사물 인터넷 환경에 적합한 커뮤니티 질의 응답 시스템 개발

김강섭*, 이호준*
*영동대학교 스마트IT학과
e-mail:hjlee@webmail.yd.ac.kr

Development of Community Question Answering System suitable for Internet of Things Environment

Gang-Sup Kim*, Ho-Joon Lee*
*Dept of Smart IT, Youngdong University

요 약

사물 인터넷(Internet of Things)의 확산으로 가까운 미래에는 사물 인터넷 환경에서 질의 응답 시스템이 활발하게 이용될 것으로 예상된다. 본 논문에서는 사물 인터넷 환경에 적합한 초소형, 저사양 하드웨어를 이용하여 커뮤니티 질의 응답 시스템(Community Question Answering System)을 구축하는 방안에 대해 살펴본다. 하드웨어는 700Mhz 싱글 코어 CPU와 512MB의 메인 메모리를 장착한 라즈베리 파이를 이용하였고, 질의 응답 시스템으로는 Apache Solr를 기본 시스템으로 활용하였다. 성능 분석 결과 실시간 응답성은 매우 훌륭하지만 정확도는 앞으로 보완이 필요한 것으로 분석되었다.

1. 서론

질의 응답 시스템(Question Answering System) 개발은 인공지능 중에서도 정보 검색이나 자연언어처리와 연관이 많은 분야로, 사용자의 질문에 대한 응답을 구축된 지식 베이스에서 자동으로 탐색하여 제시하는 시스템을 의미한다[1]. 일반적으로 질의 응답 시스템에서는 구문 분석, 의미 분석 등의 자연언어처리 과정을 통해 질문의 유형과 내용 등을 파악하고, 추론 과정을 통해 지식 베이스에서 질문의 내용에 가장 적합한 답을 찾아 이를 응답문으로 생성하게 된다. 이러한 질의 응답 시스템 중에는 IBM의 Watson[2]과 같이 대용량의 지식 베이스와 고성능의 하드웨어를 이용하여 다양한 형태의 질의에 대해 높은 정확도와 빠른 속도로 응답문을 생성하는 경우도 있지만, 애플의 Siri[3]와 같이 제한된 지식 베이스와 하드웨어 환경에서 제한된 형태의 질의에 대해 응답문을 생성하는 경우도 있다.

범용적인 질의 응답 시스템의 개발을 위해서는 다양한 형태의 정보를 담고 있는 지식 베이스의 구축이 요구되나, 특정 도메인에 한정된 질의 응답 시스템의 경우에는 온라인 게시판 등에 등록되어 있는 “질문-응답” 쌍을 이용하여 지식 베이스를 구축하는 것이 더욱 효율적이다[4, 5, 6]. 게시판 등에 사전에 작성된 “질문-응답” 쌍을 지식 베이스로 활용하는 질의 응답 시스템을 커뮤니티 질의 응답 시스템(Community Question Answering System, cQA)이라고 하는데, “질문-응답” 쌍의 경우 StackOverflow[7]와 같이 특정 분야에 특화된 경우도 존재하는 반면, 네이버 지식인 서비스[8]와 같이 다양한 분야의 내용이 존재하는 경우도 있다.

최근 사물 인터넷(Internet of Things)의 확산과 함께 사물 인터넷 기반 질의 응답 서비스에 대한 수요도 증가하고 있는데, 사물 인터넷 환경에서는 사물의 용도와 관련하여 도메인에 특화된 질의가 주된 입력이 될 것으로 예상된다. 예를 들어, 사물 인터넷 환경에서 체온계를 통한 질의-응답의 범위는 주로 이상 체온에 관한 것으로 예상되는데, 이 경우 범용적인 질의 응답 시스템보다는 커뮤니티 질의 응답 시스템이 더욱 효과적으로 질의 응답 서비스를 제공할 수 있다.

본 논문에서는 사물 인터넷 환경에 적합한 커뮤니티 질의 응답 시스템의 개발에 대해 논의하며, 실험을 위한 지식 베이스는 국립국어원 질의/응답 게시판의 “질문-응답” 쌍을 활용한다. 국립국어원 질의/응답 게시판의 “질문-응답” 쌍은 전문적인 내용을 다루면서도 충분한 크기의 데이터 확보가 가능하고, 오탈자나 문법적 오류가 적기 때문에 초기 시스템 개발의 실험 데이터로 매우 적합하다고 볼 수 있다.

2. 관련연구

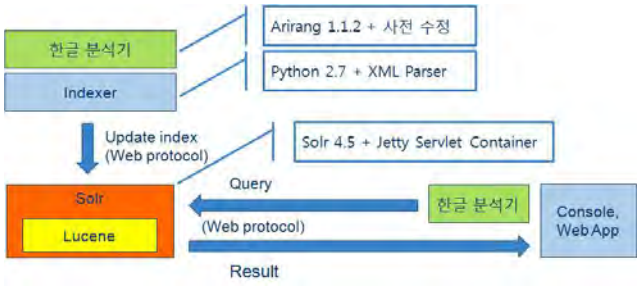
커뮤니티 질의 응답 시스템에 관한 기존 연구로는 질문과 응답의 유형을 나이브 베이즈 분류기를 이용하여 적절한 응답을 검색하는 연구[4]나 키워드의 의미를 확장하여 전문 영역을 검색하는 연구[5], 질문과 응답의 구조적 특징을 이용하여 분류 과정을 통해 적절한 응답을 제

시하는 연구[6] 등이 진행되었다. 이러한 대부분의 연구는 기존의 질의 응답 환경과 유사한 실행 환경을 설정하고 있는데, 본 연구에서는 사물 인터넷 환경에 적합한 저전력, 경량화 시스템에서 커뮤니티 질의 응답 시스템을 구현하는 방법에 대해 논의한다.

3. 사물 인터넷 환경에 적합한 커뮤니티 질의 응답 시스템 개발

사물 인터넷(IoT, Internet of Things) 환경에서는 기본적으로 저사양, 초소형 하드웨어를 기반으로 데이터 처리가 이루어지만 서비스에 대한 응답은 실시간에 가깝게 수행될 수 있어야 한다. 이러한 사물 인터넷 환경에 맞추어 본 논문에서는 라즈베리 파이 모델 B[9]를 커뮤니티 질의 응답 시스템의 하드웨어로 사용하였는데, 라즈베리 파이 모델 B는 ARM1176JZF-S 700Mhz 싱글 코어의 CPU와 512MB의 메인 메모리가 탑재되어 있다.

커뮤니티 질의 응답 시스템의 개발을 위해서 Apache Solr 4.5 버전[10]을 기본 시스템으로 사용하고 한글 처리를 위해서 Arirang 한글 분석기[11]를 적용하였다. 그림 1은 전반적인 시스템의 구조를 보이고 있는데 전체 개발은 Java와 Python으로 구현되었고, Solr의 서블릿 컨테이너로는 빠른 응답속도를 위해 Jetty를 사용하였다.



(그림 1) 커뮤니티 질의 응답 시스템 구조도

입력된 질문에 가장 적절한 응답을 제시하기 위해서는 지식 베이스의 "질문-응답" 쌍에서 입력으로 들어온 질문과 가장 유사한 "질문"을 찾고, 그에 해당하는 "응답"을 출력하면 되는데, 커뮤니티 질의 응답의 경우에는 사용자의 질문 내용이 "질문-응답" 쌍의 "응답" 부분에서만 나타날 수도 있기 때문에 "질문-응답" 쌍 모두를 입력된 질문과 비교하여 가장 유사한 "응답"을 답으로 제시한다. 유사도 비교를 위해서는 TF-IDF 방식으로 문서를 벡터화하고, 코사인 유사도로 벡터간의 유사도를 계산한다. TF(Term Frequency)는 특정 단어가 문서 내에서 얼마나 자주 등장하는지를 나타내는 값이고, IDF(Inverse Document Frequency)는 해당 단어의 일반적인 중요도를 나타내는 수치이다. 따라서 문서 d_j와 질문 q 유사도는

$$sim(d_j,q) = \frac{d_j \cdot q}{\|d_j\|\|q\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$ 이고,

$w_{t,d} = tf_{t,d} \cdot \log\frac{|D|}{|\{d' \in D \mid t \in d'\}|}$ 이며 $tf_{t,d}$는 문서 d에서 단어 t의 빈도를 의미하고 $|D|$는 문서 집합내 전체 문서의 개수, $|\{d' \in D \mid t \in d'\}|$는 단어 t를 포함하는 문서의 개수를 의미한다. TF-IDF로 문서를 벡터화 할 때, 단어는 Arirang 한글 분석기를 이용한 형태소 분석 결과를 사용하였다.

본 연구를 통해 개발된 커뮤니티 질의 응답 시스템은 리눅스 환경에서 Java JRE와 python이 설치되어 있으면 바로 실행 가능하다. 커뮤니티 질의 응답 시스템은 1) 서버 구동, 2) '질의-응답' 지식 베이스 입력, 3) 질의 입력의 과정을 거치게 되는데 '질의-응답' 지식 베이스는 다음 그림 2와 같은 형태로 입력된다.

```
<?xml version="1.0" encoding="UTF-8"?>
<add> <doc>
<field name="id">1</field>
<field name="text_hangul">이곳은 참 살 만하다.에서 살을 분석하면, 살다의 어간 살에서, 받침 ㄹ이 탈락하고, 관형사형 어미 ㄹ이 붙은 형태가 맞는지 궁금합니다. 그리고 본용언과 보조용언이 이어질 때 어미가 없어도 되는지 궁금합니다.</field>
<field name="text_hangul">생각하신 대로, 어간 끝 받침 ㄹ이 어미의 첫소리 ㄴ, ㅂ, ㅅ, 및 -오, -ㄹ 앞에서 줄어지는 경우, 준 대로 적습니다. 이에 따라, 살 만하다의 살은 살다의 어간 살- 뒤에 어미 -ㄹ이 붙은 활용형이라고 판단할 수 있습니다. 용언은 어간 뒤에 어미가 붙는 형태로 쓰입니다. 보기를 들어, 본용언과 보조 용언 짜임인 사 주었다, 가 버렸다에서 본용언인 사, 가도 사아, 가아의 준말인데, 이와 같이 용언이 준말 형태로 쓰이기는 해도, 어간만 쓰이지는 않습니다.</field>
</doc> </add>
```

그림(2) '질의-응답' 지식 베이스 예시

질의는 그림 3과 같은 형태로 입력되며, 해당 질의에 대해서 n-best 응답이 결과로 출력된다. 예를 들어 그림 3의 질의를 5,336개의 지식 베이스에서 10-best로 검색하여 출력하면 그림 4와 같다. 검색 결과는 유사도 순으로 정렬되어 출력되는데, 왼쪽은 해당 지식 베이스의 ID를 나타내고, 오른쪽은 유사도를 나타낸다.

```
<query><qnum>1</qnum>
<text>
10%감소했다를 읽을 때, 퍼센트라고 읽든 프로라고 읽든 상관없나요?
</text>
</query>
```

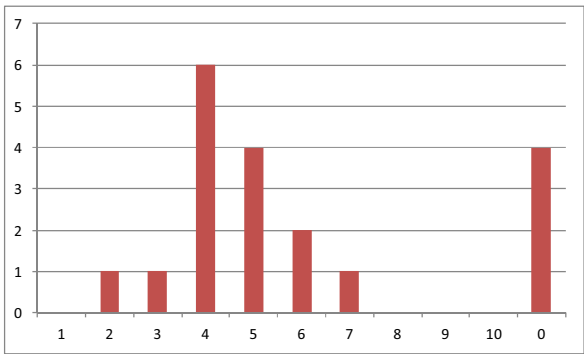
그림(3) 질의 예시

<query><qnum>1</qnum>
<rank>
561 2.732729
471 2.415414
4844 2.391138
785 2.09181
3055 1.9323312
617 1.9323312
729 1.9323312
1278 1.7933533
3608 1.6907897
5200 1.6203089
</rank></query>

그림(4) 10-best 결과 예시

실험을 위해서 국립국어원 질의/응답 게시판에서 뽑아낸 5,336개의 "질문-응답" 쌍을 지식 베이스로 활용했고, 20개의 정답 셋을 이용하여 성능을 평가하였다. 성능 평가에는 정확도뿐만 아니라, 지식 베이스 입력 시간 및 질의 응답 시간도 포함하여 본 연구의 결과물이 사물 인터넷 환경에서 실시간으로 응답을 할 수 있는가도 측정하였다.

총 20개의 질의에 대해서 10-best의 결과가 정답 셋에서 몇 번째에 나타나는지 정리해보면 그림 5와 같다. 즉 정답이 10개의 후보 중에서 나타나지 않은 경우는 총 4건 있었고, 1건은 2-best에서 정답이 검색되었으며, 1건은 3-best에서, 6건은 4-best에서 정답이 검색되었음을 확인할 수 있었다.



그림(5) 10-best 결과 분석

지식 베이스 입력 시간은 총 5,336개를 색인할 때 17분 14.995초가 걸렸고, 20개의 질의를 처리할 때에는 1.836초가 걸려서 1개의 질의를 처리할 때 92ms 정도의 시간이 필요한 것으로 평가되었다.

4. 결론

사물 인터넷(Internet of Things)의 확산으로 가까운 미래에는 사물 인터넷 환경에서 질의 응답 시스템이 활발하게 이용될 것으로 예상된다. 본 논문에서는 사물 인터넷 환경에 적합한 초소형, 저사양 하드웨어에 기반하여 실시간으로 커뮤니티 질의 응답 시스템을 구축하는 방안에 대해 살펴보았다. 성능 확인을 위해 국립국어원 질의/응답 게시판의 내용을 바탕으로 정확도와 실시간 응답성을 테스트해본 결과 실시간 응답성은 매우 훌륭하지만 정확도는 기대했던 것보다 성능이 떨어지는 것을 확인할 수 있었다. 추후 효율적인 검색 기법과 색인 방법, 한국어 문장 처리 방법 등의 연구를 통해 시스템의 성능을 향상시키고자 한다.

사사의 글

이 논문은 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 이공분야 기초연구사업의 연구 결과임 (NRF-2012R1A1A1013389)

참고문헌

[1] http://en.wikipedia.org/wiki/Question_answering
[2] Ferrucci, David, et al. "Building Watson: An overview of the DeepQA project." AI magazine 31.3 (2010): 59-79.
[3] http://www.apple.com/ios/siri
[4] 연종흠, 심준호, 이상구, "확장된 나이브 베이즈 분류기를 활용한 질문-답변 커뮤니티의 질문 분류", 정보과학회논문지: 컴퓨팅의 실제 및 레터 제16권 제1호, pp.95-99, 2010.
[5] 정옥란, 오제환, 이은석, "Q&A 커뮤니티 기반 전문 영역 검색을 위한 프레임워크", 한국전자거래학회지 제16권 제2호, pp.143-158, 2011.
[6] 배경만, 고영중, 김종훈, "커뮤니티 기반의 질의 응답서비스(cQA)에서 질문-응답 쌍의 구조적 특징을 이용한 언어 모델 기반의 주제 분류 기법", 정보과 학회논문지: 소프트웨어 및 응용 제39권 제8호, pp.664-671, 2012.
[7] http://www.stackoverflow.com
[8] http://kin.naver.com
[9] http://ko.wikipedia.org/wiki/라즈베리_파이
[10] http://lucene.apache.org/solr
[11] http://sourceforge.net/projects/lucenekorean