택배 마케팅을 위한 온톨로지 기반 잠재고객 탐색 에이전트 설계

이혜진[0], 이금우, 이현아, 김진석
한국전자통신연구원 우정기술연구센터 u-Post연구팀
{lhjin, gmoo, halee, jskim}@etri.re.kr

Design Ontology-based Agent to search hidden customer for Parcel Marketing

Hyejin Lee[0], KeumWoo Lee, Hyunah Lee, Jinseok Kim
Dept. of u-Post Research, Postal Research Center, ETRI

요 약

컴퓨터가 정보를 이해한다' 라는 말로 표현될 수 시멘틱 웹은 WWW의 발전으로 인해 축적된 방대한 데이터 속에서 우리가 원하는 ' 더' 정확한 정보를 찾아 줄 수 있는 대안으로 주목 받고 있다. 이에 대한 연구는 다양한 분야에서의 접근을 하고 있으며 그 ' 개념' 을 점점 실체화시키려고 노력하고 있으나 아직 뚜렷한 모습은 나타내고 있지 않다.

이에 본 연구에서는 시맨틱 웹의 실용화 측면에서 시맨틱 웹의 개념을 이용하여 잠재고객 탐색 에이전트를 설계하였다. 시멘틱 웹 기반의 잠재고객 탐색 에이전트는 인터넷 상의 인터넷 쇼핑몰 및 우체국 택배의 가능한 업체를 선별, 추출하여 잠재 고객을 찾아냄으로써 택배 마케팅을 위한 정보를 제공해 주기 위한 시스템이다.

본 연구에서는 택배 마케팅의 잠재 고객에 대한 정보를 검색하기 위해, 시멘틱 웹 기반의 온톨로지 생성을 위한 구체적인 도메인을 설계하고, 생성된 온톨로지를 이용하는 정보 검색 방법에 대해 소개한다.

1. 서 론

시맨틱 웹은 정보의 의미를 개념으로 정의하고, 개념간의 관계성을 명시화하는 메타데이터의 개념을 통하여, 웹 문서에 시맨틱 정보를 덧붙이고, 이를 이용하는 소프트웨어 에이전트가 이 의미정보를 자동으로 검색하여 새로운 지식을 생성하는 최적의 확장이나 공유가 가능하다[1].

이러한 시맨틱 웹 기반의 서비스는 특정 도메인 지식에 대한 명시적인 명세화 및 지식의 개념과 개념과의 관계를 정형화하는 온톨로지를 통해서 이루어진다. 온톨로지는 간단히 표현하면 단어와 관계들로 구성된 사전으로서 어느 특정 도메인에 관련된 단어들을 계층적 구조로 표현하고, 추가적으로 이를 확장할 수 있는 추론 규칙을 포함한다.

이 연구에서는 우체국 마케팅의 잠재 고객에 대한 정보를 검색하기 위해, 시맨틱 웹 기반의 온톨로지 생성을 위해 구체적인 도메인을 정의하고 생성된 온톨로지를 이용하는 검색 방법에 대해 연구한다.

이를 위해 본 시스템에서는 방대한 웹의 정보로부터 도메인에 해당하는 정보를 추출하기 위한 필터링 기반의 검색엔진에 대한 개발과 단어, URI, 정보의 상하좌우 연관 관계를 부여하기 위한 데이터 마이닝을 이용한 연관 관계 추출 기법에 대한 연구, 그리고 토픽 맵을 이용한 온톨로지 저장에 관하여 다룬다.

이에 대해 본 논문에서는 2장에서 온톨로지 기반의 잠재고객 탐색 에이전트에 대한 전체적인 설명을 하고, 3장에서는 도메인 필터링 기반의 검색 엔진에 대해 기술한다. 4장에서는 데이터 마이닝을 이용한 연관 관계 추출에 대해 설명하고, 5장에서는 온톨로지를 위한 토픽 맵에 대해 기술하고 6장에서는 본 시스템에 대한 결과와 향후 과제에 대해 다룬다.

2. 온톨로지 기반 잠재고객 탐색 에이전트

이 시스템에서는 온라인 쇼핑몰을 포함하여 택배를 이용할 가능성이 있는 회사의 검색을 위하여 Site Topic Map을 구축하고 이것을 기반으로 온톨로지를 생성한다. 이를 위해 먼저 Topic, Accurrence, Assocication에 대한 관계에 대해 다음과 같이 설계하고 도식화하여 표현하였다[2].

우선 토픽 타입으로는 잠재고객(사이트)의 물리적 위치(Location)와 품목(Item), 그리고 회사명 및 연락처와 같은 회사정보(Company Info)로 정의하였다.

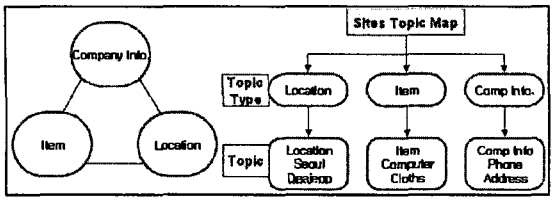

그림 1. 토픽 타입의 관계 도식화

토픽맵을 구성하는 지식과 정보와의 관계에 대해 다음과 같이 토픽, 어커런스, 어소시에이션의 관계를 도식화 하였다.

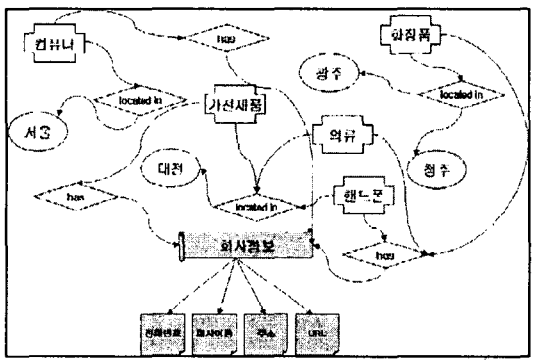

그림 2. 토픽맵을 구성하는 지식과 정보 관계 도식화

온톨로지 기반의 장재고객 탐색 에이전트는 다음과 같은 단계를 거쳐 방대한 웹으로부터 장재고객을 찾아 낸다.

1. 도메인 필터링 기반의 검색 엔진
2. 데이터 마이닝을 이용한 연관 관계 추론
3. 토픽 맵을 이용한 온톨로지 자동 생성

이 시스템의 전체적인 구성도는 다음과 같다.

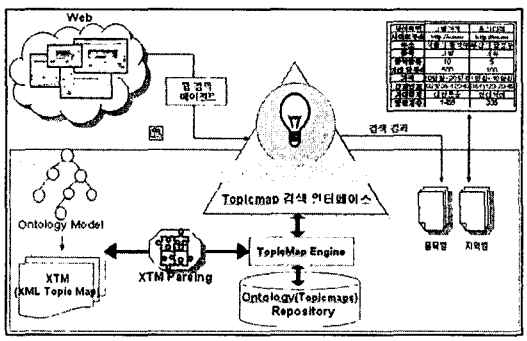

그림 3. 온톨로지 기반 장재고객 탐색 에이전트 구성도

본 논문에서는 크게 3개의 하부 시스템으로 구분하여 온톨로지 기반의 장재고객 탐색 에이전트를 설명한다.

3. 도메인 필터링 기반 검색 엔진

사용자가 요구하는 온톨로지를 생성하기 위하여 이 논문에서는 도메인 필터링 기반 검색 엔진을 이용하여 카탈로그를 구축하는 방법을 사용한다[3].

검색 엔진은 사용자를 대신하여 웹으로부터 데이터를 수집하고 온톨로지 구축에 필요한 데이터를 카탈로그에 저장하기 위하여 웹 공간을 순회하는 로봇 에이전트를 채용한다. 로봇 에이전트의 실행 과정은 다음과 같다.

사용자는 로봇 에이전트를 실행하기 위하여 초기 URI를 입력한다. 로봇 에이전트는 초기 URI로부터 사이트를 검색하여 사용자가 요구하는 도메인에 해당하는 URI 문서를 수집한다[4]. 수집된 문서는 Page, Category, Contents 정보를 추출하고, 각 데이터를 저장한다. 순회 검색을 위하여 로봇 에이전트는 수집된 URI에 링크되어 있는 관련 사이트들을 순차적으로 방문하여 다시 정보를 추출 및 저장하는 과정을 실행한다.

이를 위하여, 검색엔진은 다음과 같은 요소로 구성된다.

- **문서 수집기(Crawler)**
  : 주기적으로 초기 URI를 시작으로 순회 검색을 실행하여 웹 사이트로부터 문서 수집하는 모듈
- **색인기(Indexer)**
  : 수집한 문서들로부터 검색어를 찾아내고 색인 키에 따라 Catalog에 저장하는 모듈
- **카탈로그(Catalog)**
  : 문서 수집기가 수집한 모든 웹 문서의 내용 및 URI 정보를 담고 있는 DB

현재 테스트를 위하여 입력한 초기 URI 수는 28개이며, 이를 이용하여 추출된 Category 표본 추출 쇼핑몰 수는 41개이다. 검색 결과가 저조한 이유는 첫째, URI에 포함된 주소 및 품목명이 이미지로 처리된 경우가 많고, 둘째, 링크된 사이트가 많은 쇼핑몰의 경우 링크 오류가 발생하는 경우가 존재하며, 셋째, 링크된 사이트가 제휴회사인 경우가 많아 순환검색이 발생하기 때문이다. 이를 해결하기 위하여 다양한 URI 링크를 포함하고 있는 적절한 초기 URI를 선택하는 것이 바람직하며, 추후에는 이미지 처리를 통하여 필요한 정보를 얻을 수 있는 방안에 대한 추가적인 연구가 필요하다.

4. 데이터 마이닝을 이용한 연관 관계 추출

이 시스템에서는 검색 엔진에서 검색, 필터링한 내용에 대해 시맨틱을 부여하기 위해 데이터 마이닝을 이용하였다.

데이터 마이닝을 이용하여 단어와 단어 사이에 연관 관계를 추론하여 단어 사이의 상하 좌우 관계를 부여하여 Drill-down과 Roll-up 이 가능하도록 하였다.

이러한 관계성을 위해 본 연구에서는 Item의 1차적인 분류를 위해 계층적 클러스트링 기법인 점진적 개념 클러스터링 알고리즘을 적용하였다[5]. 이러한 분류를 통해 유사 메트리스를 구성하였다. 다음으로 Item 간의 계층적 연관 관계를 찾기 위해 Adaptive-FP 알고리즘을 이용하였다[6].

이 두 단계의 마이닝 단계를 거쳐 단어와 단어간의 계층적 개념 트리를 생성하게 된다.

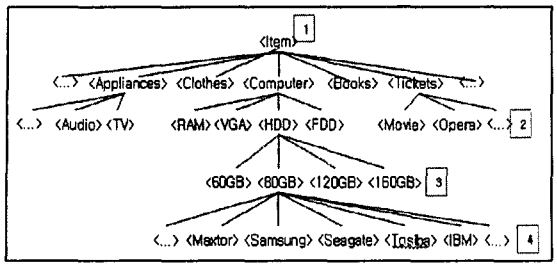

그림 4 Item의 계층적 개념 트리

이러한 계층적 연관 트리는 Topic Map을 이용하여 Location과 Item, 그리고 Company Info에 연관 관계를 부여하여 온톨로지 생성을 한다.

5. 온톨로지를 위한 Topic Map 엔진

본 연구에서는 시맨틱 웹기반에서의 ' 잘 정의된 의미' 를 컴퓨터로 하여금 이해하도록 하기 위하여 온톨로지 기반의 검색 시스템을 구축하고자 한다.
이러한 온톨로지 중 ' 우체국 마케팅' 이라는 도메인을 설정하고 XTM 기반으로 도메인에 적합한 온톨로지를 구축한다[8][9]. 온톨로지 언어로서의 XTM은 RDF를 기반으로한 DAML+OIL, OWL에 비하여, 온톨로지 간의 통합, 복잡한 정보구조를 정의 가능, 지식과 정보의 이중구조 등의 지원으로 향후 온톨로지의 구조적 확장을 진행할 수 있다. 본 연구에서는 다음 절 온톨로지 구축을 위한 절차를 다음과 같이 정의한다.

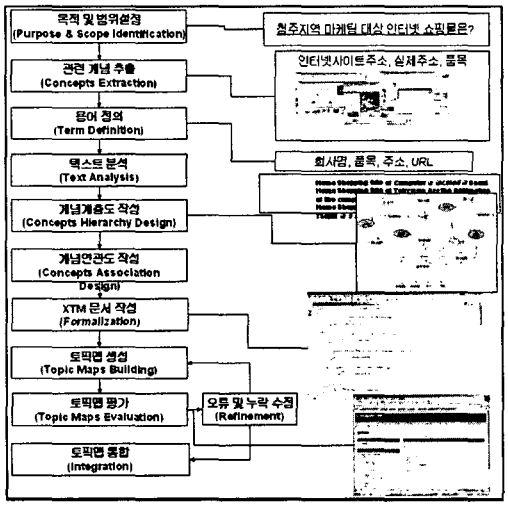

그림. 5 온톨로지 구축 절차

우체국 마케팅을 위한 인터넷 쇼핑몰을 대상으로 온톨로지 모델을 설계하여 이를 XTM 문서화 한다. 도메인에 적합하게 설계된 XTM 문서는 온톨로지 저장소(Ontology Repository) 스키마를 결정하며, TopicMap 엔진은 온톨로지 저장소의 모델을 기반으로 의미있는 검색을 하게 된다.

6. 결론

본 연구에서는 택배의 마케팅을 위한 잠재 고객의 검색을 위한 지식 도메인의 정의 및 지식을 정형화하기 위한 데이터 마이닝 기법을 적용한 온톨로지를 설계하였다.
하지만 현재 몇가지 풀어야 할 문제가 있다. 첫번째 현재 구현된 검색 엔진 모듈은 키워드 기반으로 해당 키워드를 포함하지 않는 잠재 고객 사이트에 대해서는 검색이 안된다. 따라서 이에 대한 문제 해결이 필요하다. 다음으로 도메인에 맞는 계층적 개념 연관 관계 규칙을 찾아내기 위한 알고리즘의 보완과 성능 평가가 필요하다.
현재 이 설계에 의거해 구현중에 있으며 향후 설계된 시스템을 구축하여 탐색 결과와 마케팅 시스템을 연동하여 우체국과 같은 물류/택배 회사를 중심으로 높은 활용이 예상된다.

7. 참고 문헌

[1]. T.Berners-Lee, J. Hendler, O. Lassila, " The Semantic Web" , *Scienctific American* 2001.
[2] International Organization for Standardization, ISO/IEC 13250, Information Technology SGML Applications-Topic Map, ISO, Geneva 2000.
[3] Jialun Qin, Yilu Zhow, Michael Chau, Building Domain-Specific Web Collections for Scientific Digital Libraries: A Meta-Search Enhanced Focused Crawling Method, *Proceedings of 2004 joint ACM/IEEE conference on Digital libraries*, pp. 135-141, 2004
[4] J.L. Wolf, M.S. Squillant, P.S. Yu, J.Sethuraman, L. Ozsen, Optimal Crawling Strategies for Web Search Engines, *Proceedings of the eleventh international conference on World Wide Web*, pp. 136-147, 2002
[5] P. Clerkin, P. Cunningham, C. Hayes, " Ontology discovery for the *semantic web using Hierarchical* Clustering " , Department of Computer Science Trinity Colloege Dublin. 2002
[6] Runying Mao, " ADAPTIVE-FP: AN EFFICIENT AND EFFECTIVE METHOD FOR MULTI-LEVEL MULTI-DIMENSIONAL FREQUENT PATTERN MINING " , B.Sc., Zhejiang Univesity, 1997
[7]. 2003년 3월 정보과학회지 제 21권 제3호, 시맨틱 웹에서의 온톨로지 공학, 카톨릭대학교 양정진
[8] S. Pepper, G. Moore, "XML Topic Maps(XTM) 1.0", TopicMpas.org.
[9] B. L. Grand, M. Soto, " XML Topic Maps and Semantic Web Mining," Laboratoire d'Informatique de Paris, 2001.