 가지는 문제점에 대한 인식으로 이를 극복하려는 다양한 연구들이 진행되어 왔는데, 이들 연구들은 크게 질의 재작성과 처리에 관한, ① XML문서와 사용자 질의 간의 상이(相異)한 용어(Semantic) 사용을 극복할 수 있는 방안에 대한 연구, ② XML문서와 사용자 질의 간의 상이(相異)한 구조(Structure) 형태를 극복 할 수 있는 연구로 분류 할 수 있다.

용어차이와 구조차이를 동시에 고려한 연구로써 [1]은 사용자의 질의와 문서 측의 구조와 용어가 모두 다를 때 질의와 문서 양측에서 사용한 용어의 차이를 고려하고 질의를 처리하는 알고리즘을 제공한다는 점에서 의미를 가진다고 할 수 있다. 그러나 용어 차이에 대한 고려를 단순히 기존의 워드넷(WordNet)[5]엔진을 통한 검색에 의존하는 한계점을 가지고 있고, 구조적인 차이에 관련하여 양측 노드 순서의 차이나 생략에 대한 처리 부분이 고려되지 않는 한계점을 가진다.

질의간의 상이한 구조 형태에 대한 연구인 [2]은 Flexible과 Semiflexible이라는 두 가지 알고리즘을 이용하여 XML문서와 질의 사이의 구조 차이를 비교, 질의를 문서 측 구조로 변환하여 처리하는 방법을 제시하였다. 이 기법은 질의와 문서 구조에 사용된 노드의 순서가 뒤바뀐 경우의 질의를 처리해 줄 수 있지만, 질의 측에서 사용된 노드가 문서 측에서 생략된 경우나 상이한 용어를 사용한 경우는 적절히 처리해 줄 수 없는 문제점을 가진다. [3]의 경우도 질의와 문서 구조차이에 대한 연구로써 질의를 재작성하기 위한 질의 이완의 4가지 룰을 제공한다. 그러나 구조적인 차이만 극복할 뿐 상이한 용어에 대한 고려를 하지 않는다. [3]에서 제시하는 규칙은 본 연구에서 차용되며, 자세한 설명은 3.4절에서 다루도록 한다.

최근의 연구들은 앞서 언급한 바와 같이 [3]와 [1]의 경우같이 용어 차이에 대한 고려가 부족할 뿐만 아니라, [1]와 [2]의 경우와 같이 구조 차이에 대한 고려도 미흡하다. 이들 연구들이 갖는 가장 큰 취약점은 실제로 사용자 질의가 효율적으로 정보의 낭비 없이 처리되기 위해서 이들 두 가지에 대한 고려가 함께 이루어져야 함에도 각각 따로 고려하거나, 동시에 고려한다고 해도 그 효율성이 떨어진다는 점이다. 또한 이들 기법들은 사용자 질의와 정확히 일치하는 정보만을 찾기 위한 목표를 갖고 있으므로, 문서상에 사용자 질의와 정확히 일치하는 정보가 없거나, 사용자 질의가 원하는 바를 충분히 표현하고 있지 못할 경우에 대한 고려가 미흡하다.

3. 제안기법

3.1 시스템 구조

사용자 질의와 정확하게 일치하는 정보는 물론 유사한 정보까지 찾아 사용자에게 제공하기 위한 제안 기법의 전체적인 흐름은 그림 2와 같다. 사용자는 특정한 정보를 얻기 위해 질의 하고 결과적으로 사용자 질의와 의미적으로 가까운 순서로 순위화된 문서를 제공 받을 수 있다. 단일 문서에 대해 유의어 처리를 통해 질의와 공통 노드들을 기반으로 경로들을 도출하고 도출된 경로와 질의간의 의미적인 유사도 수치를 측정한다. 의미적 유사도 수치는 상대적 위치가 같은 경로간의 공통 노드들의 점수(Position)과 상대적 위치가 다른 경로간의 공통 노드들의 점수(Node), 마지막으로 질의이완 점수(Relaxation)의 점수들로 계산되고 가장 높은 유사도 수치를 갖는 경로가 해당 단일 문서의 대표 경로가 된다. 다수의 문서들은 대표 경로가 갖는 유사도의 수치에 따라 순위화가 되고 이 결과가 사용자에게 전달된다.

3.2 유의어 처리

경로간 의미적 유사도 측정에 앞서 사용자 질의 측과 문서 측의 공통 노드에 기반한 경로 도출에 대해 설명하면 다음과 같다. 사용자 질의와 문서 측에서 도출된 경로간의 의미적으로 유사한 정도를 측정 하려면 몇 가지 기준이 사용되는데 이중 가장 선행 되어야 하는 부분이 질의 측과 문서 측의 상이한 용어 사용에 대한 처리이다. 사용자의 질의가 가지고 있는 엘리먼트 노드들에 대해 동의어, 유의어 관계에 있으면서 문서 측에 존재하는 DTD 상의 엘리먼트 노드들이 있다면 이들 노드로 구성된 문서 측의 경로는 사용자 질의와 의미적으로 연관이 있을 가능성이 높다.

이를테면, "//CS//Conference//Title"이라는 사용자 질의가 특정 문서에 대해 발생되었을 때, 우선 "CS", "Conference", "Title"이라는 세 개의 노드가 문서 측에도 존재하는지 살펴보고, 일부, 혹은 전부가 발견되지 않았을 때 발견되지 않은 이들 노드들과 동의어 유의어 관계를 가지는 용어가 문서의 DTD에도 존재하는지 찾아본다. 이때 사용되는 것이 유의어 사전 테이블인데, 유의어 사전(Thesaurus)는 검색성능을 향상시키기 위한 단어의 동의 관계, 계층 관계, 연관 관계를 구분하는 통제어휘집이라 볼 수 있다. 이를 이용해 우리는 질의 처리의 첫 번째 과정에 유의어 사전에 기반한 용어/유의어 테이블을 생성하여, 이후의 과정에서 두 단어 사이의 유의어, 동의어 관계를 고려해야 할 경우 이를 사용한다.

예를 들어, 그림 3과 같이 문서상에 "CS"와 "Conference"라는 노드는 존재 하나 "Title"이라는 노드는 존재하지 않고, 유의어 사전 테이블에서 "Title"와 동의어 유의어 관계인 다수의 용어들을 확인한 결과 "Subject"가 유의어라고 하자. 따라서 이와 같은 유의어 관계의 노드가 문서 상에 존재 할 경우 이후 단계의 작업에서는 사용자 질의에서의 모든 "Title" 노드를 "Subject"로 대체하여 고려한다. 유의어 관계를 갖는 사용자 질의 측의 노드와 문서 구조 측의 노드는 공통노드로 간주하여 공통 노드를 포함하는 모든 가능한 경로를 문서로부터 도출해낸다.

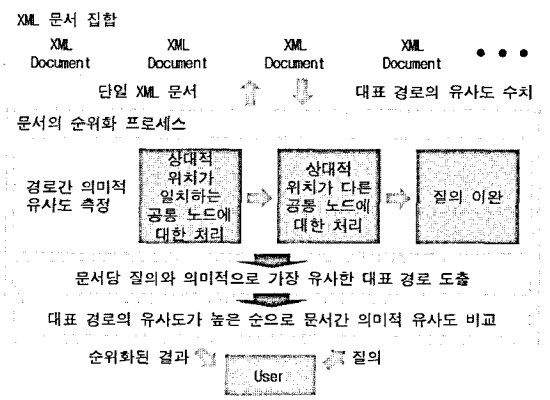


그림 2 시스템 구조

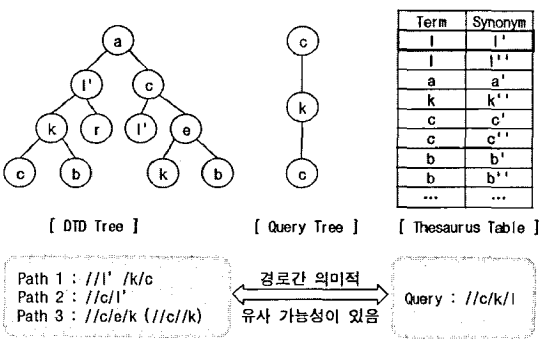


그림 3 유의어 처리

3.3 경로 유사도 측정

3.3.1 상대적 위치가 같은 경로간의 공통 노드의 처리 (Position)

사용자 질의와 문서 측에서 도출된 경로 사이의 의미적 유사도는 경로 상의 노드들의 상대적 위치를 비교하여 측정 할 수 있다. 실질적으로 질의와 문서에서 도출된 경로는 서로 다른 노드들로 구성되어 있을 수 있으나 경로간의 의미적 유사도를 측정하는 데는 경로를 구성하는 노드 중 양측이 공통적으로 가지고 있는 노드에 대해서만 고려 하는 것이 바람직하다. 양측이 공통적으로 가지고 있는 노드 중 가장 중요한 위치를 차지하는 노드는 상대적인 위치가 동일한 노드인데 따라서 경로간 의미적 유사도를 측정 할 때 가장 우선적으로 고려되어야 하는 사항이라 생각할 수 있다. 의미의 기준이 되는 경로는 사용자 질의로써 XPath 질의로 작성된 사용자 질의가 의미 하는 바를 기준으로 한다.

사용자 질의가 n개의 노드로 구성되어 있을 때, 단말 노드부터 루트 노드에 이르는 순서대로 경로 전체의 의미에 미치는 영향에 따라 각각 n점 ~ 1점에 해당하는 가중치를 가지고 있다. 이는 질의의 단말 노드가 사용자가 원하는 정보를 가장 근접하게 표현하고 있다고 보기 때문이다. 경로간 유사도 측정의 첫 번째 단계에서 특정 경로가 기준 경로에 존재하는 n개의 노드 중 k($1 \le k \le n$)개를 가지고 있다면, 그림 4과 같이 k개의 노드들이 기준 경로와 비교해 1 ~ k에 이르는 동일한 위치에 존재하는지 확인하고, 동일한 위치에 존재한다면 해당 노드가 가지고 있는 가중치를 경로 전체의 유사도를 측정하는데 적용한다.

기준 경로	//a/b/c/d/e/f/g/h/i/j
노드의 가중절수	1 2 3 4 5 6 7 8 9 10
경로1	//c/e/k/f/x/b (//c/e//f//b)
경로2	//b/f/h/l
경로3	//h/b/g/y/z/a (//h/b/g//a)
경로4	//j/g/d

기준 경로와 네 개의 경로를 노드의 상대적 위치를 기준으로 비교하면 경로2-경로3-경로4-경로1 순으로 의미적으로 유사함을 알 수 있다.

(a) //a/b/c/d/e/f/g/h/i/j
//c/e/k/f/x/b
b c e f
c e f b
기준 경로 측의 네 노드 b, c, e, f의 순서와 경로 1의 세 노드 c, e, f, b의 상대적 위치가 모두 다르므로 0점

(b) //a/b/c/d/e/f/g/h/i/j
//b/f/h/l
b f h i
b f h i
기준 경로 측의 네 노드 b, f, h, i의 순서와 경로 2의 네 노드 b, f, h, i의 상대적 위치가 모두 같으므로 2+6+8+9 = 25점

(c) //a/b/c/d/e/f/g/h/i/j
//h/b/g/y/z/a
a b g h
h b g a
기준 경로 측의 네 노드 a, b, g, h의 순서와 경로 3의 네 노드 h, b, g, a의 상대적 위치 중 b와 g만 같으므로 2+7 = 9점

(d) //a/b/c/d/e/f/g/h/i/j
//j/g/d
d g j
j g d
기준 경로 측의 세 노드 d, g, j,의 순서와 경로 4의 세 노드 j, g, d의 상대적 위치 중 g만 같으므로 7점

그림 4 Position Factor의 계산

그림 4의 (d)는 사용자 질의 측과 경로 4 측의 공통 노드가 "d", "g", "j" 세 개인 경우 양측의 세 노드의 상대적 위치를 비교 하면 "g" 노드가 양측 모두에서 두 번째 위치에 존재함을 알 수 있고 따라서 경로 4는 기준 질의와 "g" 노드의 가중 점수인 7 만큼의 유사도를 가지게 된다. 그러나 이러한 양측 경로 간의 공통 노드에 대한 상대적 위치 고려만으로는 정확한 경로간 의미적 유사도를 판단하기 어렵다. 실제 사용자 질의를 기준으로 문서 측에서 도출된 경로를 비교할 때 의미적으로 가장 큰 영향을 미치는 노드는 "사용자 질의 측과 공통적이면서 동일한 상대적 위치를 유지하는 노드"이지만 이 경우 동일한 유사도를 가지는 다수의 경로들이 발생하게 된다. 경로의 의미를 결정하는 가장 중요한 요소는 "제 위치에 있는 노드"이지만 제 위치에는 없으나 경로상에 존재함으로써 경로가 가지는 의미에 영향을 줄 수 있다는 점에 대해 처리해 줌으로서 경로 유사도 측정에 정밀을 더할 수 있다.

3.3.2 상대적 위치가 다른 경로간 공통 노드의 처리 (Node)

특정 경로를 기준으로 그와 다른 경로 간에 의미적으로 유사한 정도를 가늠하기 위해서는 앞서 언급한 양측 경로간 상대적 위치가 일치하는 노드들에 대한 고려 이외에 상대적인 위치는 다르지만 공통으로 존재 하고 있는 노드에 대한 고려를 하는 기준이 적용되어야 한다. 기준이 되는 경로가 가지고 있는 노드와 동일한 노드를 최대한 많이 가지고 있는 경로가 그렇지 않은, 말하자면 몇몇 개의 노드가 누락된 경로보다 기준 경로와 비교하여 의미적으로 더 유사할 가능성이 높다고 할 수 있다. 그러나 단순히 기준 경로가 가지는 노드의 존재 유, 무만을 가지고 경로가 가지는 의미의 유사도를 판단하기는 어렵다. 첫 번째 이유는 모든 노드들이 경로 전체의 의미를 나타내는데 있어 동일한 수준의 역할을 하고 있지 않다는 것이다. 일반적으로 경로가 가지는 의미의 가장 주요한 부분은 단말 노드에 있고, 루트 노드에 가까울수록 경로가 가지는 의미에는 적은 영향을 미치기 때문이다. 그러나 단순히 노드간의 의미적 중요도에 따른 가중치를 고려한다고 해도 경로간 유사도 측정에는 미흡함이 있다. 두 번째 이유는 기준 경로와 동일한 노드를 많이 가지고 있는 경로라 할지라도 그 순서가 마구 뒤섞여 기준 경로에서 노드들이 가지는 순서와는 상이할 경우, 동일 노드를 많이 가지고 있지는 않지만 기준 경로에서 노드들이 가지는 순서와 동일, 혹은 유사한 순서로 노드들이 배열된 경로에 비해 의미적으로 유사하다고 보기에는 무리가 있기 때문이다. 따라서 양측 경로간의 공통 노드에 대해 상대적으로 동일한 위치에 존재하고 있는 노드들에 대한 고려를 우선적으로 한 뒤, 이

단계에서 동일한 유사도 수치가 측정된 경로의 경우만 상대적으로 다른 위치에 존재하고 있는 나머지 공통 노드들에 대해서도 상대적 위치가 같은 경로간의 공통 노드의 처리(Position Factor)와 동일한 방식으로 고려해 주는 것이 바람직하다.

3.3.3 경로의 의미적 유사도

기준 경로에 대해 특정 경로들이 어느 정도의 의미적 유사성을 가지고 있는지를 판단하기 위해서 그림 4와 같은 방식으로 먼저 노드들 사이의 상대적 위치를 고려한 후 여기서 측정된 값이 동일한 경로만을 대상으로 나머지 공통 노드들을 고려한다. 따라서 이들 두 가지 기준은 순차적으로 적용되며 항상 첫 번째 기준을 우선시 한다. 노드들 사이의 상대적 위치에 대한 측정 값이 동일한 경로들의 경우는 나머지 공통 노드들을 많이 가지거나 질의 측의 단말 노드에 가까운 노드를 가질수록 기준 경로와 의미적으로 더 유사한 경로로 판별된다. 이를 도식화 하면 그림 5와 같다.

기준경로 //a/b/c/d/e
노드가중점수 1 2 3 4 5

		Position	Node
경로1	//c/b/a/e//d	2 (b)	15 (3+2+1+5+4)
경로2	//c/b/a	2 (b)	6 (3+2+1)
경로3	//c//a/e//d	0	13 (3+1+5+4)
경로4	//c//a/e/b	0	11 (3+1+5+2)
경로5	//c//d	7 (c,d)	7 (3+4)
경로6	//b/d	6 (b,d)	6 (2+4)
경로7	//b	2 (b)	2 (2)
경로8	//c	3 (c)	3 (3)
경로9	//c/a	0	4 (3+1)

경로 유사도 순위 (Ranking)

		P	N
1	//c//d	7	
2	//b/d	6	
3	//c//d	3	
4	//c/b/a/e//d		15
5	//c/b/a	2	6
6	//b		2
7	//c//a/e//d		13
8	//c//a/e/b	0	11
9	//c/a		4

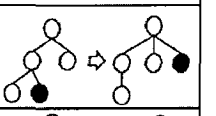
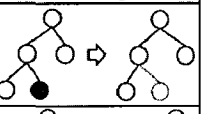
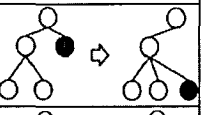
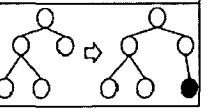
그림 5 경로간 의미적 유사도 판단

3.4 질의 이완

XPath로 작성된 사용자 질의를 문서 측에서 처리할 수 있는 형태로 바꾸어 주기 위해서 적절하게 변경 시키거나 재작성을 한다. [3]에서 제시하는 기법은 XPath 경로를 이완해 주는 것으로, XML 경로의 트리 구조를 변경 함으로써 사용자 질의를 해당 문서에서 처리하기에 적합한 형태로 변경해 줄 수 있다. [3]에서는 총 4가지 기법을 제시하고 있으며 이들의 기능과 특징을 정리하면 표 2와 같다.

여기서 사용하는 기본적인 네 가지 질의 이완 기법을 사용하여 본 연구 에서는 이 이완 기법을 ① 이전 단계까지 동일한 의미적 유사도 수치를 가진 경로들 간의 우열(偶劣)을 가리기 위한 수단으로 사용하고, ② 사용자

표 2 질의 이완의 4가지 방법

기법	설명	Tree상 변화
Subtree Promotion (SP)	특정 노드가 트리 레벨상 한 단계 위로 상승하면서 그에 자손 노드를 포함한 서브트리 전체가 트리 구조 상에서 한단계 위로 위치변경된다.	
Leafnode Deletion (LD)	트리 구조 상의 단말 노드가 트리상에서 삭제된다.	
Edge Generation (EG)	특정 노드가 트리 레벨상 한 단계 아래로 하강하면서 그에 자손 노드를 포함한 서브트리 전체가 트리 구조상에서 한 단계 아래로 위치변경된다.	
Node Generation (NG)	기존트리 구조상에 존재하지 않았던 새로운 노드가 생성된다.	

질의를 문서당 하나씩 도출된 의미적으로 가장 유사한 경로로 변형하여 처리하기 위한 수단으로 사용한다.

4. 실험 및 성능 평가

4.1 실험 환경

4.1.1 측정 기준

③ 수행 시간 비교

① 순위의 정밀도 측정(순위 단계)

√ 100개의 문서 집합을 순위화 할 때 모든 문서들이 총 몇 단계로 순위화 되는지에 대한 측정.

② 순위의 정확도 측정 (Precision계산)

√ 총 100개의 문서 집합을 대상

- auctions.dtd와 49가지의 구조적 variation들, resume.dtd와 49가지의 구조적 variation들.

√ 질의와 관련 있는 문서의 집합

- 50 개 (k=20, k=50)
- 20 개 (k=20)

4.1.2 용어 정의

정의 1. 질의와 관련 있는 DTD 어떤 DTD에서 사용하고 있는 노드들 중 일부를 사용하여 작성된 질의가 존재 할 때 해당 DTD는 질의와 관련 있는 DTD이다.

정의 2. 관련 노드 사용자 질의가 특정 DTD와 관련 있게 작성되었을 때 이 DTD측에 존재하는 노드 중 질의 작성시 사용한 노드.

정의 3. 공통 노드 두 DTD 사이에 공통적으로 사용되고 있는 엘리먼트 노드.

정의 4. 제안 기법 경로 노드들의 Position, Node, 질의 이완 단계를 모두 거쳐 순위화 한 기법.

정의 5. 제안 기법-질의 이완[3] 제안 기법 단계 중 세 번째 단계인 질의 이완을 배제 하고 순위화한 기법.

정의 6. 질의 이완[3] 제안 기법 단계 중 세 번째 단

계인 질의 이완만을 고려하여 순위화 한 기법.

질의 - 질의 1: 공통 노드 1개 관련 노드 2개
(//관련노드/공통노드)
질의 2: 공통 노드 1개 관련 노드 2개
(//공통노드/관련노드)
질의 3: 공통 노드 1 개 관련 노드 3개
(//관련노드/공통노드/관련노드)

4.2 실험 결과 및 분석

수행 속도 면에서는 그림 6에서 보여주는 바와 같이 제안 기법은 질의 이완 기법에 비해 고려하는 조건이 많으므로 약간의 시간을 더 소요하였다. 제안 기법은 세 가지 유사도 판단 단계를 거치는 동안 동점 경로에 대해서만 다음 단계의 유사도 판단을 수행하므로 첫 번째, 두 번째 단계만으로 순위가 확정되는 대부분의 경로에 대해서는 세 가지 단계 중 가장 많은 시간을 소요하는 질의 이완 단계를 제외한 제안-이완 기법의 경우처럼 상당히 적은 시간을 소요하므로 좋은 성능을 보일 것으로 예상할 수 있다.

세 기법 사이에는 순위 단계의 차이가 발생하는데 관련 노드의 수가 증가할수록 그 차이는 커진다. 특히 관련 노드의 수가 3일때, 제안 기법은 100개의 DTD가 23단계로 순위화되었고, 제안-이완 기법은 14단계, 이완 기법은 4단계로 순위화되었다. 그림 7과 같이 제안 기법이 대조군에 비해 더 많은 단계의 순위로 문서가 순위화 되므로 보다 세밀한 순위화를 제공한다.

전체 DTD가 100개일 때 질의와 관련 있는 문서와 관련 없는 문서의 개수를 각각 50:50, 20:80인 두 가지 경우로 놓고, 상위 20, 50개의 결과 중 질의와 관련 있는 문서의 비율을 계산하여 결과 검색의 정확률을 측정

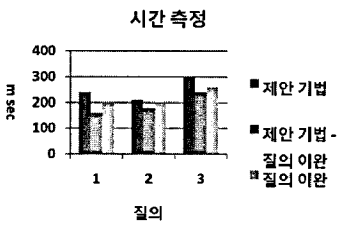


그림 6 수행시간 측정

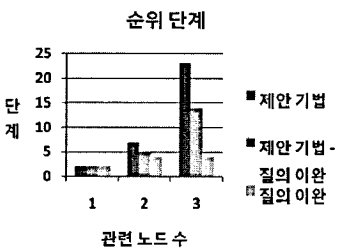


그림 7 순위단계 측정

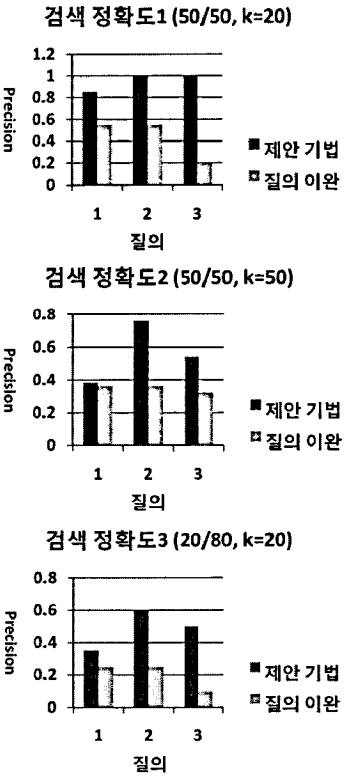


그림 8 검색 정확도 측정(Precision)

하였는데, 그림 8과 같이 검색의 성능 면에서도 모든 경우에 대조군에 비하여 더 우수한 성능을 보이고 있다.

5. 결론 및 추후 연구

본 연구는 다수의 용어적, 구조적으로 다양한 XML문서들을 대상으로 사용자가 특정한 정보를 얻고자 질의했을 때, 사용자의 질의가 가진 용어적 구조적 특성과 정확히 일치하는 문서의 정보에 대해서는 기존의 방식을 사용했을 때와 마찬가지로 최 상위 순위로 정보를 찾을 수 있게 하는 기법을 제시하였다. 정확히 일치하지 않는 문서의 경우는 사용자 질의 측과의 경로간 의미적 유사성을 측정해 사용자 질의와 의미적으로 유사한 경로를 가진 순으로 문서들을 순위화 해 준다. 이 기법은 XML문서가 가지는 특성인 용어적, 구조적 다양성을 고려하여 기존에 처리하지 못했던 사용자 질의와 문서 사이의 용어적 차이와 구조적 차이로 인해 발생했던 질의 처리의 문제를 해결함으로써 사용자 질의를 보다 확장된 개념으로 처리해 주기 위해 고안되었다. 또한 특정 데이터를 저장 관리하고 있는 환경에서의 데이터의 구조적 용어적 변경 상황에 사용자가 어떠한 질의를 작성하여 정보를 요청할 때에도 질의와 의미적 유사도를 측

정하여 질의의 의미와 유사한 정보를 가진 순으로 문서를 순위화 하여 제공할 수 있다.

측정과 검증의 기준은 크게 세 가지로 나누어 측정하였다. 첫 번째는 순위 단계에 대한 측정으로서 본 기법은 대조군보다 세밀한 순위화 단계를 제공하므로 세분화된 문서 순위화를 제공할 수 있다. 두 번째는 순위화 기법의 정확도에 관한 측정으로 본 기법은 대조군 기법보다 더 높은 검색의 정확성을 제공한다. 마지막 측정 기준은 수행 시간에 대한 측정이다. 제안 기법은 대조군의 기법보다 약간 더 많은 시간을 소요하지만 더 좋은 성능을 제공하는 장점을 가지고 있다고 볼 수 있다.

실제 본 연구는 이형적인(Heterogeneous) XML문서에 대해 사용자 질의를 유동적으로 처리 하기 위한 연구로서 웹 환경의 다양한 XML 문서 컨텐츠에 대해 고려 하였다. 특히 본 연구는 멀티미디어 애플리케이션 환경이나 지식 관리 시스템 환경에서 쌓여가는 XML데이터 환경에서의 검색 모듈에 활용 가능한 검색 기법이라 할 수 있다. 본 기법은 특정 도메인의 DTD가 갱신되면서 기존의 DTD 엘리먼트가 삭제, 혹은 새로운 엘리먼트가 생성되었을 때나, 특정 도메인에 익숙한 사용자가 다른 종류의 도메인에서 정보를 찾기 위해 본인이 익숙한 형태의 엘리먼트를 사용하여 질의할 경우 적절히 처리해 줄 수 있는 장점을 가진다. 따라서 추후 연구 과제로는 유전자 정보와 같이 특정 도메인의 정보를 기반으로 변경되고 추가되는 데이터들이 존재하는 환경에서의 문서 순위화 기법에 대한 연구를 생각해 볼 수 있다.

참 고 문 헌

[1] Y. Kanza, Y. Sagiv, "Flexible Queries over Semi-structured Data," *Proc. of 12th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp.40-51, 2001.

[2] C. X. Chen, G. A. Mihaila, S. Padmanabhan, and I. M. Rouvellou, "Query Translation Scheme for Heterogeneous XML Data Sources," *Proc. of 7th annual ACM international workshop on Web information and data management*, pp.31-38, 2005.

[3] S. Amer-Yahia, S. Cho, and D. Srivastava, "Tree Pattern Relaxation," *Proc. 8th International Conference on Extending Database Technology: Advances in Database Technology*, pp.496-513, 2002.

[4] I. Tatarinov, S. D. Viglas, K. Beyer, J. Shanmugasundaram, E. Shekita, and C. Zhang, "Storing and querying ordered XML using a relational database system," *Proc. of the 2002 ACM SIGMOD international conference on Management of data*, pp.204-215, 2002.

[5] WordNet - a Lexical Database for the English Language. http://www.cogsci.princeton.edu/wn/.

[6] toExcel, *Extensible Markup Language (Xml) 1.0 Specifications: From the W3c Recommendations*, iUniverse, Incorporated, 2000.

[7] W3C. XML path language (XPath): Version 2.0. http://www.w3.org/TR/xpath20/.

김 현 주

2007년 2월 서강대학교 컴퓨터공학과 공학사. 2009년 3월 서강대학교 컴퓨터공학과 공학석사. 2009년~현재 삼성전자 정보통신총괄 통신연구소 연구원. 관심분야는 XML, 웹데이타베이스, 정보검색 시스템

박 소 미

2008년 2월 서강대학교 컴퓨터공학과 공학사. 2008년 3월~현재 서강대학교 컴퓨터공학과 공학석사과정. 관심분야는 XML, 웹데이타베이스, 위치정보보호

박 석

1978년 서울대학교 계산통계학과 이학사 1980년 한국과학기술원 전산학과 공학석사. 1983년 한국과학기술원 전산학과 공학박사. 1983년 9월~현재 서강대학교 컴퓨터공학과 교수. 1989년~1991년, 2002년~2003년 미국 버지니아대학교 방문교수. 1997년 2월~현재 한국정보과학회 이사. 2005년, 2008년 한국정보과학회 부회장. 2004년 1월~2005년 12월 한국정보과학회 편집위원장. 1999년~2007년 DASFAA Steering Committee 멤버. 관심분야는 데이터베이스 보안, 실시간 시스템, 트랜잭션 관리, 데이터웨어하우스, 웹 데이터베이스, 프라이버시