온톨로지 기반 지식 검색 시스템 개발: KT 콜센터 사례

안세열*, 최현식**
*고려대학교 컴퓨터정보통신대학원
**고려대학교 대학원 컴퓨터전파통신공학과
e-mail : {geminigem, hyunsikchoi}@korea.ac.kr

Development of an ontology-based knowledge search system: The case of KT call center

Seyeol Ahn*, Hyunsik Choi**
*Graduate School of Computer & Information Technology, Korea University
**Department of Computer Science and Engineering,
College of Information and Communication, Korea University

요 약

콜센터의 고객문의는 복잡하여 기존 검색 시스템으로는 고객의 문제점을 신속하게 찾아 상담에 적용하는데 문제가 많았다. 온톨로지를 구축하고 시맨틱 검색을 제공할 경우 보다 보다 좋은 검색 기능을 제공할 것으로 기대되나 콜센터의 상담지식은 내용이 매우 복잡하여 그 텍스트의 내용을 완벽하게 온톨로지로 표현하는 것은 쉽지 않았다. 본 논문에서는 온톨로지 기반으로 구축된 지식베이스의 데이터 검색과 함께 그와 가장 관련성이 높은 문서를 출력하기 위해 문서를 온톨로지와 링크하여 어노테이션하는 방법을 제안한다. 본 시스템을 적용한 상담에서 상담원들의 생산성이 향상되고 고객 만족도를 높이는 결과를 확인했다.

1. 서론

최근 통신사의 경우 치열한 경쟁으로 출시하는 상품이나 마케팅 이벤트의 주기가 짧아지고 또한 그 종류가 다양해지고 복잡해지고 있다. 그 결과 상담원들이 해결해야 할 고객 문의 내용의 복잡도는 나날이 높아지고 있고, 몇 주간의 교육으로 해결할 수 없는 성격의 문의들이 점점 많아지고 있다.

오래전부터 기업의 콜센터에서는 이를 해결하는 하나의 수단으로 고객응대 지식을 체계화하고 숙련된 상담원의 지식을 공유하기 위한 지식관리시스템을 도입해왔다. 그러나 신속한 업무처리를 필요로 하는 콜센터 상담의 특성상 지식 검색의 결과가 만족스럽지 못할 경우 질의를 수정해가며 여러번 시도해야 한다. 이런 경우 대부분의 상담사는 고객의 콜을 응대하는 동안에는 지식검색을 이용하기 보다는 자신이 알고 있는 분야의 지식을 활용하거나 자신이 해결하지 못할 경우 더욱 숙련된 전문상담사에게 콜을 넘기는 사례가 빈번하게 발생한다. 이렇게 고객의 신속한 문제해결이 요구되는 콜센터에서는 상담원의 구체적인 질의에 대해서 상담원이 원하는 정보를 구체적이고 정확하게 제공하는 지식검색 시스템을 필요로 한다.

본 논문에서는 콜센터의 상담원에게 보다 수준 높은 지식검색 서비스를 제공하기 위해 온톨로지(ontology)에 기반한 지식베이스를 구축하고, 구축한 온톨로지에 기반한 콜센터 지식 검색 시스템을 설계하였다.

2. 시스템 구조

이번 장에서는 본 논문에서 소개하는 온톨로지 기반 콜센터 지식 검색 시스템의 개괄적인 구성과 처리 흐름을 살펴본다.

2.1 문서 저장소 및 온톨로지 저장소

그림 1 은 전체 시스템의 구성 및 처리 흐름을 보여준다. 먼저 기존의 지식관리 시스템을 통해서 축적된 상담지식 문서와 첨부파일, 그리고 그 문서에 대한 메타데이터 정보를 저장하고 있는 문서 저장소가 존재한다.

여기에 저장된 문서는 기존 키워드 검색을 지원하도록 인덱스 빌더를 통해서 역색인 파일을 구축한다. 여기서 키워드 인덱스를 빌드하기 위해서 Lucene[5]을 사용하였고, 키워드 추출을 위한 한글 형태소분석기는 한국의 개발자들이 루씬의 라이브러로 공개한 오픈소스 [6]를 사용하였다. 이때 형태소 분석 과정에서 추출된 어휘들은 온톨로지를 설계하는 지식 엔지니어(Knowledge Engineer)가 온톨로지 어휘를 구축할 때 참고로 사용되기도 한다.

지식베이스에는 문서 저장소에 저장되어 있는 문서 메타데이터, 문서, 첨부파일의 내용을 분석하여 시맨틱 웹 표준 언어인 RDF/OWL 로 구축한

온톨로지를 저장한다. 무료로 제공되는 트리플 저장소 중에서는 BigOWLIM 이 가장 좋은 성능을 보이며[7], 상용제품 중에서는 AllegroGraph[8]와 Oracle 11g 등이 대표적인 제품들이다. 본 연구에서는 안정성, 용량, 그리고 속도 면에서 뛰어난 성능을 보이고 있는 AllegroGraph 를 선택하였다. 온톨로지 구축 방법에 대한 자세한 내용은 3 장에서 설명한다.

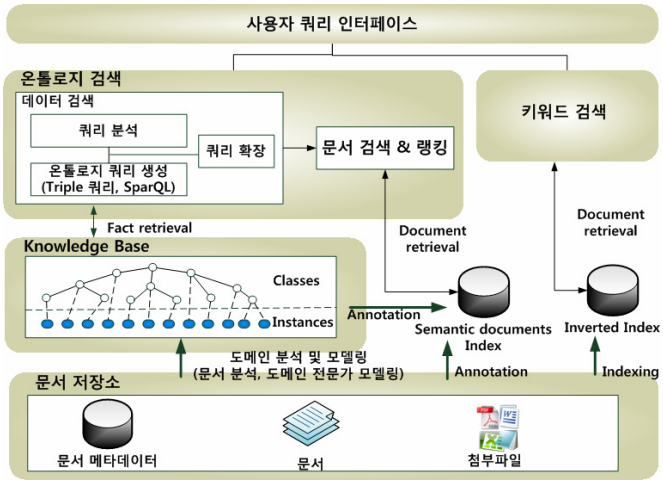

그림 1 시스템 구조

시맨틱 문서 인덱스는 온톨로지의 S-P-O 트리플과 문서의 유사도를 저장하고 있는 인덱스로서 온톨로지에서 검색한 트리플 형태의 정답과 가장 관련이 높은 문서를 사용자에게 보여주기 위해서 필요하다. 시맨틱 인덱스 구축 과정은 문서에서 온톨로지의 인스턴스를 찾아 주석을 부착하는 시맨틱 어노테이션(annotation)과 유사한 과정을 거치는데 그 자세한 설명은 3 장 2 절 에서 설명한다.

2.2 사용자 입력 질의 처리 모듈

사용자는 사용자 쿼리 인터페이스를 통해서 키워드 리스트 또는 시스템이 제안하는 트리플 형태의 자연어 질의를 할 수 있다. 시스템이 제안하는 트리플 형태의 질의는 서버의 자동완성 모듈에 의해서 제공된다. 이 자동완성 단계에서 “스마트폰”이라고 입력한 경우 그 하위 인스턴스인 “아이폰”을 함께 자동완성 제시어로 제안하도록 온톨로지 기반의 자동완성을 구현하였다. 단, 온톨로지의 그래프를 실시간으로 탐색하면서 질의를 완성하려면 지연이 발생할 수 있기 때문에, 온톨로지 기반으로 미리 구축된 자동완성 사전을 사용하였다. 사용자가 입력한 쿼리는 온톨로지 검색 모듈과 키워드 검색 모듈에 동시에 전달되어 병행 처리한다. 온톨로지에 저장된 클래스/인스턴스의 커버리지가 낮을 경우, 즉, 문서에 출현하는 어휘를 많이 포함하지 않을 경우, 온톨로지 검색 결과가 없거나 Recall 이 낮을 가능성이 높기 때문에 키워드 검색 모듈을 동시에 지원하도록 하였다.

온톨로지의 검색은 크게 두 가지의 단계로 나누어 볼 수 있다. 첫 번째 단계는 데이터 검색 (fact retrieval) 단계로서 사용자가 입력하는 쿼리를 분석하여 그 질의에 대한 답을 트리플 형태의 데이터로 추출하는 단계이다. 쿼리 분석 모듈에서는 가장 먼저 사용자의 쿼리가 온톨로지의 S-P-O 트리플 쿼리 형태로 매핑이 되는지를 판단하고, 트리플 쿼리로 정확히 매핑이 되지 않는 경우 형태소 분석 과정을 거쳐서 키워드 셋을 추출한다. 이 키워드 셋은 복수개의 가능성 있는 트리플 쿼리로 변환되어 트리플 저장소를 검색한다. 트리플 저장소에서는 AllegoGraph 에서 지원하는 RDFS++ 수준의 추론을 거쳐 트리플 형태의 정답 집합을 추출하여 사용자 질의에 대한 구체적인 답으로 제시한다. 두 번째 단계는 온톨로지에서 추출한 트리플 형태의 지식과 가장 관련이 높은 문서를 찾아서 랭킹하여 보여주는 (documents retrieval)단계이다. 이를 위해서 트리플 아이디와 문서 간에 유사도 점수를 저장하고 있는 시맨틱 문서 인덱스를 검색하여 가장 점수가 높은 순서로 문서를 출력한다.

3. 지식베이스 구축

이 장에서는 콜센터 상담지식 도메인의 지식을 온톨로지로 모델링하는 과정과 이 온톨로지를 기반으로 기존에 있던 상담지식문서를 어노테이션하여 온톨로지와 문서 간의 링크를 구축하는 방법을 설명한다.

3.1 온톨로지 모델링

온톨로지 구축 방법[1]으로는 도메인의 전문가가 수동으로 구축하는 방법(hand-crafted)과 기계학습, 자연어처리 기술분야의 알고리즘을 이용하여 자동으로 구축하는 방법이 있다. 본 연구에서는 검증을 위해 아주 작은 규모의 품질이 높은 온톨로지를 구축하는 것을 목표로 커버리지가 낮더라도 완전 수동으로 구축하였다. 온톨로지 구축의 기초 자료로는 기존 상담지식 관리시스템에서 정의하고 있는 지식분류체계와 검색로그, 그리고 5,700 건의 상담지식 문서를 활용하였다. 검색로그에서 추출한 상위 50 개 키워드에 대해서 집중적으로 middle-out 방법론[9]으로 온톨로지를 확장하였으며, 기존의 상담지식 시스템에 축적되어 있는 약 5,700 건의 상담지식 문서를 Knowledge 엔지니어가 분석하여 수동으로 구축한 후에, 도메인 전문가(콜센터 상담원 교육강사)와 수 차례 인터뷰를 통하여 수정하였다. On-to-Knowledge[2]의 온톨로지 구축 방법론에 의하면 수동으로 구축하는 경우에 온톨로지 정제(refinement) 단계에서는 도메인 전문가가 직접 온톨로지 모델링 툴을 사용할 수 있도록 교육시켜 온톨로지 엔지니어의 도움으로 도메인 전문가가 직접 온톨로지를 구축하는 것이 가장 이상적이지만, 실제 기업환경에서는 이러한 프로토타입 시스템의 검증을 위해서 그 도메인 전문가가 온톨로지 모델링을 하도록 시간과 리소스를 투자하는 것은 어려운 실정이고 가장 아쉬운 부분이다.

그림 2 는 구축된 온톨로지의 상위 레벨 스키마를 보여주고 있다. 상위 스키마에서는 상품 및 서비스, 고객, 이벤트, 요금제 등의 상위 개념과 그 개념간의 관계를 표현하고 있다. 온톨로지는 총 202 개의 클래스와 1139 개의 인스턴스로 구성된다.

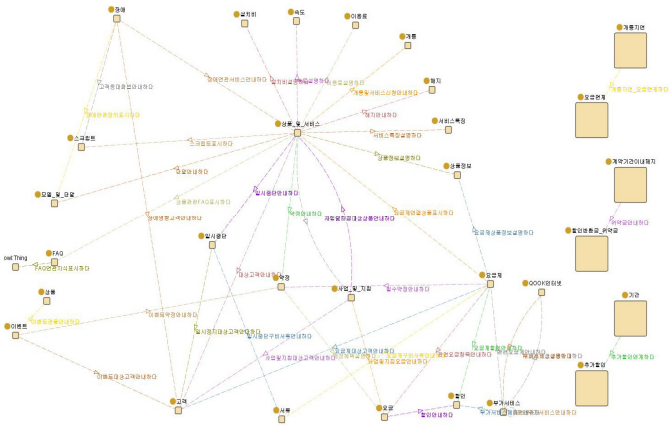

그림 2 온톨로지 스키마

3.2 시맨틱 어노테이션

여기서는 온톨로지를 구축하고 나서 온톨로지를 활용한 문서검색이 가능하도록 온톨로지와 문서 간의 링크를 만드는 방법을 소개한다.

시맨틱 어노테이션이란 원본 데이터에 추가적으로 주석을 달아서 그 데이터의 의미를 사람과 컴퓨터가 이해할 수 있도록 하는 작업이다[3]. 시맨틱 어노테이션을 자동으로 하는 기술에 대한 연구는 자연어처리, 기계학습 분야에서 많이 진행되어 왔으며 시맨틱 웹의 비전을 실현하기 위한 핵심 기술 중의 하나라고 볼 수 있다.

[3]에서 제시하는 방법은 HTML 문서에서 출현하는 인스턴스의 개체명(Entity type), 즉 온톨로지의 클래스를 자동으로 인식하여 HTML 문서 내의 인스턴스를 온톨로지에서 정의한 클래스의 인스턴스로 새롭게 추가하여 population 하거나, 존재하는 인스턴스의 URI 와 매핑하는 것이다. [10]에서는 이 온톨로지 인스턴스와 문서의 연결 정보를 별도의 RDB 에 Entity-EntityOccurence-Document 필드로 저장하였는데, 이 때 EntityOccurence 는 인스턴스가 문서에 출현한 횟수를 의미하며 문서를 랭킹하는 기준이 되었다. 그러나 이 연구에서는 IDF 와 같은 글로벌 가중치는 고려하지 않아서 문서 랭킹에 대한 사용자 만족도는 낮을 수 있다.

완전 자동화된 시맨틱 어노테이션을 목표로 하는 [3]에서는 기계학습, 자연어처리 기술을 사용하여 자동으로 문서 내의 개체명을 인식하고 온톨로지에 매핑하지만 직접 사람이 부착한 어노테이션에 비해서 정확성을 보장하기 어렵다. 그렇다고 사람이 매번 문서의 인덱싱 단계에서 개입하여 반자동으로 시맨틱 어노테이션 작업을 하기에는 많은 비용이 소모된다. 본 시스템에서는[12, 13, 14]과 같은 방향으로 접근하여 문서의 텀벡터(Term Vector) 차원을 온톨로지의 인스턴스만 포함하도록 축소시킨 후, 전통적인 IR 모델에서 널리 사용하는 TF/IDF 기반의 가중치를 계산하는 간단한 방법을 사용하였다.

최근 시맨틱 웹 등장 이후 시맨틱 웹의 온톨로지를 이용한 오픈 도메인 Q&A 분야에서는 다양한 접근이 시도되고 있지만[11] 대부분의 연구는 시맨틱웹 환경에서의 온톨로지에서 데이터를 검색(data retrieval)하는 연구에 초점을 맞추고 있고, 온톨로지 기반의 문서 랭킹에 대한 연구는 매우 적다고 조사 된바 있는데[4], 본 시스템의 방법에 의하면 간단하게 온톨로지 트리플에 대한 문서의 랭킹이 가능하다.

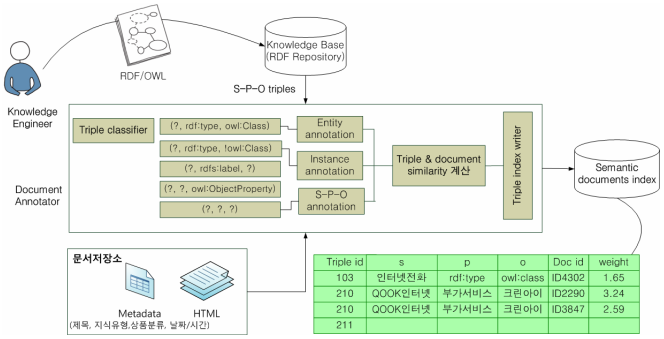

그림 3 시맨틱 문서 인덱스 구축 과정

그림 3 은 본 시스템에서 시맨틱 문서 인덱스를 구축하는 방법을 보여주고 있다. Document Annotation 모듈은 Knowledge 엔지니어가 작성한 RDF/OWL 온톨로지를 저장하고 있는 트리플 저장소의 모든 트리플을 조회하면서 각 트리플을 클래스, 인스턴스, S-P-O 세 가지 유형으로 분류한다. 트리플이 “Qook 인터넷-rdf:type-owl:Class” 또는 “크린아이-rdf:type-부가서비스” 처럼 클래스나 인스턴스의 정의를 나타내는 트리플의 경우 Subject-Predicate-Object 중에서 Subject 의 rdfs:label 을 조회하면서 각 label 이 문서에서 가지는 가중치 점수를 계산하여 그 중 최대값을 시맨틱 문서 인덱스에 저장한다. 트리플이 “Qook 인터넷-has 부가서비스-크린아이”와 같이 개체 간의 관계를 표현한 트리플의 경우에는 Subject 와 Object 의 rdfs:label 을 조회하면서 각 label 이 문서에서 가중치 점수를 계산하고, Subject 레이블 가중치의 최대값과 Object 레이블 가중치의 최대값의 평균값을 최종 가중치 점수로 인덱스에 저장한다. 여기서 Property 를 문서 가중치 점수 계산에 제외 시킨 이유는, Property 의 경우 실제 문서에 출현하지 않는 키워드를 새롭게 정의하여 (e.g. 묘사하다, 안내하다) 모델링한 경우가 많아서 Property 를 포함하면 오히려 문서 검색의 성능을 저하시키는 현상을 발견하였기 때문에 제외시켰다. 그림 4 는 위에서 설명한 시맨틱 문서 인덱스 구축 모듈의 의사코드(pseudo code)를 보여준다.

```
SemanticDocumentIndex
Input: Triple Store TS = {t1, t2,…, tn}, Document Set DS = {d1, d2, …, dm}
Output: Semantic document Index = {{tid, di, score}, …, {tid, di, score}}
For every triple ti(S,P,O) in TS
   If P=" rdf:type" && O!=" owl:ObjectProperty"
   /* S is a class or instance definition*/
```

```
    For every dj in DS
        For every rdf:label of S
          Score = Max(Similarity(S, dj))
        Store annotation in Semantic document Index
  Else if          /* S-P-O */
    For every dj in DS
        For every rdf:label of S
          Scores = Max(Similarity(S, dj))
        For every rdf:label of O
          Scoreo = Max(Similarity(O, dj))
        Score = Avg{ Scores, Scoreo}
        Store annotation in Semantic document Index
```

그림 4 시맨틱 문서 인덱스 pseudocode

4. 구현 및 동작 예제

그림 5 상담 지식 검색 시스템 UI

그림 5 는 본 논문에서 제안하는 시스템을 구현한 예제 화면이다. 사용자가 "쿡인터넷"을 입력하였을 때 왼쪽에는 온톨로지를 조회한 쿼리 확장의 결과와 가운데에 랭킹된 문서를 출력한 결과를 보여주고 있다.

사용자가 "쿡인터넷 부가서비스"로 입력을 한 경우의 예를 들어 동작 과정을 설명하면 아래와 같다. 가장 먼저 쿼리분석 모듈에 의해서 "쿡인터넷 부가서비스"는 Class + Property 패턴의 질의임을 인식하고 쿼리 패턴 별로 정의한 쿼리확장 규칙에 기반하여 다음의 트리플 쿼리 셋을 생성한다.

(S, "부가서비스", ?)
(S, "rdf:type", "QOOK 인터넷")

이 때 온톨로지 검색 결과는 다음과 같다.

("QOOK 인터넷라이트", "부가서비스", "크린아이")
("QOOK 인터넷라이트", "부가서비스", "아이디스크")
("QOOK 인터넷스페셜", "부가서비스", "크린아이")
("QOOK 인터넷스페셜", "부가서비스", "인터넷닥터")

이러한 사용자의 질의에 대한 구체적인 답이 될 수 있는 정답 집합은 검색 창 아래에 Answer 로 출력하고, 이 트리플 셋의 ID 로 시맨틱 문서 인덱스를 검색하여 이 부가서비스를 설명하는 가장 관련성이 높은 문서를 순서대로 출력하였다.

이렇게 한 개의 키워드에 대해서는 확장을 하고, 두 개 이상의 키워드 입력에 대해서는 온톨로지에서 답을 찾아 낼 가능성이 높기 때문에 결과로 추출한 트리플 집합을 결과 값으로 사용자에게 반환한다.

이 외에도 "인터넷 요금", "쿡인터넷 속도", "QOOK 인터넷 가입시 설치비", "QOOKTV 양방향 서비스". "QOOKTV 쇼핑 결재수단", "DMB 지원하는 스마트폰" 등 온톨로지에서 정답을 찾을 수 있는 쿼리에 대해서는 바로 답을 얻어내고, 더 복잡한 지식은 관련 문서로 해결할 수 있어서 기존 키워드 기반 문서 검색 시스템 보다 높은만족도를 보였다.

5. 결론 및 향후 연구 방향

본 논문에서는 콜센터의 상담 지식 도메인에 대해서 온톨로지 기반의 지식 검색 시스템을 도입하여 서비스 제공이 보다 신속하고 정확한 검색을 제공함을 보였다. 향후에는 온톨로지의 커버리지 확대, 쿼리 분석 모듈 성능 개선, 자연어 처리기술을 이용한 문서 어노테이션 성능 개선, 온톨로지 시각화, 시스템 안정화 등의 작업을 진행하고 보다 정확한 성능의 검증을 위해 콜센터 상담원들을 대상으로 필드 테스트를 진행하고자 한다.

참고문헌

[1] Grigoris Antoniou, A Semantic Web Primer, The MIT Press, 2004.

[2] Y. Sure, On-To-Knowledge: Semantic Web Enabled Knowledge Management. In Web Intelligence. Springer-Verlag, 2003.

[3] Atanas Kiryakov, Semantic Annotation, Indexing, and Retrieval, Journal of Web Semantics, 2004.

[4] C. Mangold, A survey and classification of semantic search approaches, Int. J. Metadata, Semantics and Ontolgy, Vol.2, No.1, 2007.

[5] http://lucene.apache.org

[6] http://sourceforge.net/projects/lucenekorean/

[7] K. Rohloff, An Evaluation of Triple-Store Technologies for Large Data Stores, OTM Workshops 2, Vol. 4806, Springer, 2007.

[8] http://www.franz.com/agraph/allegrograph/

[9] Y. Sure, Ontology Engineering Methodology, Handbook on Ontologies, Springer-Verlag, 2004.

[10] Borislav Popov, Co-occurrence and Ranking of Entities, http://www.ontotext.com/publications/CORE_otwp.pdf, 2006.

[11]V. Lopez,, Merging and Ranking answers in the Semantic Web: The Wisdom of Crowds, The 4th Asian Semantic Web Conference, 2009.

[12] D. Bonino, Ontology Driven Semantic Search, WSEAS Transaction on Information Science and Application 1(6), 2004.

[13] P. Castells, An adaptation of the vector-Space Model for ontology-based information retrieval, IEEE Transactions on Knowledge and Data Engineering 19(2), 2007.

[14] M. Fernandez, Semantic Search meets the Web, IEEE Semantic Computing, 2008.