웹 환경에서 온톨로지를 이용한 지역정보 융합 시스템

최영수*, 정회윤°, 노성민*, 양형정**
*°전남대학교 전자컴퓨터공학과
**전남대학교 전자컴퓨터공학과, 정보통신연구소
e-mail : {erzzr, fire_night809, archvold}@daum.net, hjyang@jnu.ac.kr

A Local Information Integration System using Ontology on Web

Young-Soo Choi*, Hoi-Yoon Jeong°, Sung-Min Noh*, Hyung-Jeong Yang**
*°Dept. of Computer Science, Chonnam National University
**Dept. of Computer Science, Chonnam National University, Information & Telecommunication Research Institute

● 요 약 ●

방대한 웹 페이지의 홍수 속에서, 의미상 서로 연관되었지만 산재되어 있는 정보들을 사용자에게 효과적으로 제공하기란 그리 쉽지 않다. 웹 페이지에서 질적으로 향상된 정보를 얻기 위해서는, 이질적이지만 서로 연관된 의미를 갖는 데이터들을 하나로 융합하는 방법이 필요하다. 본 논문에서는 이질적인 형태로 이루어진 다수의 웹 페이지들을 XML 기반의 메타데이터(metadata)로 융합하여 사용자에게 제공하는 방법을 제시한다. 본 논문에서 제안한 시스템에서 메타데이터는 온톨로지와 OWL을 이용하여 융합된다. 또한 제시한 방법을 검증하기 위해 지역정보 중 부동산을 사례를 들어 시스템을 구현한다. 구현된 시스템은 각기 다른 데이터를 가지고 있는 다수의 웹 페이지를 하나의 웹 페이지로 통합하는 과정을 거쳐 XML 문서 형태로 사용자에게 제공한다.

키워드: 시맨틱 웹(semantic web), 온톨로지(ontology), OWL, 지역정보 융합(local information integration), 메타 데이터(mata-data)

I. 서론

인터넷은 사용자에게 정보검색의 편리함과 다양한 지식을 접할 수 있는 기회를 제공하는 반면, 인터넷의 보급으로 인한 정보의 폭발적인 증가는 사용자로 하여금 원하는 정보를 얻기 위하여 많은 시간과 노력을 할애하도록 강요하고 있다. 이에 따라 웹 페이지에서 질적으로 향상된 정보를 얻기 위한 새로운 시스템이 필요하다. 이를 해결하기 위한 방법으로 시맨틱 웹은 차세대 웹의 표준으로 주목받고 있다[1].

하지만, 시맨틱 웹으로 제공된 정보라 하더라도 그 규모는 여전히 방대하다. 따라서 사용자가 웹으로부터 효율적으로 정보를 습득할 수 있도록 여러 곳에 흩어져 있지만 서로 연관된 정보를 융합하여 제공하는 방법이 필요하다.

서로 연관된 정보가 하나로 융합되어 사용자에게 편리하게 제공되기 위해서는 다양한 이질의 정보를 기술할 수 있는 메타데이터(metadata)의 도입이 필수적이다. 시맨틱 웹과 메타데이터의 결합은 개발자들에게는 체계적인 문서를 구성하는데 있어서 편리한 롤모델이 된다. 이에 따라 더욱 진보된 검색을 위하여 시맨틱 웹을 이용한 메타데이터의 구성에 관한 연구가 활발히 진행되고 있다[2].

그러나 웹에 존재하는 정보들은 각기 다른 카테고리의 다양한 의미의 언어로 구성되어 있고, 같은 카테고리의 단어라 할지라도 다른 음절을 표기하여 사용하는 경우도 많다. 또한 같은 음절이지만 다른 의미를 갖는 데이터들은 사용자들의 정보검색에 어려움을 준다. 따라서 동음이의어와 이음동의어를 구별할 수 있고, 분산된 다양한 이질의 데이터들을 하나로 통합하여 사용자들에게 제공할 수 있는 에이전트(agent)가 필요하다[3].

본 논문에서는 웹 페이지들을 대상으로 사전 학습과 특정한 태그를 이용하여 1차적으로 분류하고, 이들에 대해 온톨로지 데이터 사전을 이용해 동음이의어와 이음동의어를 구별하여 다시 카테고리로 분류한다. 그리고 분산된 이질의 데이터들을 대상으로 명칭(named entity)과 값(value)을 추출하여 통합된 XML 메타데이터를 구축하여 이를 사용자에게 제공하는 방법을 제안한다. 또한 제시된 방법의 효율성을 검증하기 위하여 지역정보 중 부동산을 사례로 들어 정보융합 시스템을 구현한다.

본 논문은 다음과 같이 구성된다. 2장에서 메타데이터들의 구성 및 정보융합 방법을 제시한 연구들을 살펴본다. 3장에서는 정보융합을 위한 메타데이터를 구성하는 보다 효과적인 시스템을 제안하

• 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음
(NIPA-2011-C1090-1011-0008)

고 부동산 사례를 들어 시스템을 구현한다. 마지막 4장에서는 결론을 도출하고 향후 연구를 제시한다.

II. 관련 연구

웹 페이지에서 질적으로 향상된 정보를 얻기 위해서는 이질적이지만 서로 연관된 의미를 갖는 데이터들을 하나로 통합하는 방법이 필요하다.

군 정보 통합을 위한 메타데이터 기반의 데이터 그리드 시스템[4]은 이질적인 데이터베이스를 기반으로 통합된 군 정보를 제공하는 시스템을 제안하였다. 정보를 통합하는 과정에서 시스템은 여러 종류의 메타데이터를 공포하고, 저장하고, 접근하기 위한 통일된 서비스를 제공한다. 또한 데이터 그리드를 사용하여 대용량의 분산된 데이터를 통합된 형태로 제공한다. 메타데이터 기반의 데이터 그리드 시스템의 전체적인 시스템 구조를 보면, 분할되어 있는 데이터들이 메타데이터 관리기에 의해서 하나로 통합되고, 통합된 데이터들은 질의 처리기를 거쳐 사용자 인터페이스로 표현된다. 그러나 이러한 방법은 사용자의 의지와는 상관없이 데이터베이스의 구성에 따라 정보가 제공되며, 웹 에이전트의 입장에서 접근했을 때는 단순히 다수의 데이터베이스를 하나의 웹 페이지로 출력하는 정도에 지나지 않는다는 한계를 보인다. 또한 동음이의어나 이음동의어에 대한 처리도 필요하다.

의미 중의성을 고려한 온톨로지 기반 메타데이터의 자동생성[5]은 데이터베이스를 기반으로 하지 않고 온톨로지를 이용하여 메타데이터를 구축하여 데이터들의 의미와 관계들을 시스템이 자동으로 판단할 수 있는 방법을 제안하였다. 메타데이터의 자동생성 시스템 구조는 전처리 단계에서 정보를 추출하고, HMM 학습[6] 결과로 생성된 온톨로지를 이용하여 의미를 결정하는 단계로 수행한다. 그러나 이 연구는 메타데이터를 구성하는 방법보다는 의미 중의성을 판별하는 방법에 초점을 두고 있다.

Corcho[7]는 메타데이터를 구성하는 방법으로 S-OGSA(Semantic Open Grid System Architecture)를 제시하였다. 이 방법은 RDF[8]와 OWL[9]을 기반으로 동음이의어와 이음동의어에 강인한(robust) 메타데이터를 구성하는 방법론으로 자연어, 태그 기반, RDF 기반의 모든 환경에 대하여 각기 다른 단계의 과정을 거쳐 메타데이터를 구성한다. 그러나 이 논문에서는 방법론만을 제시하고 실제 구축 과정이나 실험결과를 제시하지 못하였다. 또한 OWL의 특성상 특정한 카테고리 내에서 연구를 진행하는 것이 일반적임에도 불구하고 어떠한 카테고리도 지정하지 않고 연구가 진행되었다는 문제점도 있다.

본 논문에서는 위와 같은 문제점을 해결할 수 있는 시스템을 제안한다. 본 논문에서 제안한 시스템은 온톨로지와 OWL을 사용하여 기반으로 메타데이터를 구축하고 시스템의 세부 모듈의 흐름에 따라 융합된 데이터를 XML 문서 형태로 사용자에게 제공한다. 또한 제시된 방법론을 검증하기 위해 지역정보 중 부동산을 사례로 정보융합 시스템을 구현한다.

III. 지역정보 융합시스템

본 논문에서 제안하는 지역정보 융합을 위한 시스템은 전처리를 포함하여 크게 4부분으로 구성하였다. 그림 1은 제안된 시스템의 개략적인 구성도이다.

본 논문에서 제안하는 시스템은 온톨로지와 OWL 툴을 사용하여 동음이의어와 이음동의어를 구별할 수 있는 데이터 사전을 구축한다. 온톨로지(Ontology)란 정보 자원을 컴퓨터가 해석할 수 있는 시맨틱(semantic)으로 표현한 특정영역(domain)의 메타데이터(metadata)이다[10].

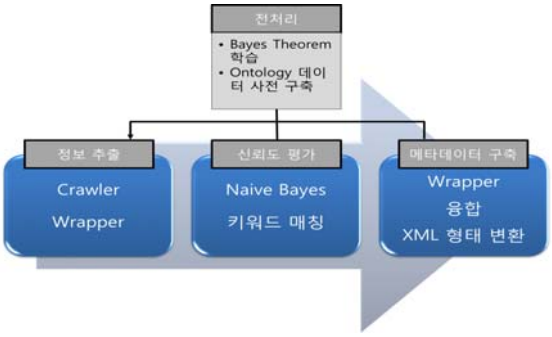

그림 1. 시스템 구성도
Fig. 1. System Architecture

본 논문에서는 앞서 제안한 시스템을 검증하기 위해 지역정보 중 부동산을 사례로 시스템을 구현한다. 그림 1의 시스템 구성도에 따라, 먼저 부동산의 데이터를 담고 있는 웹 페이지와 부동산 정보를 포함하지 않은 웹 페이지 각각 50개씩을 Training Data로 선정하여 전처리 과정과 학습을 진행하였다.

학습 결과로 얻은 특징은 표 1과 같다. 표 1에서 부동산에 관련된 특징이 다음과 같이 분류되었고, 총 출현 횟수, 문서 평균, 가중치를 종합하여 각각의 특징항목들에 대해 순위를 매겼다.

전처리 과정을 마치고, 사용자가 키워드를 입력하면 정보추출 모듈의 전 과정이 시작된다. Crawler[11]는 웹 페이지에 대한 주소를 수집하고, Wrapper[12]는 수집된 주소를 바탕으로 웹 페이지의 정보를 추출한다.

Wrapper로부터 추출된 정보들은 전처리의 학습결과로 얻은 특징을 가지고 있는지를 검사한다. 특징 검사는 그림 2와 같다. 특징을 포함하고 있는 웹 페이지는 신뢰성 판단과 메타데이터 구성의 효율을 위하여 인덱싱 단계를 거친다. 인덱싱을 거치기 전의 데이터는 그림 3과 같고, 인덱싱을 거친 데이터는 그림 4와 같다. 형태소 분석, 불용어 및 중복어 제거, Stemming[13] 등을 거친 문서가 토큰화되었다.

다음 단계로 신뢰도 평가를 수행한다. 신뢰도 평가 모듈은 수정된 웹 페이지 정보를 토대로 각 항목에 따른 카테고리로 분류한다. 항목별 카테고리 분류는 전처리 단계에서 구축된 카테고리 온톨로지 데이터 사전을 토대로 구성된다. 카테고리 분류가 완료된 웹 페이지 정보는 Naive Bayes 방법[14]을 사용하여 신뢰도를 평가한다. 그 결과는 그림 5와 같다. 선별된 문서에 대한 신뢰도는 84.726%였다.

80%이상의 신뢰도를 받은 웹 페이지 정보는 최종적으로 사용자가 입력한 키워드 매칭을 통해 적합한 문서인지 선별한다. 사용자 키워드를 포함하는 웹 페이지 정보는 메타데이터 구성 모듈로 이동하고 사용자 키워드를 포함하지 않는 웹 페이지 정보는 삭제한다. 이 작업은 Crawler에 수집된 모든 웹 페이지 주소를 검사할 때까지 계속해서 반복된다.

표 1. Training Data의 정보
Table 1. Information of the Training Data

```
//지역정보 문서 집합
String local_doc[] = new String[]{
        "융자금","기전세금","건물명","아파트","매매가", // 1번
        "연락처","소재지","주차시설","난방시설","가구", // 2번
        "건물보안","가전","매물설명","입주가능일","부동산", // 3번
        "소재지","위치정보","개인","매물","최초등록", // 4번
        "건물명","매물번호","등록자","매물종류","방향","지역", "공급면적", "기타시설", "층" // 5번
};
```

(a) 부동산 정보를 담고 있는 웹페이지에 대한 특징 추출
(a) Feature extraction for the web pages with real-estimate data

```
// 비-지역정보가 문서 집합
String nonlocal_doc[] = new String[]{
        "광고접수","중국집","법률","세무","커뮤니티", // 1번
        "비밀번호","search","쿨광고","로그인","위원회", // 2번
        "거래","회원정보","안드로이드용","아이폰용","노출", // 3번
        "그린매물","유트브동영상","프레임","활용","탈퇴", // 4번
        "유료","무통장","차익","댓글","민속신앙" // 5번
};
```

(b) 부동산 정보를 담지 않는 웹페이지에 대한 특징 추출
(b) Feature extraction for the web pages without real-estimate data

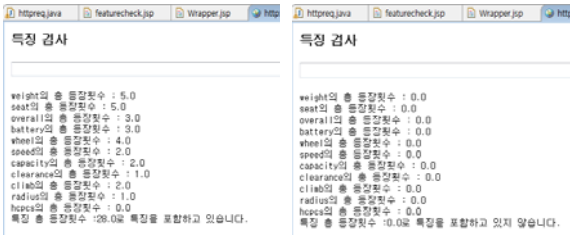

그림 2. 특징 검사
Fig. 2. Feature test

그림 3. 수정 전
Fig. 3. Before the correction

그림 4. 수정 후
Fig. 4. After the correction

그림 5. 신뢰도 평가
Fig. 5. Reliability evaluation

메타데이터를 사용자에게 출력하기 위해 신뢰도 평가 모듈에서 전송된 웹 페이지 정보는 메타데이터 구성과 융합을 위하여 각 카테고리별로 고유의 named entity를 부여한다. 각각 다른 단어로 구성된 항목들을 통합하기 위해 온톨로지 사전에 정의된 항목을 검색하고, 각 항목에 맞는 named entity를 결정하는 과정을 거친다.

다음 단계는 Wrapper를 통한 named entity를 실제 XML 형태로 구성하기 위한 태그를 부여하는 단계를 거친다. 최종적으로 사용자로부터 입력된 키워드의 매칭을 통해 XML 메타데이터를 생성하고 그 결과를 그리드 테이블의 가독성이 뛰어난 형태로 사용자에게 제공한다. 최종적으로 사용자에게 제공되는 결과는 그림 6과 같다.

지역정보 테이블

NO.	신문	매물종류	매매가	위치	매물이름	전화번호	평수	해당층
1	교차로	아파트	17,500만원	천안시	두정마을우남1,2단지	041-523-4489	105.82㎡	15
2	사랑방신문	아파트	12,900만원	북구	금호	062-417-9114	79㎡	1
3	교차로	아파트	20,500만원	동구	한화꿈에그린	010-3028-7224	112.87㎡	10

그림 6. XML 메타데이터와 최종 결과
Fig. 6. XML metada and final result

Ⅳ. 결론

본 논문에서는 웹 페이지로부터 보다 질적으로 향상된 정보를 제공하는 정보융합 시스템을 제안하였다. 제안한 시스템은 전처리, 정보추출, 신뢰도 평가, 메타데이터 구축의 체계적인 순서로 분산된 이질의 데이터들에 대하여 하나의 문서로 통합하여 사용자에게 제공한다. 또한, 제안된 방법론의 효용성을 검증하기 위하여 지역정보 중 부동산을 사례로 하여 융합 시스템을 구현하였다.

본 논문에서 제안한 방법은 첫째, 온톨로지 데이터 사전을 구축하여 각 항목 간의 동음이의어와 이음동의어를 포함한 분류를 수행하고 고유의 named entity를 부여하여 의미 중의성으로 인한 문제를 해결하였다. 둘째, 검색어를 입력하면 사용자에게 링크를 제공하는 대신에 사용자의 키워드에 맞는 통합된 정보를 제공하도록 하였다. 셋째, 사전 학습을 통해 데이터베이스를 구축하고 새로운 웹 페이지의 정보가 추가될 시에 그에 대한 새로운 학습을 수행하여 사용자 참여가 가능한 유연한 데이터베이스를 구축할 수 있다. 이러한 방법은 사용자가 원하는 정보를 얻기 위해 겪는 불편함을 해소하고 그에 따른 시간 낭비를 방지하고, 정확하고 가독성 좋은 검색결과를 제공한다.

향후 연구로는 다른 분야의 웹 페이지 정보들에 대한 메타데이터를 단순 키워드 매칭에서 벗어난, 사진이나 동영상 등과 같은 다양한 멀티미디어 정보들을 융합하여 사용자에게 제공할 수 있는 시스템으로 확장하는 것을 들 수 있다. 또한 모바일 환경에서 GPS 정보와 온톨로지 추론 기능을 활용하여 상황인지 추천 서비스를 들 수 있다.

참고문헌

[1] T. B. Lee, J. Hendler, O. Lassila, “The Semantic Web”, Scientific American, May 2001.

[2] J. C. Song, D. I. Lee, B. J. Moon, “Standardization of Semantic Web and Development Trends of Technical Factors”, National IT Industry Promotion Agency, [IITA] Weekly Technological Trends, No.1064

[3] S. Y. Park, “Comparative Evaluation of Directory Services Provided by Major Korean Search Portals: In the Field of Computer and Internet”, Korean Society for Library and Information Science, Journal of the Korean Library and Information Science Society, Vol.43, No.1, pp.215-234, 2009.3.pp.215-234, 2009.3.

[4] 나민영, “군 정보 통합을 위한 메타데이터 기반의 데이터 그리드 시스템”, 한국컴퓨터정보학회 논문지, 제13권, 제2호, pp.95-103, 2008. 3.

[5] 최정화, 박영택, “의미 중의성을 고려한 온톨로지 기반 메타데이터의 자동생성”, 정보과학회논문지, 제33권, 제11호, pp. 986-998, 2006. 11.

[6] D. R. H. Miller, T. Leek, R. M. Schwartz, “A Hidden Markov Model Information Retrieval System”, pp. 214-221, 1999.

[7] O. Corcho, P. Alper, P. Missier, S. Bechhofer, C. Goble, C., “Grid metadata management: Requirements and architecture”, Grid Computing, 2007 8th IEEE/ACM International Conference on, pp.97-104, 2007. 10.

[8] RDF, http://www.w3.org/RDF

[9] OWL, http://www.w3.org/TR/owl-features

[10] Tom Gruber, “A translation approach to portable ontology specifications”, In: Knowledge Acquisition, 1993, pp.199.

[11] Cho, J., Garcia-Molina, H., “Parallel Crawlers”, Technical Report, Stanford University, 2001.

[12] M. Hearst, “Information Integration”, IEEE Intelligent Systems, vol.13, no.5, pp. 12-24, 1998.

[13] http://tartarus.org/~martin/PorterStemmer/

[14] Pedro Domingos, Michael Pazzani, “Beyound Independence: Conditions for The Optimality of the Simple Bayesian Classifier”, International Conference on Machine Learning, 1996.